

Faculty of Natural and
Mathematical Sciences
Department of Information

King's College London
Strand Campus, London,
United Kingdom



7CCSMPRJ

Individual Project Submission 2025/04

Name: Xiang Wang
Student Number: K24020132
Degree Programme: Data Science MSC
Project Title: Data Pipeline & Decision Support Tool for Demand
Prediction in Bike Sharing
Supervisor: Dimitrios Letsios
Word Count: Word count goes here

RELEASE OF PROJECT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

- ☒ I agree to the release of my project
☐ I do not agree to the release of my project

Signature: *Signature* **Date:** May 9, 2025



Department of **Information**
King's College London
United Kingdom

7CCSMPRJ Individual Project

Data Pipeline & Decision Support Tool for Demand Prediction in Bike Sharing

Name: **Xiang Wang**
Student Number: **K24020132**
Course: **Data Science MSC**

Supervisor: Dimitrios Letsios

This dissertation is submitted for the degree of MSc in **Data Science MSC**.

Contents

1	Introduction	1
2	Background Material	2
2.1	Bike Sharing and Demand Prediction	2
2.2	Existing Approaches	2
2.3	Challenges Identified in Literature	2
2.4	My Method: Initial Station-Specific Modelling and System Architecture .	3
2.5	Scalability and Generalization to New Stations	3
3	Project Schedule	4
3.1	Timeline of Work (January–April 2025)	4
3.2	Planned Work (May–June 2025)	5
3.3	Gantt Chart: Project Timeline (Jan–Jun 2025)	6
3.4	Current Limitations and Future Work	6
	References	7
A	EDA Plots	9
B	User Interface Overview	9

List of Figures

1	Two example EDA plots showing variable relationships.	4
2	EDA Result 3	8
3	EDA Result 4	8
4	EDA Result 5	8
5	EDA Result 6	8
6	EDA Result 7	8
7	EDA Result 8	8
8	EDA Result 9	8
9	EDA Result 10	8
10	Failure Case	9
11	Success Case	9

List of Tables

1 Introduction

Bike sharing systems have become a core component of sustainable urban transportation. Their growing popularity brings operational challenges, particularly in ensuring that supply meets demand in different locations and times. Accurate demand prediction is critical to optimize fleet distribution, improve user satisfaction, and support strategic planning.

In this project, I focus on building a data pipeline and decision support tool to forecast demand in London's bike sharing system. I use publicly available datasets from Transport for London (TFL), including detailed trip records and station metadata, combined with historical weather information from the Visual Crossing API. By integrating temporal, spatial, and meteorological features, my goal is to predict both departures and arrivals of bikes at the station level.

At this stage, I have implemented and evaluated machine learning models -XGBoost, Random Forest, and Gradient Boosting - for the prediction of **daily demand**, achieving promising R^2 scores for a substantial number of stations. For stations where the model accuracy is insufficient (e.g. $R^2 < 0.65$), a fallback strategy based on historical averages is applied to ensure minimum prediction reliability.

Currently, I am extending the system to support **hourly-level prediction**, which poses greater challenges due to increased volatility and data sparsity. In parallel, I am also working to improve the underlying models and algorithms, including refining feature engineering, performing hyperparameter tuning, and experimenting with alternative modelling strategies. These ongoing efforts are aimed at enhancing the system's accuracy, robustness, and generalizability over time.

2 Background Material

2.1 Bike Sharing and Demand Prediction

Bike sharing systems (BSS) have become an essential part of modern urban transportation, offering a flexible, affordable, and sustainable alternative to traditional commuting. A major operational challenge in such systems lies in efficiently matching bike supply with demand, both spatially and temporally. Demand prediction serves as a crucial tool to support bike rebalancing, infrastructure planning, and improved user experience.

In BSS, demand can be defined as the number of trips, bike departures, arrivals, or user counts within a given period. Accurate forecasts enable service providers to reduce shortages or overflows at stations, optimize fleet distribution, and enhance operational efficiency.

Importantly, demand prediction also benefits end users. By accessing the predicted usage level of a station, users can assess whether the station is likely to meet their needs and make informed decisions, such as choosing a different station, adjusting departure times, or avoiding peak periods. This promotes smarter and more convenient commuting behavior.

2.2 Existing Approaches

Numerous methods have been proposed to address bike demand forecasting. Traditional time-series models such as ARIMA and exponential smoothing are suitable for data with clear periodic trends, but often lack flexibility in incorporating external variables such as weather or holidays [1].

Machine learning methods including Random Forest (RF), XGBoost, and Support Vector Regression (SVR) have demonstrated strong performance in handling nonlinearities and integrating diverse features such as temporal patterns, weather conditions, and spatial factors [2]. Deep learning models such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) have also been used for fine-grained prediction tasks, especially at the hourly level [3]. However, these models require large amounts of training data and are often less interpretable.

Most of the literature focuses on predicting aggregated demand across cities or major stations. These models offer scalability but often overlook the localized dynamics needed for station-level operational decisions.

2.3 Challenges Identified in Literature

Prior research identifies several recurring challenges:

- **Data sparsity and noise:** Particularly at the hourly and station level, many records consist of zeros or irregular outliers [4].
- **Temporal volatility:** Demand is highly sensitive to weather conditions, public holidays, and unforeseen events, making short-term prediction unstable [5].

- **Scalability:** Training one model per station in large-scale systems can lead to computational complexity and resource consumption [6].
- **Generalization to new stations:** Most predictive models fail to make reliable estimations for newly deployed stations with no historical data [7].

2.4 My Method: Initial Station-Specific Modelling and System Architecture

At the current stage of the project, I have adopted a simple but practical station-specific modelling approach as a starting point. Each station is treated as an independent unit, and a separate regression model is trained using its historical demand data. The system is designed to predict both daily departures and arrivals. Basic models such as XGBoost, Random Forest, and Gradient Boosting are used due to their accessibility and interpretability.

This approach primarily serves as a prototype to support a functional demo. While it enables modular model training and API integration, I acknowledge its limitations in scalability, robustness, and adaptability. The current models are not fully optimized, and some components (e.g., feature selection, hyperparameter tuning) remain relatively simple.

However, the overall system architecture has been designed with future improvements in mind. The data pipeline, station-level modelling structure, and fallback logic have all been implemented, providing a solid foundation for future development. As the project continues, I am learning new modelling techniques and exploring ways to refine and improve the models within the limits of my technical and computational capacity.

2.5 Scalability and Generalization to New Stations

A key design consideration in this project is the ability to handle new or upcoming stations that lack historical demand data. Theoretically, this can be addressed by leveraging the similarity between a new station and existing ones in terms of geographic and functional attributes.

I plan to estimate initial demand for new stations by analyzing:

- Spatial proximity to existing stations;
- The functional nature of the area (e.g., business district, park, residential);
- Urban centrality and traffic context.

Although this component has not yet been integrated into the current demo, the system's modular design makes future implementation straightforward. Preliminary spatial feature extraction has already been tested, and I expect this functionality to be added soon.

This planned capability is consistent with the project's overall direction: to develop a demand prediction framework that is accurate, modular, and adaptable to both existing and future stations.

3 Project Schedule

3.1 Timeline of Work (January–April 2025)

January 2025 – Project Preparation and Data Acquisition

- Reviewed project brief and defined the problem scope;
- Researched related work and decided to adopt a station-specific modelling approach;
- Registered API accounts and downloaded datasets from TfL and Visual Crossing;
- Cleaned and formatted raw datasets (trips, weather, station info).

February 2025 – Exploratory Data Analysis and Feature Engineering

- Conducted EDA to identify correlations between variables (e.g., demand vs. temperature, time of day);
- Visualized daily patterns, seasonal variation, and weather effects;
- Engineered base features: day-of-week, bank holiday, rolling mean, standard deviation;
- Two EDA plots are shown here; more can be found in the appendix.

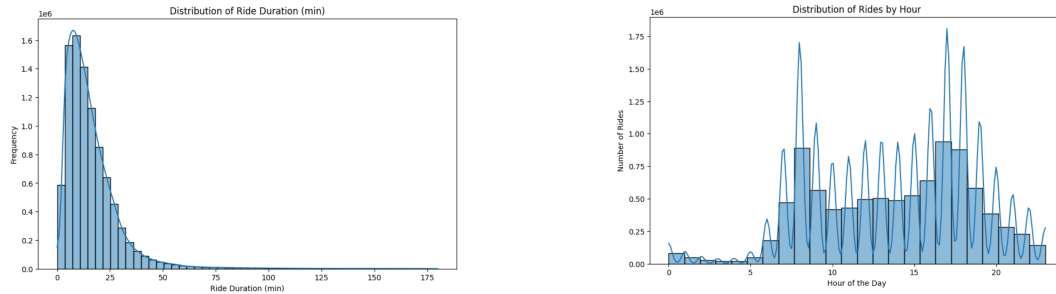


Figure 1: Two example EDA plots showing variable relationships.

March 2025 – Daily Modelling and System Architecture

- Built per-station models using XGBoost, Random Forest, and GBDT;
- Implemented a fallback strategy based on historical averages (R^2 threshold: 0.65);
- Designed and completed the data pipeline to support per-station training and prediction;
- Tested early model performance and started saving best models for deployment.

April 2025 – Demo Construction, Initial Report, and Debugging

- Built a functional demo system using Streamlit;
- Users can select a station, a target date, and demand type (departure/arrival), and view predictions;
- Integrated Folium map to locate stations and show station details;
- Wrote the MSc preliminary project report and organized key sections;
- A map is displayed at the bottom of the interface, which updates dynamically based on the selected prediction location. If the prediction is successful, an animation is triggered and the predicted demand value is clearly presented to the user. Screenshots of the interface can be found in the appendix.
- Due to the use of fuzzy matching during the station name alignment process, there are cases where similarly named but unrelated stations are incorrectly matched. For example, **King’s Cross** has occasionally been mismatched with **Brent Cross**, likely because both names contain the word "Cross" and have similar string patterns. These errors have led to incorrect mapping of trips or metadata between stations. I am currently addressing this issue by improving the matching logic and incorporating additional validation rules such as geographic distance filtering.

3.2 Planned Work (May–June 2025)

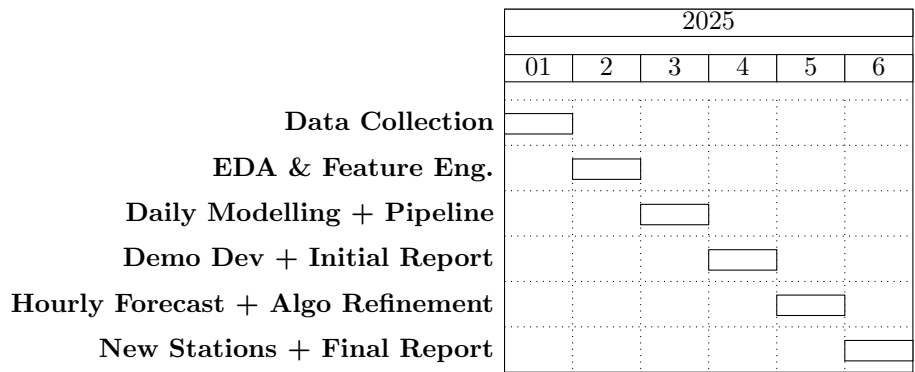
May 2025 – Hourly Prediction and Algorithm Improvement

- Extend current pipeline to support hourly-level demand prediction;
- Design new features for high-frequency data (e.g., lagged hour demand, hourly weather effects);
- Improve algorithms through deeper tuning and cross-validation;
- Experiment with alternative models such as LightGBM or CatBoost for better accuracy and performance;
- Add interpretability tools (e.g., SHAP, feature importance tracking).

June 2025 – New Station Prediction and Finalization

- Implement similarity-based estimation module for new stations with no history;
- Improve UI: add historical demand charts, user guidance, and data export;
- Polish final report with visual results and lessons learned;

3.3 Gantt Chart: Project Timeline (Jan–Jun 2025)



3.4 Current Limitations and Future Work

Although the core demo has been completed, several limitations remain:

- The current models are still at a baseline level with limited tuning;
- The hourly prediction system is in progress and has not been fully validated;
- The interface lacks robust user guidance and deeper visualization;
- The new station generalization module is still under development;
- Demo still has limitations such as lack of response messages, error handling, or latency.

I will continue refining model quality, expanding features, and improving user experience to move toward a scalable and informative decision support system.

References

- [1] H. Lim, K. Chung, and S. Lee, “Probabilistic forecasting for demand of a bike-sharing service using a deep-learning approach,” *Sustainability*, vol. 14, no. 23, 2022.
- [2] H. Mavrodiev, “Bike sharing prediction using random forest and xgboost,” *Kaggle Notebook*, 2020. Machine learning implementation, accessed: 2025-04-25.
- [3] C. Ma and T. Liu, “Demand forecasting of shared bicycles based on combined deep learning models,” *Physica A: Statistical Mechanics and its Applications*, vol. 635, p. 129492, 2024.
- [4] A. Mehdizadeh Dastjerdi and C. Morency, “Bike-sharing demand prediction at community level under covid-19 using deep learning,” *Sensors*, vol. 22, no. 3, 2022.
- [5] D. Gammelli, Y. Wang, D. Prak, F. Rodrigues, S. Minner, and F. C. Pereira, “Predictive and prescriptive performance of bike-sharing demand forecasts for inventory management,” *Transportation Research Part C: Emerging Technologies*, vol. 138, p. 103571, 2022.
- [6] A. Cortez-Ordoñez, P.-P. Vázquez, and J. A. Sanchez-Espigares, “Scalability evaluation of forecasting methods applied to bicycle sharing systems,” *Heliyon*, vol. 9, no. 10, p. e20129, 2023.
- [7] X. Yang and S. He, “Towards dynamic urban bike usage prediction for station network reconfiguration,” *arXiv preprint arXiv:2008.07318*, 2020.

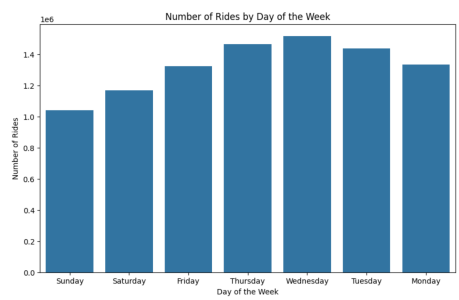


Figure 2: EDA Result 3

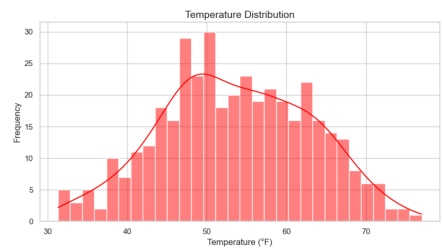


Figure 3: EDA Result 4

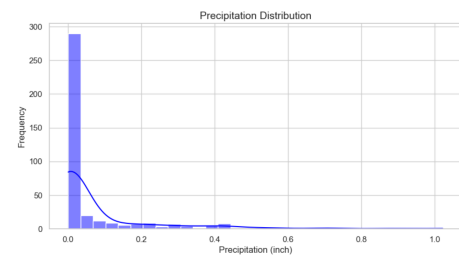


Figure 4: EDA Result 5

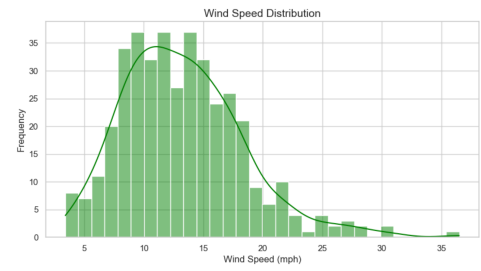


Figure 5: EDA Result 6

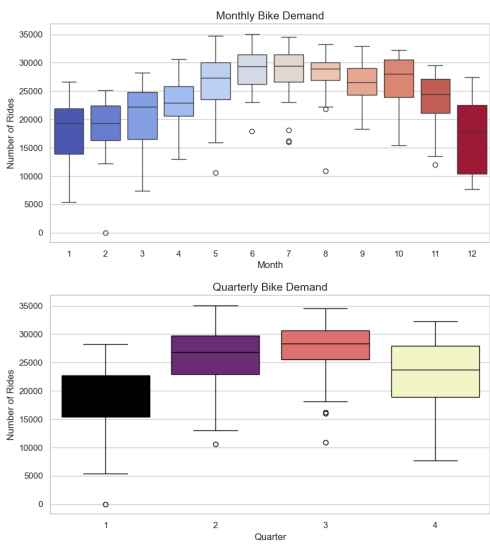


Figure 6: EDA Result 7

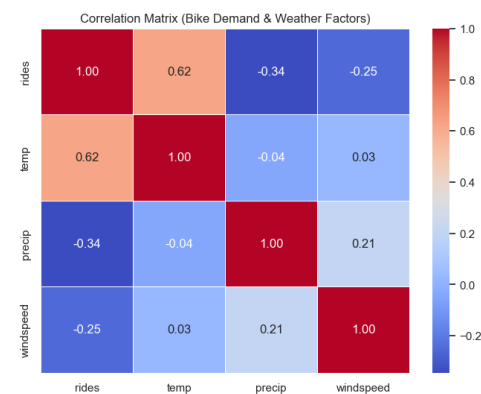


Figure 7: EDA Result 8

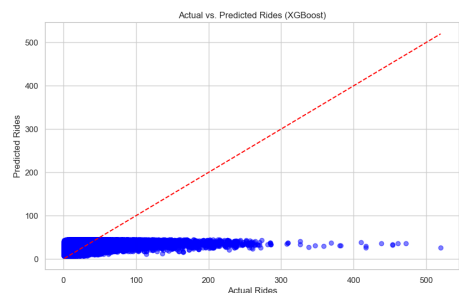


Figure 8: EDA Result 9

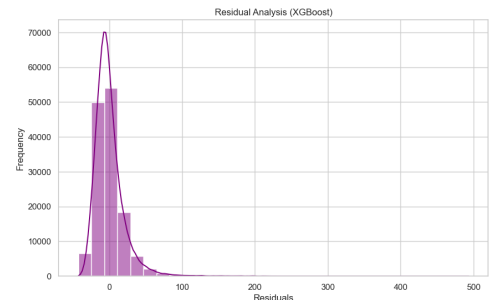



Figure 9: EDA Result 10

A EDA Plots

B User Interface Overview




Bike Demand Prediction App


Select or type a station:
vauxhall walk, vauxhall

Select target date:
2025/04/25

Prediction target:
demand_destination

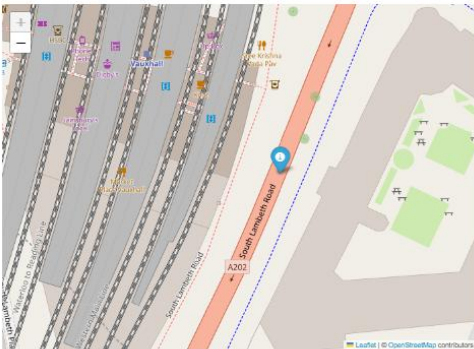
Enter your Visual Crossing API Key:
.....

 Predict Now

 Prediction failed. Possible issue with weather API or model.

Station Location on Map


vauxhall walk, vauxhall location and predicted demand



Map showing the location of Vauxhall station and the predicted demand area. The map includes labels for Vauxhall, Vauxhall Station, and the A203 road. A red line indicates the predicted demand area.

Built by Xiang Wang · Powered by Streamlit

Figure 10: Failure Case




Bike Demand Prediction App

Select or type a station:
aberdeen place, st. john's wood

Select target date:
2025/04/02

Prediction target:
demand_origin

Enter your Visual Crossing API Key:
.....

 Predict Now


 Predicted demand: 21 bikes

Figure 11: Success Case