Forecasting Dublin Bike Availability Using Machine Learning Models

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

AUTHOR

Eliana Hincapié

RedClick

dublinbikes

01. Introduction

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.

The project integrates principles of:

- Data Science
- Machine Learning
- Project Management Methodology

To address a real-world problem of growing significance.

02. Business Description

Hypothesis.
Bike availability at docking stations can be forecasted using key contextual and operational features.

Objective.
To develop machine learning models capable of accurately forecasting the availability of bicycles.

03. Technologies and tools used

Libraries.
- pandas – data manipulation and preprocessing
- numpy – numerical computing
- matplotlib & seaborn – data visualization
- scikit-learn – model implementation, training, and evaluation

Machine Learning Algorithms.
- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- XGBoost Regressor

Hyperparameter Tuning & Cross-Validation.
- GridSearchCV
- Cross-validation, with evaluation via mean cross-validated scores.

Models.
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 Score

04. Data

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 columns and sampled the dataset.
- Feature Engineering: Extracted time-based features from last_reported: hour, minute, day_of_week, day, month, year, date. Selected relevant features: capacity, time-based variables

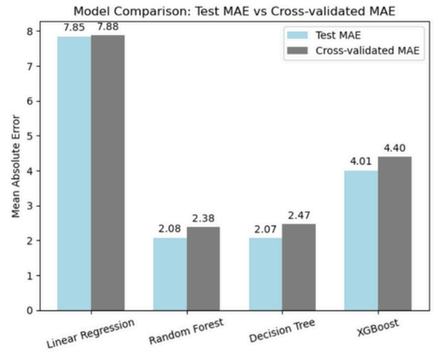
05. Results

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

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 R^2 score indicate strong generalization capabilities and make it the most reliable candidate for operational deployment.

06. Cross-Validated MAE

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



07. Challenges

	Challenge	Strategy used
Large Dataset and Processing Speed	The original dataset contained over 600,000 rows, which led to slow computations during modeling, hyperparameter tuning, and visualization.	-Sampled the dataset to a manageable size using .sample() for development and testing stages. -Removed unnecessary features to reduce dimensionality.
Datetime Feature Complexity	The last_reported column was in object format and required transformation into usable numerical time-based features.	Converted it into a datetime format using pd.to_datetime(), and then extracted relevant components to create new predictive features.
Model Overfitting	Some tree-based models, like Random Forest and Decision Tree, initially performed too well on training data but poorly during cross-validation, indicating overfitting.	-Performed cross-validation to evaluate true generalization. -Used hyperparameter tuning to reduce overfitting.

08. 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.