VEHICLE INSURANCE CLAIM FRAUD DETECTION

Reza Mosavi ,400222100

Abstract

Insurance companies have many problems today. One of the big problems of these companies today is detecting the fraud of insurance holders. In this dataset, the aim is to detect fraud in car insurance that has had an accident.Vehicle insurance fraud involves conspiring to make false or exaggerated claims involving property damage or personal injuries following an accident. Some common examples include staged accidents where fraudsters deliberately "arrange" for accidents to occur; the use of phantom passengers where people who were not even at the scene of the accident claim to have suffered grievous injury, and make false personal injury claims where personal injuries are grossly exaggerated. In this exercise, we tried to identify fraudsters with the help of machine learning methods and classification algorithms.

Introduction

We work with 33 features when working with this dataset. To implement machine learning models for data classification based on the FraudFoundP feature, we use the sklearn library in the Python programming language. The problems we face before implementing the algorithms

- Data imbalance
- Existence of incorrect values in the data

After examining these challenges, we have to encode categorical data. After the final preparation, we start training these models:

- Logistic Regression
- Support Vector Machine
- Decision Tree
- Random Forest
- KNN
- GradientBoosting

After training these models, we start to improve the results of some models For example, we use the GridSearchCV algorithm or I create over-fitting models. .

Table 1: Margin Specifications

Margin	A4 Paper	US Letter Paper
Top	37 mm (1.46 in)	0.75 in (19 mm)
Bottom	19 mm (0.75 in)	0.75 in (19 mm)
Left	20 mm (0.79 in)	0.79 in (20 mm)
Right	20 mm (0.79 in)	1.02 in (26 mm)

METHODOLOGIES

Data Analysis

first, for data analysis we should read description of dataset and know features and target of dataset in detail.

Checking Numerical Features : We have 9 numerical features in our dataset First, we draw the histogram of these data.

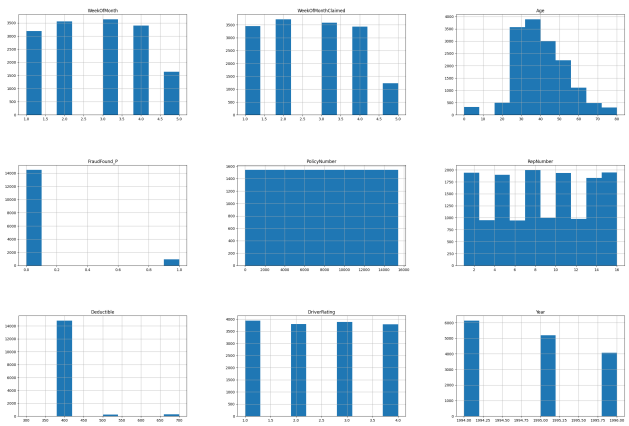

Figure 1: numerical features , It is quickly visible that there are incorrect values in Age.

It can be seen that the Age column has an incorrect value of 0. PolicyNumber has no effect because it is like an index and is just a number. The value in different DriverRating classes is almost the same. In the year column, we generally work with 3 different years. Most of the accidents occurred in the middle of the month. Most of the age signs are about 30 years old.

Now, in particular, we check the age chart and its distribution chart based on the value whose ages are recorded as zero before drawing the chart.

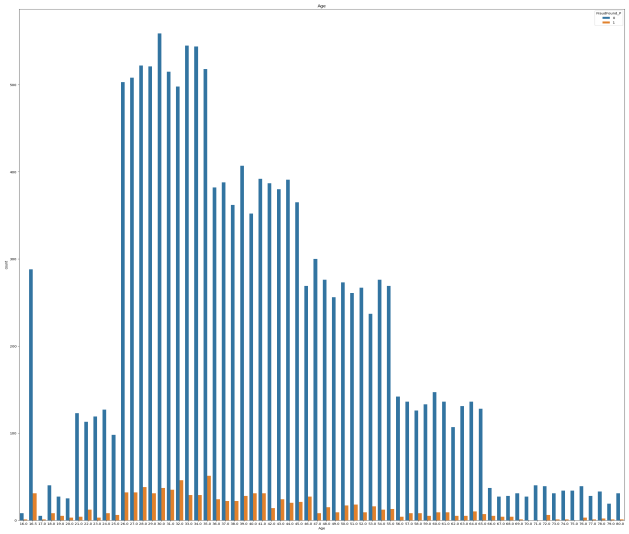

Figure 2: It can be seen that most of the accidents are around the age of 30 and FraudFoundP is more positive around the age of 30.

Now we are going to check the Fraud Found P feature. This feature is the feature based on which we want to do the classification.

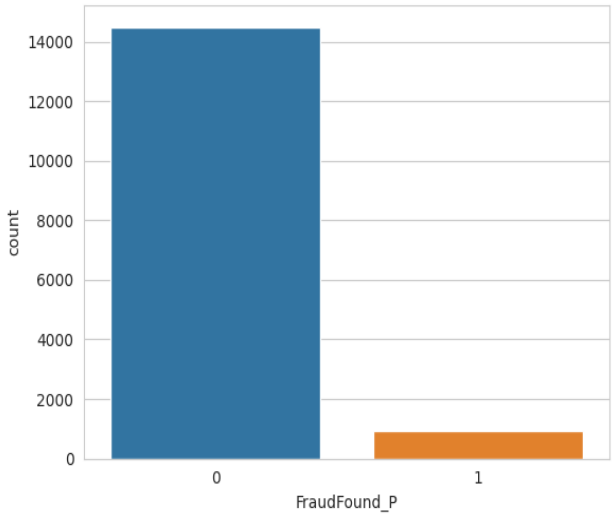

Figure 3: It can be seen quickly that the data is not balanced

In the following, I will try to balance the data in different ways. Finally, since the Policy Number feature has no effect, we will remove this feature and then correct the incorrect values of some values.

Examining the features of categorical : Now to check the feature categorical For this purpose, we draw their graphs based on FraudFoundP.

Features examined in this chart(These features are related to time):

- Month
- WeekOfMonth
- DayOfWeek
- DayOfWeekClaimed
- MonthClaimed
- WeekOfMonthClaimed

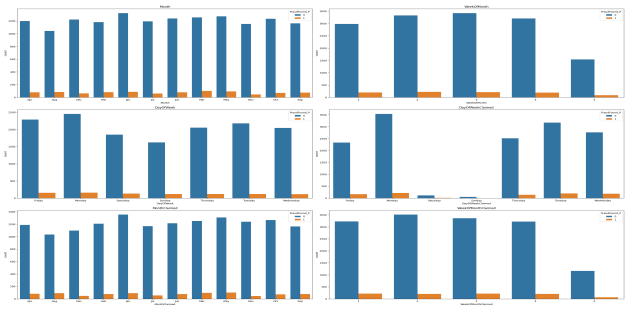

Figure 4: It can be seen that the number of accidents throughout the year in different months is almost constant

It can be seen that the number of accidents throughout the year in different months is almost constant. Every month, most accidents occur in the middle of the month. Every week, most accidents happen in the middle of the week. It is quite visible that the number of accidents claimed during Hatfa all occur in an average week. It is quite visible that the number of accidents claimed during the month is almost the same.

Now let's check some other features.

- Sex

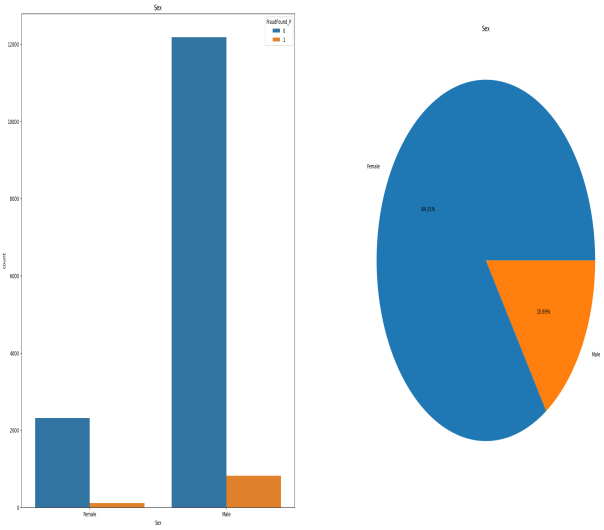

Figure 5: The dominant gender of the society is men and they have more positive FraudFoundP than women.

- DriveRate

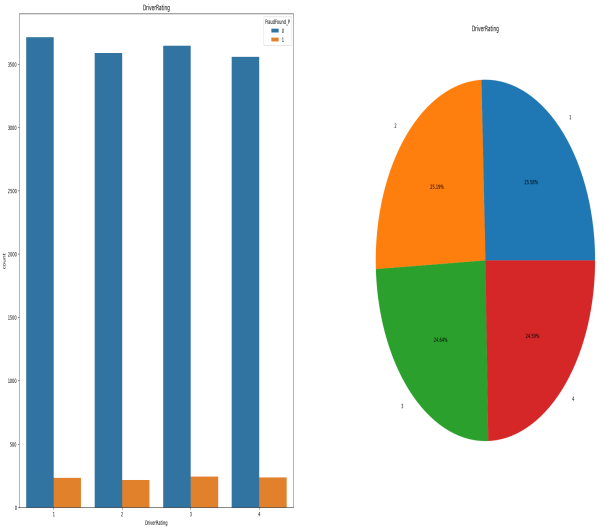

Figure 6: The accident rate of vehicles that have an insurance policy is higher and they have more positive FraudFoundP.

- VehicleCategory

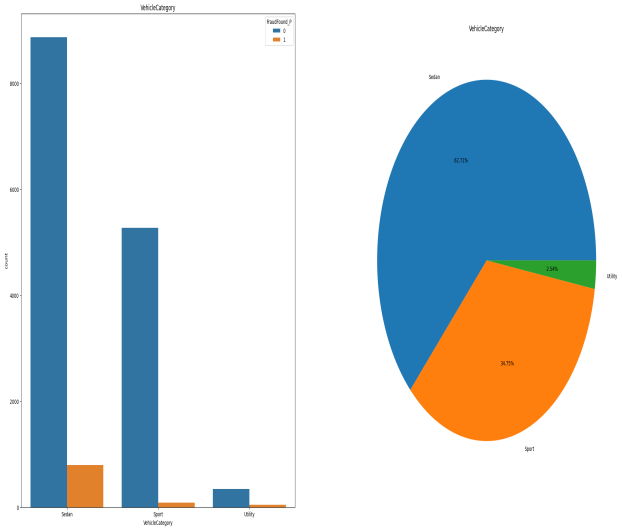

Figure 7: Most of the available vehicles are sedans and they have the highest FraudFoundP rates.

- AgentType

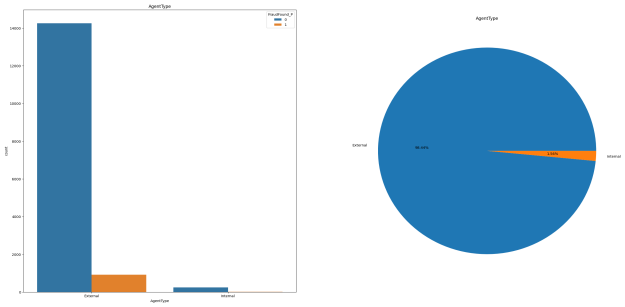

Figure 8: In general, in the agent section, most of our data is external, and all FraudFoundP is linked to external.

- Year

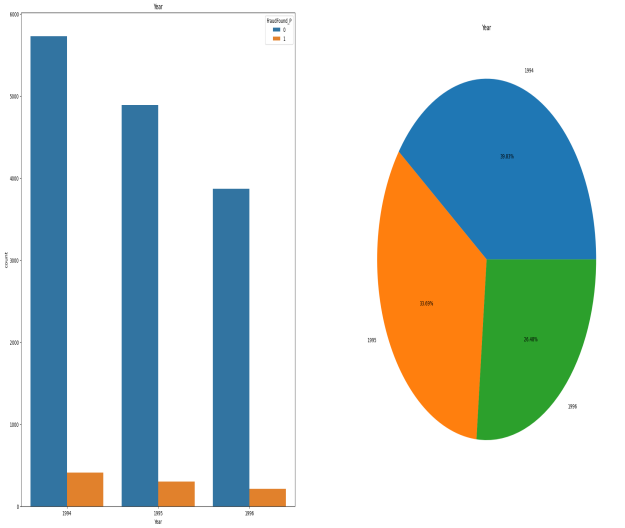

Figure 9: In general, we are dealing with 3 different years.

PreProcessing

In this section, we have to check two things. The first case is to encode categorical features. The second is data balancing.

- encoding In this section, we encode the data using the label encoding method
- Data balancing
  - Oversampling
  - Undersampling

After doing these tasks, the data is ready to train the model.

Models Results

- Logistic Regression
  - Undersampling
    f1 score : 0.23269513991163474
    roc score: 0.6709242370532693

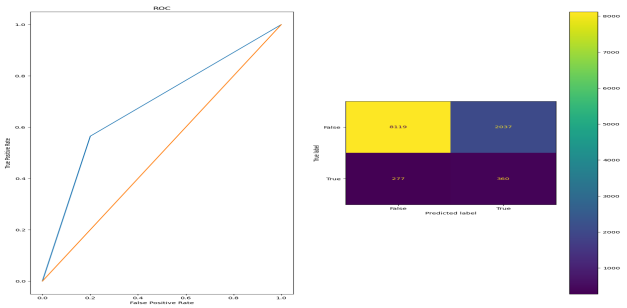

Figure 10

  - Oversampling
    f1 score : 0.22107728337236532
    rocauc score: 0.6231125679586829

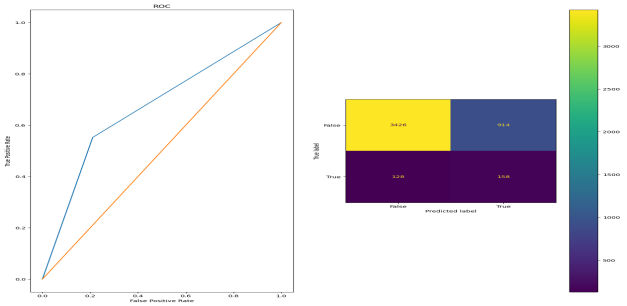

Figure 11

- svm
  - Undersampling
    f1 score : 0.23667820069204157
    roc score: 0.6851261641584222

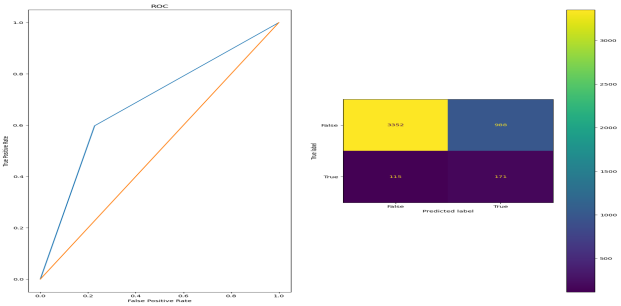

Figure 12

- Oversampling
  f1 score : 0.23804463336875661
  rocauc score: 0.6332465921175598

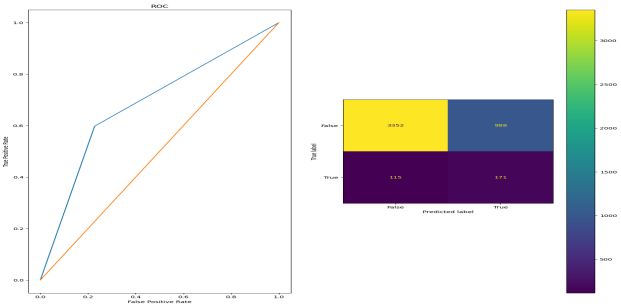

Figure 13

- svm polynomial kernel
  - Undersampling
    f1 score : 0.16927899686520378
    roc score: 0.5600737971705714

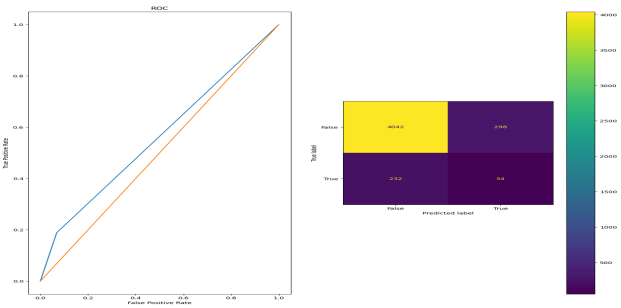

Figure 14

- Oversampling
  f1 score : 0.23169398907103822
  rocauc score: 0.6250612290934872

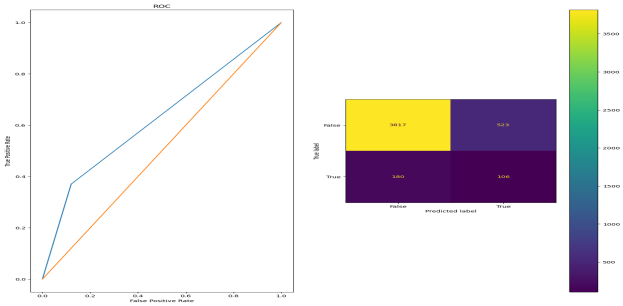

Figure 15

- Grid Search
  f1 score : 0.23847841989758592
  rocauc score: 0.6792046663014405

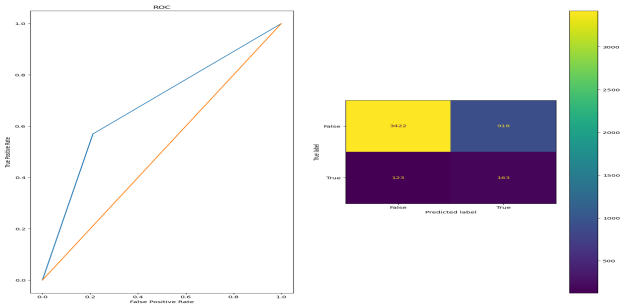

Figure 16

We use Grid Search to find the best input values of the model function. We know that this algorithm from cross validation Uses

- svm kernel RBF
  - Grid Search
    f1 score : 0.22517911975435007
    roc score: 0.6253722084367245

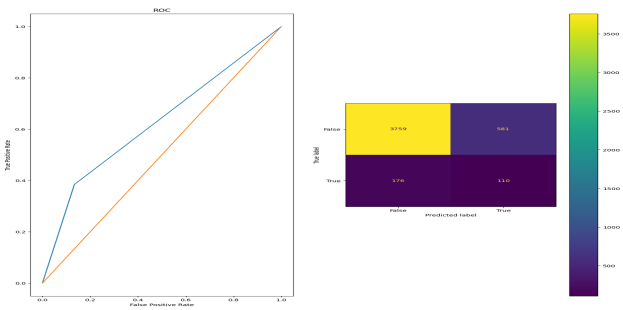

Figure 17

- Oversampling
  f1 score : 0.2130937098844673
  rocauc score: 0.5978698720634205

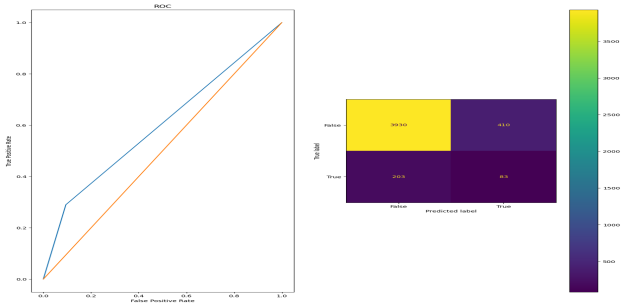

Figure 18

- Decision Tree This model is more suitable in its initial state. We will try to improve the results with pruning.

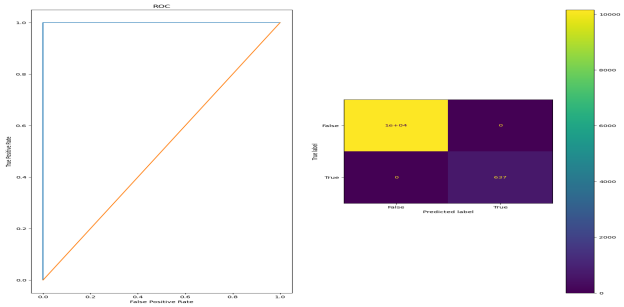

Figure 19

  - Undersampling
    f1 score : 0.22461331540013452
    roc score: 0.6728336179949084

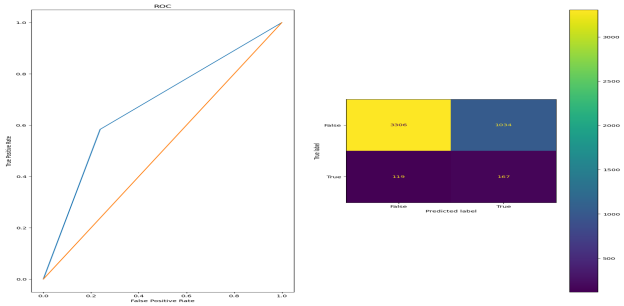

Figure 20

  - Oversampling
    f1 score : 0.2316275490573297
    rocauc score: 0.6544894156650753

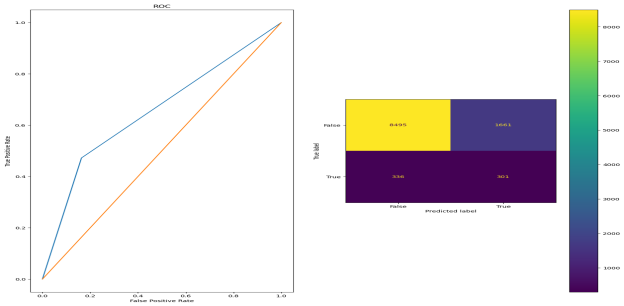

Figure 21

The result did not improve with pruning.

- GradientBoostingClassifier
  - Undersampling
    f1 score : 0.2306525037936267
    roc score: 0.664351777255003

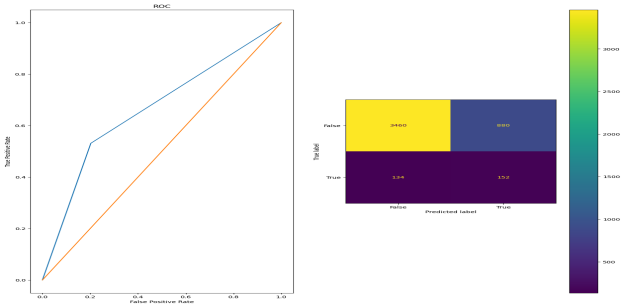

Figure 22

  - GridSearch
    f1 score : 0.2788671023965142
    rocauc score: 0.7468826958783634

Figure 23

- RandomForestClassifier
  - Undersampling
    f1 score : 0.2543933054393306
    roc score: 0.6785222841674454

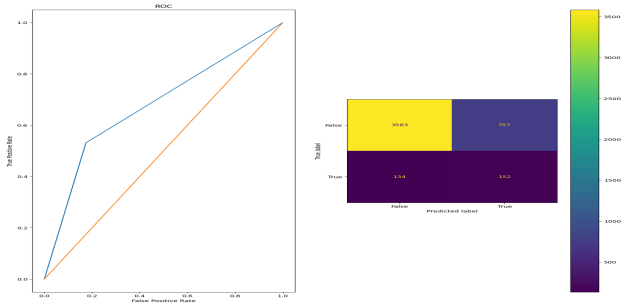

Figure 24

– Oversampling
f1 score : 0.18287937743190658
rocauc score: 0.5613152976056202

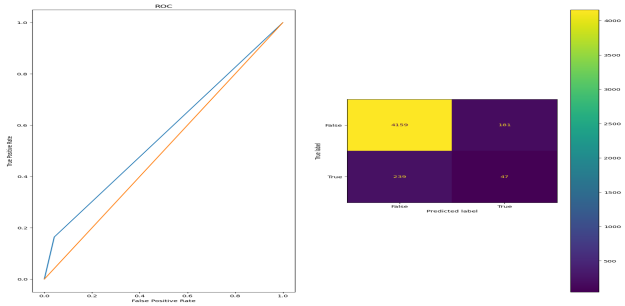

Figure 25

conclusion

The best model resulting from Grid Search is based on Gradient Boosting. Almost the results of all models are at the same level. Note that the selected metric is proportional to the imbalance of the dataset. The roc plots shown all represent model results.