Vehicle Insurance Model Training and Evaluation

Madhukar Thummala[1], Bharath Veeramallu[2], VishalReddy Yervala[3]

Asst Prof. Srilatha Doddi[4]

[1,2,3] Student, Dept. of Computer Science Engineering, Sreenidhi Institute of Science and Technology, Telangana, India.

[4] Assistant Professor, Dept. of Computer science Engineering, Sreenidhi Institute of Science and Technology, Telangana, India.

Abstract **– vehicle insurance is insurance for cars, trucks, motorcycles, and other road vehicles. Its primary use is to provide financial protection against physical damage or bodily injury resulting from traffic collisions and against liability that could also arise from incidents in a vehicle an insurance company that has provided health insurance to its customers, now they need in building a model to predict whether the policyholders from past year will also be interested in vehicle insurance provided by company. vehicle insurance model training and prediction can then accordingly plan communication strategy to reach out to those customers and optimize its business model and revenue. In this paper we develop the vehicle insurance model for the health insurance company based on the Synthetic Minority Over-sampling Technique (SMOTE) analysis and classification techniques. The proposed framework aims to minimize the human intervention, the obtained results reveal the high-performance gain achieved by XGBoost in classifying the customers based upon the response.**

Keywords: Data analysis, Synthetic Minority Over-sampling Technique (SMOTE), XGBoost, Linear SVC, Categorical NB, KNN classifier.

1.INTRODUCTION

Health insurance is a type of insurance that covers the whole or a part of the risk of the person incurring medical expenses. It is the coverage that provides for the payments of benefits as a result of sickness or injury. It includes insurance for losses from accident, medical expenses, disability, or accidental death and dismemberment. Our client is a Health Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company. An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee. For example, consider you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for up to Rs. 200,000. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2-3) would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation called sum assured to the customer. Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue. Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. In this paper we discuss about Exploratory Data Analysis (EDA) for analysing and we deal with Synthetic Minority Over-Sampling Technique (SMOTE) imbalanced classification. Imbalanced classification involves in developing the predictive models on classification datasets. The challenge of working with imbalanced datasets most machine learning models will ignore this will turn into a poor performance on the minority class, this SMOTE technique used to increase performance of the minority classes. Classification techniques used are Decision Tree classifier, Random Forest Regressor, Logistic Regression, KNN classifier, XGBclassifier, Gradient Boosting Classifier, Categorical NB, Linear SVC, among above classification Model we identify the best classification Model.

The remainder of this paper is organized as follows. Section II presents the proposed methodology. While Section. In section III implementation. Finally, conclusions are drawn in Section IV.

II. METHODOLOGY

In this section, we present the adopted methodology to process, analyse, the data and to classify using classification techniques as well as our model to classify clients of health insurance company whether interested in vehicle insurance or not. Fig.1: summarizes the proposed methodology.

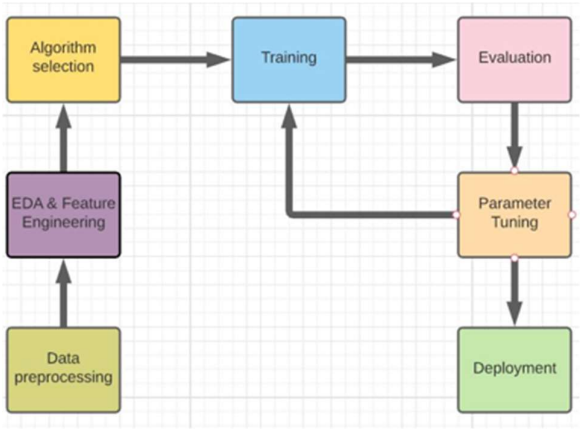


Fig. 1: Workflow of the proposed methodology.

A. Data Preprocessing

Data Preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format, techniques used for Data pre-processing are

Data Cleaning

Data cleaning is the process of detecting, rectifying, or removing inaccurate and corrupted information from the dataset or database. In addition, it recognizes inaccurate or unfinished parts of data, filling the missing ones, and removing the noisy data.

Data Integration

Data integration involves combining data residing in different sources and providing users with a unified view of these all data. Data integration may involve inconsistent data and therefor needs data cleaning.

Data Transformation.

Data transformation is the process of converting data from one from to another from. Data transformation is necessary to ensure that data from one application or database is understandable to other applications and databases.

Data Reduction

Data reduction involves in reducing the number of attributes, attribute values, number of tuples.

B. EDA & Feature Engineering

Exploratory Data analysis is the technique for analyzing datasets to summarize the main characteristics. There are many tools that are useful for EDA. Typical graphical techniques used are: box plot, histogram, multi-vari chart, run chart, parseto chart, odds ratio etc. Interactive versions of these plots are Dimensionality reduction. Univariate analysis, Bi-variate analysis, multi-variate comes under the exploratory data analysis.

Uni variate analysis

Univariate analysis is the analysis of the single variable in this paper we do analysis using Uni-variate analysis.

Bi -variate analysis

Bi- variate analysis is the analysis of the two variables, in this paper we do analysis using bi-variate analysis

Multi-variate analysis

Multi variate analysis is the analysis of the more than three variables, in this paper we do analysis using multi variate analysis

SMOTE (Synthetic Minority oversampling techn-ique)

In order to increase the performance of the minority classes we use this technique by analysing the minority classes. Imbalanced classification tech-nique is used to increase performance.

Feature Engineering is the process of using domain knowledge to extract features from raw data. A feature is a property shared by independent units on which analysis or prediction is to be done.

C. Algorithm Selection

Decision TreeClassifier

Decision tree is a supervised machine learning technique that can be used for classification and regression. Decision nodes are used to make decision and have multiple branches, leaf node of a decision tree is the outcome of the decisions. CART (classification and regression) algorithm is used to make decision tree. In this paper, we import sklearn.tree library to implement Decision TreeClssifier.

Random Forest classifier

Random forests or random decision fore-sts are an ensemble method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of over fitting to their training set. In this paper, we import sklearn.ensemble library to implement Random Forest classifier.

Logistic Regression Classifier:

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function to return a probability value. In this paper, we import sklearn.linear_model library to implement logistic regression.

XGBoost Classifier:

XGBoost is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data. In this paper we import the xgboost.XGBClassifier to implement XGB classifier

KNN Classifier

It assumes the similarity between the new data and available data and place the new data into a category it present. It is a Lazy learner algorithm because it does not learn from training set, instead it stores the dataset, at the time of classification action is performed. In this paper we import sklearn.neighbours library to implement KNN

Categorical Naïve bayes Classifier:

Categorical Naïve bayes is a variant of a native bayes that follows categorical distribution of discrete features. The categories of each feature are drawn from a categorical distribution. In this paper we import sklearn.naive_bayes library to implement CategoricalNB.

Gradient Boosting Classifier.

Gradient boosting classifiers are a group of machine learning algorithms that are combine many weak learning models together to create a strong predictive model using Gradient descent function. In this paper we import sklearn.ensemble library to implement Gradient boosting classifier.

Linear SVC

Linear SVC is to fit the data you provide returning a best fit hyperplane that categorizes the data. After getting a hyperplane you can feed some features to your classifier. In this paper we import sklearn.svm to implement Linear SVC.

Confusion matrix

Confusion matrix is used to describe the performance of the classification model. Confused matrix between actual and predicted values

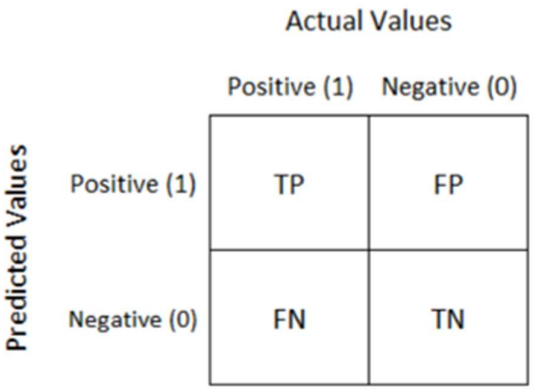


Fig 2. Confusion matrix between actual and predicted values.

III. IMPLEMENTATION

In this paper for model development, we take customer data from health insurance company the features are id, gender, age, driving license, region code, previously insured, vehicle age, vehicle damage, annual premium, policy sales channel, vintage, response.

Univariate Analysis

Univariate analysis of Gender

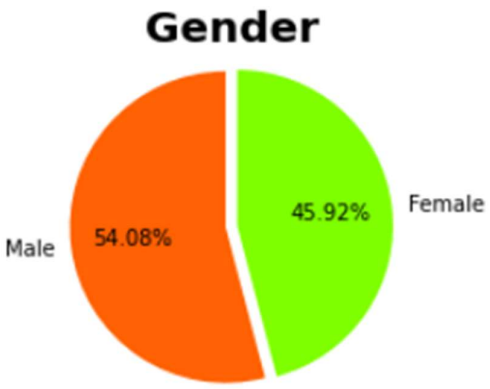


Fig 3. Based upon the gender count, we plot pie chart.

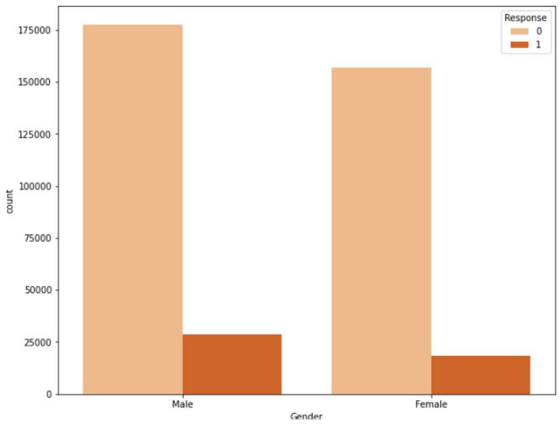


Fig 4. Based upon gender response, we plot count plot.

Univariate analysis of vehicle age

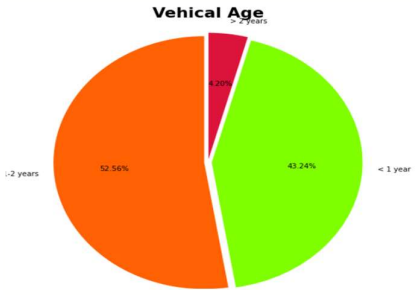


Fig 5. Based upon the vehicle count of 1-2 year, > 2 years, < 2 years, we plot pie chart.

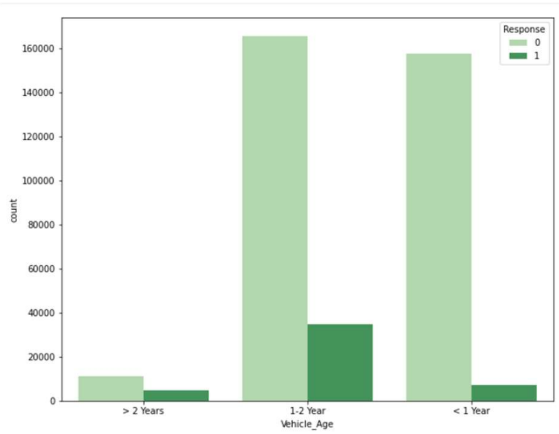


Fig 6. Based upon the vehicle age >2 years, 1-2 years, < 1 year, we plot count plot.

Univariate analysis of vehicle damage

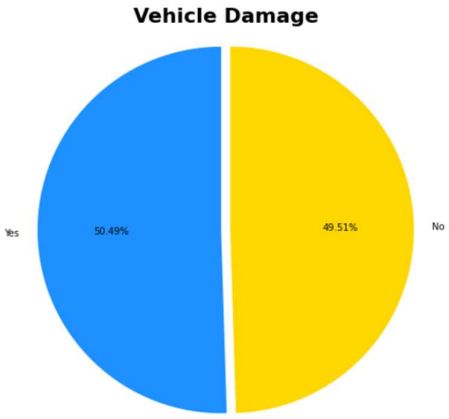


Fig 7. Based upon the vehicle damage count, we plot chart.

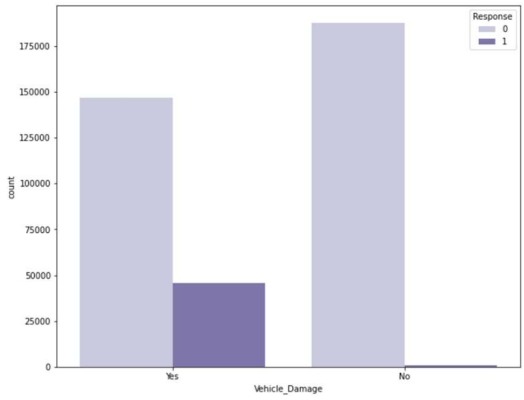
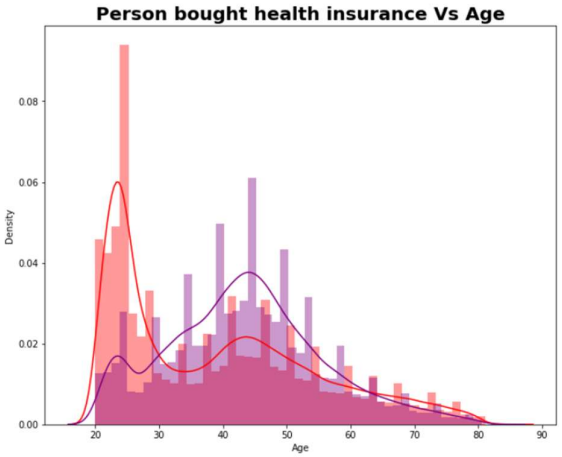


Fig 8. Based upon the vehicle damage response we plot count plot.

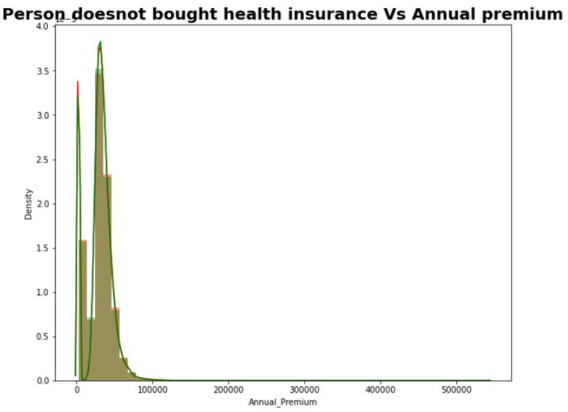
Bivariate analysis

Bivariate analysis of a person who bought health insurance versus age.



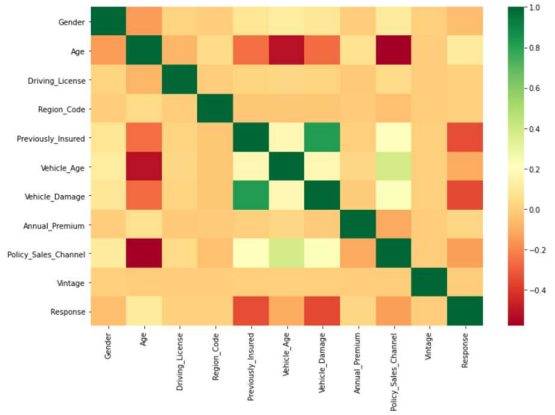
Age group of 40-50 have higher chance of buying the health insurance versus age.

Bivariate analysis of a health insurance versus annual premium.



Annual premium between 20000-50000 have higher chance buying annual premium.

Multivariate analysis

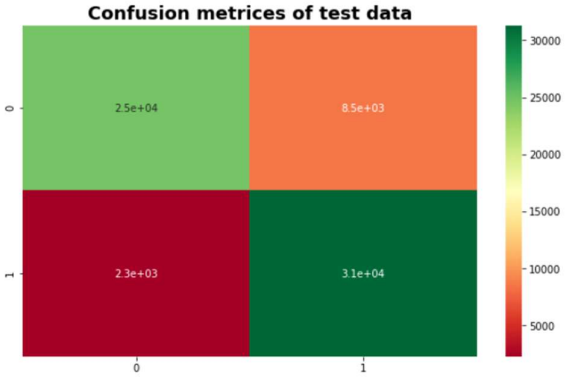


SMOTE (Synthetic Minority Oversampling Technique).

we create smote object and fit the data, imbalanced classification technique to improve the performance of minority classes.

Decision Tree classifier.

Confusion matrix for the test data after classification.



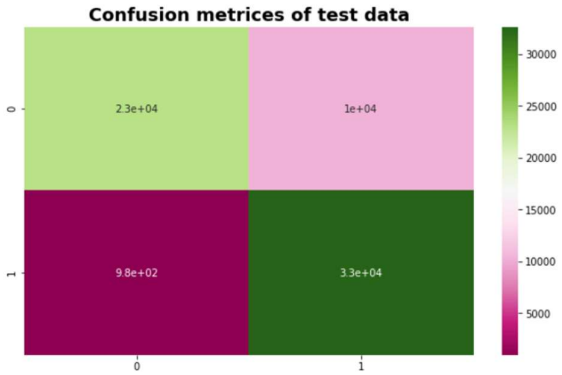
Classification report of the data.

	precision	recall	f1-score	support
0	0.92	0.74	0.82	33287
1	0.79	0.93	0.85	33593
accuracy			0.84	66880
macro avg	0.85	0.84	0.84	66880
weighted avg	0.85	0.84	0.84	66880

accuracy score is 84%

Random Forest Classifier

Confusion matrix for the test data after classification.



Classification report of the data.

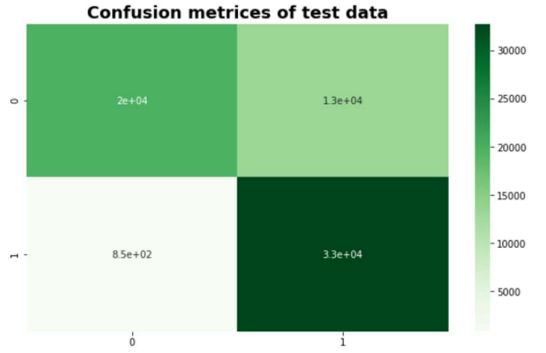
Classification report of train data

	precision	recall	f1-score	support
0	0.96	0.69	0.80	33287
1	0.76	0.97	0.85	33593
accuracy			0.83	66880
macro avg	0.86	0.83	0.83	66880
weighted avg	0.86	0.83	0.83	66880

accuracy score is 83%

Logistic Regression

Confusion matrix for the test data after logistic regression.



Classification report of the data.

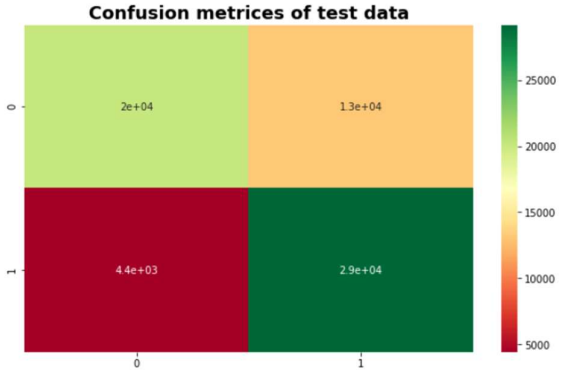
Classification report of test data

	precision	recall	f1-score	support
0	0.96	0.60	0.74	33287
1	0.71	0.97	0.82	33593
accuracy			0.79	66880
macro avg	0.83	0.79	0.78	66880
weighted avg	0.83	0.79	0.78	66880

accuracy score is 79%

KNN Classifier

Confusion matrix for test data after KNN Classifier.



Classification report of the data.

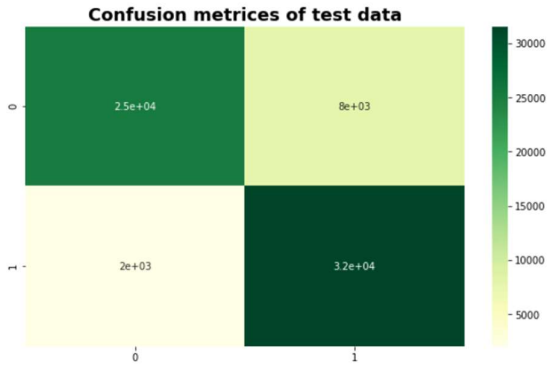
Classification report of test data

	precision	recall	f1-score	support
0	0.82	0.61	0.70	33287
1	0.69	0.87	0.77	33593
accuracy			0.74	66880
macro avg	0.75	0.74	0.73	66880
weighted avg	0.75	0.74	0.73	66880

accuracy score is 74%

XGBoost Classifier

Confusion matrix for the test data after XGBoost Classifier.



Classification report of the data.

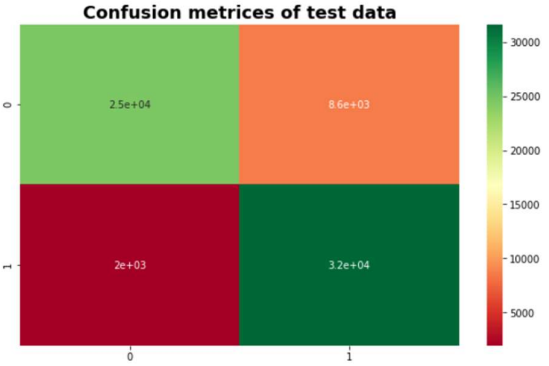
Classification report of test data

	precision	recall	f1-score	support
0	0.93	0.76	0.83	33287
1	0.80	0.94	0.86	33593
accuracy			0.85	66880
macro avg	0.86	0.85	0.85	66880
weighted avg	0.86	0.85	0.85	66880

Accuracy score is 85%

Gradient Boosting Classifier

Confusion matrix for test data after gradient boosting classifier.



Classification report of the data.

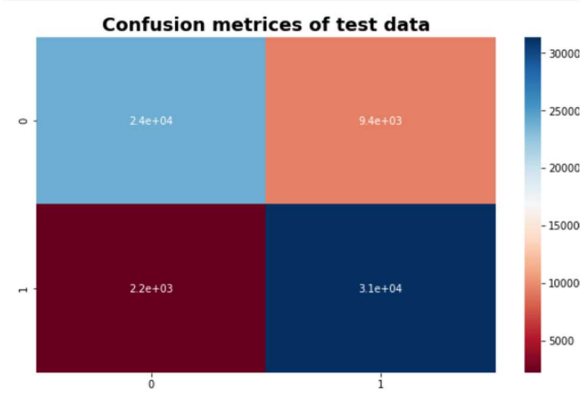
Classification report of test data

	precision	recall	f1-score	support
0	0.93	0.74	0.82	33287
1	0.79	0.94	0.86	33593
accuracy			0.84	66880
macro avg	0.86	0.84	0.84	66880
weighted avg	0.86	0.84	0.84	66880

accuracy score is 84%

Categorical NB

Confusion matrix for test data after Categorical NB.



Classification report of the data.

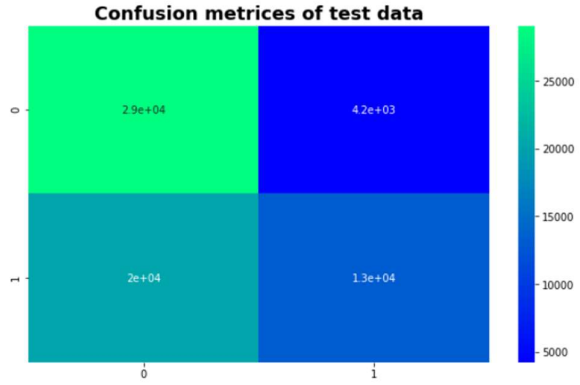
Classification report of test data

	precision	recall	f1-score	support
0	0.92	0.72	0.80	33287
1	0.77	0.93	0.84	33593
accuracy			0.83	66880
macro avg	0.84	0.83	0.82	66880
weighted avg	0.84	0.83	0.82	66880

Accuracy score is 83%

Linear SVC

Confusion matrix for test data after linear SVC.



Classification report of the data.

Classification report of test data

	precision	recall	f1-score	support
0	0.59	0.87	0.70	33287
1	0.76	0.39	0.52	33593
accuracy			0.63	66880
macro avg	0.67	0.63	0.61	66880
weighted avg	0.67	0.63	0.61	66880

accuracy score is 63%

IV. CONCLUSIONS

Using the Decision tree classifier, the accuracy score is 84%. Using the Random Forest Classifier, the accuracy score is 83%. Using Logistic Regression, the accuracy score is 79%. Using the KNN Classifier, the accuracy score is 74%. Using the XGBoost Classifier, the accuracy score is 85%. Using the Gradient Boosting classifier, the accuracy score is 84%. Using the Categorical NB, the accuracy score is 83%. Using the LinearSVC, the accuracy score is 63%.

In this study it is quite evident that XGBoost Classifier works better than other algorithms, with the test accuracy of about 85%. This solution presents an important asset for health insurance company for classifying the customers on the basis of vehicle insurance.

REFERENCES

1. *T. Badriyah, L. Rahmaniah, and I. Syarif, "Nearest neighbour and statistics method based for detecting fraud in auto insurance," in IEEE International Conference on Applied Engineering (ICAE'18), Batam, Indonesia, October 2018.*
2. *T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in ACM International Conference on Knowledge Discovery and Data Mining, (SIGKDD'16), San Francisco, California, USA, August, 2016.*
3. *Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu, and J. Peng, "XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud," in IEEE International Conference on Big Data and Smart Computing (BigComp'18), Shanghai, China, January 2018.*

4. *Heterogeneous Uncertainty Sampling for Supervised Learning. Catlett, J. and Lewis, D. (1994). In Transactions of a Machine Learning 11th International Conference, pages 148-156.*

5. *J. Brownlee, How to Develop Voting Ensembles with Python, Machine Learning Mastery, Apr 16, 2020. https:// machine-learningmastery.com/voting-ensembles-with-python/ (accessed Jul. 08, 2021).*

6. *ML – Gradient Boosting, GeeksforGeeks, Aug. 25, 2020. https://www.geeksforgeeks.org/ml-gradient-boosting/ (accessed Jul. 08, 2021).*

7. *Decision Trees Algorithms | by Madhu Sanjeevi (Mady) | Deep Math Machine learning.ai |Medium. https://medium.com/-deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1 (accessed Jul. 08, 2021).*

8. *Rohrig, K.; Lange, B., Application of wind power prediction tools for power system operations, IEEE Power Engineering Society GeneralMeeting, 18-22 June 2006.*

9. *Sharma S, Agrawal J, Agarwal S, Sharma S, "Machine Learning Techniques for Data Mining: A Survey" ,Computational Intelligence and Computing Research (ICCIC), IEEE International Conference on 26-28 Dec. 2013 Page(s):1 - 6 ,2013.*