UK Road Safety: Traffic Accidents and Vehicles

Vinita Verma
30 Aug, 2020

Business Understanding:

The project intends to predict the accident severity for given road, weather, climate and other conditions among others. The dataset is taken from kaggle- UK roads safety: Traffic accidents and vehicles.

Data understanding:

The dataset contains 1.92 million records and 34 columns. Data cleaning and preprocessed the data. The data corresponding to slight severity is 84.84%, serious severity is 13.86% and for fatal severity is 1.30%. Several processing techniques as undersampling and oversampling were tried. Our main aim is to predict the serious and fatal severity with high precision and F1 score.

Data Source:

The data collected comes from the U.K. government who amassed traffic data based on police reports. The analysis of data executed here is composed of the U.K. road accidents from 2014 to 2016.

https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents

Accidents are recorded according to these features:

- Reference Number

- Grid Ref: Easting
- Grid Ref: Northing
- Expr1
- Severity
- Day of the week
- Time (24hr)
- 1st Road Class
- Road surface
- Accident date
- Weather condition
- Lighting conditions
- Number of vehicles
- Casualty class
- Sex of casualty
- Age of casualty
- Type of vehicle

Data Pre-Processing:

•Data missing values are imputed by the most frequent value of the column

•Categorical data labelled with numerical values

•Merged similar categorical values

•SelectKBest: provides the k best features by performing various statistical tests i.e., chi squared computation between two non-negative features

•RFE(Recursive Feature Elimination): Recursively eliminates the features which does not in target variable values

•Merged Serious and Fatal classes as Serious class

Methodology:

1.Algorithms Used:

1. K- Nearest Neighbor
2. Naïve Bayes
3. XGBoost
4. Random Forest
5. GBM
6. SVM
7. Logistic Regression

2. Handling Imbalance Data:

•Over Sampling

•Under Sampling

•Mis-classification penalty

•Ensemble methods

Conclusion:

In conclusion, most of the algorithms are biased towards most frequent class. However, efficient pre-processing and corresponding imbalanced data techniques should give optimal results.