Vehicle Insurance Fraud Detection

DS105 Project Proposal

Content

- Dataset Introduction
- Problem Statement
- Anticipated Challenges

1. Dataset Introduction

- Vehicle insurance fraud involves conspiring to make false or exaggerated claims involving property damage or personal injuries following an accident
- There can be staged accidents where fraudsters deliberately “arrange” for accidents to occur
- Use of phantom passengers where people who were not even at the scene of the accident claim to have suffered grievous injury
- Make false personal injury claims where personal injuries are grossly exaggerated.

1. Dataset Introduction

- Dataset consists of vehicle and insurance-related details
- Taken from Kaggle (https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection) and is originally a real-life fraud machine learning case study used by Oracle
- Columns : 33
- Rows: 15420

1. Dataset Introduction

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15420 entries, 0 to 15419
Data columns (total 33 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Month                 15420 non-null  object
 1   WeekOfMonth           15420 non-null  int64
 2   DayOfWeek             15420 non-null  object
 3   Make                  15420 non-null  object
 4   AccidentArea          15420 non-null  object
 5   DayOfWeekClaimed      15420 non-null  object
 6   MonthClaimed          15420 non-null  object
 7   WeekOfMonthClaimed    15420 non-null  int64
 8   Sex                   15420 non-null  object
 9   MaritalStatus         15420 non-null  object
 10  Age                   15420 non-null  int64
 11  Fault                 15420 non-null  object
 12  PolicyType            15420 non-null  object
 13  VehicleCategory       15420 non-null  object
 14  VehiclePrice          15420 non-null  object
 15  FraudFound_P          15420 non-null  int64
 16  PolicyNumber          15420 non-null  int64
 17  RepNumber             15420 non-null  int64
 18  Deductible            15420 non-null  int64
 19  DriverRating          15420 non-null  int64
 20  Days_Policy_Accident  15420 non-null  object
 21  Days_Policy_Claim     15420 non-null  object
 22  PastNumberOfClaims    15420 non-null  object
 23  AgeOfVehicle          15420 non-null  object
 24  AgeOfPolicyHolder     15420 non-null  object
 25  PoliceReportFiled     15420 non-null  object
 26  WitnessPresent        15420 non-null  object
 27  AgentType             15420 non-null  object
 28  NumberOfSuppliments   15420 non-null  object
 29  AddressChange_Claim   15420 non-null  object
 30  NumberOfCars          15420 non-null  object
 31  Year                  15420 non-null  int64
 32  BasePolicy            15420 non-null  object
dtypes: int64(9), object(24)
memory usage: 3.9+ MB
```

- There are 8 continuous features and 24 categorical features
- Label – FraudFound_P (0,1)

2. Problem Statement

- In this project, we aim to help the insurance company to filter out potential fraud cases and minimise actual fraud cases
- **End Goal** : Create a machine learning model to predict if a specific vehicle insurance claim is a fraudulent one
- Supervised classification model – predict if case is fraudulent or not (Binary Classification)

2. Problem Statement

Sub-goals:

- How do the features vary for fraud cases ?
- How do the demographics (e.g. Age, Gender, Marital Status) vary with the features for fraud cases?
  - There was a decreasing trend for fraud from 1994 to 1996 - why? Were there a difference in the demographics along the years?
  - As most of the vehicles involved in fraud were priced from $20000–$29000 and mostly Sedan, what were the demographics for this group of fraudsters?
  - For the common car models in fraud cases, what were the demographics like?

3. Anticipated Challenges

- As this is a fraud dataset, dataset is highly imbalanced. Only 6% of the dataset is labeled as fraud cases. Techniques to deal with imbalanced dataset have to be applied to the dataset before ML
- There are many features with multiple categories , thus the right encoding technique will have to be applied
- Appropriate scaling technique will also have to be applied for the categorical features
- Due to the imbalanced dataset, there will be a need to ensure that the number of fraud cases are roughly balanced in the train-test split