**Predicting Car Accident Severity**

IBM Data Science

Capstone Coursera Project

Author: Noor Kouli

Data of Submission: 18/09/2020

# Introduction

Car accidents are a common problem that occur daily. They can result from speeding, tailgating, failing to use an indicator, bad weather conditions or any form of distraction. However, some accidents are more severe than others depending on the factors mentioned earlier that can lead to permanent injuries or even death. In addition, car accidents can cause heavy traffic and may block roads causing people to miss important meetings, flights and so on. Given the weather and road conditions, what if we can predict the possibility of getting into a car accident and how severe it might be? This will signal us to drive more carefully, change our route or postpone a meeting ahead of time. Not only will this solve the problem of wasted time but will also enhance awareness of our surroundings to safely arrive at our destination.

In the past, researchers and data scientists have built machine and deep learning models on different car accident scenarios since it is a very common problem they are trying to solve. For example, the automatic detection of a car accident from the driver’s phone in order to receive medical care in a timely response. This detection can help save time and allow the ambulance to be on its way as soon as the accident occurs. Another example is detecting human driver inattentiveness through capturing and analyzing the face of a driver. In addition, it can detect fatigue, drowsiness and aggressive behavior. Once detected, this deep learning algorithm with signal them to focus on the road or stop on the side.

Although there have been quite a few studies on car accidents, there is still ongoing research on how extreme weather conditions and traffic affect car accident severity. Therefore, this project encompasses a machine learning model that will help minimize car accidents by targeting car drivers in Seattle, Washington.

# Data

The data frame for this model consists of 146,059 rows and 38 columns that address collision history in Seattle, Washington from 2004 to the present day. The columns are collision type, location of accident, junction, number of people involved in the accident and so on. However, not all the 38 columns in the dataset can be used. Therefore, this dataset must be cleaned in order to begin building the model. The primary column in this dataset that we are focused on is the severity code. This severity code can be values of 0,1,2,2b and 3. 0 meaning the severity is unknown, 1 is property damage, 2 is injury, 2b is serious injury and 3 is a fatality. In addition to the severity code column, other features can be extracted from the data such as weather forecast, vehicle count, light condition, road condition and so on. However, only 3 variables will be involved in classifying a test set of car accidents into its respective severity code. Those 3 variables include the weather, vehicle count and collision type. This dataset will be divided 80% into a training set and 20% into a test set in order to achieve out of sample accuracy. The test set will help us classify accidents into a severity code number and help the driver change his route or slow down his pace depending on the value of the severity code.

# Exploratory Data Analysis

## Data Preprocessing

The first thing done was to extract the columns needed and drop all the unnecessary columns. This helped narrow down the very large data frame into 194,673 rows and 4 columns, making it easier to work with. Next, we dropped any missing values from the data frame since there’s already enough data to work with. The data shows that the target value, SEVERITYCODE, had only values two ‘1’ and ‘2’. A “value\_counts” code was used to show that there were 132480 values for ‘1’ and 57091 values for ‘2’. This meant that the data needed to be balanced in order to proceed with the data processing. Random rows with SEVERITYCODE value ‘1’ were dropped in order to match the same number of values for the value ‘2’. Lastly, the value\_counts command was used to retrieve the unique values for the variable’s collision type, weather and vehicle count in order to get an idea about the remaining data in our dataset.

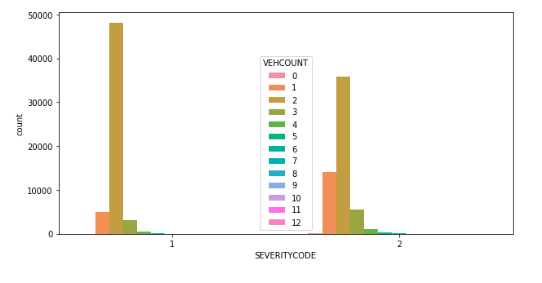
## Data Visualization

Since each variable has several unique values, the most suitable plot would be a count plot bar graph of the count of each unique value of the variable versus the SEVERITYCODE. For example, the first plot shows the count of each weather forecast for each SEVERITYCODE. The pink bar shows that the number of accidents that happen when it’s raining are mostly considered severe car accidents which cause injury (SEVERITYCODE =2). This graph also shows that most car accidents in Seattle, Washington occur when the weather is clear. There is barely any data on snow, cloudy, fog and so on. However, we can conclude that this figure does not show much correlation between the severity code and weather because most the data has collisions recorded in clear weather. Therefore, no comparison can be made with harsher weather forecasts.

Chart, bar chart

Description automatically generated

The second figure shown below displays how the vehicle count affects the SEVERITYCODE. The vehicle count involved in car accidents in this data frame ranges from 1 to 12 cars. The majority number of vehicles that get into accidents are between 2 cars. Car accidents between 2 cars can cause injury or merely property damage. However, most collisions in this data frame between 2 cars are not as severe with a SEVERITYCODE of 1. Moreover, accidents involving 1 or 3 vehicles in Seattle were more severe and likely caused injury. There is negligible data on collisions between 7 to 12 vehicles.



The last figure compares the different collision counts with each SEVERITYCODE in Seattle, Washington. This figure shows the most variability, with a good proportion of different collisions in the data frame. For instance, most parked car accidents are under the severity code of ‘1’. While, more rear ended, angled accidents and pedestrian accidents were recorded with a severity code of ‘2’ leading to injury of the passengers. In addition, there aren’t as many head on accidents recorded. However, the data is enough to make a reliable comparison and to predict future severity codes based on the collision type.

Chart, bar chart

Description automatically generated

# Machine Learning Model

For the machine learning model, **weather** and **collision type** will be the prime variables since vehicle count did not provide sufficient data for vehicle number other than 1, 2 and 3.

Since our variables had categorical variables, we needed to encode our data into 0’s and 1’s in order to build a machine learning model. This is a key step in order to build the model. Following this, the data is split into a train and test set with 80% train and 20% test. Three machine models were tested with the preprocessed data: KNN, Decision Tree and Logistic Regression. Each model was evaluated with Jaccard score and f1 score. The results show that all three models oddly gave very accurate predictions shown below.

Table

Description automatically generated

# Conclusion & Future Recommendations

Before performing the algorithms on the data, I predicted K Nearest Neighbor (KNN) to be the most accurate model since it’s one of the most often and easiest to use classification techniques. The results turned out very unexpected because all three models gave almost the same accuracy and predictions. I would conclude that the variables chosen may not have been the best to correlate with the severity code. However, it’s still no surprise that the results are very accurate. The reason is that encoding this data using get\_dummies in pandas simplified the data into a binary system with almost very similar data in the train and test sets. Moreover, the majority of data in the weather for instance was clear and therefore, most values in the test set were also probably clear. This simplified the algorithm for the three different models.

Future recommendations would include to get more data on accidents in different weather conditions other than clear skies. This also applies to other variables in the data. The more data retrieved on the unique values in a variable, the more likely the algorithm produces an output with a more realistic f1 and Jaccard score.