Final Project Report

Prediction of Accident Severity

By Shraddha Nimbalkar

IBM Data Science Professional Certificate

Section 1: Introduction

Motivation

Road accidents are a serious concern for the majority of nations around the world because accidents can cause severe injuries and fatalities. According to the World Health Organization's Global Status Report, approximately 1.25 million people deaths happened per year are because of road accident injuries, and most fatality rates were in lower income countries [1]. Our motivation is to predict the accident severity of any road, which will play a crucial factor for traffic control authorities to take proactive precautionary measures. In addition, the dataset we chose was rarely solved in prediction point of view, so we took this opportunity to predict the severity of the accident.

Objective

The purpose of this project is to predict the severity of an accident by training an efficient machine learning model with the help of existing accidents.

This machine learning pipeline will be exploring whether weather, light, and road conditions – that increase the amount of information available about the situation – can increase the accuracy of predicting the severity of an accident.

Problem Statement:

The target or label columns should be accident " severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact.

The machine learning model should be able to predict accident "severity"

Section 2: Discussion

Algorithms, technologies, and tools & classification algorithms were used and evaluated to predict the accident severity. Algorithms considered for classification were K-nearest neighbours, Naive Bayes classifier, Random Forest classifier, Logistic Regression, Gradient Boosting classifier, SVM. The first thing that came to our mind is that severity is based on different decisions like road, weather conditions. So, we tried different decision tree classifiers. We also tried ensemble methods as the data is very imbalanced.

We are using following algorithms

K-Nearest Neighbours Classifier

Naive Bayes Classifier

Random Forest classifier

Logistic Regression

Support Vector Machine

Technologies & Tools used

For developing this project, below tools and technologies have been used.

Python,Jupyter Notebook

Dataset used:

Data has been fetched from Open Data platform UK. It Captures Road Accidents in UK between 1979 and 2015 and has 70 features/columns and about 250K rows. The Data set contains column including weather conditions, Road class, road type, junction details, road surface conditions, light conditions, etc.

Data Visualization

We have created Heatmap for Plotting importance of each feature

The next step was to normalize the only features that were not categorical: age of the driver and age of the car. Normalization involves taking the logarithm of the given features. This is done to because high values for certain variables computationally skew results more in favour of that variable, than their actual contribution. In this case, age of the driver for example has values ranging from 18-88. When majority of other categorical variables are binary or limited within 1-8 categories

In this case, age of the driver and age of the vehicle were the only variables with a high numerical variance, and therefore logarithms were taken of both variables. Furthermore, taking the logarithm of both the age of the driver and age of the vehicle improved the fit by altering the scale, and making the variables more "normally" distributed**.**

After taking the log, one can notice that the values range from approximately 2.5 to 4.5. This increases the performance of machine learning algorithms, as the numerical values do not have disproportionate amounts of computing value compared to all the other categorical variables.