UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

MACHINE LEARNING ALGORITHMS FOR WEALTH DATA ANALYTICS

by

Ngoc Yen Nhi Vo

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2020

Certificate of Original Authorship

I, Ngoc Yen Nhi Vo, declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and IT at the University of Technology Sydney, Australia, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
prior to publication.

Ngoc Yen Nhi Vo

Sydney, Australia, 2020.

Acknowledgements

First of all, I express my sincere gratitude for my supervisor, Professor Guandong Xu, who had been a great mentor guiding me through my Ph.D. candidature stages. He had empowered and inspired me to work on multiple research projects which led to publications in top journals and conferences. I am grateful for the support of my co-supervisor, Shaowu Liu, who had been helpful with various technical difficulties. I am also greatly appreciate the contributions of all co-authours to my research publications that partially constructed this thesis: Professor Guandong Xu, Shaowu Liu Ph.D., Professor Xuezhong He, Professor Xitong Li, James Brownlow, Charles Chu, and Ben Culbert.

I thank Advanced Analytics Institute, School of Computer Science, Faculty of Engineering and IT, University of Technology Sydney for providing the infrastructures and computing power used in conducting empirical works. I am deeply grateful for the financial support received during my candidature time: UTS International Research Scholarship, FEIT Industry Scholarship, and Vice Chancellor Conference Fund scholarship. Part of my research was funded by Australian Research Council Linkage Project Scheme under LP170100891 and LP140100937 grants.

Finally, I am grateful for my dear family, especially my father Ngoc Bien Vo, my mother Thi Bao Di Phan, and my brother Ngoc Bao Vo, who had been supportive during my Ph.D. candidature stages, researching and writing my thesis. I am sincerely grateful for everything and especially the opportunities given in this life, from being born to the completion of this Doctor of Philosophy thesis.

List of Publications

Journal Papers

J-1. Vo, Nhi NY, Xuezhong He, Shaowu Liu, and Guandong Xu. "Deep Learning for Decision Making and the Optimization of Socially Responsible Investments and Portfolio." Decision Support Systems (2019): 113097.

J-2. Vo, Nhi NY, Xitong Li, Shaowu Liu, and Guandong Xu. "Leveraging Unstructured Call Log Data for Customer Churn Prediction." Knowledge-Based Systems (Under review).

Conference Papers

C-1. Vo, Nhi NY, Shaowu Liu, Xuezhong He, and Guandong Xu. "Multimodal mixture density boosting network for personality mining." In Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 644-655. Springer, Cham, 2018. (PAKDD 2018 conference took place in Melbourne, Australia from 3^{rd} to 6^{th} June, 2018)

C-2. Vo, Nhi NY, Shaowu Liu, James Brownlow, Charles Chu, Ben Culbert, and Guandong Xu. "Client Churn Prediction with Call Log Analysis." In Proceedings of International Conference on Database Systems for Advanced Applications, pp. 752-763. Springer, Cham, 2018. (DASFAA 2018 conference took place in Queensland, Australia from 21^{st} to 24^{th} May, 2018)

C-3. Vo, Nhi NY, and Guandong Xu. "The volatility of Bitcoin returns and its correlation to financial markets." In Proceedings of International Conference on Behavioral, Economic, Socio-cultural Computing (BESC), pp. 1-6. IEEE, 2017. (BESC 2017 conference took place in Cracow, Poland from 16^{th} to 18^{th} October, 2017)

Contents

List of Figures

List of Tables

ABSTRACT

The thesis investigates multiple machine learning algorithms with big data approach and applies cutting-edge deep analytics to tackle the challenges in financial wealth management. In general, the existing research on wealth data analytics is limited with two main challenges. Firstly, the amount of quantitative research conducted is scarce and scattered across different approaches. Partially this is due to the lack of access to the data required for the research to use a quantitative approach. Secondly, the results are rudimentary and limited to a certain aspect of wealth data analytics. This lack of integration in existing research findings is a by-product of the simplistic approaches employed in lieu of big data analytics and deep learning techniques.

This research provides a broader and comprehensive approach for quantitative research within the wealth management field from both financial and customer aspects. Particularly, this research utilizes the big data of structured demographic, behavioral, communicational data, and unstructured textual information from wealth customers, plus additional financial market and corporate responsibility data from companies. This thesis exploits deep analytics techniques to provide a better framework for decision-making support based on the constructed mathematical and computational models, combined with customer segmentation modeling and quantitative finance approach.

From the customer aspect, the thesis applies big data analytics, text mining and interpretable machine learning in customer data analytics in wealth management. The proposed approaches and models are (1) MMDB for personality mining, (2) transfer learning for customer personality prediction, (3) ensemble model with text mining for churn prediction, (4) interpretable machine learning with SHAP-MRMR+ to extract customer insight, and (5) customer segmentation and managerial implications with personality and SOM.

From the financial aspect, this is one of the first research to utilize deep learning

for socially responsible investment. The proposed framework consists of (1) text mining of CSR reports for ESG ratings, (2) ESG-based quantitative models, (3) deep learning using Multivariate BiLSTM for stock return prediction, (4) MV-ESG for ESG-based portfolio optimization, and (5) reinforcement learning for socially responsible investment.

The empirical results show the advantages and effectiveness of deep learning algorithms and big data analytics in financial wealth data analytics. Through the completion of this thesis, various aspects of wealth data analytics have been researched and integrated into sophisticated frameworks, and the information systems can provide meaningful insights for multiple stakeholders, from researchers to individual investors and fund managers.

Abbreviation

2-D	Two dimensional
AGR	Agreeableness
ARMA	AutoRegressive Moving Average
AUC	Area Under the Curve
BG	Growth of Book Value per Share
Big Five	Big Five Personalities
BiLSTM	Bidirectional Long Short Term Memory network
Borderline-SMOTE	SMOTE and Borderline samples
CL	Call Logs data set
CON	Conscientiousness
CP	Customer Profiles
CRM	Customer Relationship Management
CSR	Corporate Social Responsibility
CSR-Sent	Corporate Social Responsibility Sentiment Dictionary
DCA	Discriminant Correlation Analysis
Doc2Vec	Document to Vector
DRIP	Deep Responsible Investment Portfolio
DT	Decision Tree
DY	Growth of Dividend Yield
EBITDA	Earning Before Interest, Tax, Depreciation and Amortization
EM	Expectation Maximization
ESG	Environmental, Social and Governance
EXT	Extraversion
FI	First Impression Dataset
GARCH	Generalized AutoRegressive Conditional Heteroskedasticity
GHG	Green House Gas

GMO	Genetically Modified Organism
GP	Gaussian Process
GRI	Global Reporting Initiative
GRU	Gated Recurrent Unit network
ID	Identification
IT	Information Technology
K	thousand (quantity)
LAD	Least Absolute Deviation
LI	Lexical Information
LIWC	Linguistic Inquiry and Word Count
LR	Logistic Regression
LSTM	Long Short Term Memory network
MA	Mean Accuracy
MAE	Mean Absolute Error
MAX-ESG	Maximum ESG portfolio
MAX-S	Maximum Sharpe portfolio
MAZ	Mean Absolute Z-Score
MDN	Mixture Density Network
MMD	Multimodal Mixture Density Network
MMDB	Multimodal Mixture Density BoostingNetwork
mRMR	Minimum Redundancy Maximum Relevance
MSE	Mean Squared Error
MV	Mean Variance model
MV-ESG	Mean Variance ESG model
MV-SGP	Mean Variance Stochasitc Goal Programming model
NB	Naïve Bayes
NEU	Neuroticism
NLP	Natural Language Processing
NN	Neural Network
NSGA-II	Non-Dominated Sorting Genetic Algorithm 2
OCSVM	One Class Support Vector Machine

OPN	Openness
P/B	Price per Book
P/ESG	Price per ESG
PE	Phrase Embedding
PT	Personality Traits
RF	Random Forest
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SDGs	United Nation Sustainable Development Goals
SHAP	Shapley Addictive Explanations
SHAP-MRMR+	Combined SHAP and positive mRMR method
SLSQP	Sequential Least Square Programming
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Over-sampling Technique
SMOTETomek	SMOTE and Tomek links
SMOTTEENN	SMOTE and Edited Nearest Neighbours
SOM	Self-Organizing Maps
SRI	Socially Responsible Investment
SSS	Small Sample Size
Super	Superannuation
SVD	Single Value Decomposition
SVM	Support Vector Machine
SVM-SMOTE	SMOTE and SVM
SVR	Support Vector Regression
TF-IDF	Term Frequency - Inverse Document Frequency
TI	Term Importance
Word2Vec	Word to Vector
XGB	XgBoost Extreme Gradient Boosting method
YT	Youtube Dataset

Chapter 1

Introduction

1.1 Background and Motivation

Wealth management is an extensive researched field that attract interest from both academia and industry due to its direct socioeconomic benefits. Research in this field can potentially benefit the academia and the society in multiple folds with diverse stakeholders, ranging from financial institutions to individual consumers. Organizations and researchers in wealth management field are particularly interested in the financial and social benefits from both investment and customer portfolios. Multiple qualitative and quantitative methods have been proposed to solves challenges in wealth management research.

In the past decades, data analytics has been playing a vital role in wealth management research, especially in two aspects: customer relationship management (CRM) and financial investment management. However, many research efforts are initiated from business and finance disciplines only, which methods are mainly simple regression models and incapable of handling complex and multimodal big data in this modern age. Research in wealth management field has to expanded and shift its focus to keep up with the rapid development of technology and big data. Particularly, more research and development efforts should be spent on the top levels of data analytics, namely the algorithmic, predictive and descriptive capabilities as illustrated in Figure 1.1 below.

With the recent development in data mining and machine learning, information system research has been leveraging these advanced approaches to further improve algorithmic analytics in both customer relationship and financial management aspects within wealth management industry. This is the inspiration for conducting a multi-discipline research in machine learning algorithms in wealth data analytics

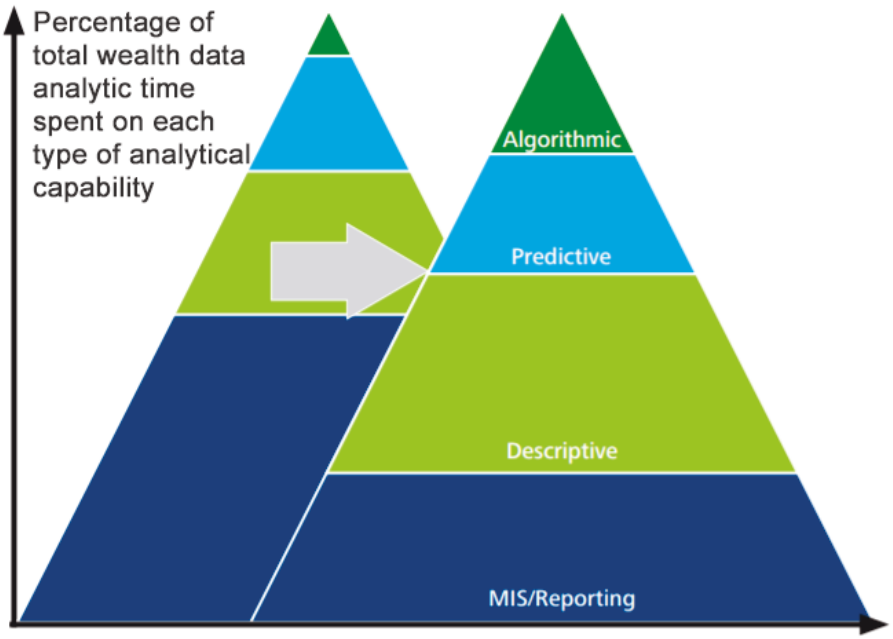


Figure 1.1 : Shifting Mix of Wealth Data Analytics

field, incorporating the domain knowledge and technical methods from both machine learning and finance disciplines.

The lack of such integrated research for wealth data analytics is the motivation for this Doctor of Philosophy thesis titled "Machine Learning Algorithms for Wealth Data Analytics". The thesis aims to use big data, particularly financial and customer data, and apply the cutting-edge deep learning analytics to tackle the research challenges in wealth management.

In general, the existing research on wealth data analytics is limited by two main challenges. Firstly, the amount of quantitative research conducted is scarce and scattered across different approaches. Partially this is due to the lack of access to the data required for the research to use a quantitative approach. Due to its high competitive nature, data in wealth management is a valuable resource which is not shared across organizations and industries. Secondly, the results are rudimentary and limited to a certain aspect of wealth data analytics. This lack of integration in existing research findings is a by-product of the simplistic approaches employed in lieu of big data analytics and deep learning techniques.

Based on the mentioned challenges in wealth data analytics, the thesis seeks to answer these core research questions:

i. Can we leverage advanced data mining techniques in order to incorporate more under-utilized data like text and other unstructured data in wealth management field?

ii. Can we integrate novel machine learning and deep learning algorithms to design better data analytics frameworks in wealth management field?

iii. Can we utilized these developed information systems in real-life business applications for the benefits of our research stakeholders?

From these research questions, the thesis seeks answers by studying all related research, developing machine learning methods and proposing quantitative models that can benefits the main stakeholders.

1.2 Research Objectives

The objectives of this thesis are to:

i. expand current literature on the methodology of big data framework in wealth data, including data preparation, cleansing, analysis and integration,

ii. conduct studies of machines learning algorithms in wealth data analytics, particularly focusing on financial services customer data and socially responsible investment data analytics,

iii. design analytics frameworks for both financial services customer and socially responsible investment data,

iv. develop and apply related machine learning algorithms for both financial services customer and socially responsible investment data, and

v. test the proposed data mining approaches and trained models on real-life datasets to infer meaningful insights and managerial implications in financial wealth industry.

1.3 Research Highlights

The expected contributions of the thesis are:

i. It contributes directly to current research literature on both theoretical and applied machine learning algorithms on wealth data analytics.

ii. The information system framework, algorithms, models developed and proposed in this thesis have a strong applicability in financial services field and can also be generalized to other industries.

iii. The research benefits multiple stakeholders: (1) academic and industry researchers from both machine learning and finance disciplines; (2) financial services firms and investment funds; and (3) individual, especially financial customers and socially responsible investors.

1.4 Thesis Structure

This thesis is organized as follows:

i. *Chapter 1*: The first chapter explores the background on wealth data analytics with a focus on two main aspects: wealth customer data analytics and socially responsible investment analytics. From a set of defined research objectives, the main research contributions are highlighted based on the methods developed in this thesis.

ii. *Chapter 2*: The second chapter presents the core theoretical foundation, including the basic background of information retrieval and data mining, machine learning and algorithms, and wealth data analytics. This introductory information will serve as the academic foundation for all methods and approaches proposed in this thesis.

iii. *Chapter 3*: This chapter consists of the initial research on customer data analytics side, utilizing personality mining for understanding financial customer behavior. A novel Multimodal Mixture Density Boosting Network (MMDB)

is developed, which can accurately predict the personality of a person using three different types of data: video, audio and text features. The personality mining model can be used in transfer learning to predict the personality traits of customers in Chapter 3.

iv. *Chapter 4*: Based on the learning from Chapter 2, the personality traits of financial services customers are derived in this chapter. We further leverage various unstructured data mining approaches and multiple machine learning algorithms to forecast the churn risks. Using an advanced interpretable machine learning approach with the proposed SHAP-MRMR+ and the SOM customer segmentation method, we can extract meaningful customer insights and marketing strategies.

v. *Chapter 5*: In this chapter, the thesis includes preliminary research on socially responsible investment. A CSR-Sent text mining model is proposed to predict the companies' ESG performance based on their corporate responsibility reports. ESG has also been tested as an indicator for the financial performance of the firms, which is a strong motivation for applying more quantitative finance methods and machine learning approaches to socially responsible investment in Chapter 5.

vi. *Chapter 6*: From the findings in Chapter 4, a novel framework for socially responsible investment is proposed in chapter 5. The Deep Responsible Investment Portfolio (DRIP) model combines quantitative finance model, deep learning and reinforcement learning, which has been fully tested and presented in this chapter.

vii. *Chapter 7*: In this concluding chapter, a brief summary of the thesis contents are given, highlighting its contributions for both theoretical and applied machine learning research, particularly in wealth data analytics.

The thesis ends with Bibliography section the together with recommendation and preliminary for future works is given in the Appendices.

Chapter 2

Theoretical Foundation

The purpose of this chapter is to provide an introductory background on the core theoretical foundation for this thesis.

2.1 Information Retrieval and Data Mining

In daily life, people tend to get information to support their decision making by asking a question or sending an inquiry. Information retrieval is a concept in the information science discipline, which refers to the process of finding and getting the requested information. In recent years, the availability of information and data has increased significantly and become overload, while people need only relevant info depending on their needs. Systems have to be developed to return only the documents, data or info that meets the inquirers' need. These are often called information systems,

With the advancement of computer science, information systems have been researched and engineered to be able to get information much faster and more accurate. Some of the most popular and advanced information-retrieval systems are search engines on the web. For example, Google crawls all the websites on the Internet, categorizes and caches them in a organized structure so that they can be served to users very precisely and quickly. After retrieving the information from the systems, the next step is to mine the data for the purpose of knowledge discovery and decision making.

In formal definition, data mining is the knowledge discovery process using a huge amount of data with the methods and techniques of computer science, statistics and artificial intelligence. Data mining is also known as Knowledge Discovery from Data, either in a structured format, e.g. data mining on multiple customer databases, or

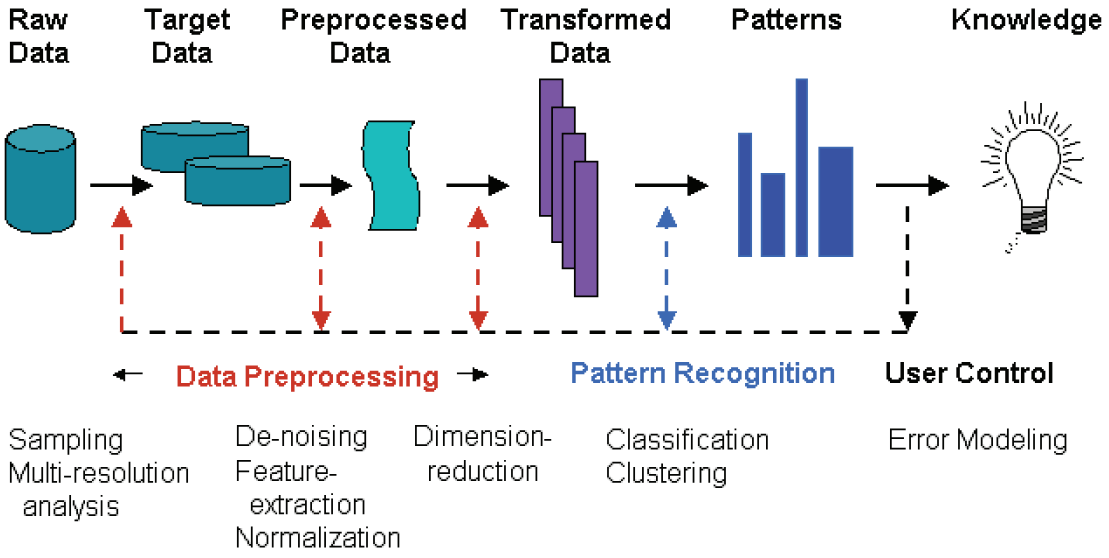


Figure 2.1 : The Data Mining Process

unstructured format, e,g, text mining on natural language documents. A challenge for many researchers is the availability of big data, particularly to develop methods to derive hidden information from these gigantic volumes of data. The whole data mining process is an iterative sequence of different data transformation and model building steps as illustrated in Figure 2.1.

In this thesis, we apply all advanced methods in information retrieval to gather both structured and unstructured data from public and private sources. We then follow the data mining process to cleanse and transform the data, e.g. anonymizing, linking, aggregating, normalizing, etc. Afterwards, the necessary features are extracted and engineered to be used in our learning models to find the hidden information, predict the patterns, discover knowledge and support the decision making of related stakeholders.

2.2 Machine Learning and Algorithms

Data Analytics and Data Science are general terms and can often be used interchangeably. Moreover, their meaning has changed with the emergence of big data challenges and complex modeling needs. Previously, research in data science was

often associated to the application of business intelligence, where knowledge were discovered from the data by the machine computation to create managerial dashboards and reports for humans. In recent years, the development of more heterogeneous datasets, which became far more complex for traditional regression models, required machine learning techniques to be applied to derive useful information.

Machine learning is an approach to data science that involves teaching computer, building and adapting models that can be used to extract meaningful insights from the data. In most scenarios, machine learning techniques involve the construction of algorithms that adapt their models to improve their ability to make predictions.

Machine learning algorithms can be divided into three main categories: supervised, unsupervised and emerging learning, which includes recent novel approaches such as deep learning, reinforcement learning, etc. (see Figure 2.2).

- Regarding supervised learning, the model is trained on known input data to predict an output using the new data in which the value is not previously known. An example of supervised learning in this thesis is to predicted the long term stock returns with the model trained on historical financial market prices. Supervised learning is the most commonly used and well developed branch of machine learning due to its high applicability in real life scenarios.

- Concerning unsupervised learning, the algorithm can discover patterns for a better understanding and representing of the data. An example of unsupervised learning we applied is the detection of customer segments that have a similar personalities and churn risks in financial services contact. Unsupervised learning algorithms harder to evaluate because the ground truth of the output is unknown. Therefore, they are often assessed based on its ability to better describe the data, determine the relationships between observations and improve the results from other techniques.

- Regarding emerging learning, we mainly use deep learning and reinforcement learning approaches. Deep learning utilizes neural networks to train the model based on large datasets. In this thesis, we use deep learning to train a long-term

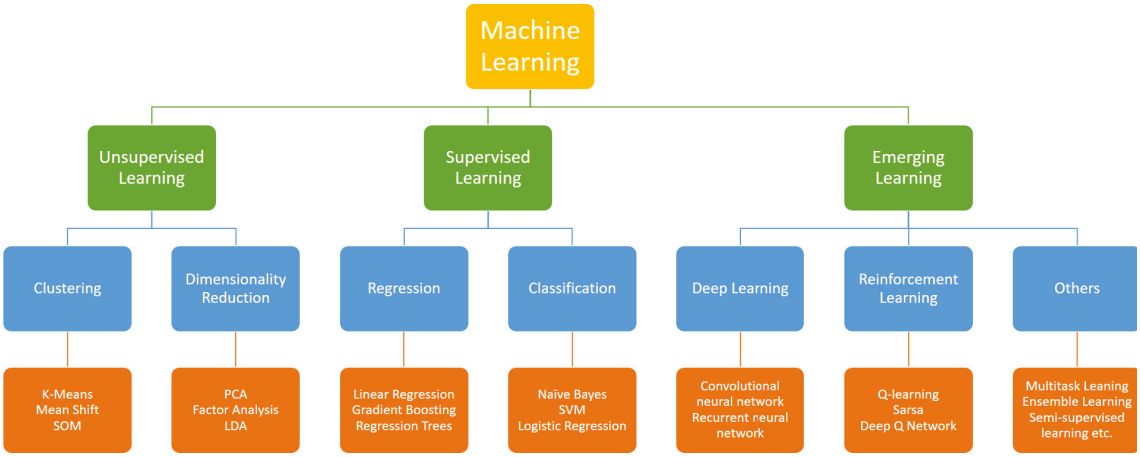


Figure 2.2 : Sample Machine Learning Algorithms

stock returns prediction model. On the other hand, reinforcement learning allows an agent to learn to perform some actions based on feedback from the environment without pre-collected datasets. An example of reinforcement learning in this thesis is a program learning to trade stocks and optimize a financial portfolio by observing the received benefits of the previous periods. Reinforcement learning built with neural networks is the latest advancement in machine learning, which attracts lots of research attention due to its potential of creating a general artificial intelligence.

In recent years, several other subcategories of emerging learning have been researched, including semi-supervised learning, ensemble learning, instance-based learning, multi-task learning, explainable machine learning and other advanced deep learning techniques. These approaches and algorithms are evolving everyday and multiple applications have been proposed. Within the scope of this thesis, not all the machine learning algorithms can be studied. However, the author believes further exploration of these advanced methods might be of an interest for future research in related topics.

There are multiple different programming languages, frameworks and software applications existed for data mining process. In this thesis, Python was determined as the main programming language to collect, explore, clean, analyze and model

the data for its popularity and extent of available libraries and machine learning algorithms. Additionally, Matlab, R and Excel were utilized for some subtasks in the data mining process.

2.3 Wealth Data Analytics

According to Goel (2009), "Wealth management is a holistic approach to understanding and providing solutions to all of the major financial challenges of an investor's financial life. From a client's perspective, this means having all financial challenges solved. From a wealth manager's perspective, it means the ability to profitably provide a wide range of products and services in a consultative way". On the other hand, Bradstreet (2009) has stated that "Wealth management is a new, discrete discipline and not just a variation on the traditional institutional investment management theme and can also be defined as an all-inclusive service to optimize, protect and manage the financial goal of an individual, household, or corporate". In this thesis, we are focusing more in individual and corporate stakeholders. However, the benefits of individual can be generalized to household stakeholders as well.

Data Analytics and Data Science are used interchangeably when it comes to the application in Wealth Management field. Wealth Data Analytics is the process that utilize qualitative and quantitative methods to analyze the data to extract managerial implications and information for better decision making. There are three main categories of data analytics in general: Descriptive Analytics, Predictive Analytics, and Prescriptive Analytics (see Table 2.1). In this thesis, we will combine methods from all three types of data analytics to apply to Wealth Management data, with a little more focus on the later two types (namely Predictive Analytics and Prescriptive Analytics).

Although the tradition data analytics models in wealth management have been successfully applied to real business scenarios for years, they has become insufficient in the current digital age, particularly in handling bigger data and more complex information systems. Together with the advancement of technology, data mining and machine learning have been incorporated and integrated into the Wealth Data

Table 2.1 : Types of Data Analytics

	Descriptive	Predictive	Prescriptive
Perspective	Retrospective	Prospective	Prospective
Objectives	- Identify past patterns - Summarize data, relationships and metrics - Describe what happened - Setup use case and thresholds - Monitor and alert	- Predict the probability of future events - Provide insights by identifying important predictors - Determine if current trends will continue and anticipate possible outcomes - Build prediction and classification models	- Identify possible paths that could lead to the optimal solution - Utilize modelling and experiments to determine possible outcomes based on varying conditions - Provide actionable information to aid in decision making process
Techniques	- Statistical Analysis - Data visualization	- Data Mining - Prediction model	- Simulations - Optimization

Analytics process in multiple aspects (see Figure 2.3). Within the scope of this thesis, we will focus on the first two aspects of wealth data analytics: customer relationship management and financial investment management.

On customer relationship management (CRM), the ability to retrieve useful information from data is essential for any wealth management firms. Demographic, socioeconomic or geographic characteristics of the customers are traditionally and widely used in machine learning models. However, this approach does not consider the customer behavior and their personalities. Customers nowadays expect to receive customized products and services based on their provided personal data. Advanced text mining and personality mining algorithms provides the ability to transform CRM in wealth management.

Moreover, interpretable machine learning techniques can explain the hidden and unknown customer insights, thus, achieve effective CRM. The data analytics with machine learning process can be integrated throughout the CRM information systems, from forecasting customer behavior to proposing marketing strategies for retention. The research work on customer management aspect is presented in Chapter 3 and Chapter 4 of this thesis accordingly.

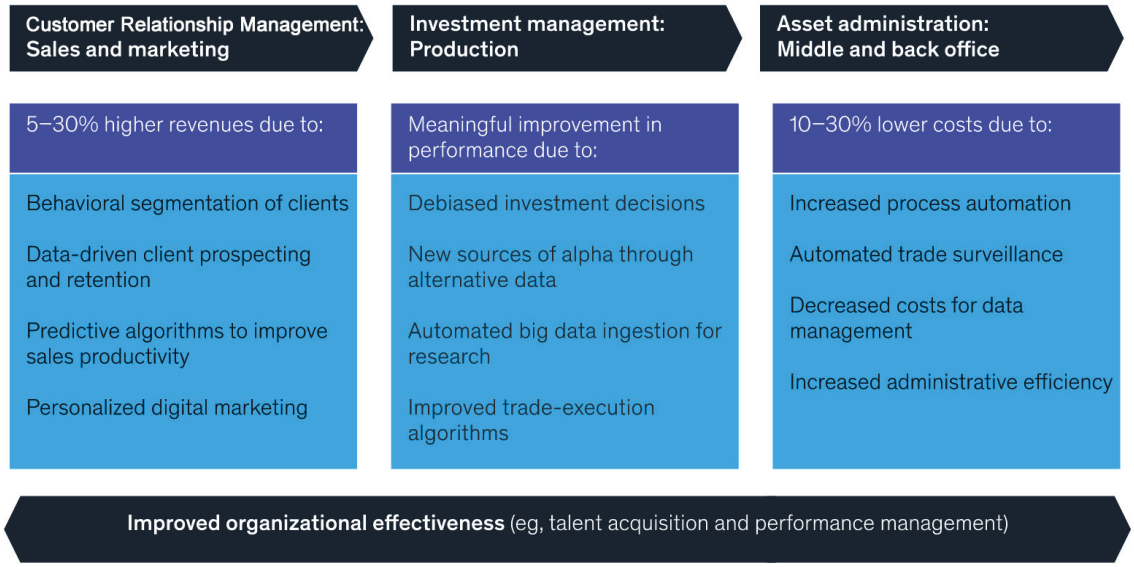


Figure 2.3 : Wealth Data Analytics

Research on investment management has focused primarily on stock trading and financial portfolio construction. The main investment management problem dealt with in the wealth data analytics literature is portfolio optimization, where a machine learning algorithm trades in several assets within an adversarial and stochastic market with the goal of maximizing returns and minimizing risks. The success of the trained model depends on future information, and it is almost impossible to compete against the unknown market conditions.

Furthermore, investors in recent years have other social goals besides the financial benefits. The focus of this thesis is on devising portfolio optimization algorithms with robust performance guarantees, which also incorporates other characteristics of socially responsible investment for better investor engagement. The research work on financial investment management aspect is presented in Chapter 5 and Chapter 6 of this thesis respectively.

Chapter 3

Multimodal Mixture Density Boosting Network for Personality Mining

3.1 Background and Motivation

Personalities denote the individual variances in characteristics patterns of thinking, feeling and behaving. People with different personalities tend to conduct themselves in varied ways and have different cognitive processes. Knowing one's traits and understanding the differences in their preferences would help with communicating and connecting to the person on a more individual level. One of the most well-known measurements of personality traits is the Five-Factor Model of Personality (Big Five) (Goldberg 1990). As showed in Figure 3.1, the Big Five model contains the five fundamental underlying personality dimensions: agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience.

These personality dimensions are stable across time, cross-culturally shared, and explain a substantial proportion of behavior (Costa and McCrae 1992). Therefore, the Big Five model has been the standard measurement for personality mining in current literature. Personality Mining is the process of identifying a person's traits by mining the information in different types of individual data. The main techniques to identify the Big Five personalities of an individual have been the qualitative methods of surveys (Goldberg et al. 2006). Recently, there have been some applications of machine learning in personality mining, mostly through text mining (Kosinski et al. 2015) using standard algorithms, e.g. support vector regression or decision tree (Farnadi et al. 2016).

Though current methodologies have showed the feasibility of personality min-

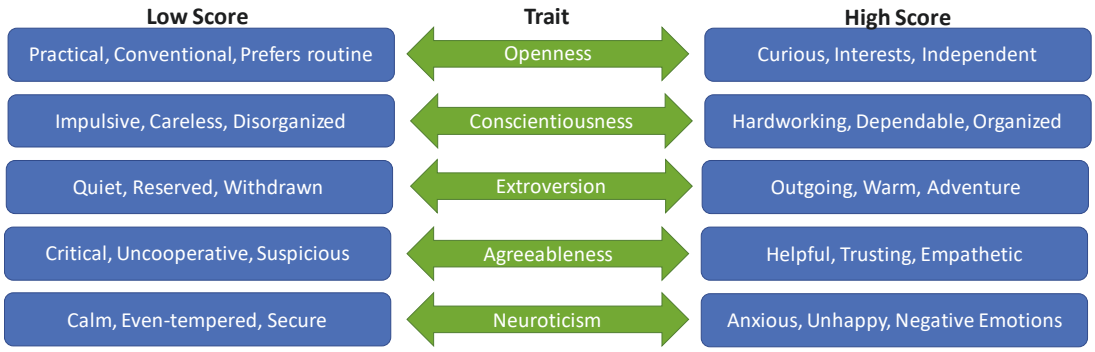


Figure 3.1 : The Five-Factor Model of Personality

ing, they have some critical limitations preventing for broader adoption. To be specific, the qualitative approaches are not practical, time-consuming, costly and might contain subjective errors. On the other hand, the standard machine learning algorithms can quickly mine the personalities of a large number of people at once without conducting surveys (Farnadi et al. 2013). However, the prediction accuracy of these quantitative techniques suffers from small data size. Considering ethical reasons, using up-sampling techniques to increase the number of observations will not be acceptable.

In addition, most researchers have been approaching the personality mining problem using textual data only (Golbeck et al. 2011; High 2012). However, human characteristics are explicitly expressing not only in the spoken words but also in their facial expression and the way they speak as well. More research has incorporated these sensory information into their predictive models. Some research have showed that sensory data would significantly improve the prediction accuracy of one's traits (Alam and Riccardi 2014). These motivates us to look for a deep learning method which can learn from multimodal data.

Realizing these research gaps, we would like to propose here one of the very first multimodal approaches in personality mining using information from videos, audio and text data. Our Multimodal Mixture Density Boosting Network (MMDB) combines advanced deep learning techniques to build a multi-layer neural network.

from small size of personality datasets. We will have an initial feature fusion layer to avoid over-weighting of one type of input data. Afterwards, we construct a combined neural network consisting of mixture density layers to avoid over-fitting and dynamic cascade gradient boosting layers to improve our prediction accuracy. In addition, our MMDB neural network has a general structure which can be applied flexibly to other similar multimodal deep learning problems.

There would be three main contributions of our research in personality mining.

- The proposed approach leverages the deep neural network to analyze multimodal data for personality prediction model.
- Our MMDB model was built to adapt both small and large dataset, which is extremely useful in psychology research where data collecting is costly.
- The final contribution is the mixture density approach which makes it easy to transfer learning cross-dataset with different input features.

3.2 Preliminary on Personality Mining

Personality mining has been mainly studied by psychologist for decades using primarily descriptive and qualitative methodologies (Pennebaker and King 1999). With the growth of data analytics using machine learning algorithms, there are more quantitative efforts to estimate one's traits. However, due to the cost of collecting data, most personality datasets are relatively small in sample size and contain only text data (Mairesse and Walker 2007). Therefore, most research in this fields is based on textual data only, which yields a low accuracy on the results. Since the availability of multimedia data (Ponce-López et al. 2016; Biel and Gatica-Perez 2013) on personality mining in recent years, we can now apply advanced neural network approach to build a better prediction model which utilizes multimodal features.

Current literatures on personality mining mainly focus on feature extraction and selection using different analysis (Sarkar et al. 2014; Vinciarelli and Mohammadi 2014). For textual features, most of current papers use Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001), Bag of Words and other text sentiment

analysis techniques. Regarding audio visual features, there are many different approaches using *Python* or *MATLAB* packages for prosody cue, speaking activity, scenery and face recognition (Pentland 2004; Bradski 2000). There is also an application of a multimodal feature extraction technique called Doc2Vec (Chen et al. 2016). The result shows improvement in prediction accuracy of some but not all Big Five traits. The variety of extracted feature sets tend to have significantly different correlation to the personality scores (Verhoeven et al. 2014), which makes it difficult to compare the methodologies and empirical results even with the same dataset. A summary of previous features and methods used in personality mining are presented in Table 3.1 below.

There have been some applications of machine learning methodologies to build prediction model (Buettner 2016). Researchers have also looked predicting personality scores both separately or together as a multivariate problem using support vector machines and decision tree algorithms with different stacking models (Farnadi et al. 2014, 2016). According to their results, the differences between univariate and multivariate model are not significant. Therefore, we will not approach personality mining as a multi-label prediction problem and will compare our model with two single stacking models from these papers. They also suggest that the cross-datasets transfer learning would not help with prediction accuracy. We will test this hypothesis again with our neural network using both two multimodal personality datasets that are publicly available (Ponce-López et al. 2016; Biel and Gatica-Perez 2013).

Personality prediction has been commonly approached as regression problem. Even though we can convert the personality scores to binary labels for classification model using a certain threshold, many researchers have proven that it is not a good practice to determine human characteristics. Most classification models have also showed a pretty low prediction accuracy around 52% to 65% only (Farnadi et al. 2013). Moreover, the sample size might not be equally distributed in each of binary classes. Therefore, we will only focus on building the regression model for personality mining within the scope of this research.

As far as we concern, there have been no application of neural network in mul-

Table 3.1 : Literature Review on Personality Mining

Paper	**Data Type**	**Corpus**	**Features**	**Algorithms**	**Tools**
Pennebaker and King (1999)	text	essays	LIWC	Correlation analysis	n/a
Gill and Oberlander (2002)	text	emails (105 students)	bigrams	Bigram Analysis	n/a
Nowson et al. (2005)	text	weblogs (410K words)	word list	Correlation analysis	n/a
Argamon et al. (2005)	text	essays	word list, conj.	SMO	Weka
Oberlander and Nowson (2006)	text	weblogs (410K words)	N-grams	NB, SMO	Weka
Mairesse and Walker (2006)	conversation extracts, text	96 persons (100K words)	LIWC, MRC, utterances	RankBoost	Weka
Mairesse and Walker (2007)	text, speech	essays	LIWC, MRC	C4.5, NB, SMO, M5	Weka
Argamon et al. (2007)	text	essays	word list, conj.	SMO	Weka, ATMan
Rigby and Hassan (2007)	text	mail. lists (140K emails)	LIWC	C4.5	Weka, SPSS
Gill et al. (2009)	text	weblogs (14.8 words)	LIWC	Linear Regression	n/a
Wang et al. (2009)	text,	weblogs (200 pairs)	lexical freq. ,TFIDF	Logistic Regression	Minitab
Yarkoni (2010)	text	weblogs (100K words)	LIWC	Correlation analysis	n/a
Iacobelli et al. (2011)	text	weblogs (3000)	LIWC, bigrams,	SVM, SMO, NB	Weka
Roshchina et al. (2011)	text	TripAdvisor reviews	LIWC, MRC	Linear, M5, SVM	Weka
Quercia et al. (2011)	meta	335 Twitter users	Twitter counts	M5 rules	Weka
Golbeck et al. (2011)	text, meta	279 FB users	5 classes (161 in total)	M5, Gaussian process	Weka
Celli (2012)	text	1065 posts	22 linguistics Features	Major-based Classification	n/a

timodal personality mining. Even though neural network still doesn't significantly outperform machine learning algorithms regarding regression problems, we believe neural network would have certain advantage in psychology fields such as human thinking and behavior. Our proposed MMDB neural network would be the first attempt to estimate one's traits using this advanced approach. It helps solve the challenge in personality mining with limitations in sample size and multimodal data. The research would contribute to the current literature with the shifting trend to use deep learning techniques.

3.3 Multimodal Mixture Density Boosting Network

We propose here a multimodal neural network that can combine different type of input data at different sizes with our Discriminant Correlation Analysis (DCA) Feature Fusion layer. Then the fused features will be used as inputs and target for layers in our Mixture Density Network to adjust for the information loss due to feature fusion without over-fitting the model. Last but not least, the output of Mixture Density Network layers will be the input for layers in our Dynamic Cascade Boosting Network to regress the final prediction with high accuracy.

3.3.1 DCA Feature Fusion Layer

The standard DCA Feature Fusion algorithm (Haghighat et al. 2016) considers the class associations in feature sets. It eliminates the between-class correlations and restricts the correlations to be within classes. DCA maximizes the correlation of corresponding features across the two feature sets and in addition, decorrelates features that belong to different classes within each feature set. It also solves small sample size (SSS) problem, where the number of samples is less than the number of features which makes the covariance matrices singular and non-invertible. Within our multimodal neural network, the DCA Feature Fusion Layer will first fuse the video, audio and text features pairwise, then it will aggregate to compute the final fused feature of three modal inputs as

$$ff_{vst} = ff_{vs} + ff_{vt} + ff_{st} \tag{3.1}$$

where $ff_{vst}, ff_{vs}, ff_{vt}, ff_{st}$ are the DCA fused features of video-sound, video-text, sound-text and video-sound-text features accordingly. We will use ff_{vst} as target scores and $ff_{vs}, ff_{vt}, ff_{st}$ as inputs in our Mixture Density Neural Network.

3.3.2 Mixture Density Network

Mixture Density Networks (MDN) (Bishop 1994) predicts not a single output value but an entire probability distribution for the output. This helps us get the inference between each fused feature and the aggregated ff_{vst}, reducing the loss of information from the DCA Feature Fusion Layer without over-fitting the model. The MDN will predict Mixture Gaussian distributions, where the output value is modeled as a sum of many Gaussian random values, each with different means and standard deviations. So for each input x, we will predict a probability weighted sum of smaller Gaussian probability distributions

$$P(Y = y|X = x) = \sum_{k=0}^{K-1} \Pi_k(x)\phi(y, \mu_k(x), \sigma_k(x)) \tag{3.2}$$

where $\phi(y, \mu_k(x), \sigma_k(x))$ is the probability distribution function (pdf) of Gaussian distribution k with predicted mean $\mu_k(x)$ and predicted deviation $\sigma_k(x)$. $\Pi_k(x)$ is the predicted weight of Gaussian distribution k, and $\sum_{k=0}^{K-1} \Pi_k(x) = 1$ to ensure that the pdf integrates to 1. Each of the parameters $\Pi_k(x), \mu_k(x), \sigma_k(x)$ will be determined by the neural network, as a function of the input x. We construct our MDN leveraging TensorFlow Slim, with three fully-connected hidden layers of 10 nodes each and Adam Optimizer for training. This feed-forward neural network will parameter $1,000$ Gaussian mixture components as outputs after $1,000$ iteration rounds.

3.3.3 Dynamic Cascade Boosting Network

Boosting algorithms have been one of the most effective machine learning methodologies for regression problems. Therefore, we believe the incorporation of boosting algorithms into our multimodal neural network would help increase the prediction accuracy for personality mining. In this model, we use gradient boosting regression

from Scikit-Learn with 100 estimators as our base learner algorithm. For testing purpose, we only use two variants of hyperparameters in our model set up, where learning rates are 0.001 and 0.01 respectively. Other parameters are the same for all models, where max depth is 1 and loss function is least absolute deviations (LAD). To avoid over-fitting, we did not perform any specific form of parameter tuning either manually or automatically.

Our dynamic boosting network, inspired by gcForest (Zhou and Feng 2017), is built with a cascade structure, where each layer is embedded with multiple boosting algorithms. These algorithms will estimate the personality scores separately, then the average scores will be evaluated using mean accuracy (MA) before constructing the next cascade layer.

$$MA = 1 - 1/N \sum_{i=1}^{N} |y_{pred} - y_{true}| \tag{3.3}$$

After feeding the output of previous cascade layer to a new layer, the network will automatically assess the prediction accuracy of the model, and the training procedure will stop if there is no significant increase in performance. For our experiment, we set the tolerance rate to zero, which means new cascade layer will be constructed even with the smallest increase in MA values. The number of layers in this dynamic cascade boosting network will be implicitly constructed depending upon how fast the model learn (see Fig. 3.2). During our experiment, the number of cascade layers constructed is ranging from 2 layers to 8 layers. Since our dynamic boosting network can reactively chooses the number of cascade layer and decides on early stopping, it can efficiently handle different dataset sizes without wasting computing power.

3.4 Experiment

To study the performance of the MMDB model, experiments were conducted on two public datasets: First Impressions dataset and YouTube Personality dataset. Comparisons were made against several baselines.

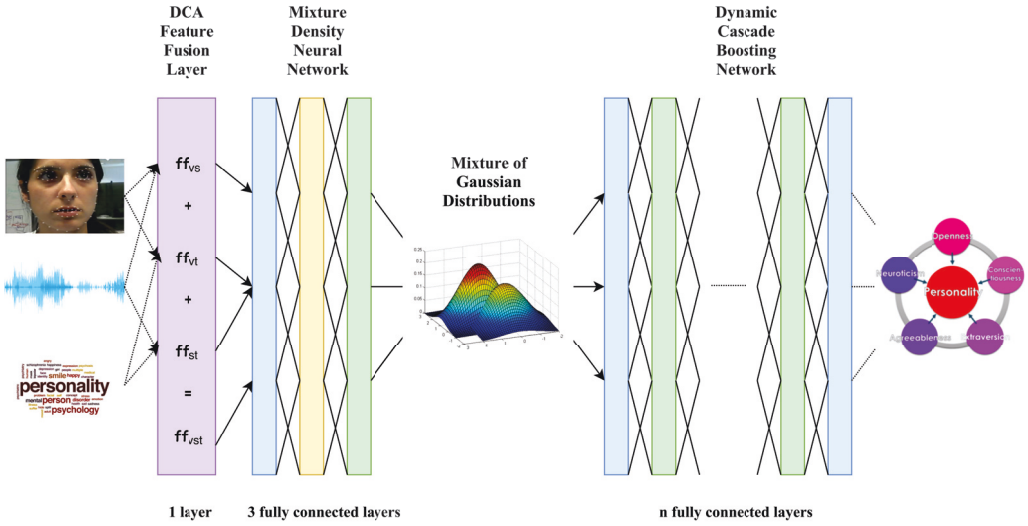


Figure 3.2 : The complete architecture of our MMDB Neural Network

3.4.1 Datasets

YouTube Personality dataset (YT)

The Youtube Personality dataset (Biel and Gatica-Perez 2013) consists of 404 YouTube clips when Video bloggers explicitly show themselves in front of the a webcam talking about a variety of topics. The text transcriptions are provided in raw text and contain 10K unique words and 240K word tokens. The personality impressions consist of Big Five scores that were collected using Amazon Mechanical Turk (AMT) and the Ten-Item Personality Inventory (TIPI) (Goldberg et al. 2006). The scores are rescaled into range $[0, 1]$.

First Impressions dataset (FI)

The First Impression dataset (Ponce-López et al. 2016) comprises $10,000$ clips with an average duration of 15 seconds. Each clip is a video of people facing and speaking English to the camera. The *gender*, *age*, *nationality*, and *ethnicity* information can be observed from clips. Beside sensory data, the dataset also contains the text transcription of the speakers' words. In total, $435,984$ words were transcribed ($183,861$ non-stopwords), which corresponds to 43 words per clip on average. Each clip is labeled with Big Five personality traits scores from $[0, 1]$.

3.4.2 Data Cleaning and Feature Extraction

As the feature set of the two datasets are different, feature importances would be varied. Therefore, we did not perform any correlation analysis and feature selection. All extracted features were the direct inputs for our multimodal network and baseline models.

Video Features

Videos in FI dataset are showing one person speaking directly to the camera, therefore we are more interested in their facial movements and gestures than other general scenic data. To extract video features, we used OpenCV (Bradski 2000) to extract the landmarks data with face detection, alignment and tracker for every single frame of each clip. From these image-base data points, we then extracted major facial features such as Yaw, Roll, Eyes and Lip movements, etc. Finally, we averaged these facial features across all frames of each clip.

Audio Features

Several speech features were extracted from audio, such as pitch and energy. These features were extracted at the interval of 5000Hz each by using the Hidden Markov Models. Similar to video features, we computed the average of speech features for each audio file, which resulted 21 audio features in total.

Text Features

The LIWC 2015 (Pennebaker et al. 2001) dictionary was used to extract text features from the transcription of each video clips. Even though there are many text analysis tools available such as Bag of Words, Word Sentiment, etc., these approaches might not suitable for our corpus with short text and variety of topics. Therefore, we used only LIWC as the standard approach, which covered many topic-related features (e.g. work, family, friend, money), sentiment (e.g. possemo, negemo) and even speech related features (non-fluent). A total 93 text features were extracted.

3.4.3 Baselines and Evaluation Metrics

Evaluation Metrics

Similar to related works, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used as our evaluation metrics:

$$\text{MAE} = 1/N \sum_{i=1}^{N} |y_{\text{pred}} - y_{\text{true}}| \tag{3.4}$$

$$\text{RMSE} = \sqrt{1/N \sum_{i=1}^{N} (y_{\text{pred}} - y_{\text{true}})^2} \tag{3.5}$$

where N is the size of data, and y_{pred} and y_{true} are the predicted and true personality scores, respectively. We also performed rank evaluation by ranking data instances for each type of personality and comparing against the true ranking using Spearman Rank Correlation Coefficient (Spearman's rho) (Spearman 1987).

Baselines

We rebuilt prediction models using some state-of-art models in current literature and used as baselines to evaluate the performances of our MMDB neural network. Specifically, we built the bench-marking models using the Gaussian Process from (Golbeck et al. 2011) and two single stacking models with base learner support vector regression and decision tree from (Farnadi et al. 2014, 2016), which were denoted as GP, SVR and DT respectively.

We designed a simple neural network (NN) for comparison with our MMDB neural network, using TensorFlow framework to construct three fully-connected ReLU layers with 10 nodes each. The deep neural network regressor used Adam Optimizer as solver and was trained for $10,000$ iterations. The input features for these baseline models included all text, auditory and visual features as in our MMDB model. We also did not perform any specific parameter tuning on these baseline models to avoid overfitting the model, which would not be good in the case of personality mining and transfer learning when applied to real business use cases.

3.5 Result and Discussion

3.5.1 Empirical Result

For individual dataset evaluation, we perform 10-fold cross validation and compare the results with baseline models using the extracted features as input. For cross-datasets evaluation and component testing, we split each dataset into train set and test at ratio 8 : 2. For FI dataset, there are 8,000 and 2,000 instances in train and test set. For YT dataset, there are 323 and 81 instances in train and test set. The 10-fold cross validation results are showed in Table 3.2, with agreeableness as AGR, conscientiousness as CON, extraversion as EXT, neuroticism as NEU and openness as OPN.

On the MAE, our MMDB neural network performs better or on the same level of accuracy to some current methods on both datasets. On rank evaluation, our model performs significantly better in most personality dimensions for both tested datasets. It has the highest Spearman's rho together with NN model for EXT on YT dataset. It performs slightly worse than SVR and NN for EXT and NEU on FI dataset. In general, our model has better prediction accuracy than other personality mining models. Especially in dealing with small data size case in psychology field like the YT dataset, our MMDB neural network helps improve prediction accuracy significantly compared to other mentioned models.

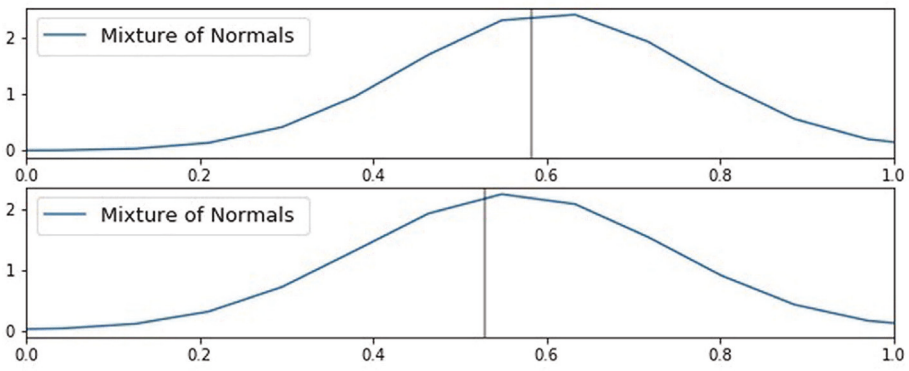


Figure 3.3 : Sample distribution predictions for Openness from MMD network

(a) (where the black vertical line is the ground truth personality scores)

Table 3.2 : 10-fold Cross-validation on YouTube and First Impression Datasets

		MAE					**rho**				
	Model	**AGR**	**CON**	**EXT**	**NEU**	**OPN**	**AGR**	**CON**	**EXT**	**NEU**	**OPN**
YouTube Dataset	**MMDB**	**.1070**	**.0980**	**.1285**	**.0973**	**.0957**	**.4352**	**.3030**	**.3233**	**.3398**	**.0949**
	NN	2.19	3.68	2.22	1.55	1.08	.1423	.0838	.2156	-.0627	.0253
	GP	.6138	.5829	.6042	.6277	.6108	.0338	-.1029	.0021	-.0223	.0488
	SVR	.1211	.1008	.1354	.1040	.0960	.1218	.1739	.0799	.0943	.0397
	DT	.1363	.1291	.1669	.1295	.1338	.2972	.2481	.1578	.1861	.0962
First Impression Dataset	**MMDB**	**.1042**	**.1228**	**.1184**	**.1186**	**.1136**	**.2616**	**.3025**	**.3356**	**.3643**	**.3118**
	NN	.1254	.1646	.1380	.1314	.1254	.1191	.1793	.1457	.2000	.0846
	GP	.5495	.5243	.4767	.5209	.5667	.0626	.0493	.0836	.0847	.0759
	SVR	.1070	.1261	.1226	.1235	.1171	.0590	.0833	.0650	.0739	.0478
	DT	.1441	.1612	.1552	.1585	.1522	.1117	.1587	.1864	.1762	.1562

Table 3.3 : Transfer learning.

	MAE					RMSE				
Direction	AGR	CON	EXT	NEU	OPN	AGR	CON	EXT	NEU	OPN
YT→FI	.1348	.1291	.1457	.1640	.1250	.1706	.1599	.1805	.2042	.1568
FI→YT	.1258	.1219	.1782	.1344	.0972	.1486	.1539	.1893	.1579	.1155

We also want to test whether we can perform transfer learning the personality mining models using cross datasets. Unlike most machine learning models where we need the same number of input features from different dataset to perform cross-data learning, our model use DCA and mixture density to fuse features and compute the inference with Gaussian distribution to create an equal input nodes for cascade boosting network. This allows our model to transfer learning easily between datasets with multimodal features.

The reported results in Table 3.3 are in line with current literature that transfer

learning does not improve prediction accuracy. This is explainable as the trait dimensions' scores were denoted by different author using various techniques and scales. However, when comparing the MAE between Table 3.3 and Table 3.2, the transfer learning from the FI to YT dataset still performs better than some baseline models trained on YT dataset for different personality dimensions. This positive result of transfer learning could help in specific case where one dataset is much smaller than the other dataset, then transfer learning would have better result than in the vice versa case.

Table 3.4 : MMDB and MMD evaluation with YouTube Personality dataset

	MAE					RMSE				
Model	AGR	CON	EXT	NEU	OPN	AGR	CON	EXT	NEU	OPN
MMDB	.10940	.10286	.15039	.09190	.09449	.13455	.13386	.16981	.12121	.11699
MMD	.11244	.11589	.15276	.13414	.11643	.14369	.14530	.18018	.17003	.14822

Last but not least, we believe the integration of gradient boosting algorithms into neural network would significantly improve the prediction accuracy, especially in the case of small sample size. We test this hypothesis using the YouTube Personality dataset by running two separated prediction models. The first one is our proposed MMDB neural network with the full layers. The second one contains only the DCA fusion layer and the mixture density neural network (MMD). The results in Table 3.4 show that the performance of MMD is quite satisfactory, which can still outperform other baseline models.

As observed in Fig. 3.4a), the predicted means of the Gaussian distributions are very close to the ground truth personality scores, which give us a lower MAE and RSME with MMD model only. However, the dynamic cascade boosting network reduces the MAE and RMSE further. This proves our hypothesis on the effectiveness of gradient boosting algorithms in these regression problems. Our proposed MMDB neural network and its component layers together can solve better the challenges in personality mining.

3.5.2 Discussion

Within the scope of this research, we have identified the current research gap in personality mining, which was dominantly using costly qualitative methods e.g. surveys. Most quantitative research only use text mining techniques to predict personality, while we believe sensory data can contain useful information about one's traits. The recent approaches using machine learning algorithms have limitations when it comes to small sample size, which have low prediction accuracy. Our proposed MMDB neural network has been proven to be an effective model in solving personality mining challenges.

The MMDB neural network is a quantitative methodology, which is the first research work to use deep learning approach in personality mining. It consists of three main components. The first one is a DCA feature fusion layer to fuse multimodal features from visual, auditory and textual data. The second component is a mixture density neural network to predict the full distribution of personality scores. This solves the common problem of over-fitting due to small sample size. Finally, a dynamic cascade boosting network will significantly improve the accuracy of finally prediction.

Our MMDB neural network has outperformed other baseline models from current literature. The experiments with cross-datasets have showed the transfer learning for personality mining is not effective in general, but can still help in case of predicting on small sample size using our model from bigger dataset. Last but not least we test the components of our MMDB neural network individually. The results confirm our hypothesis of integrating the dynamic cascade boosting will improve prediction accuracy of the mixture density network. In conclusion, our MMDB neural network has great performances, especially with small datasets in personality and psychology fields.

Since the beginning of this research, personality mining has been applied in the industry for multiple business use cases, ranging from recruitment, employee training to customer relationship management. Research in personality mining would be a significant contribution to both theoretical and applied fields. However, due

to data privacy concern, some of these personality mining datasets are no longer made available publicly. There is also no followup research with the same group of people according to the data owners. Hence, future research on personality mining, especially with multimodal data with video, audio and text, should focus more on longitudinal data in real world scenarios to further highlights its business applicability.

Regarding the technical contribution, we would also like to explore more options in future work to improve the prediction model for personality mining using particularly deep learning techniques. We would work on a more intuitive neural networks that can perform personality prediction using the videos as raw input data without the intermediate step of features extraction and feature fusion. With the current progress in computer vision and multimodal neural networks, we believe it will bring further breakthroughs in personality mining research.

The methodology and results in this chapter serve as the research foundation for the next chapter where we applies transfer learning to personality mining of customers in financial wealth industry.

Chapter 4

Unstructured Data Mining and Interpretable Machine Learning for Wealth Customer Data Analytics

4.1 Background and Motivation

Customer relationship management (CRM) has always been a core business function for any company. Among the components of CRM, increasing customer engagement and loyalty is one of the most challenging tasks. Although customer acquisition and retention are both important, prior research (Fornell and Wernerfelt 1987) has showed that acquiring a new customer is typically five times more expensive than retaining an existing customer. Because of the high cost of customer acquisition, established businesses focus more on customer retention instead of acquisition. In customer retention, predicting customer churn risk is an important task. Each single percentage increase in customer churn prediction accuracy could potentially lead to a substantial revenue saving. This is particularly true for the financial services sector in which each customer may contribute to a considerable amount of profits, while customer engagement and loyalty are relatively low (Rudin 2019).

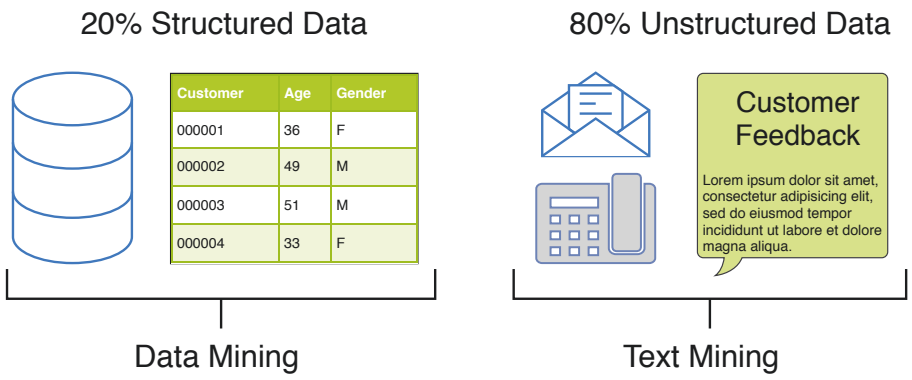
In this digital age, firms have been intensively relying on data analytics for customer churn forecast. With advances in machine learning, business intelligence applications can provide better customer insights by leveraging advanced data mining techniques powered by machine learning algorithms. Within the financial services field, a data-driven customer churn forecast model is essential for constructing efficient and effective retention strategies. Unfortunately, the lack of transparency and interpretability has limited broader adoption of machine learning models and becomes a growing concern (Adadi and Berrada 2018; Rudin 2019). In this research, we build an ensemble prediction model incorporating interpretable machine learning

techniques to analyze multi-stacking data for churn forecast and customer retention, both applicable to financial services and other industries.

Churn decision of financial services customers can be associated with various factors, including customer demographics, behaviors, affective status, etc. The nature of churn prediction is a supervised learning problem which lies in feature construction and engineering, either from structured or unstructured customer data apart from cutting-edge learning algorithms. Existing features of churn prediction models are mainly derived from demographic data (gender, employment status, educational level, etc.), transactional data (investment decision, buying insurance, etc.), social network data (financial advisor, joint account, etc.) and other behavior data (risk-averse level, investment preferences, etc.). Most of these features are structured data in table format due to the ease of data collection directly from commercial databases. In fact, structured data is often the only acceptable input for many off-the-shelf learning algorithms. However, through empirical correlation analysis of such structured features with customer churn, we observed that the correlation coefficients are actually quite low, suggesting that using structured features standalone would not guarantee the satisfactory forecast performance. This motivates us to take features from unstructured data into consideration.

The cognitive process of churn decision making is complex and can be influenced by many indicators. For example, the affective status of customers has close indications to churn decisions which are reflected by various factors, ranging from products and services satisfactory level to personal feelings and opinions. Although basic communications data from customer service call center (e.g. frequency, call lengths, etc.) has been incorporated into customer churn prediction models in telecommunications (Wei and Chiu 2002), this interaction information is still used in a limited way (most often in form of structured features), which can only slightly improve the prediction accuracy. The content of the communication, which is unstructured data, has not yet been adequately utilized to capture finer granular customer insights for the financial gains of the firms. By deriving the customers' speaking pattern and personalities from unstructured data, we aim to predict churn risk more accurately

Figure 4.1 : Structured and unstructured data



and increase customer retention rate even further.

From the methodology aspect, researchers have investigated various kinds of feature engineering techniques with multiple extraction and stacking approaches. However, any further performance lift in terms of forecast accuracy using these approaches seems to be limited. Furthermore, those table-formatted features are usually accounted for only 20% of available customer data in a typical organization, while the majority 80% of the data obtained is the unstructured information (see Figure 4.1). Examples of such information include communication via phone calls, emails, messages, and social media channels (Holzinger et al. 2013). It is a much more challenging task to mine these types of customer data as they are often not obtained and stored on a regular basis.

In recent years, content mining combined with natural language processing (NLP) on unstructured data advanced the current predictive modeling, showing very promising results (Sun et al. 2017), e.g., opinion mining and sentimental analysis (Ravi et al. 2017). However, more finely granular customer affective statues than sentiment scores, which are carried by textual features, are not sufficiently exploited, indicating the research gap for customer churn prediction. In this regard, some related works have been done, e.g. Senanayake et al. (2015) proved that a cognitive approach should be incorporated into the customer churn prediction model. Leveraging advanced text mining techniques to extract additional features such as emotions and personalities from these customer communication data further increases

the churn forecast accuracy (Coussement and Van den Poel 2009) and reveal meaningful customer intel. These works inspire the churn prediction model we propose in this research.

Our research seeks to answer three questions: 1) Is there any evidence that unstructured data can be helpful for customer churn prediction?; 2) What are the approaches and techniques suitable to utilize this unstructured data for churn prediction model?; and 3) What are the profiles and characteristics of customers with high churn risk and how we can retain these clients? To answer these questions, we evaluate the effectiveness of unstructured data using different text mining methods. Benefited by the recent advancement in NLP and interpretable machine learning, we propose the churn forecast model that takes benefit of these techniques to extract customer insights utilizing both structured and unstructured data. Particularly, we leverage the *textual features* extracted from customer call logs alongside other business data provided by our industry partner, a Superannuation management company, to enhance customer churn prediction model. The churn forecast and interpreted machine learning model results serve as a data-driven decision support that the firm would use to develop better customer retention strategies.

The contributions of our research can be broken down into four main points:

i. To the best of our knowledge, this is one of the first attempts to propose a churn prediction model incorporating unstructured text mining with structured data mining and interpretable machine learning for churn prediction in the financial services.

ii. We demonstrate that leveraging the unstructured data and interpretable machine learning can capture comprehensive and useful customer preference spectrum for churn prediction. The model has been fully tested and then deployed on large-scale real-world datasets involving two million calls from more than two hundred thousand customers.

iii. We compare multiple text mining techniques to find suitable approaches and to fully leverage the unstructured data for churn prediction. We explore three

textual feature representations and particularly personality traits in the research.

iv. We are perhaps among the first to use interpretable machine learning in churn prediction. The interpretable machine learning approach we proposed allows the evaluation of feature importance not only at the whole sample level, but also at the customer segments and even individual customer levels. The insights from interpretable machine learning are useful in developing customized retention strategies for different cohorts.

4.2 Preliminary on Wealth Customer Data Analytics

Big data has been playing a vital role in strategic business information systems (Grover et al. 2018), which attracts numerous research efforts from both academia and industry sides and mainly focuses on consumer analytics using big customer data (Kitchens et al. 2018). A strategic decision support system for customer retention is the most important key to business success and has been extensively studied in recent years. Existing churn forecast approaches have utilized multiple machine learning algorithms to increase their prediction accuracy (Almana et al. 2014). Almost all of these models are using only the structured data (Verbeke et al. 2011), and some of them have achieved great results with the tree-based algorithms and recently the neural networks approaches (Hung et al. 2006).

Different aspects of the business, even the self-service level, have been tested for churn forecast, which revealed interesting insights on customer behaviors (Scherer et al. 2015). There are also various attempts (Coussement and Van den Poel 2008) to incorporate customer call center data in such decision support information systems (Wei and Chiu 2002). However, the current state-of-art frameworks only focus on using the table-formatted data, e.g. the total number of calls, call duration (Huang et al. 2015). This common practice is mainly due to the significantly high cost of obtaining and storing the unstructured data, e.g., the transcriptions of the calls, the text of the chats. Moreover, there are also other customer privacy and ethical concerns, which require further efforts and costs for the firms to obtain and

anonymize the data. These are the main reasons why most current approaches do not consider unstructured data for their prediction models.

In the financial services field, client churn forecast has been extensively researched using different machine learning algorithms, particularly in the banking sector (Ali and Arıtürk 2014). One of the most popular and best-performing algorithms is support vector machines, especially in the case of having an imbalanced dataset of credit card customers (Farquad et al. 2014). More advanced tree-based algorithms have also been tested on electronic banking customer data (Keramati et al. 2016) and achieve some positive results. A hybrid methodology combining k-Reverse-Nearest Neighborhood and One Class Support Vector Machine (OCSVM) has been applied to solve the sample problem (Sundarkumar and Ravi 2015). Researchers also attempt to combine various fuzzy methodologies with other machine learning algorithms (Karahoca et al. 2016) to increase prediction accuracy.

Considering the Superannuation industry in particular, customer retention strategies have been constructed with decision support systems using qualitative approaches, e.g., customer survey, focus group. Recently, big data analytics and feature engineering techniques have been applied in the Superannuation industry for customer retention (Chu et al. 2016). The empirical results from their experiments show improvement in prediction accuracy. However, the results are not significantly different among multiple tested algorithms and show that these approaches cannot be further enhanced from the algorithm side using a similar type of structured data. We argue that the unstructured data which reveals other dimensions of consumer insights could be used to increase the prediction accuracy for better customer segmentation and retention strategies.

Big data text analytics has been proven to be effective in many business cases, especially in evaluating customer agility and engagement using online reviews (Zhou et al. 2018). Regarding the application of text mining in the financial services field, researchers have tested different techniques for sentiment analysis based on customer feedback and social media posts (Yee Liau and Pei Tan 2014). With the advancement of text mining research, there are various methods to be used to derive indicative

user behavioral preference. A hybrid model consisting of concept-level sentiment and fuzzy formal concept analysis has been proposed to classify opinions from customer complaints (Ponce-López et al. 2016).

In customer retention, researchers have taken into account the emotions based on text from customer emails to enhance churn prediction and achieve good results (Coussement and Van den Poel 2009). We believe there is richer lexical information carried by in the words and phrases spoken directly by customers instead of a simple binary classification of positive or negative sentimental opinions. Recent research has tried to extract multiple term-based features as input for churn forecast models [48]. However, it is still unclear from these researches how textual information improves prediction accuracy in comparison with basic features and how it can help with retention strategies.

On the other hand, personality mining has revealed significantly meaningful insights in terms of characterizing customer behavior, satisfaction (Manner 2017) and loyalty in other consumer industries (Kim et al. 2018). Researchers have proven that many personality traits are highly correlated with customer empowerment and satisfaction in the retail industry and suggest strategies based on these insights (Castillo 2017). Customers in the financial services industry have similar characters to the retail consumers, yet there is little research on customer retention in this field using quantitative personality mining. Only qualitative survey methods have been applied to bank customers (Al-Hawari 2015).

Moreover, in recent years, interpretable machine learning has been proposed to both explaining the prediction model and extracting business insights. Particularly, ensemble learning based on generalized additive models has been proposed to combine a traditional churn prediction model and explainable machine learning (De Bock and Van den Poel 2012). This method is mainly based on the improvement on the overall feature importance. On the other hand, Shapley Additive Explanations (SHAP) (Lundberg and Lee 2017) value has been applied to related tree-based methods to explain both direction and magnitude of the features for every individual customer. Applying these methods in financial services can help reveal useful

information on our different types of customers.

Realizing this research gap, we propose a more comprehensive churn prediction model leveraging both structured and unstructured big data and interpretable machine learning to extract meaningful insights and support managerial decisions. The predicted results serve as a decision support metric to construct suitable customer segmentations, while the interpretable machine learning results can help understand important customer traits for building retention strategies.

Table 4.1 presents a comparison of our work with related researches, their chosen text features and methods. The advantages and novelty of our model lie in the following aspects:

i. From the data perspective, we leverage real business datasets with four different types of financial customers, which are useful in revealing varied behaviors of different customer types. The datasets are in large scale, involving two million calls from more than two hundred thousand customers.

ii. From the text feature perspective, previous approaches focus only on sentiment scores or term-based features. Our model combines three different types of text features (term importance, phrase embedding, and lexical information) and expands to exploit personality traits to capture more hidden information with text mining.

iii. From the model perspective, instead of simply using the machine learning algorithms such as Random Forest or XgBoost to construct a prediction model, we design the ensemble approach with multi-stacking of multiple prediction models to improve both accuracy and computation speed.

iv. Our model is perhaps among the first in the existing literature to incorporate interpretable machine learning to analyze impactful features globally as a whole dataset and also locally as customer segments and individually for each individual customer.

v. In addition to advancing the technical aspects of churn prediction approaches,

Table 4.1 : Existing Text Mining Approaches in Customer Research

Paper	**Features**	**Methods**
Our research	Basic Profiles, LIWC, TF-IDF, Word2Vec, Personality Traits	Multi-stacking Ensemble Prediction model with XgBoost, Logistic Regression, Gaussian Naive Bayes, Random Forest. Interpretable Machine Learning with SHAP-MRMR+ and customer segmenting with SOM for marketing strategies.
Yee Liau and Pei Tan (2014)	Sentiment Scores, SentiStrength	K-Means clustering
Zhou et al. (2018)	TF-IDF	SVD-Based Semantic Keyword Similarity
Ravi et al. (2017)	TF-IDF, Sentiment Scores (Sentic Net3)	Fuzzy formal concept analysis and concept-level sentiment analysis
Coussement and Van den Poel (2009)	LIWC (only posemo and negemo)	Logistic Regression, SVM, Random Forest
Vo et al. (2018)	LIWC, TF-IDF, Word2Vec	XgBoost
Castillo (2017)	Big 5 Personalities, Empowerment	Qualitative Survey, Multiple Regression Analysis
Al-Hawari (2015)	Big 5 Personalities	Qualitative Survey, Statistical Analysis

we move a further step to discuss how the results from our model can be used to design customized retention strategies which can help firms retain customers of high value and high risk to churn. This discussion explicates the managerial implications of our model, comparing to the existing studies that merely focus on the technical aspects of churn prediction.

4.3 Multi-stacking ensemble model for churn prediction

The first part of our approach analyzes an integrated database of customer call log and profile data to construct churn prediction models. Our first hypothesis is that the multi-stacking prediction models using combined features can improve the churn prediction accuracy further than those using only the basic structured

customer profile data. We incorporate multiple text mining techniques on the customer call logs datasets to extract four different feature sets: *Term Importance*, *Phrase Embedding*, *Lexical Information*, and *Personality Traits*. The rationale for using different text mining approaches is to capture insights about the customers at multiple granular levels, ranging from raw term frequency vector to underlying semantics space. Term Importance captures the lexical importance of bag-of-words from a given text corpus, e.g., sentence or paragraph. Phrase Embedding would help reveal the insights of the customer decision-making process based on the words used and their linkages to each other. Lexical Information would contain insights regarding latent concepts and topic-related terms used, which illustrates customer characteristics such as money-savvy or family-oriented. Finally, Personality Traits determines different customer psychological trait spectrum, e.g., Big Five personalities, which has significant predictive power for the affective statuses and churn decision. After constructing the above textual features, we feed them into supervised prediction models alongside other features derived from structured data to predict customer churn probabilities.

4.3.1 Unstructured data mining

Term Importance

The most common technique to derive Term Importance features is the Bag-of-Words model. However, uni-gram or multi-gram models extract hundreds of thousands of textual features since we have almost one million call logs in our dataset. In our context, "term frequency-inverse document frequency" (TF-IDF) (Luhn 1957; Scherer et al. 2015) is a more suitable technique as there are many common words with less insightful meaning, e.g. "hello" or "the". This methodology captures the major textual meaning by using TF-IDF expression. We derive almost 10,000 TF-IDF features in total. Table 4.2 gives a snapshot of the Term Importance matrix extracted from our call logs dataset.

i. Term Frequency (TF): calculates the number of times each term appears in the call logs

Table 4.2 : Sample TFIDF features

Call ID	close	transfer	hello	yes	no	Churn
Call 1	1.63	0.24	0.07	0.20	0.73	1
Call 2	0	2.19	0.07	0.27	0.15	1
Call 3	1.63	0.97	0.07	0.40	0	0
Call 4	0	0	0.07	0.27	0	0
Call 5	0	0.49	0.07	0.20	0.15	0
Call 6	0	0.24	0.07	0	0.15	0
Call 7	0	0	0	0	0.15	0

ii. Inverse Document Frequency (IDF): calculates whether the term contain meaningful information or not, i.e., is a common or rare term used across all the call logs

$$\text{IDF(term,call)} = \frac{total\ number\ of\ calls}{total\ number\ of\ calls\ with\ the\ term} \tag{4.1}$$

iii. Term Frequency - Inverse Document Frequency (TF-IDF) is calculated as

$$\text{TF-IDF} = \text{TF} \times \text{IDF} \tag{4.2}$$

Phrase Embedding

Depending on the context, similar terms in English when combined in varied order can be understood completely different. Considering the Term Importance as single word feature alone would not be meaningful for our customer analysis. Therefore, we also aim to understand the semantics carried by various combination orders of terms by looking at the position and order of these phrases in a sentence, surrounding contexts, collocations, and their connections. To achieve this, we take into account the Phrase Embedding approach, which would help the forecast model to analyze the call logs transcript as sentences with contexts. We leveraged a widely used embedding algorithm, Word2Vec model (Rehurek and Sojka 2010) to extract a total of 50 word embedding features. Figure 4.2 illustrates an example of difference

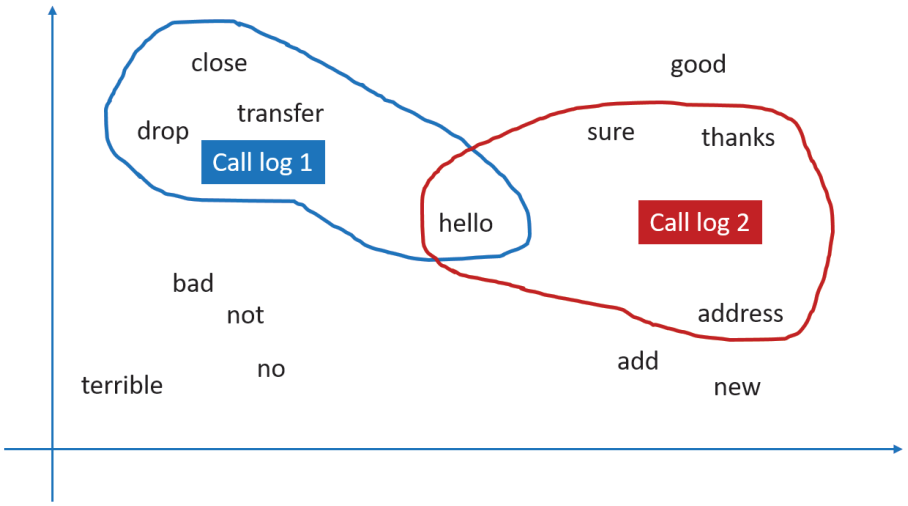


Figure 4.2 : Word Embedding model captures relationships between terms

term relationships within the Word Embedding model which could help provide more meaningful insights for our prediction model.

Lexical Information

One of the most popular NLP researches is sentiment analysis, which classifies text into a binary dimension as positive or negative emotions. We believe there are more meaningful insights in other dimensions and topics of the English language. In order to extract this information, we leverage the Linguistic Inquiry and Word Count 2015 (LIWC) to extract latent concepts and topic-related text features. The LIWC 2015 dictionary contains about $6,400$ terms and emotions. Each term has a separated corresponding dictionary entry that identifies its one or more categories. For example, the term "disappointed" belongs to five different categories: "Verb", "Overall Affect", "Past Focus", "Negative Emotion" and "Sad". If the customer said the term "disappointed", all these five lexical categories' scores would increase respectively. This is one of the most comprehensive text mining dictionaries consisting of multiple topic-based categories ("money", "leisure"), emotions ("anger", "anxious") and speech-related features ("filler, "non-fluent"), which is more suitable for our call logs dataset. The LIWC 2015 has total of 93 features. There are 12 punctuation-related features ("exclamation mark", "parenthesis", etc.). These

Table 4.3 : The Big Five Personality Traits

Trait	High Rank	Low Rank
Extroversion	Outgoing, active, seek excitement	Aloof, quiet, enjoy time alone
Neuroticism	Prone to stress, negative emotions	Emotionally stable, self-satisfied
Conscientiousness	Organized, punctual, hard-working	Spontaneous, careless, hedonistic
Agreeableness	Trusting, empathetic, compliant	Uncooperative, not listen to others
Openness	Creative, imaginative, curious	Practical, conventional, skeptical

features are only applicable for written text; hence they are omitted in our case of spoken text.

Personality Traits

Personalities are human characteristics differentiated and reflected through their cognitive and behavioral patterns. Personality Mining is an advanced data mining technique to find the traits of a person from the way he or she speaks and acts. Psychology studies have showed cognitive language spoken as unique signals for different human behaviors. Individuals having distinguished traits often present themselves and behave differently. Knowing one's traits and understanding the differences in their preferences would help with communicating and connecting to the person on a more personalized level. Our personality mining model incorporates the Five-Factor Model of Personality (Big Five) (Goldberg 1990) for personality mining task. The Big Five model contains five fundamental human traits: openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism (Table 3 presents the typical characteristics for each of the five traits). These traits, widely accepted by the psychologists as a standard measure, are proven to stay consistent despite age, gender or cultural background (Costa and McCrae 1992).

The rationale of using personality traits in churn prediction is that customers with varied personalities might make different financial decisions. From the model perspective, personality mining could be treated as a supervised learning task, and the results can explain customer behavior sufficiently. In this research, we use a

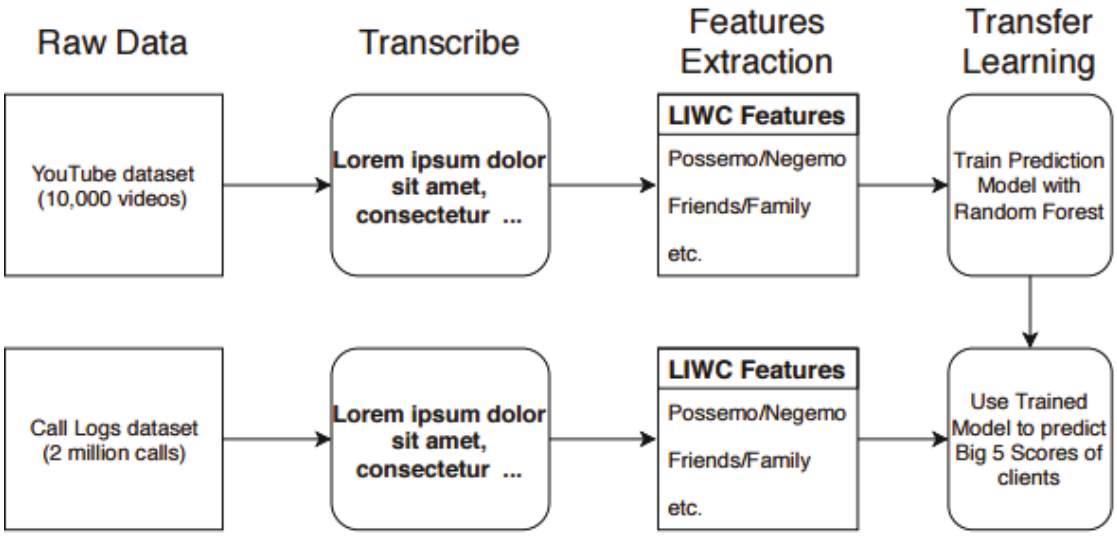


Figure 4.3 : Personality Traits Mining Methodology

random forest algorithm to train the prediction model for personality based on collected training data, e.g., the First Impression dataset. Then the trained model is transferred for use in identifying the Big Five traits of the customer based on their call logs. The process is illustrated in Figure 4.3. These five personality scores serve as input features for our prediction models and customer segmentation for suitable retention strategies.

4.3.2 Multi-stacking Ensemble Model for Churn Prediction

The customer churn forecast model automatically extracts all mentioned textual features, then combines with customer profile data from various business databases as input for the multi-stacking churn forecast models. The prediction model computes a ranked list of customers with various churn scores, upon which the company can decide which customers are at a higher risk to churn. We use the following supervised machine learning algorithms to build the churn forecast models:

Gaussian Naive Bayes (NB) Classifier

This supervised learning algorithm is built upon the famous Bayes' theorem (Bayes 1763). The computational foundation is based on the "naive" assumption that all pairwise features are not correlated with each other. We denote the churn

label as y and the dependent feature vector as x_1 through x_n where all features are pairwise uncorrelated. Bayes' theorem measures the conditional probability of churn label y as follows:

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)} \tag{4.3}$$

Since the probability $P(x_1, ..., x_n)$ is a constant based on our naive independence assumption, we can follow the classification rule:

$$P(y|x_1, ..., x_n) \propto Py \prod_{i=1}^{n} P(x_i|y) \tag{4.4}$$

$$\implies \hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y) \tag{4.5}$$

with Maximum A Posteriori (MAP) estimation (Gauvain and Lee 1994) to calculate the probability $P(x_i|y))$. Despite their naive assumptions, this classifier has worked quite effectively in many real-life business cases similar to our prediction problem. The algorithm requires a smaller sample of the training dataset to approximate the required parameters, which can be significantly fast in comparison with other advanced approaches (Zhang 2004). The algorithm serves as the baseline prediction model using the default setting from Python package Scikit-Learn (Pedregosa et al. 2011b).

Logistic Regression (LR) Classifier

Logistic Regression algorithm is derived from the statistical method (Cox 1958) of analyzing one or more uncorrelated variables in the datasets. The customer churn decision is evaluated as a dichotomous variable (in which there are only two possible outcomes of churn or not churn). The algorithm's target is to optimize the best fitting model to measure the correlation between customer churn decision and all other textual and standard account features. The computed coefficients to forecast the customer churn risk are calculated as follow:

$$\text{logit}(p) = ln\left(\frac{p}{1-p}\right) \tag{4.6}$$

where p represents the customer churn likelihood. To avoid over-fitting, our model implements the logistic regression algorithm with L2 regularization. Our binary class L2 penalized algorithm optimizes the corresponding cost function J:

$$J = \min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i(x_i^T w + c)\right) + 1\right) \tag{4.7}$$

where w represents the weight, c represents the cost, x_i and y_i represent the feature set and churn label of customer i respectively. Other setup and parameter tuning are the same as in the baseline model of similar churn prediction work Chu et al. (2016).

Random Forest (RF) Classifier

A Random Forest Classifier is a classifier based on a random forest family of classifiers based on a family of classifiers $h(x|\Theta_1), ..., h(x|\Theta_K)$ based on a classification tree with parameters Θ_k randomly chosen from a model random vector Θ. For the final classification $f(x)$ which combines the classifiers $\{h_k(x)\}$, each tree casts a vote for the most popular class at input x, and the class with the most votes wins. Specifically given data $D = \{(x_i, y_i)\}_{i=1}^n$, we train a family of classifiers $h_k(x)$. In our case, each classifier $h_k(x) \equiv h(x|\Theta_k)$ is a predictor of n and $y = \pm 1$ is the outcome associated with input x.

Extreme Gradient Boosting (XGB) Classifier

The algorithm was introduced by Chen and Guestrin (2016) in 2014 as an enhanced version of the greedy gradient boosting machine (Friedman 2001). XgBoost has become one of the most widely-used and effective algorithms in supervised machine learning. In our model, we first define the tree $f(x)$ as:

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^D \to \{1, 2, ..., T\} \tag{4.8}$$

where w represents the vector of the scores on the leaves, q is a data assigning function for the corresponding leaf, and T represents the total number of defined leaves. We also define the gradient g_i and the Hessian (second-order derivative) h_i as follow:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{4.9}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{4.10}$$

where y_i represents the churn label and $\hat{y}_i^{(t-1)}$ represents the predicted churn at time $(t-1)$. Using regularization to improve the generalization performance, the objective function at t^{th} tree is:

$$\implies obj^{(t)} = \sum_{i=1}^{n}\left[g_i w_{q(x_i)} + \frac{1}{2}h_i w^2_{q(x_i)}\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 + \alpha\sum_{j=1}^{T}|w_j| \tag{4.11}$$

where w_j represents the weight assigned to the j^{th} leaf, $\gamma = 0$, $\alpha = 0$ and $\lambda = 1$ are predefined parameters for the penalization and regularization terms respectively.

Multi-stacking Ensemble Model

We apply a multi-stacking ensemble method to build our final prediction model, as illustrated in Figure 4.4, by ensemble the predicted churn risks from single stacking models using only one type of features. This approach reduces the model training time significantly compared to the model with all features being linearly combined. The required time for training on Pension and Investment datasets are cut down to half, while the computation speeds of the other two bigger dataset models are almost three times faster. The predicted churn risk is computed as:

$$\hat{Y} = F_{f(x)} = F(Y_{CP} + Y_{TI} + Y_{PE} + Y_{LI} + Y_{PT}) \tag{4.12}$$

where $F_{f(x)}$ represents the final ensemble model, while Y_{CP}, Y_{TI}, Y_{PE}, Y_{LI}, and Y_{PT} represent the predicted churn risks using the single feature set of Customer Profiles, Term Importance, Phrase Embedding, Lexical Information, and Personality Traits accordingly.

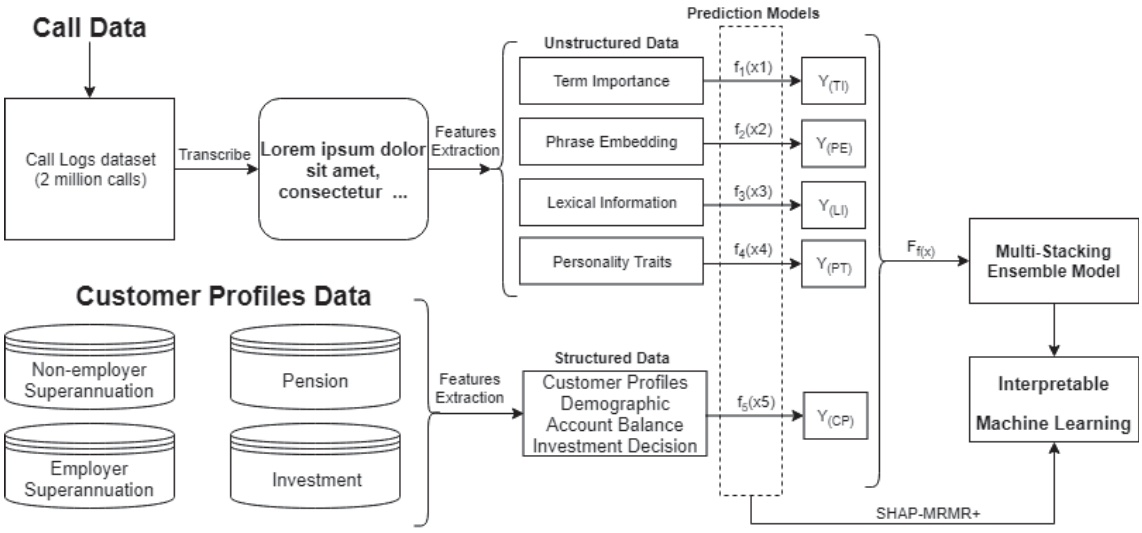


Figure 4.4 : Multi-stacking Ensemble Model and Interpretable Machine Learning

4.4 Interpretable Machine Learning for CRM strategies

From the multi-stacking ensemble prediction model, the predicted customer personalities and churn risks, we use interpretable machine learning with our SHAP-MRMR+ values to analyze the feature importance at three different levels: the whole sample, the customer segments and individual customer levels. These results help customer segmentation based on profiles and personality traits using Self-Organizing Maps (SOM) clustering algorithms. Based on the interpreted machine learning results, the company can develop the appropriate CRM strategies.

4.4.1 SHAP-MRMR+ for interpretable machine learning

SHAP (Lundberg and Lee 2017) is a unified model to interpret many machine learning models, which connects game theory with local explanations. It combines multiple methods into one consistent and locally accurate additive feature attribution with the expectation-based approach. Given a prediction $f(x)$ and $S \subseteq Z/\{i\}$ where Z is the set of all input features and M is the number of "interpretable" inputs, we calculate the Shapley values (Lundberg et al. 2018) as a weighted sum of the impact of each feature i added to the model, which is averaged over all possible orders of features:

$$\phi_i\left(f,x\right)=\sum_{S\subseteq Z/\{i\}}\frac{|S|!\left(M-S-1\right)!}{M!}[f_x(S\cup\{i\}-f_x(S)] \tag{4.13}$$

In this thesis, we propose a modified SHAP to interpret our machine learning model. Minimum Redundancy Maximum Relevance (mRMR) (Peng et al. 2005) is considered more powerful than the maximum relevance feature selection. It can select features that are mutually far away from each other but still have "high" correlation to the classification variables. We combine the Shapley value as in equation above and the positive mRMR value to compute the SHAP-MRMR+ as:

$$\text{SHAP}-\text{MRMR}_i(f,x)\ =\frac{1}{2}\left(\phi_i\left(f,x\right)+\sum_{i=1}^{N}\frac{|\text{MRMR}_i|+\text{MRMR}_i}{2N}\right) \tag{4.14}$$

where MRMR_k is the computed mRMR value. Our intuition is that the positive mRMR values indicate the most important features globally, which impacts might be minimized under the local Shapley calculation. By adding these values to the SHAP values, we can further highlight these features to support managerial decisions both globally and locally for every individual customer.

4.4.2 Customer segmentation with Personality

Our proposed model leverages the predicted customers' personalities and churn risks to construct suitable targeted marketing plans. First of all, we apply multi-filtering technique to identify the high-value customers with the highest likelihood to leave the company. We then segment this customer database into ten different categories based on whether they have high or low scores in each of the Big Five personality. It then combines with the marketing database to develop different retention strategies for each group. For example, if the customers are in the "High Openness" group, direct email marketing can be a cost-effective strategy to keep them stay with the company. On the opposite side, for customers in the "Low Openness" group, our model suggests other approaches such as promotional campaigns to engage them more. By segmenting customers and proposing different churn prevention strategies, we can improve both the effectiveness and efficiency of the CRM plans in terms of

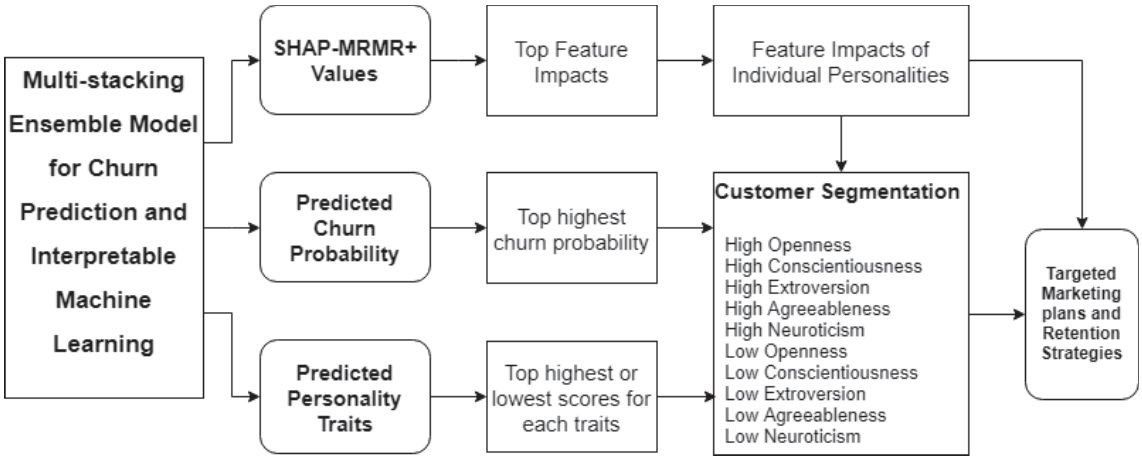


Figure 4.5 : Method for CRM strategies based on Personalities and Interpretable Machine Learning

timing, cost, and resources. Figure 4.5 illustrates the process of generating the managerial implications with combined customer segmentation using Personality Traits and interpretable machine learning with SHAP-MRMR+.

4.4.3 Customer segmentation with SOM

After the churn probabilities are obtained for individuals, we propose to use Self-Organizing Maps (SOM) (Kohonen 2013) to segment the data. The main purpose of using SOM is to map the high-dimensional vectors onto the 2-D space, which is naturally visualized. The visualization capability provides easy-to-understand knowledge and is a key advantage comparing with other alternatives. To be specific, the SOM is a centroid-based clustering algorithm that constructs a Kohonen layer with a fixed number of centroids on it. The learning process starts with assigning random weights to the centroids and gradually allocate each data record to the centroid that is the most similar. By allocating each data record, the weights of the corresponding centroid update accordingly towards that data record. This learning process repeats until the Kohonen layer becomes stable, i.e., data records are no longer moving around centroids. Once the Kohonen layer has been constructed, each centroid has a profile representing a group of similar data records, and each variable can be visualized by investigating the values distributed across the map. In

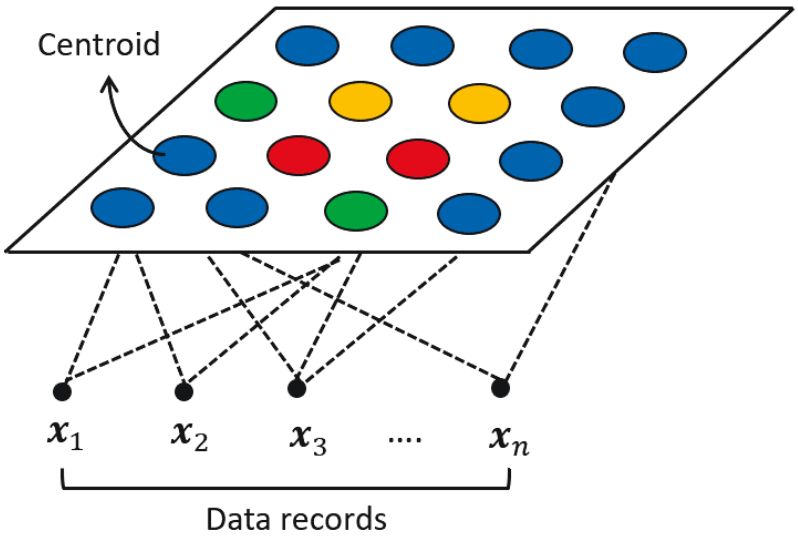


Figure 4.6 : Example of Kohonen Layer

practice, each centroid is colored based on the values of the variables. An example of the Kohonen layer is showed in Figure 4.6.

4.5 Experiment

4.5.1 Datasets

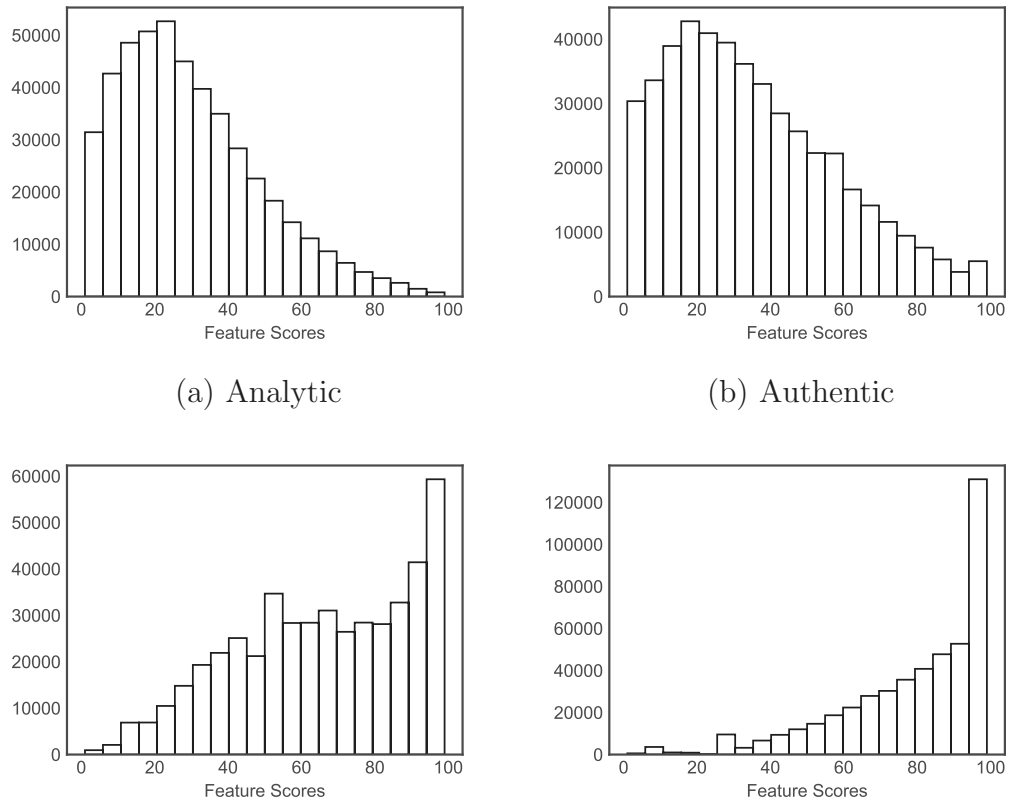
Customer Call Logs dataset (CL)

We obtained the call center data directly from an Australian financial services company. The firm has recorded over three million customer calls to their hotline for the period between 04/2011 and 04/2017. The company also outsourced the text transcriptions of the calls to a third-party service provider. For each call, conversational dialogues are separated into two monologues of the call center agent and the customer respectively. Almost two million calls can be identified using unique customer ID for mapping back to the existing client database.

Additionally, some customers call to end the services and specifically use related terms such as “close account” or “terminate account”. We exclude these calls recorded within 14 days range of an account closing request by customers to reduce bias prediction and false positive prediction. Our model uses the text transcrip-

tions of only the identified customer calls to proceed further. The final CL dataset has also been processed to remove all private names, account identification terms and sensitive financial information to ensure anonymity and data protection for the customers. Each call in our dataset contains on average 314 words in length, with the median word count is 240 words. We also look particularly at features that can provide insights into customers' personalities. The histograms of some LIWC 2015 main features suggest that the majority of the customers are not very analytical and authentic, but quite confident and emotional in their speaking tone towards the call center agents (see Figure 4.7). These features are extremely meaningful for us to distinguish customers and identify their individual personality traits.

Figure 4.7 : Histograms of LIWC 2015 main features



(a) Analytic

(b) Authentic

(c) Clout

(d) Tone

First Impressions dataset (FI)

For Personality Traits, we use a public dataset called First Impression (Ponce-López et al. 2016) to train model and then transfer learning on our private CL datasets. Though there are multiple public datasets, we focus on FI due to its nature of being spoken text, which is suitable for our context. The dataset contains ten thousand videos of people speaking directly to the camera. The mean duration of each video is about 15 seconds, 43 words spoken per clip on average. $435,984$ words ($183,861$ non-stop words) has been transcribed for all ten thousand clips. Psychology experts have labeled each video using the scoring system ranged [0,1] from the Big Five personality model.

Customer Profiles dataset (CP)

We have in total four different Customer Profiles databases provided by an Australian financial services company. The first one is Employer superannuation dataset containing institutional customers who are employers of other employee accounts. Therefore, it has more features related to the demographic data and sub-account information. We have 133 features in this dataset in comparison to 106 features for other datasets. Non-employer superannuation dataset contains individual customers who open the accounts without reference from their employers. Pension dataset includes individual superannuation customers who have reached their retirement age or opted for early retirement and their superannuation accounts become pension accounts. The three datasets are quite similar in terms of financial behaviors. The investment dataset is a little bit different with more features related to investment decisions. The distinguishable characteristics of the four datasets provide us with a better experimental set up to prove the generalizability of our methodology, that is, it would work for different types of data and customers not only in the financial services sector but also for other industry as well.

Table 4.4 presents the sizes and basic statistics of the four databases. They include several demographic features (e.g., sex, age, address) and financial performance features (e.g. annual investment rate, total balance). Combining with the CL

Table 4.4 : Statistics of the merged DAP datasets

Datasets	Number of Features	Total	Churn	Churn (%)
Non-employer Super	106	49,996	2,067	4.13%
Employer Super	133	36,608	1,959	5.35%
Pension	106	28,046	582	2.08%
Investment	106	23,422	2,098	8.96%

dataset, the merged datasets in Table 4 consist of only customers who have made phone calls to the company hotlines since 2014. The churn decisions of these customers have been labeled as 1 for churn and 0 for not churn. The binary indicator serves as the ground truth Y for testing and evaluating our churn forecast model.

We also perform correlation analysis on all CP basis features using the Pearson's r Pearson (1895), the Spearman's ρ Spearman (1904) and the Kendall's τ Kendall (1938) correlations. This is because we have a various combination set of basis features, ranging from numeric to categorical variables with different distributions, using multiple correlation coefficients would provide an unbiased analysis.

- Pearson's r: It is the most popular measure of the degree of relationship between the two linearly related features. The point-biserial correlation is calculated with the Pearson's r formula except that one of the features is dichotomous. Usually, the Pearson's r is obtained via a Least-Squares fit and a value of 1 indicates a perfect positive linear correlation, -1 is the negative one, and 0 represents no correlation between features. The following equation is used to compute the value r of Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4.15}$$

  where n represents the total number of clients in the dataset, x_i and y_i represents the input feature x and churn label y of customer i respectively, $\overline{x}$ and $\overline{y}$ represent the average values of input feature x and churn label y accordingly.

- Spearman's ρ: This is a widely-used non-parametric test measuring the degree of association between two separated variables. The Spearman ρ test is a rank-based version of Pearson's correlation coefficient, which can be used for features that are not Gaussian-distributed and have a non-linear relationship. Also, its use is not only restricted to continuous data, but can also be used in analyses of ordinal attributes. The value ρ of Spearman rank correlation is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.16}$$

  where n represents the total number of clients, while d_i represents the pairwise distances of the ranks between input feature x_i and churn label y_i of customer i.

- Kendall's τ: The non-parametric test measures the degree of a monotone correlation and the strength of dependence between two ranked features, which makes the correlation analysis more feasible for non-Gaussian distributed data. Kendall's τ can be computed for continuous as well as ordinal data. Roughly speaking, it differentiates itself from Spearman's rho by having stronger penalization of non-sequential (in context of the ranked features) dislocations. If we consider two vectors of feature x and churn label y, where the total number of clients is n, we know that the total pairings number is $n(n-1)/2$. The following equation is used to compute the value τ of Kendall rank correlation:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \tag{4.17}$$

  where n_c represents the number of concordant pairs and n_d represents the number of discordant pairs when comparing the ranked values of input feature x and churn label y for all customers in the dataset.

The result in Table 5.8 presents the top five correlated features using each analysis strategy. As we can see, these features are account balance or variables related to

the outflow of investment (e.g. outflow recency, outflow frequency, outflow amount, outflow ratio, etc.). This general financial behavior is predictable with common features in financial services industry. However, the correlation coefficients of all the basic features are really low. The scores range is from -0.085 to 0.065, while meaningful correlated features should have scores higher than 0.5 or lower than -0.5. Therefore, we believe the incorporation of text features would help the prediction accuracy even further.

Table 4.5 : Correlation analysis of basic features and customer churn label

Pearson's r		Kendall's tau		Spearman's rho	
Features	Scores	Features	Scores	Features	Scores
Outflow recency	0.0654	Outflow recency	0.0609	Account balance	−0.0851
Call recency	0.0481	Account Balance	−0.0695	Outflow recency	0.0649
Account growth	0.0471	Outflow frequency	−0.0579	Outflow frequency	−0.0607
Number of options	−0.0436	Outflow amount	−0.0552	Outflow amount	−0.0590
Saving plan N	0.0430	Outflow ratio	−0.0517	Outflow Ratio	−0.0552

4.5.2 Baselines and Evaluation Metrics

In our case, the company is specifically interested in identifying customers with the highest churn risks. They intend to target at the top 30% of clients with high churn risk. Therefore, we build our model to predict churn risk instead of binary classification. Our predicted churn p is ranging from 0 to 1 as the probability for churn. The experiments run with 10-fold cross-validation and the performance results are averaged for all folds. We use Area-Under-the-Curve (AUC) scores as our evaluation metrics and visualize the final performance assessment in the Receiver Operating Characteristic (ROC) curve (Metz 1978). The AUC score is calculated as follow:

$$AUC = \frac{1}{N_1 N_0} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \mathbf{1}_{(p_i > p_j)} \tag{4.18}$$

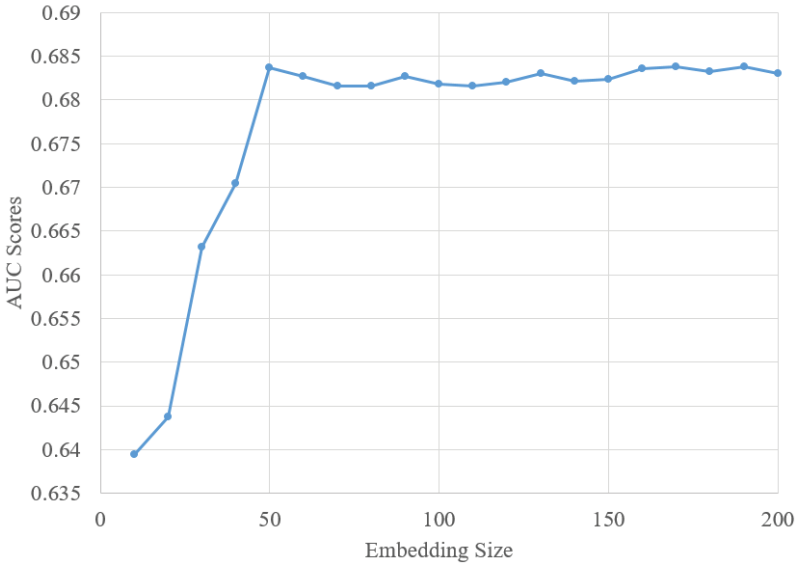


Figure 4.8 : Correlation between Word Embedding Size and AUC Scores

where N_1 represents the total number of churn customers (true label 1), N_0 represents the total number of not churn customers (true label 0), p_i and p_j represents the probability scores assigned by our model to each label respectively. **1** is the indicator function with value equals 1 if $p_i > p_j$ and 0 otherwise.

4.6 Result and Discussion

4.6.1 Empirical Result

First of all, we designed a preliminary experiment to test the correlation between the word embedding size and the AUC scores using 10-fold cross-validation with Logistic Regression algorithm on Investment dataset. The result in Figure 4.8 shows that the prediction accuracy did not improve significantly when we increase the size over 50. Therefore, we only use 50 Word2Vec features in our prediction model. Besides that, we run the preliminary experiments for all the four datasets using XgBoost, and results remain very similar in that it is sufficient to use 50 Word2Vec features in the prediction model.

We test the text features individually to confirm the effectiveness of unstructured data as input for churn prediction. We build separate models with XGB algorithm

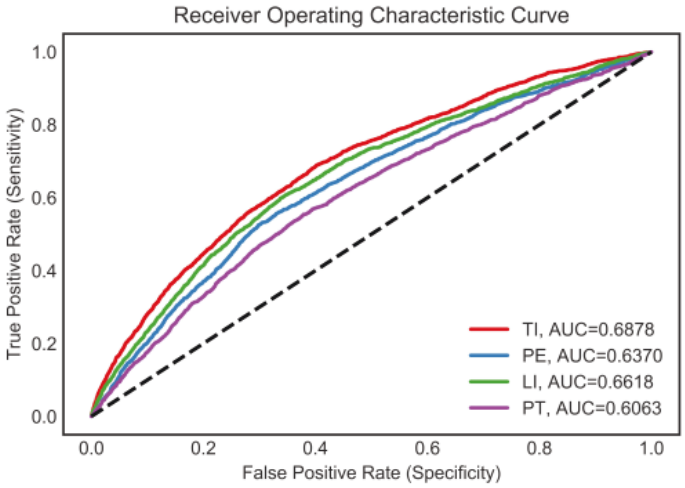


Figure 4.9 : Text Features Churn Prediction Model for Investment dataset

using Term Importance (TI), Phrase Embedding (PE), Lexical Information (LI) and Personality Traits (PT) text features. The result in Figure 4.9 shows that multiple text mining techniques can capture different information. It is also noticed that models with TI and LI features outperformed others, indicating the simpler text mining techniques might yield a sufficient result already. In the case of PT, the model can achieve comparable results even with only five features. This shows that personalities are good indicators for evaluating client churn risk. Furthermore, we believe by stacking all approaches together, the features complement each other to achieve an even better result in the final churn prediction model with AUC score 0.8124 (see Table 5).

To evaluate the performance of text mining techniques, we test separate churn forecast models using different feature sets: (1) using only structured data from the Customer Profiles (CP) datasets, (2) using only unstructured data from the Call Logs (CL) datasets, (3) using the multi-stacking ensemble feature set to build a churn forecast model. The predicted churn risks are compared against the ground truth labels to compute the AUC scores as showed in Table 4.6.

The results in Table 4.6 show that our proposed models are effective for all four datasets with different features. Especially in the case of investment data,

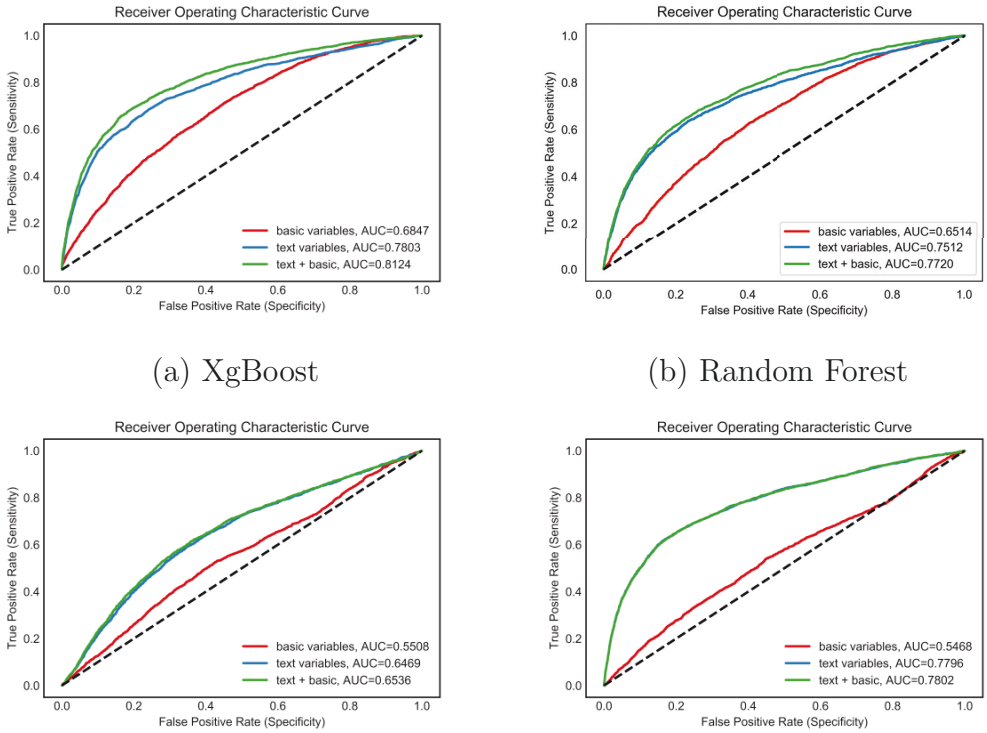
Table 4.6 : AUC results on the models' prediction accuracy

	AUC scores			
Models	Basic	Text	Stacked	Improvement
		Non-employer Super		
XGB	0.7261	0.7229	0.7818	5.56%
RF	0.7056	0.7323	0.7572	5.16%
LR	0.6024	0.7226	0.7329	13.05%
NB	0.6094	0.6451	0.6526	4.32%
		Employer Super		
XGB	0.7980	0.7562	0.8378	3.98%
RF	0.7746	0.7465	0.7930	1.84%
LR	0.6508	0.7572	0.7699	11.91%
NB	0.6756	0.6402	0.6756	0%
		Pension		
XGB	0.7843	0.7245	0.8220	3.77%
RF	0.6932	0.7518	0.7723	7.91%
LR	0.6567	0.7143	0.7336	7.69%
NB	0.6500	0.6893	0.6985	4.85%
		Investment		
XGB	0.6847	0.7803	0.8124	12.77%
RF	0.6514	0.7512	0.7720	12.06%
LR	0.5468	0.7796	0.7802	23.34%
NB	0.5508	0.6469	0.6536	10.28%

the prediction accuracy has increased substantially with an average AUC scores improvement of 15.46%. Due to the high competition in the financial services industry, "Investment" clients normally have lower levels of attachment or loyalty to the firm and churn for other companies to gain more financial benefits. Table 4.4 confirms that the customers in "Investment" dataset are the most likely to churn. It is thus particularly valuable for the company to retain "Investment" customers as they are generating higher revenue than other types of clients. Thereafter, we mainly focus on the analysis for these customers. Based on the predicted churn risk, we could help the company develop appropriate retention strategies targeting especially at the highest churn risk clients in the "Investment" datasets. The ROC curve plots indicate that using the multi-stacking feature set can further lift the performance of the forecast model compared to the one using only the structured data. The results from models with the four different classification algorithms confirmed the effectiveness of our unstructured data mining approach. Figure 4.10 visualizes the results using ROC curves of Investment model for all algorithms. The prediction performance by merely using the structured data is not good, but using the various textual features or the combination of using both structured and textual features could achieve much better prediction performance.

Logistic regression is the best performing algorithms in terms of AUC score improvement. For the Investment dataset, our forecast models achieved a significant increase of up to 23.34% when logistic regression is used. The churn prediction models using basic features cannot perform well in this Investment dataset, and the text models can provide significantly more useful information. The results prove that our advanced multi-stacking features methodology is suitable for customer churn forecast model in general. The model with the XGB algorithm achieved the best AUC score of 0.8124. Therefore, we use the predicted churn risk from this model to perform further customer segmentation and analysis for retention strategies.

Figure 4.10 : Multi-stacking Ensemble Churn Prediction Model for Investment dataset



(a) XgBoost

(b) Random Forest

(c) Gaussian Naive Bayes

(d) Logistic Regression

4.6.2 Robustness Analysis

Imbalance learning test

The focal Investment dataset is imbalanced with the number of churn customers are accounted for about 10% in total. In our methodology, we used the raw data as input for the prediction model and defined the class weights accordingly in the machine learning algorithm XgBoost. We test the efficiency of this approach by comparing the prediction results with other commonly used methods in handling imbalance data, e.g. over-sampling and under-sampling methods such as SMOTE, SMOTEENN, SMOTETomek, SVM-SMOTE and Borderline-SMOTE (Lemaître et al. 2017). The test set ratio for this experiment is 20%, where we use 80% of the data for training and for SMOTE methods, and test the prediction models on the out-of-bag 20% of data. The results in Figure 4.11 showed that our raw data with

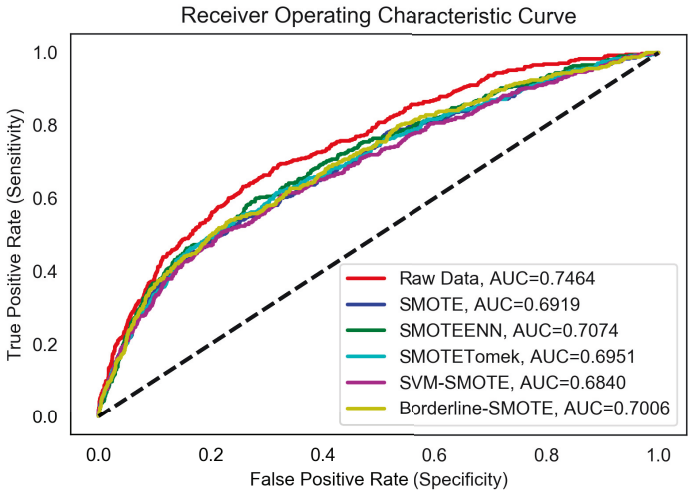


Figure 4.11 : ROC Curves of prediction models compared with SMOTE methods

class weights achieve higher AUC scores. Since XgBoost algorithm already assigns different weights to handle the imbalance class, using SMOTE methods to generate synthetic data might inject noise for the machine learning algorithm. Due to the nature of our application, the business focuses the top of the churn customer ranking list, thus AUC is a suitable choice for this study. However, the prediction model with SMOTE methods might achieve better results in other metrics such as F1 score.

Test of statistical significance

We performed the pair-wise t-test to check the performance of our prediction model using XgBoost compared to the other three algorithms, including Gaussian Naïve Bayes, Logistic Regression, and Random Forest. We conducted the test on the overall "AUC", "accuracy", "precision", "recall" and "F1" scores, and found that our model significantly outperforms other models according to all of the metrics. Table 4.7 reports the p-values of the pair-wise t-test, where p-value lower than 0.05 is considered as statistically significant at 95% confidence level. Particularly the results of the "AUC" t-test are significant in our final model, which suits the company's interest in the probability of customer churn risk.

Table 4.7 : Pair-wise t-test on model performance

	Gaussian Naive Bayes	Logistic Regression	Random Forest
AUC	<0.00001	<0.00001	0.00003
Accuracy	<0.00001	0.00209	0.00091
Precision	0.00189	0.00040	0.00344
Recall	<0.00001	0.00026	<0.00001
F1	<0.00001	0.00025	<0.00001

Hyper-parameters tuning

We tested the performance of our prediction model under various hyper-parameter settings. The list of tested hyper-parameters were: minimum sum of instance weight needed in a child (min_child_weight), minimum loss reduction required to make a further partition on a leaf node of the tree (gamma), subsample ratio of the training instances (subsample), subsample ratio of columns when constructing each tree (colsample_bytree), maximum depth of a tree (max_depth), and L1 regularization term on weights (reg_alpha). The results in Table 4.8 show the AUC scores for models with different settings and the final column indicates the best setting which we used in our final prediction model. Overall, the AUC scores did not fluctuate much under different hyper-parameter settings. This suggests that our model is robust.

Interpreting Machine Learning

We first compare our SHAP-MRMR+ with the original SHAP value for all labeled churn customers to evaluate the feature importance at a global level for the whole dataset. Figure 4.12 showed the difference in the evaluation of feature impacts between SHAP and SHAP-MRMR+ for the top 10 features. As we can see, the standard SHAP values of "WC_liwc" are scattered in four dots in extremely high values. These are the edge cases where the calls are much longer and the numbers of word count in the transcript are higher. This might introduce bias in interpreting our model. SHAP-MRMR+ lowers the ranking of this kind of features

Table 4.8 : AUC results with different hyper-parameter settings

Hyper-parameter	Tested Settings	AUC Scores			Final setting
		1	2	3	
min_child_weight	[**1**, 5, 10]	**0.7312**	0.7293	0.7257	1
gamma	[**0.5**, 1, 2]	**0.7299**	0.7291	0.7211	0.5
subsample	[0.6, 0.8, **1**]	0.7263	0.7290	**0.7312**	1
colsample_bytree	[**0.6**, 0.8, 1]	**0.7313**	0.7307	0.7312	0.6
max_depth	[3, 4, **5**]	0.7312	0.7361	**0.7362**	5
reg_alpha	[0.001, **0.01**, 0.1]	0.7301	**0.7302**	0.7288	0.01

Figure 4.12 : Compare feature impacts with SHAP and SHAP-MRMR+

and therefore, it is better than the original SHAP approach in terms of interpreting the prediction model and explaining customer insights. The SHAP-MRMR+ gives more weights to globally important features such as “account balance” and “outflow recency” from basic customer profile feature set. Hence, these features have higher impact ranks compared to those in the SHAP values ranking. Our SHAP-MRMR+ gives a lower rank for less meaningful text features such as word count (“WC_liwc”) and remove it from the top 10. According to the results of SHAP-MRMR+, seven out of the ten most impactful features are textual features from LIWC and TF-IDF feature set, suggesting that textual information is very useful for churn prediction model and needs to be taken into account.

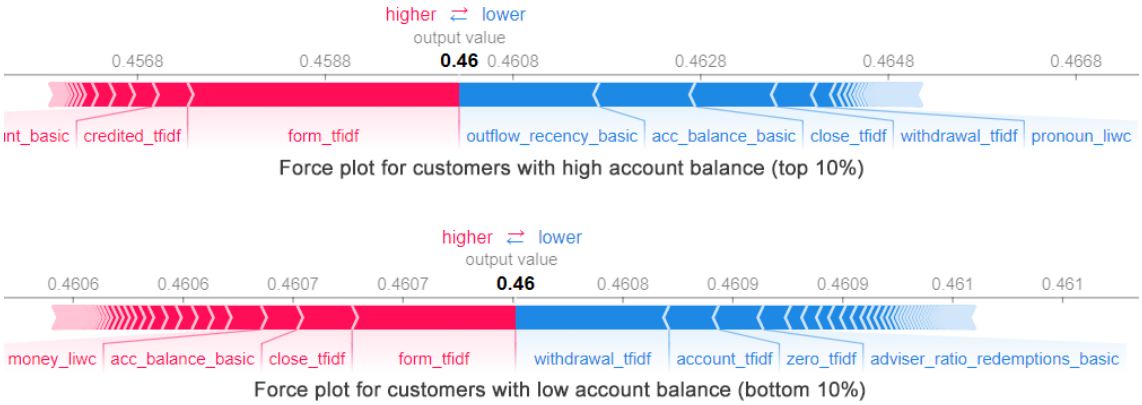


Figure 4.13 : Compare feature impacts on churn prediction for customer segments with high/low account balance

In order to visualize SHAP-MRMR+ and interpret our model for each customer segment, we use the force plot (Lundberg et al. 2018) where the red color indicates important features which can help to increase the prediction accuracy and the blue color indicates less useful features which might add noise and decrease model value. Figure 4.13 shows the results on two customer segments: high account balance (top 10%) and low account balance (bottom 10%). Our model suggests that the most impactful features to predict churn for customers with a high account balance are "form" and "credited" from TF-IDF feature set. We can infer from this result that the churn risks of "richer" investors are not dependable on their "outflow recency" and "account balance" (not very impactful features in blue color). However, they might be more concerned with the investment process and probably ask more about "form" and "credited". Meanwhile, the churn risks of investors with a low account balance are still predictable by "account balance" and related text features such as "close" from TF-IDF and "money" from LIWC feature sets. This insightful finding is not easy to identify without using the interpretable machine learning technique like SHAP-MRMR+ which we use in our research.

With the increase of churn risk prediction accuracy, the company could potentially save around five million dollars in annual revenue just by targeting the top 10% of customers with high churn risks. Therefore, our model incorporates the multi-

filtering technique to perform customer segmentation based on the probabilities of their churn decision.

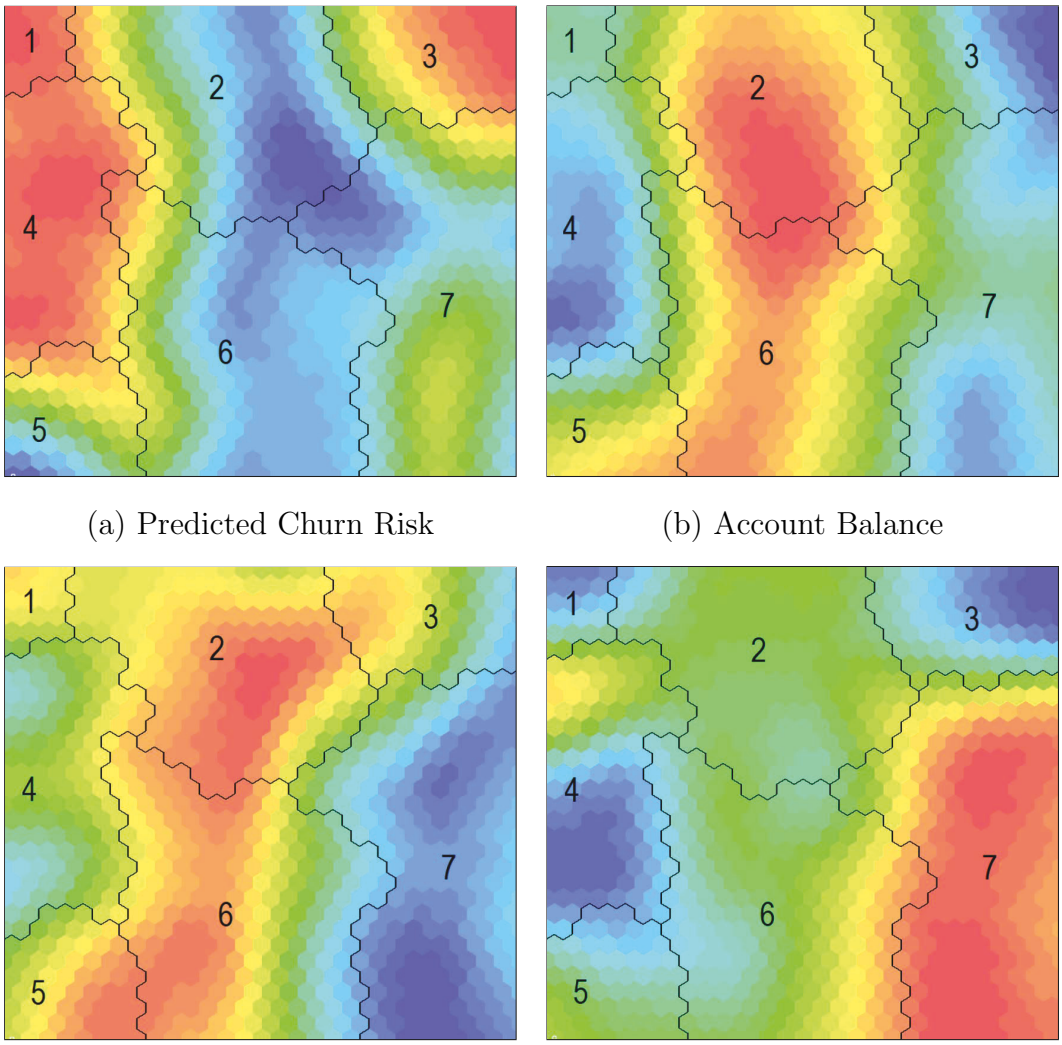
We apply SOM to cluster customer profiles to further differentiate their characteristics. The result on Employer Superannuation dataset in Figure 4.14 provide us with interesting information based on the basic account profiles. Customers in segments 1, 3 and 4 are generally having a higher churn probability. They often have low account balance with negative balance change, which means they often withdraw their funds. The lower churn risk customers in segments 2 and 6 also have a higher account tenure, which means they are long-term client and unlikely to churn. These findings are aligned with current research in financial customer data analytics, which proves our predicted churn risk are highly accurate.

To further investigate customer segmentation using unstructured information from the call logs, we also perform SOM clustering for Non-Employer Superannuation dataset. According to the heat maps in Figure 4.15, there is no clear correlation in the segments between churn risk and account balance. This result partly explains the lower AUC scores for models using only basic variables. The results with unstructured data, however, show that clients with high churn risk on the half left of the maps (segments 1, 2, 4 and 5) call in more frequently and have higher conscientiousness. This proves our approach is effective in revealing special customer characteristics that are significantly useful in retention strategies.

We apply different approaches to segmenting customers based on their profiles and personality traits. We conduct a descriptive statistical analysis on customer groups with the top 10% highest or lowest scores for each personality trait. The churn risks in Table 4.9 show that "Investment" customers with lower "Conscientiousness", "Agreeableness" and "Neuroticism" are more likely to leave the company.

This finding in Table 4.9 aligns with previous research on personality traits of the retail customers (Castillo 2017) which shows that "Conscientiousness", "Agreeableness" and "Neuroticism" are significantly correlated with customer empowerment and satisfaction. We then use this finding to perform SOM clustering to segment customers based on the predicted churn risk and the three mentioned personality traits

Figure 4.14 : SOM Segments on Employer Superannuation Dataset



(a) Predicted Churn Risk

(b) Account Balance

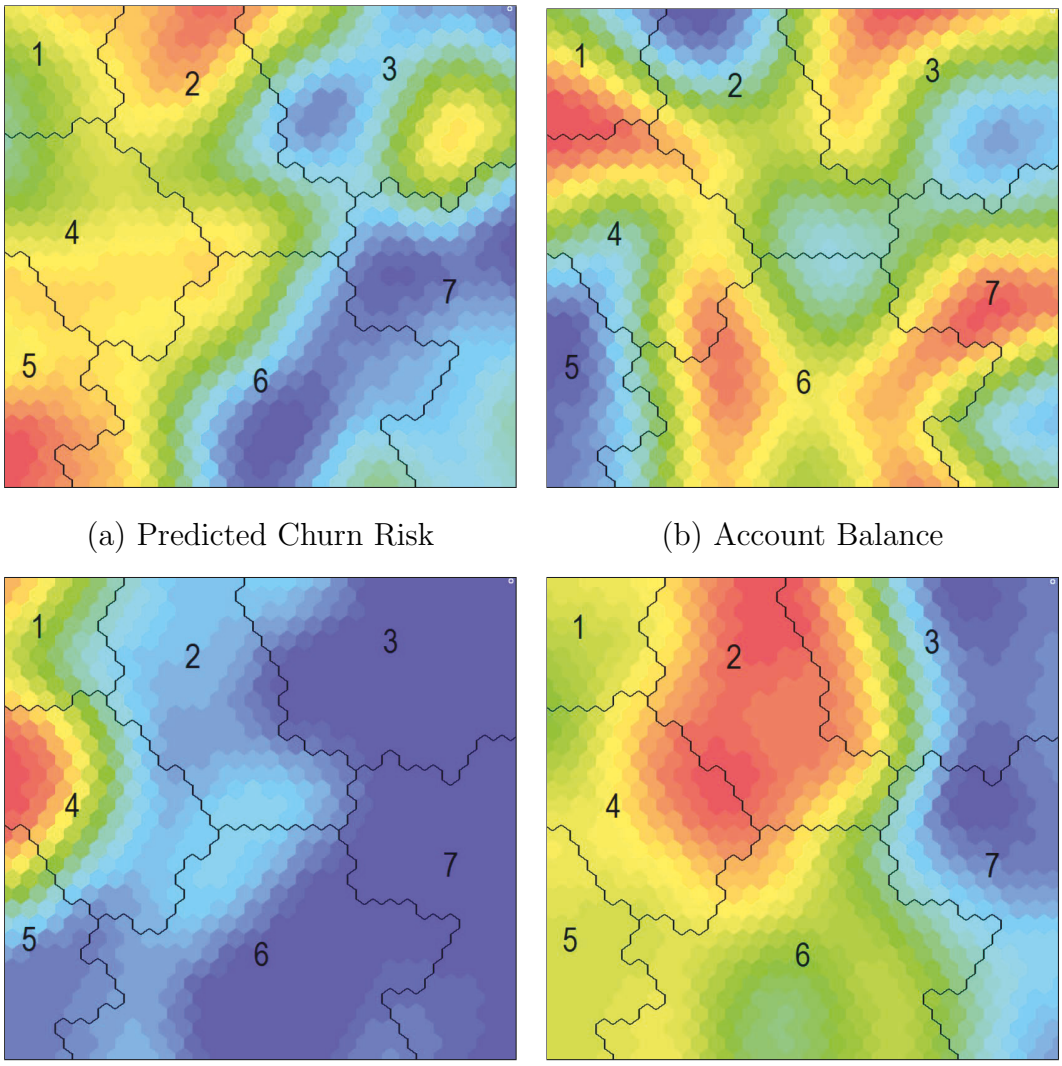
(c) Account Tenure

(d) Account Balance Change

to further test their correlation. We illustrate the SOM clustering results using the heat maps, in which the bright red color spots indicate customers with higher churn risk and personality rankings, and the cooler colors green and blue spots indicate lower churn risk and personality rankings accordingly. The heat maps with eight different SOM segments in Figure 4.16 have confirmed our hypothesis regarding the relationship between personalities and customer churn decision in our context.

The result aligns with previous studies in a similar context for bank customers (Al-Hawari 2015). Particularly for segment number 6, customers with higher churn

Figure 4.15 : SOM Segments on Non-Employer Superannuation Dataset



(a) Predicted Churn Risk

(b) Account Balance

(c) Call Frequency

(d) Conscientiousness

risk tend to have lower ranks on the three personalities “Conscientiousness”, “Agreeableness”, and “Neuroticism”. These findings generate meaningful insights for the financial services firm to characterize the personality of the customers with high churn risks. They can take risks in investment but are uncooperative and might not listen to others. Being independent in their own thinking and careless with their actions, their response to most marketing campaigns are often unpredictable and their churn decisions are also spontaneous.

Moreover, our interpretable machine learning approach can be used to explain

Table 4.9 : Predicted churn risk (%) for Customers with Top and Bottom 10% Personality Rank

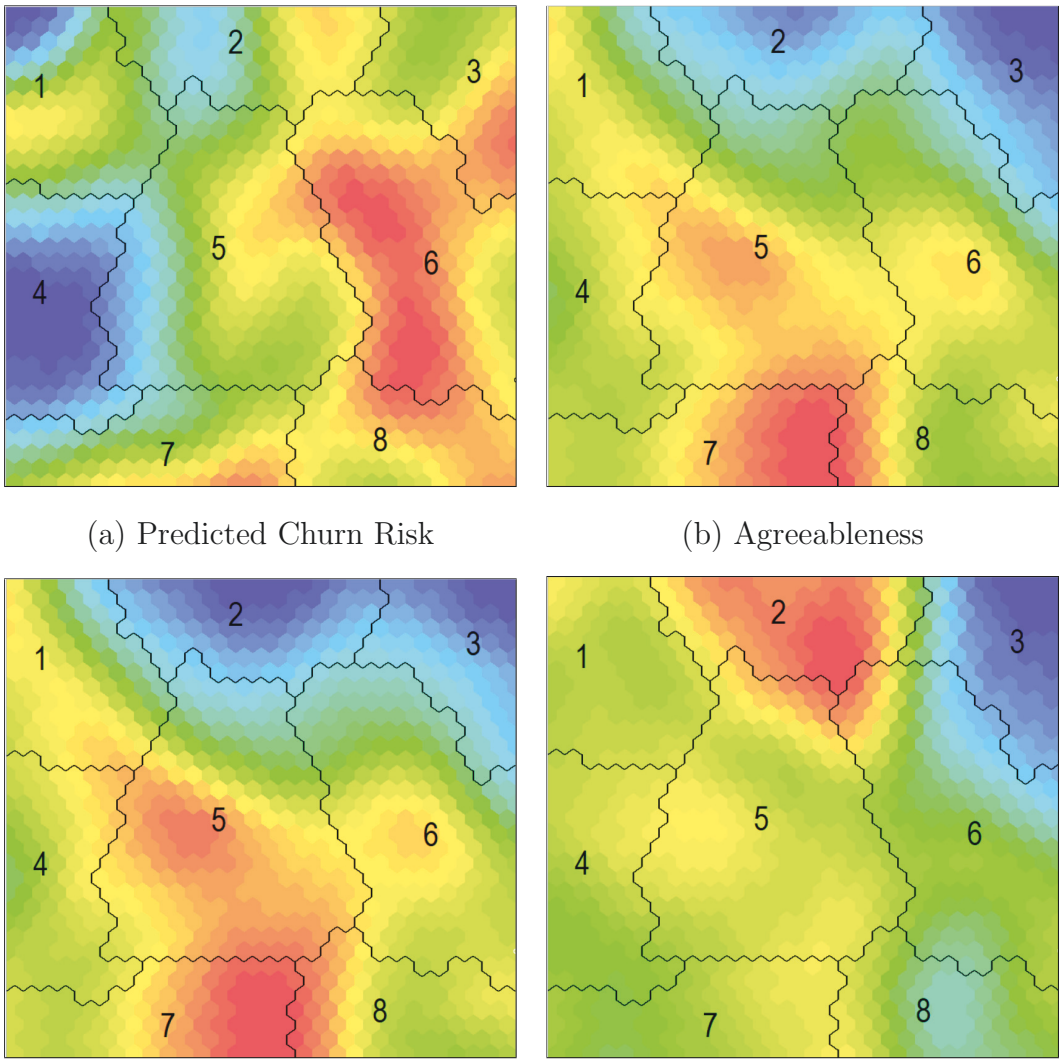
Segments	Bottom 10%	Top 10%
Openness	7.91	7.22
Conscientiousness	8.94	6.61
Extroversion	7.91	7.34
Agreeableness	9.36	7.67
Neuroticism	8.22	7.16

the features at the individual customer level. Particularly in our case, the financial services firm is interested in learning the personality traits of each churn customer and leveraging that in their retention strategies. Figure 4.17 showed the individual force plot for two churn customer A and B. "Conscientiousness", "Agreeableness", and "Neuroticism" are impactful predictors for customer A as in our SOM analysis. However, for customer B, "Neuroticism" alone is a strong enough predictor for churn risk. These customers with such a special personality are harder for the financial service firms to retain using the same retention strategies as other customer segments.

Proposed Marketing Strategies

Based on these findings, we can propose appropriate targeted marketing strategies aiming at those customers with high churn risks, especially with their individual personality profiles, to increase the engagement and loyalty of these customers. We analyze some commonly used retention strategies and adapt them to suit our customer personality profiles in order to maximize the effectiveness of each marketing campaign. Moreover, from all the personality analysis, we can further infer acquisition strategies to become more appealing the new customers with different characteristics as well. We separate these strategies into two categories for better analysis: direct and indirect marketing.

Figure 4.16 : SOM Segments with Personalities on Investment Dataset



(a) Predicted Churn Risk

(b) Agreeableness

(c) Neuroticism

(d) Conscientiousness

Direct Marketing

As the cost of direct marketing is much lower than the indirect one, companies tend to use these strategies more often. Research has showed that customers are loyal to brands that have similar personality traits to themselves (Yao et al. 2015). Based on this finding, we can tailor a marketing campaign that is personalized to their personality preferences.

- Acquisition: Regarding acquisition strategies, the first impression is considered

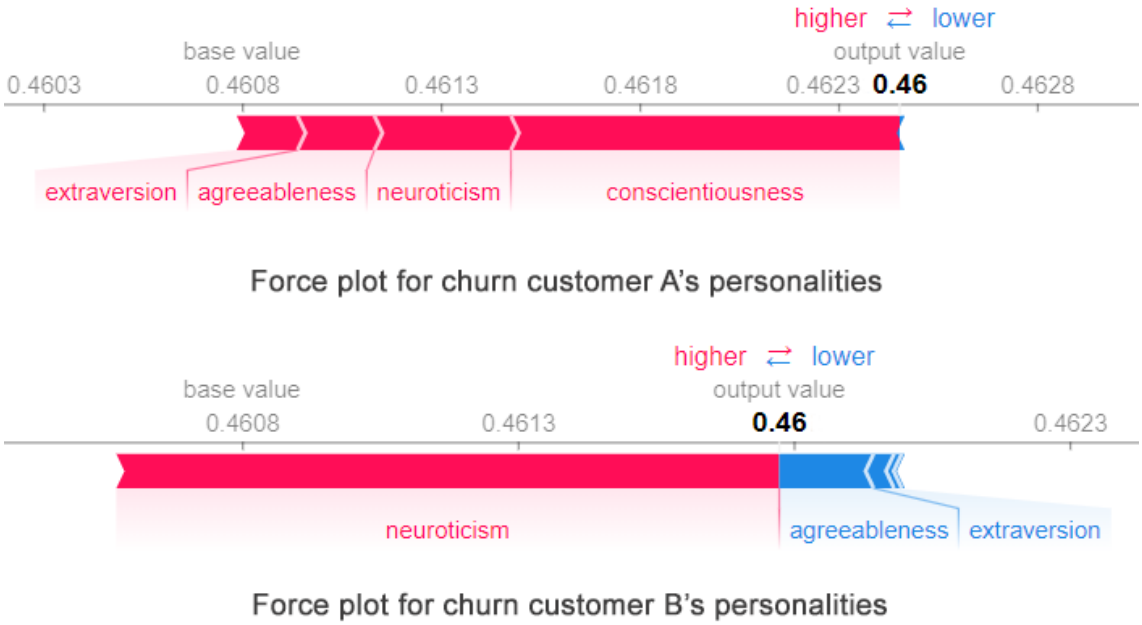


Figure 4.17 : Compare personality impacts on churn prediction for customer A and customer B

as the most important point. Companies often invested in different advertising content to target different customer segments. The image and text used in any communication channel can be personalized to represent their personality, e.g., people with high Conscientiousness scores would prefer to look at the advertisement with a portrait of a person who looks calm, confident and self-loving.

- Retention: On the retention side, these high Conscientiousness customers are independent thinkers and do not like to take advice from others. The retention marketing strategy is to made them feel like they have a choosing power, by providing them with as many investment options as possible. For these customers who also haves high Openness, the advisor can also suggest the customers to invest in emerging tech stocks that are risky but have more familiar products with them, rather than some boring government bonds. As they are self-satisfied and self-indulgent, they can also be more responsive to

the reward-based loyalty program, e.g., special fee reduction for investments on their birthdays.

Indirect Marketing

Indirect marketing is often costly but more effective for consumers with a low rank of "Conscientiousness", "Agreeableness", and "Openness". Unfortunately, in our case, the "Agreeableness" personality ranks of customers with high churn risks are generally lower, and therefore they may be less likely to respond to some direct marketing strategies, including telemarketing and commercial advertising. Therefore, indirect marketing strategies would be a better approach for these customers.

- Acquisition: One of the most popular indirect marketing strategies is word-of-mouth. Research has showed that consumers tend to take recommendations from others with the same personalities than with opposite characteristics (Adamopoulos et al. 2018). These findings help us suggest appropriate strategies for relationship marketing via word-of-mouth. We can suggest the company to identify long-term customers who have similar personality profiles and invite them to become brand ambassadors. With an incentive scheme for motivation, they can help spread the words about the firm to friends and family who have similar personalities. This strategy will be not only beneficial for existing customers, which will improve the retention rate, but also can potentially attract new investors for the company.

- Retention: Indirect marketing via social media channels is one of the most cost-effective indirect marketing approaches in recent years. Knowing a customer trait based on text mining from their social media posts will be a huge advantage for customer services. For example, a marketing agent with the same low level of "Neuroticism" will reply to calls from customers with this trait. As people are more comfortable talking to others with similar traits who use the same verbal expressions, the customer satisfaction will be higher, and their brand engagement and loyalty might increase even further in this case.

4.6.3 Discussion

Overall, the results from the experiments show that the unstructured data, e.g., customer call logs, can be used to generate meaningful insights, and interpretable machine learning should be utilized in all types of customer information systems. The Superannuation firm can potentially save millions of dollars in profit by early identifying high churn risk client and personalizing marketing strategies to achieve a customer retention rate.

This research still has some limitations due to the low quality transcriptions of the call logs by a third-party company. With potential use of better transcription and also sound features from original calls, the prediction accuracy could be significantly higher than the current state-of-art approaches. The experiment results further imply that our proposed model can be generalized to integrate other types of unstructured data, such as social media interactions.

This is the principal investigate inside the financial services industry in Australia to consolidate both structured and unstructured data to solve problems in customer retention. The empirical experiments demonstrate that unstructured information contains useful insights which can enhance the precision of the churn risk forecasting on various client datasets. It helps the firm design better-targeted marketing campaigns for both customer acquisition and retention strategies and save million dollars in profits.

In future work, more longitudinal research should be conducted to further evaluate the performance of the proposed approach. A quantitative analysis of customer personalities might be an efficient evaluation of the transfer learning model of personality mining. Moreover, customer survey should be used as an empirical method gain more insights on different investment attitudes of varied customer segments, e.g. high or low account balance customers, to understand their differences and the effectiveness of our proposed strategies.

In conclusion, the customer analytics field propels giving organizations more intends to comprehend their client on a more extensive quantitative premise utilizing

distinctive sort of information as opposed to customary ones. Inside the extent of this research, we have adopted an unconventional strategy when utilizing unstructured information from client call logs to construct a multi-stacking ensemble churn prediction model and segmenting customer using interpretable machine learning approach on their profiles and personalities.

The research also establishes an underlying framework for the utilization of different data types and interpretable machine learning in other information and business intelligence systems. The four datasets with distinguished customer profiles demonstrate our proposed approach would work for different customers, and the method can be generalized for other industries.

Chapter 5

Data Mining for Socially Responsible Investment

5.1 Background and Motivation

Financial investment has been a common mean for individuals and institutions to growth wealth. Together with rising concerns regarding environmental and social topics (e.g. climate change or equal pay), investors have been interested with not only the financial performance of the stocks but also the corporate responsibility aspect of the companies. Socially responsible or sustainable investment, first introduced by Sant (1971), has become a new investing trend in recent years with an increasing demand for incorporating Environmental, Social, and Governance (ESG) ratings into their financial portfolio management (Van Duuren et al. 2016).

In 2018, $11.6 trillion of all managed financial assets, $1 out of every $4 invested in the United States, were under ESG-focused investment funds, skyrocketing from only $3 trillion in 2010 (Connaker and Madsbjerg 2019). Many investors choose to invest their money through an ESG-focused fund, taking a higher risk in order to gain a better profit in terms of both financial and ESG performance. They expect the funds to analyze the corporate social responsibility of the invested stocks to ensure a good impact investment. However, even with the availability of ESG ratings from third-party agencies, the fund managers often make investment decisions mainly based on qualitative analytics. A more data-driven quantitative method is in need to help improve the analysis process for better fund management strategies.

In order to evaluate a company performance from ESG perspectives, agencies or fund managers have to manually read the corporate social responsibility reports, press releases, news and other media sources to rate the company in multiple categories, (e.g. green house gases emissions and water management, employee equality, board diversity) including their controversies. These metrics will then be combined

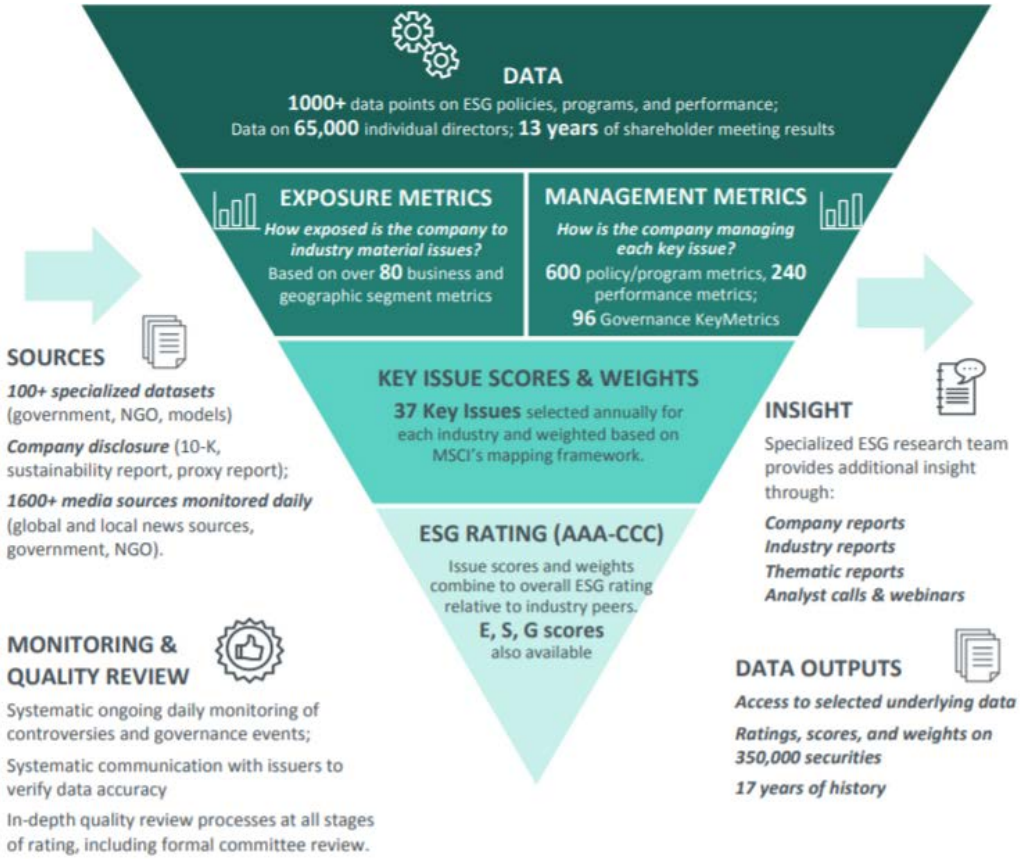


Figure 5.1 : ESG Rating Framework and Process Overview

(a) (MSCI 2018)

into the final overall ESG rating (see Figure 6.2a). This analysis process tends to be tedious and time-consuming, which often be prone to human errors.

Afterwards, the fund managers will integrate these ESG scores with other financial market analysis to make investment decisions. The common methods used by these ESG-focused funds are simply either setting an ESG scores threshold (e.g. a fund invests in large market capital companies with top 50% scores) or thematic investing in companies with a specific topic focus (e.g. a clean energy focused fund invests in solar and wind energy companies only). These approaches are heavily relied on ESG scores or the business types without a financial intuition.

Many research have been trying to test the correlation between ESG scores and

the stock market returns, in which many of them have showed positive results across regions, asset types and time periods (Friede et al. 2015). The relationship is believed to be more complex than a simple linear correlation (Sahut and Pasquini-Descomps 2015), which can be positive or negative depending on the measurements (Buallay 2019). Investors in general are still skeptical about using ESG Scores to predict financial returns. A quantitative model integrating ESG Scores as signals for returns forecast can help test this hypothesis.

Considering the current state of sustainable investment, we identify the research gaps and seek to answer the questions: 1) Can text mining and machine learning techniques help predict the ESG ratings? 2) Can ESG be used as indicator for financial performance of the firms? 3) How can fund managers use this data-driven approach to construct better sustainable investment portfolios?

In this research, we propose a data-driven prescriptive analytics approach to 1) accurately predict the ESG ratings of the companies using text mining and machine learning techniques on their annual corporate responsibility reports and 2) quantitative measure to use ESG scores as indicator to predict long term returns. The proposed data-driven methods will help the fund managers to make better investment decisions leveraging big data and machine learning. The ESG-focused funds can construct better sustainable investment portfolios as well as engage more investors' interests in the funds.

The main contributions of this research are:

- This is the first attempt to leverage text mining and machine learning to build an ESG scores prediction model and evaluate corporate social responsibility of companies.
- It propose quantitative finance models to evaluate the use of ESG score as an indicator for financial performance prediction, namely "P/ESG indicator model", "MV-ESG model" and "Combined MV-ESG model".
- The proposed models are tested using real-life datasets, showing its effectiveness and usefulness in application to all stakeholders, particularly fund

managers and individual investors.

5.2 Preliminary

5.2.1 Text Mining for Socially Responsible Investment

In the last decade, the application of data mining and machine learning in finance and management research has been increasing significantly, especially leveraging text mining approaches (Kumar and Ravi 2016). These techniques have been applied to mine textual information in company reports for risk evaluation (Bao and Datta 2014) or in regulatory disclosures for stock index forecasting (Feuerriegel and Gordon 2018). Different financial and market sentiment dictionaries have been developed from company reports (Loughran and McDonald 2016) and social media text (Chen et al. 2018), which has further advanced the research in this field.

Regarding research within sustainability field, the application of text mining has been increasing significantly in recent years, mainly analyzing textual information from company reports related to sustainable development and management (Kolk et al. 2018). (Landrum and Ohsowski 2017) are debating about the reliability of corporate social responsibility and Global Reporting Initiative (GRI) reports in terms of reflecting the current state of world sustainability. However, these reports are still considered to be the most accurate data sources for text analysis within this research topic (Székely and vom Brocke 2017).

Text mining techniques have been applied to sustainability research in various industries, e.g. maritime (Shin et al. 2018) and manufacturing and services (Park and Kremer 2017). Kim and Kim (2017) had explored sustainable supply chain of textile companies by analyzing textual information from news articles and company reports. Researchers have also been working on building text mining dictionaries for corporate responsibility (Pencle and Mălăescu 2016) and environment sustainability in IT industry (Deng et al. 2017). Since the aim of this research is to serve the broad sustainable investment across all industries, we will utilize the corporate responsibility dictionary of Pencle and Mălăescu (2016) for text mining.

The current research shows the potential of leveraging text mining and machine learning approaches within sustainability domains in general, which is the motivation for applying these techniques for sustainable investment in this paper. The proposed prescriptive analytics method will leverage the text mining and machine learning techniques to provide a data-driven approach for predicting ESG ratings and measuring the sustainable investment impacts on the both individual and fund levels. Besides the high level of industry applicability, this paper also contributes directly to the current literature in both text mining for sustainable investment and ESG-focused fund management methodologies.

5.2.2 ESG as indicator for long-term stock returns forecast

Most of the current literature on socially responsible investment field are qualitative research (Ang and Weber 2018) and limited to a certain country market (Formánková et al. 2019). Multiple researchers have investigated the correlation between ESG ratings and the performance of stocks (Kempf and Osthoff 2007) or ESG-focused funds (Brito 2018). Academic researchers are still debating about the causal relationships between corporate social responsibility and the financial performance of the companies (Bose and Pal 2012).

Despite all the controversies, investing in companies with strong performance on the environmental, social and governance is becoming significantly attractive to a growing number of highly-educated and ethical investors (Nilsson 2008). An empirical experiment by Døskeland and Pedersen (2016) showed that even traditional investors with wealth growth purpose are interested in sustainable investment options as they see them as attractive opportunities for financial gain.

Past research is mainly limited to a certain country market or region (Velte 2017; Formánková et al. 2019; Chelawat and Trivedi 2016). Many researchers have been trying to prove that ESG Scores are actually correlated to the financial performance of the companies. An aggregated evidence from 2,200 studies has showed the positive results across regions, asset types and time periods (Friede et al. 2015). However, some scholars might argue that the relationship can be either positive or negative

depending on the underlying measurements (Buallay 2019). Similarly, Sahut and Pasquini-Descomps (2015) suggested that the relationship might be non-linear.

While scholars are still debating whether ESG Scores are correlated to the financial performance of the companies, evidence has showed that it might be a significant indicator for market predictability (Khan 2018). In traditional quantitative finance model, different financial indicators have been used to predict the stock returns, e.g. Total Payouts (Straehl and Ibbotson 2017), Sales Per Share or Book Value Per Share (Pedersen 2015). We consider ESG Scores as a similar type of indicators, which can be used as predictive signals for stock returns forecast. We propose the quantitative model using the "Price / ESG Score" ratio P/ESG to test the first set of our research questions and hypotheses: H_0: ESG Scores cannot be an indicator for financial performance, versus H_1: ESG Scores can be an indicator for financial performance.

5.2.3 ESG for Portfolio Optimization

On the portfolio construction side, not many socially responsible investment models have been developed and proposed utilizing the ESG ratings (Van Duuren et al. 2016). Garcia-Bernabeu et al. (2015) have suggested a modification to the standard portfolio selection model with ESG scores. They utilize the Mean-Variance Stochastic Goal Programming (MV-SDP) model with a statistical approach for ESG screening on the stocks based on scores and controversy risk. However, they do not consider predictive analytics but only use past returns and volatility. They have not tested with real financial data thus the performance of their models are not fully evaluated.

Other scholars have proposed fuzzy multi-criteria approach to integrate ESG investors' preferences into SRI (Escrig-Olmedo et al. 2017). This research work might be beneficial to the investors who already know what they want in SRI. The vast majority of financial investors are considering SRI as a slightly more ethical mean for financial gain rather than just purely charitable purpose. These investors still want to invest in an ESG-based portfolio that can also perform well financially.

To fill the gap in the current literature, we develop a financial model to construct a socially responsible investment portfolio incorporating the Mean-Variance portfolio theory and the ESG ratings (MV-ESG). Our model is not based on ESG screening, instead, it filters and leverages the ESG ratings in a multi-objective optimization function based on the NSGA-II optimizer (Deb et al. 2000). Our "MV-ESG model" is one of the first mathematical models for constructing a socially responsible investment portfolio to achieve both better ESG ratings and a competitive financial performance.

5.2.4 ESG for Portfolio Diversification

Many ethical investors are concerned with the diversification of sustainable funds and portfolios. After the first research conclusion from Rudd (1981), researchers have expressed support for the same point of view that the integration of ESG into a financial portfolio will make it less diversified (Barnett and Salomon 2006; Renneboog et al. 2008). It is still a debate of how much worse the diversification is with the ESG-based portfolios (Bauer et al. 2007). On the other hand, there are research showing that SRI funds do not have worse portfolio diversification (Bello 2005; Renneboog et al. 2007). Additionally, Schröder (2007) observes that sustainable and standard financial indices both have comparable Sharpe ratios, similar to our results from previous part. This implies that ESG-based portfolio and conventional funds might have similar level of diversification.

We believe the observed reduction of portfolio diversification is due to the fact that most funds practice the similar screening approaches when it comes to selecting stocks for their portfolio. These screens can be classified into two main groups: negative screening and positive screening. The negative screening is when certain stocks or industries are excluded from SRI portfolios. After the negative screening, the SRI portfolios are created through financial and quantitative selection. The most common negative screens filter out companies involved in tobacco, genetically modified organism (GMO), thermal coal, etc. On a different note, the positive screening will select companies with strong focus on a specific CSR areas, e.g. renewable energy or gender equality. The positive screening is often combined with a 'best

in class' approach. The positive screen portfolio generally invest in companies that are ranked at the top in their respective industry or market sector. These screening approaches significantly limit the number of stock in the portfolio or invest in the highly correlated companies in the same industries, which worsen the portfolio diversification.

Negative and positive screens are often referred to as the first and second generation of SRI screens respectively. The third generation of screens refers to an integrated approach of selecting companies based on the ESG ratings comprised by both negative and positive screens. Following this trend of integrating ESG into portfolio construction, we proposed a weighted ESG-based portfolio optimization to incorporate the positive and negative screenings. Our proposed quantitative method seeks to enable the stock screenings based on the preferences of SRI funds and ethical investors without sacrificing the portfolio diversification.

5.3 Data Mining Methods

5.3.1 Text Mining Model

To assist sustainable investors and ESG-focused fund managers, we develop a framework to support corporate social responsibility analysis and impacts measurement. Firstly, we applying text mining techniques to extracts textual information from company reports with sentiment analysis on a dedicated corporate social responsibility dictionary (Pencle and Mălăescu 2016) in comparison with two other lexical analysis dictionaries, namely Linguistic Inquiry and Word Count (LIWC) 2015 (Pennebaker et al. 2001) and Empath (Fast et al. 2016). These stacked text features will be used as input for our ESG score estimator. We build and test the prediction models using two popular machine learning algorithms Random Forest (Pedregosa et al. 2011a) and Extreme Gradient Boosting (Chen and Guestrin 2016) Regressors.

We can use the predicted ESG scores in previous step as weighting input to measure the environmental and social impacts of sustainable investment, using the reported and estimated data from companies. We evaluate the investment impacts

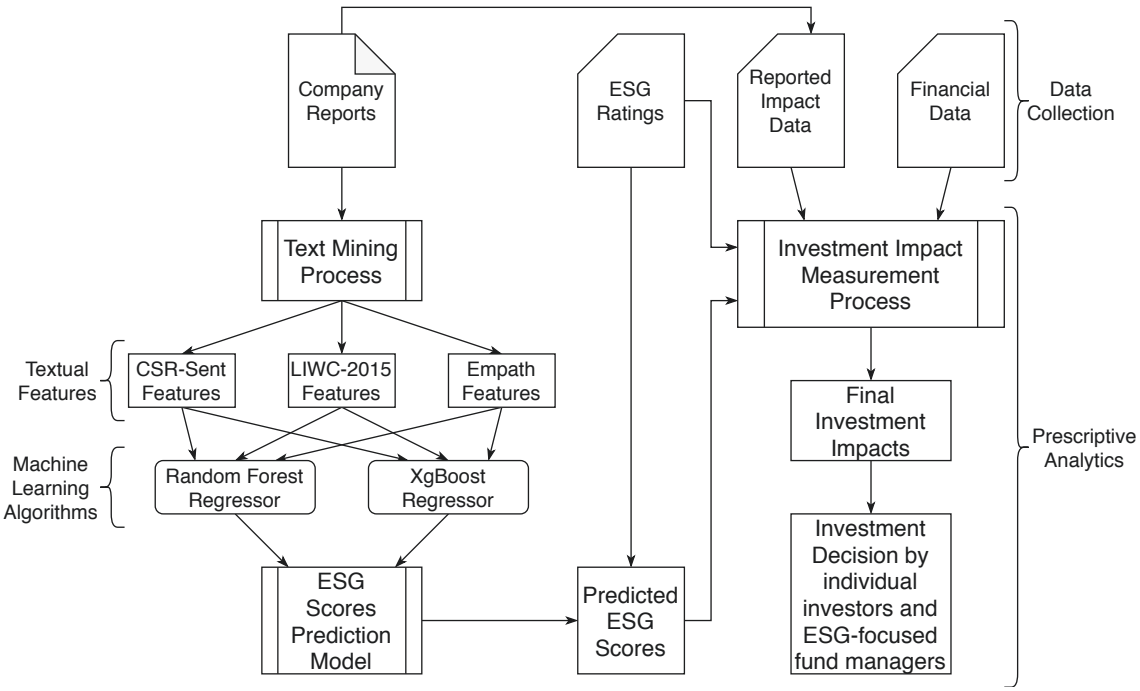


Figure 5.3 : Our Methodology Framework and Process Overview

using a nominal set of five different metrics: green house gas (GHG) emission reduction, waste recycling, energy saving, water saving and charitable giving in community support (see Appendix A). The quantified impacts measurements will serve investors and fund managers as the key sustainable metrics for making relevant investment decisions. The full framework of our methodology with process overview is illustrated in Figure 5.3.

Since the document lengths of different corporate social responsibility reports vary, the Bag-Of-Words (Harris 1954) or similar word count models would not be a suitable in this research case. Therefore, we take a alternative text mining approach using sentiment and lexical information text analysis. We extract three different type of text features: corporate social responsibility sentiment (CSR-Sent), LIWC-2015, and Empath features.

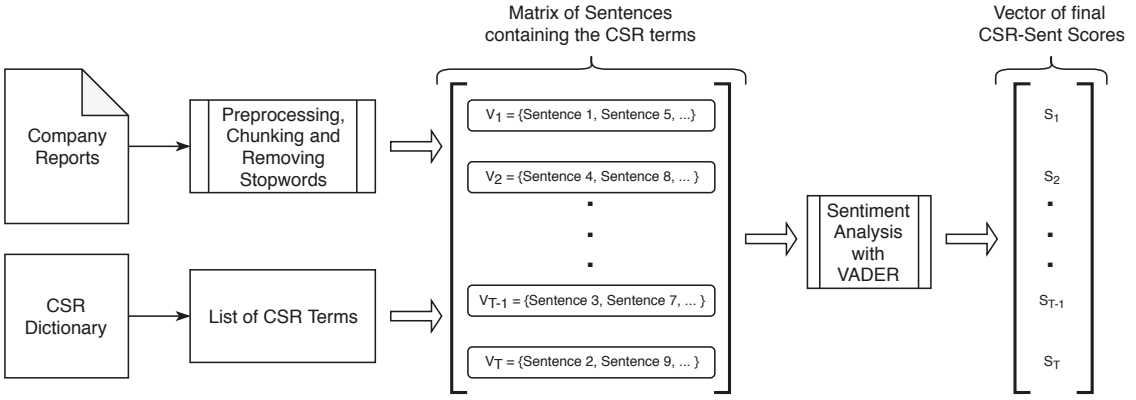


Figure 5.4 : CSR-Sent Text Feature Extraction Process

CSR-Sent Text Features

First of all, company reports are converted to a corpus dataset with text only, ignoring all photos and figures. Next in our text processing steps, we perform chunking to break down the documents into sentences and removing all common English stopwords. Afterwards, the automatic code will search through the current company report to find all the occurrences of each term in the dictionary and pull out the chunked sentences. We then use VADER (Hutto and Gilbert 2014) to analyze the sentiment and compute the CSR-Sent text feature score for that word in the sentence containing it. The final CSR-Sent text feature score for each word is averaged based on all the occurrences within each company report. The full process of CSR-Sent text feature extraction is illustrated in Figure 5.4.

The word list is generated using a multi-dimensional corporate social responsibility dictionary (Pencle and Mălăescu 2016) consisting of $1,428$ terms in total. The word list is divided into four big sustainability categories: Environment, Social and Community, Human Resources and Human Rights. For example, Table 5.1 shows the sample terms and their CSR-Sent text feature scores extracted from the latest corporate social responsibility report of Intel Corporation. The higher scores are linked to the positive sentiment about the good performance of Intel Corporation in that particular sustainable subject, and vice versa, the lower scores indicate a weaker confidence in the corresponding company performance.

Table 5.1 : CSR-Sent Text Feature Scores of Intel Corporation

Category	**Total Terms**	**Sample Term**	**CSR-Sent Score**
Environment	451	Renewable Energy	0.3094
		Preserve	0.8725
		Water	0.0304
Social and Community	361	Social	0.3669
		Responsible	0.4740
		Community	0.6060
Human Resources	319	Incentives	0.3953
		Equal Opportunity	0.9118
		Pension	0.3626
Human Rights	297	Ethical	0.7800
		Minority	0.9464
		Healthcare	0.3602

LIWC-2015 Text Features

Besides sentiment analysis, we believe there are more useful information lies within the words used in the corporate social responsibility reports. These are often called lexical information. The Linguistic Inquiry and Word Count 2015 (LIWC-2015), developed by Pennebaker et al. (2001), is one of the most comprehensive text mining tools to extract the lexical information and topic-specific features. The LIWC 2015 master dictionary consists of about $6,400$ terms, each term is linked to one or more categories.

The LIWC-2015 contains multiple topic-related features (e.g. "social" or "health") and emotions (e.g. "positive emotion", "negative emotion"). The text features related to grammar and punctuation, e.g. "verb", "adverb" or "comma", are generally not very insightful for our analysis. Some extracted features are measured in the speaking context, e.g. "non-fluent" or "filler". These speech-related text features are not relevant in our dataset of written reports. The higher scores indicates the higher frequency of feature term appearance within the report relative to the total

number of word count. Table 5.2 shows the basic statistics of a few features out of the total 93 LIWC-2015 text features extracted from our dataset consisting of 500 corporate responsibility reports.

Table 5.2 : Basic statistics of the LIWC-2015 text features

Features	**Mean**	**Std**	**Min**	**Max**
Word Count	27946.40	29389.45	196.00	228991.00
Analytic	95.02	2.34	78.12	99.00
Clout	78.84	11.00	35.09	99.00
Authentic	13.60	4.87	2.15	46.71
Tone	77.74	12.79	9.33	98.89
social	7.60	2.50	1.15	14.67
work	10.47	2.02	4.62	17.94
drives	12.36	2.72	2.55	23.59
posemo	3.65	0.83	0.51	6.46
negemo	0.67	0.33	0.00	3.36

Empath Text Features

Empath is an open-source text mining tool developed by Fast et al. (2016). The Empath model was trained using deep learning techniques with a neural embedding consisting of over 1.8 billion terms extracted from a big data corpus of modern fictions. With smaller categorical data, Empath utilize the neural embedding to find the new related words, then validates it with a crowd-powered filter. There are 200 built-in prevalidated Empath text features covering the most common topics, e.g. technology, tourism or music. Even though some of the Empath text features are highly correlated (r = 0.906) with the similar LIWC-2015 ones, we believe the set of 200 Empath text features can reveal some insightful lexical information in the reports that LIWC-2015 might have missed. Furthermore, there are some specific text features from the Empath dictionary that are closely related to our sustainable context. Table 5.3 shows some basic statistics of some of those Empath text features.

Table 5.3 : Basic statistics of the Empath text features (percentage scores)

Features	**Mean**	**Std**	**Min**	**Max**
animal	0.0628	0.1016	0.0000	1.0741
cleaning	0.1381	0.1665	0.0000	1.8979
economics	1.3195	0.6425	0.0000	3.4396
farming	0.0465	0.0963	0.0000	0.9348
giving	0.6037	0.2122	0.0000	1.5001
health	0.2281	0.2699	0.0000	1.9676
plant	0.0950	0.0983	0.0000	0.9189
pride	0.1366	0.0858	0.0000	0.8857
water	0.1181	0.1531	0.0000	1.7721
work	0.8255	0.3262	0.1245	3.0612

5.3.2 ESG Scores Prediction Model

As our main technical contribution and focus are in the text feature mining, we use two well-known regressors to build our prediction model and testing our hypotheses. We choose these two algorithms for their effectiveness and efficiency based on proven performance on other predictive analytics tasks within finance field (Chatzis et al. 2018).

Random Forest Algorithm

Random Forest Regressor, first introduced by Breiman (2001), is an estimator that can fit multiple decision trees on various dataset sub-samples. The algorithm uses averaging to improve the accuracy of the prediction and avoid over-fitting. Using the adaptation of the algorithm from Pedregosa et al. (2011a), we define the Random Forest Regressor in our ESG scores prediction model as follow.

Let $h_k(x)$ be a decision tree k with input x, each decision tree k leads to an estimator $h_k(x) = h(x|\Theta_k)$ with the parameters $\Theta_k = (\theta_{k1}, \theta_{k2}, ..., \theta_{kp})$. The ensemble of multiple decision trees forms a random forest based on a set of estimators

$\{h(x|\Theta_1), ..., h(x|\Theta_1)\}$ with parameters Θ_k randomly chosen from the model random vector Θ

Specifically given the data $D = \{(X_i, y_i)\}_{i=1}^{N}$, we train a set of estimators $h_k(X) \equiv h(X|\Theta_k)$, which we set to 100 by default. In our case, X_i is the vector of text features extracted from the latest corporate responsibility report of company i, y_i is the ESG scores of company i and $N = 500$ is the total number of companies in our dataset.

Extreme Gradient Boosting (XgBoost) Algorithm

Chen and Guestrin (2016) proposed the XgBoost algorithm in 2014 as an improved interpretation of the greedy gradient boosting machine of Friedman (2001). XGBoost has quickly been adopted and become one of the most commonly-used algorithms in supervised machine learning due to its effectiveness and efficiency. In our ESG scores prediction model, we firstly define the tree $f(x)$ as:

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^d \to \{1, 2, \ldots, T\} \tag{5.1}$$

where w is the vector of ESG scores, q is a data assigning function for the respective leaf, and T is the total number of leaves. The Hessian h_i and the gradient g_i are defined as follow:

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{5.2}$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{5.3}$$

where y_i is the actual ESG scores obtained from professional rating agency and $\hat{y}_i^{(t-1)}$ is the predicted ESG scores at the time $(t-1)$. Using regularization to improve generalization performance, the objective function at t^{th} tree is defined as:

$$\Rightarrow obj^{(t)} = \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w^2_{q(x_i)}] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j| \tag{5.4}$$

where w_j is the weight assigned to the j^{th} leaf, $\gamma = 0$, $\alpha = 0$ and $\lambda = 1$ are predefined parameters for the penalization and regularization terms accordingly.

5.3.3 P/ESG Indicator Model

The core idea is inspired by (Pedersen 2015), where the author used the "Price / (Book Value Per Share)" P/B ratio for predicting the 10-year annualized return of the S&P 500:

$$Annualized\ Return \simeq 23.4\%–4.9\% \cdot P/B \tag{5.5}$$

This theory proposes that the stock-market as a whole is not "efficient" and does not follow a purely "random walk" in the long-term. It is possible to estimate the future long-term return of the stock-market and some individual stocks from just a single indicator variable. While (Pedersen 2015) promoted Book Value Per Share as an indicator to forecast long-term stock returns, we believe ESG Scores can also be used as a forecast signal, not only in our context of social responsible investment but also in general financial investment term.

Let us define the total return r_t of a stock at time-step t as the number of shares m_t which may grow from reinvestment of dividends (taxes are ignored), multiplied by the share-price p_t:

$$r_t = m_t \cdot p_t \tag{5.6}$$

The annualized return ar_t between the start date $t = 0$ and end date $t + \Delta t$ is:

$$ar_t = \left(\frac{r_{t+\Delta t}}{r_t} \right)^{1/\Delta t} - 1 \tag{5.7}$$

We derive two formulas for the mean μ_{ar_t} and the standard deviation σ_{ar_t} of the annualized return ar_t given the P/B ratio at time-step t :

$$\mu_{ar_t} = \frac{\alpha}{P/B_t^{1/\Delta t}} - 1 \tag{5.8}$$

$$\sigma_{ar_t} = \frac{\beta}{P/B_t^{1/\Delta t}} \tag{5.9}$$

where the parameters α and β can be estimated from three factors: (1) The growth in the number of shares from reinvestment of dividends (DY), (2) the growth in Book Value Per Share (BG), and (3) the change in the P/B valuation ratio. Forecasting the future return on a stock can therefore be split into forecasting these three factors.

Similar to the "P/Book model", we define the "P/ESG model" with the mean μ_{ar_t} and the standard deviation σ_{ar_t} of the annualized return ar_t given the "Price / (ESG Score)" ratio P/ESG at time-step t as:

$$\mu_{ar_t} = \frac{\theta}{P/ESG_t^{1/\Delta t}} - 1 \tag{5.10}$$

$$\sigma_{ar_t} = \frac{\sigma}{P/ESG_t^{1/\Delta t}} \tag{5.11}$$

where the parameters θ and σ can be estimated from three factors: (1) The growth in the number of shares from reinvestment of dividends (DY), (2) the growth in ESG Score (ESGG), and (3) the change in the P/ESG valuation ratio. Since the prices of different stocks can be greatly varied (ranging from a few US dollars to a few thousands US dollars), we rescale the P/B and P/ESG ratios to range $[0, 1]$ before using them as predictive signals for the forecast models.

5.3.4 MV-ESG Model

The Mean Variance portfolio (MV) of Markowitz (1952) has always been the standard portfolio selection model. Its mathematical principle is constructed by two main components: maximizing the return r_p and minimizing the risk σ_p. The output of this optimization process is the efficient frontier, which is a set of investment portfolios with a greater return than any other with the same or less risk, and a lower risk than any other with the same or greater return. For illustration, the efficient frontier is plotted in Figure 6.4 with the risk on the horizontal axis and the return on the vertical axis.

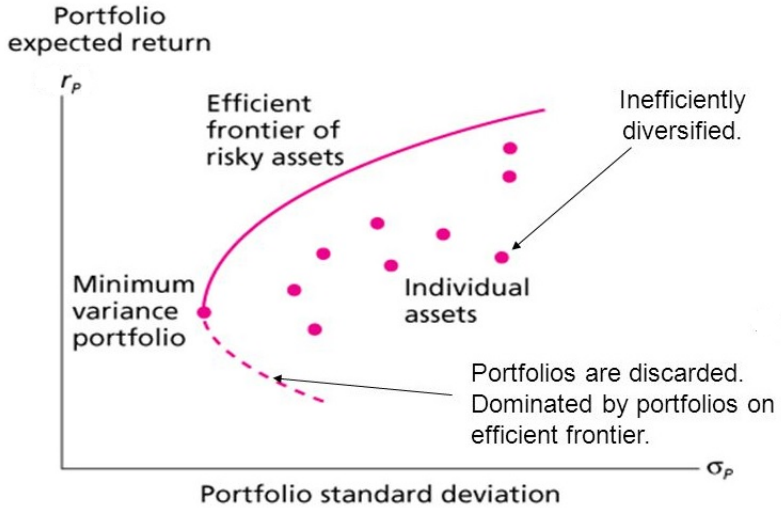


Figure 5.5 : Standard MV Portfolio with Efficient Frontier

The optimal portfolio is commonly known as the maximum Sharpe portfolio (MAX-S). For the MAX-S portfolio, considering the risk free rate r_f (normally the return on bond investment or the bank interest rate), it minimizes the negative Sharpe Ratio (Sharpe 1966):

$$min(-S_p) = min(-\frac{r_p - r_f}{\sigma_p}) \tag{5.12}$$

$$r_p = \sum_{i=1}^{N} w_i r_i \tag{5.13}$$

$$\sigma_p = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i \sigma_{ij} w_j \tag{5.14}$$

where w_i and w_j are the weights of stock i and j, with the boundary limit $w_i, w_j \in [0, 1]$, and σ_{ij} is the covariance matrix of the two stock i and j in the portfolio. The initial weight of each stock in the computation is equally allocated according to the total number of stocks N in the portfolio, $w_i(0) = w_j(0) = 1/N$.

For comparison, we construct a maximum ESG portfolio (MAX-ESG) for investors with low risk averse to compare with the standard MAX-S portfolio. In our MAX-ESG portfolio, we minimize the negative Sharpe Ratio with the portfolio ESG ratings (ESG_p) as a new variable of the objective function.

$$min(-S_p) = min(-ESG_p \frac{r_p - r_f}{\sigma_p}) \tag{5.15}$$

$$ESG_p = \sum_{i=1}^{N} wesg_i \frac{ESG_i + E\bar{S}G_i}{2} \tag{5.16}$$

where ESG_i is the combined ESG ratings of company i in the past year, $E\bar{S}G_i$ is the combined ESG ratings at the current prediction year, and $wesg_i$ is the ESG weight of stock i in the portfolio.

In the traditional MV model, r_p and σ_p are the past returns r_i and volatility σ_i, which is often called ex-post MV. In recent years, researchers and investors have been using the expected returns $\bar{r}_i$ and volatility $\bar{\sigma}_i$. This approach called ex-ante MV is more suitable for predictive analytics in real-world financial trading. In our MV-ESG model, we combine both ex-post MV and ex-ante MV for portfolio selection and replace the standard weight boundary with our ESG ones calculated based on the combined ESG ratings for each stock. Our MV-ESG model is computed using:

$$r_p = \sum_{i=1}^{N} wesg_i \frac{r_i + \bar{r}_i}{2} \tag{5.17}$$

$$\sigma_p = \sum_{i=1}^{N} \sum_{j=1}^{N} wesg_i \frac{\sigma_{ij} + \bar{\sigma}_{ij}}{2} wesg_j \tag{5.18}$$

where r_i and $\bar{r}_i$ are the ex-post and ex-ante returns, σ_{ij} and $\bar{\sigma}_{ij}$ are the ex-post and ex-ante covariance matrix of the two stock i and j in the portfolio. $wesg_i$ and $wesg_j$ are the ESG weight of stock i and j in the portfolio, with the boundary limit $wesg_1, \in [0, 1]$ for the company with the highest combined ESG score, then gradually decreasing to $wesg_N \in [0, 0]$ for the company with the lowest combined ESG score. This means the allocation of the company "N" in the portfolio is zero, indicating no investing. The initial weight of each stock in the computation is not equally allocated but assigned according to the ESG ratings.

We define three separate objectives for our ESG-based Multi-Objective Portfolio Optimization model: 1) maximizing returns, 2) minimizing risks, and 3) maximizing ESG scores. The three measures, representing these three objectives, are calculated

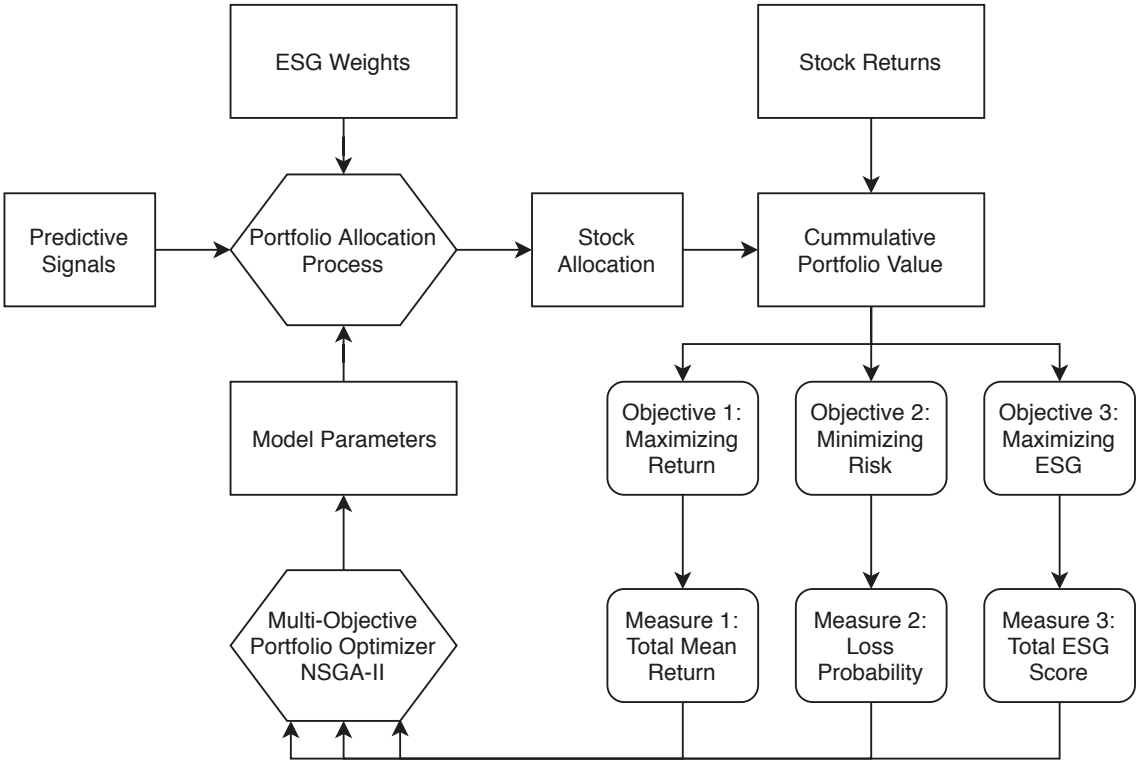


Figure 5.6 : Multi-Objective Portfolio Optimization Model

using the multi-objective optimizer known as NSGA-II (Deb et al. 2000), short for Non-Dominated Sorting Genetic Algorithm version 2. This loop is repeated until a Pareto-front of solutions is found, which optimally compromises between the two conflicting goals. The whole model framework is illustrated as in Figure 5.6 below.

5.3.5 Combined MV-ESG Model

Starting with the proposed MV-ESG Model in the previous part, we will redefine the $wesg_i$ to incorporate the positive and negative screenings into a "Combined MV-ESG Portfolio". For positive screening, we consider a subset of companies $i \in 1, ..., n$ that have strong focus on a specific CSR area. The "Positive MV-ESG Portfolio" will optimize the stock allocation $wesg_i^+$ as follows:

$$wesg_i^+ = \gamma * wesg_i \tag{5.19}$$

where $\gamma > 1$ is the nominated weighting based on the positive screening preferences, which means the investors prefer to invest more in company i. For negative screening, we also consider a subset of companies $i \in 1, ..., n$ that involved into socially negative impact businesses (e.g. gambling or adult entertainment). The "Negative MV-ESG Portfolio" will optimize the stock allocation $wesg_i^-$ as follows:

$$wesg_i^- = \Gamma * wesg_i \tag{5.20}$$

where $\Gamma = 0$ is the nominated weighting based on the negative screening preferences, which means the investors prefer not to invest in company i. We combine these positive and negative screening into a "Combined MV-ESG Portfolio" with the stock allocation $wesg_i^c$ as follows:

$$wesg_i^c = wesg_i^+ * wesg_i^- = \gamma * \Gamma * wesg_i \tag{5.21}$$

Our intuition behind the weighting approach is to adjust our "Combined MV-ESG Portfolio" to incorporate a variety of investors' preferences without compromising the financial performance or the ESG Ratings. We evaluate the model as a whole and separately as the " Positive MV-ESG Portfolio" and the "Negative MV-ESG Portfolio" to further test our hypothesis.

5.4 Experiment

5.4.1 Datasets

As of today, there are numerous ESG rating services offered by professional agencies (Schäfer 2016). Many of these data sources are accessible only by paying high subscription fees, which limits the availability of information and development of quantitative methodology. Yahoo Finance, in 2018, has made some of the ESG scores from Sustainalytics (Stay 2010) available publicly. In this research, we use Yahoo Finance website to access the public ESG Ratings dataset consisting of 500 companies for our prescriptive analysis. Table 5.4 below provides some basic statistics

on our ESG Ratings dataset, including the "Overall ESG Score", "Environmental Score", "Social Score" and "Governance Score".

Table 5.4 : Basic statistics of ESG Ratings dataset

	Mean	**Std**	**Min**	**Max**	**Max-Min Gap**
Overall ESG Score	57.97	9.02	40.49	86.47	45.98
Environmental Score	56.99	13.94	31.10	97.64	66.54
Social Score	56.21	10.70	33.88	90.35	56.47
Governance Score	62.71	7.97	38.02	88.16	50.14

Regarding dataset for text mining, we download the corporate responsibility reports of these companies from their websites. In case the company does not publish this type of report, similar sustainability or integrated annual reports will be used instead. The shortest report consists of only 196 words, and the longest one have $228,991$ words in total. The final text corpus contains a total $13,973,202$ words from these 500 reports, with an average of $27,946$ words per report. We also obtained the financial performance data, specifically the EBITDA and Price Per Earning Ratio, of all 500 stocks for investment impact measurement model.

To further test the effectiveness of ESG Score as an indicator for financial performance of the firm, we also use the real-life financial data obtained from SimFin. This dataset includes 10-year historical share prices from 2009 to 2019 of about $2,400$ stocks together with their fundamental data extracted from approximately $258,000$ financial statements. Afterwards, we map the SimFin dataset with the ESG data obtained from the Sustainability page of Yahoo Finance. We remove all the companies which do not have ESG data or have too many missing values in their financial data. Our final cleaned dataset consists of $1,121$ stocks in total.

We test the MV-ESG on that clean dataset with $1,121$ stocks across industries and countries to compare the model performance with and without the third objective of maximizing ESG scores. Since we want to ensure our portfolios do no simply invest in high return high risk stocks only, we limit the maximum loss probability

in the MV and MV-ESG portfolios to 15%. We also build a dummy portfolio with equal weights for each stock. The list of tested portfolios is in Table 5.5.

Table 5.5 : Objectives of Tested Portfolios

Portfolios	**Max Return**	**Min Loss**	**Max ESG**	**Note**
Equal Weights	No	No	No	Invest equally in all stocks
Min Prob Loss	No	Yes	No	
MV Portfolio	Yes	Yes	No	Limit max loss probability to 15%
Min Prob Loss ESG	No	Yes	Yes	
MV-ESG Portfolio	Yes	Yes	Yes	Limit max loss probability to 15%

5.4.2 Evaluation Metrics

To test ESG Score as an indicator to forecast financial performance of the firm, we test how the predicted annualized return ar_t varies with the P/B_t and P/ESG_t ratios. We are now interested in measuring how well these predicted curves fit the actual historical observations.

Let (x_t, y_t) be a pair of historical observations for time-step t so that $x_t = P/B_t$ and $y_t = ar_t$. The stochastic variable X is for the P/B ratio and the associated variable Y is for the annualized return ar. The mean of the historically observed annualized return is denoted $\bar{Y}$ and calculated as:

$$\bar{Y} = \frac{1}{n} \sum_{t=1}^{n} y_t \tag{5.22}$$

Let $\hat{\mu}_t$ be the mean annualized return estimated from the above formula for μ_{ar_t} given $x_t = P/B_t$ and some choice of parameter α. Similarly let $\hat{\sigma}_t$ be the standard deviation for σ_{ar_t} given $x_t = P/B_t$ and some choice of parameter β. We use four different measures of how well the annualized return predicted by $\hat{\mu}_t$ and $\hat{\sigma}_t$ fit the actual historical observations y_t.

Our evaluation metrics are:

- **Mean Squared Error (MSE)** is a common measure of how well predicted values fit actual observations. The MSE is non-negative with smaller values meaning a better fit. An MSE of zero means a perfect fit without any errors between actual and predicted observations. But because the errors are squared, it means that larger errors may dominate the overall error measure. Furthermore, the squared errors make interpretation difficult. The formula is:

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{\mu}_t)^2 \tag{5.23}$$

- **Mean Absolute Error (MAE)** is another common measure of how well predicted values fit actual observations. Because it uses the absolute instead of squared errors, it is easier to interpret than MSE. The MAE is also non-negative with lower values meaning better fits and zero MAE is a perfect fit. The formula is:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{\mu}_t| \tag{5.24}$$

- **Mean Absolute Z-Score (MAZ)** is probably not the right name for this statistic, but the idea is to compare both the forecasted mean and standard deviation to the actual observations, unlike the MSE and MAE which only use the forecasted mean. The MAZ is calculated as the average number of forecasted standard deviations that the actual observations are from the forecasted mean. The MAZ is also non-negative, but it is harder to interpret because we want low, but not too low, MAZ values. We can have a value of zero for the MAZ if either the forecasted mean $\hat{\mu}_t$ perfectly fits the observed data y_t , or when $\hat{\sigma}_t \to \infty$ so that MAZ $\to 0$ but that does not mean a good fit of the model to the data. The formula is:

$$\text{MAZ} = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{\mu}_t|}{\hat{\sigma}_t} \tag{5.25}$$

- **Coefficient of Determination** R^2 usually measures how much of the variance in the observed data is explained by the model. It is defined from the

Sum of Squared Errors (SSE) between the forecasted mean $\hat{\mu}_t$ and actual observations y_t, relative to the Sum of Squared Errors Total (SST) between the actual observations y_t and their mean $\bar{Y}$:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{\sum_{t=1}^{n} (y_t - \hat{\mu}_t)^2}{\sum_{t=1}^{n} \left(y_t - \bar{Y}\right)^2} \tag{5.26}$$

The R^2 usually goes from 0 to 1, with 0 meaning the model does not explain any of the variance in the data, and 1 meaning that the model fits the data perfectly and therefore explains all of the variance in the data. However, because we have a non-linear model, the SSE may be greater than the SST, so R^2 can become negative when there is great variance in the data and the non-linear model fits poorly.

To test the accuracy our ESG Score prediction model, we use the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) measurements as the evaluation metrics. The calculations of MAE and RMSE are as follow:

$$\mathbf{MAE} = 1/N \sum_{i=1}^{N} |\tilde{y}_i - y_i| \tag{5.27}$$

$$\mathbf{RMSE} = \sqrt{1/N \sum_{i=1}^{N} (\tilde{y}_i - y_i)^2} \tag{5.28}$$

where $N = 500$ is the total number of companies in the dataset and $\tilde{y}_i$ and y_i are the predicted and actual overall ESG scores of the companies i on December 2018. The lower MAE and RSME indicate the more accurate prediction and the better performing model. The original ESG Scores from Sustainalytics are ranging from 0 to 100. In our experiment, we scale the ESG scores to range 0 to 1 for better prediction, then we rescale the predicted value back to the original scale before calculating the MAE and RSME.

To evaluate the performance of our proposed "MV-ESG Model" compared to the standard "MV Model", we use a set of different metrics and graphical comparisons as below.

- **Portflio Accumulated Returns** $V_p(t)$ is the most common evaluation metrics to test trading strategies and portfolio optimization performance over a

set period of time. In our back-testing result using historical data, we will test the portfolio over 2 year period from 2017 to the end of 2018. The returns are compounded daily with a full reinvestment strategy. For simplicity purpose, we assume there is no tax, portfolio management fees or trading fees. The portfolio value at time t is calculated as:

$$V_p(t) = \prod_{i=1}^{t} r_p(i) \tag{5.29}$$

The graph starts at 1 representing the 100% initial investment amount. The higher portfolio value, which means the higher graph line, represents the better trading strategies and portfolio optimization performance.

- **Pareto Front** is the set of all Pareto efficient allocations. It is also refer to as the efficient frontier in portfolio optimization context. The Pareto Front illustrates graphically all the possible optimal combination of portfolio returns and the trade-off risks. The higher returns with lower risks indicate better portfolio optimization in terms of financial performance.

- **Return r_p, risk σ_p and Sharpe Ratio** S_p are the standard portfolio evaluation metrics. The higher return, higher Sharpe Ratio and lower risk are indications of a better portfolio's financial performance. The returns and risks are calculated as in the proposed "MV-ESG Portfolio". The Sharpe Ratio is measured with a nominal risk free rate $r_f = 2\%$

- **Portfolio ESG Score** ESG_p is the weighted average of ESG ratings from all portfolio allocation. While aboves metrics and graphical comparisons are focusing only on financial performance, the Portfolio ESG Score would provide better evaluation on the combined ratings of CSR of companies in the portfolio, which suits our context of SRI.

To further evaluate the portfolio diversification, we also use the relative stock weights $w\tilde{e}sg_i^c$ to compare the portfolio allocation. For example, the relative stock weights $w\tilde{e}sg_i^c$ of the "Combined MV-ESG Portfolio" against the "MV-ESG Portfolio" is calculated as.

5.5 Result and Discussion

5.5.1 Empirical Result

Text Analysis

First of all, we conduct an unsupervised learning on the text corpus to understand the basic nature of language used in this specific type of documents. Using WordCloud text analysis, we identify the high level concepts in our corpus and reveal some meaningful sustainability topics. Figure 5.7 shows the generated word cloud with the top 100 most-used terms in the text data after removing all common English stopwords. The bigger the font size indicate the more frequently that topic has been mentioned across all 500 company reports.

Figure 5.7 : Word Cloud of Corporate Social Responsibility Text

We can see from the word cloud that corporate responsibility are the main topics of most reports. Besides some general terms, e.g. " Social Responsiblity" and "Corporate Social", terms that are more specific to a sub-topics of sustainability, e.g. "GHG emission", "human right" or "health safety", also appear in this top 100 word list. This shows to potential of mining the text with multiple sustainability terms under different sub-topics and evaluate the performance of company in corporate social responsiblity in various categories.

To test that proposition, we perform correlation analysis on all text features using the Pearson's r (Pearson 1895). This correlation coefficient is the most well-known metric to measure the degree of correlation between the two underlying features. The point-biserial relationship is calculated with the Pearson's r formula except that one of the features is dichotomous. The Pearson's r is often obtained using a Least-Squares fit. A value of 1 indicates a perfect positive linear correlation, and a value of -1 indicates a perfect negative correlation. If the Pearson's r equals to 0, there is no correlation between the text feature and the overall ESG scores. The following equation is used to compute the value r of Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{5.30}$$

where $N = 500$ is the total number of company reports in the dataset, x_i is the text feature and y_i is the overall ESG Score for company i, $\overline{x}$ and $\overline{y}$ represent the average values of all text feature x and ESG score y respectively. Table 5.8 presents the top 15 positive correlated and the top 15 negative correlated text features from all three types of text dictionaries.

The results in Table 5.8 show that all of the top positive correlated features are extracted using the CSR-Sent dictionary. This have proven our intuition that leveraging a specialized corporate responsibility dictionary would be more suitable for any text mining task in sustainability topic. The high correlation between the overall ESG Score and the terms in sustainable topics, e.g "Emission" or "Water", also show that sustainability text features in company reports can be a significant indicators and can be used to accurately predict the ESG scores in our model.

On the other hand, most of the top negative correlated features are from the LIWC-2015 and Empath dictionaries, which means these text features are capturing different insights from the reports. This shows the possibility of improving the ESG Score prediction accuracy when we stack the text features from all three dictionaries together. Moreover, most these negative correlated features are finance-related features, e.g. "payment", "poor", "money" or "banking". The potential explanation

Table 5.6 : The top correlated text features with Pearson's r

	Text Dictionary	**Text Feature**	**Pearson's r**
Positive Correlation	CSR-Sent	Sustain	0.3622
	CSR-Sent	Engage	0.3598
	CSR-Sent	Emission	0.3431
	CSR-Sent	Sustainability	0.3348
	CSR-Sent	Water	0.3189
	CSR-Sent	Diversity	0.3092
	CSR-Sent	Goal	0.3064
	CSR-Sent	Workforce	0.3060
	CSR-Sent	Leader	0.2994
	CSR-Sent	Work	0.2967
	CSR-Sent	GRI	0.2951
	CSR-Sent	Carbon	0.2903
	CSR-Sent	Hazardous Waste	0.2861
	CSR-Sent	Reuse	0.2826
Negative Correlation	CSR-Sent	Resonable	-0.2198
	Empath	wealthy	-0.2349
	CSR-Sent	Covenants	-0.2502
	Empath	leisure	-0.2519
	Empath	real_estate	-0.2579
	Empath	money	-0.2655
	Empath	love	-0.2664
	Empath	valuable	-0.2664
	Empath	banking	-0.2714
	LIWC-2015	money	-0.2790
	Empath	poor	-0.2792
	CSR-Sent	Common	-0.2838
	LIWC-2015	discrep	-0.2839
	LIWC-2015	time	-0.2854
	Empath	payment	-0.2885

for this is that companies whose reports are focusing on financial performance are more likely to have a lower ESG ratings. This analysis result aligns with that of the general ESG rating process done by a human agent. If the company does not report much about their corporate social responsibility activities, the human agent will rate it lower than one that include a lots of information about sustainability in their reports. Overall, the correlation analysis results provide a strong support on our hypothesis of using text mining and text features to build an ESG Score prediction model in the next step of the experiment.

Evaluation of ESG Score Prediction Model

To avoid sampling and algorithm bias, we test the model using 10-fold cross validation approach with two different machine learning algorithms, e.g. Random Forest and Extreme Gradient Boosting (XgBoost). To test the effectiveness of all three text dictionaries, we build four different models using each one of the three text feature sets as input value and one final model with the stacked features. To further test our proposed methodology on different target values, we build three similar sets of models to predict the three ESG sub-scores, namely "Environmental Score", "Social Score" and "Governance Score". The performances of all models are presented in Table 5.7 below.

The empirical experiment results suggest that text mining model using XgBoost are all performing better than the ones built with Random Forest. Comparing the individual models using only on text feature set as input, CSR-Sent features lead to a more accurate prediction than LIWC-2015 and Empath ones. This findings align with the results from our previous text analysis that CSR-Sent text features are highly positive correlated to the actual ESG Scores.

For "Overall ESG Score" prediction, the model using stacked features from all three dictionaries is the best performing one with MAE = 6.0337 and RSME = 7.8551. Similarly, the prediction model with stacked features and XgBoost algorithm have the lowest MAE and RSME in predicting "Environmental Score" and "Governance Score". Even though the increase in the prediction accuracy is not a

Table 5.7 : 10-fold Cross Validation Evaluation of ESG Score Prediction Models

Prediction Model	**Algorithm**	**Input Feature**	**MAE**	**RMSE**
Overall ESG Score	XgBoost	**Stacked features**	**6.0537**	**7.8551**
		CSR-Sent	6.2323	7.9568
		LIWC-2015	6.7420	8.4078
		Empath	6.3108	8.0544
	Random Forest	**Stacked features**	**6.2671**	**8.1621**
		CSR-Sent	6.3835	8.1460
		LIWC-2015	6.8077	8.6576
		Empath	6.5413	8.2910
Environmental Score	XgBoost	**Stacked features**	**9.5829**	**12.2086**
		CSR-Sent	9.6317	12.2575
		LIWC-2015	10.3432	12.8846
		Empath	9.9543	12.5403
	Random Forest	**Stacked features**	**9.6053**	**12.1084**
		CSR-Sent	9.7339	12.5238
		LIWC-2015	10.9124	13.5261
		Empath	10.2814	12.8374
Social Score	XgBoost	Stacked features	8.0265	10.2129
		CSR-Sent	**7.9917**	**10.0440**
		LIWC-2015	8.1890	10.3075
		Empath	7.9951	10.0683
	Random Forest	Stacked features	8.3048	10.3350
		CSR-Sent	**8.2511**	**10.3070**
		LIWC-2015	8.3369	10.6424
		Empath	8.5483	10.7321
Governance Score	XgBoost	**Stacked features**	**5.9403**	**7.4560**
		CSR-Sent	5.8621	7.4161
		LIWC-2015	6.2635	7.7042
		Empath	6.2857	7.8410
	Random Forest	**Stacked features**	**6.1167**	**7.6969**
		CSR-Sent	6.1685	7.7781
		LIWC-2015	6.2073	7.9300
		Empath	6.2442	7.8739

big jump, it prove our proposed methodology of using stacked text features can help improve the model performance in most case.

In the special case of "Social Score", the stacked features model performs slightly worse than the one with CSR-Sent text feature as input only. This is due to the fact that CSR-Sent dictionary contains mostly terms that are closely related to social affairs of the company. Stacking with other features from LIWC-2015 and Empath dictionaries, which consist of other general terms and unrelated categories, might add noise to the data and worsen the prediction accuracy. This finding suggests that a development of an ESG-specialized dictionary can further improve the performance of text mining and ESG ratings prediction models.

Future research might expand to use time series data such as the multiple years of reports from a single company to conduct a longitudinal analysis of the proposed model. This would also require expanding the dataset to incorporate text data from news and social media channel sources. In that case, different ways of splitting of train and test datasets as well as suitable evaluation methods must be selected. Additional testing with different algorithms and parameter tuning might reduce the prediction errors even more significantly.

Evaluation of P/ESG Indicator Model

We conduct a quick correlation analysis to test the hypothesis from the current literature. The statistical results in Table 5.8 present the Pearson's r (Pearson 1895), Spearman's rho (Spearman 1904), Kendall's tau (Kendall 1938) test statistics together with their two-tailed p-values. The tests show there is a slight positive correlation between ESG Scores and the stock prices with Spearman's rho = 0.1992, Kendall's tau = 0.1360 and p-values close to 0. However, both statistics are lower than 0.5, which means the correlation is not significant. We cannot conclude on the direct correlation between ESG Scores and stock prices or returns. Therefore, we propose to test the correlation in an indirect methods using P/ESG ratio in the next part.

We evaluate here a test of the model's performance on forecasting the annualized

Table 5.8 : ESG Scores Correlation Analysis Results

	Pearson's r		**Spearman's**		**Kendall's**	
Metrics	r	p-value	rho	p-value	tau	p-value
Price	0.0037	9.01E-01	0.1992	1.71E-11	0.1360	9.41E-12
Daily Return	0.0002	9.95E-01	0.0314	2.93E-01	0.0182	3.62E-01
Monthly Return	0.1059	3.83E-04	0.1115	1.85E-04	0.0751	1.66E-04
Quarterly Return	0.1030	5.53E-04	0.1238	3.23E-05	0.0846	2.23E-05
Yearly Return	0.0203	4.97E-01	0.0091	7.62E-01	0.0058	7.73E-01

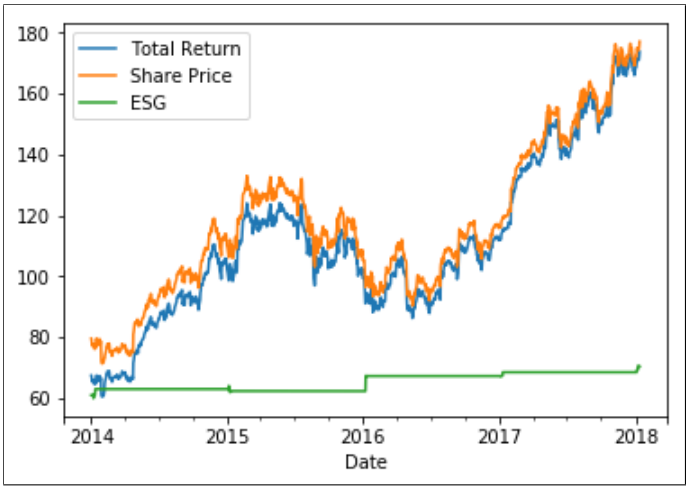
returns of the Apple Inc. (ticker: AAPL) and the T-Mobile U.S., Inc. (ticker: T) for the period between 1/1/2014 and 31/12/2018. Both companies are big technology and telecommunication company. According the Figure 5.8, the share prices of AAPL have been on the uptrend most of the time, those of T have been fluctuated over the last 5 year period. Meanwhile, the ESG scores of both stocks are flatten out, do not fluctuate much or show a clear linear correlation with the other two time series. This confirms the correlation analysis results in Table 5.8 It is hard to find a statistical significant result using correlation analysis between ESG scores and different stocks prices and returns. However, we believe the P/ESG ratio can still be a good indicator for forecasting long-term stock returns test the signals in comparison with the P/B ratio.

We test the predicted annual stock returns using the three models: "P/ESG model", "P/Book model" and the baseline. This experiment provide the evaluation on whether the P/ESG and P/Book are suitable as indicators for forecasting long term stock returns. We start by assuming our model fits the historical data just as well as the "P/Book model" and the baseline, so that the MSE values are equaled. This is our new null-hypothesis and is denoted H_0'.

We then conduct a hypothesis test by calculating the p-value, which is the probability of observing these particular MSE values for the "P/ESG model", the "P/Book model" and the baseline, if indeed the MSE values were equal. This calculation takes

Figure 5.8 : Plots of stock prices and ESG Scores time series

Apple Inc. (AAPL)



T-Mobile U.S., Inc. (T)

into account the difference in MSE values, the variance in the data, and the number of data-points. A p-value close to zero means that it is highly unlikely that we would observe those particular MSE values, if indeed they were identical, so we can reject the null-hypothesis $H_0^{'}$ and instead accept the alternative hypothesis $H_1^{'}$ that the MSE values are most likely different, and hence our model is either better or worse than the baseline. The results in Table 6.1 and Figure 5.9 shows that our "P/ESG model" outperforms the "P/Book model" and the baseline overall.

The 1 year statistics show us that the MAE and MSE of "P/ESG model" are bet-

Table 5.9 : Forecast models evaluation

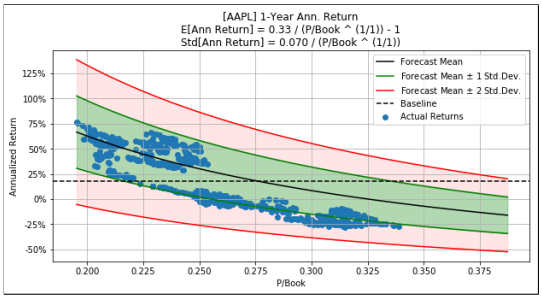
Company	**Models**	**MAE Metric**	**MSE Metric**	**MAZ Metric**	R^2 **Metric**
Apple	P/ESG model	0.1750	0.0379	0.89	0.60
	(*p-value*)	(*1.08E-153*)	(*2.66E-125*)	(*3.84E-04*)	
	P/Book model	0.1800	0.0416	0.69	0.56
	(*p-value*)	(*4.77E-65*)	(*2.49E-72*)	(*1.32E-31*)	
	Baseline	0.2870	0.0950	0.93	
T-Mobile	P/ESG model	0.0770	0.0093	0.83	0.46
	(*p-value*)	(*4.88E-43*)	(*3.88E-41*)	(*3.26E-02*)	
	P/Book model	0.6760	0.0481	0.39	-27.33
	(*p-value*)	(*0.00E-00*)	(*4.33E-277*)	(*5.41E-106*)	
	Baseline	0.1150	0.0172	0.88	

ter than the "P/Book model" and the baseline in general. Particularly in predicting returns for T-Mobile U.S. Inc., the "P/ESG model" has an MAE of 7.7%, almost 10 times smaller than the MAE 67.6% of the "P/Book model". The MAZ value of "P/Book model", which take both the predicted mean and standard deviation into account, is higher than that of the "P/ESG model" in case of Apple returns forecast and lower in case of T-Mobile one. However, the R^2 in that case is negative, which is caused by the forecasting model's poor fit to the high-variance data, so the result for "P/Book model" might not be intuitively explainable. Meanwhile, the R^2 of "P/ESG model" in both stocks are 0.6 and 0.46 respectively, which means that a significant part of the data's variance is explained by the model. This result is quite significant for such a short investment period.

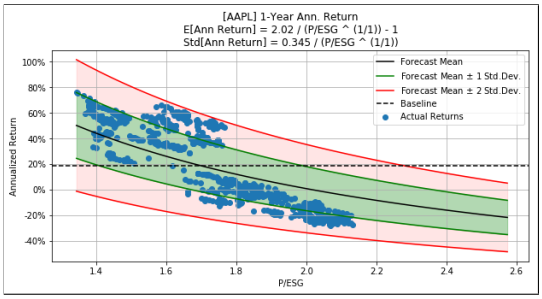
Overall, both models have lower MAZ results, which mean they fit the whole distribution slightly better than the baseline. the models using either P/B or P/ESG ratios are considerately accurate for predicting the annual returns of both stocks. The model p-values are always near zero so this result is not due to random chance. We can accept $H_1^{'}$ that the prediction accuracy of forecast model using P/ESG ratios is better. To further confirm our conclusion with the original hypothesis H_0, we repeat the correlation tests on the big dataset with $1,121$ stocks using the scaled

Figure 5.9 : Plots of forecast models evaluation

Apple Inc. (AAPL)

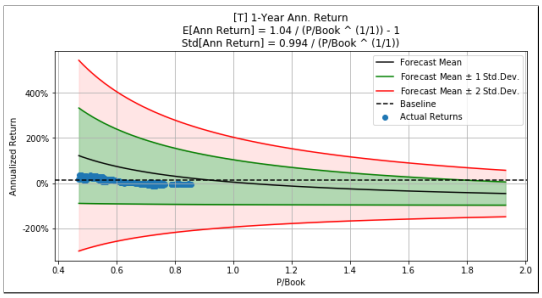


(a) P/Book model

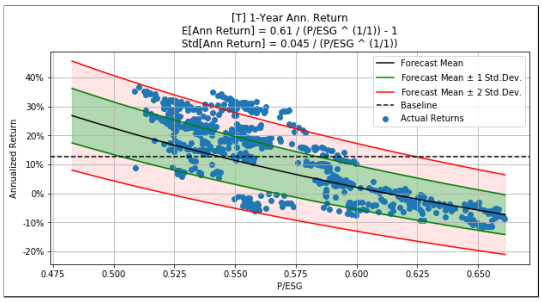


(b) P/ESG model

T-Mobile U.S., Inc. (T)



(c) P/Book model



(d) P/ESG model

P/ESG ratio. The test results are presented in Table 5.10 below.

Evaluation of MV-ESG model

We split our 10-year historical financial data into 2 subsets at the ratio 8:2, which means the model training period is from 2009 to 2016 and the testing period is from 2017 to the end of 2018. The plot in Figure 5.10 compares the accumulated portfolio returns over the testing periods from 2017 to 2018. According to the results, the "MV Portfolio" and the "MV-ESG Portfolio" achieve quite similar values with only a small gap in the final portfolio value. This shows that socially responsible investment can achieve comparatively sufficient financial performance. Furthermore, the restrained limit of selecting only stock with loss probability less than 15% also resolve the investors' concern of bearing a high risk in investing risky companies

Table 5.10 : P/ESG Scores Correlation Analysis Results

	Pearson's r		**Spearman's**		**Kendall's**	
Metrics	r	p-value	rho	p-value	tau	p-value
Price	0.2484	3.14E-017	0.6079	3.23E-114	0.4328	1.92E-104
Daily Return	0.0579	5.24E-020	0.1298	1.30E-050	0.0863	1.53E-050
Monthly Return	0.6093	7.33E-115	0.6038	2.68E-112	0.4479	1.20E-111
Quarterly Return	0.6858	1.39E-156	0.7073	9.74E-171	0.5462	4.26E-165

with higher corporate responsibility.

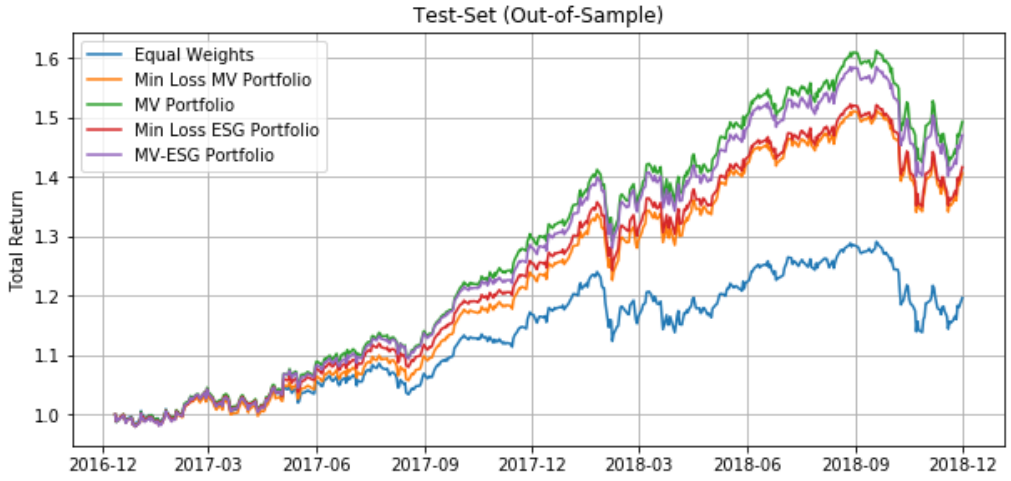
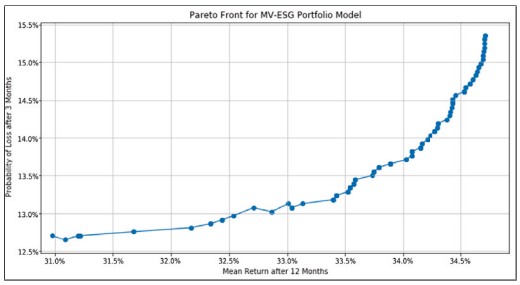


Figure 5.10 : Accumulated Portfolio Values from 2017 to 2018

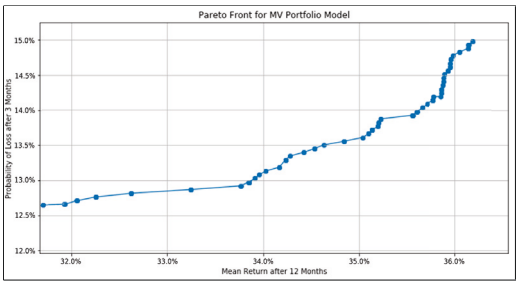
As accumulated returns might not fully reflect the financial performance, we take a further look at the Pareto Fronts of the "MV-ESG Portfolio" and "MV Portfolio" in Figure 5.11. As expected, the "MV-ESG Portfolio" appears to have less returns and bear higher risks in general. However, the differences in the Pareto Fronts are not totally distinguishable to conclude about the test significance.

Our main target with the "MV-ESG portfolio" is to achieve not only a great financial performance but also a good ESG rating. Therefore, we calculate the trade-off Portfolio ESG Scores to further evaluate the performance of the multi-

Figure 5.11 : Pareto Fronts of the "MV-ESG Portfolio" and "MV Portfolio"



MV-ESG Portfolio



MV Portfolio

objective optimization model. The results in Table 5.11 below show the returns, risks, Sharpe ratios and ESG Scores of the tested portfolios.

Table 5.11 : MV-ESG Portfolio Evaluation

	Return	**Risk**	**Sharpe Ratio**	**ESG Score**
MV-ESG Portfolio	34.6848%	13.1387%	2.4877	63.3815
MV Portfolio	35.9888%	12.9562%	2.6234	61.9099
Min Prob Loss ESG	30.9385%	12.8387%	2.2540	63.8567
Min Prob Loss	31.3214%	12.7737%	2.2954	61.7841

According to Table 5.11, we can see that the "MV Portfolio" and "Min Prob Loss" have better financial performances than the "MV-ESG Portfolio" and the "Min Prob Loss ESG" respectively. However, considering our context of SRI, a good performance portfolio has to take into account the ESG Scores. We can see an increase of 1.4716 and 2.2076 in the ESG Scores for the two "MV-ESG model" portfolios, while the decrease of Sharpe Ratios are 0.1357 and 0.0414 respectively. To most sustainable funds and ethical investors, this level of financial sacrification might be acceptable in order to achieve the addition goal of SRI.

Evaluation of Combined MV-ESG model

We plot the relative portfolio allocation of the "MV Portfolio" and "MV-ESG Portfolio" based on industry to understand the baseline of our model. According to

the result in Figure 6.8a, we can see that the "MV-ESG Portfolio" using the ESG Scores only has incorporated some sorts of positive and negative screenings without any explicit weighting rules. Particularly, the "MV-ESG Portfolio" invest significantly more in "Papers and Forestry", "Automobiles" and "Auto Components". These industries have been reported to be working on different environmental and social solutions in recent years, e.g. renewable energy, recyclable paper, electric cars, etc.

On the negative screening side, the "MV-ESG Portfolio" invests significantly less in "Steel", "Industrial Conglomerates", "Homebuilders", "Diversified Metals", "Construction Materials" and "Building Products". This negative screening is effective if the investors want to avoid these listed industries which businesses can directly harm the environment.

We further check the negative screening potential of the "MV-ESG Portfolio" by plotting the relative portfolio allocation based on negative topics. The results in Figure 5.13 show that "MV-ESG Portfolio" invest slightly less in most cases, and significantly less in "Pesticides", "GMO" and "Adult Entertainment" companies.

This shows that our "MV-ESG Portfolio" can achieve similar level of portfolio diversification as the standard "MV-ESG Portfolio" by investing in multiples companies across different industries. We can reject the null hypothesis H_0 as integrating ESG Scores does not worsen the portfolio diversification. Even though it has already practiced the positive and negative screenings by integrating ESG Scores, we want to further allow the incorporating of investors' preferences into our model to provide flexibility for further portfolio diversification. Therefore, we proposed the "Combined MV-ESG Portfolio" constructed as follows.

We initially construct the "Positive MV-ESG Portfolio" and "Negative MV-ESG Portfolio" separately. Considering an investor who cares about world hunger and medical conditions, he/she wants to invest more in companies in the industries of "Healthcare", "Pharmaceuticals", "Food Retailers" and "Food Products". The weighted positive preference parameter for these companies is set as $\gamma = 2$ which allows double the investment in these companies. This investor also want to exclude

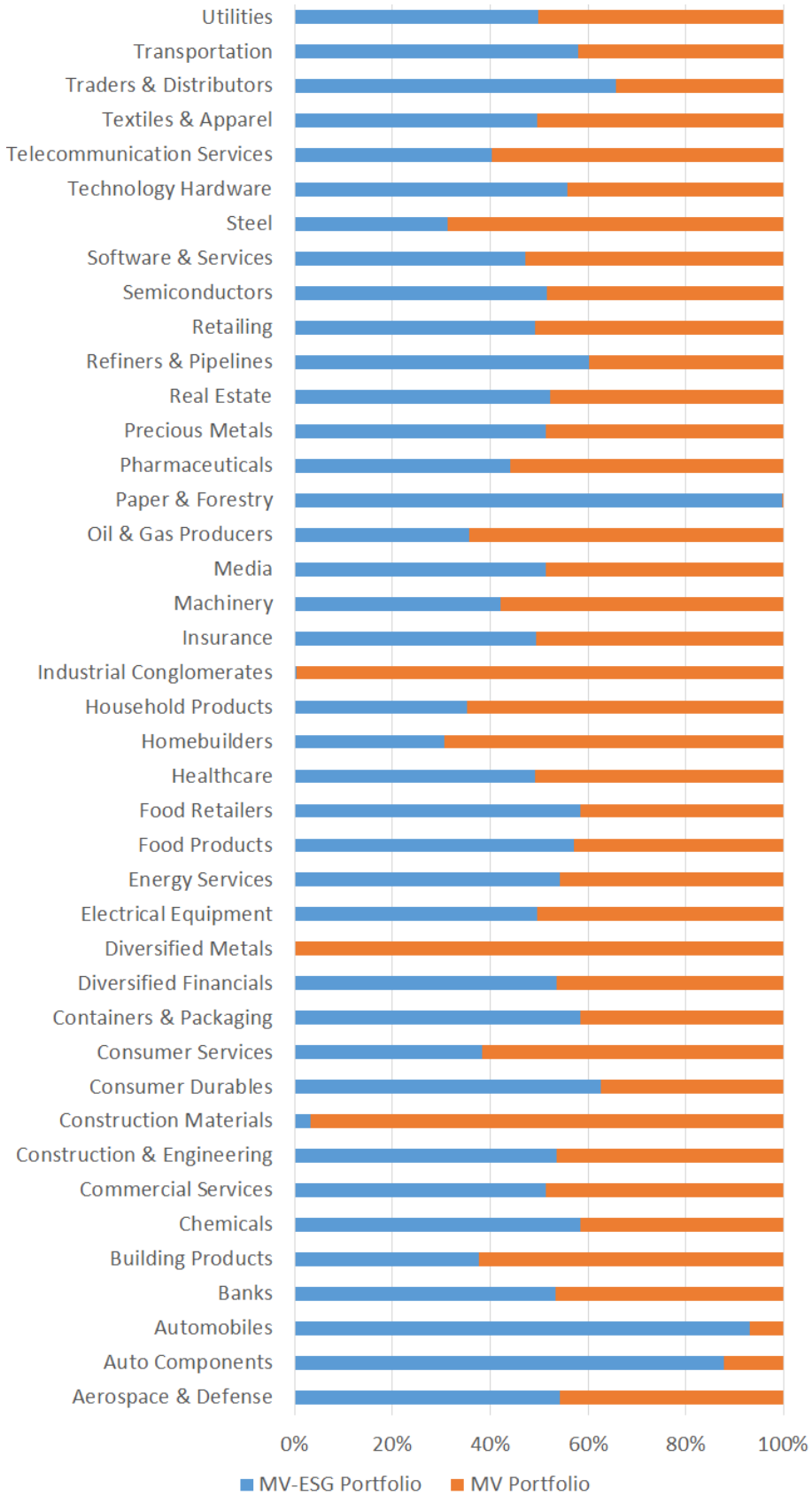


Figure 5.12 : MV and MV-ESG portfolio allocation based on industry

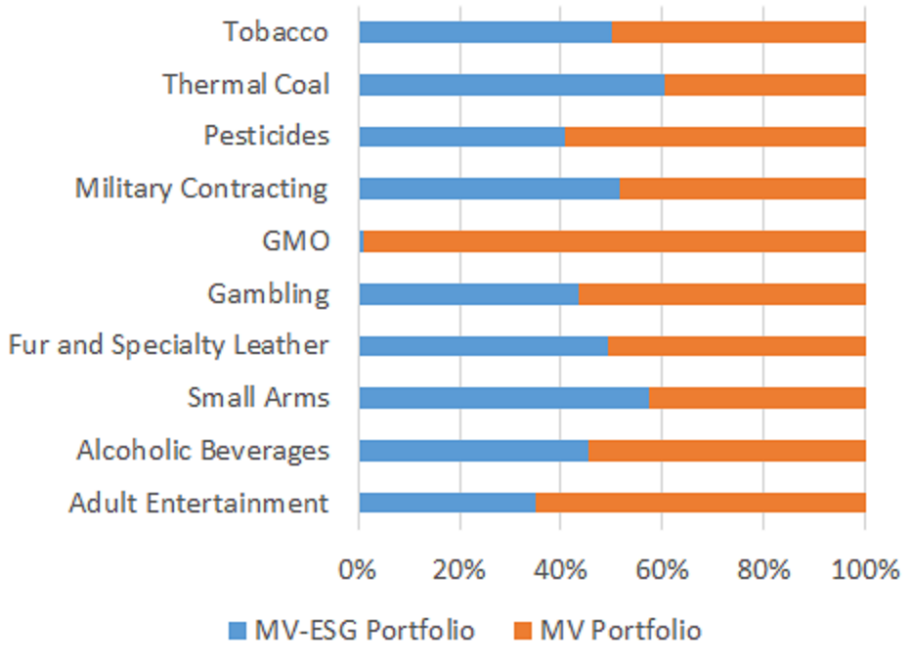


Figure 5.13 : MV and MV-ESG portfolio allocation based on negative topics

all the companies involved in any negative screening topics (see Figure 5.13. The weighted negative preference parameter for these companies are set as $\Gamma = 0$ which allows no investment in these companies.

We then combine the weighting limit of both portfolios to calculate the allocation limit and construct the "Combined MV-ESG Portfolio". The financial performances and ESG Scores of all the tested portfolio are presented in Table 5.12. Overall, the ESG Scores of these portfolios are still higher than the "MV Portfolio", which shows that the weighting approach does not affect out ESG Score objective significantly this multi-objective portfolio optimization model.

According to Table 5.12, the "Combined MV-ESG Portfolio" and "Positive MV-ESG" achieve better financial performance compared to the standard "MV Portfolio". Specifically, they have higher returns, lower risks and higher Sharpe Ratios. This result is already suggested by Garcia et al. (2017) that companies operating in more sensitive industries can have higher ESG Ratings with evidence from emerging markets. Companies in "Healthcare" and "Food Products" industries are receiv-

Table 5.12 : Combined MV-ESG Portfolio Evaluation

	Return	**Risk**	**Sharpe Ratio**	**ESG Score**
Combined MV-ESG	37.5754%	12.9562%	2.7458	63.1241
MV-ESG Portfolio	34.6848%	13.1387%	2.4877	63.3815
Positive MV-ESG	36.6256%	12.9562%	2.6725	63.1546
Negative MV-ESG	35.0014%	12.6861%	2.6014	63.1723
MV Portfolio	35.9888%	12.9562%	2.6234	61.9099

ing more attention and might be experiencing a higher growth in recent years than companies in "Industrial Conglomerates"

The "Negative MV-ESG Portfolio" still achieves a lower financial performance compared to the standard "MV Portfolio". However, it has slightly improved from the baseline model of "MV-ESG Portfolio". This result is explainable as companies involved in these negative topics also bear higher risks in finance and operations (e.g. the profit of tobacco companies might be severely affected by a new smoking in public rule). Therefore, by limiting the investment amount in these companies might actually limit the exposure for such risk and increase the financial performance of the portfolio.

We finally confirm the portfolio diversification of the "Combined MV-ESG Portfolio" by plotting its allocation by industry. The plot in Figure shows the final portfolio allocation is sufficiently diversified with investment in multiple industries. Even with the over weighting in "Healthcare" (14.4%) and "Pharmaceuticals" 11.1) according to the investors' preferences, this is still more significantly more diversified then a purely healthcare-related fund. This prove our quantitative methods are effective in portfolio optimization and diversification.

5.5.2 Discussion

In this research, we mainly focus on introducing the application of text mining and machine learning to solve the ESG rating problem. With our proposed approach,

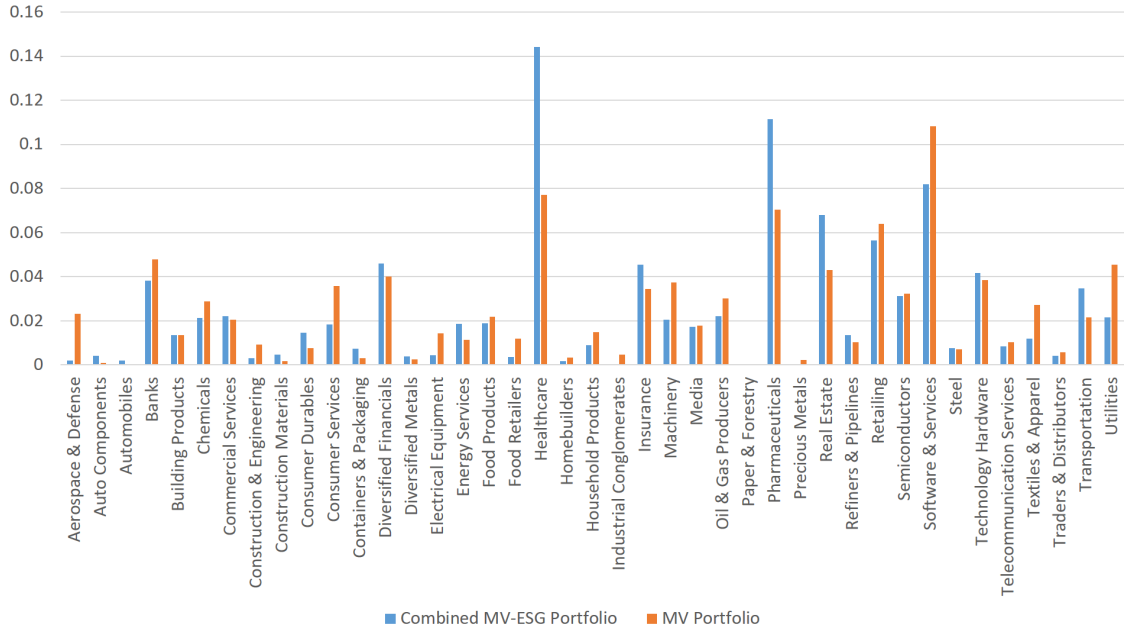


Figure 5.14 : MV and Combined MV-ESG portfolio allocation based on industry

the human agents might not need to spend too much time reading through thousands reports and rate all companies manually. With an absolute error MAE as low as 5.9 points, the human agents can utilize this model to transfer the learning from previously rated reports and predict the ESG Scores for new companies in an effective and efficient manner.

The financial analysis further show that the scaled P/ESG ratio are correlated with the stock prices and returns, except for the daily return time series. This result is significantly insightful in our case as socially responsible investors are more interested in long-term financial performance predictions than daily returns prediction. We now can reject the null hypothesis H_0 and accept H_1 that ESG Scores can be an indicator for financial performance. We can use the ESG score as signal to forecast the long term stock returns in our ESG-based multi-objective portfolio optimization model.

With the empirical experiments on "MV-ESG model", we can confirm that the performance of ESG-based multi-objective portfolio are comparatively similar to

the standard financial portfolio. After satisfying the basic financial and ESG Scores performance evaluation, we then further optimize our "MV-ESG Portfolio" in the specific context of SRI using the "Combined MV-ESG model". The results show that we can still achieve an equivalent level of portfolio diversification in ESG-based portfolios compared to normal financial ones.

Furthermore, there are not many research on quantitative models to measure the impacts of sustainable investment according to Fowler and Hope (2007). Most publication works are assessments or comparative studies of ESG-focused funds benchmarking against conventional funds using qualitative methods (Koellner et al. 2007). Research in this topic has been facing many challenges, including data availability, standardization of the evaluation metrics and lack of data-driven approaches (Peloza 2009). A context-based metrics framework has been proposed by (Vörösmarty et al. 2018) to measure the impacts on waste water treatment, renewable energy and pharmaceuticals. However, this framework will have some limitations when it comes to companies in different industries or when the companies do not report the required data. A more data-driven and industry adjusted metrics model can be more accurate in measuring the impacts of sustainable investment. We present this proposed framework in the Appendix A for potential future research.

The results in this chapter serve as the research foundation and motivation for the next chapter where we further develop deep learning and reinforcement learning methods combined with quantitative finance model for ESG-based socially responsible investment and portfolio optimization.

Chapter 6

Deep Learning for Decision Making and Optimization of Socially Responsible Investment Portfolio

6.1 Background and Motivation

Traditionally, investors have focused on the investment returns by actively looking at the financial reports to find the best performing stocks. With the recent mindset change towards sensitive topics like global warming or refugees, investors are becoming concerned with other aspects of companies rather than just earnings. They are shifting their investment towards companies which are actively doing good things for the environment, contributing to the society and operating with transparency. According to the 2018 Biennial Report On US Sustainable, Responsible and Impact Investing Trends (US SIF Foundation 2018), socially responsible investment (SRI) assets accounted for $12 trillion out of $47 trillion in total assets under professional management in the United States in 2018, representing a sharp increase of 38% since 2016.

Conventional investment and portfolio theory focuses on financial performance, i.e., the returns and risks of the portfolio (Zopounidis et al. 2015). Direct application of the theory might not be suitable for SRI because it focuses more on non-monetary objectives (Calvo et al. 2016). Therefore, socially responsible investors need a modified version of the modern portfolio theory that can serve their purpose better (Peylo 2012). Besides, SRI investors currently have to read corporate social responsibility (CSR) reports to find good companies to invest in, which is time-consuming and difficult. The lack of effective quantitative approaches for SRI makes it more difficult for not only professional investors, but also the vast majority of lay investors. There-

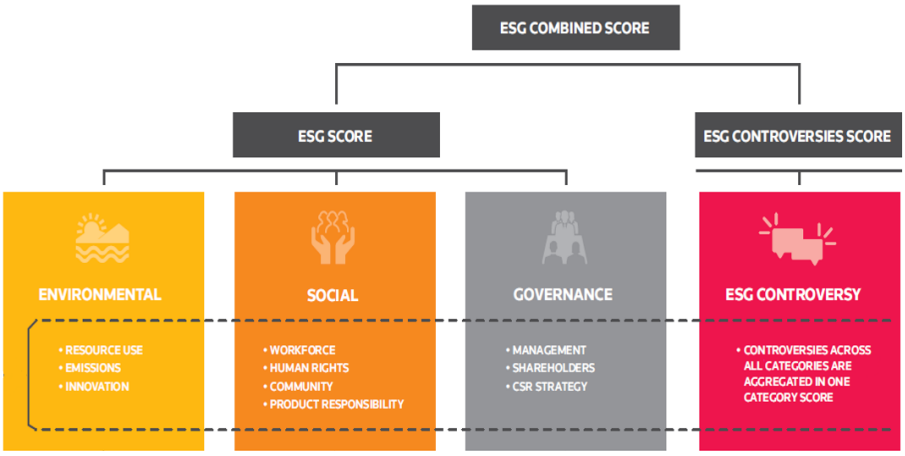


Figure 6.1 : Combined ESG ratings

(a) (Thomson Reuters 2019)

fore, this research will provide an easy and automated way of doing such investments in an ethical manner, which greatly benefits their decision-making and secures the optimal investment returns. This is one of the main motivational purposes for this research.

Recently, the Global Reporting Initiatives (GRI) and the United Nation Sustainable Development Goals (SDGs) has provided standardized metrics and frameworks for companies to disclose more information regarding their sustainability practices (Dumay et al. 2010). For example, environmental, social and governance (ESG) metrics of companies have been derived from reports and news articles (e.g. CSR reports, news articles, carbon disclosure project ratings), evaluating the company in different prospects (e.g. air emissions and waste management, employee health and safety control, board transparency and diversity) including their controversies (e.g. involvement in adult entertainment or gambling). These metrics have been consolidated into the combined ESG ratings (see Fig. 1). The availability of ESG ratings has led to an emerging research topic in SRI portfolio.

With the availability of ESG metrics, quantitative methods can now be applied effectively to address the SRI portfolio construction problem. Current data mining approaches in this research field face a number of challenges. The first challenge

is the accuracy of multivariate time series predictions. Stock return forecasts have been extensively studied with various quantitative finance and machine learning models (Henrique et al. 2019; Sermpinis et al. 2019). Most of these works have been focused on univariate time series predictions because it is expected that multivariate data would contain too much noise for the neural network to perform well (Gadre-Patwardhan et al. 2016; Moghaddam et al. 2016). However, stock movements in the financial market are highly correlated; thus a multivariate model can learn these deep insights better than a combination of univariate networks. Following recent advances in neural networks research, especially Long Short Term Memory networks (LSTM) (Hochreiter and Schmidhuber 1997), the application of deep learning in the predictive investment field has become an alternative approach to the traditional financial model. In this research, we propose a novel Multivariate Bidirectional LSTM neural network to predict multiple time series for stock returns.

The second challenge faced by current approaches in SRI is the application of multi-objective portfolio construction. Existing portfolio optimization methods are evolving around the standard Mean-Variance (MV) Portfolio (Markowitz 1952), which focuses on maximizing returns and minimizing risks. To incorporate corporate responsibility performance into our optimization problem, we introduced a modified MV model for SRI portfolio construction by integrating ESG ratings.

The third challenge for SRI portfolio is building a model that can adapt to market movements. As SRI in particular, and financial investment in general, are sensitive to market volatility, model parameters should be tuned up periodically to achieve both financial performance and ESG rating objectives. By adopting reinforcement learning techniques, we introduced a Deep Responsible Investment Portfolio (DRIP) model to retrain the prediction model and rebalance the portfolios effectively and autonomously.

An advantage of our proposed approach, which incorporates a multivariate BiLSTM neural network and MV-ESG, is that the framework can be generalized and extended to other scenarios with a similar multivariate prediction and multi-objective optimization problem. The developed deep reinforcement learning framework could

also accommodate different neural networks and AI algorithms to tackle other types of complex and highly intercorrelated problems.

The main contributions of our research are:

- A novel, deep, responsible, investment portfolio framework (DRIP) to integrate deep neural networks, multi-objective optimization, and reinforcement learning. The framework could be applied to other similar contexts of multivariate predictive analytics.
- A novel DRIP model that can forecast the returns quarterly and yearly on investment instead of just daily, which is a more realistic scenario for investors. The model has been fully tested and deployed on real-life datasets containing 100 stocks over a period of 30 years.
- The first report (to the best of our knowledge) leverages deep learning and incorporates ESG ratings into a portfolio optimization model.

6.2 Preliminary

6.2.1 Socially Responsible Investment

The optimization of financial portfolios has been researched extensively. Many approaches have been developed to build decision support systems for stock trading. This includes standard mathematical finance modeling, e.g. Mean-Variance (MV) (Markowitz 1952), AutoRegressive Moving Average (ARMA) and Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models (Francq and Zakoian 2019), text mining of financial news (Nam and Seong 2019) and social media (Ho et al. 2017).

However, limited research has been carried out on socially responsible investment. Although the socially responsible investment was proposed in the 1980s (Gray 1983), it only became a topic of interest for academia and industry in the past decade (Eccles and Viviers 2011). During this time, research has correlated

ESG ratings with the financial performance of companies (Halbritter and Dorfleitner 2015; Fatemi et al. 2018) or socially responsible funds (Kempf and Osthoff 2007; Munoz et al. 2014; Auer and Schuhmacher 2016). The availability of environmental, social and governance (ESG) ratings has enabled more research and application in this area in academia (Von Wallis and Klein 2015) and industry (US SIF Foundation 2018).

Many sustainability funds have offered portfolios with certain values to attract investors to SRI. In management funds (Siddiqui et al. 2011), there has been an increasing demand from sustainably conscious investors to have more SRI options (Nilsson 2008). Multiple sustainable indexes and funds have been constructed based on areas of investor interests (e.g. water treatment, clean tech, renewable energy, gender equality and diversity). The literature has showed that companies or sustainable funds with higher ESG ratings can outperform the lower ones financially in long-term investments (Friede et al. 2015).

The literature of qualitative research in SRI has focused on reviewing the performance of companies (Bose and Pal 2012) and socially responsible indexes or funds (Stephen 2018), and not on a data-driven approach to incorporate sustainability into an investment system. Some of the research has criticized the current stock screening process of SRI funds (Verheyden et al. 2016) and has proposed that the full integration of ESG ratings would be more beneficial (Amel-Zadeh and Serafeim 2018). These findings underpin the main motivation for our research to develop a framework with full integration of ESG ratings. Our research contributes to the current knowledge of the application of deep learning for the prediction of stock returns and ESG-based SRI portfolio optimization.

6.2.2 Deep Learning for Stock Returns Forecasting

Researchers have undertaken extensive studies to solve the time series forecasting problem of stock returns using deep learning (Di Persio and Honchar 2016; Moghaddam et al. 2016; Chiang et al. 2016). Many have suggested that different types of Recurrent Neural Networks (RNN) outperform traditional financial time

series models in different markets (Chen et al. 2015; Bao et al. 2017; Sermpinis et al. 2019). RNN contains feedback loops in its recurrent layer, which enables the storage of information in the "memory cell" over time. However, it does not perform well when the learning requires long-term temporal dependencies.

Long Short-Term Memory (LSTM) is a special type of RNN that has been proven to be effective in text mining to predict stock returns (Kraus and Feuerriegel 2017). LSTM contains "memory cells" that are able to retain information for longer periods of time (Hochreiter and Schmidhuber 1997). Consequently, LSTM often performs better in sequential data and financial time series predictions compared with RNN (Nelson et al. 2017; Jiang et al. 2019), particularly in the SRI context where investors are concerned more about long-term returns rather than the volatility of the short-term market.

Researchers have also compared the performance of different RNN architectures like LSTM and Gated Recurrent Unit (GRU) networks (Samarawickrama and Fernando 2017). Others have suggested that Bi-directional LSTM (BiLSTM) might be a better option in a similar sequence prediction problem (Chen et al. 2017). While the LSTM and GRU, with the unidirectional flow of information, might be adequate in most sequence prediction problems, the BiLSTM model reads the data one more time backward (Schuster and Paliwal 1997) which helps improve prediction accuracy, particularly in forecasting sequential data like financial time series.

Recently it was suggested that back-testing results could have given rise to false positives due to the normalization of testing data and prediction of the next time step only (Selvin et al. 2017). The next-time step prediction is only suitable for high-frequency trading strategies using intra-daily data, such as foreign exchange markets. In SRI, investors are more interested in long-term returns on investment. Conversely, research has been conducted on the long-term prediction for financial indexes with 1-year and 2-year time gaps, suggesting that long-term forecasting is possible for stock returns (Feuerriegel and Gordon 2018).

Our research contributes to the current deep learning methodologies through the design of a novel BiLSTM neural network that predicts a long-term multivariate time

series. To avoid false positive results, the financial returns data is not normalized and the model predicts multiple steps ahead. By constructing the baseline models using different types of LSTM networks as undertaken previously, we evaluate the prediction accuracy of the LSTM networks in the forecasting of SRI stock returns.

6.2.3 Portfolio Optimization

Few socially responsible investment models have been developed and proposed that utilize ESG ratings (Van Duuren et al. 2016). (Garcia-Bernabeu et al. 2015), for example, suggested a modification to the standard portfolio selection model with ESG scores. They utilized the Mean-Variance Stochastic Goal Programming (MV-SGP) model with a statistical approach for ESG screening on stocks based on scores and controversy risk. However, they did not consider predictive analytics; they only used past returns and volatility to test their hypotheses. Furthermore, they did not validate their models with real financial data.

Multiple optimization functions are available, including the Expectation Maximization (EM) algorithm (Dempster et al. 1977), quasi-Newton (Broyden et al. 1973) or Powell methods (Powell 1964). However, most of them are not multi-objective or allow the special limit conditions that are required in a complex context like in socially responsible investment. The Sequential Least SQuares Programming (SLSQP) method proposed by (Kraft 1988), for example, can be used to minimize a function of various variables with a different combination of bounds, equality and inequality constraints. However, its greedy behavior leads to a skewed distribution for the weights of stocks in the portfolio. This is not an optimum choice for investors who are worried about non-diversified portfolios with extreme exposure risk.

We have developed a financial model to construct a socially responsible investment portfolio that incorporates the Mean-Variance portfolio theory and ESG ratings (MV-ESG). Our model is not based on the ESG screening approach. Instead, it filters and leverages the ESG ratings in a multi-objective optimization function based on the SLSQP method. It also considers both past and predicted the future performance of stocks in a portfolio selection. This is one of the first mathemati-

cal models for constructing a socially responsible investment portfolio that achieves both better ESG ratings and competitive financial performance.

6.3 Deep Learning for Socially Responsible Investment Portfolio

Our DRIP framework consists of three main components: a multivariate BiLSTM neural network to predict stock returns quarterly and yearly; these predicted values are then combined with ESG ratings in our MVP- ESG model for portfolio construction; reinforcement learning techniques are then leveraged to automatically retrain the prediction models and re-balance our MVP-ESG portfolios after each period. The full reinforcement learning DRIP framework is as showed in Figure 6.5.

6.3.1 Multivariate BiLSTM for long-term returns prediction

Standard feature engineering often includes a normalization step, which transforms the data range to [0,1]. This common approach can help to improve the prediction accuracy of the neural networks. However, in the time series model, this approach implicitly tells the trained model the movement range of future stock prices, which makes out-of-bag testing results unrealistically accurate. We processed the input data for our neural networks in a different approach. In our DRIP model, we did not normalize data but instead fed the stock returns directly into the neural networks. We also trained the model to predict values with a longer time gap instead of a next period prediction, which is a more suitable scenario for stock investors in real-life trading.

Let $p_i(t)$ be the price at time t $(t = 1, ...T)$ for stock i $(i = 1, ...N)$. Δt was the time gap $(1 < \Delta t < T)$. The return $r_i(t)$ for stock i at time t was $r_i(t) = p_i(t) - p_i(t - \Delta t)$. In the DRIP model, we used the sliding window technique to perform a rolling forecast. Let δt be the sliding window size. The train features matrix $X_i(t)$ and return vector $Y_i(t)$ for stock i at time t were:

$$X_i(t) = \begin{bmatrix} r_i(t-T-\delta t) & r_i(t-T-\delta t+1) & \cdots & r_i(t-T) \\ r_i(t-T+1-\delta t) & r_i(t-T+1-\delta t+1) & \cdots & r_i(t-T+1) \\ \cdots & \cdots & \cdots & \cdots \\ r_i(t-\delta t) & r_i(t-\delta t+1) & \cdots & r_i(t) \end{bmatrix} \tag{6.1}$$

$$Y = \begin{bmatrix} r_i(t-T+\Delta t) \\ r_i(t-T+1\Delta t) \\ \cdots \\ r_i(t+\Delta t) \end{bmatrix} \tag{6.2}$$

As suggested by (Nelson et al. 2017), LSTM networks would outperform other neural networks in solving similar problems due to its information persistence characteristic. We considered three types of LSTM neural networks:

- **LSTM**, initially proposed by (Hochreiter and Schmidhuber 1997), is a special kind of RNN, which is capable of learning long-term dependencies. For each input vector x_t at time step t, LSTM network uses multiple gating functions: the input gate i_t, forget gate f_t, and output gate o_t, together with a memory cell C_t to preserve long-term information and keeps track of its flow. The forget gate f_t and input gate i_t generated at each time step t are defined as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{6.3}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{6.4}$$

  In the next step, a tanh layer generates a new memory cell $\tilde{C}_t$. LSTM then updates the old memory cell C_t and generates the output gate o_t and hidden state h_t:

$$i_t = tanh(W_C[h_{t-1}, x_t] + b_C) \tag{6.5}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{6.6}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{6.7}$$

$$h_t = o_t \odot tanh(C_t) \tag{6.8}$$

where σ is the sigmoid function and $\odot$ is the element-wise multiplication. W is the weight matrix and b is the bias vector to be learned by the LSTM at each specific gate.

- **BiLSTM** is a variation of the bidirectional RNN, firstly introduced by (Schuster and Paliwal 1997). It concatenates a forward and backward unidirectional LSTM on the stock return time series $Combined(h_t) = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. Unidirectional LSTM only preserves long-term information of the past, while BiLSTM can preserve information from both past and future by using the combined two hidden states $Combined(h_t)$.

- **GRU** is a more recent alteration of LSTM, suggested by (Cho et al. 2014). It concatenates both the forget gate f_t and input gate i_t into a single update gate z_t, and merges the cell state C_t and hidden state h_t. The architecture of GRU is simpler than the standard LSTM one. The hidden state h_t generated at each time step t is defined as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{6.9}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{6.10}$$

$$\tilde{h}_t = tanh(W \cdot [r_t \odot h_{t-1}, x_t]) \tag{6.11}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{6.12}$$

By simplifying the architecture of the LSTM, GRU may learn the data at the combined gate. However, this single update gate might not learn some hidden information effectively. Hence, the performance of GRU networks may be less effective in forecasting long-term time series.

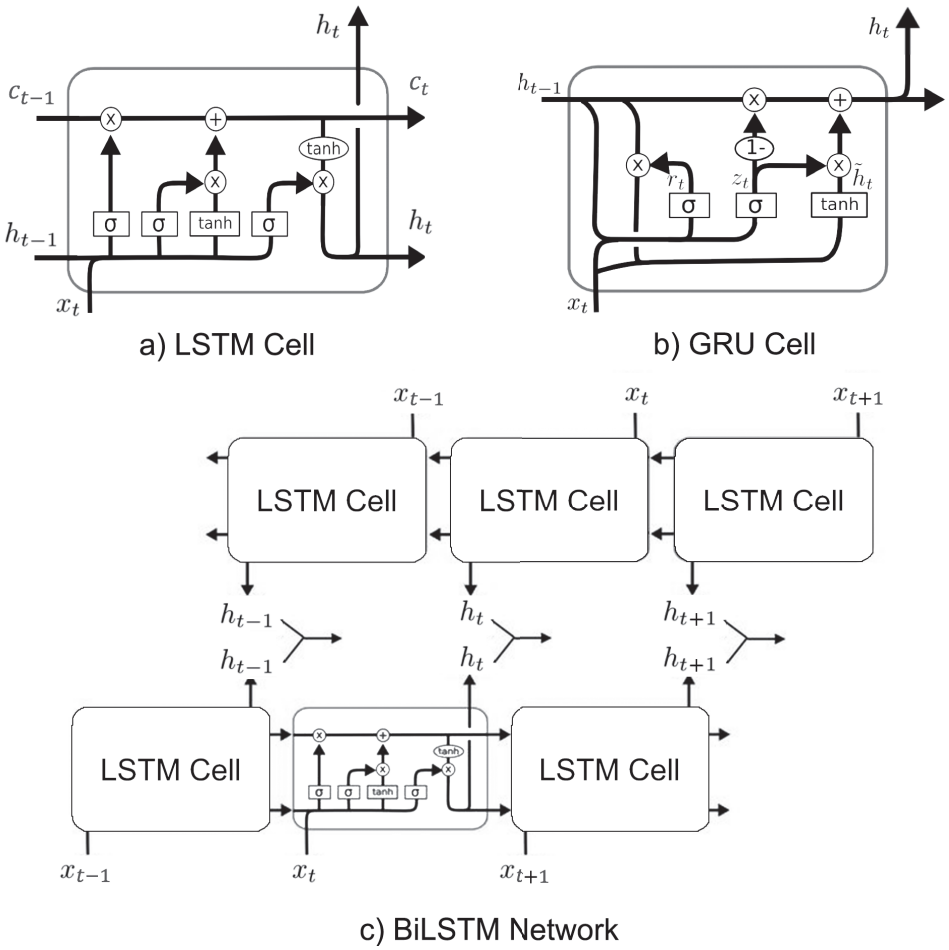


Figure 6.3 : Graphical illustration of LSTM, GRU and BiLSTM

For our DRIP model, we designed a special type of BiLSTM to perform multivariate time series prediction. The data input shape for the multivariate BiLSTM neural network was in the form of a three-dimensional matrix with sizes $(T-\Delta t-\delta t, \delta t, N)$, where N was the number of stocks in total, δt was the sliding window size, and Δt was the prediction time gap (see Figure 6.5).

We also replicated neural network models with LSTM and GRU networks as in (Samarawickrama and Fernando 2017) to predict returns for every single stock in the portfolio. We constructed the neural networks with recurrent layers using the Adam optimizer from the "Keras" package (Chollet et al. 2015). This network also contained a dense layer and a final output layer with the "linear" activation function to predict the stock returns $r_i(t+\Delta t)$ in Δt periods of time.

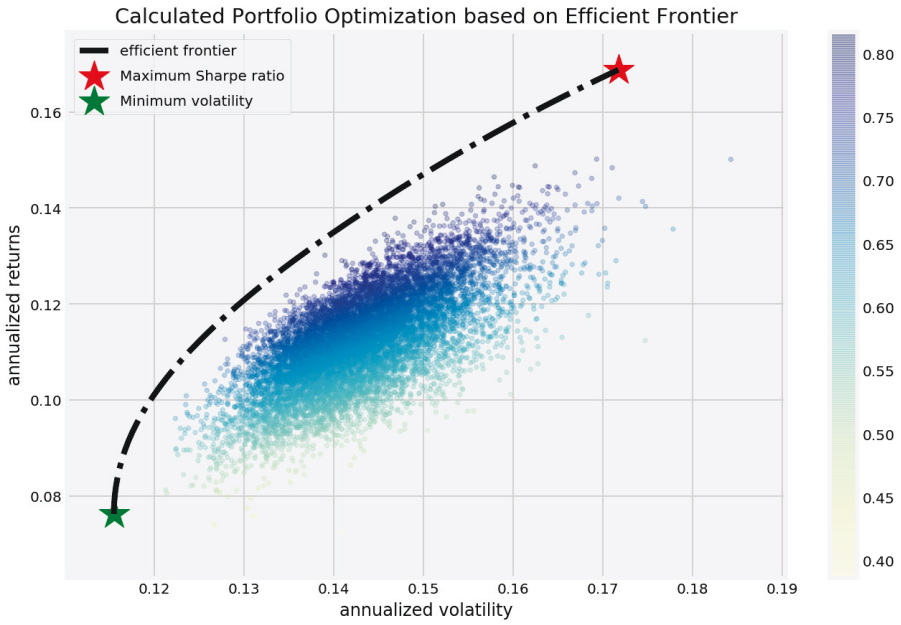


Figure 6.4 : Standard MV Portfolio with Efficient Frontier

6.3.2 MV-ESG Multi-Objective Portfolio Optimization

The Mean Variance portfolio (MV) of (Markowitz 1952) has always been the standard portfolio selection model. Its mathematical principle is constructed by two main components: maximizing the return r_p and minimizing the risk σ_p. The output of this optimization process is the efficient frontier, which is a set of investment portfolios with a greater return than any other with the same or less risk, and a lower risk than any other with the same or greater return. For illustration, the efficient frontier is plotted in Figure 6.4 with the risk on the horizontal axis and the return on the vertical axis.

The optimal portfolio based on the efficient frontier is commonly known as the maximum Sharpe portfolio (MAX-S), where the portfolio has a maximum Sharpe ratio calculated as $S_p = (r_p - r_f)/\sigma_p$. For the MAX-S portfolio, considering the risk free rate r_f (normally the return on bond investment or the bank interest rate), it minimizes the negative Sharpe Ratio (Sharpe 1966):

$$min(-S_p) = min(-\frac{r_p - r_f}{\sigma_p}) \tag{6.13}$$

$$r_p = \sum_{i=1}^{N} w_i r_i \tag{6.14}$$

$$\sigma_p = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i \sigma_{ij} w_j \tag{6.15}$$

where w_i and w_j are the weights of stock i and j, with the boundary limit $w_i, w_j \in [0, 1]$, and σ_{ij} is the covariance matrix of the two stock i and j in the portfolio. The initial weight of each stock in the computation will be equally allocated according to the total number of stocks N in the portfolio, $w_i(0) = w_j(0) = 1/N$.

In our MV-ESG model, we built a multi-objective algorithm based on the SLSQP method (Kraft 1988) with three objectives: maximizing returns, minimizing volatility and maximizing ESG ratings. This algorithm minimized: $min_{wesg}|| - G||$ with G being a three-dimensional matrix of constraints of the three objectives and $wesg_i$ being the ESG weights subject to boundary limits inferred from the companies' ESG ratings.

For comparison, we constructed a maximum ESG portfolio (MAX-ESG) for investors with low risk averse to compare with the standard MAX-S portfolio. In MAX-ESG, we minimized the negative Sharpe Ratio with the portfolio ESG ratings (ESG_p) as a new variable of the objective function.

$$min(-\tilde{S}_p) = min(-ESG_p \frac{r_p - r_f}{\sigma_p}) \tag{6.16}$$

$$ESG_p = \sum_{i=1}^{N} wesg_i \frac{ESG_i + E\bar{S}G_i}{2} \tag{6.17}$$

where ESG_i was the combined ESG ratings of company i in the past year, $E\bar{S}G_i$ was the combined ESG ratings at the current prediction year, and $wesg_i$ was the ESG weight of stock i in the portfolio.

In the traditional MV model, r_p and σ_p are the past returns r_i and volatility σ_i, which is often called ex-post MV. In recent years, researchers and investors have been using the expected returns $\bar{r}_i$ and volatility $\bar{\sigma}_i$. This approach called ex-ante MV is more suitable for predictive analytics in real-world financial trading. In our MV-

ESG model, we combined both ex-post MV and ex-ante MV for portfolio selection and replaced the standard weight boundary with our ESG ones calculated based on the combined ESG ratings for each stock. Our MV-ESG model was computed using:

$$r_p = \sum_{i=1}^{N} wesg_i \frac{r_i + \bar{r}_i}{2} \tag{6.18}$$

$$\sigma_p = \sum_{i=1}^{N} \sum_{j=1}^{N} wesg_i \frac{\sigma_{ij} + \bar{\sigma}_{ij}}{2} wesg_j \tag{6.19}$$

where r_i and $\bar{r}_i$ were the ex-post and ex-ante returns, σ_{ij} and $\bar{\sigma}_{ij}$ were the ex-post and ex-ante covariance matrix of the two stock i and j in the portfolio. $wesg_i$ and $wesg_j$ were the ESG weight of stock i and j in the portfolio, with the boundary limit $wesg_1, \in [0, 1]$ for the company with the highest combined ESG score, then gradually decreasing to $wesg_N \in [0, 0]$ for the company with the lowest combined ESG score. This means the allocation of the company "N" in the portfolio was zero, indicating no investment in that company. The initial weight of each stock in the computation was not be equally allocated but assigned according to the ESG ratings.

6.3.3 Reinforcement learning DRIP model

We combined the multivariate BiLSTM neural networks and the MV-ESG models into a single integrated reinforcement learning model named Deep Responsible Investment Portfolio (DRIP). Starting with a set of agent states S and a set of possible portfolio allocation sets A, we had the probability of the DRIP model select the specific portfolio allocation (the "action") a when in state s at time step t as:

$$\pi : S \times A \to [0, 1] \tag{6.20}$$

$$\pi(a|s) : Pr(a_t = a|s_t = s) \tag{6.21}$$

We defined a simple state-value function V_π^s as the expected reward starting with the state $s_0 = s$ and Re_t denoting the reward function calculated as the sum

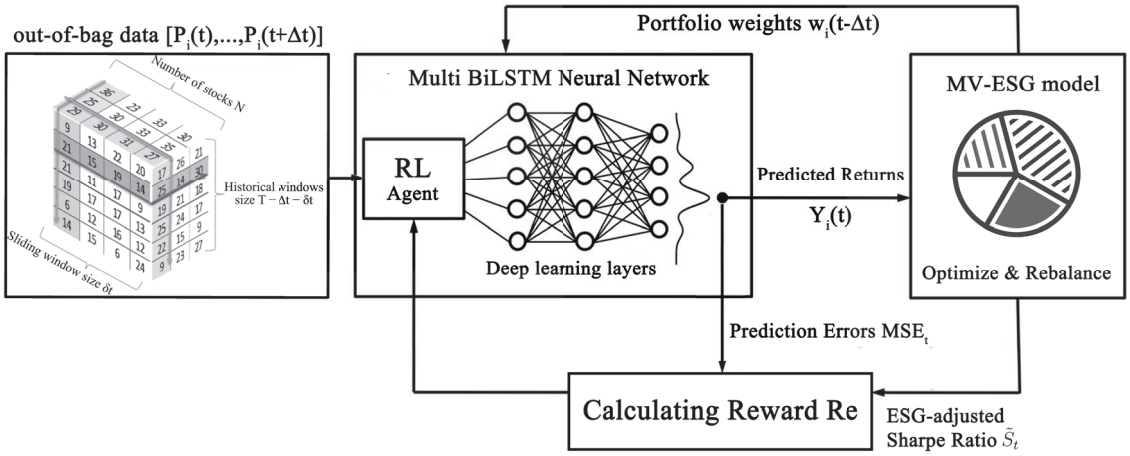


Figure 6.5 : Reinforcement Learning DRIP Model

of future discounted rewards:

$$V_\pi^s = E[Re] = E[\sum_{t=0}^{\infty} \gamma^t Re_t | s_0 = s] \tag{6.22}$$

$$Re = \sum_{t=0}^{\infty} \gamma^t \tilde{S}_t (1/MSE_t) \tag{6.23}$$

where $\gamma \in [0, 1]$ was the discount rate. $\tilde{S}_t$ was the ESG-adjusted Sharpe Ratio (Sharpe 1966), and MSE was the mean squared error of the prediction model. The DRIP model found a set of portfolio allocation to maximize the expected return.

After each time gap Δt, DRIP retrained the prediction model with new stock prices data and then, together with the portfolio performance and stock weights from the previous period of time, constructed a new portfolio with updated allocation weights. The reinforcement learning was repeated on a predefined period basis, to improve the prediction model accuracy and the performance of the portfolio over time. The design of DRIP enabled its self-learning with the least human involvement as possible. The reinforcement learning model is showed in Figure 6.5.

6.4 Experiment

We designed experiments to test our proposed model in two parts: 1) DRIP model forecasts for quarterly and yearly returns of multivariate stock time series during the three year period from 2016 to 2018; and 2) Socially Responsible Portfolios optimization using the predicted returns and reinforcement learning DRIP model framework.

6.4.1 Datasets

Currently, there are various ESG rating services available (Schäfer 2016), many of which offer a subscription fee for data access which limits its availability to the public. In 2018 however, Yahoo Finance made some of the ESG ratings obtained from Sustainalytics (Stay 2010) available publicly. In this research, we utilized Yahoo Finance to obtain both financial stock prices and public ESG rating datasets in our reinforcement learning DRIP framework.

6.4.2 Data Cleaning and Feature Extraction

We downloaded the daily closing prices of all stocks in the Standard and Poor 500 list (S&P500) from the past 30 years from 31 December 1988 to 31 December 2018. In order to ensure a sufficient number of data points, we removed all the stocks which did not have a market price on 31 December 1988, which left us with 262 companies. SRI investors do not invest in companies with low ESG ratings; therefore we used a simple stock screening process to remove these unwanted stocks. From the shortlisted 262 stocks, we selected the top 100 companies with the highest combined ESG ratings according to Sustainalytics to construct the final dataset that contained a total of 756,000 data points.

We separated the train and test datasets using an out-of-bag approach, which excludes the testing period data from the past historical data at time t to avoid feeding the model any unknown future information. Our data splitting ratio is 9:1, which meant that the training data was from the year 1989 to 2015 for each stock, and the testing data was the three-year period from the year 2016 to 2018. We also adopt the rolling forecast approach to further split the data in the testing period into validation and test sets. For quarterly return prediction, we used "Q4/2015"

and "Q1/2016" as the validation and test set for the first period. We then moved to the next quarter period until "Q3/2018" and "Q4/2018" as validation and test sets in the last period. We applied the same data splitting process to the yearly return prediction dataset.

6.4.3 Baselines and Evaluation Metrics

To test our DRIP model, we used the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) as the evaluation metrics for the absolute value prediction:

$$\text{MAE} = 1/N \sum_{i=1}^{N} |\tilde{r}_i - r_i| \tag{6.24}$$

$$\text{RMSE} = \sqrt{1/N \sum_{i=1}^{N} (\tilde{r}_i - r_i)^2} \tag{6.25}$$

where $N = 100$ was the total number of stocks in the portfolio and $\tilde{r}_i$ and r_i were the predicted and actual return of stock i for that period.

We also converted the predicted value to a binary label to evaluate the performance of uptrend or downtrend forecast using the prediction accuracy metric and the Area Under the Curve (AUC) scores with the Receiver Operating Characteristic (ROC) curve. The lower MAE and RSME together with the higher prediction accuracy and AUC scores indicate the better performance of the prediction model. Our baseline models for comparison are the LSTM and GRU neural networks as in (Samarawickrama and Fernando 2017) and a univariate standard BiLSTM model (Uni).

To evaluate the performance of our socially responsible portfolios using MV-ESG model, we compared its Sharpe Ratio against those of the standard MV portfolios and the reported financial performance from similar sustainable indexes and funds. The Sharpe Ratio was defined as $S = (r_p - r_f)/\sigma_p$ where r_p was the portfolio annualized return, σ_p was the portfolio annualized volatility, and $r_f = 2\%$ was the nominal risk-free rate. A better performing portfolio had a higher Sharpe Ratio, which yielded higher returns if the risks were similar or a lower risk if the returns

were the same.

The sustainable indexes for comparison were: Dow Jones Sustainability World Index (DJSI World), Dow Jones Sustainability World Diversified Select Index (DJSI WD), and S&P500 ESG Factor Weighted Index (S&P500 ESG). All indexes data were obtained on the 31 December 2018 from S&P Dow Jones Indices, a division of S&P Global. The sustainable Exchange Traded Funds (ETF) with their symbol codes in the brackets were: iShares Global Clean Energy ETF (ICLN), Invesco Solar ETF (TAN), iShares MSCI USA ESG Select ETF (SUSA) and Workplace Equality Portfolio (EQLT). All funds data were obtained on the 31 December 2018 from Morningstar.

6.5 Result and Discussion

6.5.1 Empirical Result

Long-term Stock Return Prediction Result

First of all, we tested the performance of DRIP model on the prediction of quarterly returns. The hyperparameters in our neural networks were set as: the number of units in the deep learning layers equaled to 100, batch size equaled to 1, the loss was the mean squared error and random seed equaled 0. We used Adam optimizer with learning rate $lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, fuzz factor $\epsilon = 1e-7$, and decay equaled to 0. We also set the number of epochs equaled to 10 with a checkpoint after each epoch and only saved the best model for prediction.

Our experiment setup was as follows: $\Delta t = 63$, $\delta t = 63$ and the time gap was set to 63 representing the total number of trading days in a quarter. This meant that the model predicted the prices and returns three months ahead in time. After each period, the model was retrained and validated with the out-of-bag three-month data and predicted the next return in 63 days. The testing data for each quarter of each year from 2016 to 2018 were referred to as "Q1", "Q2", "Q3" and "Q4" respectively.

We then tested the performance on the prediction of yearly returns with $\Delta t = 252$, $\delta t = 252$ representing the 252 trading days in a typical year. The other setup

was the same as in the quarterly returns prediction model. The empirical results in Table 6.1 showed the performance evaluation for the quarterly and yearly returns prediction models using the multivariate financial time series as input. The reported RMSE and AUC Scores were averages for 100 stocks in each time period.

Our DRIP models, which used multivariate financial returns as the input significantly outperformed the other baseline models for most prediction periods in terms of MAE and RMSE. We can conclude that the prediction model using multivariate financial returns and BiLSTM neural networks in our design was a better solution for this predictive analytic problem. Focusing on the trend prediction accuracy, except for the slightly worse results in "Q2/16" and "Q3/18", our DRIP models that used BiLSTM achieved higher prediction accuracy and AUC scores regardless of the time periods or of the quarterly or yearly returns. These results demonstrated the effectiveness of our approach, that the reinforcement learning had successfully captured the underlying hidden information in the inter-correlated multivariate series and improved itself over time.

The value predictions of quarterly returns generally had lower MAE and RSME than the yearly forecast. This result was expected as the time gap was smaller; hence, it was easier to forecast the absolute stock return values. The ROC curves in Figure 6.6 showed a performance lift in the quarterly returns prediction model compared to other baselines for the entire 3 year testing period. Conversely, the trend prediction was more accurate in yearly return models, which proved that the reinforcement learning model could filter out the market noise in short term price changes. Overall, our DRIP model effectively and accurately predicted the annual returns in all three years and the quarterly returns in 10 out of 12 testing periods. It showed that our prediction model was not over-fitted to a certain dataset period, and it could be generalized for similar applications.

Robustness Analysis

To test the robustness of our model, we first benchmarked the prediction model using different combinations of the neural network hyperparameters. We split the

Table 6.1 : DRIP Model Evaluation

	Mean Absolute Error (MAE)														
	Q1/16	Q2/16	Q3/16	Q4/16	Q1/17	Q2/17	Q3/17	Q4/17	Q1/18	Q2/18	Q3/18	Q4/18	2016	2017	2018
DRIP	**0.0547**	**0.0880**	**0.0578**	**0.0814**	**0.0600**	**0.0551**	**0.0555**	**0.0601**	**0.0692**	**0.0667**	**0.0771**	**0.0948**	**0.0754**	**0.0830**	**0.1017**
Uni	0.0656	0.0930	0.0678	0.0999	0.0633	0.0704	0.0679	0.0683	0.0714	0.0882	0.0854	0.1096	0.0821	0.0974	0.1157
LSTM	0.0771	0.0960	0.0734	0.1043	0.0.0677	0.0767	0.0707	0.0718	0.0650	0.0986	0.0788	0.1404	0.0889	0.1074	0.1122
GRU	0.0652	0.0950	0.0723	0.1139	0.0623	0.0793	0.0776	0.0729	0.0801	0.0993	0.1003	0.0936	0.0819	0.1018	0.1332
	Root Mean Squared Error (RSME)														
	Q1/16	Q2/16	Q3/16	Q4/16	Q1/17	Q2/17	Q3/17	Q4/17	Q1/18	Q2/18	Q3/18	Q4/18	2016	2017	2018
DRIP	**0.0672**	**0.1080**	**0.0754**	**0.1209**	**0.0768**	**0.0741**	**0.0750**	**0.0790**	**0.0858**	**0.0843**	**0.1202**	**0.1165**	**0.1014**	**0.1062**	**0.1273**
Uni	0.0822	0.1189	0.0863	0.1534	0.0806	0.0933	0.0887	0.0939	0.0926	0.1120	0.1333	0.1402	0.1135	0.1241	0.1461
LSTM	0.0957	0.1225	0.0911	0.1637	0.0847	0.0991	0.0944	0.1029	0.0849	0.1233	0.1238	0.1739	0.1287	0.1355	0.1425
GRU	0.0812	0.1254	0.0914	0.1708	0.0801	0.1041	0.0952	0.0981	0.1057	0.1238	0.1536	0.1191	0.1086	0.1286	0.1659
	Area Under the Curve (AUC) Scores														
	Q1/16	Q2/16	Q3/16	Q4/16	Q1/17	Q2/17	Q3/17	Q4/17	Q1/18	Q2/18	Q3/18	Q4/18	2016	2017	2018
DRIP	**0.8392**	**0.5809**	**0.8387**	**0.8286**	**0.8495**	**0.7659**	**0.7521**	**0.8170**	**0.8165**	**0.8045**	**0.7024**	**0.9407**	**0.9525**	**0.9546**	**0.8989**
Uni	0.7443	0.5803	0.7876	0.6603	0.8522	0.6052	0.6297	0.7486	0.7441	0.6477	0.6952	0.8502	0.9339	0.9397	0.8899
LSTM	0.7115	0.5946	0.7719	0.6320	0.8790	0.5069	0.5739	0.7001	0.7516	0.5692	0.7549	0.7278	0.9348	0.9106	0.8989
GRU	0.6821	0.5954	0.7522	0.5202	0.8281	0.5427	0.5632	0.7287	0.6642	0.5692	0.6884	0.8820	0.9142	0.9537	0.8720
	Trend Prediction Accuracy (%)														
	Q1/16	Q2/16	Q3/16	Q4/16	Q1/17	Q2/17	Q3/17	Q4/17	Q1/18	Q2/18	Q3/18	Q4/18	2016	2017	2018
DRIP	**77%**	**50%**	**74%**	**76%**	**80%**	**66%**	**70%**	**77%**	**72%**	**76%**	**73%**	**85%**	**92%**	**92%**	**80%**
Uni	69%	50%	73%	62%	77%	58%	65%	75%	65%	64%	70%	76%	91%	91%	78%
LSTM	64%	53%	76%	61%	75%	52%	66%	75%	66%	57%	74%	60%	92%	85%	79%
GRU	67%	57%	70%	48%	75%	56%	58%	73%	58%	58%	62%	84%	92%	92%	73%

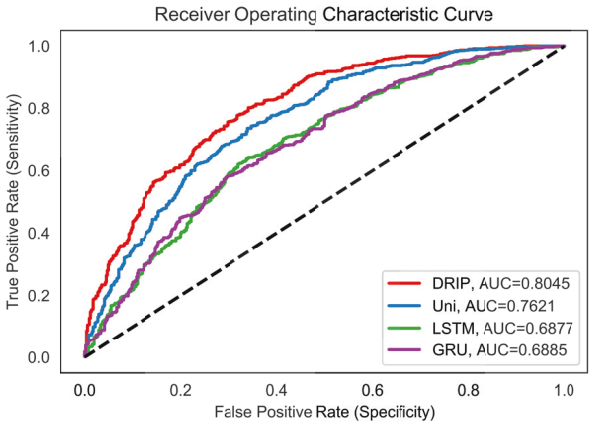


Figure 6.6 : ROC curves

dataset into train, validation and test sets with the ratio 8:1:1. The hyperparameter sets were: number of units in the deep learning layers was in $[100, 200, 300]$, batch size (BS) was in $[1, 10, 20]$ and learning rate (LR) was in [0.0001,0.001,0.01] accordingly. The MAE and prediction accuracy results of both validation and test set are presented in Table 6.2.

Table 6.2 showed that different hyperparameter sets could result in varied MAE and prediction accuracy. The gap between validation and test results are not significantly large, which indicates that our model was not overfitted. Our setting to generate the best results in the test set was: number of units equaled 100, batch size equaled 1 and learning rate equaled 0.01. In our rolling forecast and reinforcement learning model, the hyperparameters could be automatically tuned using grid search after each period.

We then used this set of hyperparameters to test the prediction model on three different datasets with 50, 100 and 200 randomly selected stocks (denoted as "Random50", "Random100", "Random200"). We also split these datasets into train, validation and test sets with the ratio 8:1:1. The results of this experiment are presented in Table 6.3.

Table 6.3 showed that our model still achieved a good prediction accuracy in randomly selected stock datasets. It is worth noticed that the MAE and the prediction

Table 6.2 : Benchmarking prediction model with multiple hyper-parameters

			Validation set		**Test Set**	
Units	**BS**	**LR**	**MAE**	**Accuracy**	**MAE**	**Accuracy**
100	1	0.001	0.05468	0.89635	**0.09335**	**0.86111**
100	20	0.001	0.04906	0.93016	0.09682	0.83270
300	10	0.001	0.04515	**0.94413**	0.09705	0.83841
300	20	0.001	**0.04514**	0.93460	0.10181	0.81857
100	10	0.0001	0.08325	0.78762	0.09744	0.81921
200	1	0.0001	0.06186	0.86063	0.09823	0.80857
300	1	0.001	0.04553	0.94000	0.09875	0.83413
100	1	0.0001	0.06901	0.82190	0.09913	0.78571
200	10	0.0001	0.07367	0.82063	0.09957	0.77841
300	20	0.0001	0.07198	0.79937	0.09988	0.79444
300	20	0.01	0.05908	0.86794	0.10055	0.80841
100	20	0.01	0.05021	0.92175	0.10056	0.82317
200	20	0.001	0.04866	0.93952	0.10097	0.78095
300	10	0.0001	0.07087	0.83746	0.10174	0.75841
100	10	0.01	0.04902	0.91206	0.10176	0.78556
100	10	0.001	0.05525	0.89238	0.10188	0.78444
200	10	0.001	0.05029	0.91651	0.10234	0.81032
200	20	0.0001	0.08412	0.79683	0.10321	0.78000
300	1	0.0001	0.05370	0.90238	0.10436	0.78762
100	1	0.01	0.06871	0.82365	0.10709	0.75921
200	20	0.01	0.05432	0.91683	0.11083	0.78190
200	1	0.001	0.04542	0.93571	0.11203	0.75000
100	20	0.0001	0.08296	0.75698	0.11208	0.70444
200	10	0.01	0.08924	0.82016	0.12027	0.74190
300	10	0.01	0.08880	0.76238	0.12863	0.74048
300	1	0.01	0.09793	0.74810	0.13317	0.65349
200	1	0.01	0.11110	0.76635	0.13671	0.74825

Table 6.3 : Benchmarking model with randomly selected datasets

	Validation set		Test Set	
Data	**MAE**	**Accuracy**	**MAE**	**Accuracy**
Random50	0.064246	0.85205	0.058508	0.779762
Random100	0.058567	0.809696	0.059457	0.755221
Random200	0.067143	0.827499	0.056684	0.818358

accuracy are not worsened for the larger dataset but varied due to the randomness of stock selection. These results indicated that our prediction model is robust and generalizable with different data sizes.

Portfolio Optimization Model Result

We used the predicted returns from the DRIP model as input for our MV-ESG model to construct socially responsible investment portfolios. We constructed the MAX-ESG portfolios using predicted returns. The nominal risk free rate was set to 2%, $r_f = 0.02$. After obtaining the stock allocation in each portfolio, we calculated the actual annualized returns and volatility using real stock prices for that period. The annualized returns, volatility, Sharpe Ratio and ESG Score given in Table 6.4 were averaged for the entire year, for each year in the testing period.

The results showed that our MAX-ESG portfolios had consistently higher ESG ratings (3 to 5 points above). Even though the MAX-S portfolios had better financial returns in 2016 and 2018, they also showed a relatively higher volatility level. Conversely, our MAX-ESG portfolios still achieved great financial returns with lower risk. The Sharpe Ratios of the MAX-ESG portfolios were higher than those of the MAX-S ones for 2017 and 2018. In 2017, the MAX-ESG portfolio achieved a better financial return 50.78% at a lower risk level 19.19%, compared with 47.76% return at 19.37% volatility in the MAX-S portfolio. These findings showed that achieving a socially responsible investment portfolio, with higher ESG ratings, and without the sacrifice of a large financial return, was achievable with our MV-ESG model.

Table 6.4 : MV-ESG Model Evaluation

	2016		**2017**		**2018**	
	MAX-S	MAX-ESG	MAX-S	MAX-ESG	MAX-S	MAX-ESG
Return	**32.73%**	28.47%	47.76%	**50.78%**	**30.33%**	26.60%
Volatility	17.22%	**14.89%**	19.37%	**19.18%**	16.84%	**14.31%**
Sharpe Ratio	**1.7845**	1.7777	2.3624	**2.5431**	1.6823	**1.7191**
ESG Score	70	**74**	70	**75**	68	**71**

We compared the performance of our final MAX-ESG portfolio with reported financial returns in 2018 obtained from similar sustainable indexes and funds. The results in Table 5 show that our portfolio outperformed other indexes and funds in terms of financial performance and achieved the Sharpe Ratio of 2.0634. Our portfolio had the best 3-year annualized return of 35.28%. Particularly in 2018, all indexes and funds had negative returns because many large stocks were in the downtrend. Our MAX-ESG portfolio was still able to achieve a positive return. This was mainly because the portfolio constructed was based on the maximization of Sharpe Ratio in the MV-ESG model, which optimally selects stocks with higher returns.

Our model's 3-year annualized volatility was in third place with 16.13%. This higher level of risk aligned with common investment knowledge on diversification (Statman 2004). Because the indexes often consist of a larger number of stocks, they generally had lower risks. However, the level diversification of our SRI portfolio was sufficient for individual investors. Our best MAX-ESG portfolio, for example, consisted of 7 stocks in 2016, 18 stocks in 2017 and 12 stocks in 2018 with the allocation as showed in Figure 6.8a. The model could be enhanced to construct a more diversified portfolio for sustainable investment funds with further constraints on weights.

Since all these indexes and funds published different types of sustainability met-

Table 6.5 : Benchmarking MAX-ESG portfolio with Sustainable Indexes and Funds in 2018

	Period Returns			**3-year Annualized**		
	2016	2017	2018	Return	Volatility	Sharpe Ratio
MAX-ESG	28.47%	50.78%	26.60%	**35.28%**	16.13%	**2.0634**
S&P500	11.29%	23.28%	-3.35%	10.41%	10.88%	0.7727
S&P500 ESG	14.52%	21.24%	-8.44%	9.11%	11.76%	0.6043
DJSI World	8.23%	27.98%	-8.03%	9.39%	11.52%	0.6418
DJSI WD	10.71%	24.00%	9.54%	14.75%	**10.48%**	1.2166
ICLN	-16.91%	21.48%	-9.02%	-1.48%	17.11%	-0.2036
TAN	-43.23%	54.39%	-25.66%	-4.83%	23.00%	-0.2971
SUSA	12.15%	22.53%	-5.65%	9.68%	11.42%	0.6722
EQLT	13.93%	21.33%	9.22%	14.83%	11.80%	1.0870

rics, we could not directly compare our portfolio ESG ratings to their benchmarks. We also could not report the net returns on investment due to the lacking of fund fees and tax calculation. In general, these results showed the effectiveness of our DRIP framework, not only for the optimization of socially responsible investment portfolios but also for financial stock investments in general.

Reinforcement Learning Result

Our reinforcement learning DRIP framework could be used to construct socially responsible portfolios with higher ESG ratings that still achieved competitive financial returns. Our DRIP model could predict multiple time steps ahead, which is an important feature for stock investors. Furthermore, the model significantly outperformed the univariate networks in both prediction accuracy and training speed with the same epoch size in terms of both prediction accuracy and training speed. In our experiments, it took one hour to perform reinforcement learning with the multivariate BiLSTM: a combination of 100 univariate neural networks with 10

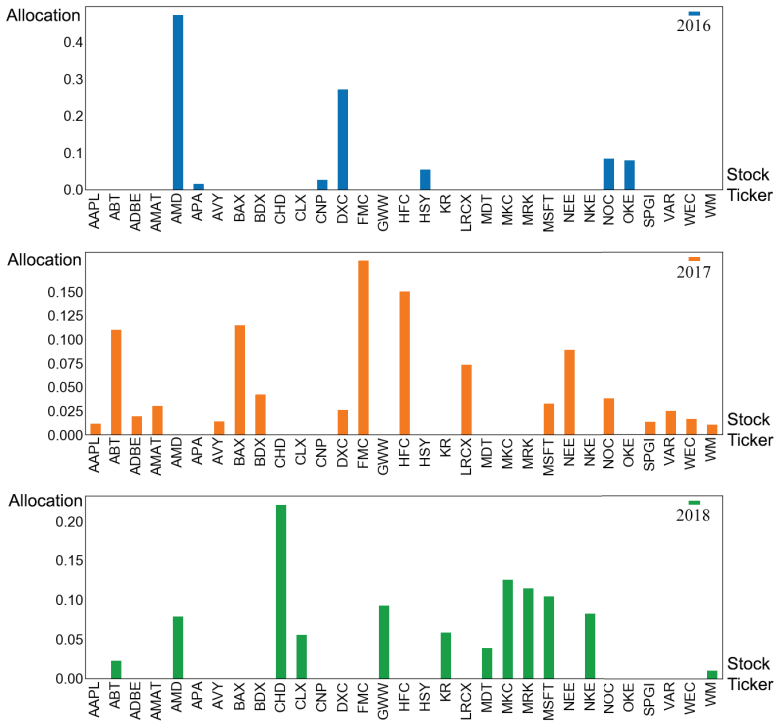


Figure 6.7 : MAX-ESG portfolio allocation

(a) (labels are trade symbols of companies)

epochs typically takes 100-times more training duration compared to our approach. This finding could lead to a better computationally efficient approach because the multivariate BiLSTM takes N times less in total training duration.

We also tested the performance of reinforcement learning by comparing the results to those without prediction model retraining and portfolio rebalancing (Non-RL). In the "Non-RL'" framework, we still used the Multivariate BiLSTM networks for prediction model and the MV-ESG for SRI portfolio optimization. However, the models were retrained after each testing period (each quarter or each year) without any pre-trained model and parameter learning from previous periods. The results in Table 6.6 showed that reinforcement learning had significantly improved the model performance in terms of both the prediction of stock returns and the optimization of portfolios. Since we were working with multivariate time series, retraining models and rebalancing portfolios were proven to be essential. Therefore, our reinforcement learning approach within the DRIP system was suitable for this time-sensitive data

Table 6.6 : Reinforcement Learning Test Results

	Prediction Model			
	MAE	RSME	AUC	Accuracy
DRIP	**0.0867**	**0.1117**	**0.9354**	**88%**
Non-RL	0.1098	0.1499	0.5515	55%
	MAX-ESG Portfolio			
	Return	Volatility	Sharpe Ratio	ESG Score
DRIP	**35.28%**	**16.13%**	**2.0634**	**73**
Non-RL	5.30%	14.00%	0.2357	68

analytics problem.

6.5.2 Discussion

Overall, our research demonstrated a promising trend in applying deep learning techniques for the selection of socially responsible investment portfolios. With the current progress in artificial intelligence, we believe it will bring further breakthroughs in socially responsible investment research. Our research will not only contribute directly to current literature in various disciplines but also translate into benefits for responsible investors, funds or indexes in markets. In the AI research field, our prediction model with a BiLSTM network could serve as a baseline for further research of long-term stock return forecasting using neural networks. In this research, we only used a single type of neural networks and structured data (stock prices and ESG ratings) as input. Studying the different variations and combination of the deep neural networks, as well as incorporating unstructured data (news, company reports or social media content) with text mining approaches in SRI, was beyond the scope of this research. However, our framework was designed with the flexibility to adopt different data mining approaches, neural networks or optimization algorithms in our future research.

Please be aware that by focusing on policies and rewards, our system might fail under extreme situations, e.g. a financial crisis. Our model's performance and applicability are subject to the hypothesis of stable company performances and normal finance market scenario. To safeguard investments in such cases, we would need additional failsafe measures when applying our model in practice.

From the financial research aspect, the MV-ESG model was one of the first to combine ESG ratings with a math finance model. Further research on how to incorporate this with other quantitative finance models such as GARCH (Francq and Zakoian 2019) would be relevant to both SRI scholars and investors. As many researchers are working on similar approaches for oil price forecasting (Kristjanpoller and Minutolo 2016), we believe a further investigation into this direction would be beneficial for SRI researchers. For simplicity purpose, we did not take into account income tax rates, inflation rates, trading fees, and other financial fund management costs. Further calculation of these fees would help the model implementation in the real-world investment scene.

Moreover, ESG ratings are not the only metrics to measure corporate social responsibility. The integration of hundreds of ESG sub-categorical ratings (e.g. greenhouse gas emissions or community support) could improve the model significantly. We have conducted experiments on our MV-ESG model as well as developing a combined MV-ESG model to utilize other metrics in the previous chapter. In our future research, we could study the potential of using deep learning approaches for a personalized stock recommendation system in the SRI context.

Last but not least, in order to incorporate the predicted ESG ratings from the CSR-SENT model as proposed in Chapter 5 into DRIP, the monthly ESG ratings of each mentioned companies are required. Unfortunately, the CSR reports of each company are only published annually. Therefore, the predicted ESG ratings from Chapter 5 are not a monthly time series. Due to the time limit of this research, the author directly uses the reported ESG ratings from Sustainalytics to evaluate the DRIP model. In future research, the author will consider expanding the text dataset to include news and social media data to predict ESG ratings as in Chapter

5 and incorporate it into DRIP.

To conclude, socially responsible investment is an emerging research topic with potential for long-term social impact. In this research, we proposed the DRIP model, which leveraged deep learning techniques to predict financial returns and construct a socially responsible investment portfolio. Validated with real-world data, our DRIP model, with a multivariate time series model, was able to accurately predict the stock returns three months ahead of time. It is possible that our framework could be generalized to build decision-support systems for similar multivariate prediction problems.

The socially responsible portfolios that we constructed using our novel MV-ESG model and reinforcement learning achieved much higher ESG ratings and a competitive financial performance overall compared with standard MV portfolio models and similar sustainable indexes and funds. With this rising trend in socially responsible investment, financial capital will diverge into good companies that contribute to a cleaner environment and a better society. This research also highlights a new direction for the use of more advanced deep learning approaches for quantitative finance research.

Chapter 7

Conclusion

This thesis provides a broader and comprehensive approach for quantitative research within the wealth management field from both financial and customer aspects. Particularly, this research utilizes the big data of structured demographic, behavioral, communicational data, and unstructured textual information from wealth customers, plus additional financial market and corporate responsibility data from companies. This thesis exploits deep analytics techniques to provide better framework for decision-making support based on the constructed mathematical and computational models, combined with customer segmentation modeling and quantitative finance approach.

The thesis has essentially achieved its outlined objectives and answers the research questions stated in Chapter 1. It expands current literature on the methodology of big data framework in wealth data, including data preparation, cleansing, analysis and integration. Studies of machines learning algorithms in wealth data analytics have been conducted, particularly focusing on financial services customer data and socially responsible investment data analytics. The research proposes analytics frameworks for both financial services customer and socially responsible investment data, and the author also develop and apply related machine learning algorithms for both financial services customer and socially responsible investment data. Finally, all the proposed data mining approaches and trained models have been tested on real-life datasets to infer meaningful insights and managerial implications in financial wealth industry.

Within the scope of this thesis, the author has covered the two main aspects of wealth data analytics: customer and investment. The methods proposed can be generalized for other research direction in wealth management industry or other

related business and finance field.

From the customer aspect, the thesis applies big data analytics, text mining and interpretable machine learning in customer data analytics in wealth management. The proposed approaches and models are (1) MMDB for personality mining, (2) transfer learning for customer personality prediction, (3) ensemble model with text mining for churn prediction, (4) interpretable machine learning with SHAP-MRMR+ to extract customer insight, and (5) customer segmentation and managerial implications with personality and SOM.

From the financial aspect, this is one of the first research to utilize deep learning for socially responsible investment. The proposed framework consists of (1) text mining of CSR reports for ESG ratings, (2) ESG-based quantitative models, (3) deep learning using Multivariate BiLSTM for stock return prediction, (4) MV-ESG for ESG-based portfolio optimization, and (5) reinforcement learning for socially responsible investment.

The theory and empirical results in this thesis expand the current literature in the multidisciplinary research area of machine learning for wealth data analytics. The methods developed and proposed in this thesis have a strong applicability in financial services field. It benefits various stakeholders: (1) multidisciplinary researchers from both machine learning and finance area; (2) financial services firms and investment funds; and (3) individual, especially financial customers and socially responsible investors.

Researchers can expand the research based on the proposed approaches to their current and future studies in related field. Financial services companies can save millions dollars in profit with better customer data analytics and CRM strategies. Customers of those firms can benefit from better services and accumulative financial gains. Both individual and institutional investors can leverage the socially responsible investment portfolio framework to improve their wealth management plan and benefits.

The thesis also serves as the initial foundation for further related research in var-

ious direction. Some of the potential future study focuses are: (1) big data mining for customer lifetime value model, (2) interpretable machine learning methods for better CRM involving not only attention but also attrition strategies, (3) quantitative models to measure investment impact as proposed in the Appendix A, (4) responsible investment portfolio to incorporate more ESG sub-categorical ratings, and/or (5) personalized recommendation system for investment customer based on their SRI preferences.

In summary, the research work presented in this thesis has showed the advantages and effectiveness of deep learning algorithms and big data analytics in wealth data analytics. Through the completion of this multi-discipline research, various aspects of wealth data analytics have been researched and integrated into a sophisticated framework, and the results of the proposed information systems can provide meaningful insights to support decision making for various different stakeholders. In recent years, efforts for research collaborations from multiple study area like this thesis have advanced methods in various traditional fields such as finance, health care, automotive, education, construction, etc. The research sets foundation and motivation for further related multidisciplinary studies, proving that integrating domain knowledge from several different research focuses can greatly advance the current literature and industry applications.

Bibliography

Adadi, A. & Berrada, M., 2018, 'Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)', *IEEE Access*, vol. 6, pp. 52138–52160.

Adamopoulos, P., Ghose, A. & Todri, V., 2018, 'The impact of user personality traits on word of mouth: Text-mining social media platforms', *Information Systems Research*, vol. 29, no. 3, pp. 612–640.

Al-Hawari, M. A., 2015, 'How the personality of retail bank customers interferes with the relationship between service quality and loyalty', *International Journal of Bank Marketing*, vol. 33, no. 1, pp. 41–57.

Alam, F. & Riccardi, G., 2014, 'Predicting personality traits using multimodal information', *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, ACM, pp. 15–18.

Ali, Ö. G. & Arıtürk, U., 2014, 'Dynamic churn prediction framework with more effective use of rare event data: The case of private banking', *Expert Systems with Applications*, vol. 41, no. 17, pp. 7889–7903.

Almana, A. M., Aksoy, M. S. & Alzahrani, R., 2014, 'A survey on data mining techniques in customer churn analysis for telecom industry', *International Journal of Engineering Research and Applications*, vol. 45, pp. 165–171.

Amel-Zadeh, A. & Serafeim, G., 2018, 'Why and how investors use ESG information: Evidence from a global survey', *Financial Analysts Journal*, vol. 74, no. 3, pp. 87–103.

Ang, W. R. & Weber, O., 2018, 'The market efficiency of socially responsible investment in korea', *Journal of Global Responsibility*, vol. 9, no. 1, pp. 96–110.

Argamon, S., Dhawle, S., Koppel, M. & Pennebaker, J. W., 2005, 'Lexical predictors of personality type', *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, pp. 1–16.

Argamon, S., Koppel, M., Pennebaker, J. W. & Schler, J., 2007, 'Mining the blogosphere: Age, gender and the varieties of self-expression', *First Monday*, vol. 12, no. 9.

Auer, B. R. & Schuhmacher, F., 2016, 'Do socially (ir) responsible investments pay? New evidence from international ESG data', *The Quarterly Review of Economics and Finance*, vol. 59, pp. 51–62.

Bao, W., Yue, J. & Rao, Y., 2017, 'A deep learning framework for financial time series using stacked autoencoders and long-short term memory', *PloS one*, vol. 12, no. 7, p. e0180944.

Bao, Y. & Datta, A., 2014, 'Simultaneously discovering and quantifying risk types from textual risk disclosures', *Management Science*, vol. 60, no. 6, pp. 1371–1391, <https://doi.org/10.1287/mnsc.2014.1930>.

Barnett, M. L. & Salomon, R. M., 2006, 'Beyond dichotomy: The curvilinear relationship between social responsibility and financial performance', *Strategic management journal*, vol. 27, no. 11, pp. 1101–1122.

Bauer, R., Derwall, J. & Otten, R., 2007, 'The ethical mutual fund performance debate: New evidence from canada', *Journal of Business Ethics*, vol. 70, no. 2, pp. 111–124.

Bayes, T., 1763, 'LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S', *Philosophical transactions of the Royal Society of London*, , no. 53, pp. 370–418.

Bello, Z. Y., 2005, 'Socially responsible investing and portfolio diversification', *Journal of Financial Research*, vol. 28, no. 1, pp. 41–57.

Biel, J.-I. & Gatica-Perez, D., 2013, 'The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs', *Multimedia, IEEE Transactions on*, vol. 15, no. 1, pp. 41–55.

Bishop, C. M., 1994, 'Mixture density networks', .

Bose, I. & Pal, R., 2012, 'Do green supply chain management initiatives impact stock prices of firms?', *Decision support systems*, vol. 52, no. 3, pp. 624–634.

Bradski, G., 2000, 'The opencv library.', *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123.

Bradstreet, D., 2009, *Wealth Management*, McGraw-Hill Education (India) Pvt Limited.

Breiman, L., 2001, 'Random forests', *Machine learning*, vol. 45, no. 1, pp. 5–32.

Brito, F. M. C., 2018, *The financial and social performance of US socially responsible mutual funds*, Ph.D. thesis, Universidade do Minho.

Broyden, C. G., Dennis Jr, J. & Moré, J. J., 1973, 'On the local and superlinear convergence of quasi-Newton methods', *IMA Journal of Applied Mathematics*, vol. 12, no. 3, pp. 223–245.

Buallay, A., 2019, 'Is sustainability reporting (ESG) associated with performance? Evidence from the European banking sector', *Management of Environmental Quality: An International Journal*, vol. 30, no. 1, pp. 98–115.

Buettner, R., 2016, 'Innovative personality-based digital services.', *PACIS*, p. 278.

Calvo, C., Ivorra, C. & Liern, V., 2016, 'Fuzzy portfolio selection with non-financial goals: exploring the efficient frontier', *Annals of Operations Research*, vol. 245, no. 1, pp. 31–46.

Castillo, J., 2017, 'The relationship between big five personality traits, customer empowerment and customer satisfaction in the retail industry', *Journal of Business and Retail Management Research (JBRMR)*, vol. 11, no. 2.

Celli, F., 2012, 'Unsupervised personality recognition for social network sites', *Proceeedings of the International Conference on Digital Society*, pp. 59–62.

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E. & Vlachogiannakis, N., 2018, 'Forecasting stock market crisis events using deep and statistical machine learning techniques', *Expert Systems with Applications*, vol. 112, pp. 353 – 371.

Chelawat, H. & Trivedi, I. V., 2016, 'The business value of ESG performance: the Indian context', *Asian journal of business ethics*, vol. 5, no. 1-2, pp. 195–210.

Chen, C.-C., Huang, H.-H. & Chen, H.-H., 2018, 'Ntusd-fin: A market sentiment dictionary for financial social media data applications', *Proceedings of the 1st Financial Narrative Processing Workshop*, .

Chen, K., Zhou, Y. & Dai, F., 2015, 'A LSTM-based method for stock returns prediction: A case study of China stock market', *Proceedings of the 2015 IEEE International Conference on Big Data*, IEEE, pp. 2823–2824.

Chen, L., Feng, G., Leong, C. W., Lehman, B., Martin-Raugh, M., Kell, H., Lee, C. M. & Yoon, S.-Y., 2016, 'Automated scoring of interview videos using doc2vec multimodal feature extraction paradigm', *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ACM, pp. 161–168.

Chen, T. & Guestrin, C., 2016, 'Xgboost: A scalable tree boosting system', *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, pp. 785–794.

Chen, T., Xu, R., He, Y. & Wang, X., 2017, 'Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN', *Expert Systems with Applications*, vol. 72, pp. 221–230.

Chiang, W., Enke, D., Wu, T. & Wang, R., 2016, 'An adaptive stock index trading decision support system', *Expert Systems with Applications*, vol. 59, pp. 195–207.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk,

H. & Bengio, Y., 2014, 'Learning phrase representations using RNN encoder-decoder for statistical machine translation', *arXiv preprint arXiv:1406.1078*.

Chollet, F. et al., 2015, 'Keras', `https://keras.io`.

Chu, C., Xu, G., Brownlow, J. & Fu, B., 2016, 'Deployment of churn prediction model in financial services industry', *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, IEEE, pp. 1–2.

Connaker, A. & Madsbjerg, S., 2019, 'The State of Socially Responsible Investing', <`https://hbr.org/2019/01/the-state-of-socially-responsible-investing`>, [Online; posted 17-January-2019].

Costa, P. T. & McCrae, R. R., 1992, 'Four ways five factors are basic', *Personality and individual differences*, vol. 13, no. 6, pp. 653–665.

Coussement, K. & Van den Poel, D., 2008, 'Integrating the voice of customers through call center emails into a decision support system for churn prediction', *Information & Management*, vol. 45, no. 3, pp. 164–174.

Coussement, K. & Van den Poel, D., 2009, 'Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers', *Expert Systems with Applications*, vol. 36, no. 3, pp. 6127–6134.

Cox, D. R., 1958, 'The regression analysis of binary sequences', *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232.

De Bock, K. W. & Van den Poel, D., 2012, 'Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models', *Expert Systems with Applications*, vol. 39, no. 8, pp. 6816–6826.

Deb, K., Agrawal, S., Pratap, A. & Meyarivan, T., 2000, 'A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii',

International conference on parallel problem solving from nature, Springer, pp. 849–858.

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977, 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22.

Deng, Q., Hine, M., Ji, S. & Sur, S., 2017, 'Building an environmental sustainability dictionary for the it industry', *Proceedings of the 50th Hawaii International Conference on System Sciences*, .

Di Persio, L. & Honchar, O., 2016, 'Artificial neural networks architectures for stock price prediction: comparisons and applications', *International Journal of Circuits, Systems and Signal Processing*, vol. 10, pp. 403–413.

Dumay, J., Guthrie, J. & Farneti, F., 2010, 'GRI sustainability reporting guidelines for public and third sector organizations: A critical review', *Public Management Review*, vol. 12, no. 4, pp. 531–548.

Døskeland, T. & Pedersen, L. J. T., 2016, 'Investing with brain or heart? a field experiment on responsible investment', *Management Science*, vol. 62, no. 6, pp. 1632–1644, <https://doi.org/10.1287/mnsc.2015.2208>.

Eccles, N. & Viviers, S., 2011, 'The origins and meanings of names describing investment practices that integrate a consideration of ESG issues in the academic literature', *Journal of Business Ethics*, vol. 104, no. 3, pp. 389–402.

Escrig-Olmedo, E., Rivera-Lirio, J. M., Muñoz-Torres, M. J. & Fernández-Izquierdo, M. Á., 2017, 'Integrating multiple esg investors' preferences into sustainable investment: A fuzzy multicriteria methodological approach', *Journal of cleaner production*, vol. 162, pp. 1334–1345.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.-F. & De Cock, M., 2016, 'Computational personality recognition in social media', *User modeling and user-adapted interaction*, vol. 26, no. 2-3, pp. 109–142.

Farnadi, G., Sushmita, S., Sitaraman, G., Ton, N., De Cock, M. & Davalos, S., 2014, 'A multivariate regression approach to personality impression recognition of vloggers', *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, ACM, pp. 1–6.

Farnadi, G., Zoghbi, S., Moens, M.-F. & De Cock, M., 2013, 'Recognising personality traits using facebook status updates', *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*, AAAI.

Farquad, M. A. H., Ravi, V. & Raju, S. B., 2014, 'Churn prediction using comprehensible support vector machine: An analytical CRM application', *Applied Soft Computing*, vol. 19, pp. 31–40.

Fast, E., Chen, B. & Bernstein, M. S., 2016, 'Empath: Understanding topic signals in large-scale text', *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, pp. 4647–4657.

Fatemi, A., Glaum, M. & Kaiser, S., 2018, 'ESG performance and firm value: The moderating role of disclosure', *Global Finance Journal*, vol. 38, pp. 45 – 64, special Issue on Corporate Social Responsibility and Ethics in Financial Markets.

Feuerriegel, S. & Gordon, J., 2018, 'Long-term stock index forecasting based on text mining of regulatory disclosures', *Decision Support Systems*, vol. 112, pp. 88–97.

Formánková, S., Trenz, O., Faldík, O., Kolomazník, J. & Sládková, J., 2019, 'Millennials' awareness and approach to social responsibility and investment—case study of the czech republic', *Sustainability*, vol. 11, no. 2, p. 504.

Fornell, C. & Wernerfelt, B., 1987, 'Defensive marketing strategy by customer complaint management: a theoretical analysis', *Journal of Marketing research*, vol. 24, no. 4, pp. 337–346.

Fowler, S. J. & Hope, C., 2007, 'A critical review of sustainable business indices and their impact', *Journal of Business Ethics*, vol. 76, no. 3, pp. 243–252.

Francq, C. & Zakoian, J.-M., 2019, *GARCH models: structure, statistical inference and financial applications*, Wiley.

Friede, G., Busch, T. & Bassen, A., 2015, 'ESG and financial performance: aggregated evidence from more than 2000 empirical studies', *Journal of Sustainable Finance and Investment*, vol. 5, no. 4, pp. 210–233.

Friedman, J. H., 2001, 'Greedy function approximation: a gradient boosting machine', *Annals of statistics*, pp. 1189–1232.

Gadre-Patwardhan, S., Katdare, V. V. & Joshi, M. R., 2016, 'A Review of Artificially Intelligent Applications in the Financial Domain', *Artificial Intelligence in Financial Markets*, Springer, pp. 3–44.

Garcia, A. S., Mendes-Da-Silva, W. & Orsato, R. J., 2017, 'Sensitive industries produce better esg performance: Evidence from emerging markets', *Journal of cleaner production*, vol. 150, pp. 135–147.

Garcia-Bernabeu, A., Pla, D., Bravo, M. & Perez-Gladish, B., 2015, 'Mean-variance stochastic goal programming for sustainable mutual funds'portfolio selection', *Rect@*, vol. 16, no. 2, p. 135.

Gauvain, J.-L. & Lee, C.-H., 1994, 'Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains', *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298.

Gill, A. J., Nowson, S. & Oberlander, J., 2009, 'What are they blogging about? Personality, topic and motivation in blogs', *Third International AAAI Conference on Weblogs and Social Media*, .

Gill, A. J. & Oberlander, J., 2002, 'Taking care of the linguistic features of extraversion', *Proceedings of the Annual Meeting of the Cognitive Science Society*, , vol. 24.

Goel, S., 2009, *Wealth Management: The New Business Model*, Global India Publications.

Golbeck, J., Robles, C. & Turner, K., 2011, 'Predicting personality with social media', *CHI'11 extended abstracts on human factors in computing systems*, ACM, pp. 253–262.

Goldberg, L. R., 1990, 'An alternative" description of personality": the Big-Five factor structure.', *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R. & Gough, H. G., 2006, 'The international personality item pool and the future of public-domain personality measures', *Journal of Research in personality*, vol. 40, no. 1, pp. 84–96.

Gray, H., 1983, *New directions in the investment and control of pension funds*, Investor Responsibility Research Center.

Grover, V., Chiang, R. H., Liang, T.-P. & Zhang, D., 2018, 'Creating strategic business value from big data analytics: A research framework', *Journal of Management Information Systems*, vol. 35, no. 2, pp. 388–423.

Haghighat, M., Abdel-Mottaleb, M. & Alhalabi, W., 2016, 'Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition', *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1984–1996.

Halbritter, G. & Dorfleitner, G., 2015, 'The wages of social responsibility—where are they? A critical review of ESG investing', *Review of Financial Economics*, vol. 26, pp. 25–35.

Harris, Z. S., 1954, 'Distributional structure', *Word*, vol. 10, no. 2-3, pp. 146–162.

Henrique, B. M., Sobreiro, V. A. & Kimura, H., 2019, 'Literature review: Machine learning techniques applied to financial market prediction', *Expert Systems with Applications*, vol. 124, pp. 226 – 251.

High, R., 2012, 'The era of cognitive systems: An inside look at ibm watson and how it works', *IBM Corporation, Redbooks*.

Ho, C.-S., Damien, P., Gu, B. & Konana, P., 2017, 'The time-varying nature of social media sentiments in modeling stock returns', *Decision Support Systems*, vol. 101, pp. 69–81.

Hochreiter, S. & Schmidhuber, J., 1997, 'Long short-term memory', *Neural computation*, vol. 9, no. 8, pp. 1735–1780.

Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A. & Hofmann-Wellenhof, R., 2013, 'Combining HCI, natural language processing, and knowledge discovery-potential of IBM content analytics as an assistive technology in the biomedical field', *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, Springer, pp. 13–24.

Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q. & Zeng, J., 2015, 'Telco churn prediction with big data', *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, ACM, pp. 607–618.

Hung, S.-Y., Yen, D. C. & Wang, H.-Y., 2006, 'Applying data mining to telecom churn management', *Expert Systems with Applications*, vol. 31, no. 3, pp. 515–524.

Hutto, C. J. & Gilbert, E., 2014, 'Vader: A parsimonious rule-based model for sentiment analysis of social media text', *Eighth international AAAI conference on weblogs and social media*, .

Iacobelli, F., Gill, A. J., Nowson, S. & Oberlander, J., 2011, 'Large scale personality classification of bloggers', *international conference on affective computing and intelligent interaction*, Springer, pp. 568–577.

Jiang, Q., Tang, C., Chen, C., Wang, X. & Huang, Q., 2019, 'Stock Price Forecast Based on LSTM Neural Network', Xu, J., Cooke, F. L., Gen, M. & Ahmed, S. E. (eds.) *Proceedings of the Twelfth International Conference on Management Science and Engineering Management*, Springer International Publishing, Cham, pp. 393–408.

Karahoca, A., Bilgen, O. & Karahoca, D., 2016, 'Churn management of e-banking customers by fuzzy AHP', *Handbook of Research on Financial and Banking Crisis Prediction Through Early Warning Systems*, IGI Global, pp. 155–172.

Kempf, A. & Osthoff, P., 2007, 'The effect of socially responsible investing on portfolio performance', *European Financial Management*, vol. 13, no. 5, pp. 908–922.

Kendall, M. G., 1938, 'A new measure of rank correlation', *Biometrika*, vol. 30, no. 1/2, pp. 81–93.

Keramati, A., Ghaneei, H. & Mirmohammadi, S. M., 2016, 'Developing a prediction model for customer churn from electronic banking services using data mining', *Financial Innovation*, vol. 2, no. 1, p. 10.

Khan, M., 2018, 'Corporate governance, esg, and stock returns around the world', *ESG, and Stock Returns around the World (November 1, 2018)*.

Kim, D. & Kim, S., 2017, 'Sustainable supply chain based on news articles and sustainability reports: Text mining with leximancer and diction', *Sustainability*, vol. 9, no. 6, p. 1008.

Kim, S.-H., Kim, M. & Holland, S., 2018, 'How customer personality traits influence brand loyalty in the coffee shop industry: the moderating role of business types', *International journal of hospitality & tourism administration*, vol. 19, no. 3, pp. 311–335.

Kitchens, B., Dobolyi, D., Li, J. & Abbasi, A., 2018, 'Advanced customer analytics: Strategic value through integration of relationship-oriented big data', *Journal Of Management Information Systems*, vol. 35, no. 2, pp. 540–574.

Koellner, T., Suh, S., Weber, O., Moser, C. & Scholz, R. W., 2007, 'Environmental impacts of conventional and sustainable investment funds compared using input-output life-cycle assessment', *Journal of Industrial Ecology*, vol. 11, no. 3, pp. 41–60.

Kohonen, T., 2013, 'Essentials of the self-organizing map', *Neural networks*, vol. 37, pp. 52–65.

Kolk, A., Kourula, A., Pisani, N., Westermann-Behaylo, M. & Worring, M., 2018, 'Embracing the un sustainable development goals? big data analysis of changes in the corporate sustainability agenda', *Academy of Management Global Proceedings*, vol. Surrey, no. 2018, p. 51, <https://journals.aom.org/doi/abs/10.5465/amgblproc.surrey.2018.0051.abs>.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V. & Stillwell, D., 2015, 'Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines.', *American Psychologist*, vol. 70, no. 6, p. 543.

Kraft, D., 1988, 'A software package for sequential quadratic programming', *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*.

Kraus, M. & Feuerriegel, S., 2017, 'Decision support from financial disclosures with deep neural networks and transfer learning', *Decision Support Systems*, vol. 104, pp. 38–48.

Kristjanpoller, W. & Minutolo, M. C., 2016, 'Forecasting volatility of oil price using an artificial neural network-GARCH model', *Expert Systems with Applications*, vol. 65, pp. 233–241.

Kumar, B. S. & Ravi, V., 2016, 'A survey of the applications of text mining in financial domain', *Knowledge-Based Systems*, vol. 114, pp. 128–147.

Landrum, N. E. & Ohsowski, B., 2017, 'Identifying worldviews on corporate sustainability: A content analysis of corporate sustainability reports', *Business Strategy and the Environment*, vol. 27, no. 1, pp. 128–151, <https://onlinelibrary.wiley.com/doi/abs/10.1002/bse.1989>.

Lemaître, G., Nogueira, F. & Aridas, C. K., 2017, 'Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning', *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563.

Loughran, T. & McDonald, B., 2016, 'Textual analysis in accounting and finance: A survey', *Journal of Accounting Research*, vol. 54, no. 4, pp. 1187–1230.

Luhn, H. P., 1957, 'A statistical approach to mechanized encoding and searching of literary information', *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317.

Lundberg, S. M. & Lee, S.-I., 2017, 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, pp. 4765–4774.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J. et al., 2018, 'Explainable machine-learning predictions for the prevention of hypoxaemia during surgery', *Nature biomedical engineering*, vol. 2, no. 10, p. 749.

Mairesse, F. & Walker, M., 2006, 'Automatic recognition of personality in conversation', *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, Association for Computational Linguistics, pp. 85–88.

Mairesse, F. & Walker, M., 2007, 'PERSONAGE: Personality generation for dialogue', *Annual Meeting-Association For Computational Linguistics*, , vol. 45p. 496.

Manner, C. K., 2017, 'Who posts online customer reviews? the role of sociodemographics and personality traits', *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, vol. 30, pp. 23–23.

Markowitz, H., 1952, 'Portfolio Selection', *The Journal of Finance*, vol. 7, no. 1, pp. 77–91.

Metz, C. E., 1978, 'Basic principles of ROC analysis', *Seminars in nuclear medicine*, , vol. 8Elsevier, pp. 283–298.

Moghaddam, A. H., Moghaddam, M. H. & Esfandyari, M., 2016, 'Stock market index prediction using artificial neural network', *Journal of Economics, Finance and Administrative Science*, vol. 21, no. 41, pp. 89 – 93.

MSCI, 2018, 'MSCI ESG Ratings Methodology', <https://www.msci.com/documents/10199/123a2b2b-1395-4aa2-a121-ea14de6d708a>, [Online; posted 1-April-2018].

Munoz, F., Vargas, M. & Marco, I., 2014, 'Environmental mutual funds: Financial performance and managerial abilities', *Journal of Business Ethics*, vol. 124, no. 4, pp. 551–569.

Nam, K. & Seong, N., 2019, 'Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market', *Decision Support Systems*, vol. 117, pp. 100 – 112.

Nelson, D. M., Pereira, A. C. & de Oliveira, R. A., 2017, 'Stock market's price movement prediction with LSTM neural networks', *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1419–1426.

Nilsson, J., 2008, 'Investment with a conscience: Examining the impact of pro-social attitudes and perceived financial performance on socially responsible investment behavior', *Journal of Business Ethics*, vol. 83, no. 2, pp. 307–325.

Nowson, S., Oberlander, J., Gill, A. J. et al., 2005, 'Weblogs, genres and individual differences', *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, , vol. 1666Stresa, p. 1671.

Oberlander, J. & Nowson, S., 2006, 'Whose thumb is it anyway? Classifying author personality from weblog text', *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 627–634.

Park, K. & Kremer, G. E. O., 2017, 'Text mining-based categorization and user perspective analysis of environmental sustainability indicators for manufacturing and service systems', *Ecological indicators*, vol. 72, pp. 803–820.

Pearson, K., 1895, 'Note on regression and inheritance in the case of two parents', *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242.

Pedersen, M., 2015, 'Strategies for investing in the s&p 500', *Strategies for Investing in the S&P*, vol. 500.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E., 2011a, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al., 2011b, 'Scikit-learn: Machine learning in Python', *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830.

Peloza, J., 2009, 'The challenge of measuring financial impacts from investments in corporate social performance', *Journal of Management*, vol. 35, no. 6, pp. 1518–1541.

Pencle, N. & Mălăescu, I., 2016, 'What's in the words? development and validation of a multidimensional dictionary for csr and application using prospectuses', *Journal of Emerging Technologies in Accounting*, vol. 13, no. 2, pp. 109–127.

Peng, H., Long, F. & Ding, C., 2005, 'Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 8, pp. 1226–1238.

Pennebaker, J. W., Francis, M. E. & Booth, R. J., 2001, 'Linguistic inquiry and word count: Liwc 2001', *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001.

Pennebaker, J. W. & King, L. A., 1999, 'Linguistic styles: language use as an individual difference.', *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296.

Pentland, A., 2004, 'Social dynamics: Signals and behavior', *International Conference on Developmental Learning*, , vol. 5.

Peylo, B. T., 2012, 'A Synthesis of Modern Portfolio Theory and Sustainable Investment', *The Journal of Investing*, vol. 21, no. 4, pp. 33–46.

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J. & Escalera, S., 2016, 'Chalearn lap 2016: First round challenge on first impressions-dataset and results', *European Conference on Computer Vision*, Springer, pp. 400–418.

Powell, M. J., 1964, 'An efficient method for finding the minimum of a function of several variables without calculating derivatives', *The Computer Journal*, vol. 7, no. 2, pp. 155–162.

Quercia, D., Kosinski, M., Stillwell, D. & Crowcroft, J., 2011, 'Our twitter profiles, our selves: Predicting personality with twitter', *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, IEEE, pp. 180–185.

Ravi, K., Ravi, V. & Prasad, P. S. R. K., 2017, 'Fuzzy formal concept analysis based opinion mining for CRM in financial services', *Applied Soft Computing*, vol. 60, pp. 786–807.

Rehurek, R. & Sojka, P., 2010, 'Software framework for topic modelling with large corpora', *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Citeseer.

Renneboog, L., Ter Horst, J. & Zhang, C., 2007, 'Socially responsible investments: Methodology, risk exposure and performance', .

Renneboog, L., Ter Horst, J. & Zhang, C., 2008, 'Socially responsible investments: Institutional aspects, performance, and investor behavior', *Journal of Banking & Finance*, vol. 32, no. 9, pp. 1723–1742.

Rigby, P. C. & Hassan, A. E., 2007, 'What can OSS mailing lists tell us? A preliminary psychometric text analysis of the apache developer mailing list', *Proceedings of the fourth international workshop on mining software repositories*, IEEE Computer Society, p. 23.

Roshchina, A., Cardiff, J. & Rosso, P., 2011, 'A comparative evaluation of personality estimation algorithms for the twin recommender system', *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, ACM, pp. 11–18.

Rudd, A., 1981, 'Social responsibility and portfolio performance', *California Management Review*, vol. 23, no. 4, pp. 55–61.

Rudin, C., 2019, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, vol. 1, no. 5, p. 206.

Sahut, J.-M. & Pasquini-Descomps, H., 2015, 'ESG impact on market performance of firms: International Evidence', *Management international/International Management/Gestiòn Internacional*, vol. 19, no. 2, pp. 40–63.

Samarawickrama, A. & Fernando, T., 2017, 'A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market', *Proceedings of the 2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, IEEE, pp. 1–6.

Sant, J. V., 1971, 'Social Responsibility and Investments', *Theology Today*, vol. 28, no. 3, pp. 369–371.

Sarkar, C., Bhatia, S., Agarwal, A. & Li, J., 2014, 'Feature analysis for computational personality recognition using youtube personality data set', *Proceedings of*

the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, pp. 11–14.

Schäfer, H., 2016, 'Corporate Social Responsibility Rating', *A Handbook of Corporate Governance and Social Responsibility*, p. 449.

Scherer, A., Wünderlich, N. V. & Von Wangenheim, F., 2015, 'The Value of Self-Service: Long-Term Effects of Technology-Based Self-Service Usage on Customer Retention.', *MIS quarterly*, vol. 39, no. 1.

Schuster, M. & Paliwal, K. K., 1997, 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K. & Soman, K., 2017, 'Stock price prediction using LSTM, RNN and CNN-sliding window model', *Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics*, IEEE, pp. 1643–1647.

Senanayake, D., Muthugama, L., Mendis, L. & Madushanka, T., 2015, 'Customer Churn Prediction: A Cognitive Approach', *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 3, pp. 767–773.

Sermpinis, G., Karathanasopoulos, A., Rosillo, R. & de la Fuente, D., 2019, 'Neural networks in financial trading', *Annals of Operations Research*, vol. Special Issue: Networks and Risk Management, pp. 11–16.

Sharpe, W. F., 1966, 'Mutual fund performance', *The Journal of Business*, vol. 39, no. 1, pp. 119–138.

Shin, S.-H., Kwon, O., Ruan, X., Chhetri, P., Lee, P. & Shahparvari, S., 2018, 'Analyzing sustainability literature in maritime studies with text mining', *Sustainability*, vol. 10, no. 10, p. 3522.

Siddiqui, A. I., Marinova, D., Hossain, A. & Todorov, V., 2011, 'Socially Responsible Investment in Australia', *Sustainability And Development In Asia And The Pacific: Emerging Policy Issues*, p. 249.

Spearman, C., 1904, 'The proof and measurement of association between two things', *The American journal of psychology*, vol. 15, no. 1, pp. 72–101.

Spearman, C., 1987, 'The proof and measurement of association between two things', *The American journal of psychology*, vol. 100, no. 3/4, pp. 441–471.

Statman, M., 2004, 'The diversification puzzle', *Financial Analysts Journal*, vol. 60, no. 4, pp. 44–53.

Stay, C., 2010, 'Corporate social responsibility', Tech. rep., Sustainalytics.

Stephen, S., 2018, 'Financial Performance of Environmentally Responsible Investment Funds: A Systematic Review', *Academy of Management Proceedings*, , vol. 2018Academy of Management Briarcliff Manor, NY 10510, p. 12451.

Straehl, P. U. & Ibbotson, R. G., 2017, 'The long-run drivers of stock returns: Total payouts and the real economy', *Financial Analysts Journal*, vol. 73, no. 3, pp. 32–52.

Sun, S., Luo, C. & Chen, J., 2017, 'A review of natural language processing techniques for opinion mining systems', *Information fusion*, vol. 36, pp. 10–25.

Sundarkumar, G. G. & Ravi, V., 2015, 'A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance', *Engineering Applications of Artificial Intelligence*, vol. 37, pp. 368–377.

Székely, N. & vom Brocke, J., 2017, 'What can we learn from corporate sustainability reporting? deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique', *PLOS ONE*, vol. 12, no. 4, pp. 1–27, <https://doi.org/10.1371/journal.pone.0174807>.

Thomson Reuters, 2019, 'Thomson Reuters ESG Scores', Tech. rep., Thomson Reuters.

US SIF Foundation, 2018, '2018 Biennial Report On US Sustainable, Responsible And Impact Investing Trends', Tech. rep., US SIF Foundation.

Van Duuren, E., Plantinga, A. & Scholtens, B., 2016, 'ESG integration and the investment management process: Fundamental investing reinvented', *Journal of Business Ethics*, vol. 138, no. 3, pp. 525–533.

Velte, P., 2017, 'Does ESG performance have an impact on financial performance? Evidence from Germany', *Journal of Global Responsibility*, vol. 8, no. 2, pp. 169–178.

Verbeke, W., Martens, D., Mues, C. & Baesens, B., 2011, 'Building comprehensible customer churn prediction models with advanced rule induction techniques', *Expert systems with applications*, vol. 38, no. 3, pp. 2354–2364.

Verheyden, T., Eccles, R. G. & Feiner, A., 2016, 'ESG for all? The impact of ESG screening on return, risk, and diversification', *Journal of Applied Corporate Finance*, vol. 28, no. 2, pp. 47–55.

Verhoeven, B., Daelemans, W. et al., 2014, 'Evaluating content-independent features for personality recognition', *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, ACM, pp. 7–10.

Vinciarelli, A. & Mohammadi, G., 2014, 'A survey of personality computing', *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291.

Vo, N. N., Liu, S., Brownlow, J., Chu, C., Culbert, B. & Xu, G., 2018, 'Client Churn Prediction with Call Log Analysis', *International Conference on Database Systems for Advanced Applications*, Springer, pp. 752–763.

Von Wallis, M. & Klein, C., 2015, 'Ethical requirement and financial interest: a literature review on socially responsible investing', *Business Research*, vol. 8, no. 1, pp. 61–98.

Vörösmarty, C., Osuna, V. R., Koehler, D., Klop, P., Spengler, J., Buonocore, J., Cak, A., Tessler, Z., Corsi, F., Green, P. et al., 2018, 'Scientifically assess impacts of sustainable investments', *Science*, vol. 359, no. 6375, pp. 523–525.

Wang, H.-X., Karp, A., Herlitz, A., Crowe, M., Kåreholt, I., Winblad, B. & Fratiglioni, L., 2009, 'Personality and lifestyle in relation to dementia incidence', *Neurology*, vol. 72, no. 3, pp. 253–259.

Wei, C.-P. & Chiu, I.-T., 2002, 'Turning telecommunications call details to churn prediction: a data mining approach', *Expert systems with applications*, vol. 23, no. 2, pp. 103–112.

Yao, Q., Chen, R. & Xu, X., 2015, 'Consistency between consumer personality and brand personality influences brand attachment', *Social Behavior and Personality: an international journal*, vol. 43, no. 9, pp. 1419–1427.

Yarkoni, T., 2010, 'Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers', *Journal of research in personality*, vol. 44, no. 3, pp. 363–373.

Yee Liau, B. & Pei Tan, P., 2014, 'Gaining customer knowledge in low cost airlines through text mining', *Industrial management & data systems*, vol. 114, no. 9, pp. 1344–1359.

Zhang, H., 2004, 'The optimality of Naive Bayes', *AA*, vol. 1, no. 2, p. 3.

Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W. & Yan, X., 2018, 'Measuring customer agility from online reviews using big data text analytics', *Journal of Management Information Systems*, vol. 35, no. 2, pp. 510–539.

Zhou, Z.-H. & Feng, J., 2017, 'Deep forest: Towards an alternative to deep neural networks', *arXiv preprint arXiv:1702.08835*.

Zopounidis, C., Galariotis, E., Doumpos, M., Sarri, S. & Andriosopoulos, K., 2015, 'Multiple criteria decision aiding for finance: An updated bibliographic survey', *European Journal of Operational Research*, vol. 247, no. 2, pp. 339 – 348.

Appendix A

Quantifying Socially Responsible Investment Impact

In this appendix, we propose a data-driven method to quantify the investment impacts regarding some nominal sustainability topics. Based on preliminary text mining research on the corporate responsibility reports of the companies, we choose five key metrics to test our approach: green house gas (GHG) emission reduction, waste recycling, energy saving, water saving and charitable giving in community support. Our method, however, can be generalized to incorporate more measuring dimensions in the future.

Let ESG_i be the ESG ratings and $EBITDA_i$ be the earnings before interest, tax, depreciation and amortization (EBITDA) of company i, $i \in \{1, 2, ..., N\}$. $\Omega_{im(k)}$ is defined as the reported total impact of company i in metric $m(k)$ with $k \in \{1, 2, 3, 4, 5\}$ being one of the chosen five metrics. We calculate the per-dollar impact $\omega_{im(k)}$ of company i as follow:

$$\omega_{im(k)} = \Omega_{im(k)}/EBITDA_i \tag{A.1}$$

For example, Intel Corporation reported that the company diverted 97% of 78.8 thousand tons of hazardous waste and recycled 85% of 108 thousand tons of non-hazardous waste in 2017. The total tons of waste recycled $\Omega_{im(k)} = 97\%*78.8+85\%*108 = 168.236$ (thousand tons). Using the reported EBITDA for Intel Corporation in 2017 is 26.58 billions USD, we can calculate the per-dollar impact $\omega_{im(k)} = 6.3294$ (kgs waste recycled per dollar investment).

The above calculation has been simplified based on the assumption that each dollar invested in the company will generate an equivalent one dollar in EBITDA. In

reality, each dollar invested will normally generate less than one dollar in EBITDA. Therefore, we use the Price Per Earning Ratio PE_i as the weighting input to calculate the adjusted per-dollar impact $\tilde{\omega}_{im(k)}$ as follow:

$$\rightarrow \tilde{\omega}_{im(k)} = \omega_{im(k)}/PE_i \tag{A.2}$$

Continuing the previous example, we have $PE_i = 11.53$ is the Price Per Earning Ratio of Intel Corporation. The adjusted per-dollar impact is calculated as $\tilde{\omega}_{im(k)} = 6.3294/11.53 = 0.5490$ (kgs waste recycled per dollar investment).

The five metrics are chosen based on the fact that their data is reported by most companies across different industries. However, various firms still do not include these numbers in their corporate responsibility reports due to their lack of data collection, management and transparency policy. To overcome the challenge of data availability, we use the ESG ratings and reported sustainability data from company i as a weighting input to estimate the impacts data for a similar company j in the same industry. For company j who does not report the respective sustainability data, we estimate the adjusted per-dollar impact $\tilde{\omega}_{jm(k)}$ for each metric $m(k)$ as follow:

$$\rightarrow \tilde{\omega}_{jm(k)} = \frac{ESG_j}{ESG_i}\tilde{\omega}_{im(k)} \tag{A.3}$$

For example, Advanced Micro Devices Incorporation is operating in the same industry with Intel Corporation, who do not report the exact number of tons of waste recycled. We have $ESG_i = 86.12$ and $ESG_j = 69.07$ as the ESG ratings for Intel Corporation and Advanced Micro Devices Incorporation. We then use the previously calculated $\tilde{\omega}_{im(k)}$ to calculate the adjusted per-dollar impact in waste recycling of Advanced Micro Devices Incorporation as $\tilde{\omega}_{jm(k)} = 0.4403$ (kgs waste recycled per dollar investment).

The intuition of our data-driven approach is that companies with higher ESG ratings are contributing more to sustainability. These companies often have higher earnings and profits, which allows them to expand the expenditure on sustainability.

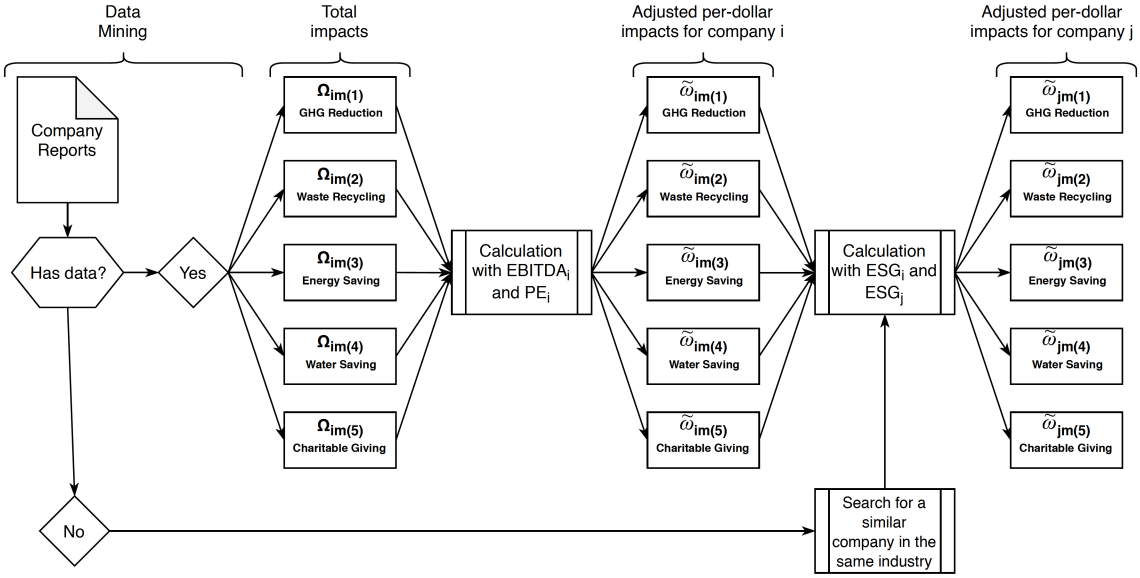


Figure A.1 : Investment Impact Measurement Process

Using the EBITDA and Price Per Earning Ratio will help measure the investment impacts more accurate. Moreover, all these data are publicly available, which enables the measuring of impacts in companies that do not publish the sustainable metrics in their corporate responsibility reports. The measurement metrics can be used directly by individual investors or ESG-focused fund to make better investment decisions. The full process of investment impacts measurement in our methodology is illustrated in Figure A.1.