Modeling Taxi Trip Distance from Trip and Fare Data,
Enriched with Zoning, Census, and Geographic Information

Anna Jurgensen

Since being able to predict the approximate distance a taxi drove could be used to help validate reported distance values (and values for the fares calculated from distance), for the capstone project using the 2013 NYC taxi trip and fare data I built and optimized a model for predicting trip distance from the information a driver would have at the start of the trip alone.

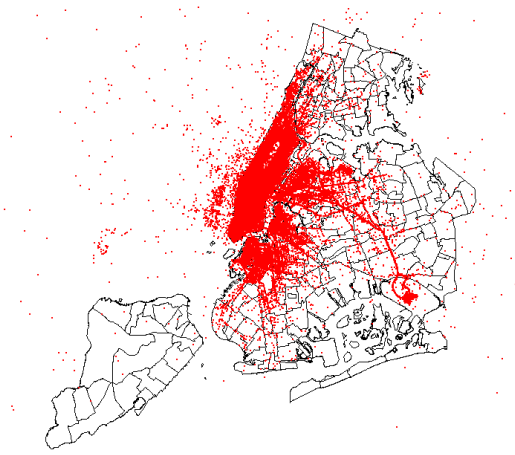
Basic Data Used & Feature Engineering

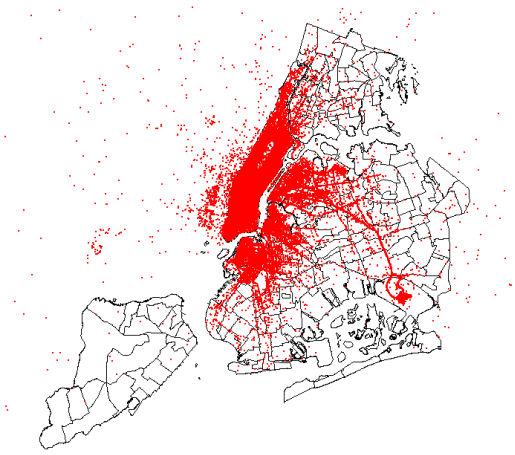
The initial one percent taxi trip and fare data received in class was the foundational data set used. This data contained the latitude and longitude of each ride's pick-up and drop-off locations, the date and time of each ride's pick-up and drop-off, the number passengers, and the distance travelled. Additionally, the data contained a number of the fields were excluded from the modeling effort since they contained information that a driver would have access to only at the finish of the drive (when the trip distance would be known anyway). In particular, the time that the trip took, the base and total fare amounts, the tax and tip amounts, and the payment type fields were all excluded from the modeling effort. Of note, while the amount of tolls is also part of the collection of payment-type data fields, a driver could know before starting the trip if the route he decided to drive would include tolls and, through GPS or an especially keen knowledge of the roads, could know the amount of tolls that would be charged. Lastly, two field of categorical data with too many different categories (the hack license and medallion numbers for associated with the taxi and driver for each ride) were simply omitted from the modeling and testing data.

The major feature engineering I did using the original data was creating new date and time features. Splitting up the original data/time information, I created a separate month feature, day feature, and an "hour of the day" feature that simplified times by lumping them into one of 24 hour-long bins. I also used the trip dates to create a day of the week feature and a binary "is holiday" feature using the US federal holidays of 2013, and from the day of the week feature I create a binary "is weekend" feature. Further, from the 24 hour bins I created "time of the day" based on my intuition about school, public transportation, office, and restaurant hours. This feature lumped several hour blocks together for a total of seven blocks: "early morning" (04:00-06:59), "morning" (07:00-10:59), "midday" (11:00-13:59), "afternoon" (14:00-17:59), "evening" (18:00-20:59), "night" (21:00-23:59), and "late night" (00:00-03:59).

Additional Data Sources & Feature Engineering

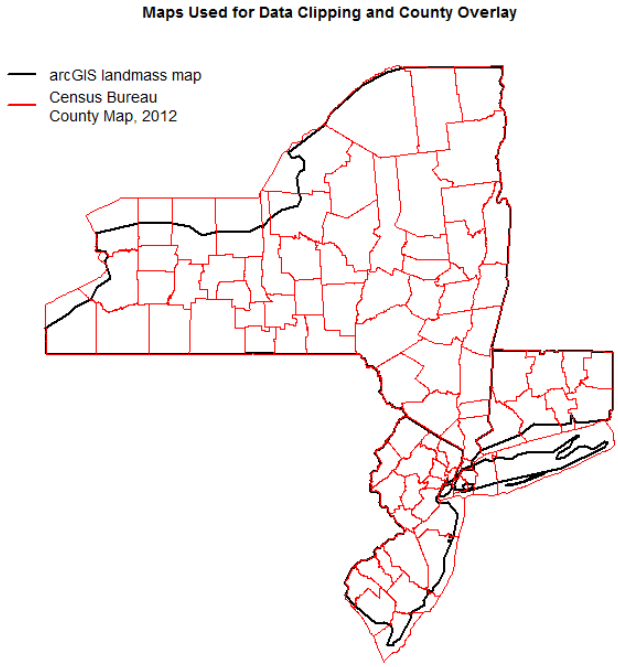
In addition to the foundational data set, data from a number of outside sources was incorporated. First, a map of the Neighborhood Tabulation Areas (NTAs) of the NYC metro area (as a shapefile) was used to spatially clip the projected pick-up and drop-off coordinates that were recorded as outside the NYC metro area. The coordinates that were clipped in this step were clipped again using a map shapefile from ArcGIS that closely follows the land mass and borders of New York and New Jersey to eliminate any unusual coordinates (such as those that appear in the middle of the ocean). Trips that had both pick-up and drop-off coordinates within the landmass of the NYC NTA map or the NY and NJ states' maps were included for use. (The effect of the two clipping steps can be seen in comparing the two maps, below, with pick-up points before and after clipping plotted on the NYC NTA map.) The NYC NTA map was used again as an overlay to determine the NTA in which a trip's pick-up or drop-off was located. Similarly, for those locations outside of the NYC metro area, the state and county maps for New York and New Jersey provided by the US Census Bureau (and available through package ***tigris*** in R), were used to determine the state and county of a pick-up or drop-off. I then created a very large volume of binary features indicating whether





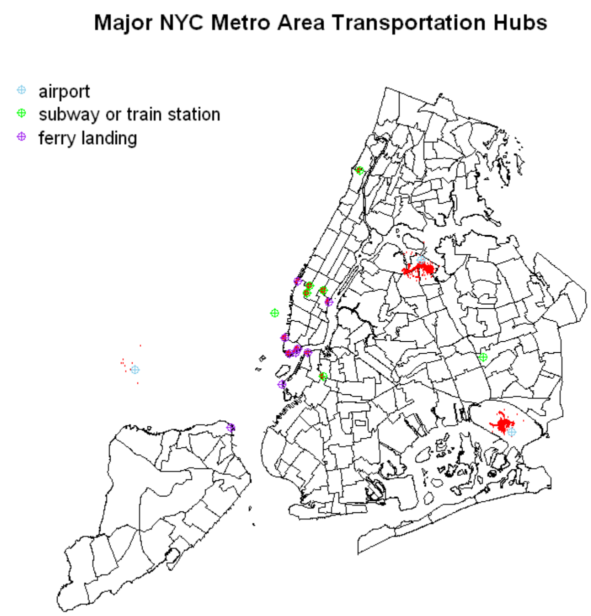
or not each pick-up and drop-off occurred in a given NTA or county. (The vast majority of these 400+ features were eliminated before modelling by simply limiting those included to those with a correlation coefficient with a magnitude of .02 or greater.)

Both of the maps used to determine the area of pick-up and drop-off locations, NTA or county, were chosen because the areas they defined were the same areas used for the American Community Survey data available from the year prior to the taxi data (2012), and so ACS data for the areas seen could then be used to create additional features.

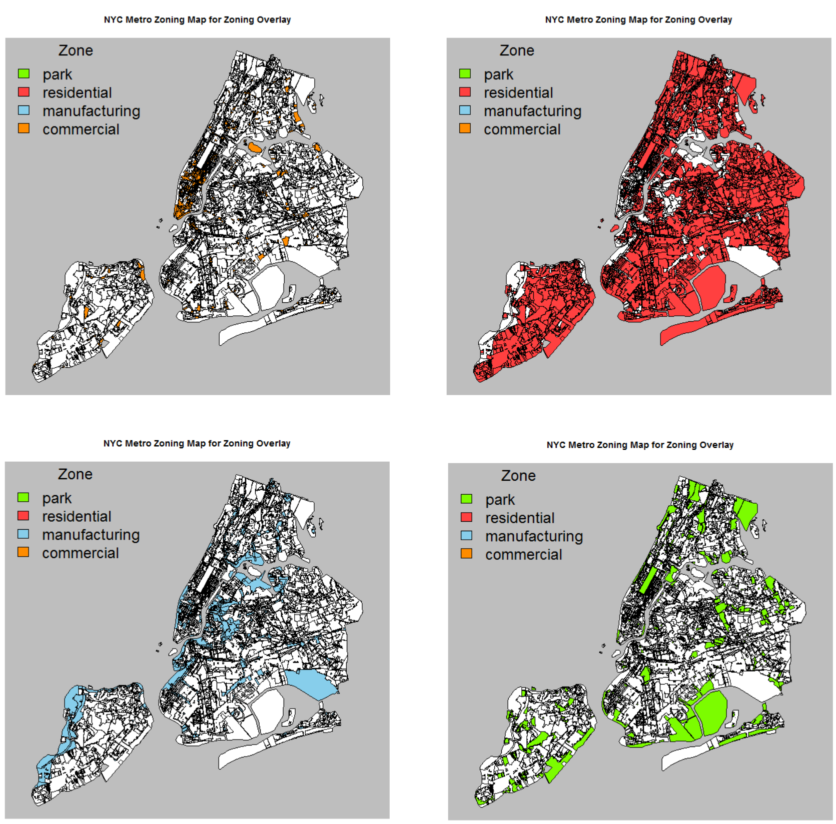


The ACS data used came from two different sources. First, NYC Open Data provided ACS data for 2008-2012 tabulated by NTA. Second, ACS data, by county, was available from the Census Bureau and in R through the package ***acs***. Using the pick-up and drop-off NTA for those that occurred within the NYC metro area, and the pick-up and/or drop-off state and county for those that occurred outside of NYC metro area, I joined ACS data that I thought would be relevant to taxi rides. In particular, I created features for each trip indicating the total number of housing units in the NTA/county in which the pick-up and drop-off occurred; the number of occupied housing units; the number of homes with access to no vehicles, 1 vehicle, 2 vehicles, or 3 or more vehicles; the total number of

workers living in the area; the percentage of workers that take taxis or motorcycles to work; and the percentage of civilian workers living in the area that work in (a) finance, real estate, or insurance, (b) information, or (c) professions, science, or management. I chose this information in particular from the vast options that the ACS data offered since I thought these would reflect the number of people using taxis, and to some degree the situations in which they use them (such as using them as a means for commuting, using them to travel to the airport for business trips, or just taking them after a late night out on the town).



In addition to using the ACS data, county, and NTA data, I created features by collecting the coordinates of transportation hubs (airports, ferry landings, and major subway and train stations, as listed by Wikipedia) and then calculated for each of these the distance from the pick-up and drop-off coordinates. Using those distances I created additional binary features that indicated for each ride whether the pick-up or drop-off was in the vicinity of any of these specific locations, within the vicinity of any type of these locations ("ground" for train and subway, "ferry" for ferry landings, and "airport" for the three major area airports), or was within the vicinity of *any* type of transportation hub. "In the vicinity of" was defined as 200m for train, subway, and ferry stations, and 2,000m for airports (pick-ups within the vicinity of these locations can be seen in red on the map above).

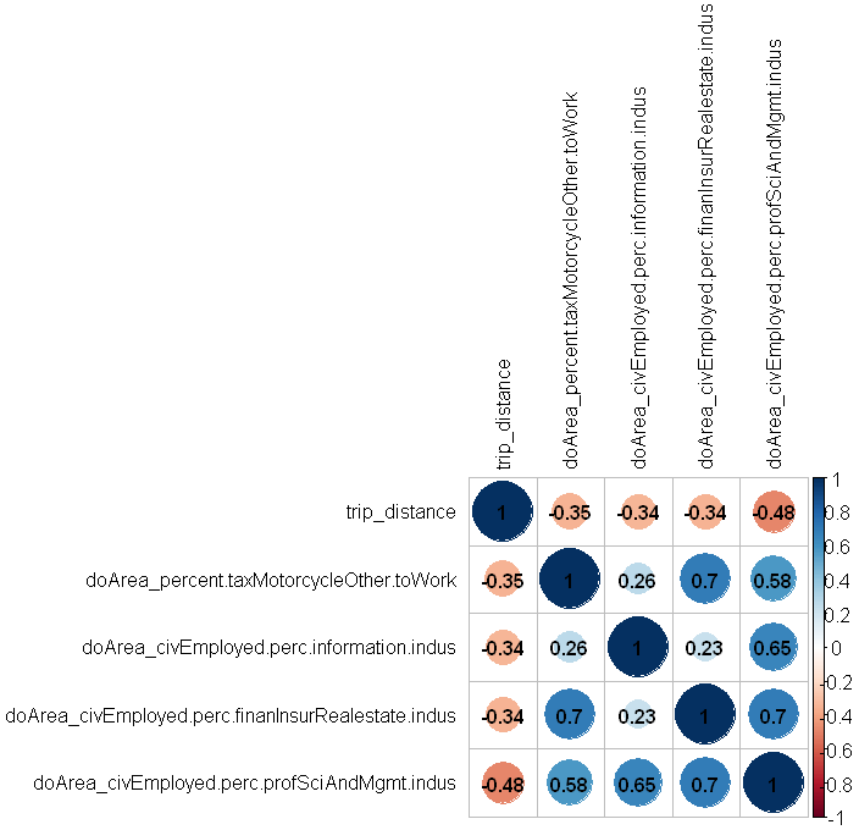


Lastly, zoning maps (again, as a shapefile) from NYC Open Data were used to create features based on pick-up and drop-off locations. From these maps, binary features indicating whether a coordinate occurred within a residential, commercial, manufacturing/industrial, or park zone.

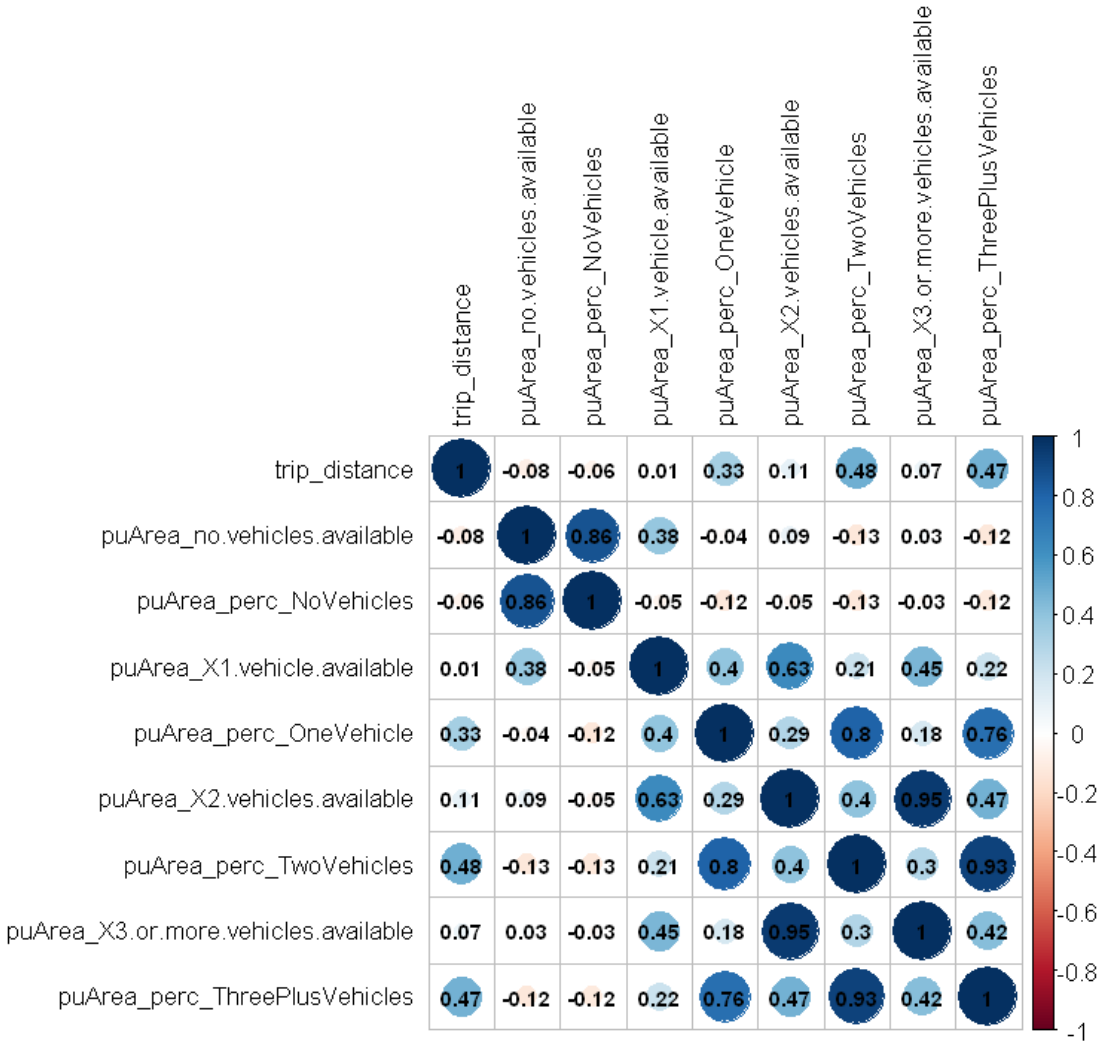
Data Exploration & Feature Selection

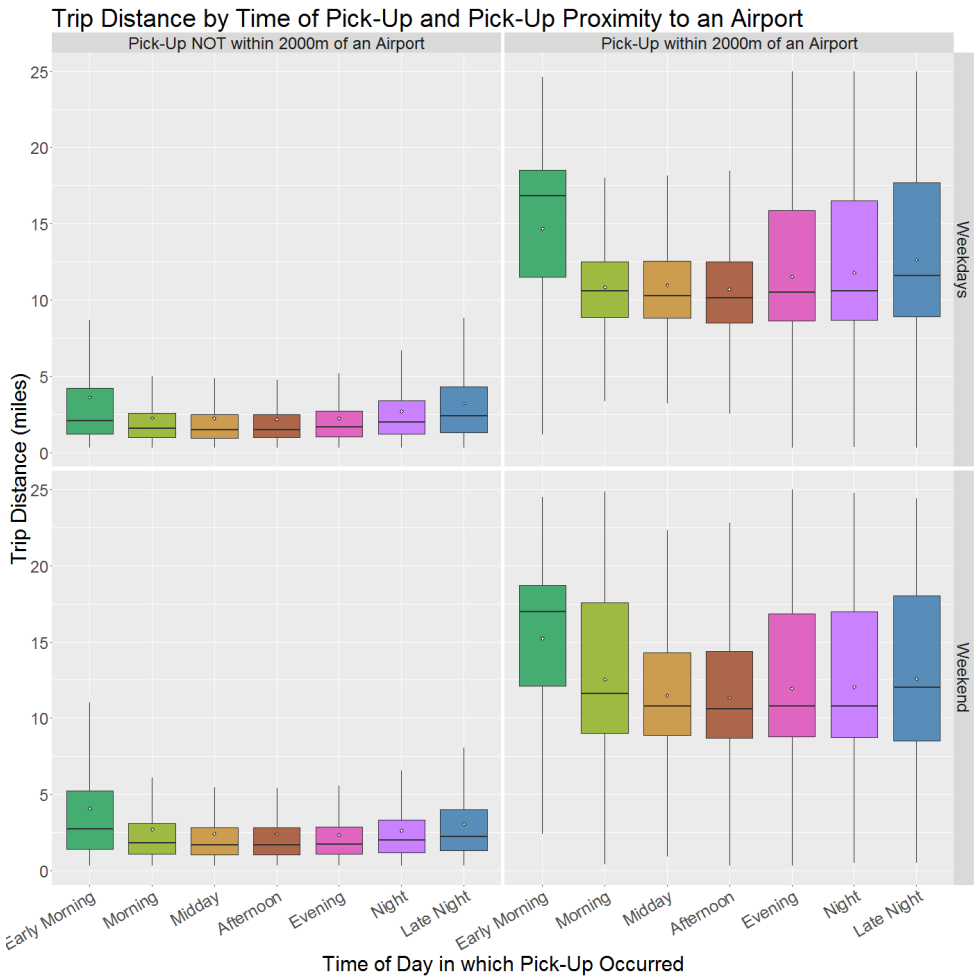
The first step in my data exploration was determining which of the over 400 binary features for the pick-up or drop-off within the different NTAs and non-NYC counties correlated with trip distance. I included for modeling features that had a magnitude of correlation coefficient of at least .02, and 62 of these features were ultimately included. Unsurprisingly, areas trip distance was positively correlated (sometimes very much so) with areas that contain an airport (such as NJ 13 and Queens 98).

Some of the features created from the American Community Survey data, such as the four features in the correlogram to the right, in particular surprised me in having a strong relationship (positive *or* negative) with trip distance. The magnitude of the correlation coefficients between trip distance and the features indicating the percentage of civilian employees living in an area, for example, are on part with the magnitude of the correlation coefficients between trip distance and being within the vicinity of an airport, one of the stronger relationships between a feature and distance seen in the data.



In one case, a reformating of the feature improved the correlation with trip distance, and subsequently improved (though slightly) the performance of the final model. Initially the performance of the features created from the ACS data indicating the number of occupied households with access to 0, 1, 2, or 3 or more vehicles was disappointing. After transforming these features from absolute volumes of households, though, into a percentage of the total occupied households for the respective area, the relationship between vehicle access and taxi trip distance became clearer. In the correlogram, below, in comparing the coefficients in the first row for "puArea_no.vehicles.availale", "puArea_X1.vehicle.available", "puArea_X2.vehicles.available", and "puArea_X3.or.more.vehicles.available" to their transformed, percentage-based versions ("puArea_perc_NoVehicles", "puArea_perc_OneVehicle", "puArea_perc_TwoVehicles", and "puArea_perc_ThreePlusVehicles"), we can see the drastic difference this simple transformation made.





In addition to the success of the features ACS (and zoning) data, the features I created from transportation hubs and the simple “time of day” and “is weekend” features I created from the original data were ultimately insightful. Looking at the plot to the right, which separates vertically pick-ups on weekdays (top) vs. weekends (bottom), separates horizontally pick-ups NOT within the vicinity of an airport (left) vs. those within it (right), and across each pane, the time of day block in which the pick-up occurred. Here, we can very clearly see that trip distances tend to be much longer and to vary much more when the pick-up

occurs within the vicinity of an airport. Also, while the pattern of trip distances across the day are roughly the same on both weekends and weekdays for those pick-ups that occur outside the vicinity of an airport, there is a very noticeable difference in trip distances between weekends and weekdays when the pick-ups are within the vicinity of an airport (especially during the morning block of time).

Model Selection

Because predicting a numeric value from a continuous scale required a regression model, I tested and compared linear regression, gradient boosted trees, bagging, and random forests. Ultimately the tree-based models had similar performance, but the random forest model had an edge over the others. Further, using the random forest model I experimented with different maximum tree depths and different subsets of the data. Ultimately, the final model used all 149 features, had a maximum, minimum, and mean tree depth of 20 (maximum tree depth was specified as 20), and "converged" at 48 trees using the stopping criterion that the 2-tree average MSE is within 0.001 of the prior 2 two-tree average. The model size was 35.574 MB, with 45,185-85,891 leaves (mean 59,043.707 leaves). The predictions using the test set of the data yielded MAE 0.372, RMSE 0.785, and R^2 0.933.

Effectiveness of the Features Created from External Data

I was particularly interested in the effectiveness of the 135 features that were ultimately added to the data from outside sources. To gauge how effective they were, I compared the performance of the final feature set in a random forest model with 20 trees and a maximum depth of 20 to the same model inputs using (a) only the features derived from the taxi trip and fare data sets, and (b) only the features using the outside data. While the performance of model using the additional data features only slightly lags behind that of the original data features only, it is within the ballpark in terms of model fit (which I found surprising). Also, this made clear that using all of these features together results in a model with better fit, so I was satisfied that the features I had created had, in fact, been effective.

Model	Number of Trees	MAE	RMSE	R^2
From Original Data only	20	0.563	0.909	0.911
Features From Outside Data Only	20	0.549	0.974	0.898
Features From Original and Outside Data	20	0.377	0.792	0.932
Features From Original and Outside Data	**48 (converged)**	**0.372**	**0.785**	**0.933**

I also compared the model to the performance metrics of using the training set's mean and, separately, median trip distances as the predicted value for all of the test observations, but found that these bars for success were particularly low (MAE 1.948, RMSE 3.044, R^2 0.000 and MAE 1.714, RMSE 3.204, R^2 0.108, respectively).