Machine Learning Summer Term 2023

COMP-7745-301

Shivaram Rudroju (U00892957)

Final Project

04 - August – 2023

Car Insurance Claim Classification Using Machine Learning

Table of Contents

Abstract

Machine learning techniques like Random Forest, XGBoost, Gradient Boosting, as well as Light GBM were researched for the prediction of insurance claims. However, with claim samples being fewer than 10%, the models' performance was hampered by the high class imbalance. The dataset was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) in order to address this problem. After oversampling, XGBoost outperformed other models, providing better insurance claim forecasting accuracy in the process.

Feature such as "is_parking_sensors," "rear_brakes_type," and "is_power_door_locks" were identified through feature importance analysis utilizing mutual information score as critical markers of insurance claims. To evaluate model performance as well as validate predictions, evaluation metrics such the ROC curve, confusion matrix, and classification report were employed. These results highlight the significance of correcting class imbalance and making use of cutting-edge machine learning methods for successful insurance claim prediction.

Keywords: Random Forest, XGBoost, Gradient Boosting, Light GBM, class imbalance, SMOTE, feature importance, mutual information score, ROC curve, confusion matrix, classification report.

Introduction

In an era of rapid increase in the number of vehicles on the road, administering auto insurance claims has become a difficult task. The implementation of machine learning offers a promising solution for handling and classifying these claims efficiently. This study centres on developing an effective machine learning model to classify auto insurance claims, such as "Accident Damage" and "Theft," thereby streamlining the claims process for faster assessments as well as enhanced customer satisfaction (Singh et al., 2019). This research seeks to improve efficiency in operations as well as minimise fraudulent claims for the benefit of insurance companies and policyholders by applying various algorithms as well as utilizing historical claim data.

• Background and Related Works

In the auto insurance industry, addressing insurance claims is an important as well as lengthy procedure. As soon as policyholders submit claims, insurance companies must precisely assess the claim's authenticity and severity in order to respond appropriately. Traditional approaches of claim classification can be prone to errors and inefficiency, resulting in processing delays as well as dissatisfied customers (Singh et al., 2019).

By developing a trustworthy and precise model, insurance companies can expedite claim processing, optimize resource allocation, as well as decrease the likelihood of fraudulent claims. Also, the project can increase customer satisfaction by ensuring a quicker and more equitable claims assessment process, ultimately resulting in an insurance industry that is more successful and efficient.

The authors (Roy and George, 2017) in their study conducted an evaluation of different models of machine learning, including the decision tree and neural networks, in order to analyze the likelihood of policyholders submitting insurance claims. Additionally, the researchers examined the potential impact of the case study on the insurance company. This study demonstrates that the neural network model outperforms the decision tree model.

In their study, the authors (Alamir et al., 2021) employ two distinct methodologies, namely XGBoost and logistic regression, in order to forecast the occurrence rate of motor insurance claims. This study demonstrates that the XGBoost model demonstrates a slight superiority over logistic regression. It is worth noting, however, that the database utilised for this study consisted of a limited sample size of 2457 observations.

- ### Objectives

The main goal of this project is to train machine learning algorithms to predict insurance claim of cars and evaluate the performance of the model with the help of different classification metrics

The objectives which will be used to achieve the following aims, which are as follows

- To explore the variables in the data and preprocess the variables for predictive modelling
- To visualize the relationship between features and understand the importance of variables with Insurance claim
- To perform data splitting and apply machine learning classifiers to train the data
- To assess the performance using evaluation metrics after predicting the models in test data
- To compare the performance of different ML models and choose the best model for Insurance claim prediction

- ### Algorithms to be used

- Random Forest
- Gradient Boosting
- XGBoost
- Light Gradient Boosting

Methodology

1. Data Exploration

➢ In this step, we will collect the dataset of auto insurance claims.
➢ We will conduct exploratory data analysis (EDA) to understand the structure and distributions of the dataset.
➢ This step involves the visualization of significant statistics as well as connections between variables.

2. Data Cleaning and Preprocessing

➢ In this step, outliers as well as missing values in the dataset will be addressed.
➢ We will encode categorical variables using appropriate methods.
➢ Normalizing numerical features and separating the data into training and testing sets are included in this step.

3. Model Selection

- In this phase, classification models such as Random Forest, Gradient Boosting, XGBoost, and LightGBM will be selected.

4. Model Training

- In this step, the training data which is 80% of the total data will be fed to all the models.
- In this phase, the selected models will be trained using the preprocessed training dataset.

5. Performance Evaluation

- In the next step, we will evaluate the performance of each model using appropriate metrics such as precision, recall, and accuracy.
- The results will be compared to determine which model performs the best.
- For insights, evaluation metrics and feature importance will be visualized.

Model Description

- Random Forest: The Random Forest algorithm is an ensemble learning technique that constructs multiple decision trees during the training process. Every individual tree is built by utilizing a randomly selected subset of the available training data and features. The ultimate prediction is determined through the process of aggregating the predictions of each individual tree, either by averaging or by voting. The algorithm exhibits robustness, effectively handles data with high dimensions, and mitigates the issue of overfitting, rendering it well-suited for both classification and regression tasks (Mantas et al., 2018).
- Gradient Boosting: It is an ensemble technique that creates a sequence of decision trees, with each successive tree aiming to correct the errors of the previous one. The boosting algorithm aggregates weak learners to form an effective learner by specifically targeting the errors made by the preceding trees. Gradient Boosting has demonstrated effectiveness in capturing complicated patterns and attaining a notable level of predictive accuracy. Classification and regression tasks are frequently performed using this method (Körner et al., 2018).
- XGBoost: XGBoost, or Extreme Gradient Boosting, stands as a meticulously optimized rendition of the gradient boosting algorithm. The gradient boosting framework has been strengthened with supplementary regularization techniques and an improved algorithm. XGBoost is renowned for its rapidity and efficacy, making it a widely chosen option for machine learning competitions and practical implementations (Hah, Kim and Ahn, 2019). The software effectively manages the presence of missing data and offers reliable estimation of feature importance.
- Light Gradient Boosting: LightGBM, referred to as Light Gradient Boosting Machine, is a gradient boosting framework designed with a primary focus on maximizing efficiency and improving computational speed. The proposed method employs a histogram-based strategy for feature partitioning, which leads to enhanced efficiency in training durations. LightGBM is known for its memory efficiency, ability to handle large datasets, and versatility in performing classification and regression tasks (Jing et al., 2020). The method provides a high level of precision and is especially valuable in the analysis of datasets with a large number of dimensions.

Results and Experiments

- ## Database

The data is collected from Kaggle which contain the following attributes

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.

Figure 1: Attributes of the data

The data is found to have imbalanced number of claimed samples where 54844 samples in the data are non claimed and 3748 samples in the data are claimed. The data is found to have no missing values in any columns.

- ## Training and Testing Logs

The training and testing logs contain the exploration of the variables with the help of different data visualization approaches as well as the machine learning results obtained from evaluation of the models in test data. For the purpose of training the models, we had splitted the data in 80:20 where 80% of the data is taken for training all the models and 20% is taken for validation of the models.

```
policy_id                           0
policy_tenure                       0
age_of_car                          0
age_of_policyholder                 0
area_cluster                        0
population_density                  0
make                                0
segment                             0
model                               0
fuel_type                           0
max_torque                          0
max_power                           0
engine_type                         0
airbags                             0
is_esc                              0
is_adjustable_steering              0
is_tpms                             0
is_parking_sensors                  0
is_parking_camera                   0
rear_brakes_type                    0
displacement                        0
cylinder                            0
transmission_type                   0
gear_box                            0
steering_type                       0
turning_radius                      0
length                              0
width                               0
height                              0
gross_weight                        0
is_front_fog_lights                 0
is_rear_window_wiper                0
is_rear_window_washer               0
is_rear_window_defogger             0
is_brake_assist                     0
is_power_door_locks                 0
is_central_locking                  0
is_power_steering                   0
is_driver_seat_height_adjustable    0
is_day_night_rear_view_mirror       0
is_ecw                              0
is_speed_alert                      0
ncap_rating                         0
is_claim                            0
dtype: int64
```

Figure 2: Missing values Exploration in data

During the exploration the data, it is found that data contain no missing values which shows that the data is completely clean and ready for predictive modelling purpose.

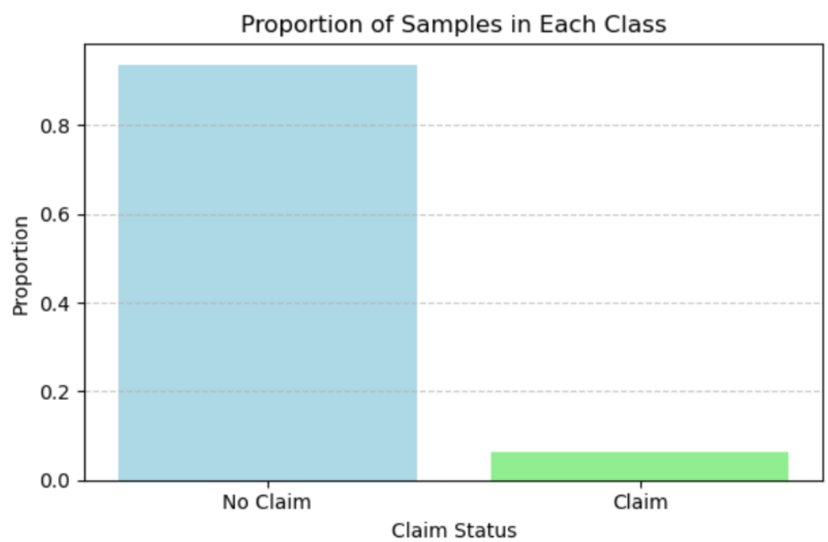


Figure 3: Proportion of Samples

From the proportion of the samples, it is observed that the claim samples are quite less compared to the non claimed samples where nearly 10% of the entire data contain claimed samples.

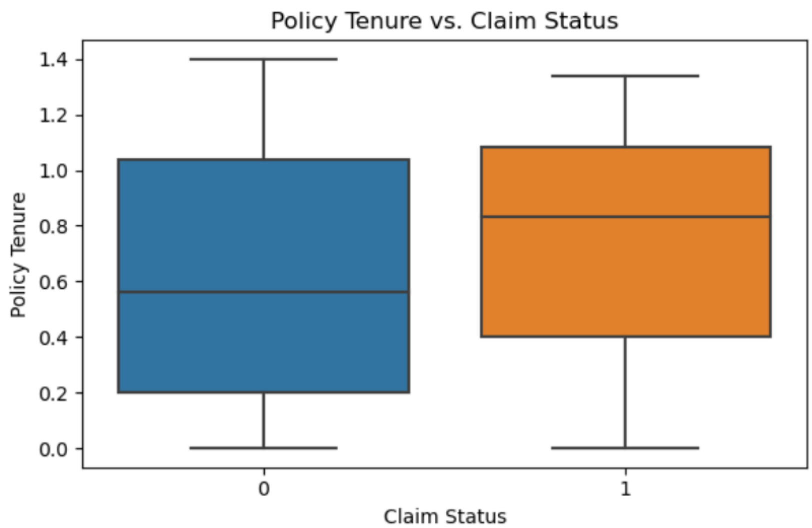


Figure 4: Policy Tenure Distribution of claim Samples

The policy tenure is visualized based on claim status where non claimed customer seem to have higher range in policy tenure but the mean distribution of policy tenure is comparatively higher in claimed customers.

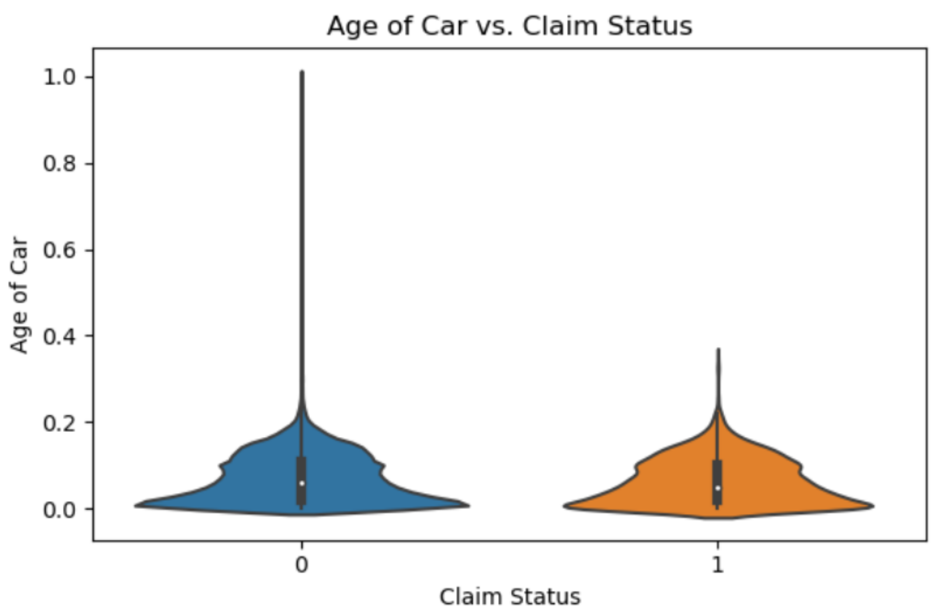


Figure 5: Age of Vehicle based on claim

The age of the vehicles also plays a vital role for Insurance claim. The visual plot shows that insurance which are claimed are generally for new cars where the age of the car is comparatively lower than the cars which are not claimed for insurance.

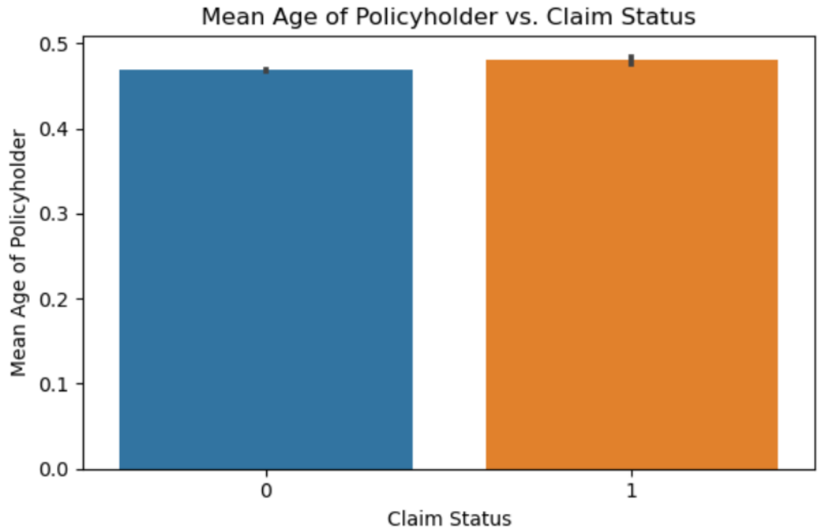


Figure 6: Age of Policy holder based on claims

The age of the policy holder is also visualized based on claims where customers who claim for insurance relatively have higher age compared to non claimed customers.

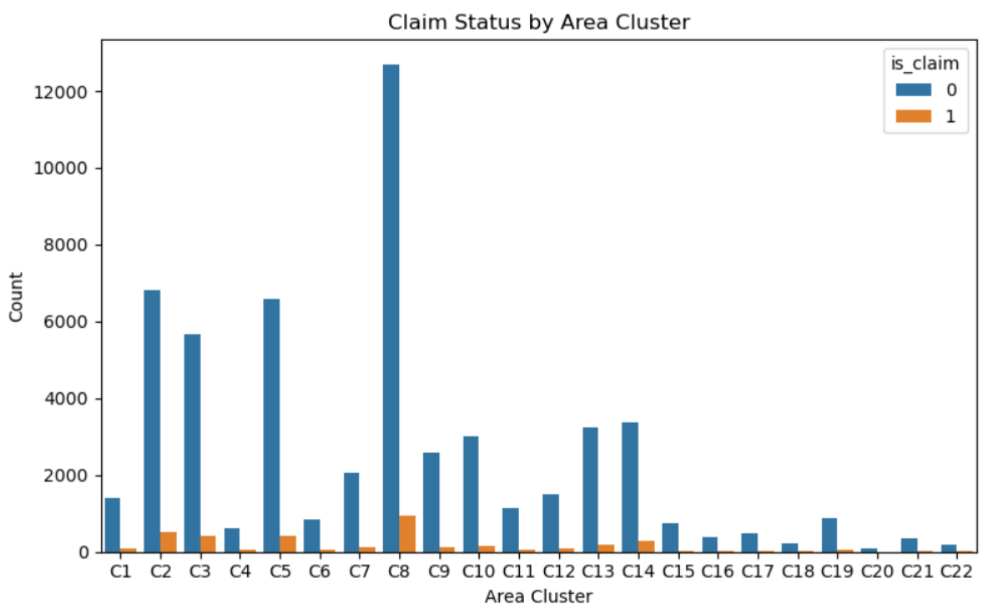


Figure 7: Claim Status by area cluster

The cluster of the areas are visualized based on claims where the highest claim is measured on cluster 8 and also claims which are not made are from the same cluster. This means the cluster contains maximum amount of customers and can be analyzed based on this plot.

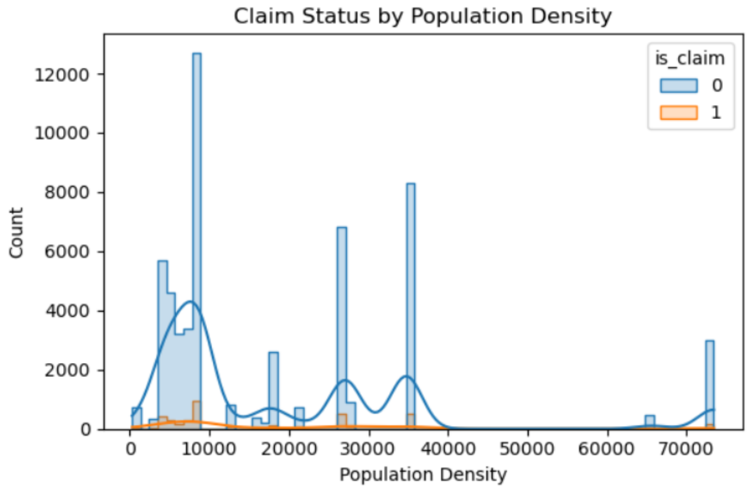


Figure 8: Population Density of claims

Population density ranging from 0 to 10000 are making highest amount of claims compared to higher densities. Claims which are basically made in areas having population of up to 10000.

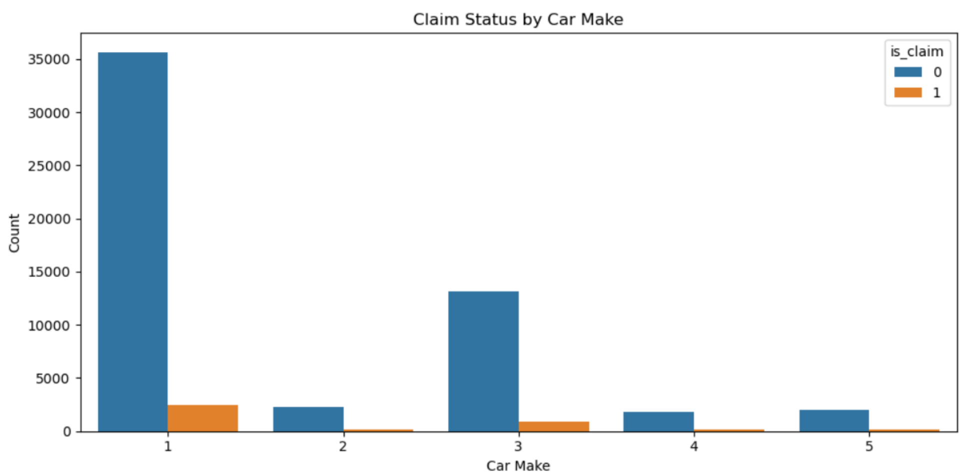


Figure 9: Claims based on make of car

From the plot, the number of customers or generally higher who bought cars of car make 1 and the claims made for the cars made in car make 1 is also higher. This plot can be useful to analyze the manufacturer of cars to determine the probability of Insurance claim.

- ## Performance of Models

The performance of the models are evaluated based on ROC curve, confusion Matrix as well as the F1 score from the classification report. F1-score gives the balance accuracy based on the false positive and false negative samples made by each model.

a. Random Forest Performance

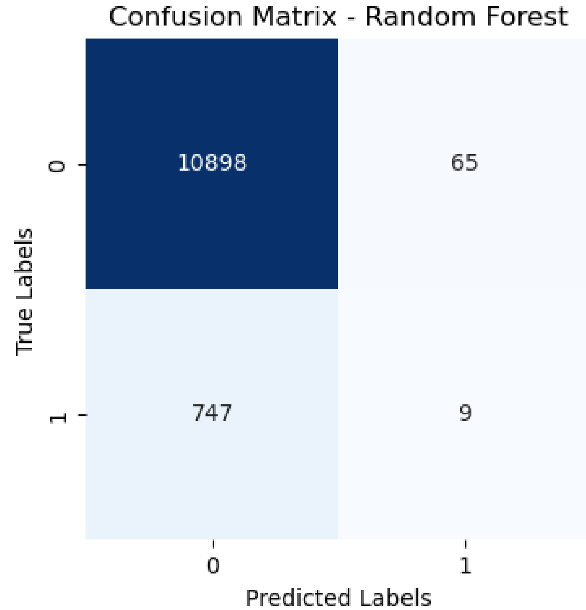


Figure 10: Confusion matrix of random forest

The confusion matrix on random forest indicates high mis classification in prediction of claimed samples where 747 samples are mis classified. Due to imbalanced samples present in the data, random forest failed to generalize on claimed samples.

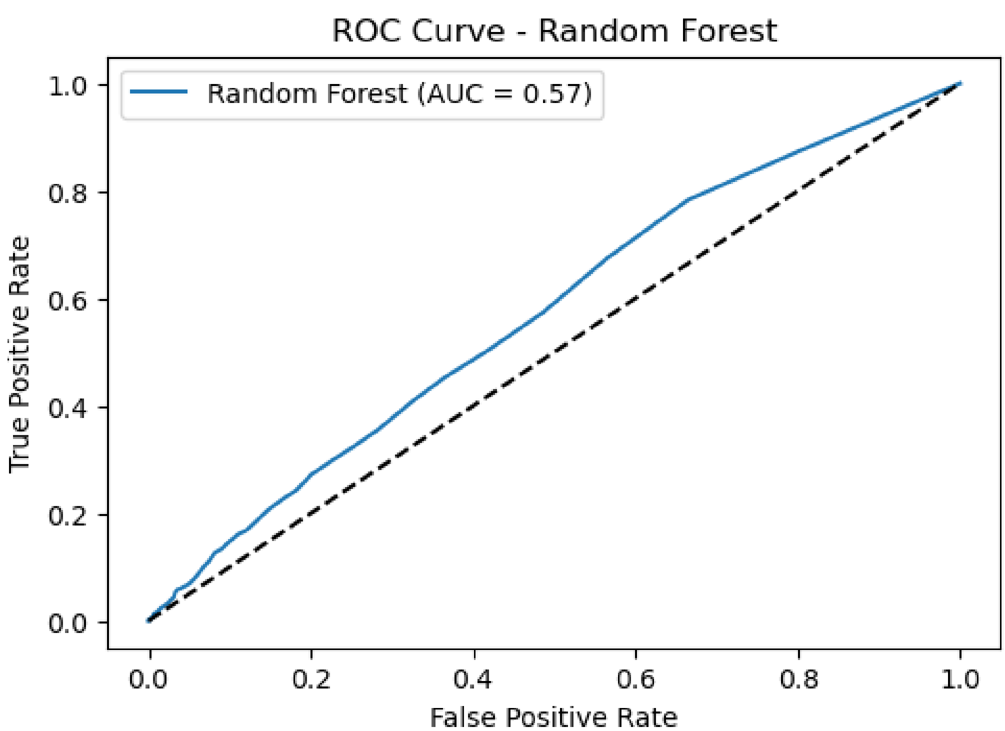


Figure 11: ROC curve of Random Forest

The ROC curve of the model is 0.57 which is also poor indicating lower number of true positive samples compared to the false positive samples.

```
Classification Report - Random Forest:
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     10963
           1       0.12      0.01      0.02       756

    accuracy                           0.93     11719
   macro avg       0.53      0.50      0.49     11719
weighted avg       0.88      0.93      0.90     11719
```

Figure 12: Classification report of random forest

Random forest produces lower prediction on samples with only 0.02 F1 score. The weighted average is 0.9 indicating model produces biased outcomes for claimed samples.

b. Gradient Boosting performance

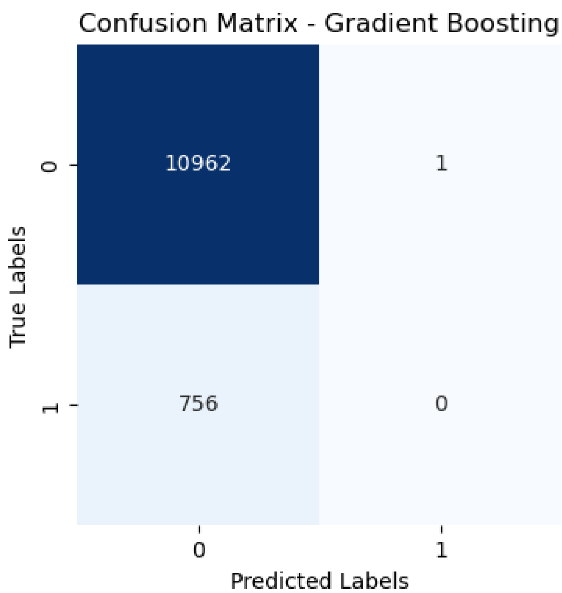


Figure 13: Confusion Matrix of Gradient Boosting

Gradient boosting performs worse than random forest in classifying claimed samples in validation data.

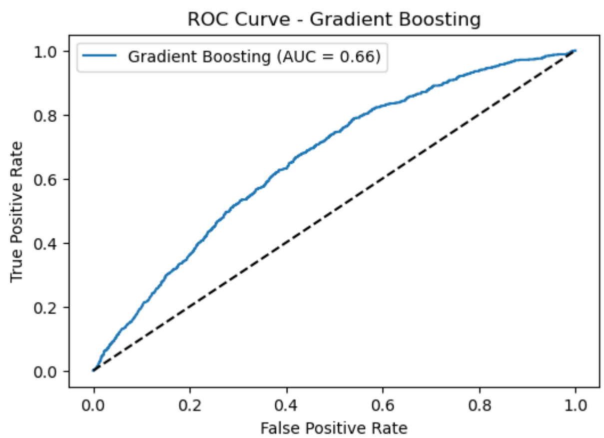


Figure 14: ROC curve of Gradient Boosting

The ROC curve indicates 0.66 which produces higher false positive temples compared to the true positive samples in the data.

Classification Report - Gradient Boosting:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	10963
1	0.00	0.00	0.00	756
accuracy			0.94	11719
macro avg	0.47	0.50	0.48	11719
weighted avg	0.88	0.94	0.90	11719

Figure 15: Classification report of gradient boosting

The classification report shows on overall F1 score of 0. 90 where the model gave poor F1 score in classifying claimed samples where 756 claim samples are fully mis classified as non claim samples by the model.

c. XGboost performance

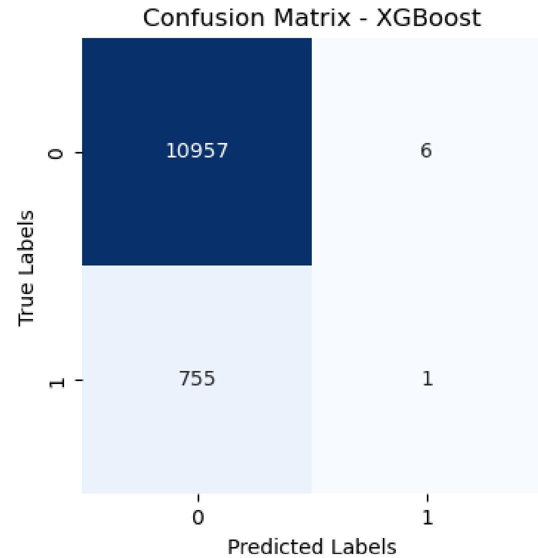


Figure 16: Confusion matrix of XGBoost model

The confusion matrix of XG boost also makes failed prediction in claimed samples where only 1 sample is correctly classified and all the other samples are misclassified by the model.

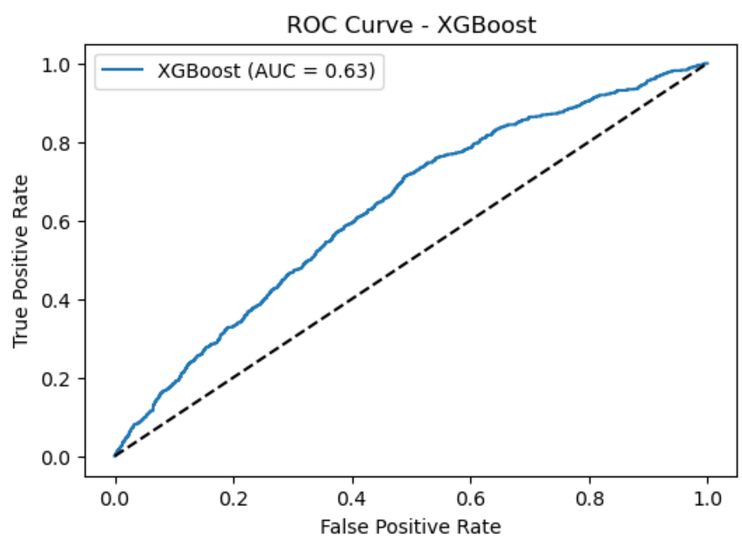


Figure 17: ROC curve of XGBoost model

The ROC curve of the model indicates a score of 0.63 indicating a number of false positive samples produced by the model compared to true positive samples.

Classification Report - XGBoost:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	10963
1	0.14	0.00	0.00	756
accuracy			0.94	11719
macro avg	0.54	0.50	0.48	11719
weighted avg	0.88	0.94	0.90	11719

Figure 18: Classification report of XG boost model

The overall F1-score is 0.9 given by the model which is similar to Gradient boosting model.

d. Light GBM Performance

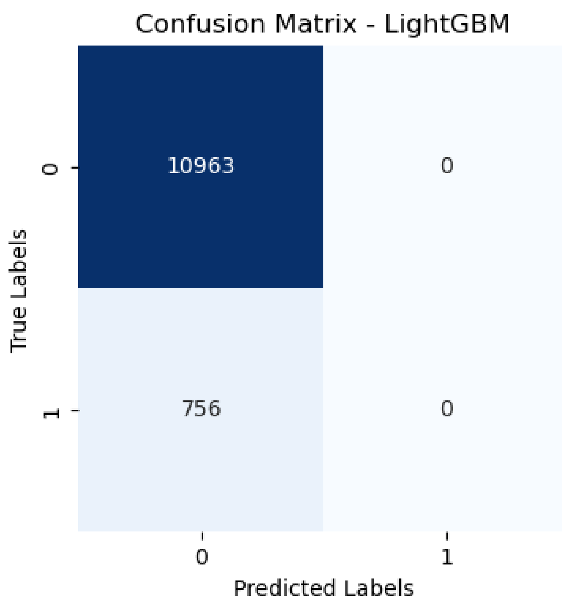


Figure 19: Confusion matrix of LightGBM method

Using light grading boosting method, the confusion Matrix shows that the model performs poorly in validation data where all the claimed samples are mis classified as non-claimed samples.

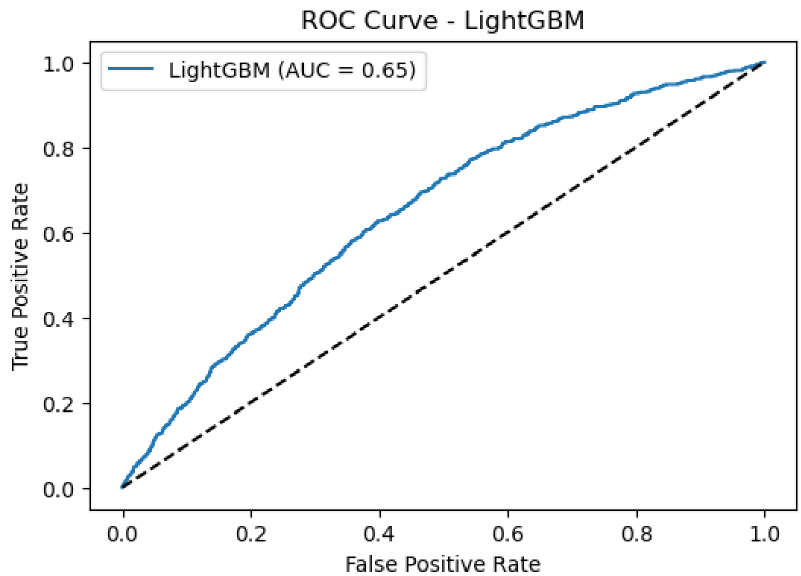


Figure 20: ROC curve of LightGBM method

The ROC curve gave an AUC score of 0.65 it is slightly lower than the XG boost model.

Classification Report - LightGBM:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	10963
1	0.00	0.00	0.00	756
accuracy			0.94	11719
macro avg	0.47	0.50	0.48	11719
weighted avg	0.88	0.94	0.90	11719

Figure 21: Classification report of Light GBM method

The classification report of the model shows poor accuracy like all the other models where they hugely suffered from imbalanced claimed samples.

- ## Performance on Oversampled data

We had applied SMOTE technique to increase the number of claimed samples in the data and perform predictive modelling to evaluate the performance.

SMOTE increases the minority class' representation in the dataset by creating synthetic samples for it. Finding the k-nearest neighbors of a sample from the minority class in the feature space is the first step in the process. Interpolation between the chosen sample and its neighbors is then used to construct synthetic samples. The synthetic samples are brand-new occurrences comparable to the minority class, but they are not exact duplicates of the observed data.

a. Random Forest Performance (SMOTE samples)

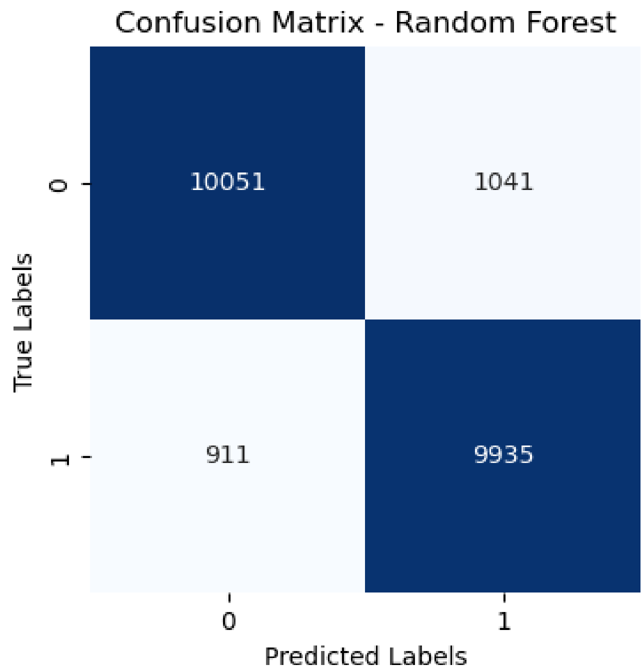
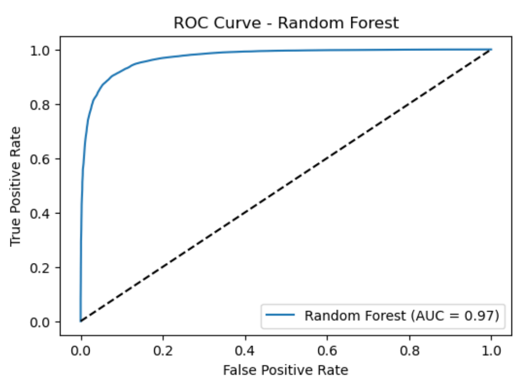


Figure 22: Confusion matrix of random forest on SMOTE samples

After oversampling of the claimed sample, random forest performed better where 911 synthetic claim sample are misclassified as non-claimed samples.




```
Classification Report - Random Forest:
              precision    recall  f1-score   support

           0       0.92      0.91      0.91     11092
           1       0.91      0.92      0.91     10846

    accuracy                           0.91     21938
   macro avg       0.91      0.91      0.91     21938
weighted avg       0.91      0.91      0.91     21938
```

Figure 23 : ROC curve and classification report of random forest on SMOTE Samples

The ROC curve shows AUC score of 0.97 indicating lower rate of false positive samples. The overall score is 0.91 showing random forest is capable to produce better results in balanced data.

b. Gradient boosting Performance (SMOTE samples)

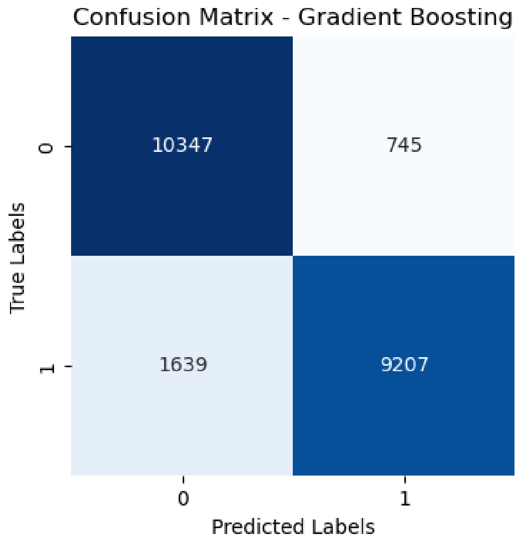
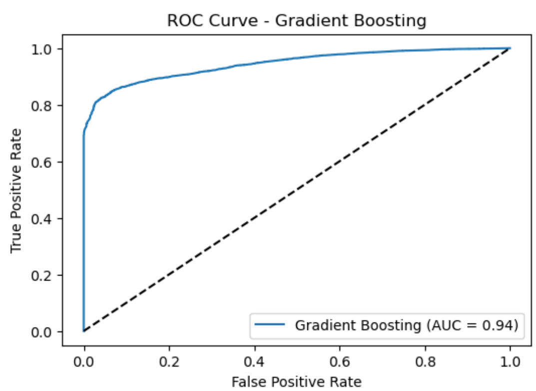


Figure 24: Confusion matrix of gradient boosting on SMOTE samples

Gradient Boosting produces high classification compared to random forest in oversampled data where around 1639 samples are misclassified on over sampled claimed data.




```
Classification Report - Gradient Boosting:
              precision    recall  f1-score   support

           0       0.86      0.93      0.90     11092
           1       0.93      0.85      0.89     10846

    accuracy                           0.89     21938
   macro avg       0.89      0.89      0.89     21938
weighted avg       0.89      0.89      0.89     21938
```

Figure 25: ROC curve and classification report of gradient boosting on SMOTE Samples

The ROC curve on gradient boosting on over sample data is giving a poor score compared to random forest and the classification report indicating an F1-score of 0.89 in prediction of claimed samples shows that Gradient boosting is not suitable for prediction of Insurance claim based on the variables present in our data.

c. XGboost Performance (SMOTE samples)

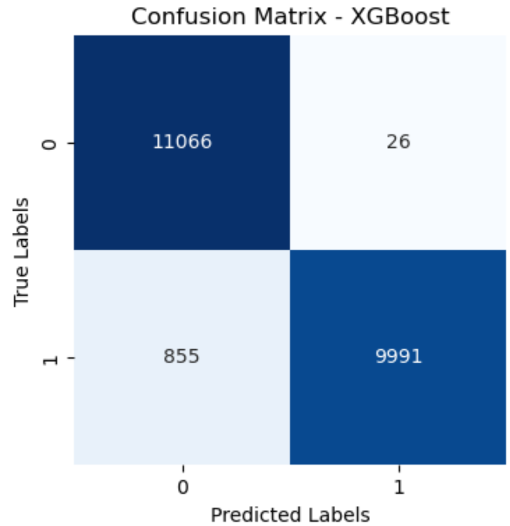
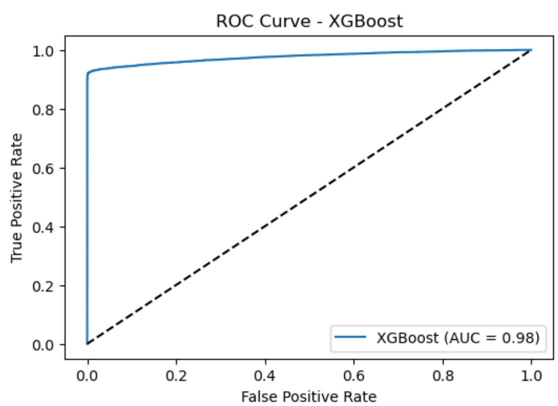


Figure 26: Confusion matrix of XG boost on SMOTE samples

The mis classification that given by the XG Boost model is better compared to random forest where it performs with high accuracy in over sample data. This indicates XGboost can perform better in prediction of insurance claim if the data is found to be balanced.



Classification Report - XGBoost:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	11092
1	1.00	0.92	0.96	10846
accuracy			0.96	21938
macro avg	0.96	0.96	0.96	21938
weighted avg	0.96	0.96	0.96	21938

Figure 27: ROC curve and classification report of XG boost on SMOTE Samples

The ROC curve gives a score of 0.98 indicating very lower amount of false positive samples given by the model. The classification report gives an overall F1 score of 0.96 showing XG boost is highly capable to predict insurance claims and it is much better performer in comparison to random forest or gradient boosting model.

d. Light Gradient boosting Performance (SMOTE samples)

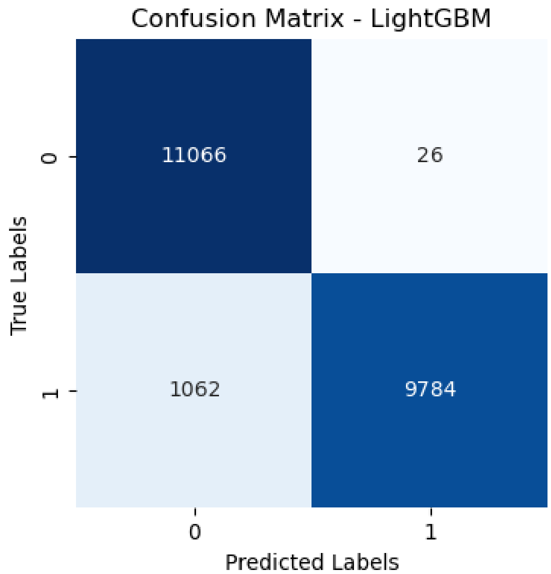
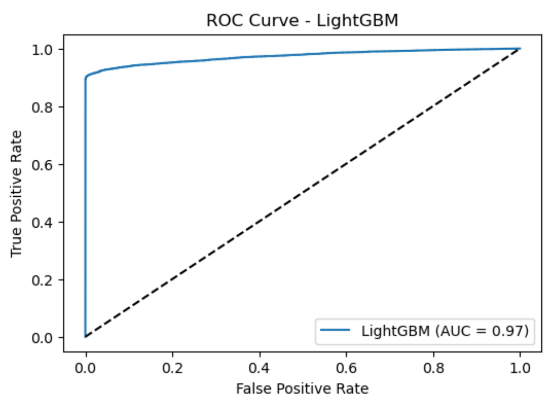


Figure 28: Confusion matrix of Light Gradient boost on SMOTE samples

From the confusion Matrix, light gradient boosting method is producing higher mis classification for claimed samples but produces better results in prediction of non claimed samples.




```
Classification Report - LightGBM:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95     11092
           1       1.00      0.90      0.95     10846

    accuracy                           0.95     21938
   macro avg       0.95      0.95      0.95     21938
weighted avg       0.95      0.95      0.95     21938
```

Figure 29: ROC curve and classification report of Light Gradient boost on SMOTE Samples

The ROC curve gave AUC score of 0.97 which shows lower false positive samples compared to the true positive samples and classification report gave an overall score of 0.95 indicating a better model than random forest but under performing compared to XG boost model.

- ## Discussion and Comparison

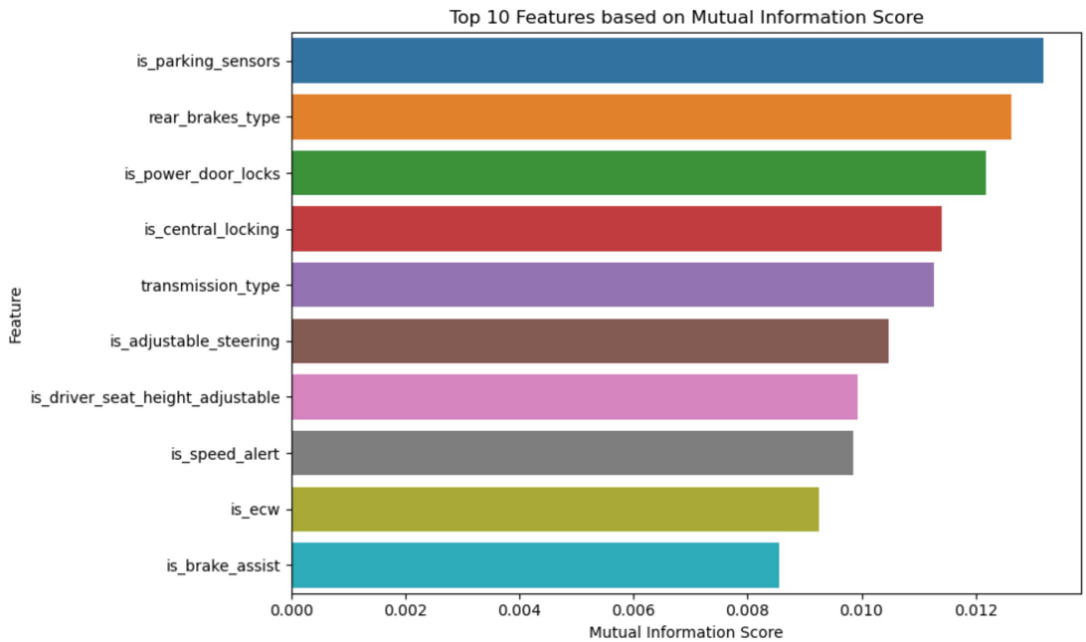


Figure 30: Feature Selection of the model

The most important features responsible for prediction of Insurance claim is determined based on the mutual information score where the highest score features is the one which is highly significant for Insurance claim prediction. The feature "**is_parking_sensors**", "**rear_brakes_type**", "**is_power_door_locks**" are some of the important features responsible for Insurance claim from the predictions made by different model. It is observed that none of the models really make good production predictions for insurance claim despite highly imbalanced samples.

After balancing the sample using **SMOTE** approach, XG boost produces better result compared to other models. XGBoost is capable to make higher insurance claim prediction based on these features on the data.

Conclusion

The challenge of predicting insurance claims using machine learning is intrinsically difficult, especially when dealing with imbalanced samples where the proportion of instances that are claimed is noticeably lower than the proportion of instances that are not claimed. The low representation of claimed samples made it difficult for conventional algorithms such Random Forest, XGBoost, Gradient Boost, as well as Light GBM to produce satisfying results in this situation.

We used the Synthetic Minority Over-sampling Technique (SMOTE) to address the problem of class imbalance and improve the accuracy of the models. By producing fake cases, SMOTE successfully raised the number of claimed samples, resulting in a more balanced dataset for training.

We saw significant increases in the XGBoost and other machine learning model performance after adding SMOTE. XGBoost outscored the other algorithms with an exceptional overall F1 score of 0.96, demonstrating its robustness in managing unbalanced datasets.

Our analysis shows that SMOTE was essential in reducing the issue of class imbalance, enabling the models to depict the patterns and quirks of the minority class more accurately (i.e., claimed occurrences). The more samples that were claimed, the better the knowledge the models had to generate correct predictions, which resulted in increased precision and recall rates, as shown by the higher F1 score.

But it's important to interpret these findings with care. Although SMOTE considerably increased model performance, it is important to keep in mind that oversampling methods like SMOTE may also increase the danger of overfitting or data redundancy. To ensure that the findings are generalizable, thorough model evaluation and validation are required.

References

https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification

Alamir, E., Urgessa, T., Hunegnaw, A. and Gopikrishna, T. (2021). Motor Insurance Claim Status Prediction using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)*, [online] 12(3). doi:https://doi.org/10.14569/IJACSA.2021.0120354.

Hah, D.W., Kim, Y.M. and Ahn, J.J. (2019). A study on KOSPI 200 direction forecasting using XGBoost model. *Journal of the Korean Data And Information Science Society*, 30(3), pp.655–669. doi:https://doi.org/10.7465/jkdi.2019.30.3.655.

Jing, L., Yong, Y., Ge, Hongmei, Li, Z. and Ruxue, G. (2020). Coal Rock Condition Detection Model Using Acoustic Emission and Light Gradient Boosting Machine. *Computers, Materials & Continua*, 62(3), pp.151–162. doi:https://doi.org/10.32604/cmc.2020.05649.

Körner, P., Kronenberg, R., Genzel, S. and Bernhofer, C. (2018). Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. *Meteorologische Zeitschrift*, 27(5), pp.369–376. doi:https://doi.org/10.1127/metz/2018/0908.

Mantas, C.J., Castellano, J.G., Moral-García, S. and Abellán, J. (2018). A comparison of random forest based algorithms: random credal random forest versus oblique random forest. *Soft Computing*. doi:https://doi.org/10.1007/s00500-018-3628-5.

Roy, R. and George, K.T. (2017). Detecting insurance claims fraud using machine learning techniques. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*. doi:https://doi.org/10.1109/iccpct.2017.8074258.

Singh, R., Ayyar, M.P., Sri Pavan, T.V., Gosain, S. and Shah, R.R. (2019). *Automating Car Insurance Claims Using Deep Learning Techniques*. [online] IEEE Xplore. doi:https://doi.org/10.1109/BigMM.2019.00-25.