Analyzing and Modeling Covid-Pandemic's Impact on Taxi Demand in New York City

Term Paper

CVEN 672

Acharya, Kushal

Ajgaonkar, Chinmay

Basit, Farooq

Jagatap, Rutuja

Koirala, Pranik

(Note: *The names are arranged in alphabetical order)*

Dec 2021

EXECUTIVE SUMMARY

Need of the Study

The covid-19 pandemic has been a disruptive event for many sectors. Similarly, the transportation industry has been hardly hit by the pandemic. The analysis of the disruption created by the pandemic in the transportation industry and understanding the recovery process is of vital importance for decision makers, operators as well as individuals. To better understand the response and behavioral pattern within varying socio-economic background and demographic, this paper investigates potential data resources and analysis techniques. The need to better understand the behavior of riders and the most important impacting factors to the transportation industry in a pandemic is important to understand future unique events and unprecedented situations like the covid-19 pandemic. The work can also be extended in predicting other transportation characteristics like for forecasting demand of transit systems or any other mode of transportation, which will help the transportation engineers and policy makers to make better investment decisions.

Methodology

This study puts the taxi industry of New York City under the microscope during the period of first and second wave of covid-19 and predicts the taxi demand in the coming third wave with some level of accuracy. In doing so this study also looks into the use of different data sources to explain the unique characteristics of the general taxi rider in an unprecedented event i.e., the covid-19 pandemic. To achieve this insight the study makes use of ensemble learning and deep learning technology combined with standard statistical analysis tools. More specifically, XGBoost and LSTM were used to predict the trend of the trip demand.

Different sources of data were explored to better understand the unique characteristic of taxi demand in a pandemic. The traditional literature dealing with forecasting trip demand were of little or no use to study this unique situation which led us to find novel sources of data such as social media's link to the perceived-sense-of-safety among people and its effect on trip demand with respect to the pandemic. This study also incorporates the qualitative information of movement restrictive policy of the city by transforming the information into qualitative data used for the analysis.

Findings

Insights from the results of ensemble learning methods and deep learning technology combined with standard statistical analysis tools tells us that with a rich data source any turbulent period can be explained. It was found that ensemble learning method gives a better understanding of the unique situation. The best performing model was XGBoost with the MAPE of 13.86 percent, while the second-best performing model was LSTM with 17 percent MAPE.

Table of Contents

1 ABSTRACT

The covid-19 pandemic has been a disruptive event for many sectors. Similarly, the transportation industry has been hardly hit by the pandemic. This study puts the taxi industry of New York City under the microscope during the period of first and second wave of covid-19 and predicts the taxi demand in the coming third wave with some level of accuracy. Different sources of data were explored to better understand the unique characteristic of taxi demand in a pandemic. The traditional literature dealing with forecasting trip demand were of little or no use to study this unique situation which led us to find novel sources of data such as social media's link to the perceived-sense-of-safety among people and its effect on trip demand with respect to the pandemic. This study also incorporates the qualitative information of movement restrictive policy of the city by transforming the information into qualitative data used for the analysis. To achieve this insight the study makes use of ensemble learning and deep learning technology combined with standard statistical analysis tools. More specifically, XGBoost and LSTM were used to predict the trend of the trip demand. This study presents findings regarding the demand of trips in pandemic, behavior of riders and the most important impacting factors to the transportation industry in an unprecedented situation like the covid-19 pandemic.

Keywords: *COVID-19, New York, Trips Forecasting, Long Short-Term Memory (LSTM), eXtreme Gradient Boosting (XGBoost), Social Media Data, Quantifying Policy, Rider behavior.*

2 INTRODUCTION

Shared mobility has gained popularity over the past few years due to its convenience aspects. It has been efficient in fulfilling the demand and services until the pandemic came into picture. The uncertainty of the situation led the service to be completely brought down to halt tand the demand was left unaddressed. Due to the nature of the disease public modes of transportation could not have provided the expected mobility and accessibility to required services. The ride hailing service was deemed to be unreliable when there was the most need for it. One of the reasons the service failed was lack of information on the demand and nature of variables affecting it. If the decision maker and planner had enough data, the situation would have been handled a lot better.

To better understand the response and behavioral pattern within varying socio-economic background and demographic, this paper investigates potential data resources and analysis techniques. The process involved determining ways to collect accurate data, finding correlation and significance of each variable to be able to predict the effect on service.

For this study an innovative data approach was implemented. Present research on ride hailing services had identified various emerging data resources and one of them was using social media platforms such as twitter, Facebook etc. Twitter is known to be a widely used platform for the reasons such as the availability and ease of use (Sayed et al.,2021). It is used to express emotions about the ongoing topics of discussing and debating policies. Over the years it has grown to be one of the trusted platforms for information (Dyer et al., 2020). Through the website available with information regarding the number of times the pandemic was discussed, it was implied to have projected the effect of pandemic on the population. The hypothesis of the intensity to which the pandemic affected the emotional state and decision making was assumed to be dependent on the frequency of its discussion on social media.

The study is divided into three major sections, starting from background study of first wave and second wave during the pandemic in terms of number of reported cases to number of deaths, followed by conducting an analysis to determine the contributing variables and concluding with appropriate methodology. The methodology explored various tools at the disposal in order to make the most out of the available information, which is discussed in detail in the subsequent sections. The following section also discusses the future implication of the study results and the immediate applications of it.

3 LITERATURE REVIEW

Since the first reported novel coronavirus disease (COVID-19) patient in late December 2019 in Wuhan, China, the disease rapidly spread throughout the world, with more than 165 million cases and 3.43 million deaths reported as of May 2021 (World Health Organization, 2021). The magnitude of this global epidemic in terms of its social, economic and health impacts has made it the most extensively reported and covered in history (Gössling et al., 2021; Rastegar et al., 2021a; Seyfi et al., 2020). Its impact of travel behavior and travel restriction is also well documented. This literature review tries to understand the impact of covid and other characteristics related to the behavior of riders in a pandemic.

In analyzing the data obtained from twitter study a model development approach was preferred. The ultimate goal of the study is to be able to predict the nature of ride sharing preferences based on the effects transportation related policies had on behavioral decision making. To obtain the policy feature the data had to be fed to a machine learning model and to determine an appropriate technique a research study into data splitting techniques suggested the appropriate methods developed to capture the essence of the data set. The Data splitting technique to fit any machine learning model is to divide your dataset into three parts such as Training set, cross validation set or also known as development set and testing set aimed towards getting an unbiased estimate of algorithm performance in real world.

(Hall et al., 1998) A research study conducted into determining the goodness of feature subsets and evaluating its effectiveness. The implications of the study suggest that machine learning algorithms automatically extract knowledge from readable information and their success is usually dependent on the quality of the they operate on.

In order to control the spread of the virus, various restrictions were put into force. In line with the study concerning specifically New York, the following section addresses the policies implemented over the span of the first and second wave of the pandemic. In early March the NY governor declared a state of emergency, which was followed by restricted use of public transport and space. Later on, it was declared to be completely shut. As much as it impacted the state's economy, it also affected access to basic necessities and emergency services. Later on as the first wave started to dissolve, the spread caught up earlier than expected. In tandem with the reported cases and deaths, the timeline of its effect on ride hailing services is the topic of interest to this study. (Mohammad et al., 2020) A recent study exploring the impact of COVID-19 on the mode of choice and travel behavior analyzed the data obtained from a survey conducted to capture the real time trip purpose, mode choice, distance travelled and frequency of trips. The study found a significant shift to private transport from public transport, not only for reasons such as the aftermath of social distancing but from a shift in perspective of the service provided as a risk factor. The outcomes of the study could be used in better planning and policy making into the future.

Social media to a large extent projects the ongoing preference of the population. It is known to provide a snapshot of a real person, hence a good data resource to test hypotheses. It could be used to analyze human connection and influence in ways linked to the interest of the field. (Morshed et al., 2021) A research study investigating potential data resources for unprecedented

events such as COVID-19 analyzed digital social media platforms to be an effective and innovative resource. This study captured the effects of pandemic on ride hailing services through tweets. The negative tweets associated with emotions such as sadness and anger were found to be most prevalent in the data set, hence projecting the behavioral mood of the potential users.

With more than 3 billion users on different platforms social media has become a key tool. Most of the activities can easily be executed through social media. One such case is spreading panic, which can easily spread through different platforms under social media. (Ahmad et al., 2020) This study carried out a survey to determine how social media caused panic during the pandemic. The survey began with an online questionnaire which was administered in Iraqi Kurdistan for this study, and 516 social media users were sampled. For data analysis, this study used a content analysis approach. Similarly, SPSS software was used to examine the data. According to the participants, social media has had a substantial impact on spreading fear and panic about the COVID-19 epidemic in Iraqi Kurdistan, potentially having a detrimental impact on people's mental health and psychological well-being. Facebook was the most popular social media platform for disseminating information about the COVID-19 epidemic in Iraq. We discovered a statistically significant positive association between self-reported social media use and the spread of fear associated to COVID-19 (R=.8701). According to our findings, the majority of young people between the ages of 18 and 35 suffer from psychological distress.

(Kadam et. al.,2020) This study showcases how social media created panic during covid-19 in India. Even before the first case of COVID-19 was discovered in India, a social media panic gripped the country, causing the market to run out of masks and sanitizers. Furthermore, false reports concerning viral transmission by air and survivability on different surfaces5 sparked a panic. Though people began wearing various types of masks such as N95, surgical, and simple cloth masks, many lacked knowledge about their proper use and disposal, as evidenced by actions such as frequent touching to mask, wearing the same mask for more than a day, reusing disposable masks, and throwing the masks on the roads or in regular dust bins. Furthermore, the unnecessary usage of N95 masks by the general public during travel and daily activities resulted in a scarcity for frontline healthcare personnel who truly need them. COVID-19 control attempts continue to be hampered by such behaviors. Because of existing medical pluralism in India, communications including false promises about the use of herbal and immunity-boosting drugs, religious and spiritual methods of preventive and treatment were extensively disseminated, adding to the confusion. Confusion was also caused by a lack of information regarding non-pharmaceutical measures such as social distance, quarantine, and isolation, as a result of which visitors from other countries and their connections experienced social stigma in the places where they remained. Fake reports such as major killings of patients in China and the prospect of prolonging the lockdown exacerbated the fear, resulting in persons abandoning quarantine or isolation centers and needless travel ahead to lockdown or even during lockdown to return to their hometown.

4 STUDY LOCATION AND DATA

4.1 Study Area

New York was the first city on the planet to be called a global megacity after it's population touched 10 million in 1950 (Chandler et al., 1987). It is still one of the world's largest mega cities with a population of 19.4 million living in the metropolitan region and 8.8 million within the municipal city (NYC) which includes boroughs of Manhattan, Queens, Bronx, Brooklyn, and Staten Island (Griffin et al., 2016). The rail and bus transit systems built for New York and other cities such as Boston and Massachusetts developed concurrently with these cities which was before automobiles became a popular choice for people to move around the city [10]. The transportation system of NYC is complex and the city has many different modes of transportation such as the subway, buses, ferry and the taxi system (yellow texis, green taxis, uber, lyft etc).

Figure 1: Highlights the five boroughs of New York City and the two major airports

4.2 Yellow Taxis of New York City

In this project we have used the data from Taxi and Limousine Commission (TLC 2021). The data available online through the TLC website and TLC Factbook which is published yearly. Licensed services in New York include Yellow Taxis, Green Taxis, High Volume For-Hire Service Vehicles, Traditional For-Hire Vehicles, Commuter Vans and paratransit (medically-related and dispatched from an affiliated base) but in our project we have only focused on Yellow Taxi data. The taxi data is available online and can be accessed through (TLC Trip Record Data - TLC (nyc.gov). The trip data was extracted from the website and the following graph was plotted which shows the trips per day from February, 2020 to July, 2021.

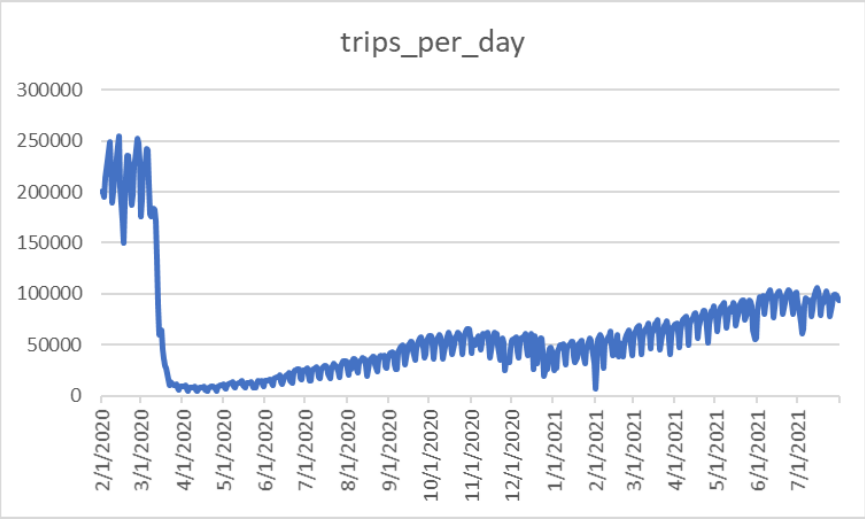

Figure 2: Plot of day versus Trips Per day

4.3 COVID-19 data and the sources used

For COVID-19 cases and death counts we used the official data provided by the New York City Department of Health and Mental Hygiene. The data obtained was analyzed and interpreted and seven-day case count average and seven-day death count averages were plotted as shown below.

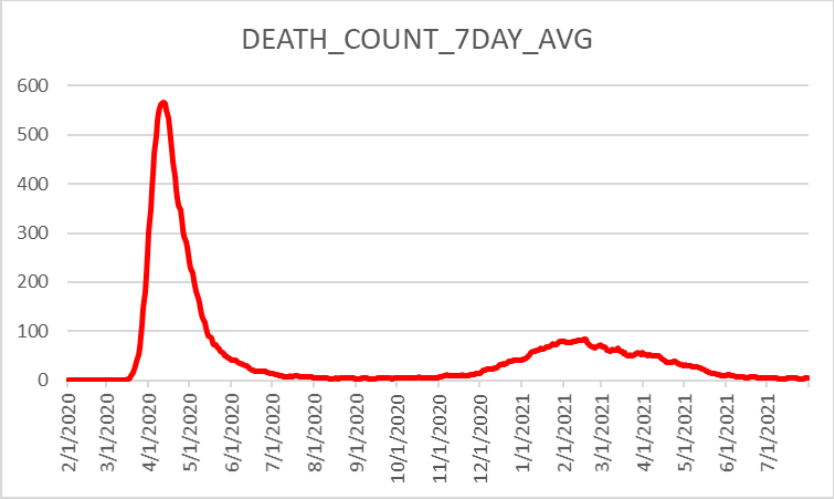

Figure 1 Plot of Days versus 7-day average death count

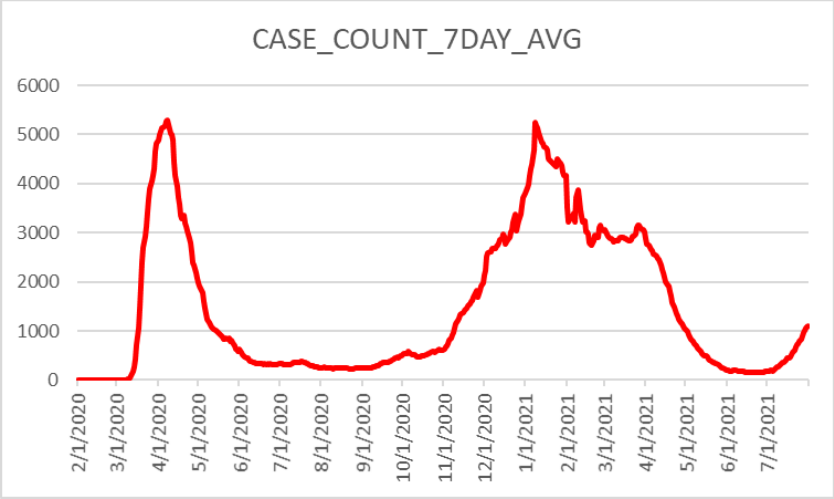

Figure 2 Plot of days versus 7-day average case count

The first COVID-19 case in New York state was confirmed on MArch 1, 2020 and on March 3, 2020 in the New York CIty. There was a steep rise in the COVID cases and they reached maximum around April and started to decline in MAy, 2020. Furthermore, COVID-19 cases did not show any significant increase till lOctober, 2020 during which they started to increase again while the city was still in phase 4 of the re-opening. After this there was a sharp increase again till cases hit a maximum again in January in all the five boroughs with an overall positivity of 10 percent.

The daily death counts in New York City closely followed the case trend (trajectory) during the first wave/peak (March 2020). Even though February and March saw an increase in the number of deaths, the increase was not as prominent.

4.4 Twitter data extraction

In this study we used the number of tweets to get an idea about the number of people who have been talking about the pandemic which in turn would give us an idea about the anticipated demand or the willingness of the people to come out and travel during the COVID-19 pandemic. We evaluated and analyzed this data to see if there is a correlation between people's tweeting behavior and their travel behavior.

The tweets data was extracted from Panacea Lab. This is an open-source platform that started dedicated data collection on March 11,2020 and recorded about 3.3 million tweets per day. The database is currently maintained by Georgia State University. Researchers developed the API to extract the number of tweets from twitter. The data was made available to the general public because of the unprecedented situation that arose in the country and throughout the world.

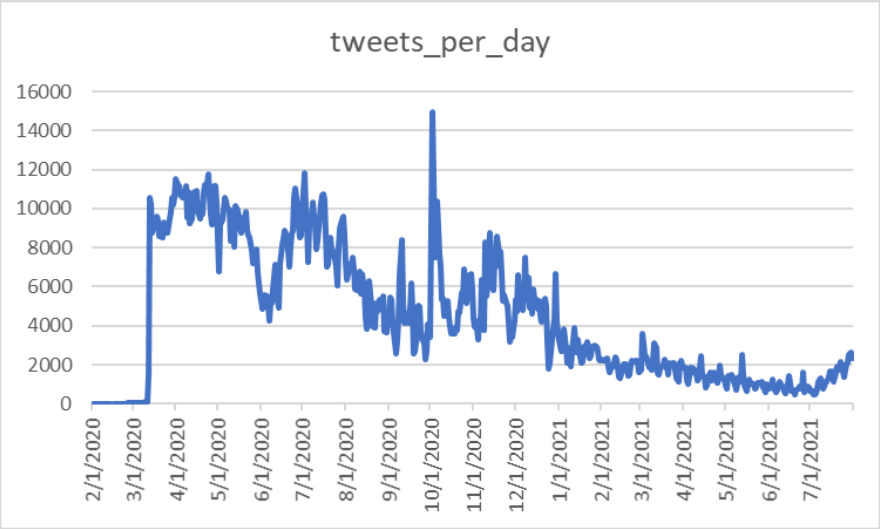

Figure 3 Number of covid mentioned tweets per day

5 METHODOLOGY

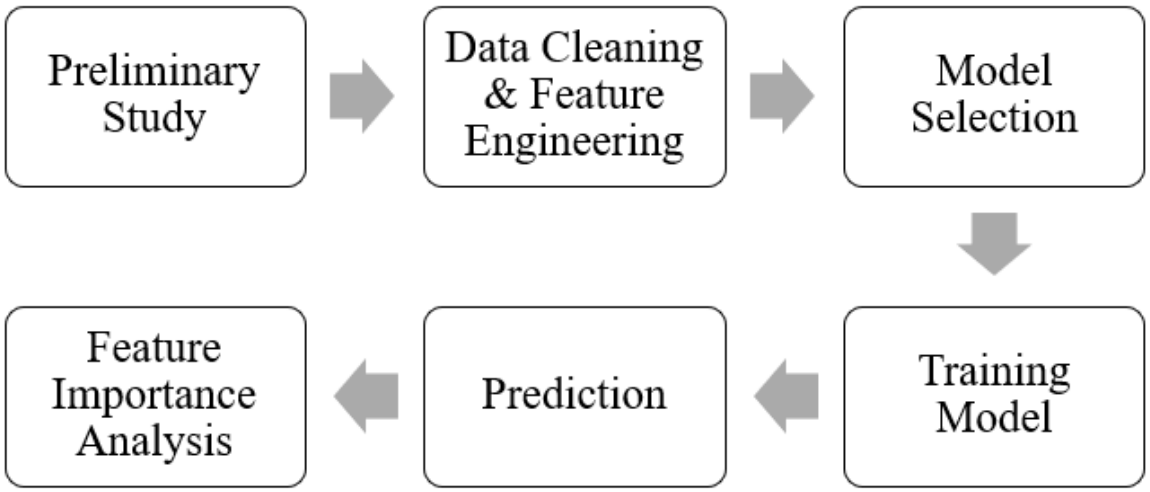

Figure 4 Methodology Flow Chart

First, we collected the data of daily taxi trips which was obtained from the NYC Taxi and Limousine Commission's website. Similarly, we collected Daily Covid Cases and Daily Covid Related Death Counts in New York City. We analyzed the relationships between these variables, which came out to be strong. However, some of the trends in the fall of trips could not be explained alone by the covid and death cases, in particular, after the starting of the second wave, instead of trips falling, they stayed almost the same and again started increasing. To this end, we thought that, maybe there are other latent variables that might explain this phenomenon, and hence explore the lockdown policy and social media. We encoded the lockdown policy ordinally (0,0.25.0.5,1) looking at the reopening trend of school, higher number denoting total opening while 0 denoting total closure. Similarly, we also extracted the number of twitters covid mentions and used it as an independent variable. We observed correlation relations between the various independent variables and the number of trips and found a strong correlation between all the independent variables and the dependent variable, thereby including them for training the model.

5.1 Preliminary Study

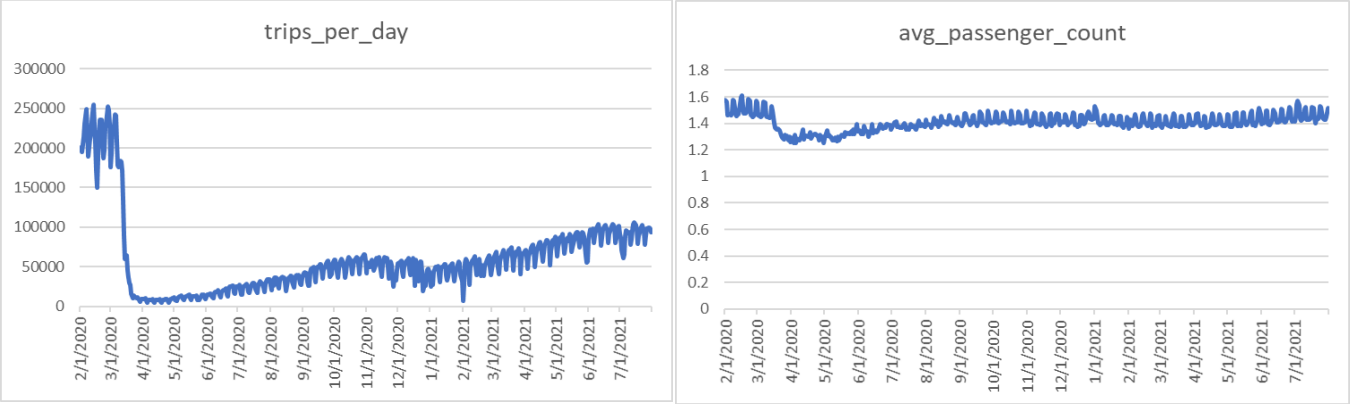

Figure 5 Plot of trips per day and average passenger count for taxi in NYC

The data collected from different sources are explored using different statistical tools. At first, the correlation between different features (data sets) were calculated and a heatmap was generated. The generated heatmap is presented below.

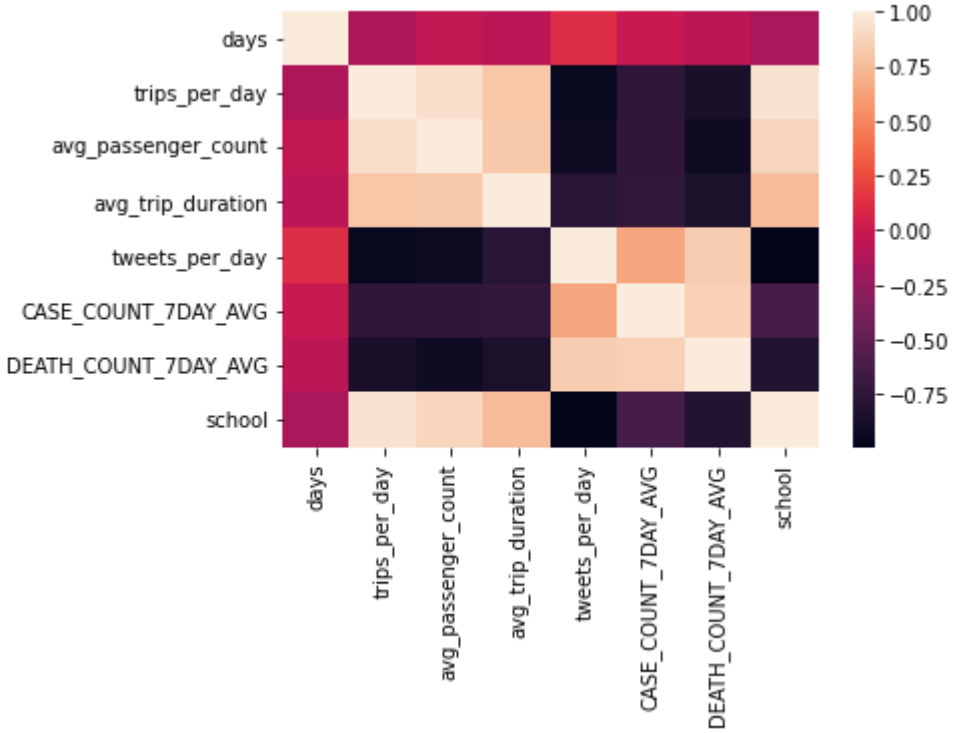

Figure 6 Correlation heatmap of different features used in Models

We can analyze the correlation between the different features looking at the hue of the color in the heatmap. Looking at the heatmap we can see covid data, policy data and tweets data highly related to the trips per day data. The darker color in the heat map represents the negative correlation whereas the light hue represents a direct positive relationship between the data. From the figure above we can see Tweets per day is inversely correlated to the trips per day. This inference is reasonable as the number of people talked about covid in twitter the number of trips

of taxi fell in NYC. The covid related tweets here represent the fear of covid in people. (Dyer et al., 2020) in their study finds that the analysis of public social media presence and expressions can give researchers a near-to-real-time monitoring of indicators of public risk perception. The impact of media reporting and public sentiments may have a strong influence on the public and private sectors in making decisions on discontinuing certain services (Depoux et al., 2020). Reviewing the findings from past literature, the use of social media data to address the perceived fear of the pandemic was used in this study. Looking at the indirect relationship between tweets and trips per day of yellow taxi, the claims of previous researchers is confirmed.

5.2 Data Cleaning and Feature Engineering

A moving average implies that it takes the previous days' values, averages them, and presents the result on a graph. A 7-day moving average is calculated by adding the last 7 days and dividing the total by 7. It will take the previous 14 days to get a 14-day average. So, for example, we have COVID data beginning December 12. The 7-day moving average requires 7 days of COVID instances, which is why it only begins on December 19. On the 19th, it totaled all of the cases between March 12 and March 19 and divided the total by 7. The spot is then plotted. The next point, on December 20, computes the average from December 13 to December 20. If you think about it, you'll see that this "moves" every day, which is why it's called a moving average. For something like new cases of COVID, you can see why this is useful: it gives you an average line over time and removes the enormous peaks and valleys to the average over time.
It is not necessary to stick with 7 days, only. There are also 21-day moving averages or 10-day moving averages, you can make it anything you want. The multiples of 7 are used since there are seven days in a week, which means that every day appears once, twice, or three times in the series.

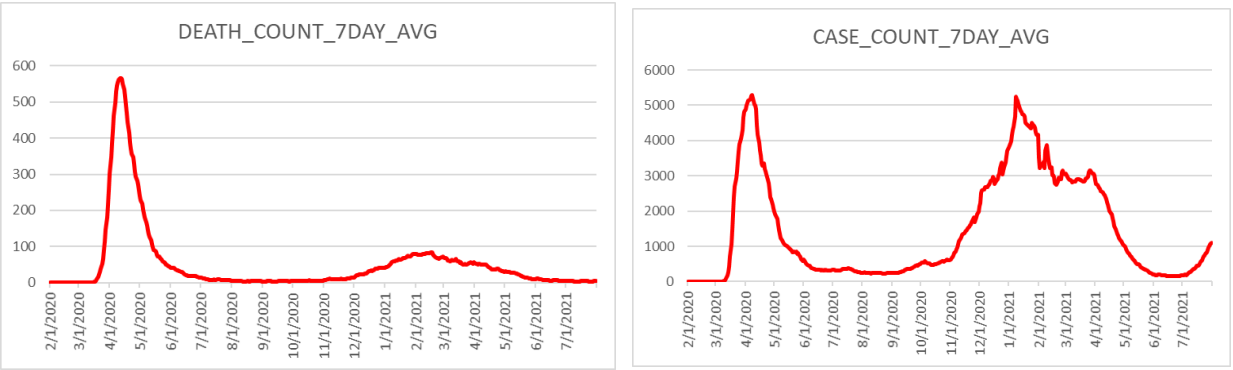

Figure 7 Plot of Days versus 7 day average death count and 7 Day average cases per day

Data smoothing is a statistical method for removing outliers from datasets in order to make patterns more visible. It is accomplished by utilizing algorithms to remove statistical noise from datasets. The use of data smoothing can aid in the prediction of trends.Data may be adjusted during compilation to decrease or remove broad variances or other statistical disturbances. Data smoothing allows traders or statisticians to look at large amounts of data – which can be difficult to analyze – to uncover trends or patterns that they would have missed otherwise. A technique like

this makes advantage of simple enhancements to better foresee varied trends. It focuses on establishing a general direction for the primary data points by eliminating volatile data points and constructing a smoother curve across data points.

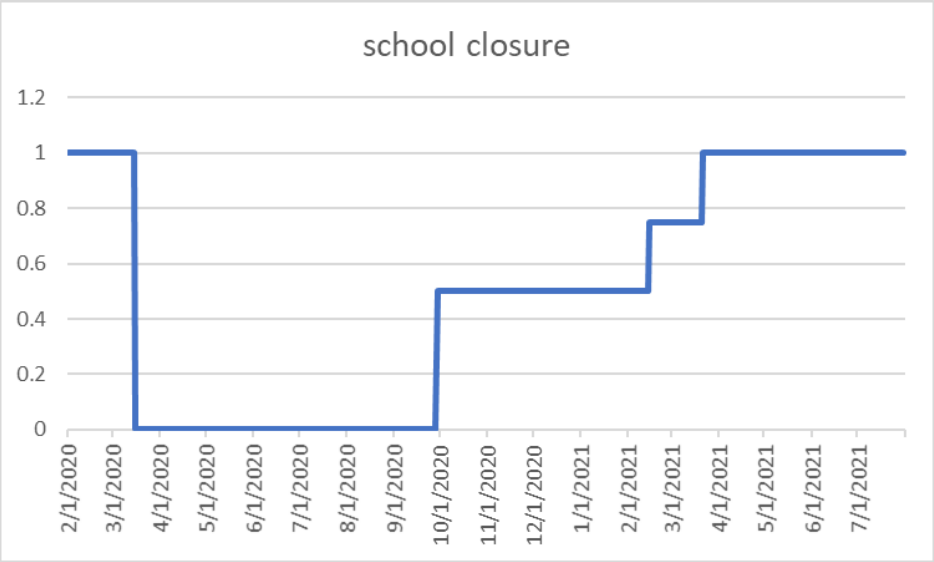

Figure 8 Quantitative graph representing school closure policy of NYC

Similarly, the school closure data which is incorporated in the study is representative of the restrictive policy for movement in the city. School closure is defined here as the period when schools were closed and gradually opened. Nytime(.)com, cnn(.)com, npr(.)com was referred to collect the timeline from shutdown to reopening information of schools for this study. The schools were closed on march 16th following a first wave of coronavirus. Along with schools, a lot of other establishments were also shut down. Here we have only used the timeline of school closures as it is more comprehensive and represents as a whole the city's policy in regards to the movement restriction in the emergency period. After a few months into the pandemic, in September the schools till elementary levels were announced to reopen. This decision was also taken as more information regarding the virus were uncovered such as its low effect on children. This information might have helped in making the decision for reopening. Likewise in February of 2021, the city announced plans to open middle schools. The decision of reopening middle school was quickly followed by opening of high schools just a month after opening middle schools.

Coding the information gathered from different sources regarding the reopening of schools we used simple mathematical relationships. We coded 1 as schools fully open and 0 as schools fully closed. Elementary, middle and high school were assigned weightage according to the number of grades in them. The figure above shows the coded data plotted against date. The school closure is positively correlated to trips per day.

5.3 Model Selection

5.3.1 SARIMA Model (6 months seasonality)

Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting. Although the method can handle data with a trend, it does not support time series with a seasonal component. An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA. SARIMA stands for Seasonal Autoregressive Integrated Moving Average. In our study we assumed a seasonality of 6 months as the waves that formed, first and second had a rough seasonality of 6 months. Seasonality was also changed keeping at 4 months, 6 months and 12 months. The best performing model was with the seasonality of 6 months. A q-q plot was also plotted to test the distribution of trips with the normal distribution. From the analysis, we could not observe any similarity of the trip distribution with the normal distribution.

5.3.2 XGBoost

XGBoost is a software library that is actually an implementation of gradient boosting created by Tianqi Chen and explained in his 2016 paper (Chen & Guestrin, 2016). It is based on the Gradient Boosting Decision Tree (GBDT) model and improves on the calculation speed of the algorithm, while optimizing its performance and efficiency, attempting to achieve the ultimate balance.

The Gradient Boosting Decision Tree being an ensemble learner differs from random forest in that it uses a method called boosting. Boosting combines weak learners sequentially so that each new tree corrects the error of the previous one. This is different from Random Forest as Random Forest builds parallel trees and finally combines the prediction of each model by majority voting.

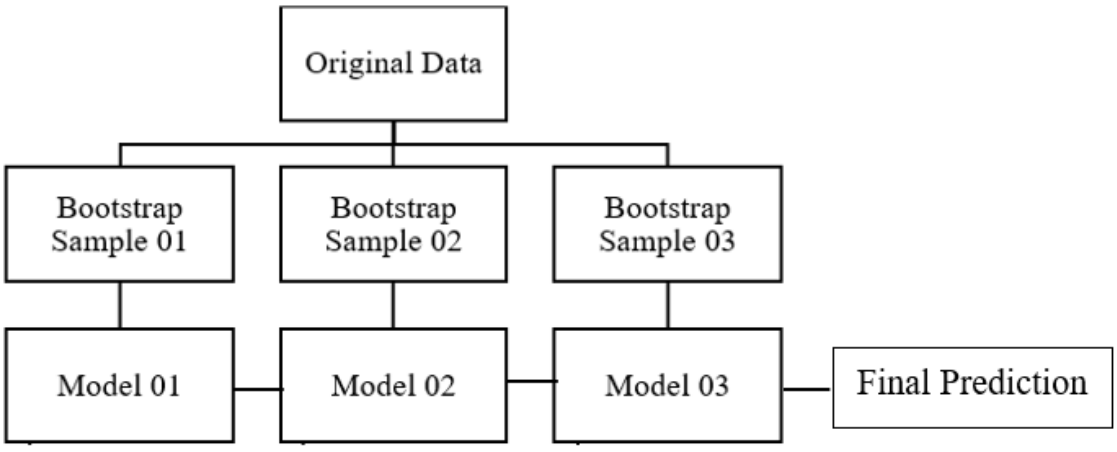

Figure 9 Workings of XGBoost

5.3.3 Long Short-Term Memory (LSTM) Model

LSTM is a deep-learning based model which is popular for time series forecasting and Natural Language Processing. One advantage of LSTM over other deep learning models is that it has a feedback circuit which can take the data from previous time steps as well as the prediction made in the prediction in the previous time step. Intuitively understanding, it makes use of the memory of the previous time step to make the prediction for future time steps. In this manner, the predictions made by the model will incorporate the nature of the data from far behind in time.

5.4 Training Model

For training the model, we took the entirety of data except for the last three months as the training set. We did this because of the time-series nature of the data. We kept the rest of the three months data for validation purposes. Initially, we trained the model on simpler models like logistic regression, and linear regression, however, their performance was not so good. Therefore, we tried a more robust and interpretable model XGBoost. We used transfer learning, i.e. we used the hyper-parameters learned from a previous trained model to train on our data.

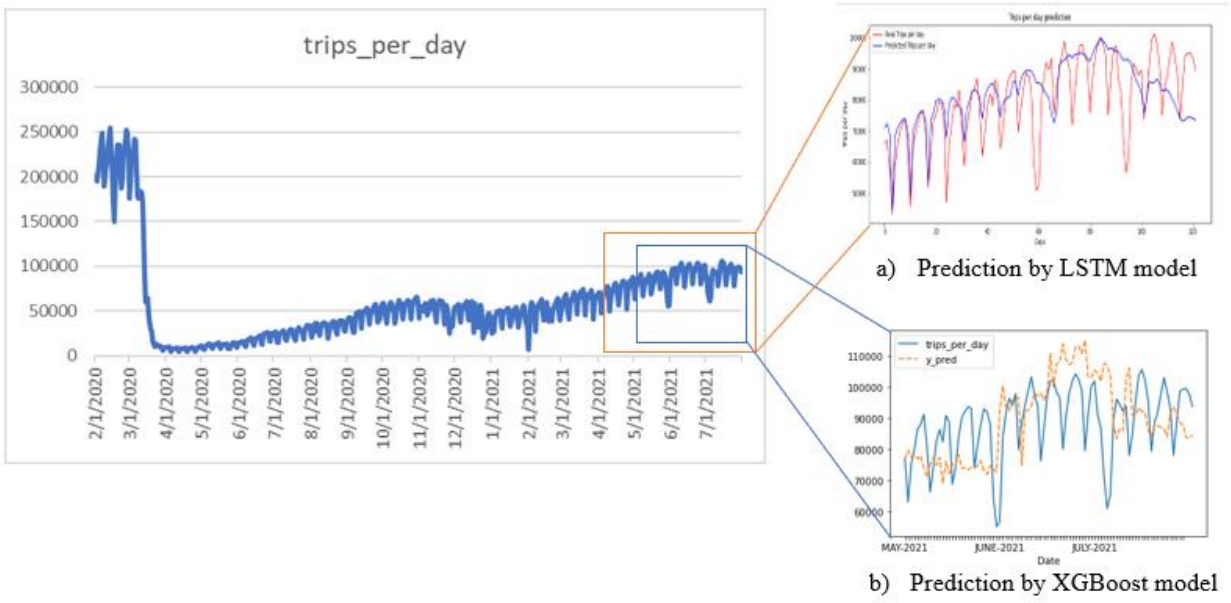

Figure 10 Graph representing the prediction done by two models

5.5 Performance measure of models

5.5.1 Out-of-Sample accuracy measure

Sometimes traditional statistical methods alone cannot judge a model accurately. The standard statistical measures cannot recognize when data is being overfitted. This is especially true for Machine Learning models and deep learning models where usually randomness in the data set is also included in the model. In such situations an out-of-sample accuracy measurement must be

done. That is, the total data is divided into an “Training” set and a “test” set or “holdout” set. Then the training set is used to train or initialize the model. Since the test dataset is not used in estimating any parameter of the model the prediction of the model is compared against the test set. For training the model, we took the entirety of data except for the last three months as the training set. We did this because of the time-series nature of the data. We kept the rest of the three months data for validation purposes.

Final Prediction set as shown in the figure below. The data from February 2020 to March 2021 was used to train the models and the rest i.e., Apr 2021 to Jul 2021 was used in test the prediction done by the models.

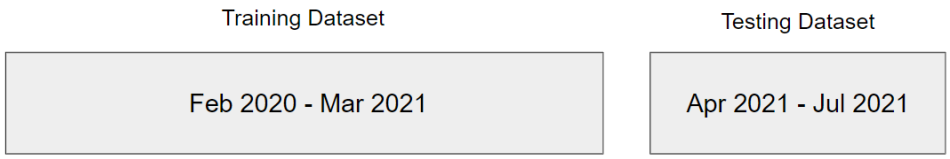

Figure 11 Figure showing the split of data for training and testing the prediction

5.5.2 Root Mean Squared Error (MSE)

Root-Mean-Square Error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. It can be calculated as the follow:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T} (x_{1,t} - x_{2,t})^2}{T}}.$$

5.5.2.1 Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is. It measures the accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus the actual values divided by the actual values

$$\mathrm{M} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

where,

- n is the number of fitted points,
- At is the actual value,
- Ft is the forecast value.
- Σ is summation notation (the absolute value is summed for every forecasted point in time).

6 SYSTEM

All executions regarding the model learning and testing processes were performed using Python programming language version 3.9. making use of the following libraries: Pandas, statsmodel, matplotlib, seaborn, scikit-learn, tensorflow, and keras. The computer specifications used in this study are: Intel® Core (TM) i7-8750H CPU @ 2.20GHz 2.21 GHz, NVIDIA GeForce RTX 2060 GPU, 16.0 GB RAM running a Windows 10 pro operating system.

7 RESULT AND DISCUSSION

7.1 Findings from SARIMA

The result of the SARIMA model was okay, but the allowable error was high. Moreover, SARIMA is a univariate model and does not take into account any features except the time period and number of trips. To this end, we tried regression analysis with popular and effective machine learning models. The Model was used just as a preliminary analysis tool to see the predictability of the trend just by using temporal data. As seen in the figure below the model did not perform well. The margin of error was high and the prediction was off from the actual trend. This finding led us to use other models and make use of other additional features to predict a more accurate trend and shed light to different factors affecting the trips of taxis in a pandemic.

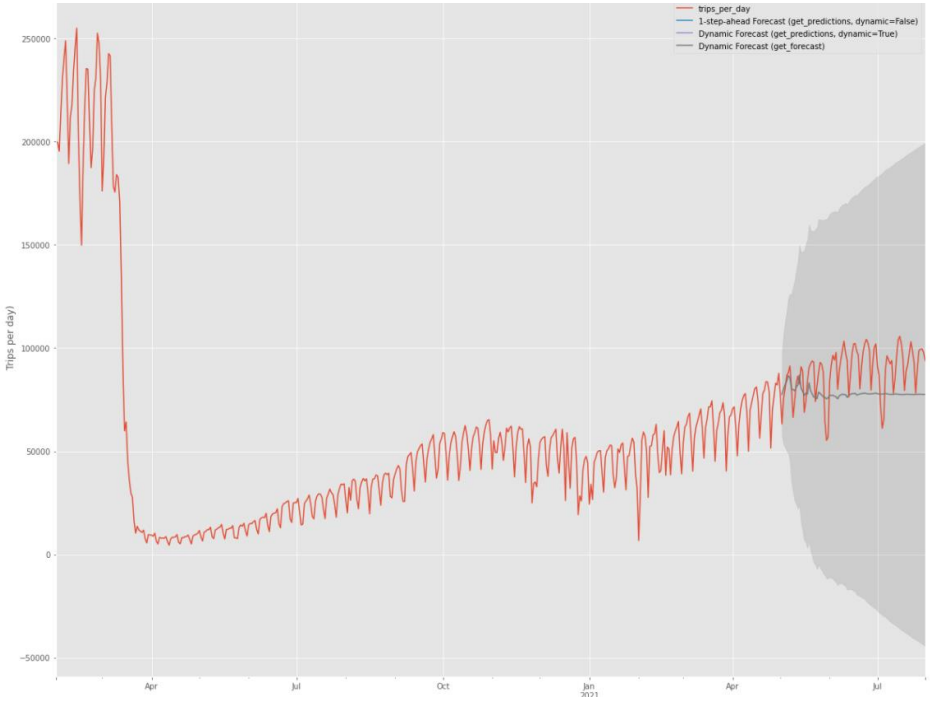

Figure 12 Prediction for the last 3 months using prior data using SARIMA model. (6 months seasonality)

7.2 Findings from LSTM

Many papers such as (Liu, et al. 2020), (Tang, et al. 2019) have made use of LSTM models for traffic characteristic prediction. LSTM which is a deep learning model which has a feedback circuit which can take the data from previous time steps as well as the prediction made in the previous time step. As seen in the figure below the LSTM model performed well in predicting the trips per day of yellow taxis in NewYork city after April 2021 to July 2021. The model used the data from different sources between the period from Feb 2020 to march 2021 to train the model and used the april to july data to predict the trips between april and july. In doing so it also made

use of every precious prediction to train the model as the prediction was being made. This dynamic nature of the model makes it a deep learning artificial intelligent model. Despite its robust mechanism and deep learning ability it produced a 17% MAPE which is not bad considering the weekly fluctuation in demand for taxis in the city. 17% MAPE means that the LSTM model was able to predict the taxi/trips demand with an accuracy of ± 17 percent, which is very not bad but our best performing model, which is discussed below, performed comparative well.

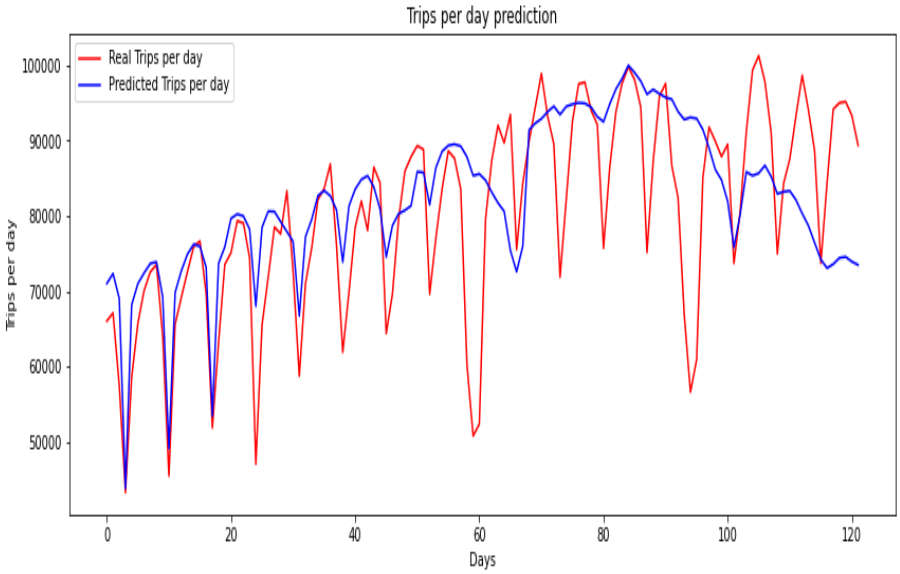

Figure 13 Predicted number of Trips vs Actual Trips using LSTM Model

7.3 Findings from XGBoost

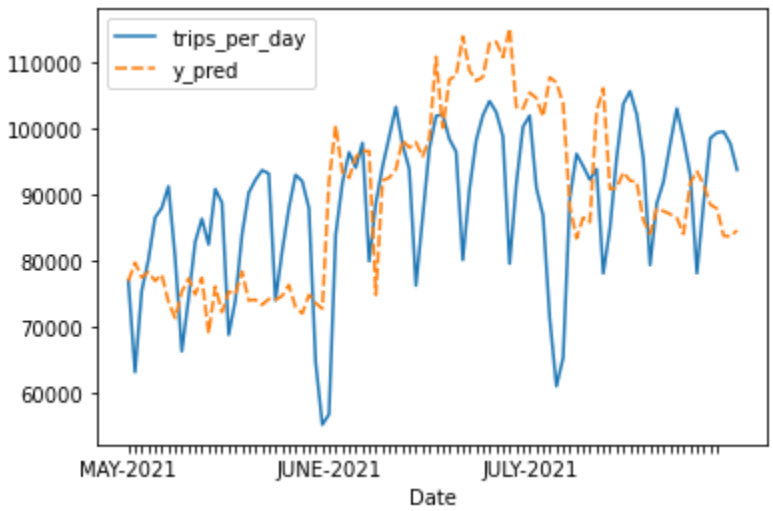

Figure 14 Predicted number of Trips vs Actual Trips using XGBoost Model

*(*Note : Yellow line represents the predicted trips and blue line represents the actual number of Trips)*

7.4 Discussion and feature importance

The best performing model was XGBoost with the MAPE of 13.86 percent, while the second best performing model was LSTM with 17 percent MAPE.

This means that our best performing model was able to predict the taxi/trips demand with an accuracy of ± 13.86 percent, which is very good given the weekly variability of the taxi demand in New York city in an unpredictable time with Covid-19. If we closely notice in the real trips per day, the trend for the amplitude of the last three peaks is declining, and the same is predicted by both of our models, although the amplitudes are slightly on the lower side.

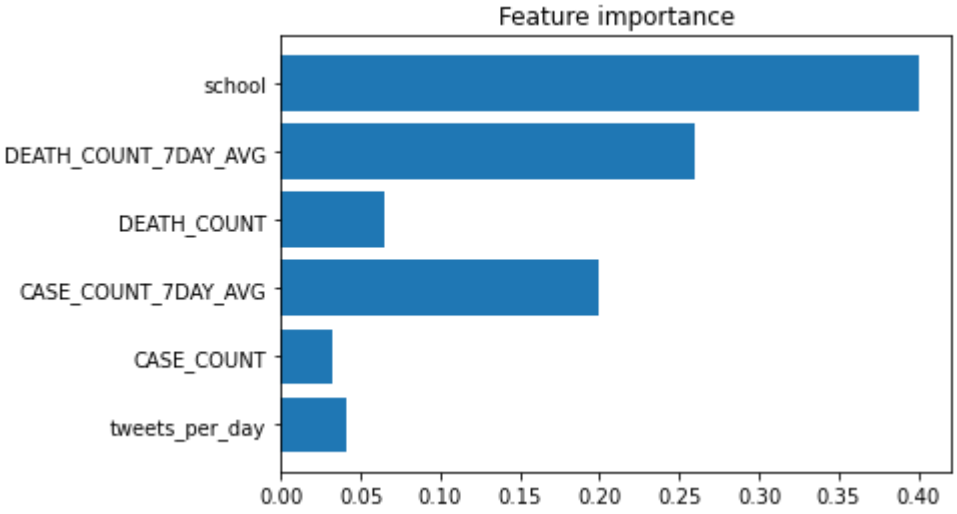

Figure 15 Feature Importance as given by the XGBoost Model

(Note : Here, school represents the school closure policies that were introduced in the city from the start of Covid-19.)

From the above feature importance observation, it is seen that the policies that were introduced during the covid times explained the trend of trips in New York the most. Similarly, the death count, covid case count and covid mentions on twitter contributed in the same order to predict the number of trips after the Covid -19.

Table 1 Showing sensitivity of the XGBoost model upon removal of various features.

MAPE % value excluding a feature from the model	
Including every feature	13.84
Excluding School (policy data)	20.298
Excluding Tweets per day	14.572
Excluding Covid cases	15.68
Excluding Covid Death	41.11
The more the MAPE increases while excluding the feature, the more the feature is important to the model	

We can observe that the variable which affected the model performance the most is the Covid Death. We already had observed a strong negative correlation between the covid deaths and number of deaths, which is further verified by the above analysis. Moreover, the policy changes that we had encoded had the second most negative impact on the model performance, followed by the covid cases, and number of tweets. The work can be extended in predicting other transportation characteristics like for forecasting demand of transit systems or any other mode of transportation, which will help the transportation engineers and policy makers to make better investment decisions.

7.5 Limitations of the study and Future Work

The study included covid case count, death count, policy changes in New York after the Covid-19 and covid mentions on twitter as the main features. Other features from various aspects of society could have been explored to see if they had any relationships with the trips. Similarly, information regarding the spread and fear of covid-19 from other social media websites could have been extracted and built as features. To understand the feature importance or the impact of features on the prediction model, statistical methods could be used. Using statistical methods would also validate the results obtained by the machine learning models, although the statistical models might not capture any non-linear relationships among the dependent and independent variables.

8 CONCLUSION

The ramifications of COVID-19 still hold a large gap in terms of the determining factors on transportation network and services. The reliability of ride hailing services still persists and moving forward to be able to predict accurately understanding the complete picture is vital. These study results provide an overview of the travel trends in the COVID era, using non-conventional sources of data. To be able to explain the aftermath of the situation, emerging data resources such as twitter were utilized to capture the policy and its timeline in combination with death cases and reported cases overlapped. The categories developed from the data showed strong positive correlation to the event. Among the predicted variables, the restrictions on social gathering through policy enforcement, and resolved uncertainties around the situation by sharing information on social media was found to explain the dependability of ride sharing services through the increase in death and reported cases during the second wave. Further on the developed model is validated to be reliable in predicting the next wave based on the correlated variables.

Insights from the results of ensemble learning methods and deep learning technology combined with standard statistical analysis tools tells us that with a rich data source any turbulent period can be explained. It was found that ensemble learning method gives a better understanding of the unique situation. The best performing model was XGBoost with the MAPE of 13.86 percent, while the second-best performing model was LSTM with 17 percent MAPE. The feature importance analysis also pointed out that policy had a huge impact in trip demand as did the covid death count. Different feature-importance analysis produced varied results implying more investigation into the best feature importance analysis tool selection.

9 REFERENCES

1. Sayed et al., “Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis”, Journal of Urban Management,Volume 10, Issue 2, June 2021, Pages 155-165.
2. Hall, M. A. & Smith, L. A. (1998). Practical feature subset selection for machine learning. In C. McDonald(Ed.), Computer Science ’98 Proceedings of the 21st Australasian Computer Science Conference ACSC’98, Perth, 4-6 February, 1998(pp. 181-191). Berlin: Springer.
3. Muhammad, A, Charitha, D. Diasb and M. D. Muleyb, “Exploring the impacts of COVID-19 on travel behavior and mode preferences”, Transportation Research Interdisciplinary Perspectives, Volume 8, November 2020.
4. S. A. Morsheda, S. Shahriar and K Raihanul, “Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis”, Journal of Urban Management, Volume 10, Issue 2, June 2021, Pages 155-165.
5. A. R Ahmad and H. R. Murad, “The Impact of Social Media on Panic During the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study”, Journal of Medical Internet Research, Vol 22, No 5 (2020): May.
6. A. B. Kadam and S. R Atre, “Negative impact of social media panic during the COVID-19 outbreak in India”, National Library of Medicine, 2020 May.
7. Chandler, T. 1987. Four thousand years of urban growth: An historical census. Lewiston, NY: St. David’s University Press.
8. Griffin, Greg Phillip, and Ipek Nese Sener. "Public transit equity analysis at metropolitan and local scales: a focus on nine large cities in the US." Journal of public transportation 19.4 (2016): 126.
9. United States Census Bureau , Population Estimates July 1, 2019.
10. Dyer, J., & Kolic, B. (2020). Public risk perception and emotion on Twitter during the Covid-19 pandemic. Applied Network Science, 5(1), 1-32.
11. Anneliese Depoux, PhD, Sam Martin, PhD, Emilie Karafillakis, MSc, Raman Preet, MPH, Annelies Wilder-Smith, MD, Heidi Larson, PhD, The pandemic of social media panic travels faster than the COVID-19 outbreak, Journal of Travel Medicine, Volume 27, Issue 3, April 2020, taaa031, https://doi.org/10.1093/jtm/taaa031
12. Liu, B., Tang, X., Cheng, J., & Shi, P. (2020). Traffic flow combination forecasting method based on improved LSTM and ARIMA. International Journal of Embedded Systems, 12(1), 22-30.
13. Tang, Q., Yang, M., & Yang, Y. (2019). ST-LSTM: A deep learning approach combined spatio-temporal features for short-term forecast in rail transit. Journal of Advanced Transportation, 2019.
14. Gössling, S., Humpe, A., Fichert, F., & Creutzig, F. (2021). COVID-19 and pathways to low-carbon air transport until 2050. Environmental Research Letters, 16(3), 034063.