EDA ANALYSIS ON FRAUD DATA

Ajulo Joseph Kolawole

13/12/2021

## 0.1 Problem Statement

You have been commissioned as an independent consult to explore motor fraud cases for an insurance company based on a given dataset. You should provide insight to the company on motor fraud cases through Exploratory Data Analysis (EDA), you may use feature engineering to aide this process. You should develop a report through R Markdown enabling your work to be reproducible.

## 0.2 Solution Summary

The data science team has identified four key areas to aid model development:

1. Develop a range of **exploratory visualisations** to understand the data after undergoing intensive cleaning procedure
2. Highlight **any patterns and anomalous behaviour**
3. **Features** that may be engineered to improve model performance
4. **Conclusions** highlighting the insights you have gained

### 0.2.1 Summery Stats and Exploratory Visualisations

A range of visualisations are developed to analyse temporal variation in building heating load.

Table 1: Data summary

|  |  |
| --- | --- |
| Name | fraud\_location\_tbl |
| Number of rows | 999 |
| Number of columns | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| factor | 2 |
| logical | 1 |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| driver | 0 | 1 | 10 | 17 | 0 | 639 | 0 |
| Location | 0 | 1 | 3 | 10 | 0 | 13 | 0 |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| passenger1 | 0 | 1 | FALSE | 2 | 0: 647, 1: 352 |
| passenger2 | 0 | 1 | FALSE | 2 | 0: 889, 1: 110 |

**Variable type: logical**

| skim\_variable | n\_missing | complete\_rate | mean | count |
| --- | --- | --- | --- | --- |
| fraudFlag | 0 | 1 | 0.1 | FAL: 899, TRU: 100 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 0 | 1 | 41.91 | 16.27 | 20 | 28 | 39 | 53 | 85 | ▇▆▅▂▂ |
| repaircost | 1 | 1 | 1321.14 | 855.87 | 500 | 500 | 1000 | 2000 | 3500 | ▇▁▃▁▂ |

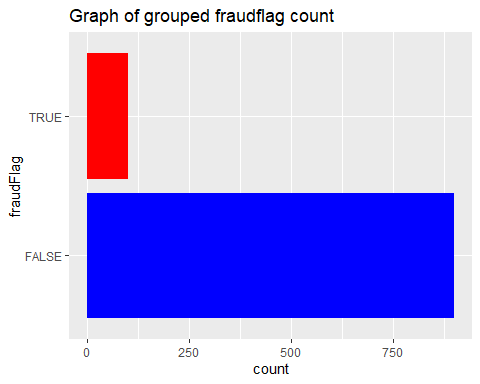


Figure 1: showing The grouped Fraudflags and their respectives count

From the figure above the FALSE has the highest count and the TRUE has the lowest count. The False indicate the drivers who are not involved in fraud while the True is the identified fraudulent drivers which are less in the data set.

## # A tibble: 26 x 3  
## # Groups: Location [13]  
## Location fraudFlag n  
## <chr> <lgl> <int>  
## 1 BARROW FALSE 84  
## 2 BARROW TRUE 2  
## 3 BLACKWATER FALSE 94  
## 4 BLACKWATER TRUE 1  
## 5 BOYNE FALSE 59  
## 6 BOYNE TRUE 2  
## 7 CAMAC FALSE 83  
## 8 CAMAC TRUE 3  
## 9 CORRIB FALSE 93  
## 10 CORRIB TRUE 1  
## # ... with 16 more rows

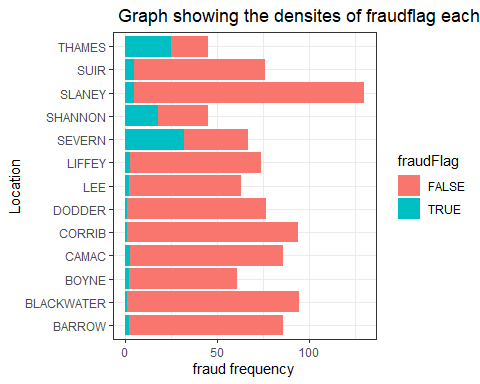


Figure 2: The Locations and Number of drivers in each

Location “Slaney” has the higest number of honest drivers and location “severn” in the other hand has the higest TRUE fraudflag count. Locations “Corrib” and “Dodder” relatively has the lowest fraud cases and also exhibited high FALSE fraudflag count. “Suir” and “Slaney” has a significant fraud rate also.

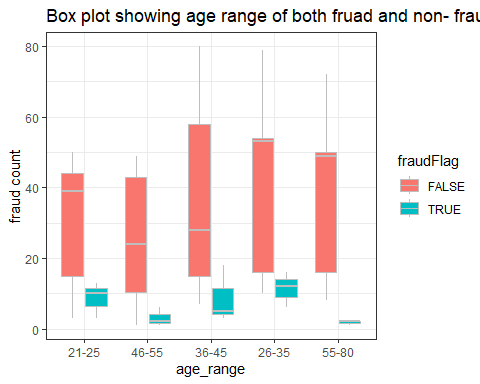


Figure 3: Plot showing Age range as an indicator of Fraud

For ‘TRUE’ fraudflag case which indicates only fraudulent drivers. The age range of 26-35 and 36-45 have the highest fraud cases. though 36-45 years range have have the highest fraud counts but more drivers of age range 26-35 have the highest fraud cases. Drivers of age 55-80 have the least fraud case , even thought they share same lowest points with 46-50 years but then more 46-55 years drivers have higher fraud cases than 46-50

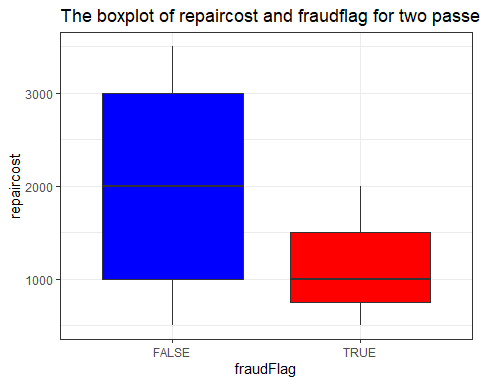


Figure 4: The visualisation of the effect of passengers on repaircost hence fraudflag

The box plot above is derived for drivers with two passengers who are likely to be involved in fraud or otherwise based on their respective repair cost. The “False” of the falseflag illustrates a driver is not involve in fraud , and in this case the drivers pays even as high as 3500 as repair cost.Though half of the drivers in this case pays between 1000 and 3000 as repair cost.

The fraudulent driver pays up to 2000 as repair cost. The median cost is about 1200 and as low as 250. we can assume based on the analysis before the plotting that we are dealing with two passengers and 50 percent of the fraudulent drivers pays above 500 and up to 1500 as repair cost.We can deduce here that the repair cost has a relatively low effect on fraud having in mind its a two passenger case.

### 0.2.2 Engineered Features

Features can be engineered based on the insights to improve predictive model performance. Feature engineering can be viewed as the process of transforming raw data into features that better represent the underlying problem to predictive models. Important features or variables are engineered from the data to better represent the underlying analysis context to the predictive models.

number of passengers was encoded from name to 0,1.

You can represent the insights from the EDA process to the respective model through a binary encoding process.

### 0.2.3 Conclusions from Exploratory Analysis

The conclusions from the analysis are as follows:

\*Locations generally all the address has both fruad and no fraud driver but location “Severn” has more higest fraud cases therefore attentions needs to be taken there.

\*Drivers age being an important factor for fraud hence more younger fraudulent between ages 26-40 exhibits highest fraud rate

\*The higher the repaircost doesnt mean the driver is frudulent. Drivers with more than two passengers involve in fraud pays approx 2000 as repaircost