Coursera Capstone

IBM Applied Data Science Capstone

Predicting Car Crash Severity

By: Kashika Dhanjal

August 2020

Introduction and Business Problem:

Driving is one of the most common forms of transportation for people across the globe, whether they are using their own car or a service such as Uber or Lyft. However, the road can be very dangerous, owing to factors as wide ranging as how many cars are on the road to the weather. If accident conditions could be identified ahead of time, then drivers may drive more safely on the road or all together avoid going a certain route or just replan a drive that they were going to make. The objective of this capstone project is to analyze and predict the conditions that lead to a severe crash with cars. Using data science methodology and machine learning techniques such as decision trees, this project aims to answer the following business question; under what conditions (both environmental and human) do severe road accidents occur?

Data:

The dataset I will be using is publicly released crash data by the city of Seattle on car accidents that have occurred in or around Seattle. The attributes that are described in this data and are relevant to this project include

- Severity Code - a measure of have severe the crash was
- Person count - how many people were involved in the crash
- Vehicle count - how many cars were involved in the crash
- Junction type - whether or not the accident occurred at a junction and what type it is
- Weather - what the weather was at the time of the crash
- Road condition - whether the road was wet, or something was on it, etc.
- Light condition - whether it was dark, or if there were streetlights present, etc.

Methodology:

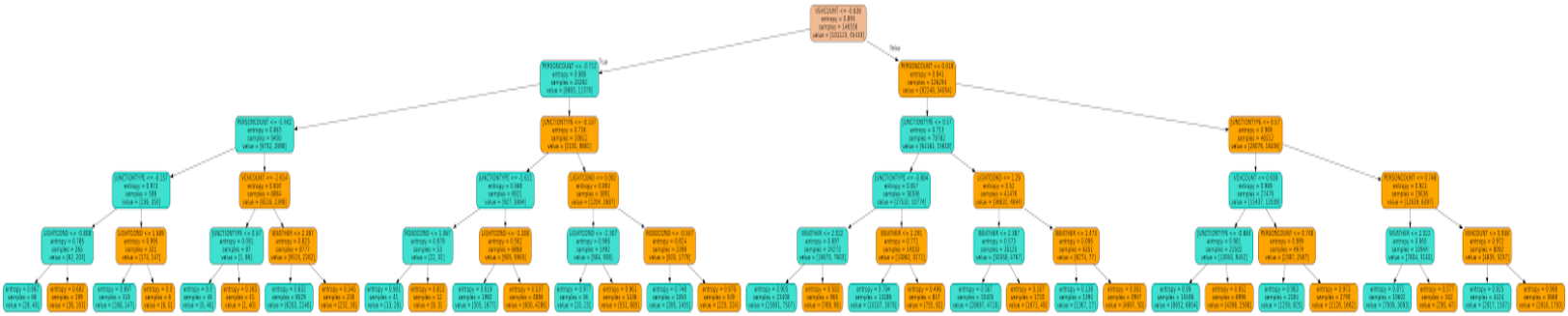
First, we have to read the csv file into a dataframe object in order to work with it. To do this we use the pandas .read_csv(filename) method and display the head of the dataframe after to make sure it was read in right. We then look through the columns of the current dataframe we have and choose which ones we want to keep. We do this by loading the names of the columns we want to keep into the array that will go be placed in df and assigned to variable df_needed. We then display the head of df_needed to make sure that this worked.

Now we will clean the data and make sure it is in the form that is needed to create a ML model. This includes getting rid of rows with NaN values and cat coding our categorical variables. First we drop rows that contain NaN values since there is no good and consistent way to replace them. Then we assign the categorical columns to type category and cat code them, a process that assigns the unique category values to numbers. This makes it so that the data is ready to be used by any of the ML models we plan to employ since these models do not understand categorical values.

Lastly, we split the data into x and y values so we can generate training and testing sets for both. Then we preprocess the x values to make sure they are in the proper format for the ML model. Then we take the x and y values and split them into a training and testing set that will be used to train and then test the model. 80% of the data was used for training, while 20% was used for testing. In numerical values, there were 146556 training samples and 36640 testing samples. For this scenario, I used a decision tree classifier to generate the ML model with a max depth of 5 and the criterion being entropy. I then used pyplotplus and collections to visualize the decision tree I had made.

Results:

The results from the decision tree showed that it was able to predict the y values from the x test data with 74% accuracy. This is pretty good since car crashes can be very unpredictable and can heavily depend on driver negligence or inebriation which were not factors. Here is the decision tree when visualized (details are hard to see because of the way the image was rendered.



Discussion:

As observed from the visualization of the decision tree, the top most part of the decision tree (the root) was how many vehicles were involved in the car crash, and from then on it went to person count. This shows that these are very important variables when it comes to the severity level of crash. While this model did seem good at predicting severity level of the car crashes, the entropy levels at some of the ending leaves was still very high which suggests more layers need to be added to try and bring it down. However, adding more layers might lead to an overfitting of the model to the specific data. Thus, we will leave the max depth at 5 levels since this seems to be the most ideal.

Conclusion:

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data,

performing machine learning by generating a decision tree with the chosen parameters and lastly providing recommendations to the relevant stakeholders i.e. car drivers and local officials about the chances of a car crash under certain conditions. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The factors that matter the most in the severity of a car crash is how many vehicles were involved and how many cars were involved. However, these findings are not concrete as the entropy levels are still high and further studies involving other factors such as the driver should be looked into.