IBM Data Science Professional Certificate

Course 9: Applied Capstone Project

Week 2 Submission

Accident severity prediction using classification algorithm for Seattle city

Student Name:	Shivpalsinh Rana
Submitted on:	21.09.2020

Table of Content

1 Introduction

In this section we will discuss the business understanding to seek clarification for the problem.

1.1 Business Understanding

Automobile is the most popular means of transport in the USA. On daily basis, hundreds of thousands of people travel from A to B using the road infrastructure. One of the less fortunate side or aspect of road transportation is the road accidents. Road accidents not only causes injuries to people, deaths but also results in loss or damage of properties worth millions of Dollars. Accidents also hinders the traffic flow which results in loss of time for many commuters. Unnecessary congestion and stop and go traffic built-up cause the vehicle to run inefficiently and emit more pollutants.

There are number of factors which influence the cause of road accidents. Some of the factors are weather condition, road condition, attention of the driver, condition of driver, light conditions etc. In this project, we try to predict the severity of the accident given the road condition, weather condition, type of accident and many more influencing factors. To effectively predict the severity, we will be using specific machine learning algorithms.

Following stack holder can be benefited from an effective machine learning model which can predict the accident severity:

1. People who are commuters and would like to plan their travel. For example, in a situation where weather is windy or rainy, the road conditions are bad and light condition is also not ideal, Commuter can obtain the severity of an accident and can avoid the traffic congestion and drive more cautiously.
2. Secondly, the government regulatory bodies can take precautionary steps, to avoid or to reduce the fatality of an accident.

In the following section we will be looking for the data which can be used to train the machine learning model and help us to solve the problem.

2 Data Understanding

Here, we will discuss about the data set selected for this project and investigate the data.

2.1 Data Source

The data set used for the project is procured from the Seattle Department of Transport and consists of the vehicle accident information. The data set is in .csv format with 38 columns and a total of 194673 entries or rows. Overall, the data seems to be rich, has many of the observations and the attributes which are suitable for training a machine learning model.

2.2 Data Description

The label for the selected .csv data set is the column ‘SEVERITYCODE’. Each row in the data set has give a severity code which is either ‘1’ or ‘2’ meaning ‘property damage only collision’ and ‘injury collision’, respectively. The rest 37 columns contain various attributes in which not all of them are useful.

By examining the number of 1s and 2s in the label of the given data set, we see that the data is has an imbalance i.e. there are more rows with 1s compared to 2s. Such unbalanced data can result in the bias behavior of the machine learning model.

As we have observed, there are many irrelevant attributes in the data set which are not useful in training a machine learning model. Moreover, like most of the data sets available, there are some columns with missing values or inconsistent values. Thus, the data must be cleaned in such a way that it avoids the irrelevant attributes and increase the accuracy of the machine learning model without any bias.