**Car Accident Severity**

*Capstone Project for IBM Data Science Professional Certificate on Coursera*

# Table of Contents

* Introduction / Business Problem
* Data Description
* Methodology
* Analysis
* Results and Discussion
* Conclusion

# Introduction / Business Problem

#### **Background and Problem**

Each year about 1.25 million people die in road traffic accidents, and an additional 20-50 million are injured or disabled. If the locations of traffic accidents could be predicted, this could have a huge beneficial impact in potentially helping to reduce the number of accidents each year. For example, routing software could avoid the most dangerous areas - particularly in the context of the coming advent of driverless cars. It could also be useful in an insurance context, in order to predict risk, as well as for governments and local road authorities looking to create more efficient systems of road maintenance and improvements. The aim of this project is to predict where traffic accidents are likely to occur. Throughout this capstone project, multiple supervised machine learning algorithms will be explored. Then, conclusion will be made by selecting the machine learning model that provides the highest prediction accuracy.

#### **Audience**

The primary audience of this study might include people who resides in Seattle, WA, and those that are interested in learning about collisions provided by SPD which are recorded by Traffic Records. The findings could also be used by Department of Transportation Traffic Management Division looking to open new methods of even a way of fostering and improving traffics within Seattle to improve safety.

# Data Description

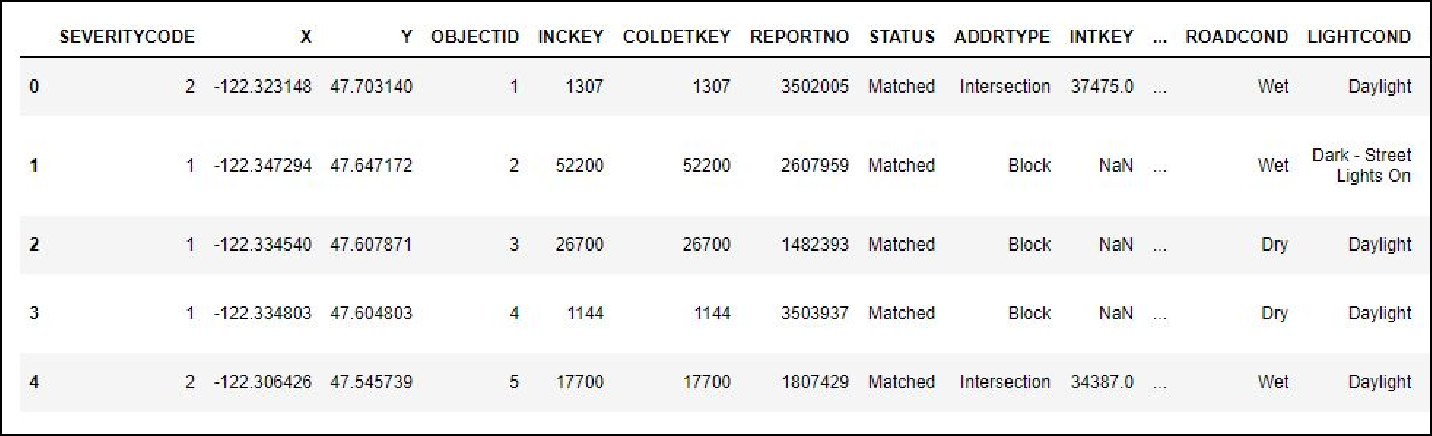
The data includes all types of collisions. Collisions will display at the intersection or mid-block of a segment during timeframe: 2004 to Present. SDOT GIS Seattle (Dataset and Metadata) provided the data set used in this study. It contains speed, light, condition of the road, severity, etc. for road accidents from 2004 to present. The intention is to use multiple machine learning methods to forecast the severity, consideing the various road conditions. The data to be used in the sample data collection provided is called "Data-Collisions.csv" will be used.

The data we will be using for the project is from **Seattle, Wasington, US**, named as “Data-Collisions.csv” provided by-"SDOT GIS Analyst”. It has stored data from the year 2004-Present. It is a large dataset with dimension **193673 x 38** to work on. It has a special column showing the Severity of the collision which can be used for training and predicting the model.

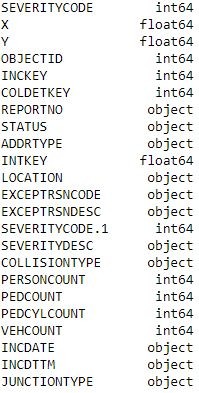
## Some Data attribute description:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data type, length** | **Description** |
| OBJECTID | ObjectID | ESRI unique identifier |
| ADDRTYPE | Text, 12 | Collision address type:Alley,Block,Intersection |
| LOCATION | Text, 255 | Description of the general location of the collision |
| SEVERITYCODE | Text, 100 | A code that corresponds to the severity of the collision |
| COLLISIONTYPE | Text, 300 | Collision type |
| PERSONCOUNT | Double | The total number of people involved in the collision |
| PEDCOUNT | Double | The number of pedestrians involved in the collision |
| PEDCYLCOUNT | Double | The number of bicycles involved in the collision. |
| VEHCOUNT | Double | The number of vehicles involved in the collision. |
| INCDTTM | Text, 30 | The date and time of the incident |
| WEATHER | Text, 300 | A description of the weather conditions during the time of the collision. |
| ROADCOND | Text, 300 | The condition of the road during the collision. |
| LIGHTCOND | Text, 300 | The light conditions during the collision. |
| PEDROWNOTGRNT | Text, 1 | Whether or not the pedestrian right of way was not granted. |
| SPEEDING | Text, 1 | Whether or not speeding was a factor in the collision. |

After importing the data from the source, taking a first look at first 5 rows of the data

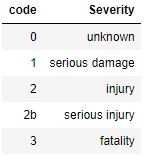


Data Visualization and selection

Let’s check the dimension of the data, which was found to be (194673, 38).

Later checking the description of each attribute, their data type and their type of data in them.

By looking at the data type we can see the data they have and what we need for the further evaluation.

The main data that we are concerned with is the severity code, this is the data that we will be predicting. Severity code is stored with a code data and this is its meaning of each numeric code.

# Methodology

The methodology is divided into three parts:

* **Data Preparation:** presents how the dataset is obtained as well as some key aspects of the data.
* **Descriptive Statistics:** provides a brief overview of the statistics and graphs used in order to have a better understanding of the collision data. The descriptive analysis is shown in the result section.
* **K-Means:** explains the parameters and how the optimal number of cluster was obtained.

Mainly, I will create a Machine Learning model for the prediction of the severity of an accident only by getting some description data of the situation and later plan our future plan.

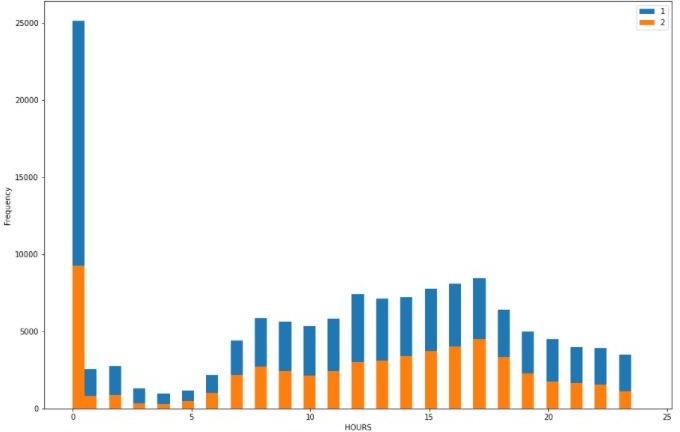
* First, separate between necessary data needed for model creating and training and the non-essential data for the model;
* Second, remove all the NaN(not a number) values and replace them with appropriate solutions;
* Third, take care with the non-accurate data types and get them back like with the date case(it's datatype is object);
* Fourth, convert the categorical data with numerical data - simple ones with just replacing commands and the difficult ones with one hot encoding method;
* Fifth, try to find some pattern by splitting the datetime into daysofweek, month and hours. directly or by plotting a graph;
* When satisfied with the data selected we will split it into train and test data;
* Train decision tree model with training and predict the test data and calculate the accuracy of our model;
* If the f1-score is satisfactory, train final model with whole data, as every data is important;

# Data Analysis

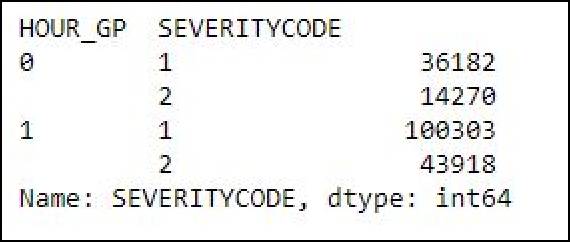
Firstly, we created a temporary copy of the original data with the necessary attributes that we thought would be necessary just by looking at their data type and the data. Later they will be processed and more will be added or removed by taking into consideration of aur needs and which seems to be best suitable for our model.

NaN values could be a real mess to a classification model, so it is mandatory to remove them ASAP. There are many ways to do this like deleting rows, replacing them with the mean of the whole column, filling them with the nearest neighbour, or in classification problems filling them with most abundant data. So , we used filling the most of the columns with the mode of that column and the columns which were having binary values like yes(‘Y) and no(‘N’), they were only filled the yes values, so the NaN were replaced with 0( later the Ys were converted to 1 too in converting categorical data to numerical). So to achieve a completely filled dataframe.

With plotting it was found the days and months have an average rate, no peaks and slopes, so they are not good to be considered. While plotting the hours it was found that



There was a peak at 12am. So the data was split in ones and zeros, where ones being the time 12am and from 9am to 6 pm, rest were turned zeros as they do not have much info.



Now again a final feature selection took place to select only necessary data

With one hot encoding the data for WEATHER, ROADCOND, AND LIGHTCOND was split into dummies

Just a section of table to show what one hot encoding did.

The data was split into X→ containing all the data that we will provide to the model for prediction y→ containing the data that our model will predict.

X data have a variety of data, so to make it normal, the data is normalized which pushes the mean of data to zero and variance to 1.

Post normalized data → array([[ 1.36398593, -0.33020207, -0.18743029, -0.16958841, 0.12553783,

-0.88333125, -0.41751024, -0.42518348, -0.2217116 , -0.1567239 ,

-0.22440165, 0.59145941, -0.01696304, -1.21707436, -0.05414257,

2.45445634, -0.00506801, -0.45298634, -0.011333 , -0.02409974,

-0.06841713, -1.4099744 , -0.07905204, -0.01813462, -0.01963186,

-0.07200071, -0.02431221, 1.76085874, -0.08920831, -0.07872239,

-0.576075 , -0.00751719, -0.1141037 , 0.77768637, -0.17682024]])

# Model creation

Decision tree data machine learning classifier was used to create a prediction model for the given data.

Because the decision tree is fast and processes a well accurate model. As our data had a dimension of 194673 x 35, training with any other machine learning was not proceeding and the notebook plus kernel kept crashing. So our model should be fast with accuracy to the decision tree provided in that environment.

Train Test split - The data was primarily splitted into train and test data. Why?

To get the best accuracy, the model should be tested on unknown data, so as our model had no biasing while predicting. So, the model is trained with train data and the prediction is done with test data which is also used for evaluating the model.

Dimension of Train set: X(155738, 35) , y(155738,) Dimension of Test set: X (38935, 35) , y(38935,)

The data is fitted and the model is trained.

**Model evaluation.**

F1 score was used for the evaluation of model and the f1-score we got is → 0.831

So our model is selected. As our model was trained only with training data so test data is as it is, now we will train our model with whole data.

After training again the data is predicted from the model only and the f1 score we got is → 0.8636 Hence our model is ready to predict any data.

# Results

From above methodology and data analysis, I could state that the data performed very well with both known and the unknown data.

* The model shows a 0.831f1 score on unknown data;
* The model shows a 0.864 f1 score on known data;

So, this model can be used to predict the accident severity just by giving the necessary data and the model will predict the severity for them. A lot of possible casualties and loss of property and life can be minimized by taking precaution measures.

# Discussion

The data was well collected and stored. The data was more than enough to predict the severity. So anyone living in Seattle or someone who is planning to visit can formally use this model to predict the accident severity of that place. They can plan their travel accordingly and mark the safe and unsafe areas. Even more attributes could be used to predict the data but they were not precise enough to come up with statistical conclusion. And even selecting more data will make our model too biased for our data plus start predicting the noise. While analyzing the data the months and days of week had not much difference to the collisions, they were just consistent and that's why they were dropped as we need a pattern to predict the right model. Adding average data will just put load on our machine.

In a future study, categories should be further tested and refined. They were constructed based on previously published literature. However, a preliminary analysis should have been performed to choose the categories that can be determinants of life expectancy. In addition to the categories used here, there can be categories of venues a positive effect on mental health, thus prolonging life, such as arts & entertainment places and spiritual center. Types of restaurants can also be considered. There can also be categories of venues that could adversely affect life expectancy that would need to be taken into account such as cigarette shops and liquor stores. The categories included in this study are still quite comprehensive as I used as many as possible from all the Foursquare categories.

Using Foursquare also presented a few challenges. There can also be redundancy in counting the venues of each category in this study. Although users do report duplicate listings, some can still exist, especially since different branches, departments, and sometimes even each hole of a golf course can have their own listings on Foursquare and counted as separate venues. Along with the fact that there are many more venues in areas with denser population, it is possible that the number of resources in the communities with many resources were over-represented. Some venues may not appear on Foursquare because of a number of reasons such as being newly opened or having some changes. There can also be a few that were missed especially in sports facilities category because there can be facilities that can only be categorized as ‘athletics & sports’ and not as any specific sports venue. However, I felt that these discrepancies are not enough to throw the whole picture.

# Conclusion

Purpose of this model was to create a model that can predict the accident severity just by knowing the basic component of that place. My analysis clearly shows that this data was excellent data which helped our data to achieve a final f1 score of 86.36%. So anyone living in Seattle or someone who is planning to visit can formally use this model to predict the accident severity of that place. They can plan their journey accordingly and mark the safe and unsafe areas further planning depends on them. This model not only will help the general public, but other concerned authorities could also see this and find the areas that are red zones at particular and save precious lives. they can even put the barricades to alert the upcoming traffic. Other concerned authorities who keep the condition of road and light could mark the places which have problems and solve them.

Final decision on journey planning will be made by public based on specific characteristics of that location and past history of collision. Stakeholders should improve the infrastructure of those communities to bring in those resources. In addition, communities with low number of resources but high collisions should be studied to see what sets them apart.