Predicting Bicycle Casualties in NYC Motor Vehicle Collisions

Joseph Vorbeck, Marvin Williams, and Krisliam Nunez

Abstract

There are multiple factors that can come into play in regards to Motor Vehicle Collisions whether it be fault in infrastructure, human error, or other components at play. This study attempts to look at data from the NYC Open Data program initiative to predict whether machine learning can be utilized to try to classify biking injuries or deaths in motor vehicle collisions. A random forest model was constructed as well as a text analysis using Natural Language Processing. The results of our forest weren't as robust, with only a 20% precision rate. Initial results were satisfactory but further work needs to be done to increase the power of the classifier.

Introduction

The intricacies surrounding daily commutes are complex for those living in high-density cities. Places such as New York City, Chicago, and Los Angeles are all subject to major traffic jams and over-populated public transit systems. With many people looking for cost-effective ways to travel, commuters have decided to look for alternative means of transportation. To avoid the complexities of car ownership while living in a big city as well as the unpredictability of public transit, many in New York City have opted for a bicycle as their primary transportation vehicle. Unfortunately, motor vehicle collisions and other incidents seem to be an inevitable result of traffic in such a compact city, and that trend seems to be on the rise.

In 2019 alone, NYC had 29 bicycle deaths, up from 10 the previous year (Clayton 2019). Some place the blame on narrow streets, inaccurate timing of traffic lights, or simply a lack of knowledge of traffic rules and regulations. Regardless of who faces the brunt of this blame, the existence of this problem is enough grounds for further investigation. This study looks to perform secondary data analysis on the NYC Motor Vehicle Collision Crashes data set to investigate whether or not there are other factors at play contributing to bicycle deaths or injuries. By using Machine Learning Models, this paper seeks to predict bicycle deaths or injuries in car crashes given a variety of other components. A text analysis will also be performed using Natural Language Processing in attempt to locate any patterns in the types of streets in which these incidents occur.

Data

The data used for analysis was provided by the NYC Open Data Initiative. It contains information on all police-reported motor vehicle collisions in New York City. The variables chosen for this analysis include the month and hour of the incident, borough and zip code it took place, number of pedestrians and motorists injured, number of pedestrians, and motorists killed, the main

contributing factor for the collision, and the type of the vehicle responsible. The target variable in this case was created by finding the collisions in which an individual on a bicycle was either injured or killed.

In order to properly pre-processes the data to be suitable for machine learning, a few steps were taken. The cardinality of all columns was initially determined so that the variables are treated correctly. Low cardinality columns were one-hot encoded. The max threshold for distinct values on categorical variables to be one-hot encoded was set at 25. With too high of a threshold, the model will be more susceptible to overfitting. High cardinality columns were target-encoded, which replaces the values in the column with the probability of it showing up against the target. It's an effective way to keep high cardinality columns in the dataset while increasing the power of the predictive model. Target encoding however lessens interpretability. The max threshold for distinct values on categorical variables to be target encoded was greater than or equal to 25 but less than or equal to 300. Rows were also filtered out where the location of the collision was unknown.

Findings

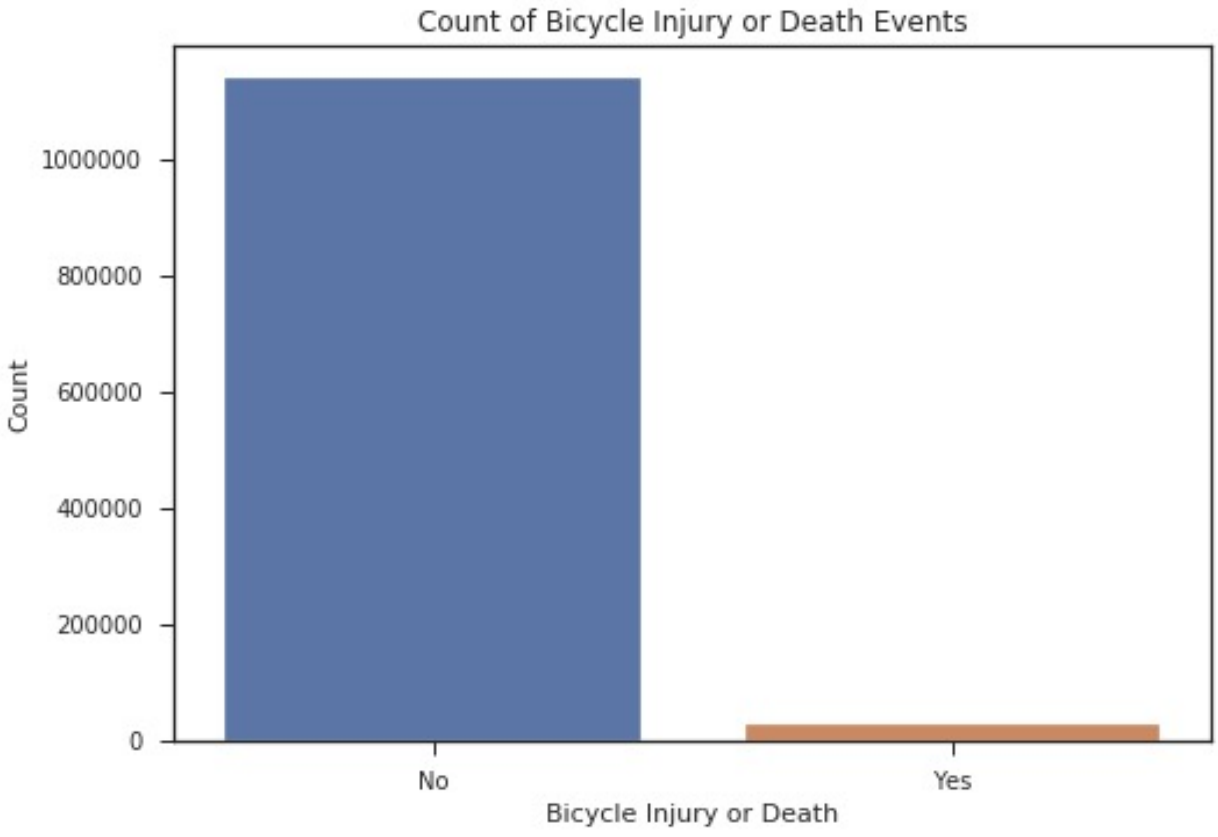

Figure 1

Figure 1 displays the breakdown of the dependent variable, which was created to analyze

whether or not 1 or more cyclists were injured or killed when involved in these collisions. These variables were collated for a more encompassing account of bicycle riders. According to the bar graph, only 27455 people in this data set were either killed or injured on a bike from a car accident. This low number was expected, as the majority of vehicles not only involved in motor vehicle collisions in New York City, but utilized generally, are automobiles.

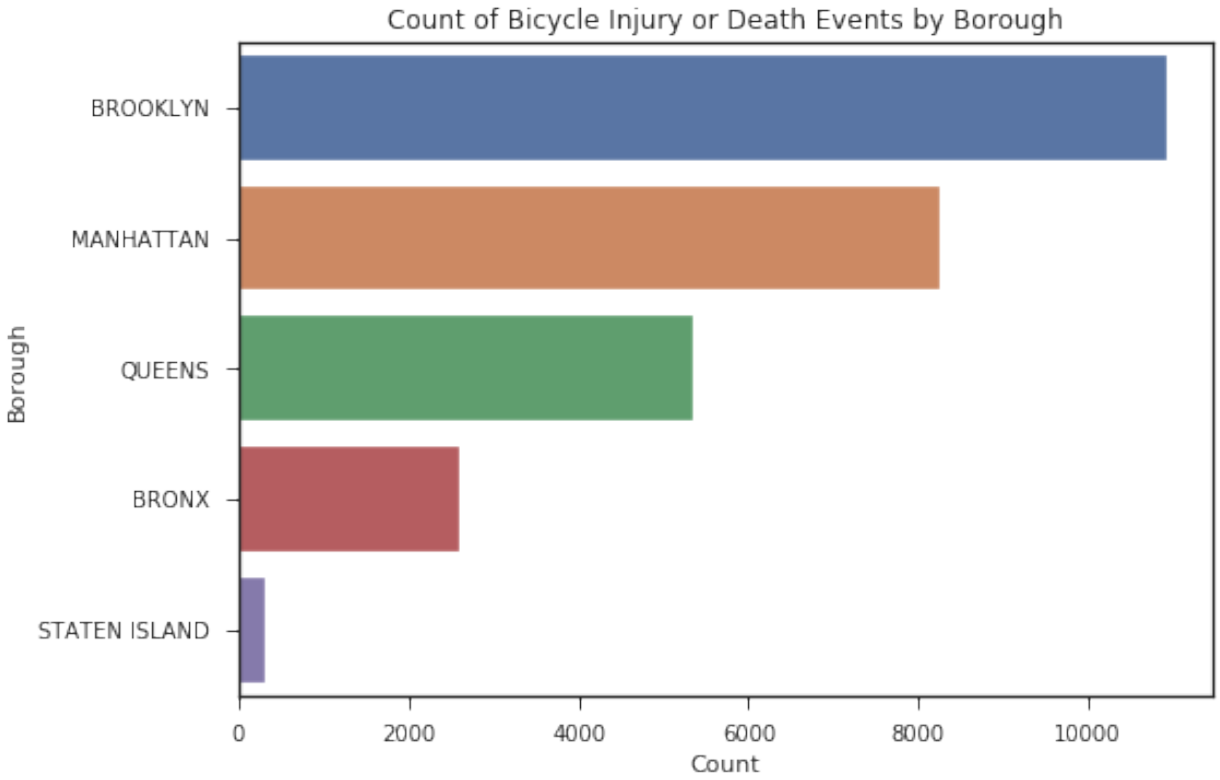

Figure 2

Figure 2 shows the breakdown of Bicycle casualties, this time according to Borough. As we can see, the majority of bicycle casualties takes place in Brooklyn, accounting for over 10,000 of these incidents. Next is Manhattan which comprises over 8,000 incidents, followed by Queens with just under 6,000 counts, Bronx with fewer than 3,000 and lastly, Staten Island with just over 200 counts. What was shown from the data coincides with what was already known, which was a noticeable amount incidents occurring in Brooklyn, which is already known to have poor infrastructure for bicycles. [1]

[1]More on Brooklyn's bicycle deaths: Fitzsimmons, Emma G. More Pedestrians and Cyclists Are Dying in N.Y.C. Drivers Are Often to Blame. www.nytimes.com/2020/03/10/nyregion/nyc-deaths-pedestrian-cycling.html

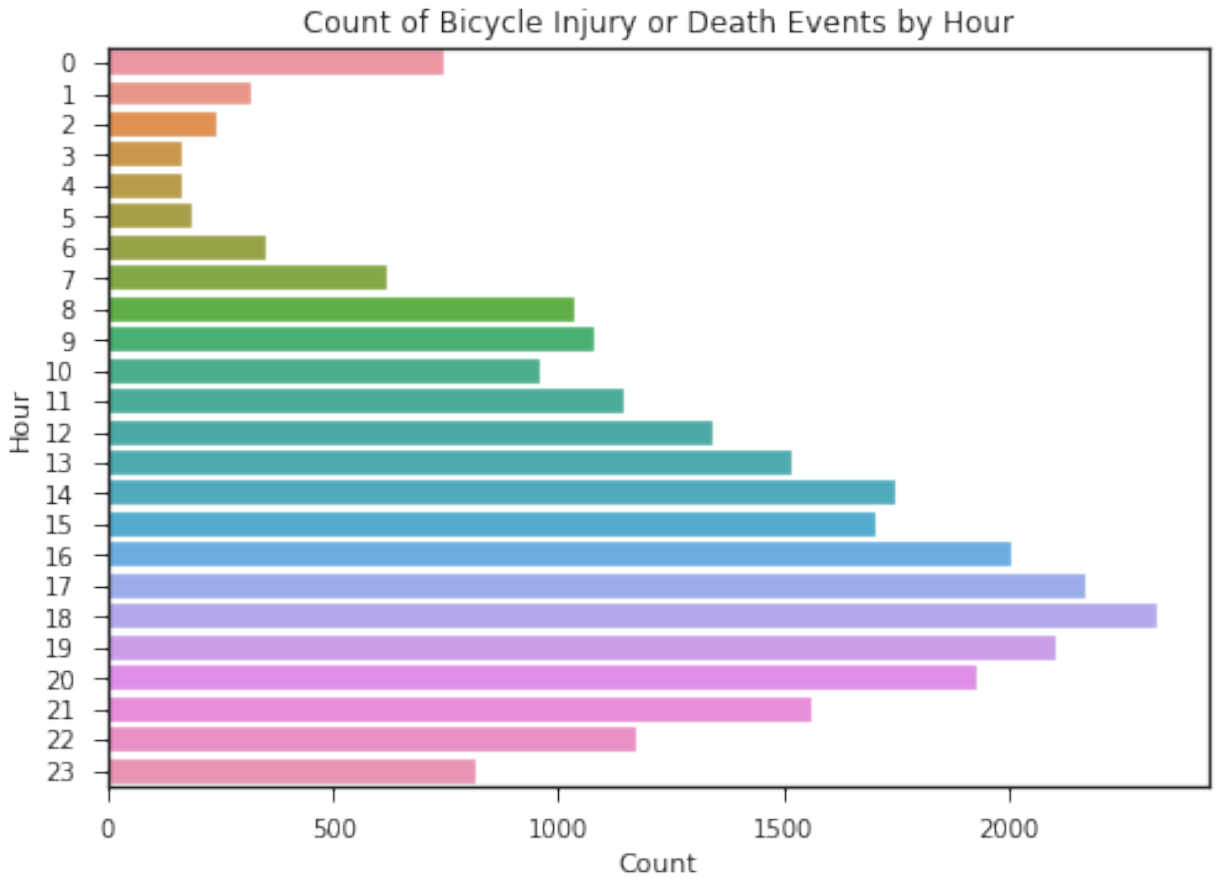

Figure 3

When analyzing casualties according to hour of day as shown in figure 3, we see an expected trend when it comes to incident counts. The highest values range from 4pm to 8pm, peaking during 5, and 6 o clock, which is the heart of NYC Rush hour. This isn't a surprise as we expect accident to be more frequent during much busier times.

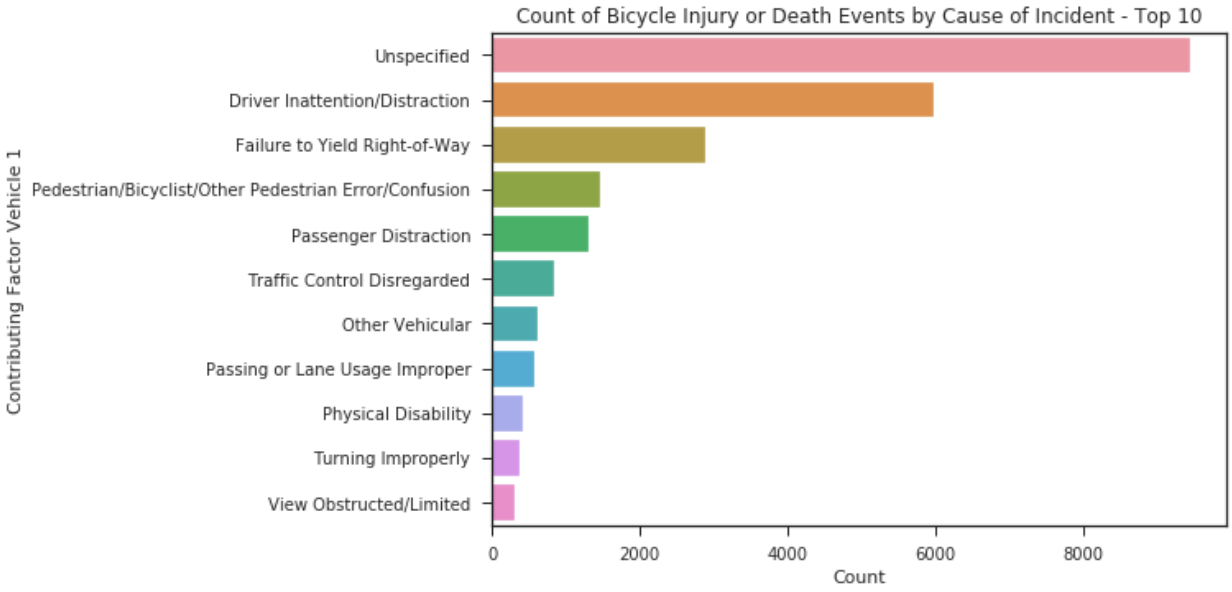

Figure 4

Figure 4 shows the Count of Bicycle injuries and deaths by cause of incident. As you can see by the figure, the majority of incidents were caused by unspecified reasons with over 9000 of the reported cases. Followed is Driver inattention or distraction with over 6000 counts, as well as Failure to yield right-of-way with about 300 counts. As the graph shows the vast majority of cases were caused by driver error, with Pedestrian/Bicycle error accounting for less than 2000 cases. Due to the high amount of those cycling in New York City, this is primarily one of the main reasons such a large investment has started being put into the bike lanes on local streets and on the side of highways.

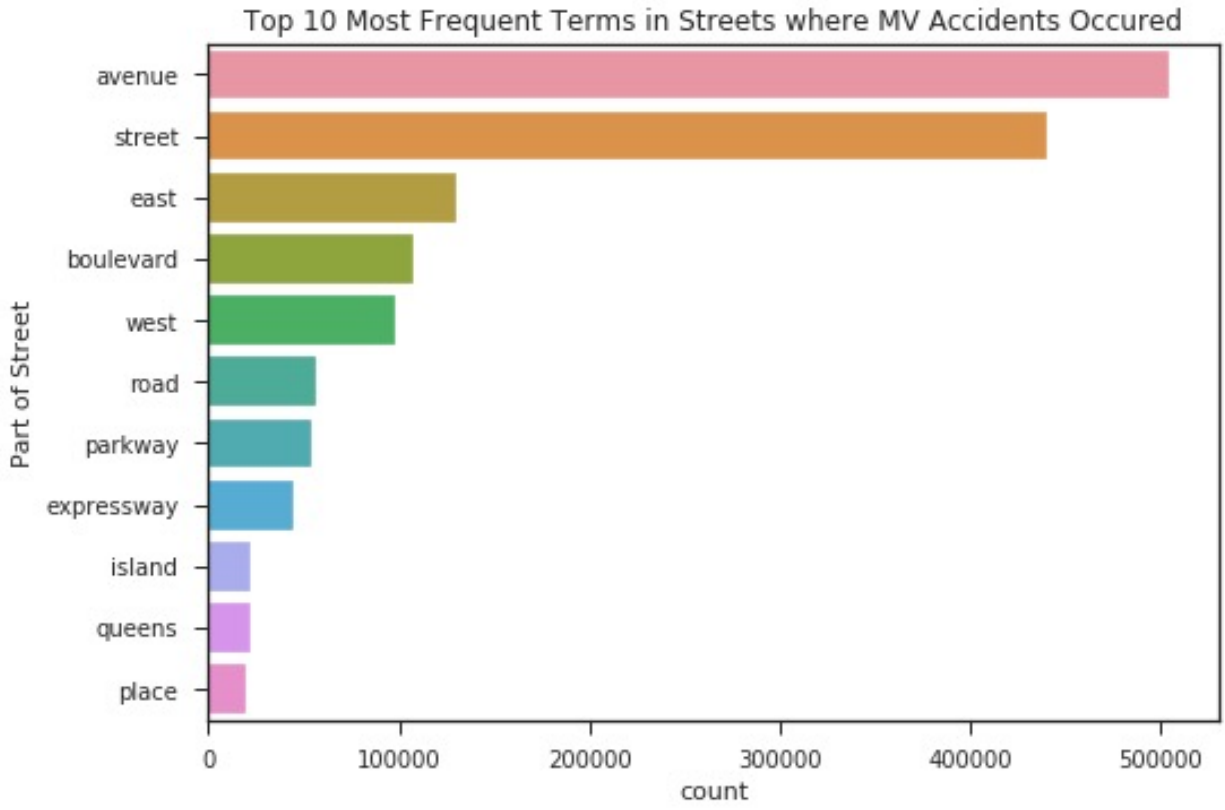

Figure 5

For our Natural language processing, a Text Analysis was performed. It decided to focus on the top ten most frequent terms in streets where these incidents occurred. In doing so, it provided us with more of a precise perspective as to where other accidents may have happened, leading us to predict where the other bicycle incidents may have occurred as well. Through our bar chart in figure 5, we discovered that “avenues” was highest reaching well over 500,000. Following is “streets” at approximately 450,000, then “east” with fewer than 150,000 counts, with both “boulevard”, and “west” being just under 125,000. Lastly, with there were less than 100,000 counts for each the following terms, respectively: “road”, “parkway”, “expressway”, “island”, “queens”, and “place”.

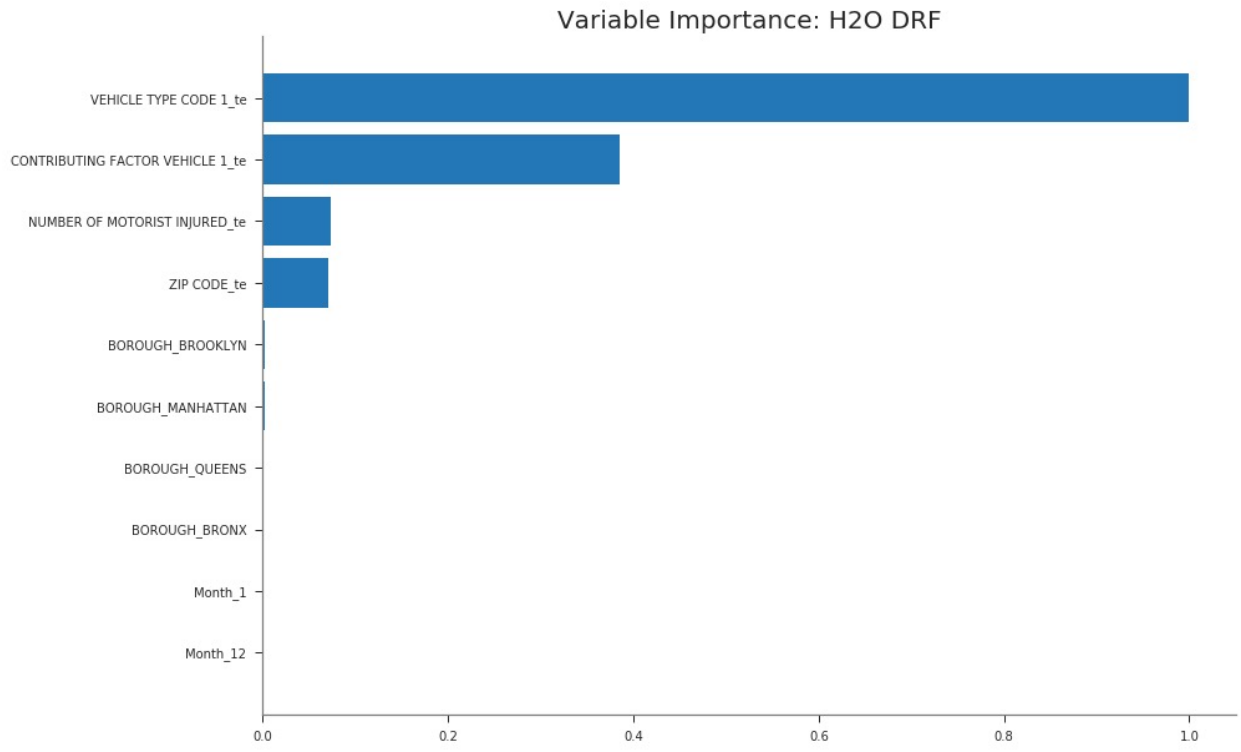

Figure 6

Model Selection

The model chosen for this analysis was a Random Forest Model. An ensemble method tends to boost the performance of a predictive model. In this study, the dependent variable is a 2% occurrence which is a very small target to hit. As random forests (and other CART algorithms) utilize sampling, the classes can be balanced so the target is distributed more evenly when the sampling is performed for each tree, which was one of the parameters of the model used here. The row sample rate was kept to 75% to assist in overfitting prevention and generalization. The column sample rate is kept at 100% to make sure all variables are considered for prediction so strong indicators are not missed. A grid search was performed on the number of trees and the depth of the trees. Generally with more trees, performance does increase but diminishing returns are reached at a certain margin, in which the increase in compute time returns minimal performance gain.

Variable Importance

Figure 6 shows the top variables for determining whether or not a motor vehicle accident resulted in a biking injury or death were the target encoded variations of type of vehicle, main contributing factor of the accident, the number of injured motorists, zip code, and the boroughs of Brooklyn and Manhattan. With the random forest model, the directionality of these variables in determining the prediction is not known but were the most determinant overall.

		Prediction outcome	
		p	**n**
actual value	**p′**	283571	948 ,
	n′	5528	1359 ,

Figure 7:
Accuracy: 0.9777767101569631
Precision: 0.19732829969507767
Recall: 0.5890767230169051

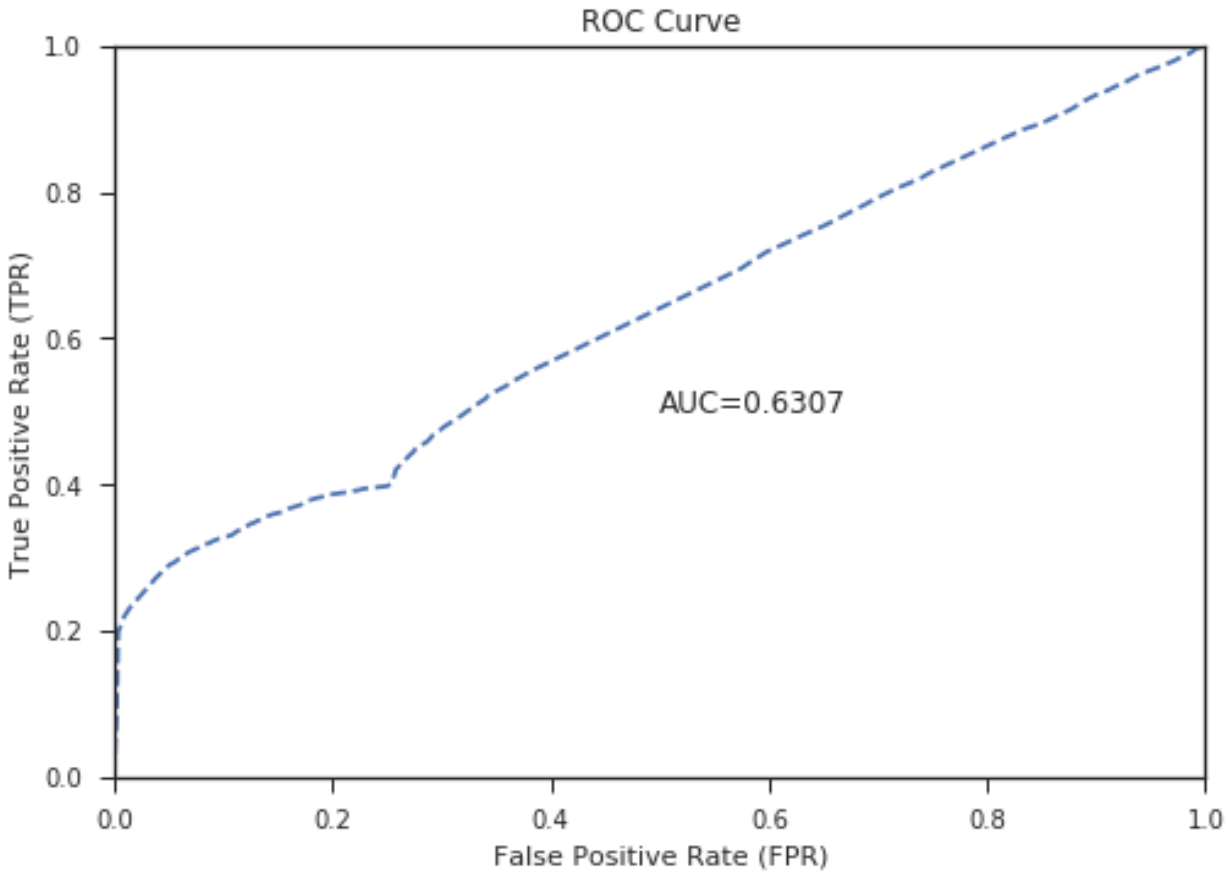

Figure 8

ROC Curve

Figure 8 shows the area under the curve (AUC) which indicates the balance between the model's false and true positive rates. At 50%, the model is no better than random chance. The AUC on the training set was 63%, indicating the model is 13% better at predicting biker injury or deaths than random chance. The model used for making predictions on the training set was chosen by it's AUC value, which was 73%.

Discussion

The random forest model used in this analysis did not end up classifying actual biker deaths or injury at a high rate (Precision = 20%), but of all the true positive cases in the test set, it accurately classified almost 60% of them (Recall = 59%). This indicates that the model classified a large portion of cases as true negatives when in reality they were actual events that occurred. To increase the lift of this model, alternative algorithms could be tested such as Gradient Boosted Trees or XGBoost, as well as additional features being created. Also, with more incidents over time, there will be more data available to train a future model on. In this instance, a handful of factors were important in resulting in biker injury or death from motor vehicle incidents in New York City, which can be used to promote any awareness campaigns and the potential to re-evaluate transportation policies in regards to city planning to hopefully give more precautions to bikers. The natural language processing performed in this analysis didn't't appear to find any unusual streets in terms of collision locations. If unusual locations were frequently mentioned, it can lend weight to policy re-evaluation and an in-depth look at why collisions occur at specific places, which hopefully could lead to increased safety measures. Although this study lacked any robust findings, it serves as a gateway for future research to promote public policy in order to increase the effectiveness of traffic regulation for cyclist.

References

[1] Fitzsimmons, Emma G. "More Pedestrians and Cyclists Are Dying in N.Y.C. Drivers Are Often to Blame." *The New York Times, The New York Times, 10 Mar. 2020,*; www.nytimes.com/2020/03/10/nyregion/nyc-deaths-pedestrian-cycling.html.

[2] Guse, Clayton "As Cyclist Deaths in NYC Hit Historic High, Safety Advocates Say Politicians Are Finally Starting to Listen." *Nydailynews.com, New York Daily News, 29 Dec. 2019,*; www.nydailynews.com/new-york/ny-bike-deaths-2019-dot-street-safety-20191229-glbw2wm5ebbf3ogstugswyqupi-story.html.

[3] Juhasz, Aubri "In NYC, Cycling Deaths Increase But Gears Turn Slowly On Safety Measures." *NPR, NPR, 14 Aug. 2019*; www.npr.org/2019/08/14/751218425/in-nyc-cycling-deaths-increase-but-gears-turn-slowly-on-safety-measures.