Vehicle Insurance Claims Prediction with Machine Learning Models

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

by:
[redacted]
DOS REIS Julien

Professor:
Dr. Bertrand Hassani

A Final Report submitted
as a requirement
in
Scoring and Machine Learning

15 January 2022

Contents

1 Introduction

Predictability of claims likelihood is an important study in the insurance industry, as the profitability of insurance/reinsurance companies rely heavily on the modelling of of future events which can be used to determine the right premium prices and manage risk. According to a report by the Insurance Information Institute (2022), although it fluctuates, both claim severity and frequency have the tendency to increase each year. From the data in the United States, it decreased from 2019 to 2020 which might be due to the COVID-19. However, property damage claims grew in frequency and severity by 16.6% and 0.25% respectively between the first quarter of 2020 and the last quarter in 2021. Claim prediction is important in the insurance industry since it provides the knowledge to create the ideal insurance policy for each potential policyholder. In the case of vehicle insurance, the cost of an insurance policy for a good driver will increase due to inaccurate claims prediction, while the cost of an insurance policy for a bad driver will decrease. Greater predictability enables the insurance sector to better adapt prices and increases the availability of auto insurance to more drivers (Fauzan & Murphy, 2017). The problem of claim prediction can be characterized as supervised learning from the perspective of machine learning. In this project, machine learning models were trained and used to predict the probability of a driver to file an automobile insurance claim using historical claims data.

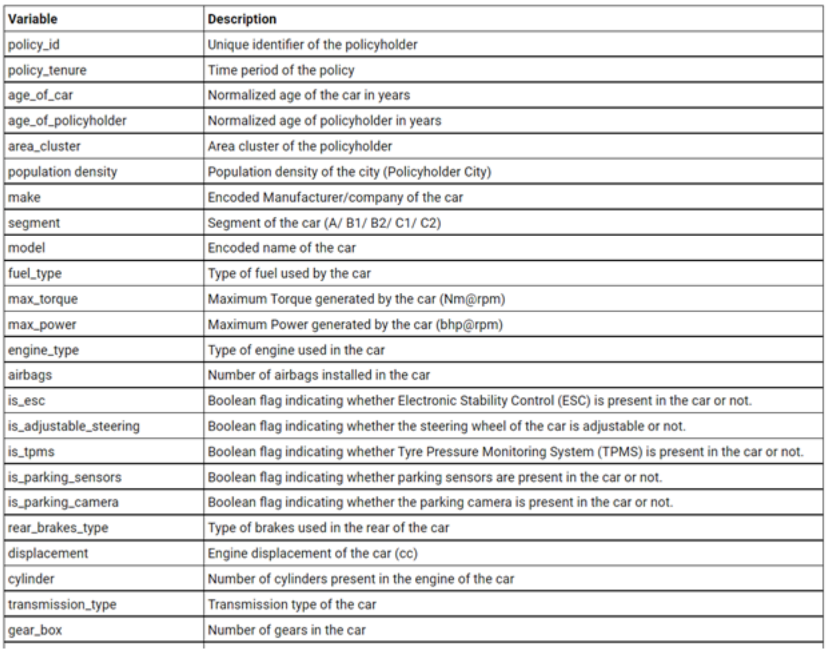
2 Methodology

As in every Machine Learning exercise, the dataset were first cleaned and prepared for analysis. Then, exploratory analysis was conducted to evaluate the features within the dataset and discover relationships between these features and claim likelihood. The champion model selected is ***Random Forest***. This was selected based on prolific examples on isurance claims prediction which are using this model. Theoretically, this classification model works very well in because of its simplicity and the foundational power behind: *wisdom of the crowds*. That is, a combination of a relatively large number of uncorrelated models(*decision trees*) will outperform any of the individual constituent models. However, Random Forest tend to overfit very easily. Despite high accuracy on the training set, it might give high variance results on novel data.

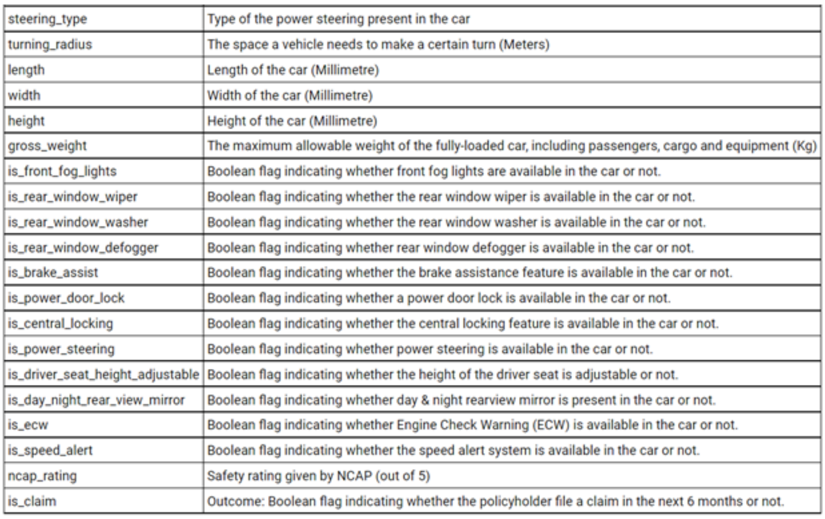
In this paper, three challenger classification models are explored: (i) Adaptive Boosting; (ii) Naive Bayes and; (iii) Logistic Regression. All models were tested using appropriate validation methods such as k-fold cross validation and evaluated using relevant metrics such as accuracy, recall, precision, and f1-score. The results of the analysis were compared with other methods such as those used in previous studies, as in ”Predicting Insurance Claims using Machine Learning Techniques" by Deshpande and Kulkarni (2015) to find the most accurate method in predicting insurance claim likelihood. Overall, this analysis will provide insights on factors that affect the likelihood of motor insurance claims and enable the ability to make more accurate predictions about claim likelihood in the future. It will also contribute to the existing literature by providing a comparison of different machine learning methods and their performance on the specific dataset.

3 Dataset

The dataset used in this study is the ***Car Insurance Claim Prediction*** obtained from an insurance company and was used in the Analytics Vidya DataVerse Hack Competition. The dataset contains information on various factors that may influence the likelihood of a driver filing an auto insurance claim, such as age, gender, location, and driving history. The goal of this study is to use this data to train a machine learning model to predict the probability of a driver filing a claim based on their characteristics. The dataset contains a total of 58592 samples and 43 variables with 18 binary classification variables, which is described below:



Variable	Description
policy_id	Unique identifier of the policyholder
policy_tenure	Time period of the policy
age_of_car	Normalized age of the car in years
age_of_policyholder	Normalized age of policyholder in years
area_cluster	Area cluster of the policyholder
population density	Population density of the city (Policyholder City)
make	Encoded Manufacturer/company of the car
segment	Segment of the car (A/ B1/ B2/ C1/ C2)
model	Encoded name of the car
fuel_type	Type of fuel used by the car
max_torque	Maximum Torque generated by the car (Nm@rpm)
max_power	Maximum Power generated by the car (bhp@rpm)
engine_type	Type of engine used in the car
airbags	Number of airbags installed in the car
is_esc	Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not.
is_adjustable_steering	Boolean flag indicating whether the steering wheel of the car is adjustable or not.
is_tpms	Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not.
is_parking_sensors	Boolean flag indicating whether parking sensors are present in the car or not.
is_parking_camera	Boolean flag indicating whether the parking camera is present in the car or not.
rear_brakes_type	Type of brakes used in the rear of the car
displacement	Engine displacement of the car (cc)
cylinder	Number of cylinders present in the engine of the car
transmission_type	Transmission type of the car
gear_box	Number of gears in the car



steering_type	Type of the power steering present in the car
turning_radius	The space a vehicle needs to make a certain turn (Meters)
length	Length of the car (Millimetre)
width	Width of the car (Millimetre)
height	Height of the car (Millimetre)
gross_weight	The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg)
is_front_fog_lights	Boolean flag indicating whether front fog lights are available in the car or not.
is_rear_window_wiper	Boolean flag indicating whether the rear window wiper is available in the car or not.
is_rear_window_washer	Boolean flag indicating whether the rear window washer is available in the car or not.
is_rear_window_defogger	Boolean flag indicating whether rear window defogger is available in the car or not.
is_brake_assist	Boolean flag indicating whether the brake assistance feature is available in the car or not.
is_power_door_lock	Boolean flag indicating whether a power door lock is available in the car or not.
is_central_locking	Boolean flag indicating whether the central locking feature is available in the car or not.
is_power_steering	Boolean flag indicating whether power steering is available in the car or not.
is_driver_seat_height_adjustable	Boolean flag indicating whether the height of the driver seat is adjustable or not.
is_day_night_rear_view_mirror	Boolean flag indicating whether day & night rearview mirror is present in the car or not.
is_ecw	Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not.
is_speed_alert	Boolean flag indicating whether the speed alert system is available in the car or not.
ncap_rating	Safety rating given by NCAP (out of 5)
is_claim	Outcome: Boolean flag indicating whether the policyholder file a claim in the next 6 months or not.

4 Data Processing and Exploratory Data Analysis(EDA)

Binary Data which are ’Yes’ or ’No’ were replaced by boolean figures ’True’ or ’False’, which are read by Python as 1 or 0, respectively. The first part of the EDA was to plot the numerical variables to represent their distributions. Figure 1 show the summary of these plots. For $'policy_tenure'$, there is no special or recognizable property for its disrtribution and it is spread between 0 to 1.25 years. The variables age_of_car and $age_of_policy_holder$ seem to be heavy-tailed and resembles a Power Law distribution while $population_density$ is highly random.

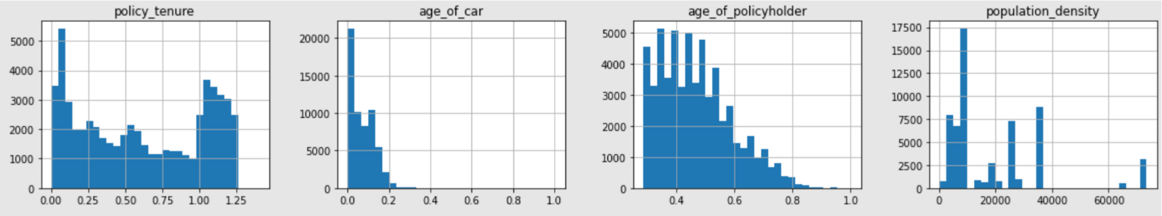


Figure 1: Distribution of Numerical Variables

The next part of the EDA was to check for outliers. Boxplots using the *1.5*Interquartile Range* criteria for each of the numerical variables was plotted to graphically search which of the variables contains outliers. Visibly, in Figure 2, there are three numeric variables which did not pass the test. The outliers in *′age_of_car′* represents 0.46% of the total dataset which is negligible and were removed. After that, the outiers under the variable *′age_of_policyholders′* respresented 0.38% of the dataset which was also negligible and were removed from the dataset. However, for the variable *′population_density′*, the amount of data which did not pass the IQR criteria is 6.24% which is a significant amount of data. With this the data needed to be transformed by taking the *Naperian logarithm.* The test was redone and the it showed that there are 1.32% remaining outliers, representing around 768 points in the dataset. These were ultimately removed from the analysis.

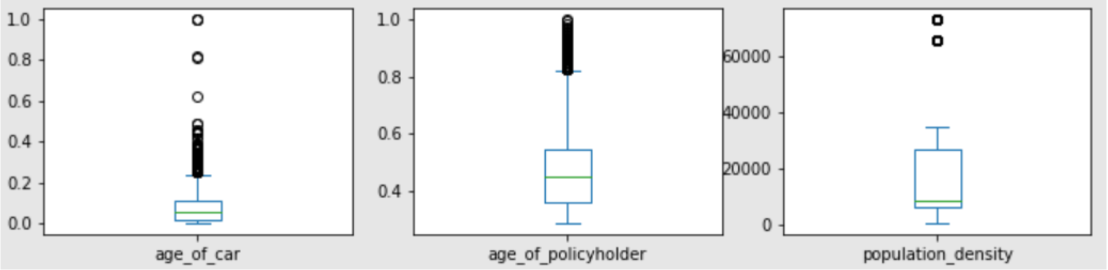


Figure 2: Boxplots of the three Numerical Variables with Outliers

Although can be considered continuous variables, *′max_power′* and *′max_torque′* were treated as categorical variables since they are given categorically (9 categories for each). EDA shows that each category of *′max_power′* correspond to each category of *′max_torque′* in terms of frequency which makes the two variables highly correlated. Correlations between variables is further explored using a correlation heatmap(Figure 3). Then, variables with more than a 0.8 as correlation coefficient were removed. After the criteria was applied the following variables were removed:

```
'is_esc', 'length', 'width', 'gross_weight', 'is_front_fog_lights',
'is_rear_window_wiper', 'is rear_window_washer',is_rear_window_deflagger',
'is_brake_assist','is_driver_seat_height_adjustable',
'is_ecw', 'displacement', 'turning radius'
```

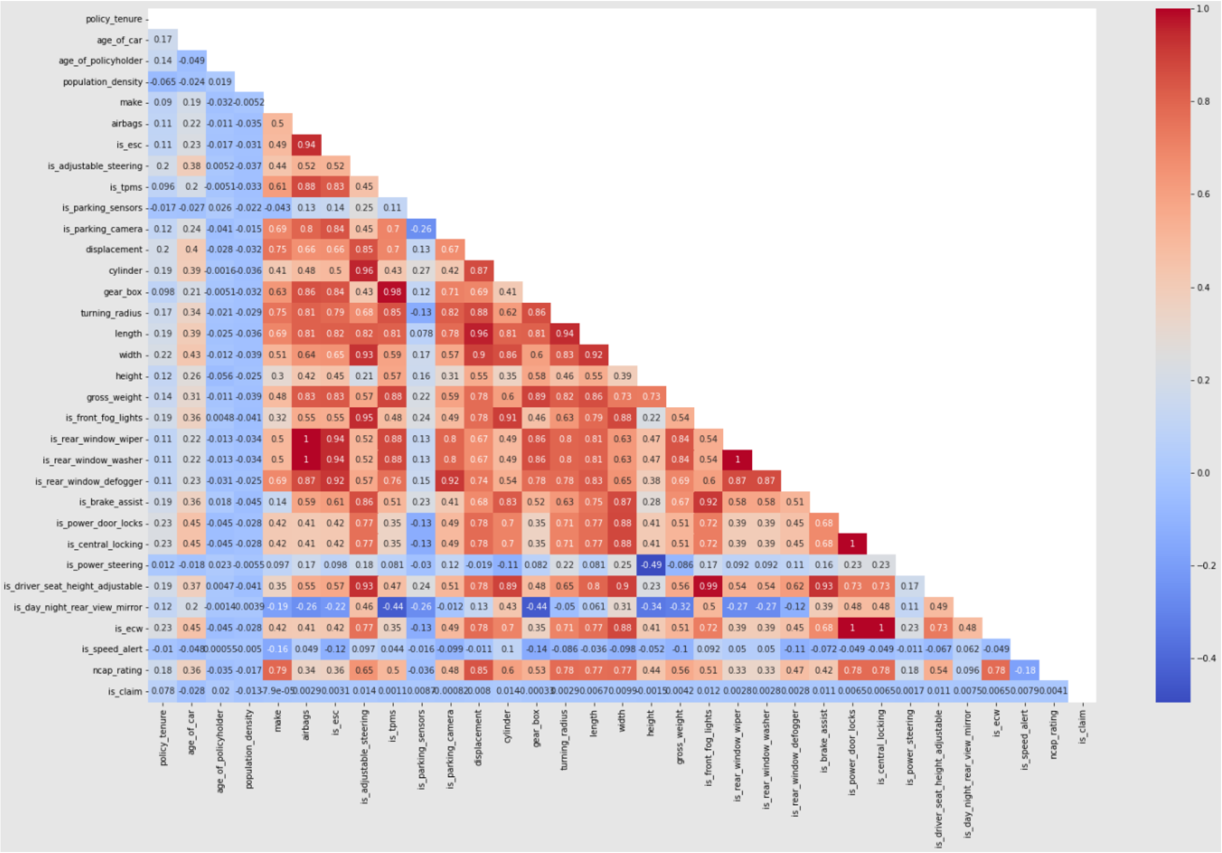


Figure 3: Correlation Heatmap between the Variables

The most important step is splitting the data frame into two parts, one with all the columns except ”is_claim” and one with only the ”is_claim” column, which is a binary variable representing whether a policy holder made a claim in the next 6 months or not. The variable X (independent) holds the dataframe with all columns except 'is_claim' and variable y (dependent) holds only the 'is_claim' column. This process is performed to create the feature set(X) and target variable(y) for the machine learning model. The result shows that there are 54844 observations of policy holders who did not make a claim in the next 6 months and 3748 observations of policy holders who made a claim in the next 6 months. This information also suggests that the proportion of policy holders who made a claim in the next 6 months, is 6.4%. It means that the data is imbalanced as the proportion of claims made is too low. If the proportion of claims made is too low, it can lead to poor model performance. The data is imbalanced as the proportion of claims made is too low. Figure 4 shows shows how much imbalance is the dataset.

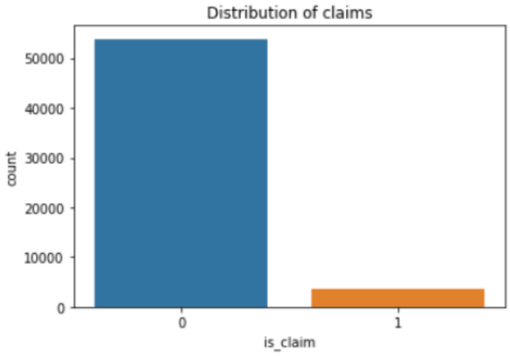


Figure 4: Distribution of Categorical Variable 'is_claim'

The SMOTEENN function, an extension of the SMOTE, was used to address this problem of imbalance dataset in the analysis. It generates synthetic samples of the minority class in order to balance the distribution. The difference with SMOTE is ENN (Edited Nearest Neighbors) which aims to avoid overfitting by removing observations too similar to their nearest neighbors. After applying the function, the distribution becomes as in Figure 5.

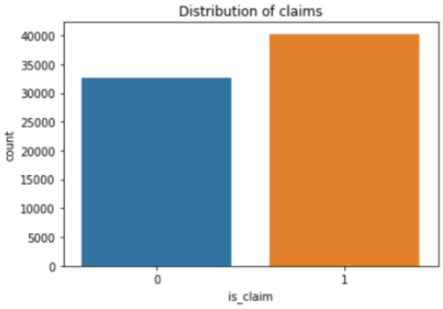


Figure 5: Distribution of Categorical Variable 'is_claim' after SMOTEENN Function

5 The Champion Model

Random Forest Model

Random forest is a method proposed by Breiman and Cutler (2001) which is an extension of the concept of bagging trees and random subspace method. Random forest is an ensemble learning 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. A random forest model is built using decision trees which is a flowchart-like tree structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree recursively in a manner called recursive partitioning. As a predictive model, random forests are often used for feature selection and to improve the predictive accuracy of a model. They can also be used for both regression and classification tasks. In order to use random forests as a predictive model, the model using a labeled dataset needs to be trained, and then use the trained model to make predictions on new, unseen data.

6 The Challenger Models

A. AdaBoost is meant to outperform the Random Forest model due to the data composition. By boosting methods, the goal is to add weight to variables that are more important in the prediction process. The AdaBoost model is creating multiple stumps with a depth of 1. Each tree will compensate for the weakness of previous trees. Hanafy (2021) in his paper: 'Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches' aims to provide a machine learning model which can predict the frequency of insurance claims while solving the challenge of the imbalanced datasets. This imbalance datasets happens since the number of occurring claims is usually significantly lower than the number of non-occurring claims. As a result, classification models tend to have a limited ability to predict the occurrence of claims. In this paper, the authors use various data level approaches to try to solve the imbalanced data problem in the insurance industry by developing 32 machine learning models for predicting insurance claims occurrence (under- sampling, over-sampling, the combination of over-and under- sampling (hybrid), and SMOTE) × (three Decision tree models, three boosting models, and two bagging models) = 32, and then compared the models' accuracies, sensitivities, and specificities to comprehend the prediction performance of the built models. The dataset they used contains 81628 claims, each of which is a car insurance claim. There were 5714 claims that occurred and 75914 claims that didn't occur. According to the findings, the AdaBoost classifier with oversampling and the hybrid method had the most accurate predictions, with a sensitivity of 92.94%, a specificity of 99.82%, and an accuracy of 99.4%. And with a sensitivity of 92.48%, a specificity of 99.63%, and an accuracy of 99.1%, respectively. This paper confirmed that when analyzing imbalanced data, the AdaBoost classifier, whether using oversampling or the hybrid process, could generate more accurate models than other boosting models, Decision tree models, and bagging models.

B. Naive Bayes. The paper "Application of k-NN and Naive Bayes Algorithm in Banking and Insurance Domain" published in 2016, describes the use of k-Nearest Neighbors (k-NN) and Naive Bayes algorithms in the banking and insurance domain. The authors of the paper applied the k-NN and Naive Bayes algorithms to datasets from a bank and an insurance company. The goal of the study was to evaluate the performance of these algorithms in classifying customers as high-risk or low-risk based on their historical data. The study found that both the k-NN and Naive Bayes algorithms performed well in classifying customers as high-risk or low-risk. The paper also highlighted that the k-NN algorithm is sensitive to the choice of k, which is the number of nearest neighbors used to classify a new observation, thus the authors did an experiment to find the optimal k value. The Naive Bayes algorithm, on the other hand, was relatively insensitive to the choice of parameters. The study suggests that these algorithms can be useful in the banking and insurance domain for identifying high-risk customers and developing targeted risk management strategies. However, it is important to note that this paper was published in 2016 and the banking and insurance industry is constantly evolving and new technologies are being developed, so it would be beneficial to also look at more recent research in the field. Additionally, the results and methods used in this paper may not be generalizable to all banking and insurance companies or datasets, and more research would be needed to confirm the results of this study.

C. Logistic Regression is one of the most popular Machine Learning algorithms. It aims to predict the output of a categorical value. So considering insurance claim predictions, with a discrete value as an output, Logistic Regression helps to solve classification problems using discrete and continuous variables. This model will try to assess which variables are impacting the dependent variable by running multiple regressions. Then, each coefficient will be taken into account which will give more importance to the variables that are impactful. These results will be computed to the test dataset so that we will be able to see how the model performed.

There are only a few studies that propose Logistic Regression as the best model to predict the occurrence of vehicle insurance claims. Pesantez-Narvaez, et.al (2019) in their paper titled "Predicting motor insurance claims using telematics data - XGBoost vs. logistic regression" found that logistic regression is a suitable model given its interpretability and good 18 predictive capacity. They created models for a binary response indicating the existence of accident claims vs. no claims can be used to identify the determinants of traffic accidents and then compared the relative performances of logistic regression and XGBoost approaches for predicting the existence of accident claims using telematics data. The dataset contains information from an insurance company about individuals' driving patterns–including total annual distance driven and percentage of total distance driven in urban areas. The logistic regression models equation as follows:

$$\frac{\ln p(x)}{1 - \ln p(x)} = \sum_i \beta_i X_i$$

where $p(x)$ is the probability of observing the event, $\beta' s$ are the model coefficients, and $X_i' s$ are ethe explanatory variables. The result proposes that the Logistic Regression model outperforms XGBoost model in predicting the occurrence of insurance claims.

7 Presentation of the Criteria of Model Selection

In this study, four different machine learning models were explored to predict whether a policy holder will make a claim in the next 6 months or not. These models are Random Forest, AdaBoost, Naive Bayes, and Logistic Regression. The aim is to compare Random Forest against 3 other models. The Random Forest model is an ensemble method that creates multiple decision trees and combines their predictions to make a final prediction. Random Forest is a powerful and widely used machine learning algorithm that can be considered for a wide range of problems, including classification and regression tasks. It is an ensemble method that creates multiple decision trees and combines their predictions to make a final prediction. Random Forest is chosen as the champion model as it can handle high-dimensional and complex datasets with a large number of features. It can also handle non-linear relationships between features and target variables, and it is less prone to overfitting than a single decision tree.Since the dataset contains a large number of features and the relationships between features and target variable is non-linear, Random Forest is one of the models to explore. Random Forest also can handle imbalanced data well, it can handle both categorical and numerical features, which makes it a versatile algorithm.

AdaBoost is also an ensemble method that combines multiple weak classifiers to make a strong classifier. Naive Bayes is a probabilistic classifier that makes predictions based on the probability of certain features belonging to a particular class. Logistic Regression is a linear model that predicts the probability of a certain event occurring. Each model was trained on the same dataset and their performance was evaluated using standard evaluation metrics such as accuracy, precision, recall, F1-score. This study aims to compare and contrast the performance of these four models to determine the best model for the given problem and dataset to find the best model for predicting the occurrence of vehicle insurance files by the policy holders.

8 Results and Discussion

Random Forest

```
   precision    recall  f1-score   support

           0       0.96      0.96      0.96      6648
           1       0.97      0.97      0.97      8128

    accuracy                           0.96     14776
   macro avg       0.96      0.96      0.96     14776
weighted avg       0.96      0.96      0.96     14776
```

The accuracy score is the proportion of correctly predicted labels out of all the predictions made. From the table above, it can be seen that, for class 0(0: not claim) the precision is 0.96, recall is 0.96, and f1-score is 0.96, which means that the model has a 96% precision. It was also confirmed the accuracy score by passing the accuracy_score(y_test,preds) function and in this case, the accuracy score is 0.9646, which means that the model has a 96.47% accuracy on the test dataset.

After applying Random Forest to the dataset, the confusion matrices were then checked. The matrices are classification reports which describes the precision, the recall, and the f1-score.

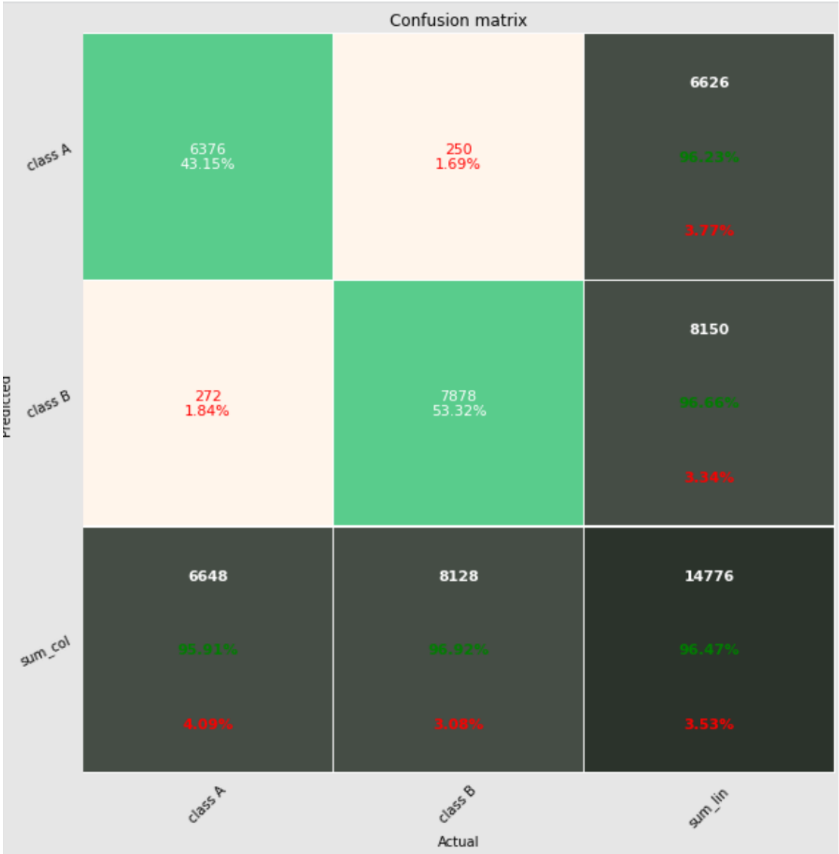


Figure 6: Distribution of Categorical Variable 'is_claim' after SMOTEENN Function

From the confusion matrix graph above, the following can be noted:

- True Positives (TP): 6376 observations are correctly predicted as the positive class, which represents 43.15% of all actual positive observations.
- True Negatives (TN): 7878 observations are correctly predicted as the negative class, which represents 53.32% of all actual negative observations.
- False Positives (FP): 250 observations are incorrectly predicted as the positive class, which represents 1.69% of all actual negative observations. It is also known as a Type I error.
- False Negatives (FN): 282 observations are incorrectly predicted as the negative class, which represents 1.84% of all actual positive observations. It is also known as a Type II error.

Based on the values of these cells, different evaluation metrics such as precision, recall, F1-score, and accuracy of the model with the train set can be calculated.

- Precision: In this case, the precision is TP/(TP+FP) = 6376/(6376+250) = 0.96.
- Recall: The recall is TP/(TP+FN) = 6376/(6376+282) = 0.96
- F1-score: F1-score is the harmonic mean of precision and recall. It ranges from 0 to 1, and a value of 1 indicates that the model has perfect precision and recall. In this case, the F1-score is 2*(PrecisionRecall)/(Precision+Recall) = 2(0.96*0.96)/(0.96+0.96) = 0.96.

In conclusion, the final accuracy of the model applied to the dataset is computed using the accuracy values of 10-fold cross validation that is equal to 0.96 which is considered high as it is close to 1. Afterwards, 3 challenger models was tested against this random forest model starting from ADABoost.

8.1 ADABoost

In this part, the input dataset splitted into a training set (80%) and test set (20%) before initializing the classifier by specifying the number of estimators (weak classifiers) as 100 and a random state as 0. The accuracy of the model on the test set using the score() method is 0.769 which means that the model correctly classifies 77% of the test set observations. This is lower than the random forest model which we proposed in the previous section.

8.2 Naive Bayes

In this section, a Multinomial Naive Bayes classifier was trained on the dataset forclassification, the MultinomialNB() function from the scikit-learn library was used, which is a probabilistic classifier that makes predictions based on the probability of certain features belonging to a particular class. The result shows the accuracy of the model on the test set using the score() method is 0.5382, this means that the model correctly classifies 53.8% of the test set observations. It can be observed that it is lower than the 2 previous models, random forest and ADABoost model.

8.3 Logistic Regression

In this part, we trained a Logistic Regression classifier on a dataset, it is using the LogisticRegression() function from the scikit-learn library, which is a linear classifier that makes predictions based on the probability of certain features belonging to a particular class. The accuracy of the model on the test set is 0.5646 which means that the model correctly classifies

56.46% of the test set observations. It is slightly higher than Naive Bayes Model but it seems still to be far lower than the proposed model (Random forest) which shows 96% accuracy.

9 Limitations

There are several limitations of the models used in this project. One limitation of the Random Forest model is that it can be computationally expensive and require a lot of memory when working with large datasets or many features. It also may not perform well when the dataset is too small or when there is a high degree of collinearity among the independent variables. On the other hand, ADABoost model's limitation is sensitivity to noisy data and outliers, which can negatively impact its performance. It also can be sensitive to the choice of base estimator, the number of estimators and the learning rate.

Naive Bayes also has a limitation where the model makes a strong assumption about the independence of the features, which may not always hold true in real-world datasets. It also may not perform well when the dataset is too small or when there is a high degree of collinearity among the independent variables. Limitation of the Logistic Regression model is that it assumes a linear relationship between the independent variables and the log-odds of the dependent variable, which may not always be the case. It also may not perform well when the dataset is too small or when there is a high degree of collinearity among the independent variables. It also can be sensitive to outliers and multicollinearity.

10 Conclusion

A. The Best Predictive Model for Predicting the Occurrence of Insurance Claims
In this project, four different models were trained to predict the likelihood of policy holders claiming insurance, namely Random Forest, ADABoost, Naive Bayes, and Logistic Regression. The models were trained on a dataset with imbalanced classes, where the majority of the observations were of the negative class. From the results, it can be seen that the Random Forest model performed the best with an accuracy of 96% on the test set. It performed the best among the four models trained in this project. With an accuracy of 96% on the test set, it was able to correctly classify 96% of the test set observations. This is significantly better than the other models, ADABoost with an accuracy of 77%, Naive Bayes with an accuracy of 53.8%, and Logistic Regression with an accuracy of 56.46%. Additionally, the Random Forest model also had a high precision and recall, indicating that it was able to correctly identify a large number of positive instances while also keeping the number of false positives low. Thus the Random Forest model is the best model to predict the likelihood of people claiming insurance among the models that have been trained and evaluated in this project.

B. Machine Learning Predictive Model from Business Point of View for Insurance Companies At the end, using the feature importance selection of Random forest, it was shown that $'policy\_tenure,'age\_of\_car', age\_of\_policyholder'$ are the most signficant features in predicting car insurance claims. fraud. This information is valuable as companies can make preventive measures or risk mitigation knowing that these features are determining factors of fraud.

A predictive model that is able to accurately predict the likelihood of people claiming insurance can greatly benefit an insurance company. First, the model can be used to identify high-risk policyholders, allowing the company to take proactive measures such as offering additional coverage or adjusting premiums to mitigate potential losses. Second, the model can also be used to identify policyholders who are less likely to claim insurance, allowing the company to offer targeted discounts and promotions to retain these policyholders and potentially increase their lifetime value. Third, the model can also be used to identify patterns and trends in claims, which can be used to improve underwriting processes and develop more effective risk management strategies. Finally, the use of the Random Forest model in this project can potentially improve the overall efficiency of the insurance company by reducing the number of false claims, and by identifying and retaining profitable policyholders. Overall, the implementation of a predictive model like Random Forest in this project, can help an insurance company to identify and manage risks more effectively, improve underwriting processes, and increase overall profitability.

11 References

Baran, Sebastian Rola, Przemysław (2022), Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. https://doi.org/10.48550/arXiv.2204.06109

Fauzan, Muhammad Arief Murfi, Hendri (2018). The Accuracy of XGBoost for Insurance Claim Prediction. International Journal of Advance Soft Compu. Application, Vol. 10, No. 2.

Gourav, Rahangdale Manish, Ahirwar Mahesh, Motwani (2016). Application of k-NN and Naïve Bayes Algorithm in Banking and Insurance Domain. International Journal of Computer Science Issues (IJCSI); Mahebourg Vol. 13 (5): 69-75.

Hanafy, Mohamed (2021), Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches. International Journal of Advanced Computer Science and Applications (IJACSA), Vol 12 (6).

Insurance Information Institute. Facts + Statistics: Auto insurance.
Accesed through https://www.iii.org/fact-statistic/facts-statistics-auto-insurance on 14 January 2022.

Pesantez-Narvaez, Jessica Guillen, Montserrat, and Alcañiz, Manuela, (2019), Predicting motor insurance claims using telematics data - XGBoost vs. logistic regression. Risks, 7, 70.

Yehuda, Kahane Nissan, Levin Ronen, Meiri Jacon, Zahavi, (2007). Applying Data Mining Technology for Insurance Rate Making: An Example of Automobile Insurance. Asia-Pacific Journal of Risk and Insurance,vol. 2, issue 1, 1-19.