Capstone Project: Applied Data Science in Car Collisions

Introduction

Driving a car and maneuvering through traffic is a basic, necessary task that millions of people around the world perform on a daily basis. Cars have been engineered to be safer and infrastructure has been made to last longer, yet vehicle crashes still account for millions of deaths around the world. Each year, more than 1.3 million people are killed on roadways around the world. In the US alone, there are 6 million car accidents a year, with those contributing to 37,000 deaths of Americans. There are any safety precautions that people can take to be safer on the road (wearing a seatbelt, staying vigilant, being a defensive driver, following the laws and regulations, etc.), but it would be also helpful to analyze the other factors in play with these millions of accidents. With my understanding of analytical concepts and machine learning algorithms, I will create a model that explains uncontrollable factors that lead to an increased probability of getting in a car accident. Public knowledge of these factors will help thousands of drivers a day make better decisions when on the road and help decrease road traffic related injuries.

The Data

The data I am using to approach this problem is a labeled data set. I am only incorporating labeled data because it is necessary for my supervised machine learning approach. The label for the dataset is severity, which describes the severity/fatality of a car accident. Each row represents a collision during the time period of 2004 to the present. The current dataset has 37 attributes, including weather conditions, road conditions, light conditions, time of day, location, weather or not the driver was speeding, etc. I hope to incorporate enough significant factors in order to build an accurate model that clearly shows what contributes to the severity of these collisions.

Methodology

In our data, we are able to analyze the collisions in Seattle over a 15 year time period. The significant attributes that I utilized in order to model the data were the number of people involved in the collision, the number of vehicles involved in the collision, the weather conditions at the time of the collision, the road conditions at the time of the collision, the lighting conditions at the time of the collision, whether or not the driver was speeding, and the type of road junction that the accident occured in. I utilized the following machine learning

techniques in order to predict the likelihood of an accident based on these conditions: KNN, Logistic Regression, and Decision Tree.

Results

I measured the accuracy of each model against one another with the jaccard prediction accuracy score. The following results were produced.

KNN:0.719815
Decision Tree: 0.742468
Logistic Regression: 0.727854

Conclusion

The accuracy of each model represented a similar score, but the decision tree model had the best accuracy of about 75%. This represents this model having the best likelihood of explaining to someone when they are most likely to get into a vehicle crash.