

# Dynamic Pricing using Machine Learning for Emerging Ride-on-demand Service

MSc Research Project  
Data Analytics (MSCDAD\_JAN24A\_O)

Muhammad Abdur Rabb  
Student ID: x23237511

School of Computing  
National College of Ireland

Supervisor: Dr David Hamill

**National College of Ireland**  
**MSc Project Submission Sheet**  
**School of Computing**



**Student Name:** ...Muhammad Abdur Rabb.....

**Student ID:** ...x23237511.....

**Programme:** ...M.Sc. Data Analytics..... **Year:** ...2024.....

**Module:** ...Research Project.....

**Supervisor:** ...Dr. David Hamill.....

**Submission**

**Due Date:** ...12/12/2024.....

**Project Title:**... Dynamic Pricing using Machine Learning for Emerging  
Ride-on-demand Service .....

**Word Count:** .....17076..... **Page Count:**.....42.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** .....Muhammad Abdur Rabb.....

**Date:** .....11/12/2024.....

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission,</b> to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project,</b> both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Dynamic Pricing using Machine Learning for Emerging Ride-on-demand Service

Muhammad Abdur Rabb  
X23237511

## Abstract

The application of machine learning presents considerable opportunities for enhancing dynamic pricing mechanisms in ride-hailing services, particularly in response to swift variations in supply and demand. This study utilised historical data from Uber rides in New York City, following the CRISP-DM framework, to examine essential factors including ride distance, time of day, and number of passengers. Among these variables, ride distance was identified as the paramount factor influencing fare, whereas the number of passengers demonstrated negligible effects. Indeed, Gradient Boosting Regressor outperformed the three models, namely Linear Regression, Random Forest, and Multi-Layer Perceptron, with a mean absolute error of 0.2550 and an  $R^2$  of 0.8036, thereby effectively modelling nonlinear relationships relevant for dynamic pricing.

Despite this, the adopted modelling is limited within this inquiry, due to the non-availability of any real-time data and other outside factors such as traffic and weather. Unfortunately, in this instance, it was necessary to rely on a historical dataset from 2009 to 2015. If these factors were taken into consideration, together with an investigation of hybrid modelling techniques, this would clearly provide more adaptability and responsiveness. In conclusion, the findings and results clearly show how machine learning can produce dynamic pricing methods that balance profitability with customer satisfaction in the ride-on-demand marketplaces.

## 1 Introduction

Ride-on-demand services like FreeNow and Uber have grown in popularity in recent years. They appeal to passengers due to their convenience and flexibility, affordable pricing, while also attracting drivers who prefer the flexibility of using their own cars.

These services use dynamic pricing to increase revenue. In other words, they change prices based on things happening right now, based on supply and demand, for e.g., how many drivers are available, how many people need rides, seasonal conditions, events, fuel pricing and subsequent traffic. This is different from static pricing, where prices stay the same no matter what (Banerjee, Riquelme, and Johari, 2015). Since cities have busy traffic and changing conditions, static pricing doesn't work well anymore. To address this, companies can use machine learning, which helps them look at a lot of data and find patterns.

Static pricing doesn't fit well with changing situations like peak hours, driver availability, and traffic. Because these things keep changing, using dynamic pricing makes

sense to help companies make more money and have enough drivers available (McGuire, 2015). Machine learning assists companies check how many people need rides, how many drivers there are, and how traffic is. It lets them set prices in real-time, so they can charge fair prices and still have enough drivers when people need rides.

Even though dynamic pricing has been used for a while, there's still room for improvement using better machine learning. Current studies say there is a need for models that can work with real-time data and adjust prices to maximize profit while keeping customers happy. This study aims to look at machine learning models that can predict the best prices using historical data.

**RQ:** *“How optimized multi-variable dynamic pricing strategies can be developed using machine learning for an industry like the ride-on-demand service where demand and supply can fluctuate rapidly on a daily basis for time of day, distance to be covered, and number of passengers with competition dynamics while still assuring profitability & customer satisfaction?”*

**SQ1:** *“How do different time-of-day segments (peak vs. off-peak hours) affect dynamic pricing model, which machine learning algorithms can be used to capture these temporal variations for the purpose of optimizing price?”*

**SQ2:** *“What influence does the distance to cover and number of passengers in a dynamic pricing model, also which machine learning techniques are best suited for predicting fare changes on this basis?”*

Although previous studies have investigated various aspects of dynamic pricing in ride-on-demand, considerable gaps still remain. Many studies aim at the optimization of a single variable or a small set of variables, which often do not consider how important variables like time of day, day of the week, and ride distance interact in creating pricing strategies. Moreover, the ability of advanced machine learning methods, such as Gradient Boosting, to capture such complex relationships is yet unexplored. Another major shortcoming is that real-world constraints—like supply side variations—have been dealt with poorly so far, especially during severe mismatches in demand and supply. This gap is filled with a multi-variable machine learning approach by integrating temporal trends, ride distances, and supply-demand dynamics into dynamic pricing strategy optimization. This study develops a detailed framework for creating adaptive, robust pricing models of ride-on-demand services by using advanced predictive techniques and overcoming the limitations inherent in existing methodologies.

The objective of this study is to investigate the dynamic pricing in case of ride-on-demand services with a brief review over previous works and what advantages it offers over the prevailing static pricing that we have. The study will also break down the data to identify trends by time of day, trip distance, as well as by passenger count in order to determine how and when these play a role in pricing. A critical step is to prepare the data for analysis by cleaning it, filling in missing information, treatment of outliers and preparing it to be compatible for ML models. Then to build & test machine learning models (like regression based and decision tree) to predict the best fares. The models will be evaluated on accuracy — Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. Finally, a report detailing each of these steps and results, as well as recommend areas for further research/comparison and how to optimize ride-on-demand services pricing strategies.

This study can provide a range of benefits, including the possibility to offer ride-on-demand businesses an intelligent approach to pricing, as well as the opportunity to help businesses gain maximum profits by using adaptive pricing strategies.

This report is divided into four major sections. Related work on dynamic pricing of ride-on-demand services is presented in Chapter 2. In Chapter 3, the methodology of this research is discussed, which includes details of machine learning models, as well as their specifications. In Chapter 4, the validation of the models using various metrics in terms of performance, like accuracy, is discussed in detail. Finally, the findings obtained from this study, along with the conclusion and discussion for future work and enhancing solutions to implement a better dynamic pricing practice, are discussed in Chapter 5.

## **2 Review of Literature**

The current review chapter will assess previous dynamic pricing models, machine learning techniques, and metrics for ride-on-demand services. Dynamic pricing models allow price adjustments according to the real-time situation. Various dynamic pricing strategies that have evolved over the years from basic descriptive analytics to complex optimization on real-time personalized data will also be discussed and evaluated. Finally, several machine-learning approaches will be discussed, as such methods can be adopted to improve the estimates and ride pricing in general. Moreover, concepts coming under coverage will include but are not limited to neural networks and time series. Finally, the chapter focuses on how the accuracy of such models is sustained through evaluation metrics, such as mean absolute error or R-squared, displaying how the model is performing. The information to be covered will give a general overview of the relevant research, particularly within a data-driven arena related to the ride-on-demand industry.

### **2.1 Dynamic pricing models**

The wide adoption of dynamic pricing models changed the dynamics of ride-on-demand services, where the price can be changed with respect to immediate fluctuation in supply and demand. This section presents the use of dynamic pricing based on research in the literature, methodologies used and obtained results in order to analyse the research question.

The study by Guo et al. in 2017 focused on the empirical analysis of pricing mechanisms concerning ride-on-demand services, in which data was obtained from a leading Chinese ride-on-demand provider. The authors focused on the role of dynamic pricing variables, which change both in location and time to regulate the supply-demand balance. Simultaneously, the solution that Chen et al. proposed in 2020, basically analyses how to develop optimisation strategies for dynamic pricing. The authors have emphasised the need for real-time data and machine learning models that can effectively predict demand and ensure the introduction of relevant features. Thus, Guo et al. suggested the need to understand the essence of dynamic pricing, whereas Chen et al. took it forward by emphasising the introduction of machine learning-based optimization strategies. Considering the transition of the literature from descriptive analytics to predictive ones, this study tries to develop an optimal dynamic pricing strategy by incorporating machine learning.

In the context of using dynamic pricing to influence consumer behaviour and improve service efficiency, Luo et al. in 2017 performed both theoretical modelling and empirical analysis. The authors concluded that besides improving the operation effort of services, dynamic pricing also increased passenger satisfaction due to reduced waiting time. However, later research by Sun et al., 2020 discussed the integration of dynamic pricing with real-time traffic and environmental information to further create refinement in the pricing model. Furthermore, these studies indicate that dynamic factors are sustainable in view of higher accuracy and functional efficiency in the dynamic pricing model. It would, therefore, probably not be surprising that drivers behind dynamic pricing are multivariate in nature: while the discussion by Luo et al. centres on consumer behaviour, Sun et al. emphasise the use of external data. How real-time variables in the form of traffic and environmental conditions improve predictive accuracy and general efficiency in dynamic pricing models is, therefore, of prime importance to investigate.

Battifarano and Qian (2019) presented a model on surge multiplier prediction for Uber and Lyft using L1 regularization with clustering techniques. This real-time spatiotemporal predictive model produces an accurate forecast of surge price up to two hours in advance, defeating the traditional models. Their model focuses on the prediction of gaps between demand and supply but does not consider complications that might be involved in optimizing dynamic pricing, such as fluctuation of fares with regard to distance or time factors, including peak and off-peak periods. This paper has incorporated optimization of pricing in the model. Machine learning models will be applied to predict fares, factoring in several aspects: trip distance, temporal variation, and supply-demand gap discrepancies. This distinction represents an important limitation to their study, since the lack of price integration confines the applicability of their findings only to theoretical price mechanisms.

Chen and Sheldon (2015) analysed the impact of surge pricing on the behaviour of drivers. According to them, surge pricing significantly incentivizes drivers to work during periods of high demand. Their work revolved around how surge pricing controls labour elasticity in the gig economy, hence some of the earlier theories regarding income targeting were proven wrong. Though the findings are vital in understanding supply-side dynamics, these findings fail to relate such labour behaviour to the pricing of fares. While the present study follows their earlier work in so far as it looks at supply-demand dynamics, it also embeds them in methodologies for fare optimization. Whereas Chen and Sheldon's work had a focus on behavioural economics, the present study deploys machine learning methods to develop practical, data-driven price determination strategies that improve profitability and operational efficiency.

This review presented how dynamic pricing cannot be taken away from ride-on-demand services, and the models can further be extended with the use of optimisation methods. The ideas ranged from analysing the pattern of the price to the use of real-time data in the optimization hence framing the current research on dynamic pricing. Ultimately, this forms the basis of the authors' own research, in which a dynamic pricing strategy has been considered by applying machine learning techniques to devise a strategy that would consider a set of variables in real-time against shifting levels of demand and supply.

The following section compares different machine learning models, which will determine the optimal model for dynamic pricing of the ride-on-demand service.

## 2.2 Machine learning applications in dynamic pricing

Machine learning (ML) has been integrated into the fare predictions of ride-on-demand services for serving dynamically priced models more accurately and effectively, particularly in recent times. The following review selected the research works that have employed machine learning in dynamic pricing for a wide variety of contexts in order to establish if, indeed, the introduction of machine learning techniques can improve dynamic pricing strategies, hence answering the main question that forms the core of this research.

An early study by Guo et al. (2018) proposed a neural network to analyse multisource urban data to predict dynamic prices. The model, considering complex, high-dimensional characteristics from the input data, resulted in a very high accuracy of prediction and gave significant improvement over traditional baseline models. Alternatively, a more recent inquiry by Nalamothu (2023) compares different ML models, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest, to discover which one is best for the dynamic pricing prediction of ride-on-demand services. It followed that Random Forest was the optimal choice, followed by SVM and KNN. Comparing the neural network approach of Guo et al. with the approach of Nalamothu evaluating multiple models yields important insights into the strengths and weaknesses of various ML techniques. These studies have brought forth the foresight that model selection will be tantamount to the complexity of data and specific requirements of prediction.

Arora et al. (2021) performed linear regression for dynamic pricing in on-demand ride services based on various features like travel distance and time. Good accuracy obtained shows that even linear regression can provide distinct advantages within his area. As such, exploratory data analysis was conducted to get the most useful patterns and trends of the data to choose the best model. Conclusions have shown it was well-suited for introductory dynamic pricing tasks because of the simplicity and interpretability of the variables when the relationship between them is in a simple form. Further studies by Faghih et al. (2020) combined linear regression with the ARMA model with a view to considering other factors, such as weather and demands for high service. In this instance, Timeseries analysis was adopted to catch seasonal trends and has provided improved results as compared to using only linear regression. Their findings prove that combined, machine learning methods with time-series analysis solve the problem of temporal dependencies in dynamic pricing, with the introduction of the ARMA model best suited to discover the time-dependent relationship.

El Youbi et al. (2023) carried out a comprehensive comparative study of Gradient Boosting Machine, Random Forest, and Neural Networks in modelling performance, with particular emphasis on applicability to dynamic pricing strategies in e-commerce. Their findings indicated that the Gradient Boosting Machine (GBM) outperformed the rest of the models by achieving an  $R^2$  value of 0.92, thereby completing the ability of modelling complex, non-linear relationships in large datasets. The study has emphasized the utility of hyperparameter tuning for model performance optimization in dynamic environments, where variables such as competitor actions and customer behaviour drive pricing decisions. Despite the very interesting insights that are brought out in the work of these researchers on the potential of GBM, the domain remains limited to e-commerce, where price optimization is driven by market trends and purchasing patterns of individuals. In contrast, this paper extends

their approach to the ride-on-demand sector, including time-dependent and ride-specific parameters such as distance, time of day and changes in demand and supply. This study thus adapts GBM to these specific problems and thereby demonstrates the adaptability and effectiveness of the algorithm in solving dynamic pricing problems in mobility services.

Saadi et al. (2022) present an in-depth analysis that involves spatiotemporal demand prediction in ride-hailing based on machine learning techniques. The work focused on the assessment of changes in short-term demand based on such influences as meteorological conditions, traffic volume, and dynamic pricing. Single decision trees, bagged decision trees, random forests, boosted decision trees, and artificial neural networks are different models applied in their work for which their respective performances were methodically compared. Among them, boosted decision trees attained an improved predictive performance for the least value of root mean square error (RMSE), avoiding overfitting, hence very suitable for short-term demand forecasting. This research points out the importance of spatio-temporal dynamics in demand fluctuations of different districts and times of a day for the effective management of ride-hailing services. However, this framework was only limited to the prediction of demand trends and did not include optimization of pricing strategies. This present research, on the other hand, uses the same temporal dynamics but takes one step further by embedding these into a holistic pricing framework. By correlating demand forecasts with fare modifications, this research establishes a connection between understanding variations in demand and implementing effective pricing strategies aimed at revenue maximization while reconciling discrepancies between supply and demand. This distinction underlines the contribution of this present study to address both the operational and strategic aspects of dynamic pricing in the context of ride-on-demand service.

The work of Yamuna et al. (2024) studied different machine learning models for dynamic pricing at e-commerce companies to study the effects on profit maximization and customer satisfaction by making real-time adjustments. Methodologies ranged from competitor pricing to demand fluctuation and seasonality, hence representing the flexibility of machine learning in dynamic and competitive market settings. They applied reinforcement learning and Gradient Boosting to demonstrate how such models can efficiently balance profitability and customer retention while price setting. However, their study largely ignored contextual factors like supply-side constraints, which in ride-on-demand services are very important as the availability of drivers and real-time supply-demand imbalances strongly impact pricing strategies. Building on the work of Yamuna et al., this study modifies dynamic pricing frameworks to capture ride-specific variables such as distance, temporal variations, and peak-hour demand patterns addressing unique challenges in this domain.

These reviewed studies have provided unique insights into ML techniques applied to the dynamic pricing problem in ride-hailing services. While neural networks and model comparison methods have their own merits, so does time-series integrated analysis use of linear regression. All the works identify the problem of robust machine learning models that will be capable of dealing with real-time data and complex variables for optimal pricing solutions. This will align with the purpose of the research, which will be to develop an efficient dynamic pricing model using advanced ML techniques.



The subsequent section presents the performance measures that will be used in testing the efficacy and robustness of the proposed machine learning models in dynamic pricing for ride-on-demand services.

## **2.3 Evaluation metrics in machine learning for dynamic pricing**

In dynamically priced ride-on-demand services, the evaluation metric becomes very important, in indicating whether the performance of the machine learning models is accurate or not. The section that follows will review relevant studies, putting more emphasis on an appropriate metric which best suits the research question.

Chai and Draxler (2014) provided an exhaustive review of the statistical measures adopted for model predictions. From a practical viewpoint, both mean absolute error (MAE) and root mean squared error (RMSE) are useful. Though the latter is more general in the literature, its value is misleading quite often due to its sensitivity to outliers. Therefore, it would seem that MAE has the obvious interpretation as an average error, hence an edge against outliers in applications. Indeed, Willmott and Matsuura (2005) further offer more weight to this view when they hold that MAE is, in theory, a more logical and unambiguous estimate of average error than RMSE. They caution against using the RMSE because its scaling with error variance might result in bias that could mislead the interpretation of model performance. This becomes even more critical in dynamic pricing, where the choices of the most representative performance metrics should not be too sensitive to outliers in precision and robustness.

In fact, early studies by Chicco et al. (2021) stated that the coefficient of determination can be used instead of symmetric mean absolute percentage error (SMAPE), mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), and root mean squared error (RMSE), provided a regression analysis has been applied. It will provide information about how well the model explains the variability of the response data, a very important aspect in dynamic pricing models. Similarly, Hodson describes the same use for RMSE and MAE in 2022. According to their study, RMSE works best when the form of error takes a normal distribution, and for non-normal distributions of error, MAE will be more apt. Hodson concludes that neither measure is intrinsically better than the other. The suitability depends upon the type of distribution in error for that particular usage. This view will be necessary for dynamic pricing models since it gives light to the choice of a good metric for evaluation, bringing more intuition about the error behaviour.

These studies further emphasize applying appropriate assessment measures with regard to dynamic pricing models. MAE indeed seems robust, performing particularly well in cases of non-normal distribution of errors, while R-squared again provides valuable information with respect to variance explanation. Such insights will iteratively help to improve the performance of dynamic pricing models that needs to be evaluated.

In other words, there are still gaps in the literature today with regard to dynamic pricing models and even machine learning techniques. Most of the studies done in the past focused on single variables or single models that cannot capture the dynamics in real time, considering supply, demand, traffic, and competition. Other areas of further study call for the development of hybrid models that combine different machine learning techniques. This paper tends to bridge those gaps by proposing a new multidimensional dynamic pricing model for the ride-

on-demand economy with the power of machine learning to improve profitability, keeping customer satisfaction in mind.

The next chapter describes the methodology and specification of the research, incorporating how this study was carried out based on the literature review.

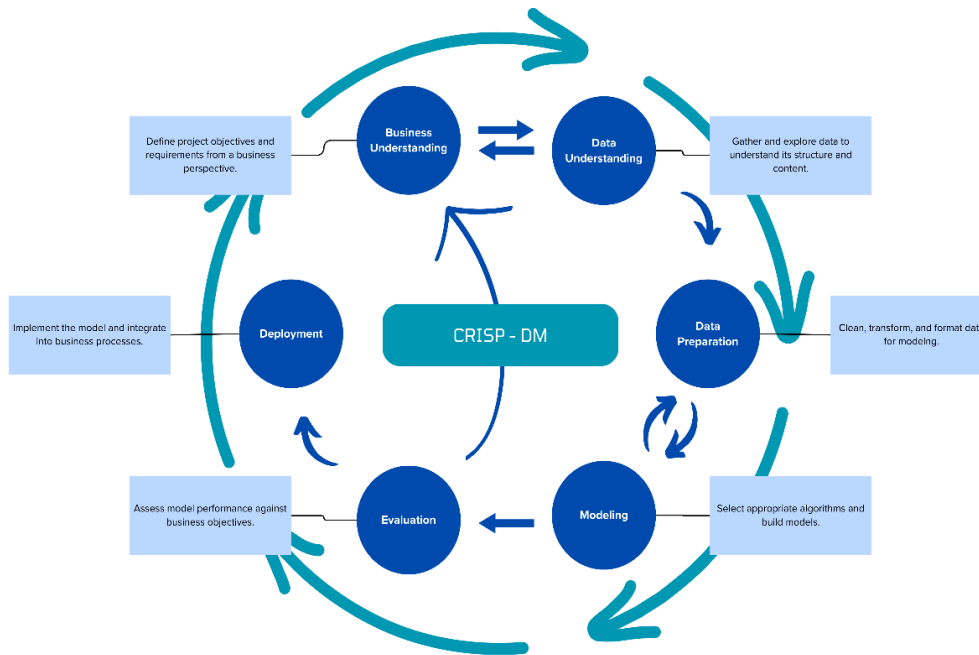
### **3 Research Methodology**

The key basis for adopting CRISP-DM within this research study is that it introduces a robust, structured process in which, logically, there will be clear routes from problem identification to model deployment: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. All of these involve repetition for each step since sometimes steps require revision after new facts have been learned. In practice, this makes the CRISP-DM most effective in volatile contexts like the dynamic pricing in ride-on-demand services, where real-time information often changes significantly. The cyclical nature of the CRISP-DM ensures that data undergoes constant review, hence always accommodative to new information, a great advantage in handling big and dynamic datasets. (Saltz, 2021)

Other methods have been explored such as the KDD, or Knowledge Discovery in Databases, and SEMMA: Sample, Explore, Modify, Model, Assess. These have proven to be less flexible than the Cross-Industry process. KDD focuses on finding trends within data. Due to the inability to iterate further back than model evaluation, this makes it unsuitable for projects that need to refine the models iteratively, like in dynamic pricing (Fayyad et al., 1996). The SEMMA methodology, formulated by SAS, emphasizes exploratory data analysis and the construction of models; however, it is deficient in the thorough comprehension of business contexts and the deployment stages that are essential for practical applications (SAS Institute, 2008). In contrast, CRISP-DM adopts a comprehensive framework that integrates business objectives and deployment, thereby guaranteeing that the machine learning models created are not only technically robust but also consistent with organizational aims (Saltz, 2021).

More applicable for this research, CRISP-DM gives the opportunity to consider at early stages the business contexts, supported by deep data and model exploration. Considering dynamic pricing, changes have to be instant, given that the demand and supply conditions change rapidly; the adaptability and the iterative feedback mechanism within CRISP-DM enable the continuous improvement of the pricing models. This will make sure the model continuously meets the business needs by optimizing performance (Saltz, 2021; Rathore et al., 2024). The model implementation calls for adaptability most in dynamic settings; thus, as in the case of the ride-on-demand service, approaches like KDD and SEMMA cannot compete against real-time market fluctuations.

CRISP-DM stands for Cross Industry Standard Process for Data Mining. It describes one widely used model in breaking down data science projects into six steps. Each step brings structure and a guideline to keep the project on track and assure that the outcome is meaningful. Specific details of each step are delineated as follows:



**Figure 3.1: The CRISP-DM Process.**

**i) Business understanding**

This is the very foundational step. It stipulates the business objectives, stating the problems that are to be solved. The aim during this stage is to have a clear vision with respect to what the success of the projects would look like, and whether the data analytics methodology is in line with the aim of the organization. During this stage, a broader view of insight into the business is necessary.

**ii) Data understanding**

Where the business objectives are well-defined, the subsequent process becomes that of data exploration. The gathering of data, familiarization with its structure, and identification of key variables are included in this. The quality assessment of the data should be performed on the assumption that the sooner incompleteness or inconsistency is detected, the fewer problems will emerge during further project work.

**iii) Data preparation**

In this step data cleaning and preparation are performed, by treating missing values, outliers, and sometimes inconsistencies, and at times making new features out of existing ones, usually improving the performance of a model. Actually, it is one of the most time-consuming steps; however, it lays the bed for the success of the entire project.

**iv) Modeling**

With clean data, various machine learning or statistical models are created and compared. This is the area where choices regarding which algorithm is best should be made, together with their optimization, in order to determine the best model for the project. Many different models may be considered in finding which best actually makes the most accurate or valuable predictions on the data.

**v) Evaluation**

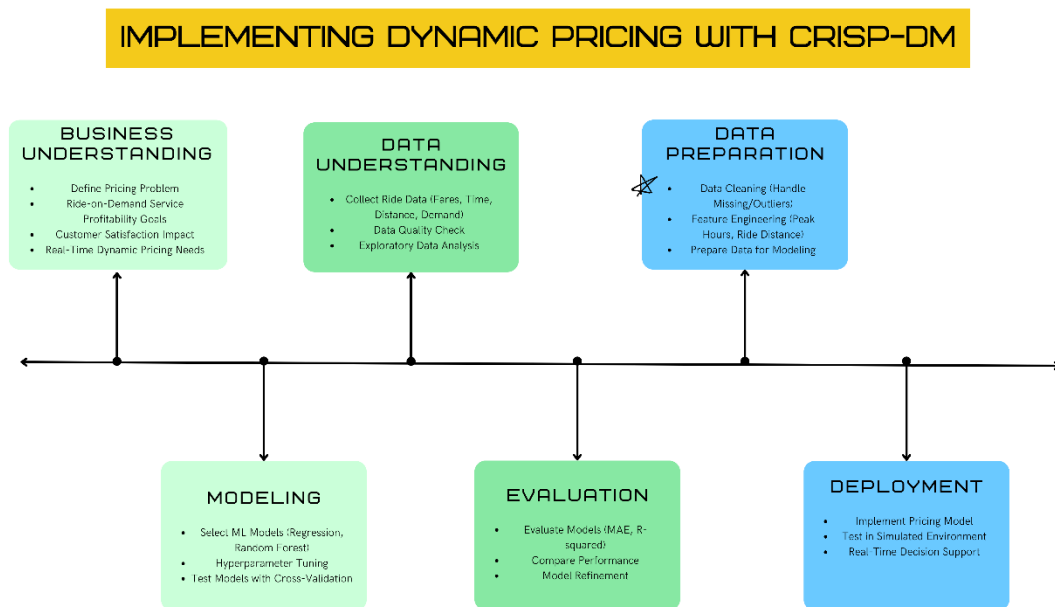
After modelling, the next in line is model evaluation. The developed models have to be checked against the business goals during this step and their performance in terms of some key metrics. Otherwise, if the results cannot give satisfaction, it might be necessary to tune the model and probably try new ones.

#### vi) Deployment

At this stage, which is the end of model development, it is now ready to be deployed, either embedding the model in a real-world application or running simulations so as to test how the model would work in practice. This is, therefore, done in order to exploit the insight extracted from the data for better decision-making or the automation of a certain process.

### 3.1 Implementation of dynamic pricing with machine learning

Following is the detailed methodology of implementation of dynamic pricing with machine learning:



**Figure 3.2: The CRISP-DM Process for Dynamic Pricing**

#### 3.1.1 Business understanding

The Business Understanding phase of CRISP-DM is crucial for setting the foundation of a project by aligning data-driven goals with business objectives. For ride-hailing services, the main challenge is optimizing pricing to balance supply and demand efficiently. Dynamic pricing allows for real-time adjustments, preventing mismatches in availability and fare pricing. One such approach, certain studies have pointed out, would increase overall service reliability and customer satisfaction since driver supply would accord with dynamic pricing according to the situation (Yan et al., 2020). The prices of fares can also be estimated by machine learning algorithms through demand fluctuation and supply constraints for profitability and enhancement in user experience (Ashlagi et al., 2018).

Key focus is to determine how several variables interact with each other to influence price while keeping the system fair for customers and profitable for the business. Predicting

how much and when the price would change, based on historical data, market dynamics, and real-time factors, is key. Refining these predictions through machine learning will help achieve the aim of optimizing pricing for profitability without compromising service quality.

### 3.1.2 Data understanding

Following the CRISP-DM methodology, Data Understanding should be done to develop preliminary insights into the data and detect any problems that might pop up in order to adjust the approach in the further process.

Analysis and exploration in this research are going to be done using Python, Jupyter Notebook, and VSCode, supported by libraries like pandas, numpy, matplotlib, and seaborn. This is deep exploration into the dataset for key variables of interest: ride distance, fare amount, and duration of the ride, with deep understanding of any pattern, inconsistency, or outliers that could influence model predictions. The data quality checks include the detection of missing values to be imputed or excluded, and outliers to be excluded or transformed. EDA is done to illustrate key relationships and correlations that give further understanding of those variables causing dynamic pricing, for example, how peak hours or long-distance rides affect the fares. It is a good framework to build correct models because the step ensures that the data is well understood and prepared for preparation and modelling in later stages.

The present section will proceed with discussing the collection of data, description, and quality assessment procedures that are supposed to precede preliminary exploratory analysis.

#### i) Data collection

The data set used for this analysis is a publicly available one, containing historical records of Uber rides. The data is very important to understand the dynamic pricing mechanism, as it contains important information about every ride regarding time, place, and fare amount-which are the keys to understanding the basics of any pricing model.

- **Source:** Data for this research is taken from a publicly sourced dataset on Kaggle entitled "Uber Fares Dataset". The dataset is actually designed for research studies to be performed on Uber ride pricing and contains all the necessary information with regard to time, location, and the amount of fare. It is commonly used to analyze patterns in Uber rides, including fare fluctuations, ride demand, and other features that impact dynamic pricing. The dataset is licensed under the CC0: Creative Commons Public Domain License. As such, it indicates that the dataset has been dedicated to the public domain, and users can, without restriction under copyright or database law, copy, modify, distribute and perform the work, including for commercial purposes, without having to obtain permission. Regarding this dataset, no copyright restrictions exist; the dataset is available for all possible uses.

#### ii) Data description

- **Variables:** The dataset contains various feature types relevant in dynamic pricing research. Some of the essential variables are:
  - a. **Time of Ride (pickup\_datetime):** this includes the date, time, and even day of the week that the ride was begun. This would give them an idea of peak hours for demand and how those might impact prices of their services.

- b. **Ride Distance (calculated from pickup and dropoff coordinates):** The critical determinant of pricing of fares is the distance that lies between the pickup and drop-off points.
- c. **Pickup and Dropoff Locations (pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude):** Geographical coordinates are used to identify the origin and destination of the trip for a spatial analysis of fare variation.
- d. **Fare Amount (fare\_amount):** the target variable for dynamic pricing models is the fare amount for every single ride.
- e. **Passenger Count (passenger\_count):** The number of passengers in the vehicle could affect the price, especially with regard to ride-sharing or group rides.
- **Volume and data types:** Dataset comprises precisely 200,000 records and 9 variables. A dataset of this size applies to dynamic pricing research because large sets of data are needed to observe how demand and price are changing over time. A dataset with this magnitude in size, its dimensions, and arrangement allows for the implementation of proper statistical analysis and the application of machine learning algorithms; hence, detailed analyses can be carried out in fare trends and determinants.

The dataset contains several kinds of variables, in which numeric variables consists of fare\_amount, pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude, and passenger\_count, which are categorized in a category of float64 and int64.

Categorical/object variables include key and pickup\_datetime. A variable of this type requires parsing to do some more analysis; more so, the pickup\_datetime may also be changed to datetime format.

```

>>> print(df.shape)
Success: print(df.info())
✓ 0.0s

(200000, 9)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0            200000 non-null  int64
1   key                   200000 non-null  object
2   fare_amount           200000 non-null  float64
3   pickup_datetime       200000 non-null  object
4   pickup_longitude      200000 non-null  float64
5   pickup_latitude       200000 non-null  float64
6   dropoff_longitude     199999 non-null  float64
7   dropoff_latitude      199999 non-null  float64
8   passenger_count       200000 non-null  int64
dtypes: float64(5), int64(2), object(2)
memory usage: 13.7+ MB
None

```

**Figure 3.3: Dataset shape and variable types.**

- **Time period of data:** It contains data from January 1, 2009, through June 30, 2015. It was derived by first converting pickup\_datetime into a datetime format and then finding the minimum and maximum date. This in turn could be useful later as basis for analysis of fare and other time dependent variables.

```
df['pickup_datetime'] = pd.to_datetime(df['pickup_datetime'])

sorted_data = df.sort_values(by='pickup_datetime')

start_date = sorted_data['pickup_datetime'].min()
end_date = sorted_data['pickup_datetime'].max()

print("Time period for dataset is from", start_date, "to", end_date)
```

✓ 0.0s

Time period for dataset is from 2009-01-01 01:15:22+00:00 to 2015-06-30 23:40:39+00:00

Figure 3.4: Time period of data.

### iii) Data analysis and quality checks:

- **Irrelevant columns:** In the following figure 6, there are top five rows of dataset, along with variable names. The first step in data analysis and quality checks is to remove unwanted variables, that do not contribute to any analysis. In this data, column “Unnamed 0” and “key” do not provide any information.

```
df = pd.read_csv('uber.csv')
df.head()
```

✓ 0.4s

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	24238194	2015-05-07 19:52:06.000000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
1	27835199	2009-07-17 20:04:56.000000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
2	44984355	2009-08-24 21:45:00.000000001	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
3	25894730	2009-06-26 08:22:21.000000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3
4	17610152	2014-08-28 17:47:00.0000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5

Figure 3.5: Top five rows in dataset along with variable names.

- **Identifying missing values:** In this dataset, missing values were identified by utilizing the `isnull()` function. Careful study showed that it had only one missing value in the `dropoff_latitude` and `dropoff_longitude` columns, which was further confirmed by segregating the row for further detailed study. This single missing value is very important since the latitude is one of the important features for identifying the location of drop off, and its presence or absence might affect the accuracy of fare prediction.

```
df.isnull().sum()
```

✓ 0.0s

fare_amount	0
pickup_datetime	0
pickup_longitude	0
pickup_latitude	0
dropoff_longitude	1
dropoff_latitude	1
passenger_count	0
dtype:	int64

Generate + Code + Markdown

```
df[df['dropoff_latitude'].isnull()]
```

✓ 0.0s

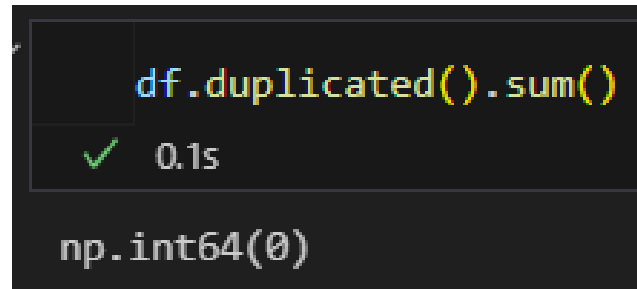
fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
87946	24.1	2013-07-02 03:51:57 UTC	-73.950581	40.779692	NaN	0

Figure 3.6: Identifying missing values.

- **Identifying duplicate values:** Verification of duplicate entries was essential to ensure that the quality data used for analysis is correct. Duplicate records may appear because of errors in data collection, repeated transactions, and many other problems related to data integration. This may further distort the analysis. For instance, where some rows are

duplicated, it increases the frequency of a particular data point, which is biased in drawing insight into specific analyses, especially in dynamic pricing data analysis.

This problem was solved by implementing the function `duplicated()`, which can be used to find duplicated rows in a dataset. The function returns a Boolean series indicating whether each row is a duplicate; summing these thus gives the total number of duplicates present. Since the output of this operation returned that there were no duplicates in the dataset, nothing further was needed here.



```
df.duplicated().sum()
0.1s
np.int64(0)
```

**Figure 3.7: Identifying duplicate values.**

- **Outliers' detection:** Generally, outliers in any given dataset are a subset of data values that lie considerably away from the other observed points. Outlier detection reaches anomalies and extreme values which may influence further data analysis or model performance.
  - a. **Longitude and latitude outlier detection:** Since this information is on New York City, the longitude and latitude coordinates of pick-up and drop sites are supposed to fall within the geographical limits of New York City, that is, between 40.4774 to 40.9176 in latitude and between -74.2591 to -73.7004 for longitude. Outliers were identified for values outside of this range.
  - b. **Passenger count:** The passenger count normally varies within a range of 1 to 6 passengers in one car. All the instances beyond this, that is 0 or more than 6, fall into outliers since they are not typical data for an Uber ride.
  - c. **Pickup and dropoff datetime:** Though not in the numerical sense, any date anomalies or impossibilities were treated as outliers. For instance, all those rides which were taken outside the known timeframe of the dataset, that is, before 2009 or after 2015, are indicative of incorrect input of data.

**iv) Exploratory data analysis:**

Exploratory data analysis had been performed at the Data Understanding stage to get the general view of the dataset and outline probable problems. EDA allows to discover patterns, trends, and anomalies in the data, which is fundamentally important for further preparation steps and modelling. For this, EDA involved a few steps: the preprocessing of data by feature engineering and transformation, in that the raw data lacked some of the basic variables necessary for thorough analysis.

Specifically, the ride distance variable was not included within the original dataset which is very important variable in describing the relationship of the distances driven with the fare prices. Thus, in the pre-processing stages, the ride distance was calculated using the Manhattan distance formula from the coordinates of `pickup_longitude` and `dropoff_longitude`. This derived feature allowed a more in-depth analysis of how distance influences fare pricing, directly addressing one of the key aspects of the research.



Similarly, no meaningful temporal analysis could have been conducted on the raw variable of `pickup_datetime`; hence, from `pickup_datetime`, the year, month, day, hour, and weekday features were extracted to allow for time-based explorations. Such features would be required in explaining how the fares change during the day, particularly in identifying peak vs. off-peak hours, which is most relevant to the research questions on temporal variations in dynamic pricing.

Although all of the above steps fall technically under Data Preparation but had to be done for doing proper EDA. Otherwise, the relationships involving variables like fare, distance, and time would not be studied comprehensively, and derivation of meaningful insight would also be limited.

The key steps in EDA included:

- **Descriptive statistics:** Basic summary statistics were calculated for important variables like `fare_amount`, `ride_distance`, and `passenger_count`. The summary statistics gave an overview of the structure of the data and allowed the identification of anomalies, such as extreme outliers or unusual distributions.
- **Data visualization:** Visualizations were then used to explore the distribution of key variables:
  - a. Histograms were used to visualize the distribution of fare amounts, emphasizing possible skewness or outliers.
  - b. Scatter plots were created to study relationship between ride distance and fare to demonstrate the impact of distance on pricing.
  - c. Box plots helped in finding some outliers in `fare_amount` and `passenger_count` that might potentially bias the analysis.
- **Correlation analysis:** A correlation matrix was developed to study the relationships among key variables: `fare_amount`, `ride_distance`, and time-related attributes. It was necessary to understand how different variables interact with one another and impact fare pricing; it provided insights that directly supported the research questions.
- **Time-Based exploration:** Temporal trends were explored by analysing the fluctuations in fare amount depending on the hour of the day and day of the week utilizing the time-based features extracted previously. This analysis allowed for the identification of possible differences between peak and off-peak hours, reinforcing the development of optimized dynamic pricing model.

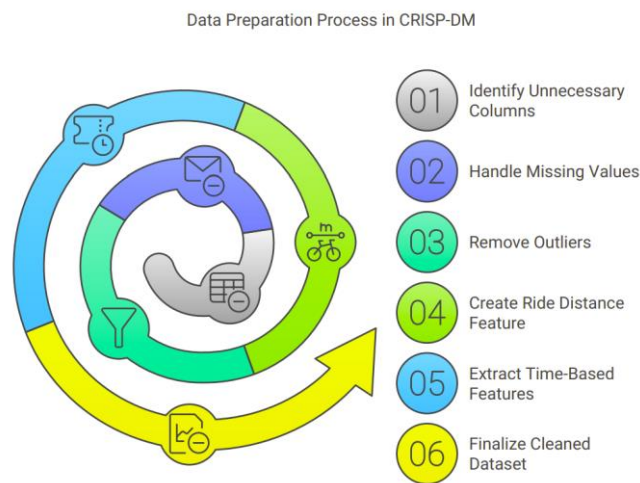
Preliminary EDA was done during the Data Understanding phase to understand the dataset, identifying potential issues such as missing values or anomalies. Basic visualizations and summary statistics have been generated in an attempt to understand the distribution of key variables of interest, such as `fare_amount` and `passenger_count`. Though these initial observations were useful to detect the anomalies, deep analyses involving integration of correlation analysis and temporal exploration needed further cleaning of data that is thoroughly discussed in Data Preparation phase. This ensured that all the variables related to ride distance and time-related features are suitably prepared for deeper analysis and model development.

The EDA process helped in understanding patterns, relationships, and anomalies in the data, hence laying a suitable foundation for data preparation and model development. With the

extraction of ride distance and time-based features, EDA facilitated the suitability of the dataset in answering basic research questions on fare dynamics and pricing strategies.

### 3.1.3 Data preparation

Data Preparation in CRISP-DM means the raw dataset will be transformed into a clean and structured format ready to go, which includes handling missing values, addressing outliers, and performing necessary feature engineering. In this respect, the data preparation was an important part of this research to ensure relevance and precision of the dynamic pricing model by refining key variables such as fare amount, ride distance, and time-based features. Procedures include cleaning the data, extracting features, their transformation to be appropriate for goals of research, and development of optimized pricing strategies. Several key steps were taken in this phase:



**Figure 3.8: Data preparation process in CRISP-DM for dynamic pricing of ride on demand service.**

**i) Dropping unnecessary columns:**

Initially, the columns labelled Unnamed: 0 and key were recognized as nonessential for the analysis and subsequently eliminated. The removal of these unimportant columns was crucial for streamlining the dataset and focusing on relevant variables.

**ii) Handling missing values:**

There was one missing value for which, instead of imputation, the record was removed. Given the size of the dataset, removing this row did not impact the overall dataset. This is decided as part of the approach to avoid the possible biases or inaccuracies that can result from imputation methods themselves and further affect the analyses, especially those on latitude and longitude as sensitive variables.

**iii) Outlier removal:**

Several outlier detection and removal processes were performed:

- **Fare amount:** Fare amounts ranging from 4 to 130 have been kept because values outside this interval could reflect extreme outliers or inaccuracies. This choice was based on an analysis of the distribution of fare amounts, where values outside this range were observed to be highly unusual or unrealistic for typical Uber rides.
- **Ride distance:** Rides that have been measured at less than 1 km or over 80 km were excluded from use. This is based on the assumption that any ride below 1 km is quite

infrequent and might represent unusual or erroneous data points (e.g., rides not captured properly). Every ride longer than 80 km was excluded as well, which depicts outliers not characteristic of typical urban ride on demand service.

- **Geographical boundaries:** Longitudes and latitudes were bounded within known geographic boundaries for New York City: the purpose being to make sure analyses considered valid pickup and drop-off locations only. Any points beyond these defined limits were considered invalid and hence removed.
- **Passenger count:** The range of from 1 to 6 passengers was taken because this range precisely reflects the range that is possible for the number of passengers which typical ride on demand cars can carry. Rides having zero passengers and above six were treated as outliers, likely the result of data entry errors.

iv) **Feature engineering:**

New features were created to enhance the analysis of fare dynamics and temporal trends.

- **Ride distance calculation:**
  - a. One of the most important variables underlying the fare dynamic, ride distance, did not exist in the raw data. A function was used to calculate the distance between the pickup and drop-off points latitude and longitude coordinates using the Manhattan distance formula.
  - b. Other options considered were Google Maps API, Here API, and OSRM (Open Source Routing Machine), but each came with some issues related to financial costs, rate limit, and complicated configuration for large amounts of data points. These methods were really quite accurate in the calculation of road-based distances, but scaling and financial demands make them impractical for this research. Simpler methods like the Haversine formula and Euclidean distance were also considered but were less accurate for New York City's grid-like layout. Hence, the Manhattan distance formula was used for its balance between simplicity and accuracy. This avoids the cost and complexity of API requests for a much more realistic approximation of real-world distances and thus making it the most efficient and scalable solution for the dataset.
- **Time-based feature extraction:** Some temporal features were extracted from the variable `pickup_datetime`, which enabled the study of time-of-day and time-of-week temporal trends in fare prices. These features were useful to understand the changing fare rates as time of day and day of a week. Converting the `pickup_datetime` to a New York time zone was also important to accurately capture the local time dynamics, which also ensured that any temporal analysis aligns with real-world conditions.

v) **Feature selection:**

In addition to feature engineering, careful consideration was given to the selection of variables to be included in the modelling phase.

- **Target variable:** The target variable for prediction is labelled as `fare_amount`, as it represents the pricing mechanism of ride on demand services and remains the main focus of this research.
- **Selected features:** Selection of features in the list below has been based on its relevance to the research questions and their importance as a determinant of the fare dynamics:

- a. **Ride distance:** The ride distance, one of the strongest predictors of fare, was key component for modelling. That is highly correlated to the fare amount, with a correlation coefficient of 0.89, which makes it essential for understanding the base pricing structure of ride on demand services.
- b. **Passenger count:** Although the feature passenger count was explored in exploratory analysis and did not relate to the fare, it was retained to verify its contribution within the model. In case there are subtle patterns or interaction in feature which might be influencing the price, even if the direct effect seems small.
- c. **Time-based features:**
  - 1- **Hour:** This feature depends on the time of day, that is, whether it is peak or off-peak condition. Including this feature allows the model to capture these temporal variations, which are important for understanding dynamic pricing.
  - 2- **WeekDay:** Patterns of ride demand and pricing are influenced by both weekdays and weekends. Incorporating the day of the week allows the model to account for these behavioral variations.
  - 3- **Month:** This feature was included to capture potential seasonal variations in fare pricing, which may be relevant depending on the time span of the dataset.
- **Excluded Features:** Geographical coordinates such as pickup\_longitude and dropoff\_latitude were excluded since already a feature represented the distance of the ride, and that provided enough information. Even though the dataset spanned several years, the feature Year was excluded because it cannot be useful in short-term fare prediction. This study is focused on the prediction of fares by relying more on the short-term temporal features: the hour and day of the week.

At this point, after all the above-mentioned steps, the data reduced from 200,000 rows to 173,557 rows. Although the reduction is significant, it retained massive data for meaningful analysis keeping records valid and representative of normal conditions for the ride on demand service.

#### vi) **Data Transformation:**

Once the feature selections were made, the dataset had to undergo various steps of necessary transformation to meet the requirements of machine learning algorithms, assuming particular forms of distributions and feature formats. These transformations include scaling, encoding, and cyclical feature encoding.

- **Scaling of continuous variables (ride\_distance, fare\_amount):** Scaling prevents features with large magnitudes/units/ranges from having an undue impact on the model.
  - a. **Why it was necessary:** Some variables are of larger magnitudes like ride\_distance and fare\_amount, while most models are sensitive to large values if not normalized.
  - b. **Impact on the model:** Without scaling, large values like that for ride\_distance could dominate the learning process, making the models predict in a biased manner to not correctly reflect other feature's relative importance.
- **Cyclical encoding of temporal features (Hour, WeekDay):** Time features like hours and days are fundamentally cyclic, and this cyclical nature is not captured by a standard encoding.
  - a. **Why it was necessary:** Since time is not always a linear variable, treating it as such would distort proximity between successive periods such as hour 23 and hour 0.

- b. Impact on the model:** Transformed sine and cosine features helps the model understand temporal patterns better to make improved fare predictions concerning peak and off-peak hours.
- **Encoding of categorical variables (Month):** While months were in numeric format ranging from 1 to 12, it was critical not to treat them as ordinal.
  - a. Why it was necessary:** Ordinal treatment of the months would range from 1 to 12; implying that month 12 has an intrinsic value that is "greater" or "higher" than month 1. This will create unwanted relationships, making December "larger" than January.
  - b. Impact on the model:** One-hot encoding allows the model to capture the pattern of each and every month without making any wrong numeric sequence. That kind of approach depicts the month-wise trend, capturing seasonality accurately without feeding the model with any numeric hierarchy.
- **Encoding of categorical variables (passenger\_count):** In the passenger\_count feature, there was an obvious ordinal structure in the values that go from 1 to 6, representing increasing quantities
  - a. Why it was not transformed:** One-hot encoding would make each passenger count into a separate, unrelated category. It removes the indication of gradualness, and essentially tricks the model to take away the inherent hierarchy within the numbers of passengers themselves.
  - b. Impact on the model:** Leaving the passenger count in their raw format also allowed the model to understand it correctly as an ordinal feature for the proper capture of a relationship between increasing passenger numbers and fare dynamics without overcomplicating it.

Features in the Dataset

```

fare_amount
passenger_count
ride_distance
Hour_sin
Hour_cos
WeekDay_sin
WeekDay_cos
Month_1
Month_2
Month_3
Month_4
Month_5
Month_6
Month_7
Month_8
Month_9
Month_10
Month_11
Month_12

```

**Figure 3.9: Final features in the dataset after all previous steps.**

### 3.1.4 Modeling

The prime focus of this research was on providing a predictive model for dynamic pricing in a ride-on-demand service, where fare changes based on different contextual factors like demand-supply dynamics, time of day, ride distance, and other temporal patterns. Accurately predicting fares in such a dynamic environment required models capable of capturing both linear and non-linear relationships, as well as interactions between these

variables. Several machine learning models were selected based on their different capabilities in handling different natures of data and the relationship exist between the variables concerned.

**i) Models for dynamic pricing:**

The four models selected for this study include Linear Regression, Random Forest, Multi-Layer Perceptron, and Gradient Boosting Regressor. Each model was chosen based on the potential to capture various aspects of the data and to provide a robust dynamic pricing model.

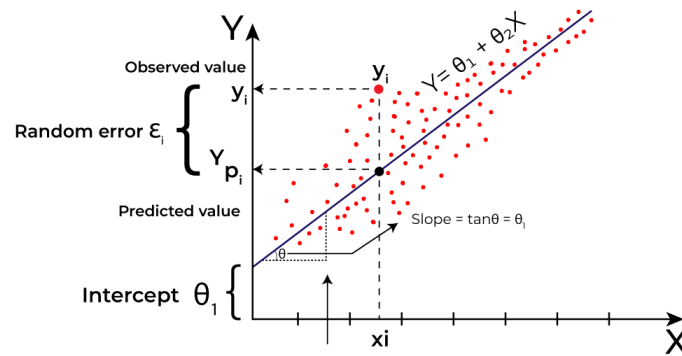
- **Linear regression:** Linear regression is one of the basic models that assumes a direct and proportional relationship between predictors and target variables; though simple, it is a baseline model offering interpretability and allowed the validation of another more complex model with respect to it. The equation for a simple linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

Where

- y is the target variable (fare amount),
- $\beta_0$  is the intercept,
- $\beta_1, \beta_2$  and  $\beta_n$  are the coefficients for each feature  $x_1, x_2$  and  $x_n$
- $\epsilon$  represents the error term.

In ride-on-demand services, linear regression had its limitation in terms of handling nonlinear patterns, but it was helpful to model the general linear trend of fluctuation in fares, such as proportionality in increases of fare with longer ride distances. If this model performed well, that would mean the relationship between the features and fare amount was mostly linear - a situation unlikely in dynamic pricing.



**Figure 3.10: Illustration of linear regression in which  $\theta_1 = \beta_0$  and so on. (Source: GeeksforGeeks, 2023)**

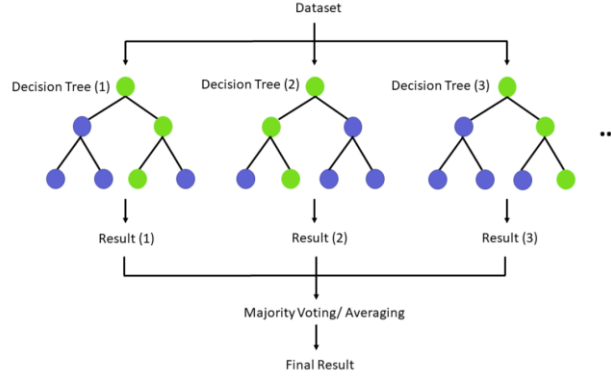
- **Random forest:** Random Forest is an ensemble method that builds multiple decision trees and averages their predictions. Each tree splits data based on the value of features to minimize impurity using the Gini impurity or entropy for classification or MSE (mean square error) for regression. The prediction from random forest is:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^M T_m(x)$$

Where

- $\hat{y}$  is the predicted fare,
- M is the total number of trees,
- $T_m(x)$  is the prediction from the mth tree.

In ride-on-demand pricing, Random Forest helped capture complex nonlinear interactions between time variation and variation in passenger count on fare. Each of the trees in the forest learned specific patterns in the variables, while aggregation of predictions over all trees actually resulted in a better overall prediction. However, given that the Random Forest did not account for sequential dependencies, it might fail to capture some of the temporal dynamics in the patterns of pricing.



**Figure 3.11: Illustration of random forest model with multiple decision trees.**

- **Multiple layer perceptron:** MLPs are neural network models that can learn any complex and nonlinear pattern in data. Because of their hidden layers with nonlinear activation functions such as the ReLU or sigmoid, MLPs can universally approximate a wide variety of functions, thereby making this a very flexible model to carry out prediction tasks. The equation for the output of each neuron is:

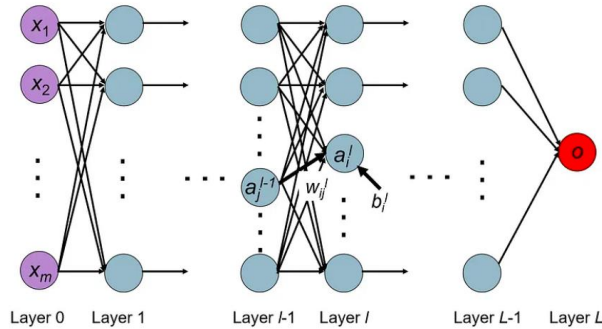
$$a^{(l)} = \sigma(w^{(l)}a^{(l-1)} + b^{(l)})$$

Where

- $a^{(l)}$  is the activation at layer  $l$ ,
- $w^{(l)}$  is the weight matrix,
- $b^{(l)}$  is the bias term,
- $\sigma$  is the activation function.

The network uses backpropagation to adjust weights, minimizing a loss function (e.g., mean squared error) to improve predictions.

For dynamic pricing, MLP had the advantage of complicated patterns in fare fluctuations caused by a combination of factors, such as time, distance, and passenger count. That, however, requires more data and computational resources in order to converge to an optimal solution and is less interpretable compared to the tree-based model, which may restrict practical usage in this case.



**Figure 3.12: Structure of a multi-layer perceptron (MLP). (Source: Towards Data Science, 2023)**

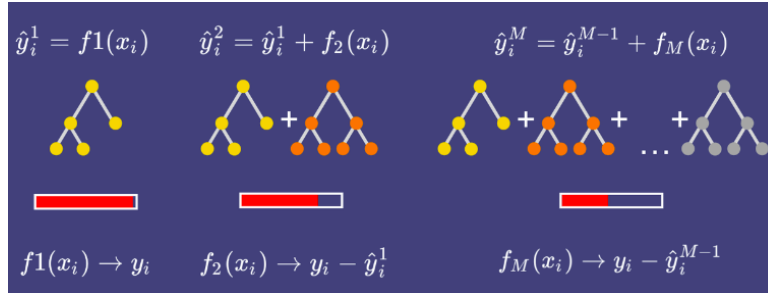
- **Gradient boosting regressor:** Gradient Boosting Regressor combines boosting with gradient descent optimization so that it builds an ensemble of trees sequentially. This method aims to minimize the prediction error; thus, it constructs each tree so that it corrects the residuals of the previous ones, which works well enough to capture both linear and nonlinear patterns:

$$\hat{y}^{(m)} = \hat{y}^{(m-1)} + \alpha f_m(x)$$

Where

- $\hat{y}^{(m)}$  represents the prediction after the  $m^{\text{th}}$  boosting step,
- $\hat{y}^{(m-1)}$  is the prediction from the previous step,
- $\alpha$  is the learning rate controlling the contribution of each new tree,
- $f_m(x)$  is the  $m^{\text{th}}$  tree fitted to the residuals of the previous ensemble.

Gradient Boosting was particularly suitable for dynamic pricing, as it captured those very small demand-driven shifts in price by building on previous errors. It did so through iterative model improvements and treated the fluctuations of fare, which depends on factors such as high demand and distances quite well. This iterative reduction of error rates ensured that each new tree in it targeted the difficult cases to predict, hence giving a much finer model responding to the variability of ride-on-demand services.



**Figure 3.13: Gradient Boosting process showing sequential trees (Source: Thorat, 2023)**

This selection strategy was designed to find the most suitable model for dynamic pricing, balancing accuracy, interpretability, and efficiency.

## ii) Hyperparameter tuning for gradient boosting regressor:

After initial model evaluation, Gradient Boosting Regressor was the most promising, since it could grasp the underlining complexity of the data and give correct predictions in the ever-dynamic environment. To further tune the Gradient Boosting model, Grid Search with 10-fold Cross-Validation was performed to find the best hyperparameters. The following hyperparameters and ranges were selected based on common practices in gradient boosting and their impact on model performance:

- **Learning rate (learning\_rate):** The learning rate regulates the step size of each iteration and defines the speed at which the model gets adapted to residual errors. While a low learning rate enables gradual learning of the model with reduced chances of overfitting, a higher rate speeds up convergence.

The values tested for learning\_rate were [0.01, 0.1, 0.2]. A value of 0.01 tests a conservative learning rate to prioritize generalization over quick adaptation, while 0.1, the default in gradient boosting, balances generalization and convergence speed. The value of 0.2 represents a faster learning rate to see if quicker adaptation improves accuracy in dynamic pricing.



- **Number of estimators (n\_estimators):** It controls the number of boosting rounds (trees) added to the ensemble. More trees available allow the model to capture finer detail but at the cost of a higher risk of overfitting.

The values tested for n\_estimators were [50, 100, 200]. The value of 50 test how a small number of trees can influence the performance of the model and may help in preventing overfitting. The value of 100 being a moderate value has a good balance between computation efficiency and model complexity. The value of 200 allows for a larger number of trees to model finer relationships in data.

- **Max depth (max\_depth):** It limits the depth of each tree within the ensemble and, in turn, limits model complexity. Deep trees are able to capture more complex patters but at the expense of overfitting to training data.

Values tested for max\_depth were [3, 5, 7]. A depth of 3 was used to favor shallow trees, which do not easily overfit and give better generalization. A depth of 5 allowed for moderately deep trees to model more interaction between variables, while a depth of 7 tested the impact of deeper trees to see whether they might capture any useful pattern present in high variability fare predictions.

By applying 10-fold-cross-validation it was ensured that the best set of hyperparameters generalize well to random subsets of the data and avoid overfitting. This method allowed for a robust selection of parameters, focusing on those that minimized the mean absolute error (MAE) across different folds, aligning with the objective of minimizing fare prediction errors.

### iii) **Final model selection:**

Based on the hyperparameter tuning results, the Gradient Boosting Regressor with learning\_rate = 0.1, max\_depth = 5 and n\_estimators = 100 was chosen as the final model. This gives a good balance between model complexity and adaptability, correctly predicting fares under dynamic conditions with no overfitting.

## 3.1.5 Evaluation

For this study, three relevant metrics were considered to evaluate the predictive accuracy and robustness of the models by using the MAE, RMSE, and R<sup>2</sup>. These three metrics have been chosen to fully understand model performance in the context of dynamic pricing for ride-on-demand services with regard to the accuracy of model.

### i) **Mean absolute error (MAE):**

MAE represents the average absolute difference between the predicted and actual fare amounts. It is calculated as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Where

- $y_i$  is the actual fare,
- $\hat{y}_i$  is the predicted fare,
- $n$  is the number of observations.

The main metric chosen here was the MAE, since it gave the average prediction error, which is quite relevant for dynamic pricing. In the case of pricing, the model errors had to be

smaller as that kept the pricing consistent with the real time and minimum dissatisfaction among the customers or losses in revenues from incorrect pricing.

It meant with a lower MAE, the model was consistent with its prediction of fares close to the actual amounts, hence it was stable in pricing adjustments. A model with a smaller MAE was better positioned to make accurate fare predictions, essential for optimizing ride-on-demand pricing strategies.

**ii) Root mean squared error (RMSE):**

RMSE is the square root of the average squared differences between the predicted and actual values, calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Where

- a.  $y_i$  is the actual fare,
- b.  $\hat{y}_i$  is the predicted fare,
- c.  $n$  is the number of observations.

RMSE was added to handle the sensitivity of fare predictions for larger deviations. In dynamic pricing, seldom large errors can be very dangerous since the significantly mispriced fares may contribute to very dissatisfied customers or operational inefficiencies.

The lower values of the RMSE show that the model has accurate average predictions and also fewer chances of substantial mispredictions. This is particularly important for ensuring that fare adjustments remain consistent and reliable across a range of pricing scenarios, especially during peak or off-peak times.

**iii) R-Squared ( $R^2$ ):**

$R^2$ , or the coefficient of determination, represents the proportion of variance in the target variable (fare amount) that is explained by the model. It is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Where

- a.  $y_i$  is the actual fare,
- b.  $\hat{y}_i$  is the predicted fare,
- c.  $\bar{y}$  is the mean of the actual fare,
- d.  $n$  is the number of observations.

$R^2$  was chosen to complement the MAE and RMSE metrics because it gives the overall fitness of the model. In this instance, for dynamic pricing, the high value of  $R^2$  will mean that the model is therefore able to capture the varying factors such as time, distance, and demand pattern.

The larger the  $R^2$ , the more the model captures and explains the underlying trends in fare changes. This aligns with the goal of research of creating a dynamic pricing model that accurately reflects real-time conditions.

Together, MAE, RMSE, and  $R^2$  were a good balance of metrics that worked well together. The MAE and RMSE were focused on prediction accuracy, with the RMSE providing

an indication of larger errors that the model produced. Meanwhile,  $R^2$  served to confirm that the model captured the primary drivers of fare variability. This provided the guarantee that the selected model was not only accurate but also aligned with the practical requirements of dynamic pricing.

### **3.1.6 Deployment**

In this deployment phase, the primary objective was to determine the ability of the model in adaptability and efficiency in the prediction of fares, using simulated real-world dynamic pricing for ride-on-demand services. This phase did not involve a live production deployment; instead, it relied on hypothetical but realistic scenarios to investigate how well it was able to capture the variations in fare predictions according to different ride conditions.

To ensure accurate testing, different scenarios for testing were developed based on changing the attributes of the ride - time of day, day of the week, month, number of passengers, and distance. Each scenario was carefully designed to reflect typical factors affecting fare amounts, such as peak or off-peak hours and seasonal trends. For example, one scenario involved predicting the fare at 10:55 pm on a Tuesday in July for a 10-kilometer ride with five passengers.

#### **i) Data transformation for scenario testing:**

A transformation function was created so that the real-life case data could fit what was expected by the model at the time of making a prediction. This function accepted raw attributes from each scenario (including ride distance, hour, weekday, month, and passenger count) and transformed them according to the steps followed during the pre-processing of the training of the model.

#### **ii) Input array construction and model prediction:**

Once the above transformations were applied, these variables were then combined in a structured array format to match the expected input of the model. The array included the scaled distance, the passenger count, the sine and cosine transformations of hour and weekday, and the twelve one-hot encoded month features. This final input array was then used to generate a fare prediction from the trained model.

#### **iii) Model evaluation and analysis of predictions:**

The predicted fare was then compared to expected patterns for the given scenario. The analysis was based on how well the model could dynamically react to major pricing drivers, such as temporal variations (e.g., peak vs. off-peak pricing), seasonal variations (e.g., possible fare increases in months of high demand), and distance sensitivity. This gave insight into the robustness and effectiveness of the model in reacting under real-world conditions.

The multi-scenario testing gave a broad view of how the model would perform under dynamic pricing conditions. This is a very critical phase, as it showed how well the model could react to different conditions—a very good assessment of its performance and dynamic price capability before any real deployment.

The following section presents the results and discussions, where in-depth detail on the findings acquired during exploratory data analysis and model performance evaluation is given. This chapter discusses how the results obtained by the exploratory data analysis and model

performance evaluation answer the research questions. In addition, it provides insight into dynamic pricing mechanisms and how effective the models are.

## 4 Results and Findings

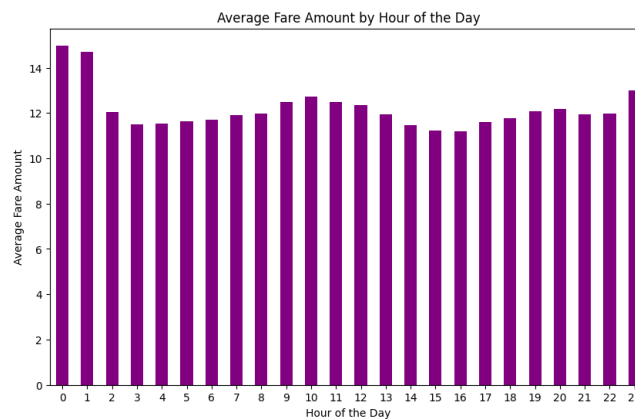
The findings presented in the following chapter represent the major results of the research, including exploratory data analysis and model performance evaluation. Furthermore, this section will elaborate on how the findings address the research question and objectives, pointing out positive and negative findings. Moreover, it highlights findings in terms of the existing literature to compare these with the existing study.

### 4.1 Key insights from exploratory data analysis (EDA)

Important insight into the variable-variable relationship was obtained in the exploratory data analysis phase and hence the basis for the development of robust dynamic pricing models. The following key patterns and trends were identified:

#### 4.1.1 Fare trends and ride volume by hour analysis

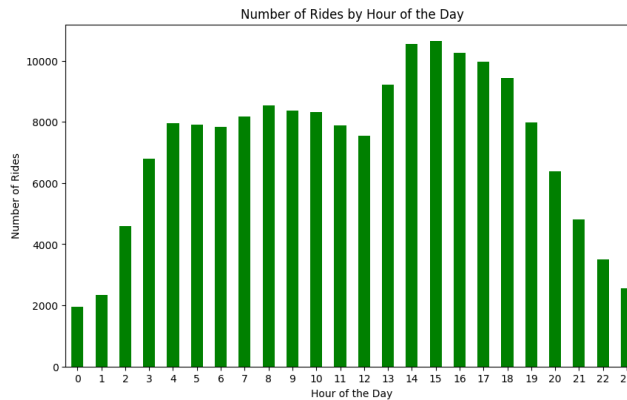
The trend of fare throughout the day follows quite distinct patterns related to the supply and demand variability throughout the day. Figure 4.1 clearly illustrates that the average fare ranges very much throughout each hour of the day.



**Figure 4.1: Average fare amount by hour of the day.**

- We can see the trend of higher fares between 12 AM and 2 AM, where the average fare went up to approximately \$14.98 at midnight. Of course, that would be seen with fewer driver availability, resulting in surge pricing to incentivise drivers.
- Morning fares increase moderately from 9 AM onwards up to 10 AM because of commuting demand. However, the predictable demand and enough drivers mean that the fare spikes are less dramatic compared to late night.
- Fares during evening commutes, 5 PM to 7 PM, do not change much and show a moderate increase similar to morning fares, as there is a good balance of supply and demand, which prevents any spike in fares.

The number of rides varies throughout the day, influenced by daily habits and travelling needs. Looking at these patterns in Figure 4.2 we can see how the number of rides and fares are influenced by driver availability and demand.



**Figure 4.2: Number of rides by hour of the day.**

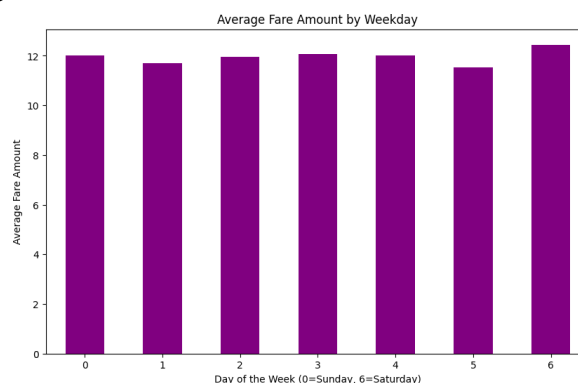
- The number of rides between 2 pm and 4 pm are high. Though this is not the peak time of day, it can be driven mostly by activities such as business trips, shopping, and personal errands. Since the supply of drivers is balanced with the demand for rides, high volumes do not force the fares up.
- Low volumes of rides are seen in the late-night hours from 12 AM to 2 AM, but with much higher fares, indicating a high supply-demand imbalance in such hours.

These findings further indicate that dynamic pricing for ride-hailing services is more about the supply of drivers rather than the increase in demand. Increased fares during late nights, when volumes of rides are low, actually establish the fact that shortages in supply could be a stronger driver of increased fares compared to demand itself. On the other hand, peak-hour pricing shows efficient demand and supply balance, with less severe dynamic pricing due to better resource allocation.

The much smaller fare variation between peak and off-peak periods would suggest that dynamic pricing applies mostly when a supply-demand mismatch is obviously large, which is during the late-night off-peak periods. In other words, compared to fluctuations in demand alone, the dynamic pricing systems normally would be much more sensitive with respect to fluctuations in driver availability.

#### 4.1.2 Weekly fare trends and ride volume analysis

The analysis of fare trends throughout the week also highlights some variation in prices, which is depicted in Figure 4.3 below.

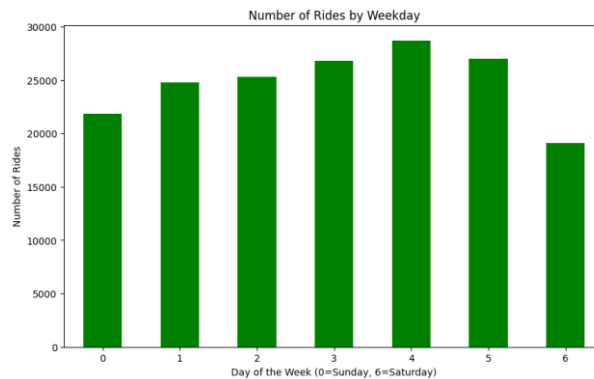


**Figure 4.3: Average fare amount by weekday.**

- The Saturdays have the highest average fare, at \$12.44, driven by increased weekend travel and fewer drivers, which results in surge pricing to equilibrate the demand-supply gap.

- The lowest fare days are Fridays, surprisingly, at \$11.53, because of a higher driver supply anticipating weekend demand and with ride requests more uniformly distributed.
- Midweek, it increases slightly to \$12.08 on Wednesdays and to \$12.01 on Thursdays, due to greater business travel and commutes.
- On Sundays, it is also comparatively high at \$12.00, reflecting leisure travel and lower driver supply.

The pattern of ride volume by weekday indicates that demand is far greater on the weekdays because of commuting and business travel demands.



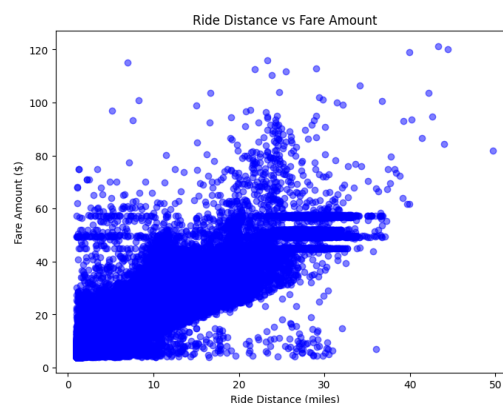
**Figure 4.4: Average number of rides by weekday.**

- Thursdays have the maximum rides-28,705-which indicates the trend in the middle of the week. During weekends, however, rides decreased considerably, having a minimum of 19,128 on Saturdays and 21,855 rides on Sundays because of the shift to leisure from home-to-work travels.

These results further support the hypothesis of the weekday analysis: dynamic pricing in ride-on-demand service is more supply-side constrained than driven by a simple variation in demand. Although demand clearly drives the price of fares, data shows large increases in fares are much more likely to happen when there's a shortage of drivers, an observation given particular notice on weekends. In this regard, driver availability acts as one crucial determinant of price stability.

#### 4.1.3 Influence of ride distance and passenger count on fare

Ride distance is a key feature of pricing models in ride on demand service. Figure 4.5 shows how ride prices depend on ride distance.



**Figure 4.5: Scatter plot of distance of ride vs fare amount of ride.**

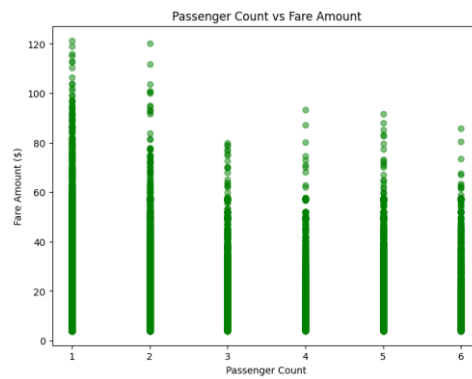
- A correlation coefficient of 0.89 implies a strong positive relationship between the ride distance and the fare amount, where with distance travelled, the amount of fare increases directly.

	ride_distance	fare_amount
ride_distance	1.000000	0.890602
fare_amount	0.890602	1.000000

**Figure 4.6: Correlation coefficient between ride\_distance and fare\_amount.**

- This trend is visually confirmed by the scatter plot in Figure 4.5, with longer rides distinctly yielding higher fares. Such would make sense in standard pricing models, where distance is a major factor

The analysis demonstrates that the number of passengers exerts a nearly negligible influence on fare prices, as illustrated in Figure 4.7.



**Figure 4.7: Scatter plot of passenger count vs fare amount.**

- The relationship between the number of passengers and the fare amount is only 0.0136, suggesting almost no relation among the variables. This is further supported by the scatter plot that shows no clear pattern between the variables.

	passenger_count	fare_amount
passenger_count	1.000000	0.013643
fare_amount	0.013643	1.000000

**Figure 4.8: Correlation coefficient between passenger\_count and fare\_amount.**

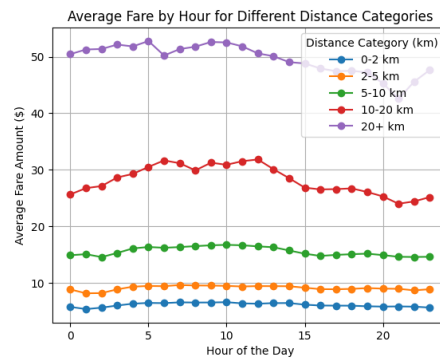
- The ride-on-demand services do not charge per head but according to distance travelled and time of day.
- This makes sense because operationally, more passengers do not cost the drivers much, nor does it cost much to the service.

These findings are in line with typical mechanisms of pricing in ride on demand services, whereby ride distance is a critical determinant of fare pricing, whereas passenger count is pretty irrelevant. The high correlation of distance to fare shows how well ride-hailing platforms optimize pricing for the effort and cost involved in longer trips.

#### 4.1.4 Average fare by hour for different distance categories

This analysis of fare variation at different times of the day, grouped by ride distance, serves as a basis for studying how ride distance and temporal factors jointly determine fare prices. As expected, this has shown clear trends in the adaptive pricing strategies of the ride-hailing platforms. Table shown in figure 4.10 summarizes the average fares by hour of day

across the different classes of distance, giving a quantitative perspective to the trends shown in figure 4.9.



**Figure 4.9: Average fare by hour for different distance categories.**

distance_category	0-2 km	2-5 km	5-10 km	10-20 km	20+ km
Hour					
0	5.807434	8.889985	14.932921	25.673714	50.442072
1	5.419493	8.208444	15.118139	26.781860	51.267102
2	5.672241	8.244396	14.602834	27.143220	51.353600
3	6.055785	8.908970	15.333448	28.668254	52.134059
4	6.371112	9.377974	16.133848	29.290084	51.804593

**Figure 4.10: Table of average fare by hour for different distance categories (km)**

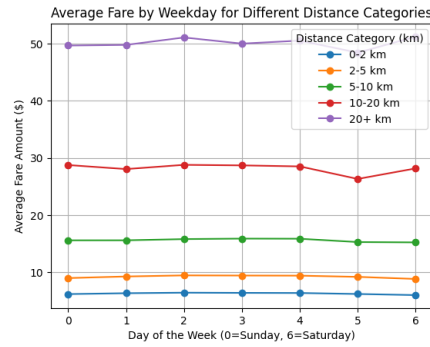
- Short journeys (0-2 km) have constant prices throughout the day, ranging from \$5.41 at 1 AM to \$6.37 at 4 AM. The rate increases because there are fewer drivers in the early morning.
- Medium-distance rides show moderate fare variation, such as rides that are 2-5 km varies from \$8.20 at 1 AM to \$9.42 at 3 AM, while 5-10 km cruises vary from \$14.60 at 2 AM to \$16.13 at 4 AM. It shows marked growth during the early morning hours, such as from 3 to 5 AM due to scarcity and stabilization in the daytime, from 6 AM to 7 PM.
- Long-distance rides, both 10-20 km and over 20 km, have the most significant growth in fares during this time of the day. Fares for 10-20 km-long rides increase from \$25.67 at midnight to \$29.29 at 4 AM, while trips over 20 km are over \$50, peaking at \$52.13 at 3 AM, reflecting increased operating costs and fewer drivers on the road.

Analysis also reveals that in all ranges, the fares tend to stabilize during the day—that is, between 6 AM and 7 PM—at which the fare structure balances driver supply against passenger demand and prevents sudden increases in fares. However, the fares for the above 20 km travel range always maintain their high value, reflecting the basic costs involved in longer travel.

#### 4.1.5 Average fare by weekday for different distance categories

The fare trend analysis against different days of the week conveys important information with respect to ride distance on how dynamic pricing works at different lengths of the ride. The results, visualized in figure 4.11, and summarized in figure 4.12 in tabular form show how ride distance influences fares and whether the variations on a daily basis will bring significant changes in pricing.





**Figure 4.11: Average fare by weekday for different distance categories (km)**

distance_category	0-2 km	2-5 km	5-10 km	10-20 km	20+ km
WeekDay					
0	6.182455	8.983602	15.578990	28.746761	49.644939
1	6.333246	9.259876	15.584227	28.038999	49.755397
2	6.427787	9.448649	15.798541	28.782074	51.055251
3	6.393484	9.423165	15.877815	28.686404	49.979872
4	6.370968	9.404104	15.859476	28.508639	50.503884

**Figure 4.12: Table of average fare by weekday for different distance categories (km)**

- Short rides are 0-2 km and throughout the week have fairly constant pricing, ranging from \$6.18 on Sunday to a high of \$6.43 on Tuesday, illustrated in Figure 4.12. This is because demand for short rides is relatively stable and driver supply is ample, so prices remain low across all days.
- Medium-distance rides, between 2 to 5 km and 5 to 10 km also display consistent prices. The fares for rides between 2-5 km range from \$8.98 on Sunday to a high of \$9.45 on Tuesday. Rides from 5 to 10 km are similarly ranged from \$15.58 to \$15.88 during this timeframe. Figure 4.11 shows these consistent trends likely from predictable commutes.
- The long-distance rides are a bit more varied: 10-20 km and 20+ km. For 10-20 km trips, Sunday fares begin at \$28.50 and increase on Tuesday to \$28.78. Rides of 20+ km peak on Tuesday to \$51.05 from \$49.64 on Sunday. This small variation perhaps might reflect shifting demand or supply on particular days for longer trips.

Results highlight how the length of a ride is the main basis of calculating fares, with very minimal influences of variability related to weekdays. Such stability suggests that dynamic pricing schemes have been more sensitive to other factors, such as time of the day or length of a ride, rather than to daily shifts in passenger demand.

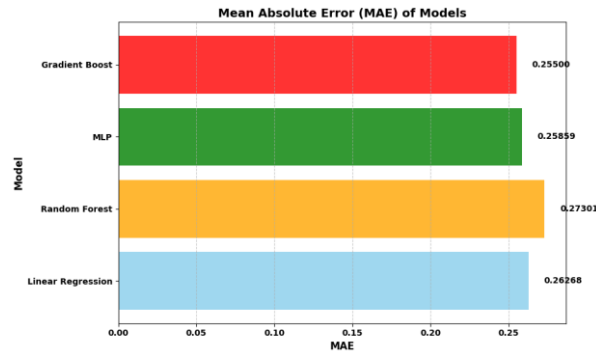
## 4.2 Evaluation of machine learning models

The machine learning models were evaluated in terms of the overall performance by three major metrics: Mean Absolute Error, Root Mean Squared Error, and  $R^2$ . These metrics provide insights regarding model accuracy, robustness, and explanatory power and therefore enable an in-depth assessment of models regarding their suitability for dynamic pricing. Every metric represents another predictive performance aspect; therefore, it allows more options to compare the comprehensive performances of different models.

### 4.2.1 Based on mean absolute error (MAE)

The MAE has been used to evaluate the models because it calculates the average difference between the predicted and actual fares. Thus, low MAE values mean higher

accuracy, which is very important in fulfilling the goal of this research-reliable dynamic pricing.



**Figure 4.13: Comparison of machine learning models based on MAE**

Gradient Boosting had a minimum MAE of 0.2550, hence is most accurate in predicting fares. This probably has an iterative refinement process that can best capture non-linear relationships within, making it most suitable for dynamic pricing.

MAE of the MLP was 0.2586, which is slightly higher than the gradient boosting. Although extremely powerful in modelling complex patterns, this performance suggests that it does not generalize as well as Gradient Boosting, partly because of its higher sensitivity to hyperparameter tuning.

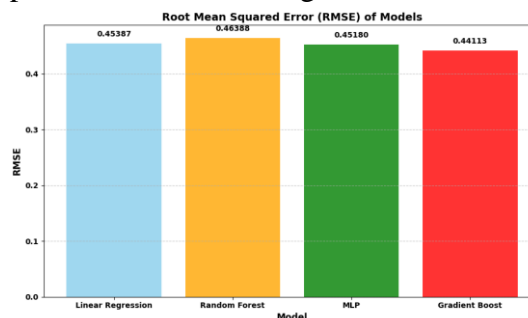
The best MAE for the Random Forest was 0.2730, reflecting the poorest accuracy of all models. While very robust at handling nonlinearities, without the capability to refine residual errors, it can't be as effective in capturing dynamic relationships.

Linear Regression had an MAE of 0.2628 and, therefore, outperformed Random Forest. It has a linearity assumption that restricts capturing the higher-order nonlinear interactions of variables; hence, the model is of minimum suitability in dynamic pricing.

Gradient Boosting outperformed all other models in the lowest MAE to manifest its capability in modelling effective dynamic fare adjustments. MLP and Linear Regression came in as competitive but less accurate, while Random Forest underlined the limitation of ensemble models without sequential refinement on dynamic pricing tasks.

#### 4.2.2 Based on root mean square error (RMSE)

RMSE is the metric that represents the average magnitude of the prediction errors, but it puts greater emphasis on the bigger errors due to the squaring of the residuals. The smaller the value of RMSE is, the higher the capability of the model in reducing large deviations in prediction. It is thus an important metric in testing the robustness of fare predictions.



**Figure 4.14: Comparison of machine learning models based on RMSE**

The Gradient Boosting algorithm performed very well with an RMSE of 0.4411, showing how capable it is in keeping huge prediction errors at their lowest. It works iteratively to refine residuals and hence is very reliable for dynamic pricing tasks with huge fare changes.

MLP yielded an RMSE of 0.4518, relatively higher than that of Gradient Boosting. While it models nonlinear relationships quite well, the higher RMSE suggests it sometimes has more significant deviations, a likely consequence of its non-iterative settings.

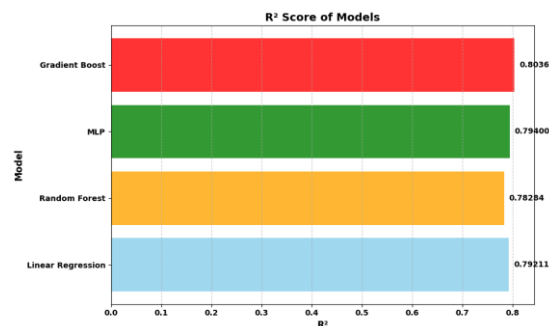
Linear Regression gave an RMSE of 0.4539, indicating moderate performance. It represents the general trend well but cannot handle nonlinear interaction and hence can be prone to larger errors in dynamic conditions.

The random forest gave the highest RMSE-0.4639, showing lower effectiveness in minimizing big errors. Averaging across trees smooths the predictions but at a cost, limiting the ability to handle outliers or extreme values common in dynamic pricing.

Gradient Boosting had the lowest RMSE, reflecting that it provides more accurate and robust fare predictions. The moderate performance was by MLP and Linear Regression, while Random Forest had the largest errors. That also points to the necessity of using advanced ensemble approaches, such as Gradient Boosting, which helps to minimize high-value deviations in the predictions.

#### 4.2.3 Based on $R^2$ score

The  $R^2$  score, or coefficient of determination, gives a measure of the variability of the target variable explained by the model. The best performance is represented by a value of  $R^2$  close to 1, where it means that the model can explain, with good approximation, the underlying pattern of data.



**Figure 4.15: Comparison of machine learning models based on  $R^2$**

Among the different models, Gradient Boosting yielded the highest  $R^2$  of 0.8036, indicating that it explains more than 80% of the variance in fare predictions. This points out an extraordinary capability in the modelling of complex and dynamic relations in the current dataset; hence, the best model for dynamic pricing.

The  $R^2$  score of the MLP was 0.7940, hence strong performance in explaining the variability. Being able to handle nonlinear relationships makes it very apt at modeling complex trends in the data, for example, those arising from the temporal and distance-based effects. Still, with this excellent performance, Gradient Boosting outperformed MLP, due to the simple reason that the former handles residual errors in a much more sequential way.

Linear Regression is next at 0.7921 to explain most of the variance. This clearly indicates that it does well in the general capturing of trends, such as the linear relationship between the distance and fare. Its inability to capture nonlinear interactions, such as time-based demand fluctuations interacting with each other, makes it less effective for dynamic pricing.

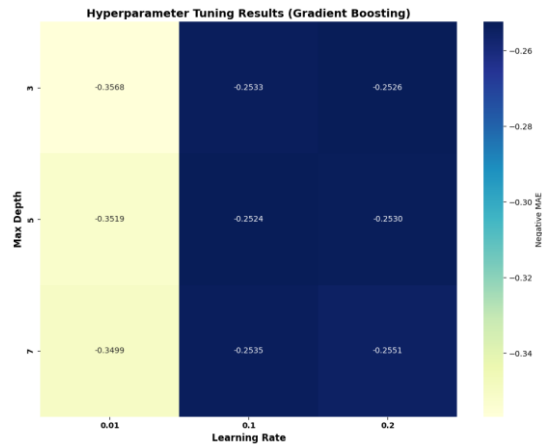
The lowest  $R^2$  score obtained by the Random Forest model is 0.7828, indicating a limited explanatory capability compared to the other models. While it captures non-linear interactions well, averaging the effects over many trees can sometimes result in a loss of ability to model finely detailed variation in highly variable data like temporal impacts. Such failure to model detailed variance is probably one of the reasons for the lower  $R^2$  score.

Gradient Boosting had the highest  $R^2$ , which explained the variance in fares very clearly. Next were MLP and Linear Regression: MLP because of its non-linear pattern grasping capability, and Linear Regression because of the representation of the overall trend. Random Forest came last, probably because it failed to capture the detailed variability required for dynamic pricing. From these observations, it follows that any effort dealing with dynamic and complex data has to be iteratively refined and nonlinear in nature.

Among them, Gradient Boosting has proved to be the most accurate and reliable model. It outperformed all other models in every metric: MAE, RMSE, and  $R^2$ . A 10-fold Grid Search Cross-Validation was done to optimize this model further by tuning its hyperparameters like the number of estimators, maximum depth, and learning rate. In this way, a very robust model was obtained, suitable for dynamic pricing.

### 4.3 Results of hyperparameter tuning for gradient boosting regressor

Optimizing the hyperparameters of the Gradient Boosting Regressor was done using grid search supported by 10-fold cross-validation. Indeed, this made the model perform better for all metrics with the optimal parameters of learning rate = 0.1, max depth = 5, and n\_estimators = 100.



**Figure 4.16: Heatmap of negative MAE for Gradient Boosting hyperparameter tuning.**

The performance on the optimized model gave a MAE of 0.2544, a RMSE of 0.4408, and an  $R^2$  score of 0.8039 which showed good accuracy and robustness. These results really underline that hyperparameter tuning was effective in further improving the performance of dynamic pricing tasks of the model.

A Mean Absolute Error of 0.2544 means that the model's fare predictions, on average, deviate about \$0.25 per ride, which is very negligible and acceptable in most business scenarios. That would guarantee pricing accuracy which balances customer satisfaction and revenue optimization.

## 4.4 Simulated deployment and testing results of model on real-world scenarios

The following different types of analyses (table 4.1) were run, related to the efficiency of a dynamic pricing model for scenarios developed to answer the research question: How does dynamic pricing respond to variations in ride parameters, time factors, and exogenous factors such as seasonality or driver supply? All of these scenarios were developed with due care in order to shed sufficient light on how the model performs under realistic and variable conditions.

**Table 4.1: Fare Predictions Across Simulated Real-World Scenarios.**

Scenario	Scenario A	Predicted Fare (A)	Scenario B	Predicted Fare (B)
1. Same Ride with Different Passengers	Monday, 8:00 AM, 10 km, 1 passenger	\$26.79	Monday, 8:00 AM, 10 km, 4 passengers	\$27.15
2. Off-Peak vs. Peak for the Same Ride	Monday, 10:00 AM, 5 km, 2 passengers	\$12.20	Monday, 6:00 PM, 5 km, 2 passengers	\$11.04
3. Weekday vs. Weekend Morning	Thursday, 8:00 AM, 10 km, 3 passengers	\$27.25	Saturday, 8:00 AM, 10 km, 3 passengers	\$23.64
4. Weekday vs. Weekend Late Night	Thursday, 11:00 PM, 10 km, 3 passengers	\$20.30	Saturday, 11:00 PM, 10 km, 3 passengers	\$19.96
5. Seasonal Pricing	April, 10:00 AM, 7.5 km, 1 passenger	\$17.69	December, 10:00 AM, 7.5 km, 1 passenger	\$18.36
6. Supply-Driven Pricing	Saturday, 1:00 AM, 15 km, 3 passengers	\$29.94	Saturday, 3:00 AM, 15 km, 3 passengers	\$30.98

### 4.4.1 Same ride with different passengers

- **Scenario A:** A ride with 1 passenger (Monday, 8:00 AM, 10 km) resulted in a predicted fare of \$26.79.
- **Scenario B:** Keeping all other parameters constant and increasing the passenger count to 4 increased the fare to \$27.15.

The minimal price increase would then suggest that the model does not give much weight to the number of passengers in determining the fares. This agrees with the industrial facts, as additional passengers are not always increasing the fare unless it reaches the limit capacity. The slight difference may, however, incorporate small operational costs such as vehicle wear or comfort adjustments.

### 4.4.2 Off-peak vs. peak for the same ride

- **Scenario A:** A ride during off-peak hours (Monday, 10:00 AM, 5 km, 2 passengers) costs \$12.20.
- **Scenario B:** At peak evening hours-6:00 PM on Monday, with the same parameters, the fare was slightly lower at \$11.04.

The fare went down during peak hours, contrary to surprise. This unexpected outcome of a normally quite predictable trend might indicate either that the model fails to capture a rise in demand typical for peak hours, or it overcompensates for drivers' availability or efficiency of routes at peak hours.

### 4.4.3 Weekday vs. weekend morning

- **Scenario A:** A ride that occurred on Thursday morning at 8:00 AM, 10 km long and with 3 passengers, cost \$27.25.
- **Scenario B:** A similar ride on the weekend, for the same time and settings, cost a little cheaper, priced at \$23.64.

The fare reduction on weekends corresponds to a drop in demand for commuting rides on non-working days. This is a suggestion that the model is appropriately incorporating weekday versus weekend effects. It raises, however, questions of whether the model might be undervaluing weekend demand for leisure or recreational trips through such a significant differential pricing.

#### 4.4.4 Weekday vs. weekend late night

- **Scenario A:** For a weekday late-night ride, Thursday at 11:00 PM, traveling 10 km with 3 passengers, the fare was \$20.30.
- **Scenario B:** During the weekend on Saturday, given the same timing and parameters, the price dropped slightly to \$19.96.

The price structure for late-night hours showed minimal differences between weekdays and weekends. This might indicate that the model is not fully capturing the increased demand related to nightlife travel on weekends. Also, the persistence of this fare level might indicate homogenous late-night pricing without accounting for changes in demand or constraints in supply, particularly over weekends.

#### 4.4.5 Seasonal pricing

- **Scenario A:** A spring ride in April, 10:00 AM, 7.5 km distance, with 1 passenger was \$17.69.
- **Scenario B:** The same ride in the holiday season-that is, in December under the same parameter-was priced slightly higher at \$18.36.

This model fits really well to describe the small price increase around the holidays, presumably due to anticipated demand. The difference is such a small one that it raises questions as to whether or not this model accounts for large increases in demand, seen more normally over holidays.

#### 4.4.6 Supply-driven pricing

- **Scenario A:** A late-night ride with moderate driver availability costs \$29.94 for a Saturday at 1:00 AM for 15 km with 3 passengers.
- **Scenario B:** Further shortening of driver supply (Saturday, at 3:00 AM, with the same parameters) gave a more-than-regular price increase: \$30.98.

It reflects the supply constraint through the surge in fare, but that marginal increase is rather minute, reflecting cautious adjustment, and may not capture the urgency/willingness to pay when there is a severe driver shortage.

### 4.5 General observations and critical insights

The model is sensitive to the parameters of passenger count, day of the week, and seasonality. Yet it makes rather conservative adjustments and thus cannot be sufficiently responsive in case of a sharp supply-demand imbalance. Price cuts during peak hours and late nights at weekends reflect the inadequacy of accounting for temporal surges of demand either for the lack of peak hour training data or over-compensation of driver supply. Moreover, minor

fare variations associated with differing passenger numbers suggest a diminished emphasis on this parameter, which is consistent with industry standards; however, this may neglect instances where the presence of extra passengers substantially impacts expenses or capacity.

Results have shown both strengths and weaknesses concerning the model's dynamic fare adaptation capabilities: it integrates some demand-related factors quite well, like differentiating between weekdays and weekends or capturing the seasonality of demand but does poorly on other factors like peak-hour demands or nightlife on weekends. These results underline further refinements that the model needs, especially in terms of capturing demand spikes and optimizing fare adjustments for real-world conditions.

## **4.6 Comparative analysis with literature**

This study presents the results that are consistent with the existing literature on dynamic pricing in ride-on-demand services and extends it. A detailed comparison is as follows:

### **4.6.1 Dynamic pricing models**

Guo et al. (2017) have stressed real-time information on dynamic pricing, considering location and temporal variables as critical variables. Likewise, the present study has established that peak and off-peak temporal variations create great impacts that bring about variation in fares; this therefore justifies their argument that dynamic pricing is multivariate and has to change in real time if it is to perform optimally.

Sun et al. (2020) added external factors such as traffic and environmental data in order to further refine the dynamic pricing models. While this study did not involve any traffic or environmental conditions variable, it also showed a strong dependence of fare on the distance, an essential parameter to improve dynamic pricing.

### **4.6.2 Machine learning applications in dynamic pricing**

Guo et al. (2018) have shown that neural networks are efficient for dynamic pricing predictions with remarkable precision using multi-source data from cities. Although this study found the most powerful model was Gradient Boosting Regressor, results support their conclusion that advanced machine learning methods can handle such complex and high-dimensionality data to reach superior performance.

Nalamothu (2023) compared several machine learning models and obtained the best performance from Random Forest among KNN and SVM. Although this study determined Gradient Boosting Regressor to be the best model, both these works highlight the model selection strategy to be adapted depending on the problem at hand and data complexity.

### **4.6.3 Insights on temporal and spatial dynamics**

The study by Luo et al. (2017) established how dynamic pricing can reduce waiting times and hence improve service efficiency. This study further supports the previous conclusions on surcharges during peak hours and pricing based on distance, in the sense that these approaches guarantee the optimality of resource allocation with customer satisfaction

## **4.7 Limitations and implications**

This section highlights the key limitations that arose during the study and their broader implications for further research and practical application in dynamic pricing of ride-on-demand services.

#### **4.7.1 Limitations**

The dataset used in this study ranges from 2009 to 2015; as such, there is a limitation in the data itself. Information on recent trends and developments within ride-on-demand services is not given, such as real-time incorporation of traffic, shifting consumer behaviors, and updated regulations, which are highly important for dynamic pricing today.

Also, the distance here is calculated using the Manhattan formula based on grid-like city layouts and doesn't consider real roads, which can blur the exact prediction of fares for those cities where road structures are disorganized. Besides, real factors such as live driver availability, traffic disruptions, or disturbances in weather conditions are not included in this dataset and hence shrink the scope of analysis.

#### **4.7.2 Implications**

Despite these limitations, these findings set a very strong foundation upon which dynamic pricing models can be built. The results underline the high importance of the temporal and spatial variables, hence laying a foundation for incorporation into further studies of even more advanced features, real-time traffic data, and environmental factors. Applying this knowledge with real-time and full-scope datasets could extend the scope of pricing models and their robustness in order to arrive at more accurate fare predictions and closeness to the current market conditions.

The present study underlined certain limitations of traditional data and methods; therefore, it requires the development of new solutions such as hybrid models or simulations able to represent reality in all its complexity. This would also contribute to enhancing dynamic pricing strategies toward an optimum balance between profitability and customer satisfaction.

The findings are critically evaluated in the next chapter related to the research objectives and the existing literature. Furthermore, this chapter covers discussion related to the limitation of the study, practical implications, and future directions for research.

## **5 Discussions**

This study was focused on the application of machine learning in enhancing the pricing strategy of ride-on-demand services. Results have underlined important insights into fare determination strategies and showed some areas in which the methodologies of this study could be improved. This section evaluates the results of this work, discusses implications, and relate it to previous research and literature. Finally, discussion of limitations in the present study is given, together with opportunities for further research.

### **5.1 Evaluation of Results and Model Performance**

The testing of various machine learning models provided enormous insight into how each of them performs in dynamic pricing, hence showing the accuracy and robustness of each. The best performance was from the Gradient Boosting Regressor, with the lowest MAE of 0.2550, lowest RMSE of 0.4411, and highest  $R^2$  score of 0.8036. These findings are thus in tune with various earlier studies such as El Youbi et al. (2023), that identified the Gradient Boosting to be more robust at uncovering nonlinear associations in dynamic environments such as e-commerce. The present study extends these insights to the ride-on-demand environment



and thus shows the adaptability of the model to different variables such as ride distance and peak-hour demand.

The performance of the MLP was also quite impressive: an MAE of 0.2586, an RMSE of 0.4518, and an  $R^2$  score of 0.7940. This result supports the observations of Guo et al. (2018), where the efficiency of neural networks in high-dimensionality data was presented. However, probably the reason for the relatively lower performance compared with Gradient Boosting is the non-iterative refinement nature of MLP.

Linear Regression was a decent baseline model with MAE, RMSE, and  $R^2$  scores of 0.2628, 0.4539, and 0.7921, respectively. It had modelled the linear trend in data-for example, the proportionality between the ride distance and its fare-which was observed by Arora et al. (2021). It could not model the complex relationships-for example, temporal demand-supply fluctuation-which was expected from any model designed for dynamic pricing.

The poorest performance was represented by Random Forest, which had the value of 0.2730 MAE, 0.4639 RMSE, with a low  $R^2$  score of 0.7828. Though very strong in the capturing of nonlinearities, lacking this sequential refinement, as identified by Nalamothu (2023), it is poorly positioned to handle dynamic patterns, especially those where temporal variations might play a key role.

The results highlight the importance of the iterative refinement mechanism intrinsic to Gradient Boosting in achieving accurate fare estimates. This ability to iteratively reduce residual errors makes the model particularly suited to dynamic pricing problems with rapidly shifting demand and supply conditions. Although hyperparameter tuning increased the model's accuracy, much future work can be done by using more variables to deal with this constraint.

## 5.2 Strengths of the Methodology

The use of the CRISP-DM framework gave a proper flow to this study by ensuring that iterations comprehensively prepared the data, modelling, and evaluation. The major strengths in the methodology are the rigorous processes involved in data preparation, including handling missing values and removing outliers, that are part of feature engineering. Similarly, the derivation of ride distance and cyclical encoding of temporal features ensure models effectively capture the recurring patterns in fare dynamics. These steps thereby laid a very strong foundation for predictive models to be developed.

Although these are the benefits, there were limitations in terms of geographical and temporal dimensions. The data was on essentially Uber historical trips in New York City from 2009 to 2015, which might not reflect today's market or regional variations. Such limitations may be avoided in future studies through the use of a set of diversified datasets, which may add further depth in insight into dynamic pricing mechanisms.

## 5.3 Limitations and Opportunities for Improvement

The models captured many features of dynamic pricing, certain limitations arose. Passenger number had little impact on fare predictions, as would be expected from industry practices that focus on the distance and time taken rather than the number of passengers carried. However, this limits the applicability of the model in ride-sharing situations, where the passenger number could indirectly influence pricing. Besides, exogenous variables such as traffic and weather data were not considered, which again restricts the capability of the model

to adapt to the real world. These could be included further in the models for better precision and applicability.

The results show that the model cannot support high supply-demand gaps and especially during high demand periods. Further improvements can be done on the model by expanding the dataset with more recent and diversified information and adding real-time features that could make the model respond promptly to such fluctuations. Further, as much as the Gradient Boosting Regressor performed well, studying hybrid approaches might further improve the prediction performance.

## **5.4 Contribution to the Field and Future Directions**

This research contributes to the literature the potential of machine learning in optimizing dynamic pricing strategies for ride-on-demand services. It also contributes to basic insights from a theoretical and practical point of view by focusing on the relevance of a set of variables and using state-of-the-art machine learning approaches. Nevertheless, observation of fare adjustments in some extreme cases and limitations of the dataset used in this research suggest further research is needed.

Future research should be directed more toward real-time data, inclusions of exogenous variables, geographical diversification of data, and studies related to the implementation of hybrid modelling techniques. This will surely enhance adaptability and robustness in dynamic pricing models for an optimum balance between profitability and customer satisfaction.

Hence, the current study demonstrated very clearly the feasibility of using machine learning methods in developing dynamic pricing models for ride-on-demand services. While the results provide a good starting point for improvements in pricing strategies, at the same time, they highlight areas for further refinement, particularly regarding real-time variations and addressing constraints due to the scope of data.

## **6 Conclusion**

In this research following research question was addressed: “How can optimized multi-variable dynamic pricing strategies be developed using machine learning for the ride-on-demand industry, where demand and supply fluctuate rapidly due to factors like time of day, ride distance, passenger count, and competition?” This research was based on historical ride data analysis for the identification of major features that influence dynamic pricing and developing predictive models. The research followed the CRISP-DM approach, involving data preparation and feature engineering, comparing machine learning models. Among them all, Gradient Boosting Regressor worked out to be the best one that returned an MAE of 0.2550 with a good  $R^2$  value of 0.8036. The key takeaways include the fact that ride distance remains the paramount determinant, the negligible role played by the number of passengers, and time variations with surcharging at late-night hours. Such results expose the potential of machine learning in developing more advanced pricing methodologies and give useful suggestions to ride-on-demand industry.

While the research was efficient in achieving its objectives, certain limitations remain that could be the start for further investigation. The dataset, which falls within historic Uber

rides in New York City, prohibits generalization to different locations or today's current market position. Further research should incorporate current traffic and other weather conditions into the model for improved adaptability and accuracy in predictions. Also, enhancing the models to handle demand spikes at peak times and exploring hybrid approaches might significantly improve their performance. The economics implications for this study are also significant; this could give a great opportunity for ride-on-demand companies to consider adaptive price strategies which will maximize profitability without sacrificing customer satisfaction. Such strategies may create scope for the development of much more efficient and responsive dynamic pricing systems and enhance competitiveness within the industry.

## References

Arora, K., Kaur, S., and Sharma, V., 2020. Prediction of Dynamic Price of Ride-On-Demand Services using Linear Regression. *International Journal of Computer Applications and Information Technology*, 13(1), pp.376-389.

Ashlagi, I., et al. (2018). Pricing in ride-hailing services: Matching and dispatch optimization. *Uber Research*.

Banerjee, S., Riquelme, C. and Johari, R. (2015) 'Pricing in Ride-Share Platforms: A Queueing-Theoretic Approach', SSRN Electronic Journal [Preprint]. Available at: <https://doi.org/10.2139/ssrn.2568258>.

Battifarano, M. and Qian, Z. (2019) 'Predicting real-time surge pricing of ride-sourcing companies', *Transportation Research Part C*, 107, pp. 444–462. Available at: <https://doi.org/10.1016/j.trc.2019.08.019>.

Chen, M.K. and Sheldon, M. (2015) 'Dynamic Pricing in a Labor Market: surge Pricing and Flexible Work on the Uber Platform', *UCLA Anderson School of Management and University of Chicago*. Available at: [https://www.anderson.ucla.edu/faculty\\_pages/keith.chen/papers/uberpricing.pdf](https://www.anderson.ucla.edu/faculty_pages/keith.chen/papers/uberpricing.pdf)

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp.1247-1250. Available at: <https://doi.org/10.5194/gmd-7-1247-2014>.

Chen, X., Zheng, H., Ke, J., and Yang, H., 2020. Dynamic optimization strategies for on-demand ride services platform: Surge pricing, commission rate, and incentives. *Transportation Research Part B: Methodological*, 138, pp.23-45. Available at: <https://doi.org/10.1016/j.trb.2020.05.006>.

Chicco, D., Warrens, M.J., and Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623. Available at: <https://doi.org/10.7717/peerj-cs.623>.

El Youbi, R., Messaoudi, F. and Loukili, M. (2023) 'Machine Learning-driven Dynamic Pricing Strategies in E-Commerce', 14th International Conference on Information and

Communication Systems (ICICS). IEEE, pp. 1–10. Available at: <https://doi.org/10.1109/ICICS60529.2023.10330541>.

Faghih, S., Shah, A., Wang, Z., Safikhani, A., and Kamga, C., 2020. Taxi and Mobility: Modeling Taxi Demand Using ARMA and Linear Regression. *Procedia Computer Science*, 177, pp.186-195. Available at: <https://doi.org/10.1016/j.procs.2020.10.027>.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.

GeeksforGeeks. (2023). ML - Linear Regression. Retrieved from <https://www.geeksforgeeks.org/ml-linear-regression/>.

Guo, S., Chen, C., Wang, J., Liu, Y., Xu, K., and Chiu, D.M., 2018. Dynamic Price Prediction in Ride-on- demand Service with Multi-source Urban Data. In: *Proceedings of the 15th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '18)*. ACM, New York, NY, USA, pp.1-10. Available at: <https://doi.org/10.1145/3286978.3286992>.

Guo, S., Liu, Y., Xu, K., and Chiu, D.M., 2017. Understanding Ride-on-demand Service: Demand and Dynamic Pricing. In: *Proceedings of the First International Workshop on Pervasive Smart Living Spaces (PerLS)*. IEEE, pp.1-8. Available at: <https://doi.org/10.1109/PerLS.2017.7946784>.

Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14), pp.5481-5487. Available at: <https://doi.org/10.5194/gmd-15- 5481-2022>.

Luo, Q., and Saigal, R., 2017. Dynamic Pricing for On-Demand Ride-Sharing: A Continuous Approach. *SSRN Electronic Journal*. Available at: <https://ssrn.com/abstract=3056498>.

McGuire, K. (2015) *Hotel Pricing in a Social World*. 1<sup>st</sup> edn Wiley. Available at: <https://www.perlego.com/book/997049/hotel-pricing-in-a-social-world-driving-value-in-the-digital-economy-pdf>.

Nalamothu, P.P., 2023. Comparative Analysis of Regression Models for Price Prediction of Ride-On- Demand Services. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 11(V), pp.1687-1700. Available at: <https://doi.org/10.22214/ijraset.2023.51770>.

Rathore, B. et al. (2024) 'Predicting the price of taxicabs using Artificial Intelligence: A hybrid approach based on clustering and ordinal regression models', *Transportation Research Part E: Logistics and Transportation Review*, 185, p. 103530. Available at: <https://doi.org/10.1016/j.tre.2024.103530>.

Saadi, I., Wong, M., Farooq, B., Teller, J. and Cools, M. (2022) 'An investigation into machine learning approaches for forecasting spatio-temporal demand in ride-hailing services', *Transportation Research Part C*. Available at: <https://arxiv.org/abs/1703.02433>.

Saltz, J., 2021. CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. In: 2021 IEEE International Conference on Big Data (Big Data), 2021. IEEE, pp.2337-2344. Available at: <https://doi.org/10.1109/BigData52589.2021.9671634>.

SAS Institute. (2008). SEMMA Methodology. Retrieved from [<https://support.sas.com/>].

Sun, Z., Xu, Q., and Shi, B., 2020. Dynamic Pricing of Ride-Hailing Platforms considering Service Quality and Supply Capacity under Demand Fluctuation. Complexity, 2020, Article ID 836434. Available at: <https://doi.org/10.1155/2020/836434>.

Thorat, P., 2023. Mastering Gradient Boosting: A Machine Learning Guide. Available at: <https://www.linkedin.com/pulse/mastering-gradient-boosting-machine-learning-guide-pratik-thorat/>.

Towards Data Science, 2023. Multi-Layer Perceptrons. Available at: <https://towardsdatascience.com/multi-layer-perceptrons-8d76972afa2b>.

Yamuna, G., Dhinakaran, P., Vijai, C., Kingsly, J., Raynukaazhakarsamy, R. and Devi, S.R. (2024) 'Machine Learning-Based Price Optimization for Dynamic Pricing on Online Retail', 9th International Conference on Science Technology Engineering and Mathematics (ICONSTEM). IEEE. DOI: 10.1109/ICONSTEM60960.2024.10568763.

Yan, C., Zhu, H., Korolko, N., & Woodard, D. (2020). Dynamic pricing and matching in ride-hailing platforms. Naval Research Logistics, 67(8), 705-724.

Willmott, C.J. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, pp.79-82. Available at: <https://doi.org/10.3354/cr030079>.