Enhancing Taxi Operations : Advanced Demand Forecasting and Dynamic Pricing Strategy Development

A Final report for the BDM capstone Project

Submitted by

Name: Surya Vikram

Roll number: 22f3002751

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents :

Declaration Statement

I am working on a Project titled "Enhancing Taxi Operations : Advanced Demand Forecasting and Dynamic Pricing Strategy Development". I extend my appreciation to Sugam Sawaari, for providing the necessary resources that enabled me to conduct my project.

I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered from primary sources and carefully analyzed to assure its reliability.

Additionally, I affirm that all procedures employed for the purpose of data collection and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through thorough analytical procedures.

I am dedicated to adhering to the principles of academic honesty and integrity, and I am receptive to any additional examination or validation of the data contained in this project report.

I understand that the execution of this project is intended for individual completion and is not to be undertaken collectively. I thus affirm that I am not engaged in any form of collaboration with other individuals, and that all the work undertaken has been solely conducted by me. In the event that plagiarism is detected in the report at any stage of the project's completion, I am fully aware and prepared to accept disciplinary measures imposed by the relevant authority.

I understand that all recommendations made in this project report are within the context of the academic project taken up towards course fulfillment in the BS Degree Program offered by IIT Madras. The institution does not endorse any of the claims or comments.

Signature of Candidate:

Surya Vikram

Name: Surya Vikram

Date: 4th July, 2024

1 Executive Summary and Title

This Business Data Management (BDM) Capstone Project aims to streamline the operations of **Sugam Sawaari**, a local **taxi service** based in Masaurhi, Patna, that primarily offers transport solutions to residents in the region. The taxi service currently faces three major challenges: **inefficient demand forecasting**, **static pricing strategies**, and **high operational costs**, leading to suboptimal resource utilization and ultimately impacting profitability.

To address these challenges, the project adopts a comprehensive data collection and preprocessing approach, focusing on **key metrics** such as **demand**, **route distances**, **pricing**, and **weather conditions**. The analysis is specifically tailored to identify and rectify inefficiencies, aligning solutions with the unique operational needs of Sugam Sawaari.

For **demand forecasting**, a combination of **Time Series Models** including **SARIMAX, Exponential Smoothing,** and **XGBoost** is employed to create an ensemble model. This approach enhances the accuracy of future demand predictions, enabling Sugam Sawaari to effectively prepare for peak and off-peak periods.

To overcome the challenge of **static pricing strategies**, a **Random Forest Regressor** and a **Regularized Multiple Linear Regression (OLS)** model are utilized. These models facilitate the development of a dynamic pricing strategy that adjusts fares based on factors such as demand, distance, and special events (e.g., marriage seasons), thereby optimizing profitability while maintaining customer satisfaction.

In tackling **high operational costs**, the project includes a **Route Optimization** analysis, focusing on the most frequently traveled destinations. This optimization helps minimize fuel consumption and reduce travel time, significantly lowering operational expenses. Additional analysis is conducted to pinpoint and eliminate inefficiencies, further enhancing overall operational efficiency and profitability.

The report concludes by summarizing the results and providing actionable recommendations for Sugam Sawaari to implement in their business operations, thereby streamlining their processes and eliminating potential bottlenecks.

2 Detailed Explanation of Analysis Process/Method

2.1 Data Collection

The data collection process involved gathering information from four sources covering the period from August 2022 to July 2024. The primary data source, data.csv, contains 722 entries with fields such as date, demand, source, destination, distance, profit, and a marriage indicator. Secondary data sources include fuel.csv, detailing daily fuel prices; weather.csv, capturing weather data such as temperature and precipitation around Patna; and location.csv, which provides geographical coordinates for specific locations in the area.

2.2 Data Preprocessing

The preprocessed_data.csv file consolidates the original CSV files by aligning them on the common DATE field. Missing values were addressed using a two-step approach: first, a K-Nearest Neighbors (KNN) imputer was used to maintain data relationships, followed by mean imputation to fill any remaining null values, particularly in the DISTANCE and PROFIT fields.

2.3 Feature Engineering

To enhance the analysis, a FUEL_COST field has been introduced to estimate the fuel cost for each trip in rupees, calculated based on the trip distance (in kilometers), the taxi's mileage (15 kilometers per liter as specified by the owner), and the fuel cost per liter (in rupees). The fuel cost is then used to calculate EFFICIENCY, defined as the ratio of profit to fuel cost (both in rupees). This unitless efficiency metric acts as a key performance indicator, offering insights into the financial performance of each trip in relation to its fuel consumption and identifying the most profitable trips.

2.4 Time Series Analysis

Prior to conducting the analysis, the DEMAND was aggregated to a monthly level to smooth out daily fluctuations, revealing clearer long-term trends and seasonal patterns.

2.4.1 Seasonal Decomposition

Figure 1: Seasonal Decomposition of Monthly Demand

The trend component of the seasonal decomposition reveals a gradual yearly increase in demand. The seasonal component exhibits a clear yearly pattern, consistent across the observed time series. The additive nature of the decomposition suggests that these components combine to create the observed fluctuations in demand, with the seasonal effects maintaining a consistent magnitude over time, independent of the trend.

2.4.2 Augmented Dickey-Fuller Test for Stationarity

The ADF test results confirm the stationarity of the time series, with an ADF statistic of -3.4855 and a p-value of 0.0084, well below the significance level of 0.05. This allows us to confidently reject the null hypothesis of a unit root. Stationarity is essential in time series analysis as it ensures that the statistical properties of the series remain constant over time, making it suitable for many forecasting models. Additionally, it indicates that shocks to the system are temporary and that the series tends to revert to its mean. The series' stationarity, despite its clear seasonal patterns, implies that these patterns are consistent and predictable, providing a solid foundation for accurate forecasting.

2.4.3 ACF and PACF Plots

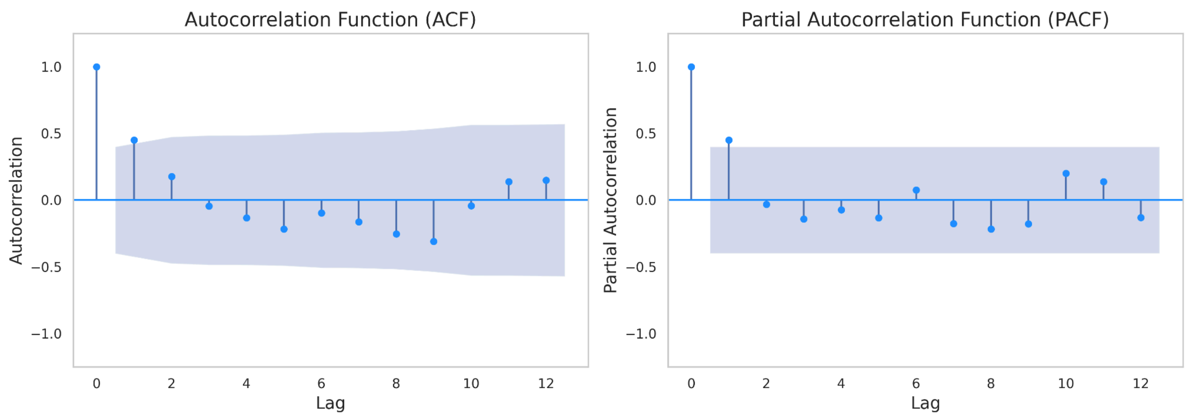

Figure 2: Correlation Plots for Monthly Demand

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots offer key insights for model selection. The ACF reveals strong autocorrelation at lag 1, indicating short-term dependence, and a prominent spike at lag 12, suggesting yearly seasonality. The PACF also shows significant spikes at lags 1 and 12, confirming monthly dependence and annual cycles. With few significant spikes beyond lag 1 in the PACF, an AR(1) process may effectively capture the non-seasonal component. Overall, the demand's trend and seasonal patterns suggest using the SARIMA model for forecasting.

2.4.4 Modeling Process

The SARIMAX model, with its non-seasonal component (1,0,1), effectively captures the short-term autocorrelation structure indicated by the PACF plot. The seasonal component (1,0,1,12) specifically addresses the monthly seasonality. This model is particularly strong in its ability to capture both trend and seasonality within a statistically rigorous framework.

Exponential Smoothing is well-suited for capturing the additive trend and seasonal components that align with the patterns seen in the decomposition plot. This model excels at handling evolving trends and seasonality, providing flexibility in how these components change over time. It serves as a robust alternative to SARIMAX.

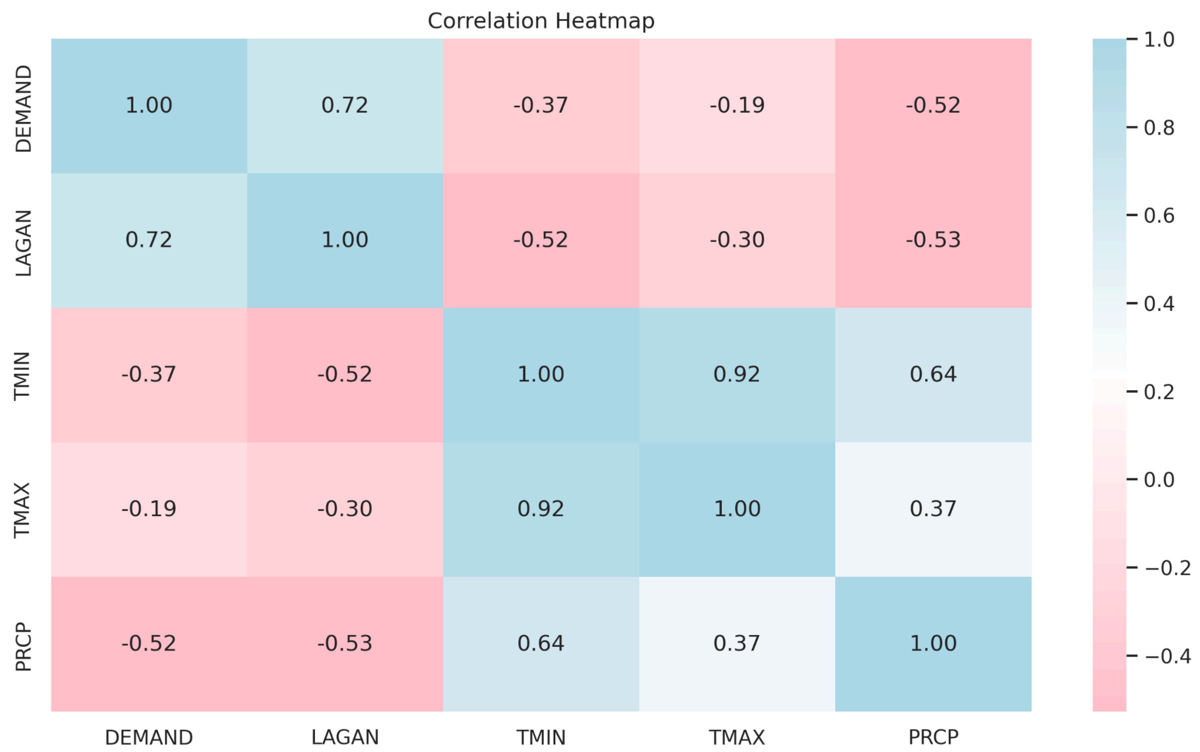

Figure 3: Correlation Heatmap highlighting the influence of various factors on Demand

XGBoost includes variables such as LAGAN, TMIN, TMAX, and PRCP to detect complex non-linear relationships, resulting in a training RMSE of 0.0009, but a higher test RMSE of 3.3851, indicating overfitting. This suggests a need for further tuning and regularization to improve generalization.

2.5 Regression Methods

Before fitting the regression model, the data preparation phase involves aggregating two essential features to a monthly level: DEMAND and LAGAN.

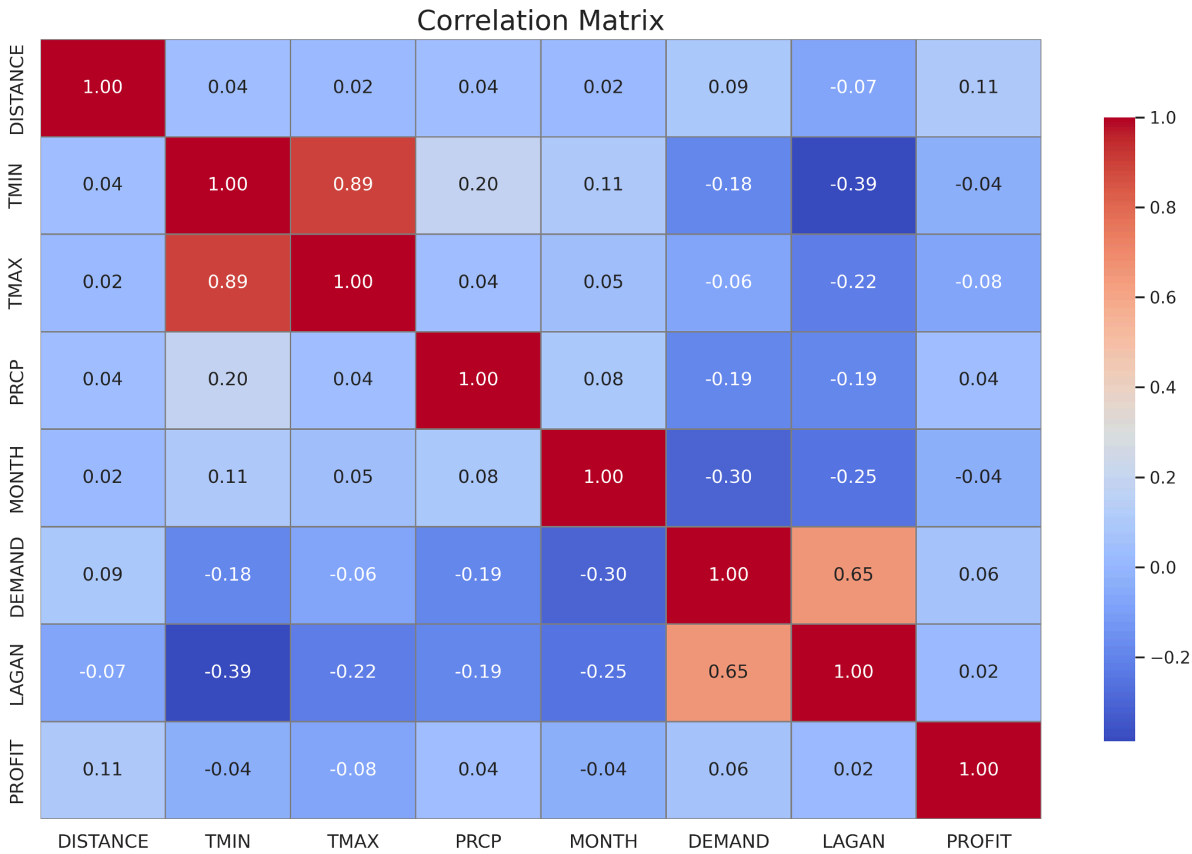

Figure 4: Correlation Heatmap illustrating the relationship between Profit and other key features

Based on the correlation matrix, which showed a high correlation of 0.65 between LAGAN and DEMAND, LAGAN was excluded to avoid multicollinearity. The dataset was split into 80% training and 20% testing sets. The MLR-OLS model was fitted on the training data to predict profit, capturing linear relationships between features and PROFIT. Despite a high correlation of 0.98 between predicted profit and distance, indicating good model performance, the linear approach may lead to customer dissatisfaction. Therefore, the model will be regularized to set a profit lower bound, and a Random Forest Regressor will be used to capture non-linear relationships and enhance the dynamic pricing strategy.

2.6 Analyzing Routes to the Prime Destination

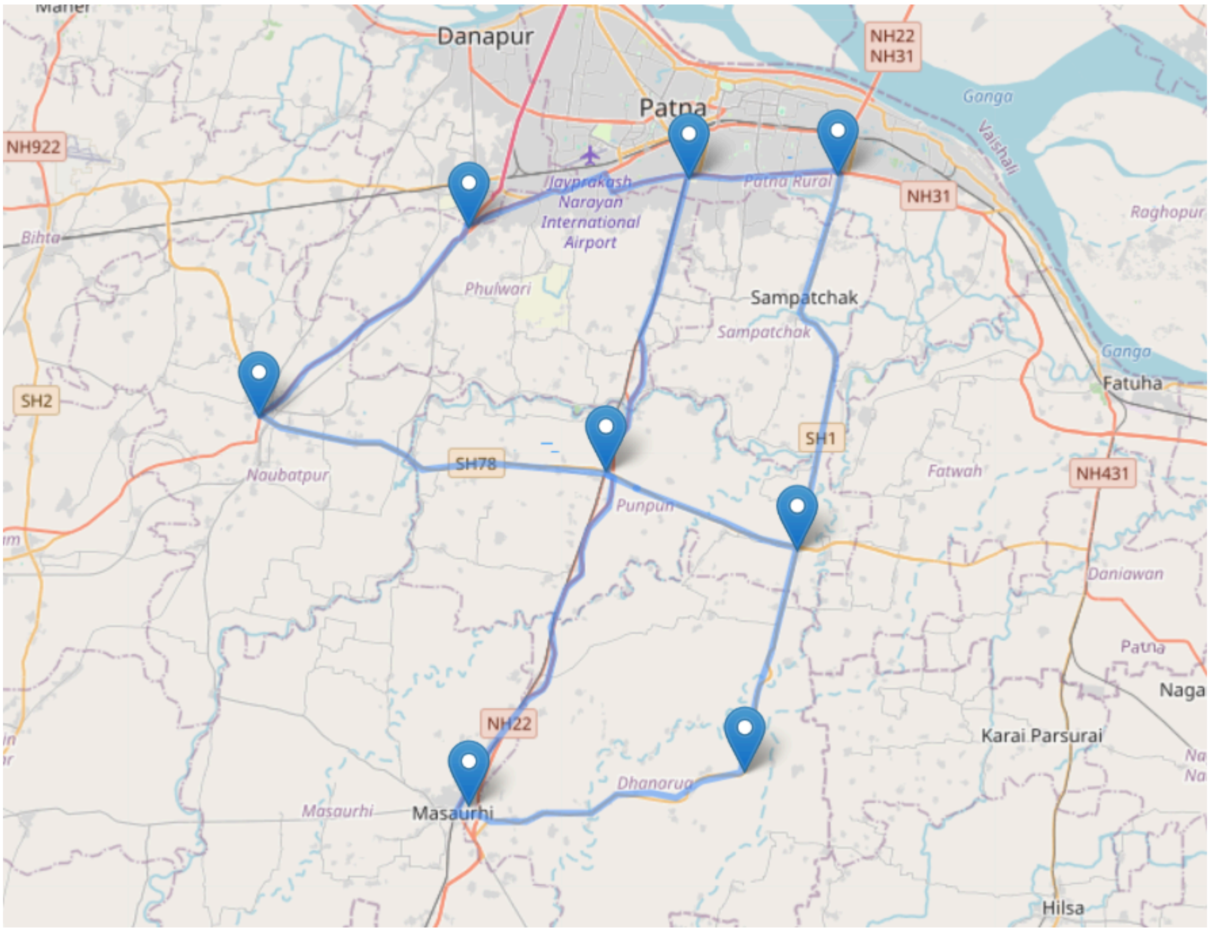

Figure 5: A map of Patna, Bihar, showing eight vertices marked with pins and blue edges

From the earlier wordcloud and frequency analysis of TO and FROM features, it is evident that Masaurhi serves as the primary hub for taxi services, originating 95% of all rides, while Patna emerges as a significant destination, accounting for 27% of all journeys. The transportation network is visually represented using Folium, with routes weighted by haversine distances to facilitate shortest path calculations via Dijkstra's algorithm. While the system accommodates flexible departure points near Masaurhi, it prioritizes a custom mapping solution over Google Maps due to past navigational issues, thereby ensuring more reliable service delivery. This data-driven approach enables the optimization of routes and resource allocation, potentially improving overall efficiency and customer satisfaction.

3 Results and Findings

3.1 SARIMAX

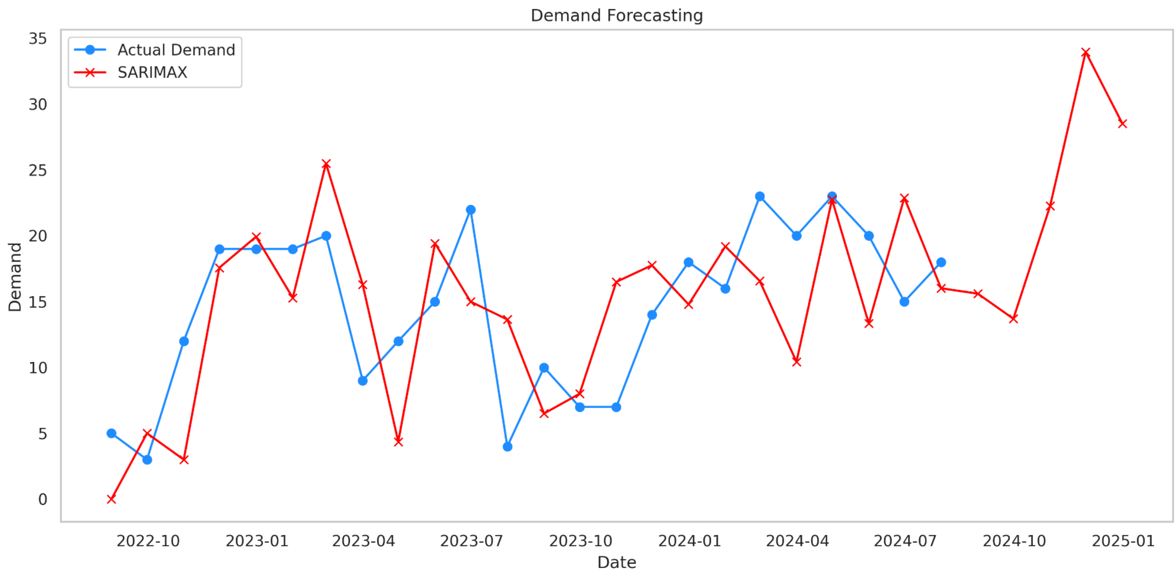

Figure 6: Temporal Analysis of SARIMAX Predictions and Observed Demand

The initial Seasonal Autoregressive Integrated Moving Average model with exogenous variables (SARIMAX), specified as **SARIMAX**(order=(1, 0, 1), seasonal_order=(1, 0, 1, 12)), yielded promising results with an R-squared score of 0.8642. However, upon closer examination, it became apparent that the forecasted demand was nearly constant for the five-month period from August 2024 to December 2024. This was evidenced by the low variance in the forecast (0.223) and a mean forecast value of 15.80, suggesting that the initial approach might not effectively capture seasonal variations or trends in the demand data.

To address these limitations, a refined model was developed by incorporating exogenous variables, particularly 'LAGAN', which represents auspicious marriage occasions known in advance. The parameters were adjusted to include differencing, resulting in a new specification: **(order=(1, 1, 1), seasonal_order=(1, 1, 1, 12))**. This adjustment demonstrated a marked improvement in capturing trends and seasonality, producing varied forecasts with a mean of approximately 22.81 and a variance of 73.06.

The inclusion of differencing in both the regular and seasonal components, along with the 'LAGAN' exogenous variable, significantly enhanced the model's ability to account for trends and known future events impacting demand.

While the improved model showed substantial enhancements in trend capture and variability, it was noted that some predictions might be extreme. To mitigate this, consideration is being given to reducing the weight of this forecast within an ensemble approach, aiming to balance its trend-capturing capabilities with the stability of other forecasting methods.

3.2 Holt-Winters Exponential Smoothing

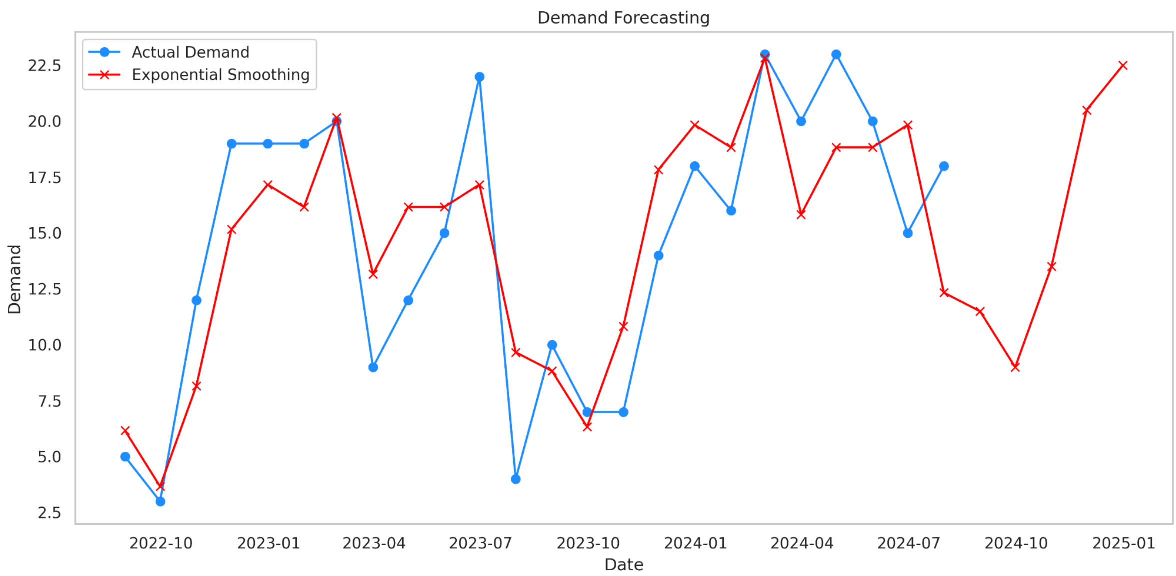

Figure 7: Temporal Analysis of Exponential Smoothing Predictions and Observed Demand

The Holt-Winters Exponential Smoothing model was implemented as a univariate time series forecasting method, complementing the multivariate SARIMAX approach. This additive model, with both trend and seasonal components and a seasonal period of 12 months, was fitted to the entire dataset. The model achieved an R-squared score of 0.6978, indicating its ability to explain approximately 69.78% of the variance in the observed demand. It offers several advantages, including its simplicity, computational efficiency, and effectiveness in capturing both trend and seasonality without requiring exogenous variables.

3.3 Hyperparameter-Tuned XGBoost

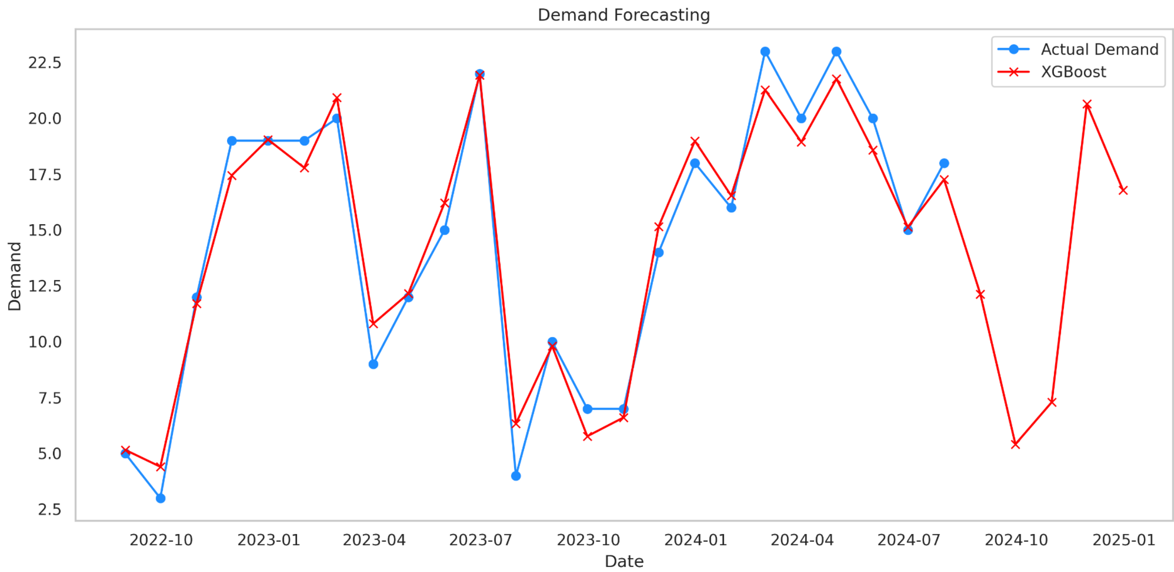

Figure 8: XGBoost Model Performance - Actual vs. Predicted Demand

To address the issue of overfitting observed in the initial XGBoost model, which achieved an R-squared of 1.00 on the training set but only 0.59 on the test set, several regularization techniques were applied. The model was tuned by reducing the number of trees (n_estimators) to 50 and decreasing the learning rate to 0.1, which helped in controlling the learning process and avoiding overly complex models. The maximum tree depth was limited to 3 to prevent the model from learning too many details specific to the training data. Additionally, subsampling was introduced, where 80% of the data and 80% of the features were used at each iteration, promoting model generalization. Further regularization was applied using a gamma value of 1, which added constraints on the minimum loss reduction required to make a split. L1 (reg_alpha=0.5) and L2 (reg_lambda=1) regularizations were also implemented to penalize large coefficients, thereby reducing model complexity. These adjustments led to a significant improvement in the model's performance, reflected by a more balanced R-squared of 0.96 on the training set and 0.86 on the test set, with corresponding RMSE values of 1.07 and 1.95 indicating a well-regularized and generalizable model.

3.4 Advanced Demand Forecasting Model

Each model presents unique strengths: SARIMAX offers a statistically grounded approach to time series forecasting, Exponential Smoothing provides flexibility in trend and seasonal modeling, and XGBoost leverages additional variables for potentially more accurate predictions. The correlation matrix reveals that while each model has a strong correlation with the actual demand, they also capture different aspects of the underlying data. This diversity in model behavior supports the decision to use an ensemble approach.

$$\text{Ensemble Forecast} = 0.3\times\text{SARIMAX}+0.4\times\text{XGBoost}+0.3\times\text{Exponential Smoothing}$$

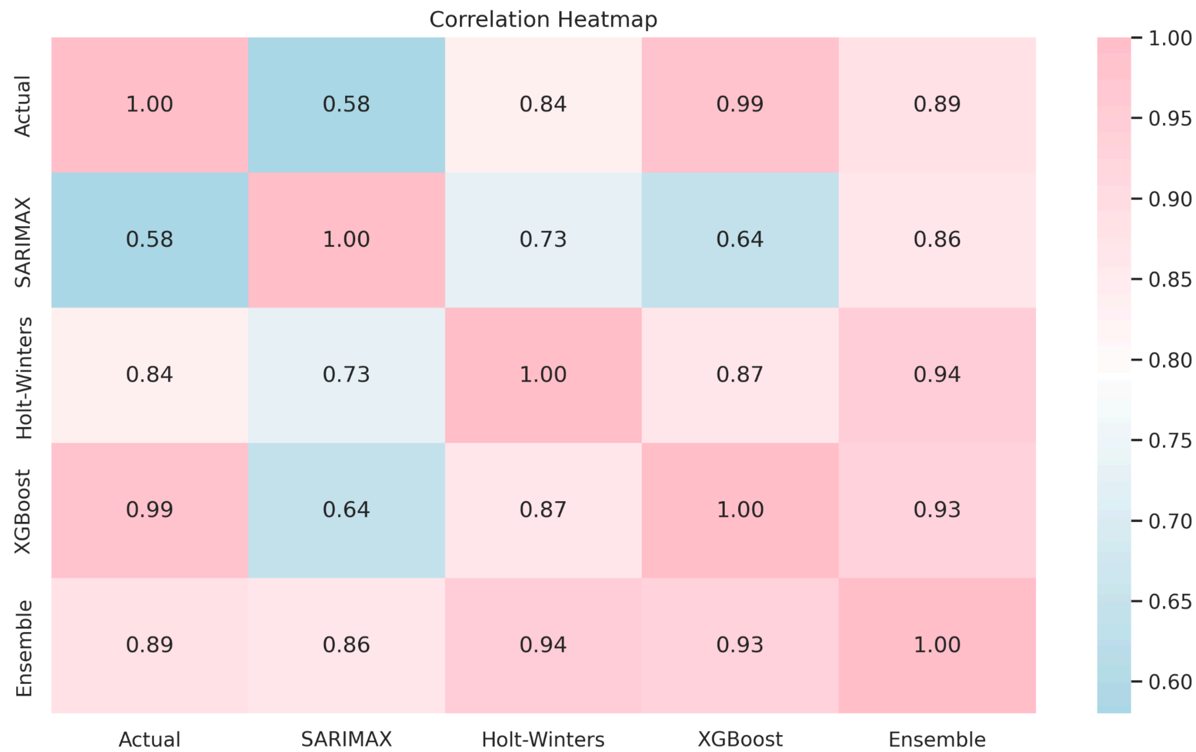

Figure 9: Correlation Heatmap of Actual Demand and Model Predictions

The ensemble method combines each model's strengths using fine-tuned weights for optimal balance. SARIMAX provides a reliable statistical base, while Exponential Smoothing captures trend and seasonality. XGBoost, despite its overfitting tendency, effectively captures yearly patterns and complex non-linear relationships.

3.5 Regularized MLR OLS Results

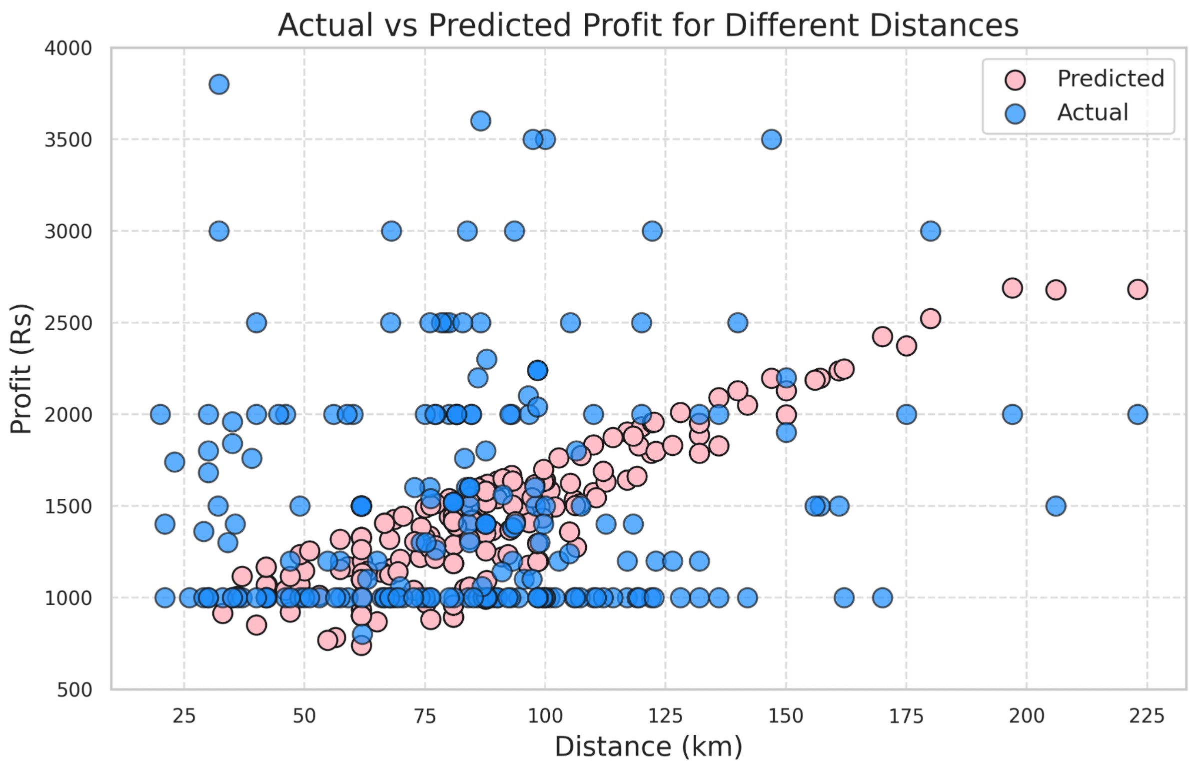

Figure 10: Predicted MLR OLS Profit vs Actual Profit (Blue)

The Ridge regression model, with a regularization strength of 100, is utilized to establish a lower bound on pricing. The evaluation of the model highlights that demand is significantly more important (32.72) compared to distance (9.83). This disparity indicates that demand is the primary driver of the pricing strategy, suggesting that the pricing model relies heavily on anticipated demand levels rather than the distance traveled.

The model is intentionally designed to underprice, thereby establishing a lower bound. This strategy facilitates setting minimum price thresholds and assists in identifying routes or scenarios that may not be profitable. Consequently, informed decisions can be made regarding viable routes and necessary price adjustments or additional considerations. Scatter plot analysis comparing actual profits to predicted profits shows that actual profits are often higher than the predictions made by the model. This pattern aligns with the model's approach of setting a lower bound on prices.

3.6 Random Forest Regressor

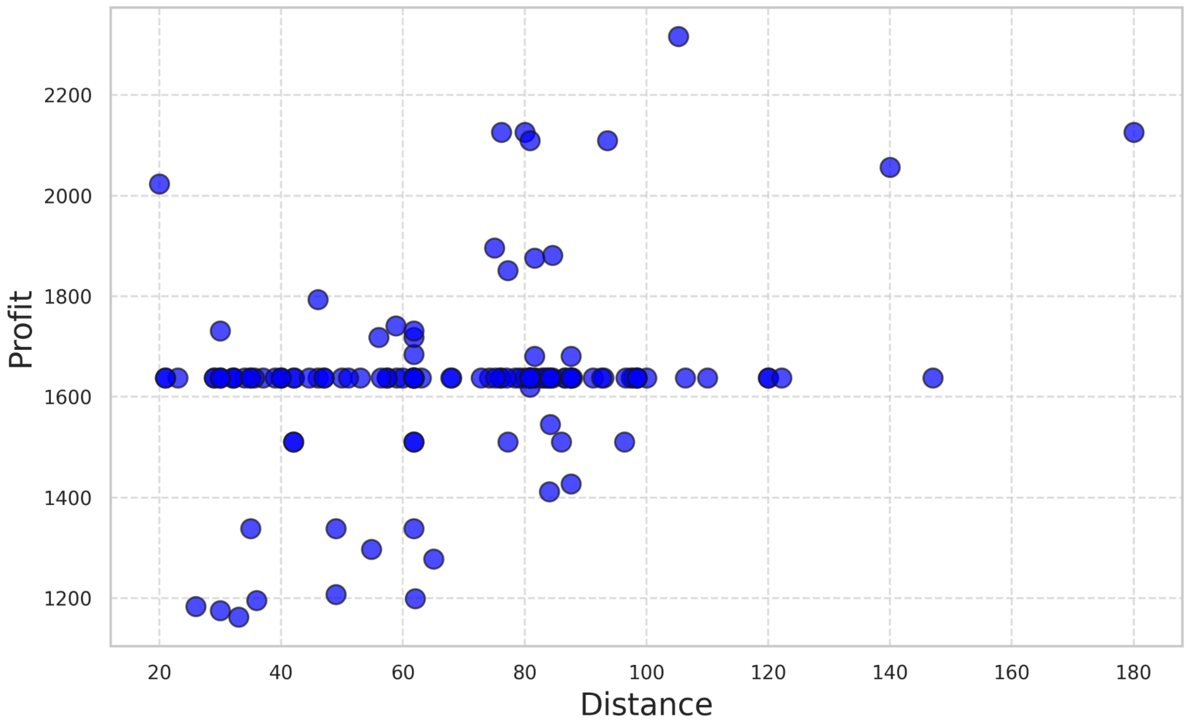

Figure 11: Scatter Plot of Predicted Profits Against Distances

The model serves as an overpricing predictor, trained on rides with an efficiency greater than 2.5. This threshold lies between the 25th and 100th percentiles of efficiency, effectively eliminating low-profit and static pricing rides. The model incorporates both numeric features (DISTANCE, MONTHLY_LAGAN, MONTHLY_DEMAND, TMIN, TMAX, PRCP) and categorical features (FROM, TO), with the latter processed using one-hot encoding. A pipeline approach is employed, combining preprocessing steps with a Random Forest model consisting of 100 estimators.

Analysis of the model reveals a moderate positive correlation (0.311) between predicted profit and distance, indicating that longer distances tend to yield higher predicted profits. The scatter plot of predicted profit versus distance visually confirms this trend, while also highlighting the variability in predictions. This overpricing model, when used in conjunction with the underpricing Ridge regression model, provides a range for pricing decisions. It identifies high-potential scenarios for maximizing revenue, enabling more dynamic and flexible pricing strategies.

3.7 Dynamic Pricing Model

The dynamic pricing model combines the strengths of both the underpricing (Ridge regression) and overpricing (Random Forest) models to create a flexible pricing strategy. This ensemble approach allows for real-time adjustments based on market conditions, demand fluctuations, and operational factors. The core of this model is represented by the following equation:

$$\text{Ensemble Prediction} = w_1 \cdot \text{UnderPricingModel} + w_2 \cdot \text{OverPricingModel}$$

$$\text{where } 0 \leq w_i \leq 1 \text{ for all i, and } \sum_{i=1}^{n} w_i = 1$$

Where w_1 and w_2 are adjustable weights, each constrained between 0 and 1, and their sum must equal 1. These weights can be dynamically tuned to shift the pricing strategy between conservative (favoring the underpricing model) and aggressive (favoring the overpricing model) approaches.

This dynamic model offers several advantages. It provides a robust pricing range, with the underpricing model establishing a lower bound to ensure profitability, and the overpricing model identifying opportunities for higher margins. The ability to adjust weights in real-time allows for rapid response to changing market conditions. For instance, during peak demand periods, the weight of the overpricing model (w_2) could be increased to capitalize on higher willingness to pay. Conversely, during off-peak times or in highly competitive scenarios, the weight of the underpricing model (w_1) could be increased to maintain market share. This adaptive approach enables a balance between maximizing revenue and ensuring competitiveness across various operational scenarios.

3.8 Optimal Route Classification and Findings

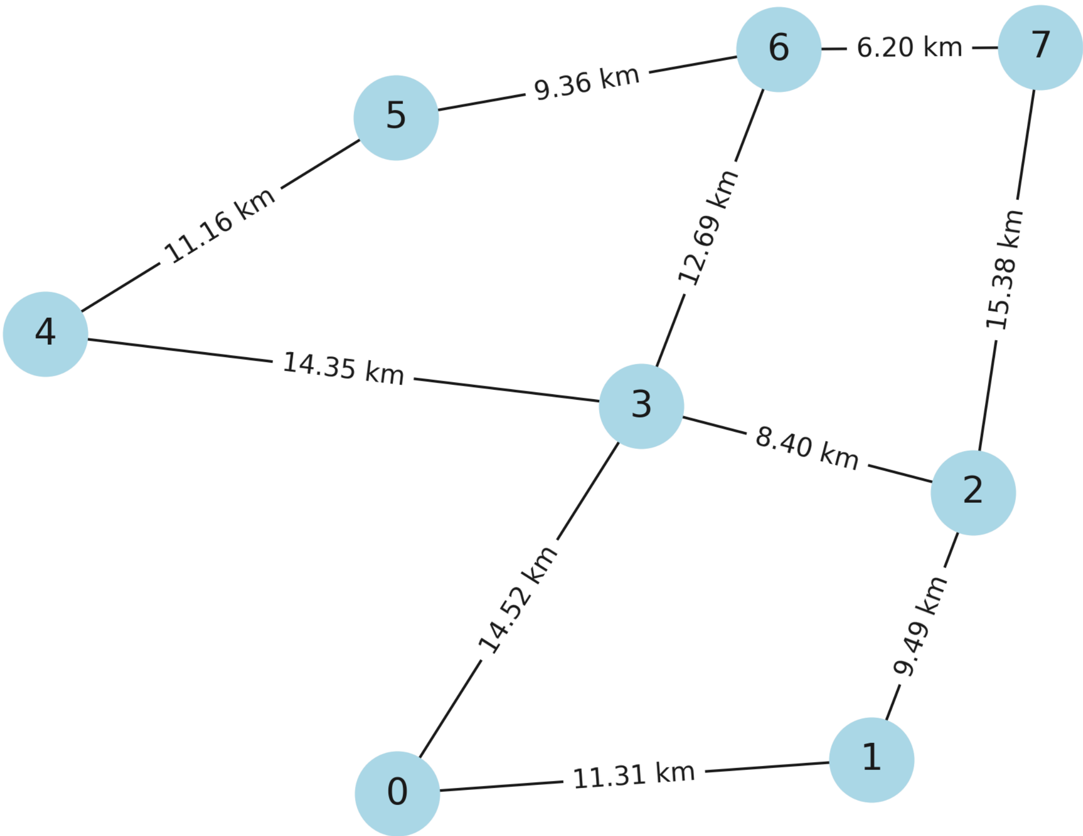

Figure 12: Network Graph of Routes Connecting Masaurhi and Patna

This figure illustrates a network of eight key locations, labeled from 0 to 7, connected by roads of varying distances. The locations are represented by light blue circular nodes, while the roads are depicted as black lines connecting these nodes. Each road is annotated with its length in kilometers, ranging from 6.20 km to 15.38 km. The graph shows a non-uniform distribution of nodes, with some areas more densely connected than others. Notable features include:

- **Origins of Rides:** Nodes 0, 1, and 3 can be the origin of rides around Masaurhi, reflecting key points of departure in the network.
- **Patna Connections:** Nodes 5, 6, and 7 represent areas in Patna, which are connected to the Masaurhi region via three different paths.

- **Central Hub:** A central node (3) connects to five other nodes, serving as a major hub in the network, enhancing connectivity and route flexibility.
- **Peripheral Nodes:** Nodes 4, 5, and 7 are peripheral with fewer connections, indicating potential endpoints or less traveled routes.
- **Variety of Path Lengths:** The graph provides multiple route options between most pairs of nodes due to the variety of path lengths.

The shortest path algorithm employed is Dijkstra's algorithm, a graph search method that solves the single-source shortest path problem for graphs with non-negative edge weights. It maintains a set of vertices with known shortest distances from a source and iteratively expands this set until the destination vertex is included. In this context, the algorithm is applied to a graph representing a road network, where vertices (0-7) correspond to key locations, and edges represent roads weighted by the Haversine distance between endpoints. The source and destination are mapped to the nearest vertices by calculating the Haversine distance from the actual locations to these vertices. NetworkX's implementation of Dijkstra's algorithm then determines the shortest path between these mapped start and end vertices. The total route distance is computed by summing the distance from the actual source to the start vertex, the path length within the graph, and the distance from the end vertex to the actual destination. This path is represented as a list including the actual source and destination along with the intermediate vertices, allowing for efficient route calculation.

The Haversine Formula is used to calculate the distance (in kilometers) between two points along the edges.

$$d = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

where:

-ϕ_1 and ϕ_2 are the latitudes of the two points in radians.

- $\Delta\phi$ is the difference between the latitudes of the two points ($\phi_2 - \phi_1$).

- $\Delta\lambda$ is the difference between the longitudes of the two points ($\lambda_2 - \lambda_1$).

- R is the Earth's radius (mean radius = 6,371 km).

4 Interpretation of Results and Recommendations

4.1 Recommendation 1: Forecasting Model and Maintenance Schedule

Interpretation of Results:

The ensemble forecasting model combines SARIMAX, Exponential Smoothing, and XGBoost, each offering unique strengths in capturing different aspects of the underlying data:

- **SARIMAX** provides a statistically grounded approach to time series forecasting.
- **Exponential Smoothing** offers flexibility in trend and seasonal modeling.
- **XGBoost** leverages additional variables for potentially more accurate predictions, despite its tendency to overfit.

The correlation matrix reveals that while each model correlates strongly with actual demand, they capture distinct aspects, supporting the use of an ensemble approach. Fine-tuned weights (XGBoost: 0.4, SARIMAX: 0.3, Exponential Smoothing: 0.3) ensure an optimal balance, with the ensemble model reliably predicting demand fluctuations.

Recommendations:

1. **Schedule Maintenance During Low-Demand Months:**
   1.1. Utilize the ensemble model to forecast low-demand periods.
   1.2. Plan maintenance activities during these months to ensure high vehicle availability during peak demand periods.
2. **Retrain the Model with Sufficient New Data:**
   2.1. Regularly monitor the demand data and retrain the ensemble model as new data accumulates to maintain accuracy and adapt to changing demand patterns.

By following these recommendations, the taxi service can enhance operational efficiency, reduce downtime, and improve overall profitability by strategically aligning vehicle maintenance with demand forecasts.

4.2 Recommendation 2: Dynamic Pricing Strategy

Interpretation of Results:

The dynamic pricing model blends the advantages of underpricing (Ridge regression) and overpricing (Random Forest) to create a flexible pricing strategy. It adjusts in real-time based on market conditions and demand changes. By adjusting the model's weights, it can switch between a lower pricing strategy to attract more customers and a higher pricing strategy to increase revenue during peak times. This approach ensures a balanced and responsive pricing strategy.

Recommendations:

1. **Peak Demand Periods (Waiting Strategy):**
    - 1.1. During peak demand periods, such as the Lagan months, delay offering advance bookings at lower fares to align prices more closely with the Random Forest Regressor's upper bound.
    - 1.2. Increase the weight of the overpricing model to capitalize on higher customer willingness to pay and maximize profitability.
2. **Low Demand Periods:**
    - 2.1. During low demand periods, adjust pricing towards the MLR-OLS model to ensure affordability and attract more customers.
    - 2.2. Increase the weight of the underpricing model to maintain market share and competitiveness.
3. **Real-time Monitoring and Adjustment:**
    - 3.1. Continuously monitor market conditions and demand levels to dynamically adjust the weights between the underpricing and overpricing models.
    - 3.2. Utilize real-time data to inform pricing decisions, allowing for quick and effective responses to changes in customer demand and competitive pressures.

This dynamic pricing strategy provides a balanced and flexible approach to pricing, enabling the taxi service to optimize revenue generation while maintaining customer satisfaction and competitiveness across various market scenarios.

4.3 Recommendation 3: Optimal Route Selection Algorithm

Interpretation of Results:

An in-depth data analysis reveals Masaurhi as the central hub of taxi operations, with 95% of rides starting from this location, and Patna as a key destination, attracting 27% of all journeys. The route network visualization, developed using Folium, accurately maps the most popular routes to Patna by representing key locations as vertices and roads as edges, weighted by haversine distances. This model effectively uses Dijkstra's algorithm to calculate the shortest paths, offering a more reliable and efficient mapping solution than third-party tools like Google Maps. Due to previous navigation issues with Google Maps, this customized approach is prioritized to reduce errors and ensure smoother service delivery.

Recommendations:

1. **Prioritize Custom Mapping Solutions:**
    - 1.1. Continue to use the customized mapping solution that leverages Dijkstra's algorithm for route selection, ensuring accuracy and reliability in navigation.
    - 1.2. Avoid relying on Google Maps to prevent navigation into impassable or unsuitable routes, thus improving service quality and safety.
2. **Monitor and Update Route Data:**
    - 2.1. Regularly update the route network data to reflect changes in road conditions, traffic patterns, and customer preferences and adjust weights accordingly.
    - 2.2. Use feedback from drivers and passengers to continuously improve the route selection algorithm and mapping solution.

This strategy synthesizes data-driven insights with practical considerations, fostering a more robust and dependable transportation framework that enhances customer satisfaction and service efficiency.