Optimizing Taxi Service Operations Using Machine Learning and Big Data Techniques

Shamsa Halima Kasozi Nantale
x23344083
Data Intensive Scalable System
National College of Ireland
https://youtu.be/phh73X72Wzc

Abstract

The aim of this project is to predict taxi fares, high pick up demand locations and trip durations using machine learning and big data techniques in large taxi trip record. Apache Spark is used for distributed data processing on the dataset, which consists of about 400,000 rows, and the results are stored and retrieved using Azure Blob Storage and MySQL. Both the fare prediction and high demand pickup location identification are done by using a Gradient Boosting Regressor (GBT) model. By analysing the factors such as trip distance, pickup hour and passenger count, the analysis reveals how fare prediction depends on these factors, and also the high traffic zones for optimized fleet management. This methodology not only demonstrates the feasibility of applying the big data analytics in the urban mobility but also proves that the integration of the advanced machine learning techniques and the distributed data processing could improve the operational efficiency and the service quality in taxi services.

***Keywords*— Big Data Analytics, Machine Learning, Apache Spark, Fare Prediction, Urban Mobility Optimization**

1 Introduction

The objective of this study is to improve the urban taxi service using big data analytics and machine learning models in fare prediction, high demand location identification and trip duration analysis. The study involves 400,000 rows of data related to taxi trip, which is processed using Apache Spark and the store for efficient data management is Azure Blob Storage. The idea is to use real time features, such as pickup hour, trip distance and passenger count, to build a scalable solution for the dynamic pricing, operation optimisation and efficiency in the taxi industry.

Research Question.

- How can machine learning models predict taxi fares based on features like trip distance and passenger count?

- What factors influence high-demand pickup locations, and how can they be identified?
- How can trip duration and fare be correlated with trip distance to optimize pricing?
- How can big data processing tools like Azure Blob Storage and Apache Spark handle large datasets for real-time taxi data analysis?

This study shows how big data processing and machine learning can optimize taxi services by enabling real-time fare prediction, improving trip duration estimates, and identifying high-demand pickup locations. This allows integration of large datasets with Azure Blob Storage and Apache Spark for efficient handling of large datasets and a scalable solution for smart city transportation systems. In contrast to the work done so far, this approach overcomes limitations by using real time data features and is scalable for predicting dynamic and accurate features.

2 Related Work

Recent studies about exploiting machine learning and big data methods to improve the performance of urban transportation systems are now addressing taxi and ride-hailing services. For example, (Munawar & Piantanakulchai 2025) mentioned how machine learning can be used to predict passenger demand in autonomous taxi transportation systems in smart cities by highlighting the role of predictive models in optimizing the operation. (Pakdel et al. 2025) also investigated the use of Long Short Term Memory (LSTM) networks to predict customer demand, adding customer information to improve prediction accuracy. These are the studies that emphasize the increase in the demand prediction interest and the inclusion of deep learning and real time data in it to improve the decision making.

Taxi demand prediction models were examined and contrasted by (Saputra & Sihabuddin 2024), who evaluated different machine learning approaches for demand forecasting to enhance urban mobility optimization. (Rhouas & Hami 2025) also studied New York taxi data using Ordinary Least Squares (OLS) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithms to predict revenue from the traditional optimization techniques applied to big data problems.

Although these studies provide useful knowledge, they usually focus on a restricted or particular facet of urban mobility and predict fares. On the other hand, this project builds on the previous studies by combining both real time elements and state of the art machine learning techniques (such as Gradient Boosting Regressor) to predict taxi fares and trip demand from wide variety of factors including pickup location, trip distance, time of day and number of passengers. Spatial features such as pickup and drop-off locations can typically be well incorporated in existing models if a large enough dataset is available, primarily due to scalability and efficiency limitations.Similar to this approach, (Haery et al. 2024) demonstrated the effectiveness of combining GIS with machine learning techniques to analyze Uber data in New York City, finding that

environmental factors significantly influence travel demand prediction accuracy.

This study uses Apache Spark for distributed data processing and the Azure Blob Storage for data storage and management in efficient manner to deal with the large datasets and performing real time predictions, breaking the scalability and feature integration barriers as reported in other studies. In this way, the model can affordably and precisely estimate fares in high demand areas and during peak hours, thus providing a broader and time real solution in improving the taxi service efficiency.

3 Methodology

This project used different technologies to process and analyze a large set of taxi trip records. This can be divided into four steps of the methodology: data storage, data processing, model training, and result storage. To implement these steps, Azure Blob Storage, MySQL and Apache Spark (PySpark) were chosen as they are able to efficiently handle large-scale data processing. Each of the methodology's components is broken down below:

About dataset:

The dataset consists of about 400,000 rows of taxi trip records with each row containing the pickup and dropoff times, trip distance, fare amount, tip amount, payment type, and location identifiers. Since the volumes of data are so large, this data serves as an ideal source to study trip patterns, pricing models, and location based demand to understand and optimize taxi service operation. Using these datasets, trip distance can be encoded and it leads to valuable insight on the effect of trip distance on fare prediction, trip distance distribution across pickup locations and trip duration with respect to trip distance, thus providing new fleet and dynamic fare strategies.

3.1 Data Storage with Azure Blob Storage

Azure Blob Storage is a scalable cloud storage solution for storing large amounts of structured and unstructured data and it was used to store the dataset. The reason for choosing Azure Blob Storage as it is secure, cost effective, and can scale up as needed. A CSV file of the taxi trip dataset was uploaded into a container in Azure Blob Storage, making it easy to access and process further.

Using Azure Storage Blob SDK for Python, a connection was established to the Azure Blob Storage service, and the dataset was downloaded to a local system for analysis. Through this process the data was securely retrieved and was ready for processing on a distributed computing platform.

3.2 Data Processing with Apache Spark (PySpark)

Apache Spark was used for processing the data, and specifically to handle large datasets in a distributed manner PySpark library was used. The reason to choose spark is that it has ability to perform data transformations and machine

learning tasks in parallel, and is therefore an perfect choice for processing big data.

The processing of dataset was using the following steps:

- Data Loading: The dataset was loaded into a Spark DataFrame using the spark.read.csv() function.
- Data Cleaning and Transformation: Various data preprocessing tasks like casting column to proper data types, checking for null values and handling outliers using Interquartile Range (IQR) method were done. Even with a large dataset, capabilities of PySpark's parallel processing made the tasks very efficient.
- Feature Engineering: The pickup hour was derived from the pickup timestamp to be used as a feature. This was an important feature for predicting fare amounts based on time of day.
- Data Aggregation: The data was aggregated by using the information about pickup location ID and the number of trips for each location. This helped to determine high demand.
- Model Training: Using features like passenger count, trip distance and pickup hour, the fare prediction was done with the use of Gradient Boosting Regressor (GBT) model. Root Mean Square Error (RMSE), and R squared (R2) were used to evaluate the performance of the model.

3.3 Data Storage and Retrieval with MySQL

The results were stored in a **MySQL** relational database for future querying and reporting. Once the data was processed and analyzed **MySQL** was then selected because of its robustness in handling structured data and providing fast queries. The results were stored in separate tables, including:

- Fare prediction results
- High-demand pickup locations
- Trip summaries (such as average trip duration and fare)

SQL queries were used to insert the processed data into the appropriate tables. Functions were created to fetch data from MySQL and export them to **CSV** format for further analysis or reporting in future. This facilitated smooth transition from data processing to result storage and retrieval.

3.4 Automation and Scalability

Automation was used to ensure that the entire process was smooth. To automate the steps of downloading the data, processing it, training the machine learning model and storing the results in MySQL, I created python scripts. These scripts were run within an Apache Spark environment, a natural place to run such scripts as it can handle large datasets in parallel. This approach was

scalable and efficient for processing large amounts of data in real time or batch processing.

Tools and Technologies

- **Azure Blob Storage**: Used for secure and scalable cloud storage of the dataset.
- **Apache Spark (PySpark)**: Utilized for distributed data processing and machine learning tasks.
- **MySQL:** Employed to store processed results and facilitate querying for reporting and further analysis.
- **Python:** The programming language used to integrate the workflow, from data retrieval to processing, model training, and result storage.

The combination of Azure Blob Storage, PySpark and MySQl was able to process large datasets efficiently. This methodology allowed to go deep in the dataset, fare prediction, high demand pickup locations and trip duration analysis. This method is very effective for managing and processing big data in transportation industry because of the combination of Apache Spark's scalability and MySQL's flexibility for storing results.

4 Results and Discussion:

Analysis 1: Fare Prediction Using Gradient Boosting Regressor

Fare prediction plays a very important role in this taxi service industry as it helps both the businesses and the passengers to know how much they are supposed to pay for a ride. The aim of this analysis was to predict the fare amount on some factors such as trip distance, passenger count and other attributes which are typically available during a ride. There are subsets of features that we can model using GBT (Gradient Boosting Regressor) to give us the relationship between those features and fare amount. Since we are dealing with big data, this analysis is useful to make accurate and scalable predictions on a large dataset of taxi trips. The predictive model can be used to learn to automatically tune prices in real time, optimize routes, improve customer satisfaction through better fare estimates before or during a ride.

Gradient Boosting Regressor (GBT) was chosen because of its able to handle complex relationships in the data. Using historical data from a large dataset with trip information like trip distance, passenger count, and total amount, as well as the pickup hour extracted from the timestamp, the model was trained. The results of the model had a Root Mean Squared Error (RMSE) of 6.4894101 and a R-squared (R2) value of 0.90817975, which indicates that the model fits the data very well. The R2 value is, indeed, noteworthy because it indicates that about 90.8% of the variance in the fare amount can be explained by the

model, which indicates a high level of accuracy. Also, the RMSE value, which is the model error in units of fare amount, is low, so the predictions are not far from the actual values.

This analysis is important for big data processing since large amount of data such as taxi trip records can be processed efficiently using distributed computing platforms like Apache Spark. By using the big data tool power the analysis can be scaled to manage large volumes of data and be run to predict fares on millions of trips in real time or batch processes. In practice, such models are needed by companies like Uber, Lyft or local taxi services to accurately and real time predict fares in order to optimize the operations and improve the customer experience. Additionally, integrating this analysis with additional data sources, like traffic or weather, can further enhance this with more dynamic fare estimates on the status of the current conditions.

Analysis 2: High-Demand Pickup Locations

The aim of this analysis was to understand the pickup location demand, i.e. how often taxis are called for trips at different locations. By using the pickup location IDs (PULocationID) to analyze the traffic patterns of different locations in the city by counting the number of trips for each, valuable insights into the traffic patterns of various locations in the city were gained

The trip frequency across a large array of pickup locations is presented in the first plot, labelled "Number of Trips by Pickup Location." This indicates that most pickup locations feature small traffic counts, however, there are patterns with significant deviations of trip counts to suggest that these locations are high demand zones. Since these locations aren't the typical action locations, these locations are possibly in places that have more foot traffic and are more urban like locations airports, train stations and commercial hubs. The large spikes in the plot reveal these popular areas.

The second plot, "Top 10 Pickup Locations with Highest Number of Trips", focuses on top 10 most popular pickup locations by the number of trips. The bar plot also shows a few locations with very high trip counts, which means they are high demand pickup points. The findings can be very useful for taxi fleet management optimization. This allows taxi companies to identify these areas, in turn, to make sure that vehicles are distributed to meet demand and to reduce passenger waiting times. Pricing strategies could also be altered to match up to the increased demand of these areas, leading to higher charges on peak hours.

This analysis shows that big datasets can be processed using Apache Spark tools to identify trends in high demand zones and that overall this is how the data driven insights can be meaningfully applied. This analysis is important to know how to spend resource in increasing the efficiency of the service and how to make the passenger experience better.

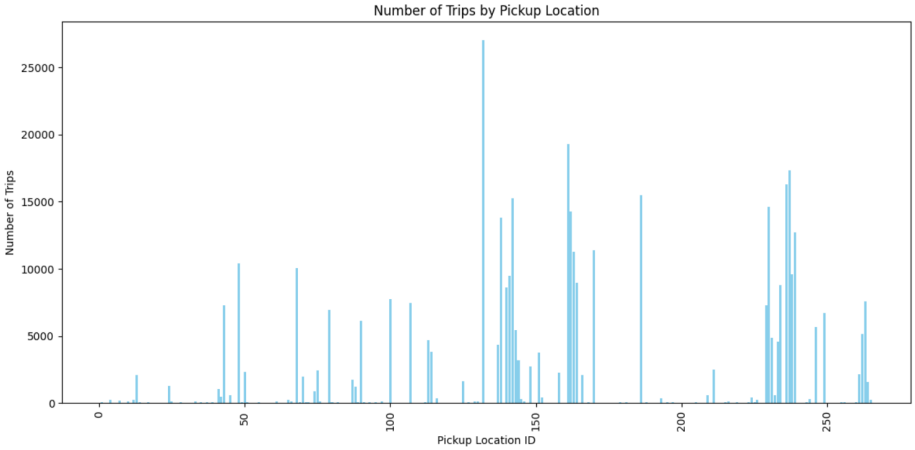

Figure 1: Number of Trips by Pickup Location

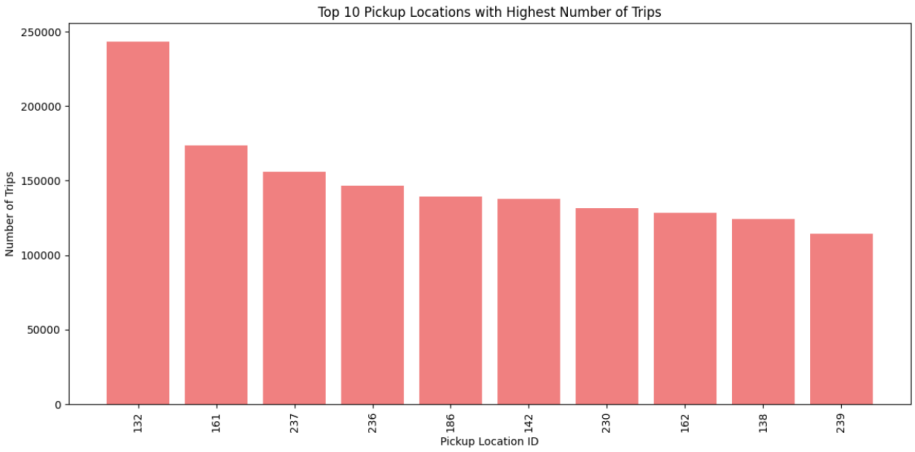

Figure 2: Top 10 Pickup Locations with Highest Number of Trips

Analysis 3: Trip Duration and Fare Analysis

This analysis was done in order to examine the relationship between trip distance and trip duration and fare amount. Optimizing taxi services, fare estimation, and a better customer experience depends on these relationships. The duration of the trip and the amount of the ticket can be investigated in terms of the distance traveled to see how the time of travel and the distance traveled in the trip affect pricing and service efficiency.

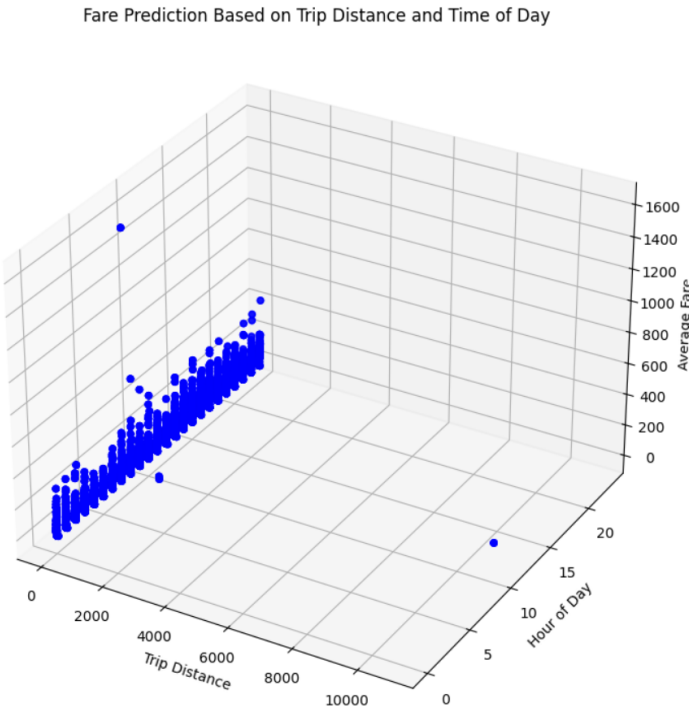

Figure 3: Fare Prediction Based on Trip Distance and Time of Day

The 3D scatter plot shows the relationship between trip distance, pickup hour (time of day) and the average fare. In addition, it allows us to see how the fare changes with trip distance and time of day. It is expected that the average fare increases as trip distance increases, and the plot confirms that. Additionally, it indicates that some hours of the day (i.e., the one with the higher values of Axis Hour of Day) may have outliers in terms of fare and for reasons that are different from the trip distance alone. This visualization is in line with the idea that trip distance and time of day can have a big impact on the fare prediction.

The scatter plot shows the relationship between trip distance and trip duration. The figure clearly shows that the average trip duration increases with the trip distance for most trips. This is logical, longer trips would need more time

to complete. however the plot shows a few outliers with very long duration for short distances that may indicate something strange such as traffic delays or detours. Outliers provide valuable information for understanding possible disruptions on trips.

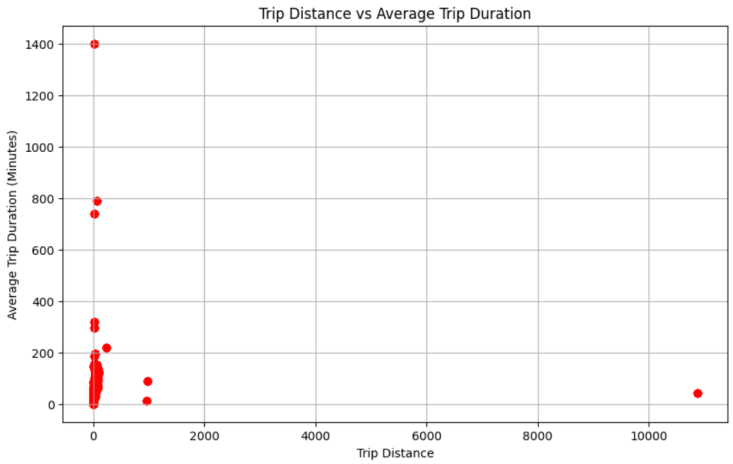

Figure 4: Trip Distance vs Average Trip Duration

5 Conclusion:

The combination of big data technologies, taxi service operations, and the use of machine learning models showed the effectiveness of improving an operation. For the task of fare prediction, it uses Gradient Boosting Regressor and studies high demand pickup locations that can help taxi fleet management be more effective and increase customer satisfaction. The solution is scalable due to its use of Apache Spark for distributed data processing and Azure Blob Storage for efficient data management, which allows for real-time processing of large datasets.

Additionally, integration with spatial features (pickup and dropoff locations), real time variables (pickup hour, trip distance, number of passengers) makes it possible to predict more accurately and dynamically the fares. By adopting this approach, the limitations of existing studies are overcome by providing a more comprehensive and scalable solution to real time taxi service optimization, that is, very pertinent to smart cities and dynamic pricing systems.

References

Haery, S., Mahpour, A. & Vafaeinejad, A. (2024), 'Forecasting urban travel demand with geo-ai: a combination of gis and machine learning techniques utilizing uber data in new york city', *Environmental Earth Sciences* **83**(20), 594.
URL: *https://doi.org/10.1007/s12665-024-11900-y*

Munawar, A. & Piantanakulchai, M. (2025), 'Machine learning-driven passenger demand forecasting for autonomous taxi transportation systems in smart cities', *Expert Systems* **42**(3), e70014.
URL: *https://doi.org/10.1111/exsy.70014*

Pakdel, G. H., He, Y. & Chen, X. (2025), 'Predicting customer demand with deep learning: an lstm-based approach incorporating customer information', *International Journal of Production Research* **0**(0), 1–13.
URL: *https://doi.org/10.1080/00207543.2025.2468885*

Rhouas, S. & Hami, N. E. (2025), 'Analysis of big data from new york taxi trip 2023: revenue prediction using ordinary least squares solution and limited-memory broyden-fletcher-goldfarb-shanno algorithms', *International Journal of Electrical & Computer Engineering* **15**(1), 711–718.
URL: *https://doi.org/10.11591/ijece.v15i1.pp711-718*

Saputra, R. & Sihabuddin, A. (2024), 'Optimizing urban mobility: A comparative analysis of taxi demand prediction models', *Ingénierie des Systèmes d'Information* **29**(5).
URL: *https://doi.org/10.18280/isi.290522*