EDA and Feature Engineering for Insurance Fraud Detection

RENAN R. FIORAMONTE

15/12/2021

1 Problem Statement

I have been tasked to identify key features of fraud for a insurance company. Exploratory Data Analysis (EDA) is a key process in the development of the model to understand the range of improper activities which a person may commit in order to achieve a favorable outcome from the insurance company.
This could range from staging the incident, this takes into consideration two analysis:
1. Areas where the accident happened and the number of passengers; 2. Age analysis to determine if there is a particular group whom is more likely to commit fraud; 3. the extent of damage, analysing how often claims were ruled as fraud in relation to the repair cost.

2 Solution Summary

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

1. Undertake some **data preparation** tasks to ensure the quality of data, this includes using **Feature engineering** to improve model performance;
2. Develop a range of **exploratory data visualisations** to understand the factors of insurance fraud;
3. **Conclusions** to summarise findings.

3 Data Preparation

A range of data preparations tasks were necessary to ensure accuracy, completeness and consistence of the data allowing us to use the highest quality data in our analysis.

The raw data set was messy and inconsistent, issues such as special characters and string mixed with numerical values in the same column throughout the data set as you can see below:

	driver	age	address	passenger1	passenger2	repaircost	fraudFlag
1	JOSEPH MCGRATH	24	3 CO$RIB VIEW	JOSEPH GRIFFIN	<NA>	approx 3k	FALSE
2	MARY BRENNAN	53	21 BLACKWATER VIEW	<NA>	<NA>	approx 2k	FALSE
3	JOSEPH COLLINS	48	7 SLANEY LODGE	<NA>	<NA>	approx 2k	FALSE
4	ROBERT WALSH	40	12 LIFFEY GROVE	<NA>	<NA>	approx 500	FALSE
5	KEVIN OCONNELL	27	21 BLACKWATER GLADE	<NA>	<NA>	approx 500	FALSE
6	BRIAN CULLEN	38	9 BARROW GLADE	MICHAEL BROWNE	<NA>	approx 2k	FALSE
7	HELEN PHELAN	34	6 BAR&OW GLADE	<NA>	<NA>	approx 500	FALSE
8	SEAN KAVANAGH	34	20 LEE LODGE	<NA>	<NA>	approx 2k	FALSE
9	PATRICK GRIFFIN	33	4 SHANNON COURT	<NA>	<NA>	approx 2k	FALSE
10	MARK MCNAMARA	21	16 SLANEY GLADE	JOHN BUTLER	<NA>	approx 500	FALSE

The data preparation phase consisted of the following tasks:

- Clean the unnecessary information. An example of unnecessary data can be seen in the repair cost column. I.e.: “approx” and “k” after the cost.
- Remove special characters and unnecessary information in the address column, for example, the number and the second part of the address. It was required in order to be able to analyse fraud pattern though addresses.
- Special characters had to be dealt with in 3 rows in the driver’s column as well.

3.1 Engineered Features

The first step was to create binary encoded columns. i.e.: “1” for yes and “0” for no. Predictive models need to have the data coded this way to make accurate predictions. Some of the EDA were taken from binary encoded columns.
The outcome of the task mentioned above is shown below:

	passenger1	passenger2	fraudFlag	driver_alone
1	1	0	0	0
2	0	0	0	1
3	0	0	0	1
4	0	0	0	1
5	0	0	0	1
6	1	0	0	0
7	0	0	0	1
8	0	0	0	1
9	0	0	0	1
10	1	0	0	0

The second step was to categorize via binning the age column, creating a new column called “age_category”. This would help us understanding if there is a particular group whom is more likely to commit fraud.
The output of the task mentioned above is shown below:

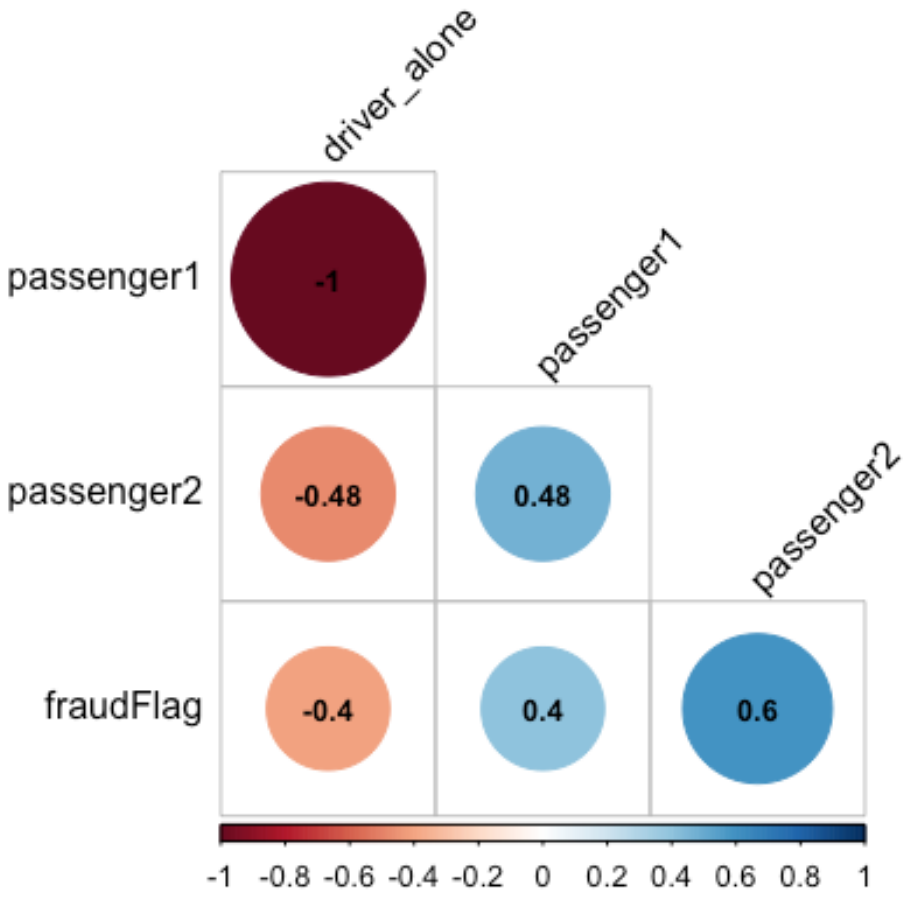
	age	age_category
1	24	20 - 29
2	53	50 - 59
3	48	40 - 49
4	40	40 - 49
5	27	20 - 29
6	38	30 - 39
7	34	30 - 39
8	34	30 - 39
9	33	30 - 39
10	21	20 - 29

Upon finishing all the tasks in the **Data preparation** and **feature engineering**, the data set is now in a great format to start modelling.
The outcome of the tasks performed can be seen below:

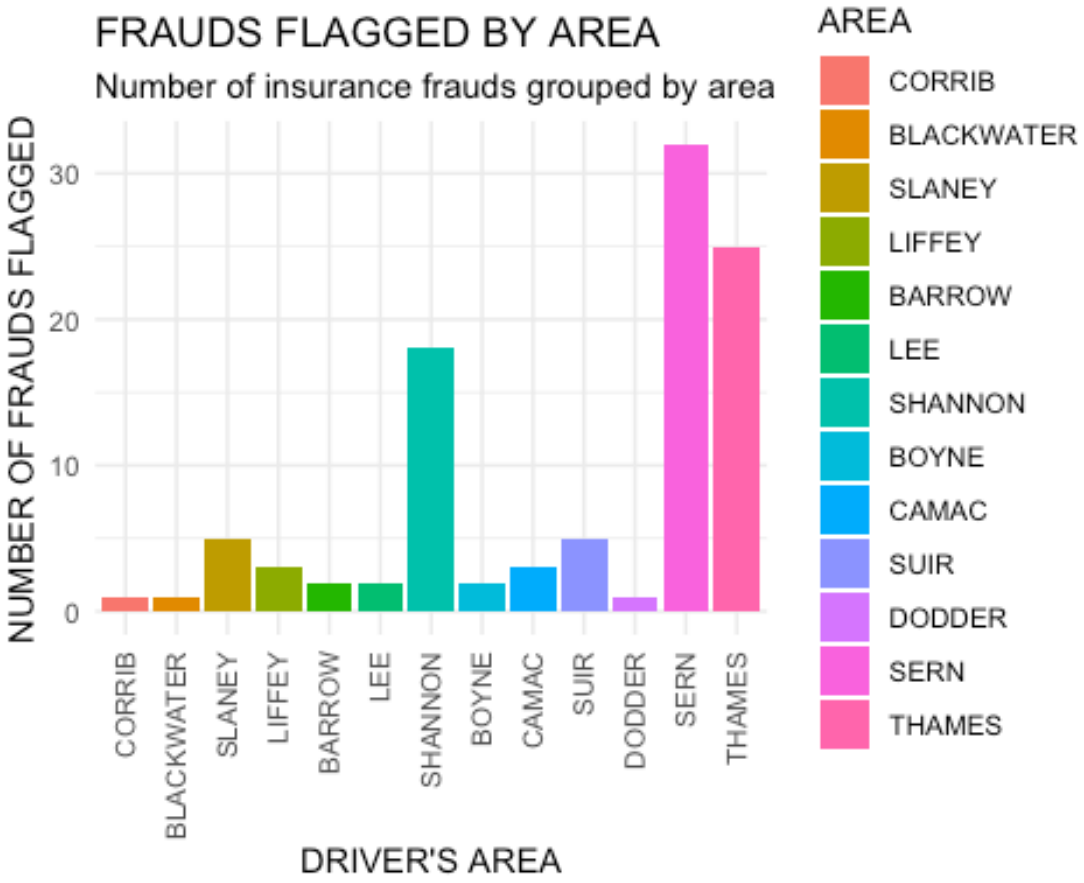
	driver	age	address	passenger1	passenger2	repaircost	fraudFlag	age_category	driver_alone
1	JOSEPH MCGRATH	24	CORRIB	1	0	3000	0	20 - 29	0
2	MARY BRENNAN	53	BLACKWATER	0	0	2000	0	50 - 59	1
3	JOSEPH COLLINS	48	SLANEY	0	0	2000	0	40 - 49	1
4	ROBERT WALSH	40	LIFFEY	0	0	500	0	40 - 49	1
5	KEVIN OCONNELL	27	BLACKWATER	0	0	500	0	20 - 29	1
6	BRIAN CULLEN	38	BARROW	1	0	2000	0	30 - 39	0
7	HELEN PHELAN	34	BARROW	0	0	500	0	30 - 39	1
8	SEAN KAVANAGH	34	LEE	0	0	2000	0	30 - 39	1
9	PATRICK GRIFFIN	33	SHANNON	0	0	2000	0	30 - 39	1
10	MARK MCNAMARA	21	SLANEY	1	0	500	0	20 - 29	0

4 Exploratory Data Analysis

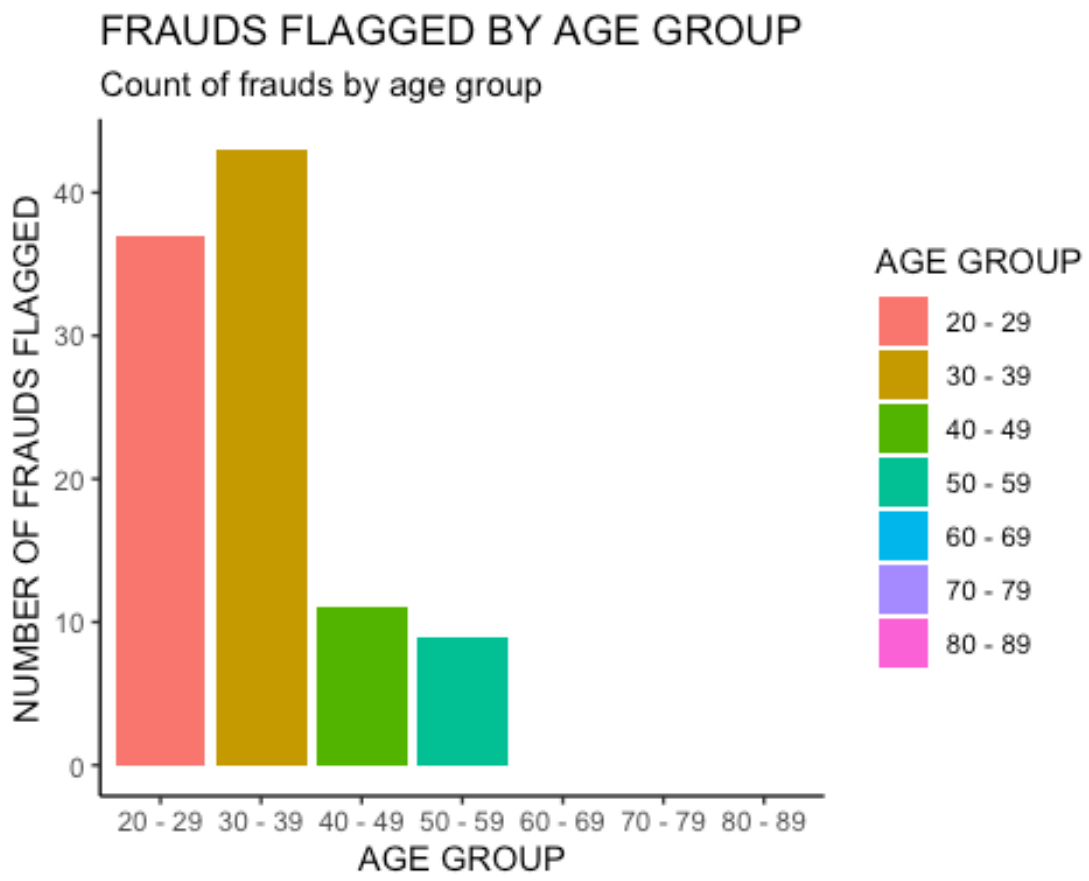
Exploratory Data Analysis (EDA) is critical and allows us to perform initial investigations on data so as to discover patterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representations. The visualizations shown in this section will identify patterns of the frauds detected in the data set.



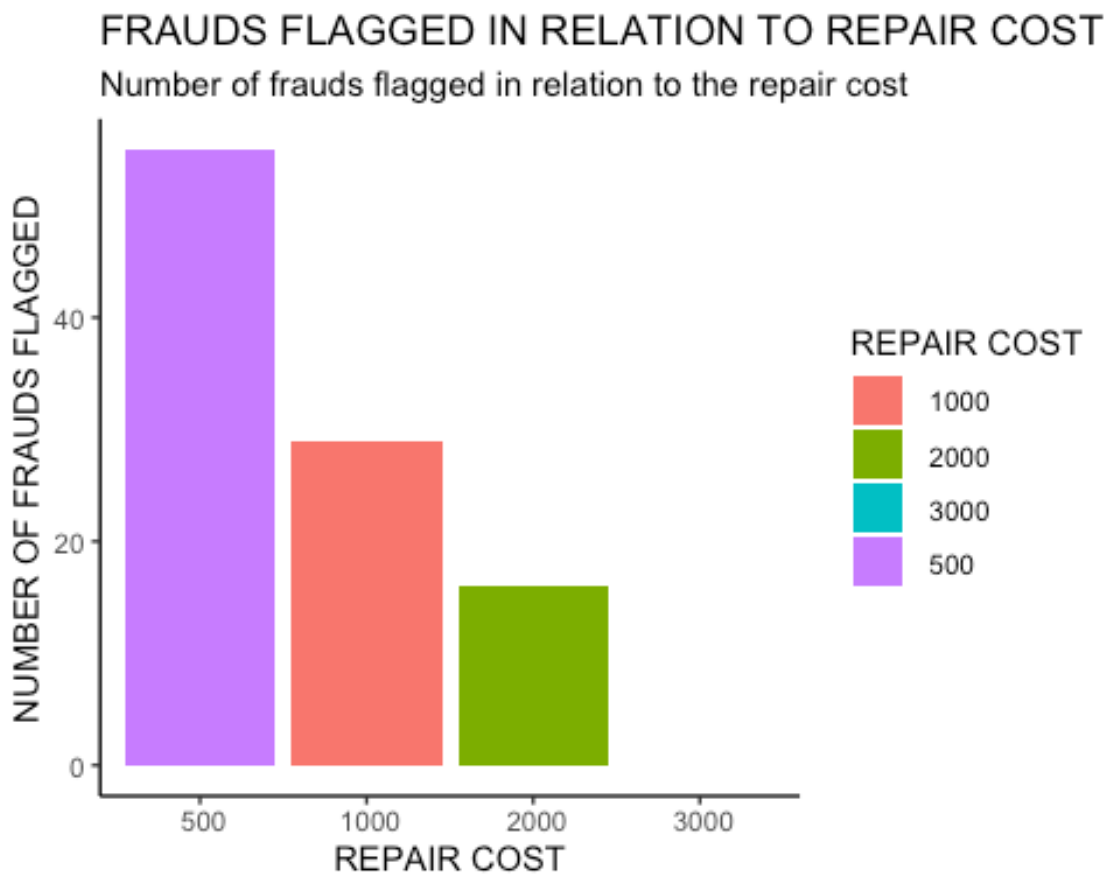
- There is a strong correlation between passenger2 and fraudFlag, which indicates that most of the frauds detected were from claims where there were 2 passengers in the car. Other correlations that exist are weaker but still notable.



* The graph above shows the amount of claims made by drivers from certain areas, it clearly shows that drivers from **Sern** holds the larger share, accounting for **32%** of the total of frauds flagged, followed by **Thames** with **25%** and **Shannon** with **18%**, highlighting that for the future claims made from drivers from this area will require further investigation.



As the graph above shows, the larger share of the frauds belongs to the **30 to 39** years old group, accounting for **43%** of the total, follwed by the **20 to 29** years old with **37%**, **40 to 49** years old with **11%** and then **50 to 59** years old with only **9%** of the total. Normally in the insurance world, the policy would be much more expensive for younger people, but with this number of frauds being commited by an older group, it requires some reevaluation.



* The graph above highlights the fact that the vast majority of frauds were flagged where the repair cost amount was **500€** with **55%** of the total, followed by **1000€** with **30%** and then **2000€** with **16%** in third place. Interesting though, **no frauds** were flagged where the repair cost amount was **3000€**. This could be due to the fact that perhaps, in claims where the repair cost is higher than a certain amount, a further investigation is required. This highlights the fact that the vast majority of the frauds were in the claims where the repair cost was low.

5 Conclusions

The conclusions from the analysis are as follows:

- Only 10% of the total number of claims in the data set were ruled as fraud.
- The majority of the frauds were detected when drivers were from Sern, Thames and Shannon accounting for 75% of the total number of frauds flagged.
- There is a higher chance of the claim being ruled as fraud if there are two passengers. The more passenger there are, the higher is the chance of fraud.
- The people who are more likely to commit fraud are aged between 30 and 39 years old accounting for 43% of total, followed by the 20 to 29 years old group with 37%.The other age groups numbers are low but still notable.

- The majority of the frauds were commited where the repair cost was 500€ and there were no frauds identified where the repair cost was as high as 3000 and as mentioned previously, this could be due to the fact that perhaps, there some further investigations in high value claims.