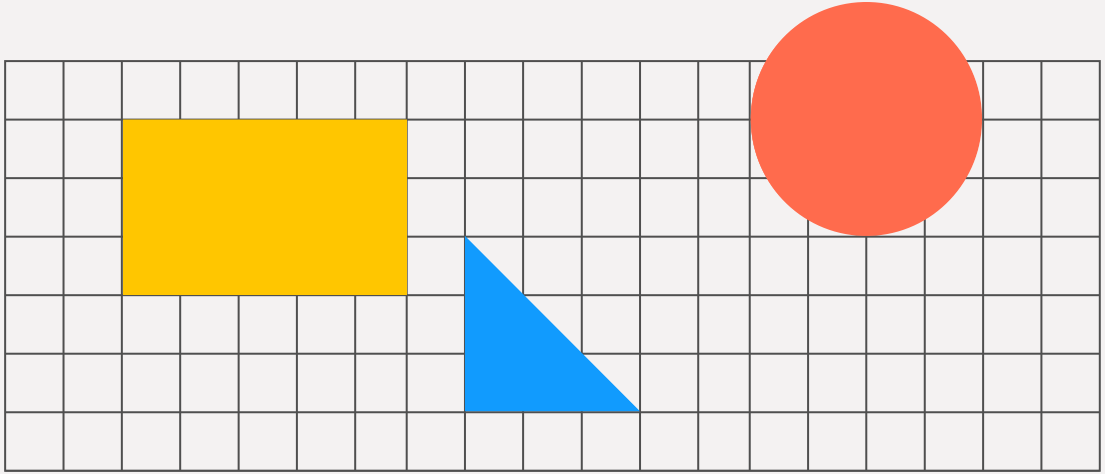
▶ **UpGrad Assignment** April- 2025

Fraudulent Claim Detection

Submitted by:
Kante Shiva Chandra
Krishkumar V



PROBLEM STATEMENT

Outline

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process. This would minimise financial losses and optimise the overall claims handling process.

Business Use Case

Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.

Based on this assignment, you have to answer the following questions

- How can we analyse historical claim data to detect patterns that indicate fraudulent claims?
- Which features are most predictive of fraudulent behaviour?
- Can we predict the likelihood of fraud for an incoming claim, based on past data?
- What insights can be drawn from the model that can help in improving the fraud detection process?

APPROACH

1. Data Preparation
2. Data Cleaning
3. Train-Validation Split
4. EDA on Training Data
5. Feature Engineering
6. Model Building
7. Prediction and Model Evaluation

1. Data Preparation

There are few '?' entries in the data, which are address by replacing them with the NULL Values.

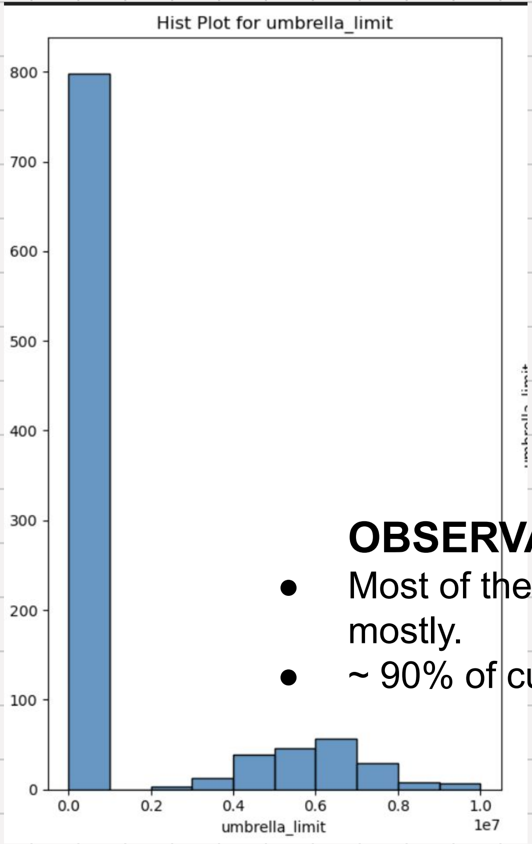
2. Data Cleaning

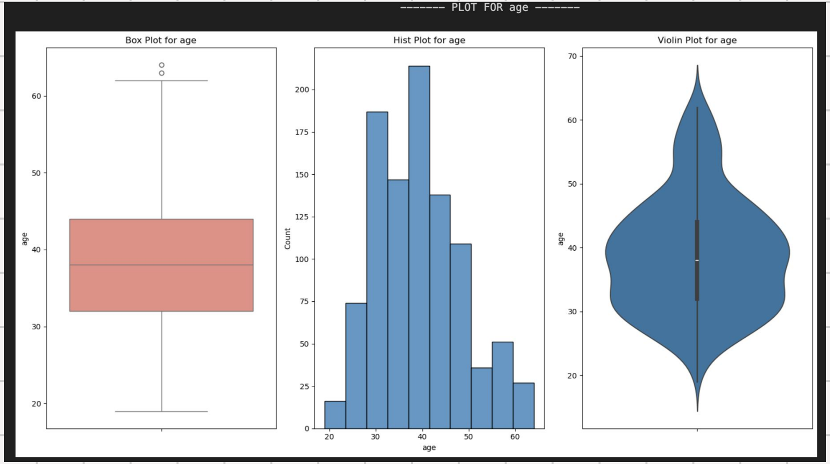
- Handled null values in the feature property_damage, police_report_available, collision_type with most frequently repeated value i.e. mode.
- We can see a negative value in the `umbrella_limit` feature. So we dropped that row from the dataset.

3. Train Test Validation Split

- Divided the data into 70%-30% split using the starify method as the data is imbalance.

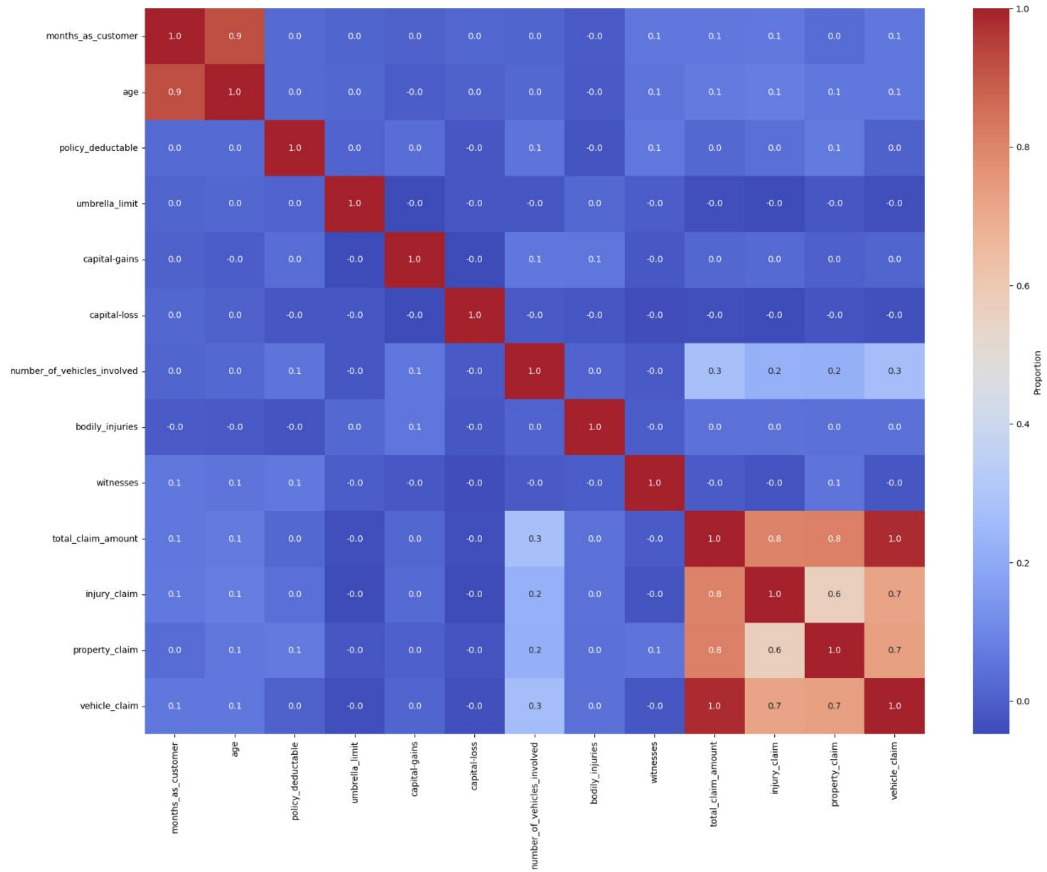
4. EDA on Training Data





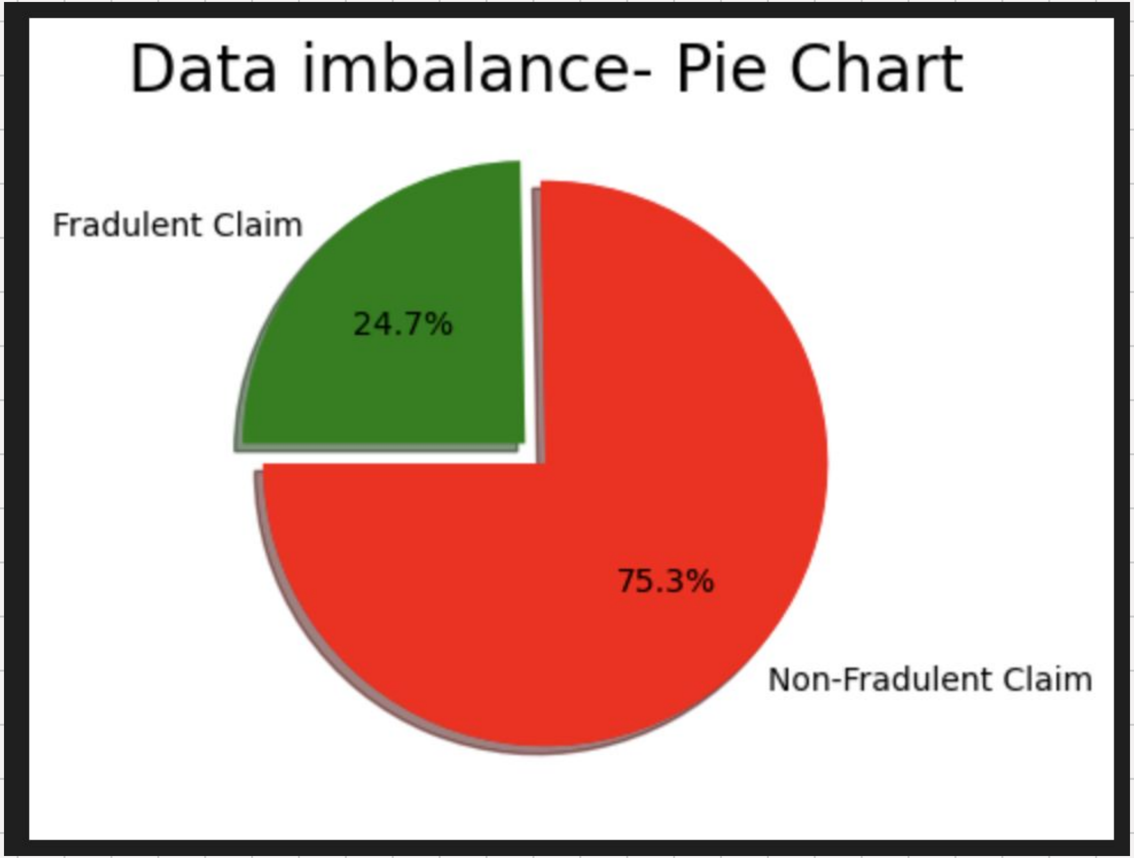
OBSERVATIONS

- Most of the customer who took insurance are between 30's and 50's as they use the cars mostly.
- ~ 90% of customers didn't took any extra capping limit on the insurance.



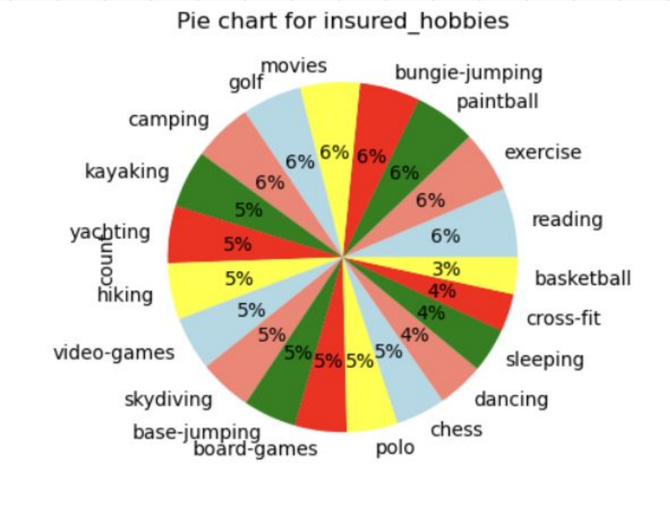
OBSERVATIONS

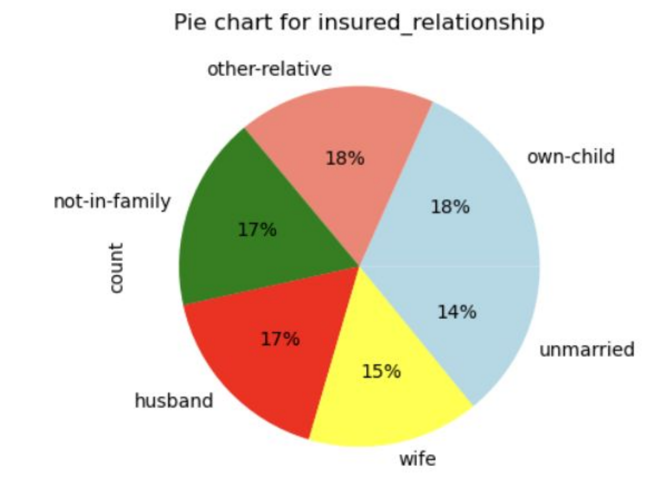
- Age and months_as_customer are highly correlated. we dropped Age column for the better model prediction.
- We saw there is a high correlation between total_claim_amount with the injury_claim, property_claim & vehicle_claim. So we dropped the total_claim_amount.

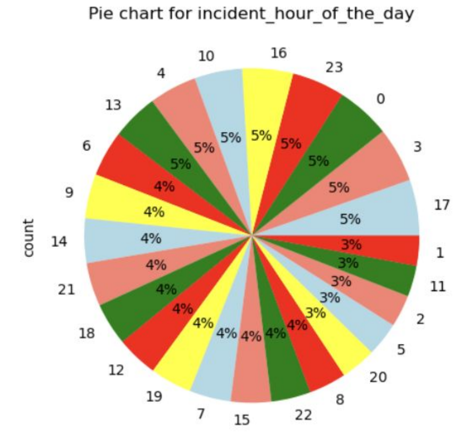


OBSERVATIONS

- The fraudulent rate is 24.7%, meaning only 24.7% of claims raised are turned out to be fraudulent (minority), while 75.3% are not fraudulent claims (majority). This indicates a class imbalance in the data.

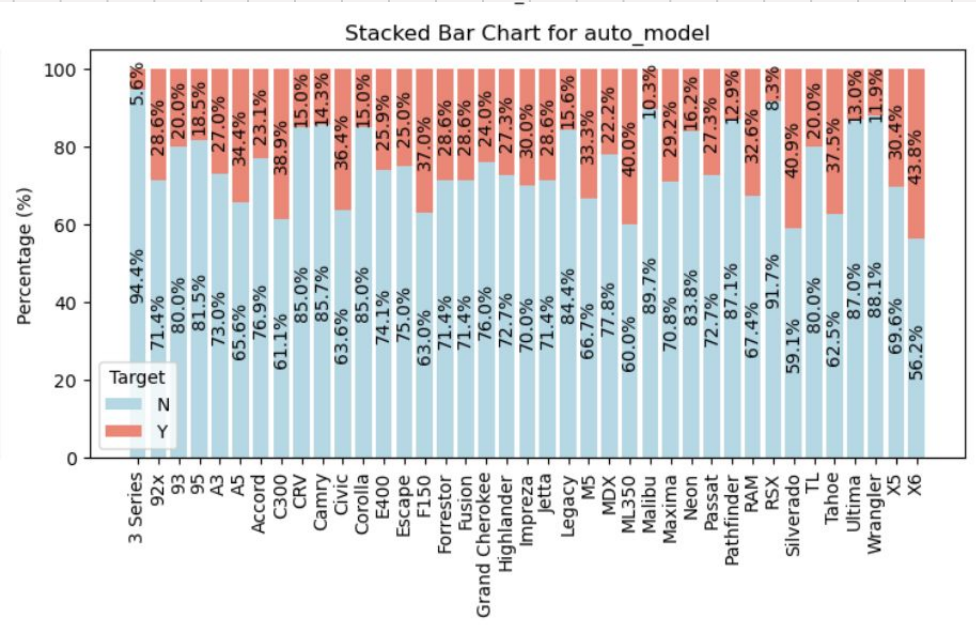


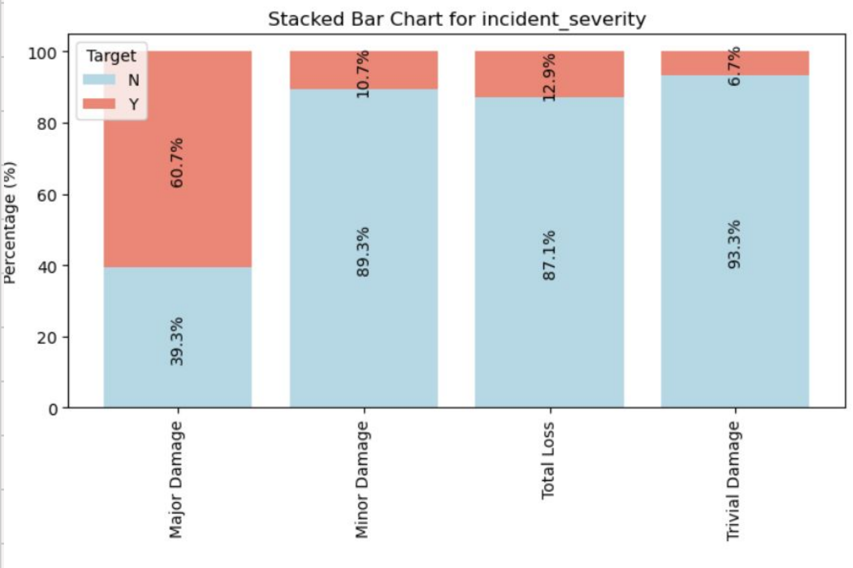




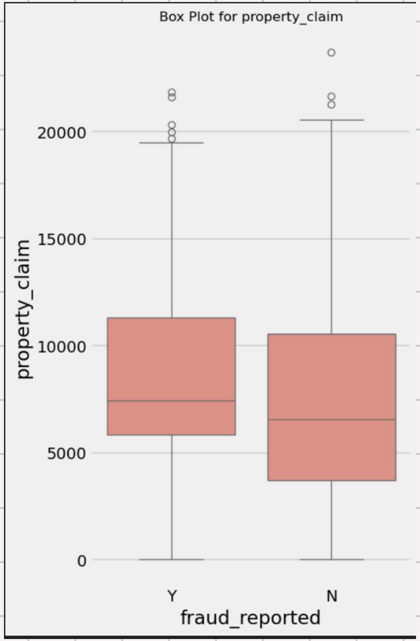
OBSERVATIONS

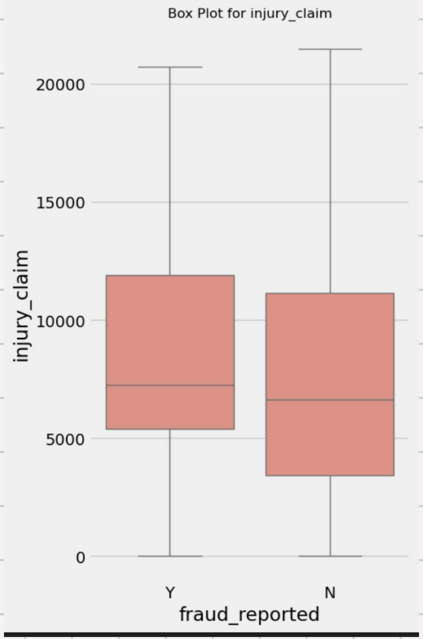
- The data is evenly distributed across the columns: auto_year, auto_make, incident_hour_of_the_day, insured_hobbies, and insured_occupation.

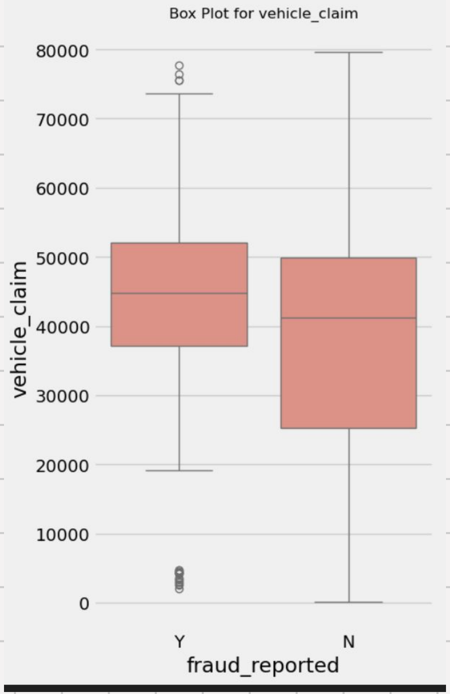




- Certain auto models such as ML350, Silverado, and X6 show a higher proportion of fraudulent claims.
- In cases of Major Damage, around 60% of the claims turn out to be fraudulent.







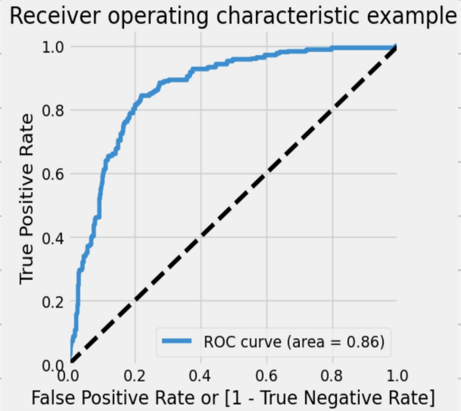
- The Median amount for vehicle claim, property claim, injury claim and the combined filed total claim is higher for the Fraudulent claims compare with non-fraudulent claims

4. Feature Engineering

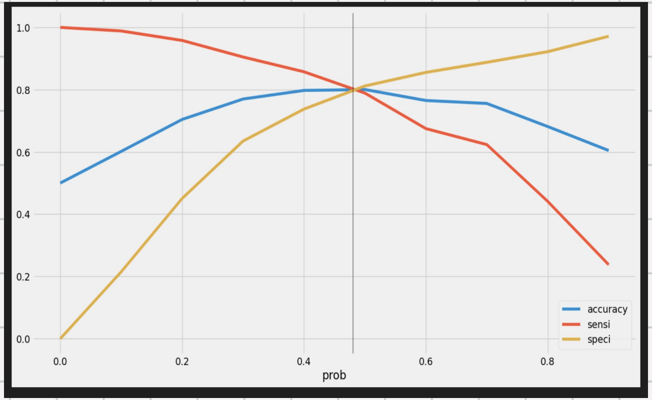
- Dropped auto_year as we extracted vehicle_age column out of it.
- Dropped incident_hour_of_the_day as we extracted Incident_period_of_day column out of it.
- Dropped umbrella_limit as we extracted Is_additional_umbrella_present column out of it.
- Dropped auto_model as we have lot of unique values with each of them in small chunks we bucketed them on the auto_make into **Luxury Cars**, **Mainstream**, **Niche Cars.**
- Dropped insured_hobbies can bucketed them into **Adventure**, **Fitness** & **Indoor** Categories.

5.1 Logistic Regression Model Build

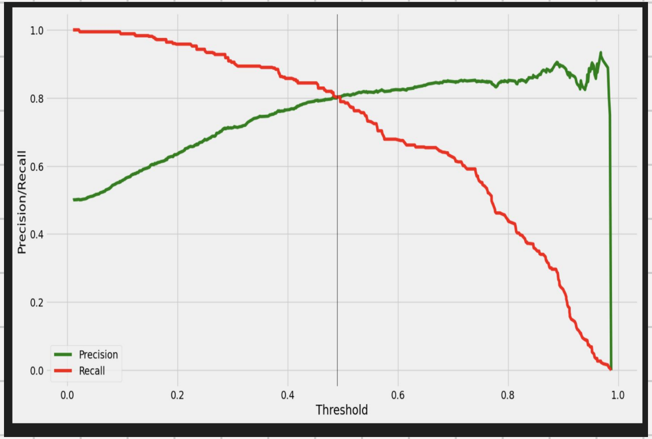
ROC Curve



Accuracy, Sensitivity & Specificity



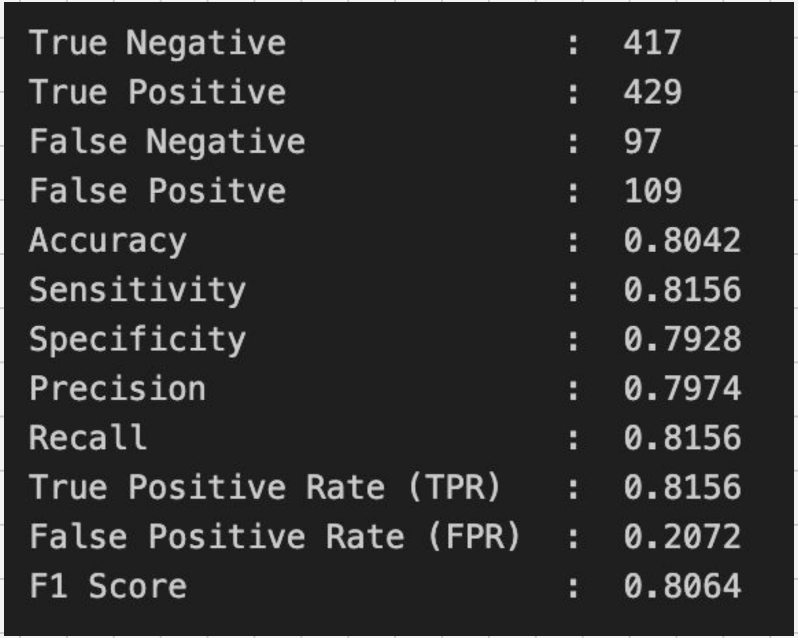
Precision & Recall



- For ROC in train data, Area under ROC curve is 0.86 out of 1 which indicates a good predictive model.
- 0.48 is the approx point where all the curves meet, so 0.48 seems to be our Optimal cutoff point for probability threshold using Accuracy, Sensitivity and specificity.
- The intersection point of the curve represents the threshold where the model balances precision and recall. This value helps optimize model performance based on business requirements. From the curve above, the optimal probability threshold is approximately 0.49.

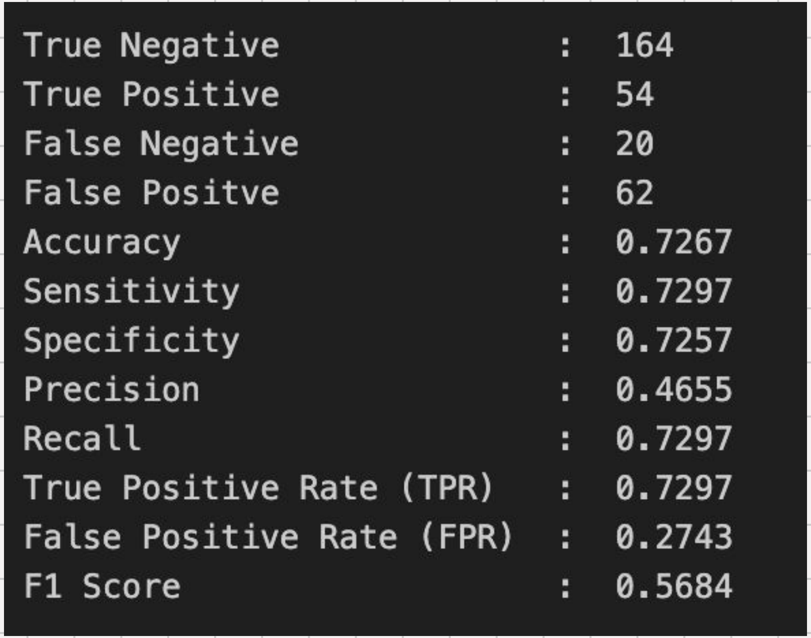
5.1 LR Model Evaluation

Training Results



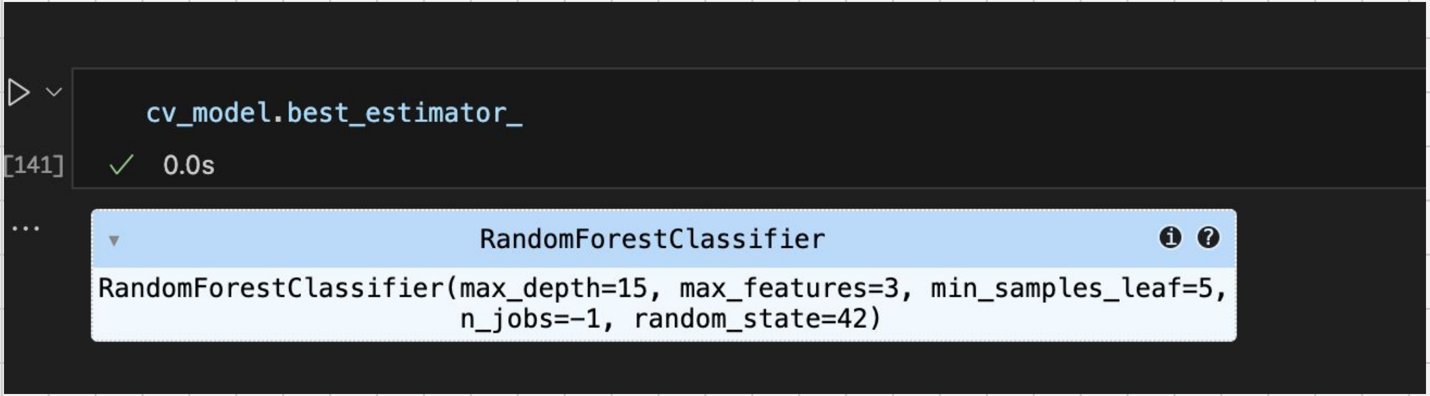
```
True Negative              :  417
True Positive              :  429
False Negative             :  97
False Positve              :  109
Accuracy                   :  0.8042
Sensitivity                :  0.8156
Specificity                :  0.7928
Precision                  :  0.7974
Recall                     :  0.8156
True Positive Rate (TPR)   :  0.8156
False Positive Rate (FPR)  :  0.2072
F1 Score                   :  0.8064
```

Test Results



```
True Negative              :  164
True Positive              :  54
False Negative             :  20
False Positve              :  62
Accuracy                   :  0.7267
Sensitivity                :  0.7297
Specificity                :  0.7257
Precision                  :  0.4655
Recall                     :  0.7297
True Positive Rate (TPR)   :  0.7297
False Positive Rate (FPR)  :  0.2743
F1 Score                   :  0.5684
```

5.2 Random Forest Model Build



```
cv_model.best_estimator_
```

[141] ✓ 0.0s

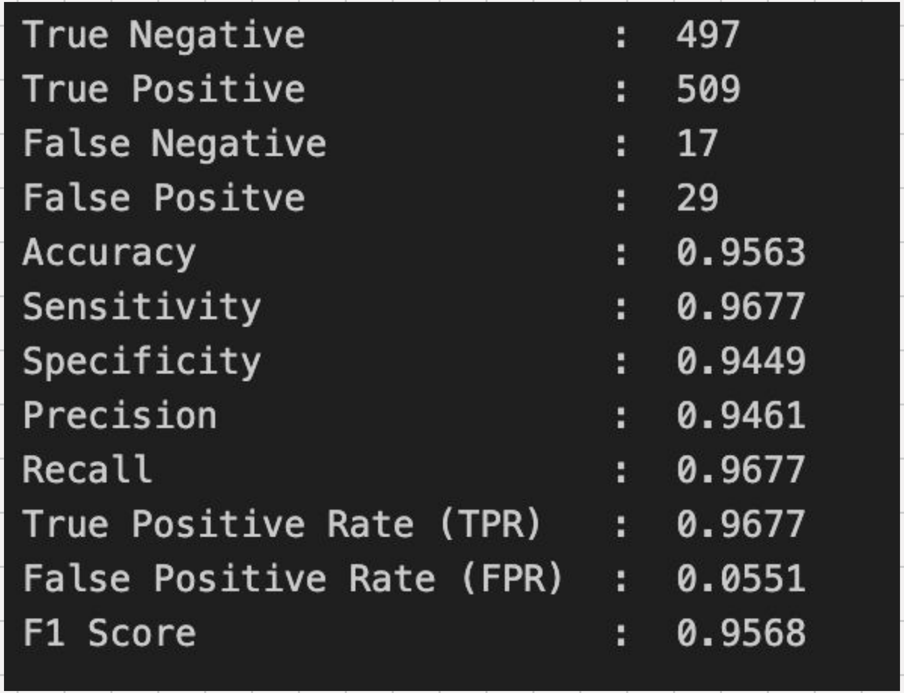
RandomForestClassifier

```
RandomForestClassifier(max_depth=15, max_features=3, min_samples_leaf=5,
                       n_jobs=-1, random_state=42)
```

- The Best Random forest model using the Grid Search CV is with max depth of 15, 3 max features with 5 min sample leafs.

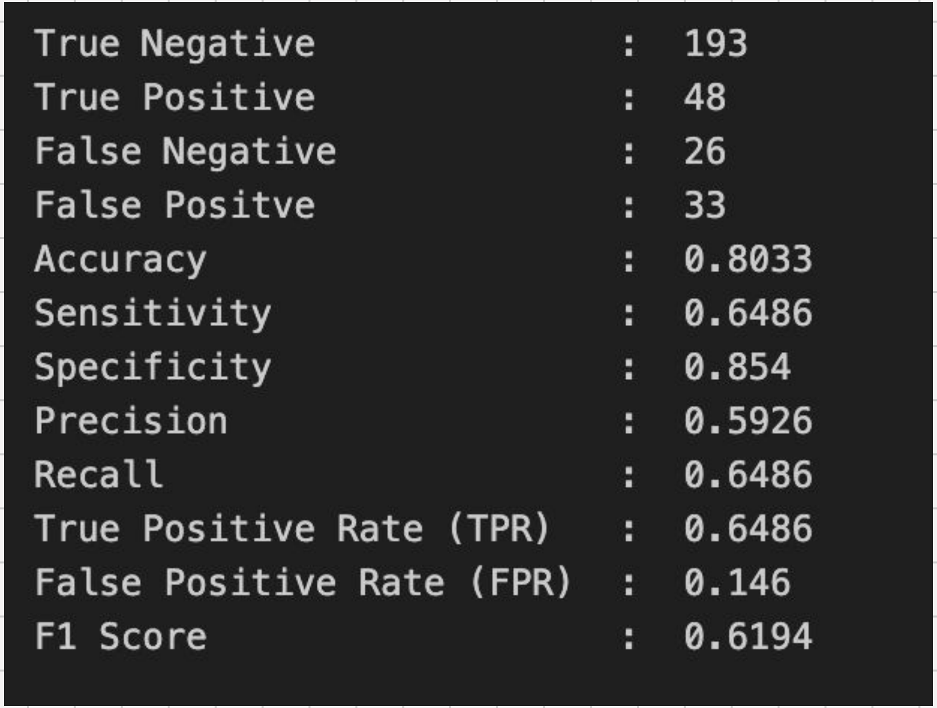
5.1 Random Forest Model Evaluation

Training Results



```
True Negative              :  497
True Positive              :  509
False Negative             :  17
False Positve              :  29
Accuracy                   :  0.9563
Sensitivity                :  0.9677
Specificity                :  0.9449
Precision                  :  0.9461
Recall                     :  0.9677
True Positive Rate (TPR)   :  0.9677
False Positive Rate (FPR)  :  0.0551
F1 Score                   :  0.9568
```

Test Results



```
True Negative              :  193
True Positive              :  48
False Negative             :  26
False Positve              :  33
Accuracy                   :  0.8033
Sensitivity                :  0.6486
Specificity                :  0.854
Precision                  :  0.5926
Recall                     :  0.6486
True Positive Rate (TPR)   :  0.6486
False Positive Rate (FPR)  :  0.146
F1 Score                   :  0.6194
```

Evaluation & Conclusion

1. **Identify Suspicious Claim Patterns**
   Analyze features like `vehicle_claim`, `property_claim`, and `injury_claim` — unusually high amounts or frequent claims may indicate fraud.
2. **Watch for Red Flags in Incident Details**
   Claims marked as `Total Loss` or `Minor Damage`, especially without witnesses, are more likely to be fraudulent.
3. **Spot High-Risk Customer Profiles**
   Short customer tenure, specific hobbies (`Mental/Indoor`, `Fitness/Active`), and certain vehicle brands are often linked to fraud.
4. **Leverage Predictive Features**
   Key features like `capital-gains`, `capital-loss`, `policy_deductable`, and `policy_csl_250/500` have strong influence on fraud prediction models.
5. **Use ML Models to Score New Claims**
   A trained Random Forest model can assign fraud probability scores to new claims, flagging high-risk cases based on patterns in historical data.
6. **Strengthen Internal Checks**
   Apply stricter reviews for high-value claims with Total Loss/Minor Damage tags, especially if witnesses are missing or claim patterns look suspicious.
7. **Enable Early Detection with Insights**
   Use behavioral indicators like hobbies and vehicle type, along with policy structure, to trigger early fraud alerts.

Thank You