Driver Insurance Claim Prediction

Yash Hatekar
Rishika Samala
Diksha Singh

ABSTRACT

Auto Insurances have become a key part of owning a vehicle nowadays. It has become a necessity for vehicle owners. But despite being a necessity people don’t weigh the importance of getting the best insurance suitable for them. The traditional way of insurance brokers handling the insurance has reduced and online purchases of insurances have increased. Therefore the ability of companies to contact the customers directly has reduced and the best insurance should be quoted at a good reasonable price for a particular customer efficiently and accurately based on his safety and priorities. We aim to solve this problem using Machine Learning methods. These methods help in a good interpretation and comprehension of the data of customers and thus companies can efficiently provide the best insurance to their customers which is beneficial to both of them. We experimented with various methods like Logistic Regression, Support Vector Machine, XGBoost, Adaboost with Decision Tree and Random Forest, NaiveBayes, KNN, Bagging with Decision Tree classifiers and Artificial Neural Networks to predict claim occurrence. Furthermore, we evaluated all these models' performances. The results showed that AdaBoost with Random Forest and Bagging Classifier with decision tree as base estimators perform better than other methods. We achieved accuracy, f1 score, and ROC score of 0.969, 0.969, and 0.969 respectively for Adaboost with Random Forest and 0.954, 0.955, 0.954 for Bagging Classifier with Decision Tree as base estimator.

I. KEYWORDS

Insurance, Customer, Driver, Classification, Machine Learning models, F1 score, ROC.

II. INTRODUCTION

Auto insurance is a non life insurance. When a person has insurance he can claim it in case of a road accident, which helps to reduce the costs of property damage, liability costs and medical costs. When a policyholder files a request for claim in such cases, the company checks and validates the request and then takes a decision of the payment to the customer. The company considers various factors in this process which estimates if the customer will pay the insurance further after issuance of the claim. One of the main factors is the credit history of the customer. It is observed that people with low credit history are the ones who will skip the payments and commit frauds once they claim the insurance. Also one other factor will be the location of the customer. If the location is densely populated with congestion, the probability of accidents at the place is high and thus the customers can be charged a high price. Also other factors like the driving pattern of the customer. If the driving is not rash then such customers can have a low premium amount. There are applications these days on the phone which track the driving car patterns and give the data to the company to understand if the driver could commit more accidents. There are more other factors. These all factors need to be studied clearly to quote a best insurance price. If a good driver has been quoted a very high price for insurance he may choose other insurance which is better for him thus the company loses clients. On the other hand, if the driver's history is not good, he may skip the payments which may incur losses to the company. At this point, the need for Machine Learning comes into the picture. The model should predict which customers are likely to make a claim and thus increase the insurance premium for them and decrease the insurance premium for customers with less chance to make a claim. Our aim is to help companies by giving them the best model to estimate the claims occurrence.

III. RELATED WORK

There has been a lot of auto insurance related work going on in the field using machine learning. There are few papers which we came across. They are, An analysis of customer retention and insurance claim patterns using data mining: A case study. [1], Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection.[2], Research on Probability-based Learning Application on Car Insurance Data.[3], A proposed model to predict auto insurance claims using machine learning techniques.[4]. We observed that these papers have addressed the customer retention patterns, forecasting profitability, prediction of insurance fraud and some work on prediction of claims occurrence which have simple models. As part of this project we wanted to improve on the claim occurrence concept and provide companies with a customer insight so as to provide much more customer catered insurance.

IV. DATA DESCRIPTION

Our dataset is considered from Kaggle Competition[5]. It is the data by PortoSeguro, one of Brazil's largest auto and homeowner insurance companies. Dataset consists of 595212 instances and 59 features. All Personally Identifiable Information has been masked and converted to either binary, numerical or categorical features.

Features Description:

- bin - features with binary values (0 or 1),
- cat - features with categorical values
- Features with tags as ‘ind’, ‘reg’, ‘car’, ‘calc’:
  - ind - personal info of customer (name, etc.)
  - reg - region/location info customer.
  - calc - calculated features.
- Values of -1 indicate missing values.
- Target variable gives if a claim is filed or not

V. DATA PREPROCESSING

For preprocessing we started with cleaning of the dataset :

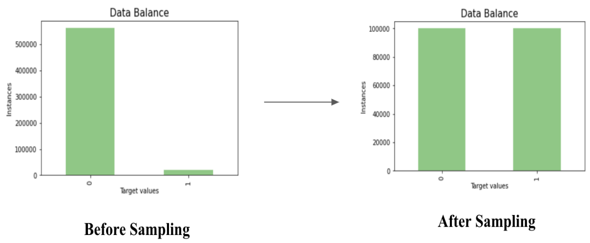
- Missing values represented by -1 are converted into NaN.
- Cleaned data by removing NaN values.
- Highly correlated and zero correlation features were removed.
- Performed one hot encoding on categorical features.

VI. FEATURE ENGINEERING

Once we had cleaned the dataset, the next step was to scale the features and we split the dataset into training and testing data. We performed standardization of feature data by making data with zero mean and unit variance. This helps in the model perceiving all the attributes to be the same and not prioritizing any one based on different data ranges. For the division into training and testing datasets, we have used sklearn train and test split with 70% and 30% of data division.

VII. EXPLORATORY DATA ANALYSIS

First we applied Logistic regression on the unsampled data and we observed that even though the accuracy is good, i.e. 96.4%, the F1 score is 0. So, this F1 score of 0 shows that the dataset is imbalanced, i.e. the effect of samples with class 0 is higher as compared to class 1. So in order to balance the data we applied downsampling and upsampling techniques on our dataset. Since our dataset is very imbalanced with 20K records for claims being filed and 500K records for claims not being filed, sampling is very important. Without sampling the models generated would be very biased to the majority class and not provide accurate or reliable outputs. So we have upsampled the claim being filed data and downsampled the claim not being filed data to eliminate any biases.



Before Sampling

After Sampling

Fig. 1: Number of Instances of target classes before and after sampling

VIII. EXPERIMENTS

After sampling, we experimented training different models on our dataset. We used various metrics like Accuracy, Precision, Recall, F1score, Confusion matrix, ROC score, ROC-AUC curve to measure the performance of the models. And also plotted a roc-auc graph for better visualization. The claim prediction is done to predict if the insured customers will file a claim or not. The output will be based on the target variable which is 0,1 which are categorical values which reflect that customer 'will not file a claim' or 'will file a claim' respectively. Therefore this problem is a binary classification. The purpose of our machine learning models are to predict the probability of claim occurrence.

IX. METHODS

- **Logistic Regression:**

We implemented Logistic Regression with penalty l2, lbfgs solver and with maximum iterations as 100.

- **Linear SVC:**

We used Grid Search CV for choosing the best hyperparameters for Linear SVC and found at max iterations of 2 it was giving a good F1 score.

- **K-Nearest Neighbor:**

We applied the K nearest neighbors model with number of neighbors as 3, distance metric as Minkowski, weights as uniform, leaf size as 30.

- **Gaussian Naive-Bayes:**

Applied Naive Bayes Classifier with variance smoothing as 1e-09.

- **XGBoost Classifier:**

We applied the XGBoost ensemble classifier with default parameters.

- **AdaBoost Classifier with Decision Tree:**

We applied Adaboost classification using Decision tree as base estimator with estimators as 100 and learning rate as 1.0.

- **AdaBoost Classifier with Random Forest:**

We applied Adaboost classification using Random Forest as base estimator with estimators as 100 and learning rate as 1.0. Following is the ROC-AUC curve generated for it as shown in Fig. 2-

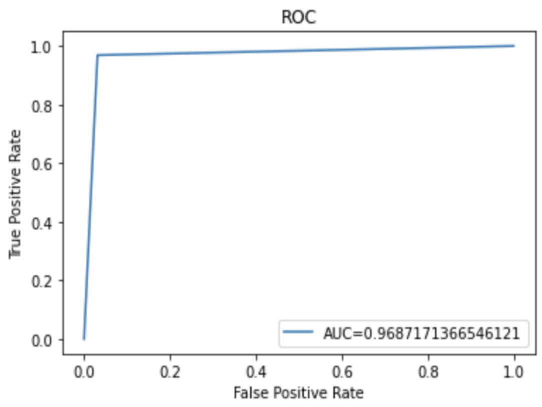


Fig. 2

- **Bagging Classifier with Decision Tree:**

We tried an ensemble bagging classifier with base estimator as Decision Tree and number of estimators as 100, max samples as 1.0. Following is the ROC-AUC curve generated for it as shown in Fig. 3 -

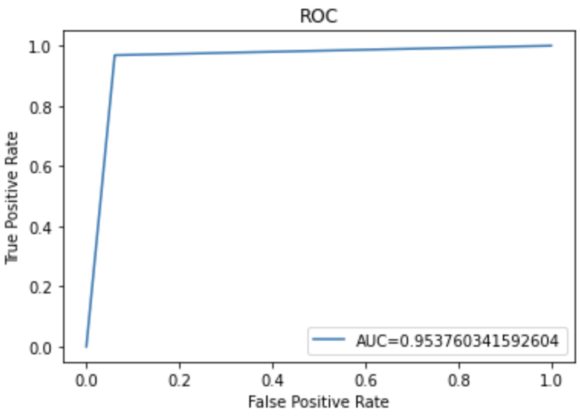


Fig. 3

- **Artificial Neural Network:**

We experimented with different layers and settled with 4 dense layers and one output layer with 512, 256,128, 64 and 1 number of neurons in each layer respectively. We used the initializer ‘RandomNormal’ and used activation ‘Relu’ and ‘Sigmoid’ functions. The learning curves for the ANN are shown in Fig.4. The ROC-AUC curve for ANN is shown in Fig. 5.

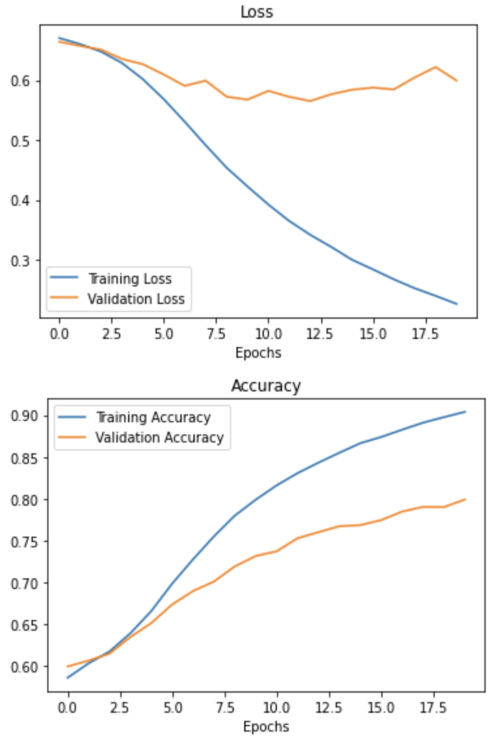


Fig. 4

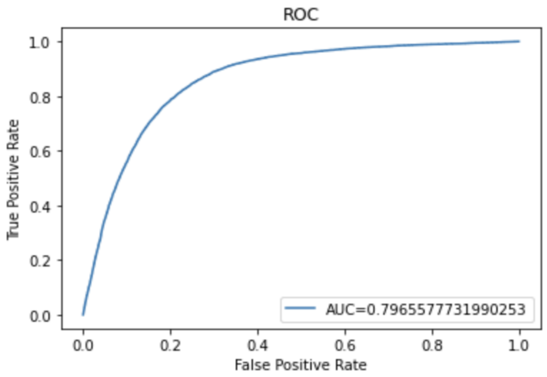


Fig. 5

X. RESULTS AND DISCUSSION

In this section we show the results obtained from different classification models and compare them to analyze the best model for the given data. All the models performance is evaluated based on the confusion matrix, precision, recall, F1-score, and ROC score.

The comparison table is shown below(see Table 1).

Model	**Recall**	**F1 score**	**ROC score**
Logistic Regression	0.561	0.581	0.594
Linear SVC	0.558	0.579	0.593
XGBoost	0.579	0.594	0.603
AdaBoost (Decision Tree)	0.574	0.59	0.599
AdaBoost (Random Forest)	0.969	0.969	0.969
KNN	0.908	0.812	0.789
Naive Bayes	0.385	0.47	0.564
Bagging (Decision Tree)	0.968	0.955	0.954
ANN	0.868	0.811	0.796

Table 1. Model Performance

Model	Training Accuracy	Testing Accuracy	Precision
Logistic Regression	59.3%	59.3%	0.603
Linear SVC	59.3%	59.3%	0.603
XGBoost	60.5%	60.3%	0.611
AdaBoost (Decision Tree)	59.6%	59.8%	0.606
AdaBoost (Random Forest)	100%	96.9%	0.969
KNN	89.6%	78.9%	0.735
Naive Bayes	56.4%	56.3%	0.603
Bagging (Decision Tree)	100%	95.4%	0.941
ANN	90%	79.69%	0.76

Table 1. Model performance (contd.)

Table 1 shows the performance of all classifiers

We have used F1 score and ROC score as the metric to rank our models. F1 score is used as it is inclusive of both precision and recall values.We have used F1 over accuracy as it provides more insight on unevenly distributed classes. ROC score is used since we care about both claims being filed and claims not being filed. F1 score along with ROC score provide a good insight on the performance of the model.

The range of testing accuracy values for all our classifiers was between 56.3% and 96.9%. AdaBoost and Bagging Classifiers were the best models, with a high F1 score of 0.969 and 0.955. Naïve Bayes showed the lowest accuracy of 56.3% and f1 score of 0.47.

As per our analysis, Adaboost performs better than other models as it sequentially improves over multiple iterations by adjusting the weights of incorrectly classified instances such that subsequent classifiers focus more on difficult cases. This decreases bias and improves the overall accuracy and f1 score.

On the other hand, our analysis on Bagging shows that it runs the classifier on multiple subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. This decreases variance and solves overfitting issues and improves the overall accuracy and f1 score.

XI. CONCLUSION

In this project, we implemented different machine learning models to predict the occurrence of insurance claims by customers. We started with linear regression and observed it was not performing well on the data. Next we tried on different models and finally settled with Ensemble techniques like Adaboost and Bagging Classifier as they are giving good performance and are most reliable among all the nine models. These results help companies to provide custom tailored insurance.

XII. FUTURE WORK

We would like to analyze other ML and deep learning models performance using this dataset. We also think that with better hyper parameterization neural networks can be improved to provide better and reliable results. We should also check whether Adaboost with Random forest as estimator and bagging classifier provide similar reliable results with other similar datasets.

XIII. REFERENCES

[1] https://www.tandfonline.com/doi/abs/10.1057/palgrave.jors.2600941

[2] https://www.sciencedirect.com/science/article/pii/S1319157817301672?via%3Dihub

[3]
https://www.researchgate.net/publication/322478575_Research_on_Probability-based_Learning_Application_on_Car_Insurance_Data

[4]
http://www.jatit.org/volumes/Vol98No22/8Vol98No22.pdf

[5]Kaggle:
https://www.kaggle.com/competitions/porto-seguro-safe-driver-prediction/data?select=train.csv

[6]AdaBoost:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[7]BaggingClassifier:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html

[8]
https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc