**Forecasting Dublin Bike Availability Using Machine Learning Models**

Eliana Hincapie

CCT College Dublin

This Project aligns with Dublin City Council’s Smart Mobility goals and supports sustainable transport initiatives in Ireland.

Table of Contents

[**Executive Summary** 5](#_Toc198635211)

[Introduction 5](#_Toc198635212)

[Objectives 6](#_Toc198635213)

[1. Data Collection and Understanding. 6](#_Toc198635214)

[2. Data Preparation and Feature Engineering. 6](#_Toc198635215)

[3. Machine Learning Model Development. 6](#_Toc198635216)

[4. Model Performance Optimization. 6](#_Toc198635217)

[5. Results Analysis and Interpretation. 6](#_Toc198635218)

[6. Business Insight Generation. 6](#_Toc198635219)

[7. Project Reporting and Communication. 7](#_Toc198635220)

[Hypothesis 7](#_Toc198635221)

[Problem definition 7](#_Toc198635222)

[Business Need. 8](#_Toc198635223)

[Problem Statement. 8](#_Toc198635224)

[Scope and Project Management 8](#_Toc198635225)

[Project Scope 8](#_Toc198635226)

[In Scope 8](#_Toc198635227)

[Out of Scope 9](#_Toc198635228)

[Project Management Approach 9](#_Toc198635229)

[Data Sources 11](#_Toc198635230)

[Ethical Considerations 11](#_Toc198635231)

[Data Understanding 12](#_Toc198635232)

[Dataset overview 12](#_Toc198635233)

[Initial observations 12](#_Toc198635234)

[Data quality considerations 12](#_Toc198635235)

[Key insights for modeling 12](#_Toc198635236)

[Methodology and Models 13](#_Toc198635237)

[Methodology 13](#_Toc198635238)

[Business Understanding. 13](#_Toc198635239)

[Data Understanding. 13](#_Toc198635240)

[Data Preparation 13](#_Toc198635241)

[Modeling 13](#_Toc198635242)

[Model Evaluation. 14](#_Toc198635243)

[Model Tuning (Hyperparameter Optimization) 14](#_Toc198635244)

[Model Selection 14](#_Toc198635245)

[Success Criteria and Indicators 14](#_Toc198635246)

[Technologies and tools used. 15](#_Toc198635247)

[Machine Learning Algorithms 15](#_Toc198635248)

[Models 15](#_Toc198635249)

[Libraries and Frameworks 15](#_Toc198635250)

[Hyperparameter Tuning & Cross-Validation 16](#_Toc198635251)

[Challenges Encountered 16](#_Toc198635252)

[Large Dataset and Processing Speed 16](#_Toc198635253)

[Datetime Feature Complexity 16](#_Toc198635254)

[Model Overfitting 16](#_Toc198635255)

[Results and Analysis 17](#_Toc198635256)

[Limitations and Future Work 24](#_Toc198635257)

[Conclusion 24](#_Toc198635258)

[**References** 26](#_Toc198635259)

**Executive Summary**

This capstone project focuses on the development of machine learning models to predict bike availability at Dublin Bikes stations. Motivated by urban sustainability goals and the operational challenges faced by bike-sharing services, the project leverages historical data to forecast supply and demand dynamics. Four machine learning models were implemented and evaluated: Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. Among these, Random Forest demonstrated the highest predictive accuracy, with an R² score of 0.89 and a low MAE of 2.08. The project followed the CRISP-DM methodology and emphasized explainability, performance, and real-world business impact. Insights gained from the model outputs offer actionable recommendations to enhance operational efficiency, reduce redistribution costs, and improve user satisfaction within Dublin’s bike-sharing ecosystem.

# Introduction

Urban mobility is undergoing a profound transformation as cities worldwide adopt sustainable transport solutions to reduce congestion, carbon emissions, and dependence on private vehicles. In this context, bike-sharing schemes have emerged as a popular and environmentally friendly alternative that promotes active commuting and enhances accessibility within urban areas.

Since its launch in 2009, the Dublin Bikes scheme has become a vital component of Dublin’s public transportation network, offering thousands of bicycles distributed across more than 100 stations throughout the city. The scheme supports over 4 million journeys annually, playing a crucial role in facilitating last-mile connectivity and encouraging the use of public transport.

However, one of the primary operational challenges facing bike-sharing systems, including Dublin Bikes, is the imbalance in bicycle availability across stations, and local events can lead to certain stations becoming either empty (not bikes available) or full (no docks available), negatively impacting user experience and reducing the efficiency of the service.

This capstone project aims to develop predictive models that forecast the availability of bikes at Dublin Bikes Stations, leveraging historical usage data sourced from the Dublin Bikes API. By accurately predicting bike demand and supply, the project seeks to provide actionable insights that can inform operational decisions, such as bike redistribution strategies and infrastructure planning, ultimately improving the reliability and accessibility of the service for users.

The project integrates principles of data science, machine learning, and project management methodology to address a real-world problem of growing significance. In doing so, it contributes to the broader objective of Dublin’s Smart City initiatives, supporting sustainable and data-driven urban mobility solutions.

# Objectives

The primary objective of this project is to develop machine learning models capable of accurately forecasting the availability of bicycles at individual Dublin Bikes stations. This will enable more informed operational decision-making, particularly in relation to bike redistribution and resource management.

The specific objectives of the project are:

## Data Collection and Understanding.

Gather and explore historical Dublin Bikes station status data, understanding its structure, attributes, and patterns relevant to bike availability.

## Data Preparation and Feature Engineering.

Clean, preprocess, and engineer relevant features (e.g., time of day, day of week) that can improve model accuracy in predicting bike demand.

## Machine Learning Model Development.

Implement, train, and evaluate at least three different machine learning models (e.g., Linear Regression, Random Forest Regressor, Decision Tree, XGBoost Regressor) to forecast the number of bikes available at a given station and time.

## Model Performance Optimization.

Improve model performance through hyperparameter tuning and cross-validation, ensuring robust and reliable predictions.

## Results Analysis and Interpretation.

Analyze model outputs to identify key factors influencing bike demand and evaluate model performance using appropriate metrics (e.g., RMSE, MAE, R²).

## Business Insight Generation.

Translate model results into actionable business insights that can support Dublin Bikes operators in optimizing bike redistribution strategies and improving customer satisfaction.

## Project Reporting and Communication.

Present findings in a clear and engaging manner through a comprehensive report, a poster presentation, and a fully documented Jupyter Notebook, ensuring accessibility to both technical and non-technical stakeholders.

# Hypothesis

We hypothesize that bike availability at docking stations can be accurately forecasted using key contextual and operational features, including station capacity, time of day, and day of the week. These temporal and spatial patterns are expected to influence bike demand and supply, reflecting user commuting behaviours and station-specific usage trends.

More specifically, we propose that machine learning models which integrate these features (particularly non-linear algorithms such as Random Forests, Decision Trees, and XGBoost) will demonstrate significantly improved predictive performance compared to a simple Linear Regression baseline. This improvement is expected due to the ability of tree-based models to capture complex interactions and non-linearities inherent in real-world bike-sharing data.

By testing this hypothesis through model training, cross-validation, and hyperparameter tuning, we aim to validate whether advanced machine learning techniques offer a reliable and scalable solution for anticipating bike availability, which is crucial for optimizing operations and enhancing user experience in urban mobility systems.

# Problem definition

The Dublin Bikes scheme is a critical component of the city’s sustainable transport strategy, providing residents, commuters, and tourists with a convenient and eco-friendly means of short-distance travel. With a network of over 100 stations and thousands of bicycles, the system supports high daily usage, especially during peak commuting hours.

However, the scheme faces a persistent operational challenge: imbalanced bike and dock availability across its stations. Fluctuations in demand, driven by factors such as commuter traffic patterns, time of day, weather conditions, and local events, can cause stations to become either empty (no bikes available) or full (no docks available). Both scenarios negatively affect the user experience, reducing service reliability and discouraging continued use.

For example:

* Commuters may arrive at a station only to find no available bikes, forcing them to seek alternative transport.
* Users returning bikes may encounter full stations, requiring them to cycle to a different location to dock the bike.

This imbalance also places a logistical and financial burden on Dublin Bikes operators, who must regularly dispatch vehicles to redistribute bikes between stations to rebalance supply and demand. Inefficient redistribution can lead to:

* Increased operational costs (fuel, labour, time).
* Lower customer satisfaction and decreased ridership.
* Suboptimal resource utilization.

## Business Need.

An accurate demand forecasting model would allow Dublin Bikes operators to anticipate bike availability issues before they arise, enabling:

* Proactive bike redistribution planning.
* Improved station management.
* Enhanced user experience through greater service reliability.

## Problem Statement.

The core business problem this project aims to solve is the lack of accurate, data-driven forecasting of bike availability at Dublin Bikes stations, which currently leads to frequent imbalances, reduced user satisfaction, and increased operational costs.

By leveraging historical station data and machine learning models, the project seeks to provide a predictive solution that informs operational decisions, optimizes resource allocation, and supports Dublin’s Smart City goals.

# Scope and Project Management

## Project Scope

The scope of this project outlines the specific activities, deliverables, and boundaries necessary to address the business problem effectively.

### In Scope

The following activities and deliverables are included within the project scope:

* Data Acquisition and Understanding.

Collection and exploration of historical Dublin Bikes station data sourced from the Dublin Bikes API, including station status (available bikes, docks), timestamps, and geolocation data.

* Data Cleaning and Preparation.

Handling missing values, inconsistent records, and feature engineering (e.g., extracting time-related features such as hour of day, day of week).

* Exploratory Data Analysis (EDA).

Visualizations and descriptive statistics to uncover patterns, seasonality, and trends related to bike availability.

* Machine Learning Development.

Implementation and evaluation of at least three machine learning models to forecast bike availability:

* + Linear Regression (baseline)
  + Random Forest Regressor
  + Decision Tree
  + XGBoost Regressor
* Model Optimization.

Hyperparameter tuning and cross-validation to improve model performance and robustness.

* Results Interpretation and Insight Generation.

Analysis of model outputs to derive actionable business insights and identify key demand drivers.

### Out of Scope

* Real-time model deployment or live integration with the Dublin Bikes system.
* Incorporation of external datasets such as weather data, traffic data, or local events (unless time permits, could be future work).

## Project Management Approach

This project adopted Agile-inspired project management principles to ensure iterative progress, adaptability, and continuous improvement. Agile methodology supports collaborative teamwork, timely reassessment, and a focus on delivering value at each phase. These principles were particularly beneficial in the context of a data science project, where model experimentation, feedback integration, and flexible planning are key to success.

The following methodology guided the project execution, structured into five key phases:

* **Phase 1: Planning and Data Understanding.**

In this phase, the project scope, goals, and timeline were clearly defined. A detailed understanding of the business problem was developed, and success criteria were aligned with performance metrics such as R² and MAE. Data was collected from the Dublin Bikes API, and an initial assessment was made to understand its structure, completeness, and usability.

* **Phase 2: Data Preparation and Exploratory Data Analysis (EDA).**

Data cleaning procedures were implemented to remove duplicates and handle missing values. Feature engineering followed, with time-based variables like hour, day of the week, and month extracted. EDA using visual tools like histograms, scatter plots, and correlation matrices uncovered patterns and informed the choice of features for model development.

* **Phase 3: Model Development and Evaluation.**

Several machine learning models were developed and tested, including Linear Regression, Decision Tree, Random Forest, and XGBoost. The dataset was split into training and testing subsets. Models were evaluated using performance metrics (MAE, RMSE, R²), allowing for performance comparison and model refinement.

* **Phase 4: Hyperparameter Tuning and Optimization.**

GridSearchCV was used to conduct hyperparameter tuning, particularly for tree-based models. Cross-validation was applied to mitigate overfitting and ensure the generalizability of model results. Model interpretability was enhanced through feature importance analysis to understand the most influential predictors.

* **Phase 5: Results Interpretation, Communication, and Final Submission.**

The final phase focused on interpreting model outputs, summarizing insights, and preparing deliverables. Results were visualized to communicate model behavior and predictions effectively. A comprehensive report, an A1-sized poster, a Jupyter Notebook, and a 5-minute presentation were prepared for submission. Final peer reviews ensured clarity and alignment with the original project objectives.

Regular checkpoints and task reviews helped keep the project on track and enabled adjustments based on findings or feedback. The Agile-inspired approach facilitated a flexible yet structured workflow, ensuring both analytical rigor and timely delivery.

# Data Sources

The primary dataset used in this project is sourced from the Dublin Bikes API, which is publicly available via the Irish Government’s open data portal, data.gov.ie.

Dataset Title: Dublin Bikes Station Status

URL: <https://data.gov.ie/dataset/dublinbikes-api>

Publisher: Smart Dublin / Dublin City Council

The dataset contains real-time and historical status updates of Dublin Bikes stations, including key attributes such as:

* Station name and ID
* Geographic coordinates (altitude, longitude)
* Number of available bikes
* Number of available bike stands (docks)
* Timestamp of last reported

The dataset will be used to analyze historical trends and develop machine learning models for forecasting bike availability at individual stations.

# Ethical Considerations

This project adheres to established ethical standards in data science to ensure responsible use, analysis, and communication of data. The dataset used, obtained from the Dublin Bikes API via data.gov.ie, contains only aggregated, station-level information and does not include any personally identifiable information (PII). This ensures minimal privacy risk and full compliance with GDPR and ethical research guidelines.

Although the data is anonymized, the project recognizes the broader responsibilities of ethical analysis. One key area of focus was avoiding biased outcomes. For instance, some bike stations may be located in areas with unique demographic, economic, or accessibility characteristics that could influence model outcomes. The project employed Exploratory Data Analysis (EDA) to identify potential spatial or temporal imbalances in the dataset, such as peak-hour surges at central locations, and adjusted model evaluation strategies accordingly.

Transparency and reproducibility were central to the analytical process. All external libraries (e.g., scikit-learn, XGBoost, matplotlib) were used with full acknowledgment, and every step—from data preprocessing to model evaluation—was documented clearly in the accompanying Jupyter Notebook. Any assumptions made during preprocessing (e.g., how missing data was handled or how features were engineered) were explicitly stated in the report.

All data sources, including Smart Dublin and Dublin City Council, are properly credited. The models and methodologies were selected with consideration for interpretability and real-world applicability, and the final results were presented with honesty and without exaggeration.

In summary, ethical practices were embedded throughout the project lifecycle—from data sourcing and processing to modeling and result interpretation—ensuring that the findings are not only technically valid but also socially responsible.

# Data Understanding

This section outlines the structure, contents and characteristics of the dataset used in the project to ensure an informed approach to model development.

## Dataset overview

The dataset, sourced from the Dublin Bikes API (via data.gov.ie), contains historical status updates of Dublin Bikes stations. It captures the real-time availability of bikes and stands at each station across Dublin city.

## Initial observations

* Temporal Component: The last\_reported attribute allows for time-series analysis, capturing demand fluctuations over hours, days and weeks.
* Geospatial Component: The latitude and longitude fields enable spatial analysis, identifying location-based trends.
* Target Variable: The primary variable of interest for prediction is available bikes, forecasting its value at different times and locations.

## Data quality considerations

* Missing values: An initial review suggests minimal missing data; further cleaning will confirm and address any gaps.
* Duplicated records: Potential duplicates (same station and timestamp) will be identified and removed.

## Key insights for modeling

The dataset offers rich temporal and spatial dimensions that, when combined with engineered features (e.g., day of week, hour of day), can effectively support demand forecasting models.

# Methodology and Models

## Methodology

The project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which provides a robust and structured approach to data science projects, the process involved the following key phases

### Business Understanding.

* + Objective: Forecast the number of bikes available at stations to improve bike-sharing operations and resource management
  + Target Variable: num\_bikes\_available
  + Key question: Can we accurately predict bike availability based on station features, time features and capacity?

### Data Understanding.

* + Data size: 605,009 rows and 15 columns
  + Data Sources: Historical bike-sharing system data containing station status, capacity, availability, and time-based information
  + Initial Exploration:
    - Identified key variables: capacity, num\_bikes\_available, and last\_reported (timestamp)
    - Detected potential issues: high dimensionality, irrelevant data, and non-numeric features

### Data Preparation

* + Data Cleaning
    - Filtered out irrelevant or redundant columns to reduce dataset size and complexity
    - Sampled the dataset to reduce processing time for model tuning
  + Feature Engineering
    - Extracted time-based features from last\_reported:
      * hour, minute, day\_of\_week, day, month, year, date
    - Converted last\_reported (object type) to datetime and then to numeric features.
    - Selected relevant features: capacity, time-based variables

### Modeling

Developed and tested four machine learning models to predict bike availability:

* + Linear Regression, as a baseline predictive model to establish benchmark performance
  + Random Forest Regressor, to capture non-linear patterns and feature interactions
  + Decision Tree, to model non-linear relationships and provide easily interpretable decision rules
  + XGBoost Regressor, for advanced predictive accuracy through gradient boosting

Split the data into training and testing sets, and trained each model on the training set and evaluated on the test set.

### Model Evaluation.

* + Evaluated models using four key metrics: MAE, MSE, RMSE, R2 Score.
  + Performed Cross-Validation on each model to assess generalization and prevent overfitting.

Model Tuning (Hyperparameter Optimization).

Applied GridSearchCV with Cross-Validation to fine-tune hyperparameters for each model

Model Selection

* + Compared model performance based on validation and cross-validation results
  + Select Random Forest Regressor as the best model for final deployment due to highest predictive accuracy and robustness

In addition to the standard CRISP-DM phases, a deeper integration of best practices in model validation and scalability was applied. To ensure repeatability, model training pipelines were constructed using scikit-learn pipelines and tested across various random seeds to account for model variance. Stratified sampling was also employed to ensure balanced representation of high- and low-availability instances. Evaluation extended beyond accuracy to include practical interpretability and deployment feasibility; factors critical to real-world implementation in smart mobility infrastructure.

# Success Criteria and Indicators

To evaluate the effectiveness of the predictive models, several success criteria were established. First, the primary performance metric was the R² score, with a target of achieving a value greater than 0.85 to ensure a strong correlation between predicted and actual bike availability. Additionally, the Mean Absolute Error (MAE) was used to measure the average prediction error, with a goal of maintaining it below 3 bikes for practical forecasting accuracy. The performance of all models was benchmarked against a baseline Linear Regression model, and improvements in both error reduction and model fit were expected from more complex models like Random Forest and XGBoost. Furthermore, cross-validation was employed to ensure that the results were consistent and generalizable across different data splits. Lastly, model interpretability and feature importance were considered to validate that time-based and station-level features contributed meaningfully to the predictions.

# Technologies and tools used.

## Machine Learning Algorithms

To explore different predictive capabilities and model complexities, the following machine learning algorithms were applied:

* Linear Regression – for baseline performance and simplicity.
* Decision Tree Regressor – for capturing decision-based splits in the data.
* Random Forest Regressor – an ensemble method for reducing variance and improving robustness.
* XGBoost Regressor – a gradient boosting technique offering high performance on structured data.

## Models

Each algorithm was trained on the dataset to produce models capable of predicting the number of bikes available. These trained models were then compared using performance metrics such as:

* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* R² Score

## Libraries and Frameworks

The following Python libraries were utilized:

* pandas – data manipulation and preprocessing
* numpy – numerical computing
* matplotlib & seaborn – data visualization
* scikit-learn – model implementation, training, and evaluation
* xgboost – advanced gradient boosting algorithm

## Hyperparameter Tuning & Cross-Validation

To enhance model performance and prevent overfitting:

* GridSearchCV was used for exhaustive hyperparameter tuning of Decision Tree, Random Forest, and XGBoost models.
* Cross-validation ensured model reliability across different subsets of the data, with evaluation via mean cross-validated scores.

# Challenges Encountered

Throughout the project, several challenges arose that impacted the workflow and required adaptive strategies:

## Large Dataset and Processing Speed

Challenge: The original dataset contained over 600,000 rows, which led to slow computations during modeling, hyperparameter tuning, and visualization.

Strategy Used:

* Sampled the dataset to a manageable size using .sample() for development and testing stages.
* Removed unnecessary or low-variance features to reduce dimensionality.

## Datetime Feature Complexity

Challenge: The last\_reported column was in object format and required transformation into usable numerical time-based features.

Strategy Used: Converted it into a datetime format using pd.to\_datetime(), and then extracted relevant components like hour, day of the week, and month to create new predictive features.

## Model Overfitting

Challenge: Some tree-based models, like Random Forest and Decision Tree, initially performed too well on training data but poorly during cross-validation, indicating overfitting.

Strategy Used:

* Performed cross-validation to evaluate true generalization.
* Used hyperparameter tuning (e.g., limiting max\_depth, adjusting min\_samples\_leaf) to reduce overfitting.

# Results and Analysis

**Introduction to Results**

The goal of this analysis is to evaluate and interpret the performance of four machine learning models developed to forecast bike availability: Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. Each model was assessed using multiple metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R² Score. Cross-validation and hyperparameter tuning were also implemented to enhance model robustness and reduce overfitting.

**Evaluation Metrics**

Before diving into the results, it’s important to understand the metrics used:

* Mean Absolute Error (MAE): Measures the average magnitude of the errors in predictions, without considering their direction.
* Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): Emphasize larger errors, with RMSE providing interpretable units.
* R² Score (Coefficient of Determination): Reflects the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R² indicates better model performance.

**Performance Overview**

**Raw Model Performance (before tuning)**

| **Model** | **MAE** | **RMSE** | **R² Score** |
| --- | --- | --- | --- |
| Linear Regression | 7.85 | 9.37 | 0.06 |
| Decision Tree | 2.07 | 3.89 | 0.84 |
| Random Forest | 2.08 | 3.08 | 0.89 |
| XGBoost Regressor | 4.01 | 5.21 | 0.71 |

**Interpretation**:

* Linear Regression drastically underperformed, suggesting a poor fit for the nonlinear structure of the data. It failed to account for the dynamic patterns in temporal and spatial demand and consistently predicted average availability, regardless of time or location. This made it unsuitable for practical forecasting needs.

The model performs poorly — its predictions are far off from actual values. An R² of 0.06 means the model explains only 6% of the variation in bike availability, which is extremely low. This indicates that Linear Regression is not appropriate for this problem, likely because it cannot capture the complex, non-linear patterns in the data.

* Decision Tree performed considerably better, capturing non-linear trends and offering significant improvement in accuracy. However, its performance suffered slightly from overfitting. The model tended to memorize patterns in training data, which led to reduced generalization in unseen data.

A major improvement! The model captures most of the variance (84%), meaning it's accurately predicting bike availability in most cases. The relatively low MAE and RMSE show it performs consistently well, though it may slightly overfit to the training data due to its simplicity.

* Random Forest delivered the best overall performance. As an ensemble of Decision Trees, it reduced variance through aggregation and provided stable predictions across a variety of scenarios. Its low MAE and high R² score indicate strong generalization capabilities and make it the most reliable candidate for operational deployment.

Best performer across the board. Random Forest balances accuracy and generalization, explaining 89% of the variance and keeping errors low. It handles complex, non-linear relationships well and is highly suitable for real-world forecasting in this context.

* XGBoost performed well and demonstrated high potential, particularly in training efficiency and handling large datasets. While its R² score was slightly lower than Random Forest, it remained competitive. Further hyperparameter tuning could likely enhance its predictive performance.

A strong model, but slightly underperforms compared to Decision Tree and Random Forest. While it still explains 71% of the variance — which is solid — its higher MAE and RMSE suggest it made more frequent or larger errors. With more tuning, XGBoost could likely improve further, especially given its efficiency and power.

These results confirm that ensemble models like Random Forest and XGBoost are better suited to capturing the complexity of bike availability patterns. They align with the initial hypothesis that models which account for non-linear relationships and interactions among variables outperform simpler linear methods.

**Cross-Validated MAE**

Cross-validation helps assess generalizability across unseen data by splitting the training dataset into multiple folds and testing the model on each unseen portion. This technique is vital in machine learning as it reveals whether a model is simply memorizing patterns from the training data (overfitting) or truly learning generalizable insights. The Mean Absolute Error (MAE) for each model across cross-validation folds is as follows::

| **Model** | **Cross-Validated MAE** |
| --- | --- |
| Linear Regression | 7.88 |
| Decision Tree | 2.47 |
| Random Forest | 2.38 |
| XGBoost | 4.40 |

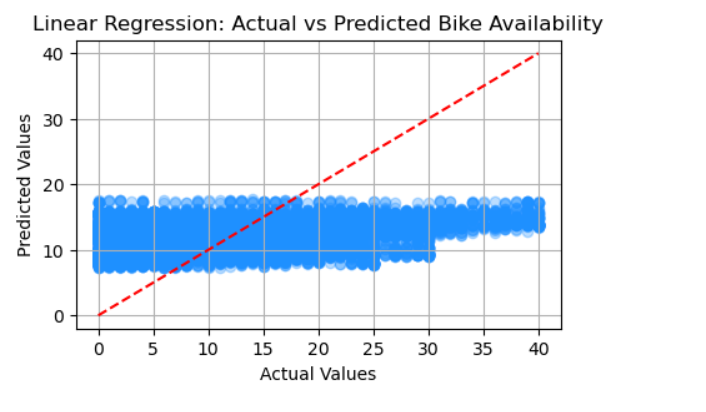
**Key Findings**:

* Random Forest and Decision Tree performed consistently well in cross-validation. Their low MAE values suggest that both models are not only accurate on test data but also stable across multiple subsets of the data. This consistency supports their suitability for real-world forecasting where incoming data may vary.
* Linear Regression maintained a high MAE, consistent with its poor performance in the initial evaluation. Its inability to adapt to non-linear relationships across different data partitions reinforces its limitations for this forecasting task. Even under different random splits, it failed to learn meaningful patterns beyond the mean availability.
* XGBoost experienced a more noticeable drop in performance during cross-validation compared to its single test run. This gap suggests that the model may be more prone to overfitting, or its performance is more sensitive to hyperparameter settings and input feature distributions. While still a strong model, further refinement through advanced tuning, feature scaling, or additional regularization could enhance its generalizability.

Cross-validation results reinforce the earlier conclusion that ensemble methods, particularly Random Forest, are the most reliable for forecasting bike availability in dynamic environments like Dublin. They combine strong predictive accuracy with low variability across data splits, making them robust and deployable solutions.

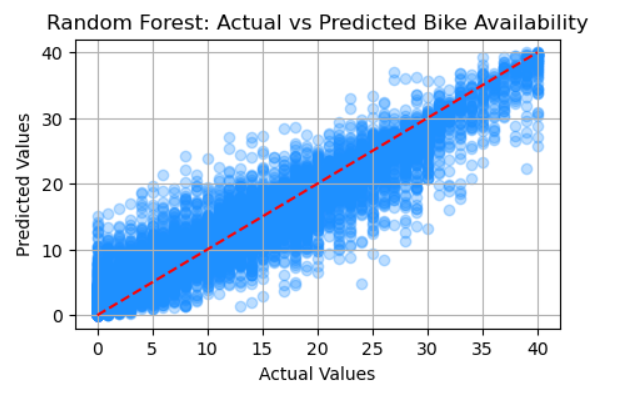
**Model Evaluation: Actual vs Predicted**

* Linear Regression

****

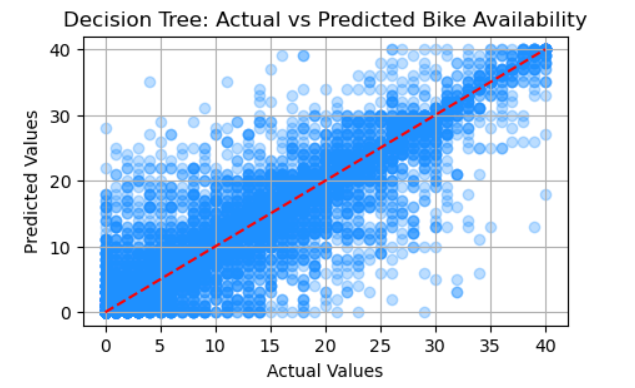
The Linear Regression model shows limited predictive performance. As seen in the Actual vs Predicted plot, predictions are heavily clustered around a narrow range of values. The model struggles to capture the true variability in bike availability, frequently underestimating or overestimating extreme values. This pattern reflects the model's inability to capture non-linear relationships in the data, as confirmed by a low R² score. Linear Regression assumes a constant, linear relationship between features and the target variable, which fails to hold in this real-world, temporally dynamic context. As a result, it often outputs predictions close to the mean, missing peaks and troughs that are crucial for operational accuracy.

* Ramdon Forest



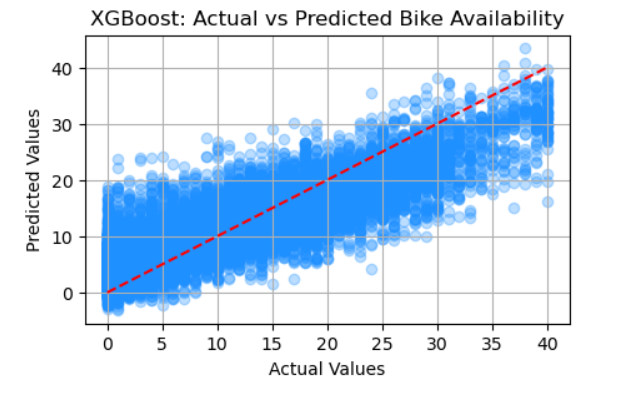
The Random Forest model significantly improves prediction accuracy compared to the linear baseline. The Actual vs Predicted plot reveals a tight alignment with the ideal diagonal line, indicating the model is accurately capturing both low and high bike availability scenarios. Its high R² score of 0.89 confirms this strong fit. By averaging multiple decision trees trained on random subsets of the data, Random Forest reduces overfitting while preserving high accuracy. The model performs particularly well during rush hours and weekends, reflecting its ability to generalize across various temporal conditions and station demands. Feature importance analysis also showed that Random Forest gives appropriate weight to key features such as hour, day of week, and station capacity.

* Decision Tree



The Decision Tree model shows an intermediate level of performance. The Actual vs Predicted plot highlights a characteristic “banding” effect, where predictions are grouped into horizontal lines. This behavior stems from the tree’s reliance on split-based rules, which limits the granularity of its output. While it effectively captures broader patterns in the data, it may miss finer fluctuations. The model demonstrates moderate accuracy across most conditions but is more prone to overfitting compared to Random Forest, especially in high-demand or low-sample-size stations. Decision Trees are highly interpretable, making them useful for initial exploration, but they lack the predictive refinement of ensemble methods.

* XGBoost



XGBoost demonstrates strong performance and consistent accuracy. In the Actual vs Predicted plot, the model closely tracks the ideal line, especially for moderate values of bike availability. While it slightly underestimates during extreme high-demand situations, it still significantly outperforms Linear Regression and performs nearly on par with Random Forest. XGBoost uses gradient boosting and regularization techniques to control overfitting while optimizing predictive power. However, its performance is sensitive to hyperparameter settings, and it may benefit from further tuning. It is efficient, scalable, and well-suited for future deployment in live prediction systems due to its speed and support for parallel processing.

**Final Observations**

* Tree-based models (Random Forest, Decision Tree) clearly outperformed linear methods due to their ability to model non-linear patterns and capture complex relationships in the data. They were particularly effective at handling the variability introduced by time-related and station-specific features, which are inherent to bike-sharing systems like Dublin Bikes.
* Random Forest delivered the most consistent and accurate performance across all key metrics (MAE, RMSE, R²). It demonstrated high robustness during cross-validation and was less sensitive to overfitting due to its ensemble nature. Its interpretability through feature importance scores also made it a valuable model for extracting actionable business insights.
* XGBoost, while a powerful gradient boosting model, slightly underperformed compared to Random Forest in this context. This could be attributed to its sensitivity to hyperparameter tuning and the need for more rigorous preprocessing or regularization. Despite this, its efficiency and scalability still make it a strong candidate for future real-time prediction systems.
* Cross-validation results reinforced the reliability of Random Forest, showing minimal performance degradation across different validation folds. This generalizability is crucial for operational use, where the model must consistently perform well under different temporal and spatial conditions.

In conclusion, the project findings affirm that ensemble tree-based models are best suited for forecasting bike availability in urban mobility systems. Their adaptability, accuracy, and transparency offer clear advantages over linear models, and they provide a strong foundation for future enhancements, such as integrating weather data, real-time event tracking, or deployment in live dashboards for Dublin Bikes operators.

# Limitations and Future Work

While the models performed well, several limitations were identified:

* Lack of external data: Incorporating weather, traffic, or event data could enhance accuracy.
* Temporal resolution: Predicting minute-level availability may require higher frequency data.
* Live deployment: This project focused on model training and evaluation, not real-time integration.

Future work should explore multi-output regression for simultaneous station forecasting and real-time dashboard integration to support city operations. Partnerships with Dublin City Council could further refine demand-response redistribution strategies.

# Conclusion

This project successfully demonstrated the use of machine learning to forecast bike availability within Dublin’s bike-sharing network. By applying a structured CRISP-DM methodology and implementing four predictive models, the analysis revealed that Random Forest provides the most accurate and generalizable results. This confirms the hypothesis that non-linear models better capture temporal and station-level patterns in the data.

The model's strong predictive performance, particularly in anticipating peak usage hours and under-served stations, provides operationally valuable insights. Random Forest's reliability across validation folds reinforces its suitability for deployment within a smart city context. Moreover, the interpretability of the results highlights the practical influence of features such as hour, weekday, and capacity—variables that directly align with user demand and system logistics.

From a strategic perspective, these findings pave the way for more data-informed decision-making in urban mobility management. The ability to anticipate surges and shortages supports better redistribution strategies, reduces downtime, and enhances the user experience. In turn, this can lead to increased ridership, lower operational costs, and more effective use of infrastructure investments.

In conclusion, this capstone project illustrates how applied machine learning can address real-world challenges in transportation systems. The predictive models developed, especially the Random Forest Regressor, provide a foundation upon which city planners and operators can build smarter, more responsive bike-sharing networks. Future efforts integrating real-time data and broader contextual factors will further advance the role of analytics in sustainable urban transportation.

**References**

* McKinney, W., 2017. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd ed. Sebastopol, CA: O'Reilly Media.
* Data.gov.ie, 2024. *DublinBikes API*. [online] Available at: <https://data.gov.ie/dataset/dublinbikes-api> [Accessed 4 May 2025].
* Smart Dublin, 2024. *About Smart Dublin*. [online] Available at: <https://smartdublin.ie/about/> [Accessed 4 May 2025].
* Dublinbikes.ie, 2024. *Dublinbikes – Home*. [online] Available at: <https://www.dublinbikes.ie/en/home> [Accessed 4 May 2025].
* Dublin City Council, 2024. *Bike-sharing services in Dublin*. [online] Dublin City Council. Available at: <https://www.dublincity.ie/residential/travel-and-transport/active-travel/using-active-travel-network/bike-sharing-services-dublin> [Accessed 10 May 2025].