# Automated Underwriting in Life Insurance: Predictions and Optimisation (Industry Track)

Rhys Biddle, Shaowu Liu, Guandong Xu

Advanced Analytics Institute, University of Technology Sydney.
rhys.biddle@student.uts.edu.au,
shaowu.liu@uts.edu.au, guandong.xu@uts.edu.au

**Abstract.** Underwriting is an important stage in the life insurance process and is concerned with accepting individuals into an insurance fund and on what terms. It is a tedious and labour-intensive process for both the applicant and the underwriting team. An applicant must fill out a large survey containing thousands of questions about their life. The underwriting team must then process this application and assess the risks posed by the applicant and offer them insurance products as a result. Our work implements and evaluates classical data mining techniques to help automate some aspects of the process to ease the burden on the underwriting team as well as optimise the survey to improve the applicant experience. Logistic Regression, XGBoost and Recursive Feature Elimination are proposed as techniques for the prediction of underwriting outcomes. We conduct experiments on a dataset provided by a leading Australian life insurer and show that our early-stage results are promising and serve as a foundation for further work in this space.

## 1 Introduction

The concept of an insurance fund is to create a pool of wealth such that an unfortunate loss incurred by the few can be compensated by the wealth of the many [9]. It is clear that the success of this concept relies on accurately making the distinction between the few, those deemed to have a high chance of claiming, and the many, those deemed to have a low chance of claiming. The process of making such decisions is called underwriting.

The goal of underwriting from the perspective of the insurer is to accurately assess the risk posed by individual applicants, where risk in life insurance can be considered as the likelihood of an injury, sickness, disease, disability or mortality. A direct outcome of the underwriting process is the decision to accept or decline an individuals access into the insurance fund and what financial cost they should incur in exchange for access. Most individuals will incur a standard cost for access to the fund but some may incur a penalty, known as a *loading*, that should ideally reflect the level of risk they pose and their likelihood of claiming. In addition to a loading an individual may be granted special access to the fund but with certain claiming constraints attached, known as an *exclusion*. An exclusion is applied to prevent a specific individual claiming as a result of a particular event, back injury for example, but still allowing them access to the fund and rights to claim for other events they are not excluded from. Correctly identifying risks and applying the relevant loadings and exclusions during the underwriting process is fundamental to maintaining the wealth of a life insurance fund.

Underwriting is a tedious and labor intensive process on behalf of both the applicant and the underwriter. An applicant must fill out a highly personal questionnaire that delves into almost all aspects of their life which can be up to 100 pages long and consist of over 2 and a half thousand questions, an imposing amount of paperwork that can turn individuals off pursuing insurance cover. The industry is well aware of the growing market group of millennials, forecast to reach 75 percent of the workforce by 2025, who prioritise fast and seamless digital user experiences [1]. Current underwriting processes do not allow for these kinds of digital experiences and insurers are aware that significant time and costs must be allocated to transforming current practices in order to capture the attention of this growing market [1]. In addition to being tedious on behalf of the user this questionnaire must be closely examined by a team of skilled underwriters who must follow guidelines mixed with intuition to arrive at a decision, resulting in a process that takes many weeks to complete. The mixture of guidelines and intuition is evident in the common phrase in the industry that underwriting is both an art and a science [3]. It has been reported in industry trend analysis that there is a need to improve the quantitative methods that make up the science aspect of the underwriting process in order to maintain the relevancy of the industry. Fortunately, with recent advances in machine learning and pattern recognition, it becomes possible to make significant improvements to the decision making process.

Existing research on automated underwriting in life insurance sector is lacking in broad and deep coverage. A combination of a neural network and fuzzy logic is presented without any experimental validation in [11], the strengths of the proposed approach is only informally justified with no proofs. The prototyping and implementation of a knowledge based system given requirements gathered from a survey of 23 individuals from Kenyan life insurance companies was proposed in [9], however, no experimental validation provided. We believe that this lack of depth and coverage is due to the difficulty in gaining access to real world life insurance datasets and the proprietary nature of any endeavor to implement such a system. There is also a noted distinction between life and non-life insurance industries in the literature due to the differing complexity and sensitivity of the information involved. It has been shown in [13], that classical data mining algorithms, specifically Support Vector Regression(SVR) and Kernel Logistic Regression (KLR), can successfully classify the risk and accurately predict insurance premiums in the automotive industry using real-world data. In this paper, we aim to propose a prediction framework based on state-of-the-art machine learning algorithms and evaluate on 9-years of life insurance data collected by a major insurance provider in Australia.

The rest of the paper is organized as follows. In Section 2, we introduce the basic concepts of automated underwriting followed by problem formulation. Section 3 is devoted to describing our methodology. Experiment results are presented in Section 4, and Section 5 concludes.

## 2 Preliminary

This section briefly summarises necessary background of automated underwriting and problem formulation that form the basis of this paper.

### 2.1 Automated Underwriting

Automation of the underwriting process can assist in a number of ways and benefit all parties involved. Timeliness of the underwriting process can be significantly improved, instances of human error can be reduced and misunderstandings or knowledge gaps in the underwriters can be filled. The current underwriting completion time frame of weeks can be reduced significantly with the assistance of automated decision making tools. Most applications go through the underwriting process with no exclusion or loading applied, underwriters spend a lot of time dealing with these cases that could be streamlined and allow that time to be spent focusing on the more complex cases. In some instances rule-based expert systems have been crafted to identify and process these simple applications but they are complex and cumbersome to update in light of new information [3]. The breadth and detail covered by the thousands of questions within the questionnaire requires a considerably deep and wide knowledge base to be able to deeply understand the answers and the implications for risk. In addition to gaining a thorough understanding of these numerous knowledge areas, an ambitious task alone, there is the added difficulty of being able to identify the complex relationships between the diverse knowledge areas and how they can be used to forecast risk. The use of machine learning and pattern recognition tools can assist the underwriter in increasing their knowledge base and identifying these complex relationships.

### 2.2 Underwriting Outcome

One of the most important underwriting outcomes is identifying *exclusions*. An exclusion inhibits an individual from making a specific claim due to information gathered from the applicants questionnaire. The reason for exclusions are numerous, in the thousands, and considerably specific. This specific nature of the exclusion is necessary when evaluating any claim made by an individual. If an applicant has a history of left knee medical issues and experiences frequent pain or limitations as a result of these issues than they may be excluded from making any claims that related to their left knee. As well as numerous exclusions targeting specific claims they may also have a temporal condition attached, such as a 90 day or 12 month exclusion from a particular claim. Exclusions allow an insurance fund to tailor products to each individual applicant by choosing what specific risks they are willing to accept and provide cover for and those which they are not.

### 2.3 Exclusions Prediction Problem

The prediction of an exclusion can be approached as a supervised binary classification problem. We have a dataset $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i$ is a feature vector for applicant $i$ and $y_i$ is a binary label indicating the presence of a particular exclusion for applicant $i$. The feature vector $\mathbf{x}_i$ consists of the responses to all binary questions filled out by applicant $i$, some continuous features such as age and sum insured amounts. The questionnaire covers a large range of information about each applicant including family and medical history, occupation details, finances as well as leisure activities. In the current process of underwriting a team experts comes up with $y_i$ for each exclusion. We propose to learn a function $f$ that can accurately predict $y_i$ given $\mathbf{x}_i$ using the labels provided by the expert underwriters to evaluate the performance of $f$. There are a few properties of this problem that make it an interesting supervised classification problem. Firstly the questionnaire has been designed and refined over the years to catch as many risky applicants as possible yet make it streamlined for the applicant. This results in a questionnaire that contains conditional-branching, which is the creation of unique pathways through the questionnaire depending on responses to particular questions. A result of this conditional branching is that the responses to the questionnaire are considerably sparse because only a small subset of the questions need to be answered by all applicants, i.e., the majority of $x_i^j = 0$ for some questions $j$. Questions are designed to catch exclusions so for any exclusion we expect a small subset of feature vector $\mathbf{x}_i$ to be very strong features for the predictive task and the large majority to be redundant. In addition to this sparsity we have the added issue of class imbalance due to the specificity and rarity of exclusions. As mentioned previously exclusions must be detailed enough so that the insurer can cover themselves at claim time resulting in thousands of different and highly specific exclusion types.

# 3 Methodology

We propose to address the problems identified in the underwriting process and the gaps in the existing research by implementing and evaluating two learning models to the problem of exclusion prediction on a real-world dataset. There are two key goals of this work, the prediction of exclusions and providing recommendations for questionnaire optimisation. In building our methodology both predictive accuracy and high interpretability of results are equally important. This limits our choice of data preparation methods and learning algorithm as addressed in the following sections.

## 3.1 Feature Selection

Reducing the size of the feature space is an important first step in learning problems and provides many benefits. A reduction in the size of feature vectors decreases the learning time, can improve accuracy and avoid overfitting [8]. A decrease in learning time is due to the smaller size of the training data after the reduction of the feature space. Overfitting is a well known pitfall and occurs when a model learns the training data so well that the predictive capabilities on new unseen data begins to suffer [14]. A large feature space with numerous redundant features can lead to overfitting and a reduction of these redundant features is a strategy to combat this pitfall [14]. In addition to this a model trained on a large feature space is complex and can be difficult to interpret.

There are two main approaches to feature space reduction, transformation-based and selection-based methods. Transformation-based methods perform a transformation of the initial input feature space to a smaller space [12, 7] where as selection-based methods look to find an optimal subset of the original feature space [14].

Transformation-based methods are unsuitable for our work because they would destroy the one-to-one relationship of feature to question response. Preservation of this one-to-one relationship is key for us to assess the impact of individual questions and the respective response provided by an applicant.

There are numerous approaches that can be taken for feature selection methods. Filter methods involve ranking features under a chosen criterion and specifying a threshold at which to remove features from the feature space for training. Wrapper methods use the prediction results of a learning algorithm to identify and select features that are deemed important by the learning algorithm. In Embedded methods the feature selection process is part of the learning algorithm and it is difficult to separate the two processes. We have chosen to implement a wrapper method in our learning pipeline.

For the wrapper method we have chosen Recursive Feature Elimination (RFE) [8, 6]. RFE is an iterative wrapper method that consists of training a classifier on numerous feature subsets and provides feature rankings for each subset. The three steps for RFE are: i) train a classifier using feature set $\mathbf{f}$; ii) get feature importances from trained classifier, rank them; iii) remove a of subset the worst performing features for $\mathbf{f}$. There are two main parameters to be set for RFE, the size of the desired feature subset at the end of the algorithm and the number of features to remove at each iteration. The size of the desired feature subset can be found via cross-validation. RFE can be fit across all training folds in the cross-validation loop and the feature subset that gives the best averaged results across all testing folds can be selected as the optimal feature subset.

### 3.2 Prediction Models

**Logistic Regression and Regularisation** Logistic regression was chosen as a baseline method because linear models are a favored tool in the insurance sector because of the simple implementation, interperatability and their connection with traditional statistics [10]. Logistic Regression is a popular statistical method used for modeling binary classification problems by prescribing a weight to all input features to perform a linear separation of the two classes. There is no feature selection inherent in the construction of Logistic Regression model however the addition of $l1$ regularisation addresses this. Logistic Regression with the addition of $l1$ as penalty term is referred to as Lasso Regression. The addition of this penalty term in Lasso Regression performs feature selection because it shrinks the weights of unimportant features to zero.

**Gradient Boosting Trees** Gradient Boosting methods [5, 2] are tree-based ensemble methods for supervised learning that are founded on the hypothesis that numerous weak learners provide more accurate predictions than a single learner [10]. A weak learner in this context is a simple model that can be considered to be only slightly better than random guessing. The simplest approach to combining all the predictions from the individual learners to arrive at a single prediction is via a voting procedure. A prediction by each weak learner is considered a vote and all of these are tallied up and the label predict by most weak learners is chosen as the final prediction. A motivation for using tree-based ensemble methods in insurance is that the decision making process is made up a large number of simple conditional rules, if applicant ticks "yes" to question A but "no" to question B then accept, which can be learnt by the numerous different weak learners in the ensemble [10]. Interpretability of Gradient Boosting methods in comparison to other learning techniques of similar power and complexity, such as Neural Networks and Support Vector Machines, is another motivation for using it in our work. Gradient Boosting methods provide clear and intuitive metrics for each input feature that indicate their importance in the resulting prediction, this aligns with our goal for providing recommendations for questionnaire optimisation. The nature of tree construction in gradient boosting means that all variables are candidates for splitting the tree and are evaluated. A direct result of this is that feature selection is inherent within the ensemble construction and is capable of dealing with redundant features. In this work, the XGBoost [2] implementation is employed.

### 3.3 Proposed Pipelines

We propose to use four separate feature selection and classification pipelines for implementation and evaluation. Firstly a pipeline of RFE with a standard Logistic Regression model as the learning algorithm for the RFE process. Cross-validation will be used to select the ideal number of features and the Logistic Regression model will be fit on the reduced subset produced by the RFE procedure. Our second pipeline will consist of Lasso Regression with the cross-validation used to select the ideal strength of the $l1$ penalty term. Another pipeline will be XGBoost with cross-validation to select the ideal number of weak estimators and the learning rate. Lastly a pipeline of RFE with XGBoost as learning algorithm.

## 4 Experiment

In this section, we introduce the experimental settings and a large-scale data collection from Australian insurer, followed by experiment results and discussions.

### 4.1 Dataset Preparation

We have been given access to data from a leading insurer in the Australian life insurance sector dating from 2009 to 2017. As with any real-world dataset a considerable amount of effort was needed to prepare the data for modelling. There were several key issues that needed to be addressed before modeling could be performed on the entire dataset. Firstly the questionnaire data and the underwriting data had been stored and maintained by separate entities due to privacy concerns. In addition to this the data had changed hands several times across this time period due to organizational takeovers and vendor changes. Questionnaire data was most impacted by these changes and underwent four major changes in this time period. There were numerous changes that were made to the applicant data in how it was stored, such as different attributes and data types with no master data dictionary available to resolve these changes. In this time period the questionnaire itself had also changed with the addition, removal and modification of the questions contained within. These issues are currently being resolved and as a result the latest version of the questionnaire data, 2014-2017, has been used for modeling.

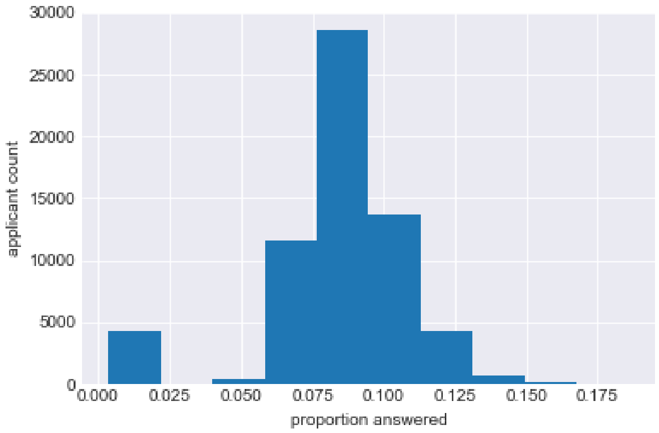

**Fig. 1.** Histogram of response proportion on questionnaire

A straightforward match between the questionnaire data and underwriting data was not possible for privacy reasons and as a result we had to come up with a process to merge the separate data stores. We used three attributes relating to the applicant found in both application and underwriting data. These were related to the suburb, age and gender of the applicant. In such a large dataset we found numerous applicants sharing these traits so we used the date in which each applicant was entered into the two separate systems to resolve any ambiguous cases. Through this process we were able to identify approximately 60 thousand individuals from the 2014-2017 period.

As can be seen in Fig. 1 the response rates to questions are considerably low, the majority of applicants fill out less than 10 percent of the entire questionnaire, due to the conditional-branching structure of the questionnaire. This results in sparse feature vectors for the majority of applicants. As well as the sparse feature vectors the data exhibits sparsity in relation to the application of exclusions resulting in extreme class imbalances when predicting the binary problem of exclusion application. There are over 1 thousand different exclusions applied in the data set. Many of these exclusions are extremely rare occurring far too infrequently, single digit frequency counts, and thus

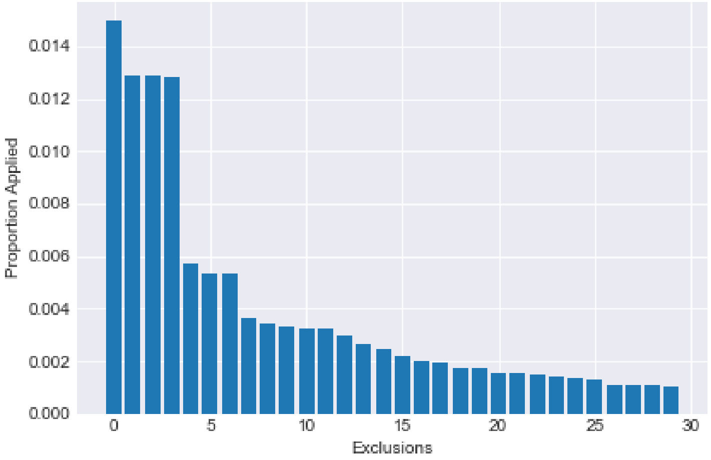

**Fig. 2.** Histogram of application rate for the 30 most common exclusion codes in descending order

not included in experiments. The most frequently applied exclusion is applied to only 1.5 percent of all applications, see Fig. 2.

### 4.2 Experimental Settings

We ran our proposed pipeline on the 20 most frequent exclusion codes. Our experiments were conducted using the scikit learn library for the python programming language. For the two pipelines that utilised RFE as feature transformation the following was implemented. Nested cross validation (CV) loop containing RFE with CV as the outer loop and Grid Search with CV as the inner loop to optimise the hyper-parameters of the learning algorithm. The embedded feature selection approaches required no such nested loop as there was no need for the transformation before prediction. CV was set to 5 stratified folds for all experiments, seed was kept the same for all experiments to ensure of the same dataset splits. For all learning algorithms the sampling procedure was weighted to account for the class imbalance.

### 4.3 Evaluation Metrics

We used area under the Receiver Operating Curve (ROC) [4] to evaluate the performance of our approach. An ROC curve is a plot of the rate of true positive predictions, correctly predict an positive example, against the rate of false positive predictions, predict a positive label when in fact negative. This is plotted across all thresholds for the prediction score of the model. The area under the ROC curve (AUC) is a single metric that can be interpreted as the ability of a model to correctly rank a positive example as more likely to be positive than a negative example. AUC is a common tool for comparing models in supervised binary classification problems.

### 4.4 Results and Discussions

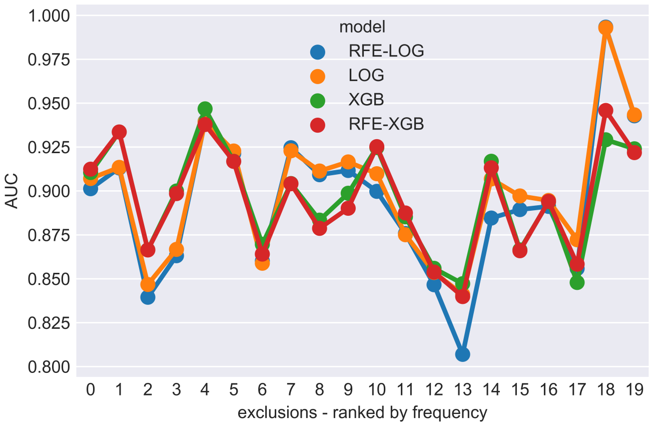

**Fig. 3.** Prediction results on the $20$ most common exclusion codes, ranked in descending order. The results from four pipelines in this figure i) RFE-LOG : Recursive Feature Elimination with Logistic Regression ii) LOG : Logisitic Regression with $l1$ regularisation iii) XGB : Extreme Gradient Boosting, iv) RFE-XGB : Recursive Feature Elimination with Extreme Gradient Boosting

**Predictions** The prediction results vary considerably between exclusion codes, see figure 3. The worst average AUC across all models is $0.83$, while the best average AUC is $0.96$. In all but five of the top 20 exclusions setting a strong $l1$ regularization penalty on Logistic Regression provides greater or equal predictive accuracy when compared to using RFE with CV as a feature selection phase before Logistic Regression with no $l1$ penalty as shown in Fig. 3. However the mean difference in AUC between Logistic Regression with $l1$ and RFE and Logistic Regression with no penalty is $-0.006$ which is insignificant. The mean difference in AUC between XGBoost and RFE with XGBoost is even more insignificant at only $-0.0006$. XGBoost and Logistic Regression with $l1$ regularisation deal adequately with the feature selection process requiring no need for the prior feature selection step. Logistic Regression with $l1$ is the best performing model with an average AUC $0.0035$ units greater than XGBoost the next best model. There is little to separate these models in terms of predictive performance. The number of relevant features needed by each model shows a clear gap between the models. XGBoost uses far fewer features to get similar accuracy as shown in Fig. 4. This has

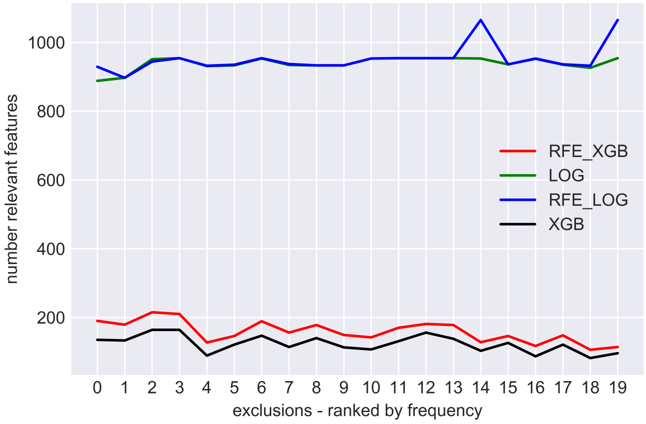

**Fig. 4.** Number of features used by the four modelling pipelines i) RFE-LOG : Recursive Feature Elimination with Logistic Regression ii) LOG : Logisitic Regression with $l1$ regularisation iii) XGB : Extreme Gradient Boosting, iv) RFE-XGB : Recursive Feature Elimination with Extreme Gradient Boosting

implications for our recommendations for the questionnaire optimisation. Logistic Regression on average uses 4 times as many features as XGBoost with a similar prediction accuracy on average. Our recommendations for questionnaire optimisation is based on the feature importance as given bu the XGBoost model.

**Question Optimisation using Feature Importance** We further explore the trained model and discover insights of feature importance, i.e., the importance of each feature (question) played for each exclusion. The result is shown as heatmap in Fig. 5 where x-axis shows the selected exclusions and y-axis shows the features. Each cell is colored from blue to red where red indicating the feature is high relevant to deciding the corresponding exclusion. For example, the heatmap shows that the question “Alcohol” is commonly used for deciding the exclusion “lose of income”. Despite of the red cells, the blue cells are also important for optimising the questions. For example, the question “Asthma Medication” has shown no relevance to any of the exclusions, which suggests this is a potential redundant question. Note that due to the large number of questions and exclusions, only a small fragment of the full heatmap is shown here.

## 5 Conclusions

In this paper, we implemented and evaluated a number of different machine learning algorithms and feature selection methods to predict the application of exclusions in life insurance applications. The results show that this simple approach performs well and

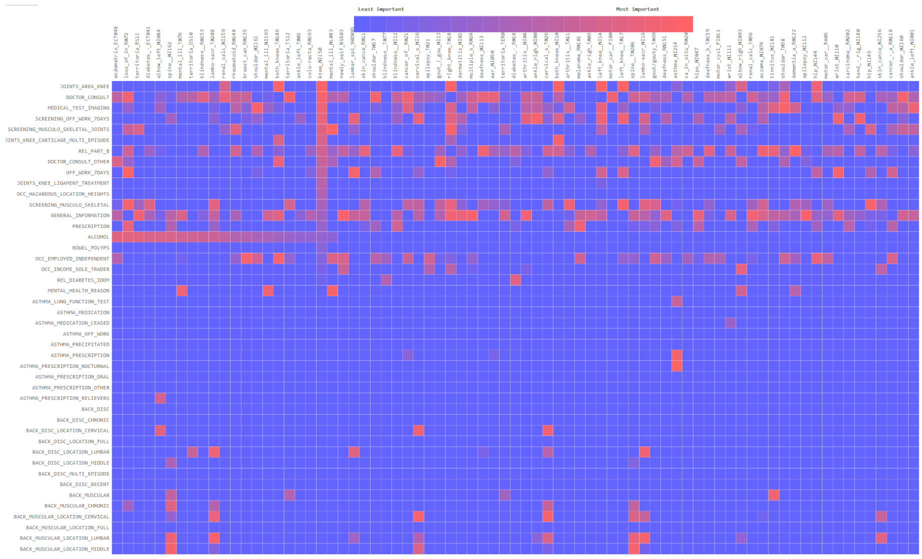

**Fig. 5.** Feature importance as heatmap

can add value to the insurer. XGBoost is the most ideal model due to the need for the significantly smaller number of features needed to produce similar accuracy. For future work we would like to look into implementing a cost-sensitive approach to the prediction of exclusions. Data from claims made by applicants along with the current data would be needed to completely understand the cost of the underwriting decisions. We currently have not processed enough of the dataset to utilize the claims data making this approach unfeasible at the moment. Given that we only have last 3 years worth of usable data at present moment the number of claims for this period is too small to be of any use. Another direction for future work is the incorporation of the free text responses provided by the applicants into the feature set.

## References

1. Howlette B., Rajan M., and S. P. Chieng. Future of life insurance in australia. Technical report, PricewaterhouseCoopers, 2017.
2. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.
3. Gandhi D. and Kaul R. Life and health - future of life underwriting. *Asia Insurance Review*, pages 76–77, Jun 2016.
4. Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
5. Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

6. Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90, 2006.
7. Qinghua Hu, Jinfu Liu, and Daren Yu. Mixed feature selection based on granulation and approximation. *Knowledge-Based Systems*, 21(4):294–304, 2008.
8. Guyon I., Weston J., Barnhill S., and Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, pages 389–422, 2002.
9. Joram M. K., Harrison B. K., and Joseph K. N. A knowledge-based system for life insurance underwriting. *International Journal of Information Technology and Computer Science*, pages 40–49, 2017.
10. Guelman L. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, pages 3659–3667, 2012.
11. Arora N. and Vij S. A hybrid neuro-fuzzy network for underwriting of life insurance. *International Journal of Advanced Research in Computer Science*, pages 231–236, 2012.
12. Jensen R. and Shen Q. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. *IEEE Transactions on Knowledge and Data Engineering*, pages 1457–1471, 2004.
13. Kacelan V., Kacelan L., and Buric M. N. A nonparametric data mining approach for risk prediction in car insurance: a case study from the montenegrin market. *Economic Research-Ekonomska Istraivanja*, pages 545–558, 2017.
14. Rodriguez-Galiano V.F., Luque-Espinar J.A., Chica-Olmo M., and Mendes M.P. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Science of the Total Environment*, pages 661–672, 2018.