Predicting Severity of Car Accidents - Applied Data Science IBM Capstone Project

Aviad Klinger

September 2020

1. Introduction

Car accidents are a huge problem in the modern world. The Global Status report on road safety collected data from 180 countries and found that 1.25 million people die internationally every year from car accidents. Many governments collect accident records and make these publicly available to encourage researchers to try and find models that can explain and predict what conditions cause car accidents and hopefully help prevent them. In this project I pose the car accident injury prediction as a supervised learning classification problem and offer our model to all governments.

2. Data

In this project I use the collisions records dataset provided by the Seattle Police Department at Seattle.gov. The dataset contains 195,000 records of all types of collisions from the year 2004 to present, updated weekly. It includes 37 features that describe the type of collision or accident that occurred, such as: accident location, number of people involved, number of vehicles involved, time of day, weather conditions etc. The target variable is 'SEVERITYCODE' which is a binary variable that determines wither the collision involved injury/death or it was a property damage only collision.

3. Methodology

In this section I will discuss and describe the exploratory data analysis that was conducted, and the modeling process that led to choosing the best model that fit the data.

3.1. Data cleaning and wrangling

Out of the 37 initial features that were available in the dataset, 25 features were removed from consideration as having predictive potential because of different reasons. For example, variables 'X', 'Y' and 'LOCATION' were removed because the exact location of the collisions is will prevent the model from generalizing to other locations. Variables 'OBJECTID', 'INCKEY' and 'INTKEY' were removed because they are just serial numbers with no correlation to the collisions. Other variables, such as 'PEDCYLCOUNT', 'PEDROWNOTGRNT' and 'EXCEPTRSNDESC' were removed due to too many missing values or an uneven category balance that made them extremely uninformative.

The variables 'UNDERINFL' and 'SPEEDING' were corrected such that their missing values were interpreted as 'N' because all the other values were 'Y', relating to the yes or no answer to the question 'was the driver drinking' and 'was the driver speeding', respectively.

A new feature was created from the variable 'INCDATE', by extracting the day of the week the accident occurred. After that the variable 'INCDATE' was dropped from the dataset.

After discarding several redundant features, such as 'EXCEPTRSNDESC' and 'SEVERITYDESC', I inspected the correlation of the independent categorical variables using Cramer's V association (Figure 1), and found that 'ST_COLCODE' and 'COLLISIONTYPE' were perfectly correlated and so the former was dropped from the dataset. Also the variables 'UNDERINFL', 'SPEEDING' and 'weekday' were dropped due to extremely low association to the target variable 'SEVERITYCODE'.

At the end of the analysis 8 features were selected.

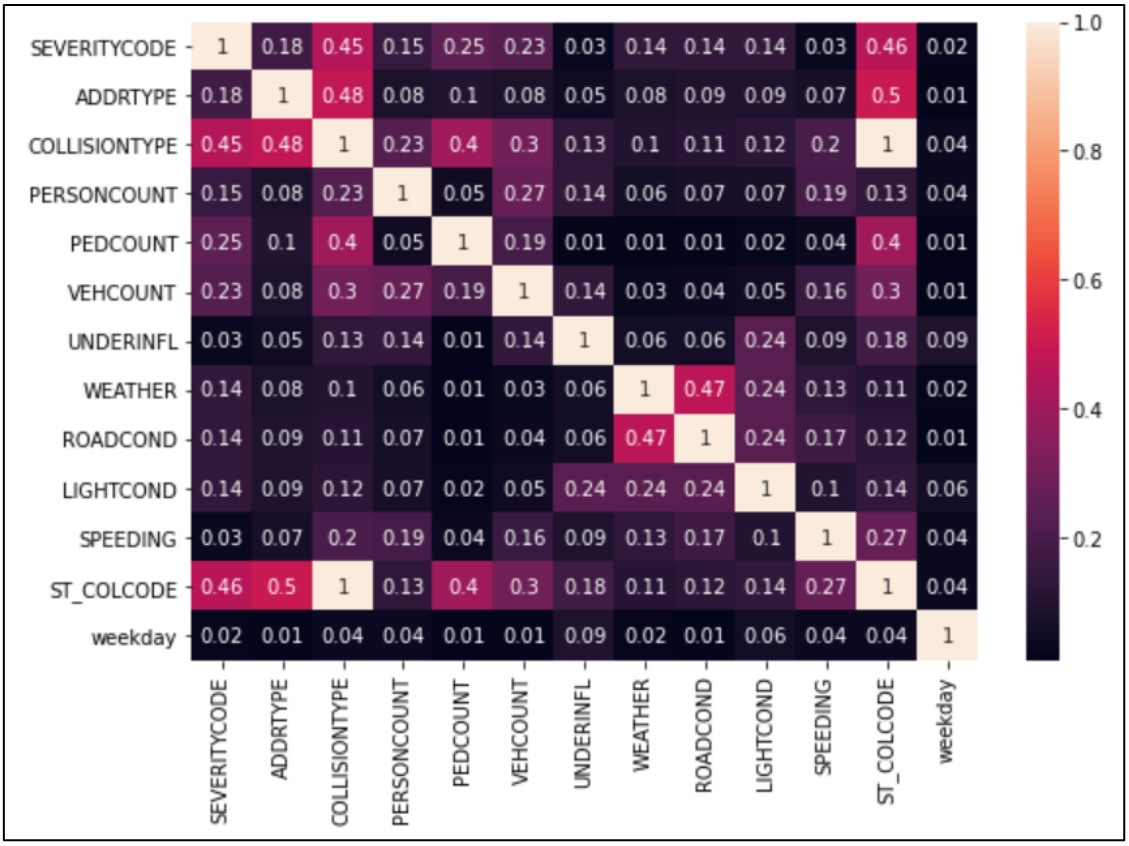

Figure 1. Cramer's V association matrix for categorical variables

3.2. Relationship between severity and the independent variables

Severity vs. Address type shows that the chance of injury is much higher if the collision is in an intersection, and on the other hand the chance of property damage is much higher if the collision is in a block (Figure 2).

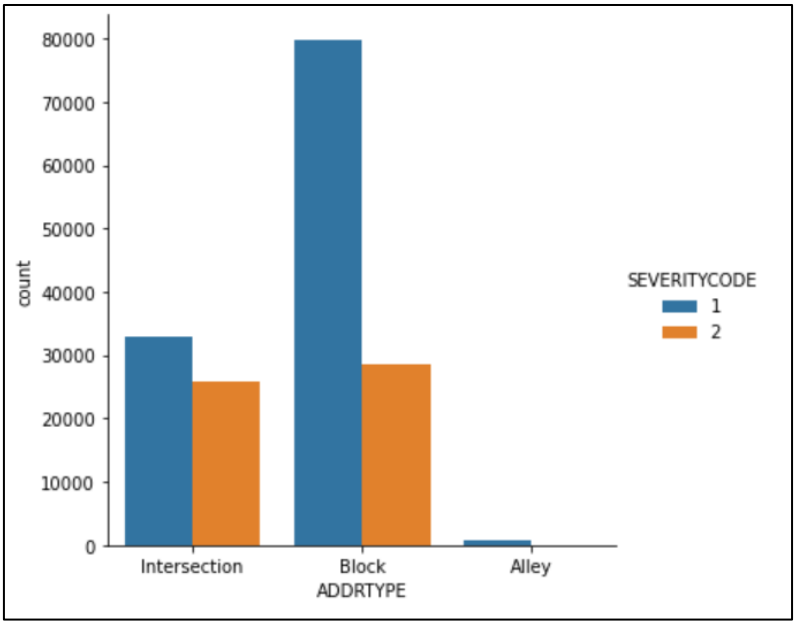

Figure 2. Bar plot of Severity and Address type

Severity vs. Collision type shows that the chance of injury is much higher if collision is from an angle or rear-ended, and on the other hand the chance of property damage is much higher if the collision is on a parked car (Figure 3).

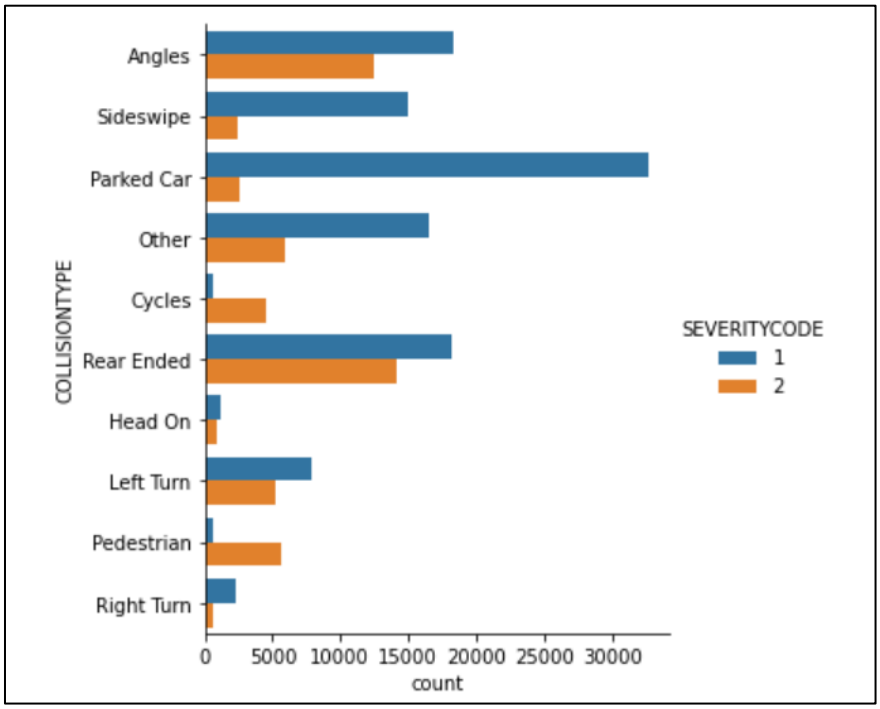

Figure 3. Bar plot of Severity and Collision type

Severity vs. Number of pedestrians involved shows that pedestrians are seldom involved in collisions, but if they are then the chance of injuries is high (Figure 4).

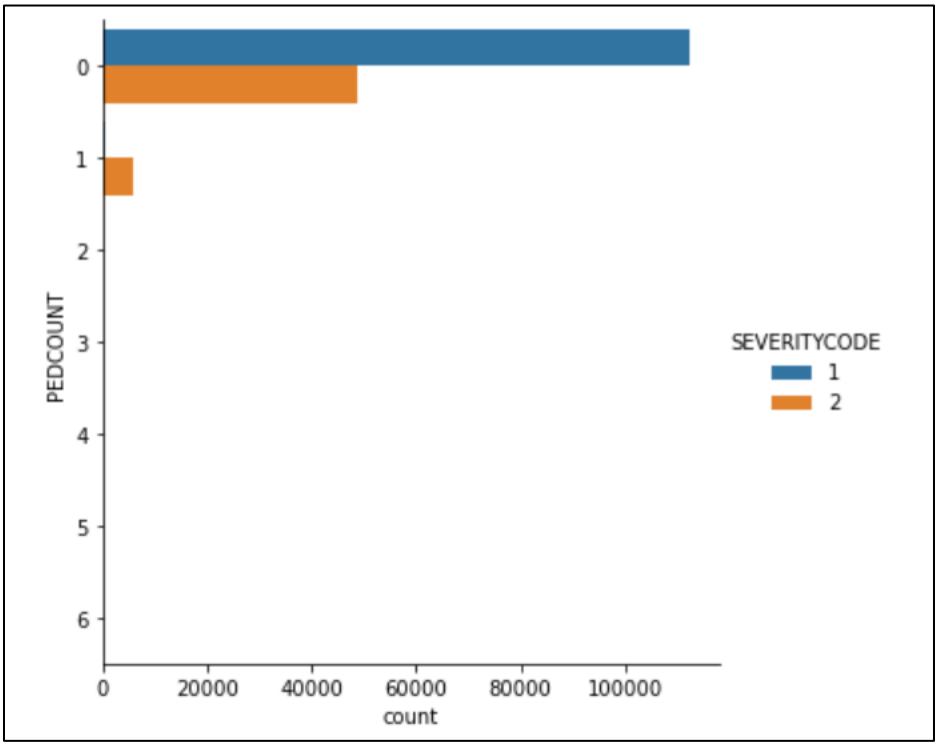

Figure 4. Bar plot of Severity and Number of pedestrians involved

Severity vs. Vehicle count shows that most collisions involve two vehicles, but the chance of injury is lower in these cases relative to higher or lower vehicle count. The dataset contained values for up to 12 vehicles, but collisions involving 5 vehicles and over were so sparse that I decided to group them to a category 4+ (Figure 5).

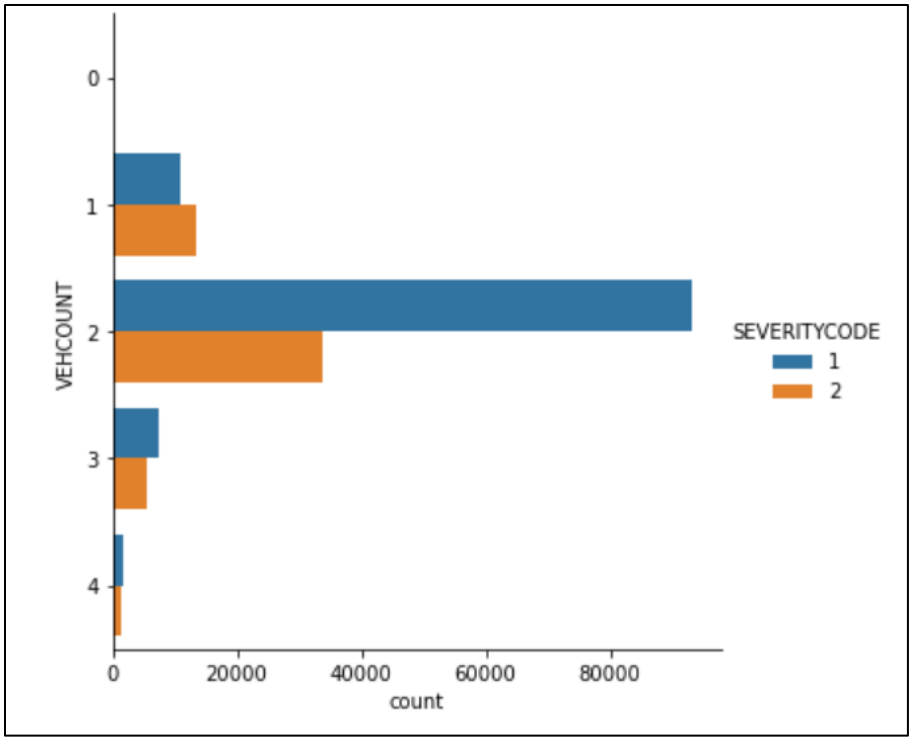

Figure 5. Bar plot of Severity and Vehicle count

Severity vs. Weather conditions shows that counter-intuitively most collisions occurred in clear weather, with a varying ratio of injuries to property damage. This variable also contained additional values that were with extremely low frequency, such as ‘Severe Crosswind’ and ‘Partly Cloudy’, so I decided to group them to the Unknown category (Figure 6).

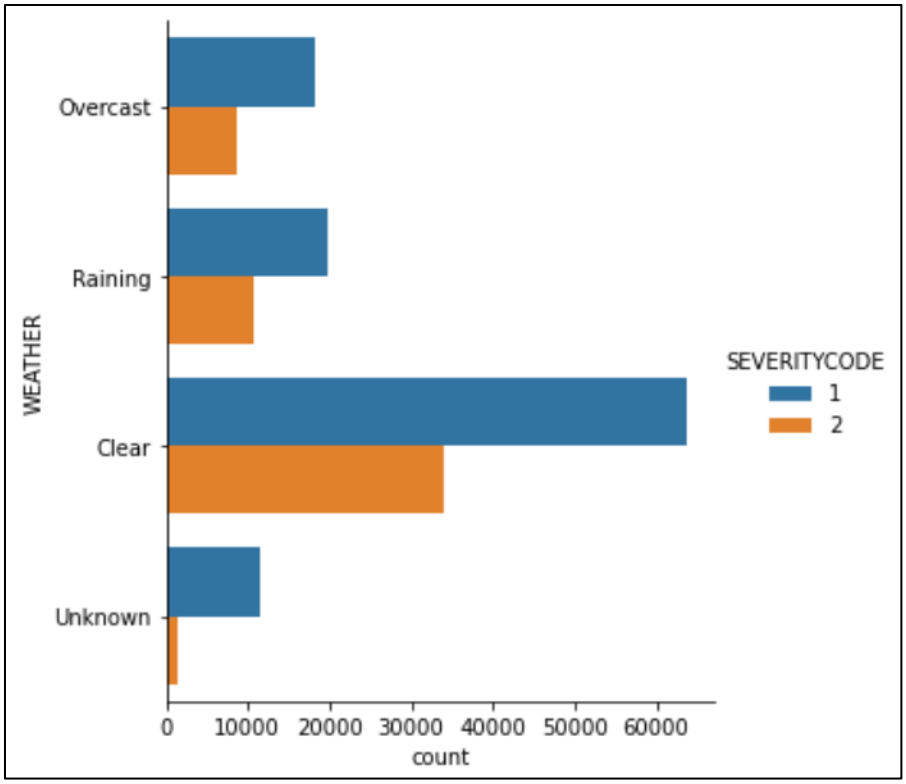

Figure 6. Bar plot of Severity and Weather

Severity vs. Road conditions also shows counter-intuitively that most collisions occurred in dry road conditions, with a varying ratio of injuries to property damage. This variable also contained additional values that were with extremely low frequency, such as ‘Oil’ and ‘Sand/Mud/Dirt’, so I decided to group them to the Unknown category (Figure 7).

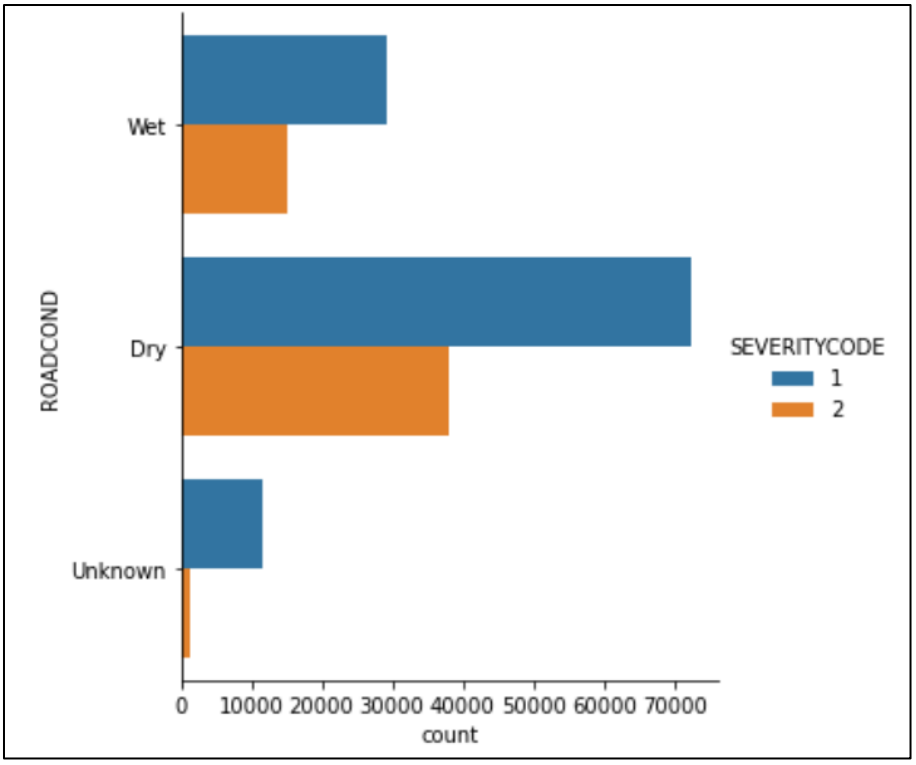

Figure 7. Bar plot of Severity and Road Conditions

Severity vs. Light conditions shows yet again counter-intuitively that most collisions occurred in day light, with a varying ratio of injuries to property damage. This variable also contained additional values that were with extremely low frequency, such as ‘Dusk ‘ and ‘Dark - Street Lights Off’, so I decided to group them to the Unknown category (Figure 8).

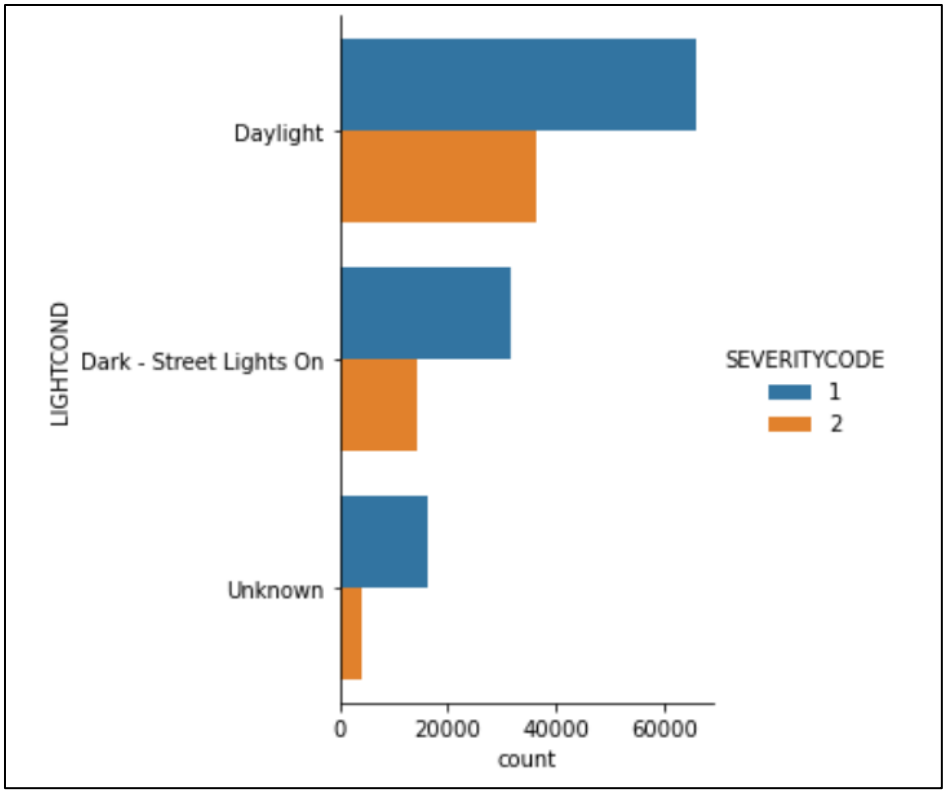

Figure 8. Bar plot of Severity and Light Conditions

3.3. Predictive modeling

This project is a classic supervised learning classification problem, and so I decided to examine four popular and well known classification models: Naïve Bayes, Logistic Regression, Random Forest and XGBoost. Even though the severity target variable is unbalanced I have decided, based on online inquiry, not to artificially balance it. The reason is that with such a large dataset a ratio of 30:70 in enough for the model to learn effectively. I chose ROC_AUC as the metric here because I think it is the best way to make sure the model balances the two categories of the target variable. I used 10-fold cross validation to choose the best model.

4. Results and Discussion

Injuries caused by car accidents present an important challenge with many real life implications. In this paper, I described my use of various machine learning approaches to

the complex problem of predicting car accident severity with the data provided through the Seattle Police Department.

The results indicate that it is difficult to create accurate standard machine learning models for predicting car accident severity. Since the dataset was skewed, optimizing for accuracy tended to produce models that classified most accidents as Property Damage Only. I instead used area under the Receiver Operating Characteristic (ROC) curve, as the performance metric.

The most successful model was the XGBoost model that achieved an AUC of 0.78 on average in the 10-fold cross validation process, surpassing with a small margin the Logistic Regression model that achieved an AUC of 0.77. The Random Forest model obtained a slightly lower AUC of 0.77 and the Naive Bayes model got the lowest score AUC of 0.75.

The most influential features in the models were quite intuitive - light conditions, road conditions and weather conditions involved, as is apparent from the feature importance plot (Figure 9). But there were a few surprises when it turned out that speeding, drunk driving and the day of the week were not good predictors for accident severity.

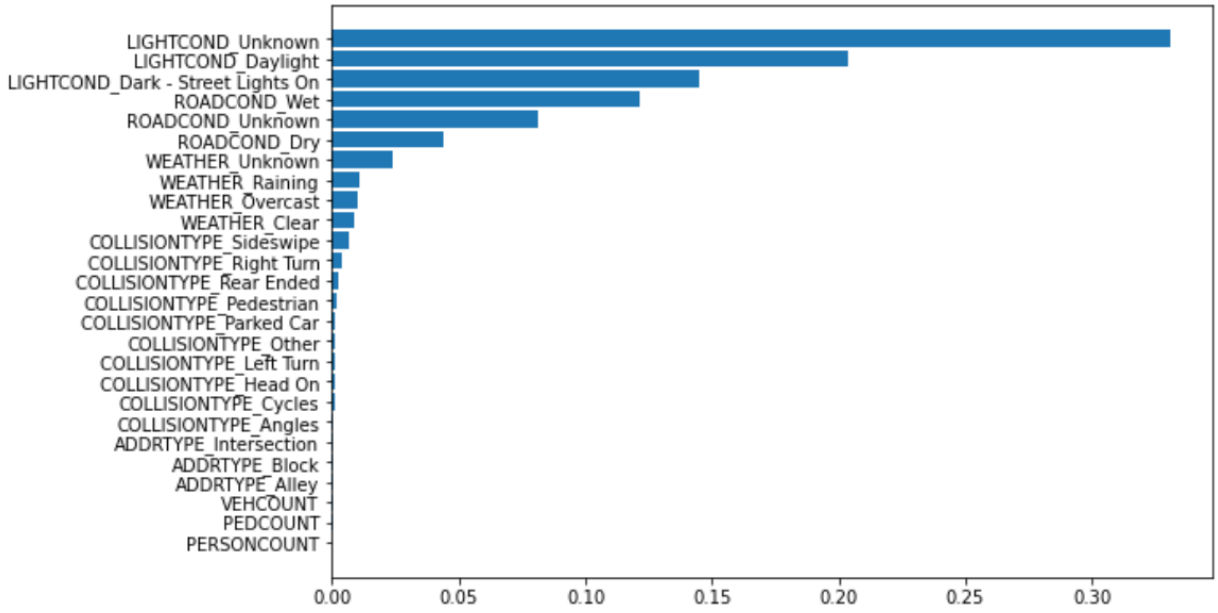

Figure 9. Feature importance plot of Independent variables

The results show that the data doesn't provide much insight as to a clear difference between accidents that end with property damage and those that end with injury or death.

The benchmark of ~70% that is the rate of Property Damage Only accidents out of all accidents is improved by the model to 75% in the test set, an improvement but not a substantial one.

5. Conclusion

The purpose of this project was to identify car accident severity in order to determine what conditions cause car accidents and hopefully help prevent them or reduce their severity. The results show that although it is difficult to distinguish between Property Damage Only accidents and accidents with injuries, it is possible to ascertain what attributes are important and can indicate the possibility of an occurrence of a more fatal accident.