Predicting the car crash severity

CHAU IP SING

1 October, 2020

1. Introduction

This project is going to predict an car incident on the road by the different attributes, such that driver can avoid to drive in the same way. This project will build a predicting system to identify the factors which are effecting the traffic crash severity. If there are AI deploy in the car in future, the car can give advance to driver, such that to avoid accident happens.

1. Data acquisition and cleaning

The data source is downloaded from the course. The dataset has incident records start in 01/01/2004 and end in 24/04/2020 in the area of Seattle.

The dataset has 38 attributes and 194673 rows of records.

There are missing data in different attributes, such as X, Y, ADDRTYPE, INTKEY, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, COLLISIONTYPE, JUNCTIONTYPE, INATTENTIONIND, UNDERINFL,WEATHER, ROADCOND, LIGHTCOND, PEDROWNOTGRNT, SDOTCOLNUM, SPEEDING, ST\_COLCODE and ST\_COLDESC.

As some of the attributes are description or conclusion from the SDOT. I finally determine to use only the attributes of 'SEVERITYCODE','X','Y','ADDRTYPE','COLLISIONTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','INCDATE','INCDTTM','JUNCTIONTYPE','SDOT\_COLCODE','UNDERINFL','WEATHER','ROADCOND','LIGHTCOND','SPEEDING','ST\_COLCODE','HITPARKEDCAR'.

As the data of 'WEATHER','ROADCOND','LIGHTCOND','JUNCTIONTYPE','COLLISIONTYPE','ADDRTYPE' are objects, I change them to new columns by get\_dummies function.

Also, the incident time of the date may be also affect the severity and the number of incidents occurrence, so I extract the incident time to be hour in an additional column for easy handling.

Chart, histogram

Description automatically generated

As expected, the number of incidents is rather low in deep night and strange that there is no incident records in the period of 11:00-12:00 for those filtered records.

Finally, there are 7297 rows of records are useful.

1. Exploratory Data Analysis

For exploratory the data, first of all, I use the correlation to check the relationship of the data to each others.

A close up of a logo

Description automatically generated

According to the above data, it is not obvious to see the relationship of severity to any of the attributes. On the other hand, some attributes have close relationship as expected, e.g. number of person in the incident to the number of vehicles.

Although it is not obvious, it is higher corelations on the severity due to parked car or cycles compare to others collision type.

Also, all incidents are having SPEEDING issue.

In the very beginning, I think that there may be some hotspots of incident location. But, after plotting the locations on the map, it seems it is quite huge of locations or even distributed on the same road, such that there are a number of different incidents. For the interest of driver, it seems not much useful when driving in high speed.

Map

Description automatically generated

1. Predictive Modelling

As the target is to predict the severity class of a incident, I choose the KNN algorithm. I split my dataset into 80% for training data and 20% for testing data. I find that when using k = 17, the accuracy is highest on hitting 0.619.

Chart, line chart

Description automatically generated

1. Results and Observation

In this study, I identify the SPEEDING, ADDRTYPE, COLLISIONTYPE, LIGHTCOND, WEATHER, VECHICE COUNT, PEDESTRIAN COUNT attributes will mainly effect on the severity classification.

1. Conclusions

I was able to achieve ~62% of accuracy on the classification. Although there are many different external attributes, the driver’s driving year experience may be a good indicator to affect the car incident. So, if possible, to include the driver experience year is a good choice to improve the prediction accuracy.