Analysis of UK Traffic Accidents

1. Introduction

1.1 Background

Road accident Causing enormous losses it did not affect only those who are victims. But it also causes impact to the economy and society. Because of the deaths and injuries both physically and mentally from the accident Causing the victims and their families to lose their productivity Affect the overall productivity of the country

Accidents also generate other costs such as legal costs. Costs from impact on traffic conditions, etc. The economic and societal impact of traffic accidents cost British citizen million British pounds every year.

1.2 Problem

It would be great if we can find what are the most common causes, to decrease or prevent the accidents. we might be able to make well-informed actions and better divide financial and human resources.

1.3 Interest

The data comes from government website www.data.gov.uk. UK police forces collect the accidents data using the form called Stats19. This project aim to analysis on U.K accidents data from year 2015 to predicting the accident severity.
Government who want to decrease prevent cost. Insurance company would be very interested in accurate prediction of the accident severity, for Insurance premium and risk management. And a driver who want to increase they chances of staying safe on the road.

2. Data acquisition and cleaning

2.1 Data sources

Road Accidents and Safety Data comes from government website www.data.gov.uk. This data is consist of 3 files that are Accident Circumstances, Vehicle and Casualty. Every column of the dataset is in numerical format. With a variable lookup to explain each numerical category in accidents dataset was provided on the www.data.gov.uk website as well.

This dataset can answer some questions like

- what the most accident occour on the days of a week?
- what time that had the highly accident? -
- what the most age of driver that involved in accident?

we can find this answer by using data Visualization

2.2 Data cleaning

Data downloaded and 2 files Accident Circumstances and Vehicle were joined into one table. Because I think that was the internal and external factors to cause accident. There is one problem with the datasets. This data had 2 types of missing values '-1' and 'Nan'. I did not imputing any mean or median value as the dataset are big enough to execute analysis.

2.3 Feature selection

After data cleaning, there were 70,383 samples and 53 features in the data. When I inspected the correlation of independent variables, most of them not have strong correlations between each other feature. I also consider some features to predictors for machine learning algorithm. After all, 11

features were selected as day of week, speed limit, light conditions, weather conditions, road surface conditions, did police officer attend scene of accident, vehicle type, sex of driver, age of driver, engine capacity (CC), age of vehicle