Modelling and Analysis of The Impact of COVID-19 on Taxi Industry in New York City

Mieshan Liu
University of Melbourne
Student ID:850076

Muhammad Danish Naseem
University of Melbourne
Student ID:878924

Yichuan Zhang
University of Melbourne
Student ID:904722

Yuchen Xie
University of Melbourne
Student ID:904926

Abstract

COVID-19 is probably the largest challenge in the year 2020. This research project aims to analyze the current and future situations of New York City (NYC) taxi industry under COVID-19 period. Besides pure analysis, this research project also provides suggestions for the potential stakeholders of the taxi industry in combination of government policies. In our search, by using empirical models such as Random Forest, XGboost and Time Series, we investigated questions, such as how does COVID-19 impact on the trip distance and taxi demand, and what will the taxi industry be like in the long term. Inside of the research, DBSCAN and Isolation Forest were used to proceed data pre-processing. Furthermore, Multivariate visualizations based on models and processed data-set showed: (1) There is a significant change in people's travel modes before and after COVID-19 period, especially passenger count and travel distance; (2) there is a significant drop in trip distance from March, and it will hard go back to the original level before COVID-19 period; (3) for taxi demand, there is a significant drop after March, whereas taxi demand is gradually increasing since June.

1 Introduction

1.1 Background

New York City is an international metropolis with the largest population in the US which is also one of the centres of the world in economy, commerce, finance, media, and politics. The COVID-19 is a coronavirus disease which spread suddenly around the globe at the end of 2019. Approximately 170,000 people were tested to have a positive result by the end of March of 2020 including around 7,000 deaths in about 150 countries [Alipio, 2020]. And on March 11, 2020, the COVID-19 was outbreak a pandemic declared by the World Health Organization. The taxi market was influenced significantly by the COVID-19 pandemic in terms of taxi demand and other aspects [Hongyu Zhenga, 2019]. However, in addition to the significant influence, the analysis of the impact of COVID-19 to the taxi industry should not be separate from the analysis of policies. Based on the lockdown policy and reopening strategy policy published by the US government, passengers' demand and the amount of total payment will be considered to analyze the trends of the taxi industry [see Novel Coronavirus, 2020]. In this research project, the aim is to find the impact of the unexpected situations on these kinds of attributes of the taxi industry and provide a couple of suggestions to drivers and companies to promote the recovery from the outbreak.

Preprint. Under review.

1.2 Literature review

So far, there is not much research focus on analyzing the impact of COVID-19 on the taxi industry. To the best of my knowledge, Hongyu'group in Northwestern University(USA) were the first to use spatio-temporal analysis to analyze the impact of COVID-19 on taxi industry, and put forward theoretically feasible operational strategies for the current situation of the taxi market in combination with government policies [Hongyu Zhenga, 2019]. It mainly pointed out taxi demand reduced more than 85% in Shenzhen, and it barely recovered to the original level even after reopening of the city. Besides Hongyu's group, Haleh Ale-Ahmad [2020] analysed the Chicago city and the research showed the extreme lockdown policy will reduce the taxi demand by 95%, and there is a further reduction in the number of operating taxis. Lastly, to the reopening stage of the city, Yue Hu [2020] mainly evaluated whether changing in transportation mode would deteriorate the traffic environment in the case of reopening in the late COVID-19 period by using the BPR model.

2 Data Specification and Preparation

2.1 Data Characteristics

Data-sets	Source	Sizes	No.Instance	NO.Features	Missing Values
Taxi Trip 2020-02	TCL	584 MB	6299354 rows	18	48834 rows
Taxi Trip 2020-03	TCL	278 MB	3007292 rows	18	37487 rows
Taxi Trip 2020-04	TCL	21.7 MB	237993 rows	18	19513 rows
Taxi Trip 2020-05	TCL	31.6 MB	348371 rows	18	58891 rows
Taxi Trip 2020-06	TCL	50.3 MB	549760 rows	18	50717 rows
NYC COVID data	NYCH	30.4KB	212 rows	43	0 rows

Table 1: Data Characteristic For Each Data-Set

The data-sets used in this research come from three resources. For the taxi records, Yellow taxi trip records were used, and it comes from Limousine Service Trip Record Data [Taxi and Service, 2020]. For the COVID data, it mainly comes from New York City Health Department [NYC Health Department, 2020] and The COVID Track Project [The COVID Tracking Project, 2020]. The major reason that we chose the taxi records in New York City is because of the availability of data. Taxi and Limousine Service Trip Record Data (TLC) website records the trip records from 2009 to 2020, which also contains Yellow taxi records, Green taxi records, and For-Hire Vehicle Trip Records. For the Yellow taxi trip data that we selected ranged from February 2020 to June 2020. We chose Yellow taxi data for the following reasons: Firstly, the yellow taxis form majority of New York City's taxi market. It contains a much higher number of records than the green taxi and FHV taxi. Note that we are focusing on the taxi demand, therefore more data will be useful for our analysis afterwards, since it contains more information of the NYC taxi market. Secondly,COVID-19 in New York City began in March and continues to today. In this case, we chose the data from March to June (TLC only released data up to June) and added the data from February as a normal month before COVID period,which allowed us to make a comparison.

2.2 Data Preparation

For the preprocessing section, there are few situations that we need to consider. Firstly, we assume that the covid-19 data is cleaned data which does not need any extra preprocessing steps but the raw data of New York City taxi records has not been fully preprocessed and cleaned. The taxi records data-set contains many inaccurate and illogical data points, therefore data preprocessing was essential and necessary. The preprocessing steps are therefore data cleansing, outlier detection, and feature engineering.

For the data cleansing, the first step was to remove the inaccurate data and missing values, and number of missing values is shown in Table 1. The inaccurate data defined by the distribution of data and verified with reliable sources such as the Trip Record User Guide [New York City Taxi and Limousine Service, 2019]. For example, all the negative values for the attributes like "tip amount", "fare amount", "trip distance" were removed, because these values should not be negative. We also removed some extremely large values by plotting the boxplot graph and then verified with size checking. Taking the

attribute fare amount as an example, we noticed that a little points go over 1000$ on the boxplot and obviously taxi trips which cost over 1000$ are unusual. Therefore, we marked these points as outliers and removed them. Next, because of the large data-set with more than 10 million sizes, removal of those values will not influence the data, but also help to generalize a more accurate model. Then we removed all the rows which contained missing values, table above shows the number of rows that were removed in each dataset.

For the outlier detection, two methods (DBSCAN and Isolation Forest) were used in this research. To make sure there are no potential points that break the accuracy of our model, Density-based spatial clustering of applications with noise (DBSCAN) method was mainly used to detect the outliers for the datasets. DBSCAN is an unsupervised, clustering method used to detect outliers which have nonparametric distributions in many dimensions. 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 was 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 [Heinrich Jiang, 2020]. In this case, the advantage to using this method is that outliers can be excluded based on all continuous attributes, instead of a specific attribute. Comparison between the two methods will be discussed later.

Result: Remove outliers for 9 million data
Recursion DBSCAN Algorithm(df1, df2);
while *still data in the df1* **do**
    if *length of df1 > 100000* **then**
        sampling 100000 date from df1;
        indentify outliers by using DBSCAN;
        exclude outliers in df2;
        drop the 100000 sampling df from df1;
        recall the Recursion DBSCAN Algorithm
    else
        sampling df = df1;
        indentify outliers by using DBSCAN;
        exclude outliers in df2;
        break
    end
    return df2
end

Algorithm 1: How to write algorithms

Since the total data-set had a size of more than 9 million data points, taking into account the high complexity and long prepocessing time of DBSCAN algorithm, a self written recursion function is designed to exclude outliers by sampling from the whole data-set. The algorithm is shown above.

The last step was to check the data size, the original data which consists of data from February to June has around 10 million rows, after the pre-processing steps the data remains around 9.8 million rows and this was an acceptable result.

3 Visualization and Analysis

Visualization of the data assists in the observation of trends or correlations between features.To better understand how COVID-19 impacted the taxi industry, we may want to know whether there is any difference of taxi demand and trip distance before and after the Covid-19 period.

From the figure below. There exists a negative relationship between the number of patients and the demand for taxis, in other words as the number of patients increase at the beginning of COVID, the demand drops significantly. But when the situation became better, there was no obvious increase in demand. Similarly according to Hongyu Zhenga [2019], the findings in Shenzhen also show that many residents prefer to replace taxis by using personal vehicles.

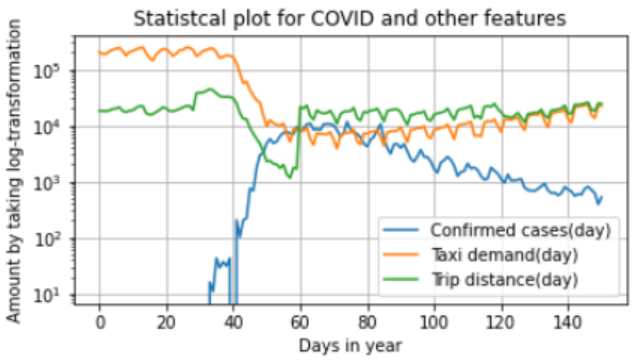


Figure 1: The Time Trend of The Data in COVID and Taxi Industry

4 Modelling and Results

In this research project, modelling will be helpful in predicting the specific attribute based on prepared data. To some extent, the goal of this research project is to explore a potential regression model for predicting the main indicator which can represent changes in the taxi industry, such as in trip distance and taxi demand. Furthermore, besides the prediction, forecasting the future trend based on past data will be useful to understand future situations in the next few months.

4.1 Machine learning Models

The Random Forests and XGBOOST are utilized for prediction of trip distance, fare amount, and tip amount. The goal is to identify the best possible prediction model for the given attributes.

The Random Forest Model, utilizes regression trees as base learners. The regression algorithm of Random Forests is based on bagging. In bagging, a collection of learners is used wherein random samples from the training data-sets are drawn with replacement and each learner is trained on these random samples [Peter, 2012]. The final prediction of the collective of learners is decided by averaging of results. Bagging has been known to reduce variance in the final model and also helps to reduce overfitting [Yiu, 2020]. For our analysis two types of Random Forests Models are used, such that one model utilizes a single tree and another model utilizes ten trees.

Like Random Forests, the XGBOOST model is an ensemble learning method.In this manner, for regression base learners are regression trees but for boosting, the trees are built sequentially and each tree attempts to reduce the error of the previous tree [Peter, 2012]. Initial trees built have high bias which with each iteration is lowered. Thus, any bias in prediction is minimized [Brownlee, 2020].

Initially, to assess the best training and testing split each model for prediction of trip distance, fare amount and tip amount is run. Root Mean Squared Error (RMSE) is used for evaluation and to see whether there is any overfitting in training and testing sets, in addition to measure the model, RMSE from different models will be compared with each other.

$$RMSE_{fO} = [\sum_{i=1}^{N}(z_{fi} - z_{oi})^2/N]^{1/2}$$

RMSE is an approach to measure how far from the regression line data points are [Chai and Draxler, 2014]. In this case, $\sum$ is summation, $(z_{fi} - z_{oi})$ is total difference, and N is sample size.

In terms of the results for XGBOOST model, training and testing are quite similar therefore there is little indication of any overfitting. There is a slight jump in RMSE for both training and testing as training set size increases. However, for greater validity of results, it is better to have a large training wherein results based on small training dataset are not suitable for analysis.

The results are presented in the tables below wherein it is clear that training and testing RMSE is close enough that there is little sign of overfitting any more. Furthermore, there is a similar performance for each model of each attribute such that there are only small differences seen. The XGBOOST Model has a mean RMSE of 0.297 and 0.085 for prediction of fare amount and trip distance respectively, which is lower than that of the other models. However, for prediction of tip amount, Random Forest

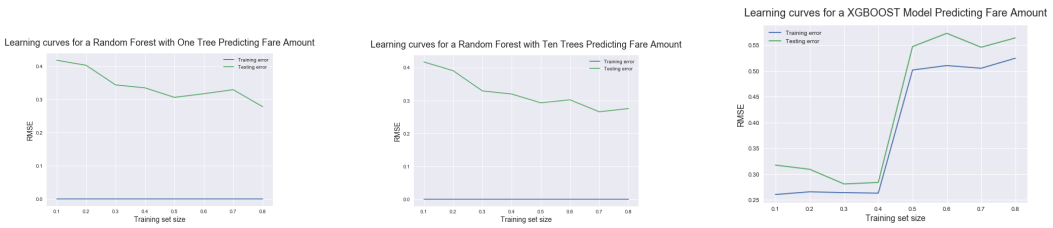


Figure 2: Learning Curves for Fare Amount Prediction Models

Table 2: Five Fold Cross Validation Results

Performance of Each Model			
Machine Learning Model	Tip Amount RMSE	Fare Amount RMSE	Trip Distance RMSE
XGBOOST	0.183	0.297	0.085
Random Forest Single Tree	0.15	0.35	0.095
Random Forest Ten Trees	0.147	0.316	0.108

has a lower RMSE at 0.147. Therefore depending on which attribute is being predicted, a particular model may be more suitable than others. In general, as any difference seen is marginal, all three models perform well in prediction of the three given attributes.

4.2 Time Series Models

Moreover, this research project intends to investigate what will the taxi industry be like in the future. For securing our prediction,Time Series models are used to predict the main indicators from taxi industry, such as taxi demand and trip distance. As the visualization part from previous assignment showe, we found there was a significant drop in taxi demand and travel distance after March, and a slight increase occurred during May and June. However, it will be difficult to make sure every main indicator in taxi industry will increase. Hence, Time Series models will be used to forecast the trend in following months based on the past data from 01-02-2020 to 30-06-2020. According to our previous assignment, an increasing trends are expected to find for the attributes, taxi demand and trip distance.

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

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.

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

Akaike information criterion (AIC) measures goodness of fit of the model according to number of parameters and in-sample predictor error [Bozdogan, 1987], which k represent the number of parameters, and we will use this to measure the goodness of the time series models.

In terms of the methodology of Time Series Modelling, there are four steps. Firstly, the Adjusted Dickey-Fuller Test (ADFT) will be used to test the stationarity of data. Secondly, data manipulation and data transformation will be used if necessary, such as log-transformation and difference of data. Thirdly, autocorrelation and partial autocorrelation function plots will be introduced to select the parameters of the Time Series model. Lastly, to measure the goodness of model, RMES and AIC

score are the expected measurements. Our assumptions of the Time Series model will be data is time correlated only, and other interrelation between data should not be considered.

Performance of Time Series Models				
Time Series Model	Dep.Variable	RMSE	AIC	Log Likelihood
ARMA(1,0)	trip$_distance$	0.175**	58.293	-26.147
SARIMA (1,1,2)	taxi$_demand$	1923.84	3248.45**	-1618.229

Table 3: Results of Time Series Models

For forecasting the daily trip distance in each day. From the table above, since the log transformation of the data has an ADFT test score around 0.011, 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 (1, 0) model was selected with an 58.293 AIC score, and 0.1750 RMSE score, which is an ideal model. Forecasting the taxi demand will be more complex, since there is a huge drop in daily taxi 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, 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 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, we found that the data had different degrees of seasonal trends, so we locked the SARIMA 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.

5 Discussion

We have analyzed the change in the taxi market in NYC during the coronavirus pandemic by using three different models. In general, all model fits our expectation, but there are some limitations as well. In this case, some inspirations and interpretations for the models, and suggestions to potential stakeholders will be briefly summarized below.

For the data processing section. Although the DBSCAN method is a useful method to detect outliers, we also found that this method required a lot of computing power and time to process the dataset. For large datasets with 9 million sizes, it is almost impossible to implement the DBSCAN method to process all the data at once. In order to solve this problem, we found that the isolation forest method could be another alternative method. From our results, the Isolation forest method can be done within less processing time and also requires less computing power to achieve a similar result with the DBSCAN method. More importantly, the isolation forest could process all data at once. However, the disadvantage of this Isolation Forest method is that it will reduce the data size by 10%. In this research, more information was required, therefore DBSCAN was used, instead of Isolation Forest. However, we encourage future research to try other methods for outlier detection.

To better understand the result of machine learning models, this research used learning curve for analyzing the performance, and hyperparameters tuning were used for adjusting the problem of overfitting. As the figure shows above, both types of Random Forest models show a high degree of overfitting such that training RMSE is much lower than testing RMSE for all training and testing splits. To reduce overfitting and thereby get training and testing RMSE that are close to each other, hyperparameter tuning is conducted on both random forest models. A similar performance is seen for prediction of trip distance and tip amount as well.

The three main hyperparameters tuned wherein a substantial effect is seen are ‘max_depth’, which is the maximum number of levels in each decision tree, ‘min_samples_leaf’ which is the minimum number of data points allowed in each leaf node of a tree and ‘min_samples_split which is the minimum of data points in a node before the node is split. The optimum values for the given

hyperparameters are “min_samples_split” with 10, “min_samples_leaf” to be 2, and “max_depth” is set at 30.

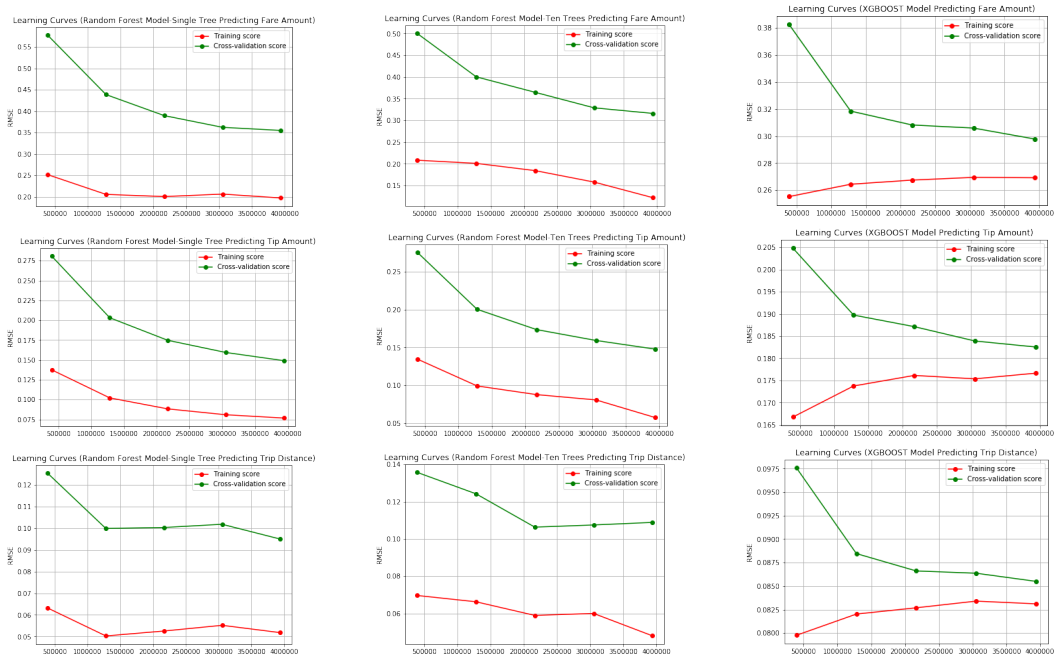
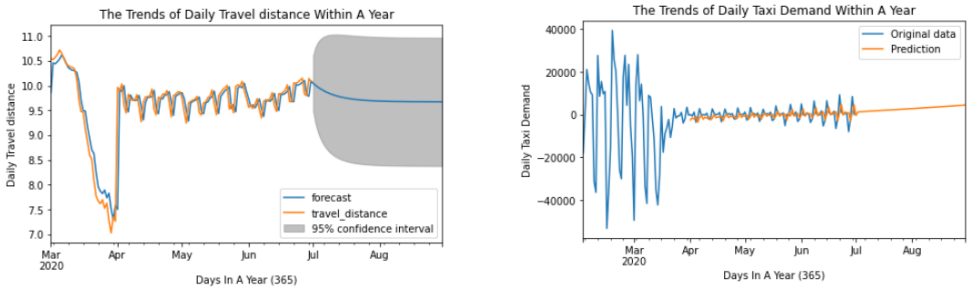


Figure 3: Learning Curves for Each Model

After hyperparameter tuning, the performance of each model is evaluated. This is done by random sampling of half of the given data after which five-fold cross-validation is conducted. The results are presented in the tables below wherein it is clear that training and testing RMSE is close enough that there is little sign of overfitting any more. Furthermore, there is a similar performance for each model of each attribute such that there are only small differences seen. The below figures present the learning curve for prediction of fare amount, trip distance and tip amount Similar performance is seen for prediction of each attributes. Specifically, although there is still a gap in between cross validation error and training error, the ponderance of overfitting has decreased.



(a) Daily Travel Distance (b) Daily Taxi Demand

Figure 4: Daily Travel Distance And Daily Taxi Demand

For the Time Series model, many interesting findings and insightful results were observed from above figures. At first, although the city is reopening right now and the restriction policy is softer than before, at least in the following months, the model did not predict that the travel distance will increase, but there would be a slight decline instead. This indicates that the attitude from passengers to take taxis will not go back to the normal level. In other words, people might still prefer to stay near the home and not willing to go far away. Secondly, taxi demand is complicated to forecast, in the

case, there is a slight increase in tax demand, which probably means reopening of the city leads to taxi demand gradually returning to the original levels before COVID period as a minority of people have to work by taking taxis. However, we still need a further investigation on the case that demand is shown to increase, but trip distance is declining. Thirdly, to potentially improve the performance of SARIMA model, it might be prudent to reconsider the selection of model, and more efficient data processing step might be necessary to exclude more noise. For the method to improve the model based on previous research, implementation of the genetic algorithm (GA) optimization algorithm is worth to try, especially to deal with the problem of traffic flow, as it gives a higher accuracy based on artificial neural networks (ANNS) [Luo et al., 2018]. Lastly, besides doing the research about trip distance and demand only, we recommend future research to investigate other indicators within taxi industry, such as trip duration and tipping rate.

Moreover, suggestions to the potential stakeholders should go alone without understanding the policies from government. Similar to the findings in Shenzhen, two research projects indicate that the trend of demand is highly correlated with the published policy [Hongyu Zhenga, 2019]. As pointed out in the paper, Shenzhen has had a strict lockdown for a couple of weeks and reopened since the mid-February which lead to a huge decrease in trip distance for all trips. New York's governor, Andrew Cuomo, published the lockdown rules on 20th March to control the spread of the virus by limiting the residences' outgoing. The policy required all non-essential businesses to be closed and residents to stay home if their work is non-essential until the coronavirus has been well controlled in the state. Therefore, in this period, In our project for New York City, we found that in March and April, the trip distance has dramatically decreased.

In some of the main cities in China, Chris states that the ridership restored at a rate around 50% in the first 1 and 2 month of reopen [Roberts, 2020]. And according to Hongyu's team's finding, after reopen, the demand recovers the fastest when most people rush to work [Hongyu Zhenga, 2019]. Therefore, we need to consider how to promote the recovery of the taxi industry in NYC.

Taxi drivers suffering financially were given unemployment benefits and immediate cash subsidies by the New York government [Roberts, 2020]. This helps a part of taxi drivers to maintain their life in this period so that after the city reopens, they can still go back to work rather than just change jobs to have a stable income. However, this method only protected the taxi drivers themselves, it did not help to increase the demand of taxi to a normal level which means that it does not promote the recovery of the whole industry.

Therefore, as the suggestions from this research, the government should reduce the fare paid by passengers by preparing a fund to pay part of the total amount in each trip. We can believe that as the cost to take the taxi is offset by government's fund, the consumer confidence will be promoted. More professional suggestions are encouraged to be investigated in the future.

6 Conclusion

In conclusion, this research project has analyzed some of the main attributes from taxi records datasets and used 3 empirical models to fit the data. According to the trends, the impact of COVID-19 can be basically divided into 3 stages: pre-epidemic, epidemic, and the recovery stage. In the first two stages, there exists a sharp decline in the demand which might be resulted by the rise of patients, and after the most serious time, the demand will recover slowly. We can indicate from the positive prediction provided by the time series model that in the third stages, the whole industry began to recover at a very low speed. It is also reasonable to believe that the earnings of taxi drivers will not return to a standard level in a short time. Moreover, for the stimulus policies, besides the funding from the New York government, the government also needs to adjust the balance in the supply and demand within the taxi industry. In this case, the main indicators such as taxi demand will be more stable in a relatively long time.

Ultimately, the contribution of this research is limited by the selection of data processing methods and options for modelling, we strongly encourage the future study to investigate other attributes within the taxi industry, and design a more valid strategy that might potentially increase the efficiency of the taxi industry under unexpected situations.

References

Mark M. Alipio. 2019-ncov scare: Situation report, role of healthcare professionals and clinical findings. Technical report, 2020.

Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

Jason Brownlee. A gentle introduction to xgboost for applied machine learning. Available at https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/, 2020.

Tianfeng Chai and Roland R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? *GMDD*, 7(1):1525–1534, 2014.

Hani Mahmassani Haleh Ale-Ahmad. Impact of covid-19 on taxi operation in chicago. Available at https://www.transportation.northwestern.edu/news-events/articles/2020/taxi-operations-during-covid-19.html, 2020.

Jakub Łącki Heinrich Jiang, Jennifer Jang. Faster dbscan via subsampled similarity queries. Available at https://arxiv.org/abs/2006.06743 (2020/06/11), 2020.

Yu (Marco) Niea * Hongyu Zhenga, Kenan Zhanga. The fall and rise of the taxi industry in the COVID-19 pandemic: A case study. *SSRN Electronic Journal*, (3674241), 2019.

Xianglong Luo, Liyao Niu, and Shengrui Zhang. An Algorithm for Traffic Flow Prediction Based on Improved SARIMA and GA. *KSCE Journal of Civil Engineering*, 22(10):4107–4115, 2018.

New York City Taxi and Limousine Service. TLC Trip Records User Guide. Available at https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf, 2019.

Novel Coronavirus. New york state on pause. Available at https://coronavirus.health.ny.gov/new-york-state-pause, 2020.

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

Bühlmann Peter. Bagging, boosting and ensemble methods. *Handbook of Computational Statistics*, 985(1022), 2012.

Chris Roberts. Without government rescue, new york city yellow cabs could soon be history. Available at https://observer.com/2020/03/coronavirus-new-york-city-yellow-taxi-cab-need-government-stimulus/, 2020.

New York City Taxi and Limousine Service. Yellow Taxi Trip Records. Available at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (2020/10/18), 2020.

The COVID Tracking Project. COVID-19: Data. Available at https://covidtracking.com/, 2020.

Tony Yiu. Understanding random forest. Available at https://towardsdatascience.com/understanding-random-forest-58381e0602d2, 2020.

Samitha Samaranayake Dan Work Yue Hu, William Barbour. Impacts of covid-19 mode shift on road traffic. *arXiv*, 5(01610), 2020.