Machine Learning

Fraud Detection Dataset EDA

Mobin Nesari
Dr. Hadi Farahani
May 8, 2023

Introduction:

Vehicle insurance fraud is a type of crime in which individuals collaborate to make deceitful or exaggerated claims regarding property damage or personal injuries that arise from an accident. This wrongful act can take many forms, such as staged accidents where criminals intentionally orchestrate a collision, phantom passengers who falsely claim to have suffered injuries despite not being present during the accident, and fraudulent personal injury claims where the extent of the injuries is exaggerated beyond reality.

Dataset Information:

The shape of this dataset is 15420×33 which indicates 15420 records each has 33 features. Here, features with their name, count of non-null values and datatype are shown:

#	Feature	Non-Null Content	Datatype
0	Month	15420	object
1	WeekOfMonth	15420	int64
2	DayOfWeek	15420	object
3	Make	15420	object
4	AccidentArea	15420	object
5	DayOfWeekClaimed	15420	object
6	MonthClaimed	15420	object
7	WeekOfMonthClaimed	15420	int64
8	Sex	15420	object
9	MaritalStatus	15420	object
10	Age	15420	int64
11	Fault	15420	object
12	PolicyType	15420	object

13	VehicleCategory	15420	object
14	VehiclePrice	15420	object
15	FraudFound_P	15420	int64
16	PolicyNumber	15420	int64
17	RepNumber	15420	int64
18	Deductible	15420	int64
19	DriverRating	15420	int64
20	Days_Policy_Accident	15420	object
21	Days_Policy_Claim	15420	object
22	PastNumberOfClaims	15420	object
23	AgeOfVehicle	15420	object
24	AgeOfPolicyHolder	15420	object
25	PoliceReportFiled	15420	object
26	WitnessPresent	15420	object
27	AgentType	15420	object
28	NumberOfSuppliments	15420	object
29	AddressChange_Claim	15420	object
30	NumberOfCars	15420	object
31	Year	15420	int64
32	BasePolicy	15420	object

According to dataset information, there are no missing values and we have a clean dataset in missing value. Due to this fact, missing values section will not be covered in this EDA.

Data Visualization:

In this section, some plots will be shown and then argue about plots and their relation and correlation.

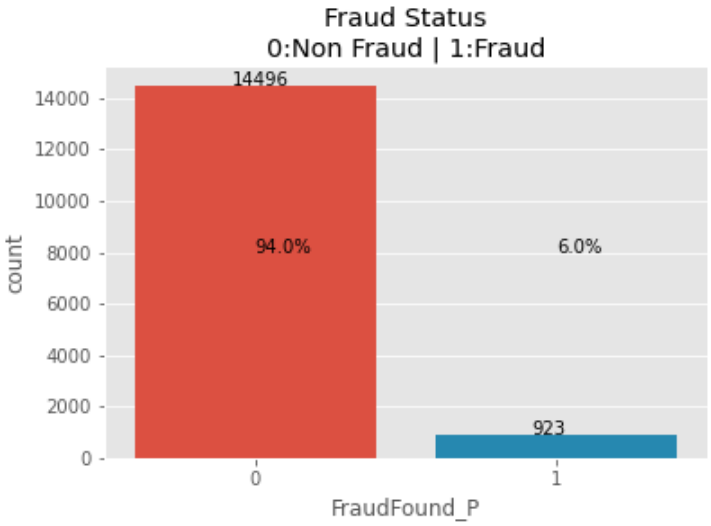


Fig 1: Count of Target Value

In this plot, as you can see, count of target value 'FraudFound_P' has been shown. Intuitively, we can see that our dataset is imbalanced so we can't rely on accuracy for model criterion and should consider other metrics like F1 or ROC-AUC metrics. Additionally, we can use sampling methods to balance dataset for training machine learning models.

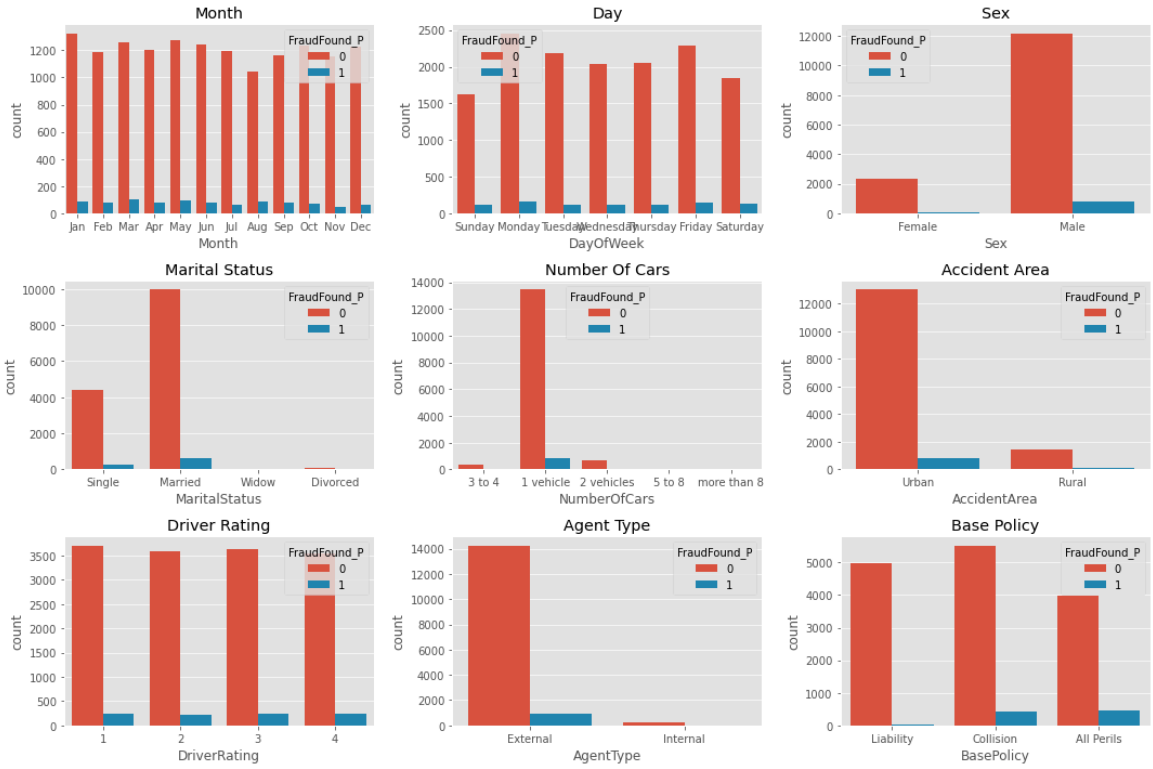


Fig 2: Count Plots for Several Important Features

In Fig 2, some important features which are indicated by scientist has been shown. According to this plot, men had more fraud records than women. Most of frauds has been found in urban area. Those who are chosen for fraud sue, have one vehicle. Most of them has been evaluated by external agent. Others don't show a significant information about feature correlation with target value.

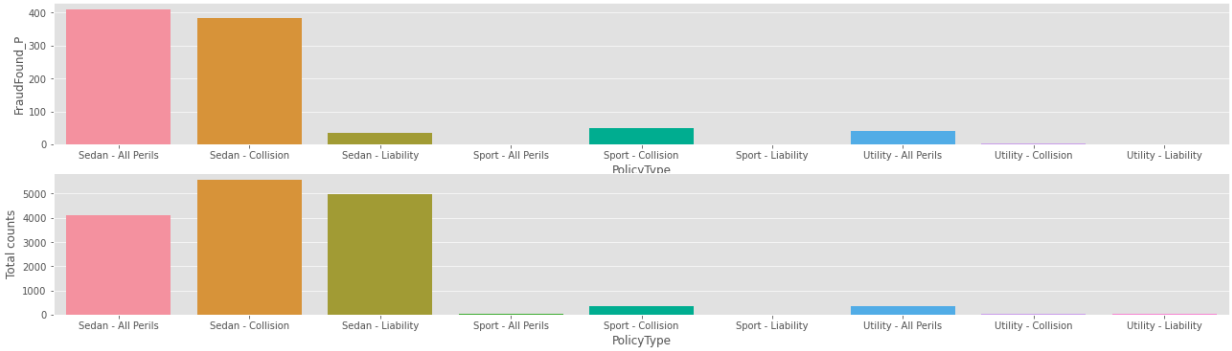


Fig 3: Number of Fraud Cases Among All Types of Cars

In Fig 3, A distribution of car types for fraud cases has been shown which indicates the most used cars in positive frauds are sedans all perils, then sedan collision. But total counts sedan collision is on top of the list. This fact points out that sedans are the main car type in fraud cases.

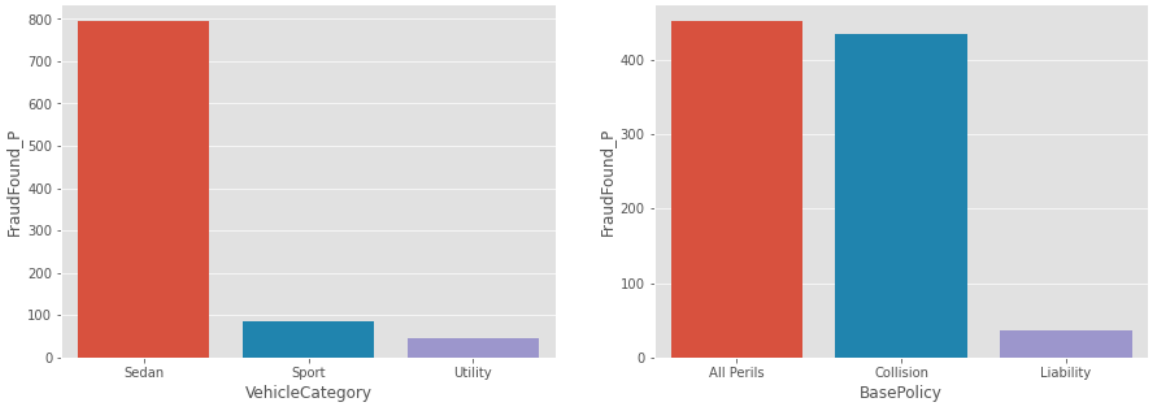


Fig 4: Number of Different Car Types and Different Base Policies

Fig 4 is proving our statement in recent paragraph about car types and indicates this fact that all perils policy is the main policy among fraud cases and liability is the least.

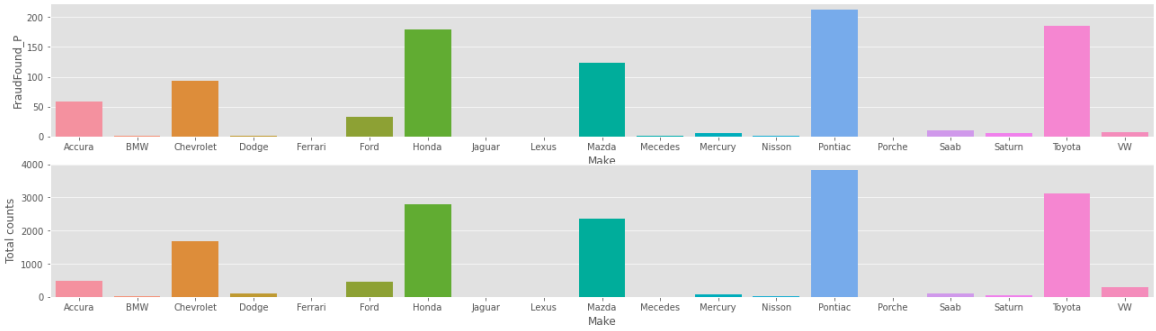


Fig 5: Number of Fraud Cases in Different Car Companies

In Fig 5, a distribution over all recorded fraud cases over different car companies has been shown which indicates critical companies over different companies. This fact plays a crucial role in determining if a record is fraud or not. According to this plot, most of positive fraud cases involved Pontiac cars and more luxurious car brands like Ferrari, Jaguar, Lexus and Porsche are not involved in fraud cases. Obviously, this fact is correct because rich people are less likely to involve in fraud cases.

Correlation Matrix:

Correlation matrix is a widely used tool in machine learning and data science to understand the relationship between variables in a dataset. A correlation matrix summarizes the pairwise correlations between all the variables in the dataset, typically represented as a square matrix with each row and column representing a variable. The values in the matrix range from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation between two variables. Correlation matrices are often used for feature selection, identifying highly correlated features that can be removed to reduce redundancy and improve model performance. They also help in identifying relationships between variables that can inform insights and decision-making.

In Cramers correlation matrix shown in Fig 6, the correlation between features have been shown. Some features like month and monthclaimed because of redundancy have strong correlation which should be removed in model training to avoid model bias. But some correlations like correlation between vehicle category and base policy reveal obvious facts. For instance, rich people get more expensive policies than other people which is indicated by their luxurious cars.

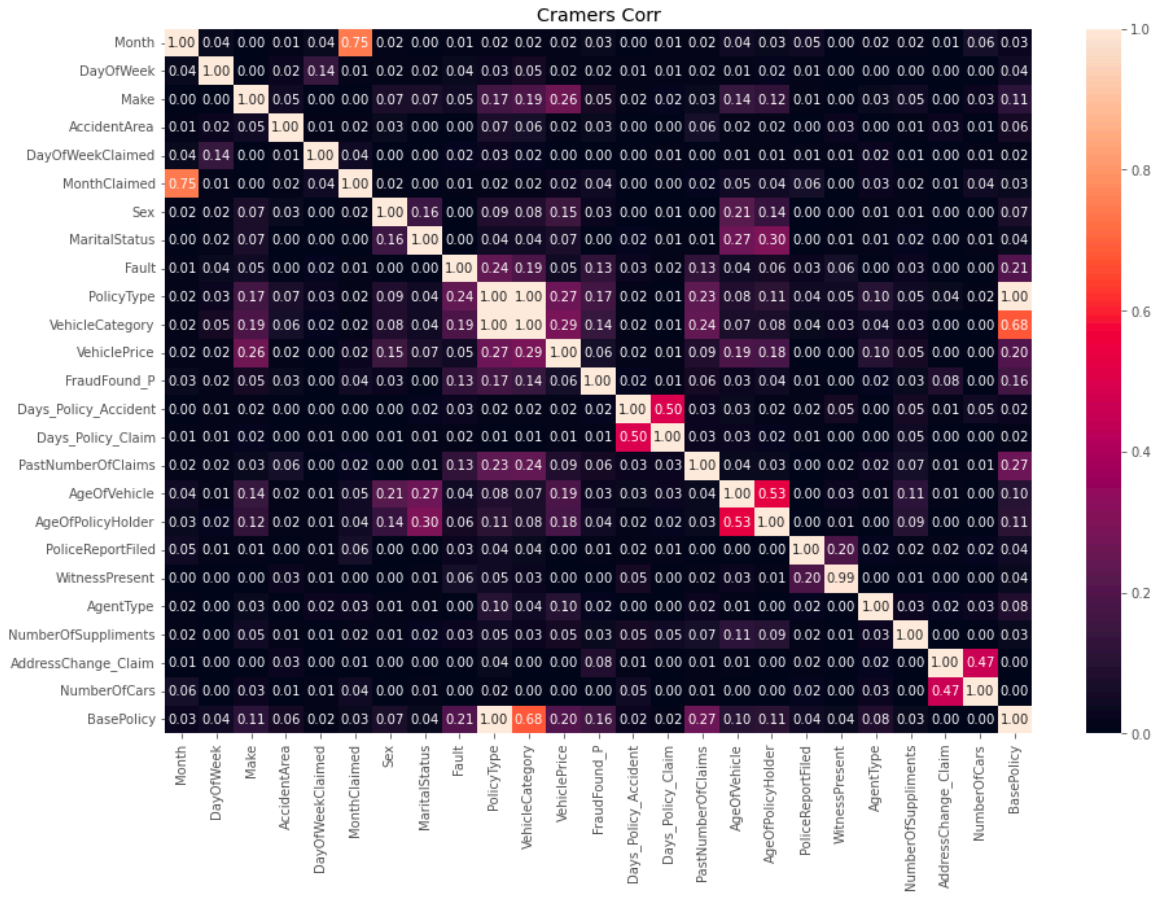


Fig 6: Cramers Correlation Matrix Over Features

Covariance Matrix:

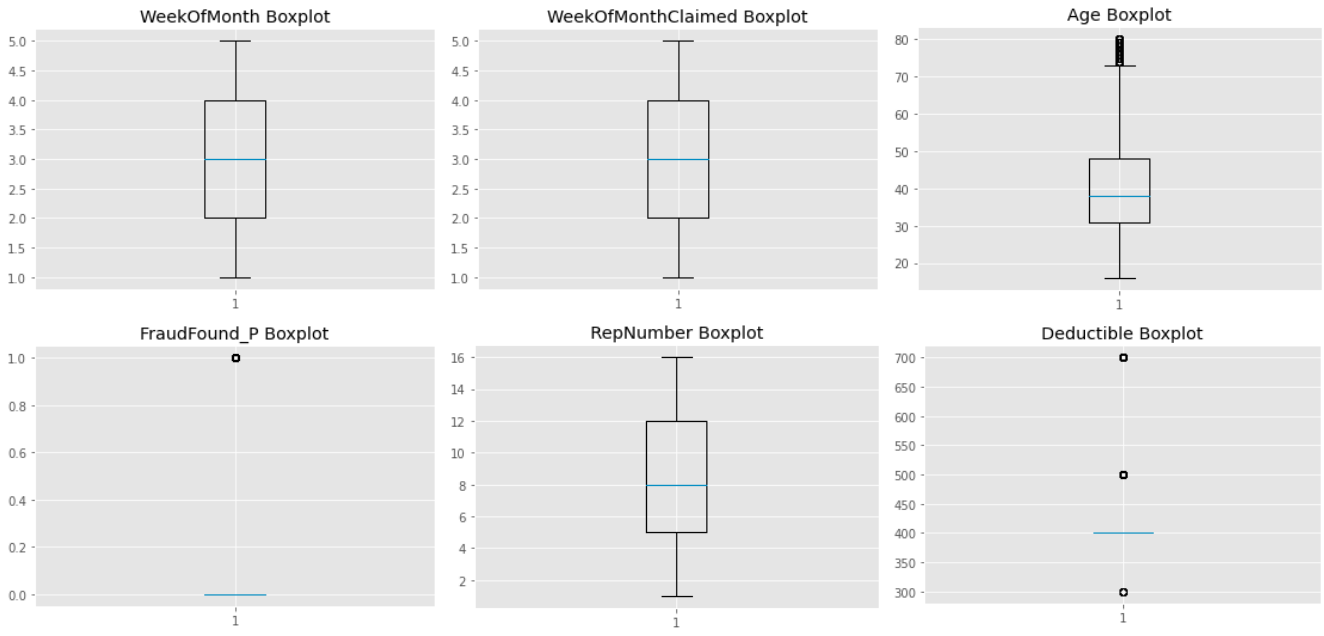
In data science, the covariance matrix is a key tool used to understand and describe the relationships between multiple variables in a dataset. The covariance matrix measures the degree to which two or more variables vary together, providing a measure of the strength and direction of their relationship. Specifically, the covariance matrix contains all of the pairwise covariances between the variables in a dataset, along with their variances on the diagonal. By analyzing the covariance matrix, data scientists can identify patterns and correlations in their data that can inform further analysis, modeling, and decision-making.

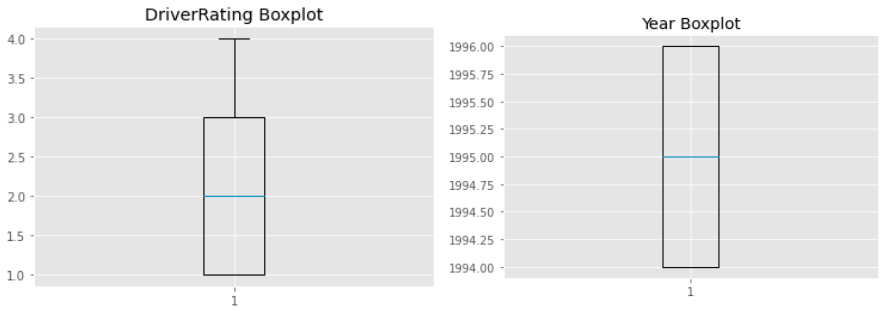
Fig 7: Covariance Matrix of Dataset

In Fig7, you can see covariance matrix of given dataset. As you can see in this dataset, features are very related and due to this fact, trained models will be more confident in their decision making.

Outliers:

In machine learning, outliers are data points that deviate significantly from the majority of other data points in a dataset. While outliers may seem like useful pieces of information at first glance, they can actually have a negative impact on a machine learning model's performance. This is because outliers can skew the data and cause the model to make inaccurate predictions. Therefore, it is often recommended to get rid of outliers before training a machine learning model. By removing outliers, we can improve the accuracy of our model and ensure that it is making predictions based on the majority of the data, rather than being influenced by a few anomalous data points.





As you can see, In this dataset, according to IQR criterion to detect outliers, some features have outliers in their record which will be deleted for training phase.