Predicting the Fare Amount of Taxi Pick-ups in New York City

Alireza Saberi, Jenna Wong, and Liu Yuguang
McGill University

Abstract— **The goal of this study was to predict the fare amount of a specific taxi pick-up in NYC. We used data from a random sample of 200,000 taxi trips that occurred during February and March 2013 in Manhattan, NYC and considered seven predictors that contained information about the location, weather, and time of the pick-up. We used two different linear regression methods (standard linear regression and ridge regression) to predict the outcome. To select the optimal parameters for each regression method, we performed 2-fold cross-validation on 50% of the data (training set) and evaluated the performance of the final models on the remaining 50% (testing set). Overall, we found that both regression methods performed similarly and could not predict the outcome very well. The root mean squared error (RMSE) of the predictions was approximately 7.5 in the testing set, which was quite large compared to the average taxi fare amount of $10.75. Although the NYC taxi dataset could not answer our prediction question very well, the information contained in the NYC taxi datasets could be useful for other future applications, especially when linked with additional data sources.**

To access the study dataset, go to: https://github.com/alireza-saberi/Applied_ML-COMP-598

I. Context and Problem Description

New York City (NYC) is the busiest and most populous city in the United States [1], and demand for taxis is high [2]. On average, NYC's yellow taxis provide 485,000 trips to 600,000 passengers daily [3]. All yellow taxis in NYC are licensed and regulated by the NYC Taxi and Limousine Commission (TLC), which also sets taxi fare rates [3].

In the past, taxi drivers owned their own licenses (medallions), which could provide a stable middle class income [2]. Nowadays, however, since the cost of owning a NYC taxi medallion is extremely expensive (possibly over $1 million), most drivers cannot afford to purchase their own medallions [2]. Instead, most medallions are owned by large fleet companies that rent out their vehicles to drivers. This shift has consequently made it more difficult for taxi drivers to earn a stable income since drivers are restricted to their rental hours and must pay rental fees [2]. Moreover, drivers' hourly fare revenues are affected by the fluctuating demand for taxis in NYC by day of week and time of day [3]. Thus, it could be of value to taxi drivers to be able to predict the fare amount of a specific taxi pick-up. This ability could help taxi drivers more strategically choose their pick-ups as one way of maximizing their fare revenues.

In 2008, the TLC introduced the Taxi Passenger Enhancement Program (TPEP), which allowed passengers to pay by credit card and also enabled the collection of detailed electronic trip data for all taxi trips, including information on pick-up and drop-off times and locations and fare amounts [3]. As described on his blog [4], data junkie Chris Whong was able to obtain these data from the TLC for all trips from January to December 2013, which he subsequently made publicly available [5]. Upon discovering these data, we set out to predict the fare amount (before tips and excluding tolls and taxes) of a specific taxi pick-up in NYC. We hypothesized that important predictors of fare amount could include the pick-up location, month, day of week, time of day, weather, number of passengers, and average fare in the previous month for the same community district and weekday.

II. Related Work

We searched the Internet for evidence of other similar datasets containing electronic trip data in the public and private domain. We found that in November 2013, the Massachusetts Institute of Technology (MIT) held a 'Big Data Challenge' to predict taxi demand (measured as the total number of taxi trips) in various locations of downtown Boston within a two-hour time window [6]. The purpose of the contest was to better understand public transportation patterns in the city [6]. To predict taxi demand, contestants were provided with historical data from over 2.3 million taxi rides (including the latitude and longitude co-ordinates of pick-ups and drop-offs) as well as other datasets containing information on weather, events, and social media data (e.g. twitter). However, as the contest ended in January 2014, these data are no longer available. Besides the NYC taxi data and the MIT Big Data Challenge dataset, we could not find any other publicly available datasets containing taxi trip information. However, similar datasets do exist in the private domain in cities where taxis are equipped with wireless credit card processing systems. For example, the city of Chicago recently used electronic trip data from over 10.6 million taxi trips to conduct a study of taxi fare rates [7]. One purpose of this study was to use taxi data to predict the annual revenue of, annual costs incurred by, and annual net income of taxi drivers, and explore how fare adjustments, policies and other external factors would consequently affect these outputs [7]. The MIT Big Data Challenge and the Chicago Taxi Fare Rate Study show how electronic taxi trip data can be used to answer interesting and different prediction questions. We hypothesize that the range of applications for electronic taxi trip data will likely increase in the future as more and more cities implement wireless credit technology, thus making these types of data more prevalent.

III. Dataset Creation and Description

The dataset for this project was created using the raw NYC taxi data from Chris Whong (5), as well as historical weather data and geographical data for NYC. We first linked the raw trip and fare datasets, and then used polygon-mapping software along with the geographical co-ordinates of the pick-up locations to determine the borough and community district where each pick-up occurred. We also determined the average temperature in NYC on each day of the prediction period and linked these temperatures to the study dataset by pick-up date. Due to the vast amount of data (more than 14 million trips per month) and the amount of computing time required to map pick-up locations to a specific borough and community district in NYC, we limited the prediction period to taxi trips in February and March 2013 (because taxi demand in NYC is typically highest in the spring months [3]) and took a random sample of 100,000 taxi trips per month. We also limited the predictions to pick-ups in the Manhattan borough only, since over 90% of all pick-ups occurred in Manhattan. Details of the dataset creation process are included in Appendix A.

The prediction outcome (Y) in this study was the fare amount of a taxi pick-up in dollars (continuous). Since we were primarily interested in predicting the minimum gross revenue that a taxi driver could expect from a pick-up, the fare amount excluded tips (since tipping is variable between passengers), and tolls and taxes (because the passenger pays these costs). We considered seven features (or inputs) to predict the outcome: month of the pick-up (binary), day of the week (categorical), community district in which the pick-up occurred (categorical), time of day (continuous, to the nearest half hour), number of passengers (continuous), average fare in the previous month in the same community district and on the same day as the index pick-up (continuous), and the average daily temperature on the pick-up day (continuous). Details of the analytical dataset are included in Appendix B.

IV. Methods

We used two linear regression methods to predict taxi fare amount as a function of the seven predictors: standard linear regression and ridge regression. In all models, we represented the binary and categorical predictors using a dummy encoding scheme (i.e. n-1 binary variables for a categorical predictor with n levels, where the reference level has a zero value for all n-1 variables) and solved for the least-squares solution by both using the closed-form equation and implementing gradient descent. For gradient descent, we first scaled the continuous predictors by subtracting the mean and dividing by the range to increase the convergence rate. To estimate the least-squares solution for standard regression, we used a Robbins-Monroe learning rate of $\alpha_k = 0.1/(k + 1)$ for iteration k and ended the descent when the absolute difference between the weight vector from iteration k and k+1 was less than 1.0 x 10^{-4}. For ridge regression, we used a Robbins-Monroe learning rate of $\alpha_k = 1/(k + 250)$ and ended the descent when the absolute difference between the weight vector from iteration k and k+1 was less than 4.4 x 10^{-4}. To evaluate the models, we calculated the root mean squared error (RMSE). We chose to use the RMSE as the loss function instead of the 'raw' sum of the squared errors (SSE) because the RMSE was a scaled measure that allowed us to compare the performance of models that used different training and validation set sizes. Also, we felt that the RMSE was more meaningful since the large size of our dataset consequently produced SSEs that were extremely large.

For each regression method, we explored increasing degrees of model complexity by adding polynomial terms to the continuous predictors, varying the order-d of the models from d=1 to 4 (e.g. $x + x^2$ for d=2). We did not add polynomial terms to the categorical predictors since raising a binary variable to any power d would return the same value (and thus create redundant columns). To guard against over-fitting, we split the dataset into a 50% training set and a 50% testing set. We used 2-fold cross-validation in the training set to select the optimal "hyper-parameters" for each regression method (e.g. model complexity, lambda). We performed cross-validation by dividing the training set into two equal parts and repeating the following process twice: 1) using one part to estimate the weight vector (i.e. least-squares solution), 2) applying the weight vector in the entire training set, and 3) calculating the RMSE separately for the two parts, where $\text{RMSE}_{\text{valid}}$ was the error in the held-out subset and $\text{RMSE}_{\text{train}}$ was the error in the subset used to estimate the weight vector. The training error was then calculated as the average of $\text{RMSE}_{\text{train}}$ from the two rounds, and the estimated true prediction error was calculated as the average of $\text{RMSE}_{\text{valid}}$ from the two rounds. For the ridge regression analysis, we used the scaled continuous inputs and conducted 2-fold cross-validation to test 11 different values for the shrinkage factor, lambda, where λ = {0.00001 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000}. We selected the optimal degree of model complexity and the optimal shrinkage factor by identifying the value of d and λ that produced the lowest RMSE in the validation set, respectively.

We tested the performance of the final model parameters by fitting the least-squares solution for the final regression models using the training set and then applying the weights in the testing set. To show how the regression error changes with increasing amounts of training data, we varied the percentage of the training set used to estimate the weights from 10% to 100% of the entire training set.

V. Results

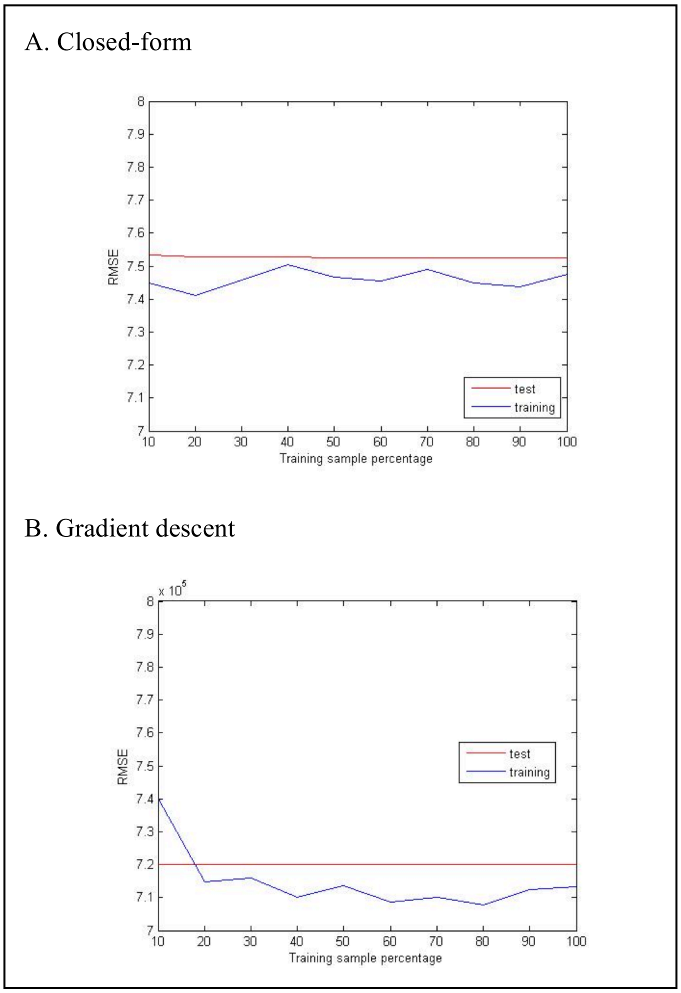
The final dataset contained a total of 185,472 taxi trips. The average fare amount of a taxi trip in Manhattan during February and March 2013 was $10.75 in the training set and $10.77 in the testing set. In both the training and testing set, the median fare amount was $8.50 (interquartile range $6.50 to $12.50). The total number of taxi pick-ups was fairly stable in each month, but was highest on Fridays and Saturdays and in the evenings after approximately 6pm.

A. *Standard linear regression*

Table 1 shows the RMSE in the training and validation set for a standard linear regression model when we increased the degree of model complexity from order-1 to order-4. The RMSE values we obtained using the closed-form equation and gradient descent were fairly similar and both indicated that increasing the order-d of the model slightly but continually decreased the RMSE in the training and validation set. Thus, we did not observe any evidence of over-fitting up to d=4 and concluded that d=4 was the best degree of model complexity out of the four orders considered.

TABLE I. ROOT MEAN SQUARED ERROR (RMSE) FROM STANDARD LINEAR REGRESSION FOR INCREASING LEVELS OF MODEL COMPLEXITY IN THE TRAINING AND VALIDATION SET

Model Complexity (Order-*d*)	$\text{RMSE}_{\text{Training}}$	$\text{RMSE}_{\text{Validation}}$
Closed-form equation		
1	7.4804	7.4842
2	7.4770	7.4816
3	7.4737	7.4804
4	7.4723	7.4798
Gradient descent		
1	7.5147	7.5203
2	7.4994	7.5038
3	7.4862	7.4922
4	7.4856	7.4902


A. Closed-form

B. Gradient descent

Figure 1. Root Mean Squared Error (RMSE) for an Order-4 Standard Linear Regression Model When Using Increasing Amount of Training Data (*Panels A and B show the training and testing error when using the closed-form equation and gradient descent to solve for the least-squares solution, respectively).*

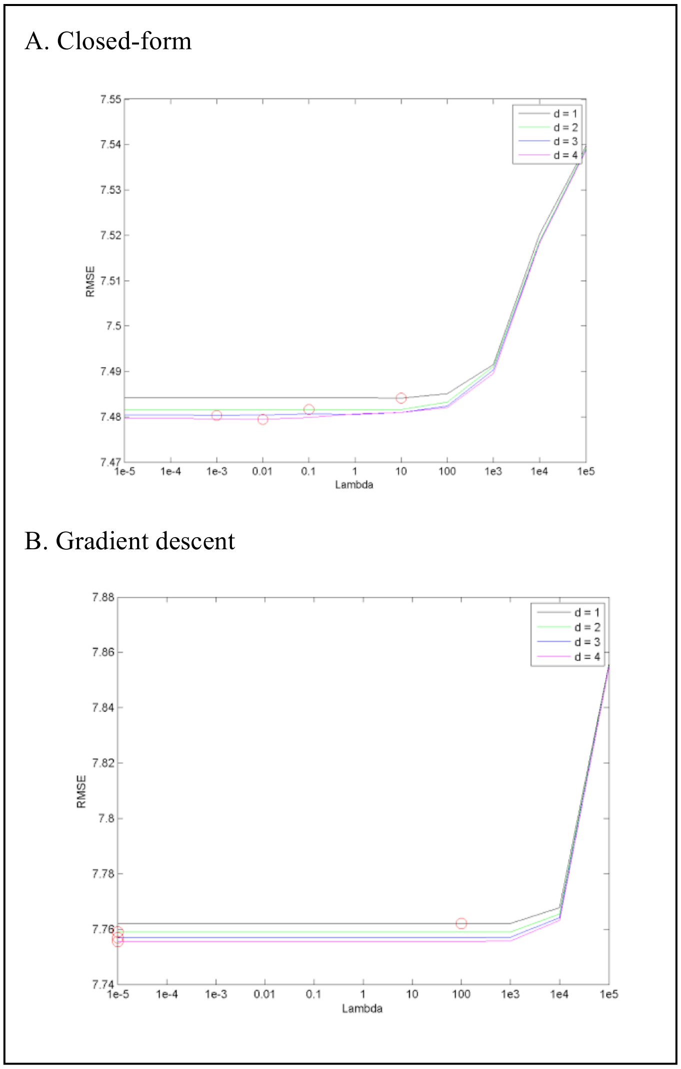
Our final choice of model complexity for the standard linear regression model was order-4 (i.e. all continuous predictors were modeled using polynomial terms up to the 4th degree). Figure 1 shows the RMSE of this model in the training and testing set when increasing amounts of the training data were used to estimate the model parameters. We expected the RMSE to decrease smoothly in the training and testing set with increasing amounts of training data, with the RMSE decreasing more drastically in the training set. Contrary to our expectations, we found that the RMSE decreased minimally in the RMSE in the testing set and was quite unstable in the training set. Overall, the magnitude of the RMSE in the testing set indicated that the predictive performance of the model was not very good. The RMSE, which can be interpreted as the average absolute difference between the predicted and actual fare amount, was approximately 7.5 in the testing set. This is quite large considering that the average taxi fare amount in the testing set was approximately $10.75.

B. Ridge Regression

For each degree *d* of model complexity, we used ridge regression to estimate the weight vector for the predictors, varying the shrinkage or penalty parameter, λ, logarithmically from 0.00001 to 100000. Figure 2 shows the change in RMSE in the validation set for increasing values of λ within each degree of model complexity when using the closed-form equation and gradient descent, respectively, to solve the least-squares equation. Both methods showed that the RMSE decreased slightly at very small values of λ, but after a certain threshold, the shrinkage penalty was too drastic and the RMSE increased. At all degrees of model complexity, we could not reproduce the optimal value of λ (i.e. the value that produced the lowest RMSE in the validation set) from the closed-form equation using gradient descent (Table 2). However, the general trends we observed using gradient descent were similar. We found that λ was largest for the least complex model (order-1) and smaller for the more complex models (order-2 to order-4), indicating the need for a larger amount of shrinkage with a simpler model. When each degree of model complexity was modeled at its optimal λ, an order-4 model had the lowest RMSE in the validation set; thus, we concluded that for ridge regression, λ = 0.01 was the optimal shrinkage factor and *d*=4 was the optimal level of model complexity (we chose to use the optimal λ obtained using the closed-form equation).

TABLE II. ROOT MEAN SQUARED ERROR (RMSE) FROM RIDGE REGRESSION IN THE TRAINING AND VALIDATION SET AT THE OPTIMAL SHRINKAGE FACTOR (λ) WITHIN EACH DEGREE OF MODEL COMPLEXITY

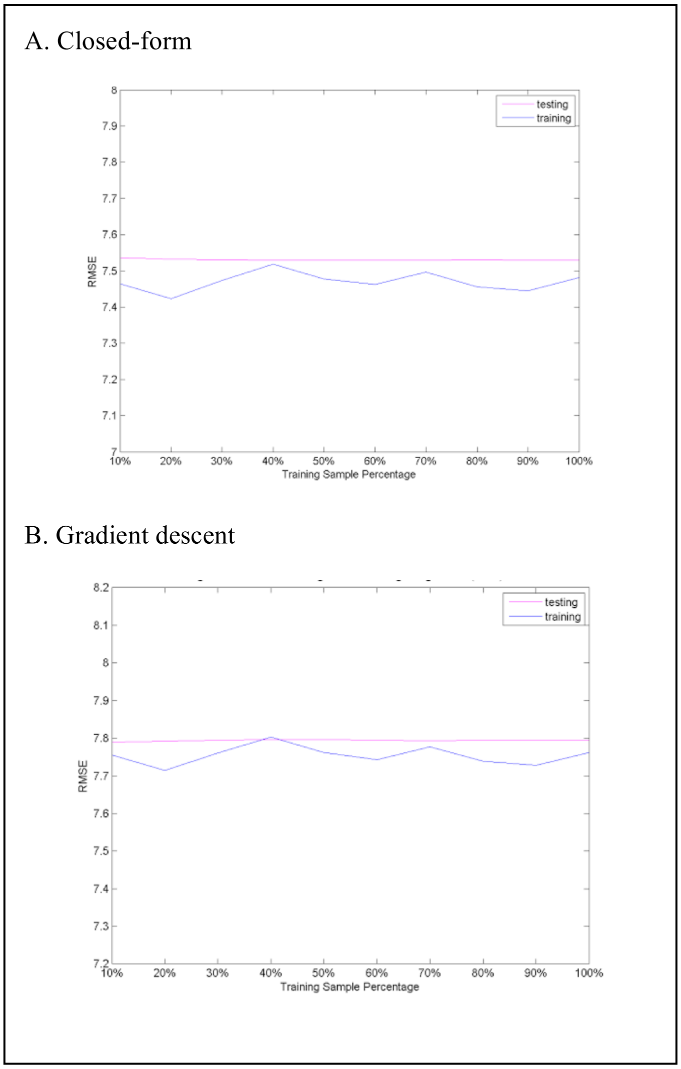
Model Complexity (Order-*d*)	Optimal λ	$\text{RMSE}_{\text{Training}}$	$\text{RMSE}_{\text{Validation}}$
Closed-form equation			
1	10	7.481	7.484
2	0.01	7.478	7.481
3	0.001	7.475	7.480
4	0.01	7.474	7.479
Gradient descent			
1	100	7.7617	7.7620
2	0.00001	7.7586	7.7589
3	0.00001	7.7566	7.7569
4	0.00001	7.7553	7.7556


A. Closed-form

B. Gradient descent

Figure 2. Root Mean Squared Error (RMSE) in the Validation Set from Ridge Regression for Increasing Amounts of Shrinkage (Lambda) Within Each Degree (*d*) of Model Complexity (*Panels A and B show the regression error when using the closed-form equation and gradient descent to solve for the least-squares solution, respectively. The circles indicate the lambda value that produced the lowest RMSE in the validation set).*

Figure 3 shows the change in RMSE for an order-4 ridge linear regression model with λ = 0.01 using increasing amounts of training data. Again, the results were contrary to our expectations. The RMSE remained fairly stable in the testing set with increasing amounts of training data, while the RMSE was unstable in the training set. Overall, the magnitude of the RMSE was similar to that observed with standard linear regression, indicating that the models had fairly poor predictive ability.


A. Closed-form

B. Gradient descent

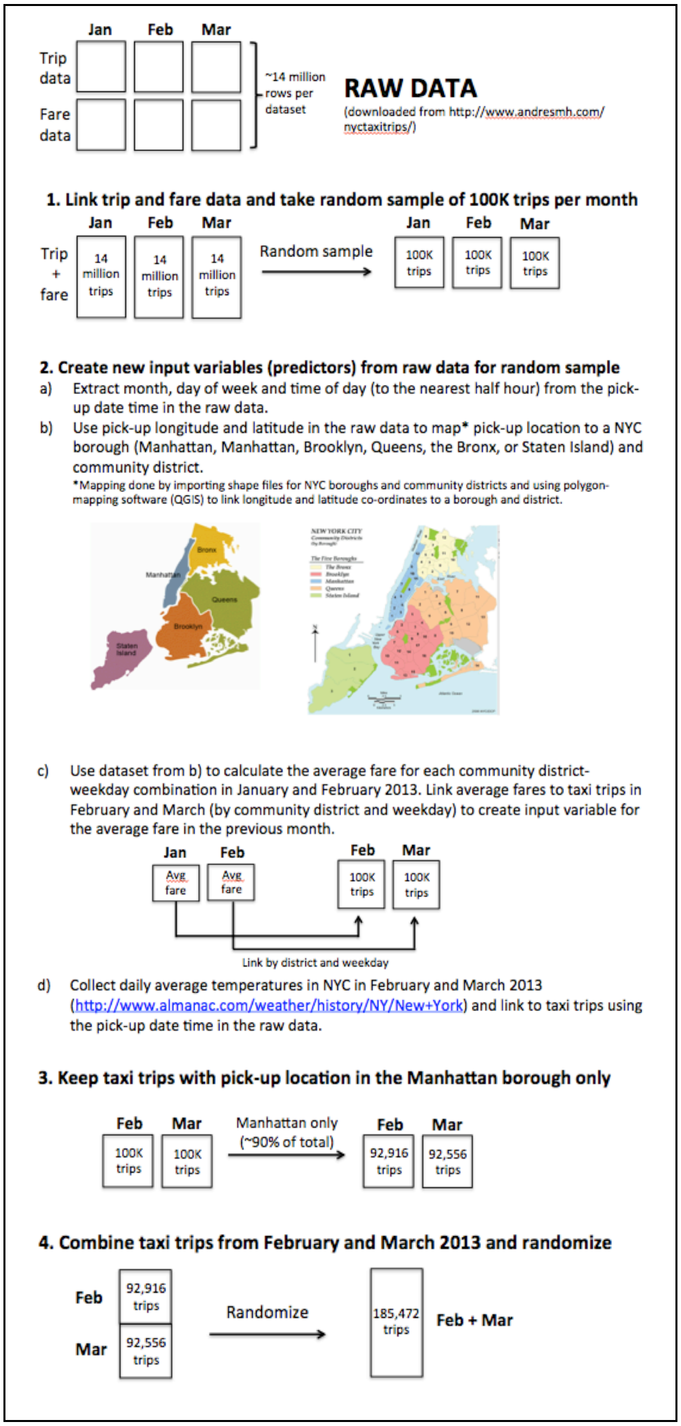
Figure 3. Root Mean Squared Error (RMSE) for an Order-4 Ridge Regression Model Using λ = 0.01 and Increasing Amounts of Training Data (*Panels A and B show the training and testing error when using the closed-form equation and gradient descent to solve for the least-squares solution, respectively).*

VI. Discussion

In this study, we used data from a large random sample of almost 200,000 NYC taxi trips to predict the fare amount of a taxi pick-up in Manhattan between the months of February and March. We considered seven predictors (month, day of week, community district, time of day, number of passengers, average fare in the previous month in the same district and on the same day, and average temperature on the pick-up day) and used two different linear regression methods (standard linear regression and ridge regression) to predict the outcome. In the end, both types of regression methods performed similarly and did not predict the outcome very well. This was evidenced by a fairly large RMSE (approximately 7.5 for both regression methods) when compared to the average taxi fare of $10.75. One reason that the regression models did not predict the outcome well is that are that a linear assumption (i.e. the assumption that the predictors are linearly associated with the outcome) is a poor approximation for this prediction question. Another reason is that the inputs we considered were not informative enough to answer the prediction question. We hypothesize that our results were due more to the latter, as our inputs may have been too ‘high-level’ and additional information about the characteristics of the passenger (e.g. gender, age) would help improve the predictions.

Although the NYC taxi dataset could not answer our prediction question very well, we feel that this dataset contains valuable, high quality information that could be used to better answer other types of prediction question. For example, the fact that the dataset contains complete information about the exact time and geographical co-ordinates of all taxi pick-ups and drop-offs in NYC could be used to predict taxi demand in various areas of NYC at various times. This information could be used to identify underserviced areas of NYC that could benefit from more taxis or more alternative modes of transportation. When linked with additional data sources such as datasets on social media and local events, the NYC taxi dataset could also used to predict how social media and local events impact transportation patterns and taxi demand in NYC.

Appendix A – Dataset Creation Process



(downloaded from http://www.andresmh.com/nyctaxitrips/)

1. Link trip and fare data and take random sample of 100K trips per month

2. Create new input variables (predictors) from raw data for random sample

a) Extract month, day of week and time of day (to the nearest half hour) from the pick-up date time in the raw data.

b) Use pick-up longitude and latitude in the raw data to map* pick-up location to a NYC borough (Manhattan, Manhattan, Brooklyn, Queens, the Bronx, or Staten Island) and community district.

*Mapping done by importing shape files for NYC boroughs and community districts and using polygon-mapping software (QGIS) to link longitude and latitude co-ordinates to a borough and district.

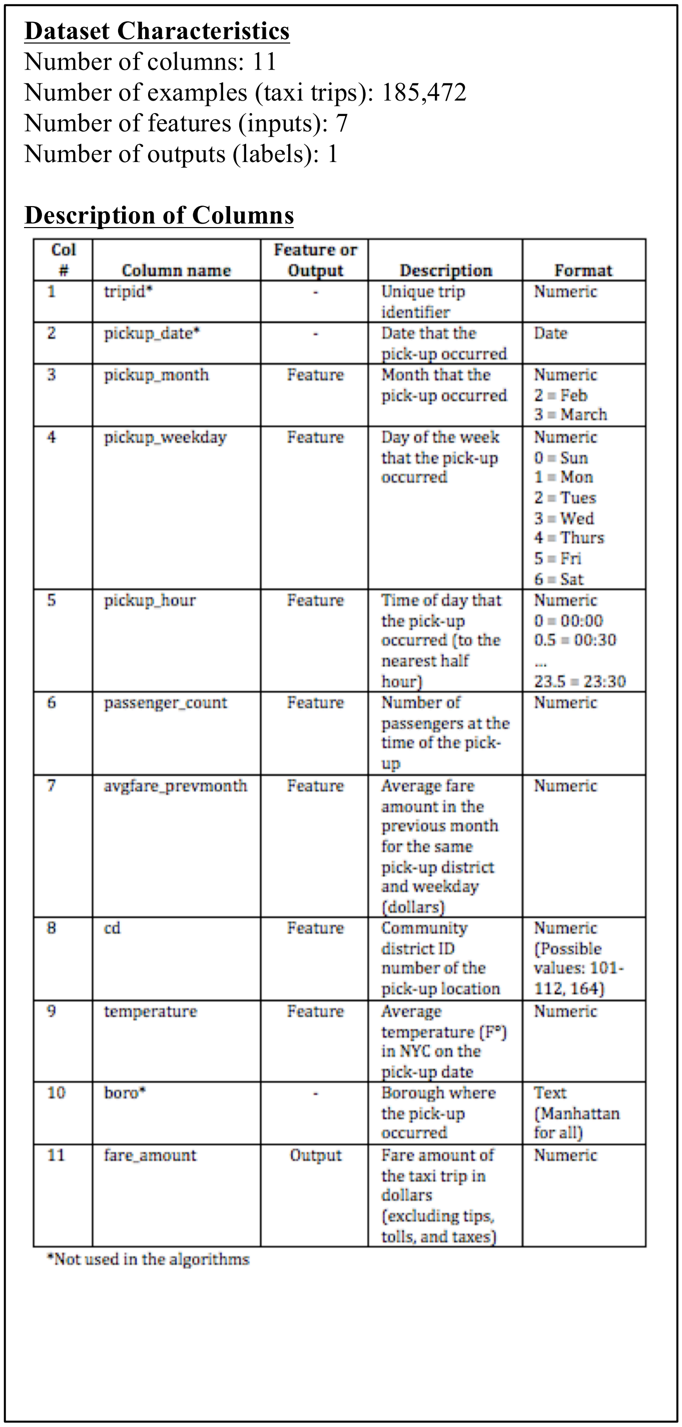
c) Use dataset from b) to calculate the average fare for each community district-weekday combination in January and February 2013. Link average fares to taxi trips in February and March (by community district and weekday) to create input variable for the average fare in the previous month.

d) Collect daily average temperatures in NYC in February and March 2013 (http://www.almanac.com/weather/history/NY/New+York) and link to taxi trips using the pick-up date time in the raw data.

3. Keep taxi trips with pick-up location in the Manhattan borough only

4. Combine taxi trips from February and March 2013 and randomize

APPENDIX B – DATA DICTIONARY



Dataset Characteristics

Number of columns: 11
Number of examples (taxi trips): 185,472
Number of features (inputs): 7
Number of outputs (labels): 1

Description of Columns

Col #	Column name	Feature or Output	Description	Format
1	tripid*	-	Unique trip identifier	Numeric
2	pickup_date*	-	Date that the pick-up occurred	Date
3	pickup_month	Feature	Month that the pick-up occurred	Numeric 2 = Feb 3 = March
4	pickup_weekday	Feature	Day of the week that the pick-up occurred	Numeric 0 = Sun 1 = Mon 2 = Tues 3 = Wed 4 = Thurs 5 = Fri 6 = Sat
5	pickup_hour	Feature	Time of day that the pick-up occurred (to the nearest half hour)	Numeric 0 = 00:00 0.5 = 00:30 ... 23.5 = 23:30
6	passenger_count	Feature	Number of passengers at the time of the pick-up	Numeric
7	avgfare_prevmonth	Feature	Average fare amount in the previous month for the same pick-up district and weekday (dollars)	Numeric
8	cd	Feature	Community district ID number of the pick-up location	Numeric (Possible values: 101-112, 164)
9	temperature	Feature	Average temperature (F°) in NYC on the pick-up date	Numeric
10	boro*	-	Borough where the pick-up occurred	Text (Manhattan for all)
11	fare_amount	Output	Fare amount of the taxi trip in dollars (excluding tips, tolls, and taxes)	Numeric

*Not used in the algorithms

We hereby state that all the work presented in this report is that of the authors.

REFERENCES

[1] U.S. Census Bureau, U.S. Department of Commerce. (2011). "Population distribution and change: 2000 to 2010." P. Mackun and S. Wilson.

[2] Guiffo, John. "NYC's new green taxis: what you should know." *Forbes.com.* Published 30 Sept 2013. Web. Accessed 19 Sept 2014. <http://www.forbes.com/sites/johngiuffo/2013/09/30/nycs-new-green-taxis-what-you-should-know/>

[3] NYC Taxi and Limousine Commission. (2014). "2014 taxicab factbook."

[4] Whong, Chris. "FOILing NYC's taxi trip data." *Chris Whong.* Published 18 Mar 2014. Web. Accessed 13 Sept 2014. <http://chriswhong.com/open-data/foil_nyc_taxi/>

[5] Monroy-Hernandez, Andres. "NYC taxi trips data from 2013." GitHub Pages. Accessed 14 Sept 2014. <http://www.andresmh.com/nyctaxitrips/>

[6] BigData@CSAIL."MIT Big Data Challenge." Accessed 20 Sept 2014. <http://bigdatachallenge.csail.mit.edu>

[7] City of Chicago, Business Affairs and Consumer Protection. (2014). "Taxi fare rate study." Nelson\Nygaard Consulting Associates Inc.