IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

End-to-End Predictive Analysis on Uber's Data.

Mani Chandana Kadarla
Computer Science and EngineeringVardhaman College of Engineering.
Hyderabad, India.

Mr.V.N.L.N Murthy
Assistant Professor Computer Science and Engineering.Vardhaman College of Engineering
Hyderabad,India.

Nanditha Kalal
Computer Science and EngineeringVardhaman College of Engineering.
Hyderabad, India.

Chaitanya Kumar Kattasavugari
Computer Science and Engineering Vardhaman College of Engineering Hyderabad , India.

***Abstract*—We have choosen this project as it mainly focuses on the Uber data, which has become one of the most trending apps these days. Companies have used data analytics to improve and expand their performance since decades. Visualisation and data analysis have helped us in many ways, including detecting new trends, examining correlations and patterns in the data, doing in-depth research, and, the cherry on top, drawing conclusions from these patterns. For all of the benefits that this notion offers, it is necessary that we study it in depth over time. Using Machinelearning algorithms in making the right analysis from the data ,that helps in making a decision.The varying cost of Uber rates is probably going to be significantly influenced by the weather. Various climate factors will have different impacts on the price rise. and at various levels: We believe that weather conditions like cloudiness or clarity do not have the same impact on inflation rates as weather conditions like snow or fog. Addressing the day of the week, recognising weekends and weekends is essential as individuals frequently engage in distinct activities, visit different locations, and retain a different mode of transportation during weekends and weekends.**

***Index Terms*—Machine Learning , Data visualizations, Linear Regression, Decision Trees, Random Forest, Gradient Boosting.**

I. INTRODUCTION

Uber is the market leader in the provision of new transporta- tion choices in the modern era. As a result, Uber's core busi- ness is networking, and all of the company's recent initiatives may be seen as simply providing a platform for the pertinent supply and the noteworthy demand to collide. People use analytics, a field that is rapidly expanding, in their businesses to help them succeed. This project focuses primarily on data visualisation and will improve our grasp of how to use the machine learning algorithms to comprehend the data and to cultivate an intuition for comprehending the travellers.[1] As we are considering the NYC uber's data of the prediction of price and availability of the cabs.As the large no.of usage isin NYC area , the larger data involves the complexity of the

analysing and predicting the features. Understanding market segmentation, also known as customer segmentation, might help with this problem. A method of customer segmentation characterised as a game where a child sorts balls and cubes according on their colours or shapes. Customer segmentation,to put it simply, is the process of separating customers into different groups and markets according to various criteria and traits. The Uber data is not as precise as the taxi data, and oddly enough, Uber only offers time and location for pickups and not drop-offs. Nevertheless, I wanted to offer a combined data set that included all of the taxi and Uber data that was currently accessible. Uber determines areas of the city that have incredibly high demand by analysing historical data. The usage of uber cabs has a rapid increase in the world. It considers all the factors effecting the usage. The factors that are effecting are the weather conditions , pollution, traffic conditions, price. Data analysis is essential to business growth in this fiercely competitive environment. reports on data analysis, different kinds of analysis, and Business must generate report documents so that they have references for atypical activities and initiatives, notably when making a choice for the company's future operations. The Uber pickup data and the meteorological data should be merged in the Excel files for the analysis. If you can organise all the data according to the activity you'll be doing, you can

construct a data analysis accordingly. We studied the process of producing visualization tool using the Uber data analysis Python project. The only transportation firm to evaluate and disseminate actualsustainability statistics is Uber.

II. LITERATURE REVIEW

In this paper we have performed various machine learning algorithms in the prediction of the data and the future scope of getting the accurate analysis on the data.On comparing the accuracy obtained from the machine learning algorithms it makes us to choose the best algorithm that helps in predicting the future of the company.

Mostly the classification and regression algorithms are used while predicting the future. As we know there are numerous algorithm that provide prediction. Mostly used regression models are Linear Regression , logistic regression , K-means , KNN algorithm and Decision Trees and Random Forest algorithms and more.

The prediction is done based on the most usage of the cabs in particular area,in an hour,in a day,in a week,in a month and predicting the price accordingly by considering the maximum features for increase in the usage. This helps in analysing the price.

On analysing the different researches held on uber's data , the approaches have been different based on the requirements.Some of them are held on the data storage ,some of them are on the development on the uber facility.

1) The same projects are interested in two related papers: one on urban transportation development and the other on case studies of Uber data in various cities. Ggplot2 has been around for more than ten years and has been used by millions of people to create plots. It is a data visualization-focused R package.The quality and appeal of your graphics can be significantly increased, and you'll be able to create them with a great deal more speed. The construction of practically any type of chart is possible with ggplot2. It is a framework for declaratively image editing and is based on the grammar of graphics. The original plot object is created using ggplot, which is almost always followed by + to add additional partsto the plot.

As we know, that uber required a large amount of memory to store the data , so as to analyze the large, store the data and retrieve the data, the traditional approach has its limitations. So as to provide the fastest accessibility and easy storage,Map Reduce concept has been introduced.

2) MapReduce is picking up steam because the Apache Hadoop and Spark parallel computing platforms enable programmers to utilise mapReduce to run models over enormous distributed knowledge sets and employ cutting- edge applied math and machine learning approaches to attempt predictions, discover patterns, uncover correlations, etc.

As we know that , uber nearly considers all the factors that effect the future scope and it's important to maintain the factors at an accurate points so to reach the breakeven points. Mainly, considering the factors such as weather , price and availability of cabs in a particular area , the predictions has been done.

The leads to the spread in evaluating Uber's future potential is price research. New York is the ideal testing setting due to its sizable market for online auto services, intricate traffic patterns, and astonishing diversity in terms of culture, economy, and functionality.[3] Studying Uber pricing and the different elements that influence it in NYC is not a standalone study, but rather a mission that offers insights for metropolitan places all around the world. These insights will continue to be crucial as urban development and population decline.

Online transit agencies like Uber and Lyft are an example of the "Sharing Economy," which is an essential balance for the limited room and resources in urban areas called on by the rise of modernization. Despite the unprecedented advancements and gains this new era has brought about in many areas, new economics theories and models must be developed to address the numerous problems it encounters. In order to find out how much and how the urban lifestyle, rush hours, and commute patterns impact online vehicle services, field studies must be conducted and novel mathematical and mathematical models must be established.

III. PROPOSED WORK

As the previous predictions on the price are done using the basic machine learning algorithms which are less effective and have the least the prediction and the future scope isn't clear. Machine learning provides the effective prediction on every large data.The study details about the various types and the algorithms that are effectively used.

A. Machine Learning

With in broader subject of artificial intelligence, machine learning is a sub field that gives accurate models to develop predictions. It is usually referred to as a type of predictive modelling or predictive analytic, and historically, it has been described as a computer's capability to learn without being personalised to do so.

B. Machine intelligence Types

There are three distinct categories of methods used for machine learning i.e., Supervised, Unsupervised and Reinforcement learning.These provide the techniques to the data analysts to learn the different methods and predict the future accordingly.

1) Supervised Learning: The machine is trained on a set of data inputs and outputs with the objective of comprehending a general rule that maps the given inputs to the given outputs. The two fundamental sub types of supervised learning are classification, which comprises projecting a class label, and regression, which entails deriving a numerical value.

2) Unsupervised Learning: Instead of getting such type of direction, the learning algorithm works to individually recognize the pattern or structure in the data. Unsupervised learning can be divided into two primary categories: clustering, which involves defining groups in the

data set that have similar traits, and density estimation, which requires studying the data set's statistical distribution. The data can be represented using unsupervised learning techniques, and projection can be used to simplify the data by lowering its dimensions.

3) Reinforcement Learning: A issue will be provided to the computer and its algorithms in a dynamic environment, and as they try to achieve a specified goal, they will receive advice (rewards), which will support their goal-seeking and learning efforts. Reinforcement learning, which uses algorithms like Q-learning, temporal- difference learning, and deep reinforce- ment learning, is illustrated through the game Alpha-Go.

C. Regression

A statistical model recognised as regression develops the relationship between the independent and dependent variables. There are independent and dependent variables while using regression. The dependent variable, which is a feedback vari- able, can be determined using the independent variable. The independent variable is also known as the predictor variableas a response.

1) LINEAR REGRESSION: Linear Regression is a su- pervised learning that builds the linear relationship between the independent and a dependent variable. The further study of linear regression involves the types of linear regression models.Efficiently the Simple and Multiple linear regression models are the most used algorithms.

Simple Linear Regression model performs between the sin- gle independent variable and the single dependent variable. Significantly, it is denoted as

$$y = ax + b \qquad (1)$$

where independent variable x and dependent variable y are present.

Multiple Linear Regression model defines the complex re- lation between the multiple dependent variables and a single independent variable.

$$y = ax_1 + bx_2 + \ldots \qquad (2)$$

The large data is considered to analyse the atmost prediction from it and assist in making the appropriate decision.

In this section,we have explained the proposed method the project which gives the accurate prediction.The following steps helps in processing the work. The following steps are:

1) Data collection
2) Data cleaning
3) Data Segmentation
4) Model Training
5) Prediction and analysis.

D. Data Collection

As the work defines the uber cabs details that gives the complete details of a ride. The dataset consists of the start date - which determines the date and time of booking the cab,end date - which determines the drop off time and date of the ride, category- which determines the personal and the professional use of the ride,start place - pickup location,stop place-dropping location , miles - distance travelled and purpose - travelled based on the work.

As the analysis is on huge dataset , the dataset has been divided into two categories , which determines the basic details and the other gives the information about price , weather conditions such as wind, temperature, humidity etc. This gives a detailed explanation on the dataset providesefficient information to analyze the patterns in the data.

```
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   START_DATE  1156 non-null   object
 1   END_DATE    1155 non-null   object
 2   CATEGORY    1155 non-null   object
 3   START       1155 non-null   object
 4   STOP        1155 non-null   object
 5   MILES       1156 non-null   float64
 6   PURPOSE     653 non-null    object
dtypes: float64(1), object(6)
```

Fig. 1. The columns present in the dataset.

E.Data Cleaning

- As the data differs everyday, it is highly impossible to find the complete relation at a time. So by dividing the data into minor segments , we can identify the relations easily.
- Any data that has been stored in a dataset it set to have null values, some errors. These errors can create a disturbance in finding

the patterns and doesn't effect in giving the accurate prediction.

- Data cleaning, which entails locating and deleting any missing, duplicate, or irrelevant data, is an essential stage in the machine learning (ML) pipeline. Data accuracy, consistency, and error-freeness are the main objectives of data cleaning since inaccurate or inconsistent data might impair the performance of the ML model.
- It improves the quality of the dataset.

F. Data Segmentation

- The enormous data that contains all the information of every ride makes the pattern identification difficult and makes the process complex.So as to make it easy , the segmentation of the data is used.
- Every detail of the cab ride is taken into consideration. Like maximum usage of the cabs in the area , at what time of the day , what day of the week.
- The distance of the ride is also calculated.This is gives the result mostly cabs are used to travel near by loca- tions.Mostly people prefer cabs to avoid pollution and traffic so they cabs for even a small distance.
- The data also provides the information on which area the maximum usage of the cabs is identified. It also helps in allotting the required number of cabs in that particular area.
- Most of the cabs are used based on which purpose.

The data segmentation is represented using the graphs.Generally in python language, the graphs are represented by implementing the matlabplot.It provides allthe information graphically , which makes the interpretation easy. It also helps in easy prediction of the result.

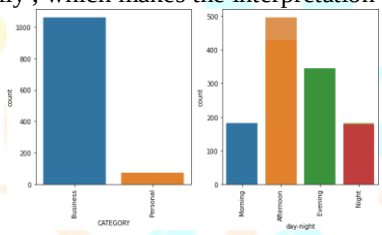


Fig. 2. Gives the category of the maximum usage of cabs and at what timeof the day.

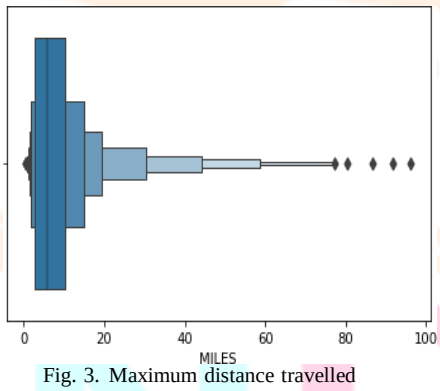


Fig. 3. Maximum distance travelled

The segmentation provides the deeper knowledge on the dataand gives the optimized data and effective communications.

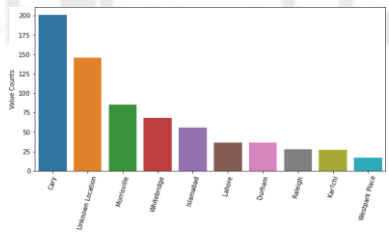


Fig. 4. Usage of cabs in particular area.

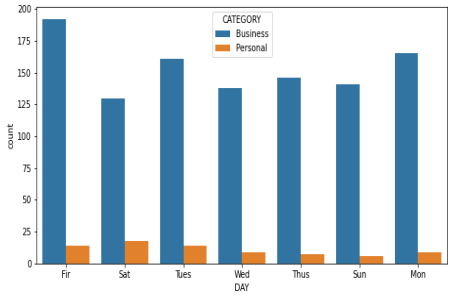


Fig. 5. On which purpose

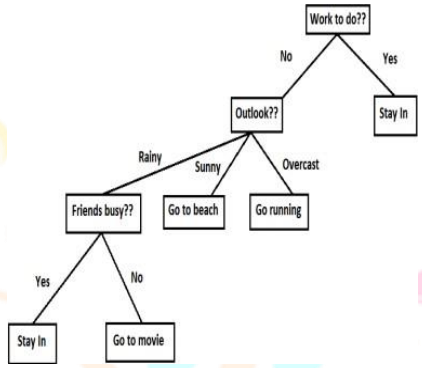


Fig. 6. General example of making a decision.

G. Model Training

- The machine learning is a vast language that itself contains a larger algorithms that works in a way more precised way.The most effective predictions are done using Decision Trees, Random Forest, Gradient Boost.
- These are the regression methods for analysing the data.
- These regression models have their own process flow which create a major difference while predicting.

1)DECISION TREES:

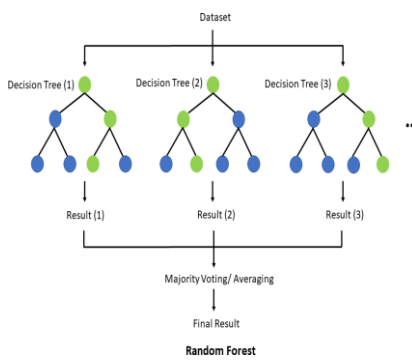
- This is a Supervised learning which performs on a supervised learning.
- Decision Trees are the graphical representation the datain a tree structure where the decisions are made out ofthe given conditions.
- The decision node and the leaf node are the two typical elements of the tree.

Decision Node - It can have several branches and is used to make choices.
Leaf Node - It is the outcome of the picks andlacks additional lines.

Advantages

- It is simple to understand because it follows the same logic that a person would use to reach any choice in the real world.
- It can be very beneficial for problems regarding decisions.
- It is advantageous to take into account every situationthat might be caused by a problem.
- It requires less data cleaning than other algorithms.

While performing decision, the data has to be more clearwithout any errors.So the decision are accurate.
Here , our project mainly works on the price of the uber data, so as to know the highest accuracy after considering all the factors that effect the price the decision is finalized.
The features that effect the price are weather conditions such as humidity , sunny , outlook , rainy etc., traffic , pollution , distance. So these help in calculating the price for a particular ride.



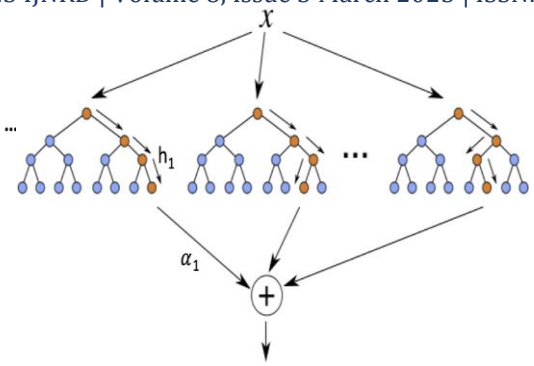


Fig. 7. Random forest example Fig. 8. Gradient boost algorithm

2) RANDOM FOREST: The random forest utilizes the prediction from each decision tree and makes its forecast of the ultimate result on the majority votes of predictions rather than relying solely on one decision tree. Random Forest is a classifier that uses multiple decision trees on various sections of the supplied dataset and averages the outcomes to enhance the dataset's predictive accuracy.

Advantages

- This helps in reducing the over-fitting.
- Random Forest takes a way more minimum time in comparing the features.

Generally , the random forest is a extension of decision.The large data set it divided into smaller groups of data.On each group the algorithm is applied.This works precisely on the data.It is used both classification and regression models.

Steps involved in building the model:

- 1. Select a random K points in the dataset.
- 2. Create the decision trees connected to the selected datapoints (Subsets).
- 3.The decision tree N that you want to construct shouldbe chosen.
- 4.Repeat step 1 and 2
- Locate each decision tree's predictions and group them according to the group that gets the most votes when dealing with new data points.

3) GRADIENT BOOST ALGORITHM: One popular method of learning ensemble modelling strategies called "boosting" is used to create strong classifiers from a varietyof weak classifiers.

It starts by developing a primary model using training data sets that are readily available, and then it finds any errors in the base model. A secondary model is created when the error has been located, and a third model is then added to the process.

In this manner, adding additional models is continued until we have a complete set of training data from which the model can accurately predict.

In the history of machine learning, AdaBoost (Adaptive boost-ing) was the first boosting method to merge many weak

classifiers into a single strong classifier. It focuses mostly on categorization problems, such binary classification.

Steps involved in Gradient Boost are:

- Imagine starting a dataset with a variety of data elements.
- Give each data point the same weight at this time.
- Let's say that this weight is one of the inputs to thealgorithm..
- Locate the data elements that were incorrectly classified.
- In step 4, make the data elements' weights heavier.
- If the result matches your expectations, end the process;if not, repeat steps 2 and 3.

The major difference between the Decision trees and Random Forest are:

Decision trees are much easier when compared to a random forest. A decision tree combines some choices, whereas a random forest combines many decision trees. Consequently,it is a lengthy and complicated procedure.

A decision tree is quick and effective with large data sets, particularly a linear one. The random forest algorithm requires a lot of training. When setting up an undertaking, more than one model may be needed. As a result, random woodlandsare more common.

The major difference between the Random Forest and Gradient Boost are :

Random forests and gradient boosting trees are distinct from one another in two significant respects. We systematically train the earlier, fixing the flaws of the prior trees with each new tree. In comparison, we construct every tree in a random forest independently. As a consequence, while we can train a forest in parallel, we are unable to train the gradient-boosting trees in parallel.

The other significant distinction is how choices are made. Each tree in a random forest is independent, allowing the trees to select any sequence for their outputs. On the other hand, the gradient boosting trees run in any sequence that is predetermined and cannot be

changed. As a consequence, they only accept sequential evaluation.

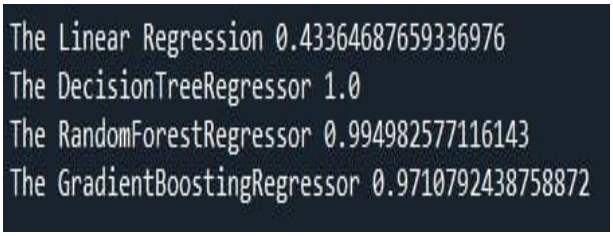


Fig. 9. The accuracy score of the algorithms

H. Prediction and Result

On applying various machine learning methods to the Uber dataset for analysis. methods for decision trees, random forests, and gradient boost. These algorithms use the accuracy number to predict the outcome after training on the dataset.
After training the dataset , each of the algorithm gives a prediction score on the price.

The result shows that the linear regression has the least prediction from the dataset because it performs the basic operations and the errors are not neglected.
Each of the algorithm Decision tree , Random Forest and Gradient boost divides the dataset and analyzes the data without errors and predicts with accurately.
As seen Decision trees and Random forest almost predict the accuracy nearly equally. As each node is considered in taking the decision , the prediction is almost perfect and helps in making the exact decision.

IV. Future Scope

1. This project can further be extended for the entire uber data, as we have only predicted and analysed the data from only one particular area.
2. This project can be used to get the uber data and analyse iton a particular scenario.
3. This also helps in estimating the price of uber based on thesource area and destination area.
4. It also helps in predicting the number of drivers required for a particular area.columns. To determine the best number based on results, we will test training these four models on 5, 10, 15, 20, and all25 features. The accuracy's obtained are

1. Linear Regression: 0.43(approximately)2.Decision Tree: 1
3.Random Forest: 0.99(approximately) 4.Gradient Boosting: 0.97 (approximately)
From the above mentioned information, Decision trees aremore preferred than any other machine learning algorithms.

References

[1] https://ggplot2.tidyverse.org/
[2] Majaski, C, Uber vs. Yellow Cabs in New York City: What's the Dif- ference? Retrieved from https://www.investopedia.com/articles/personal- finance/021015/uber-versus-yellow-cabs-newyork-city.asp, 2019
[3] https://www.researchgate.net/publication/338438722 Modeling and Analysis of Uber's Rider Pricing
[4] . Banu GR. A Role of decision Tree classification data Mining Tech- nique in Diagnosing Thyroid disease. International Journal of Computer Sciences and Engineering. 2016;4(11):111-5.
[5] https://www.researchgate.net/publication/330138092 Study and Analysis of Decision
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] Uber Technologies, Inc., "Facts and figures," 2018.
[8] A. Ley and P. Newton, "Creating and sustaining liveable cities," in developing living cities: From analysis to action. World Scientific, 2010,pp. 191–229.
[9] R. Cervero, Transit-oriented development in the United States: Experi- ences, challenges, and prospects. Transportation Research Board, 2004, vol. 102.
[10]https://growvation.com/paritoshsankhla/project/uber-data-analysis/5e95ee80-9455- 4473-acaf-b670fe2abc8b
[11]]https://www.skyfilabs.com/project-ideas/uberdata-analysis
[12]]https://iedu.us/tag/project-in-r-uber-dataanalysis-project/

V. CONCLUSION

Uber's usage between September 2014 and August 2015 was the subject of a dataset published by the NYC Taxi and Limousine Commission (TLC) at the start of 2017. The data includes characteristics that are different from those in the collection that was previously made public and in-depth examined by FiveThirtyEight and the Kaggle community.
This project gives the clear description on "End-to-End predic- tive analysis on Uber Data". The Classification and Grouping methods are the only ones that the current model emphasises. However, we will train four ML models, including gradient boosting, which outperforms the current model in terms of accuracy, and compare their results. Moreover, a feature se- lection method like RFE would be beneficial for performing the best analysis given that our dataset contains 25 feature