Predicting Fraudulent Car Insurance Claims

By, Jenny Robbins

Problem Statement

- Car insurance fraud can be difficult to detect due to the wide range of fraudulent activities that can occur and because they are far less common than legitimate claims.
- A predictive model for determining if claims are fraudulent will give insurance companies more confidence in pursuing a challenge to the claim and hopefully preventing more loss from these individuals in the future.

1. Data Wrangling

The data was obtained from kaggle via a scientific paper “Minority report in fraud detection: classification of skewed data”, which provided all necessary data for this project.

➔ **Data Collection**

A suitable dataset was found.

➔ **Data Visualization**

The dataset was looked at to see what steps were necessary for cleaning.

➔ **Data Cleaning**

Missing data was dealt with and the data was prepared for data exploration.

2. EDA

Visualization of relationships in the dataset.

- This resulted in further exploration into the 'Age' and 'ClaimSize' columns due to their distributions.
- Non-numerical columns were visualized through tables identifying fraud rate of each feature.
- Of the 11,565 rows, only 685 of them represented claims that were found as fraudulent.
- 5.9% of the claims were fraudulent, while 94.1% were non-fraudulent.

Fig 1. Age Distribution

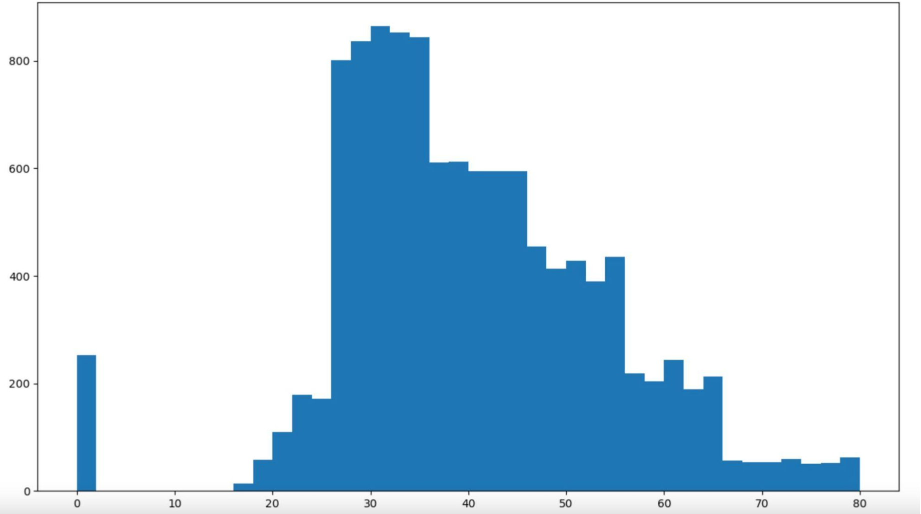


Fig 2. Age Distribution for fraud and no fraud claims

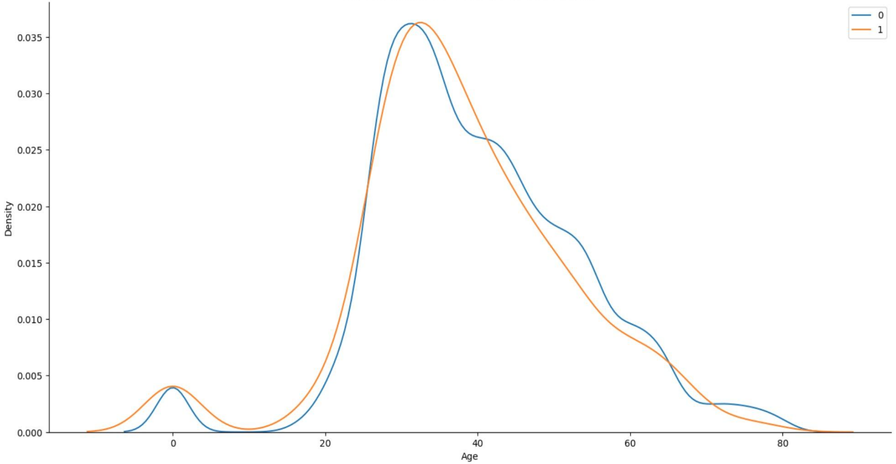


Fig 3. ClaimSize Distribution

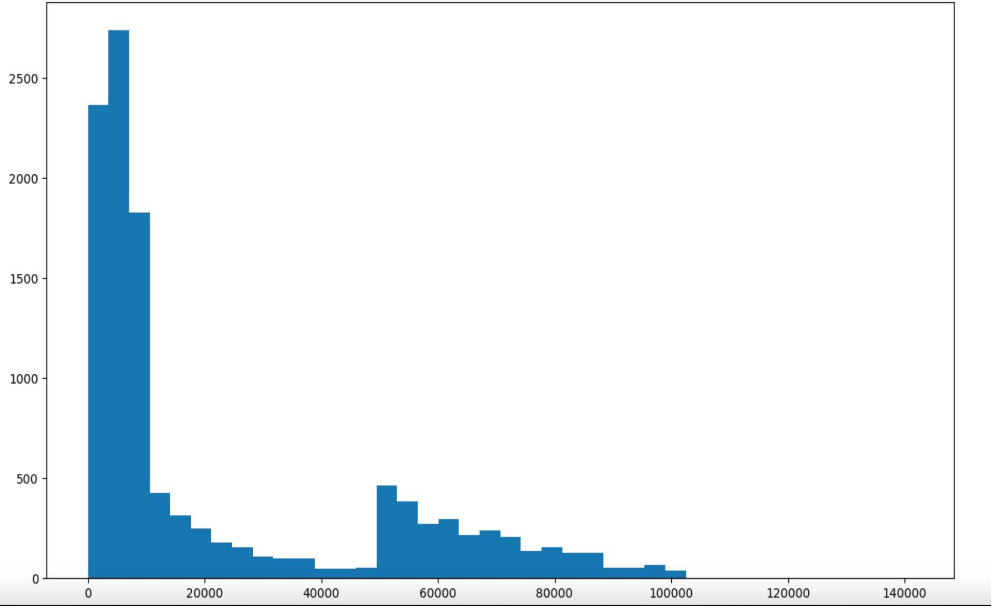


Fig 4. ClaimSize Distribution for fraud and no fraud claims

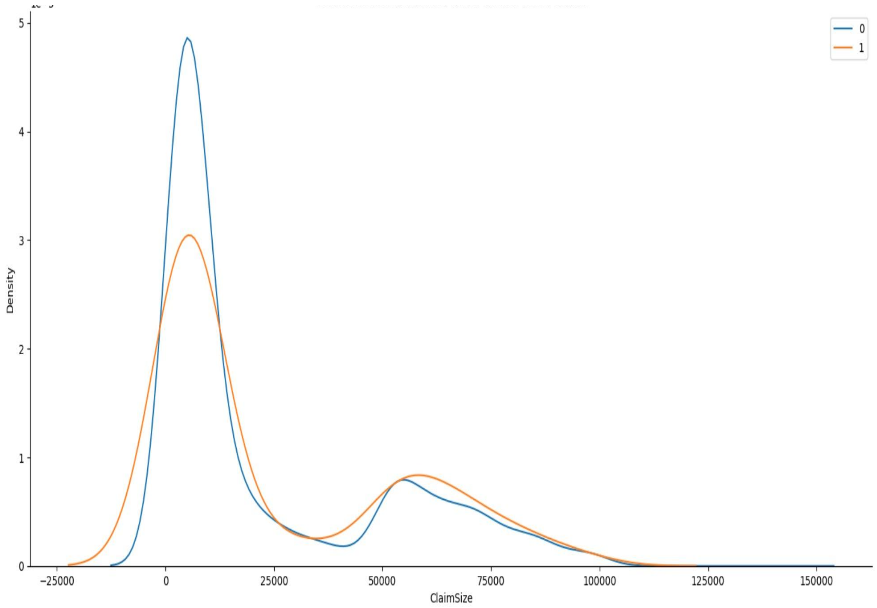
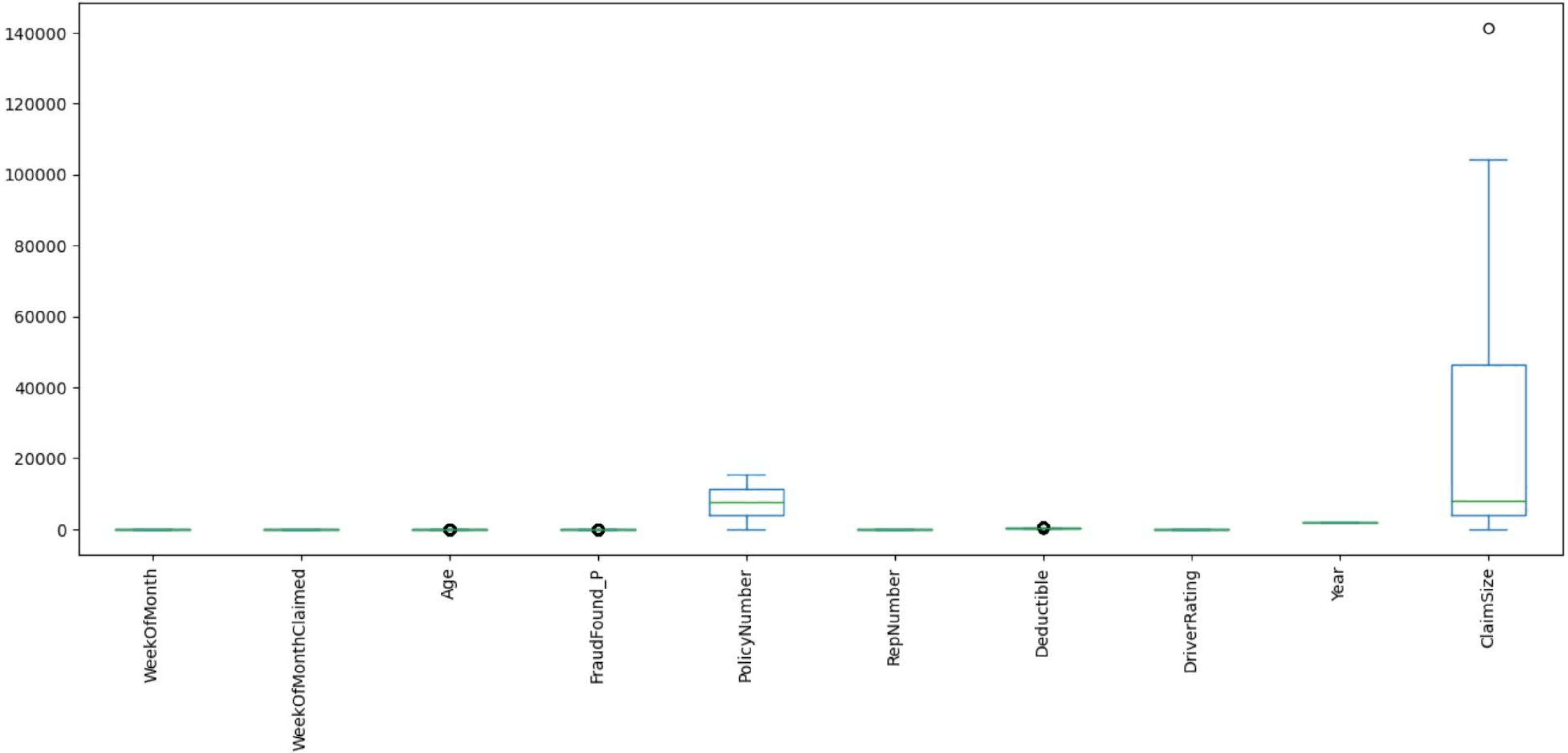


Fig 5. Outlier Analysis

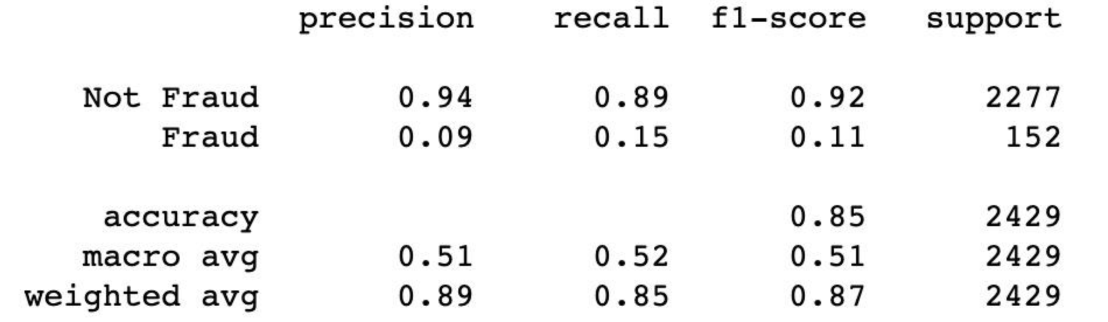


3. Modeling

The data was pre-processed for modeling and dummy features were created.

- Non-numerical features such as ‘MonthClaimed’, ‘DayOfWeekClaimed’, ‘VehiclePrice’, ‘AgeOfVehicle’, and ‘AgeOfPolicyHolder’ were converted to numerical values.
- Logistic regression, decision tree, and random forest were tested with the data.

Logistic Regression

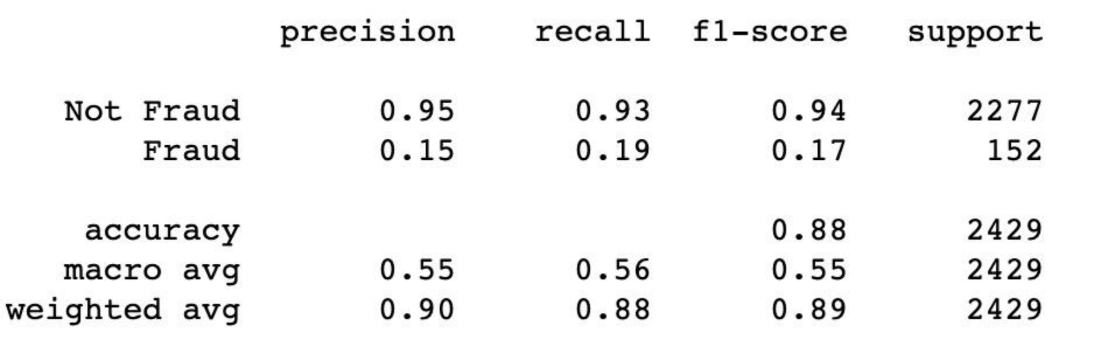


```
              precision    recall  f1-score   support

   Not Fraud       0.94      0.89      0.92      2277
       Fraud       0.09      0.15      0.11       152

    accuracy                           0.85      2429
   macro avg       0.51      0.52      0.51      2429
weighted avg       0.89      0.85      0.87      2429
```

Decision Tree



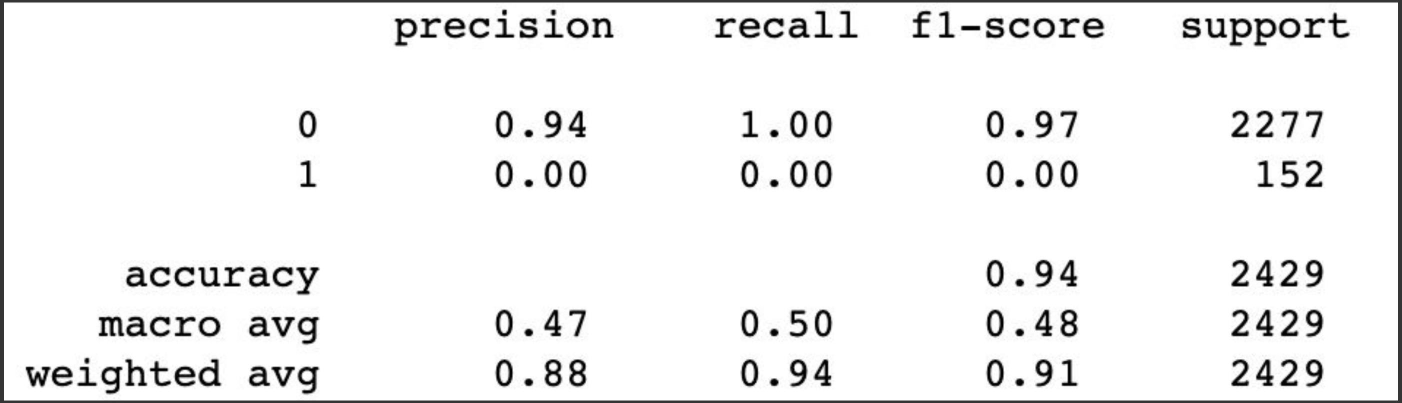
```
              precision    recall  f1-score   support

   Not Fraud       0.95      0.93      0.94      2277
       Fraud       0.15      0.19      0.17       152

    accuracy                           0.88      2429
   macro avg       0.55      0.56      0.55      2429
weighted avg       0.90      0.88      0.89      2429
```

Best Model:

Random Forest



```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      2277
           1       0.00      0.00      0.00       152

    accuracy                           0.94      2429
   macro avg       0.47      0.50      0.48      2429
weighted avg       0.88      0.94      0.91      2429
```

Future Work

Larger Dataset

Modeling on larger dataset

More context for fraudulent claims