Big Data Analysis of The Impact of COVID-19 on NYC Taxi Market: A Discussion of Change In Travel Model and Resilience of Taxi Market Under Global Emergencies

Yichuan Zhang* 904722
*University of Melbourne MAST30034 APPLIED DATA SCIENCE
†ASSIGNMENT 2

***Abstract*—Big data analysis is a field with increasing demand in the 21th century. This research aims to analyze and forecast the impact of COVID-19 on the New York City (NYC) taxi market in terms of the change in passengers' travel mode and mark resilience by applying visualizations and empirical models, such as Random Forest and Time Series. In our search, we are interested in the questions, such as whether the COVID-19 changed passengers' behaviors during the first half of 2020 and how the taxi market will perform in the following month in terms of taxi demand. Within the research, the method and procedure to proceed data preprocessing and feature engineering were provided. Multi-variate visualizations based on big data shows that people's travel patterns can be divided into three types, and that travel patterns after April significantly differs than those before the COVID-19. Lastly, machine learning and time series models are used to model and analyze future trends. Among them, the OLS model confirmed the significance of the impact of COVID-19 on taxi industry, and the Random Forest model showed a high predictive accuracy. In the time series model, through the prediction analysis of ARMA and SARIMAX models, this research evaluates that the taxi market is slowly warming up but may reach a new steady state.**

***Index Terms*—Big data analysis, Machine learning, Consumer behaviors, DBSCAN algorithm, Random Forest, Time Series.**

I. Introduction

A. Background

In New York City (NYC), the busy buzzing city life endows this city with a representative status in this world. As the world's largest urban circle, New York City is the city with the largest population in the United States and an international metropolis with great influence over the world's economy, commerce, finance, media, politics, education, and entertainment. As one of the main transportation modes for people to travel, the taxi market is often more sensitive to big events than any other markets, especially in this outbreak, the taxi market can often release more information than just travel mode. The big data analysis to the taxi industry is shown to be important to analyze the use of taxi and demand supply [1]. Therefore, it is valuable and insightful to investigate the impact of any change from the taxi industry on consumer behaviors.

During this time, due to the COVID pandemic, different countries have adopted various degrees of policies to alleviate the harm of the COVID pandemic, and there is no doubt that the COVID pandemic will have a significant impact on the taxi market. Although many countries encourage the "stay-at-home" policy, the United States has not followed the extreme lock-down policy at this stage, especially the NYC government is planned to restart the city by reopening all the services in four phases[2].

In the case of sudden global emergence, it has raised a new challenge to the transportation system, especially for the busy city with a larger population, such as NYC. This profound and consistent challenge might change the people's lifestyle and impact the taxi industry in terms of demand and service hours, which might hardly return to the normal level at the per-COVID stage [3].

So, the question arises naturally:" *1. Is there any impact on consumer behaviors due to Covic-19, how can we visualize it and prove it? 2. How can we show and forecast trend of taxi demand, and how is the resilience of taxi industry?*". The potential stakeholder in this research could be taxi companies in NYC, or those taxi drivers who want to know the current situation of taxi industry and future tendency in NYC, since our research will be related with analysis of the passpengers' travel under COVID-19 and also statsical analysis of future trends. Although the extreme policies have not been taken by the NYC government [2]. this study remains committed to finding the impact of the COVID pandemic outbreak on the New York taxi industry.

II. Data Specification and Preparation

A. Data Characteristics

Data-sets	Source	Sizes	Row number	No.Feature
Taxi Trip 2020-02	TCL	584 MB	6299354 rows	18
Taxi Trip 2020-03	TCL	278 MB	3007292 rows	18
Taxi Trip 2020-04	TCL	21.7 MB	237993 rows	18
Taxi Trip 2020-05	TCL	31.6 MB	348371 rows	18
Taxi Trip 2020-06	TCL	50.3 MB	549760 rows	18
NYC COVID data	NYCH	30.4KB	212 rows	43

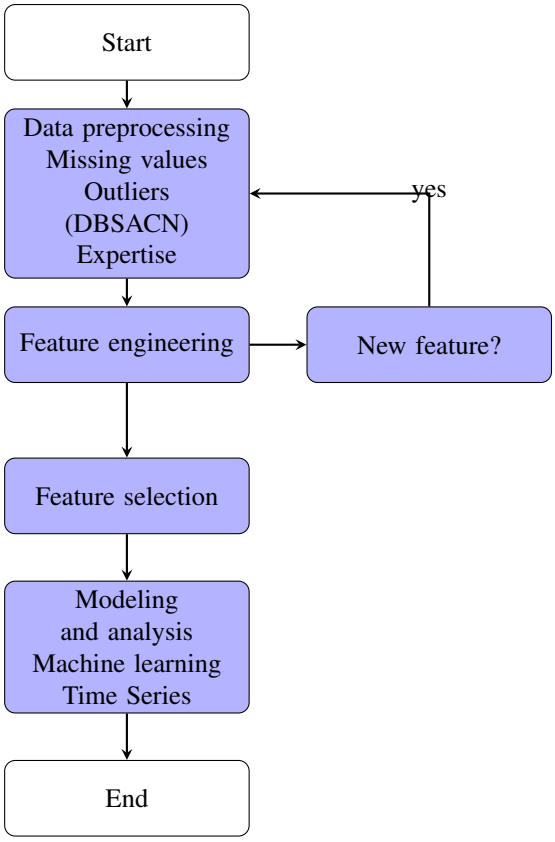
TABLE I
Data Characteristic For Each Data-Set

The data-sets used by this research are New York City Taxi and Limousine Service Trip Record Data (TLC)[4], Covid-19 data from NYC Health (NYCH)[5], and The COVID Tracking Project (CTP)[6]. The data in TLC records information for each taxi and for-hire vehicle trip, it includes features such as pick-up time, trip distance. The data in NYCH and CTP records the features such as case number, death number, and test number in NYC by borough.

Further, in the aspect of data usage. The trip records were selected across the range from 01-02-2020 to 30-06-202, which takes seconds as a unit. There are no further records released after June. The TCL data-set has a total size of 1 GB. Correspondingly, the COVID data-sets were selected across the range from 01-03-2020 to 30-06-2020, which take days as a unit. Since the NYC Health started to count the COVID-19 confrimed cases from March, therefore we assume the date before March did not get affected by COVID-19, and we treat the trip records in February as per-COVID stage. Consider the COVID-19 is the event in 2020, therefore, data before 2020 is not necessary to be included.

Now the TCL data set has over 10 million sizes with no specific order, and COVID data set has around 120 rows of data in term of days.

B. Data Preparation



In the data preparation stage, this paper presents the process of data cleaning, feature engineering, and feature selection, respectively. Firstly, we assume COVID data given by the official department is true and no needs to pre-process. For TCL trip records, this phase shows how raw data is converted into usable data for modeling as well as prediction. Especially, the DBSCAN clustering algorithm is used for identifying the outliers.

In the part of data cleaning, it mainly shows the method to deal with missing values, outliers, and indescribable strange data points. The first way to deal with missing values is to remove all missing values. Given that the data set is tens of millions, direct removal of missing values will not lead to excessive loss of information (average of 40,000 numbers of data were removed from each month's records). Secondly, considering the data-set contains numerous outliers in various ways, and it is also hard to define whether all features fit the shape of normal distribution, therefore the DBSCAN clustering algorithm is used to remove the outliers instead of using Z-score.

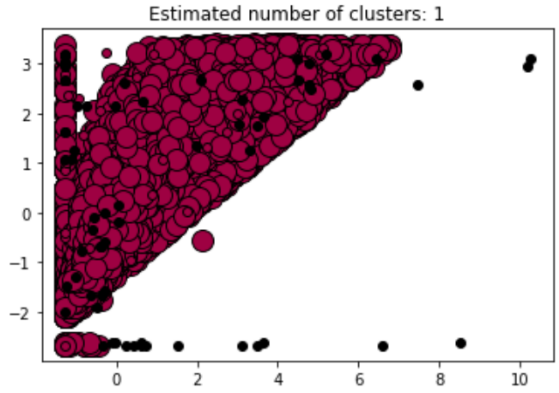


Fig. 1. Cluster and Outliers by Using DBSCAN Algorithm.

Density-based spatial clustering of applications with noise (DBSCAN) is a grouping algorithm, it calculates the distance and marks the outliers in a low-density region. Considering the high complexity of DBSCAN algorithm, the whole data-set could not be processed at once, therefore a self-written recursion function is designed to process the data-set by sampling method, and the complexity is reduced to $O(nlog(n))$ from $O(n^2)$ with no loss of accuracy[7]. From the figure above, it indicates that there is only one cluster but some outliers are represented as block dots. However, some outliers were not easy to be detected even by this algorithm, in some special cases. Some outliers have to be identified by the researchers' expertise. According to our knowledge and policies in NYC, the speed limit should be strictly below the 70KM/H and the trip duration might not exceed a full day. Therefore, any instance contains a speed over 70KM/H, or the trip duration of over 300 minutes was removed.

In the part of feature engineering, it mainly focuses on integrating different data-sets and generating more valuable features based on the raw features. Firstly, within the Taxi trip (TCL) data set. The features such as "Speed" and "Time duration" were easily generated, and time was specified in "Month", "Week" and "Day" by using DateTime object. In the next step, daily and monthly count for each feature was aggregated by grouping the data. Lastly, the feature "Cases confirmed" from the COVID data-set was integrated into the Taxi trip data-set.

In the part of Feature selection, it demonstrates the statistical method used for converting categorical variables into dummy variables, and the filter method for the non-significant variables. Within data, it contains various types of data types. Even though most of them are "int" and "float" type, whereas categorical data, such as "payment type" cannot be used

directly during modeling. Therefore a less than full rank model was constructed by adding constant at the first column and converting all the categorical data into dummy variables.

III. Visualization and Analysis

Visualization is an extremely important part of data science, which can not only show the information contained in the data itself to readers in a very rapid and intuitive way, but also guide the direction of future analysis. This study will use the "Seaborn" visualization tool to provide an overview of people's travel patterns, and we are mainly concerned about whether there is any difference in people's travel patterns before and after the Covid-19, and whether those differences are in line with our expectations. The data used for visualization will certainly be all 9 million datasets, so this is a big data visualization, but we have an important assumption that the data published by TCL is real and complete. In the following visualization, the project will present and analyze the values of "Daily COVID-19 confirmed cases", "Trip distance", "Trip duration", "Tip amount by customer", and "Taxi demand" in terms of different hours, days, and months.

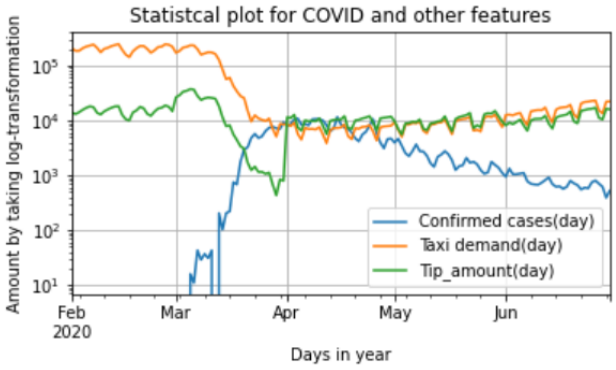


Fig. 2. Daily Confirmed Cases, Taxi Demand And Tip Amount

The figure above gives a general trend information by taking the log-transformation on y-axis, which combines daily COVID-19 confirmed cases, daily taxi demand, and daily tip amount. The figure indicates that from March, the daily confirmed cases from March climbed quickly, and reach the peak point in early April, and then slowly decreased. At the same time, the daily demand for taxis market has also decreased significantly while the number of confirmed diagnoses was rising, and after April the demand is slowly increasing with the number of confirmed diagnoses decreasing.

From the above figure, we find that passengers' travel modes can be divided into three types. In February and March, passengers' behaviors before the Covid-19 can be divided into the first category, in April and May, when the Covid-19 was most severe, passengers' behaviors can be divided into the second category, and in June, when the Covid-19 resumes passengers' behaviors will be the third category. Plot reveals that before the outbreak, passengers would exercise farther areas on weekday, while the distance traveled after the outbreak is not as long as before, and after June passengers

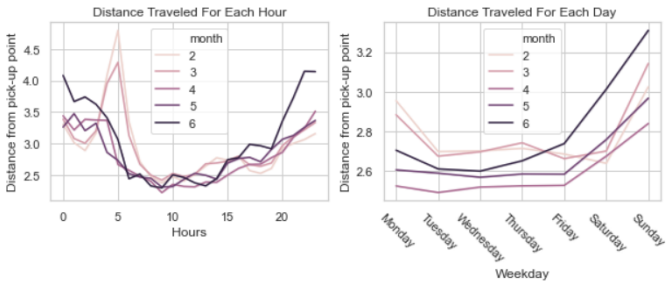


Fig. 3. Total Distance Traveled In Each Day Form February to June.

were more likely to exercise farther places on Saturday and Sunday.

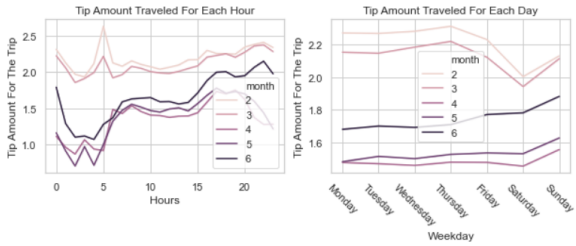


Fig. 4. Total Tip Amount In each day Form February to June.

Furthermore, this research will discuss tip and total amount uniformly because they are very similar. From the figure below, passengers would give more tips at 5 am on weekday before the outbreak, but after the outbreak, passengers would prefer to give more tips after 10 am on weekends. And after the outbreak, there was a large decrease in the level of overall tips given.

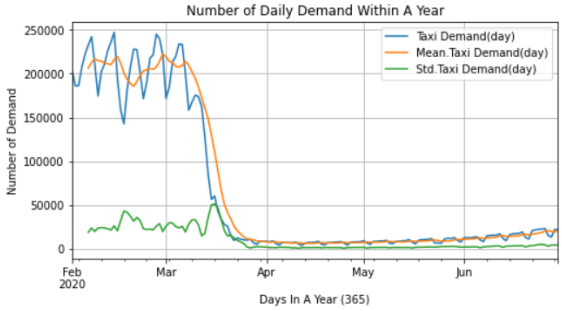


Fig. 5. Daily Taxi Demand in each day form February to June.

Moreover, the most dramatic float is the demand graph, which shows the demand for each day. For daily demand, from the plot, there was a huge drop between March and April. But after decompose the overall data (see Appendix fig.11), the data showed that taxi demand would start slightly to rise slowly from April.

In addition to this, we analyzed what was the difference between the number of passengers riding taxi before the outbreak occurred. Through histograms, the data clearly showed

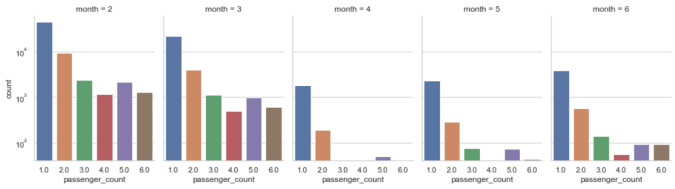


Fig. 6. Passenger Count Form February to June.

that there was no significant gap between February and March, but there was a significant reduction in trip with more than two passengers starting from April.

Overall, data visualization tells us that there is a big gap in passengers' travel patterns before and after the Covid-19, and there are significant changes in both passenger number, taxi demand, and tip amount. Although this study has shown the significant impact of outbreaks on the taxi market, more findings are worth quantifying and confirming using empirical models.

IV. Modelling and Results

In this study, we are interested in building models based on our data. On the one hand, this study wishes to explore a potential regression model for predicting the tip amount by customers and total charge by taxi driver for future study. On the other hand, for investigating the resilience of the taxi industry. It is important to forecast the future trends based on the existing data. The features that the research forecasted were daily total charge by taxi drivers and daily taxi demand.

The method for selecting the regression model that this research used is separated into three parts. While consider using an empirical model to justify the Covid-19 does influence the taxi industry to some degree, we consider using the linear model at first, which all the features were planned to include. Further, a self-written function is designed to process the backward eliminations for excluding the insignificant variables. This step does not only show the relevancy and importance for each feature, but also proceed the feature selection for future use. Furthermore, considering the time cost, not all 9 million data will be used to build models, the research will randomly select 200,000 data. Moreover, the training set and the testing set will be split with ratio 8 to 2.

Firstly, starts from the OLS model, one of the assumptions is the correlation between each parameter is independent, as the figure shows (see Appendix fig.13), general it fits our assumption except their might be a correlation between trip distance trip duration and speed. However, after using PCA, there is no correlation between each feature at all.

From the table above, it includes the descriptive information of OSL model and the significane of feature COVID-19 has been proved (see Appendix fig.9). In the OLS model, the tip amount is the response variable, and the rest of them are explanatory variables. Overall, the result looks fine for this model, R square (0.861) performs pretty well, about 86.1% variance was explained by this model. Without counting the removed variables, there is 45 number of variables remain in

	Machine Learning Models		
Measures	OLS model	Random Forest	
RMSE	0.87**	0.44**	**
F-test	3805	**	**
R suqre	0.861	0.97	
Complexity	$O(v)$	$O(v * nlog(n))$	
	Time Series Models		
Measures	ARMA(3,0)	SARIMAX(1,1,2)	
ADFT test	0.006**	0.08**	**
AIC	85.23**	3248.45**	**
RMSE	2.47**	1923.84**	**

TABLE II
Table of Partial Results From Modeling

the OLS model. As expected, the variable "Cases" (daily confirmed corona cases) presents a very strong significance to the response variable with a p-value less than 0.01. Surprisingly, the variable "Death" (Death under COVID 19) was removed, and some variables related to the time were left. In the case of the evaluation of the model, we chose Root Mean Square Error (RMSE) and variance score as the measures. Within this model, the RMSE from the original dataset gives 0.87, and the variance score is 0.86, which gives same values from the new data-set after feature selection. Even though the RMSE and variance scores are pretty good, the exact same scores from the selected feature model indicates the usefulness of feature selection to improve the accuracy is doubtable.

Within the Random Forest model, the RMSE score (0.44159) and variance score (0.97) are less than the baseline model (OLS model), which gives a better prediction result. However, there is one problem that it is unavoidably to meet the problem of long waiting time and higher computing complexity. Random Forest is an ensemble model of decision trees, and it has the complexity of O $(v * nlog(n))$, which is not friendly to high dimensional data. Another problem incurred due to dilemma between overfitting and maximum accuracy[8].In this case, Random Forest gives higher accuracy with and lower RMSE score, comparing to OLS model.

On the other hand, this research is interested in investigating the resilience of the taxi industry. As the section "Visualization and Analysis" demonstrates, there is a huge drop in taxi demand and daily taxi charges in March and April, and passengers are not willing to give higher tips under the period of Covid-19. However, during May and June, there is a slightly increasing trend in taxi demand and daily charges. It is hard to assert every attribute in our data-set will have an increment. Therefore, this research is going to use the Timer Series model to forecast the future trends in July and August based on the period from 01-02-2020 to 30-06-2020, in terms of daily tip amount and daily demand. Based on our inference from the last section, increasing trends are expected to be observed for both features.

In the aspect of the method of Time Series, the Adjusted Dickey-Fuller Test (ADFT) will be introduced in the first hand

for testing the stationarity of the trend of data. Secondly, data will be transformed or differentiated if necessary. Thirdly, Plots of autocorrelation and partial autocorrelation functions will be used to reference the parameters of the selection model. Lastly, based on the third point, RMES and AIC score will be used as the measurement for the selected model. Our assumption to time series model would be, constant variance, stationarity of data, and normal distributed.

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j}$$

Above is the ARMA model with general (p,q) term, where ϕ is auto-regressive term, and θ is the moving average term.

$$\theta_q(B)\Theta_Q\left(B^s\right)a_t = \phi_p(B)\Phi(P)\left(B^s\right)(1-B)^d\left(1-B^s\right)^D Y_t$$

Above is the SARIMA model with general (p,d,q) term, wher d represents the order of sensonal difference, p and q represent auto-regressive and moving average orders, respectively.

$$AIC = -2Log(Likehood) + 2K$$

Above is the AIC measurement, which k represent the number of parameters, and we will use this to measure the goodness of the time series models.

ARMA Model					
Dep.Variable	total_amount				Model ARMA(3,0)
NO.observation	151				RMSE 2.47**
Method	css-mle				S.D of innovation 0.309**
AIC	85.231		BIC		100.317
Log Likelihood	-37.651		HQIC		91.360
Const					
coef 11.4820	std_err	0.216	p-value	0.00**	
Const ar.L1					
coef 0.9254	std_err	0.810	p-value	0.00**	
Const ar.L1					
coef 0.0252	std_err	0.111	p-value	0.820	
Const ar.L1					
coef -0.0622	std_err	0.081	p-value	0.443	

TABLE III
TABLE OF PARTIAL RESULTS FROM ARMA MODEL

For forecasting the daily total charge by taxi drivers. From the table above, since the log transformation of the data has an ADFT test score below 0.01, it reveals that the data is stationary and no need for difference. According to the ACF and PACF plots (see Appendix fig.12), the ARMA (3, 0) model was selected with an 85.231 AIC score, and 2.47949 RMSE score, which is an ideal model.

Forecasting the total demand will be more complex, since there is a huge drop in total demand in March. Therefore, the ADFT test score appears very large ($p - value = 0.401$), and could not reject the null hypothesis. The data is non-stationary. After the difference and log transformation, we found that the ADFT test value ($P - value = 0.0852$) is still more than 0.05, and the second-order difference does not give a good result either. In this case, the data with a first-order difference plus log-transformation will be used. According to the model description, we found that the AIC score (around 3248) is very large, probably because of too many parameters. Then, through the ACF and PACF plots (see Appendix fig.13), we found that the data had different degrees of seasonal trends, so we locked the SARIMAX model. Because of the complexity of the model, the parameters of AR and MA are hard speculated, so we choose to refine the model by testing all the parameters between the ranges 1 to 4, and the determine of the choice of the model will according to the AIC and RMSE score. The final SARIMAX (1, 1, 2) model was selected with a 1923.84 RMSE score (see Appendix fig.10).

In general, two time series models fit our assumptions motioned above (see Appendix fig 14), and the model for tip amount follows our expectation, whereas the model selection of total demand needs to further investigate, the discussion of potential improvement will be discussed in the next section.

V. DISCUSSION

Combined with our research and findings in visualizations and model building, some inspirations and interpretations will be discussed in this section. Overall, all studies met expectations. The OLS model not only did feature selection, but also confirmed that "Coronavirus confirmed cases" has a significant effect on tip amount. Random forest gives quite good results in prediction, despite the time of model training and the risk of overfitting. When it comes to the time series model, with respect to tip amount, the model also gives a reasonably good score, except that the model with respect to total demand needs further investigation. Although the model basically meets our expectations and the goal of this research, we still firmly believe that there are other better alternative models, which will also be discussed in this section.

From the figure shows below, as the data size increase, the OLS model will finally reach a stable state with around 0.85 accuracy, and Random Forest will achieve almost 0.97 accuracy, which has the suspicion of overfitting. It shows that in the model of OLS, there is no potential risk of overfitting, and the fitting time increases significantly with the number of model instances. On the other hand, in the Random Forest model, the accuracy of the training set has been above the cross-validation set, which represents the overfitting phenomenon. As the suggestion for dealing with overfitting, in regression case, it would be better the look at correlations between variables and think of the sensibility of the model [9]. Another phenomenon in the Random Forest model is that the fitting time increases greatly with the increase of instance, but the training time is also positively correlated with the accuracy rate. In this case, the number of instances might need to be increased in the future. Considering the size of

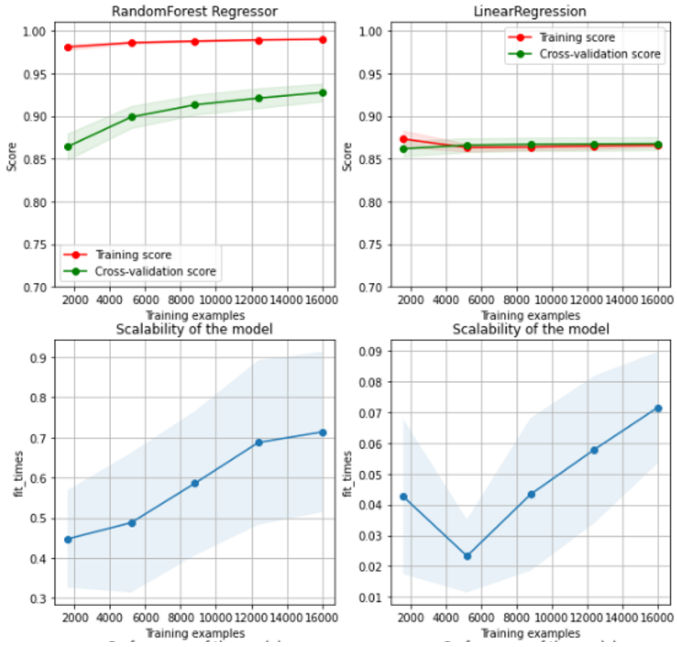


Fig. 7. Descriptive Information of Learning Curve.

the data and the special diagnosis of high dimension, we can choose models that process high dimensional data such as SVM or actively use Principal component analysis (PCA) to reduce the dimension. Moreover, given the million sizes of the data volume, the use of deep learning to process big data is recommended in future studies.

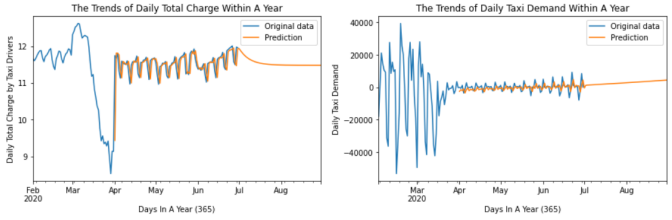


Fig. 8. Daily Taxi Charge and Taxi Demand form February to June.

Discussed in the time series model. There are a lot of very inspiring discoveries that have been released. Firstly, we found that even if people's lives have gradually returned to normal, at least in the July and August, the data do not show that the total amount charged by taxi drivers will increase, but have different degrees of decline. This means that passengers' attitude towards the use of taxi still needs time to return to normal levels. Secondly, when predicting total demand, we found a significant increase in total demand, which represents a gradual return to previous levels in the frequency of people's travels. Although demand is shown to improve, the tip amount is declining, and we may need to study this interesting phenomenon more deeply. Thirdly, when discus's about the improvement of SARIMAX model, it might need to reconsider the choice of model and redesign the method for data processing. To potentially improve the performance of SARIMA model, it wroth to implement the genetic algorithm (GA) optimization algorithm, especially in the case of traffic flow, since it almost has equivalent accuracy with artificial neural networks (ANNS) [10]. Lastly, in addition to the research about taxi charge and demand, future researches are also encouraged to pay more attention on other aspects of prediction and analysis, such as trip duration and distance.

VI. Conclusion

In this research, we studied the impact of the Covid-19 on the New York taxi market from the perspective of passengers as well as future trends. The research found, people's travel modes in 2020 can be divided into three categories, pre-epidemic, epidemic, and epidemic recovery stage. After the outbreak, people prefer to go out on weekends rather on weekdays, and there is a significant decline in the taxi demand before and after the epidemic. Secondly, the distance traveled by people, the level of tips given, and the number of passengers also have decreased in various degrees, with people more inclined to travel with of two or less after the COVID-19. Moreover, after fitting the machine learning and time series models, the OLS model confirmed the significance of the COVID-19 to the taxi market, and the Random Forest model was found to have a high prediction rate for tip giving. It is worth noting that the analysis of time series models tells us that the daily taxi demand will increase from June, but daily charge by taxi drivers is still decreasing and quickly stabilizing. Lastly, there are some recommendations for the taxi companies and taxi divers. For taxi companies, they could increase the taxi supply on weekend, and guradully increase the taxi supply in next few months, but might not expect the increase in fare amount. For taxi drivers, if they wish to earn higher tips, they could spend more time on running the taxi after 10 am on weekend. However, as discussed before, the total avenger level for the taxi industry has dropped, and hardly return to the normal level, in a short time.

This research not only shows data acquisition, data processing, but also demonstrates models and analysis based on existing data. In big data, data cleaning is a very important part. In this study, the DBSCSN clustering algorithm is used to detect outliers, and data cleaning is also continuously carried out in the process of data integration. Although the Random Forest model gives a very low RMSE score for our data-set, we still look forward to future studies that could try to use other machine learning models given the time and the possibility of overfitting. Secondly, considering that for big data processing, this study did not use the Spark distributed system as well as deep learning. Whereas we believe that it is worthy to try the deeping learning model combining Apache Spark and the advanced machine learning architecture, such as deep multi-layer perceptron (MLP), it is proved to have advantages over tradition big data analytic methods with lesser computational complexity and with significantly higher accuracy[11]. On the other hand, although the time series model gives a positive prediction of future demand for taxi industry in New York City, the model shows a downward trend in the prediction of

daily charges by taxi drivers, which may represent that even if demand can return to normal quickly, the whole industry may still need more time to reach a new steady state.

REFERENCES

1 Tang, Y., "Big data analytics of taxi operations in new york city," *American Journal of Operations Research*, vol. 09, no. 04, pp. 192–199, 2019.

2 Government, N., "What you need to know about COVID-19," Available at https://www1.nyc.gov/site/coronavirus/index.page (2020/10/18).

3 Hongyu Zhenga, Y. M. N. ., "The fall and rise of the taxi industry in the COVID-19."

4 Taxi, N. Y. C. and Data, L. S. T. R., "Yellow Taxi Trip Records," Available at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (2020/10/18).

5 NYC Health Department, "COVID-19: Data," Available at https://www1.nyc.gov/site/doh/covid/covid-19-data-boroughs.page (2020/10/18).

6 The COVID Tracking Project, "COVID-19: Data," Available at https://covidtracking.com/ (2020/10/18).

7 Heinrich Jiang, J. , "Faster dbscan via subsampled similarity queries," Available at https://arxiv.org/abs/2006.06743 (2020/06/11).

8 Fawagreh, K., Gaber, M. M., and Elyan, E., "Random forests: from early developments to recent advancements," *Systems Science Control Engineering*, vol. 2, no. 1, pp. 602–609, 2014.

9 Hawkins*, D. M., "The problem of overfitting," *ChemInform*, vol. 35, no. 19, 2004.

10 Luo, X., Niu, L., and Zhang, S., "An Algorithm for Traffic Flow Prediction Based on Improved SARIMA and GA," *KSCE Journal of Civil Engineering*, vol. 22, no. 10, pp. 4107–4115, 2018.

11 Gupta, A., Thakur, H. K., Shrivastava, R., Kumar, P., and Nag, S., "Mobile big data analytics using deep learning and apache spark," *IEEE Network*, vol. 30, no. 3, pp. 22–29, 2016.

APPENDIX

OLS Regression Results

Dep. Variable:	tip_amount	R-squared:	0.861
Model:	OLS	Adj. R-squared:	0.861
Method:	Least Squares	F-statistic:	2.812e+04
Date:	Wed, 07 Oct 2020	Prob (F-statistic):	0.00
Time:	23:16:39	Log-Likelihood:	-2.5693e+05
No. Observations:	200000	AIC:	5.139e+05
Df Residuals:	199955	BIC:	5.144e+05
Df Model:	44		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	-1.2836	0.011	-113.846	0.000	-1.306	-1.261
Cases	8.657e-05	1.8e-05	4.822	0.000	5.14e-05	0.000
trip_distance	-0.0950	0.002	-42.966	0.000	-0.099	-0.091
fare_amount	-0.5122	0.001	-438.753	0.000	-0.515	-0.510
Time_duration	0.0144	0.000	31.409	0.000	0.014	0.015
total_amount	0.5413	0.001	593.394	0.000	0.539	0.543
Speed	0.0149	0.001	21.979	0.000	0.014	0.016
vendor_id_2.0	0.0385	0.004	9.123	0.000	0.030	0.047
month_4	0.0827	0.016	5.269	0.000	0.052	0.113
month_5	0.1237	0.012	10.157	0.000	0.100	0.148
month_6	0.0737	0.009	7.906	0.000	0.055	0.092
dayofweek_2	-0.0130	0.006	-2.208	0.027	-0.025	-0.001
dayofweek_3	-0.0133	0.006	-2.283	0.022	-0.025	-0.002
dayofweek_5	0.1236	0.006	20.534	0.000	0.112	0.135
dayofweek_6	0.1283	0.007	19.488	0.000	0.115	0.141
hour_1	-0.0624	0.015	-4.030	0.000	-0.093	-0.032
hour_2	-0.0820	0.018	-4.444	0.000	-0.118	-0.046
hour_3	-0.1036	0.021	-4.855	0.000	-0.145	-0.062
hour_4	-0.1740	0.024	-7.190	0.000	-0.221	-0.127
hour_5	-0.0713	0.022	-3.239	0.001	-0.114	-0.028
hour_6	0.1205	0.014	8.430	0.000	0.092	0.148
hour_7	0.1410	0.011	12.616	0.000	0.119	0.163
hour_8	0.1426	0.010	14.019	0.000	0.123	0.163
hour_9	0.1586	0.010	15.425	0.000	0.138	0.179
hour_10	0.1424	0.010	13.805	0.000	0.122	0.163
hour_11	0.1497	0.010	14.763	0.000	0.130	0.170
hour_12	0.1560	0.010	15.666	0.000	0.137	0.176
hour_13	0.1452	0.010	14.927	0.000	0.126	0.164
hour_14	0.1550	0.010	16.236	0.000	0.136	0.174
hour_15	0.1278	0.010	13.404	0.000	0.109	0.147
hour_16	-0.2186	0.010	-22.615	0.000	-0.238	-0.200
hour_17	-0.2248	0.009	-24.175	0.000	-0.243	-0.207
hour_18	-0.2212	0.009	-24.522	0.000	-0.239	-0.203
hour_19	-0.1979	0.009	-21.149	0.000	-0.216	-0.180
hour_20	-0.0524	0.010	-5.343	0.000	-0.072	-0.033
borough_Brooklyn	0.7140	0.020	35.062	0.000	0.674	0.754
borough_EWR	2.1481	0.311	6.896	0.000	1.538	2.759
borough_Queens	0.5259	0.010	50.231	0.000	0.505	0.546
payment_2.0	-1.1430	0.005	-213.640	0.000	-1.154	-1.133
payment_3.0	-0.9962	0.031	-31.819	0.000	-1.058	-0.935
payment_4.0	-1.0429	0.052	-20.178	0.000	-1.144	-0.942
passenger_2.0	-0.0258	0.006	-4.591	0.000	-0.037	-0.015
passenger_3.0	-0.0375	0.010	-3.622	0.000	-0.058	-0.017
passenger_4.0	-0.0758	0.015	-4.978	0.000	-0.106	-0.046
passenger_7.0	4.6765	0.875	5.347	0.000	2.962	6.391

Omnibus:	67953.311	Durbin-Watson:	2.017
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3104866.514
Skew:	-0.903	Prob(JB):	0.00
Kurtosis:	22.218	Cond. No.	6.02e+04

Fig. 9. Descriptive Information of OLS Model.

SARIMAX Results

Dep. Variable:	demand_diff_1	No. Observations:	150
Model:	SARIMAX(1, 1, 2)	Log Likelihood	-1618.229
Date:	Wed, 07 Oct 2020	AIC	3248.457
Time:	22:37:26	BIC	3266.481
Sample:	02-02-2020 - 06-30-2020	HQIC	3255.780
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	7.7006	132.857	0.058	0.954	-252.694	268.095
drift	0.2246	1.950	0.115	0.908	-3.598	4.047
ar.L1	0.0889	0.103	0.866	0.386	-0.112	0.290
ma.L1	-0.6326	0.109	-5.823	0.000	-0.845	-0.420
ma.L2	-0.3674	0.104	-3.524	0.000	-0.572	-0.163
sigma2	1.113e+08	7.01e-10	1.59e+17	0.000	1.11e+08	1.11e+08

Ljung-Box (Q):	227.45	Jarque-Bera (JB):	233.15
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.03	Skew:	-0.64
Prob(H) (two-sided):	0.00	Kurtosis:	8.99

Fig. 10. Descriptive Information of SARIMAX Model.

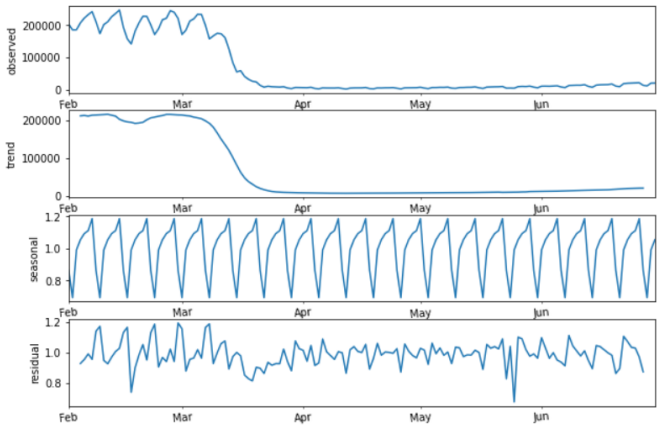


Fig. 11. Season Decomposed Plot of Daily Taxi Demand.

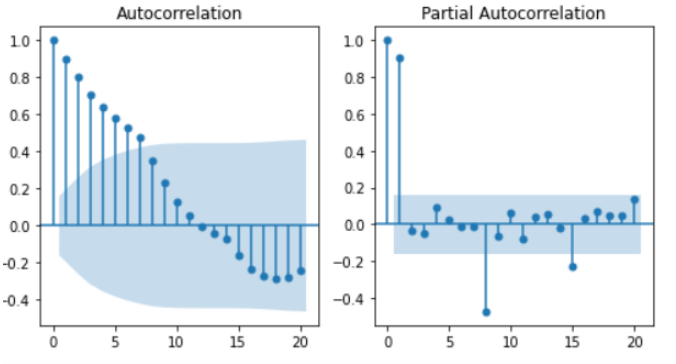


Fig. 12. The ACF And PACF Plot For ARMA Model.

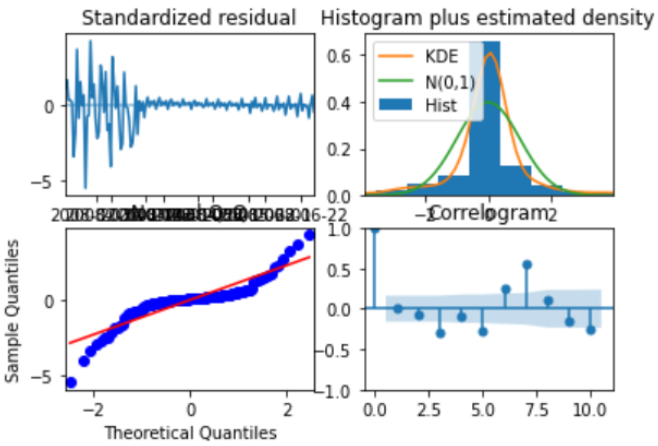


Fig. 14. Diagnostic Plot For Time Series Models

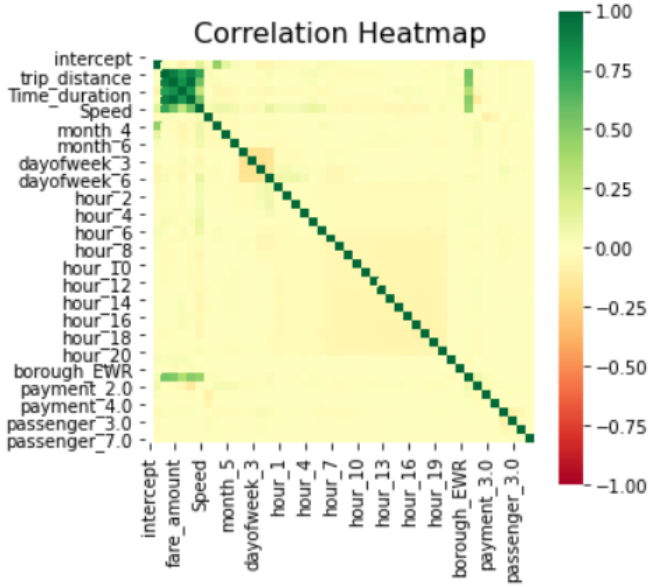


Fig. 13. Correlation Between Each Feature.