Capstone2_Final Report

Vehicle Insurance Claim Fraud Detection

Introduction

Vehicle Insurance Fraud is a major area of losses to the auto insurance companies. Study conducted by Verisk, indicated that insurance companies had to bear losses to the tune of $29 billion a year.[1] According to Progressive, auto insurance fraud is committed when someone falsely claims about an event and get monetarily compensated.[2] In this project an attempt has been made to explore major contributing factors which have led to fraud along with developing models which can distinguish fraudulent transactions. For this model data is obtained from Vehicle Insurance Claim Fraud Detection[3]

Problem Statement

How can Data Science be leveraged to detect Vehicle Insurance Fraudulent claims to reduce losses on account of exaggerated and false claims about accidents, property damage and physical injury by developing machine learning models which can improve the accuracy as measured by F1 SCORE to above 0.8 in 6 months**.**

Scope

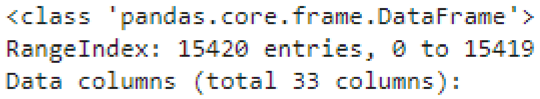
The only available data used for analysis is a single CSV file which is used for developing models and exploratory data analysis. Additionally, the model can be used only if the new data has same feature space as original csv file.

Constraints within solution space

The major constraint is availability of only one csv file along with the fact that data is highly imbalanced consisting of mainly categorical data making it difficult to generalize the model. Further, economic background and previous offence data is not available for fraudulent claimants

Dataset

Original dataset consists of 15420 rows and 33 features. Out of 33 features, only age in years is discrete variable whereas remaining features are categorical variable.

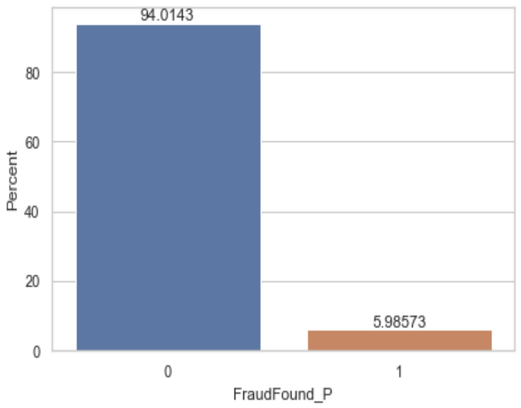


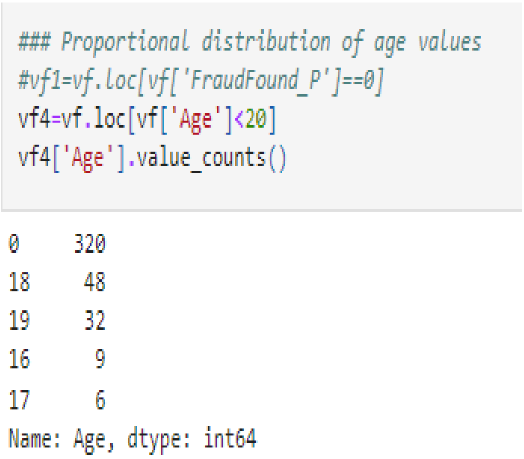
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15420 entries, 0 to 15419
Data columns (total 33 columns):
```

Data Wrangling

During this process, data is checked for missing values, type of data and data discrepancies. The data looked clean without any missing values. The column Policy Number looked randomly distributed and the values indicated it is just a serial number. Target variable ‘FraudFound_P’ looked imbalanced as 94.01% transactions were non-fraudulent. Additionally, age column had

320 values which were 0 and most likely they represented under-age drivers. Other features did not show any major issues.





```
### Proportional distribution of age values
#vf1=vf.loc[vf['FraudFound_P']==0]
vf4=vf.loc[vf['Age']<20]
vf4['Age'].value_counts()
```

```
0     320
18     48
19     32
16      9
17      6
Name: Age, dtype: int64
```

Fig.1 Probablity distribution of target Fraud_Found column

Fig.2 Distribution of age column < 20 years

Exploratory Data Analysis

Eight feature has 50.30% missing values and that column is dropped. Additionally the resorts without any

price information were also dropped. Further, Adult Weekend had more missing values and showed linear

relationship with Adult Weekday prices, it is decided to keep Adult Weekend Price as our target variable

and AdultWeekday column is dropped

24 were numer

●The data set is highly imbalanced

●It does not provide any measurable metrics about the visitors during weekdays and weekends to see the

impact of pricing

strategy.

References:

1. https://www.iii.org/article/background-on-insurance-fraud
2. https://www.progressive.com/answers/car-insurance-fraud
3. Vehicle Insurance Claim Fraud Detection