**CAR SEVERITY REPORT**

Mohit Singh

4 September 2020

**Table of contents**

* Introduction:
* Business Problem -
* Data Understanding
* Methodology
* Analysis
* Results and Discussion
* Conclusion

**Introduction: Business Problem**

**BACKGROUND**

Accidents in traffic lead to associated fatalities and economic losses every year worldwide and thus is an area of primary concern to society from loss prevention point of view. Modelling accident severity prediction and improving the model are critical to the effective performance of road traffic systems for improved safety. In accident severity modelling, the input vectors are the characteristics of the accident, such as driver behaviour and attributes of vehicle, highway and environment characteristics while the output vector is the corresponding class of accident severity.

There are two main engineering approaches for dealing with traffic safety problems: the reactive approach and the proactive approach. The reactive approach, or retrofit approach, consists of making the necessary improvements to variable, for instance, existing hazardous sites in order to reduce collision frequency and severity at these sites. The proactive approach, on the other hand, includes a collision prevention approach, like, preventing a potential unsafe road conditions from occurring in the first place. We focus on proactive approach which involves prediction of accident severity and working backwards, the concerned entity implements appropriate remedial measures to improve road safety. By recognizing the key factors that influence accident severity, the solution may be of great utility to various Government Departments/Authorities like Police, R&B and Transport from public policy point of view. The results of analysis and modeling can be used by these Departments to take appropriate measures to reduce accident impact and thereby improve traffic safety. It is also useful to the Insurers in terms of reduced claims and better underwriting as well as rate making.

**INTEREST**

In this project we will try to find the fatalities or severity occurred in an accident. Specifically, the report will be helpful to government officials and highway authorities in predicting the fatalities of an accident according to the given weather conditions, road conditions and visibility conditions of streetlights and collision address type. When the condition are bad, alert drivers to remind travelling people to be more careful and divert the route if any accident on their way of destination.

We will use data science powers to generate a model which will predict car accident probability using the conditions mentioned in the business understanding section.

**DATA UNDERSTANDING**

Based on the definition of our problem, factors that will influence our decisions. Our predictor variable or attribute which will be 'SEVERITYCODE' as mentioned in our metadata of certain city.

* 3- fatality
* 2b - serious injury
* 2 - injury
* 1 - prop damage
* 0 - unknown

Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND' in this project.

WEATHER- Weather condition during collision

ROADCOND - Condition of road during collision

LIGHTCOND - Condition of streetlights during collision

Collision address type (ADDRTYPE)

- Alley

- Block

- Intersection

According to collision address type we will define if any accident occurred we will be able to take the other route if possible.

**Data preparation**

The data preparation includes all the required activities to construct the final dataset which will be fed into the modelling tools. Data preparation can be performed multiple times and it includes balancing the labelled data, transformation, filling missing data, and cleaning the dataset. For this project we will be using the data set of SEATTLE city of WASHINGTON. The dataset which we are considering in unbalanced so we will drop the unwanted variables from our main dataset and create a new one considering the variables or attributes which are mentioned in the Data Understanding section. In this project the dataset we are using is collected from sample dataset on COURSERA from where I have acquired this CAPSTONE project. Here is the link - You can find the Example Dataset by [Clicking here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). You can also find the Metadata by [Clicking here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf)

**METHODOLOGY**

**Loading the dataset and extraction of data**

Here I am going to use some of the libraries of python like numpy, pandas and matplotlib. So here is first look of our dataset after opening it in python.

A screenshot of a cell phone

Description automatically generated

I also perform some of **exploratory data analysis** of our dataset.Here is Descriptive summary of our dataset some of statistical features of our dataset.A screenshot of a cell phone

Description automatically generated

After this step I tried to extract main features of data set. These feature help me to build the predictor model of severity based on these conditions.Firstly I removed the **NaN values** from dataset and an undesired variable **Unknown** variable from our dataset.We fill the empty cells of our dataset with this variable and drop the rows containing this **Unknown** vaiable. After this we got our dataset where our predictor variable SEVERITYCODE .

#### Feature engineering

I have some of the feature engineering to change the categorical variable so they can be fed into our model. In this project we have used LABEL encoding.A screenshot of a social media post

Description automatically generated I have converted the categorical variables using **Label Encoder** as sklearn libraries do not work on categorical variables. I have also performed some of data visualization techniques for analysis.

A screenshot of a cell phone

Description automatically generated

Here I am going to show certain effects of features on different SEVERITYCODE using bar graph

This graph shows that category 1 of SEVERITYCODE is maximum when the collision address is **BLOCK**. It means road is blocked when accident has done some property damage.

A screenshot of a cell phone

Description automatically generated

This graph shows that most of severity is type 2 which is **INJURY** the road condition is DRYA screenshot of a cell phone

Description automatically generated

This graph shows that when the LIGHT condition is daylight when the severity of type 1.

## Model Development

## A model can be thought as a mathematical equation used to predict a value given one or more other values, relating one or more independent variable to dependent variables. In development we are going to use Decision Tree machine learning which is supervised as our data is labelled. We will our dataset in training and testing subset.Here we are using 30% of data in test case of .the model.

## DECISION TREE

## A decision tree model gives us a layout of all possible outcomes so we can fully analyse the consequences of a decision. It context, the decision tree we observe the different condition effecting the SEVERITYCODE.

## Here we are going to divide our sample space into training and testing set using the test/train function repeating the process for 3 times as random seat value is 3. We also check the dimensions of our training and testing set as they match.

## A screenshot of a cell phone Description automatically generated

## MODEL PREDICTION

## Here we are going to make our prediction using testing or Out of Sample set.

## A screenshot of a social media post Description automatically generated

## 

## Here we are using Model Evolution in which we are measuring the accuracy of model by comparing our model prediction with our testing set result and our model gives 67% accuracy.

## Here is our DECISION TREE MODEL

## 

The above diagram is Decision Tree of car severity dataset with ADDRTYPE as its main attribute. The Attribute is selected as the it give less entropy and more information gain.

## Discussion & Conclusion

In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algoritim, so label encoding was used to created new classes that were of type int8; a numerical data type.

After solving that issue we were presented with another - imbalanced data. As mentioned earlier, class 1 was nearly three times larger than class 2. The solution to this was downsampling the majority class with sklearn's resample tool. We have change the categorical value.

Once we analyzed and cleaned the data, it was then fed through ML model;Decision Tree because of its binary nature.

Evaluation metrics used to test the accuracy of our model. Choosing values evalution metrices to improve our accuracy to be the best possible.

Based on historical data from road block conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2). and address collision type is best predictor for model.