PREDICTION OF AUTO INSURANCE CLAIMS

Giselle Fernandes 2308

Swizel Monteiro 2314

ABSTRACT

One of the main challenges faced by the insurance companies is to determine the proper insurance premium for each risk represented by customers. Risk differs widely from for each client, and a careful understanding of various risk factors assists the likelihood of insurance claims based on historical data. The main objective of this research is to build a precise model to predict car insurance claims through machine learning techniques. Real-world datasets often have missing values and hence can cause bias in results. A kaggle car insurance dataset was used, the research was carried out by using Logistic regression, K Nearest Neighbours(KNN), Decision Tree (DT), Random Forest, Adaboost, Gradient Boost, XGBoost and SVM to develop the prediction model. The experimental results showed that the Random Forest, SVC, and the Boosting Algorithm (AdaBoost, Gradient Boost and XGBoost) obtained acceptable results. With a focus on advanced statistical methods and machine learning algorithms that are the most suitable method for handling missing values and large data. XGBoost, which is a new ensemble learning method, proved to be very suitable for both data characteristics. The XGBoost model achieved the best accuracy among the models, with an accuracy of exactly 85.45% & AUC of 0.92.

INTRODUCTION

Auto-insurance claim is a request for financial coverage caused by automobile related loss or sustained damage from a policyholder. Car insurance is required for drivers in almost every state, drivers do need to show proof that they can afford to pay the cost of an accident if it's their fault. Most drivers have car insurance because it is the law, but that doesn't mean one should only buy the minimum required coverage.

Claim prediction is an important process in the insurance industry because the insurance company can prepare the right type of insurance policy for each potential policyholder. Inaccuracies in the prediction of vehicle insurance claims will raise the price of the insurance policy for the good driver and reduce the price of the policy for the driver who is not good. More accurate prediction capability allows the insurance industry to adjust pricing better and makes car insurance coverage more accessible to more drivers.

From a machine learning point of view, the problem of claim prediction can be categorized as supervised learning. Given the historical claim data, we need to build a machine learning model that can predict if a driver will initiate an auto insurance claim.

One of the main challenges faced by the insurance companies nowadays, is to define a proper premium for each risk represented by those customers, the majority of insurance companies keep the data on the history of its operations in a data warehouse. These huge quantities of data are hiding very important knowledge, which could contribute to increasing profitability. This historical data provides the greatest source of information on claim exposure and is the starting place for insurance claim modeling. In this context, machine learning mainly contributes to creating more accurate predictive models to solve such problems.

The main objective of the research is to build a precise model to predict car insurance claims through machine learning techniques to help insurance companies to improve their pricing decision. With a focus on a proposed approach to handle missing data using advanced statistical methods and machine learning algorithms that are the most suitable method for handling missing values. We use Logistic Regression, Decision tree, Random Forest, KNN, and the Boosting algorithms (Adaboost, Gradient Boost, XGBoost) to develop the prediction model.

LITERATURE SURVEY

The rate of the insurance premium in many insurance companies is calculated with only two factors, price of the car and the rate of loss. Different demographic factors, car specifications, and the record of damage caused by the car owner, apart from the policyholder's age, marital status, and gender, vehicle type, the location of the car owner's residence, the driving pattern, and the claim history need to be considered in determining the risk in car insurance. The lack of such factors leads to computing unfair rates of Insurance premiums, because in these cases instead of the customer, the vehicle is insured. That's why most insurance companies experience great losses as far as car insurance.

Previous research has tackled the problem of claim prediction using many other machine learning algorithms. A comparison between three machine learning methods, i.e., neural networks, decision tree, and multinomial logistic regression was carried out. Their results indicated that neural networks were the best predictor to predict whether a policyholder files a claim or not.

The volume of the historical claim data is usually large. Moreover, there are many missing values for many features of the historical claim data. The above previous works did not consider both big volume and missing value issues in their works. Therefore, we focus on examining the machine learning methods that are the most suitable method for the problem of claim prediction with big training data and many missing values.

RESEARCH AND METHODOLOGY

To build the claim predictor, we obtained the data set through the Kaggle site namely Car Insurance Data. The data is split into train, test and validation datasets for training, hyperparameter tuning and testing the accuracy of the models. The training data is used to build a model that can predict whether or not a

driver will make an insurance claim using a very complex dataset consisting of 18 features (columns) and 10,000 entries(rows).

DESCRIPTION OF DATASET

Name	Description
Target	Whether or not a claim was filed for that policyholder last year
Age	Age of the client
Gender	Male / Female
Driving Experience	Years of experience
Education	Graduate or Undergraduate
Income	Class of income
Credit Score	Credit-based insurance score to calculate policy premiums
Vehicle Ownership	Owner of vehicle or not
Vehicle year	Year the vehicle was manufactured
Married	Married / Single
Children	Has children/No
Postal code	Postal address code
Annual Mileage	Annual mileage of vehicle
Vehicle type	Type of vehicle
Speeding Violations	No. of reported speeding violations
DUIS	Arrested previously of driving under the influence of alcohol
Past Accidents	No. of past accidents that were reported under the vehicle

EXPLORATORY DATA ANALYSIS

In this section, we will be doing some basic Exploratory Data Analysis to get the "feel" of the data, we load the dataset into the data frame, view the columns and rows of the data, investigate the dataset to discover patterns, and anomalies (outliers), perform descriptive statistics to know better about the features inside the dataset, write the observations, find missing values and duplicate rows. We check the distributions and the correlations etc. between the different columns. Various graphical representations were made to understand the data better.

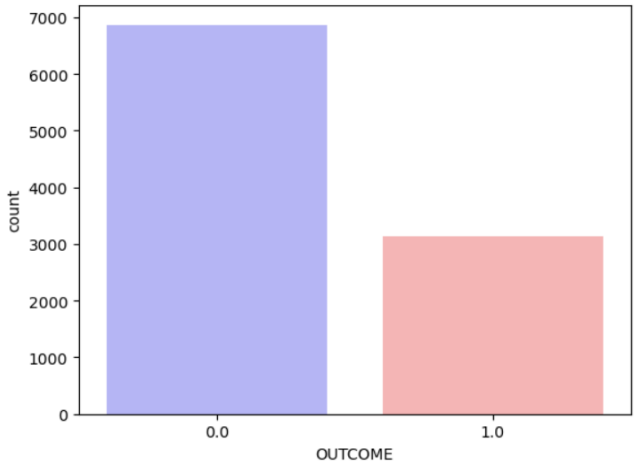

Fig 1: Percentage of people who apply for a loan: 31.33%,
Percentage of people who did not apply for a loan: 68.67%

It is observed that 31.33% persons have claimed insurance; And 68.67% did not claim insurance, so the data problem is somewhat imbalanced. The features are analyzed by drawing plots. The correlation matrix is created to evaluate the relationship between the various features in the dataset. The correlation coefficient measures the strength of the relationship between the relative movements of two variables.

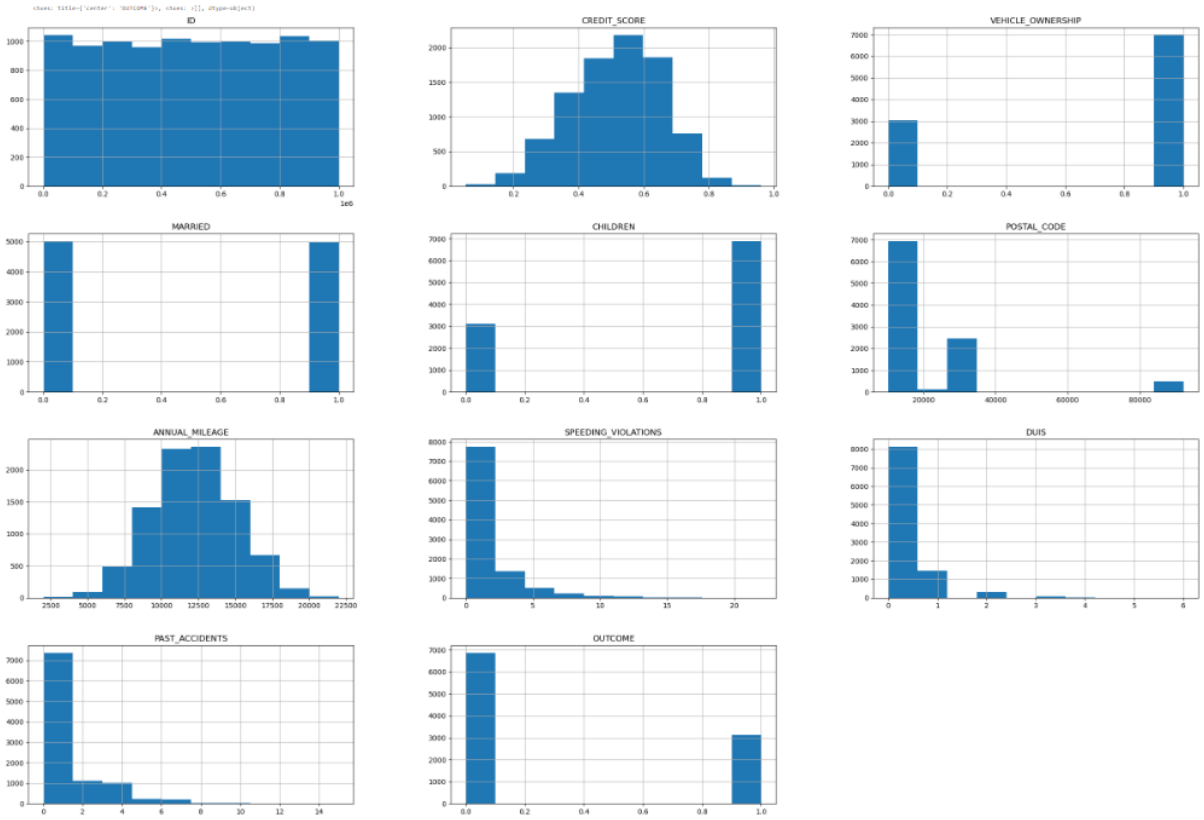

Fig 2 : Plots to analyze the dataset

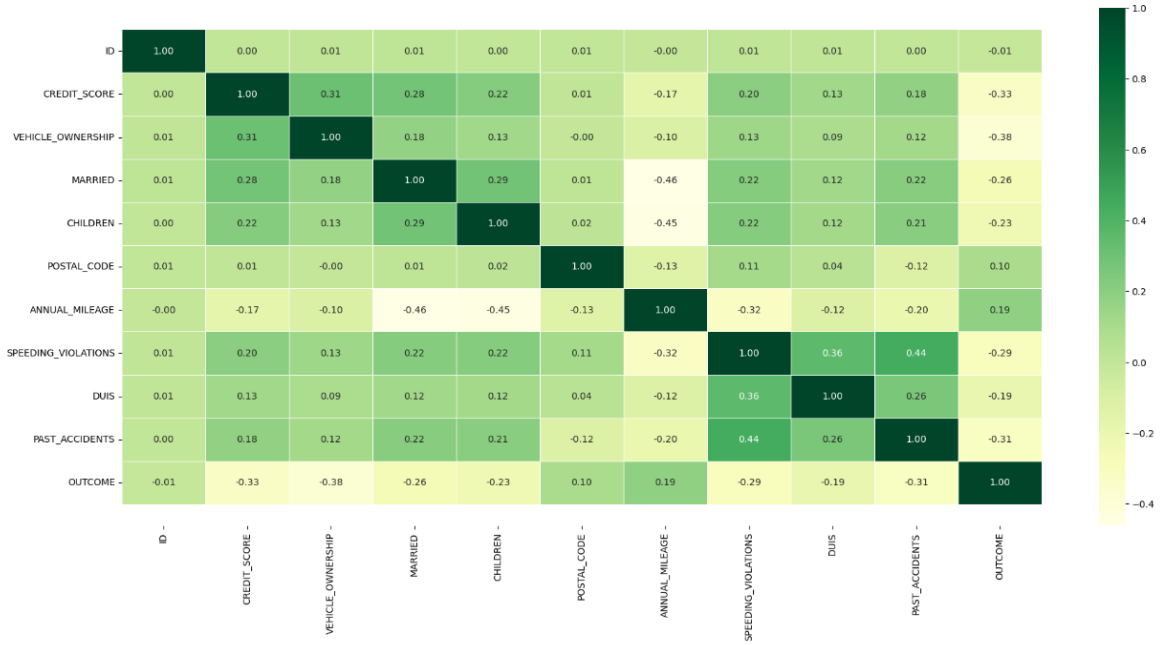

Fig 3 : POSTAL_CODE, ID are the least correlated whereas all other variables show a significant correlation with the OUTCOME(Insurance Claim).

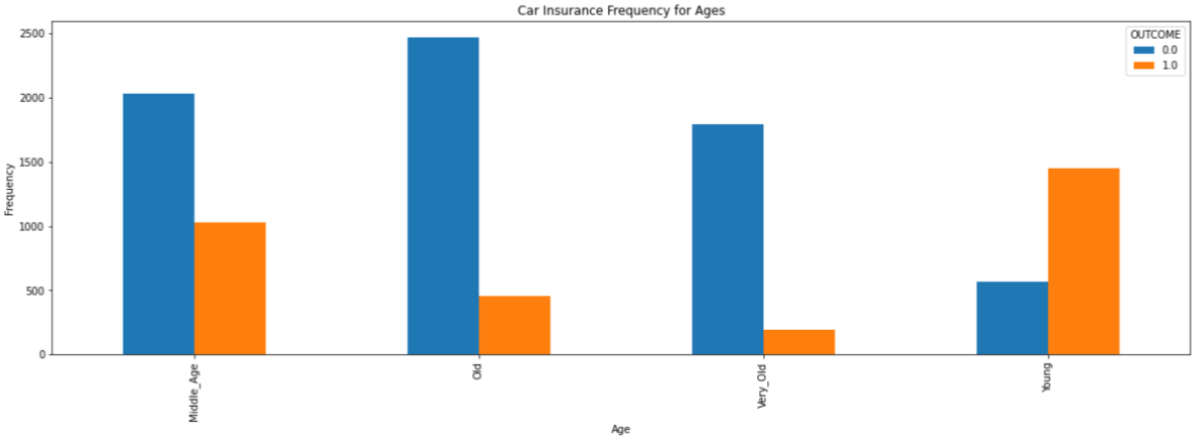

Fig 4 : Young people are more likely to claim loan

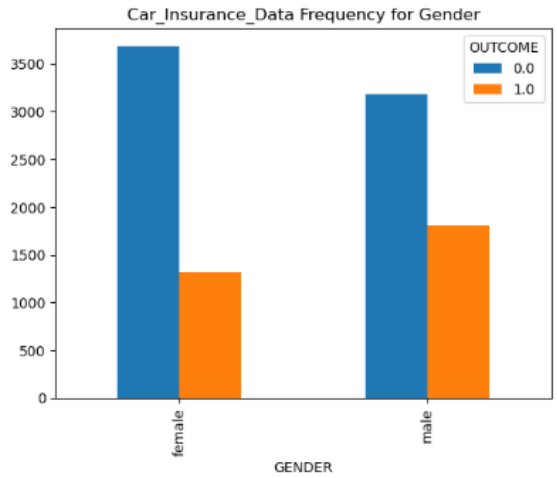

Fig 5: Male are most likely to claim loan than Female

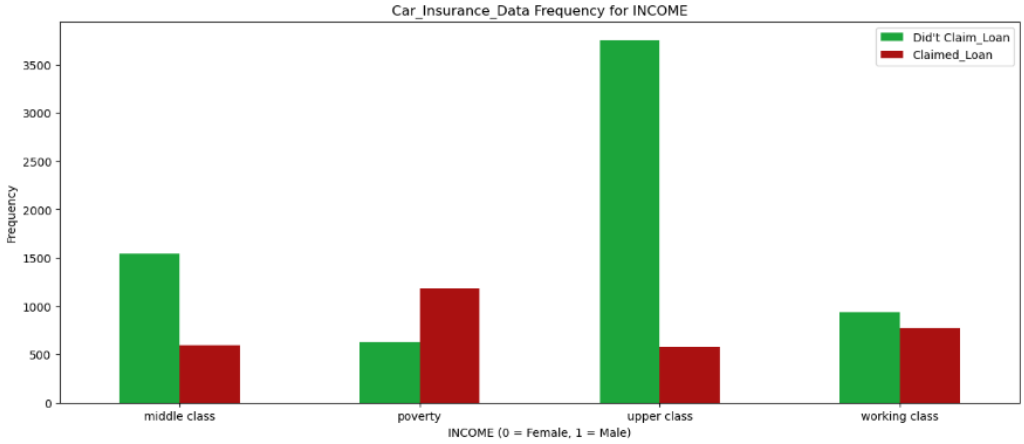

Fig 6 : The higher class is the least likely to claim insurance

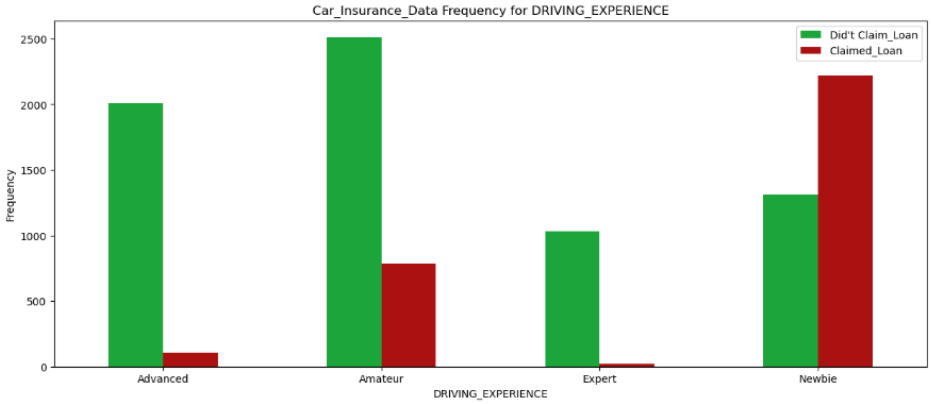

Fig 7: Newbie driver has claimed loan by a wide margin

FEATURE SELECTION

Feature selection is made based on the insights gained from the charts above, as well as the pairwise correlations between independent features and the target column and the average of absolute pairwise correlation values between a feature and all other features of the dataset. In particular, features that show little correlation to the target column and low pairwise correlation to other features will not be considered in the model. It is observed that POSTAL_CODE, ID are the least correlated whereas all other variables show a significant correlation with the OUTCOME variable (Insurance Claim).

DATA PREPROCESSING

The training data is used to build a model as a predictor of probabilities a person will file a claim next year. To improve the predictive effect of our proposed model, the raw data, in which there are missing values, wrong formats, is inconsistent. Therefore, it is important to preprocess the data before developing the predictive model. The following steps have been done to achieve enhancement in this section. SMOTE oversampling method is used for imbalanced classification data.

Handle missing data

There are many missing values existing in the dataset about driver's information, the following is an analysis of missing data. The dataset contains missing values in 2 columns in the Car insurance dataset. There are 982 cases of missing value in one column called Credit Score and 957 in Annual Mileage. The fillna() method replaces the NULL values with a specified feature mean value in the respective columns.

CHOOSING A MODEL

Here we assess the performance of the models ,we use this to choose a better performing model. The Car insurance data set is divided into two parts, 80 percent of which is used as a training set and 20 percent of which is the testing set. The training data is used to model a fitted and logical model. As for testing data, it is utilized to calculate the accuracy of the prediction model. Here widely used classification models are used, namely Logistic regression, K Nearest Neighbours(KNN), Decision Tree (DT), SVM.

There are some machine learning paradigms relevant in big data, especially volume context. They include ensemble learning and online learning. Ensemble learning breaks the large volume of claim data into small ones, training models on a small subset, and combining results with high accuracy. Ensemble learning algorithms including Random Forest, Adaboost, Gradient Boost and XGBoost were also implemented to predict the claim.

HYPERPARAMETER TUNING

The GridSearchCV is a library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit the estimator (model) on your training set. So, in the end, the best parameters can be selected from the listed hyperparameters. After Grid Search, we got the best parameters for all the models. We then tune the hyperparameters to see how it performs.

EVALUATING THE MODELS

After conducting significant data analysis, we experimented with various classification models and ensemble learning methods to see how well they performed on the dataset. We assess the models performance in depth and conduct cross validation. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. The confusion matrix is displayed to evaluate the performance of the models. With accuracy, roc, precision, and recall score, quite decent results were obtained.

RESULTS

We compare the performance of the models used. The results in the table below show that Random Forest, SVC, and the Boosting Algorithm (AdaBoost, Gradient Boost, XGBoost) are models which are best fit for our dataset. After tuning the hyperparameter, XGBoost algorithm has the highest accuracy of exactly 85.45% & AUC of 0.92 and hence outperforms other models on datasets. We can also conclude that the ensemble learning methods give better accuracy than the other models. We can now choose to work with the XGBoost model or any of the ensemble methods to predict the claim of insurance.

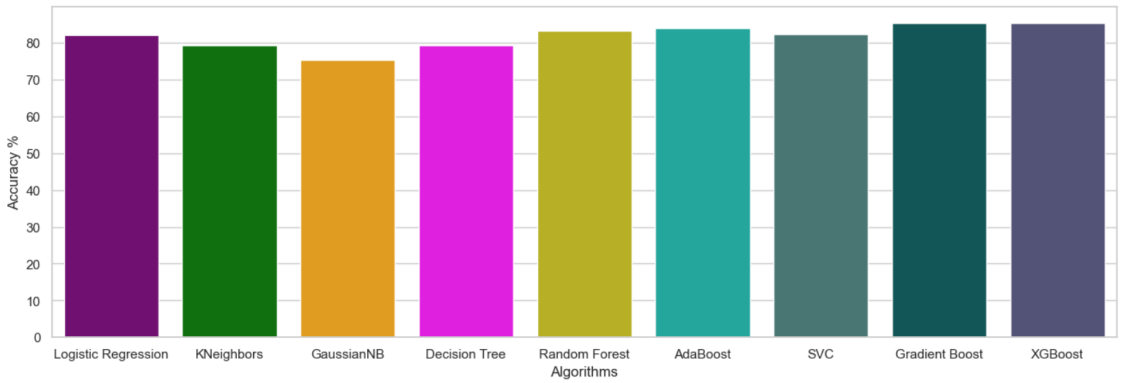

Fig 8: Comparison of performance between models on the dataset

	Model	Accuracy
8	XGBoost	85.50
7	Gradient Boost	85.35
5	AdaBoost	84.00
4	Random Forest	83.30
6	SVC	82.40
0	Logistic Regression	82.15
3	Decision Tree	79.35
1	KNeighbors	79.30
2	GaussianNB	75.35

Table 1: XGBoost algorithm has the Highest Accuracy of exactly 85.45%

CONCLUSION

Claim prediction is an important process in the insurance industry. The volume of the historical claim data is usually large. Moreover, there are many missing values for many features of the data. Therefore, we need machine learning models that can handle both data characteristics. Here, we apply and analyze the accuracy of new ensemble learning called XGBoost for the problem of claim prediction. We also compare the performance of XGBoost with that of other ensemble learning, i.e., AdaBoost, Random Forest, Gradient Boost and other classification models. Our simulations show that XGBoost gives better accuracy than other models. This model hence provides automobile insurance companies with the most accurate prediction as the best benchmark model. The scope of this work as well as its results is limited to the information and knowledge embedded at the provided dataset from Kaggle.

REFERENCES

https://www.i-csrs.org/Volumes/ijasca/11_IJASCA_The-accuracy-of-XGBoost_159-171.pdf
https://www.jatit.org/volumes/Vol98No22/8Vol98No22.pdf

Research Papers

1. A proposed model to predict auto insurance claims using machine learning techniques
2. The accuracy of XGBoost for insurance claim prediction

Dataset

Car Insurance Data | Kaggle

Code