Taxi Availability Prediction in Singapore

Group 23:
Peng Ziwei, Erik Naeslund, Tan Kai Xin, Nicole

November 18, 2024

Contents

1 Introduction **2**

2 Problem Description **2**

3 Datasets **2**

4 Methodology **2**

5 Results and Analysis **3**
5.1 Rainfall Dataset 3
5.2 UV Index Dataset 3
5.3 Carpark Availability Dataset 4
5.4 Combined Model 4

6 Reflections and Conclusions **4**

7 Proposed Practical Action **5**

1 Introduction

Taxi availability is a critical factor for urban mobility in Singapore, where high population density and dynamic environmental conditions often lead to unpredictable demand-supply mismatches. This study aims to develop a predictive model for taxi availability by leveraging data on rainfall, UV index, and carpark occupancy. The model provides actionable insights to improve commuter satisfaction, optimize taxi fleet management, and reduce environmental emissions.

The relevance of this study lies in its alignment with Singapore's Smart Nation initiative [1], which seeks to enhance mobility through data-driven solutions. By integrating diverse datasets, this project addresses the challenges of urban transportation and contributes to a sustainable and efficient mobility ecosystem.

2 Problem Description

The primary objective of this study is to predict taxi availability in real-time using external factors such as rainfall, UV index levels, and carpark occupancy. These factors may influence commuter behavior and transportation demand. For instance, heavy rainfall often leads to increased reliance on taxis, while high UV index levels discourage outdoor activity, driving similar demand surges. Additionally, carpark availability serves as a proxy for urban congestion, indirectly impacting taxi utilization.

Singapore's compact geography and advanced data infrastructure make it an ideal setting for implementing such predictive models.

3 Datasets

This project utilizes four key datasets. Rainfall data, updated every five minutes, provides precipitation readings aggregated to hourly intervals. UV index data captures data poitns hourly. Carpark availability data tracks HDB carpark occupancy updated every minute and aggregated hourly, serving as a proxy for urban congestion. Lastly, taxi availability data, retrieved every 30 seconds, provides real-time information on taxi distribution across the city.

To simplify analysis, the study assumes uniform distribution of rainfall and UV levels across Singapore and relies on aggregated carpark and taxi data to represent broader urban patterns. These assumptions, while practical, may limit the granularity of the findings.

4 Methodology

The experimental design involves modelling taxi availability using individual datasets and an integrated approach. First, rainfall data is used to examine how precipitation affects taxi demand through a multi-layered perceptron model. Similarly, the UV index dataset is analysed to understand its impact on outdoor mobility and taxi usage. Car park availability data is modelled to investigate the relationship between occupancy ratios and taxi utilization, particularly in congested areas.

Finally, an integrated model combines all datasets to predict taxi availability for August 2024. The model is trained on July 2024 data using a multi-layered perceptron,

allowing for a holistic analysis of the interplay between environmental and urban factors. Evaluation metrics such as accuracy, mean absolute error (MAE), and correlation coefficients are used to assess model performance.

5 Results and Analysis

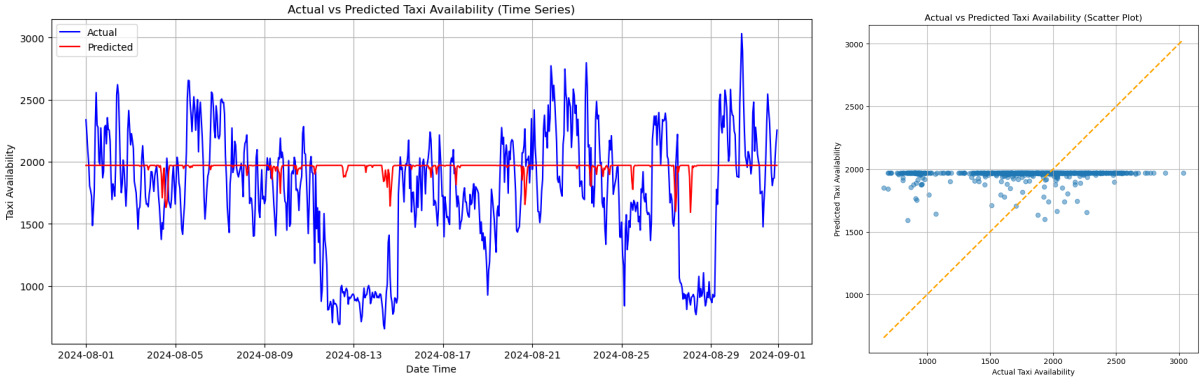
This section summarizes the analysis of taxi availability predictions using rainfall, UV index, carpark availability datasets, and their combined integration. A summary of performance metrics for all models is presented in Table 1. Visual representations of the actual vs. predicted values, including time series and scatter plots, are provided in Figure 1 to Figure 4.

Table 1: Performance Metrics for Taxi Availability Prediction Models

Dataset	**Mean Squared Error (MSE)**	**R^2 Score**	**Pearson**	**Spearman**
Rainfall	271,570.76	-0.1601	0.0885	0.1316
UV Index	231,815.92	0.0098	0.1147	0.1183
Carpark Availability	309,045.38	-0.3201	0.0809	0.1018
Combined Model	262,615.71	-0.1218	0.1226	0.1586

5.1 Rainfall Dataset

The rainfall dataset (Figure 1) captures foundational trends in taxi availability during weather-driven demand shifts. However, the relatively high MSE and low correlations (Table 1) suggest limited predictive accuracy when used in isolation.

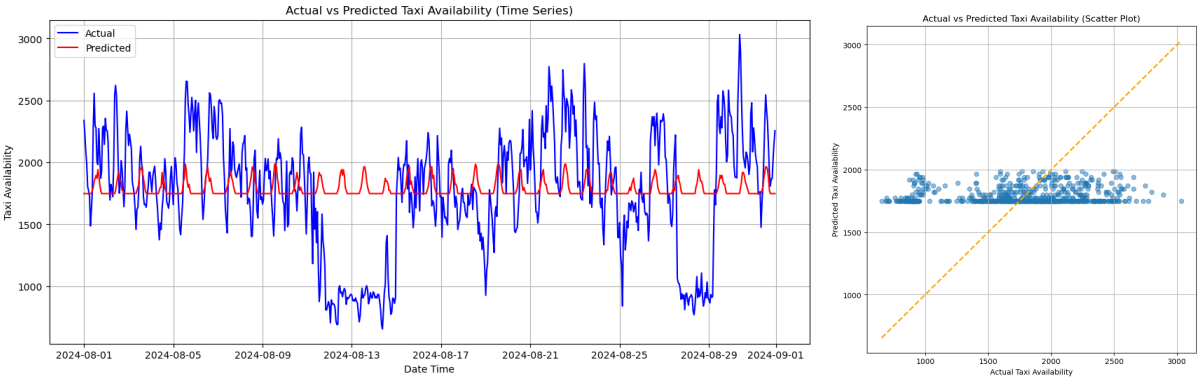


(a) Time Series Plot (b) Scatter Plot

Figure 1: Actual vs Predicted Taxi Availability Using Rainfall Dataset

5.2 UV Index Dataset

The UV index dataset (Figure 2) adds nuanced insights into demand variations during high UV periods. As seen in Table 1, it has a lower MSE and improved correlations compared to the rainfall dataset.

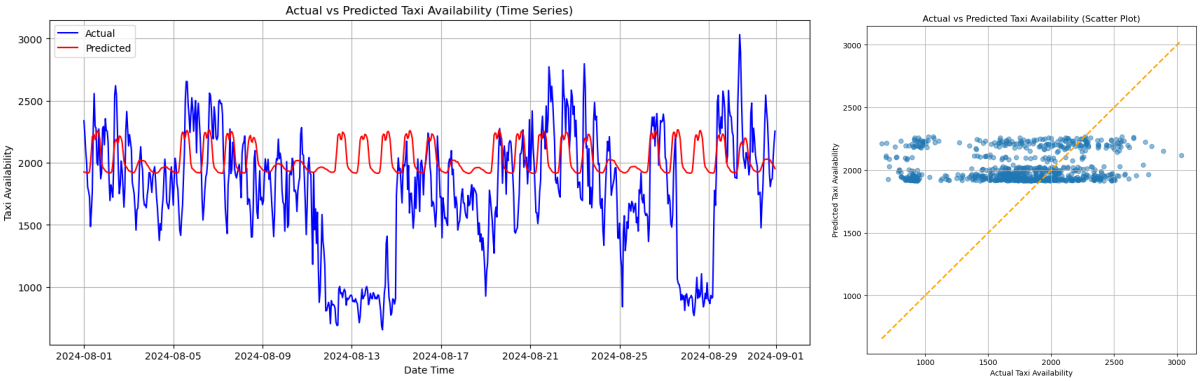


(a) Time Series Plot (b) Scatter Plot

Figure 2: Actual vs Predicted Taxi Availability Using UV Index Dataset

5.3 Carpark Availability Dataset

The carpark availability dataset (Figure 3) highlights potential inverse relationships between carpark occupancy and taxi availability. However, as shown in Table 1, it has the highest MSE and lowest correlations among all datasets.



(a) Time Series Plot (b) Scatter Plot

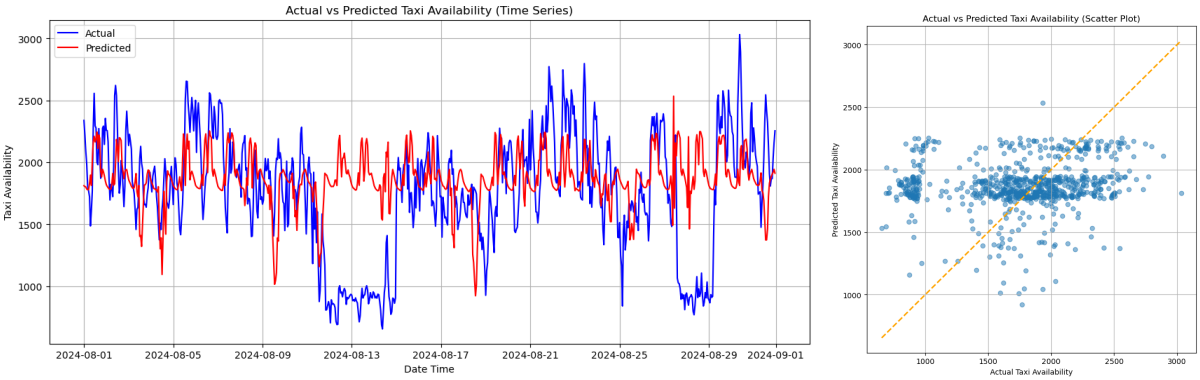
Figure 3: Actual vs Predicted Taxi Availability Using Carpark Availability Dataset

5.4 Combined Model

The combined model (Figure 4) outperforms individual datasets, achieving the lowest MSE and highest correlations (Table 1). This demonstrates the importance of integrating multiple data sources.

6 Reflections and Conclusions

The results (Table 1) demonstrate the significant improvement offered by integrating datasets. While individual datasets like rainfall and carpark availability highlight specific trends, their limitations in predictive power emphasize the necessity of combining diverse data sources. Visual analysis in Figure 1 to Figure 4 shows that the combined model effectively captures general trends and reduces prediction errors.



(a) Time Series Plot (b) Scatter Plot

Figure 4: Actual vs Predicted Taxi Availability Using Combined Datasets

It is important to note that our analysis works with correlation rather than causation. The datasets utilized in this study are highly correlated with factors such as time of day and type of day (e.g., weekdays versus weekends). While these temporal patterns provide valuable predictive signals, they may also limit the ability to establish direct causal relationships between variables.

Future work should focus on enriching datasets with additional factors such as public events, traffic congestion, and socioeconomic variables to enhance prediction accuracy. The insights from this study underline the importance of data-driven approaches in urban planning.

7 Proposed Practical Action

A predictive system based on the combined dataset can benefit various stakeholders:

- **Commuters**: Access to real-time forecasts for better trip planning.
- **Taxi Operators**: Optimized fleet distribution to reduce idle time and improve efficiency.
- **Urban Planners**: Insights into mobility trends to guide infrastructure development and policy decisions.

By aligning with Singapore's Smart Nation vision, such a system could serve as a scalable model for other urbanized cities, promoting sustainable and efficient transportation systems.

References

[1] Government of Singapore. Smart nation singapore, 2024. Accessed: 2024-11-17.