UBER DATA ANALYSIS USING MACHINE LEARNING STRATEGIES

Vyshnavi Basude (00793667), Manasa Sukavasi (00795174), Pooja Donuru (00789033)
Department of Computer Science
University of New Haven
300 Boston Post Rd, West Haven, CT 06516
{vbasu1, msuka1, pdonu1@unh.newhaven.edu}

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. 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 analyse 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. Machine learning algorithms predict Uber demand surges by analysing historical ride data, weather patterns, events, and other relevant factors. They optimize dynamic pricing in real-time by adjusting rates based on demand predictions, helping balance supply and demand, maximizing earnings for drivers, and ensuring efficient service for passengers.
Keywords: Uber, Data analysis, Real time analysis, Machine Learning, Modelling, Optimization.

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.

In this ambitious machine learning project, we set out to revolutionize the way Uber manages its demand surges and dynamic pricing in real-time. Armed with a robust dataset encompassing historical ride data, weather patterns, events, and other influential factors, we embark on a comprehensive analysis to extract valuable insights and drive data-driven decisions.

Our approach is grounded in a multifaceted toolkit of machine learning techniques, including time series analysis, regression models, classification algorithms, and time series forecasting. These methodologies collectively empower us to predict demand surges with precision, understand the intricate relationships between variables, and categorize rides based on critical factors. This project's foundation also involves sophisticated feature engineering to extract meaningful insights from diverse data sources, ensuring that we capture the full spectrum of influencing factors.

Crucially, our dynamic pricing algorithms, fortified by reinforcement learning and optimization strategies, will enable Uber to make real-time adjustments that maintain a delicate balance between supply and demand, thereby maximizing driver earnings and ensuring an efficient service for passengers.

This paper is structured as follows: Section II provides details on the literature survey. Section III explains the proposed method. In Section IV the details about ex-perimetral results have been provided; followed by the conclusion, future work, and references. The appendix provides the link to the GitHub repository.

II. RELATED WORK

In 2016 Kai Zhao et al. addresses a pressing concern in urban transportation planning - the accurate prediction of taxi demand, particularly at a high level of spatial granularity. This study builds upon a substantial body of prior research in transportation modeling, urban analytics, and machine learning techniques. Historically, demand prediction models have operated at coarser spatial resolutions, potentially limiting their effectiveness in dense, dynamic urban landscapes. Zhao et al.'s focus on fine-grained spatial resolution is a natural evolution in this field, aiming to provide more precise and actionable insights for urban planners and transportation authorities. Moreover, this paper may contribute to the broader discourse on the application of machine learning and data-driven techniques in urban planning and transportation. Zhao et al. are likely positioned within this larger trend, seeking to push the boundaries of predictability by exploiting the wealth of data now available. In conclusion, "Predicting Taxi Demand at High Spatial Resolution" by Zhao et al. contributes to an evolving discourse in urban transportation research.

Guda and Subramanian's 2019 paper, "Your Uber Is Arriving," provides a thorough analysis of strategies employed by on-demand platforms, particularly Uber, to effectively manage their dynamic workforce. The authors focus on surge pricing, a real-time fare adjustment mechanism, as a pivotal tool in balancing supply and demand. This dynamic pricing strategy incentivizes drivers during peak periods, ensuring reliable service for consumers. Complementing surge pricing, the paper underscores the significance of forecast communication and worker incentives. Timely and accurate forecasts equip workers with crucial information about anticipated demand patterns, enabling them to make informed decisions about their availability. Platforms use incentive schemes to motivate favourable actions, such as accepting ride requests during high-demand periods or relocating to areas with increased demand. The authors likely draw from a body of literature surrounding incentive design in the gig economy, delving into the efficacy of different incentive structures in influencing worker performance and satisfaction. This research not only advances our understanding of labour market dynamics in the digital age but also provides practical guidance for platform operators seeking to optimize their operations and enhance the experiences of both workers and consumers in the on-demand economy.

In 2018 Abel Brodeur and Kerry Nield's "An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC," investigates the influence of weather disruptions on transportation choices in New York City. By scrutinizing traditional taxis, Lyft, and Uber, the study provides valuable insights into how external factors, like weather, shape consumer behavior and competition dynamics within the ride-hailing industry. The study stands out for its methodological rigor. Brodeur and Nield employ meticulous data collection and employ robust statistical modeling, ensuring the reliability and credibility of their findings. This rigorous approach allows for a nuanced understanding of how weather conditions impact ride service preferences, contributing significantly to the discourse on urban mobility. One of the key contributions of this research lies in its implications for urban transportation and the broader sharing economy. Understanding how weather affects choices between traditional taxis and app-based services has direct relevance for policymakers and urban planners. The study also provides valuable insights into platform competition. Brodeur and Nield discern how traditional taxis and app-based services respond differently to weather shocks. This analysis illuminates the strengths and vulnerabilities of each mode of transportation, offering crucial information for industry stakeholders and policymakers. It adds a nuanced perspective to the ongoing discourse on the evolution of transportation services in the digital era. However, it's important to acknowledge potential limitations. The study focuses on a specific city (NYC) and a particular set of weather events. While this specificity allows for in-depth analysis, it may limit the generalizability of the findings to other contexts. Future research could expand the scope to include different cities and a broader range of weather conditions, providing a more comprehensive understanding of consumer behavior in varying urban environments.

In Junzhi Chao's study on Uber's pricing examines the complex mechanisms behind the platform's fare structure. It operates within the context of the disruptive rise of ride-hailing services, focusing on Uber's innovative pricing models. These may encompass dynamic pricing, considering variables like demand, distance, and time. The study likely employs empirical modeling techniques, such as regression analysis, to distil real-world data into meaningful insights. It also delves into consumer behavior and price sensitivity in response to pricing changes, offering valuable insights for Uber's business strategy and broader considerations in urban mobility. Additionally, the research may hold policy implications for regulating the ride-hailing industry, addressing issues of consumer protection and fair competition. Chao's work contributes to the broader discourse on pricing strategies in the sharing economy and urban transportation, potentially paving the way for future research avenues exploring the impact of pricing on driver behavior, competitive dynamics, and broader societal implications. Overall, this study significantly advances our understanding of the intricate dynamics at play in ride-hailing services, offering insights with relevance for both industry practices and policy considerations in urban mobility.

In 2022 E. R. G et al. presents a pioneering approach to improve cost prediction in ride-hailing services. Leveraging machine learning algorithms, the authors propose an automated system that enhances the accuracy of estimating ride fares. This innovation holds paramount importance in the context of urban transportation, where transparency and reliability in pricing are crucial factors influencing user decisions. The application of machine learning in this context represents a significant departure from conventional methods of cost estimation. Unlike static rate structures or basic distance-based calculations, machine learning offers a dynamic and adaptable approach. By training on diverse and extensive datasets, the model learns intricate patterns and dependencies, leading to more accurate predictions. Moreover, the proposed system's practical implications are substantial. Accurate cost predictions in ride-hailing services directly impact user behaviour and satisfaction. When users can anticipate costs reliably, it enhances their confidence in the platform, fostering trust and loyalty. This, in turn, contributes to the long-term sustainability and success of ride-hailing companies. However, it's essential to acknowledge potential challenges. The quality and diversity of the training data, as well as the incorporation of dynamic factors like surge pricing or traffic conditions, are critical considerations. E. R. G et al.'s paper represents a significant advancement in the domain of ride-hailing services. By introducing an automated cost prediction system based on machine learning, the authors address a pertinent concern in the industry. The research not only advances theoretical understanding but also offers practical benefits for users and service providers alike.

III. THE PROPOSED METHOD

Machine learning algorithms power Uber's dynamic pricing strategy by processing historical ride data, weather forecasts, event calendars, and traffic conditions. These algorithms predict when and where demand for rides will increase, allowing Uber to allocate more drivers to high-demand areas and optimize pricing accordingly. Surge pricing incentivizes drivers to meet surging demand while ensuring efficient service for passengers. This data-driven approach maximizes driver earnings during peak times and enhances overall user satisfaction by reducing wait times and providing reliable transportation options.

Data Mining Techniques:
The Following are the Data Mining techniques to forecast future sales prediction:

- Logistic Regression
- Naïve Bayes
- Decision Tree Regression
- KNN

Dataset:

The dataset is called the " Uber and Lyft Dataset Boston, MA" and it contains information.
about the ride information between Uber and Lyft. Here are some details about the dataset:
The selected dataset contains extensive details regarding Uber rides taken over a two-month period in Boston, Massachusetts.
It provides not only ride details but also contextual features such as the weather conditions, time specifics, and environmental settings during each trip. Accessibility: The dataset is publicly available on Kaggle, a platform for predictive modeling and analytics competitions. Users can freely download this dataset after creating an account on Kaggle.
The data in this dataset has been collected using APIs provided by both Uber and Lyft, amalgamated with weather data APIs to provide context regarding environmental conditions during each ride.
The dataset is likely to have categorical data representing the type of ride (e.g., UberX, UberXL), although this is inferred and not explicitly mentioned in the provided data.
Exploring Uber and Lyft, ride sharing details, pricing, temperature, hours to Drive Sales. Identification of Deviations in sales and supply chain based on Customer reviews and sentiment analysis.

Preprocessing:

Before we do all our analysis, we must ensure, that our dataset contains balanced target classes, the appropriate dimensions, and the most representative observations. In our investigation, we used feature selection and instance selection as preprocess approaches. The purpose of feature selection is to remove unnecessary or redundant characteristics in order to minimize the dimensionality of the dataset.
We create a train/test split of the dataset with the cab type column as target. For all the techniques which we have used below, we utilized the two datasets i.e., Test and Train. From these two we have used 70% of the Training data and 30% of the Test data throughout all the techniques.
The models are evaluated based on the accuracy and performance given by the model for our dataset.

A summary statistics and exploratory data analysis was performed to gain insights of the data that included varied visualization techniques such as scatter plot, bar graph and histogram, 2-Y Axis Plot, Pie Chart, Co-relation. Scatter plots help in visualizing the relationship between two numeric variables. To observe the data distribution and a relationship (if any), Scatter plot of Price vs Distance

Results of plot price against distance

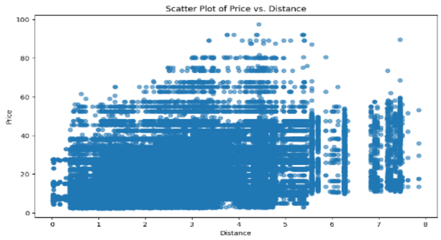

Fig 1:Scattter plot between Price vs Distance

Bar graphs are great for visualizing the distribution of categorical data. Given Uber's popularity, rides are almost equally frequent at all day hours (and night). However, more rides are ordered towards midnight or during business hours in the afternoon.
Interestingly, more rides are ordered on the weekdays of Monday and Tuesday than on most. This might indicate active business meetings or late-night outings, as seen in the previous graph.

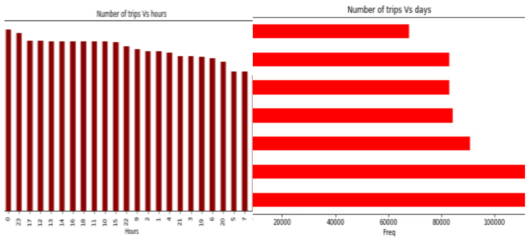

Fig 2: No.of Trips vs Hours vs Days

A histogram is a graphical representation of the distribution of a dataset. It's a way to visualize the frequency (or count) of different values within a dataset and understand the underlying pattern or structure.
In our exploration technique, we have plotted an histogram for Price Distribution of cab type between Lyft and Uber, so that we coud easily predict the pricing and get an easy estimation which cab type is cheaper.

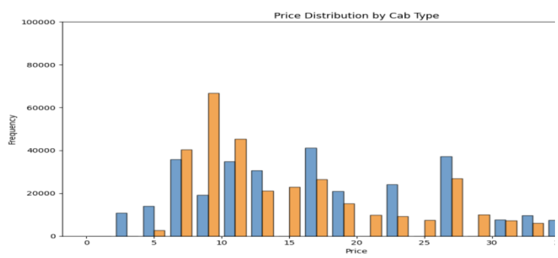

Fig 3: Histogram of price vs Cab Type

IV. THE EXPERIMENTAL RESULTS

The data analysis provided a detailed understanding of the data distribution. A significant finding is that the Uber has more sales compared to Lyft, in terms of price, days, temperature or humidity and we obtained 50% accuracy. Plotting the Logistic Regression Graph, which helps is better understanding for the different parameters and we can see that ROC area is about 0.5

```
Accuracy: 0.5055879494655005

Classification Report:
              precision    recall  f1-score   support

        Lyft       0.48      0.43      0.46     61339
        Uber       0.52      0.58      0.55     66257

    accuracy                           0.51    127596
   macro avg       0.50      0.50      0.50    127596
weighted avg       0.50      0.51      0.50    127596
```

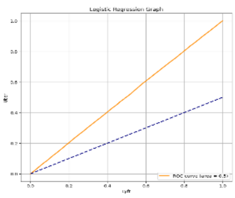

Fig 4: Logistic Regression on the dataset

A decision tree is a hierarchical tree-like structure used in machine learning for classification and regression tasks, making decisions based on feature attributes. By using decision tree we obtained 52% accuracy.

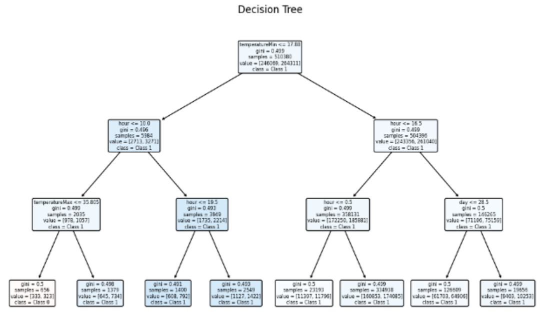

Fig 5: Decision Trees on the dataset

k-Nearest Neighbours (kNN) is a simple and intuitive machine learning algorithm used for both classification and regression tasks, making predictions based on the majority class or average value of its k nearest neighbours in the feature space and we used the n_neighbours value and we obtained 50 percent accuracy.

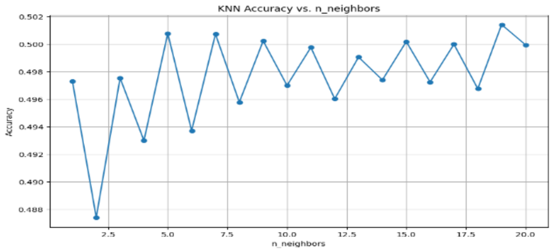

Fig 6: KNN on the dataset

Optimization strategies:

To observe how our model is reacting to different kinds of optimization technique we have tried to do the Boosting, GridSearchCV, Randomized Search .
Boosting is an ensemble learning technique where multiple weak learners are combined to form a strong learner. It is done sequentially, with each model correcting the errors of its predecessor.
Popular boosting algorithms include AdaBoost, Gradient Boosting (e.g., XGBoost, LightGBM), and CatBoost. These algorithms are widely used for classification and regression tasks.
For our model We have used
XG Boost and ADA Boost.

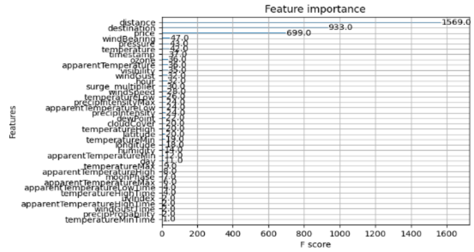

Fig 7: XG Boost - Feature Importance

Using the XG Boost algorithm, we were able to optimize the model to 95%. The critical factors influencing the decision-making process of the model include destination ,price, and distance. Through comprehensive analysis, it is evident that these features play a pivotal role in shaping the model's decisions.

```
accuracy: 0.9759240101570582
```

```
if isinstance(base_classifier, DecisionTreeClassifier):
    feature_importance = adaboost_model.feature_importances_
    plt.bar(range(len(feature_importance)),feature_importance)
    plt.show()
```

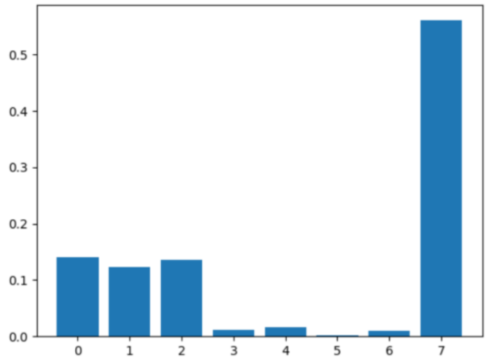

Fig 8: ADA Boosting

Above Figure is the result of ADA boosting which we applied on Decision Trees and have obtained the highest accuracy upto 97 %

V. DISCUSSION

Our analysis encompassed various machine learning models, including decision trees, k Nearest Neighbours (kNN), Logistic Regression and Naïve Bayes.
The dataset was split into a 70% training set and a 30% testing set for robust model evaluation. To facilitate model comprehension, categorical variables were converted to numerical values using dummy encoding, enhancing data accuracy.
Following data collection, preprocessing was initiated. This critical step involved handling missing values and potentially applying feature scaling for algorithms sensitive to feature magnitudes. Feature selection was imperative to retain essential characteristics, with techniques like Principal Component Analysis (PCA) employed to mitigate overfitting and enhance model performance.
The choice of model hinged on an in-depth understanding of the data and task requirements. Logistic Regression, due to its simplicity and interpretability, served as an excellent starting point. However, its linear nature could limit effectiveness for complex

relationships. Decision Trees and their ensembles. K-Nearest Neighbours, while intuitive, posed computational demands and performed sub-optimally with high-dimensional data.

In the pursuit of enhancing prediction performance, a comprehensive optimization strategy was employed across various machine learning models, including Decision Trees, k Nearest Neighbours (kNN), Logistic Regression, Naïve Bayes, and ADA Boost. The powerful XG Boost algorithm sequentially combined weak learners, achieving an impressive 95% accuracy. Grid search was then employed for a Decision Tree model, elevating accuracy from 52% to 83%. Subsequently, randomized search for the KNN model resulted in a substantial accuracy increase from 50% to an impressive 91%. A noteworthy development includes the implementation of ADA Boost, yielding a commendable accuracy of 97.59%. The accompanying graph illustrates the importance of various elements in the decision-making process, showcasing feature prioritization. The X-axis, representing the index of features, aligns with their respective importance. This visual insight enhances our understanding of the ADA Boost algorithm's decision-making dynamics. The overall analysis encompassed model selection, preprocessing steps such as dummy encoding and feature scaling, and continuous monitoring for accuracy maintenance.

VI. CONCLUSION AND FUTURE WORK

In conclusion, selecting the final model and its parameters required a meticulous assessment of the data, task intricacies, and trade-offs between predictability, and simplicity. Continuous monitoring and updates were deemed essential to maintain model accuracy and relevance. This dynamic approach ensured the most effective solution for the sales prediction of the uber and Lyft rides and we have got above 50 percent accuracy in all these techniques, but we can try to optimize them, out of all four decision tree was have good accuracy.

This vital approach ensured above 80% accuracy across techniques, with ADA Boost standing out at 97.59%. This strategic optimization underscores the pivotal role of hyperparameter tuning and thoughtful model selection in achieving excellence in predicting sales for ride-sharing services like Uber and Lyft.

This collaboration could extend to incorporating technical and fundamental analytical methodologies to enhance the refinement of these techniques. Future iterations of these methods will involve exploring additional tuning approaches to assess potential improvements in results. However, the challenge lies in acquiring and analyzing identical data, a concern shared by numerous companies. Many businesses in the market specialize in facilitating the aggregation of data in the form of prices surges from diverse sources and in different formats, seamlessly integrating it into your preferred data storage system. Future research in this area could explore methods for refining machine learning algorithms and integrating additional variables to further enhance prediction accuracy.

APPENDIX

Link to the project repository on GitHub
https://github.com/VyshnaviBasude/Team-TrailBlazers

REFERENCES

[1] K. Zhao, D. Khryashchev, J. Freire, C. Silva and H. Vo, "Predicting taxi demand at high spatial resolution: Approaching the limit of predictability," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016, pp. 833-842, doi: 10.1109/BigData.2016.7840676.

[2] Guda, Harish & Subramanian, Upender. (2019). Your Uber Is Arriving: Managing On-Demand Workers Through Surge Pricing, Forecast Communication, and Worker Incentives. Management Science. 65. 10.1287/mnsc.2018.3050.

[3] E. R. G, S. M, R. R. R, S. G. M, S. S. R and K. K, "An Automated Cost Prediction in Uber/Call Taxi Using Machine Learning Algorithm," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 764-767, doi: 10.1109/ICACITE53722.2022.9823852.

[4] Abel Brodeur, Kerry Nield,An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC,Journal of Economic Behavior & Organization,Volume 152,2018,Pages 1-16,ISSN 0167-2681,https://doi.org/10.1016/j.jebo.2018.06.004.

[5] Chao, Junzhi. (2019). Modeling and Analysis of Uber’s Rider Pricing. 10.2991/aebmr.k.191217.127.