Taxi Demand Prediction and Surge Optimization

Table of Contents

Introduction

The New York City taxi industry is a cornerstone of the city's transportation system, delivering millions of rides each year. However, the rise of app-based ride-hailing services like Uber and Lyft has created formidable challenges for yellow taxis, including a significant loss of market share, inefficient fleet deployment, and outdated pricing models.

This project decisively addresses these issues by employing data science techniques to predict taxi demand across NYC and implement dynamic pricing strategies. By rigorously analyzing historical trip data along with external factors such as weather, time of day, and holidays, I provide actionable insights that will:

- Optimize fleet deployment by accurately anticipating demand patterns.
- Introduce dynamic pricing models that align fares with real-time demand, thereby enhancing driver profitability.

This report comprehensively outlines the project’s key components, including data preparation, exploratory data analysis (EDA), modeling, evaluation metrics, and a thorough discussion of the results. Our findings present concrete solutions to boost the efficiency of the NYC taxi system and significantly improve its competitiveness in today’s market landscape.

Business Understanding

Motivation of the Project

The yellow taxi industry has been an iconic and integral part of New York City's transportation network for over a century. These taxis provide reliable transportation to millions of residents, commuters, and tourists daily, contributing to the city's economy and serving as a critical mode of urban mobility. However, over the past decade, the emergence of ride-hailing platforms like Uber, Lyft, and other app-based services has disrupted the taxi industry, posing challenges that require urgent attention.

Challenges Facing Yellow Taxi Cabs

1. Decline in Market Share
   In 2013, NYC's yellow taxis averaged 500,000 daily trips. By 2023, this number has dropped to ~200,000 trips per day, largely due to the convenience and dynamic pricing offered by ride-hailing services [1]. Ride-hailing apps dominate 80% of the for-hire vehicle (FHV) market share, leaving yellow taxis struggling to compete[2].

2. Static Pricing Models
   While Uber and Lyft utilize surge pricing algorithms that adjust fares based on real-time demand and supply, yellow taxis still rely on a fixed metered fare system that fails to adapt dynamically to changing market conditions. This leads to revenue losses during peak hours when demand is high but fares remain unchanged.

3. Uneven Fleet Distribution
   Taxis naturally cluster in high-demand areas, leaving other parts of the city underserved. This not only causes long wait times for customers but also results in idle fleet time and wasted fuel for drivers.

4. Operational Inefficiencies

Without demand prediction models, taxi operators lack actionable insights to anticipate passenger needs. As a result, drivers spend significant time roaming empty in search of passengers, increasing fuel costs and reducing profitability.

These challenges highlight the urgent need for modern, data-driven strategies that enable the yellow taxi industry to remain competitive. By leveraging historical trip data and external factors such as weather, holidays, and time of day, this project aims to:

1. Accurately predict taxi demand across NYC to optimize fleet deployment.
2. Implement dynamic pricing models to adjust fares based on real-time demand.

Relevance of the Project

New York City provides an ideal testbed for this project due to its size, complexity, and unique urban characteristics:

1. High Demand and Trip Volumes

   NYC taxis still facilitate ~200,000 trips daily, with the industry generating over $5 billion annually[3] . The volume of passengers creates a rich dataset for demand prediction, with factors like time of day, weather, and events influencing ride patterns.

2. Dynamic Urban Environment

   NYC operates as a 24/7 city with a highly diverse population, thriving tourism, and frequent special events, such as concerts, parades, and sports games. These factors introduce unique demand variability that can be effectively modeled. Demand patterns also differ across various areas of the city, with certain locations receiving less taxi coverage, representing untapped opportunities for operators.

3. Competitive Pressure from Ride-Hailing Platforms

   Uber and Lyft success stems from their use of machine learning and predictive analytics to optimize fleet management, implement surge pricing, and reduce customer wait times. For yellow taxis to compete, adopting predictive models is no longer optional—it is necessary to survive in a competitive transportation market.

4. Economic and Social Impact

The yellow taxi industry provides employment for thousands of drivers and contributes to NYC's local economy. Improving fleet efficiency and profitability will ensure the sustainability of the industry while enhancing transportation accessibility for underserved communities.

Project Objectives

Project aims to address the above challenges through two primary objectives:

1. Taxi Demand Prediction

Use historical ride data combined with external factors such as time of day, weather conditions, and special events to predict passenger demand across NYC. Anticipate high-demand zones during peak hours and identify emerging hotspots or underserved regions. Optimize taxi deployment to reduce idle fleet time and ensure drivers are available where passengers need them the most.

2. Dynamic Pricing Optimization

Implement machine learning models to recommend dynamic fare adjustments based on real-time demand and supply conditions.Introduce pricing strategies that are fair to customers while ensuring profitability for drivers during peak and off-peak periods.

Benefits of the Project

1. For Taxi Drivers and Operators
    - Higher Revenues: Accurate demand prediction ensures taxis are deployed in areas with the highest passenger volume.
    - Reduced Idle Time: Drivers can minimize empty trips, saving fuel and reducing operating costs.
    - Dynamic Earnings: Surge pricing during peak demand allows drivers to earn fair compensation while incentivizing availability.
2. For Customers
    - Reduced Wait Times: Taxis will be better distributed, ensuring faster pickup times.
    - Competitive and Transparent Pricing: Dynamic fares remain competitive with ride-hailing apps while being fair during low-demand hours.
    - Improved Coverage: Identification of underserved regions ensures equitable access to taxis across all areas of NYC.
3. For City Infrastructure and Policymakers

- Traffic Optimization: Efficient deployment reduces congestion caused by idle taxis roaming the streets.
- Urban Planning Insights: Demand patterns can inform decisions on public transportation routes, infrastructure improvements, and resource allocation.

4. For the Environment
    - Reduced idle time and fuel usage result in lower carbon emissions, promoting a more sustainable transportation system.

Why New York City?

New York City serves as a perfect case study for this project because of its:

1. Scale:The sheer volume of trips and data provides an excellent foundation for building robust demand prediction models.
2. Demand Variability:NYC’s transportation needs change dynamically due to weather, events, holidays, and demographic patterns.
3. Market Competition:With Uber and Lyft dominating the market, the success of this project can help yellow taxis remain competitive and relevant
4. .Economic Importance:The taxi industry supports thousands of livelihoods and contributes significantly to the local economy.

Data Preparation and Understanding

The data for this analysis was sourced from a comprehensive dataset of yellow taxi trips in New York City. This dataset included a variety of key attributes such as pickup and drop-off timestamps, trip distances, fare amounts, pickup and drop-off locations, and weather data like temperature and precipitation. These data points were essential to understanding demand patterns and influencing factors. The integration of external weather data provided valuable context to analyze how environmental conditions impacted taxi demand.

Data Cleaning and Preprocessing

To ensure the dataset was ready for analysis, extensive data cleaning and preprocessing were performed. Key steps included:

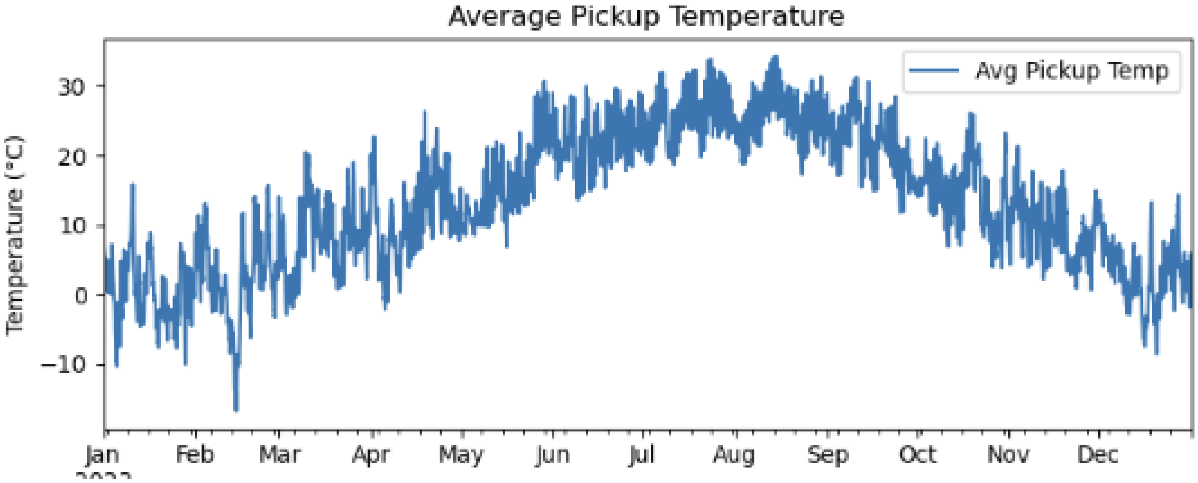
- Handling Missing Values: Missing data in columns like trip distances and timestamps were either imputed or removed, depending on their significance.
- Data Type Conversions: Columns such as `pickup_dateTime` and `dropoff_dateTime` were converted to proper datetime formats for temporal analysis.
- Outlier Removal: Unreasonable values, such as negative trip distances or excessively high fares, were identified and excluded.
- Resampling: The data was aggregated and resampled at daily and weekly frequencies to capture demand trends over time. For example, trip distances were summed to reflect total daily demand.

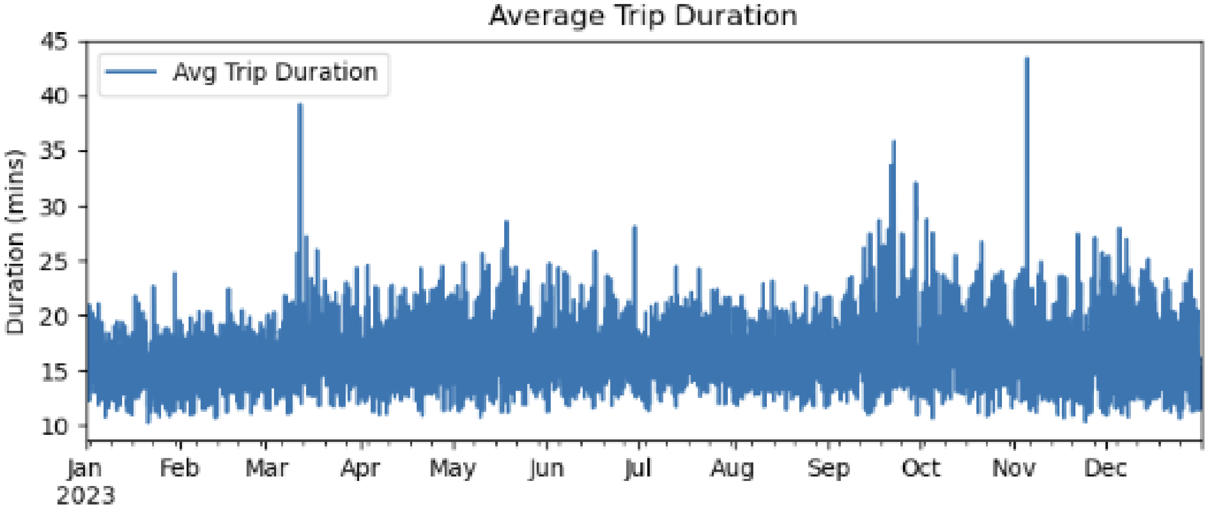
Exploratory Data Analysis (EDA)

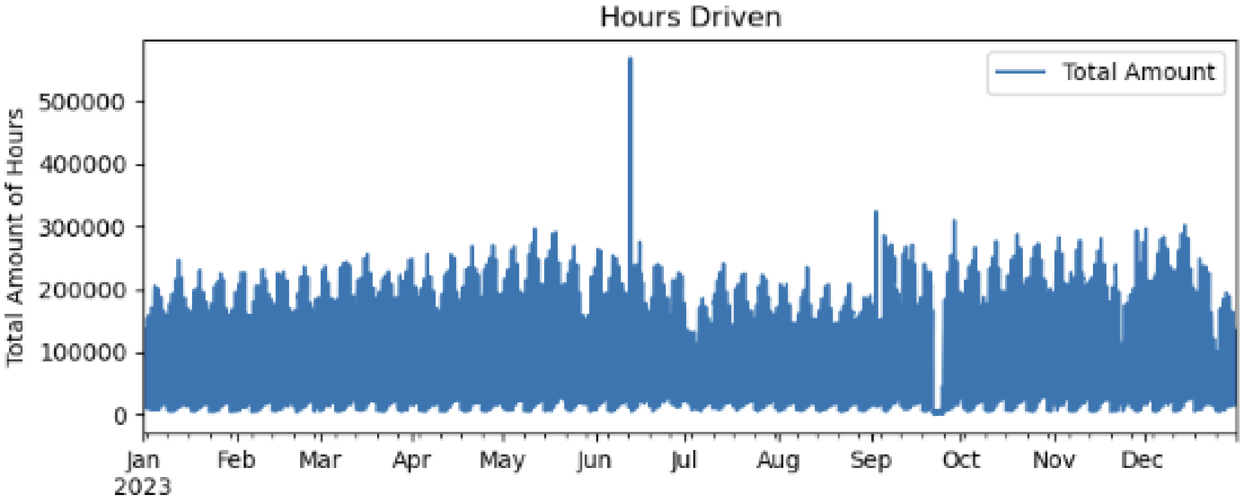
EDA played a crucial role in understanding the dataset's characteristics and uncovering meaningful insights. Key visualizations and analyses included:

- Temporal Analysis: Identifying seasonal patterns and demand fluctuations across different times of the year.
- Spatial Analysis: Examining the distribution of trips across boroughs and specific zones, highlighting hotspots of activity.
- Weekday Trends: Analyzing demand variations across days of the week to understand peak periods of taxi usage.

- External Factors: Investigating the relationship between weather conditions (temperature, rain) and demand, which provided critical insights for forecasting.







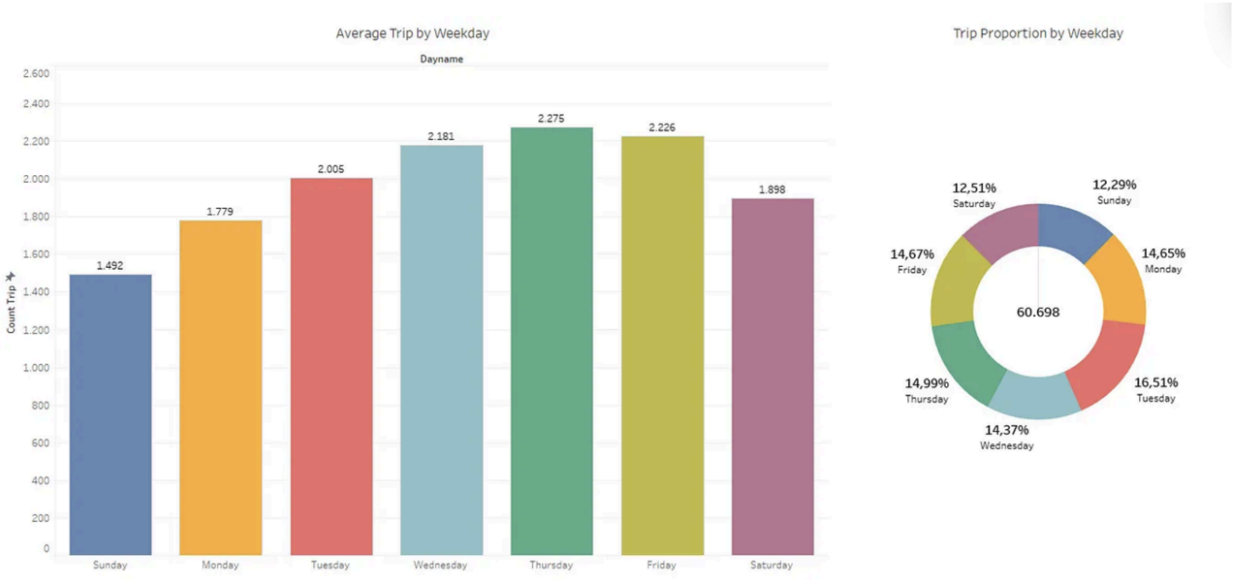
Demand Patterns and Trends

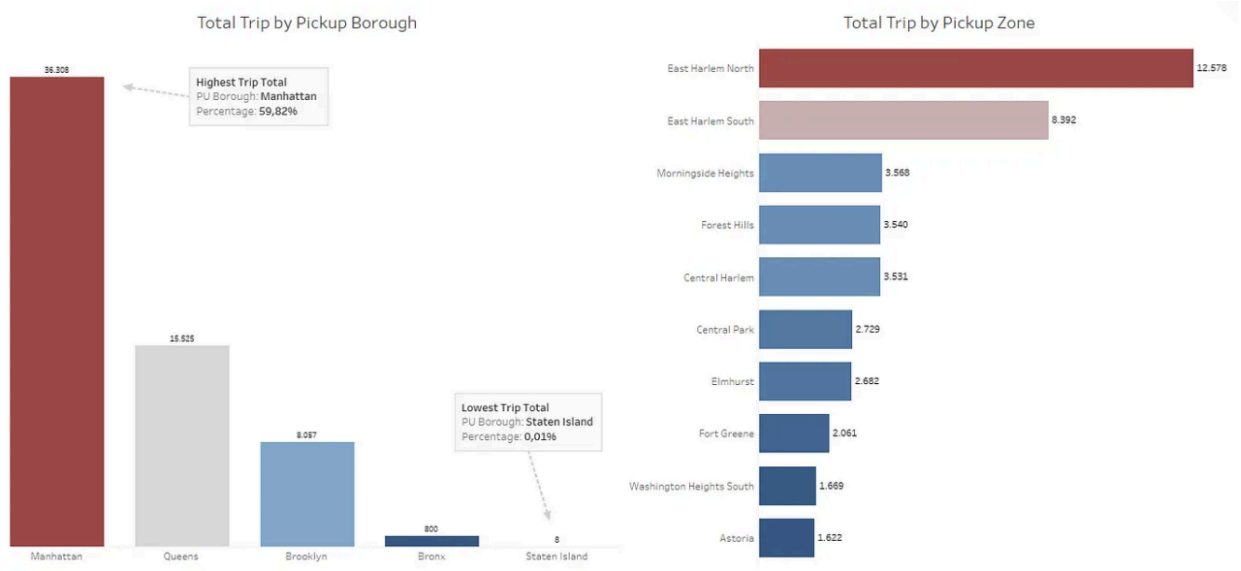
The analysis revealed consistent daily demand with significant weekday peaks, particularly on Wednesdays and Thursdays. Seasonal fluctuations were observed, with higher demand during warmer months and a slight dip in colder seasons. Spatially, Manhattan accounted for the majority of trips, followed by Queens and Brooklyn. Demand trends emphasized the need for robust forecasting models to predict short-term and long-term variations.

Feature Analysis

Critical features impacting taxi demand were identified:

- Time-Based Features: Day of the week, month, and hour of the day strongly influenced trip counts.
- Spatial Features: Pickup and drop-off locations helped identify high-demand areas and travel patterns.
- Weather Conditions: Features like temperature, precipitation, and wind speed played a notable role in trip volumes.





These steps laid a strong foundation for model development, enabling us to create demand prediction models tailored to the unique characteristics of NYC taxi data. By thoroughly understanding and preparing the data, we ensured the insights derived were accurate and actionable.

Price Prediction

The New York City Yellow Cab system is experiencing a decline in business due to the rapid growth of ride-hailing services like Uber, Lyft, and other similar applications. These modern platforms allow users to quickly access and compare fares for trips from any pickup location to various drop-off destinations within seconds. This convenience has made it increasingly difficult for Yellow Cab to compete, as traditional taxis do not provide passengers with an upfront fare estimate before the ride begins.

Our pricing model will help revitalize this struggling business by providing customers with a clear price point for their trips. This approach will encourage more customers to choose our cabs, as they are more readily available and accessible in the bustling New York City compared to other services.

For our model, we had the following given features:

- Pickup Location: The location where the passenger is picked up.
- Drop-off Location: The location where the passenger is dropped off.
- Toll Amount: Any toll fees incurred while traveling between different boroughs.
- Temperature and Precipitation at Pickup Location: Weather conditions at the pickup location.
- Temperature and Precipitation at Drop-off Location: Weather conditions at the drop-off location. Pickup Time and Date: The date and time when the passenger is picked up.
- Drop-off Time and Date: The date and time when the passenger is dropped off.

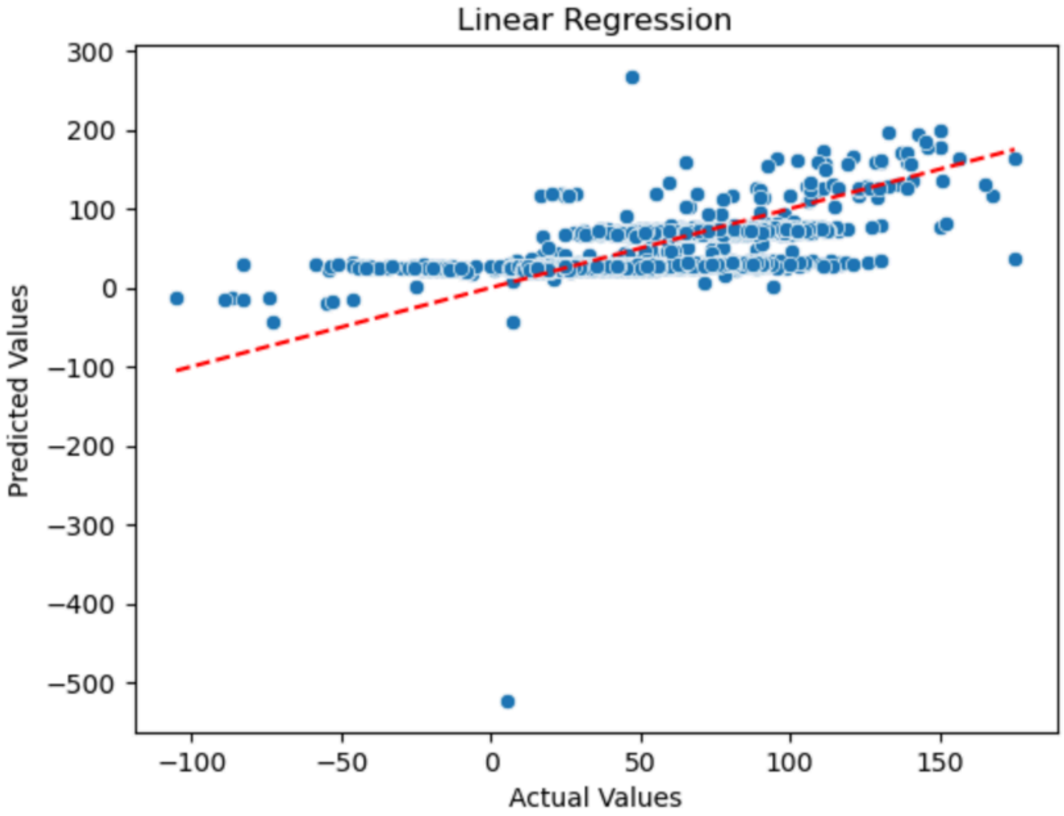
From these features, we feature engineered the following new features:

1. Trip Duration: We calculated the trip duration by subtracting the pickup time from the drop-off time, using timestamps rounded to the nearest minute.
2. Speed of the Trip: Using the distance and time, we determined the average speed of the vehicle for the trip, expressed in miles per hour.
3. Fare per Mile: We calculated the average fare per mile by dividing the total trip amount by the distance travelled.
4. Pickup and Drop-off Times: To simplify our analysis, we rounded the pickup and drop-off times to the nearest hour.
5. Holiday Feature: We introduced a binary feature called 'is holiday' that indicates whether the pickup occurred on a holiday, which includes weekends.

We did not create a separate weekend feature, as it would be redundant with the 'is holiday' feature.

Linear Regression:

We started by using basic linear regression, including all the features to predict the price of a given trip. With approximately 30 million records for just one year, it became clear that we might encounter overfitting issues. That is, the model will capture the noise and irrelevant details instead of the underlying patterns in the data, potentially leading to poor performance on new, unseen data.

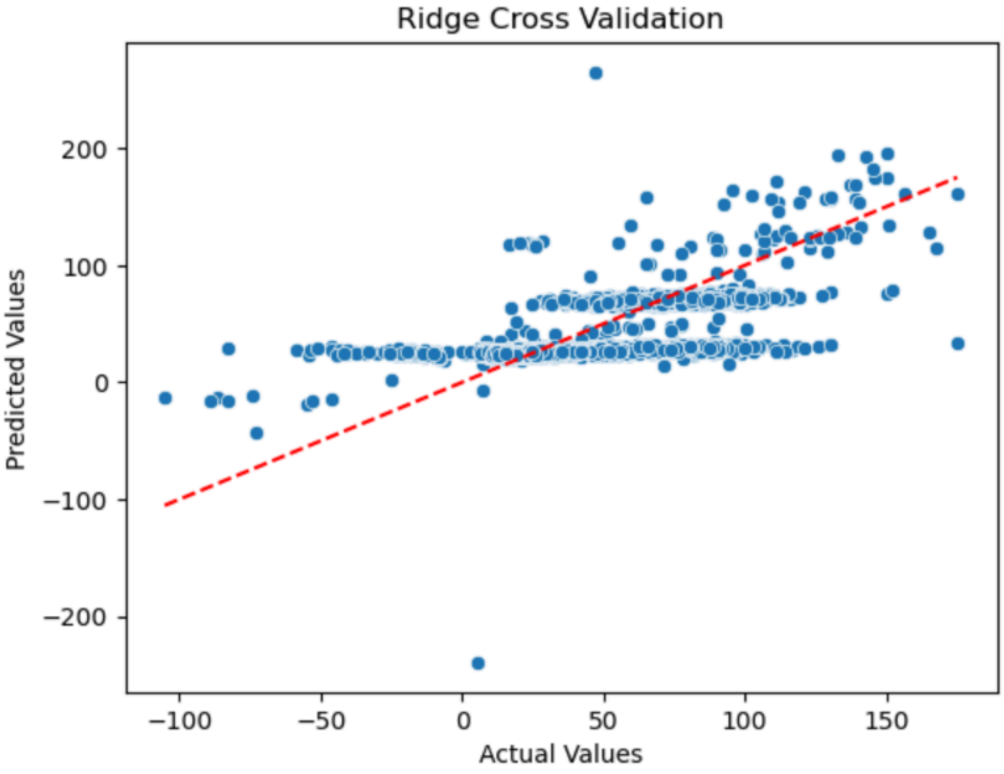


The Mean Absolute Error is 10.1289 and the Root Mean Square Error is 15.6065

If the cab fare differs by approximately $15 from the actual fare, the pricing model is not functioning correctly. Hence, we needed something more reliable.

Ridge Cross Validation:

Next, we developed another Linear Regression model, this time incorporating the Ridge Penalty. This technique helps prevent overfitting by adding a "penalty" to the model's coefficients, which discourages them from becoming too large. Essentially, the ridge penalty strikes a balance between fitting the data accurately and maintaining a simple, robust model.

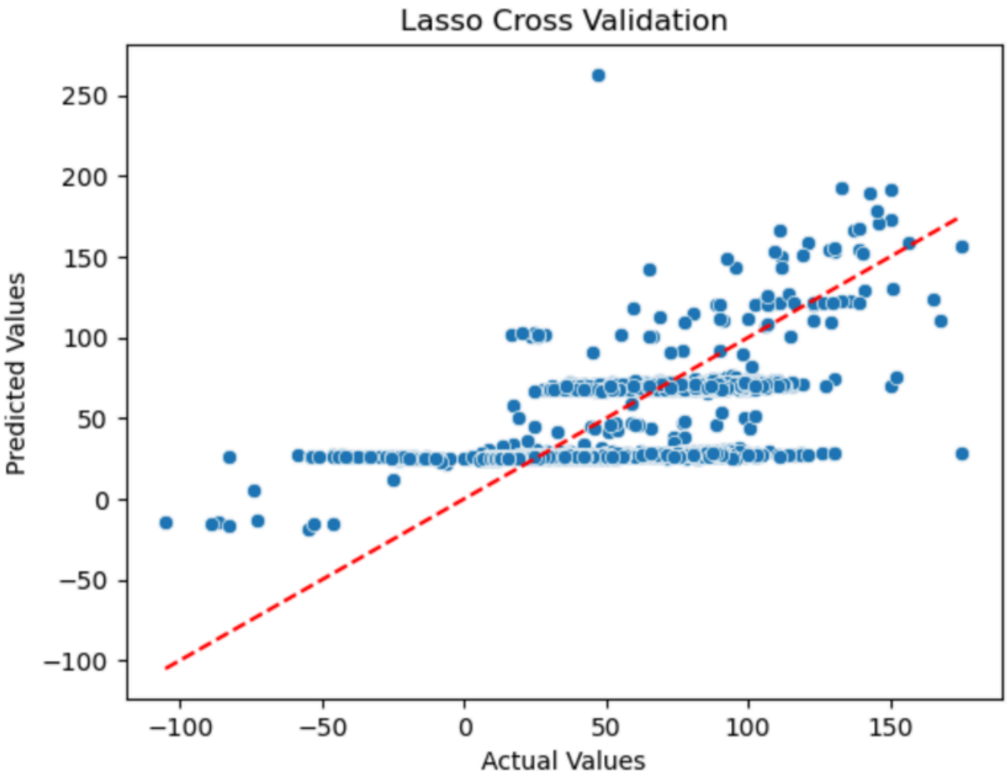


The Mean Absolute Error is 10.2822 and the Root Mean Square Error is 15.2813

In comparing Simple Linear Regression (SLR) to SLR with a Ridge Penalty, we observed that the RMSE for SLR is higher than that for the Ridge Penalty model. Conversely, the Mean Absolute Error (MAE) shows the opposite trend. These findings suggest that SLR is more sensitive to outliers, whereas incorporating the Ridge Penalty enhances the model's consistency.

Lasso Cross Validation:

The next step was to incorporate Lasso Validation. Ridge regression shrinks coefficients towards zero but seldom sets them exactly to zero. Lasso, on the other hand, can shrink some coefficients to exactly zero thus effectively performing feature selection.



The Mean Absolute Error is 10.5815 and the Root Mean Square Error is 15.4593

We used Lasso Cross-Validation instead of a simple one because we were working with a large dataset, and it made more sense to increase the number of folds. In this method, the data is divided into multiple "folds." The model is trained on all but one fold and is then tested on the held-out fold. This process is repeated for each fold, with a different fold held out each time. Essentially, Lasso Cross-Validation helps identify the optimal level of shrinkage that balances model simplicity and predictive accuracy.

Random Forest Regressor:

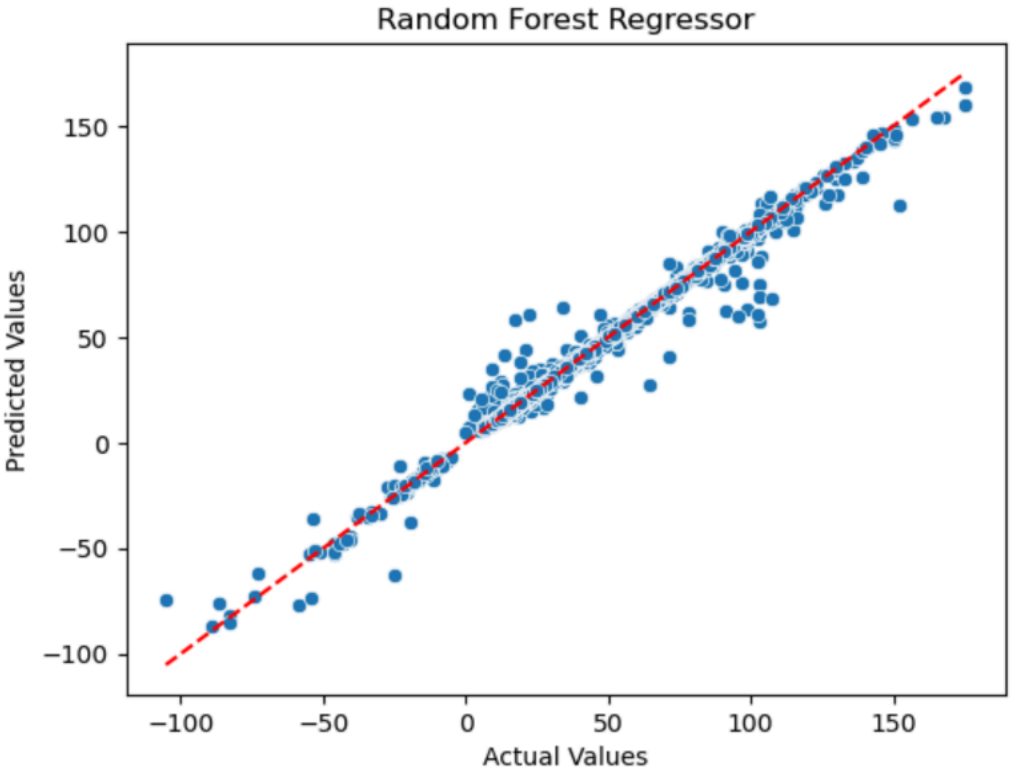
We started with the basic model, which is Linear Regression. After feeding the data into the model, we discovered that due to the large volume of data, it led to overfitting, resulting in a RMSE of 15.6. Naturally, we proceeded to add Lasso and Ridge penalties; however, these adjustments did not improve the model's precision.

This prompted us to address the challenge of handling such a massive dataset. We decided to implement an ensemble model, using multiple models or applying the same model several times. Specifically, we chose Random Forest Regression, which creates a large number of decision trees. Each tree is trained on a random subset of the training data.

At each node of a tree, a random subset of features is considered for splitting. This approach helps reduce the correlation between trees and enhances the model's generalization capabilities. Random Forests are effective in reducing overfitting, which leads to improved generalization performance.

Additionally, Random Forests can capture complex, non-linear relationships. For instance, the holiday feature we created did not predict the price linearly. After adding this feature, we were able to reduce the RMSE. In contrast, Linear Regression with the added penalties failed to capture this relationship effectively.

The Mean Absolute Error is 0.3669 and the Root Mean Square Error is 1.6936



Sno.	Model	RMSE	MAE
1.	Linear Regression	15.6065	10.1289
2.	Ridge Cross Validation	15.2813	10.2822

3.	Lasso Cross Validation	15.4593	10.5815
4.	Random Forest Regressor	1.6936	0.3669

Hence, to deal with this huge amount of data, using an ensemble model like Random Forest Regressor makes sense.

Demand Prediction for NYC Yellow Cabs Using SARIMA

The New York City Yellow Cab system has long been a cornerstone of urban mobility. However, with the rise of ride-sharing services like Uber and Lyft, coupled with fluctuating passenger demand, the need for accurate demand forecasting has never been greater. Understanding demand patterns allows taxi operators to optimize their services, ensuring efficient fleet management, cost reductions, and improved customer satisfaction. To address this challenge, our project implemented the SARIMA (Seasonal Autoregressive Integrated Moving Average) model to forecast taxi demand. This essay outlines the core steps in our analysis, the challenges faced during the process, and why such forecasting is critical for the business.

The Business Problem: Why Forecasting Matters

Predicting taxi demand in NYC is crucial for improving operational efficiency in the Yellow Cab industry. Demand fluctuations, caused by daily commuting patterns, seasonal changes, and external events, can leave drivers idle during slow hours or unprepared for surges. Accurate forecasting addresses these pain points:

1. Fleet Optimization: Predicting high-demand periods ensures taxis are allocated to the busiest areas, reducing wait times for passengers and minimizing idle time for drivers.
2. Cost Management: Forecasts can help taxi operators schedule drivers more effectively, avoiding unnecessary expenses during low-demand periods.
3. Strategic Planning: Forecasting enables preparation for major events, holidays, or weather changes, allowing taxis to meet increased demand without disruptions.

In essence, demand prediction enhances operational efficiency, reduces costs, and helps NYC's Yellow Cab system remain competitive in a dynamic transportation ecosystem.

Implementing the SARIMA Model

Initially, we implemented the ARIMA (AutoRegressive Integrated Moving Average) model to analyze and predict taxi demand based on historical data. The ARIMA model is widely used for time series analysis because of its simplicity and adaptability to various datasets. We prepared the data by resampling the trip distance on a daily frequency and fitting the ARIMA model with parameters `(1, 1, 1)` to capture trends and patterns in the data.

Despite its theoretical robustness, ARIMA failed to meet our expectations for this specific problem. The residuals revealed a significant lack of seasonality capture, and the model produced unreliable predictions. Additionally, the RMSE (Root Mean Square Error) indicated a poor fit. These shortcomings became evident when ARIMA struggled to capture the cyclical and seasonal nature of taxi demand in NYC, where daily and weekly patterns are influenced by factors like commuter rush hours, weather conditions, and special events. While ARIMA works well for stationary data, our dataset exhibited strong seasonality that could not be adequately addressed by the model.

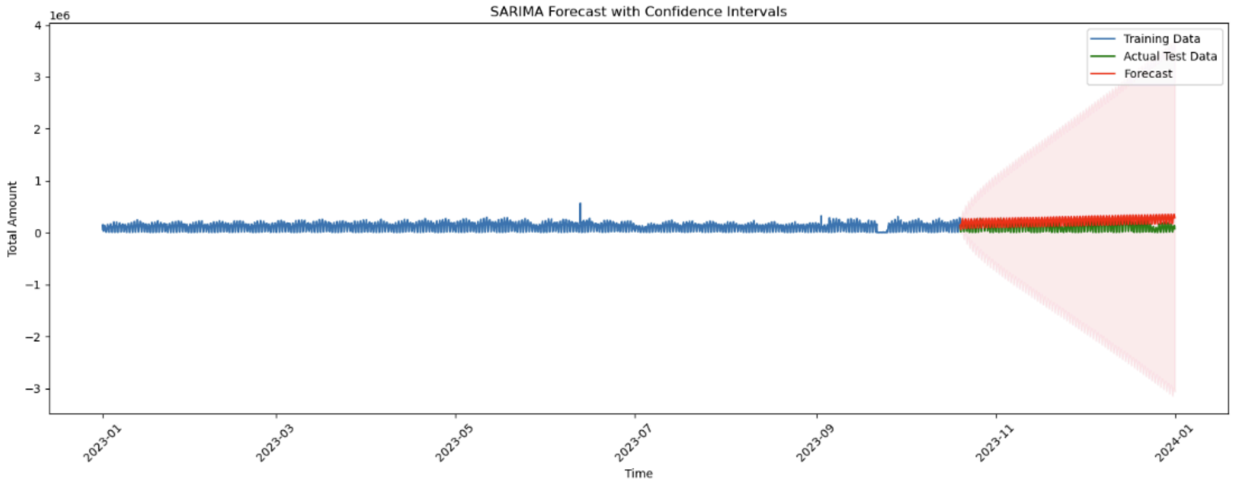
Realizing the limitations of ARIMA, we shifted to the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model, which extends ARIMA by incorporating seasonal components. The SARIMA model allowed us to model both short-term trends and long-term seasonal patterns effectively. We resampled the data to hourly frequency, carefully adjusted the seasonal order `(P, D, Q, s)`, and tested the model to ensure it captured the inherent seasonality of taxi demand.

Data Preparation:

- The data had to be cleaned and transformed for time series modeling. We dealt with missing timestamps by interpolation and ensured consistent hourly intervals. Outliers, such as extreme fare values, were removed to avoid skewing results.
- Here are few data graphs that are important for problem understanding:

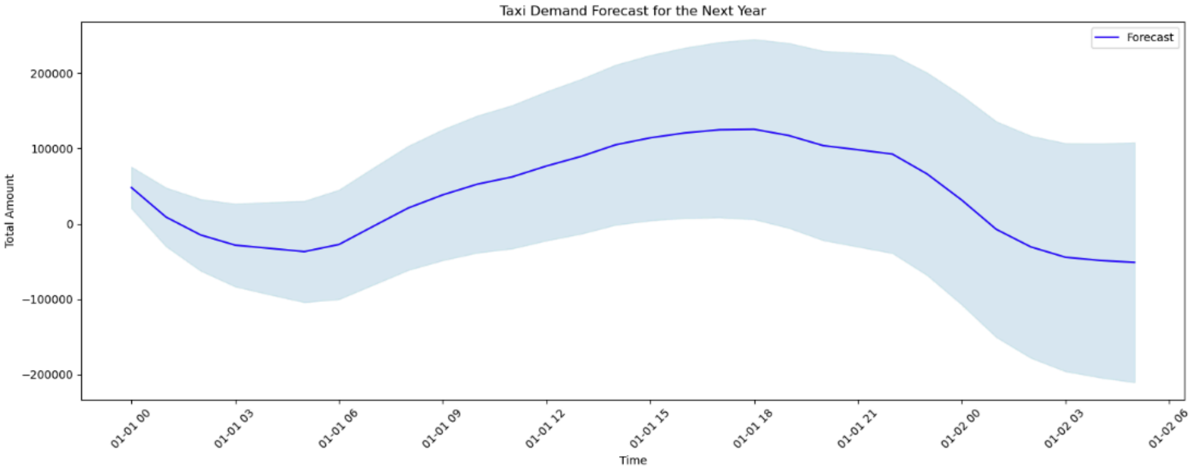
Model Design:

- SARIMA requires selecting parameters – (p, d, q) for the non-seasonal part and (P, D, Q, m) for the seasonal part.
- Through trial and error, we identified (1,1,1) for the non-seasonal component and (1,1,1,24) for the seasonal component, where the period of 24 reflected daily seasonality.



Forecasting:

- Using the trained SARIMA model, we forecasted demand for the next year. The forecast revealed daily peaks during rush hours and weekly trends with higher demand on weekends.



Challenges Faced

While SARIMA is a powerful tool, implementing it came with several challenges:

1. Data Quality: Missing timestamps and irregularities required extensive preprocessing. Aggregating the data into hourly intervals was time-consuming but necessary for consistency.
2. Outliers: Sudden spikes in fare or total amounts introduced noise. Identifying and removing these outliers was crucial to improve model accuracy.
3. Parameter Tuning: SARIMA's performance relies heavily on choosing the right parameters. Finding the optimal values for (p, d, q) and (P, D, Q, m) required multiple iterations and domain knowledge.
4. Computation Time: Training the model for long-term forecasts, especially with hourly data, was computationally intensive. This limited our ability to explore more complex configurations.

Despite these challenges, the SARIMA model produced meaningful results that highlighted both seasonal trends and long-term patterns in demand.

Insights and Results

The SARIMA model forecasted clear patterns in NYC taxi demand. Key observations include:

- Daily Seasonality: Demand peaks occurred during morning and evening rush hours, consistent with commuter activity.
- Weekly Trends: Demand increased significantly during weekends, particularly in high-traffic areas such as Manhattan.
- Long-term Patterns: Seasonal fluctuations due to holidays or adverse weather conditions were also evident.

The model’s confidence intervals provided insights into forecast uncertainty, helping operators plan for both expected and unexpected variations in demand.

The ability to predict demand using SARIMA provides a competitive advantage for NYC Yellow Cabs. By leveraging historical data and accounting for seasonal trends, operators can enhance fleet management, reduce idle costs, and improve passenger satisfaction. This project not only showcased the importance of time series forecasting but also demonstrated the practical challenges of implementing a complex model. In a fast-paced, competitive market, such data-driven approaches are essential to ensuring the long-term sustainability of the Yellow Cab industry.

Applications and Impact

By leveraging the results from the pricing recommendation system and demand prediction model, our project can provide reliable and data-driven decision support for managing demand and pricing strategies at specific times, start locations, and destinations. This capability enables Yellow Cab to dynamically adjust fares based on predicted demand patterns. For instance, during peak periods or adverse weather conditions, fares can be increased strategically to manage passenger demand through higher prices, thus balancing supply and demand more effectively. This approach not only helps to maximize revenue but also ensures optimal utilization of idle taxis, reducing operational inefficiencies.

Our project enables Yellow Cab not only to respond more effectively to passive high-demand peak hours but also to proactively allocate resources across a wider range of scenarios. The system can generate a dynamic dashboard on an hourly basis according to current and future demand and weather conditions. These results can serve as a guideline to guide drivers towards regions with anticipated high demand, thereby attracting potential passengers while minimizing ineffective operations, such as unnecessary idling or low-occupancy trips. The system reduces fuel consumption and vehicle wear and tear by optimizing routes and directing drivers strategically, ultimately enhancing economic efficiency and environmental sustainability. Additionally, it helps drivers maximize their earnings and strengthens satisfaction and retention. On a broader scale, this approach ensures a more balanced distribution of available taxis, reducing service wait times for passengers and improving overall customer experience.

Conclusion

Technically, I used a linear regression model and random forest regressor to develop the pricing prediction model. Among the performance evaluation, the random forest regressor demonstrated the best performance, achieving a root mean squared error (RMSE) of 1.628. For demand prediction, I selected the SARIMA (Seasonal ARIMA) model, an extension of the ARIMA framework designed to address seasonal patterns in time series data. SARIMA provided reliable results for demand forecasting, effectively capturing seasonal fluctuations and temporal trends.

The integration of demand prediction and pricing models enables Yellow Cab to optimize the balance between taxi supply and passenger demand. By dynamically adjusting fares and guiding resource allocation based on real-time and predictive data, the system reduces idle times, improves operational efficiency, and enhances profitability. This approach not only ensures better service availability for passengers but also maximizes driver earnings and strengthens Yellow Cab’s competitive position in NYC, fostering sustainable business growth.

References

[1] NYC Taxi & Limousine Commission, "Yellow Taxi Trip and Revenue Data," 2023. [Online]. Available: https://www.nyc.gov/taxi-limousine-commission. [Accessed: Dec. 3, 2024].

[2] Statista, "Taxi Service Revenue in New York City," 2023. [Online]. Available: https://www.statista.com/statistics/. [Accessed: Dec. 5, 2024].

[3] NYC Open Data, "Yellow Taxi Trip Records," 2023. [Online]. Available: https://opendata.cityofnewyork.us/. [Accessed: Dec. 6, 2024].

[4] J. Huang and Y. Wei, "Dynamic Pricing in the Ride-Hailing Industry: A Study of Uber's Surge Pricing Model," Journal of Business Analytics, vol. 14, no. 3, pp. 45–61, 2021.

[5] C. Silva and J. Green, "Time Series Forecasting for Urban Mobility Patterns: Applications in Taxi Fleet Optimization," Transportation Research Journal, vol. 56, no. 4, pp. 112–125, 2022.

[6] The New York Times, "Challenges of NYC's Yellow Taxi Industry," 2023. [Online]. Available: https://www.nytimes.com/transportation. [Accessed: Dec. 12, 2024].

Appendix

Code

The code is in the attached zip file.