Maximizing profits for NYC yellow taxi drivers using NYC Yellow Taxi Dataset

ZYMIC: Zheng Chai, Yuan-Yao Chang, Michael Chon, Chong Zhang
Final Paper for COSC588/ANLY502
Georgetown University

ABSTRACT

In 2016, there were approximately 144 million taxi trips carried out by NYC yellow taxis throughout the city. Knowing where the pick-ups occur the most will likely increase a chance of picking up passengers, which will eventually lead to the higher earning potential for NYC yellow taxi drivers. In this paper, we explored the NYC taxi dataset and found the best hours, days, and months for NYC yellow cab drivers to work. We performed K-means clustering analysis on the entire dataset to find centroids, representing popular pick-up locations with a highest probability of pick-ups. In addition, we implemented a random forest model on the dataset and achieved a training accuracy of 99.10% and a test accuracy of 93.82%.

I. INTRODUCTION

In New York City, the yellow taxis are an important iconic symbol that represents the city. They are still often used as a means of public transportation to travel within the city. However, NYC taxi market has deteriorated for the traditional yellow cabs since the Uber and other ridesharing applications introduced in the city. NYC yellow cabs are still beating Uber and Lyft in the number of trips made in NYC but the share of trips shrank to 65% in April 2016 from 84% in April 2015.[1]

Understanding and exploiting the pick-up patterns in NYC, especially in the current situation, can be very crucial to NYC yellow taxi drivers because they will have a higher chance of picking up passengers, which will lead to the higher earning potential. Our project is to explore the NYC taxi dataset, find hidden pickup patterns in the dataset, and provide recommendations on hours, days, months, and locations for the drivers to have an opportunity to earn more profits.

II. PRIOR WORK

In *How does taxi driver behavior impact their profit? Discerning the real driving from large scale GPS traces,* this paper was published at Ubicomp/ISWC'16 and explored the large-scale GPS dataset to provide useful recommendations to taxi drivers and passengers in Bangkok, Thailand. The objective of this paper was to find patterns among the taxis and discover the earning potential of taxi drivers based on spatial and temporal profiles. First, the authors used their cost-distance algorithm to calculate the new taxi cost of each individual trip of thousands of taxis for 5 months. They analyzed the dataset to understand distance profit and service area in timely basis for analysis. For conclusion, the authors had suggested a few recommendations to the taxi drivers in Bangkok, Thailand; for instance, they had proposed that the suggested working hours are 8:00AM - 2:00PM and 6:00PM - 10:00PM where the taxi

[1] Holodny, Elena. "Uber and Lyft are demolishing New York City taxi drivers" *http://www.businessinsider.com/nyc-yellow-cab-medallion-prices-falling-further-2016-10*, Business Insider, 12 Oct, 2016

Does the model take into account the increase in competition?

drivers would get more profits working during those time period.

An article called *An effective taxi recommender system based on a spatio-temporal factor analysis model* discussed how mining historical GPS trajectories of taxis provided useful information for taxi drivers to make more profits. In the paper, the authors proposed a taxi recommender system for determining the next pickup locations. First, the authors collected the location clusters with a grid-based algorithm. They had analyzed the time distribution of passenger pickups and the distribution of revenue each day. They built a location to location model, OFF-ON model, based on the pickup and drop-off information from the GPS dataset to obtain the average revenue and the pick-up probability of each location. Finally, they constructed a taxi recommender model. For conclusion, they claimed that the model yielded an average revenue that is 62% higher than the average revenue found on each weekday.

III. METHODOLOGY

A. DATASET

There are many publicly available datasets regarding NYC taxi from the website[2] managed by the city of New York. The data used in the attached datasets were collected and provided to the TLC by technology providers authorized under the Taxicab Passenger Enhancement Program (TPEP). The trip datasets were not created by the TLC and TLC does not hold a responsibility for the accuracy of these data.[3]

The dataset used for this particular project is exactly the same dataset that can be retrieved from NYC OpenData but the dataset was obtained from a public AWS s3 bucket account.

The dataset contains the data of every single trip made by the yellow cabs in New York City from January-2016 to December-2016. The format of the dataset was in a csv (comma separated values) format which is 2GB for each month. We have combined 12 csv files (one-year worth data) by loading them on Spark and combine them into one SparkSQL table.

In the dataset, each data point contains information regarding every single trip made by a yellow cab in NYC. The attributes of the dataset are VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, pickup_longitude, pickup_latitude, etc. The exact description of each attribute in the dataset is provided on the pdf file[4] available online.

If a trip was made by a yellow cab and the trip was recorded in the dataset, the attributes of that data point trip provided lots of information about the trip. For example, tpep_pickup_datetime means the time when a customer was picked up, tpep_dropoff_datetime means the time when the customer was dropped off, pickup_longitude and pickup_latitude represent the exact coordinates of the customer pick-up location on a map, etc.

---

[2] NYC OpenData, *https://data.cityofnewyork.us/*

[3] *https://data.cityofnewyork.us/Transportation/2013-Yellow-Taxi-Trip-Data/7rnv-m532*

[4] "Data Dictionary - Yellow Taxi Trip Records" *http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf*

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
passenger_c	trip_distance	pickup_longi	pickup_latitu	RatecodeID	store_and_f	dropoff_long	dropoff_latit	payment_ty	fare_amount	extra	mta_tax	tip_amount	tolls_amoun	improvemen	total_amoun
5	0.96	-73.979942	40.7653809	1	N	-73.966309	40.7630882	1	5.5	0.5	0.5	1	0	0.3	7.8
2	2.69	-73.972336	40.7623787	1	N	-73.989429	40.7498894	1	21.5	0	0.5	3.34	0	0.3	25.64
1	2.62	-73.968849	40.7645302	1	N	-73.974648	40.7916412	1	17	0	0.5	3.56	0	0.3	21.36
1	1.2	-73.993935	40.741684	1	N	-73.997665	40.747467	1	6.5	0.5	0.5	0.2	0	0.3	8
2	3	-73.988922	40.7269897	1	N	-73.975594	40.6968689	2	11	0.5	0.5	0	0	0.3	12.3
1	6.3	-73.974083	40.7629128	1	N	-74.012802	40.7022095	1	20.5	0.5	0.5	4.35	0	0.3	26.15
6	0.63	-73.968315	40.7553291	1	N	-73.962082	40.7588149	1	4	0.5	0.5	1.06	0	0.3	6.36
2	1.91	-73.994209	40.7461014	1	N	-74.00425	40.7218094	1	8	0.5	0.5	1.86	0	0.3	11.16
1	4.5	-74.00676	40.7189064	1	N	-73.988083	40.7728539	1	16.5	0.5	0.5	3.56	0	0.3	21.36

Screenshot of the original csv

B. DATA PREPARATION / CLEANING

The dataset contained many erroneous data points and outliers in various features. For instance, some data points had their total_fare_amount less than $2.50, which is impossible, because the initial charge for NYC yellow cabs is $2.50 and total_fare_amounts less than $2.50 had to be considered as an error and were excluded from the dataset. In addition, some data points had pickup_latitude and pickup_longitude equal to dropoff_latitude and dropoff_longitude, which could be considered as that taxis picked up customers and dropped them off at the same locations the taxis picked up. These instances were considered as an error; those data points were excluded from the dataset.

The first half of the entire dataset had the columns: "pickup_latitude", "pickup_longtitude", "dropoff_latitue", and "dropoff_longtitude", which were changed to "PULocationID" and "DOLocationID" in the other half of the dataset because the other half did not contain the attributes, "pickup_latitude", "pickup_longtitude", "dropoff_latitue", and "dropoff_longtitude". To handle this difference, first, we joined the second half of the dataset with "taxi_zone_lookup" table, which was provided by NYC public data, and then unioned the first half and the other half of the entire dataset. Afterwards, we dropped some unnecessary, such as "Store_and_fwd_flag" and "VendorID" and create new attributes for analysis, such as, "pickup_hour_group", "month", "pickup_weekdays". In addition, for analysis, we changed the datatype of some columns from string to float by using a function called, "withColumn".

How many excluded?

pickup_weekday	if_weekday	fare_amount_group	tip_amount_group	month	pickup_hours	dropoff_hours
Sat	Weenkend	5-10	1-2	1	6-8	6-8
Sat	Weenkend	5-10	Invalid	1	6-8	8-10
Sat	Weenkend	10-15	2-3	1	6-8	8-10
Sat	Weenkend	2.5-5	Invalid	1	6-8	6-8
Sat	Weenkend	2.5-5	1-2	1	6-8	6-8

Screenshot of the cleaned dataset

IV. ANALYSIS

A. LINE GRAPHS

After data preparation and cleaning, we used SparkSQL to calculate the number of pickups, tips occurrence, total amount of tips, and total amount of fares. We used a python module called pygal to generate line graphs to see how each day in a week varies in

terms of the numbers of pickups and tis occurrence in every 2-hour period in 2016.

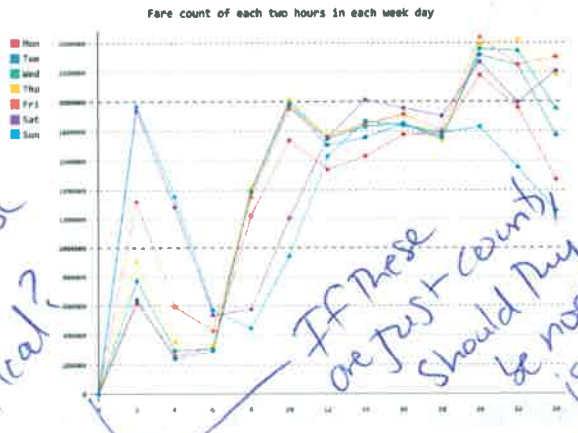

Fig. 1: Line graph displaying the frequencies of pickups occurred between every 2-hour period on each day of week

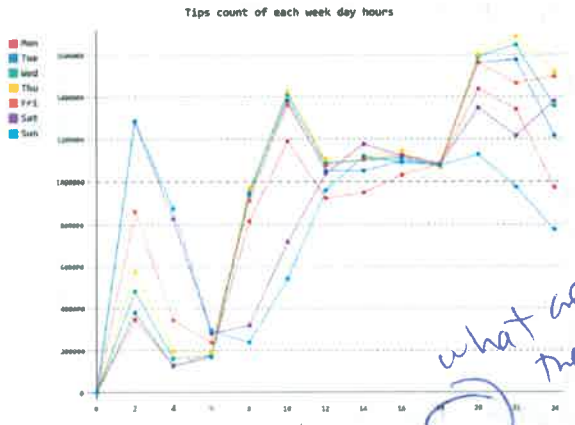

Fig. 2: Line graph displaying the frequencies of tips occurred in every 2-hour period on each day of week

Fig.1 [5] represents the number of pickups occurred in a 2-hour time period. In Fig.1, the purple and blue lines, representing Saturday and Sunday respectively, have the highest peaks during 0:00 - 2:00, meaning a lot of pickups occurring in that time period. From 6:00 - 10:00, the rest of the colors other than purple and blue shows very steep increase, indicating many pickups happening in morning rush hour. Later on, during 18:00 - 20:00, the numbers of pickups increase even higher due to evening rush hour and drop substantially except Thursday, which showed a slight increase.

Fig.2 [6] represents the number of tip occurrences in a 2-hour time period. Fig.2 is very similar to Fig.1 in appearance. However, noticeable differences are that the highest pickups occurred during 18:00 - 20:00 on Friday and that the highest number of tip occurrences happened during 20:00 - 22:00 on Thursday.

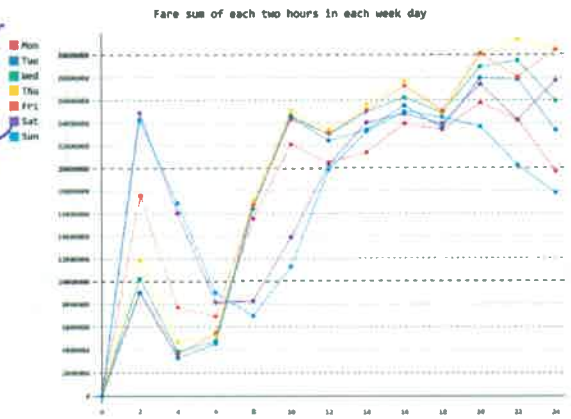

Fig.3: Line graph displaying the amounts of fares earned in every 2-hour period on each day of week

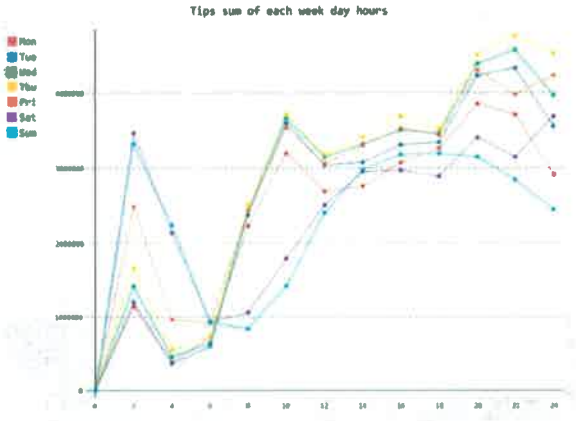

Fig.4:Line graph displaying the amounts of tips earned in every 2-hour period on each day of week

[5] Fare count: https://s3.amazonaws.com/StevenChang/Massive+Data/hours_count.svg

[6] Tips sum: https://s3.amazonaws.com/StevenChang/Massive+Data/tips_hours_count.svg

Fig.3 [7] represents the total fare amount occurred in each 2-hour time period. Fig.3 shows the same trends as the two previous line graphs. However, it shows that the total fare amounts are generally higher on Thursday and Friday especially, around 18:00 - 20:00. However, it shows that on the yellow line, representing Thursday, is far above Friday during 20:00 - 22:00.

Fig.4[8] represents the total tip amount. From 0:00 - 6:00, the purple and blue lines are far above the other lines, indicating that the total tip amounts are higher on weekend than weekdays. As shown by the graph, starting from 6:00 to the rest of a day (24:00) the yellow line relatively stays above the other lines, meaning people pay a higher amount of tips to taxi drivers.

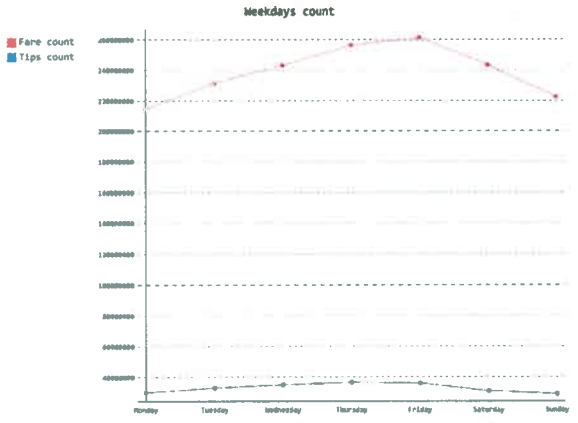

Fig.5: Line graph displaying the number of pickups and tips occurred on each day of week

paper shows your thought process. I really want the analysis

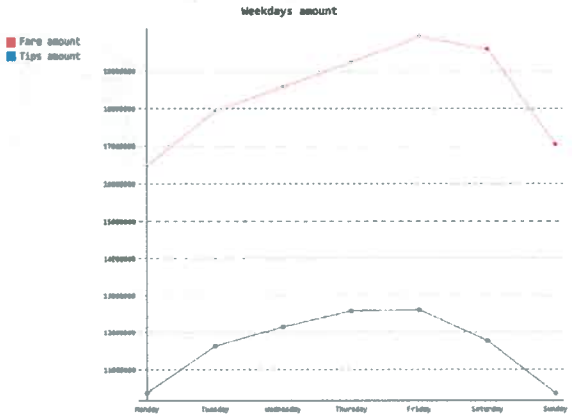

Fig.6: Line graph displaying the total amounts of fares and tips earned on each day of week

For Fig.5[9] and Fig.6[10], we grouped the entire dataset by day of a week and found the numbers of pickup and tip occurrences and the total amounts of fares and tips. According to the line graphs, Friday is the most profitable day for NYC yellow cab drivers to work on. We further looked into the data as follows:

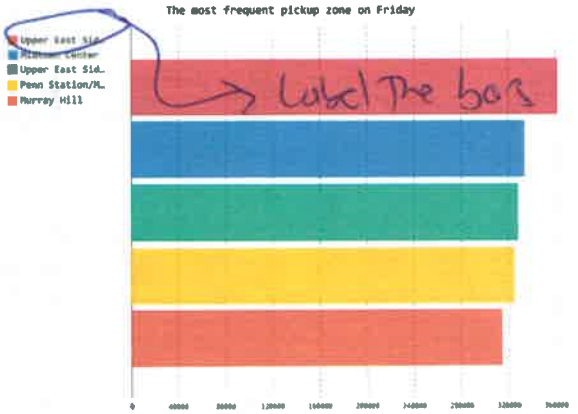

Fig.7: Bar graph displaying the top five pickup locations on Friday

[7] Fare sum: https://s3.amazonaws.com/StevenChang/Massive+Data/hours_sum.svg

[8] Tips sum: https://s3.amazonaws.com/StevenChang/Massive+Data/tips_hours_sum.svg

[9] Weekday count: https://s3.amazonaws.com/StevenChang/Massive+Data/weekdays_count.svg

[10] Weekday sum: https://s3.amazonaws.com/StevenChang/Massive+Data/weekdays_amount.svg

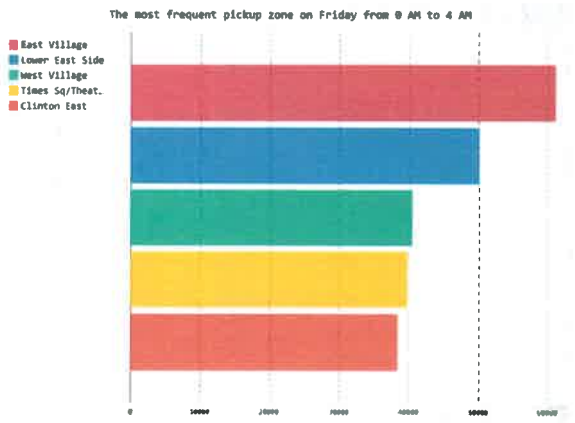

Fig.8: Bar graph displaying the top five pickup locations in every 4-hour period on Friday

Fig.7 represents the top five pickup locations for NYC yellow cab drivers. The locations are 1) Upper East Side South, 2) Midtown Center, 3) Penn Station/Madison Square West, 4) Upper East Side North, and 5) Murray Hill.

Fig.8 [11] displays how top five pickup locations vary in each 4-hour time period on Friday. From 0:00 - 4:00, East Village is found to be very popular for picking up passengers. From 4:00 - 8:00, Penn Station/Madison Square is busier for picking up passengers. From 8:00 - 20:00, Upper East Side is extremely pickup locations for drivers because the numbers of pickups occurred in Upper East Side are so much higher than the numbers of pickups occurred in other places. From 20:00 - 24:00, East Village becomes a popular pickup location for drivers.

B. K-MEANS CLUSTERING

After data cleaning, we took out the errors in the following attributes in the dataset, "pickup_latitude", "pickup_longitude", "the dropoff_latitude", and "dropoff_longitude". Then, we decided to do K-means clustering on pickup latitudes and longitudes using KMeans from sklearn.cluster. Below is the graph of our K-means clustering as Fig.9. Apparently, there is a huge cluster in the center which represents all the points in the NYC metropolitan area.

how does this cleaning bias results?

We cleaned the dataset again by taking out all the other points which is far away from the center we treated those as outliers. We selected four geo-locations, (41, -74.5), (41, -71.5), (40.5, -74.5), and (40.5, -71.5), and set up a boundary for the area, which included entire NYC areas and some parts of northeast New Jersey, including Newark Airport.

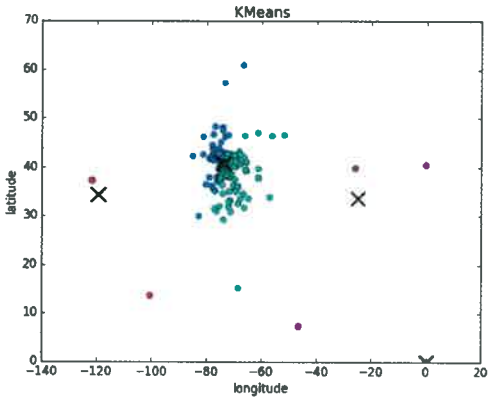

Fig.9: K-means clustering plot on pickup_latitude and longitude before cleaning

Once the cleaning steps were done, we generated K-means clustering again. Below are our results; the result on the left as Fig.10 is pick-up and the result on the right as Fig.11 is drop-off. Nearly, all of the centroids are on the left side of the clustering, which means that all of the pick-ups and drop-offs are concentrated around the Manhattan island.

[11] Friday top 5 most pickups in separated by 4 hours a group: https://s3.amazonaws.com/StevenChang/Massive+Data/Friday_hours_group.gif

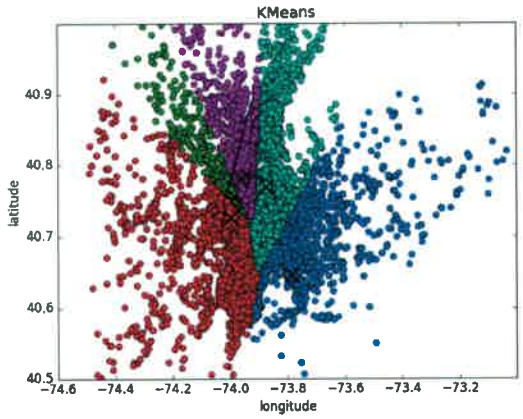

Fig.10: K-means clustering plot on pickup_latitude and longtitude with 5 centroids after cleaning

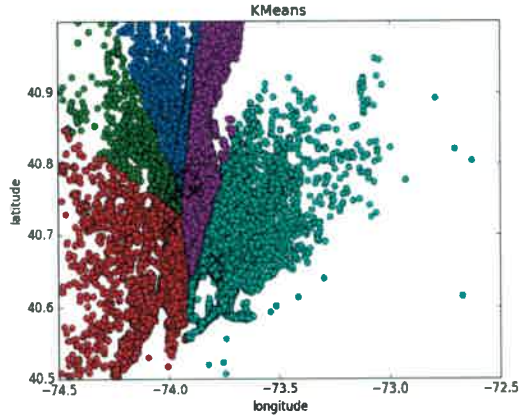

Fig.11: K-means clustering plot on dropoff_latitude and longtitude with 5 centroids after cleaning

All of the K-means clustering were done by choosing k = 5. Also, we have done the clustering by using k starting from 1 to 20[12], and drew the centroids on the Google Map to have a better visualization. For the maps below, 1, 5, 15 and 20 (Fig.12, Fig.13, Fig.14, Fig.15) are selected to be the number of centroids, k. Nearly, all the centroids were accumulated in Manhattan; as for Newark airport, the centroid appeared around Newark airport when k was equal to 17.

[12] K-means 1 to 20: https://s3.amazonaws.com/StevenChang/Massive+Data/Kmeans_1-20.gif

After generating k-means clustering plots from k = 1 to 20, we decided to use k equal to 5 as our means to find popular pick-up locations because we found out that the three of the centroids appeared in all three distinct parts of Manhattan, each representing as 1) Uptown, 2) Midtown, and 3) Downtown. It is simpler to summarize and provide a popular pick-up location in each part of Manhattan.

Fig.12: K-means clustering plots with k = 1 on Google map

Fig.13: K-means clustering plot with k = 5 on Google map

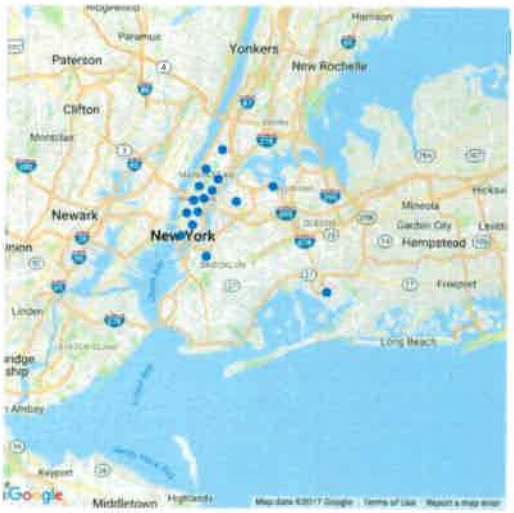

Fig.14: K-means clustering plot with k = 15 on Google map

Fig.15: K-means clustering plot with k = 20 on Google map

The following 5 centroids are found for Pick-ups and Drop-offs: [13]

Pick-up	JFK	LGA	SoHo	Time Square	Metropolitan Museum
Drop-off	Cooper Hewitt Smithsonian Design Museum	Bryant Park	94b E Broadway(Near Chinatown)	Holiday Inn beside JFK	Gorman Playground (beside LGA)

[13] Kmeans pickup and dropoff: https://s3.amazonaws.com/StevenChang/Massive+Data/kmeans.gif

C. HEATMAP

Based on the day of a week graph results shown above, Friday is the day with the most pickups and the highest income potential for NYC yellow cab drivers. We separated the entire Friday into 6 time groups using SparkSQL, each representing 4-hour (0-4, 4-8...etc.).

We used a python module called gmaps to generate heatmaps to understand the density of pickups happened in Manhattan. According to the heatmaps as in Fig.16, Fig.17, and Fig.18, in the morning, customers hail more taxis more frequently around the south of Uptown East. By afternoon, customers hail more taxis around Midtown and Upper East side. At night, more taxis have been hailed in downtown compared to the afternoon.[14]

Fig.16 – Heatmap displaying pickups occurred between 4:00AM – 8:00AM on Friday

Fig.17: Heatmap displaying pickups occurred between 12:00PM – 4:00PM on Friday

Fig.18: Heatmap displaying pickups occurred between 16:00PM – 10:00PM on Friday

As clearly seen in the heatmaps, it is interesting that many pickups occurred on avenues than on streets.

D. RANDOM FOREST

The original dataset, from July to December, used location_ids representing zones in the city instead of the combination of latitude and longitude. We decided to use a Random Forest model to predict possible pickups on Friday since we know that

[14] Heatmap for Friday: https://s3.amazonaws.com/StevenChang/Massive+Data/HeatMap_zoom.gif

Friday is the most profitable day to work on. By using the model, it will support our recommendations for Friday, proving its accuracy and reliability.

We chose Random Forest algorithm because it is one of the most accurate learning algorithms. It fully utilizes the datasets, and it can estimate the weighting schema of variables in the classification[15]. We split the dataset from January to June to train the model using SparkSQL. This subset of the original dataset was randomly split into two parts, 80% of which was used for training, and the remaining was used for testing. We used RandomForestRegressor from sklearn.ensemble to build our random forest model. We ended up with a model that has a training accuracy of 0.9910 and a test accuracy of 0.9382. The details of the inputs and outputs of our random forest model are shown in Table 1, followed by Fig.20 and Fig.21.

For demonstration, we predicted possible pick-up locations in New York City using the random forest model. The result is shown in Fig.22.

We generated the heatmaps as Fig.19 for the whole day and for every 4-hour period[16] from the result. We were also able to put pins on the heatmaps, representing the district IDs from the original dataset from July to December. The heatmap generally covered the areas around the pins. However, some of the pins might seem to be on the edge of the red areas because one pin only represents one point in a zone and it does not cover an entire area of the zone pointed by the pin.

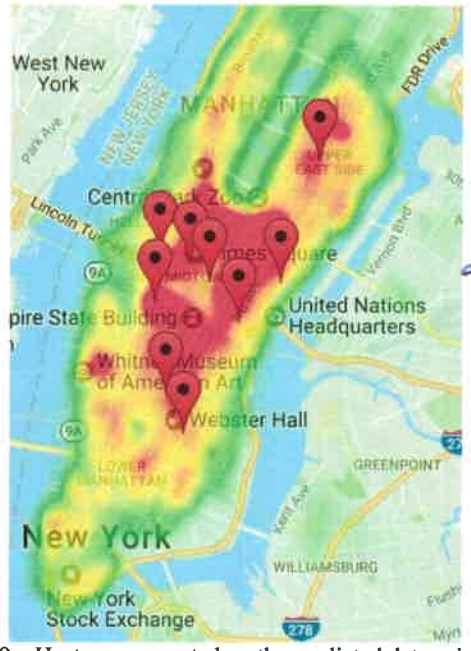

Fig.19 – Heatmap generated on the predicted data using the random forest model with the pins representing actual pickup locations

At the same time, we can also easily expand our predictions to a whole week. Because our dataset for training and testing remains the same, and the split methods are also the same, we will always have random forest models with similar training accuracy and testing accuracy. When we eliminate the restrictions on time variables when generating new longitude and longitude, we will have a result for the whole week.

[15] Vikram Jha. Random Forest algorithm. January 25, 2012.http://amateurdatascientist.blogspot.com/2012/01/random-forest-algorithm.html

[16] Four-hourly divided heatmap for New York Taxi Friday pickups https://s3.amazonaws.com/StevenChang/Massive+Data/predictHeatmap.gif

Random Forest Model	
Input (Dataset from 2016-1 to 2016-6)	Output (Prediction result)
Geohash • Translated into latitude and longitude with Geohash library before training and testing	latitudes and longitude • Predicted future pickup locations
time_num • Normalized into each 30-minute, divided by 24*60	time_num • Predicted future pickup times • Can be translated into time
time_cos • Cosine value of time_num	time_cos • Cosine value of time_num
time_sin • Sine value of time num	time_sin • Sine value of time num
day_num • Day of the week as a numerical feature α going from 0 to 1 • day_num = (α + time_num*60) / 7	day_num • Predicted future pickup days • We set restrictions on this variable to only predict for Friday • Can be translated into day of week
day_cos • Cosine value of day_num	day_cos • Cosine value of day_num
day_sin • Sine value of day_num	day_sin • Sine value of day_num
week_end • 0 if weekday, 1if weekend	week_end • 0 if weekday, 1if weekend
pickups • Total pickups occurred with same time_num, day_num and geohash	pred_pickups • Total pickups predicted for this time_num, day_num and the combination of latitude and longitude

Table 1: Table displaying the inputs and the outputs of the random forest model

How can the model predict so many things? Is it predicting for each row?

Input:

geohash	time_num	time_sin	time_cos	day_num	day_sin	day_cos	pickups	weekend
dr5ru73	0.28125	0.98078528	-0.1950903	0.04017857	0.24977648	0.96830352	2011	0
dr5ru73	0.29166667	0.96592583	-0.258819	0.04166667	0.25881905	0.96592583	1838	0
dr5ru73	0.27083333	0.99144486	-0.1305262	0.03869048	0.24071208	0.97059657	1759	0
dr5ru73	0.30208333	0.94689013	-0.3214395	0.04315476	0.26788899	0.96346369	1756	0
dr5rzjy	0.45833333	0.25881905	-0.9659258	0.06547619	0.39989202	0.91656226	1735	0
dr5ru63	0.3125	0.92387953	-0.3826834	0.04464286	0.27683551	0.96091732	1692	0
dr5ru63	0.32291667	0.89687274	-0.4422887	0.04613095	0.28580784	0.95828695	1682	0
dr5rzjy	0.42708333	0.44228869	-0.8968727	0.06101191	0.37402858	0.92741718	1643	0
dr5rzjy	0.46875	0.19509032	-0.9807853	0.06696429	0.40844426	0.91278327	1624	0

Fig.20: Screenshot of the inputs of the random forest model

Output:

latitude	longitude	time_num	time_sin	time_cos	day_num	day_sin	day_cos	weekend	pred_pickups
40.7558441	-73.99086	0.3125	0.92387953	-0.3826834	0.61607143	-0.6863466	-0.7456422	0	1368.93877
40.769577	-73.863144	0.79166667	-0.9659258	0.25881905	0.88452381	-0.9165623	-0.399892	0	1362.86129
40.769577	-73.863144	0.78125	-0.9807853	0.19509032	0.88303571	-0.9127833	-0.4084443	0	1362.36554
40.769577	-73.863144	0.80208333	-0.9469301	0.32143947	0.88601191	-0.9202611	-0.3913048	0	1321.43178
40.769577	-73.863144	0.71875	-0.9807853	-0.1950903	0.87410714	-0.8884456	-0.4589819	0	1297.18141
40.769577	-73.863144	0.8125	-0.9238795	0.38268343	0.8875	-0.9238795	-0.3826834	0	1290.97277
40.769577	-73.863144	0.70833333	-0.9659258	-0.258819	0.87261905	-0.8841154	-0.4672686	0	1276.77371
40.769577	-73.863144	0.82291667	-0.8968727	0.44228869	0.8888881	-0.9274172	-0.3740286	0	1242.57797
40.769577	-73.863144	0.69791667	-0.9469301	-0.3214395	0.87113095	-0.8797079	-0.4755145	0	1241.58977
40.769577	-73.863144	0.77083333	-0.9914449	0.13052619	0.88154762	-0.9089245	-0.4169608	0	1236.36711

Fig.21: Screenshot of the outputs of the random forest model

Fig. 22: Screenshot of predicted, possible pickup locations using the random forest model

V. CONCLUSION

Based on our analysis, we can provide the following strategies for NYC yellow cab drivers to earn more profits.

1. From the line graphs, the profitable days for NYC yellow cab drivers to work on are Thursday, Friday, and Saturday. Here is the table for hours that they should work.

Friday	Saturday	Sunday
10:00AM - 12:00AM	10:00AM - 2:00AM	2:00PM - 12:00AM

We found out that more tips and pickups happened on the three days and around those hours suggested in the table.

2. In general, on weekdays in the morning rush hour between 6:00AM - 10:00AM and the evening rush hour between 6:00PM - 8:00PM (18:00 - 20:00), many pickups and tips occurred at a higher frequency and total fare amounts are higher around those hours.

3. On weekend, it is more profitable to work around 0:00AM - 2:00AM because many pickups and tips occurred at a higher frequency and total fare amounts are higher around those hours due to nightlife in NYC on the weekend.

4. Based on our k-means clustering plot for pickup locations, we have found the five centroids, which represent an average pickup location in each cluster. The five centroids are 1) JFK airport, 2) LaGuardia Airport, 3) SoHo, 4) Time Square, and 5) Metropolitan Museum. From the above analysis of the k-means clustering, there were three centroids in Manhattan, meaning that a chance of picking up passengers in Manhattan is far higher than picking up passengers outside Manhattan. So, if NYC yellow cab drivers want to pick up passengers outside Manhattan with a higher probability, these drivers should go to JFK airport and LGA airport according to our k-means clustering plot.

5. Based on the line graphs, overall, Friday is the most profitable day to work on because the sum of fares and tips amounts on Friday are higher than the sums on other days. If a yellow cab driver wants to drive in Manhattan on Friday, the top five pickup locations are 1) Upper East Side South, 2) Midtown Center, 3) Penn Station/Madison Square West, 4) Upper East Side North, and 5) Murray Hill. For the other days of week, NYC yellow cab drivers can use our random forest model to predict possible pickup locations in the near future.

VI. FUTURE WORK

1.Stochastic Simulation:
We could have used the model would be to do a Monte Carlo simulation of taxis driving around the city, and then we could see if our virtual taxi can get or not get the fares.

2. Routes:
We could have optimized various routes to maximize profits.

3. Datasets:
We could have used other datasets such as Green Taxi and Uber or even weather data to compare and find competitive strategies.

VII. APPENDIX

Hours distance total amount:
https://s3.amazonaws.com/StevenChang/Massive+Data/hours_distance_totalamount.svg
Fare months count:
https://s3.amazonaws.com/StevenChang/Massive+Data/months_count.svg
Fare month sum:
https://s3.amazonaws.com/StevenChang/Massive+Da

ta/months_sum.svg
Tips months count:
https://s3.amazonaws.com/StevenChang/Massive+Data/tips_months_count.svg
Tips months sum:
https://s3.amazonaws.com/StevenChang/Massive+Data/tips_months_sum.svg
K-means pickups:
https://s3.amazonaws.com/StevenChang/Massive+Data/kmeans_pickup.html
K-means drop-offs:
https://s3.amazonaws.com/StevenChang/Massive+Data/kmeans_dropoff.html

VIII. REFERENCE

Vikram, Jha. Random Forest algorithm. January 25, 2012.
http://amateurdatascientist.blogspot.com/2012/01/random-forest-algorithm.html

Ren-Hung Hwang, Yu-Ling Hsueh, and Yu-Ting Chen. 2015. An effective taxi recommender system based on a spatio-temporal factor analysis model. Information Sciences 314: 28-40.

Meng Qu, Hengshu Zhu, Junming Liu, and Guannan Liu. 2014.
A Cost-Effective Recommender System for Taxi Drivers. In Proceedings of Hui Xiong
KDD'14, 45-54.
http://dx.doi.org/10.1145/2623330.2623668.