International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

A NOVEL APPROACH TO ANALYZE UBER DATA USING MACHINE LEARNING

[1]Akula Prasanth Kumar, [2]Kanchi Aashritha, [3]Marri Johnwesley,
[4]Jonnadula Narasimharao
[1,2,3] B. Tech Students, Department of Computer Science and Engineering,
CMR Technical Campus, Medchal, Hyderabad, Telangana, India.
[1]prashanthkumar2701@gmail.com, [2]aashritha0923@gmail.com, [3]Johnwesleymarri143@gmail.com,
[4]Assistant Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India.
jonnadula.narasimharao@gmail.com

ABSTRACT: Uber is a digital aggregator application platform, connecting passengers who need a ride from one place to another with drivers that are willing to serve them. Riders create the demand; drivers supply the demand and Uber acts as the facilitator to make this happen seamlessly on a mobile platform through its engineering. Data analytics has helped companies optimize and grow their performance for decades. Data analytics and visualization has aided us with several benefits, few of them being identifying emerging trends, studying relationships and patterns in data, analysis in depth and cherry on top are the insights we draw from these patterns. It is requirement of time that we study these concepts in thoroughly for all this benefits it provides. Hence in this work, a Novel approach to analyze uber data using Machine Learning is presented. Uber Data Analysis task permits us to recognize the complicated facts visualization of this large organization. It is developed with the assist of python programming language. Uber Data Analysis project enables us to understand the complex data visualization of this huge organization and it also help us to understand the about the Architecture, models, and implementation of fair price prediction in uber by applying the Linear Regression and Random Forest Regression Algorithms.

KEYWORDS: Uber, Big data, Data analysis, Real time analysis, Machine Learning, Linear Regression.

I. INTRODUCTION

Data is crucial in today's business and technology environment. There is a growing demand for Big Data applications to extract and evaluate information, which will provide the necessary knowledge that will help us make important rational decisions.

These ideas emerged at the beginning of the 21st century, and every technological giant is now exploiting Big Data technologies. Big Data refers to huge and broad data collections that can be organized or unstructured. Big Data analytics is the method of analyzing massive data sets to highlight trends and patterns. Uber uses real-time Big Data to perfect its processes, from calculating Uber's pricing to finding the optimal positioning of taxis to maximize profits. The Uber platform connects you with drivers who can take you to your destination or location.

Uber has emerged as leading company in the provision of new transportation options within the contemporary world. Uber, then, is primarily in the business of networking, and all the company's emerging operations can be conceptualized in terms of simply providing a medium through which the relevant supply can meet up with the relevant demand [1]. With the assist of visualization, businesses can avail the advantage of appreciation the complicated information and acquire insights that would assist them to craft decisions. This assignment is associated to large facts as we are inspecting very big quantity of records to be aware of the instinct of uber customers. This challenge will exhibit the plotting of daily, month-to-month and every year rides of uber in a complete city. On-demand, app-based trip

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

offerings like Uber and Ola have end up an necessary section of today's transportation gadget with its exibility and speedy responsiveness. Compared with regular taxicabs, Uber- like taxis have loggers to reveal and document time out statistics such as pickup place and outing distance, which can be a precious facts supply for information discovering [3].

According to IBM reports, every day "2.5 quintillion bytes of data" is generated. Furthermore, Uber provides over 20 million rides within a single day while all the processing is done with real-time analysis. Therefore the performance is very important for these businesses. Yet, processing of the Big Data analysis is a very challenging task for companies. Distributed computation and parallel processing methods can also simplify the processing of massive data because they vary from conventional approaches. It eliminates latency and data rate constraints. Not only Big Data, the distributed environment also very important for real-time Big Data analysis.

Uber's data science department also performs an in depth analysis of public transit networks in various cities so that they can concentrate on cities with weak transit systems and make effective use of the data to improve customer service experiences. Uber is currently using Big Data for different scenarios. Among them, identifying popular Uber locations is very important for the efficiency of their business. If they can efficiently and accurately predict the popular Uber locations, they can increase their profit. There are different areas and different technologies to study for this scenario such as Batch Processing, Stream Processing, Docker and Kubernetes, Spark, etc.

Real-time data analysis is very challenging for the implementation because we need to process data in real-time, if we use Big Data, it is more complex than before. Implementation of real-time data analysis by Uber to identify their popular pickups would be advantageous in various ways. It will require high-performance platform to run their application [5].

Machine learning is a branch of artificial intelligence that is concerned with the construction of models from data & presenting it in an understandable format. It is an application of artificial intelligence that deals with training the machine based on algorithms making it self-learn a pattern of solving problems. The machine then gets trained with a new input data given to it and trains the new data to the already trained data. Recently, the field of machine learning has seen a rise in the popularity of probabilistic and statistical models. [2].

Hence in this work**,** a novel approach to analyze uber data analysis using machine learning is presented. The rest of the work is organized as follows: The section II describes the Literature Survey. The section III demonstrates the novel approach to analyze uber data analysis using machine learning. The section IV details the result analysis of presented approach. The section V presents the conclusion.

II. LITERATURE SURVEY

Shashank H et. al., [6] presents Data Analysis of Uber and Lyft Cab Services. This project gives us basic understanding of how one can use machine learning in order to predict the cab fare from given source to destination before starting the cab ride. The model created is able to give us the predictions which are not exactly equal to the actual the price fluctuation is around the difference of ten to twenty rupees compared to the actual price. Since the model is good but not the best, one can improve the predictions of the model by using the Fine-tuning technique. If fine

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

tuning is applied to the existing model, then we are able to get higher accuracy than this model.

P. Devika, Y. Prasanna , P. Swetha ,G. Akhilesh Babu et. al., [7] presents Uber Data Analysis using Map Reduce. Uber Data is used for analyzing the vehicle with most popular trips. As map reduce is used to process huge amounts of data, we are using map reducing model to analyze uber data and give insights about the most used vehicle, number of trips it has covered. The main objective of this project is to investigate no of trips so as to produce data for the company to take care of the records and helps to company in creating huge information for long run endeavor.

Daksh Shah, Aravinda Kumaran, Rijurekha Sen, Ponnurangam Kumaraguru et. al., [8] describes Travel time estimation accuracy in developing regions: An empirical case study with Uber data in Delhi-NCR. This paper investigates the quality of travel time estimates in the Indian capital city of Delhi and the National Capital Region (NCR). Using Uber mobile and web applications, we collect data about 610 trips from 34 Uber users. They empirically show the unpredictability of travel time estimates for Uber cabs. Authors also discussed the adverse effects of such unpredictability on passengers waiting for the cabs, leading to a whopping 28.4% of the requested trips being cancelled. The empirical observations differ significantly from the high accuracies reported in travel time estimation literature.

Nikolay Laptev, Jason Yosinski, Li Erran Li, Slawek Smy et. al., [10] presents Time-series Extreme Event Forecasting with Neural Networks at Uber Motivated by the recent resurgence of Long Short Term Memory networks, a novel end to- end recurrent neural network architecture is presented that outperforms the current state of the art event forecasting methods on Uber data and generalizes well to a public M3 dataset used for time-series forecasting competitions.

III. A NOVEL APPRAOCH TO ANALYSE UBER DATA

A novel approach to analyze uber data analysis using machine learning is presented in this section. The Fig. 1 shows the system architecture of presented a novel approach to analyze uber data analysis using machine learning.

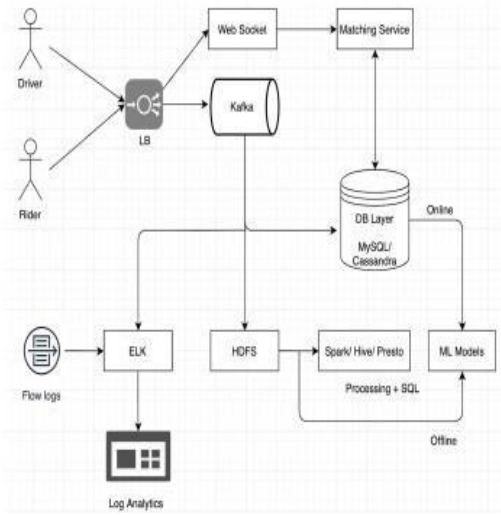


Fig. 1: System Architecture

This work is about on world's largest taxi company Uber inc. In this assignment, we're looking to predict the fare for their future transactional cases. Uber delivers service to lakhs of customers daily. Now it becomes really important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes really important to estimate the fare prices accurately.

Predict the price of the Uber ride from a given pick up point to the agreed drop off location. Perform following tasks: i) Pre-process the dataset, ii) Identify outliers, iii) Check the correlation, iv) Implement linear regression and random forest regression

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

models. Evaluate the models and compare their respectivescoreslikeR2, RMSE, etc.

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. A real-world data generally contains noises, missing values, and may be in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. It involves below steps: i) Getting the dataset; ii) Importing libraries; iii) Importing datasets; iv) Finding Missing Data; v) Encoding Categorical Data; vi) Splitting dataset into training and test set and vi) Feature scaling.

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to in consistent data entry, or erroneous observations. To ensure that the trained model generalizes well to the valid range of test inputs, it's important to detect and remove outliers.

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast tour target variable. Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.

Positive Correlation: Two features (variables) can be positively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) also increases.

Negative Correlation: Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increase then the value of the other variable (s) decreases.

No Correlation: Two features (variables) are not correlated with each other. It means that when the value of one variable increase or decrease then the value of the other variable (s) doesn't increase or decreases.

After data Preprocessing an important step comes and that is modelling also known as Model Selection. Model Selection is the process of selecting a model among many models for a predictive problem. Our problem is to predict the fare amount. This is a Regression problem. In Regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

Machine Learning models: Uber employs a lot of ML at different stages to predict user behaviour and optimize their processes. These models run by calling services to access features stored in different data layers. There are hundreds of ML models which predict a bunch of metrics while running in offline or online mode. Offline ML models are the ones where the models can run on HDFS vi a Spark/ Hive and store the results back.

Uber also employs a lot of Online ML models to feed low latency predictions. Within Online, there are some Batch models which need features to be Pre-

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

computed and stored so that they can be "looked up" when the app needs it (e.g. The average meal preparation time for a restaurant for last one month). In such cases, pre-computes happen in HDFS and the results get stored in Cassandra gets accessed at run time.

Linear regression: Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis regression algorithm shows a linear relationship between a dependent(y) and one or more independent(y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it find show the value of the dependent variable is changing according to the value of the independent variable.

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. A forest comprises numerous trees, and the more trees more it will be robust. First, start with the selection of random samples from a given dataset. Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree. Voting will be performed for every predicted result.

RF and LR are used to predict the fair price. In addition this approach Analyses various parameters such as (i) Trips by the hours in a day (ii) Trips during months in a year. At the end, visualizations are created for different timeframes of the year.

IV. RESULT ANALYSIS

A novel approach to analyze uber data analysis using machine learning is implemented using python in this section. The result analysis of presented approach is evaluated in this section. In this analysis a data visualization and fair price prediction project is build with the help of Linear Regression and Random Forest Regression Algorithms using Machine Learning and its libraries. Analyse various parameters like (a) Trips by the hours in a day (b) Trips during months in a year. At the end, visualizations are created for different timeframes of the year. In this paper, we analyze over 14 million yellow and Uber taxi pick-up samples in NYC. We find that there is a high predictability of uber demand (up to 83% in average), which indicates strong temporal correlation of human mobility. We also examine which commonly used predictive algorithm could approach the maximum predictability. The Fig. 2 shows the trip versus frequency.

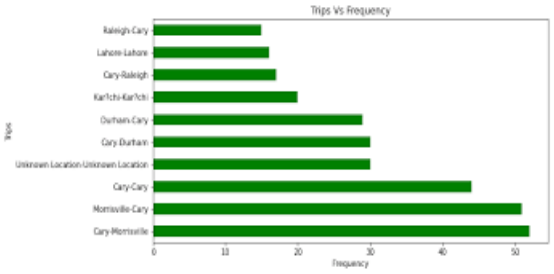


Fig. 2: Trip versus Frequency

The Fig. 3 shows the trips by day and month.

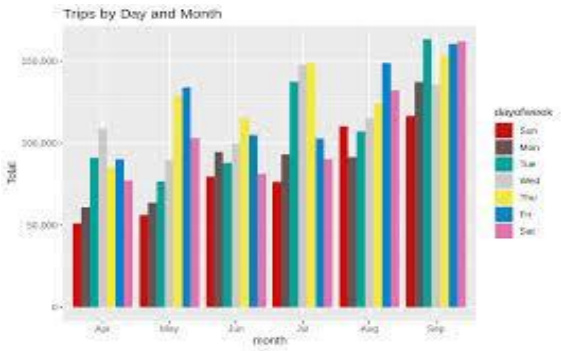
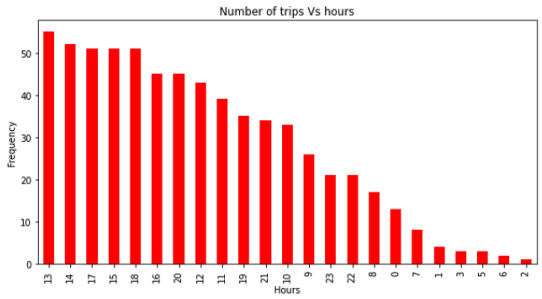


Fig. 3: Trips by day and month

The Fig. 4 shows the number of trips vs hours.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362



7.3 Number of trips vs hours

Fig. 4: Number of Trips vs hours

This analysis has explained how time affects customer trips. Find the days on which each basement has a greater number of active vehicles can tap growing markets in suburban areas where taxi services are not available. Estimated Time of Arrival can be reduced with increase in the number of Uber drivers which successively will make Uber more liked by the customers and hence, the company will get more revenue and drivers will also be profited. Based on the data, destination is determined that the people travel the most, which generates high air line revenues for travel, formed on booked trip count.

V. CONCLUSION

In this work, A novel approach to analyze uber data analysis using machine learning is presented. In this analysis Random Forest and Linear regression algorithms are sued to predict the price. This work is about on world's largest taxi company Uber inc. This project is all about understanding one such data set of Uber and to understand the use of data analytics and visualization and fare pricing for Services. It is generated with the help of Python programming language using libraries such as Pandas, Sklearn, Numpy by performing Linear Regression and Random Forest Algorithm in This approach has analysed parameters like (i) Trips by the hours in a day (ii) Trips during months in a year. At the end, visualizations were created for different timeframes of the year. From the results it is determined that there is a high predictability of uber demand (up to 83% in average), which indicates strong temporal correlation of human mobility.

VI. ACKNOWLEDGEMENT

We thank CMR Technical Campus for supporting this paper titled “A Novel Approach to analyze Uber Data Using Machine Learning”, which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head Of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work

VII. REFERENCES

[1] Mrunal Patil, Vidya Kumari, Adarsh Patil, Laxmikant Ahire, Umakant Mandawkar, “Uber Data Analysis Using Ggplot”, Journal of Engineering Sciences,Vol 12, Issue 7, July/ 2021, ISSN NO: 0377-9254

[2] Yash Indulkar, Abhijit Patil, “Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis”, 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021

[3] Rahul Pradhan, Praveen Kumar Mannepallli and Vikram Rajpoot, “Analysing Uber Trips using PySpark”, IOP Conf. Series: Materials Science and Engineering 1119 (2021) 012013, IOP Publishing, doi:10.1088/1757-899X/1119/1/012013

[4] Rishi Srinivas, B.Ankayarkanni, R.Sathya Bama Krishna, “Uber Related Data Analysis using Machine Learning”, Proceedings of the Fifth International Conference on Intelligent Computing and

Control Systems (ICICCS 2021), DOI: 10.1109/ICICCS51141.2021.9432347
[5] T.M Gunawardena; K.P.N Jayasena, "Real-Time Uber Data Analysis of Popular Uber Locations in Kubernetes Environment", 2020 5th International Conference on Information Technology Research (ICITR), DOI: 10.1109/ICITR51448.2020.9310851
[6] Shashank H, "Data Analysis of Uber and Lyft Cab Services", International Journal of Interdisciplinary Innovative Research &Development (IJIIRD), ISSN: 2456-236X, Vol. 05 Issue 01, 2020
[7] P. Devika, Y. Prasanna , P. Swetha ,G. Akhilesh Babu, "Uber Data Analysis using Map Reduce", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-4, November 2019
[8] Daksh Shah, Aravinda Kumaran, Rijurekha Sen, Ponnurangam Kumaraguru, "Travel time estimation accuracy in developing regions: An empirical case study with Uber data in Delhi-NCR", 2019 IW3C2 (International World Wide Web Conference Committee), ACM, doi:10.1145/3308560.3317057
[9] Lingxue Zhu, Nikolay Laptev, "Deep and Confident Prediction for Time Series at Uber", 2017 IEEE International Conference on Data Mining Workshops (ICDMW),DOI: 10.1109/ICDMW.2017.19
[10] Nikolay Laptev, Jason Yosinski, Li Erran Li 1 Slawek Smyl, "Time-series Extreme Event Forecasting with Neural Networks at Uber", ICML 2017 Time Series Workshop, Sydney, Australia

Akula Prashanth Kumar is currently pursuing B. Tech final year in the stream of Computer Science and Engineering in CMR Technical Campus, Medchal, Hyderabad, Telangana, India.

Kanchi Aashritha is currently pursuing B. Tech final year in the stream of Computer Science and Engineering in CMR Technical Campus, Medchal, Hyderabad, Telangana, India.

Marri Johnwesley is currently pursuing B. Tech final year in the stream of Computer Science and Engineering in CMR Technical Campus, Medchal, Hyderabad, Telangana, India.

Mr. Jonnadula Narasimharao is currently working as an Assistant Professor in the Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad. He obtained his Bachelor Degree in B.E – Computer Science and Engineering from DMI College of Engineering, Anna University and Master's Degree in M. Tech – Computer Science and Engineering from TRR Engineering College , JNTU Hyderabad. He is pursuing his doctorate in the field of Image Processing at Madhav University, Pindwara, Rajastan. His areas of Specialization include Image Processing, Deep Learning, Machine Learning, IOT, Data Mining and Networks. He has more than 15 years of Teaching Experience. He has published his Research work in reputed international journals with high impact factor in Elsevier, Springer, web of science, Scopus. He has certified in NPTEL, Coursera courses. In addition, he has presented papers in both international and national conferences. He has an Indian patent in his expertise areas from Computer Science and Engineering field. He is a life member of the professional body - Indian Society for Technical Education (ISTE).