Fraudulent Claim Detection

Case Study Report

Ameya Parab; Amar Nath Kumar

4/23/2025

This report contains the overall approach of the assignment, covering the problem statement, methodology, techniques used and key insights

Problem Statement:

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 minimize financial losses and optimize the overall claims handling process.

Data Preparation and Cleaning

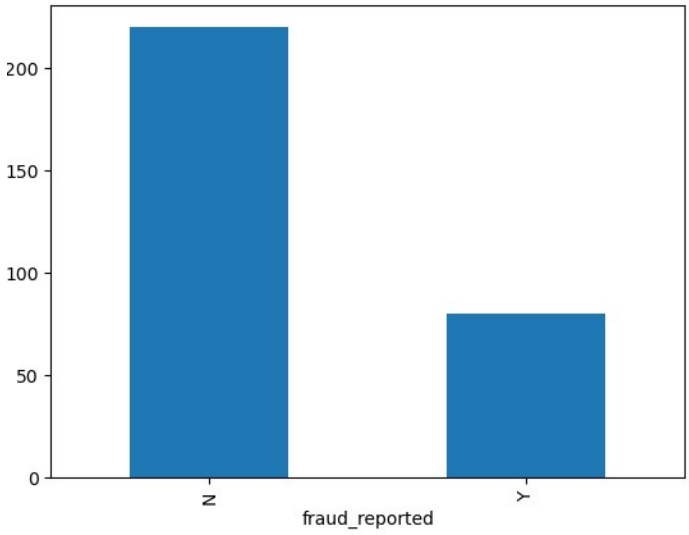
- Empty feature like "_c39" was removed.
- Value "None" in the feature "authorities_contacted" was considered as a NULL value by program but it is actually a very important piece of information. It is expected from the user to contact authorities in case of any damage or theft. Hence, "None" was replaced by "None Contacted" so that it will be used for analysis.
- Features like "auto_model" and "auto_year" are dropped from the analysis. "auto_make" feature covers "auto_model" and "auto_year" seems to be irrelevant

Splitting Data for Training and Testing

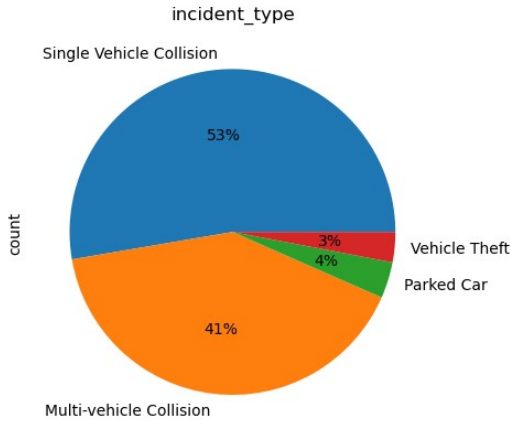
70% Data is reserved for training and 30% data is reserved for testing.

Exploratory Data Analysis

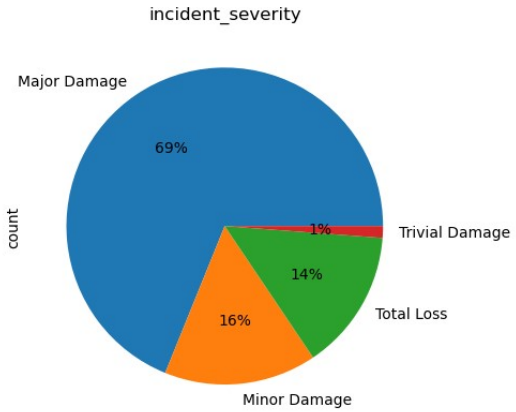
- No outliers were found in the univariate analysis of numeric features of both train and test data
- After performing correlation analysis of the numeric variables of both train and test data, it was found that features "months_as_customer" and "age" highly correlated. Same can be said for features "total_claim_amount", "injury_claim", "property_claim" and "vehicle_claim"
- Class imbalance was observed for a target variable "fraud_reported"



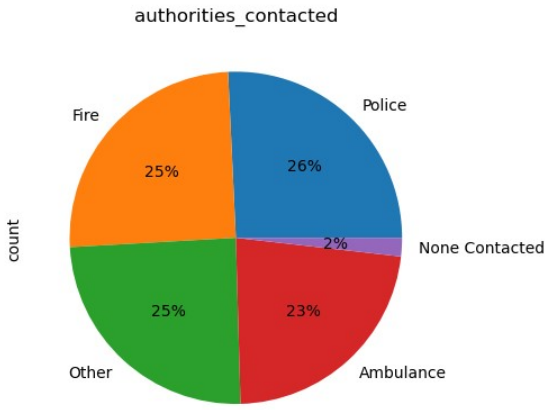
- Analysis is carried for categorical data to see their contribution to the fraudulent cases.
- Vehicle collision cases seem to contribute heavily in fraudulent claims.



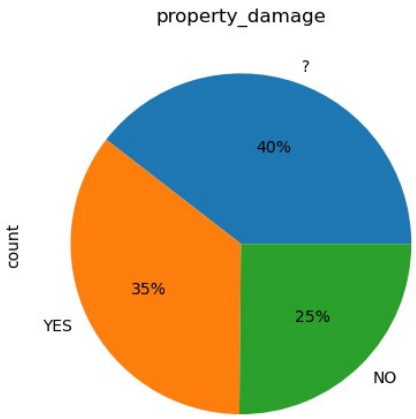
- Cases of Major Damage and Total Loss seem to contribute heavily in fraudulent claims.



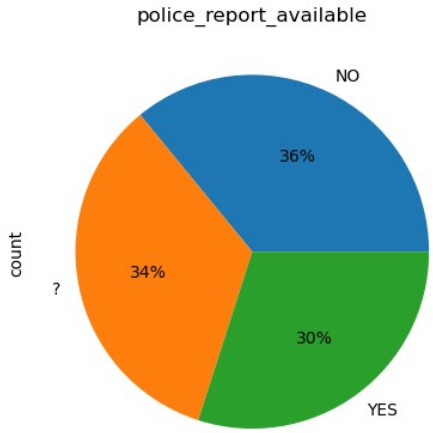
- Cases in which no authority was contacted seem to have little contribution in the fraudulent cases.



- Cases in which Property damage is reported seem to contribute less in fraudulent claims



- Cases in which Police report is available seem to contribute less in fraudulent claims



- Feature analysis is observed to be similar for both training and test data

Feature Engineering

- Oversampling technique is used to address the class imbalance issue for target variable.

```
Before oversampling
 fraud_reported
N                533
Y                167
Name: count, dtype: int64


After oversampling
 fraud_reported
N                533
Y                533
Name: count, dtype: int64
```

- A new feature “capital_gain_loss” is created by summing up “capital-loss” and “capital-gains”.
- Following features were removed due to multicollinearity issue and lack of relevance.
    1. policy_number
    2. incident_date
    3. age
    4. injury_claim

5. property_claim
6. vehicle_claim
7. policy_bind_date
8. insured_education_level
9. insured_occupation
10. insured_hobbies
11. insured_relationship
12. incident_location
13. incident_state
14. capital-loss
15. capital-gains

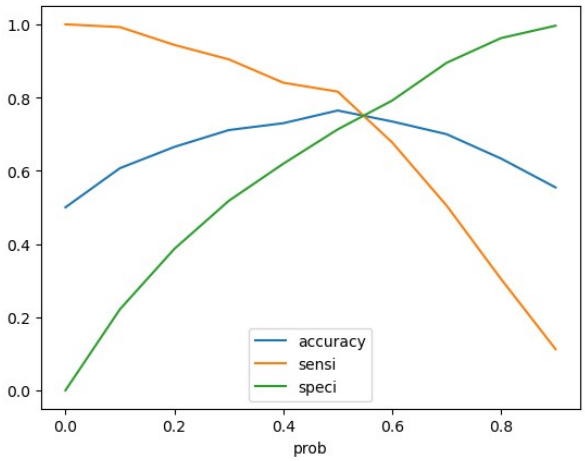
- For a feature “incident_severity”, category “Trivial Damage” is combined with “Minor Damage” and “Total Loss” is combined with “Major Damage”
- Categories were renamed so that dummies will not have duplicates.
- Dummy variables were created for test and train data set.
- Numeric features were scaled using MinMax scaling method

Model building and evaluation

Logistic Regression and Random Forest methods were used for predictions

Logistic Regression:

- RFECV method is applied to identify the most relevant features
- VIF of features was assessed to check the multicollinearity after building the model. More features were dropped because of high VIFs
- Model was initially built with cut off value of 0.5 but the optimum cut off value was found out to be 0.55 based on the below plot.

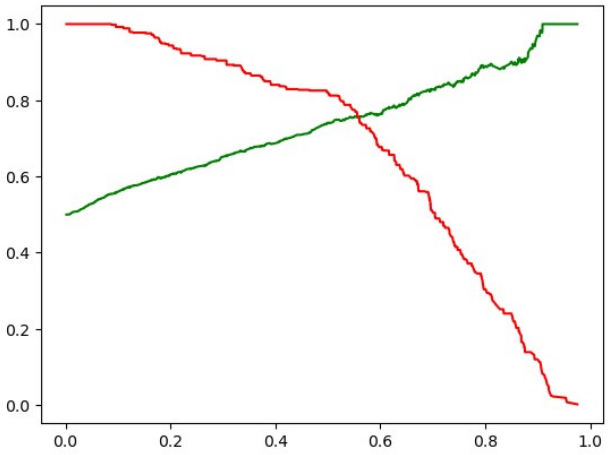


- Prediction were made on the Training data with cut-off 0.55 and following metric
    - Training Accuracy = 0.76

- Training Sensitivity, Specificity, Precision, Recall and F1-score

```
Sensitivity =  0.8161350844277674

Specificity =  0.7129455909943715

Precision =  0.7397959183673469

Recall =  0.8161350844277674

F1_score =  0.776092774308653
```

- Cut off value is adjusted to 0.58 based on below Precision-Recall curve



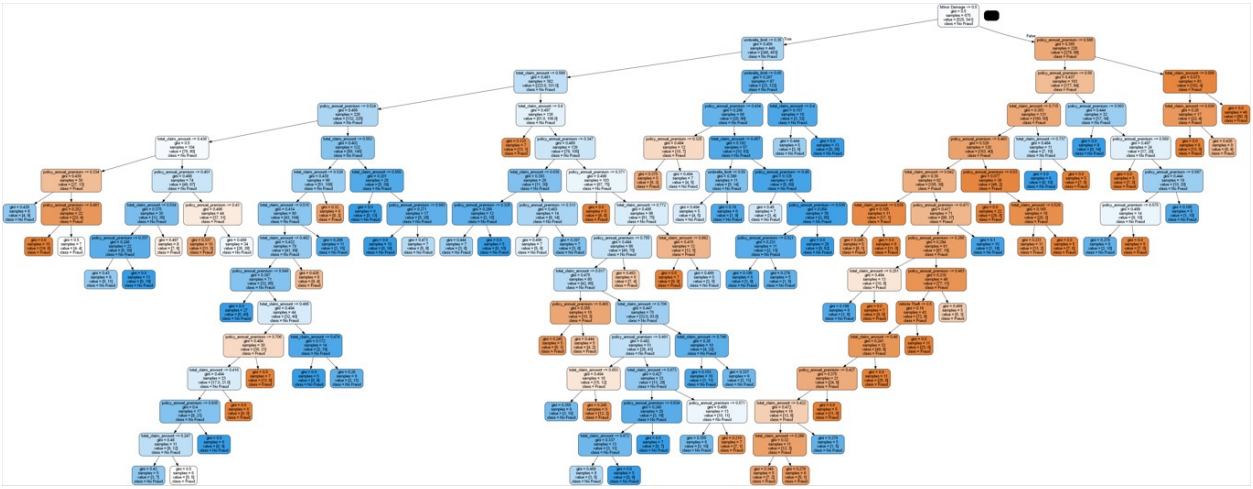
- Predictions were made on the test data with cutoff value of 0.58 with following metric
  - Testing Accuracy = 0.65
  - Testing Sensitivity, Specificity, Precision, Recall and F1-score

```
Sensitivity =  0.325

Specificity =  0.7772727272727272

Precision =  0.3466666666666667

Recall =  0.325

F1_score =  0.335483870967742
```

Random Forest:

- Base Random forest model was built with following high importance features
  - Minor Damage
  - Vehicle Theft
  - None Contacted
  - total_claim_amount
  - policy_annual_premium
  - umbrella_limit

- Following is the sample tree after hyper parameter tuning



Vehicle damage condition is the most important feature

- Predictions were made on Training data with following metric
    - Training Accuracy = 0.9
    - Training Sensitivity, Specificity, Precision, Recall and F1-score

```
Sensitivity =  0.9512195121951219

Specificity =  0.8592870544090057

Precision =  0.8711340206185567

Recall =  0.9512195121951219

F1_score =  0.9094170403587444
```

- Predictions were made on the test data with following metric
    - Testing Accuracy = 0.85
    - Testing Sensitivity, Specificity, Precision, Recall and F1-score

```
Sensitivity =  0.5

Specificity =  0.9727272727272728

Precision =  0.8695652173913043

Recall =  0.5

F1_score =  0.634920634920635
```

Conclusion:

- Random Forest model has performed better than Logistic Regression in terms of Accuracy, Sensitivity and Specificity and this model will be used for further analysis
- Sensitivity is the critical parameter as detecting fraud is very important.
- Though specificity is very high for Random forest, sensitivity is 0.5 which is not the in the comfortable zone
- There is scope of improvement for sensitivity by more feature engineering.