Synopsis

Problem Statement

Over 200,000 accidents occurred in just the five boroughs of NYC in 2018. While some car accidents are considered true accidents' over 90% are caused by human error according to Bttlaw.com. With such a high rate of accidents considered to be caused by no external factors, would we still be able to predict the severity of an accident just by the supporting data? If so, this may support the idea that the human error fault may not be as high as commonly accepted. Items such as confusing signs, poorly lit roads, or bad drainage can all cause factors that may not be seen on the surface.

By using NYC traffic data, we wanted to see if we could predict the severity of a car accident based on the data provided in the traffic report.

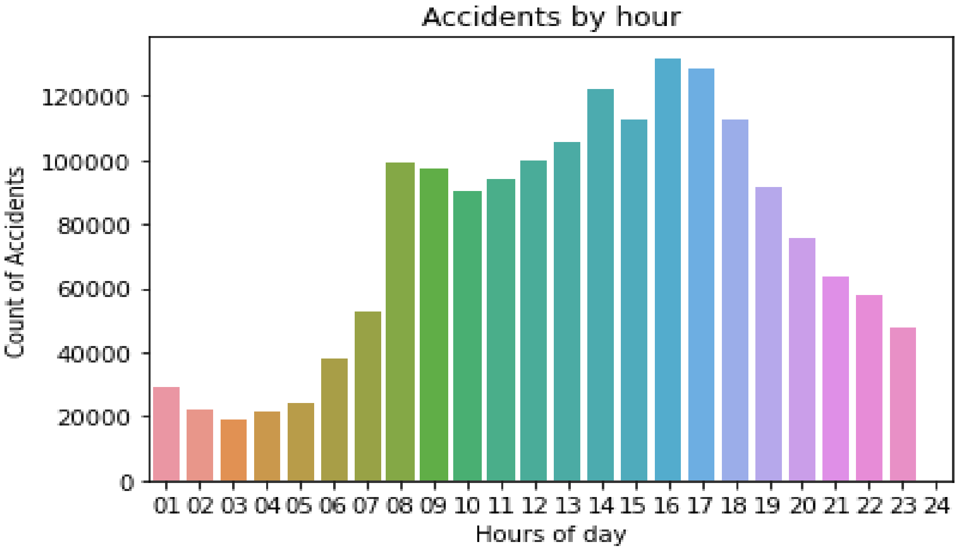
Data Wrangling

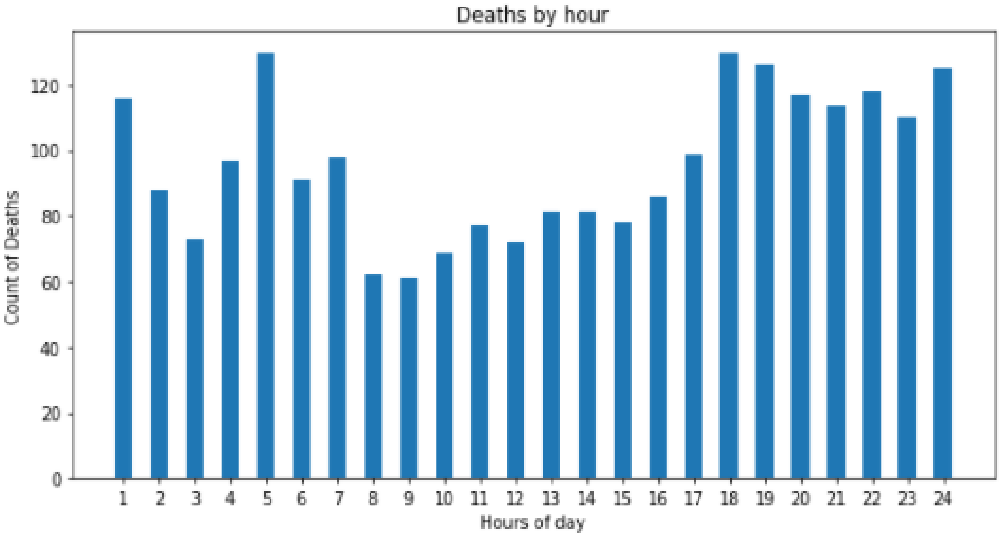
The NYC traffic data started with 1.7 million rows and 29 columns. In this section I was focused on making the data useable for further analysis, so I started with changing the data type objects to their appropriate option. We had time, categorical, numerical, and location data to complete further analysis on. Next, I wanted to clean the null data from our dataset. I focused on the columns with the most missing values first and removed anything with more than 50% missing data. Including but not limited to vehicle type codes, secondary contributing factors and off-street names. For the remaining data, I considered inputting the most common variable or random variables for missing values but decided against it. All the remaining data was categorical and random variables would only give my data more noise. I also strayed away from using the most common variable because it would then overrepresent the data.

Exploratory Data Analysis

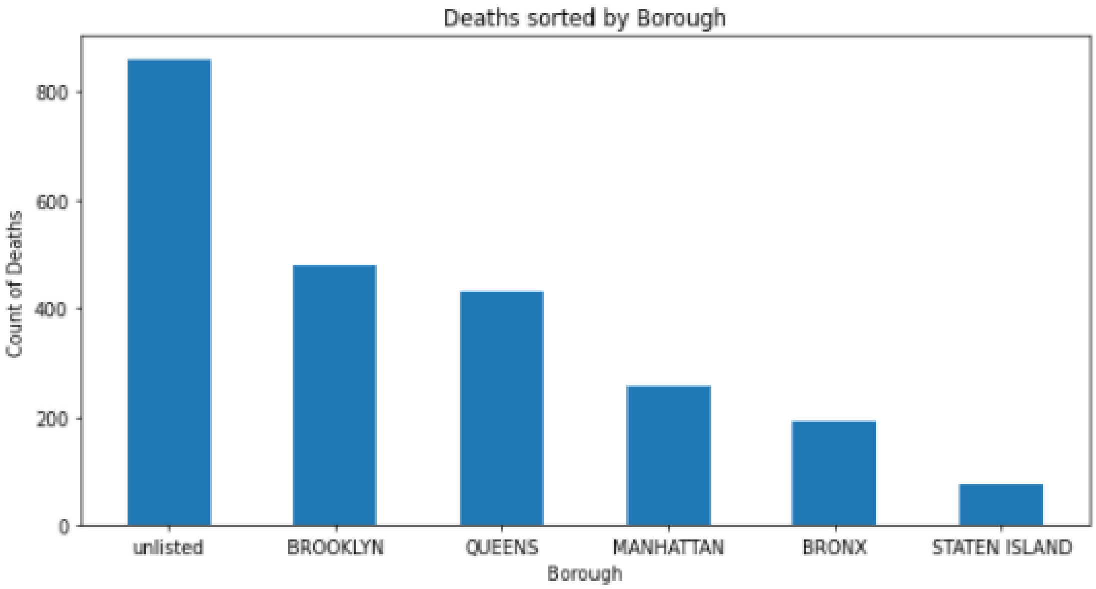
In this section I was focused on getting familiar with the data and looking for anything that stood out to me upon initial observation.

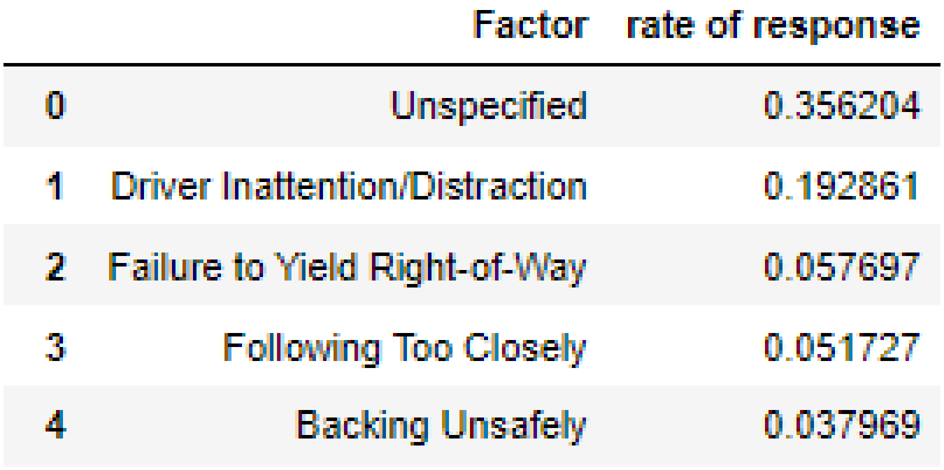
From the graphs below one can see that while total accidents peak around 4 p.m. deaths reach peak around 5 a.m. and 8 p.m. While human behavior plays a huge part in the type of accident, we can begin to draw our ideas that maybe lighting may effect the severity of the car crash as well as the types of things that are occurring during those times of day.





An unfortunate discovery is that most of our categorical data is unlabeled, this can also be seen in contributing vehicle factor where the largest input by more than double was “unspecified.” As well as locational data such as Borough where most of the data is “unlisted.”





	Factor	rate of response
0	Unspecified	0.356204
1	Driver Inattention/Distraction	0.192861
2	Failure to Yield Right-of-Way	0.057697
3	Following Too Closely	0.051727
4	Backing Unsafely	0.037969

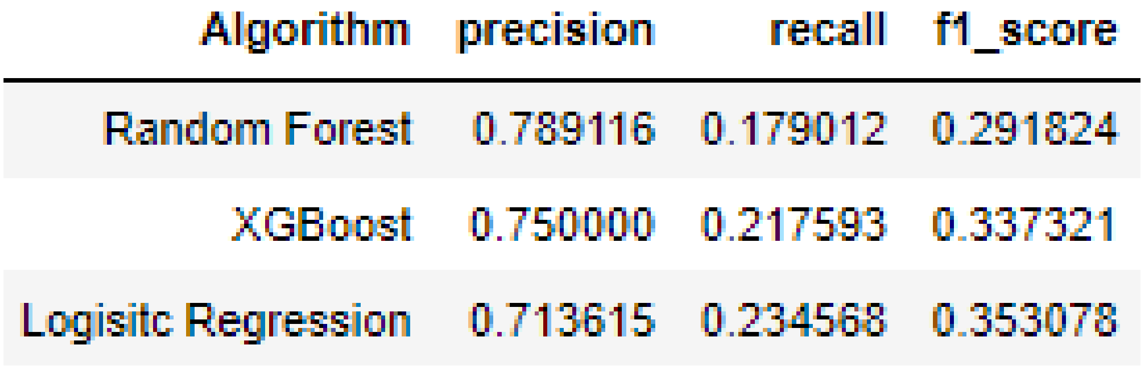
Preparation and Analysis

Preparation was generally easy in our case since our variables were categorical the data did not need to be normalized. I did implement dummy variables for categorical options, which increased the width of our data tremendously and caused some problems later.

Next we focused on Model selection where I decided on three models to use, XGboost, Random Forest, and logistic regression. Random forest works well with categorical data and is notorious for working well with data with high number of features which seemed like a good fit for this study. A major downside is that random forest can be considered a black box mechanism where much of what goes on in the model cannot be explained or controlled easily. Next, logistic regression was a safe pick for this problem since it is easy to implement, and the model is straight forward in production. Logistic regression does have its downsides, it is known for having a hard time working with many features at once and has a higher tendency to overfit data. Lastly, Xgboost seemed to be the most promising model coming into the study, it handles missing values well (which we had a lot), handles large data well, and has good execution speed. The major downside of XGBoost would be the high number of hyperparameters that could complicate things and lead to over-under fitting if one is not careful. We ran simple model, fit train scenarios, before doing a random search across hyperparameters to find the best options. (for more in information about the hyperparameters please check the model metric file)

The best metrics to rate these models would be precision, recall and f1-score since we a looking at a categorical problem with a simple yes/no answer. We focused more on recall because we wanted to predict as many correctly as possible like other high-risk situations such as fraud where False negatives can be more detrimental than False positives.

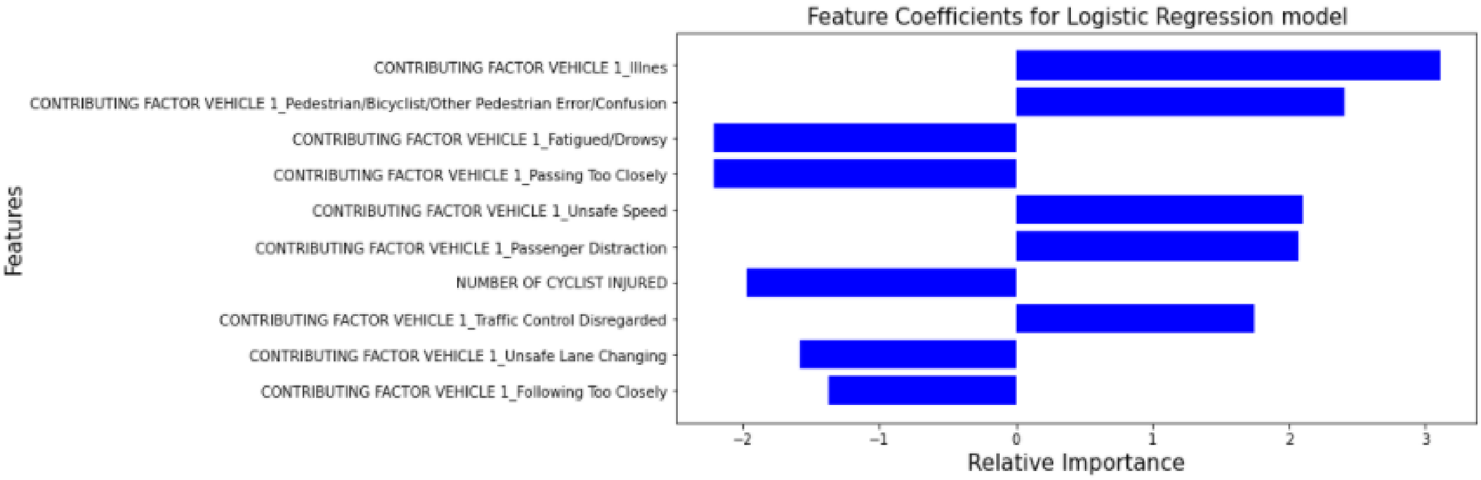
Our highest recall was done by our logistic regression model with an abysmal 23%. XGboost and Random Forest registered a recall of 22% and 18%, respectively. These models even after hyperparameter tuning and cross validation still fail to provide an algorithm that can predict fatality of car accidents at an acceptable rate.

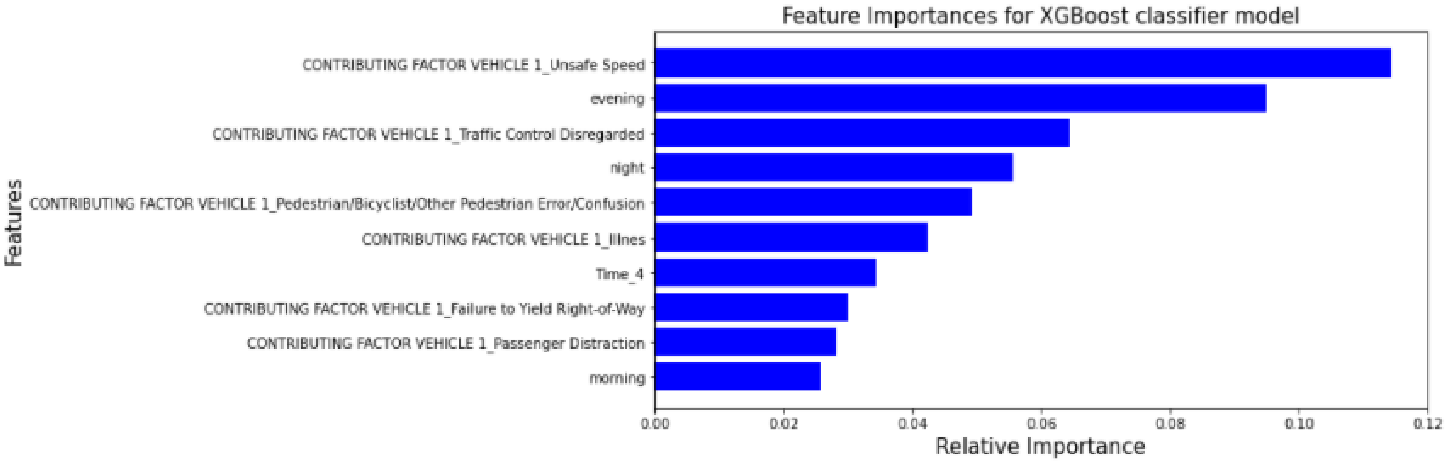


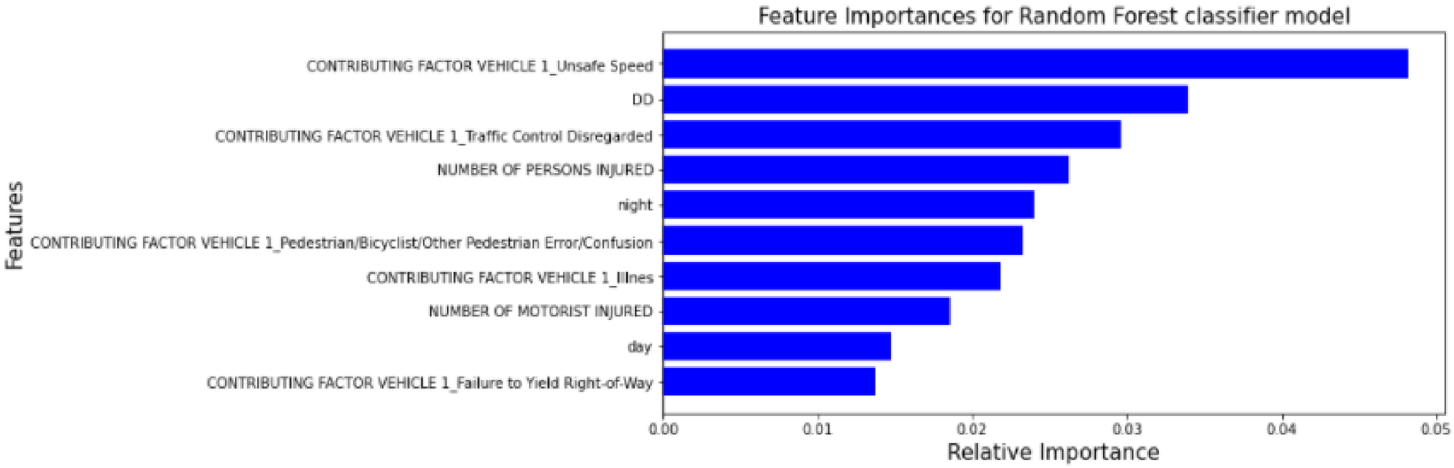
Algorithm	precision	recall	f1_score
Random Forest	0.789116	0.179012	0.291824
XGBoost	0.750000	0.217593	0.337321
Logisitc Regression	0.713615	0.234568	0.353078

While our models were not very good at encompassing all the car accidents that caused deaths, we can look at their feature importance's to see what they weighed heavily, to maybe give us a clue on what features created the highest chance for a death. Looking at all three models for their feature importance's. It was clear that "unsafe speed" was the largest factor in this study. Other factors such as "evening" and "pedestrian confusion" also weighed heavily across the board. While these importance's make sense, we must remember that the model was not very good at predicting the outcomes in the first place. Therefore, these feature importance's may not seem as important as we think.

Below are the feature importance's for all three models and the weight of the top features







Takeaways

While my test did not end with a model to predict car accident deaths, there are ways to improve on what was captured in this study. Locational data would have been a very interesting item to include, but the data provided was very sparse and during EDA it looked like only certain parts of the city would input location (this could be due to how each precinct is run and what is mandatory on the police reports). Finding streets or highways that are locations with high severity of car accidents, could provide immediate reason to increase safety in those areas and look further into why. Another item that, now thinking back on was obvious, would be to take a closer look at days before and after holidays and see the rates on those days since those were the days of most travel. I think one of the most interesting ideas to come from this would be using the weather to predict accidents, especially in New York where there are all four seasons. Including weather as variable would not only increase our model's accuracy but could also help locate areas that are more dangerous on days with certain conditions. This could lead to adding new signs to slow down or improving drainage systems in those locations

Future Research

This data would be great to compare against car accidents for AI controlled cars. The clearest hurdle that autonomous cars must overcome is gaining the trust of the public and government. Using this data and comparing it to autonomous car accident rates will give people a solid baseline to compare on. Furthermore, this data could be split into timeframes based on new car requirements, such as back-up cameras or auto-breaking assist, to really see if those functions make a difference.