National College of Ireland

A Comparative Study of Ensemble Techniques and Individual Classifiers in Predicting Insurance Claim

MSc Research Project
FinTech

Ifeoma Njoh-Paul
Student ID: X18199721

School of Computing
National College of Ireland

Supervisor: Mr Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ifeoma Njoh-Paul
Student ID:	X18199721
Programme:	Msc FinTech **Year:** 2019/2020
Module:	Research Project
Supervisor:	Mr Victor Del Rosal
Submission Due Date:	17 August 2020
Project Title:	A Comparative Study of Ensemble Techniques and Individual Classifiers in Predicting Insurance Claim
Word Count:	6,378 **Page Count** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Ifeoma Njoh-Paul
Date:	17 August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comparative Study of Ensemble Techniques and Individual Classifiers in Predicting Insurance Claim

Ifeoma Marian Njoh-Paul
X18199721

Abstract:
The insurance industry has grown rapidly and is significantly playing an important role in the economy of a country. However, one lingering issue faced by insurers is being able to correctly predict if a policyholder will lay a claim so as to determine a fair price to be charged for purchasing an insurance policy. The goal of this research is to make a comparison between individual classifiers and ensemble techniques to determine which provides the best predictive results. The Knowledge Discovery in Database (KDD) process was adopted to gain insight and business knowledge from the dataset. Four individual classifiers, Support Vector Machine, Linear Discriminate Analysis, Logistic regression and Artificial Neural Network along with two ensemble techniques, Extreme Gradient boosting and stacking were used, the research discovered that the ensemble techniques used gave a better predictive result than all the selected individual classifiers. XGBoost had an accuracy of 96% while stacking algorithm had 76%. The performance metrics chosen for this research was accuracy, sensitivity and AUC.
Keywords: Insurance claim, Ensemble, Prediction, Stacking, Support Vector Machine, Linear Discriminate Analysis, Extreme Gradient Boosting, Logistic regression, Artificial Neural Network.

1 Introduction

1.1Background of the Study

The insurance industry has grown rapidly and is significantly playing an important role in the economy of every country. There are many insurance companies with their different focal points, this includes vehicle insurance, life insurance, travel insurance etc. Despite these various types, one similar problem between them is “insurance claims”. Insurance claims are requests made by the policyholder to the insurance company in order to collect compensation for a loss suffered (Baesens, et al., 2016), these claims can be very costly and are significant to insurance companies. To be cost effective and maximize profitability, it is important for companies to identify the risk factors involved. Risk in this case is defined as the probability that a certain event will occur in the future thereby causing harm or loss (Mustika, Murfi and Widyaningsih, 2019).

Insurance companies have a massive database that stores customer’s personal information, policy and claims information. This historical data is useful for analysts to gain insights on customer’s behavior, identify risk factors and build models for prediction purposes (Weerasinghe and Wijegunasekara, 2016). Analyzing risks can be a challenging task for insurance companies because risks differ from customer to customer. Traditionally, the process of filing an insurance claim for a policyholder is done manually and this makes the process very strenuous, Furthermore, many insurers assumed a normal distribution payments for insurance claim but this method was not very flexible and could not be used in rigid

situations (Wüthrich, 2018). Due to the limitations of the previous technique and to promote flexibility and ease, machine learning techniques were adopted to deal with both structured and unstructured information. Machine learning as defined by (Mohammed, Khan and Bashier, 2016) is the use of intelligent software to make business decisions, this software makes use of statistical learning methods and learns from the company data. The inability of human beings to generate insights and make decisions on a huge amount of data gave rise to innovative tools and novel techniques that can be used to make effective decisions (Wüthrich, 2018).

One major advantage of this machine learning technology is its ability to make predictions based on the machine's ability to make predictions from the data that has been inputted in it. For insurance companies, claim prediction is a vital business process, this is because it helps the insurers to construct the right insurance policy for each policy holder at a fair price (Wüthrich, 2018). When a claim is accurately predicted, this benefits both the policy holder and the insurer. Where it predicts that a policy holder will not lay a claim, the policy holder will be charged a low fee for the insurance policy and this will increase accessibility, cause customer retention and increased customer base for the insurer.

1.2 Statement of the Problem

The insurance process is a contractual agreement between two individuals, the insurer and the insured which forms the basis by which premiums are received by the insurers to serve as compensation for a loss suffered due to an uncertain event (Mustika, Murfi and Widyaningsih, 2019). Many individuals have had to face the burden of paying the unnecessarily high fees to purchase an insurance policy. For example, a good and experienced driver paying the same high insurance fee as a new and inexperienced driver. Being able to build a predictive model that can accurately classify if an individual will make a claim or not and correctly charge customers a fee that equally represents their risk is a problem that insurers have been trying to solve. Many insurers have developed various machine learning algorithm that can help correctly predict if a customer will present a claim in the future thereby reducing the claim cost (Burri, Burri, Bojja and Buruga, 2019). Through proactive management with the help of this technology, insurers can gain insights that will help to significantly reduce their cost.

1.3 Research Objective

The main objective of this research is to compare different individual classifiers and ensemble techniques to determine which model gives the best predictive accuracy. This involves examining which model best predicts the positive and negative class using a pre-processed insurance dataset. The motivation for this comparative study using ensemble models is due to its recent use in the industry and to confirm existing literatures view that ensemble techniques give a better predictive result.

1.4 Research Question

Predicting insurance claim is an important business process for insurance companies, there are many existing research in this area but this study focused on answering the question below;

- To what extent does ensemble techniques provide a significant improvement in predicting insurance claims when compared with other individual traditional models?

1.5 Outline of the Paper

This study is outlined as follows, in section 2, an extensive literature review was conducted to gain insight on the findings of past researchers, section 3 gives a step by step explanation of the methodology, section 4 present the design specification and explains all selected models for the analysis, section 5 shows the graphical outputs that describes the results from the implementation of the methodology, section 6 gives a comprehensive analysis and evaluation of all model results, discussion and interpretation of the results, analysis of and implications of findings and then the final section,7, gives an overview of the study including discussing the conclusions and proposing suggestions for future work

2 Related Work

Due to the relevance of this topic, many researchers have studied and come up with various solutions to help curb the lingering issues in the insurance sector. This section discusses the methods used by other researchers, reviews their machine learning techniques and identifies loopholes. Protection claims are one of the significant components in the administrations aspect of an insurance company, the seriousness of the case alludes to the measure of assets to be spent on fixing the damages. For insurance industries using the machine learning technology has the most important advantage of data set facility. Every type of data whether it is structured, unstructured or semi structured can be modified using machine learning. The use of machine learning is dependent around the worth chain, through cutting edge prescient accuracy, risk related, cases, and client conduct.

To measure the insurance claim, a lot of factors effect on this, so we need maximum information to create a technique. In this way a fitting technique is required to deal with this issue, one of the machine learning strategies can be actualized using random forest. The research by (Baesens, et al., 2016) applied the vehicle protection model random forest to gauge the whole of this insurance claim prediction. Likewise, an investigation is performed of the impact of the quantity of highlights utilized on model precision. The result shows that in instances of estimation of insurance claim the random forest model can be applied. Outfit strategies are learning calculations that make a lot of classifiers and afterward recognize new information focuses by taking a weighted vote of their forecasts. The underlying troupe

strategy is Bayesian assert maturing yet later calculations contain blunders in the remedy of execution coding Bagging and boosting (Guelman, et al., 2012).

(Weerasinghe and Wijegunasekara,. 2016) performed a comparative study of three machine learning algorithms to predict insurance claims, the researcher used multinomial logistic regression, neural network and decision tree and divided the data into train, validation and testing set, the result of the analysis shows that Neural network gives a better accuracy by 61.7% compared with decision tree 57.05% and logistic regression by 52.39%. (Wüthrich, 2018) performed a simpler analysis using regression trees to give a deep understanding of the data selected while (Chen and Guestrin, 2016) tried to predict insurance claims using XGboost model and compared the result with neural network and Adaboost. XGboost still had the highest result using the normalized gini of the algorithms (Chen and Guestrin, 2016). (Wagh and Kamalja, 2017), predicted vehicle insurance claim frequency, insurance claim frequency is the number of times a claim has occurred in a specific period. The dataset used had 7,483 rows and was gotten from an insurance company in Singapore, the researchers used eight (8) regression models to make a comparison on which has a good fit for the prediction measured by the AIC,BIC, log-likelihood. A non-parametric approach was proposed by (Baudry and Robert, 2019) revealed that Extra trees algorithm used gives a small standard deviation figure and provides an almost unbiased estimators. Despite the various machine learning techniques used by the researchers, it was observed that many of them did not perform an in-depth pre-processing analysis before the modeling process. A thorough preparation of the data is usually important to improve the data quality and produce a more reliable result. According to (Bedia, Tauler and Jaumot, 2018) one of the major the benefits of data pre-processing is that it helps to reduce or completely eliminate small data that contributes to experimental errors and noise.

The utilization of choice trees has become an inexorably well-known option prescient strategy for building characterization and relapse models, due for its numerous potential benefits. Different advantages of univariate decision tree models are the result of the various multivariate response and their extension which might include ranking variables of essential explanatory, distribution free feature and high predictive accuracy (Balasubramanian, M.V, 2019).

Hybrid approach was used by (Sundara Kumar, 2018) when working on the problem of dealing with data imbalance, the research used Reverse Nearest Neighborhood and One Class support vector machine (OCSVM) to rectify the problem. Several tests which had more than ten folds of cross validation was done for fraud detection database as well as the database of credit card churn prediction. Different validation methods include SVM, LR, GMDH, DT, MLP, PNN etc. Using decision tree and support vector machine 90.74% and 91.89% of high sensitivity value was achieved. However, for credit card churn prediction data base DT gave 91.2%, SVM gave 87.7% while the GMDH gave 83.2% accuracy.

A systematic approach which is used in creating the models for inputting the data sets and their classification id know as a classifier or classification technique. The learning algorithm

generated model should be able to fit the input data well and should be able to predict the records and class labels that were unknown before. It is therefore, important key objective in learning algorithms to generate the models which are good and have capability to generalize the class labels. The precision or blunder rate figured from the test set can likewise be utilized to think about the overall execution of various classifiers on a similar space. Two types of models are used for the prediction of insurance claims which includes ensemble modeling and individual classifiers (Lessmann, 2015).

For the credit scoring applications ensemble models have been used many times, as they are more stable and accurate in predicting the results and liabilities (Lessmann, 2015). The variance and bias are also known to be reduced by using ensemble models (Tsai C-F, 2016). The ensemble models include the databases like gradient boosting and random forest e.t.c while the single classifiers include logistic regression and neutral network.
When two or more classifiers of different kinds are used, they are known as the ensemble methods. This will enable us to have more accuracy as a set of classifiers are being used to predict the accuracy and giving high performance instead of a single classifier being used. For example, if there are 10 classifiers having 10 percent of accuracy then using all classifiers will give more accurate results and will predict more liabilities then having a single classifier giving us 90% of the accuracy. So, for that different classifiers of different competence regions are being used to gain full strength than a single one.

Ensemble classifier pool various base model forecasts. Much exact and hypothetical proof has proposed that the blend of models improves prescient exactness (Finlay, 2016).The stowing calculation, for instance, gets autonomous base models from bootstrap tests from the first information (Breiman, 2016). Consequently, boosting calculations lets a troupe extend in a reliant manner. Iteratively, they include base models which are prepared to stay away from the current troupe's blunders (Freund and Schapire, 2016).

Investigation of the impact of the quantity of highlights utilized on model precision is directed. The result shows that the Random Forest model can be applied in instances of expectation of insurance claim, Just by utilizing 1/3 of the general highlights, the precision of the Random Forest model can create exactness that is equivalent to that acquired when utilizing all highlights which is around 99%. This outcome affirms the adaptability of Random Forest, particularly as far as the quantity of highlights. Henceforth, the Random Forest model can be utilized as an answer for big data issues identified with information volume (Breiman, 2016).

Predicting the frequency of motor insurance claims lies at the core of premium calculation, but with the advent of modern artificial intelligence approaches, the issue of selecting an acceptable model has yet to be completely answered. (Chen and Guestrin., 2016) compared two different methods logistic regression and XGBoost which are used for the predictive performance of insure information in the form of telematics and drivers.

(Maria Fernanda, et al., 2016) explored the conduct of various oversampling strategies through various classifiers and assessment measurements. The strategies are Random oversampling, SOMO and SMOTE. A genuine information from a Colombian insurance agency was utilized in the examination in foreseeing fake cases for its obligatory auto item. They finished up from the examination and plainly exhibited the benefits of utilizing oversampling for imbalanced conditions yet in addition the significance of contrasting diverse assessment measurements and classifiers with acquire precise suitable ends and practically identical outcomes. This outcome was useful for choosing the technique to use for imbalanced dataset.

Due to the literature review conducted, it is evident that for successful prediction, a massive amount of data is required from an insurance company's database. However this will mean that the computation time will be very high. Many researchers have used several modelling techniques to determine which is best for the prediction process but one thing that was common with many of the reviewed papers is that vital pre-processing steps were not conducted before modelling. The result of this review discovered that there is no existing research on comparing various ensemble techniques with individual classifiers to determine which performs best for insurance claim prediction.

3 Research methodology

This research will follow the Knowledge discovery in database process (KDD), it involves gaining insight on large data extracted from a company's database for the purpose of making business decisions. This method was selected over cross-industry standard process for data mining (CRISP-DM) and sample, explore, modify, model, access (SEMMA) due to its completeness and accuracy (Palacios, Toledo, Pantoja and Navarro, 2017). The KDD process involves several stages and this is explained in details below;

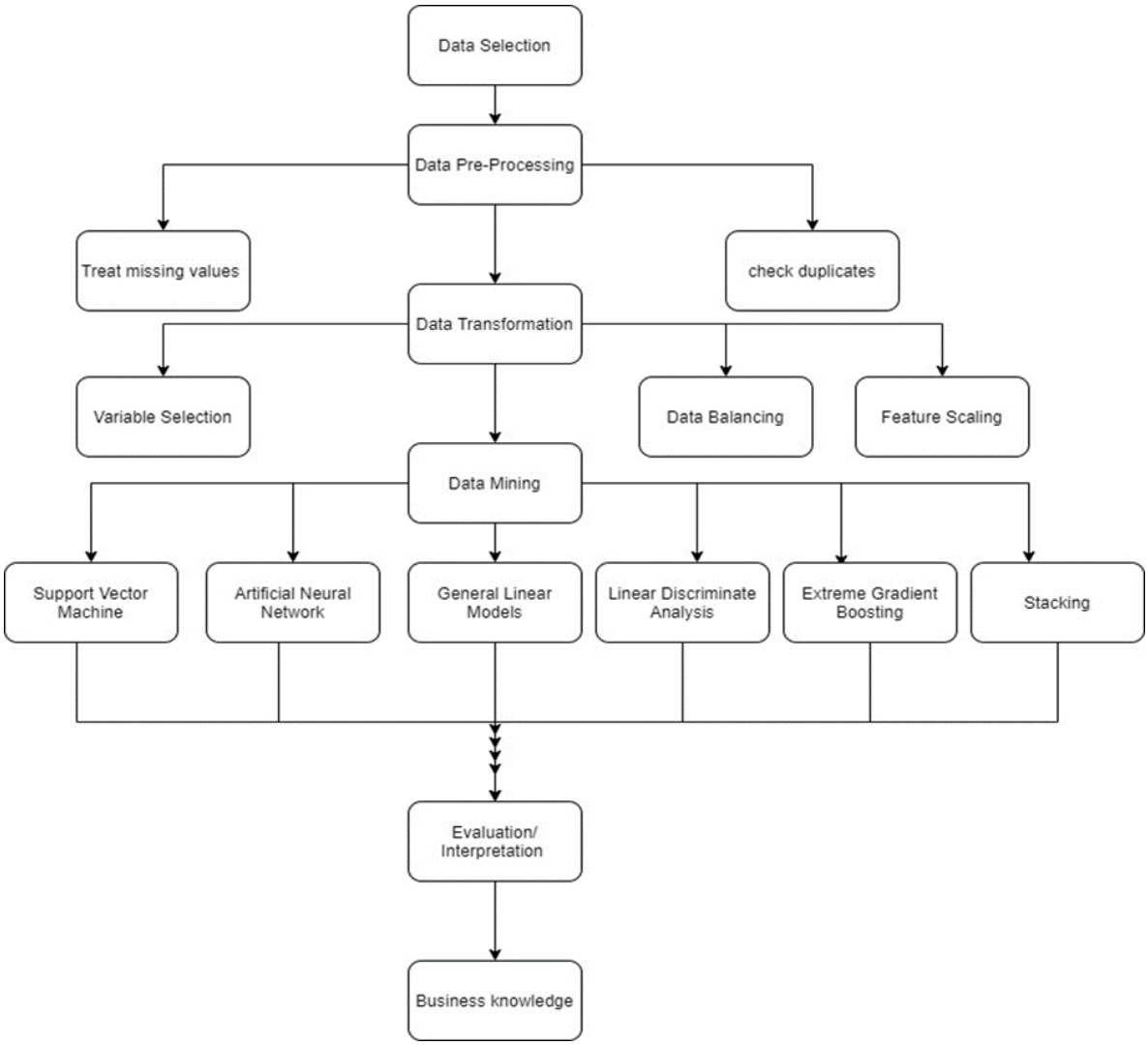


Figure 1: The Knowledge Discovery Database process

3.1. Data Selection

This is the first stage of the KDD process, the data selected for this research is a real world anonymized insurance dataset from a large insurance company in Brazil downloaded from Kaggle website. It is a public dataset used for an online competition. The dataset consists of 595,212 observations and 59 variables that represents the individual/driver, type of vehicle, region and previously calculated values. The dataset consists of both binary and categorical features, the target variable has two labels, "0" represents "no claim" and 1 represents "there is claim".

3.2. Data Pre-processing

The next stage is cleaning the data, this is important to prevent errors by removing noisy and irrelevant data in other to increase the reliability of the data. This process involves treating for missing values by completely removing them or using an imputation technique such as Multivariate Imputation by Chained Equation (MICE), mode for categorical variables and

mean for continuous variables, the research also checked for duplicates in the dataset. The dataset for this research contains missing values, these missing values are represented with -1 in the dataset, and this will be re-classified to NA's and then imputed using mean and mode for continuous and categorical variables respectively. The MICE package was tested on this dataset but due to the size of the dataset, it took too much computation time.

3.3. Data Transformation

This process is important to prepare the data for the mining process. At this stage, it is required to use techniques such as high dimensionality reduction, feature engineering and variable selection. However, this research will use three techniques;

3.3.1. Variable Selection

Variable selection is an important step when building a predictive model, this is because real world data usually contain many variables/features that are irrelevant for model building. Some of the advantages are (Dietterich, 2016);

- It helps to reduce the computation time
- It increases the performance of the model built
- It helps to prevent overfitting

This process can be conducted using statistical techniques such as correlation, backward elimination, forward selection, Chi-square test, and stepwise selection and other the use of machine learning to identify important variables, this research will fit a Gradient Boosting Model (GBM) using the caret package and then obtain the important variables using 'VarImp' function for the predictive model. The Boruta technique is another important method, it is an ensemble technique that runs without tuning of parameters and as a result produces numerical approximations of each variable's importance to the model (Kursa, Jankowski and Rudnicki, 2010).

3.3.2. Feature Scaling

Feature scaling is another important step that helps to improve the quality of the data, it treats datasets with different scale, units or range. In many real world datasets, there are variables with different scales/range, this will cause variables with higher weight to have a higher magnitude and cause a bias in the performance of the selected machine learning algorithm. Therefore, this step ensures that each variable gives equal contribution and increases the quality of the data for better prediction. (Singh and Singh, 2019) performed a comparative analysis to check the impact on feature scaling on the performance of a classification dataset and discovered that un-normalized data performed poorly compared with when normalized using pareto scaling. Feature scaling can be done using mainly various methods; normalization or standardization. This study will normalize the dataset to have equal contribution to the target variable and prevent bias or over representation. The formula for normalization is expressed below;

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.3.3. Data Balancing

When building a predictive model, it is important to make sure that the dataset is balanced. A dataset is described as imbalanced where the one variable class is overrepresented and the other is under represented (Gu, Cai, Zhu and Huang, 2016). If not treated, an imbalanced dataset may cause overfitting of the model and also traditional algorithms will not have a good performance (Gu, Cai, Zhu and Huang, 2016). The dataset for this research is highly imbalanced and this can be seen in figure 2 below. The various techniques to treat imbalance problems are; undersampling, oversampling, Synthetic Minority Oversampling Technique (SMOTE), Self-Organizing Map Oversampling (SOMO) technique. The goal of this process is to minimize skewed distribution by eliminating instances from the majority class or bring in a synthetic individuals to the minority class. The hybrid technique in caret package will be used for this research, it is an sampling technique that helps to produce new class of data by interpolating the many minority class (Gu, Cai, Zhu and Huang, 2016). According to researchers, it has proven to be a very successful technique and has birthed other approaches to treat class imbalance (Fernandez, Garcia, Herrera and Chawla, 2018). Data balancing will be performed on the training dataset before the modeling process begins.

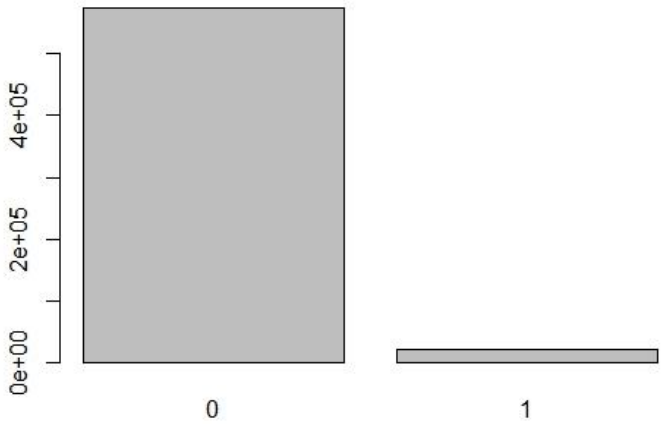


Figure 2: Visualization of the imbalanced dataset

3.4. Data Mining

Having cleaned and prepared the data for the mining process, the next stage is choosing strategies based on the objective of the research. Data mining involves using algorithms to gain insight on a large set of data—checking for pattern and relationships. There are many

data mining techniques for both classification and regression problems, this research has chosen to use five techniques to perform a comparative study between ensemble algorithm techniques and individual classifiers. The individual classifiers chosen are Support Vector Machine, Artificial Neural Network and Logistic Regression, Linear Discriminate Analysis while the ensemble techniques are Extreme Gradient Boosting and Stacking. This will be explained in details below.

3.4.1 Data Splitting

Before the modeling process the data will be split into 70:30 for training and testing set. The train data was then balanced to avoid bias. The train data was used for building the models and then the test data was used to predict.

4 Design Specification

To properly implement all the selected models and make comparison, it is important to have an in-depth understanding of each models. The models selected for this study can be referred to as classifiers or classification models. The study deals with supervised learning because the target variable is categorical having two labels.

4.1 Support Vector Machine (SVM)

This is a type of supervised machine learning algorithm that is useful for solving both classification and regression problems. It is used for recognizing subtle patterns in a large and complicated dataset, it also has a solid theoretical background based on risk minimization (Pavlidis, Wapinski and Noble, 2016). One advantage it has over other types of algorithm is that it works well with high dimensionality data, this is because the technique does not depend on the whole training data but a subset of it called the support vectors (Ruping, 2011). Due to the different types of datasets available, the concept of kernel function was introduced, they are;

- Sigmoid kernel: Can be used in place of Neural Network
- Polynomial Kernel: Mostly used for image processing
- Gaussian kernel: Where prior knowledge about the data is not available

4.2 Logistic Regression (LR)

This is a widely used machine learning algorithm on a dataset that has one or more predictor variable which is responsible for the occurrence of a particular outcome. The proposed outcome/result is given by target variable in the dataset. The target variable is usually categorical represented by 0 which usually means and 1 which means yes. This technique is usually used to solve classification problems, it is based on the probability concept and it uses the sigmoid function to model the dataset. A sigmoid function helps to map the real figures in

the dataset and assigns a new value approximately between 0 and 1 but it's usually never exactly 0 or 1. There are various assumptions of this model according to (Park, 2013);

- There is linearity between logit and predictor variable
- There is no multicollinearity among the predictors
- There is no extreme values among the predictors

The logistic function is written as; $P(Y = 1 \mid X = xi) = \frac{1}{1+\exp -(\beta 0 + \sum p\, i=1\ \beta i x i)}$

Where; Y is the target variable showing if there is a claim or no claim

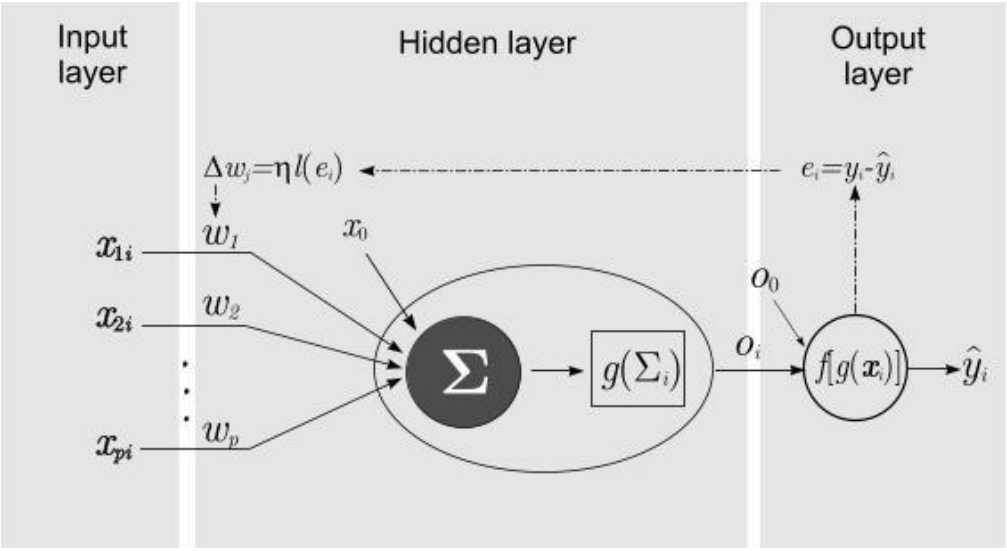
Xi are the random variables to predict the occurrence of a future event

β0 is a constant value

β*i* are coefficients.

4.3 Artificial Neural Network

This is a computational algorithm inspired by the central nervous system of animals, it a type of machine learning model that also identifies patterns in a large dataset. A neural network contains many connected nodes called the neurons (Sharma, 2017). There are two major ways by which a neuron can be activated, one is through weighted connections from neurons that were previously active and another is through sensors that exist in the environment. Synapses are what connects the neurons and allow it to pass signals. The algorithm works like the human brain, it comprises of a large amount of connected processing units working together to help process information and generate meaningful result. This technique can be used to solve both regression and classification problems and does not require a one hot encoding like other traditional machine learning algorithms. Before this algorithm can run successfully, the data has to be normalized. This study already normalized the data in the preprocessing stage. A pictorial visualization can be seen below;



Source: (Maroco et al., 2011)

4.4 Linear Discriminate Analysis (LDA)

Amongst the many models for solving classification problems, LDA which is also referred to as Fisher Discriminate Analysis is one that handles class infrequencies, where each class is unequal (Prince and Elder, 2007). This technique has been used by many researchers for mobile robotics, facial recognition and also for data mining (Pang, Ozawa and Kasabov, 2016), it can be used to identify which variable discriminates between two or more classes and then build a classification model which gives a prediction. (Pohar, Turk and Blas, 2016)

4.5 Ensemble techniques

According to researchers, ensemble techniques are more stable and has a better predictability power than individual classifiers, it helps to reduce model bias and variance (Tsai C-F et.al., 2011). This technique works as a learning algorithm to build a set of classifiers and then derive new data points by taking predictions based on their weighted vote. The foundation of this technique started with the Bayesian averaging but new methods have developed overtime, this includes bagging, boosting, stacking etc. This research will consider stacking method.

4.5.1 Extreme Gradient Boosting

This is a type of ensemble learning algorithm and is the short name for Extreme Gradient Boosting, it can be used to treat classification, regression, ranking and prediction problems. It is a relatively new decision tree boosting technique and works well on large dataset for prediction and handling missing values unlike random forest and NN that do not work well with missing values. This model which was introduced by (Chen and Guestrin, 2016), they described that the model first introduces weak learners and later increases the performance of the trees by ensembling which helps to minimize the regularized objective function.

4.5.2 Stacking

This is a popular ensemble model that helps to improve accuracy by using a combination of individual classifiers. This method has been not been used as much compared to boosting and bagging methods due to its complexity and difficulty level. Stacking works by learning from a single classifier and then combines the result with the predictions of other classifiers used in the model (Wang, Hao, Ma and Jiang, 2011).
The visualization of the modelling process can be seen in the figure 3 below;

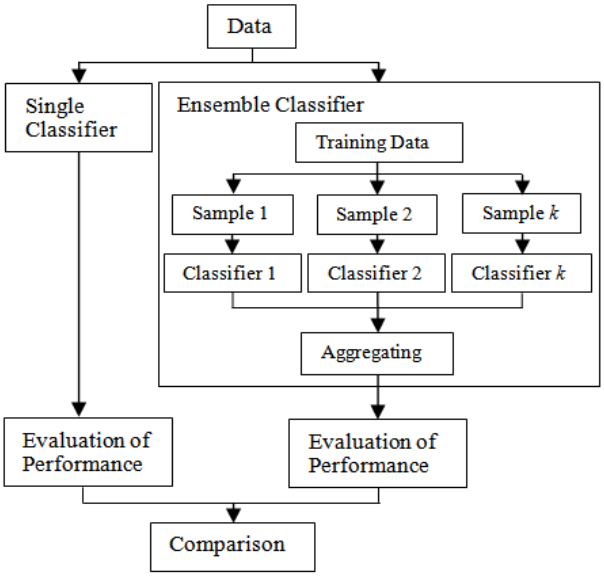


Figure 3: The single classifier and ensemble comparison

4.6 Evaluation

After building the selected models, the next step is to evaluate its performance and check how well the model fits and make a decision on whether to implement it or not. This is where performance measures come in. This study will use confusion matrix to check the metrics of the models and then ROC/AUC curve.

4.6.1 Confusion Matrix

To measure the performance of the predictive models, this study used the confusion matrix to check the relevant metrics. The accuracy and sensitivity will be used to check the performance of the model. Accuracy measures the total number of predictions in the test set that are correct, True positives and True negatives shows the proportion of values that are correctly predicted while False positives and False negatives are the number of positive and negative instances that were incorrectly classified (Andjelkovic Cirkovic, 2020). Accuracy is usually the most popular measure but in the case of an imbalanced data, it is geared towards the majority class and can be bias. But in this study, the data is balanced and the accuracy should give a fair representation. Sensitivity is another metric that will help to check the correctness of predicting individuals that will file for a claim. The formula for accuracy and sensitivity can be seen below;

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

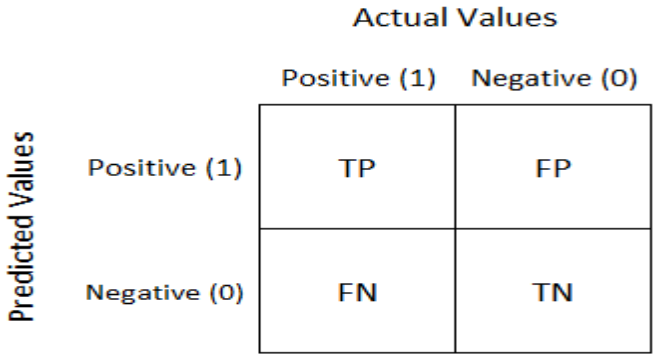


Figure 4: The confusion matrix

4.6.2 ROC Curve

This is known as Receiving Operating Characteristics Curve, this is a graphical representation that shows the performance of the classification model built. This curve is mostly used for binary classification problems. This graph is obtained by plotting the false positive figures on the x-axis and true positive figures on the Y-axis. The AUC is also shown in the graph to measure the performance of the classification model. An AUC figure should be >=0.5. Any figure below 0.5 means that the model is not predicting correctly.

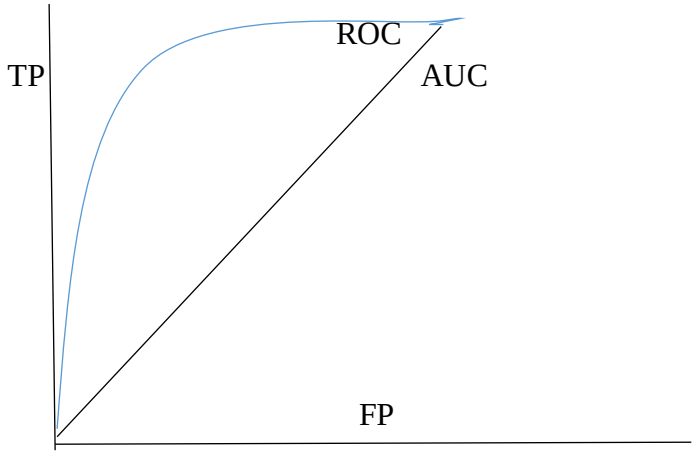


Figure 5: The ROC & AUC plot

5 Implementation

This section will show the graphical results from implementing the methodology process.

5.1 Data Balancing (Hybrid method)

The caret package was used to balance the dataset, ensuring that the two levels in the target class are equally represented. Using the hybrid approach instead of undersampling ensures that relevant data is not lost. This method made use of the caret function “Ovunsample” using

(method= “both”). This ensures that the underrepresented class is oversampled and the overrepresented class is undersampled. The result of this can be seen in Fig 6 below;

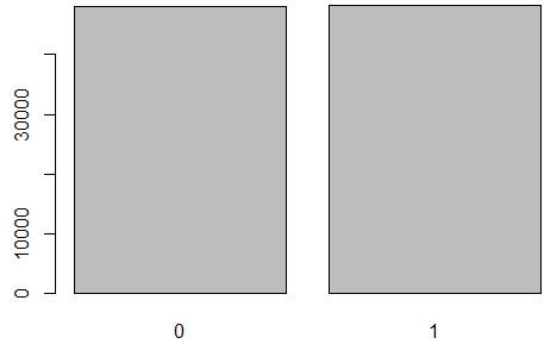


Figure 6: The balanced training dataset

5.2 Variable Selection through Gradient Boosting

The VarImp() function was used after building a model on the full dataset to identify which variables are most important and highly correlated with the target variable. This method is a boosting method of selection that uses the technique as a single tree but it adds up each important variables over the boosting iteration. The result of this algorithm was compared with other statistical methods and discovered similar results. It is important to have a good understanding of the industry that is being researched on other to take appropriate decisions. The result of the machine learning technique can be seen in fig 7 below.

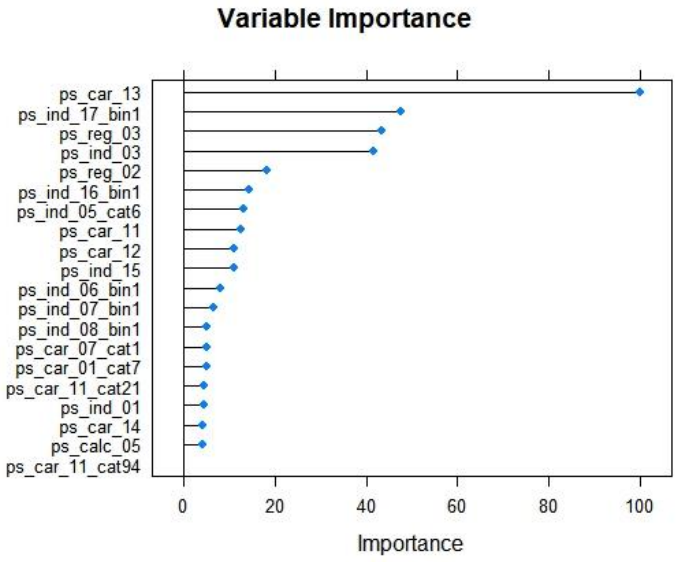


Figure 7: Variable importance through GBM

5.3 Artificial Neural Network

The diagram in Fig 8 shows the calculated neural network, the model built has a hidden layer made up of just one neuron. The black web (lines) shows the connections of various weights while the blue line shows the bias term.

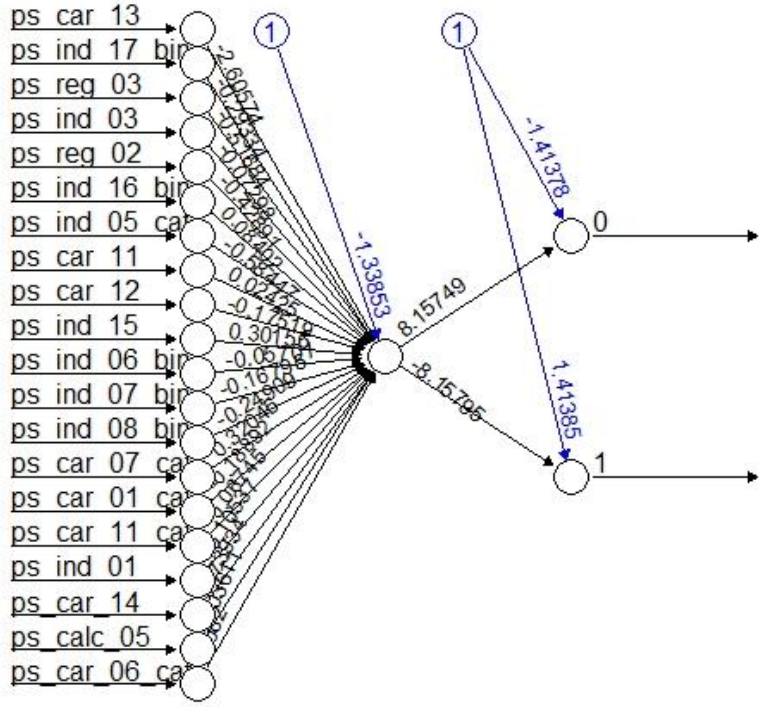


Figure 8: Neural Network

6 Evaluation

This section presents the results of the selected models and give a comprehensive analysis of each results

6.1 Data Mining

A comparison graph of the ROC was plotted to show the results of the four individual classifiers, they are all on the same range having approximately 57% as its AUC. While the Ensemble technique, stacking was significantly higher than all individual classifiers having 76% as its AUC. The comparison of all the performance measures shows that the ensemble techniques has a highest occur This can be seen in Fig 9,10 and 11 below;

6.2 Support Vector Machine

This was the first model used for the prediction exercise. The test dataset was used for the prediction and confusion matrix was computed as seen in table 1 below. The table shows the two possible classes which is 0 for “no claim” and “1” for there is claim. The model predicted 22,180 correctly out of 29,999 observations in the test set, the accuracy was 74%, sensitivity

was 75% and specificity which is the ability to predict the negative class (no claim) is 33% and AUC was 54%.

Table 1: confusion matrix for SVM

Predicted Class	Actual Results	
	0	1
0	21827	729
1	7090	353

6.3 Generalized Linear Model

The second model, also using the testing set for prediction was able to predict the 18,859 correctly therefore having an accuracy of 64% as seen in table 2. The SVM performed significantly better than the GLM model. The sensitivity was 63%, specificity 51% and 57% for AUC.

Table 2: Confusion matrix for GLM

Predicted Class	Actual Results	
	0	1
0	18303	526
1	10614	556

6.4 Artificial Neural Network

The third model, artificial neural network was able to correctly predict 17,986 correctly out of 29,999 observations as seen in table 3. The model had an accuracy of 60%. The model identifies 17402 of negative class (no claim) making the specificity 5% while the sensitivity was at 97%.

Table 3: Confusion matrix for ANN

Predicted Class	Actual Results	
	0	1
0	17402	11515
1	498	584

6.5 Linear Discriminate Analysis

The fourth model tested, LDA was able to classify 18,934 correctly therefore having an accuracy of 63% as seen in table 4, the model identifies 18,380 as the negative class having a specificity as 51% and sensitivity as 64%.

Table 4: Confusion matrix for LDA

Predicted Class	Actual Results	
	0	1
0	18380	528
1	10537	554

6.6 Extreme Gradient Boosting

The fifth model used which is an ensemble model had the best accuracy of 96% after correctly predicting the classes by 28916 out of 29,999 as seen in table 5 below. The sensitivity is 1% and the AUC is 52%.

Table 5: Confusion matrix for Xgboost

Predicted Class	Actual Results	
	0	1
0	28916	1082

1	**1**	**0**

6.7 Stacking

Stacking was done by splitting the dataset into three datasets, training, testing and validation in the ratio 70:15:15 respectively. The training data was balanced using a hybrid method in the caret package with the function "ovunsample" and then the four algorithms, NN, SVM, LDA and GLM were stacked together for the modeling process. The validation set was used to predict the results and a new data frame was derived from the combination of predicted results. The stacking model was built using a regression model because according to (Weerasinghe and Wijegunasekara, 2016) regression algorithm performs better than tree models when in building a stack model. The model correctly predicted 10,950 out of 14,550 as seen in table 6 thereby making the accuracy 75%, sensitivity 75% and AUC 76%.

Table 6: Confusion matrix for Stacking

Predicted Class	**Actual Results**	
	0	**1**
0	**5435**	**1797**
1	**1803**	**5515**

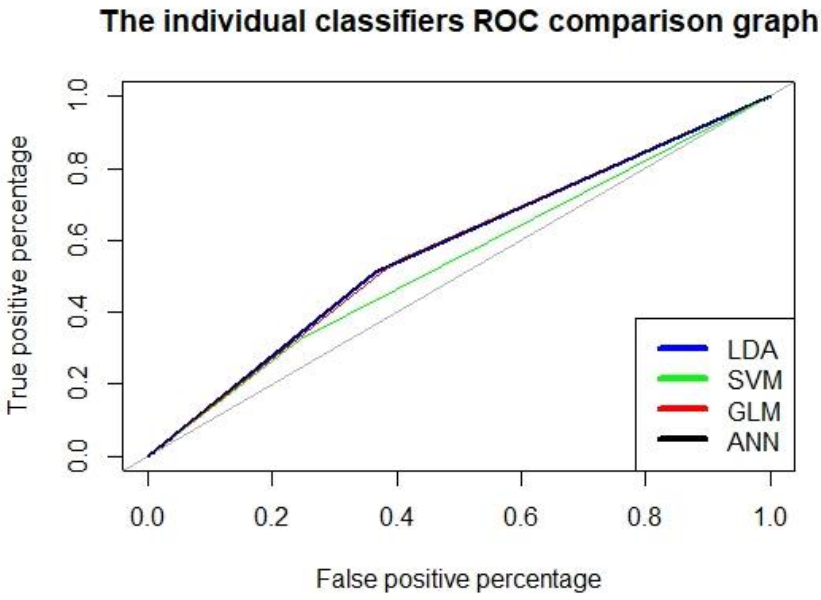


Figure 9: ROC curve comparison graph

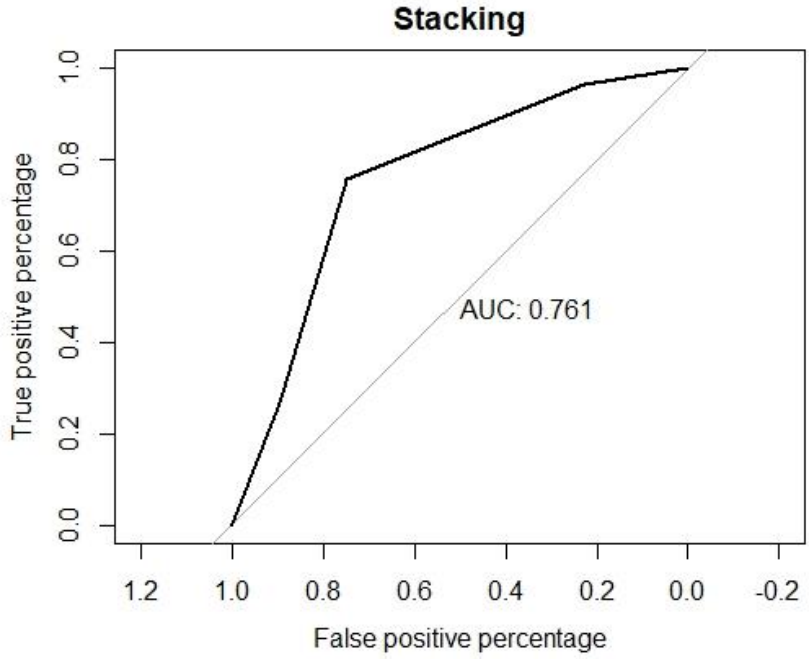


Figure 10: Stacking model AUC graph

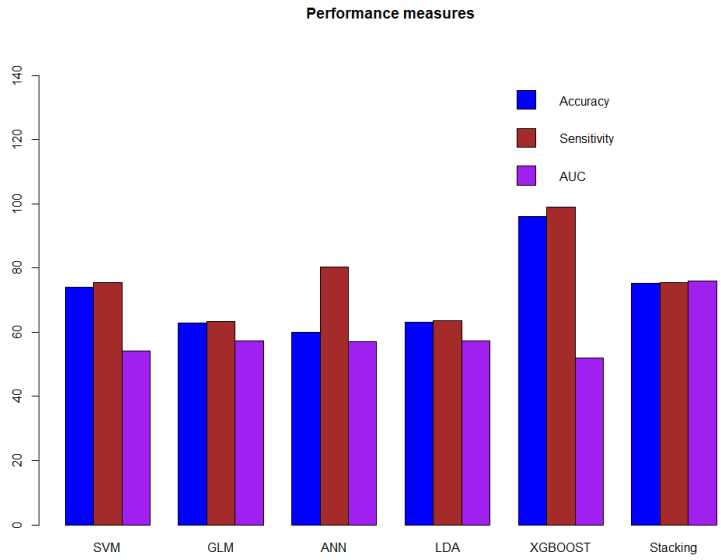


Figure 11: Performance measures of selected models

6.8 Discussion

In predicting insurance claim, the stacking model and Xgboost which are the two ensemble techniques used in this research have proved to have the best results amongst other selected individual models. The Xgboost had the highest accuracy with 96%. This research went a step further and examined the impact of feature selection in the prediction process, it made comparison between including it in the pre-processing stage and eliminating it and discovered that the results was in support of literatures by (Dietterich,2016) and (Tsai, 2016) that feature selection is an important step that helps to improve predictive results. Details of the result can be seen in table 12 below. However, better results might have been obtained for feature selection if Boruta was used, this is because Boruta package works well with datasets that contain over 50 predictors or independent variables (Speiser, Miller, Tooze and Ip, 2019).
After improving the data quality using pre-processing methods, the models selected were able to correctly predict true positives and false positives to a good extent but was best using the stacking model. Overall, this analysis met its objective and proved that the ensemble techniques are better than traditional models in predicting insurance claims. This research will therefore contribute to the existing body of knowledge based on its comparisons and findings.

6.8.1 Limitations of the Research

- Due to the sensitive nature of the insurance industry, getting a real insurance dataset was a challenge. The dataset used for this research was real but it was 100% anonymized and variables were not clearly defined to ensure privacy of information.
- Time constraints was another challenge as this research could not use some preferred models/statistical methods on the full dataset. Furthermore, a larger dataset can be used without random sampling with the aim of getting a higher result.

Table 12: Comparison of performance metrics with and without feature selection

Without feature selection	Accuracy	AUC	Sensitivity/Recall
SVM	65%	58%	66%
ANN	59%	50%	63%
GLM	61%	50%	62%
LDA	62%	57%	63%
Stacking	65%	66%	67%
XGboost	86%	63%	60%
With Feature selection			
SVM	74%	54%	75%
ANN	60%	57%	97%
GLM	62%	57%	63%
LDA	63%	57%	64%
Stacking	76%	75%	75%
XGboost	96%	52%	100%

7 Conclusion and Future Work

Predicting insurance claim is an important process for insurers, it is important for them to determine which policyholder will default or not. However, this process can be quite cumbersome and needs the right techniques and steps to ensure that the data quality is good enough to build an excellent classifier. The higher the accuracy of model built, the better for the insurer to determine a fair cost for the intending policyholder. The objective of this research was fulfilled by comparing two ensemble techniques with four traditional models on a dataset than includes both classification and continuous variables to determine which gives a better predictive accuracy based on the different performance metrics selected. The research question was answered with results proving that ensemble machine learning techniques, XGboost and stacking, gives the best results compared to all other individual models, their accuracy is 96% and 76% respectively. For future research, it will be useful to use a redefined dataset that gives a detailed explanation of each variable for better understanding, specifically when implementing feature engineering and variable selection.

Furthermore, it is important to try other ensemble techniques such as Adaboost to determine if accuracy will be improved. For big insurance firms, an improvement in prediction metrics by as little as 1% can save them a lot of cost and improve efficiency, this study is therefore in favor of ensemble techniques being explored more by insurers.

References:

Andjelkovic Cirkovic, B., (2020). Machine learning approach for breast cancer prognosis prediction. *Computational Modeling in Bioengineering and Bioinformatics*, pp.41-68.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J., (2016). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*(6), pp.627-635.

Balasubramanian, M.V., (2019). *Ensemble modeling & prediction interpretability for insurance fraud claims classification* (Doctoral dissertation, Dublin Business School).

Baudry, M. and Robert, C., (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, 35(5), pp.1127-1155.

Bedia, C., Tauler, R. and Jaumot, J., (2018). Introduction to the Data Analysis Relevance in the Omic Era. *Comprehensive Analytical Chemistry*, pp.1-12.

Breiman, L., (2016). Consistency for a simple model of random forests.

Caropreso, M.F., Matwin, S. and Sebastiani, F., (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text databases and document management: Theory and practice*, *5478*, pp.78-102.

Chen, T. and Guestrin, C., (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,

Dewi, K.C., Murfi, H. and Abdullah, S., (2019), October. Analysis Accuracy of Random Forest Model for Big Data–A Case Study of Claim Severity Prediction in Car Insurance. In *2019 5th International Conference on Science in Information Technology (ICSITech)* (pp. 60-65). IEEE.

Fernandez, A., Garcia, S., Herrera, F. and Chawla, N., (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, pp.863-905.

Finlay, S., (2016). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2), pp.368-378.

Freund, Y., Schapire, R.E., Singer, Y. and Warmuth, M.K., (1997), May. Using and combining predictors that specialize. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing* (pp. 334-343).

Gu, Q., Cai, Z., Zhu, L. and Huang, B., (2016). Data Mining on Imbalanced Data Sets. *2016 International Conference on Advanced Computer Theory and Engineering*,

Guelman, L., Guillén, M. and Pérez-Marín, A.M., (2012), May. Random forests for uplift modeling: an insurance customer retention case. In *International Conference on Modeling and Simulation in Engineering, Economics and Management* (pp. 123-133). Springer, Berlin, Heidelberg.

Kursa, M., Jankowski, A. and Rudnicki, W., (2010). Boruta – A System for Feature Selection. *Fundamenta Informaticae*, 101(4), pp.271-285.

Kumar, M.N., Koushik, K.V.S. and Sundar, K.J., (2018). Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *3*(3), pp.162-167.

Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C., (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), pp.124-136.

Li, Y., Yan, C., Liu, W. and Li, M., (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. *Applied Soft Computing*, *70*, pp.1000-1009.

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I. and de Mendonça, A., (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, 4(1).

Mohammed, M., Khan, M. and Bashier, E., (2016). Machine Learning.

Mustika, W., Murfi, H. and Widyaningsih, Y., (2019). Analysis Accuracy of XGBoost Model for Multiclass Classification - A Case Study of Applicant Level Risk Prediction for Life Insurance. *2019 5th International Conference on Science in Information Technology (ICSITech)*,

Pang, S., Ozawa, S. and Kasabov, N., (2016). Incremental Linear Discriminant Analysis for Classification of Data Streams. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 35(5), pp.905-914.

Palacios, H., Toledo, R., Pantoja, G. and Navarro, Á.,(2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Advances in Science, Technology and Engineering Systems Journal*, 2(3), pp.598-604.

Prince, S. and Elder, J., (2016). Probabilistic Linear Discriminant Analysis for Inferences About Identity. *2007 IEEE 11th International Conference on Computer Vision*,

Rodrigues, L.A. and Omar, N., (2014). Auto claim fraud detection using multi classifier system. *Journal of Computer Science & Information Technology*, *14*.

Ruping, S., (2016). Incremental learning with support vector machines. *Proceedings 2001 IEEE International Conference on Data Mining*,.

Sharma, S., (2017). *Artificial Neural Network (ANN) In Machine Learning*.

Speiser, J., Miller, M., Tooze, J. and Ip, E., (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 134, pp.93-101.

T. G. Dietterich, (2016). Ensemble methods in machine learning. *In Proceedings of the First International Workshop on Multiple Classifier Systems, pages 1–15*, Berlin, 2016. Springer

Tsai, C., (2016). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22(2), pp.120-127. Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), pp.124-136.

Tsai, C.-F. , & Hsiao, Y.-C. (2011). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, 50 ,258–269
Wang, G., Hao, J., Ma, J. and Jiang, H., (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1), pp.223-230.

Weerasinghe, K. and Wijegunasekara, M., (2016). A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims. *European International Journal of Science and Technology*, 5(1), pp.47-54.

Wüthrich, M., (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6), pp.465-480.

Wagh, Y. and Kamalja, K., (2017). Modelling auto insurance claims in Singapore. *Sri Lankan Journal of Applied Statistics*, 18(2), p.105.

Wüthrich, M., (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6), pp.465-480.

Yau, K., Yip, K. and Yuen, H., (2016). Modelling repeated insurance claim frequency data using the generalized linear mixed model. *Journal of Applied Statistics*, 30(8), pp.857-865.