Exploring Bicycle Collision Factors in New York City

Adam Ziel and Harry Gordenstein
Machine Learning
Professor Bargshady
April 14, 2018

Motivation

In 2015 there were 45,000 reported collisions between motor vehicles and a cyclists[1]. As bike couriers in downtown Boston, we want to explore the significance of certain riding conditions as they pertain to vehicle-cyclist crash likelihoods. From personal experience, it was intuited that weather and traffic conditions would likely be useful features in determining the presence of this correlation. The base collision data was taken from NYPD Motor Vehicle Collisions, provided through NYC OpenData[2]. With this, supplemental data was appended from the datasets Citi Bike Daily Ridership and Membership Data, Bicycle Counts, NOAA Online Weather Data, TLC Trip Record Data, and Uber Pickups in NYC[3,4,5,6,7].

Dataset Assembly

The following sections outline the justification and methodology of combing each discrete dataset into a final, aggregated repository. It was decided to find one contiguous year where each dataset had clean data available. It was found that June 2015 - July 2016 was the optimal span given the available data.

Collision Reports

Overview: Vehicle-related collision reports affecting pedestrians, cyclists, and other vehicles.

Justification: This was deemed the clearest metric for measuring danger posed to cyclists.

Caveats: This data does not include unreported accidents, as well as accidents not involving vehicles (e.g. a cyclist slipping out on black ice).

Resolution: The data was deemed the best option available.

Bicycle Traffic

Overview: Citi Bike Daily Ridership and Membership Data gave daily totals for ridership. Bicycle Counts gave daily totals for bike riders across various bridges throughout the city.

Justification: We wanted a metric to display the total number of bikers on the road at any given time, or at least a relative density.

Caveats: Citi Bikes are geared towards people who don't own bikes, and therefore may not be reflective of the total bike population. Similarly, the bridge data does not account for bikers who do not cross any bridges in a given day. Bridge data was not available for a contiguous year.

Resolution: As shown in Figure 1, when scaled, the two datasets were correlated within 93% over a 30 day sample. Therefore, it was decided that the Citi Bike data would suffice for measuring the general bike ridership on any given day.

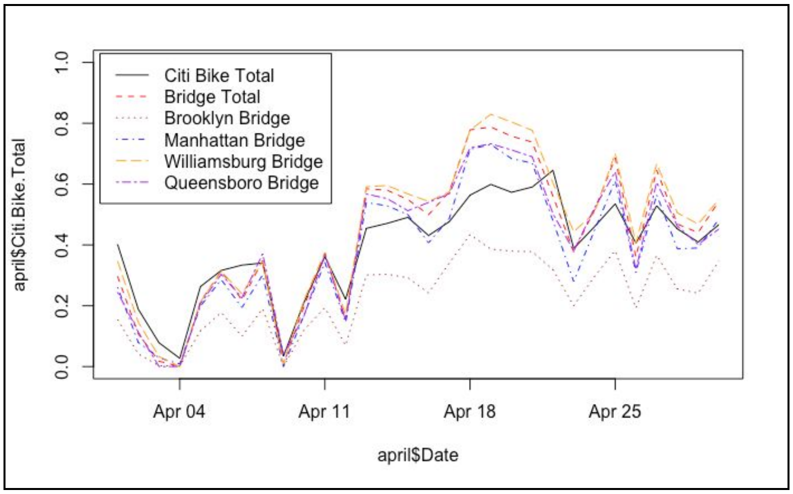


Figure 1: Correlation between Bridge ridership, and Citi Bike ridership

Uber Traffic

Overview: Uber Pickups in NYC recorded Uber trip data, and TLC Trip Record Data recorded Yellow Cab data.

Justification: We wanted a metric to display the total number of Ubers on the road at any given time. Ubers can often pose threats to cyclist as they frequently park in bike lanes when picking up or dropping off passengers.

Caveats: Uber data was not available for a contiguous year. Yellow Cab data was not formatted in terms of daily totals.

Resolution: The correlations between the two datasets were observed in Figure 2. While their long term correlations differ due to the changing climate taxi usage, it was decided that short term correlations would be strong (e.g. spikes in usage for both companies on a high-traffic night like New Years Eve). Therefore, Uber traffic could be replaced with Yellow Cab traffic, and still preserve the same intended information.

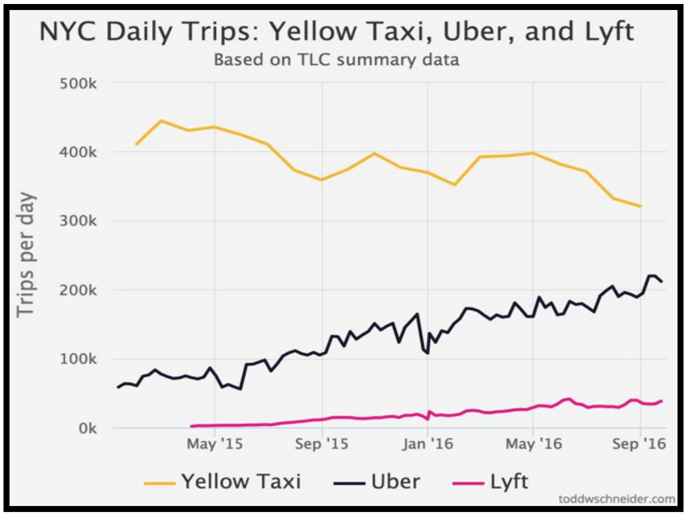


Figure 2: Trends in Uber, Lyft, and Yellow Cab ridership

Weather

Overview: NOAA Online Weather Data provided various weather metrics per day in NYC.
Justification: We wanted a metric to display weather conditions that may factor into a crash.
Caveats: The selected dates included a very anomalous, mild winter, in addition to an outlying data point which corresponded to SNowzilla - NYC's 2nd largest snowstorm on record, which took place January 23rd, 2016.
Resolution: While it would have been more optimal to cover a wider range of dates to soften the effects of 2015-2016's erratic weather patterns, all outlying points were maintained, so long as data was available for the remaining features (i.e. the blizzard caused similar outliers and/or missing data points in the Yellow Cab data, as well as the Citi Bike data.

Figure 3 gives an overview of the features selected for the final, aggregated dataset.

Feature	**Data Type**	**Description**
Month	Integer	Numeric month (i.e. Jan = 1 … Dec = 12)
Weekday	Integer	Numeric day of the week (Monday = 1…Sunday = 7)
DailyBikeRides	Integer	An approximate number of cyclists on the road
DailyCabRides	Integer	Total Yellow Cab trips
MaxTemp	Numeric	Maximum temperature
MinTemp	Numeric	Minimum temperature
AvgTemp	Numeric	Average temperature
Precip	Numeric	Total precipitation
SnowFall	Numeric	Total snowfall
SnowDepth	Numeric	The total snow depth (included from past snow storms)
NumCyclistsInjured	Numeric	Number of reported cyclist collisions

Figure 3: Feature descriptions

In addition, two features were extrapolated to give better metrics on the number of reported bike collisions, as seen in Figure 4.

Feature	Data Type	Description
CollisionsPer10000Cyclists	Numeric	Number of collisions per 10,000 cyclists
CollisionRisk	Factor	Binary factor for ‘Low’ and ‘High’ risks

Figure 4: Extrapolated feature descriptions

CollisionsPer10000Cyclists was chosen as a better regression output as it it effectively normalized the collision data with the number of total riders (i.e. 20 collisions implies different risks if the rider sample size is 50 vs. 50,000). CollisionRisk split the data at the arbitrary threshold of 5 collisions per 10,000 cyclists, as seen in Figure 5. Anything above this threshold was deemed High risk, and everything below Low risk. This was chosen to shift the problem from regression to classification. Due to immense amount of noise present in a given traffic scenario (e.g. dozens of independent actors, texting while driving, sun glare, alcohol, cyclist error etc.) it was deemed unrealistic to predict exact number of expected collisions. Instead, the classification approach softened the decision margin, therefore accounting for much of the present noise.

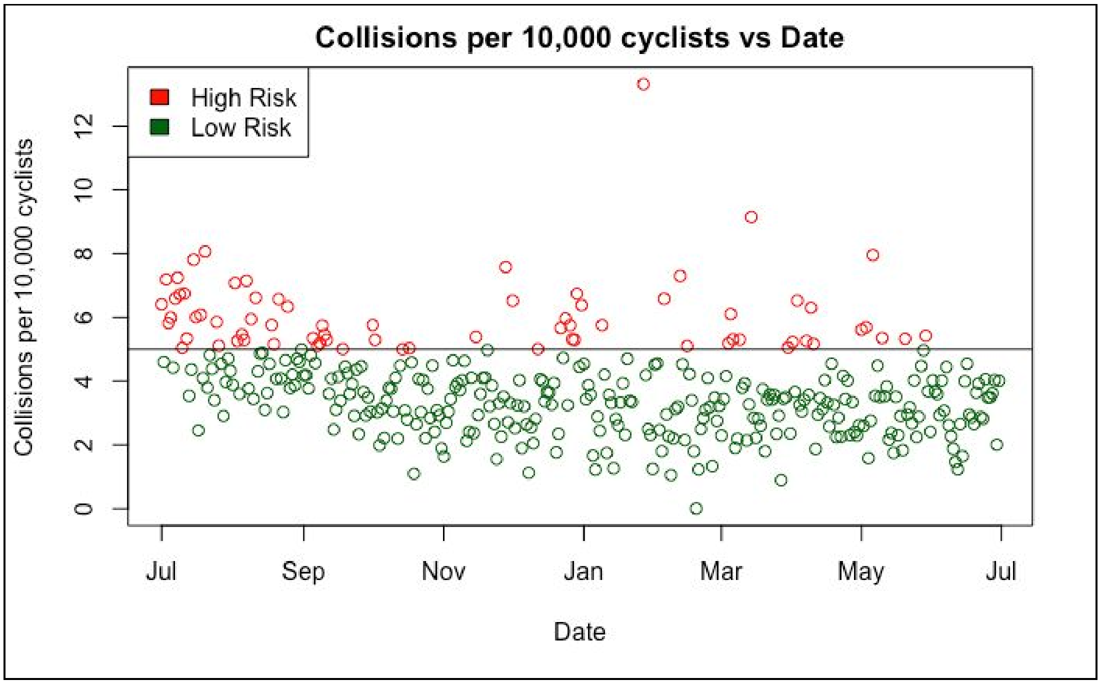


Figure 5: Binary split to separate data into Low and High risks

Figure 6 shows a correlation plot of the initial feature set. MinTemp and MaxTemp are 96% correlated, MaxTemp and AveTemp are 99% correlated, and MinTemp and AveTemp are 99% correlated. This suggests that only one of those features is necessary for the model. We can also see that number of cyclist injured is highly correlated at 71% for Min and Max Temperature and 72% for Average Temperature. This means they’ll definitely be important features. The 63-70% correlation between daily bike rides and the temperatures shows that weather is a big determining factor for how many people want to ride.

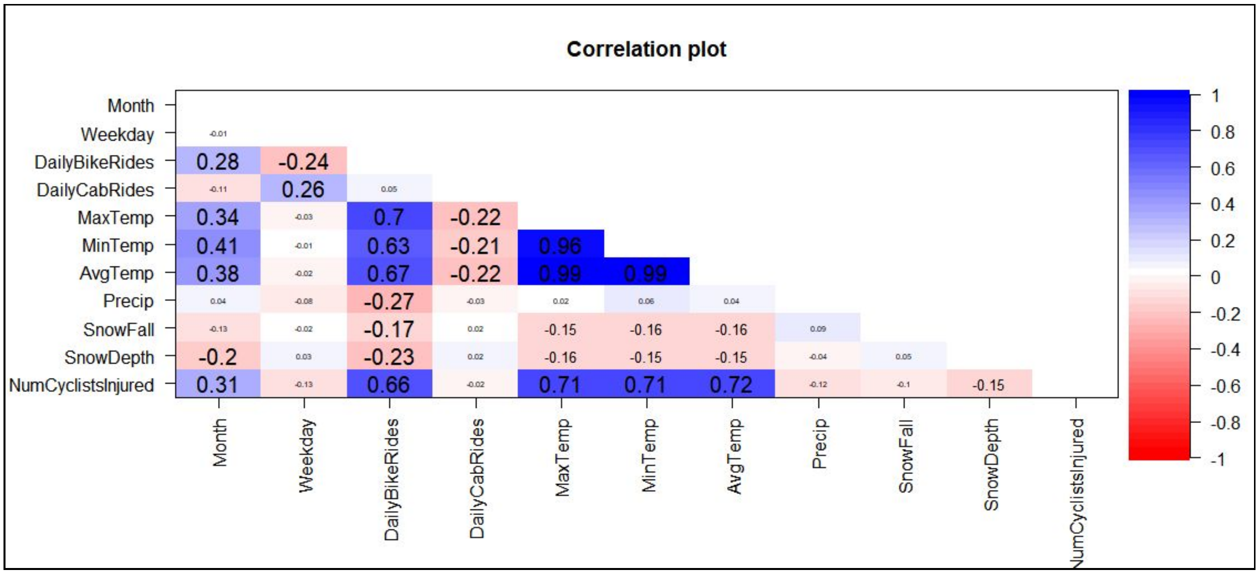


Figure 6: Correlation plot of initial features

Feature Reduction

Regsubset

We then put the data through the `regsubset()` function from the `leads` package to find the best subset of our factors. The nvmax parameter is set to 11 so that we use the full dataset if it is necessary. Within the summary of the model fit we find what number of variables optimize 5 measures of the model. For R Squared (rsq) the model is optimized at 7 factors {Weekday, DailyBikeRides, DailyCabRides, MinTemp, SnowFall, SnowDepth, and AveTemp} where rsq = 0.3. Adjusted R squared (adjr2) is optimized at 6 factors {Weekday, DailyBikeRides, DailyCabRides, MinTemp, SnowDepth, and AveTemp} where adjr2 = 0.28. Mallows' Cp (Cp) has an optimal value of 7.8 with 7 factors {Weekday, DailyBikeRides, DailyCabRides, MinTemp, SnowFall, SnowDepth, and AveTemp}. Schwartz's information criterion, BIC (bic) is optimized with 3 factors {DailyBikeRides, MinTemp, and AveTemp} giving a value of -87. These values aren't great however that is not a concern because we are more interested in the subsets that we get. Since none of the models contained Month, MaxTemp, or Precip they are not adding anything if we include them in our final model. Figure 7 shows the output for the r squared value from summary and the features that are included in the calculation.

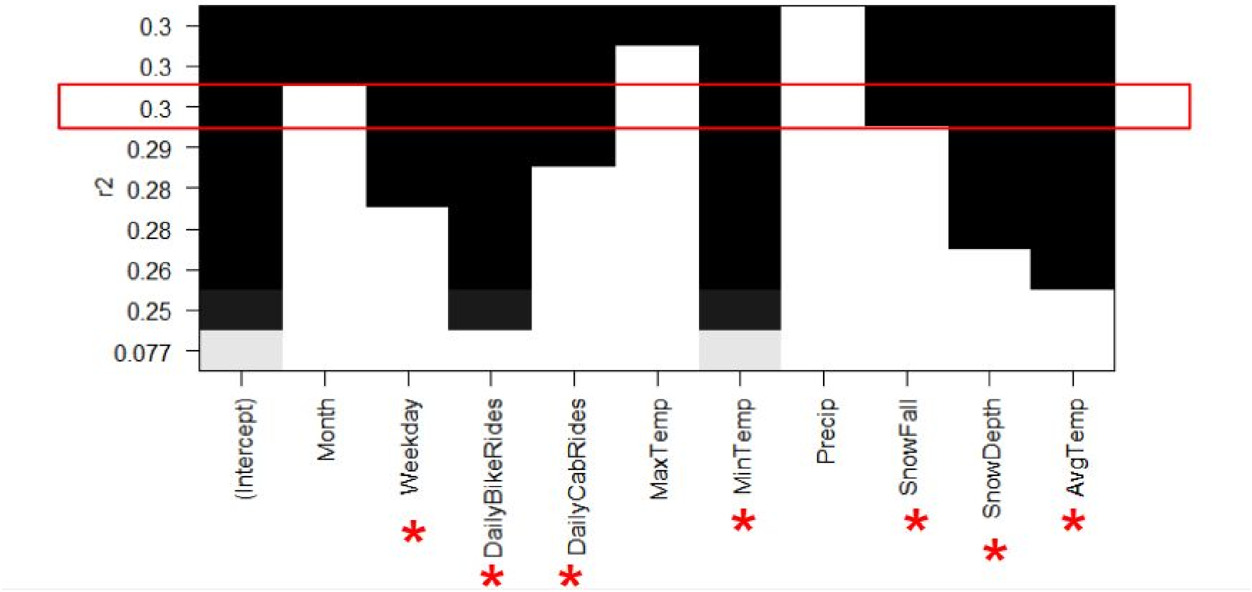


Figure 7: r Squared values from regsubset() displaying important features in black

Lasso

We further explore the important features by using `glmnet`, specifically running lasso (Least Absolute Shrinkage Selector Operator). Lasso produces a sparse model that is good at choosing factors in datasets with high noise. Lasso chooses {Month, DailyBikeRides, MinTemp, SnowFall, and SnowDepth} at the min lambda. and chooses only DailyBikeRides, MinTemp, and SnowDepth when lambda is at 1 standard error away from the min. Figure 8 shows where the minimum lambda and lambda at 1 standard error as the left and right vertical dotted lines respectively. Figures 9 and 10 show the coefficient values calculated at lambda equal to the minimum and one standard error away, values that shrink to zero are represented by a '.' meaning they are not necessary to predict the outcome of our data. Any remaining values have an importance relative to their coefficient values.

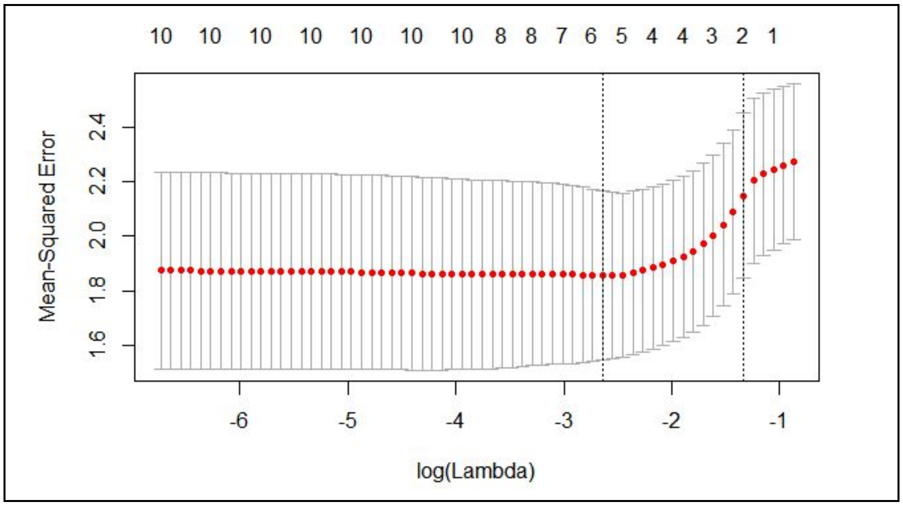
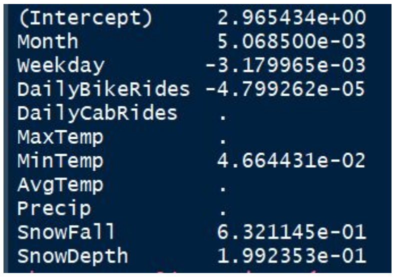
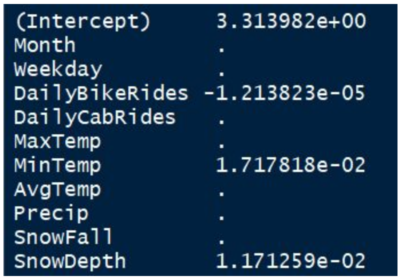


Figure 8: plot of the lambda values from the lasso model



```
(Intercept)      2.965434e+00
Month            5.068500e-03
Weekday         -3.179965e-03
DailyBikeRides  -4.799262e-05
DailyCabRides    .
MaxTemp          .
MinTemp          4.664431e-02
AvgTemp          .
Precip           .
SnowFall         6.321145e-01
SnowDepth        1.992353e-01
```

Figure 9: coefficients of lasso where lambda produces the smallest mean-squared error



```
(Intercept)      3.313982e+00
Month            .
Weekday          .
DailyBikeRides  -1.213823e-05
DailyCabRides    .
MaxTemp          .
MinTemp          1.717818e-02
AvgTemp          .
Precip           .
SnowFall         .
SnowDepth        1.171259e-02
```

Figure 10: Coefficients of lasso where lambda is 1 standard error above the smallest mean-squared error

PCA

To further explore feature significance, PCA was performed on the initial feature set. Figure 11 shows the variances corresponding to each set of Principal Components, and Figures 12 & 13 show the most significant in the first two sets of Principal Components, respectively.

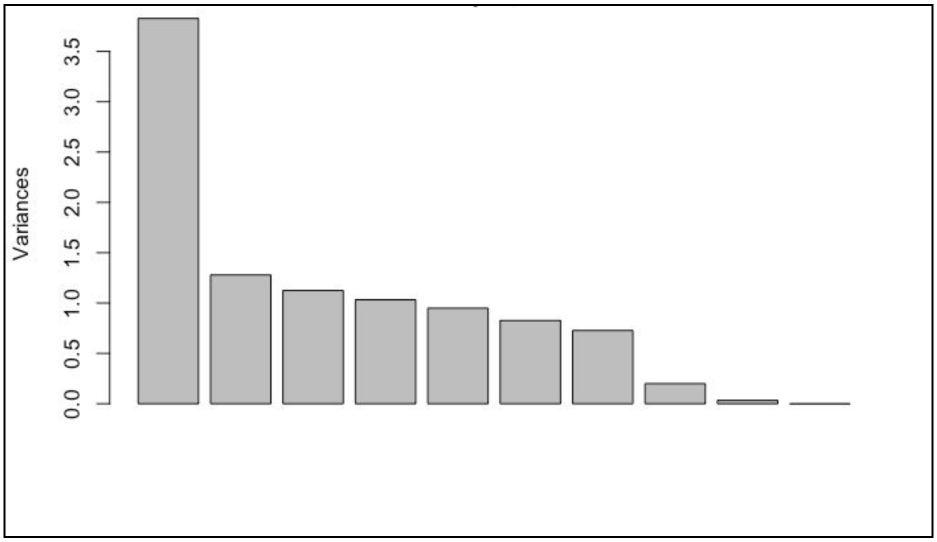


Figure 11: Variances for each set of Principal Components

Feature	PC1
AvgTemp	-0.4956175
MaxTemp	-0.4914201
MinTemp	-0.4891009
DailyBikeRides	-0.3968601
Month	-0.2560892

Figure 12: Most significant features in PC1

Feature	PC2
DailyCabRides	-0.6012972
Weekday	-0.5420302
Precip	-0.4932681
SnowFall	0.2499456
DailyBikeRides	-0.1932672

Figure 13: Most significant features in PC2

As expected based on the correlation plot in Figure 6, the three temperature features, MinTemp, MaxTemp, and AvgTemp, dominate the first set of Principal Components. Since their weights are all essentially identical, it was decided to drop MinTemp and MaxTemp, and then run PCA a second time.

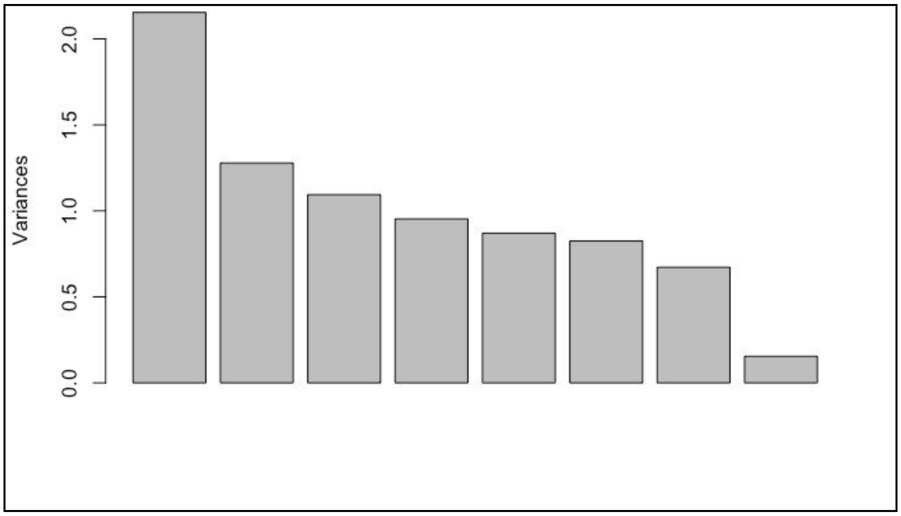


Figure 14: Updated variances for each set of Principal Components

Feature	PC1
DailyBikeRides	-0.5681321
AvgTemp	-0.5650564
Month	-0.4117587
SnowDepth	0.2713128
SnowFall	0.2286779

Figure 15: Updated most significant features in PC1

Feature	PC2
DailyCabRides	-0.6012972
Weekday	-0.5420302
Precip	-0.4932681
SnowFall	0.2499456
DailyBikeRides	-0.1932672

Figure 16: Updated most significant features in PC2

Omitting MinTemp and MaxTemp yielded a much more balanced feature representation in the first set Principal Components. They were replaced with SnowFall and SnowDepth in the updated trial. Both PC2's remained the same for both trials.

Random Forest Classification

After exploring feature significance with a variety of methods, it was decided to run the dataset through Random Forest to evaluate classification viability. As a result of the high correlations found between all the temperature-based features, MinTemp and MaxTemp were omitted from the input dataset.

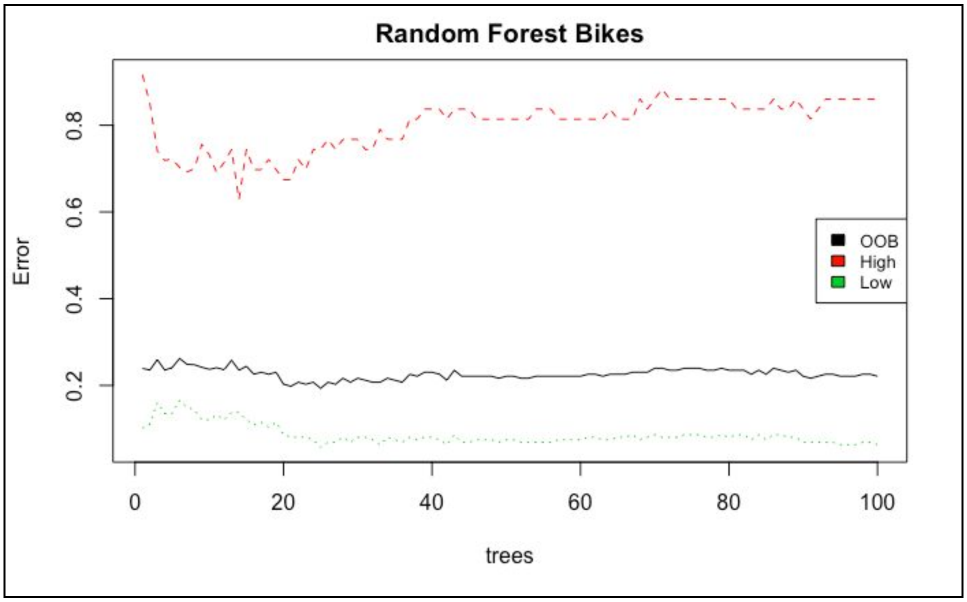


Figure 17: Random Forest plot

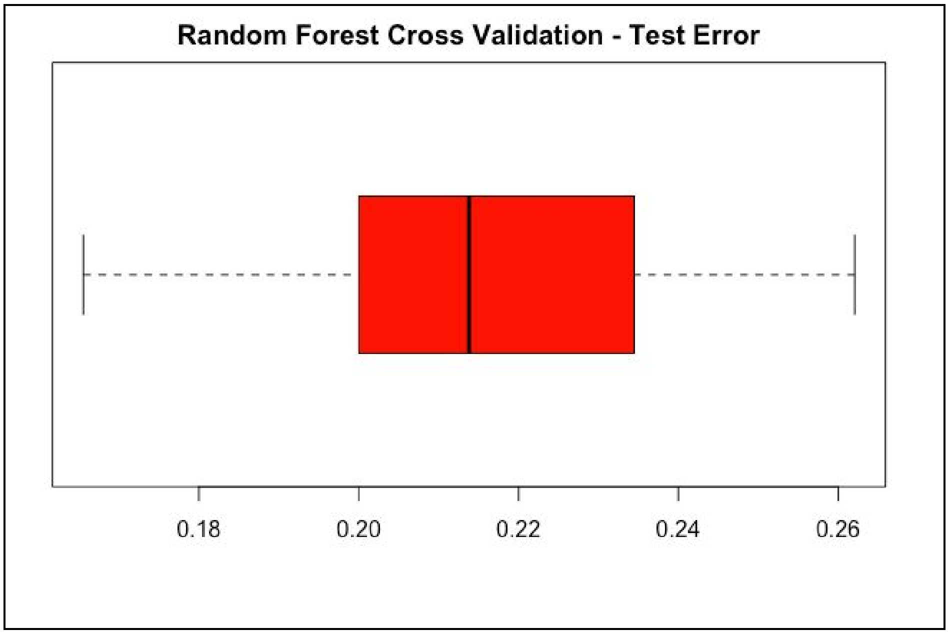


Figure 18: Random Forest test error

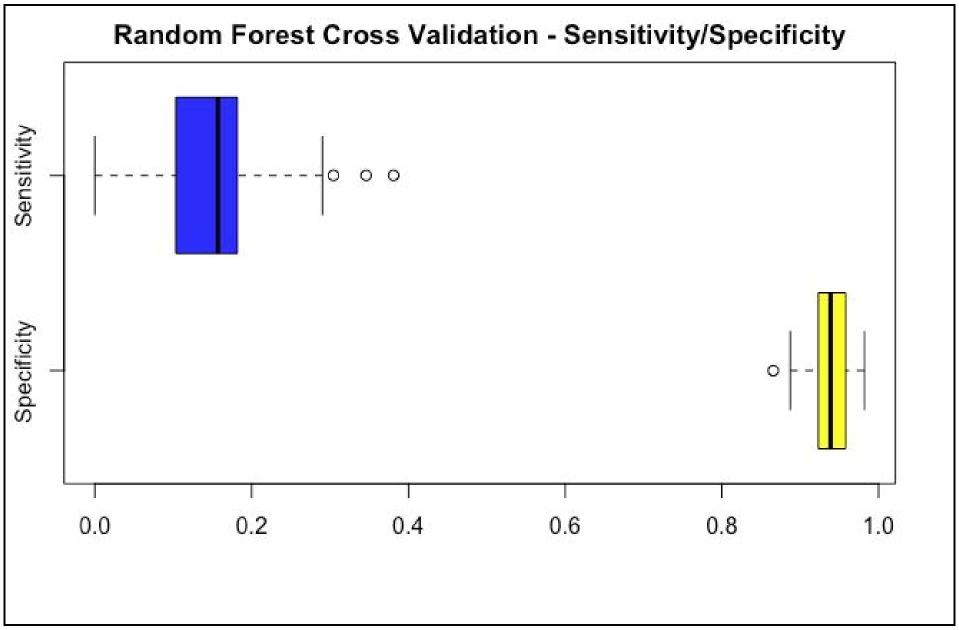


Figure 19: Random Forest sensitivity and specificity

The approximate error rate of 0.21 shown in Figure 18 was quite positively surprising given the small dataset, as well as all the forecasted noise in any given crash scenario. However, despite the model's favorable specificity, the extremely low sensitivity of the model was not permissible given the nature of the problem - it isn't as important to predict safe days accurately, instead it's much more valuable to the cyclist to predict the unsafe days.

To explore possible ways to mitigate this issue of glaringly low sensitivity, we looked at the way we initially split the data into Low and High collision risks. In the original arbitrary separation level of 5 collisions per 10,000 cyclists, the dataset only provided 71 high-risk days, and 291 low-risk days. The poor sensitivity may have been a result of simply not having many opportunities of predicting high-risk days. Therefore, the split was re-performed at the median value of collisions per 10,000 cyclists, 3.59, as shown in Figure 20.

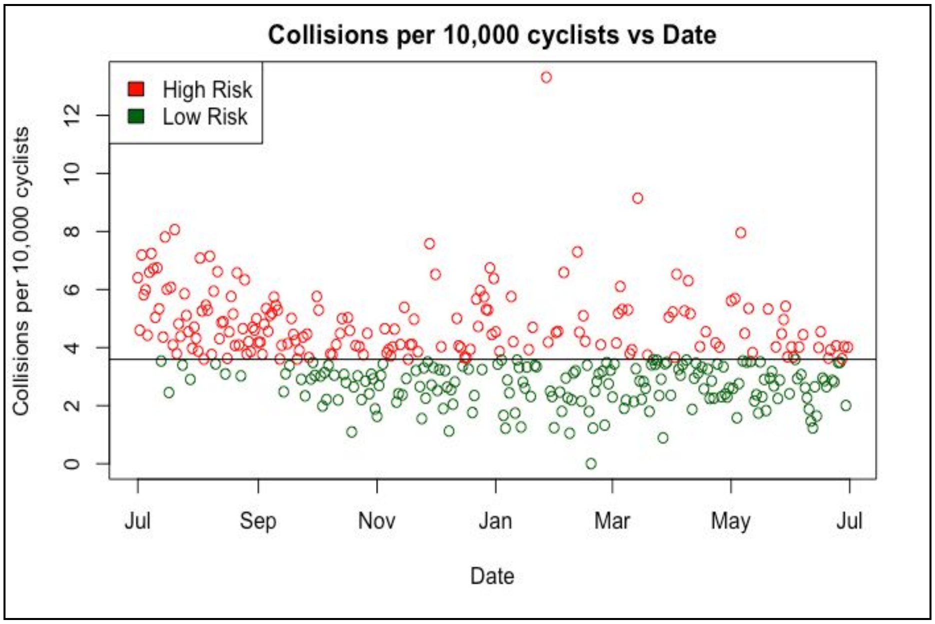


Figure 20: Reevaluation of factor thresholding

With this new separation point Random Forest was then ran again.

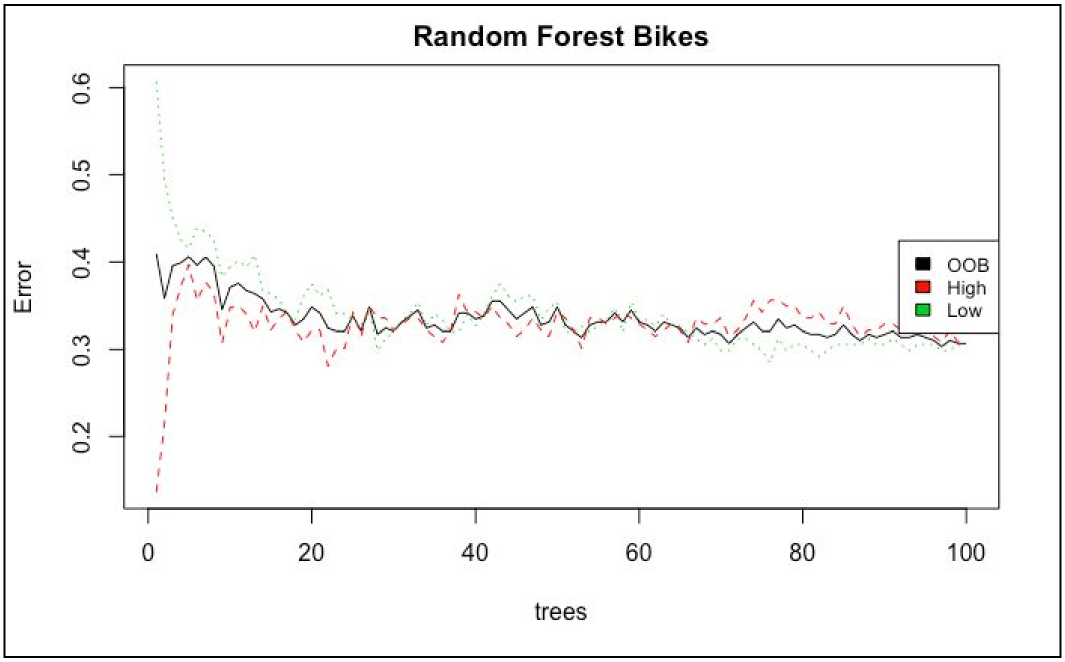


Figure 21: Updated random forest plot with new separation point

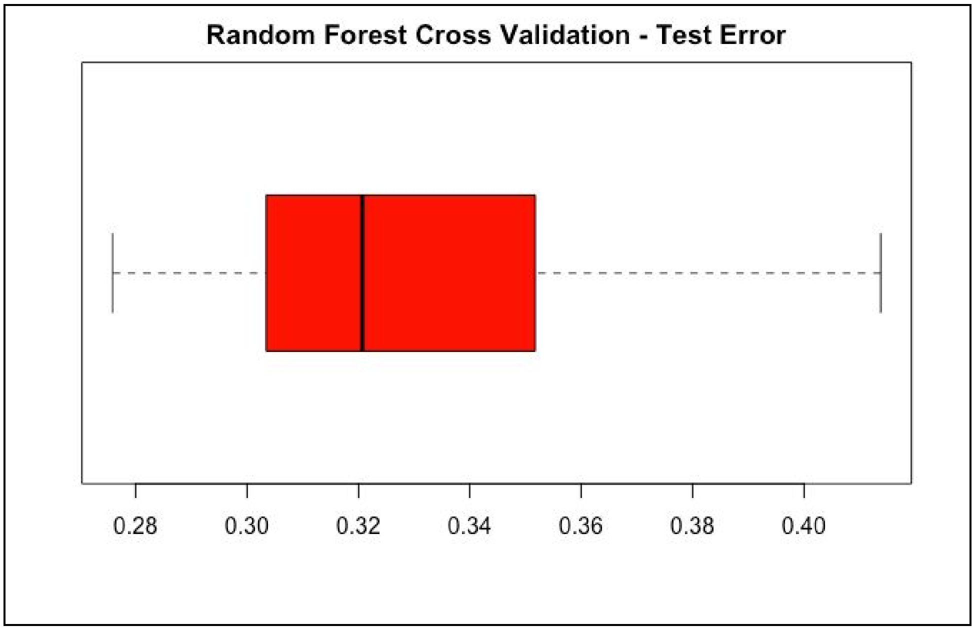


Figure 22: Updated test error

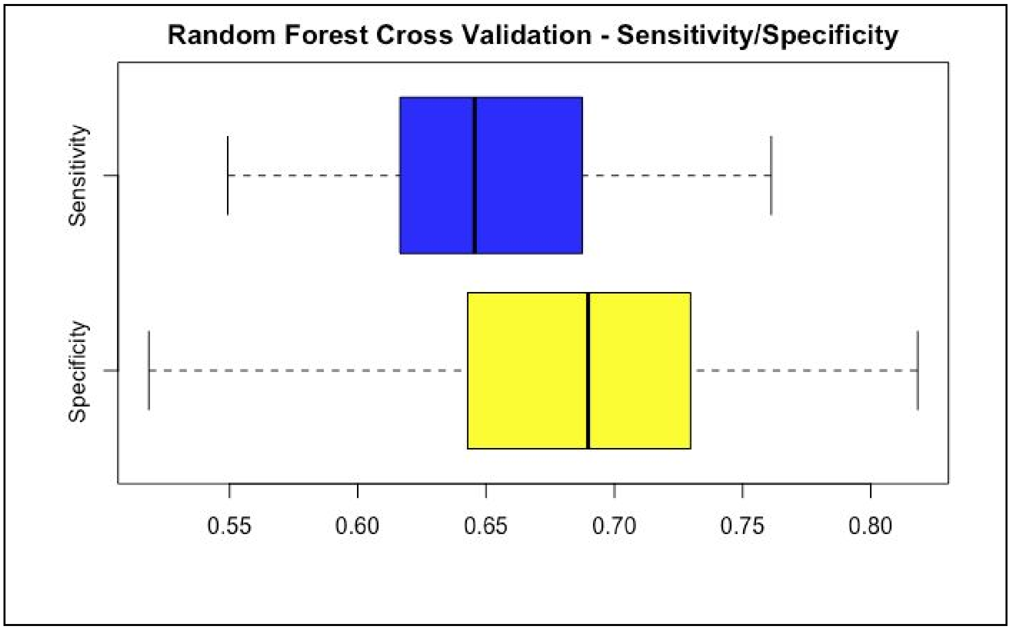


Figure 23: Updated sensitivity and specificity

Given that this new separation point was less intuitive, it make sense that was hit was taken on test error. However, it indicates that with more examples of high-risk observations, sensitivity improves as desired.

Conclusion

Our dataset was able to predict crash likelihoods with an error rate of approximately 32.1%. While this is not stellar by any means, we are optimistic on its potential for improvement given the ad hoc nature of the dataset, as well as the noisiness of the problem domain. Through several feature selection algorithms we found that our initial assumptions, about precipitation and Uber traffic causing crashes, were wrong. The biggest determining factors, from lasso where lambda is 1 standard error away from the minimum, were the number of cyclists on the road and the temperature, which are about 67% correlated alongside snow depth. Using best subset regression we found that these features were important as well as weekday, daily cab rides and snow fall. From a higher level, we were able to appreciate the overhead involved in building a novel, experimental dataset.

Future Work

The primary future goal is to aggregate more observations in the dataset. From there, the separation of the continuous crash data into discrete factors would want to be explored in more depth. Next, the results of the different feature reduction algorithms would want to be applied more rigorously in the classification process. Lastly, more classification algorithms would ideally be tested.

References

[1] Pedestrian and Cyclist Crash Information http://www.pedbikeinfo.org/data/factsheet_crash.cfm

[2] NYPD Motor Vehicle Collisions
https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

[3] Citi Bike Daily Ridership and Membership Data https://www.citibikenyc.com/system-data

[4] Bicyle Counts http://www.nyc.gov/html/dot/html/bicyclists/bike-counts.shtml

[5] NOAA Online Weather Data http://w2.weather.gov/climate/xmacis.php?wfo=okx

[6] TLC Trip Data http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

[7] Uber Pickups in New York City
https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city