Coursera Capstone Project.

IBM Applied Data Science Capstone Project

Car Accident Severity – Seattle

By: Nysa P Ginu

October 2020

1. Introduction:

1.1. Background:

There was a time when only the rich and powerful people owned a car. But in today's world, almost everyone is able to afford a car. With the increase in the number of cars on the road, The number of accidents happening is also in an upward trend. Moreover, It is at an all time high. Around the world, Most of us drive cars closely each single day. We avail car to commute to work, visit friends and family members and pick up provisions. It can be simple to consider for granted how risky driving car can be still at low speeds. Car accidents put our well being and health at risk. Every car accident is different. Nobody knows how an accident will happen or the outcome of an accident. All accidents are defined by the severity. All accidents impact a person. Nobody understands the impact of an accident on a person until it happens to them. So, being able to understand and prevent the causes of accidents becomes of utmost importance.

1.2. Business Problem:

The objective of this capstone project is to analyze the factors leading to a collision from a dataset of Car accident Severity in the city of Seattle, USA and using Machine Learning Methodology, create a system that can successfully predict the Severity of the car accident. This will help in answering the business problem: What are the conditions that would lead to a car accident and how severe it would be?

1.3. Target Audience:

This project would be very useful for Civil Authorities who are looking for a way to curb the increasing number of accidents in Seattle. This project is timely as the city is currently suffering from an increase in the number of accidents.

2. Data

2.1. Data Sources:

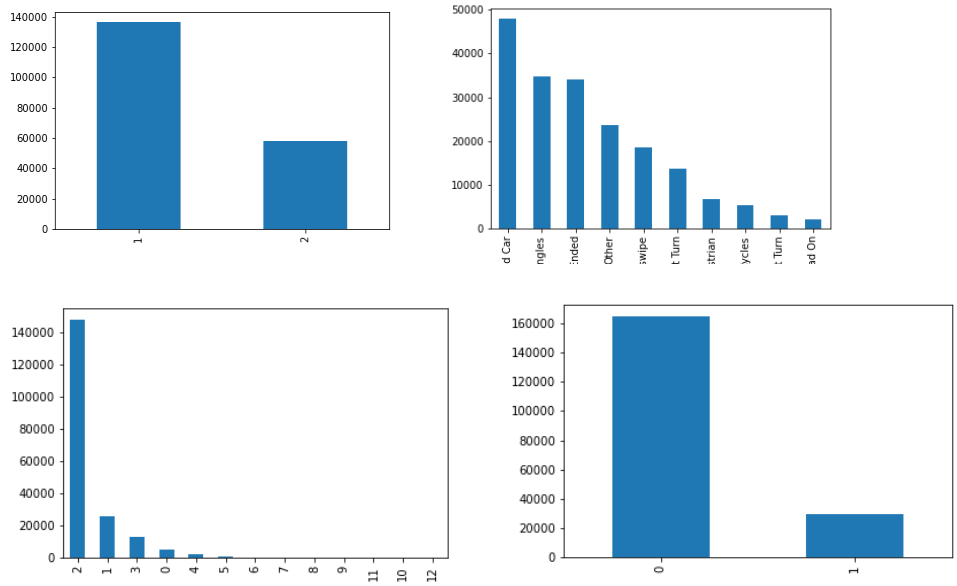
The dataset taken is the one given in the earlier modules of the course. This dataset can be found here.

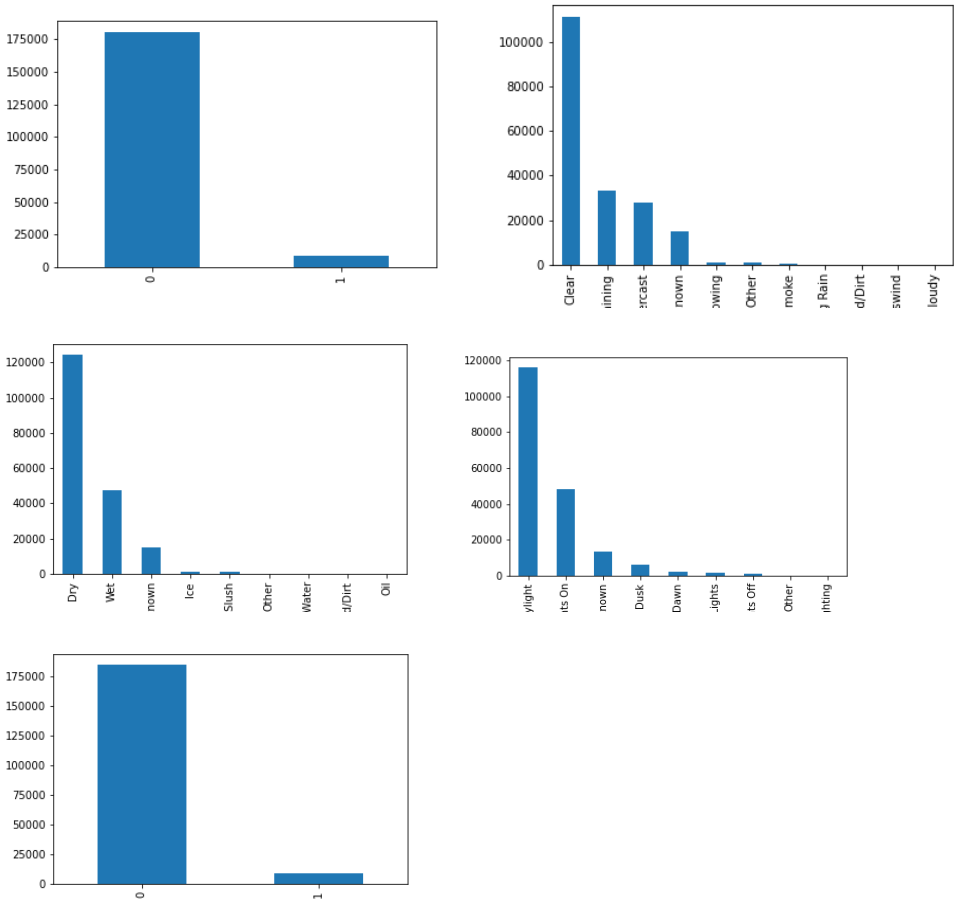
2.2 Data Cleaning:

The data downloaded can be loaded into a dataframe. On observing the dataframe, I felt that a lot of the attributes(columns) are redundant i.e. repeating the same information or are irrelevant to the problem in hand. Therefore, I selected only a few attributes for working on the project i.e. SEVERITYCODE, COLLISIONTYPE, VEHCOUNT, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, SPEEDING. I compiled them into another dataframe named “new”.
Again a few of the columns had variable values like INATTENTIONIND, UNDERINFL, SPEEDING. So we will convert them into a more correct form.
Most of the columns of the dataframe are of the type object, when we need it to be numerical. Therefore, The dataframe in this form is not suitable for analysis. To make it suitable, we will label encoding on the data.

2.3. Feature Selection:

During this step, all the redundant data were removed and made into a new dataframe which had the following attributes: SEVERITYCODE, COLLISIONTYPE, VEHCOUNT, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, SPEEDING.





Upon visualizing these attributes, we could see that, INATTENTIONIND, UNDERINFL, SPEEDING didn't really play any important role in the accidents. Therefore, these attributes were also dropped.

3. Methodology

After performing all the above steps, our data is ready to be fed into machine learning models. The models that we used are:

K-Nearest Neighbor(KNN):

KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

Logistic Regression:

Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression

Decision Tree:

A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

4. Conclusion

In this study, I analyzed the factors that can cause an accident. I identified that most crashes happened in clear, dry and bright conditions while the attentiveness, under the influence or speeding did not play a vital role. Most days are clear, dry and bright, so it's no surprise that car crashes occur under these conditions. I also created a model to predict the severity of the car accident using machine learning algorithms.