Complex & Intelligent Systems (2018) 4:145–154
https://doi.org/10.1007/s40747-018-0072-1

ORIGINAL ARTICLE

Risk prediction in life insurance industry using supervised learning algorithms

Noorhannah Boodhun[1] · **Manoj Jayabalan**[1]

Received: 7 February 2018 / Accepted: 2 April 2018 / Published online: 7 April 2018

Abstract
Risk assessment is a crucial element in the life insurance business to classify the applicants. Companies perform underwriting process to make decisions on applications and to price policies accordingly. With the increase in the amount of data and advances in data analytics, the underwriting process can be automated for faster processing of applications. This research aims at providing solutions to enhance risk assessment among life insurance firms using predictive analytics. The real world dataset with over hundred attributes (anonymized) has been used to conduct the analysis. The dimensionality reduction has been performed to choose prominent attributes that can improve the prediction power of the models. The data dimension has been reduced by feature selection techniques and feature extraction namely, Correlation-Based Feature Selection (CFS) and Principal Components Analysis (PCA). Machine learning algorithms, namely Multiple Linear Regression, Artificial Neural Network, REPTree and Random Tree classifiers were implemented on the dataset to predict the risk level of applicants. Findings revealed that REPTree algorithm showed the highest performance with the lowest mean absolute error (MAE) value of 1.5285 and lowest root-mean-squared error (RMSE) value of 2.027 for the CFS method, whereas Multiple Linear Regression showed the best performance for the PCA with the lowest MAE and RMSE values of 1.6396 and 2.0659, respectively, as compared to the other models.

Keywords Life insurance underwriting · Machine learning · Predictive analytics · Correlation · Principal components · Feature selection · Dimensionality reduction

Introduction

The big data technologies revolutionize the way insurance companies to collect, process, analyze, and manage data more efficiently [1,2]. Thus, proliferate in various sectors of insurance industries such as risk assessment, customer analytics, product development, marketing analytics, claims analysis, underwriting analysis, fraud detection, and reinsurance [3,4]. Telematics is a typical example where big data analytics is being vastly implemented and is transforming the way auto insurers price the premiums of individual drivers [5].

Individual life insurance organizations still rely on the conventional actuarial formulas to predict mortality rates and premiums of life policies. Life insurance companies have recently started carrying out predictive analytics to improve their business efficacy, but there is still a lack of extensive research on how predictive analytics can enrich the life insurance domain. Researchers have concentrated on data mining techniques to detect frauds among insurance firms, which is a crucial issue due to the companies facing great losses [6–8].

Manulife insurance company in Canada was the first to offer insurance to HIV suffering applicants through analyzing survival rates [9]. Analytics help in the underwriting process to provide the right premiums for the right risk to avoid adverse selection. Predictive analytics has been used by Property and Casualty (P&C) insurers for over 20 years, primarily for scoring disability claims on the probability of recovery. Predictive analytics approach in life insurance

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40747-018-0072-1) contains supplementary material, which is available to authorized users.

✉ Manoj Jayabalan
manoj@apu.edu.my

Noorhannah Boodhun
noorhannah@hotmail.com

[1] Asia Pacific University of Technology and Innovation, Technology Park Malaysia, 57000 Bukit Jalil, Kuala Lumpur, Malaysia

mainly deals with modeling mortality rates of applicants to improve underwriting decisions and profitability of the business [10].

Risk profiles of individual applicants are thoroughly analyzed by underwriters, especially in the life insurance business. The job of the underwriter is to make sure that the risks are evaluated, and premiums as accurately as possible to sustain the smooth running of the business. Risk classification is a common term used among insurance companies, which refers grouping customers according to their estimated level of risks, determined from their historical data [11]. Since decades, life insurance firms have been relying on the traditional mortality tables and actuarial formulas to estimate life expectancy and devise underwriting rules. However, the conventional techniques are time-consuming, usually taking over a month and also costly. Hence, it is essential to find ways to make the underwriting process faster and more economical. Predictive analytics have proven to be useful in streamlining the underwriting process and improve decision-making [12]. However, extensive research has not been conducted in this area. The purpose of this research is to apply predictive modeling to classify the risk level based on the available past data in the life insurance industry and recommend the most appropriate model to assess risk and provide solutions to refine underwriting processes.

Literature review

Over the years, life insurance companies have been attempting to sell their products efficiently, and it is known that before an application is accepted by the life insurance company, a series of tasks must be undertaken during the underwriting process [13].

According to [14], underwriting involves gathering extensive information about the applicant, which can be a lengthy process. The applicants usually undergo several medical tests and need to submit all the relevant documents to the insurance agent. Then, the underwriter assesses the risk profile of the customer and evaluates if the application needs to be accepted. Subsequently, premiums are calculated [15]. On average, it takes at least 30 days for an application to be processed. However, nowadays, people are reluctant to buy services that are slow. Due to the underwriting process being lengthy and time-consuming, customers are more prone to switch to a competitor or prefer to avoid buying life insurance policies. Lack of proper underwriting practices can consequently lead to customers being unsatisfied and a decrease in policy sales.

The underwriting service quality is an essential element in determining the corporate reputation of life insurance businesses and helps in maintaining an advantageous position in a competitive market [16]. Thus, it is crucial improving the underwriting process to enhance customer acquisition and customer retention.

Similarly, underwriting process and the medical procedures required by the insurance company to profile the risks of the applicants can be expensive [17]. Usually, all the costs to perform the medical examinations are initially borne by the firm. Underwriting costs are fully paid from the contract and can last 10–30 years. In case, where there is a policy lapse, the insurer incurs great losses [18]. Therefore, it is imperative to automate the underwriting process using analytical processes. Predicting the significant factors impacting the risk assessment process can help to streamline the procedures, making it more efficient and economical.

A study by [19] shows that low underwriting capacities are a prominent operational problem among insurance companies surveyed in Bangladesh. Another threat to the life insurance businesses is that they can face adverse selection. Adverse selection refers to a situation where the insurers do not have all information on the applicant, and they end up giving life insurance policies to customers with a high-risk profile [20]. Insurance firms with competent underwriting teams stress on making the least possible losses. In other words, the insurers strive to avoid adverse selection as it can have powerful impacts on the life insurance business [21]. Adverse selection can be avoided by correctly classifying the risk levels of individual applications through predictive analytics, which is the goal of this research.

Methods and techniques

The research approach involves the collection of data from online databases. The hypotheses about possible relationships between variables would be investigated using defined logical steps. The research paradigm deals with a positivist approach, as it is mainly a predictive study involving the use of machine learning algorithms to support the research objectives.

Figure 1 shows the data analysis flow chart. It gives an idea of the stages that have been going through systematically to build the prediction models.

Description of data set

The data set consists of 59,381 applications with 128 attributes, which describe the characteristics of life insurance applicants. The data set comprises of nominal, continuous, as well as discrete variables, which are anonymized. Table 1 describes the variables present in the data set.

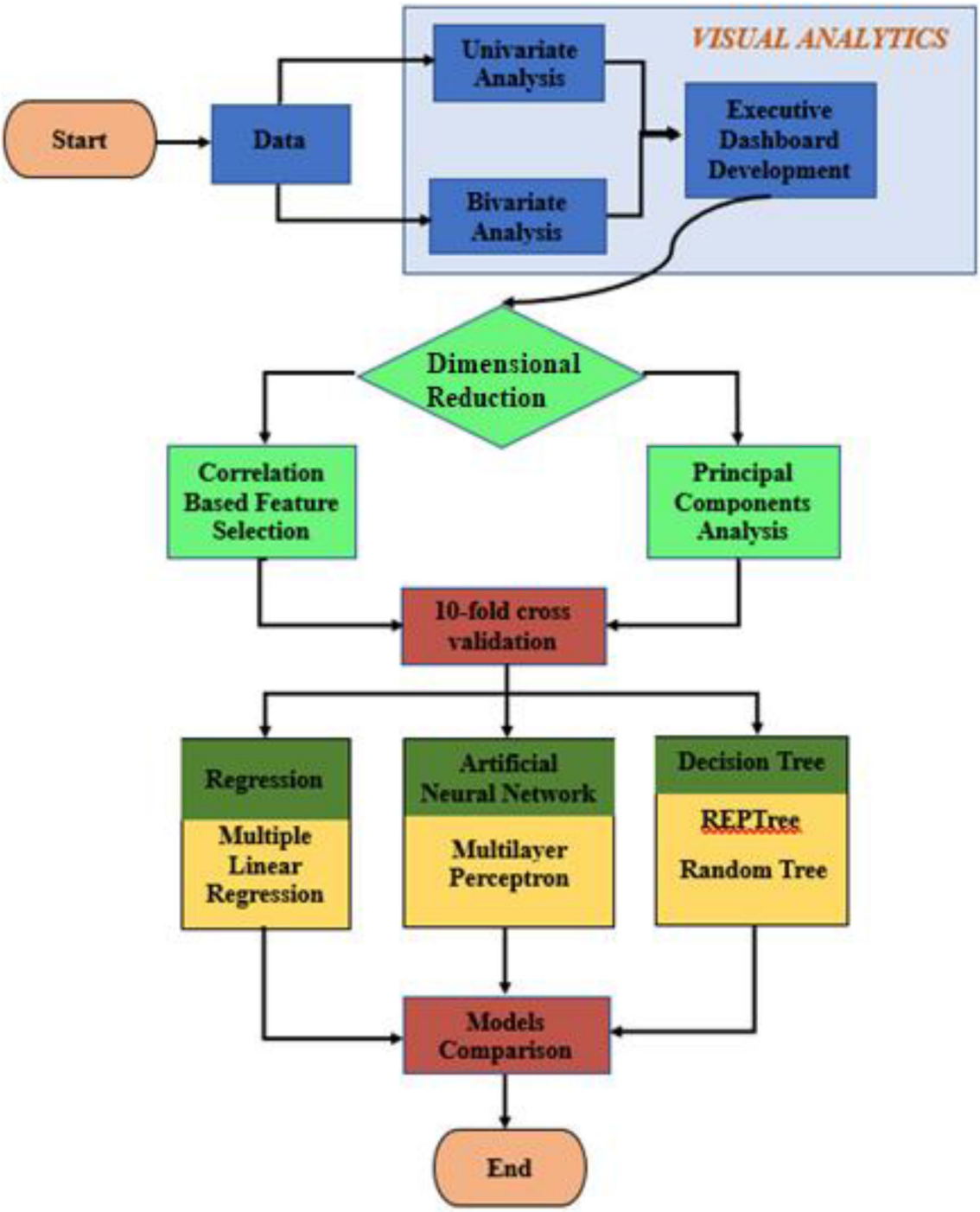

Fig. 1 Data analysis approach

Data pre-processing

Data pre-processing, also known as the data cleaning step, implicates that noisy data or outliers are removed from the target dataset. This step also encompasses the development of any strategies needed to deal with the inconsistencies in the target data. In case of discrepancies, specific variables will be transformed to ease analysis and interpretation. In this step, the data gathered from Prudential Life Insurance will be cleaned to treat missing values to make the data consistent with analysis. Prudential Life Insurance data set has attributes with a remarkable amount of missing data. The missing data structure and mechanism will be studied to decide the suitable imputation method for the data set. Usually, there exist three

Table 1 Data set description

Attributes	Type	Description
Product_Info_1-7	Categorical	7 normalized attributes concerning the product applied for
Ins_Age	Numeric	Normalized age of an applicant
Ht	Numeric	Normalized height of an applicant
Wt	Numeric	Normalized weight of an applicant
BMI	Numeric	Normalized Body Mass Index of an applicant
Employment_Info_1-6	Numeric	6 normalized attributes concerning employment history of an applicant
InsuredInfo_1-6	Numeric	6 normalized attributes offering information about an applicant
Insurance_History_1-9	Numeric	9 normalized attributes relating to the insurance history of an applicant
Family_Hist_1-5	Numeric	5 normalized attributes related to an applicant's family history
Medical_History_1-41	Numeric	41 normalized variables providing information on an applicant's medical history
Medical_Keyword_1-48	Numeric	48 dummy variables relating to the presence or absence of a medical keyword associated with the application
Response	Categorical	Target variable, which is an ordinal measure of risk level, having 8 levels

mechanisms of missing data, namely, Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) [23].

MCAR This is the case when the distribution of the missing values does not show any relationship between the observed data and the missing data. In other words, the missing values are like a random sample of all the cases in the feature.

MAR This mechanism requires that the missingness may be dependent on other observed variables, but independent of any unobserved features. In other words, missing values do not depend on the missing data, yet can be predicted using the observed data.

MNAR This mechanism, on the other hand, implies that the missing pattern relies on the unobserved variables; that is, the observed part of the data cannot explain the missing values. This missing data mechanism is the most difficult to treat as it renders the usual imputation methods meaningless.

Data exploration using visual analytics

The Exploratory Data Analysis (EDA) will comprise of univariate and bivariate analyses. The EDA would allow the researcher to understand the different distributions that the features exhibit. On the other hand, for the bivariate analysis, the relationships between the different features and the response attribute, risk level, would be analyzed. Therefore, it would help to understand the extent to which the independent variables are capable of impacting the response variable significantly. Due to page limitation, the results of EDA not discussed here. The interested reader can refer the attached supplementary data analysis.

Visual analytics will be performed on the data set to gain insights into the data structure. The data will be visualized using charts and graphs to show the distribution of the data set to have a better knowledge of which prediction models will be more suitable for the data set.

The interactive dashboards are very helpful to business users to understand their data. The dashboard will comprise of several graphs relating to the data set on one screen. As such, trends and patterns in the data set can be studied while showing the relationships between different attributes. In short, a summary of the data can be seen in one view.

Dimensionality reduction

The dimensionality reduction involves reducing the number of variables to be used for efficient modeling. It can be broadly divided into feature selection and feature extraction. Feature selection is a process involved in selecting the prominent variables, whereas the feature extraction applied to transform the high dimensional data into fewer dimensions to be used in building the models. Thus, dimensionality reduction is used to train machine learning algorithms faster as well as increase model accuracy by reducing model overfitting [24].

There are several techniques available for feature selection classified under the filter methods, wrapper methods, and embedded methods. The filter method uses a ranking to provide scores to each variable, either based on univariate statistics or depending on the target variable. The rankings can then be assessed to decide whether to keep or discard the variable from the analysis [25]. The wrapper method, on the contrary, takes into account a subset of features and com-

pares between different combinations of attributes to assign scores to the features [26]. The embedded method is slightly more complicated, since the learning method usually decides which features are best for a model while the model is being built [27]. Attributes can be selected based on Pearson's correlation, Chi-square, information gain ratio (IGR), and several other techniques [28,29].

In contrary, the feature extraction process derives new features from the original features, to increase the accuracy via eliminating redundant features and irrelevant features. This research limits itself on two methods, namely the correlation-based feature selection method and principal component analysis-based feature extraction method. The discussion about these methods is presented in the below subsections.

Correlation-based feature selection

Correlation-based feature selection (CFS) evaluates subsets of attributes based on a hypothesis, which is a useful subset of features contains highly correlated features with the class, yet uncorrelated to each other [30]. This feature selection method is easy to understand and fast to execute. It removes noisy data and improves the performance of algorithms. It does not require the analyst to state any limits on the selected number of attributes but generates the optimal number of features by itself. It is usually classified under the filter method.

The correlation values for the feature selection are not only calculated based on Pearson's correlation coefficient but are based on the measures namely, minimum description length (MDL), symmetrical uncertainty, and relief [31,32]. CFS requires the nominal attributes in a data set to be discretized before calculating the correlation. Nonetheless, it works on any data set, independent of the data transformation methods used [31]. In a study, [33] found that CFS was more accurate compared to IGR. Similarly, [34] concluded that they obtained the highest accuracy for their classification problem using a CFS as compared to other feature selection methods.

Principal components analysis feature extraction

Principal components analysis (PCA) is an unsupervised linear feature extraction technique aimed at reducing the size of the data by extracting features having most information [35]. PCA uses the features in the data set to create new features, known as the principal components. The principal components are then used as the new attributes to create the prediction model. The principal components have better explaining power compared to the single attributes. The explaining power can be measured by the explained variance ratio of the principal components. This value shows how much information is retained by the combined features [36].

PCA works by calculating eigenvalues of the correlation matrix of the attributes. The variance explained by each newly generated component is determined and the components retained are those which describe the maximal variation in the data set. Scholars like [37] and [38] conducted studies using PCA, and they concluded that the PCA method is useful when used with predictive algorithms.

Comparison between correlation-based feature selection and principal components analysis feature extraction

PCA creates new features by combining the existing ones to create better attributes, while correlation feature selection only selects the best attributes as they are, that is, without the creation of new ones, based on the predictive power. While PCA does some feature engineering with the attributes in the data set, the resulting new features are more complicated to explain, as it is difficult to deduce meanings from the principal components. CFS, on the other hand, is relatively easier to understand and interpret, as the original features are not combined or modified.

In this research, four machine learning algorithms are implemented on CFS and PCA. Following the implementation of the algorithms, the accuracy measures will be compared to evaluate the effectiveness of both feature reduction techniques.

Supervised learning algorithms

This section will elaborate on the different algorithms implemented on the data set to build the predictive models. The techniques namely, Multiple Linear Regression, REPTree, Random Tree, and Multilayer Perceptron.

Multiple linear regression model

Multiple linear regression shows the relationship between the response variable and at least two predictor variables by fitting a linear equation to the observed data points. In other words, the equation is used to predict the response variable based on the values of the explanatory variables collectively [39].

Multiple linear regression models are evaluated based on the sum of squared errors which shows the average distance of the predicted data points to the observed data values. The model parameter estimates are usually calculated to minimize the sum of squared errors, such that the accuracy of the model is increased. The variables significance in the regression equation are determined by statistical calculations and are mostly based on the collinearity and partial correlation statistics of the explanatory features [40].

REPTree algorithm

The REPTree classifier is a type of decision tree classification technique. It can build both classification and regression trees, depending on the type of the response variable. Typically, a decision tree is created in case of discrete response attribute, while a regression tree is developed if the response attribute is continuous [41].

Decision trees are a useful machine learning technique for classification problems. A decision tree structure comprises of a root node, branches, and leaf nodes aimed at representing data in the form of a tree-like graph [42]. Each internal node represents the tests performed, and the branches are representative of the outcome of the test. The leaf nodes, on the other hand, represent class labels. Decision trees mainly use the divide and conquer algorithm for prediction purposes. Decision trees are a widely used machine learning technique for prediction and have been implemented in several studies [43–45]. The advantage of using decision trees is that they are easy to understand and explain.

REPTree stands for Reduced Error Pruning Tree. It makes use of regression tree logic to create numerous trees in different iterations. Mostly, this algorithm is used as it is a fast learner, which develops decision trees based on the information gain and variance reduction. After creating several trees, the algorithm chooses the best tree using the lowest mean-square-error measure when pruning the trees [46].

Random Tree

The Random Tree is also a decision tree algorithm, but it is different from the previously explained REPTree algorithm in the way it works. Random Tree is a machine learning algorithm which accounts for k randomly selected attributes at each node in the decision tree. In other words, random tree classifier builds a decision tree based on random selection of data as well as by randomly choosing attributes in the data set.

Unlike REPTree classifier, this algorithm performs no pruning of the tree. The algorithm works in a way that it conducts backfitting, which means that it estimates class probabilities based on a hold-out set. In [47], the authors used the random tree classifier in their research together with CFS and concluded that the classifier works efficiently with large data sets. Likewise, [48] investigated on the use of random trees in their work and the scholars were able to achieve high levels of model accuracy by modifying the parameters of the random tree classifier.

Artificial neural network

The artificial neural network is an algorithm, which works like the neural network system in the human brain. It is comprised of many highly interconnected processing elements, also known as the neurons. The neurons are usually organized in three layers, which are the input, hidden, and output layers. The neurons keep learning to improve the predictive performance of a model used in problem-solving. This adaptive learning capability of the model is very beneficial for developing high accuracy prediction models given a data set for training [49]. Artificial neural networks are widely utilized in numerous domains, for instance for speech and image recognition, machine translation, artificial intelligence, social network filtering, and medical diagnosis [50–52]. The neural network model makes use of backpropagation to classify instances. Backpropagation refers to a supervised learning method which calculates the error of each neuron after a subset of the data is processed and distributes back the errors through the layers in the network. The neural network can also be altered when it is trained [53].

Experiments and results

Data pre-processing

The data set has 59,381 instances and 128 attributes. The data pre-processing step carried out using R programming to detect the missing data.

Missing data mechanism

The attributes that are showing more than 30% missing data would be dropped from the analysis [54]. Therefore, attributes, Employment_Info_1, Employment_Info_4, Employment_Info_6, and Medical_History_1 are the only features, which are retained for further analysis. These four attributes will need to be treated to impute their missing values.

The data were tested for MCAR using the Little's test [55]. The null hypothesis is that the missing data are MCAR. However, a significance value of 0.000 was obtained which implies that the null hypothesis was rejected. Thus, the Little's test revealed that the missing data are not entirely at random. If the data are not MCAR, they can be MAR or MNAR. Usually, there is no such reliable test to determine directly if the data are MAR, because this requires acquiring some of the missing data, which is not possible when using secondary data sets. To understand the missing value mechanism, patterns in the data set can be examined.

Figure 2 depicts the plot for the missing value in the data set, with the variable having most missing values on the top of the y-axis and least missing values on the bottom.

The visualization of the missing data structure suggests a random distribution of the missing value observations. The pattern of missing data and non-missing data is scattered

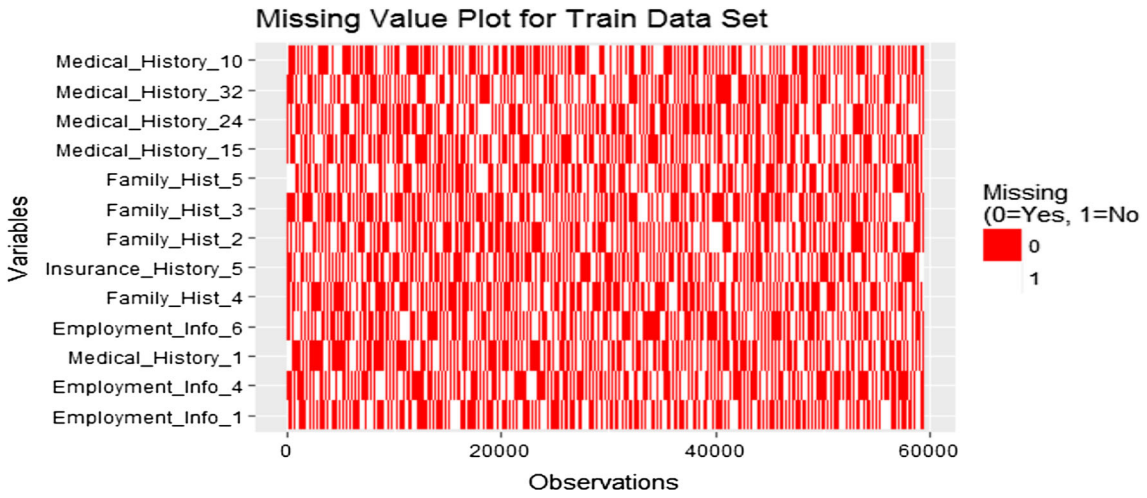

Fig. 2 Missing value plot for train data

throughout the observations. Therefore, the data set in this study is assumed to be MAR and treatment of the missing values will be based on this assumption.

Missing data imputation

If the data are assumed to be MAR, the multiple imputation is an appropriate technique to replace the missing values in the features. Multiple imputation is a statistical technique which uses available data to predict missing values. Multiple imputation involves three steps, namely, imputation, analysis, and pooling as determined by [56]. Multiple imputation is more reliable than single imputation, such as mean or median imputation as it considers the uncertainty in missing values [57,58].

The steps for multiple imputation involve:

- Imputation: This step involves the imputation of the missing values several times depending on the number of imputations stated. This step results in a number of complete data sets. The imputation is usually done by a predictive model, such as linear regression to replace missing values by predicted ones based on the other variables present in the data set.
- Analysis: The various complete data sets formed are analyzed. Parameter estimates and standard errors are evaluated.
- Pooling: The analysis results are then integrated together to form a final result.

The MICE (Multivariate Imputation via Chained Equations) package in R has been utilized to do the multiple imputations [59]. The missing data were assumed to be MAR. The categorical variables were removed and only numeric attributes were used to do the imputation.

Executive dashboard

The cleaned data set was used in Microsoft Power BI to create dynamic visualizations to gain better insights about the data. Power BI is an influential analytical tool offering a friendly interface, whereby interactive visualization can be easily created to ease interpretation and do efficient reporting. The resulting cleaned data set consisted of 118 variables and 59, 381 instances.

Figure 3 shows the dashboard, which has been created using the Prudential insurance data set. The dashboard shows several graphs that are interactive with each other. This dashboard mainly presents the distribution of demographic variables in the data set with the response variable. For instance, BMI, age, weight, and family history and how they vary with the different risk levels. Such a dashboard provides insights into the customer data. Thus, the life insurance company knows its applicants better and has better engagement with them.

Comparison between feature selection and feature extraction

The experiment was carried out using Waikato Environment for Knowledge Analysis (WEKA). The correlation-based feature selection was implemented using a BestFirst search method on a CfsSubsetEval attribute evaluator. Thirty-three variables were selected out of a total 117 features, excluding the response variable in the data set.

The PCA was implemented using a Ranker search method on a PrincipalComponents, attribute evaluator. The PCA feature extraction technique provides a rank for all the 117 attributes in the data set. The technique works by combining the attributes to create new features, which can predict the target variable in a better way. Furthermore, the selec-

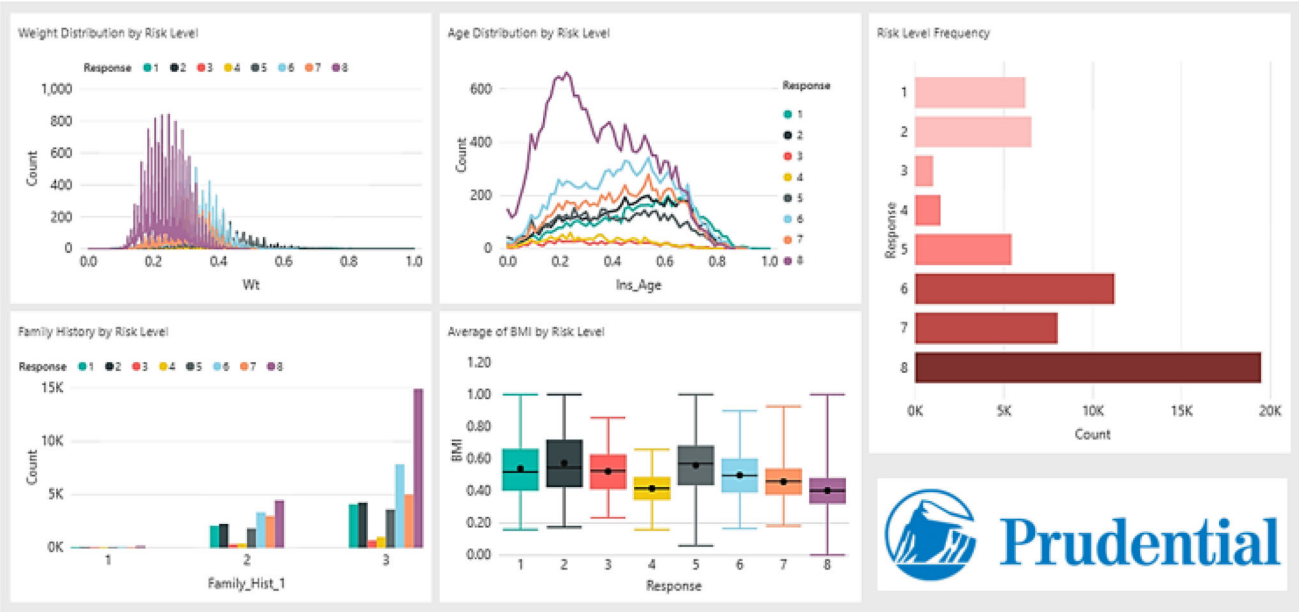

Fig. 3 Life insurance dashboard

Table 2 Comparison of algorithms between CFS and PCA

Algorithms	Error measures			
	CFS		PCA	
	MAE	RMSE	MAE	RMSE
Multiple linear regression	1.5872	2.0309	1.6396	2.0659
Artificial neural network	1.7859	2.369	1.7261	2.3369
REPTree	1.5285	2.027	1.6973	2.1607
Random Tree	1.7892	2.7475	2.0305	2.9142

tion was conducted to choose optimum variables with better predictive capabilities based on the standard deviation.

A cut-off threshold of 0.5 has been used to decide on the number of principal components to retain from the data set. In other words, only those attributes which standard deviation value that is half of that of the first principal component (2.442) would be retained. Therefore, those principal components with a standard deviation of 1.221 or more were retained, resulting in 20 attributes.

Following the dimensionality reduction, the reduced data set was exported and used for building the prediction models using machine learning algorithms discussed in the previous section. Model validation has been performed using a k-folds (tenfold) cross-validation.

Four models were developed using multiple linear regression, artificial neural network, REPTree, and random tree classifiers on the CFS and PCA. The error measures are shown in Table 2.

For the CFS, the model developed using REPTree classifier shows the highest performance with the lowest mean absolute error (MAE) value of 1.5285 and lowest root mean square error (RMSE) value of 2.027 as compared to the other models. However, for the PCA, the model developed with multiple linear regression shows the best performance with the data set by having the lowest MAE and RMSE values as 1.6396 and 2.0659, respectively. Moreover, random tree classifier shows the highest error values for both feature selection techniques.

Comparing between the feature selection and feature extraction techniques, CFS shows that most of the models achieved lower errors compared to PCA. Multiple linear regression, REPTree, and random tree classifiers show better performance when used with CFS, while artificial neural network shows a better performance with PCA.

Conclusions

This research has specific implications for the business environment. Data analytics is now the trend that is gaining significance among companies worldwide. In the life insurance domain, predictive modeling using learning algorithms can provide the notable difference in the way which business is done as compared to the traditional methods. Previously, risk assessment for life underwriting was conducted using complex actuarial formulas and usually was a very lengthy process. Now, with data analytical solutions, the work can be done faster and with better results. Therefore, it would enhance the business by allowing faster service to customer, thereby increasing satisfaction and loyalty.

The data obtained from Prudential Life Insurance were pre-processed using R programming. Missing values were detected using Missing At Random (MAR), and the multiple imputation methods were used to replace the missing values. Those attributes have more than 30% missing data which were eliminated from the analysis. Furthermore, a dashboard was built to show the effectiveness of visual analytics for data-rich business processes.

The research demonstrated the use of dimensionality reduction to reduce the data dimension and to select only the most important attributes which can explain the target variable. Thirty-three attributes were selected by the CFS method, while 20 features were retained by the PCA.

The supervised learning algorithms namely, Multiple Linear Regression, Artificial Neural Network, REPTree, and Random Tree were implemented. The model validation was performed using tenfold cross-validation. The performance of the models was evaluated using MAE and RMSE. Findings suggested that the REPTree algorithm had the highest accuracy with lowest MAE and RMSE statistics of 1.5285 and 2.027, respectively, for the CFS method. Conversely, for the PCA method, Multiple Linear Regression showed the best performance with MAE and RMSE values of 1.6396 and 2.0659, respectively. Ultimately, it can be concluded that machine learning algorithms can be efficient in predicting the risk level of insurance applicants.

Future work relates to the more in-depth analysis of the problem and new methods to deal with specific mechanisms. Customer segmentation is the division of the data set into groups with similar attributes can be implemented to segment the applicants into groups with similar characteristics based on the attributes present in the dataset. For example, similar employment history, insurance history, and medical history. Following the grouping of the applicants, predictive models can be implemented to contribute to a different data mining approach for the life insurance customer data set.

The dashboards can be extended depending on the availability of the data. For instance, financial dashboards can be built showing the premiums received and claims paid by the firm within a given period to ease profit and loss analysis. Another report can be of sales showing policy sales by different customers and time of the year, so that marketing strategies could be improved.

References

1. Sivarajah U, Kamal M, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. J Bus Res 70:263–286
2. Joly Y, Burton H, Irani Z, Knoppers B, Feze I, Dent T, Pashayan N, Chowdhury S, Foulkes W, Hall A, Hamet P, Kirwan N, Macdonald A, Simard J, Hoyweghen I (2014) Life Insurance: genomicsStratification and risk classification. Eur J Hum Genet 22:575–579
3. Umamaheswari K, Janakiraman D (2014) Role of data mining in Insurance Industry. Int J Adv Comput Technol 3:961–966
4. Raj A, Joshi P (2017) Changing face of the Insurance Industry. [Online]. https://www.infosys.com/industries/insurance/white-papers/Documents/changing-face-insurance-industry.pdf
5. Fan C, Wang W (2017) A comparison of underwriting decision making between telematics-enabled UBI and traditional auto insurance. Adv Manag Appl Econ 7:17–30
6. Goleiji L, Tarokh M (2015) Identification of influential features and fraud detection in the Insurance Industry using the data mining techniques (Case study: automobile's body insurance). Majlesi J Multimed Process 4:1–5
7. Joudaki H, Rashidian A, Minaei-Bidgoli B, Mahmoodi M, Geraili B, Nasiri M, Arab M (2016) Improving fraud and abuse detection in general physician claims: a data mining study. Int J Health Policy Manag 5:165–172
8. Nian K, Zhang H, Tayal A, Coleman T, Li Y (2016) Auto insurance fraud detection using unsupervised spectral ranking for anomaly. J Fin Data Sci 2:58–75
9. Bell M (2016) Is analytics the underwriting we know? [Online]. https://insurance-journal.ca/article/is-analytics-changing-the-underwriting-we-know/
10. Fang K, Jiang Y, Song M (2016) Customer profitability forecasting using Big Data analytics: a case study of the insurance industry. Comput Ind Eng 101:554–564
11. Cummins J, Smith B, Vance R, Vanderhel J (2013) Risk classificaition in Life Insurance, 1st edn. Springer, New York
12. Bhalla A (2012) Enhancement in predictive model for insurance underwriting. Int J Comput Sci Eng Technol 3:160–165
13. Mishr K (2016) Fundamentals of life insurance theories and applications. In: 2nd ed, Delhi: PHI Learning Pvt Ltd,
14. Wuppermann A (2016) Private information in life insurance, annuity and health insurance markets. Scand J Econ 119:1–45
15. Prince A (2016) Tantamount to fraud? Exploring non-disclosure of genetic information in life insurance applications as grounds for policy rescission. Health Matrix 26:255–307
16. Chen TJ (2016) Corporate reputation and financial performance of life insurers. Geneva Papers Risk Insur Issues Pract 41:378–397
17. Huang Y, Kamiya S, Schmit J (2016) A model of underwriting and post-loss Test without commitment in competitive insurance market. SSRN Electron J
18. Carson J, Ellis CM, Hoyt RE, Ostaszewski K (2017) Sunk costs and screening: two-part tariffs in life insurance. SSRN Electron J 1–26
19. Mamun DMZ, Ali K, Bhuiyan P, Khan S, Hossain S, Ibrahim M, Huda K (2016) Problems and prospects of insurance business in Bangladesh from the companies' perspective. Insur J Bangladesh Insurance Acad 62:5–164
20. Harri T, Yelowitz A (2014) Is there adverse selection in the life insurance market? Evidence from a representative sample of purchasers. Econ Lett 124:520–522
21. Hedengren D, Stratmann T (2016) Is there adverse selection in life insurance markets? Econ Inq 54:450–463
22. The Kaggle Website. [Online]. https://www.kaggle.com/c/prudential-life-insurance-assessment/data/

23. Nicholson J, Deboeck P, Howard W (2015) Attrition in developmental psychology: a review of modern missing data reporting and practices. Int J Behav Dev 41:143–153
24. Hoque N, Singh M, Bhattacharyya DK (2017) EFS-MI: an ensemble feature selection method for classification. Complex Intell Syst
25. Haq S, Asif M, Ali A, Jan T, Ahmad N, Khan Y (2015) Audio-visual emotion classification using filter and wrapper feature selection approaches. Sindh Univ Res J 47:67–72
26. Ma L, Li M, Gao Y, Chen T, Ma X, Qu L (2017) A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation. IEEE Geosci Remote Sensing Soc 14:409–413
27. Mirzaei A, Mohsenzadeh Y, Sheikhzadeh H (2017) Variational relevant sample-feature machine: a fully Bayesian approach for embedded feature selection. Neurocomputing 241:181–190
28. Kumar V, Minz S (2014) Feature selection: a literature review. Smart Comput Rev 4:211–229
29. Novakovic J, Strbac P, Bulatovic D (2016) Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav J Oper Res 21:119–135
30. Hira Z, Gillies D (2015) A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinf 2015:1–13
31. Hall M (2000) Correlation-based feature selection for discrete and numeric class machine learning, Working Paper Series. Hamilton, New Zealand: The University of Waikato
32. Doshi M, Chaturvedi S (2014) Correlation based feature selection (CFS) technique to predict student performance. Int J Comput Netw Commun 6:197–206
33. Chinnaswamy A, Srinivasan R (eds) (2017) Performance analysis of classifiers on filter-based feature selection approaches on microarray data. Bio-Inspired Computing for Information Retrieval Applications. United States of America, IGI Global
34. Hernández-Pereira E, Bolón-Canedo V, Sánchez-Maroño N, Álvarez-Estévez D, Moret-Bonillo V, Alonso-Betanzos A (2016) A comparison of performance of K-complex classification methods using feature selection. Inf Sci 328:1–14
35. Sharifzadeh S, Ghodsi A, Clemmensen L, Ersbøll B (2017) Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection. Eng Appl Artif Intell 65:168–177
36. Taguchi Y, Iwadate M, Umeyama H (2015) Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. BMC Bioinf 16:1–26
37. Shi X, Guo Z, Nie F, Yang L, You J, Tao D (2016) Two-dimensional whitening reconstruction for enhancing robustness of Principal Component Analysis. IEEE Trans Pattern Anal Mach Intell 38:2130–2136
38. Yi S, Lai Z, He Z, Cheung Y, Liu Y (2017) Joint sparse principal component analysis. Pattern Recogn 61:524–536
39. Forkuor G, Hounkpatin O, Welp G, Thiel M (2017) High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple Linear Regression models. PLOS One 12
40. Chatterjee S, Hadi A (2015) Regression analysis by example, 5th edn. Wiley, USA
41. Naji H, Ashour W, Alhanjouri M (2017) A new model in Arabic text classification using BPSO/REP-Tree. J Eng Res Technol 4:28–42
42. Gokgoz E, Subasi A (2015) Comparison of decision tree algorithms for EMG signal classification using DWT. Biomed Signal Process Control 18:138–144
43. Bhaskaran S, Lu K, Aali M (2017) Student performance and time-to-degree analysis by the study of course-taking patterns using J48 decision tree algorithm. Int J Model Oper Manag 6:194
44. Sudhakar M, Reddy C (2016) Two step credit risk assessment model for retail bank loan applications using Decision Tree data mining technique. Int J Adv Res Comput Eng Technol 5:705–718
45. Joshi A, Dangra J, Rawat M (2016) A Decision Tree based classification technique for accurate heart disease classification and prediction. Int J Technol Res Manag 3:1–4
46. Kalmegh S (2015) Analysis of WEKA Data Mining algorithm REPTree, Simple Cart and RandomTree for classification of Indian news. Int J Innov Sci Eng Technol 2:438–446
47. Gupta A, Jain P (2017) A Map Reduce Hadoop implementation of Random Tree algorithm based on correlation feature selection. Int J Comput Appl 160:41–44
48. Gupta S, Abraham S, Sugumaran V, Amarnath M (2016) Fault diagnostics of a gearbox via acoustic signal using wavelet features, J48 Decision Tree and Random Tree Classifier. Indian J Sci Technol 9:1–8
49. Demuth H, Beale M, Jess O, Hagan M (2014) Neural network design, 2nd edn. ACM Digital Library, USA
50. Ata R (2015) Artificial neural networks applications in wind energy systems: a review. Renew Sustain Energy Rev 49:534–562
51. Chowdhury M, Gao J, Chowdhury M (2015) Image spam classification using Neural Network. In :International Conference on Security and Privacy in Communication Systems, pp 622–632
52. Tkac M, Verner R (2016) Artificial neural networks in business: two decades of research. Appl Soft Comput 38:788–804
53. Dongmei H, Shiqing H, Xuhui H, Xue Z (2017) Prediction of wind loads on high-rise building using a BP neural network combined with POD. J Wind Eng Ind Aerodyn 170:1–17
54. Mertler C, Reinhart R (2016) Advanced and multivariate statistical methods, 6th edn. Routledge, New York
55. Li C (2013) Little's test of missing completely at random. Stata J 13:795–809
56. Rubin D (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
57. Garson G (2015) Missing values analysis and data imputation. Statistical Associates Publishing, Asheboro
58. Lee K, Roberts G, Doyle L, Anderson P, Carlin J (2016) Multiple imputation for missing data in a longitudinal cohort study: a tutorial based on a detailed case study involving imputation of missing outcome data. Int J Soc Res Methodol 19:575–591
59. Cheng X, Cook D, Hofmann H (2015) Visually exploring missing values in multivariable data using a graphical user interface. J Stat Softw 68:1–23