Journal of Applied Technology and Innovation
vol. 1, no. 1, (2017), pp. 58-73

A Review of Data Analytical Approaches in the Insurance Industry

Noorhannah Boodhun
Faculty of Computing, Engineering & Technology
Asia Pacific University of Technology and Innovation
57000 Kuala Lumpur, Malaysia
noorhannahb@gmail.com

Abstract - Insurance companies are facing a growth in their transactional data. Valuable information can be gained from the data collected by using analytical approaches. The key areas in which data analytics can be useful to insurance firms is customer level analytics, risk assessment and prediction as well as fraud detection. Commonly used techniques to identify patterns in the data sets are clustering and classification to be able to predict future occurrences of events. The main purpose of implementing analytics among the insurance firms is to better understand their customers, minimize their losses and gain a competitive advantage in the market. A review of several algorithms used to analyse insurance data is provided in this paper together with some evaluations of the different approaches.

Index Terms – Data analytics; insurance industry; risk prediction; fraud detection

1. Introduction

In the global era, demand of insurance is increasing and therefore insurance companies have a large amount of data at their disposal. The insurance sector generally comprises of life, health and non-life insurance. Insurance can be defined as the hedging of the risk of an uncertain loss (Robson, 2015). Insurance companies are expanding all over the world and like other industries, they also must bear the cost of expansion. The insurance market faces threats like fraud and abuse cases as well as it encounters the risk of customer attrition (Sithic & Balasubramanian, 2013; Mohammadi, Albadvi & Teymorpur, 2014). Therefore, it is critical for the insurance companies to come up with intelligent solutions to sustain their businesses. A good approach is to implement analytics to their rich amount of data to enable better decision making. Analytics provide companies with the abilities to gain useful insights from their customer database and hence can allow the insurance companies to gain competitive advantage over their adversaries (Umamaheswari & Janakiraman, 2014). Nowadays, more businesses are shifting towards the use of data analytical techniques to improve operations (Sivarajah et al., 2017). With the revolution of technology, insurance industries have experienced many changes in the way they operate. Technological innovations, like data warehousing has allowed them to cut down their costs of storing and accessing data. However, the conventional techniques of analysing data may not be reliable in today's world, where the existence of big data is prominent. Hence, it is essential for the insurance sector to benefit from the availability of modern data mining techniques and machine learning algorithms to scrutinize their available data.

The several data analytical approaches in different insurance industries is studied in this paper. The importance and efficacity of the techniques are elucidated and the objective of this research is to provide a review of the various methods used to analyse a broad variety of insurance data as an endeavour to improve the industry. The scope focuses on three key features, specifically, customer analytics, risk prediction and fraud detection in the insurance sector. The paper is divided as follows. Section 2 describes the application of data analytics in the insurance industry, followed by Section 3 which highlights the discussion of the literature whereby prominent limitations in the approaches for data analytics in the insurance industry are elaborated. Finally, Section 4 outlines the conclusion.

2. Data Analytical Approach in the Insurance Sector

In this section, the different data analytical approaches in the insurance sector will be discussed. The areas discussed will be customer analytics, risk prediction and fraud detection.

2.1 Customer Analytics

Customer analytics refers to studying the behaviour of customers in order to anticipate what they want from the company. It is usually carried out by using data mining techniques, specifically, predictive analytics with the objective to increase customer loyalty and enhance customer retention (Zhou & Chen, 2014). Data mining is a powerful tool allowing insurance companies to concentrate on valuable information gathered about the behaviour of their clients. It involves classification and segmentation of customers according to various criteria in order to model their behaviour and purchase patterns. As such, insurance firms can profile their customers based on their data and eventually target them with appropriate policies which they might be keen to buy (Desik et al., 2016). Insurance firms are highly dependent on customers, thus doing analytics enhances the marketing strategies of the firms and aids in cost reduction (Ravasan & Mansouri, 2015). Customer level analytics offers the potential of accurately identifying the specific policies to sell to particular customers. Researchers propose various algorithms, such as classification, clustering, association, regression, and several others, that have been widely studied and implemented on insurance data to create prediction models (Rahman et al., 2017; Hanafizadeh & Paydar, 2013).

Golmah (2014) carried out an analysis on customer segmentation in the automobile insurance industry. A case study of 23 automobile insurance companies in

Iran was used to analyse customers' choices and personalize marketing campaigns accordingly. Data mining tool used was Self Organizing Maps (SOMs), which applied unsupervised classification of the insurance customers to understand their behaviour in order to increase customer loyalty. Data collection was done by using questionnaire and the target respondents were automobile insurance customers. The questionnaire comprised of four sections collecting demographics data, automobile characteristics data, automobile insurance information and factors impacting insurance selection. Out of the 2000 questionnaires distributed via email, 1260 effective responses were used for the analysis after data cleaning. The SOM technique was implemented using MATLAB software to cluster customers into distinct groups. The prediction accuracy of the SOM was assessed using statistical measures like mean absolute error (MAE) and root mean square error (RMSE). The calculated MAE and RMSE values were 0.04 and 0.06 respectively, demonstrating that the model was precise. Moreover, a U-matrix map was used to visualize distinct clusters in the data set. The results established four clusters of customers with a set of characteristics. The clusters were those to whom cost of policies matter, quality of service matter, style of payment is crucial and those who chose their insurance policies in regard to their family and friends. The SOM method of clustering customers proved to be useful on one data set and can therefore be applied to other insurance data sets to improve marketing strategies.

Another research dealing with the segmentation of customers for the analysis of behaviour is done by Jandaghi & Moradpour (2015). They used the data mining technique, fuzzy clustering on life insurance customers. Secondary data was collected from the database of Pasargad life insurance company for a period of April to October 2014 and consisted of 1071 customers. Age, gender, number of children, job, relationship between insured and insurer, insurance term, payment method, premium, number of supplementary coverage were among the efficient attributes used for the analysis, after interviewing experts for the selection of right attributes. R statistical software was employed to carry out the data mining task of fuzzy clustering method. Findings suggested an optimal number of two clusters, which were labelled 'investment policy' and 'life security'. Customer segmentation is concluded to be an appropriate technique to improve sales and performance of the insurance company. However, a prominent limitation of this research is that the authors did not justify the accuracy of their model by any statistical means of evaluating performance of a method.

Roodpishi & Nashtaei (2015) conducted a research using insurance data in order to perform market basket analysis to better understand customer demand. Data from a car insurance company in Anzali city of Iran was used to perform the analysis. The data set was for the year 2014 and was collected from 300 insurance clients using questionnaires. The primary technique employed to mine the data was the K-means clustering algorithm which clustered the data into five optimal clusters based on the demographic attributes present in the data set, namely gender, age, occupation, education level, marital status, place of residence and income. The data mining software SPSS Clementine was the tool utilised to perform the clustering method. Further in the research, association rules were performed on each cluster using causal extraction algorithms to identify hidden patterns in the data set and to predict customer behaviour. Moreover, the performance evaluation criteria used to validate the modelling was the sum square error (SSE). A useful implementation of the study is to target potential clients to sell appropriate policies. However, the data set used was limited to only one region which is why the findings cannot be generalised for the whole country.

Fang, Jiang & Song (2016) identified critical factors to predict customer profitability in the insurance industry using big data analytical approach. Data was obtained from a commercial industry company in Taiwan. Customer profitability is the segmentation factor to predict valuable and nonvaluable clients. The data set consisted of 25,000 observations with several attributes such as premium paid, claim amount, payment type, guarantee years, age, gender, occupation, region. The research aimed to develop a new model for the prediction of customer profitability. Liability reserve was the new variable added to distinguish the model from existing traditional customer profitability models, which only deal with premium and claim amounts. The new model was tested using 7:3 ratio of training and testing data set. Random forest regression was applied to forecast the insurance customer profitability. Moreover, the model was compared against linear regression model, decision tree, support vector machine (SVM) and generalised boosted model. Random forest regression provided higher accuracy compared to the other models, accounting for 99.03% of variation in the model. However, a pertinent drawback in the model was that it did not include customer income data, which could affect the accuracy of the model.

A study conducted by Delafrooz & Farzanfar (2016) employed clustering method in the insurance industry to determine the lifetime value of customers. Customer lifetime value is another way of assessing the behaviours of customers by segmenting them and therefore evaluating their profitability for the company in order to improve customer relationship management systems. The customer lifetime value is calculated using metrics like learning cost, customer loss rate, discount rate, costs of maintenance, periodic income, period of time and profit margin. The customers are segmented according to the percentage of customer lifetime value, most precisely a benefit segmentation is carried out. Data used in this research was obtained from the companies' database and comprised of 10 types of insurance companies whose customers have been studied for a period of four years. Based on the customer lifetime value calculations, third party insurance customers had the highest amount of customer lifetime compared to freight insurance which showed the lowest amount. Results show that the customers were segmented into classes, namely 'gold', 'silver', 'bronze' and 'tin', with gold being those with the highest profitability.

Ansari & Riasi (2016) used neural networks and linear regression analysis to identify factors impacting the customer loyalty among start-up insurance companies. The approach was to evaluate the performance of both techniques in analysing customer loyalty. A total of 389 customer data from 10 start-up insurance companies in Iran were collected using survey technique. Three models were developed for the linear regression analysis using target variables perceived value, customer satisfaction and customer loyalty for each. Besides, artificial neural networks were created using MATLAB software and several possible networks were tested to find the best suited one for the data set. Both techniques suggested that perceived value and customer satisfaction were significant predictors of customer loyalty. However, findings obtained from the artificial neural network models had lower error rates compared to the regression models. Hence, it could be concluded that artificial neural networks is best suited for analysing customer loyalty in the insurance market.

2.2 Risk Prediction

Insurance companies highly practice risk management because insuring a policyholder means that the risks are transferred from the customer to the firm. Insurance firms profile and assess the risks of customers to deduce the amount of premium that must be paid (Polyakova, 2016). Traditionally, risk is calculated using actuarial formulas, yet now, with the presence of data analytics more research is focusing on the deduction of new formulas with more accurate factors impacting the risk amount of a customer. Nonetheless, risk prediction based on business risk can also be explained as the risks involved with customer churn, which is mostly the focus of this paper. The literature present on customer churn analysis is considerable. Customer churn analysis is a main concern among numerous industries, for instance the retail industry, telco industry and banking institutions (Almana, Aksoy & Alzahran, 2014; Vafeiadis et al., 2015; Amoo et al., 2015). Some extensively used techniques to evaluate the risks involved in the insurance companies are decision trees, neural networks, Naïve Bayes and regression analysis. The predictive analysis of whether a customer is likely to churn is fundamental for any company so as it can ameliorate its customer relationship management to deter the customers from terminating their relationship with the company (Rodriguez & Shin, 2013).

Günther et al. (2014) performed a study in which they created a statistical model for the prediction of customer churn. They obtained data from Gjensidige, which is Norway's largest non-life insurance company. The data involves health, home and car insurance policy holders. The data is of time series nature and is for a 19 months' period, which dates from November 2007 to May 2009. The data set contains numerous attributes, specifically, yearly premium, age of customer, gender, partner, discount programme, car insurance policy, home insurance policy, number of home insurance policies, health insurance policy and customer lifetime. A random sampling techniques was employed to select 160,000 customers from the database, which eventually decreased to an effective sample size of 127,961 customer data after data cleaning and reduction. Furthermore, a generalised additive model (GAM) approach was used to identify linear relationships between the explanatory attributes and some attributes were log transformed to meet linearity conditions. The model was based on the linear regression model and the analysis identified significant attributes which were used in building the prediction model. The receiver operating characteristic (ROC) curve was utilised to evaluate the prediction performance. Moreover, the regression model was validated using a new data set for the period June 2009 to January 2010 and the prediction performance was tested for each month. Results elucidated that the model was accurate.

Likewise, Goonetilleke & Caldera (2013) conducted a research on customer churn prediction. However, the study was designated to life insurance domain, using data from a life insurance company in Sri Lanka for policies starting from the year 2002 to 2003. The authors used classification techniques, namely decision trees, neural networks and logistic regression method to analyse the data. Initially four class labels were established for the policies, explicitly, 'lapse', 'open', 'paid-up' and 'mature' to determine the status of the policies in the data set. The approach involved selection of attributes with the help of domain experts followed by initial exploration of the data set to find any significant patterns in visualizations like stacked bar charts and by statistical tests like chi-squared testing. The set of attributes dealt with were demographic information, policy details and number of other policies the customer has with the company. After initial data cleaning and pre-processing, 21 variables were selected. Yet, other techniques, namely, correlation-based feature selection subset evaluation, gain ratio and information gain were employed to rank the most importance attributes and eventually only 8 most important were used for the data mining task. A prediction model was built using the Weka package software by applying decision trees, neural networks and logistic regression. The different models were evaluated using the performance criteria, prediction accuracy, ROC curve and AUC value. Besides, cost sensitive learning strategies were employed to the data set to take care of skewness in the class distribution. Upon application of the strategies, findings suggested a 90% accuracy in capturing customer attrition by the prediction model.

Rodpysh (2013) came up with a formula to calculate the customer churn index using insurance data and applied data mining techniques to evaluate the formula. The objective of this investigation was to propose a formula to calculate the optimal customer churn in order to be able to assess risky situations. Data used in this study originated from a third-party insurance database in Guilan province for the period July to September 2011. The proposed churn formula comprised of six variables, namely, satisfaction from competing companies, overall satisfaction of company's products, probability of discontinuation of services in the future, likelihood of using competing company's products, likelihood of using the company's products and service and company recommendations. It was explained that customers with a calculated index of 0.5 and above have a high probability of churn. The formulation was compared to five other indices of churn, namely, 'may terminate the contract and service', 'extension of contract and services', 'advised the company to others', 'satisfaction of rival companies' and 'overall satisfaction of the company' using six different classification methods to determine customer churn and the accuracy of each model was assessed. The six classification models employed were QUEST decision tree, C5.0 decision tree, CHAID decision tree, CART decision tree, Bayesian Network and Neural Network. The models were evaluated using the methods overall accuracy, profit function, ROC curve and Lift Index. Findings suggested a better accuracy in predicting customer churn using the proposed formula instead of individual indicators.

Kašćelan, Kašćelan & Novović Burić (2016) used a nonparametric data mining approach for risk prediction in the car insurance industry. Data used was collected from third party motor insurance company in Sava Montenegro. The data analysed was for the period 2009 to 2011 and consisted of 35,521 policies data with attributes such as region, age, sex, type of vehicle, number of claims per policy, years of policy ownership and mean claim size. The research aimed to demonstrate the effectiveness of using data mining techniques to predict risk instead of the standard techniques used for risk assessment. K-means clustering method was applied to form clusters of similar data points and 12 clusters were obtained. Support vector regression (SVR) was performed on the clusters with average claim cost being the target variable. SVR was used to estimate claim size and the performance evaluation statistics showed that the technique had relative error of less than 10% among most clusters. Logistic regression was applied using number of claims as the target variable, to estimate the probability of a claim occurring. Logistic regression model had an accuracy of around 80%. The proposed approach revealed that data mining techniques are fairly accurate when predicting risk in the car insurance industry, even with small data sets. However, the techniques were not tested against other data mining algorithms which could have provided better results.

Similarly, Rose (2016) used a nonparametric approach to analyse insurance data. However, the framework proposed was machine learning algorithms. The aim of the research was to conduct machine learning for determining a simple predicting formula for plan payment risk adjustment and evaluate the likelihood of improving risk adjustment. Data originated from Truven MarketScan database from which a random sample of 250,000 data was chosen and was for the period 2011-2012. The data consisted of enrolment and claim data from private health plans and the attributes used for the research involved age of client, gender, region, inpatient diagnosis categories and 74 created Hierarchical Condition Category (HCC). Analysis involved methods like linear regression, penalised regression, artificial neural networks, decision trees, random forests. To assess the performance of the algorithms on a data set a cross-validation method was applied, which refers to assigning measures of performance to each algorithm to validate whether it is useful when applied to novel data. The optimal choice of prediction function was denoted as the super learner and has been developed by the super learning algorithm, which is the incorporation of important components from multiple algorithms to make up a single combined algorithm. Findings suggested that the super learner model outperformed the other algorithms studied.

2.3 Fraud Detection

Currently, insurance fraud and abuse is a growing concern among the insurance industries. Fraud refers to intentional misrepresentation of certain information to make false claims. Insurance fraud is detrimental to the insurance firms as they incur great losses when false claims are made. The fraudulent cases can be found among any type of insurance, yet it shows more occurrence in the healthcare insurance sector since physicians have been found falsifying medical reports for policyholders. Statistics reveal that the United States of America incurs over 30 billion dollars every year due to health insurance frauds (Rawte and Anuradha, 2015). Data analytics is providing great support in insurance fraud detection, since recently companies are facing an upsurge of transactional data that is being collected daily (Bănărescu, 2015). Over the past years, fraud detection methods have received much interest among researchers (Sithic & Balasubramanian, 2013; Muhammad, 2014; Li e al., 2016; Wang et al., 2017). Many algorithms have been applied to provide models where fraud and abuse cases can be spotted efficiently as an attempt to improve business processes and reduce unnecessary losses.

Joudaki et al. (2016) identified potential fraud and abuse from health insurance data set by using data mining approach. Data was collected from the Social Security Organisation (SSO) from Iran and consisted of general physicians' drug prescription claims data for 612,804 outpatient drug prescription claims in year 2011. The aim was to detect potential suspects among the physicians who would be involved in fraudulent activities. Indicators of fraud and abuse were developed whereby 13 indicators were found significant to be able to classify the suspects. A hierarchical clustering method was used to identify the segment of physicians who were suspects for fraudulent practices. Discriminant analysis was carried out to verify the reliability of the clusters formed. It was found that the indicators employed to perform the clustering showed satisfactory performance as when tested on a new sample data set, fraud suspects were detected with 98% accuracy and abuse suspects with 85% accuracy. The research demonstrated the effectiveness of applying data analytical approach to health insurance data and this concept can be useful to streamline auditing practices towards suspect groups instead of all physicians.

Another research on fraud and abuse detection in the health care insurance is undertaken by Kose, Gokturk & Kilic (2015) who developed an interactive system called eFAD suite to detect fraudulent activities. The framework was designed to be independent of actors who involve individuals like insured or physicians and commodities, which are the medications that the insurance company pays for. The framework is based on interactive machine learning approach which means that the system is flexible to changes and allows experts to interact with the data. Transactional data was used to identify suspicious fraudulent and abuse cases. To achieve a proactive analysis, which refers to the online detection of fraud and abuse, and enhance performance, a two-stage data warehousing method was employed. The authors used pairwise comparison method for weighting the actors and commodities and clustering of similar actors was done by expectation maximization method. The risk associated with claims and actors were determined using the z-score of the attributes in the study. Finally, a dashboard was created for visualization purpose. The visualization tool gives analysis results based on the input attributes and risk factors. The model accuracy when tested gave an overall accuracy statistic between 70.8% to 89.6%.

Thornton et al. (2014) used a different methodology to analyse fraudulent activities from healthcare insurance data, which is the outlier detection method. Outlier detection method is an unsupervised anomaly detection method commonly used to detect frauds. The analysis was performed using the dental providers case study which comprised of data in a state Medicaid program from the United States from July 2012 to May 2013, amounting to 650,000 dental claims. Attributes that were involved are claim data, healthcare provider and patient data and were analysed using an iterative outlier detection technique. Fourteen specific fraud metrics in the dental domain, were identified based on literature, discussions with domain experts and analysed cases from the U.S Federal Bureau of Investigation. The analysis techniques used included multivariate analysis, time series analysis and box-plot analysis and the outlier detection methods applied were regression analysis, K-means cluster analysis, trend deviations and peak deviations. Oracle SQL Procedures and R language scripts were utilised to conduct the analysis. The findings suggested that 12 out of 17 (71%) suspected dental providers were subjected to formal investigation by officials for committing fraudulent practices.

Similarly, Nian et al. (2016) employed an unsupervised anomaly detection method to identify fraudulent activities in the auto insurance industry. Nonetheless, they proposed a new technique which is the ranking method for anomaly detection based on spectral analysis. The proposed method involved a ranking scheme which identified the top ranked case as being the most suspicious. The method detects anomaly in the dependence among attributes in a dataset by using similarity kernels. Spectral analysis was employed to generate anomaly ranking and assist visualizations. For this analysis, to assess the performance of the proposed method, an open access auto insurance claim data set from Angoss Knowledge Seeker Software was used with 15,420 claims data from January 1994 to December 1996. The data set comprised of over 30 attributes, both ordinal and categorical. Together with the proposed method of spectral ranking analysis, one-class support vector machine (OCSVM), local outlier factor (LOF) and supervised random forest (RF) techniques were tested on the data set to detect fraudulent cases. OCSVM was obtained from the LibSVM Library while LOF was from Ddtools Library and RF from MATLAB software. The ROC curves and AUC were used as performance assessment criteria to compare the accuracy of the models. Results suggested that the unsupervised spectral ranking analysis method provided more precision in determining fraud cases compared to the other fraud detection methods.

Another research related to automobile insurance fraud detection is done by Hargreaves & Singhania (2016). They used the same data set as the previous mentioned paper, which is from the Angoss Knowledge Seeker software comprising of 31 attributes in regard to auto insurance claim data. However, their approach in detecting fraud was different. Hypothesis testing was conducted on the attributes in the data set to identify the significant variables for determining fraud cases. The Chi square test and independent sample t-test were employed to determine any correlation between the categorical attributes and to compare means of the continuous variables respectively. 20 out of 31 variables were found significant and were used to profile fraudulent and non-fraudulent claims. A set of characteristics was identified for the fraud group based on the significant variable identification. Furthermore, a total of 20 business rules were derived to aid the identification of future fraudulent claims. Four rules were derived based on demographic characteristics of a fraud profile, ten rules were identified to test the claim characteristics of a fraud profile, another four rules were based on the vehicle characteristics and finally four rules were based on the policy type to identify fraudulent claims.

Goleiji & Tarokh (2015) used two methods to select significant factors pertaining to automobile insurance fraudulent practices. The feature selection methods used to choose the influential attributes are genetic algorithm method and correlation method. The genetic algorithm feature selection method uses learning algorithm and is similar to Naïve Bayesian method. The selection of feature based on correlation is the Pearson test which evaluates the strength of association between attributes. The data set used in this study was obtained from an automobile insurance industry comprising of 27 attributes. According to the genetic algorithm, 14 variables were crucial while for the correlation method, 13 variables were found important. Decision tree algorithm and Naïve Bayesian techniques were applied to both set of attributes and fraudulency and non-fraudulency was considered as the target variable. Results suggested that among both feature selection methods and data mining techniques, decision tree models applied on the set of attributes chosen by the genetic algorithm demonstrated the highest accuracy (93.89%).

3. Discussion

3.1 Most Popular Approach

Based on the researches reviewed in the previous section, it can be found that for customer level analytics, the most popular approaches were clustering and classification to separate the customers into various groups in order to understand their behaviour. The reason why their behaviour is studied is because the companies want to know their customers better to offer them with the right products in order to maximize their satisfaction and to ensure their retention. For risk prediction purposes, the most common approach was again classification, specifically decision trees and neural networks. Such algorithms provide a better insight on the prediction of customer churn. It is crucial for insurance firms to identify reasons and likelihood of their customer attrition rates because it can help them strive in the competitive market. As for fraud detection, the literature suggests that researchers are coming up with novel algorithms based on the existing traditional methods like clustering, classification and linear to get more accurate prediction models to detect fraudulent cases.

3.2 Popular Evaluation Criteria

Most of the studies applied performance evaluation criteria to measure the accuracy of their model. Out of the literature reviewed in this paper, root mean square error (RMSE), mean square error (MSE), receiver operating characteristic (ROC) curve and area under curve (AUC) were the most applied statistical measurements for testing the accuracy of the models used.

3.3 Limitations

A few researchers did not elaborate on their source and attributes of data set. This does not allow the analysis to be reproduced with other data mining techniques using the same data set. However, the reason why the data source is not revealed is due to privacy and data protection policies of the insurance firms. The companies deal with sensitive customer data and are restricted by laws to not disclose any personal information. Certain studies also faced hindrance in their analysis because they were not given the right to access the required data or the data provided to them were masked for privacy reasons. Moreover, few papers failed to evaluate their models based on statistical proofs and performance evaluation criteria, which makes their research not robust enough to extract valuable information. Table 1 gives an overview of the different literature reviewed in this paper and their main approaches to data analytics in the insurance sector.

Table 1: Summary of approaches

Area	Research	Insurance Industry	Approach	Algorithms	Evaluation Criteria
Customer Analytics	Golmah (2014)	Automobile	Classification	Self-Organizing Maps (SOMs)	Mean absolute error (MAE), root mean square error (RMSE)
	Jandaghi & Moradpour (2015)	Life	Clustering	Fuzzy clustering	No statistical evaluation
	Roodpishi & Nashtaei (2015)	Automobile	Market basket analysis	Causal extraction algorithms, k-means clustering	Sum square error (SSE)
	Fang, Jiang & Song (2016)	Healthcare	Regression	Random forests regression	Mean square error (MSE)
	Delafrooz & Farzanfar (2016)	Life and non-life	Segmentation	Benefit clustering	No statistical evaluation
	Ansari & Riasi (2016)	Not disclosed	Classification, regression	Neural network, linear	MSE, RMSE

				regression, ANOVA testing	
Risk prediction	Goonetilleke & Caldera (2013)	Life	Classification	Decision tree, neural network, logistic regression	Receiver operating characteristic (ROC) curve, area under curve (AUC)
	Rodpysh (2013)	Third party	Classification	Decision tree, Bayesian network, neural network	Profit function, ROC curve, lift index
	Günther et al. (2014)	Automobile, home, healthcare	Regression	Generalised additive models (GAM)	Receiver operating characteristic (ROC) curve
	Kašćelan, Kašćelan & Novović Burić (2016)	Automobile	Clustering, Regression	K-means clustering, support vector regression (SVR), Kernel logistic regression (KLR)	Absolute error (AE), RMSE, relative error (RE), normalised absolute error (NAE)
	Rose (2016)	Healthcare	Non-parametric machine learning	Artificial neural network, decision tree, random forest	Cross-valided R-square
Fraud detection	Thornton et al. (2014)	Healthcare	Outlier detection	Multivariate analysis, time series, box plot analysis, regression, k-means clustering, trend and peak deviations	Standard deviation
	Kose, Gokturk & Kilic (2015)	Healthcare	Machine learning	Proactive and retrospective analysis, clustering	AUC
	Goleiji & Tarokh (2015)	Automobile	Genetic algorithm method, correlation	Decision tree, Naïve bayesian	Total accuracy percentage

Joudaki et al., (2016)	Healthcare	Classification	Hierarchical clustering	Discriminant analysis
Nian et al. (2016)	Automobile	Spectral ranking analysis	One-class support vector machine (SVM), local outlier factor (LOF), supervised random forest	ROC curve, AUC
Hargreaves & Singhania (2016)	Automobile	Statistical hypothesis testing	Chi-squared test, independent sample t-test	p-value significance

4. Conclusions

Data analytics is being implemented worldwide across various industries. Data mining and machine learning techniques have immense potential in offering businesses with a competitive edge over their adversaries. Extensive literature exists on numerous domains where different analytical techniques were applied. The objective of this paper was to conduct a review of the literature on the analytical techniques employed in the insurance industry. Three major areas were discussed namely customer analytics, risk prediction and fraud detection. Finally, a critical evaluation was made based on the literature studied and a summary was provided on the different techniques implemented by the researchers.

References

Almana, A., Aksoy, M. & Alzahran, R. (2014) A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry. *International Journal of Engineering Research and Applications*. 4(5). p.165-171.

Amoo, A., Akinyemi, B., Awoyelu, I. & Adagunodo, E. (2015) Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industr. *Journal of Emerging Trends in Computing and Information Sciences*. 6(11). p.630-641.

Ansari, A. & Riasi, A. (2016) Modelling and evaluating customer loyalty using neural networks: Evidence from startup insurance companies. *Future Business Journal*. 2(1). p.15-30.

Bănărescu, A. (2015) Detecting and Preventing Fraud with Data Analytics. *Procedia Economics and Finance.* 32(1). p.1827-1836.

Delafrooz, N. & Farzanfar, E. (2016) Determining the Customer Lifetime Value based on the Benefit Clustering in the Insurance Industry. *Indian Journal of Science and Technology*. 9(1).

Desik, P., Samarendra, B., Soma, P. & Sundari, N. (2016) Segmentation-Based Predictive Modeling Approach in Insurance Marketing Strategy. *IUP Journal of Business Strategy.* 13(2).

Fang, K., Jiang, Y. & Song, M. (2016) Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. *Computers & Industrial Engineering.* 101. p.554-564.

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 Journal of Multimedia Processing.* 4(3). p.1-5.

Golmah, V. (2014) A Case Study of Applying SOM in Market Segmentation of Automobile Insurance Customers. *International Journal of Database Theory and Application.* 7(1). p.25-36.

Goonetilleke, T. & Caldera, H. (2013) Mining Life Insurance Data for Customer Attrition Analysis. *Journal of Industrial and Intelligent Information.* 1(1). p.52-58.

Günther, C., Tvete, I., Aas, K., Sandnes, G. & Borgan, Ø. (2014) Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal.* 2014(1). p.58-71.

Hanafizadeh, P. & Paydar, N. (2013) A Data Mining Model for Risk Assessment and Customer Segmentation in the Insurance Industry. *International Journal of Strategic Decision Science.* 4(1). p.52-78.

Hargreaves, C. & Singhania, V. (2016) Analytics for Insurance Fraud Detection: An Empirical Study. *American Journal of Mobile Systems, Applications and Services.* 1(3). p.227-232.

Jandaghi, G. & Moradpour, Z. (2015) Segmentation of Life Insurance Customers Based on their Profile Using Fuzzy Clustering. *International Letters of Social and Humanistic Sciences.* 61. p.17-24.

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. *International Journal of Health Policy and Management.* 5(3). p.165-172.

Kašćelan, V., Kašćelan, L. & Novović Burić, M. (2016) A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market. *Economic Research-Ekonomska Istraživanja.* 29(1). p.545-558.

Kose, I., Gokturk, M. & Kilic, K. (2015) An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing.* 36. p.283-299.

Li, Y., Yan, C., Liu, W. & Li, M. (2016) *Research and Application of Random Forest Model in Mining Automobile Insurance Fraud*. In Proceedings of the 12th International

Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. IEEE.

Mohammadi, V., Albadvi, A. & Teymorpur, B. (2014) Predicting Customer Churn Using CLV in Insurance Industry. *Shiraz Journal of System Management.* 2(1). p.39-49.

Muhammad, S. (2014) Fraud: The Affinity of Classification Techniques to Insurance Fraud Detection. *International Journal of Innovative Technology and Exploring Engineering*. 3(11). p.62-66.

Nian, K., Zhang, H., Tayal, A., Coleman, T. & Li, Y. (2016) Auto Insurance Fraud Detection using Unsupervised Spectral Ranking for Anomaly. *The Journal of Finance and Data Science.* 2(1). p.58-75.

Polyakova, M. (2016) Risk Selection and Heterogeneous Preferences in Health Insurance Markets with a Public Option. *Journal of Health Economics*. 49. p.153-168.

Rahman, M., Arefin, K., Masud, S. & Sultana, S. (eds.) (2017) Analyzing Life Insurance Data with Different Classification Techniques for Customers' Behavior Analysis. In *Advanced Topics in Intelligent Information and Database Systems.* 1st ed. Springer International Publishing, p.15-25.

Ravasan, A. & Mansouri, T. (2015) A Fuzzy ANP Based Weighted RFM Model for Customer Segmentation in Auto Insurance Sector. *International Journal of Information Systems in the Service Sector.* 7(2). p.71-86.

Rawte, V. & Anuradha, G. (2015) *Fraud Detection in Health Insurance using Data Mining Techniques*. In International Conference on Communication, Information & Computing Technology. IEEE.

Robson, J. (2015) General Insurance Marketing: A Review and Future Research Agenda. *Journal of Financial Services Marketing.* 20(4). p.282-291.

Rodpysh, K. (2013) Providing a Method for Determining the Index of Customer Churn in Industry. *International Journal of Information Technology, Control and Automation.* 3(1). p.61-70.

Rodriguez, S. & Shin, H. (2013) Developing Customer Churn Models for Customer Relationship Management. *International Journal of Business Continuity and Risk Management.* 4(4). p.302.

Roodpishi, M. & Nashtaei, R. (2015) Market Basket Analysis in Insurance Industry. *Management Science Letters.* 5(4). p.393-400.

Rose, S. (2016) A Machine Learning Framework for Plan Payment Risk Adjustment. *Health Services Research*. 51(6). p.2358-2374.

Sithic, H. & Balasubramanian, T. (2013) Survey of Insurance Fraud Detection Using Data Mining Technique. *International Journal of Innovative Technology and Exploring Engineering.* 2(3). p.62-65.

Sivarajah, U., Kamal, M., Irani, Z. & Weerakkody, V. (2017) Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research.* 70(C). p.263-286.

Thornton, D., van Capelleveen, G., Poel, M., van Hillegersberg, J. & Mueller, R. (2014) *Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data.* In Proceedings of the 16th International Conference on Enterprise Information Systems. Scitepress. p.684-694.

Umamaheswari, K. & Janakiraman, D. (2014) Role of Data mining in Insurance Industry. *An International Journal of Advanced Computer Technology.* 3(6). p.961-966.

Vafeiadis, T., Diamantaras, K., Sarigiannidis, G. & Chatzisavvas, K. (2015) A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Simulation Modelling Practice and Theory.* 55. p.1-9.

Wang, S., Pai, H., Wu, M., Wu, F. & Li, C. (2017) The Evaluation of Trustworthiness to Identify Health Insurance Fraud in Dentistry. *Artificial Intelligence in Medicine.* 75. p.40-50.

Zhou, L. & Chen, Q. (2014) Customer Segmentation, Return and Risk Management: An Empirical Analysis Based on BP Neural Network. *Journal of Chemical and Pharmaceutical Research*. 6(6). p.698-703.