Saaesha Bhimjiani

Professor Koehler

Data Bootcamp

18 December, 2025

Final Project: Citi Bike Riders' Bike Type Preference

1. Introduction

The purpose of this project is to predict whether a Citi Bike user on a given day will choose to ride an electric or a classic bike. This behavior is important for CitiBike to be able to predict so that they can better manage their fleet operations, such as by gaining a better understanding of where they should install more docks. The dataset used in this project includes Jersey City ridership information from December 2024 to November 2025, which provides the most recent twelve months' worth of information and covers seasonal variation while being a more manageable file size compared to NYC trip data. Because the goal is to predict behavior, only information available from the beginning of the ride was used. Variables considered include starting station, starting hour, weather and precipitation at start time, and member type.

My results show that these features are meaningful predictive indicators. A simple logistic regression improved performance over a baseline classifier, while a random forest and boosted ensemble model performed best overall. An artificial neural network implemented via PyTorch did not do better than a boosted ensemble, but still performed competitively and demonstrated that nonlinear modeling could capture additional structure in the data, even with a simplified feature set.

2. Data Source Overview

The primary data source is Citi Bike trip-level data for Jersey City, downloaded as monthly CSV files. To make the project reproducible for others, I uploaded the CSV files to a GitHub repository and loaded them directly in Colab using the raw GitHub URLs. The raw trip data include identifiers, timestamps, rideable type, membership status, and station/location variables.

Because my goal was to predict bike type using only information available at the start of the ride, I dropped all end-of-trip variables (e.g., end station and end coordinates). This is important because including end-of-trip variables would create leakage.The model would “learn” from information that would not realistically be available at prediction time. I also converted the start timestamp (started_at) into a datetime and constructed additional time features such as date and hour to make modeling easier.

As an outdoor activity, weather likely has a relationship with ridership decisions. Thus, I sourced hourly weather data from Open-Meteo’s archive API for the Jersey City area and chose to include hourly temperature and hourly precipitation in my analysis.

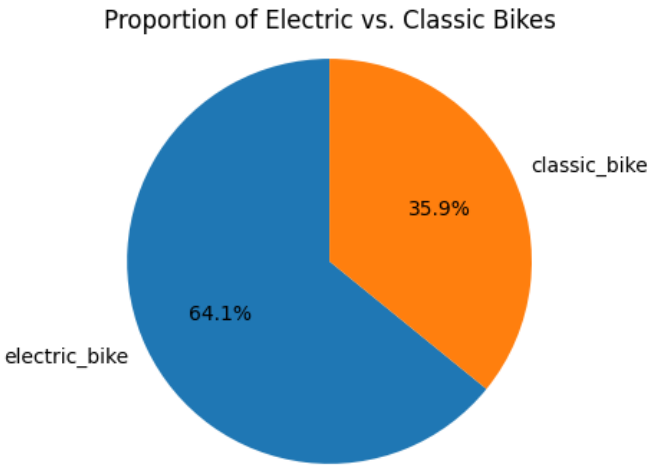
Finally, to create the dataset used in this analysis, I parsed the weather timestamps, created matching merge keys (date and hour_of_day), and performed a many-to-one merge from rides to weather (validate="m:1") to match each ride to the weather at the hour it started. This is consistent with the “start-of-ride information” requirement.

3. Exploratory Data Analysis

The purpose of exploratory data analysis in this project is to assess whether the starting variables demonstrate a structure that could plausibly inform a predictive model of bike type choice. Because electric bikes constitute a majority of rides in Jersey City (64.1%), the goal of this exploration is to 1) understand the class balance, 2) determine which variables to include,

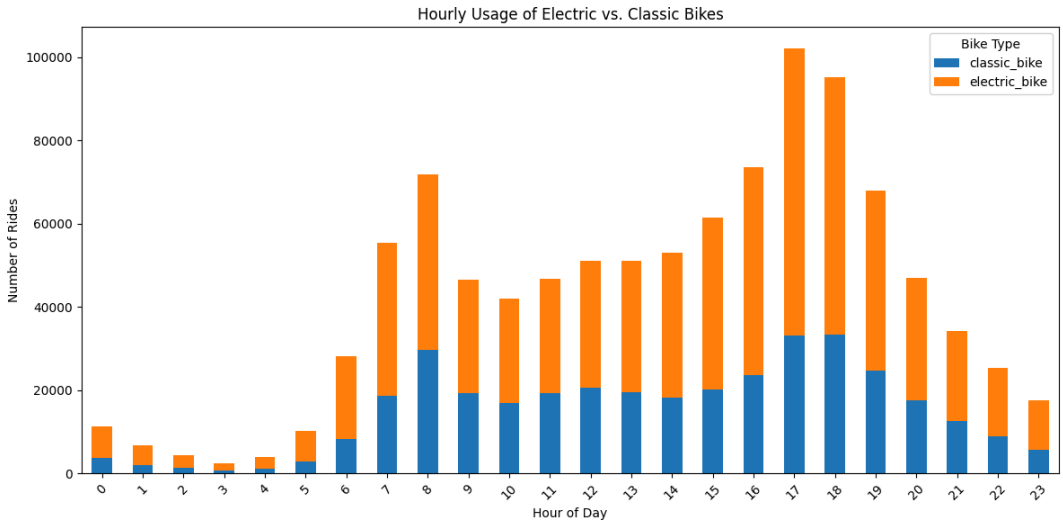
and 3) anticipate whether nonlinear models might benefit our understanding of CitiBike user decisions.

3.1. Bike Type Distribution



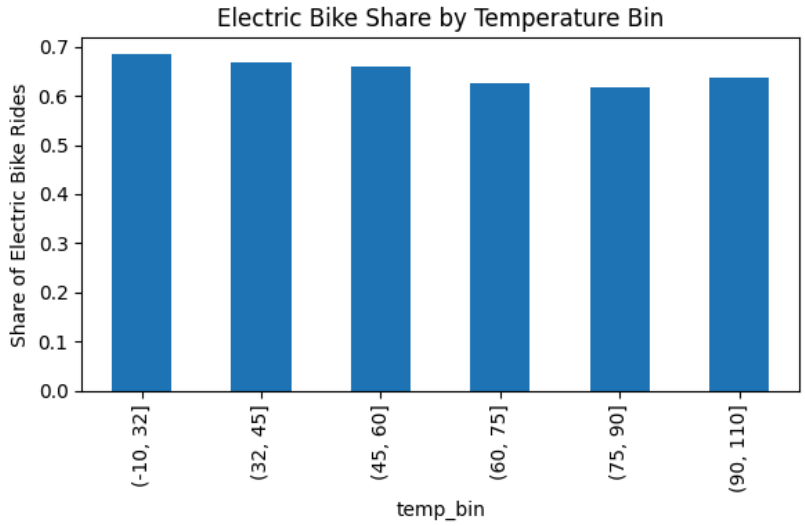
Electric bikes represent a clear majority of rides in the dataset. For modeling purposes, this indicates that a simple classifier could do well if a metric that rewards predicting the majority class is used. For this reason, I will evaluate models using AUC and confusion matrices in addition to summary metrics like F1.

3.2. Time-of-Day Patterns



Ride volume changes substantially through the day. This hourly plot shows clear peaks in the morning and late afternoon/early evening, which is consistent with commuting patterns. Electric bikes dominate at most hours, but the overall volume and the mix of rides still shift across time. This motivates including the hour_of_day variable in the model, because even if time does not perfectly align with bike type, it provides context that can interact with weather and rider type.

3.3. Weather and Bike Choice



Electric bike share changes slightly across temperature bins. Though there is not a significant difference, this pattern suggests that weather contributes incrementally instead of as a strict cutoff. Since weather effects are likely small and may depend on other factors (for example, precipitation could matter more at certain times of day or for certain rider types), this motivates trying nonlinear models later in the analysis. Even modest differences in EDA can translate into meaningful predictive gains when combined with other features.

3.4. Station-level Variation

Electric bike share varies substantially across start stations. Some stations show electric shares above 80%, while the lowest reaches 46%. This indicates that location has a significant relation to the observed bike type. This supports including start_station_id in the feature set. At the same time, station-level differences likely reflect both rider behavior and supply-side conditions (such as where e-bikes are more available). For this reason, station features are best interpreted as relevant to predictive modeling.

3.5. EDA Summary

Overall, EDA suggests that bike type choice is related to time-of-day patterns, weather conditions, and spatial differences across stations. No single factor deterministically predicts bike type, but the structure in these variables supports the use of multivariate models—and provides a rationale for testing nonlinear approaches that can capture interactions among features.

4. Models and Methods

The predictive task is a binary classification problem: the model predicts whether a ride begins on an electric bike (1) or a classic bike (0). The feature set is restricted to variables available at the start of the ride: start station, start coordinates, hour of day, weather (temperature and precipitation), and rider type (member vs. casual). This restriction ensures that the models reflect a realistic forecasting problem rather than relying on information that would not be known when a ride starts. Model performance is evaluated using AUC, F1 scores, and confusion matrices. AUC measures how well the model ranks rides by likelihood of being electric. This is particularly useful because electric rides are the majority class and AUC is less sensitive to the specific classification threshold. F1 scores balance precision and ercall for predicting electric rides at a default threshold, while confusion matrices show the actual counts of correct and incorrect predictions per class, making it easier to see what kinds of errors the model is making.

4.1. Time Based Train/Test Split

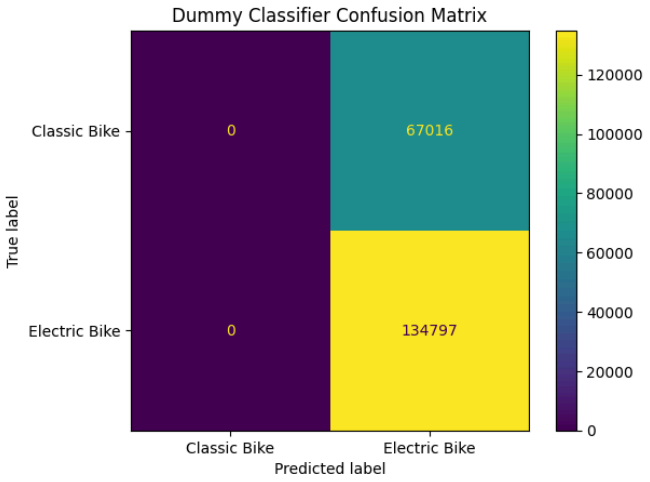
To evaluate the models fairly, I use a time-based split: the first 80% of rides (chronologically) are used for training, and the last 20% are used for testing. This mirrors how the model would be used in practice (training on past behavior and predicting future behavior) and it also helps avoid leakage from time patterns. The electric-bike share is slightly higher in the test set (0.668) than in the training set (0.635), indicating mild drift over time, which makes the test evaluation more realistic.

4.2. Preprocessing

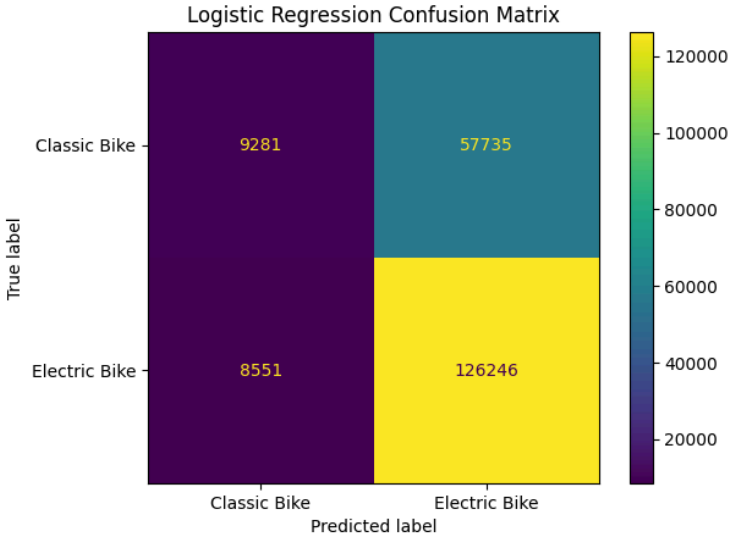
The modeling pipeline uses a ColumnTransformer to apply consistent preprocessing across models. Numeric variables are median-imputed and standardized, and categorical variables are imputed using the most common value and one-hot encoded. One-hot encoding is important for start_station_id because station IDs are categories, not numeric values. This also allows the models to learn station-specific patterns. I used the same preprocessing pipeline for all models to ensure comparisons reflect differences in model structure rather than differences in data preparation.

4.3. Model Sequence and Rationale

A series of models were used on the dataset in order to assess which one worked best at predicting rider decisions. In this section, the use and performance of each model is explained.

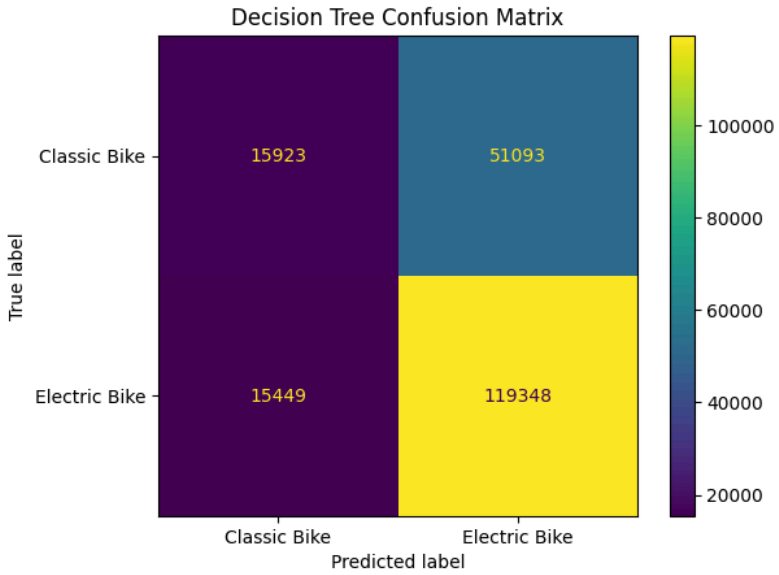


First, I chose to use a dummy classifier as a minimum benchmark with which I would assess the more complex models. The dummy had an AUC of 0.5, indicating no discrimination between classes, and an F1 score of 0.8. The F1 is likely high only because this classifier predicted only electric bikes for a dataset that is majority electric bikes. This demonstrates that for this project, the AUC will be more valuable to assess the quality of a predictive model that goes beyond chance.

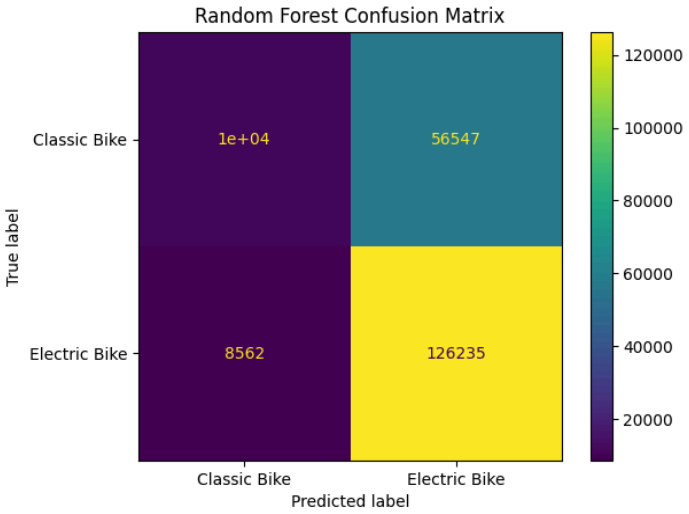


Next, I implemented a logistic regression model to assess whether a linear combination of start-of-ride variables can meaningfully predict bike type. This model improved in AUC, with an

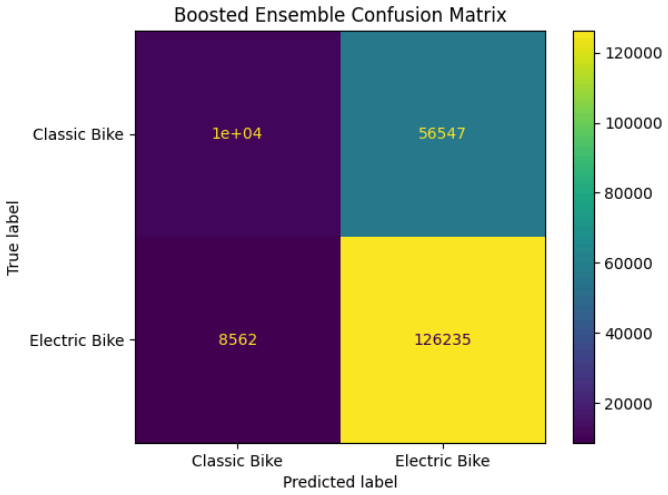
area of 0.649. This model also provides a useful reference point with which to compare nonlinear models; if the latter only slightly improves performance, this would suggest that relationships are mostly linear.



To assess whether there is a nonlinear relationship, a decision tree was introduced. This represents rules such as, “If precipitation is high and the hour is early morning (e..g 8am), predict classic bikes more often.” Unlike logistic regression, a tree can naturally model interactions between variables. The AUC slightly improved to a 0.650, while the F1 decreased further to a 0.782. This is expected as the dummy classifier was bound to overfit given the dominance of electric bikes.

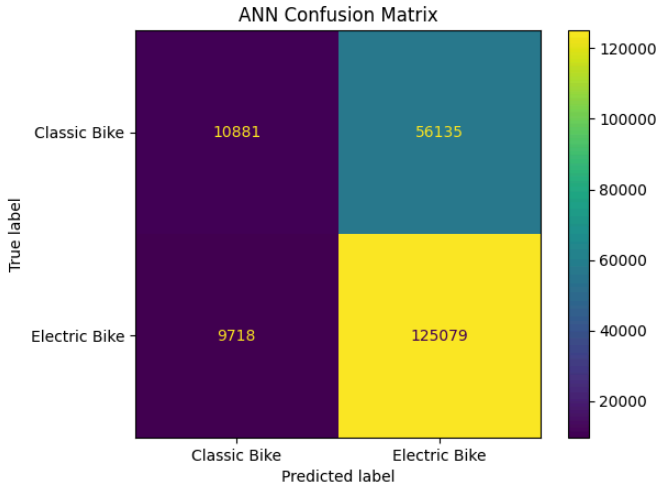


Because decision trees are known to be unstable and can overfit, a random forest was implemented next. This model builds many trees and averages their predictions, which reduces variance and tends to generalize better. It did better than the decision tree on both AUC (0.664) and F1 (0.795). This model preserves the decision tree's ability to model nonlinear interactions while improving reliability and out-of-sample performance.



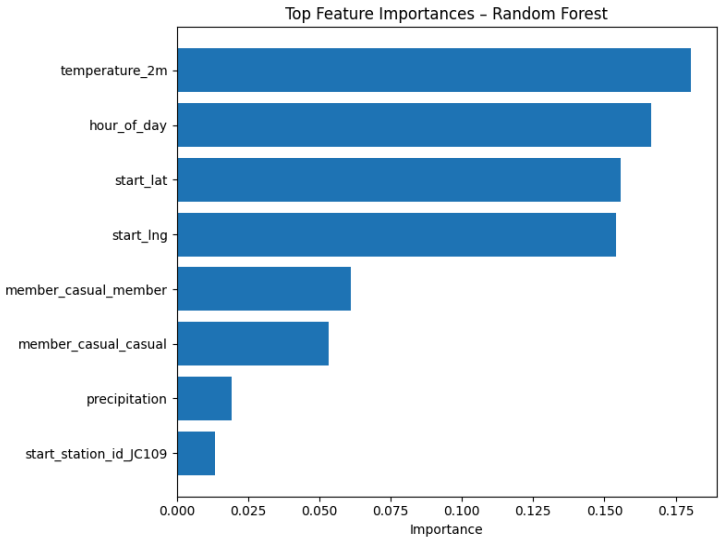
Boosting was added because it often produces stronger discrimination than random forests in problems where many small signals matter. Instead of averaging independent trees, boosting builds trees sequentially, where each new tree focuses on correcting the mistakes of

previous trees. This was thought to be well suited to the Citi Bike dataset because bike choice likely depends on multiple contextual factors that each provide incremental predictive value. The boosted ensemble did, in fact, do better than all of the previous models on AUC, reaching 0.667. However, it did slightly worse than the random forest model on F1, achieving 0.788. This is likely because it was more conservative in predicting electric bikes.



Finally, an artificial neural network was implemented to see if any significant improvements could be gained in the prediction of rider decision making. The neural network achieved an AUC of 0.657 and an F1 score of 0.792, performing competitively despite using a reduced feature set and excluding station ID (one-hot encoded data would have created too much complexity). This result suggests that nonlinear structure exists in time, weather, and geographic variables even without station-level indicators. At the same time, its performance remains below the boosted ensemble and random forest, which is consistent with the fact that tree-based ensembles are often particularly strong for tabular data with mixed feature types. The ANN therefore functions as a successful demonstration of neural network modeling in this context, while reinforcing that ensembles remain the best-performing approach for this problem.

4.4. Feature Importance & Interpretation (Random Forest)



The random forest was chosen for feature importance analysis as the most balanced model for performance on both the F1 and AUC metrics. Feature importance from the random forest indicates that temperature, hour-of-day, rider type, and location all contribute to prediction of bike type usage. Station features such as start latitude and longitude ranked highly, consistent with the EDA showing large station-level differences. However, station importance should not be interpreted as pure preference: it likely reflects where e-bikes are more available, along with spatial and operational patterns in the system. Temperature and hour-of-day proved to be the strongest predictive features, as expected due to seasonal and commuter patterns.

Conclusion & Next Steps

This project shows that bike type choice in Jersey City can be predicted with some meaningful accuracy using only information available at the start of a ride. A baseline model relying on class prevalence fails to discriminate between outcomes, while logistic regression demonstrates that start-of-ride contextual variables contain predictive signals. Nonlinear models provide incremental improvements, with boosted ensemble and random forest methods performing best overall. The boosted model achieves the strongest discrimination (highest

AUC), while the random forest produces the strongest balance at a default threshold (highest F1). A compact neural network performs competitively even without station ID, indicating that nonlinear patterns exist in time, weather, and geography, though tree ensembles remain most effective for this tabular prediction task.

Future work could improve both interpretability and robustness. First, incorporating a proxy for station-level e-bike availability (for example, station-hour electric share in the recent past) would help separate supply constraints from rider preferences. Second, threshold tuning could align predictions with an operational objective, such as improving classic-bike detection. Third, additional temporal features (weekday/weekend and month) could capture seasonality more directly. These steps would strengthen the model's ability to generalize under temporal drift and would better connect predictions to operational decisions.