Car Accident Severity Analysis: Seattle, Washington

(Applied Data Science Capstone)

The project aims to understand the factors which play a role in the severity of accidents using Machine Learning Models.

SUBMITTED BY : SIDDHARTH M DHELIA

DATE: 10th OCTOBER 2020

GITHUB: https://github.com/siddhelia/Coursera_Capstone

Contents

1.Introduction

Motor vehicle accidents continue to be one of the leading causes of accidental deaths and injuries in the United States. It is estimated more than six million car accidents occur each year in the U.S according to the NHTSA and about 6% of all motor vehicle accidents in the United States result in at least one death.

Roughly 27% of all vehicle accidents result in nonfatal injuries. However, some of these injuries can cause tremendous pain or lead to permanent disabilities.

In this project we will leverage the accident data of "Seattle city" to predict the different accidents' severity.

Our project can be a valuable asset to Governments, states, provinces and municipalities which could use our model to not only prevent road accidents,but also to identify key factors that can lead to a road accident,and consequently, help elaborate new policies.

Goverment agencies can also use this data to warn you, given the weather and the road conditions about the possibility of you getting into a car accident and severity of it.This can enable the driver to drive more carefully or even change his travel plans.

2.Data

Based on definition of our problem, factors that will influence our decission are:

* Number of people and number of vehicles involved in the accident.

* Location,whether,road and light conditions are also infudencial factors.

* Speed of car and junctions are another factors we would consider.

The Following data source will be needed to extract/generate the required information:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The dataset contains details of car accidents which have taken place within the city of Seattle,Washington in the past fifteen years(2004-2020).The dataset is highly extensive and contains all details of above mentioned factors.

Data Cleaning

There are a lot of problems with the data set keeping in mind that this is a machine learning project which uses classification to predict a categorical variable. The dataset has total observations of 194673 with variation in number of observations for every feature. First of all, the total dataset was high variation in the lengths of almost every column of the dataset. The dataset had a lot of empty columns which could have been beneficial had the data been present there. But since we have have large enough data base,missing entries where removed .Also columns which had no impact in severity of the accidents where omitted from the report and only select few which had most impact on severity of accidents were considered in the study. In order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had ‘Other’ and ‘Unknown’ in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

Feature Selection

A total of 8 features were selected for this project along with the target variable being Severity Code.

FEATURE VARIABLES	DESCRIPTION
UNDERINFL	Whether or not the driver was under influence. (Y/N)
WEATHER	Weather Condition During Collison. Clear/Rain/Snow
ROADCOND	Road Condition During Collison. Dry/Wet/Snow
LIGHTCOND	Light Condition During Collison. Day/Lights-off/Lights-On
SPEEDING	Whether driver was above speed limit causing accident (Y/N)
JUNCTIONTYPE	Whether Accident occurred at Intersection .(Y/N)
NO. OF PERSONS INVOLVED	Number of people involved in the accident.(1,2,3)
NO. OF VEHICLES INVOLVED	Number of Vechicles involved in the accident.(1,2,3)

The models aim was to predict the severity of an accident, considering that, the variable of Severity Code was in the form of 1 (Property Damage Only) and 2 (Injury Collision) which were encoded to the form of 0 (Property Damage Only) and 1 (Injury Collision). Furthermore, the Y was given value of 1 whereas N and no value was given 0 for the variables, “Speeding” and “Under the influence”. For lighting condition,Day Light was given 0 ,with Street Light ON as 1 and Street light off as 2. For “Road Condition”, Dry was assigned 0, Wet was assigned 1 and Snow/Ice was given 2. As for “Weather Condition”, 0 is Clear, rain/overcast is 1, Snow/hail is 2. If the accident was caused at intersection the value given was 0 and if not than 1 in “Junction Type”.Also no of persons and vehicles involved where also considered.A visual representation of each feature Variable –“value count” is given in the following page.