CAPSTONE PROJECT: NEW YORK CITY TAXI TRIP AND FARE DATA ANALYSIS

Team 15: Gagan Kohli, Ranjani Kamath, Ravindra Kishore
Course: DATASCI 450

Introduction

New York City's yellow taxi cabs are literally everywhere in busy Manhattan, cruising the streets looking for fares. They are usually easy to hail on the street or from a cabstand in front of major hotels. Though they can be costly to ride, taxis are usually the most convenient method of traveling throughout Midtown.

The taxicabs of New York City are widely recognized icons of the city and come in two varieties: yellow and green. Taxis painted canary yellow (medallion taxis) are able to pick up passengers anywhere in the five boroughs (Manhattan, the Bronx, Queens, Brooklyn, and Staten Island). Taxicabs are operated by private companies and licensed by the New York City Taxi and Limousine Commission (TLC). It also oversees over 40,000 other for-hire vehicles, including "black cars", commuter vans and ambulettes. Taxicab vehicles, each of which must have a medallion to operate, are driven an average of 180 miles per shift.

We obtained two datasets for yellow taxi trips taken in NYC for the entire year of 2013. Our goal is to find interesting information from this data stash.

Team member credits:

Gagan Kohli: Owned the weather data extraction, cleaning & transforming for New York city for year 2013. He merged NYC_Taxi_Trip, NYC_Taxi_Fare with weather data. He also introduced extra attributes to dataset (is_rain, is_night, is_weekend, is_holiday etc.) to make feature set rich and help in feature engineering done by Ravindra and Ranjani.

Ranjani Kamath: Took the lead on data exploration, platform selection for running the analysis (R studio, Jupyter notebook, Azure ML Studio, etc.), fine tuning and running of all models, and final report creation

Ravindra Kishore: Took the lead on forming the teams and helping coordinate meetings with other teams to help gain better insights and ideas on the project. He helped come up with the initial problem definition. He also helped with feature and model selection and execution for predictive analysis.

Merging of datasets and exploratory analyses were run on Jupyter notebook. Machine learning models for predictive analysis were run on Azure Machine Learning Studio.

Team members met on an average of twice a week in person or via teleconference/ google hangouts to discuss the progress of the project and to plan the next steps.

Problem Definition:

Tipping behavior: 1. Predict whether a tip will be paid or not. 2. Can we predict the amount of tip paid?

Travel behavior: Does weather have any impact on distance travelled or tips paid? Can we predict the trip duration?

About the Dataset:

This data was originally obtained by Chris Whong. He obtained a year's worth of NYC taxi fare data (Jan to Dec 2013), from the New York City TLC by Freedom Of Information Act (FOIA) request. There are two datasets for NYC Taxi Data 2013: trip data and fare data. Each individual trip can be uniquely identified by columns medallion, hack_license, and pickup_datetime. A detailed explanation of some of the columns in the data set can be found here. The size of the original datasets in year 2013 (all 12 months) is about 50 GB in total, with over 173 million rows.

In order to minimize the needs on high performance computing infrastructure such as computers with big memory or Hadoop cluster for this project, we sampled 1% out of the entire datasets. When we sampled the datasets, we made sure that each row in the trip dataset has a one-to-one mapping in the fare dataset,

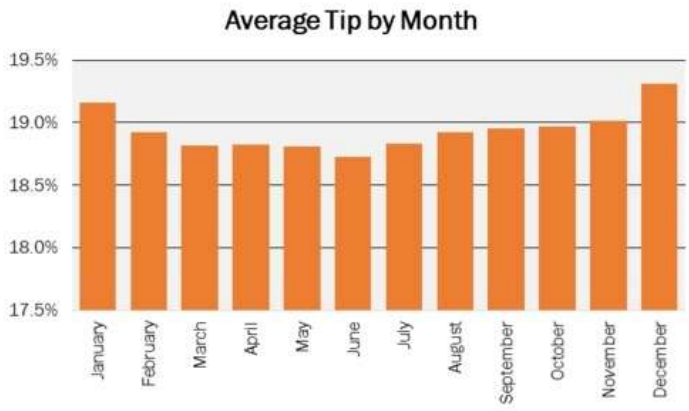
for each unique combination of the three key columns: medallion, hack_license, and pickup_datetime. The smaller datasets have 1703987 rows each.

A third dataset consisting of weather information for NYC in 2013 was obtained. Source of weather dataset for NYC: National Climatic Data Center, U.S. Department of commerce. Additional attributes like day of week (Mon, Tue, etc.), month names, day of the year (1 to 365), Holiday or not, Night or not, season (spring, summer, etc.) weather of the day (Cloudy, Rainy, Foggy, etc.) were added in order to better understand the riding behavior.

Some early insights:

Minimum tip paid: 0%, maximum tip paid: 99.88%. We also found some negative values in fare amount. 918680 rides (54%) were paid by credit card. 778188 rides (46%) were paid by cash. Rides paid for by card were more likely to be tipped than those paid for by cash. Out of the 778188 rides paid for by cash, only 75 (0.0001%) paid a tip, whereas, out of the 918680 rides paid for by credit card, 890698 (96.95%) paid a tip. Most common tip amount was 20%. We also did not find any relationship between number of taxi trips and average temperature during the day

The average taxi fare in 2013 was $14.74, with a standard deviation of $11.94. About 43% of all taxi fares were $10 or less, and nearly 83% are less than $20. 2.89% of fares were greater than $50 and 0.07% were greater than $100. The most expensive trip recorded in this sample was a whopping $500, paid in cash. We found some super short trips where minimum total fare of $3.00 was paid (0.14%). There were few cases where there were no passengers or more than 6 passengers. Some 126 taxis logged more than 2 hours per trip. There were 3 cases where the trip distance was more than 500 miles! Also, the average speed per trip was found to be 13.02mph.

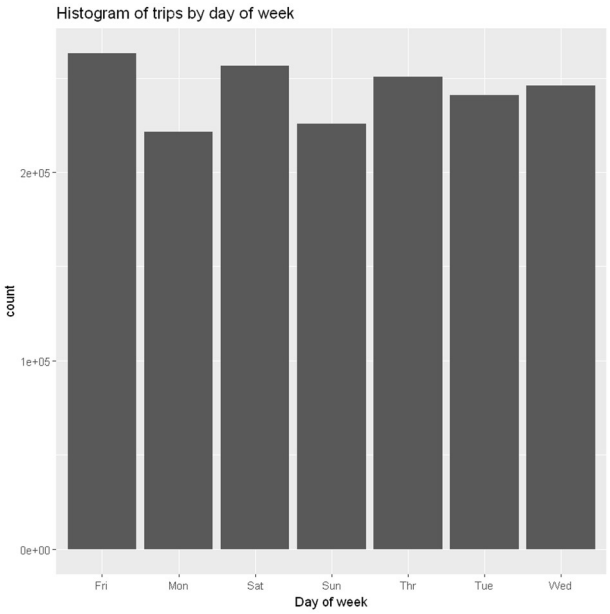


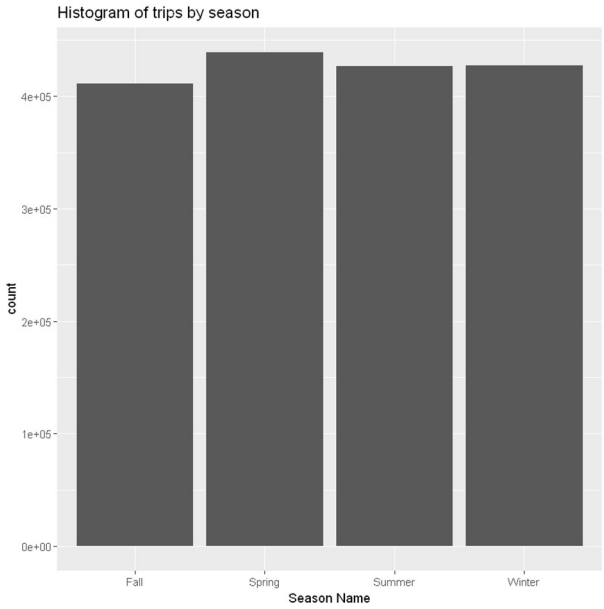
People seemed more likely to tip most generously during holidays (December) than any other time (chart above)

We found two vendors, CMT & VTS serving the NYC area. This data subset covered 11836 unique taxis (identified by medallion numbers). The average tip was $1.36 with a standard deviation of $2.17 for all rides. For trips paid for by credit card, the average tip was $2.52 with a standard deviation of $2.40. As a fraction of the total fare, the average tip was 8.1% for all rides, and 15.01% for all rides that were paid for by credit card. 47.37% of the total riders did not tip at all, most being cash paying riders.

We wanted to know if time of day, weather of the day, time of the year, etc., affects riding and/or tipping behavior. From our exploratory analysis, we did not notice any strong relationships between weather (temperature, visibility, fog, etc,) and riding behavior like trip distance or tips paid.

However, from our initial analysis, we found that more trips are taken on Friday than any other day of the week, implying people tend to travel more around weekends. Also people seemed to hail more rides in spring, with March having the highest number of trips among all months. Below is a graphical visualization of the same:



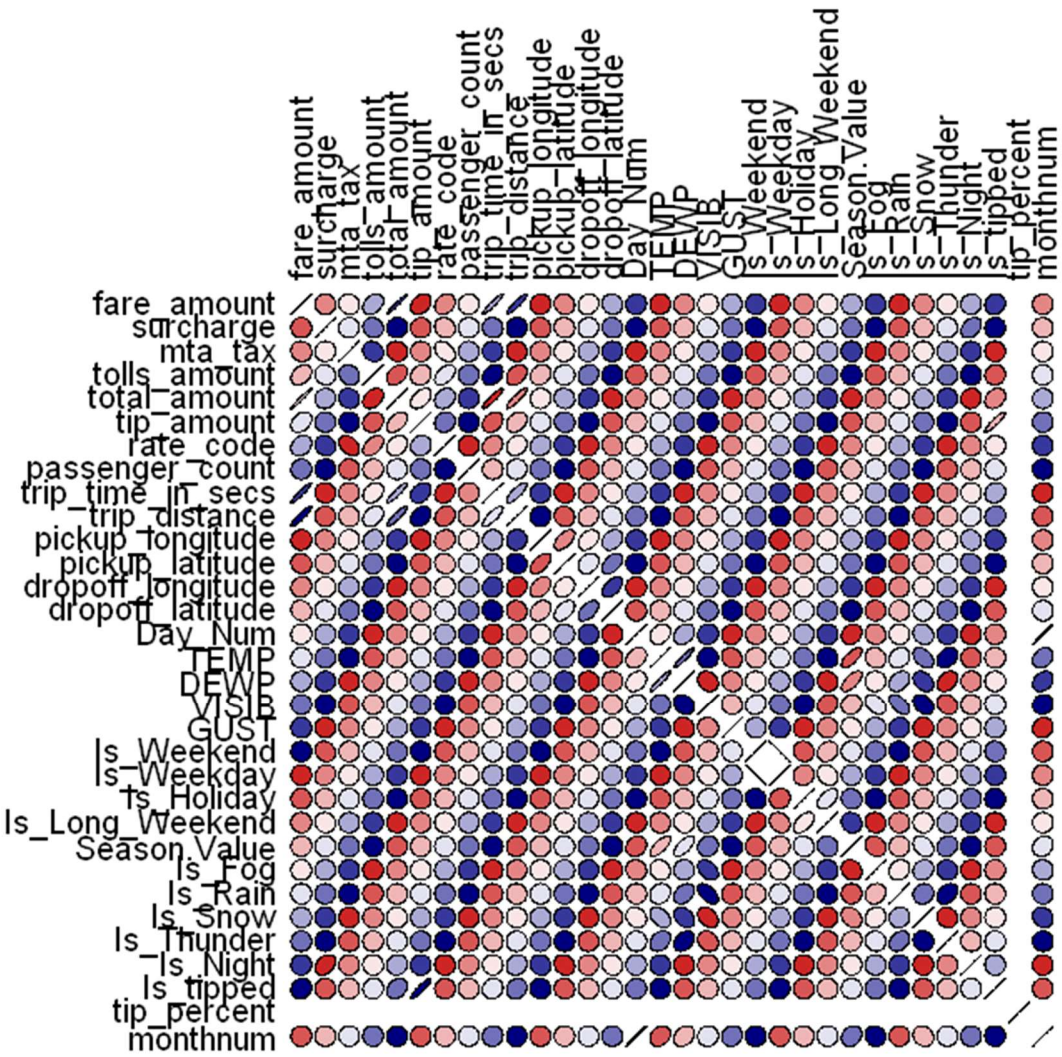


Feature Engineering and Feature selection

Spearman Correlation between numeric columns was calculated to help understand underlying relationships between attributes (table below):

	fare_amount	surcharge	mta_tax	tolls_amount	total_amount	tip_amount	rate_code	passenger_count	trip_time_in_sec	trip_distance	TEMP	VISIB	Is_Weekend	Is_Holiday	Season.Value
fare_amount	1.000	-0.017	-0.088	0.340	0.978	0.323	0.241	0.025	**0.945**	**0.915**	0.031	0.006	-0.005	-0.008	0.026
surcharge	-0.017	1.000	0.039	-0.083	0.058	0.025	-0.129	0.020	-0.045	0.035	0.015	-0.017	-0.154	-0.045	0.013
mta_tax	-0.088	0.039	1.000	-0.152	-0.087	-0.034	-0.377	0.002	-0.029	-0.038	-0.004	-0.001	-0.001	-0.003	-0.003
tolls_amount	0.340	-0.083	-0.152	1.000	0.349	0.158	**0.477**	0.019	0.297	0.330	0.012	0.001	-0.020	0.003	0.012
total_amount	0.978	0.058	-0.087	0.349	1.000	0.468	0.240	0.019	**0.921**	**0.900**	0.031	0.003	-0.025	-0.014	0.027
tip_amount	0.323	0.025	-0.034	0.158	**0.468**	1.000	0.090	-0.029	0.297	0.306	0.007	-0.008	-0.041	-0.017	0.009
rate_code	0.241	-0.129	-0.377	0.477	0.240	0.090	1.000	0.011	0.191	0.199	0.004	0.002	-0.001	0.006	0.004
passenger_count	0.025	0.020	0.002	0.019	0.019	-0.029	0.011	1.000	0.028	0.031	0.003	0.012	0.056	0.009	0.002
trip_time_in_secs	0.945	-0.045	-0.029	0.297	0.921	0.297	0.191	0.028	1.000	**0.812**	0.032	-0.005	-0.046	-0.027	0.027
trip_distance	0.915	0.035	-0.038	0.330	0.900	0.306	0.199	0.031	0.812	1.000	0.023	0.014	0.040	0.009	0.021
TEMP	0.031	0.015	-0.004	0.012	0.031	0.007	0.004	0.003	0.032	0.023	1.000	0.004	-0.019	-0.066	**0.717**
VISIB	0.006	-0.017	-0.001	0.001	0.003	-0.008	0.002	0.012	-0.005	0.014	0.004	1.000	0.104	0.082	0.115
Is_Weekend	-0.005	-0.154	-0.001	-0.020	-0.025	-0.041	-0.001	0.056	-0.046	0.040	-0.019	0.104	1.000	-0.086	-0.016
Is_Holiday	-0.008	-0.045	-0.003	0.003	-0.014	-0.017	0.006	0.009	-0.027	0.009	-0.066	0.082	-0.086	1.000	-0.081
Season.Value	0.026	0.013	-0.003	0.012	0.027	0.009	0.004	0.002	0.027	0.021	0.717	0.115	-0.016	-0.081	1.000

The following is a visual representation of the above correlations:



Pearson correlations were initially calculated. These did not seem to explain the relationships much. We noticed that tip amount is not linearly correlated with any variables except Fare Amount and Total Amount. So we calculated the mutual information in addition to Spearman correlation. The mutual information seems to explain some relationship between tip amount and total trip time & distance as seen below:

Mutual information between tip amount and total trip time =0.316
Pearson Correlation between tip amount and total trip time =0.0134
Spearman Correlation between tip amount and total trip time =0.297

Mutual information between tip amount and total trip distance=0.3411
Pearson Correlation between tip amount and total trip distance=-0.00017
Spearman Correlation between tip amount and total trip distance=-0.306

So from this we can safely say that trip distance or trip duration can be used to model tipping behavior. Also, from the above table, we can see a weak correlation between trip duration (trip_time_in_secs) and rate code. So this could be one of the features used to model trip duration.

Feature and model selection:

Azure Machine Learning studio was used to select features (columns) to be used for model building. Filter based feature selection using Spearman correlation or Fisher score method gave us the best set of features for our models.

1. Building a model to predict time taken for each trip: Target column - trip_time_in_secs.

For this, several models like linear regression and neural network were tried with different parameter settings and several feature combinations. We did not find any feature combinations or any models to be able to accurately predict the time taken on any trip. So we log transformed the target column [lntrip_duration = log(trip_time_in_secs)]. We also added another new feature called speed {trip_distance/(trip_time_in_secs/3600)} which is the speed in MPH.

New target column: lntrip_duration

We selected features using Fisher Score Method for this regression problem, which gave us the best set of features for prediction. An 80/20 split was used between training and test data and numeric columns were Zscore normalized

Final features: ***lntrip_duration ~ trip_distance + passenger_count + rate_code + speed + surcharge + ls_Night + monthnum***

These 7 features helped predict the trip duration with coefficient of determination = 0.98 and least amount of errors with a Neural Network model.

Below is a comparison of 3 different models for the above features. Default set of model parameters were used in all 3 cases:

	Bayesian LR	Linear Regr	Neural Net
Negative Log Likelihood	530655.70664		
Mean Absolute Error	0.45652	0.45652	0.03419
Root Mean Squared Error	0.59910	0.59910	0.11597
Relative Absolute Error	0.57636	0.57636	0.04317
Relative Squared Error	0.35892	0.57636	0.01345
Coefficient of Determination	0.64108	0.64108	**0.98655**

2. Building a model to predict if a tip was paid or not: Target column (label) – ls_tipped

For this exercise, the following 3 features helped predict the outcome with nearly 95% accuracy for two-class Decision forest model - total_amount, fare_amount and trip_distance. This is the best accuracy achieved for the least amount of features. Tenfold cross validation was used to train and test the models. When we removed fare_amount, the accuracy went down to 80%. With 6 or more features, the accuracy went up to 99%, but we believe there might be some over-parametrization and/or overfitting involved in this case.

Final features: ***ls_tipped ~ total_amount + fare_amount + trip_distance***

Below is a comparison of 5 different models for the above features. Default set of model parameters were used in all 5 cases:

	Averaged Perceptron	Bayes Point Machine	Boosted DT	Decision Forest	SVM
Accuracy	0.885	0.916	0.944	**0.950**	0.636
Precision	0.890	0.938	0.976	0.977	0.663
Recall	0.892	0.899	0.916	0.927	0.619
F1 Score	0.891	0.918	0.945	0.951	0.640
AUC	0.946	0.962	0.990	0.992	0.695

The best model to predict if a tip will be paid is the two-class Decision Forest model, followed by Boosted Decision Tree model. It has the best accuracy & AUC at threshold = 0.5

3. Building a model to predict how much tip might be paid: Target column - tip_amount

An 80/20 split was used between training and test data and numeric columns were Zscore normalized. For this case we found that the columns total_amount, fare_amount, trip_distance could predict the tip amount with 0.85 Rsq coefficient of determination for Bayesian Linear Regression. Adding two more features trip_time_in_secs and tolls_amount increased the coefficient of determination of the same model to 0.97. Adding more features resulted in a slight decrease in MEA, RMSE and relative squared errors, but the coefficient of determination did not increase significantly.

Final features: ***tip_amount ~ total_amount + fare_amount + trip_distance + trip_time_in_secs + tolls_amount***

Below is a comparison of 5 different models for the above features. Default set of model parameters were used in all 5 cases:

	Bayesian LR	Boosted DT	Decision Forest	Linear Regr	Neural Net
Negative Log Likelihood	-34900.53687		28985274602		
Mean Absolute Error	0.141948	0.133067	0.235664	0.141955	0.129194
Root Mean Squared Error	0.161214	0.287761	0.497155	0.161218	0.153479
Relative Absolute Error	0.224458	0.210415	0.372647	0.224468	0.20429
Relative Squared Error	0.02599	0.082806	0.247163	0.025991	0.023556
Coefficient of Determination	0.97401	0.917194	0.752837	0.974009	**0.976444**

The best model to predict tip amount paid is the Neural Network Regression model. It has the least amount of errors & the best coefficient of determination. The next best being Bayesian Linear Regression and ordinary least squares Linear Regression. We could use any one of these three models to accurately predict the tip amount that will be paid.

Conclusion: We found several interesting bits of information by analyzing this huge data subset of 1.7 million rows. Adding weather and season data helped us figure out some more underlying patterns although the correlations were very weak to discern any link between weather and travel behavior like distance traveled or tips paid. However, several machine learning models were compared for making different predictions and the best one for each kind of prediction was picked. The following were the best models:

Problem definition	Model	features	Feature selection by	Evaluation method
Predict trip duration	Neural Net	*Log(trip_time_in_secs) ~ trip_distance + passenger_count + rate_code + speed + surcharge + Is_Night + monthnum*	Fisher score method	80/20 Split betw training & test data
Predict if a tip was paid or not	Two-Class Decision Forest	*Is_tipped ~ total_amount + fare_amount + trip_distance*	Spearman Correlation	10-Fold Cross Validation
Predict the amount of tip paid	Neural Net	*tip_amount ~ total_amount + fare_amount + trip_distance + trip_time_in_secs + tolls_amount*	Spearman Correlation	80/20 Split betw training & test data

Given the time, resources and the size of the data, these were a few analyses we could complete. More insights on travel behavior could perhaps be gained by obtaining traffic & events information, zipcode, income information by neighborhood, etc. This will be an interesting exercise for the future.