International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Uber Data Analysis Using Machine Learning

Syeda Hiba Fathima[1], Tharun M Kumaran[2], Sumedha N[3] Ms. Sapna R[4],

Final Year Students, Department of Computer Science Engineering, Presidency University, Bengaluru, Karnataka, India[1,2,3].

Assistant Professor Department of Computer Science Engineering, Presidency University, Bengaluru, Karnataka, India[4].

Abstract:

Uber is one of the fastest-growing companies in the world and is defined as a P2P platform. The platform that links you to drivers who can take you to your destination The paper explains the workings of an Uber dataset from Boston in 2018. In this paper, we experimented with a real-world dataset and explored how machine learning algorithms could be used to find patterns in the data. We mainly analysed the price prediction of different Uber cabs by using machine learning algorithms like linear regression, decision tree, random forest regression, and gradient boosting regression, and finally chose the one that proved best for the price prediction. We must choose an algorithm that improves accuracy and reduces overfitting.

Keywords: Machine Learning, Dataset, Price Prediction, Boston, Linear Regression, Decision Tree, Random Forest Regressor, Gradient Boosting Regressor, flask.

1. Introduction:

Uber Technologies, Inc., commonly known as Uber, was a ride-sharing company. It was founded in 2009 by Travis Kalanick and Garrett Camp, successful technology entrepreneurs. Together, they developed the Uber app, which connects riders and local drivers. The service was initially launched in San Francisco and eventually expanded to Chicago in April 2012, proving to be a highly convenient and great alternative to taxis and poorly funded public transportation systems. Over time, Uber has expanded into smaller communities and has become popular throughout the world. On December 13, 2013, the USA named Uber its tech company of the year.

In this project, we use supervised learning, where we have a training set and a test set. The training and test sets consist of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. We applied machine learning algorithms to make a prediction of price in the Uber Boston dataset. Several features will be selected from the 55 columns of the dataset. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large amounts of data.

The important packages used in the project are Pandas, Numpy, Seaborn, etc.

2. Literature Review:

The past few years have seen tremendous growth in Uber-related data analysis using machine learning. The rise of Uber as a global alternative has attracted a lot of interest recently. Researchers are coming up with various methods to analyse Uber-related data based on different factors. Our work on Uber's predictive pricing strategy is still relatively new. In this research, "Uber Data Analysis," we aim to analyse Uber's price. We are predicting the price of different types of Ubers based on different factors. Some of the other factors that we found in another study are:

Abel Brodeurand and Kerry Nield (2018) analyse the effect of rain on Uber rides in New York City. After Uber rides entered the market in May 2011, passengers and fares decreased in all other rides, such as taxi rides. Also, dynamic pricing makes Uber drivers compete for rides when demand suddenly increases, i.e., during rainy hours. With increasing rain, Uber rides are also increasing by 22%, while the number of

taxi rides per hour increases by only 5%. Taxis have not responded differently to increased demand in rainy hours than in non-rainy hours since the entrance of Uber. [1]

Junfeng Jiao's 2018 investigation of Uber's surge multiplier in Austin, Texas, found that during times of high usage, Uber will enhance their prices to reflect this demand via a surge multiplier. According to communications released by Uber in 2015, this pricing is meant to attract more drivers into service at certain times while also reducing demand on the part of riders. (Chen & Sheldon, 2016) While some research is mixed, in general, surge pricing does appear to control both supply and demand while keeping wait times consistently under 5 minutes. [2]

Anna Baj-Rogowska (2017) analyses the user's feedback from social networking sites such as Facebook in the period between July 2016 and July 2017. Uber is one of the fastest-growing companies in the so-called sharing economy. It is also a basis for the ongoing evaluation of brand perception by the community and can be helpful in developing a marketing strategy and activities that will effectively improve the current rating and reduce possible losses. So, it can be concluded that feedback should be an important instrument to improve Uber's market performance today. [3]

Ahmed, M., has shown that by using detailed data on taxis at the travel level and on the rental vehicle and data on complaints about the level of new complaints at the level of incidents, we study how Uber and Lyft have damaged the quality of taxi services in New York City. The overall effect of the organisations based on the scenario and the riding administrations was enormous and widespread. One of these effects is the expansion of the rivalry between Uber and Lyft over the quality of taxi administration. They use a new set of complaint data to measure the lack of quality of service that has never been analysed before. Focus on the quality dimensions generated by most of the complaints we demonstrate. The increased competition for these shared travel services has had an intuitive impact on the behaviour of taxi drivers [4].

Faghih, S.S., said that the demand for electronic mail services is growing rapidly, particularly in large cities. Uber is the first and most famous email company in the United States and New York City. A comparison between the demand for yellow and Uber taxis in New York in 2014 and 2015 shows that the demand for Uber has increased. To study the forecast performance of the models, you choose data for a typical day. Our goal in this document is to describe how these models can be used for forecasting Uber demand. The Uber data contains information about the position and time of the pick-ups and returns of each trip during a day. According to the available data, the Uber historical data for April 2014 [5].

Some papers make a comparison between the iconic yellow taxi and its modern competitor, Uber. (Vsevolod Salnikov, Renaud Lambiotte, Anastasios Noulas, and Cecilia Mascolo, 2014) identify situations when UberX, the cheapest version of the Uber taxi service, tends to be more expensive than yellow taxis for the same journey [6]. Our observations show that it might be financially advantageous on average for travellers to choose either Yellow Cabs or Uber, depending on the duration of their journey. However, the specific journey they are willing to take matters.

3. Proposed Methodology:

Based on the problems of forecasting errors and risk of overfitting due to large datasets. The data analysed and sent to the company is resulted as inefficient and ineffective. Thus, to overcome the problem we are going to predict the price of cab using Supervised Learning Machine Algorithm.

3.1 Flowchart:

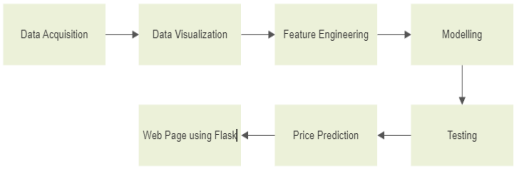

Figure 1. Data Flow of the project

3.2 Data Preparation:

The data we used for our project was provided on the Kaggle website. The original dataset contains 693071 rows and 57 columns, which contain the data for both Uber and Lyft. But for our analysis, we just need the Uber data, so we filtered out the data according to our purpose and got a new dataset that has 322844 rows

and 56 columns. The dataset has many fields that tell us about the time, geographic location, and climatic conditions when the different Uber cabs opted in.

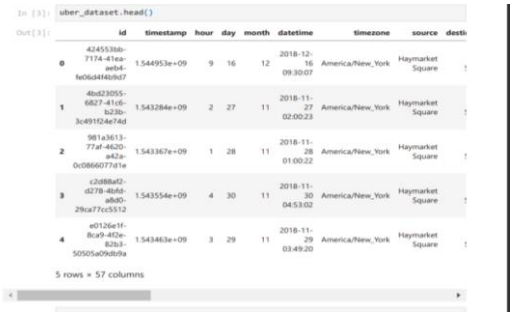

Figure 2. Data Head

3.3 Data Visualization:

Data visualisation is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualisation tools provide an accessible way to see and understand trends, outliers, and patterns in data.

To do this, we must import Matplotlib and Seaborn Library and plot different types of charts like strip plots, scatter plots, and bar charts.

3.3.1 Strip-plot:

A scatter plot that differentiates different categories and is very simple to understand. We have plotted different graphs, as mentioned below:

1) Strip-plot between Name and Price

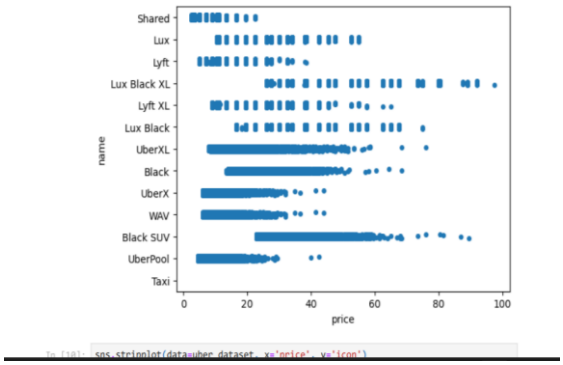

Figure 3. Strip Plot of Name and Price

From the above chart, Shared trip was cheapest among all, and Black SUV was most expensive. Taxi has no graph means no dataset.

2) Strip-plot between Icon and Price

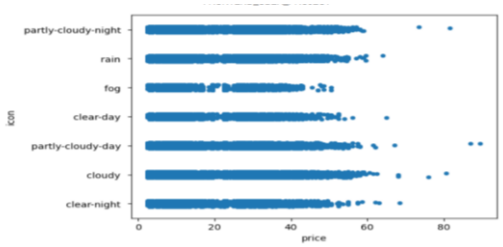

Figure 4. Strip Plot of Icon and Price

From the above chart, we analyse that in cloudy-day weather price was the highest while in foggy weather price was minimum.

3.3.2 Bar-plot :

A bar-plot or bar graph is a chart or graph that presents categorical data in rectangular bars to the values that they represent.

Representing the data in Bar-graph:

1) Bar-Chart of Month

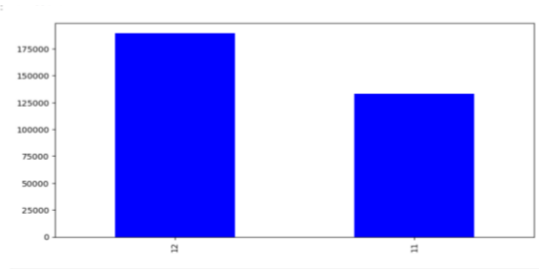

Figure 5. Bar-Chart of Month

2) Bar-Chart of Source

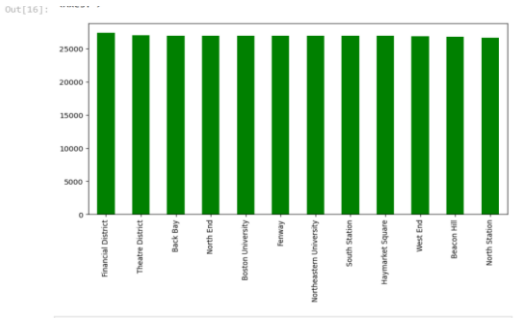

Figure 6. Bar-Chart of Source

3) Bar-Chart of Car Name

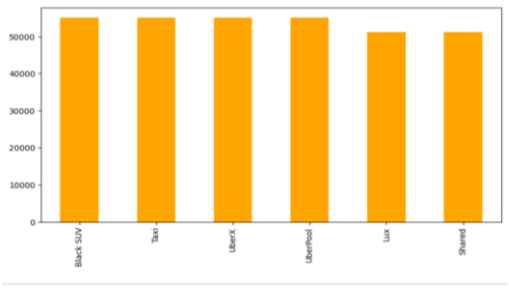

Figure 7. Bar-Chart of Car Name

4) Bar-Chart of Icon

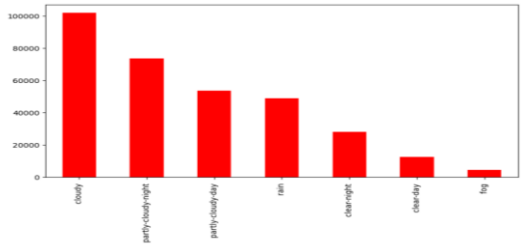

Figure 8. Bar-Chart of Icon

3.4 Feature Engineering:

The most crucial stage of data analytics is feature engineering. By picking the appropriate features for the model and getting the features ready in a way that is appropriate for the machine learning model, it enhances the performance of the machine learning model.

3.4.1 Label Coding:

To make categorical values machine-readable, label encoding refers to transforming them into numeric form. Since most machine learning algorithms don't understand data that combines categorical and continuous variables, they perform better when the data is represented as a number. As a result, we employ class mapping and label encoding to determine which categorical value is encoded into which numerical value.

3.4.2 Filling NAN Values:

We employ the function isnull() to examine missing data in a Pandas data frame. Here, we discover that our dataset's price column contains 55095 Nan values. The fillna() function is now used to fill these null values. Because price cannot be provided in a float, we convert missing values to integers and fill them with the median of the remaining dataset values. We now create a bar chart of the price-value count for visualisation purposes.

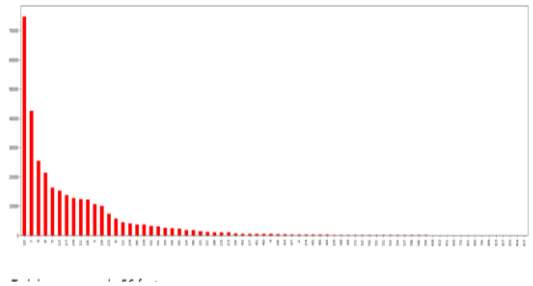

Figure 9. Bar Chart of Price

3.4.3 RFE (Recursive Feature Elimination):

It is a well-liked feature selection technique because it is simple to set up, straightforward to use, and efficient at choosing the features (columns) in a training dataset that are more or more useful for predicting the target variable.

When utilising RFE, there are two crucial configuration choices: the number of features to choose from and the algorithm that will be used to aid in feature selection. Recursive feature elimination is being implemented using Scikit-Learn via the Sklearn.feature_Selection.RFE class.

In order to apply RFE to our dataset using a linear regression model, we first split it into training and test sets. Then, we noticed that different numbers of features (k values) resulted in different accuracies, as shown below:

Serial No.	No. of Feature (K)	Accuracy
1	56	0.8054834220
2	40	0.8050662132
3	25	0.8055355151
4	15	0.8050457819

Figure 10. RFE Features

We can see that 25 features are the best features provided by RFE and have the maximum accuracy. Now that we are only using it for future work, our dataset has been reduced from 56 features to 25 features.

3.4.4 Drop Useless Columns:

The method drop () eliminates rows or columns in accordance with particular column names and associated axes. After using RFE, we have 25 of the best features, but there are still a lot of features that don't directly affect the price. We discard those, leaving eight features in the dataset.

3.4.5 Binning:

The process of binning involves converting numerical variables into their corresponding category counterparts. We define a range, often referred to as a bin, and any data value inside the range is made to fit into the bin, which is binning, during data smoothing (to make data appropriate) in this process.

After eliminating pointless features, some features are still out of range, so we use binning to bring all the features into range and obtain our final dataset, which is then used for

modelling.

	month	source	destination	product_id	name	surge_multiplier	icon	uvIndex
0	1	5	7	4	2	0	5	0
1	0	5	7	5	1	0	6	0
2	1	0	8	4	2	0	3	0
3	0	0	8	5	1	0	0	2
4	1	6	11	0	5	0	4	0

Figure 11. Final Dataset after Feature Engineering

3.5 Modelling:

A machine-learning algorithm is trained to predict labels from features, then tuned for business requirements and validated using holdout data. Using the provided dataset to train the model, we can then use that model to predict the response values for the same dataset and assess the model's accuracy.

Several models, including linear regression, decision trees, random forests, and gradient boosting, are implemented in this project using Scikit-Learn.

3.5.1 Linear Regression:

A supervised machine learning approach called linear regression produces continuous output within the specified range. It is a statistical method for modelling the connection between the characteristics of the input and the output. The qualities of the input are known as independent variables, while the characteristics of the output are known as dependent variables. By multiplying the input features by the output's ideal coefficients, we want to be able to predict the value of the output based on the features of the input.

3.5.2 Decision Tree:

A decision tree is a supervised machine learning algorithm used to categorise or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.

3.5.3 Random Forest:

Leo Breiman and Adele Cutler, the inventors of the widely used machine learning technique known as random forest, combined the output of various decision trees to get a single outcome. Given that it can deal with both classification and regression issues, its adoption has been fueled by its simplicity and adaptability.

3.5.4 Gradient Boosting:

A machine learning approach called gradient boosting is used, among other things, for classification and regression tasks. It provides a prediction model in the form of an ensemble of decision trees, like weak prediction models. Gradient-boosted trees is the name of the resulting algorithm when a decision tree serves as the weak learner. The modelling is done in the following steps: First, we divide the dataset into a training set and a testing set.
Then the model is trained using the training set.
Finally, we run the model on the testing set and assess how well it works.
Following the application of these models, we obtain the accuracy of:

Serial No.	Models	Accuracy
1	Linear Regression	0.747545073
2	Decision Tree	0.961791729
3	Random Forest	0.962269474
4	Gradient Boosting Regressor	0.963187213

Figure 12. Modelling Accuracy Table

3.5.5 K-fold cross validation:

The dataset is divided into a K number of folds during cross-validation, which assesses the model's performance when faced with fresh data. K is the number of groups into which the data sample is divided.

3.6 Testing for the proposed method:

A subset of the training dataset called testing is created to test all potential combinations and provide an estimate of how effectively the model trains. The primary goal of machine learning is to model data and anticipate the results using a variety of techniques. But we ought to pick the algorithm with the best accuracy. The correctness of the regression problem is assessed using mean squared error (MSE), mean absolute error (MAE), and root mean square error (RMSE). Both the mean_absolute_error technique and the mean_squared_error method from Sklearn can be used to implement these.

3.6.1 Mean Absolute Error (MAE):

The mean of all absolute errors is known as the mean absolute error (MAE). Similar to RMSE, MAE (which ranges from 0 to infinity; lower is preferable) just averages the absolute difference of the residuals rather than squaring the difference of the residuals and taking the square root of the outcome. The output of MAE is calculated by averaging the errors of all samples in a dataset.Hence, MAE = true values minus predicted values.

3.6.2 Mean Squared Error (MSE):

It is the mean of square of all errors. It is the sum, overall the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points. MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data.

3.6.3 Root Mean Squared Error (RMSE):

RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model. RMSE (ranges from 0 to infinity, lower is better), also called Root Mean Square Deviation (RMSD), is a quadratic-based rule to measure the absolute average magnitude of the error.

In our project, we perform testing on two models: Linear Regression and Random Forest.

Linear Regression Model Testing:

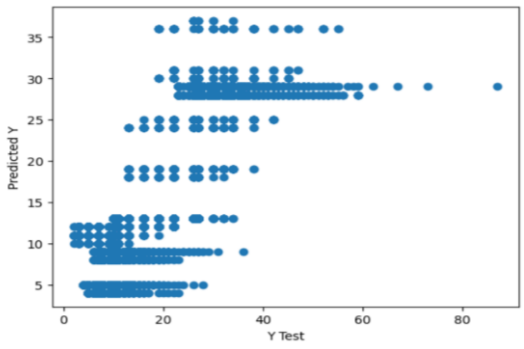

Figure 13. Scatter Plot for Linear Regression

After implementing the scatter plot between predicted and tested values, we find errors like MSE, MAE, and RMSE. Then, we represent a distribution plot of the difference between actual and predicted values using the Seaborn library. A distplot or distribution plot represents the overall distribution of continuous data variables.

Serial No.	Models	Accuracy
1	Mean Absolute Error	3.40607721
2	Mean Squared Error	20.0334370
3	Root Mean Absolute Error	4.47587277

Figure 14. Error table for Linear Regression

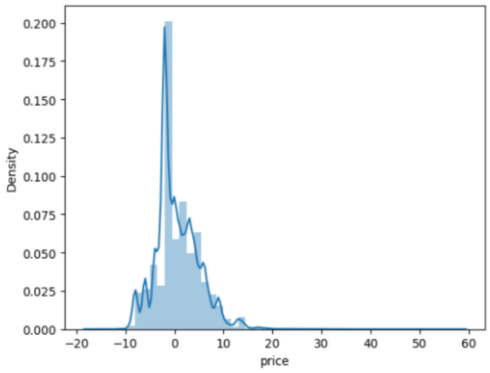

Figure 15. Dist Plot for Linear Regression

Random Forest Model Testing:

Similarly, we implement scatter plot, dist plot, and find all three errors for random forest also.

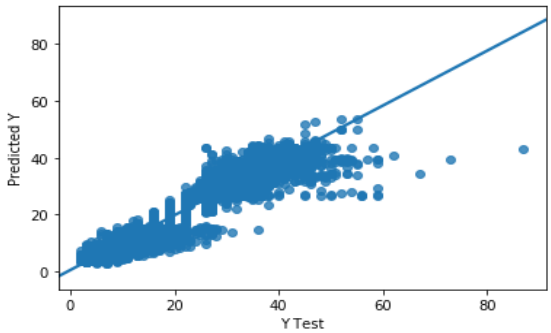

Figure 16. Scatter Plot for Random Forest

Serial No.	Models	Accuracy
1	Mean Absolute Error	0.99813700
2	Mean Squared Error	2.94465361
3	Root Mean Absolute Error	1.71599930

Figure 17. Error table for Random Forest

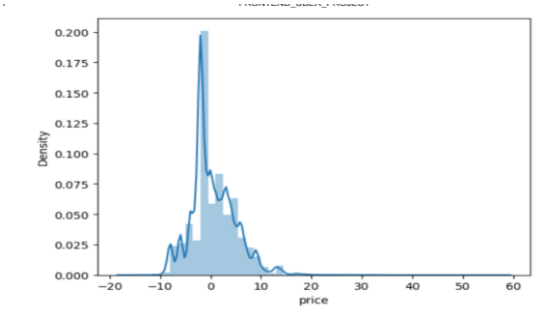

Figure 18. Dist Plot for Random Forest

3.7 Price Prediction Function:

After finding the errors for both the linear regression and random forest algorithms, we built a function named "predict_price," whose purpose is to predict the price by taking four parameters as input. These four parameters are cab name, source, surge multiplier, and icon (weather). As the dataset trains on continuous values and not categorical values, these values are also passed in the same manner, i.e., in integer type. We create a manual for users that gives instructions about the input, like what you need to type for a specific thing and in what sequence. In our function to forecast the price, we employ a random forest model. The first step is to find every desired row containing the input cab name and retrieve that row's number. The new dataset's length is then determined by creating an array x, whose initial values are all zero. We first create a blank array, and then we assign the source, surge multiplier, and icon input values to the appropriate indices. After that, we determine whether or not the count of all desired rows is greater than zero. If the requirement is met, we use the predict function and trained random forest algorithm to provide the price, set the index of the x array's data to 1, and return the price otherwise.

It somehow works like a hypothesis space because it gives an output for any input from input the space.

3.8 Web Page:

Here, we are going to design a UI with flask using HTML, CSS. Which will take the four parameters cab name, source, surge multiplier, and icon (weather) as input and predict the price.

4. CONCLUSION

Before working on features, we first need to know about the data insights that we get from EDA. Apart from that, we visualise the data by drawing various plots, due to which we understand that we don't have any data for the taxi's price, the price variations of other cabs, or different types of weather. Other value-count plots show the type and amount of data the dataset has. After this, we convert all categorical values into continuous data types and fill in Nan by the median of other values. Then the most important part of feature selection came, which was done with the help of recursive feature elimination. With the help of RFE, the top 25 features were selected. Among those 25 features, there are still some that we think are not that important to predict the price, so we dropped them and left with 8 important columns.

We apply four different models to our remaining dataset, among which Decision Tree, Random Forest, and Gradient Boosting Regressor prove best with 96%+ accuracy on training for our model. This means the predictive power of all three algorithms in this dataset with the chosen features is very high, but, in the end, we go with random forest because it is not prone to overfitting and design a function with the help of the same model to predict the price.

5. REFERENCES

[1] Abel Brodeurand & Kerry Nield (2018) An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC, Journal of Economic Behaviour & Organization, 2018, vol. 152, issue C, 1-16.

[2] Jun Feng Jiao (2018) Investigating Uber price surges during a special event in Austin, TX, Research in Transportation Business & Management Volume 29, December 2018, Pages 101-107.

[3] Anna Baj-Rogowska (2017) Sentiment analysis of Facebook posts: The Uber Case, 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS).

[4] Ahmed, M., Johnson, E.B. and Kim, B.C., 2018. The Impact of Uber and Lyft on Taxi Service Quality Evidence from New York City, available at SSRN 3267082.

[5] Faghih, S.S., Safikhani, A., Moghimi, B. and Kamga, C., 2017. Predicting Short-Term Uber Demand Using Spatio-Temporal Modeling: A New York City Case Study, arXiv preprintarXiv:1712.02001.

[6] Anastasios Noulas, Cecilia Mascolo, Renaud Lambiotte, and Vsevolod Salnikov (2014) OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs, arXiv:1503.03021.

[7] Modeling and Analysis of Uber's Rider Pricing Proceedings of the 2019, International Conference on Economic Management and Cultural Industry (ICEMCI 2019).

[8] Majaski, C, Uber vs. Yellow Cabs in New York City: What's the Difference? Retrieved from https://www.investopedia.com/articles/personal-finance/021015/uber-versus-yellow-cabs-new-york-city.asp, 2019.

[9] Chen, Y., Liu, Q., Chen, W., & He, J. (2018). Predictive modeling of Uber surge pricing using spatiotemporal analysis. ISPRS International Journal of Geo-Information, 7(11), 431.

[10] Ayodele, T., Chikalov, I., & Markov, I. (2019). Dynamic pricing model for Uber-like taxi service using multi-attribute utility theory. Transportation Research Part C: Emerging Technologies, 107, 1-14.

[11] Zhang, J., Wang, X., Zhao, Y., & Huang, Y. (2017). Taxi demand prediction with big data: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 19(12), 3815-3824.

[12] He, M., Xu, Y., Ding, X., Zhang, X., & Hu, Z. (2017). Taxi travel time prediction with LSTM recurrent neural

networks. IEEE Transactions on Intelligent Transportation Systems, 18(4), 847-856.
[13] Yang, Z., Wang, Y., & Wu, F. (2018). Uber demand forecasting with recurrent neural networks. Transportation Research Part C: Emerging Technologies, 89, 167-183.
[14] Gao, X., Zhu, X., Chen, Q., & Chen, D. (2019). Deep spatiotemporal residual networks for citywide crowd flows prediction. IEEE Transactions on Intelligent Transportation Systems, 21(8), 3214-3223.
[15] Borràs, J., Pérez-Bellido, A. M., del Castillo, J. M., & de la Prieta, F. (2019). Predicting prices in Uber-like systems using time series analysis. Applied Soft Computing, 79, 230-242.
[16] Khosravi, A., Nahavandi, S., Creighton, D., & Fidge, C. (2019). Real-time prediction of surge pricing in ride-hailing services using LSTM recurrent neural networks. Information Sciences, 491, 232-243.
[17] Yao, H., Wu, Y., Ke, X., & Chen, W. (2018). Taxi travel time prediction with data fusion and recurrent neural networks. Transportation Research Part C: Emerging Technologies, 92, 415-430.
[18] Xu, Z., Zhang, H., & Li, H. (2020). A hybrid model for taxi demand prediction using deep learning and gradient boosting. IEEE Access, 8, 45418-45429.