Accident Severity Prediction

Pradiv Gnanaraj

13/10/2020

1. Introduction

1.1 Background

Wheels getting invented during the 3,500 B.C from caveman technology to 1886 when Carl Benz applied for vehicle powered by gas engine until today with Self driven cars, travel becomes unavoidable. With the increased number of vehicles used for road transport, travelling has become more frequent. Resulting in accidents from time to time. The number of Fatal cases seems to increase in number every year. The road safey team from every country have passed laws and rules to reduce death by road accident. However, this seems to be still a problem.

1.2 Problem

Data that might include another point of view and possibly a solution to reduce the number of deaths is what required to be created and following by every government. Data which includes the road, weather condition, light conditions can be learnt which you be possible method to reduce the accidents. This project aims to predict the natural conditions to reduce the accident severity based on these data.

1.3 Interest

How well this model could be if we can know the possibility in predicting the accident before we could travel. This will safe life and definitely will be implemented by governments.

2. Data acquisition and Cleaning

2.1 Data Sources

In order to work with data which has already collected these information, I found and good dataset which was uploaded to Kaggle as two files. First file itself for the dataset and the second file was a key to the true dataset, explaining the attributes and values more in detail.

Primarily Captures Road Accidents in UK between 1979 and 2015 and has 70 features/columns and about 250K rows.

2.2 Data Cleaning

There were number of items to be checked with the uploaded dataset. I started with finding the attributes and get a statistical summary of all numeric-typed attributes. Then, I wanted to ensure that the data is converted to a format which is best understand during analysis. Prior goal during this process is to : handle all the missing values, correct the data format, standardize and normalize the data.

I created a small script – a python for loop to find the total number of missing data. The missing data was quite large nearing 14,06,982 belonging to various attributes. All of the missing data in the dataset were of dtype:int64 this lead to dealing with the missing data.

I could split the missing data into following types,

- Data to be dropped by Row.
- Data to be dropped by whole column.
- Replace the data with mean.
- Replace the data with frequency.
- Replace the data based on its other functions.

Dropped Values	Replaced Values	Reasons
location_easting_osgr, location_northing_osgr, longitude, latitude, time		For the following set of columns the missing rows are very less, therefore we will drop the following rows
	casualty_class, sex_of_casualty, age_of_casualty, age_band_of_casualty, casualty_severity, pedestrian_location, pedestrian_movement,	Replaced the missing values with the mean
	casualty_class, bus_or_coach_passenger, pedestrian_road_maintenance_worker, casualty_type, casualty_home_area_type	Replaced the missing values with the mean
	casualty_imd_decile	Replaced the missing values with the mean
casualty_reference		drop this column as we have much clearer data from the casualty_severity column
lsoa_of_accident_location		drop this column as we are focusing on the accident serverity.

Indicator Variable(s)

To use categorical variables for regression analysis in the later part of model development

Selecting only the required columns for the serveirty analysis.

Splitting the variables in to three variables categories

- Accident Circumstances
- Vehicle type
- Casualty type

Before starting to understand the (linear) relationship between an individual variable and the target variable, I worked with descriptive statistical analysis. And also examining the value counts for these attributes could be a good predictor.

Before starting to feed the data to the model (to be developed). Finding the Pearson Correlation Coefficient and the P-value between attribute(s) and the target to find the linear distribution.

Conclusion

The correlation between weather_conditions, light_conditions, speed_limit, road_surface_conditions, special_conditions_at_site, day_of_week and casualty_severity attributes are sigificant.

The linear distribution are weak for the following:

weather_conditions

ligh_conditions

speed_limit

road_surface_conditions

special_conditions_at_site

day_of_week

However, a weak correlation can be statistically significant, as the sample size is large enough

Gauss Markov assumptions- Collinear variables change at the same time, and therefore it is difficult to assess each variables distinct effect on the outcome variable.