**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

[Introduction 4](#_Toc198069863)

[Objectives 5](#_Toc198069864)

[1. Data Collection and Understanding. 5](#_Toc198069865)

[2. Data Preparation and Feature Engineering. 5](#_Toc198069866)

[3. Machine Learning Model Development. 5](#_Toc198069867)

[4. Model Performance Optimization. 5](#_Toc198069868)

[5. Results Analysis and Interpretation. 5](#_Toc198069869)

[6. Business Insight Generation. 6](#_Toc198069870)

[7. Project Reporting and Communication. 6](#_Toc198069871)

[Hypothesis 6](#_Toc198069872)

[Problem definition 6](#_Toc198069873)

[Business Need. 7](#_Toc198069874)

[Problem Statement. 7](#_Toc198069875)

[Scope and Project Management 7](#_Toc198069876)

[Project Scope 7](#_Toc198069877)

[In Scope 7](#_Toc198069878)

[Out of Scope 8](#_Toc198069879)

[Project Management Approach 8](#_Toc198069880)

[Data Sources 9](#_Toc198069881)

[Ethical Considerations 9](#_Toc198069882)

[Data Understanding 9](#_Toc198069883)

[Dataset overview 10](#_Toc198069884)

[Initial observations 10](#_Toc198069885)

[Data quality considerations 10](#_Toc198069886)

[Key insights for modeling 10](#_Toc198069887)

[Methodology and Models 10](#_Toc198069888)

[Methodology 10](#_Toc198069889)

[Business Understanding. 10](#_Toc198069890)

[Data Understanding. 10](#_Toc198069891)

[Data Preparation 11](#_Toc198069892)

[Modeling 11](#_Toc198069893)

[Model Evaluation. 11](#_Toc198069894)

[Model Tuning (Hyperparameter Optimization) 12](#_Toc198069895)

[Model Selection 12](#_Toc198069896)

[Success Criteria and Indicators 12](#_Toc198069897)

[Technologies and tools used. 12](#_Toc198069898)

[Machine Learning Algorithms 12](#_Toc198069899)

[Models 13](#_Toc198069900)

[Libraries and Frameworks 13](#_Toc198069901)

[Hyperparameter Tuning & Cross-Validation 13](#_Toc198069902)

[Challenges Encountered 13](#_Toc198069903)

[Large Dataset and Processing Speed 13](#_Toc198069904)

[Datetime Feature Complexity 14](#_Toc198069905)

[Model Overfitting 14](#_Toc198069906)

[Results 14](#_Toc198069907)

[**Reference** 15](#_Toc198069908)

# 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 will adopt Agile-inspired project management principles, ensuring iterative progress, flexibility, and continuous improvement. The following methodology will guide project execution:

* Phase 1: Planning and Data Understanding

Define project scope, collect and explore dataset, and establish initial hypotheses and success criteria.

* Phase 2: Data Preparation and EDA

Clean and preprocess data; conduct exploratory analysis to inform model selection.

* Phase 3: Model Development and Evaluation

Implement and compare machine learning models; optimize using hyperparameter tuning and cross-validation.

* Phase 4: Results Interpretation and Reporting

Analyze model results; document insights, write final report, and design poster presentation.

* Phase 5: Final Review and Submission.

Conduct peer review of deliverables, finalise documentation, and ensure submission of all required materials.

# 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 key ethical principles to ensure responsible data use and analysis; the Dublin Bikes dataset contains only aggregated station-level data with no personally identifiable information, minimizing privacy risks.

All sources, including Smart Dublin and Dublin City Council, will be properly credited.

Potential biases (e.g., geographic or temporal imbalances) will be examined through Exploratory Data Analysis. Model performance will be evaluated across diverse contexts to ensure fairness and transparency.

All external sources, tools, and methods will be cited appropriately, and all findings will be reported transparently and honestly.

# 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

# 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.
* Decision Tree performed well but was slightly overfitted in some folds.
* Random Forest outperformed all others in terms of accuracy and generalization.
* XGBoost performed well but slightly under Random Forest, possibly due to needing further tuning.

**Cross-Validated MAE**

Cross-validation helps assess generalizability across unseen data:

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

**Key Takeaways**:

* Random Forest and Decision Tree still performed well under cross-validation, showing they are not highly overfitted.
* Linear Regression’s MAE remained high, confirming its inadequacy for this problem.
* XGBoost showed a greater drop from test to cross-validation MAE, indicating it may need deeper tuning or feature engineering.

**Final Observations**

* Tree-based models (Random Forest, Decision Tree) greatly outperformed linear methods due to the nonlinear nature of the problem.
* Random Forest delivered the best performance across all metrics.
* XGBoost, although powerful, underperformed slightly, likely due to parameter sensitivity and data preprocessing.
* Cross-validation confirmed **Random Forest's generalizability** with minimal performance drop.

Reference

Apellidos, n. s. (Año). Título del artículo. *Título del diario*, Páginas desde - hasta.

Apellidos, n. s. (Año). *Título del libro.* Nombre de la ciudad: Nombre del editor.