Causal Analysis of Customer Churn Using Deep Learning

1st David Hason Rudd
Faculty of Engineering and IT
The University of Technology Sydney
Sydney, Australia

2nd Huan Huo
Faculty of Engineering and IT
The University of Technology Sydney
Sydney, Australia

3rd Guandong Xu
Faculty of Engineering and IT
Advanced Analytics institute (AAi)
Sydney, Australia

Abstract—**Customer churn describes terminating a relationship with a business or reducing customer engagement over a specific period. Two main business marketing strategies play vital roles to increase market share dollar-value: gaining new and preserving existing customers. Customer acquisition cost can be five to six times that for customer retention, hence investing in customers with churn risk is smart. Causal analysis of the churn model can predict whether a customer will churn in the foreseeable future and assist enterprises to identify effects and possible causes for churn and subsequently use that knowledge to apply tailored incentives. This paper proposes a framework using a deep feed-forward neural network for classification accompanied by a sequential pattern mining method on high-dimensional sparse data. We also propose a causal Bayesian networks to predict cause probabilities that lead to customer churn. Evaluation metrics on test data confirm the XGBoost and our deep learning model outperformed previous techniques. Experimental analysis confirms that some independent causal variables representing the level of super guarantee contribution, account growth, and customer tenure were identified as confounding factors for customer churn with a high degree of belief. This paper provides a real-world customer churn analysis from current status inference to future directions in local superannuation funds.**

Index Terms—**Churn Analysis, Causality Analysis, Deep Neural Network, Data Mining**

I. INTRODUCTION

Businesses rely heavily upon retaining satisfied customers in the global marketplace, which account for a large portion of their revenue. As the market becomes increasingly saturated, businesses have learned to emphasize maintaining existing clients. Although obtaining new clients was critical for initial business success, retention policies should have equal weight. Many previous studies have identified retention rate's substantial effect on the market, but clients are always able to churn away from a business, resulting in potential losses for the organization. However, customers usually offer some warning before being churned, hence churn prediction systems primarily focus on customer behaviour to identify specific customers who are likely to churn out and indicate reasons for the churn. Such factors would aid

marketing to develop effective retention strategies, increasing overall customer lifetime value, and assisting in growing the company's market value. For example, Bank or financial organization churn customers who close their accounts and had fewer transactions/interactions over time [1]. Churn rate,

$$\text{ChurnRate} = \text{LostCustomers}/\text{InitialCustomers}$$

is the percentage of customers a company loses in a fixed period, and represents the business pulse [2]. Voluntary churn is where a consumer wants to leave the company on their own. This could be due to dissatisfaction with product, or perhaps they feel they are not receiving that value they expected. In contrast, involuntary churn is where a customer leaves the company for unavoidable reason(s), such as payment problems due to expired account details, network issues, or inadequate cash; and incidental churn occurs due to location or financial situation changes. Deliberate churn arises due to customer desire to change innovation and price, where the most common reasons are poor service quality, noncompetitive pricing, and missing customer expectations. Companies can utilize churn analysis to determine the individual user risk levels and develop appropriate focused retention initiatives.

The proposed framework applies churn analysis for a local superannuation fund. Data preparation and analysis verified that most customer accounts were active for less than one year, for several possible reasons, including some customers being involuntary churners who were subsequently enrolled in different superannuation fund(s) operated by their employers. In contrast, voluntary churn occurs when a customer decides to cease the account for personal reasons, including perceived quality, technology, and price.

As our main contribution to this study, we built a specific churn propensity modelling work-frame to analyse customer financial behaviour data, then generalised the model for any superannuation fund(s) businesses. First, we analysed 12-month time window data to predict churners for the next six months. Since mining data

from the 12-month observation window allowed us to extract the latent factors cause of churn. Then, we scaled up the model by exploiting Bayesian networks to describe how deliberate churn occurs due to multiple causes and showed that hypothesis tests on common features greatly influence prediction results. finally, we proposed a specific causality analysis method that can be applied to other similar datasets employed in most superannuation funds.

It should be noted that this study only focuses on the performance of a DFF NN to address churn prediction on a massive and high dimensional sparse dataset that is usually created in financial networks. We conducted preprocessing phase in the proposed model for the baseline classifiers to address the sparse and skewed data and then compared the results with the output of proposed DFF NN. These very high dimensions data are primarily generated in subscribed-base businesses mostly in superannuation funds due to employing interval-based features used in customer relationship management (CRM) systems.

The remainder part of the paper proceeds as follows: Section II reviews recent relevant churn studies, and Section III gives preliminary knowledge on the application of the deep learning method in churn prediction problems. Section IV discusses specific research and analysis methods employed to predict underlying reasons for churn including problem definition and churn propensity modelling workflow. Section V analyses gathered data to address research questions: a causal analysis of churn through the financial data collected by superannuation funds and present experiments causality analysis outcomes. Finally, Section VI summarises and concludes the paper and discusses implications from the findings for real-world applications.

II. RELATED WORKS

A. Customer churn prediction and causal analysis

Most previous studies focused on determining churn variables for a particular dataset rather than customer churn causation analysis.

CHAMP (Churn analysis, modelling, and prediction) is an integrated system for forecasting consumers cancelling their cellular phone service [3]. Alyuda NeuroIntelligence employs neural networks (NNs) for data mining to forecast customer churn at banks [4]. Integrating textual data using customer churn prediction (CCP) algorithms adds value [5], and combining different classifiers, e.g. gradient boost, oversampling, and contrast sequential pattern mining on single year observation windows has been shown to be a practical strategy to deal with highly skewed data collected from superannuation funds [6]. Different churn prediction techniques have been evaluated to identify optimal approaches [7].

Hidden churn is a common problem for superannuation funds, where customer accounts become dormant once mandatory employer payments cease. Various remedies for insufficient consumer interaction have been proposed, addressing unbalanced and fully leveraged data problems, and multiple classifiers have been developed from sampled datasets [6]. Deep learning techniques can handle very large datasets compared with standard machine learning (ML) approaches, and combining deep learning and convolutional neural networks (CNNs) have successfully forecast churn [8].

Recent causal inference developments highlight fundamental alterations required to move from standard statistical analysis towards multivariate causal analysis [3]. New PC-Stable approaches effectively learn causal structures using DFS data, allowing temporal causal modelling for enormous time information datasets [9]. Directed acyclic graphs have been proposed to represent causal relationships in Bayesian networks [10]. A likelihood of churn anticipates, as are the driving factors on banking data. Subsequently, Shah et al [11] trained a model to produce suitable weights for features that predict whether a customer would churn. They contrasted churn definitions commonly used in business administration, marketing, IT, telecommunications, newspapers, insurance, and psychology. Many studies have and described churn loss, feature engineering, and prediction models using this approach [2].

The above approaches have been variously adopted for current churn prediction methods, but most are employed for collecting data from telco, banks, and retail markets. In contrast, the proposed approach was tested using a massive financial dataset with high-dimensional sparse data. To our best knowledge, no previous study considered causality analysis for customer churn in superannuation companies. Motivated by the studies mentioned above, We proposed a new approach that combines deep learning and Bayesian theory to improve churn prediction modelling systems.

III. PRELIMINARY KNOWLEDGE

Churn prediction models will provide earlier churner identification and hence assist customer intervention program development. This is basically a classification problem, i.e., to categorise each customer as potential churner or non-churner. Several machine learning techniques for churn prediction have produced verifiable results from interpretable models, such as boosting technique, a nonparametric method, logistic regression. However, this approach is only valid predictability under circumstances where the size of the customer database is very small, varying sample size, and fails for larger datasets [7]. Therefore, we propose a deep learning (DL) algorithm to deal with massive financial data volumes since deep

learning (feature transformation) differentially weights features using historical data. Deep learning algorithms follow human brain architecture using artificial neural networks (ANNs) [12]. Feed-forward NNs use output from one layer as input for the next layer with no loops between layers [13]. Therefore, we employed the deep learning feed-forward (DFF) algorithm for experiments, trading-off between reduced computational load and highest accuracy for different algorithms. General advantages for DFF NNs can be summarized as

1) superior accuracy from DL;
2) includes more variables than classical ML;
3) DL algorithms can extract patterns while avoiding blind spots from extensive customer demographics, behavioural variables, and billions of customer engagement logs.
4) reduces time-consuming feature engineering and manual financial data analysis.

We employed a DFF NN approach not only due to DL useful features, but also we only need changing the last layer(s) to update the proposed model for future data, skipping time-consuming training models. One significant weakness for deep learning is that is remains a black-box model, i.e., DL does not express uncovered patterns in the underlying data in easily understandable ways. However, we conducted experimental analyses to identify churn effects and possible causes using Bayesian networks to address NN model complexity.

IV. METHODOLOGY

A. Problem definition

Data mining can identify useful knowledge in terms of pattern extraction from different sources, and various feature engineering tools can extract hidden patterns from massive datasets [12], [14]. Figure 1 shows that this study included 12 datasets from a superannuation company for real-world experiments to verify the proposed model's effectiveness. We define a customer as a churner if they closed their account during the subsequent 6 month time window. Therefore, we use binary outcome for each customer [0 or 1], where 1 means the account closed and 0 that it was not close in the subsequent 6 month time window.

We applied two main inclusion criteria for account tenure and balance to satisfy model development. First, we only retained data for customers with more than 6 month account tenure; and second, we removed account balances below $1500 to improve prediction since predicting churn probability for inactive accounts has low value for superannuation funds. Figure 2 shows features in the observation window to predict which user will be churned or non-churned in the next 6 month outcome window.
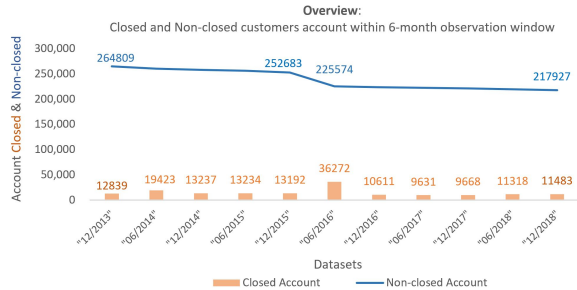

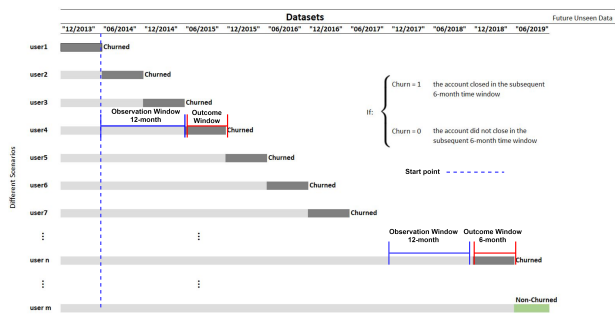Fig. 1: Closed and non-closed customer accounts in time periods of 6-month.

Fig. 2: Extracting feature values by sliding observation window on datasets

B. Churn propensity modelling

We compared the proposed DFF NN algorithm approach with seven current best-practise classifiers to evaluate the proposed framework effectiveness on the datasets with 12 month observation window and 6 month test window, as discussed in problem definition section. The workflow of proposed model shown in Figure 3. Feature engineering algorithms were employed for training (80%) and test (20%) sets to derive 193 features from 124363 total examples. Binary classification generally produces severely skewed data distributions, and hence we employed the synthetic minority oversampling technique (SMOTE) [17] in pre-processing to synthesize new examples for the minority class to ensure equal sample count (14031) for each class (1 or 0) in the training set. Experimental results confirmed improved performance for the proposed model using the sampling method.

Furthermore, We employed a majority voting ensemble from the scikit-learn library [18] to combine predictions from multiple models, which can helpful to improve model performance for classification tasks. Ensemble hard and soft voting were both applied on the comparison supervised models. Hard voting counts each individual classifier votes and the majority wins; whereas soft voting weights each prediction by classifier importance, and the target label with the greatest sum of weighted probabilities wins [18].
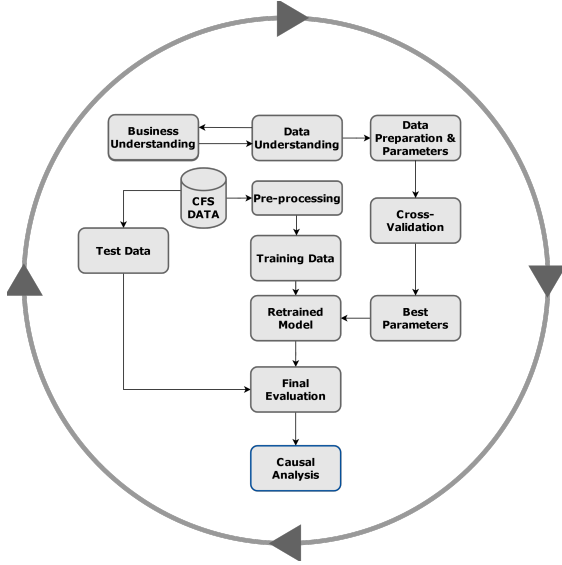
Fig. 3: Churn propensity modelling

TABLE I: Network Architecture

Network type	Deep feed-forward neural network
Hidden layers	4
Dense activation 1,2,and 3	tanh, and 2 x relu
Dropout 1 and 2	0.3 and 0.0
Output activation function	Sigmoid
Learning rate	0.000474718
Epochs	100
Optimization algorithm	ADAM

We defined a specific ANN architecture to address the research problem. Table I shows the DFF ANN's architecture optimised based on the hyperparameter tuning analysis. Finally, in the last step of the churn propensity model, we considered features that have the most predictive power on classifier outcome as causal variables. Then, we used a directed acyclic graph (DAG) to trace the latent variables and unknown parameters. Therefore, we take advantage of both Machine learning (ML) and Bayesian systems in this study. In ML, we predicted the target or independent variable by observing dependent variables, selected the most influential features to enhance prediction accuracy, and used them in our proposed causal Bayesian network. Thereinafter we employ a different approach for causal analysis. We built a dependency structure between dependent features, i.e., a causal graph, based on statistical relationships between independent features. Different methods to identify causal traces from large, labelled datasets enabled predicting more flexible causal relations [19].

V. EXPERIMENT SETUP AND RESULTS

A. Datasets

Experiments were conducted on 12 datasets for members holding accounts provided by a local finance com-
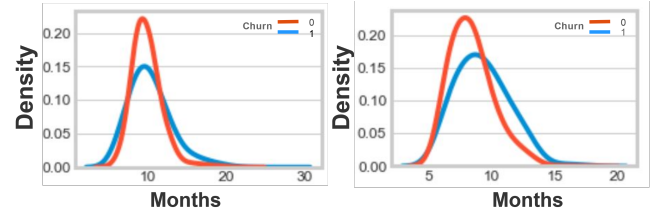

Fig. 4: Distribution of customer and account tenure

pany. Datasets included customer account, demographics, customer engagement, and financial data. Each dataset included approximately 250,000 examples with 88 features (71 numerical and 17 nominal). Data cleaning used criteria detailed in Section IV to improve prediction accuracy.

Figure 4 visualises customer and account tenure distribution data that are almost same on the 12 datasets. It can be inferred that approximately 90% customers had 5–12 month account tenure, with average tenure ≈ 7.6 month. Thus, most customers closed their accounts in less than one year. Therefore, we adopted the rational approach to only retain customer accounts that were open for more than six months, i.e., we eliminated all recently opened accounts.

B. Evaluation metrics

Measuring model performance is essential for ML, and several techniques evaluate model effectiveness for regression and classification tasks. We used area under the curve (AUC) and recall to visualise overall performance of a classifier. AUC represents how capable the model is to discriminate classes [15], with larger AUC indicating better distinguishability between churned and non-churned customers, and AUC ≈ 1 represents good separability. AUC can be defined as [6]

$$AUC = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} 1.pi > pj, \tag{1}$$

where i denotes to whole data points from 1 to m when the churn label 1 in churned class p_i and j denotes to all data points represent from 1 to n when the churn label 0 in non-churned class. Both p_j and p_j are probabilities outcome related to each class, here 1 is an indicator when the condition $p_i > p_j$ is true.

C. Correlation analysis

Relationships between variables can be measured as their correlation defined as [16]:

$$Cor(X,Y) = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}, \tag{2}$$

We used correlation analysis to identify potentially meaningful connections between variables, selecting
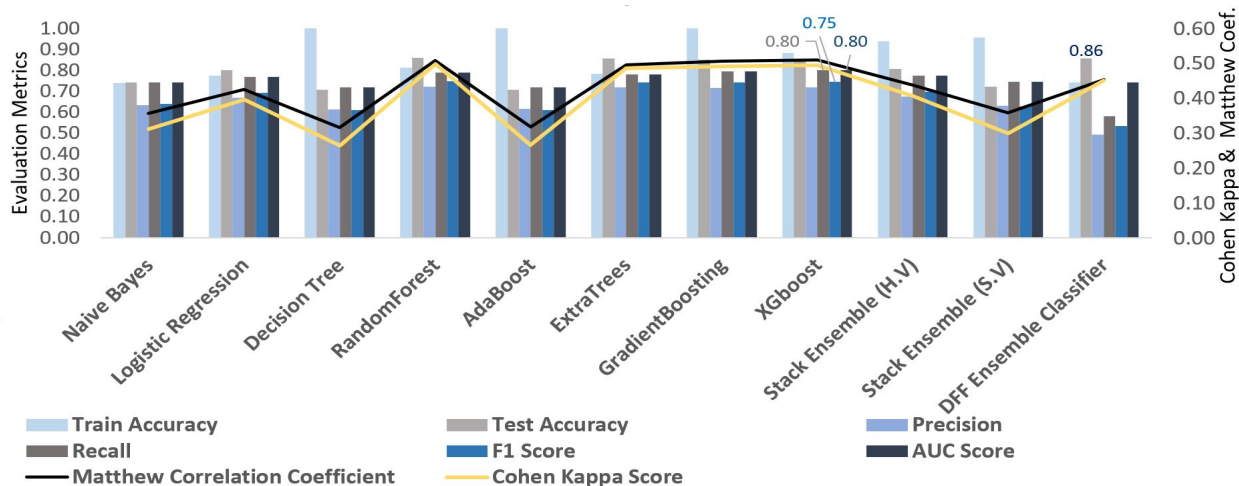
Fig. 5: Proposed DFF NN and current best-practise model outcomes

highly correlated relationships for subsequent causal discovery and hypothesis testing, as described in Section IV.

D. Causality analysis experiment setup

We employed Bayesian causal graphs to encode assumptions and determine dependency levels between features, using the DoWhy python package [20]. DoWhy performs causal discovery on all potential ways to identify a desired effect based on the Bayesian causal graph model, exploiting graph based criteria to find possible interpretation methods [20].

E. Prediction results

Figure 5 shows experiment outcomes conducted on the most updated observation and outcome windows. XGBoost outperformed the other algorithms with AUC = 80%; whereas the proposed DFF NN and Random Forest (RF) have almost the same performance, with improved AUC by 7.5% compared with logistic regression. Thus, DFF NN outcomes were comparable with current best-practise classifiers and achieved maximum prediction accuracy on test data. These findings provide a solid evidence base for exploiting DL reduced time-consuming feature engineering requiring expert knowledge on these specific financial datasets.

Descriptive approaches in statistical analysis define feature weights reflecting their contribution to push model output from its base (average output on the training dataset) to more meaningful outputs. Figure 6 shows SHAP, representing feature impacts on model output, where red features increase and blue reduce prediction outcomes relative to the base value, e.g. feature acc_balance_change_amount reduced and sg_recency increased prediction outcomes.
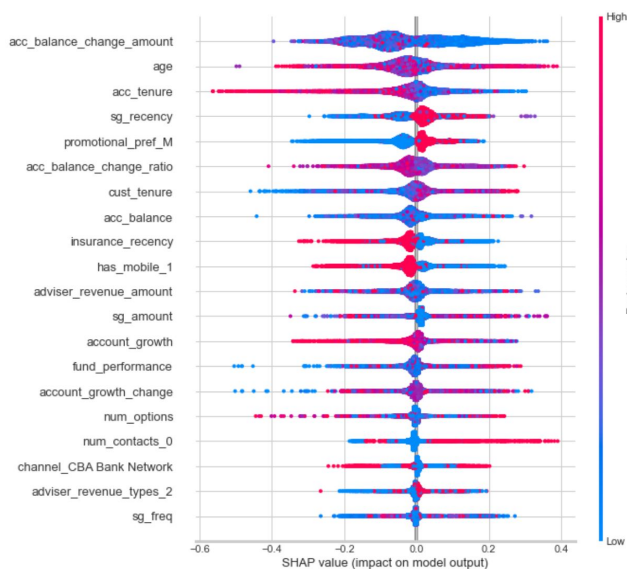

Fig. 6: SHAP visualisation on prediction outcomes. The age, acc_tenure, and sg_recency are the most influential features pushing the prediction higher, unlike feature acc_balance_change_amount that is pulling the model's prediction lower.

F. Causality analysis results

Figure 7 shows empirical results and related causal graphs based on assumptions that would affect churn as follows:

- High limit account balance might affect the churn as users with a lower limit (account balance) might not be loyal to the superannuation funds compared to customers with a high account balance.
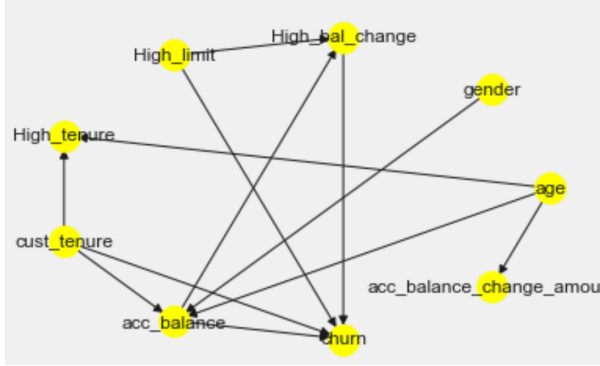- The account balance change amount could affect the customer tenure. The customer tenure is often

Fig. 7: Causal graph exhibiting assumptions encoded in the causal model

TABLE II: Causal analysis of customer churn

Variable	Estimate Mean Value	Pobability of churn
acc_tenure	-0.0262	decreased by ∼3%
annualrpt_pref	0.1444	increased by ∼14%
stmt_pref	-0.1427	decreased by ∼14%
acc_balance	-0.0916	decreased by ∼9%
account_growth	0.0331	decreased by ∼3%
cust_tenure	-0.0279	decreased by ∼3%
sg_recency	0.1563	increased by ∼15%
promotional_pref	-0.0864	decreased by ∼8%

based on account balance, after all. The account balance amount itself might affect the churn. An account balance should not be below $1500 in most superannuation funds, and a low account balance generally represents an inactive customer who may be willing to churn.

- A cascade relation between account balance and gender shows that gender would affect account balance then indirectly affect churn.
- Account growth and high balance change could directly affect churn. No variation account growth indicates stopped business for most accounts with consequential increased churn propensity.

We identified treatment causal effects on churn outcome based on the initial assumptions by holding other potential effects constant while changing the target treatment. Linear regression estimation indicates that effect $= -0.026298$ corresponds to churn probability reducing by $\approx$ 2.5% when the customer has higher account tenure. To test our assumption so that if the assumption is correct, the new estimation effect should not significantly change. Therefore, we applied random common cause (RCC) [20] to refute obtained estimates by adding independent random variable res_random as a common cause in the dataset. Outcome from the refuting method $= -0.026299$, almost identical to the estimation result. Thus, we can confirm the assumption was correct that high account tenure was a causal feature for churn outcome.

The treatment's causal effect on the outcome is based on the change in the value of the treatment variable. How strong the effect is a matter of statistical estimation. There are many methods for the statistical estimation of the causal effect. In this study, we used the "Propensity Score-Based Inverse Weighting" method [21] and concluded the estimations and churn probability results in Table II. For instance, the mean estimation for variable sg_recency is ~ 0.15, where it is equivalent to saying

that the probability of churn is increased by $\sim 15\%$ when the customer has higher days since the last day of super guarantee SG contribution. The mean estimate of ~ 0.03 for the account_growth variable can be concluded that churn probability increases by $\sim 3\%$ when the customer has a negative account growth rate.

The python implementation for the proposed framework and result visualisation, including hyper-parameter sensitivity analysis of DFF NN in hiplot, and our causal analysis in detail, is available in GitHub (https://github.com/DavidHason/Causal_Analysis), to simplify reproducing and improving this study experiment results.

VI. CONCLUSION

Losing customers is inevitable for most businesses, but churn can be managed at acceptable levels by investing in customers with churn risk. A new churn propensity model was designed and integrated with the Bayesian causal network. Unbalanced churned and non-churned classes were levelled in pre-processing, and then accuracy was compared between the proposed DFF NN approach and ten current best-practise churn prediction classifiers. Although XGBoost achieved superior AUC, DFF NN obtained comparable AUC with highest accuracy of all considered models on test data.

We analysed possible customer churn causes for a particular financial dataset created at superannuation fund(s) corporations. Causal analysis results confirmed variables representing recent SG contribution, annual report preference changed, account growth and balance amount were identified as confounding factors for customer churn with high degree of belief. Churn rate can be reduced by $\sim 3\%$ for customers with active account > 1year, consistent with expert knowledge.

A natural progression of this work is extending pattern mining techniques with smaller outcome windows that should be investigated to obtain more efficient prediction results in future studies. Also, different methods to identify causes of churn based on counterfactual causal analysis should be investigated.

VII. ACKNOWLEDGMENT

REFERENCES

[1] Domingos, B. Ojeme and O. Daramola, "Experimental Analysis of Hyperparameters for Deep Learning-Based Churn Prediction in the Banking Sector", Computation, vol. 9, no. 3, p. 34, 2021.

[2] Ahn, J. Hwang, D. Kim, H. Choi and S. Kang, "A Survey on Churn Analysis in Various Business Domains", IEEE Access, vol. 8, pp. 220816-220839, 2020.

[3] Pearl, J. (2009). Causal inference in statistics: An overview. Statistics surveys, 3, 96-146.

[4] Bilal Zorić, A. (2016). Predicting customer churn in banking industry using neural networks. Interdisciplinary Description of Complex Systems: INDECS, 14(2), 116-124.

[5] De Caigny, A., Coussement, K., De Bock, K. W., Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. International Journal of Forecasting, 36(4), 1563-1578.

[6] Culbert, B., Fu, B., Brownlow, J., Chu, C., Meng, Q., Xu, G. (2018, May). Customer churn prediction in superannuation: a sequential pattern mining approach. In Australasian Database Conference (pp. 123-134). Springer, Cham.

[7] Tamaddoni, A., Stakhovych, S., Ewing, M. (2016). Comparing churn prediction techniques and assessing their performance: A contingent perspective. Journal of service research, 19(2), 123-141.

[8] Mishra, A., Reddy, U. S. (2017, December). A novel approach for churn prediction using deep learning. In 2017 IEEE international conference on computational intelligence and computing research (ICCIC) (pp. 1-4). IEEE.

[9] Mohan, R., Chaudhury, S., Lall, B. (2021). Temporal Causal Modelling on Large Volume Enterprise Data. IEEE Transactions on Big Data.

[10] Yimin Huang and Marco Valtorta Identifiability in causal Bayesian networks: A sound and complete algorithm. , In Proceedings of the Twenty-First National Conference on Artificial Intelligence, pages 1149-1154, Edinboro, Scotland, 2006. AAAI Press.

[11] M. Shah, D. Adiga, S. Bhat and V. Vyeth, "Prediction and Causality analysis of churn using deep learning", 6th International Conference on Computer Science, Engineering and Information Technology (CSEIT-2019), 2019.

[12] Zaïane, O. R. (1999). Principles of knowledge discovery in databases. Department of Computing Science, University of Alberta, 20.

[13] Nielsen, M. A. (2015). Neural networks and deep learning (Vol. 25). San Francisco, CA: Determination press.

[14] Karp, A. H. (1998). Using logistic regression to predict customer retention. In Proceedings of the Eleventh Northeast SAS Users Group Conference. http://www. lexjansen. com/nesug/nesug98/solu/p095. pdf.

[15] Bradley, A.P.: Use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145–1159 (1997)

[16] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning - with Applications in R — Gareth James — Springer.

[17] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[18] Kramer, O. (2016). Scikit-learn. In Machine learning for evolution strategies (pp. 45-53). Springer, Cham.

[19] Lopez-Paz, D., Muandet, K., Schölkopf, B., Tolstikhin, I. (2015, June). Towards a learning theory of cause-effect inference. In International Conference on Machine Learning (pp. 1452-1461). PMLR.

[20] Sharma, A., Kiciman, E. (2020). DoWhy: An end-to-end library for causal inference. arXiv preprint arXiv:2011.04216.

[21] Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.