IBM – Data Science

Applied Data Science Capstone

The aim of this project is to define a problem statement, which will be analyzed and solved, further enhancing the knowledge and skills and attaining a certificate as a Data Scientist through IBM Coursera.

Introduction

The 8th largest cause of Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030.

Factors like the weather condition, roadworks, traffic jams need to be analyzed inorder to accurately predict the severity of accidents.

Analyzing of this data could provide non-profit, public organizations to allow effective and efficient allotment of resources and help prevent future incidents. Furthermore, the knowledge of all incidents can be effectively transmitted and can be warned to drivers so as to make them aware in the route or make them avoid the route.

City governments should be interested in accurate predictions of accidents on the route, so as to reduce time interval leading to a reducing of a significant amount of people each year. Technologies companies can play a key role to improve road safely.

The study consists of a notebook. This notebook contains the steps and transformations performed for the feature selection. The information on the raw data in the following kaggle page. Kaggle datasets usually contain an extended descriptions of different aspect of the car accidents, which are utilized to perform the analysis.

Predicting Traffic Accident Severity

Data Description -

Data Cleaning -

EDA -

Model Development -

Random Forest -

Logistic Regression -

KNN -

SVM -

Results -:

Algorithm	Jaccard	f1-score	Precision	Recall	Time(s)
Random Forest	0.722	0.72	0.724	0.591	6.588
Logistic Regression	0.661	0.65	0.667	0.456	6.530
KNN	0.664	0.66	0.652	0.506	200.58
SVM	0.659	0.65	0.630	0.528	403.92

Within this problem precision stands for the % of predicted severe accidents. The recall is % of proper prediction of severe accidents. Making recall more important than the precision and help equip and allocate adequate resources. Thus the recall becomes more important than precision as a high recall will favor that all required resources will be equipped up to the severity of the accident. Accuracy with The KNN, SVM and logistic regression models have around the same accuracy, however the computational time differs for one when compared to the other two models. The Random Forest is the best model, in the data set and the similar as the log. res. it improves the accuracy ranging between 0.66 - 0.72 and the recall ranging between 0.45 - 0.59.