THE CAPSTONE PROJECT REPORT

PREDICTING THE SEVERITY OF THE COLLISION

By Nguyễn Tường Quang

September 7, 2020

1. Introduction

Traffic accident is something no one wants, but also a part of modern life. Traffic accidents do not happen naturally, but have causes. If an accident can be predicted, from the initially assumed causes, we can adjust everything related to reduce the accident.

This is the goal of this project.

Under the aspect of data science, we should must collect data related to vehicle accidents in history, as much as possible, and then analyze the data to build a suitable predictive model.

If it is possible to build a good predictive model, that is, it is possible to establish a link between the input causes and the severity of the accident, we will be able to predict the accidents in future.

2. Data

2.1 Data requirements

The prerequisite must be real data, recording many aspects related to accidents, such as location, date and time, road conditions, weather conditions, lighting conditions, number of people involved in an accident, number of vehicles and type of vehicle crashed ... And, ultimately and most importantly, the consequences of the accident. The aftermath of an accident must be fully recorded on the damage to property as well as human life. This is the target of classification in this project.

The next condition is that the data must be large enough, preferably recorded for many consecutive years in the same region or city, country ...

The third condition is that data must be recorded in a structured and clear way.

2.2 Data sourses

There are many datasets that meet the above requirements. For this project, we used the City of Seattle dataset, recorded from 2004 to 4/2020. This dataset is shared by IBM Data Science course.

2.3 Data descriptions

The dataset consists of nearly two hundred thousand rows, with 38 columns. In which, the first column is about the severity of the collision, which is the target for classification in the model. Consists of two serious levels: 1-prop damage and 2-injury.

The remaining 37 columns, in addition to the separate record columns of the state, are the remaining data that can be used as inputs to build the model. These columns are entirely categorical variables. Include:

- Type of collision

- Total number of people involved in a collision

- Number of pedestrians involved in a collision

- Number of cyclists involved in a collision

- Number of vehicles involved in a collision

- Traffic configuration where collision occurred (junctiontype)

- Weather conditions at the time of the collision

- Road conditions at the time of the collision

- Light conditions at the time of the collision

- The date and time of the collision

Which features to choose as input for the predictive model will depend on the relationship between that feature and the target (i.e the severity of the collision). This will be achieved after exploring data analysis and preprocessing step.