Coursera Capstone

Kashika Dhanjal

Introduction and Business Problem

- Driving is one of the most common forms of transportation
  - Lots of crashes
- Many factors in accidents
- Business Problem: 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. 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 - accident occurred on a junction and type
- Weather - what the weather was at the time of the crash
- Road condition - whether the road was wet, or something was on it,
- Light condition - whether it was dark, or if there were streetlights

Methodology

- Read csv file
- Clean data
  - Remove nan values
  - Cat coding categorical values
- Generate ML model
  - Decision tree with 6 levels
  - Criteria is entropy

Results

- R^2 score of 0.74
  - Good
    - Car crashes unpredictable
    - Didn’t take driver factors into account

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 and performing machine learning by generating a decision tree with the chosen parameters. 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.