International Conference on Advancements in Computing Technologies - ICACT 2018 ISSN: 2454-4248
Volume: 4 Issue: 2 99 – 103

An Analysis of Customer Retention and Car Insurance Claim using Supervised Learning

Sushmitha.S.P[1], Renuka Devi D[2]

PG Scholar, Department of Computer Science, Stella Maris College, Chennai, India[1]
Assistant Professor, Department of Computer Science, Stella Maris College, Chennai, India[2]

Abstract: Insurance sector by nature has scrupulous collection of data. Large insurance business data are used to predict the potential clients by applying big data analytics techniques. Big Data Analytics is used to discover the prospective client, allowing insurance business to the elevated stage. The key objective of this paper is to analyze and comprehend the purchase plan to uncover the business opportunities. In this paper, we propose different classification algorithms which are applied on large-scale insurance data to improve the performance and predictive modeling. The experiment is carried out on the Kaggle datasets. The performance metrics used are accuracy, precision, recall and F-Measure. The performance efficacy of different classifications algorithms is also compared.

Keywords*: Supervised learning, Big Data, Classification Algorithms, Sampling Technique*.

*****

I. INTRODUCTION

A huge amount of data is generally referred as Big Data. It is enormous in size, diverse variety and has highest velocity of data arrival. This huge information is useless unless the data is examined to uncover the new correlations, customer experiences and other useful information that can facilitate the organization to take more informed business decisions. Big data is widely applied in all sectors like healthcare, insurance, finance and many more.Traditional marketing system of insurance is offline based sales business, where in general, companies vend the insurance policies by calling and visiting the customers. This fixed marketing system also achieved good results in past time. But currently many new private insurances companies also have entered into the marketplace which gives healthier competition. On other hand, eagerness of people to pay for the insurance service is also enlarged. Therefore, understanding the need and purchase plan of clients is exceptionally essential for insurance companies to elevate the sales volume.

Big data technology ropes the insurance companies' transformations. Due to lack of principle and novelty of traditional marketing, badly structured insurance data, unclear customers purchasing characteristics leads to imbalanced data, which brings the intricacy of classification of user and insurance product recommendation.Decision making task is complicated with imbalanced data distribution. To solve this problem, we usually use few resampling methods which will construct the balanced training datasets. This will improve the performance of predictive model.Main purpose of this paper is to identify the potential customer with help of big data technology. This paper outlines the fineapproach for identifying the probable clients. We propose supervised learning algorithmensemble decision tree (Random forest and XGBoosting).This paper is organized as follows. Section II introduces the current research status of machine learning; Section III puts forward the classification model and intelligent recommendation algorithm based on XGBoost algorithm for insurance business data, and analyzes its competence; Section IV outlines the experiment and result.Section V outlines conclusion and future work.

II. RELATED WORK

The classification problem for US bank insurance business data has imbalanced data distribution. This means ratio between positive and negative proportion are extremely unbalanced and the prediction models generated directly by supervised learning algorithms like SVM, logistic regression are biased for large proportion. Example, the ratio between positive and negative classes is 100:1. Therefore, this can be seen as such model does not help in prediction.Imbalanced class distribution will influence the performance of classification problem. Thus, some techniques should be applied to deal this issue. One approach to handle the problem of unbalanced class distribution is sampling techniques [2]. This will rebalance the dataset.

Sampling techniques are broadly classified into two types. They are under sampling and over sampling. Under sampling technique is applied to major class for reduction process (e.g. Random Under Sampling) and over sampling is another technique applied to add missing scores to set of samples of minor class (e.g. Random Over Sampling(ROS)).The drawback of ROS is redundancy in dataset this will again lead to classification problem that is classifier may not recognize the minor class significantly. To overcome this problem, SMOTE (Synthetic Minority Over Sampling) is used. This will create additional sample which are close and similar to nearest neighbors along with samples of the minor class to rebalance the dataset with help of K-Nearest Neighbors (KNN) [2].Sampling method is divided into non-heuristic method and heuristic method. Non-heuristic will randomly remove the samples from majority class in order to reduce the degree of imbalance[10].

Heuristic sampling is another method which will distinguish samples based on nearest neighbor algorithm [7].Another difficulty in classification problem is data quality, which is existence of missing data. Frequent occurrence of missing

data will give biased result. Mostly, dataset attributes are dependent to each other. Thus, identifying the correlation between those attributes can be used to discover the missing data values. One approach to replace the missing values with some probable values is called imputation [6].One of the challenges in big data is data quality. We need to ensure the quality of data otherwise it will deceive to erroneous predictions sometimes. One significant difficultyof data quality is missing data. Imputation is method for handling the missing data. This will reconstruct the missing data with estimated ones. Imputation method has advantage of handling missing data without help of learning algorithms and this will also allow the researcher to select the suitable imputation method for particular circumstance [3].There are many imputation methods are existing for missing value treatment (Some widely used data imputation methods are Case substitution, Mean and Mode imputation, Predictive model). In this paper we propose the predictive model for missing value treatment.

There are a range of machine learning algorithms to crack both classification and regression problems. Machine learning is practice of designing the classification which hascapability to repeatedly learn and perform without being explicitly programmed. Machine learning algorithms are classified into three types (Supervised learning, Unsupervised learning,Reinforcement Learning).In this paper, we propose supervised machine learning algorithms to create the model. Some of the supervised learning algorithms are listed below: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc [8]. Decision tree in machine learning can be used for both classification and regression. In decision examination, a decision tree can be used to visually and unambiguously represent decision.The tree has two significantentities preciselyknown as decision nodes and leaves. The leaves are the verdict or the final result and the decision nodes are wherever the data is split. The classification tree is type of decision tree where the outcome was a variable like ‘fit’ or ‘unfit’. Here the decision variable is categorical.

One of the best ensemble methods is random forest. It is used for both classification and regression [5]. Random Forest is collection of many decision trees; every tree has its full growth. And it has advantage of automatic feature selection [4].Gradient Boosting is to consecutively decreasefault with each consecutive model, until one final model is produced. The keyintend of every machine learning algorithms is to construct the strongest predictive model while accounting for computational effectivenesson top. This is where XGBoosting algorithm engages in recreation. XGBoost (eXtreme Gradient Boosting) is a direct application of Gradient boosting for decision trees. It gives furtherregularizethe model formalization to manage over-fitting, which gives improved performance [8].

III. METHODOLOGY

Classification Model: Traditional sales approach of insurance product is offline process and it has the following disadvantages: (1) lack of customer evaluation system, don't know the characteristics influence weight of the potential customers; (2) the data accumulated in this way usually has serious ruinous, indirect influence the accuracy of classification model [4].For a bunch of classification models, distribution of class and correlation features affects the forecast results. Imbalanced data classification and independent attributes of insurance dataset will have serious deviation in classification model result. We can handle this kind of issues with different sampling method and supervised learning algorithms.

In this paper, we proposed different sampling approach with supervised learning algorithms on car insurance dataset to build the best predictive model. Imbalanced data classification problem is resolved with sampling techniques and then we construct the model with supervised learning algorithms using training dataset. Finally, predictive model is validated with test dataset and performance of algorithms is evaluated using confusion matrix method with test dataset. Precision, Recall and F-measure is used to evaluate efficacy of the algorithms. The proposed model is shown in the Fig.1.

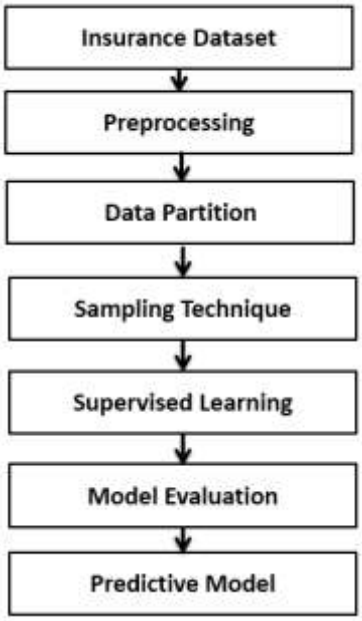

Fig.1.ProposedModel

A. Dataset

The experiment is carried out on the Kaggle datasets: The Home of Data Science & Machine Learning and implemented in R.This dataset is collected from one of the banks in the US. In addition tocommon services, this bank also provides car insurance services. Thisbank arrangespromotions like campaign every year to catch the attention of new customers. The bank has provided details about potential customers’ data, and bankstaff’ call duration time forpromotion available car insurance decision.

B. Preprocessing

Data is usually collected for unspecified applications. Data quality is one of the major issues that are needed to be addressed in process of big data analytics. Problems that

affect the data quality are given in the following: 1.Noise and outliers 2. Missing values 3. Duplicate data. Preprocessing is a method used to make data more appropriate for data analytics. Data Cleaning is a process to handle the misplaced data. We have used analytical model for imputation method to envisage the misplaced values using the non-missing data. Here, we used KNN algorithm to estimate the missing data. This will estimate missing data with help of the recent neighbor values. Data transformation is one of the methods in preprocessing to normalize data.Normalization is a process in which we modify the complex dataset into simpler dataset. Here, we used Min-Max normalization to normalize the data. It will scale the data between the 0 and 1.

$$z = \frac{x - \min(x)}{\min(x) - \max(x)} \quad (1)$$

Where, x is the vector that we going to normalize. Then *min* and *max* are the minimum and maximum values in x given its range. Once the dataset is pre-processed, now it is ready for data partition.

C. Data Partition

In this step, the data is split into separate roles of learn (train) and test rather than just working with all of the data. Training data will contain the input data with expected output. More or less 60% of the original dataset constitutes the training dataset and other 40% is considered as testing dataset. This is the data that validate the core model and checks for accuracy of the model. Here, we partitioned the original insurance dataset into train and test set with probability of 60 and 40 split.

D. Supervised Learning Algorithms

Supervised learning is machine learning technique. This infers function and practice with training data without explicitly programmed. Learning is said to be supervised, when the desired outcome is already known. After partition, next step is to build the model with training sample set. Here, our target variable is chosen first. We selected our target variable as car insurance and other attributes in dataset is taken as predictors to develop the predictive model. Now, to predict,a model to envisagethe potential clients of car insurance service during campaign has to be designed. In this problem, we need to separate out clients who buy car insurance and who were not buy the insurance in the campaign based on extremelyconsiderablekey variables.In this paper, we used random forest and extreme gradient boosting algorithm to visualize the model and the accuracy of the model is evaluated.

Decision Tree

Decision Tree can be used for both classification and regression. Unlike linear models, tree based predictive model gives high accuracy. Decision tree is frequently used in classification problem. It will separate out the clients based on predictor variables and identify the variable, which creates the best uniform sets of clients. In this, our decision variable is categorical.

Random Forest

Random forest is one of the frequently used predictive model and machine learning technique. In a normal decision tree, one decision tree is built to identify the potential client but in case of random forest algorithm, numbers of decision trees are built during the process to identify the potential client. A vote from each of the decision trees is considered in deciding the final class of an object.Sampling is one of the methods in preprocessing. This will select the subset of original samples. This is mainly used in case of balance the data classification. In our model, we have used under sampling approaches to balance the datasampling. It will condense the majority group to make their occurrence closer to the infrequentgroup. Original insurance data is balanced with under sampling. So further we will use this sample in Random Forest Algorithms to build the model. This randomly generates the n number of trees to build the effective model.

Extreme Gradient Boosting

Another classifier is extreme gradient boosting. The XGBoost has an immensely high predictive model. This algorithm works ten percent faster than existing algorithms. It can be used for both regression, classification and also ranking.One of the most interesting factsabout the XGBoost is regularized boosting method. This helps to lessen overfit modeling. Over-fitting is the occurrence in which the learning model tightly fits the given training data so much that it would be inaccurate in predicting the outcome of the test data. In our model, first we used over sampling method to balance the classification. Sampling technique can be used to get better forecast performance in the case of imbalanced classes using R and caret package. Over sampling will randomly duplicate samples from the class with few instances. Here, we used over sampling method with train set to improve the performance of model. Now, balanced samples are collected. We will pass these samples to XGBoost as train set and built the model. XGBoost built the binary classification model with insurance data. After this, model is validated with test set. This produces much better prediction performance compared to random forest algorithm.

E. Model Evaluation

Performance analysis of classification problems includes the matrix analysis of predicted result. In this paper, we have used following precision, recall and F-measure metrics to evaluate the performance of classification algorithms. This is shown in the table 1.Precision is the fraction of predicted occurrence that is related. It is also called positive predicted value. Recall is part of related instances that have been repossessed over the totalquantity of relatedinstance. F1-Measure is the weighted harmonic mean (Number of interpretation, divided by the sum of reciprocals of the interpretation) of the precision and recall and correspond to the overall performance.

$$TPR = Recall = \frac{TP}{P} = \frac{TP}{(TP + FN)} \quad (2)$$

 ISSN: 2454-4248
99 – 103

$$Precision = \frac{TP}{(TP + FP)} \quad (3)$$

$$F - measure = \frac{2 * P * R}{(P + R)} \quad (4)$$

Where, TP – True positive ,FP – False Positive, TN-True Negative, FN-False Negative.

Table 1: Confusion Matrix

		True Condition	
	Total Population	**Condition Positive**	**Condition Negative**
Predicted Condition	**Predicted positive**	**True Positive (TP)**	**False Positive (FP)**
	Predicted negative	**False Negative (FN)**	**True Negative (TN)**

IV. EXPERIMENT AND RESULTS

KNN is used for missing data treatment and after preprocessing, the model is created with XGBoost and random forest for business case. The comparison accuracy of both models isgiven in the table 2.

Table 2: Performance comparison of XGBoost and Random forest algorithm

Algorithm	Precision	Recall	F1	Accuracy
Random Forest	0.81	80	0.76	0.76
XGBoost	0.86	0.86	0.86	0.86

From the Table 2, the result shows that XGBoosting algorithm out performed than random forest. The performances of the two algorithms areshown in the Fig.7.

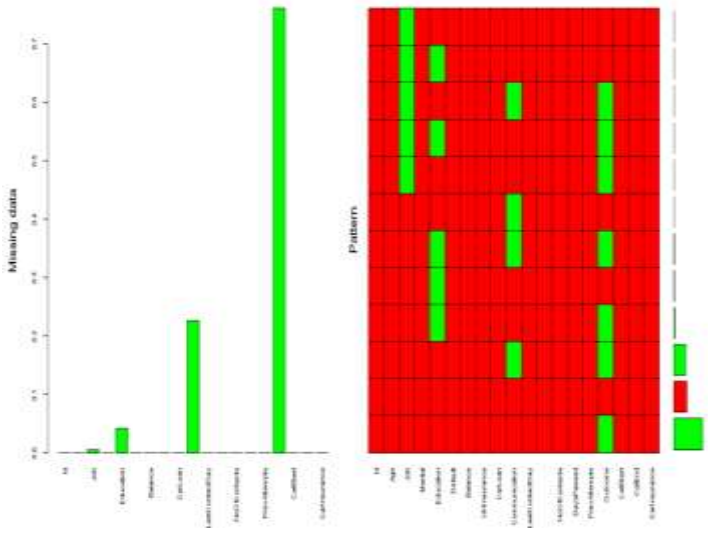

Fig.2. Effect of Missing values before Imputation

The effect of missing value imputation for the proposed algorithm is shown in the Fig.2. The important feature selection is depicted in the Fig.3 and Fig.6.

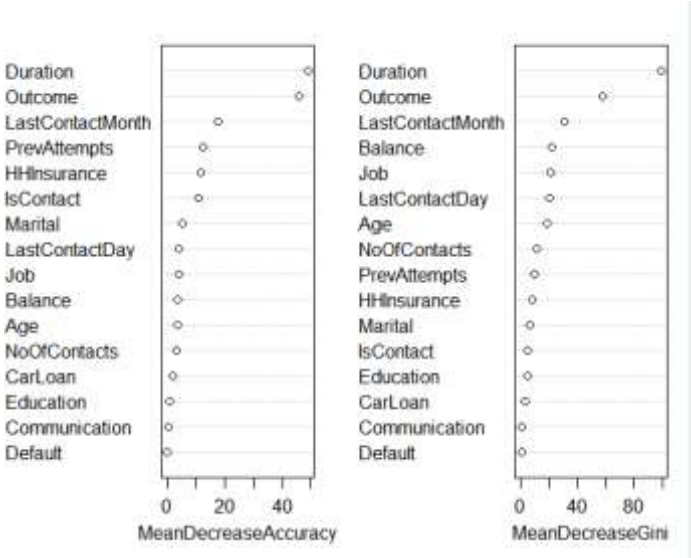

Fig.3. Important Features that impact on target variable using Random Forest Algorithm

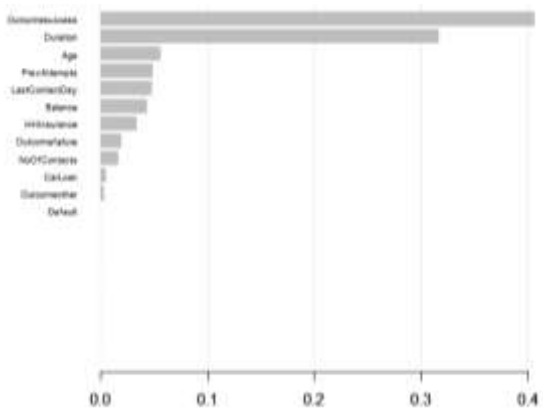

Fig.6. Important Features that impact on target variable using Gradient Boosting Algorithm.

 ISSN: 2454-4248
99 – 103

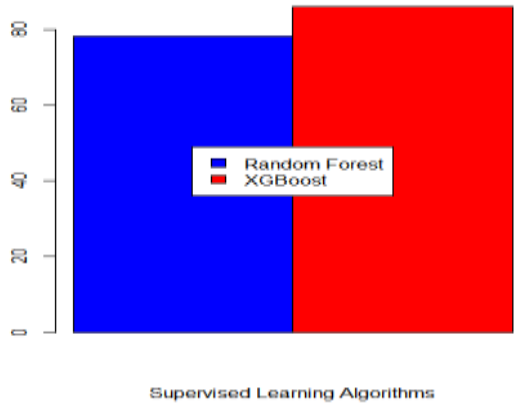

Fig.7. Overall Performance Analysis

CONCLUSION

This paper analyzed the imbalance distribution of insurance business data, and presented the preprocessing algorithms of imbalance dataset. The random forest and XGBoost learning algorithmsare implemented in R. The experiment results showed that the XGBoost algorithm out performed than other decision tree algorithm called Random Forest. Our future works include combining proposed algorithm with deep learning.

REFERENCES

[1] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, ``SMOTE-RSB:A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory,"Knowl. Inf. Syst., vol. 33, no. 2, pp. 245_265, 2012.

[2] Maryam Farajzadeh-Zanjani, RoozbehRazavi-Far, MehrdadSaif," Efficient Sampling Techniques for Ensemble Learning and Diagnosing Bearing Defects under Class Imbalanced Condition".

[3] Gustavo E. A. P. A. Batista and Maria Carolina Monard," An Analysis of Four Missing Data Treatment Methods for Supervised Learning".

[4] Weiwei Lin, Ziming Wu, Longxin Lin, Angzhan Wen, And Jin Li," An Ensemble Random Forest Algorithm for Insurance Big Data Analysis",2017.

[5] EeshaGoel, Er. Abhilasha," Random Forest: A Review",2017.

[6] Conception Of Data Preprocessing And Partitioning Procedure For Machine Learning.Avaliable:http://www.academia.edu/9517738/conception_of_data_preprocessing_and_partitioning_procedure_for_machine_learning_algorithm.

[7] Down-Sampling Using Random Forests, Avaliable:https://www.r-bloggers.com/down-sampling-using-random-forests/

[8] Boosting in Machine Learning and the Implementation of XGBoostAvaliable: https:// towardsdatascience.com/boosting-in-machine-learning-and-the-implementation-of-xgboost-in-python.

[9] Tianqi Chen and Tong He,"xgboost: eXtreme Gradient Boosting", January 4, 2017.

[10] JormaLaurikkala ,"Improving Identification of Difficult Small Classes byBalancing Class Distribution",2001.