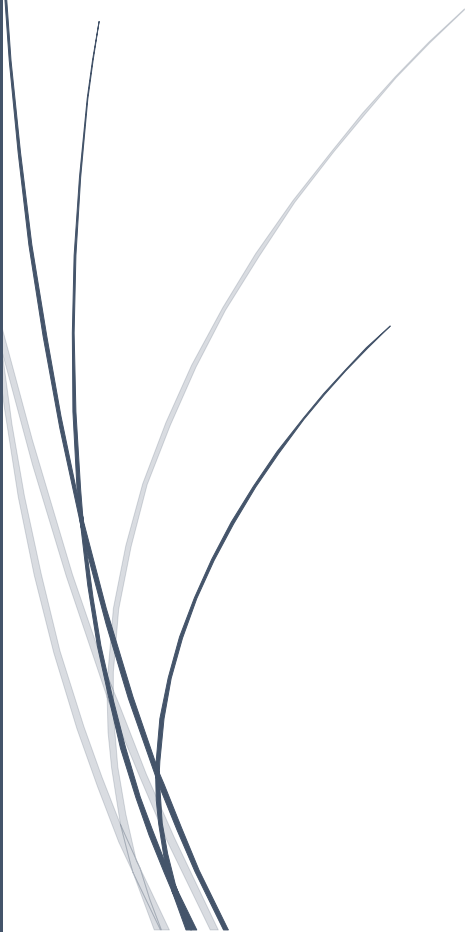
A dark blue vertical bar runs down the left side of the slide. A blue arrow points to the right from this bar, containing the date.

18/01/2021

Predicting Uber Demand in New York City

Several thin, curved lines in shades of blue and grey originate from the bottom left and sweep upwards and to the right.

DEVILLARD Aurélien – PFAU Maurice
AIX-MARSEILLE SCHOOL OF ECONOMICS (AMSE)

1 Table of contents

1	Table of contents	1
3	Executive Summary	3
4	Introduction	4
4.1	Presentation of Uber	4
4.2	Building a prediction model for New-York City	4
4.3	Overview of results and already existing studies	5
4.4	Structure of the report	5
5	Literature review	6
6	The Data	7
6.1	Different sources of the database	7
6.1.1	Uber pickups database	7
6.1.2	Uber zone lookup	7
6.1.3	Weather data	7
6.1.4	NYC official neighborhoods and suburbs delimitation	7
6.1.5	NYC sports event	8
6.1.6	National Holidays	8
6.2	Databases pre-processing and cleaning	8
6.2.1	Merging all databases into one database	8
6.2.2	Visualization of trips per borough on a map	10
6.2.3	Splitting the data	10
6.3	Final database description	10
6.4	Statistics description and data visualization	12
7	Models and results	14
7.1	Regression tree	14
7.2	Deep Learning with Neural Networks	15
7.3	Random Forest	16
7.4	Results of the models	17
8	Conclusion	21
9	List of references	22
9.1	Data sources:	22
9.2	Literature references:	22
9.3	Image sources:	22

10	Appendix	23
10.1	Appendix 1: Result visualization for Chinatown	23
10.2	Appendix 2: Result visualization for East Village	23

3 Executive Summary

The question submitted was to firstly get insights on the behavior of the demand for Uber trips, and then build a model able to precisely predict the number of pickups for a given location and time in New York City. To answer this question, we use the data provided by FiveThirtyEight, who obtained the data from the NYC Taxi & Limousine Commission (TLC) by submitting a Freedom of Information Law request on July 20, 2015. It covers Uber pickups in New York City from January 2015 to June 2015, with the date and the location associated to each one of those pickups. We complete this data with weather data and sports event data, that we found on different sources on the internet.

From the exploratory analysis of the data, we find out that there is clearly a seasonality in Uber demands. Indeed, the demand is at its highest at the end of the week, with the busiest day being Saturday. Regarding the hour of the day, each day can be split into two big periods. From midnight to 9 AM, the demand is lower with spikes at midnight and between 8 and 9 AM. Then, the demand keeps increasing until the evening (between 6 and 10PM), when the peak of demand is highest. As a comparison, there are twice more trips at 7PM than at 10PM.

Besides, the size of the demand clearly depends on the location: the neighborhoods of the city-center (especially Manhattan) concentrate way more pickups than the other neighborhoods of the city. Also, the points of interest such as the JFK airport seem to be real determinants of the demand. Indeed, in the Queens borough, the number of pickups is quite low except in the zone where the airport is located.

The modelling of the demand through machine learning models gives unbalanced results, depending on the neighborhood in which we want to make the prediction. We choose the random forest regression as a machine learning technique to build our prediction model, as it is the most adapted from the results we obtain. This model gives a higher R-square (0,85) and a lower Mean Squared Error than the other models we trained.

The most important explanatory variables for predicting the demand of the Uber Trips are the location (the neighborhood of New York) as a main part of trips occur in Manhattan, and the hour of the day. The addition of variables such as sport event and the weather are not so important for the predictions. About the impact of the weather, this is in line with the existing literature, which is not consensual on its impact for Uber demand.

We can see it by the difference in the results for models trained with and without those added explanatory variables, as they both have the same predictive power. This is also confirmed when we look at the feature importance for our final random forest regression model.

Then, the study we lead here can be exploited by Uber to better understand the demand and predict the number of trips. For the busiest neighborhoods, such as Midtown South in Manhattan, the quality of the predictions is quite good, as we can capture the global trend and magnitude of the demand. But, in the less busy neighborhoods, the number of trips is not high enough to be able to produce exploitable predictions for Uber. From an economic point of view, Uber can use these insights to give incentives for drivers to provide their services at the right time and location. Indeed, as we can predict the demand for the next hour, this gives the possibility to react quickly by trying to increase the number of available drivers.

4 Introduction

4.1 Presentation of Uber

Uber is a company that provides a platform to establish a link between a driver and a passenger, it is the biggest competition to the classical taxi system in big cities. This platform is available through a smartphone application that both the driver and the user must have at hand.

The way this app works is that the client chooses a trip to a destination, with the starting point usually being his current location. The drivers in the area will see the pickup request on their phone and can then choose to accept the trip and go to the pickup location. The contact between driver and passenger is established through the phone app. At the end of the drive both give a rating on a scale from 1 to 5. Thus, every passenger and driver have an average rating which is visible while ordering or accepting a trip.

Let us show an example to better illustrate. A user of the Uber app is at location A and wants to go to location B, he will enter his destination into the app which will start alerting the closest available drivers around point A. Once a driver accepts the trip, the driver will drive to the pickup location to find the passenger, pick him up and bring him to location B. Both then get to rate the drive on the app, which will be added to their average rating as a passenger or driver.

Uber's pricing system is dynamic, as a trip's price can change given the demand, time of the day for example. Drivers can see on the map where prices might be higher, and thus are more likely to go to these areas. Uber drivers only make money while driving passengers, every moment in between is unpaid.

4.2 Building a prediction model for New-York City

The prediction of the location of Uber pick-ups is a crucial point for the company as it is the heart of its business. If the company knows where the customer will probably ask for a trip, Uber can encourage drivers to be present at the right place in the city, and at the right time. The goal is that Uber drivers spend a maximum of their working time driving customers, and that the waiting time, or the reaching time to clients is minimized. In the current situation, drivers may be tempted not to go to the less interesting parts of the city. What we want to deliver is a model that will help Uber predict the trend of the demand at different times of the day, and for some of the neighborhoods of the city.

The main goal is to increase the productivity of the workers at Uber and the satisfaction of the clients at the same time. Indeed, we know that drivers are highly tempted to work only during peak hours, and only in some districts. Thus, this can result in a higher waiting time for the consumers asking for rides in less profitable areas. Besides, if the company knows when and where the demand for Uber will be, this information can be transmitted to the drivers.

Studying New-York City is pertinent for the following reasons: this is a place where there is a huge competition with the famous yellow taxis of New York, but also other vehicle-for-hire competitors like Lyft. Besides, this city represents a huge market with more than 14 million trips accomplished by uber between January 2015 and June 2015, the period of our study.

4.3 Overview of results and already existing studies

The model used data provided by Uber, from January to June 2015. Some other studies identify a correlation between the number of rides and the weather in the city. Consequently, we have included weather data in our models.

Globally, we find out that predictions from our machine learning model are exploitable only in some parts of the city: indeed, as some zones are characterized by a very low demand of Uber pickups, there is no clear pattern of the demand to make precise predictions.

4.4 Structure of the report

The report will be structured as follows: we first introduce the treatment of the data and the explanatory variables used. Then, we provide some statistics and visualization of the demand of Uber in New-York City. We also briefly present the theoretical machine learning models used in our study, followed by the results obtained. The illustration of the predictions is made on a few neighborhoods that were selected by us, to show the behavior of our model in neighborhoods with different dynamics of Uber pickups.

5 Literature review

As vehicles for hire economy has been growing in the last years, recent studies have tried to understand the factors of the demand and predict pickups thanks to statistical modelling. Ye, Tianyu (2019) worked on New York City and Washington D.C data, and they find that the “temporal information provides more predictive power” than the weather features. On the other hand, Brodeur and Nield (2018) show that the number of rides is correlated to the weather, essentially if it is raining. There is no clear consensus on the importance of the impact of weather on the Uber demand in the literature.

Some also try to build models to predict the global demand for Uber Pickups thanks to Deep Learning methods (Wang et al, 2018.) and make their predictions with Long Short-Term Memory (LSTM) Networks. Their model is used to make global predictions of the demand in the whole city of New-York, not neighborhood per neighborhood, with hour-to-hour predictions. They have also included weather data as predictive variables in their study.

We work on the same database as the one used above, but with some modifications. Indeed, we bring a wide range of explanatory variables such as the weather and the occurrence of sports events (sports games in the City). Besides, the predictions of the number of Uber pickups we are trying to make are precise to the hour, and specific to each borough.

6 The Data

6.1 Different sources of the database

All the data used in this report is freely available on the web, with the links available in the list of references section at the end of the report.

6.1.1 Uber pickups database

Firstly, we have the database given by Uber, covering all the Uber pick-ups between January 2015 and June 2015 in New-York City. In this database, we have a table with each line corresponding to a trip that was realized, with its exact time of pick-up, and the location of the pick-up (borough and precise neighborhood).

6.1.2 Uber zone lookup

Each pick-up location is associated to a specific borough and zone of New York. Thus, we have merged the database containing the name of the borough and the subzone. There are 265 different values for the zones, and seven unique values for the boroughs.

6.1.3 Weather data

To obtain the weather data, we use the data from the website “openweathermap.org” which gives us the weather for the whole city, hour per hour. The measures are made from a station located in the heart of New York City, in Central park. We choose to keep the weather description and the temperature, as they are the most likely to influence the demand. Indeed, customers may replace a walking journey by Uber journey because of the meteorologic conditions.

As it is initially a weather-oriented dataset, the description of the weather is precise: we decide to group the different descriptions to create a dummy variable, stating if it is raining or not.

6.1.4 NYC official neighborhoods and suburbs delimitation

The database contains the exact and official delimitation of the neighborhoods of New York City. The file associates each neighborhood with a zone delimited by latitude and longitude coordinates. There are 195 unique values for the name of the neighborhoods. We use the database to plot the map of New York and the number of pick ups associated to each neighborhood of the city.

6.1.5 NYC sports event

To add another explanatory variable to our model, we look for all the big sport events that happened in New York City during the period we study. We search for all the games of basketball, baseball, and hockey as those are the most attended sports in the USA. For each game, we have the hour of beginning, and we deduce the hour at which the match ends, based on the average length of a game. Indeed, there may be a rise in the demand for Uber rides when the people exit the stadium. For basketball, baseball and hockey, the average length of a game is respectively 2h30, 3h and 3h. This new explanatory variable is equal to 0 if there is no game at a given time and borough in New York, and to 1 otherwise. The aim is to take in account big sportive events for the demand, as those are predictable events.

We have 3 sports stadiums in our zones:

- Maddison Square Garden for basketball and hockey in Midtown South, Manhattan
- Barclays Center for basketball in Park Slope-Gowanus, Brooklyn
- Yankee Stadium for baseball in West Concourse, Bronx

6.1.6 National Holidays

We create a dummy variable indicating if a day is a holiday or not, to take in account the possible effect of a holiday on the behavior of the demand. There are 4 national holidays during the period of our study:

- New Year's Day, January 1st, 2015
- Martin Luther King Day, January 19th, 2015
- President's Day, February 16th, 2015
- Memorial Day, May 25th, 2015

6.2 Databases pre-processing and cleaning

6.2.1 Merging all databases into one database

To start, we need to merge the data from the different databases. We first merge the databases with the uber pickups and the database "uber zone lookup", to associate each pickup to its exact location (borough and neighborhood). After having merged those databases, we have the following boroughs associated with the total number of trips in each of them:

	Total number of trips
Manhattan	10371060
Brooklyn	2322000
Queens	1343945
Bronx	220146
Staten Island	6959
Unknown	6264
EWR	105

Figure 1 Total Number of Uber pickups per borough, between January and June 2015

We see that there are data that are categorized in an “Unknown” borough. Indeed, we have two location identifier that are associated with unknown locations. Those locations concern 6264 pickups. As it represents a small number of observations compared to the whole dataset (more than 14 million trips), we decide to drop those observations as we cannot use them in our prediction analysis. Besides, we have 105 locations associated with the EWR borough. This corresponds to Newark Liberty International Airport, which is in New Jersey, and not in New York City. Considering the neglectable number of pick-ups associated to this location, we also decide to drop this data.

Then, we finally consider the 5 main boroughs of New York City: Manhattan, Brooklyn, Queens, the Bronx, and Staten Island. Each of those boroughs is divided in neighborhoods, with 262 neighborhoods in total.



Figure 2 Map of the neighborhoods of New York City

After this, we add the other explaining variables to our database which are: the national holidays and the holding of a sport event.

6.2.2 Visualization of trips per borough on a map

Finally, to be able to visualize the location of the trips on the map of New York City, we create a heatmap, showing the number of trips on the whole period for each neighborhood. To do this, we use the database from “New York official neighborhoods and suburbs delimitation”, which geographically delimitates the neighborhoods. Yet, to use it, we need to harmonize the names of neighborhoods with the one we have in our initial pick-up dataset. Indeed, the pickup locations from our initial dataset sometimes corresponds to the same neighborhood, but the name can be slightly different. Besides, the pickup locations are more precise than the official neighborhoods, meaning that we must group some sub-neighborhoods together for the data visualization.

6.2.3 Splitting the data

For the training of our models, we need a training subsample and a testing sample, to compare which model gives the best results and thus, to select the best possible one. We proceed to do two splits of the data, each for a different purpose:

- We split the data randomly into two samples, one training sample with 80% of the data which represents 540486 observations and one testing sample with 20% of the data which represents 135122 observations.

With these two samples we can compare every combination of hyperparameters for each model, and every type of model by training on the train sample and testing on the test sample.

- We create a second split, independently from the previous one, to test and visualize the predictive power of our model. Indeed, once the optimal model is chosen, it is interesting for us to know what the model is capable of. Thus, we choose to create the train and test samples ourselves, by training the model on every observation available except for the first week of March, which we will use as a test sample. We choose a week in March as it is closest to all the other months available in the data.

6.3 Final database description

The dependent variable of our model, Y , is the number of Uber pickups at a given neighborhood and a given date and hour: $Y_{\text{date, hour, neighborhood}}$

Below is the dictionary of the explanatory variables used.

Name of the variable	Type	Description	Values associated
zone	Characters	Neighborhood of the pick-up location	Name of the neighborhood (258 unique neighborhoods)

locationID	Numerical	Unique identifier associated to a neighborhood of the pick-up location	Integer between 1 and 263
month	Numerical	Number of the month in the year at pick-up	1 for January 2 for February, etc.
day	Numerical	Number of the day in the month at pick-up	Integer between 1 and 31
hour	Numerical	Hour of the day at pick-up	Integer between 1 and 24
temperature	Numerical	Temperature in New York City in Celsius at pick-up	Float of the temperature in degree Celsius
week_end	Numerical	Dummy variable to differentiate days of the week at pick-up	Equal to 1 if Friday, Saturday, Sunday 0 otherwise
sport_event	Numerical	Dummy variable if a sport event ended at the time and the location of the pick-up	1 if a sport event occurred, 0 otherwise
borough	Characters	Name of the borough where the pick-up occurred	Name of the borough (Manhattan, Bronx, Staten Island, Queens, Brooklyn)
nb_trips	Numerical	Number of trips which occurred in a given neighborhood at a given time	Integer
weather_cat	Numerical	Dummy variable if it is raining or not	1 if it is raining, 0 otherwise
holiday	Numerical	Dummy variable to indicate if it is a holiday	1 if it is a holiday, 0 otherwise

Figure 3 Dictionary of the dataset

6.4 Statistics description and data visualization

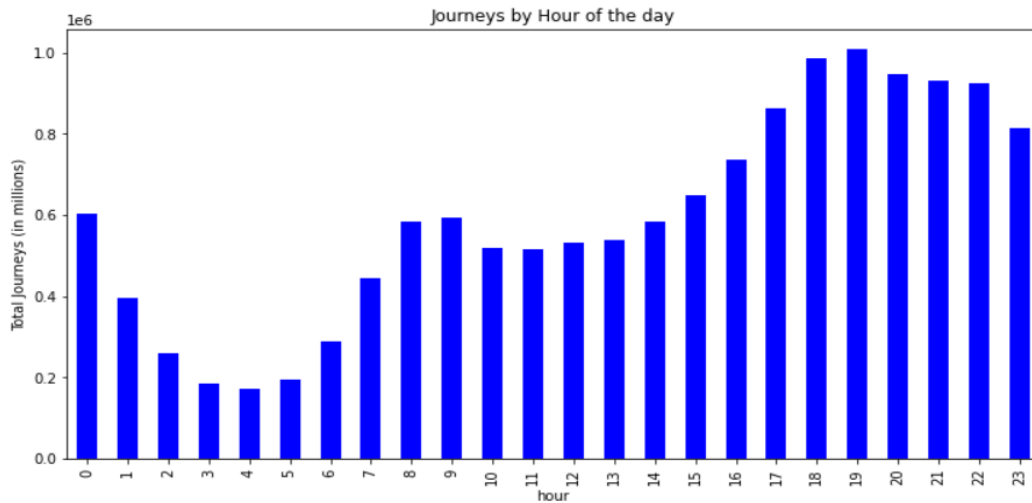


Figure 4 Total number of Uber pickups per hour during the period of study

On the Figure 4, we clearly see a seasonality in the demand depending on the hour of the day, with periods with a high demand (evening) and some with a low demand (middle of the night). The relation is logical as clients must be at work during the day, and they will ask for Uber rides during their free time during the evening, to reach different places in the city.

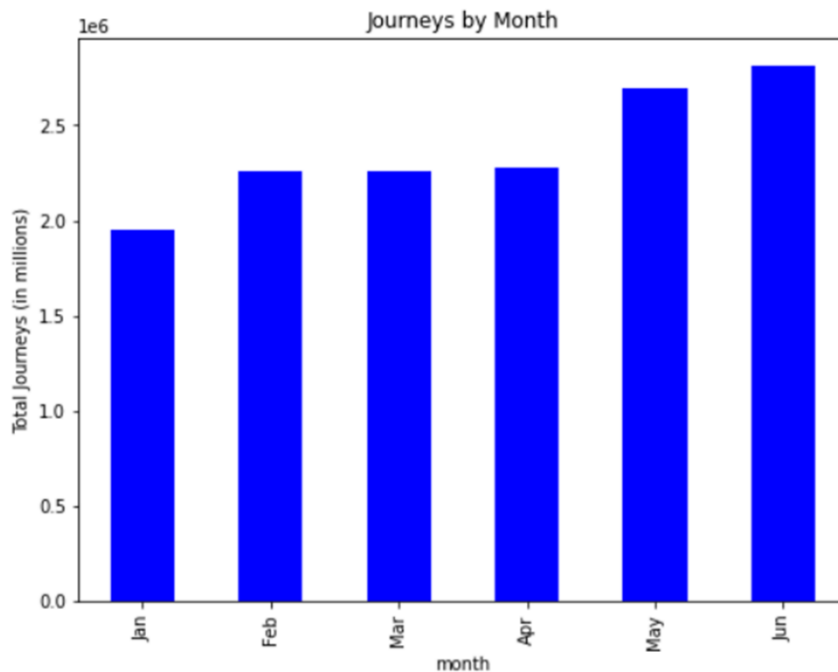


Figure 5 Total number of Uber pickups per month during the period of study

From the bar plot above, there seems to be an increase of Uber demand over time. Indeed, the number of trips per month is more than 30% more important in June than it is in January.

As we try to predict the number of trips for short-term (for instance the next week), this trend is not a problem for us.

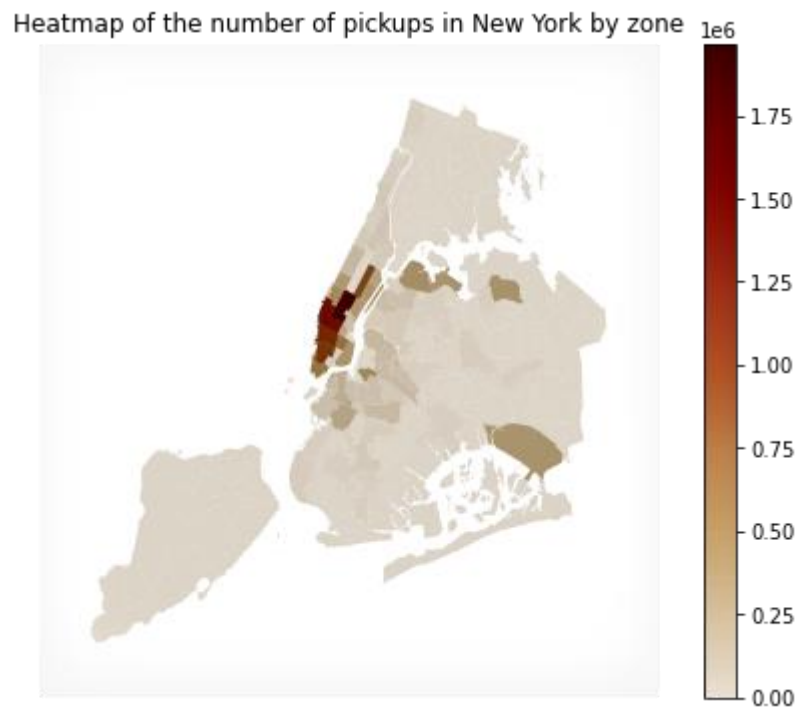


Figure 6 Heatmap of the location of Uber pickups by neighborhood

From the heatmap (Fig 6), we clearly see that the occurrence of Uber pick-ups is concentrated in the neighborhoods of Manhattan. The zone which shows some activity on the South-East of the map corresponds to the JFK airport location. From this, we already expect better predictions for Manhattan than for the other boroughs.

7 Models and results

To be able to predict the Uber pickups in the city, we implement two different machine learning models that are supposed to be efficient in the case of regression and prediction. All the work has been done with the Python software.

To choose the optimal model, we test three types of models:

- Regression Tree
- Artificial Neural Network
- Random Forests

For all comparisons of models, we look at three indicators measured on the test sample:

- The mean squared error (MSE), which we will want to be as low as possible.
- The mean absolute error (MAE), which we will also try to minimize.
- The R^2 score, which we will want to be as close to 1 as possible. It corresponds to the part of the variance explained by our model.

7.1 Regression tree

The regression tree is a machine learning method that follows a process called binary recursive partitioning: it splits the data into partitions, using every binary split possible on an explanatory variable. The best split will be retained, it is the one that leads to the lowest Residual sum of squares, corresponding to the following formula:

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2.$$

With J the region in which we assign, the observation, y the observation, \hat{y} the mean value in the region.

On the following example (Fig 7), we consider the value of the 24th explanatory variable (x_{24}) to make the first split: based on its value, the data is split into the left or the right branch.

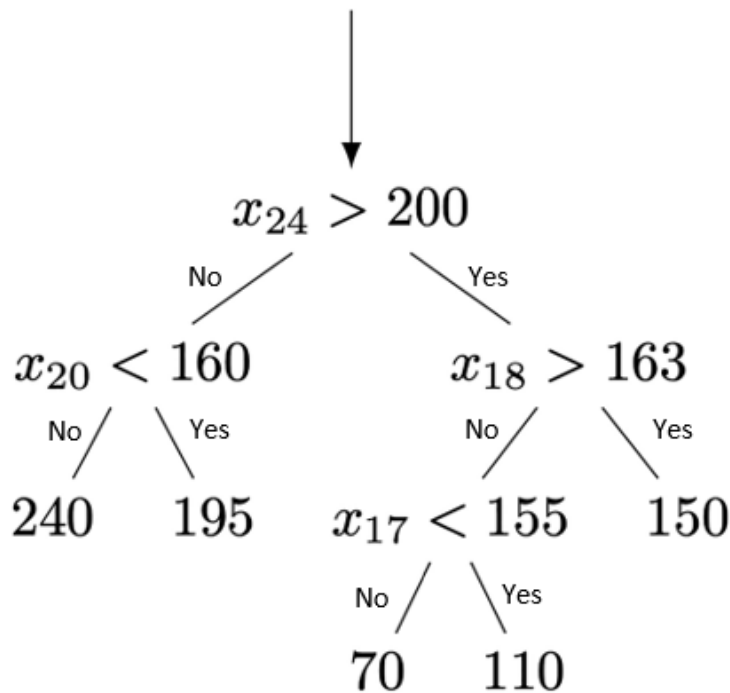


Figure 7 Simple illustration of a regression tree, with x the explanatory variables

7.2 Deep Learning with Neural Networks

The artificial neural networks (ANN) are a computing system inspired by the neural network from the brain, they are based on a collection of nodes. A neural network is then characterized by the number of layers (group of neurons), and how many neurons they are made up of.

To sum things up, a classical neural network can be defined with an input layer (corresponding to data we feed it with), hidden layers that apply complex mathematical computations and transformations, and an output layer giving the final predictions (Fig 8).

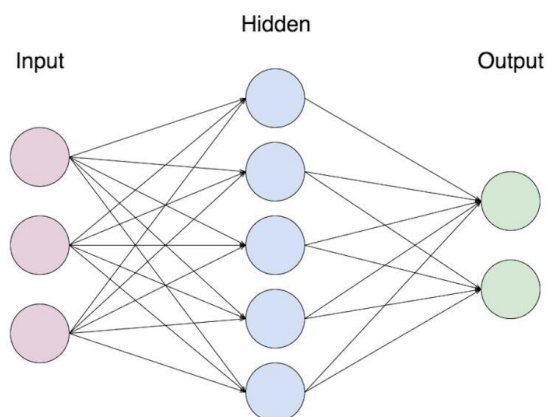


Figure 8 Simplified representation of a neural network

In a neuron (Fig 9), we multiply each input by a weight (w), and we add a constant (b). Then, we apply a transformation to this number, the most often not linear, this transformation corresponding to the activation function of the neuron. We can choose the activation function we use, based on the type of problem we face.

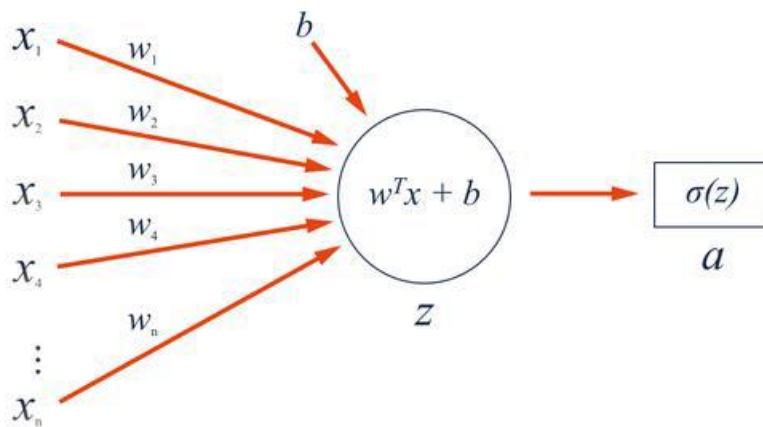


Figure 9 Computations made in the neuron of a Neural Network

7.3 Random Forest

The random forest algorithm is an extension of the regression tree model, introducing randomness. Basically, the random forest creates N number of regression trees, and takes the mean of the result at the end. It works better than the single tree because it trains on more samples, but it takes longer to run.

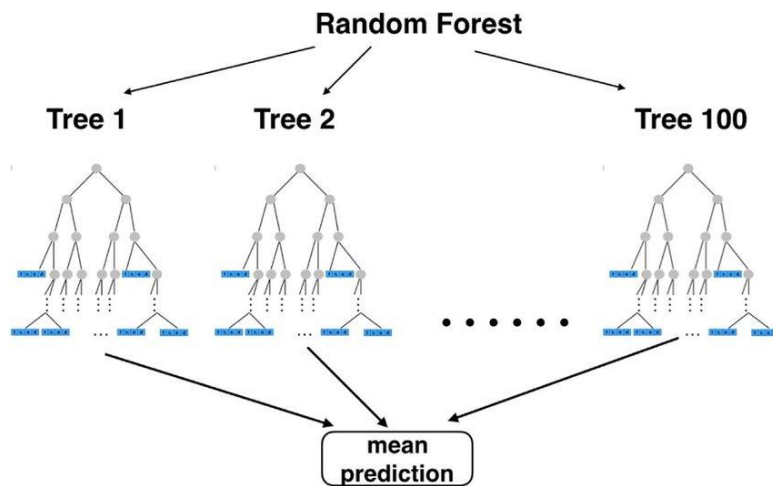


Figure 10 Illustration of a Random Forest

In the figure above we can see how the random forest predicts a result. For our model we choose 100 trees in our forest and get the mean prediction as a result.

7.4 Results of the models

We proceed in two steps to select the best model. First, we select the best hyperparameters for each type of model by measuring the error and R square indicators for different values of hyperparameters. Then, once we have the best hyperparameter combination for each machine learning model, we compare them to keep the best one:

	Regression Tree	Artificial Neural Network	Random Forest
Mean Squared Error	254.22	550.09	85.25
Mean Absolute Error	6.80	13.83	4.21
R2 Score	0.56	0.05	0.85

Figure 11 Result scores for each model on the test sample

As we can see above, one model is clearly better for its predictive power and measured error, it is the random forest regression. The score of the Neural network is a bit surprising, but it may come from the limitation of our computation machines.

To better understand the chosen model, we proceed to plot the importance of each explanatory variable in the model:

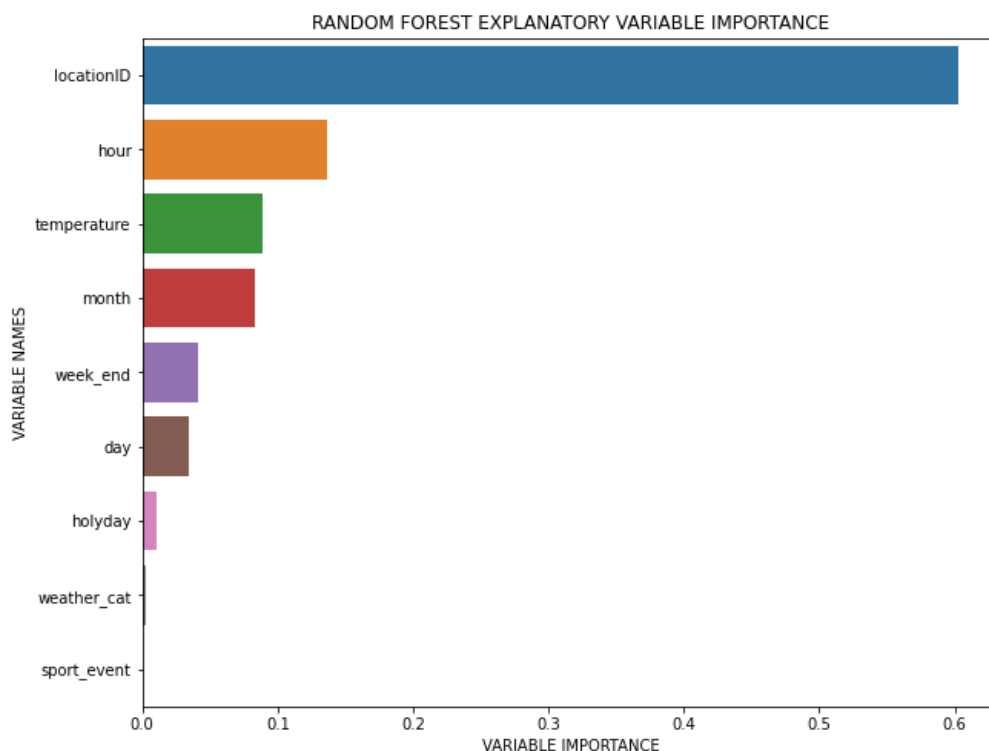


Figure 12 Plotting the importance of each of the model's explanatory variables

In the figure above we can see which variables are the most important for the model. This means that at each split of the data in the trees, these are the variables that are most looked at to try and reduce the mean squared error (as it is how the splits are made). We can see here that the location (locationID) is by far the most important, while additional variables such as the weather category (0 for no rain, 1 for rain) and sports events (0 if there is no sport event at the

time at the location, 1 if there is) seem to not be considered. When removing these variables from the model, the results are not improved, thus we decided to keep them.

We try to check the importance of the weather and sports variables by predicting a number of trips for given values for the explanatory variables, and by measuring the difference in output for both possible states of each of the binary variables:

- For the sports events we select a day in the test set with a game that day, and only have the corresponding variable vary, but we get the same result in predicted number of trips with or without a sports event.
- We proceed the same way for the rain variable and get the same null impact as the sports variable.

With the chosen model, we are trying to show some predictions on the first week of March in certain districts. For this we use the model trained on all observations except the week in question.

To present our prediction results, we decide to plot the predictions over some representative neighborhoods. Indeed, as there are more than 250 different neighborhoods, providing predictions for all those neighborhoods would not be pertinent. We choose the following ones:

- Midtown South: this neighborhood located in Manhattan is one of the busiest ones in terms of number of pickups.
- JFK Airport: in this neighborhood there is an international airport, it is in the borough of the Bronx. Thus, the activity of the airport should be predictable.
- Spuyten Duyvil - Kingsbridge: located in the Bronx, this a neighborhood where there are not many trips recorded. We want to look at the prediction performance of our model on this kind of zone.

Some additional predictions for other Zones are available in the appendix.

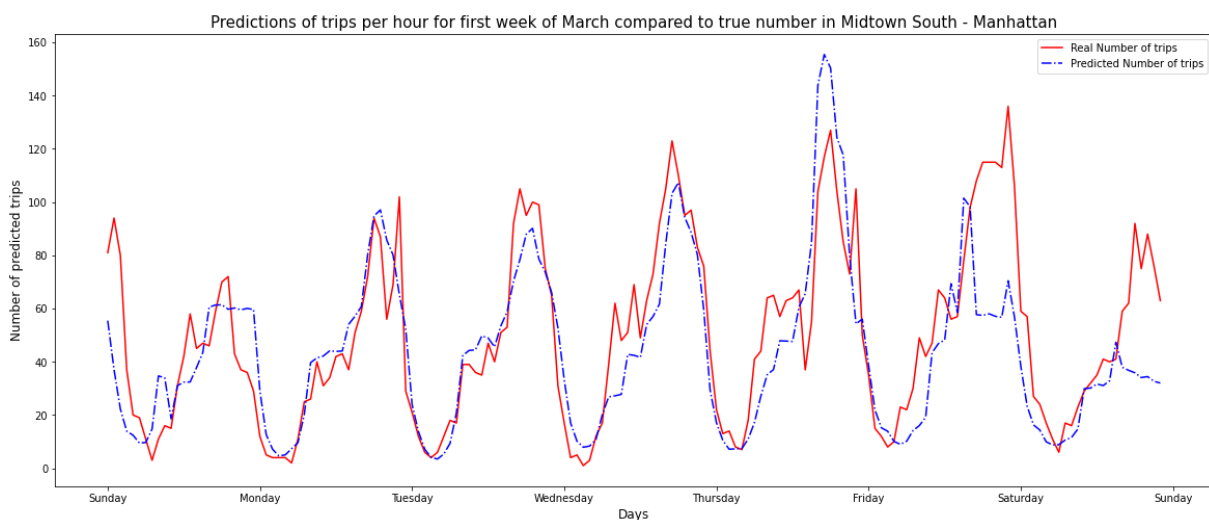


Figure 13 Model predictions for Midtown South

We have here above a confrontation of our model's predictions with the actual data for the first week of March. We can see that our model seems to capture the tendencies quite well, the

accuracy scores are higher than on the overall model with a mean squared error of 465.92 and a mean absolute error of 15.38. These high values for the error compared to the global model can be explained by the fact that Manhattan is one of the busiest boroughs, by far, and thus each prediction error is exacerbated. We also have an R^2 of 0.58, so our model can explain parts of the data's volatility.

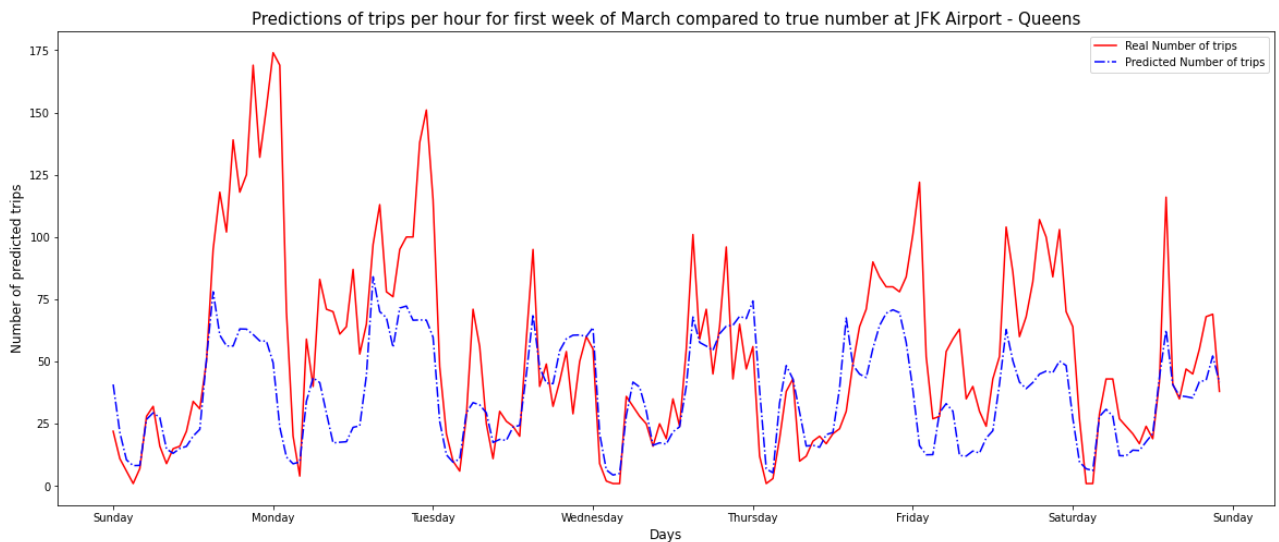


Figure 14 Model predictions for JFK Airport

JFK Airport in Queens is one of the hubs in New York City, it is thus interesting to see how our model performs there, we observe that while it follows less the real data than in Midtown it still can capture some of the observable tendencies. We have a mean squared error of 1992.90 and a mean absolute error of 27.99 which are high, but explainable by the underestimated spikes at the beginning of the week. We also have an R^2 of 0.44.

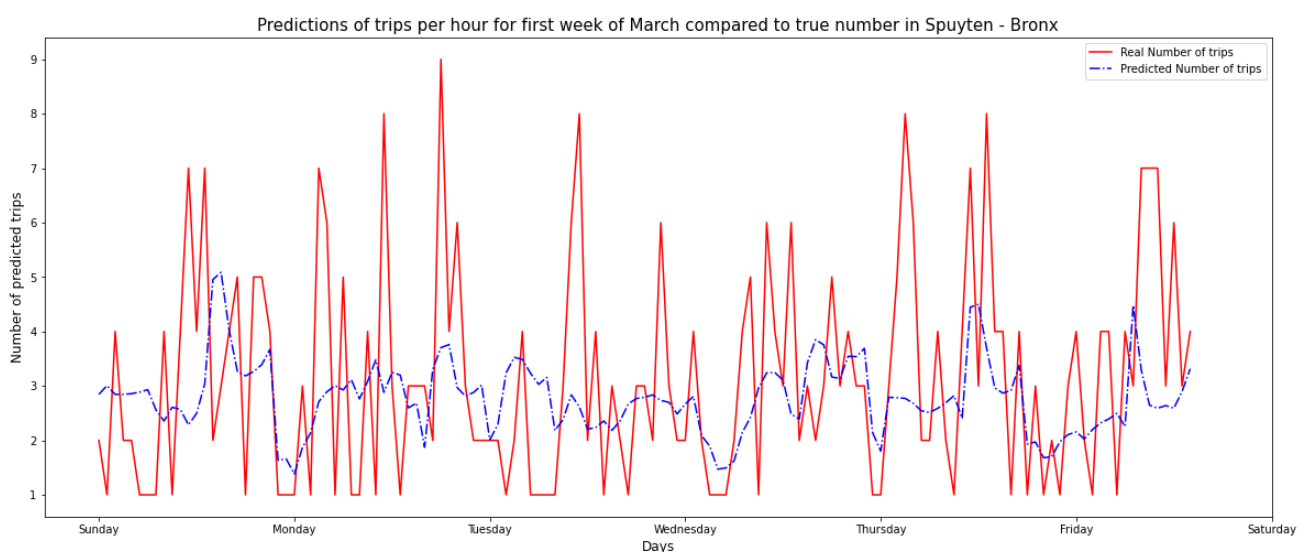


Figure 15 Model predictions for Spuyten

We have shown what our model is like on busy zones of New York, here above we have Spuyten Duyvil, a very quiet zone in the Bronx, as can be seen on the scale, the number of trips per hour is never above 10. We see that here our model has more issues as it clearly underestimates the variations in the data. This is especially reflected in the R^2 here which is 0.02. The error scores are quite low as the total number of trips is very low, with a mean squared error of 3.87 and a mean absolute error of 1.52.

8 Conclusion

To conclude, this study brings a new machine learning model to predict the Uber demand in New-York City. Yet, this model is exploitable only for estimating the demand in some zones of the city. The machine learning model the most appropriated for this issue is the random forest regression.

The predictions can be used by Uber to provide incentives for drivers. Indeed, if the company detects that there are not enough drivers present in the city at a given time when the demand is supposed to be high, the company can send notifications to drivers that are on a break so that they can work and take advantage of the demand. This notification for the drivers could be accompanied by a financial incentive.

An extension of this work could be to focus more precisely on Manhattan which is where most of the trips are originating, and try to use more precise data, like geographic coordinates if they are available to try and get closer results on location. The next step for Uber might also be to try and model trips with a starting point, as well as an arrival point, to better adjust prices, as well as locations for the drivers to go to.

9 List of references

9.1 Data sources:

Uber trips: <https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city?select=Uber-Jan-Feb-FOIL.csv>

Weather data: <https://openweathermap.org/>

Basketball events: <https://www.basketball-reference.com/>

Baseball events: <https://www.baseball-reference.com/>

Hockey events: <https://www.hockey-reference.com/>

New York suburbs: <https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq>

9.2 Literature references:

Brodeur, Nield (2018): “An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC”, *Journal of Economic Behavior & Organization*, vol. 152, issue C, 1-16

Wang, Hao, Wu, Qi, Barth (2048): “Predicting the Number of Uber Pickups by Deep Learning “, Transportation Research Board 2018, Volume: 18-06738.

Ye, Tianyu (2019): “Mind the Gap: a Case Study of Demand Prediction and Factors Affecting Waiting Time at Uber NYC and Washington D.C.”, UCLA Electronic Theses and Dissertations.

9.3 Image sources:

Fig 7:

https://www.researchgate.net/figure/Illustration-of-a-decision-tree-used-for-regression-Intermediate-nodes-represent-tests_fig2_337698181

Fig 8:

<https://medium.com/@jamesdacombe/an-introduction-to-artificial-neural-networks-with-example-ad459bb6941b>

Fig 9:

<https://www.innoarchitech.com/blog/artificial-intelligence-deep-learning-neural-networks-explained>

Fig 10:

https://www.researchgate.net/figure/fig-A10-Random-Forest-Regressor-The-regressor-used-here-is-formed-of-100-trees-and-the_fig3_313489088

10 Appendix

10.1 Appendix 1: Result visualization for Chinatown

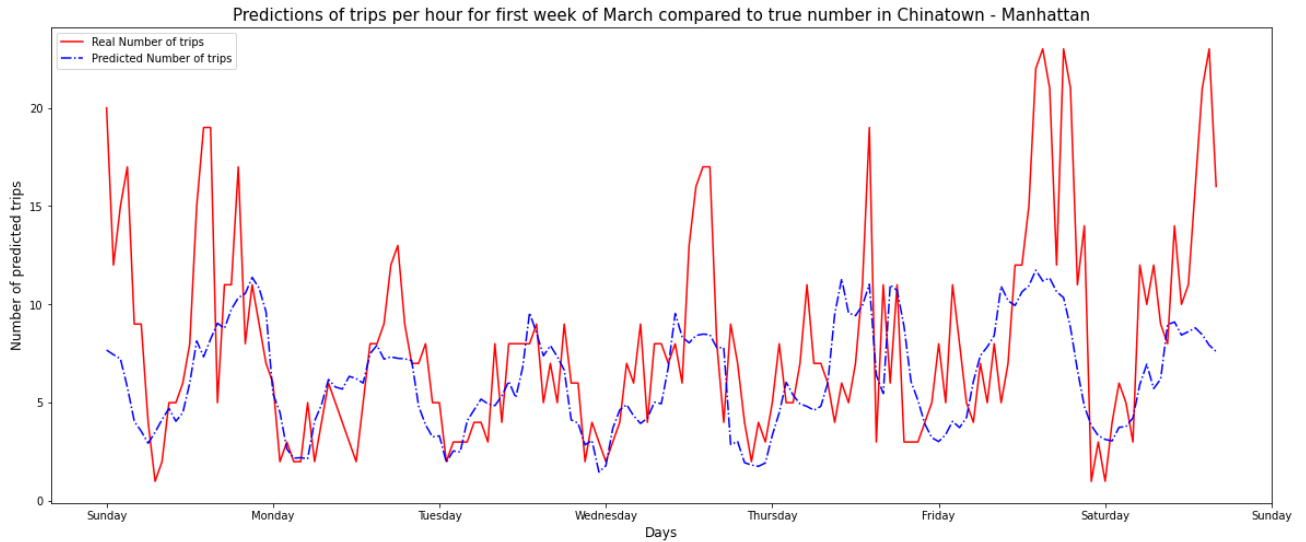


Figure 16 Model predictions for Chinatown

Chinatown model scores:

- MSE: 465.92280318992334
- MAE: 15.381032953395374
- R2: 0.5817835622914361

10.2 Appendix 2: Result visualization for East Village

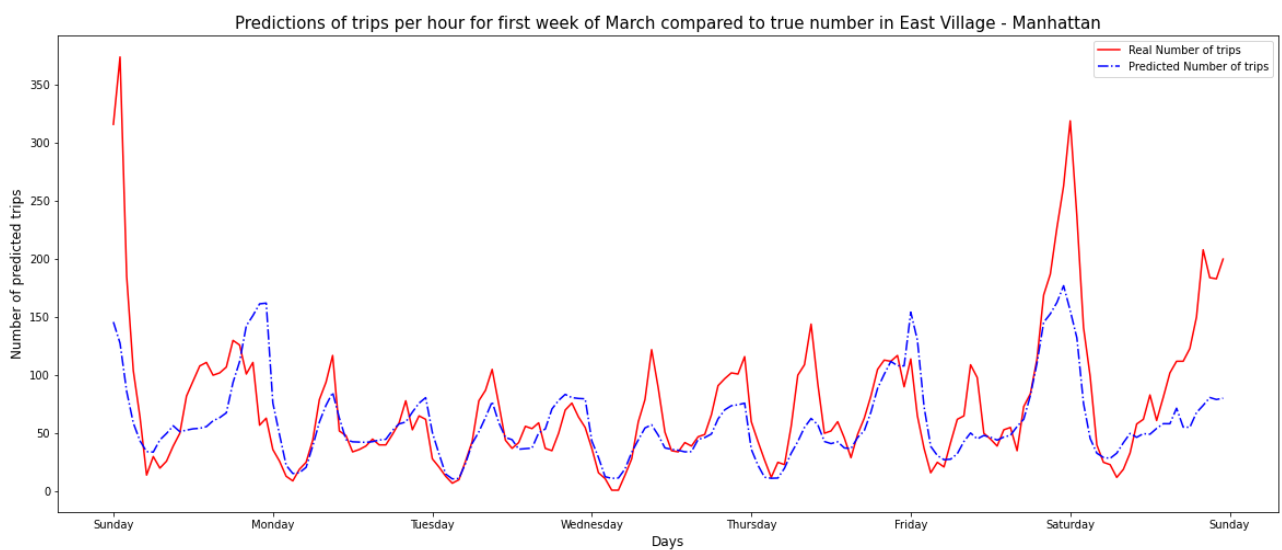


Figure 17 Model preiction for East Village

East Village model scores :

- MSE: 1992.898497415552
- MAE: 27.985246270177445
- R2: 0.43956574349481015