Date: 12/16/2024 NetID: ryz215

Final Project

Introduction to Problem & Data

Problem Statement:

CitiBike's bike-sharing data offers valuable insights into how users use its bike-sharing service, particularly their choices between classic and electric bikes. The goal of this project is to develop a predictive model that can identify which type of bike a user is likely to choose based on trip characteristics, such as ride duration, ride distance, time of day, and user type (member or non-member). By leveraging these features, this project aims to uncover patterns that influence bike selection and can provide actionable insights for CitiBike to improve its operations. Successfully developing a classification model could enable CitiBike to improve management and operational efficiency. CitiBike can optimize bike distribution and make improvements on its membership offerings. Additionally, understanding these trends can enhance user satisfaction by reducing wait times and better meet customer bike preferences. Furthermore, these insights gained may help inform future strategies for bike-sharing systems in other cities.

Dataset Description:

Data for this project is sourced from CitiBike's publicly available system data, accessed in csv format from their website. It provides comprehensive information about bike-sharing trips, including ride type, ride duration, distance, start and end locations (coordinates), and user type. The dataset will require some cleaning since the original csv contains about 1 million lines and some null values. There may be challenges in building an accurate classification model due to the range of factors not in the dataset influencing bike type selection, such as user preferences and trip characteristics. However, I believe that certain predictive variables provided in the data, like ride duration, distance, member type, and time of day, will help determine the type of bike chosen.

Data Pre-Processing & Preliminary Examination:

1. Take down 10000 random rows to save down as a new csv to work with
2. Create a column for ride duration by subtracting started_at from ended_at (ride_minutes)
3. Create a column for ride time of day by categorizing started_at (tod)
4. Create a column for ride distance in miles using start/end lat/long (ride_miles)
5. Drop columns we will not be using and cap data where ride_minutes is less than 3 hours

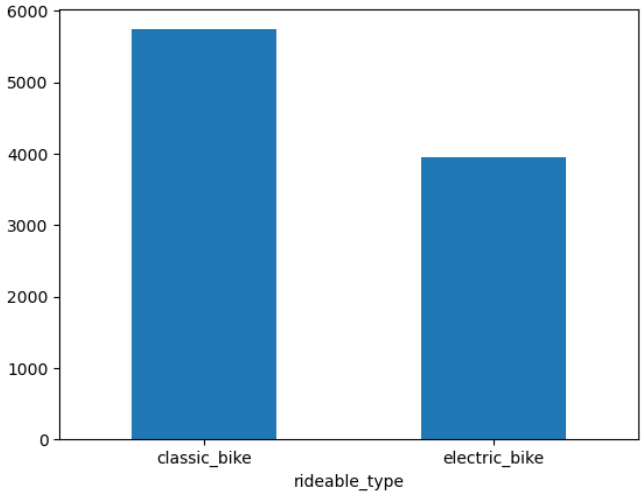
	rideable_type	member_casual	ride_minutes	tod	ride_miles
0	classic_bike	member	7.239250	evening	0.865933
2	classic_bike	member	5.189167	evening	0.619814
3	electric_bike	member	42.894217	afternoon	1.897988
4	electric_bike	member	14.600767	evening	2.014889
5	classic_bike	member	8.864317	afternoon	0.964408

The revised dataset that I'll be working with contains 9691 data points on CitiBike rides in May 2023. The rides range from 1 minute to 174 minutes, and are on average 1.22 miles long.

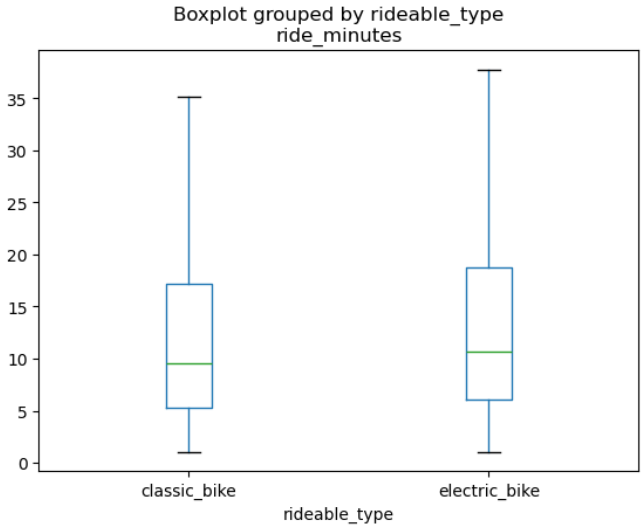
Exploratory Data Analysis

Descriptive Statistics and Initial Visualizations

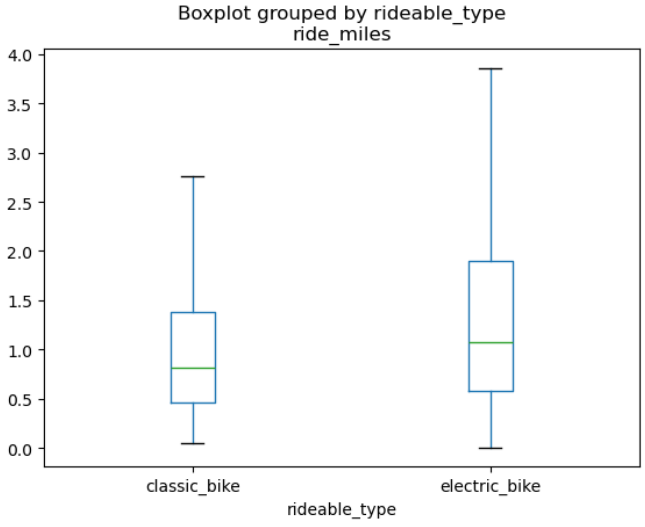
There are more classic bikes in this dataset (even after randomly sampled), potentially because of the higher price of using ebikes.



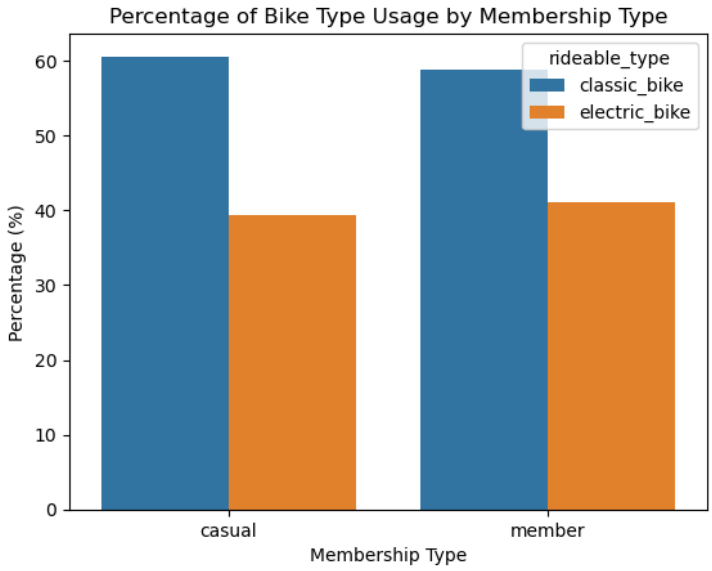
This boxplot shows the distribution of ride minutes grouped by bike type. Both bike types have similar ranges for ride durations, but the median ride time for ebikes is slightly higher than for classic bikes. Additionally, ebikes have a slightly wider IQR for ride times.



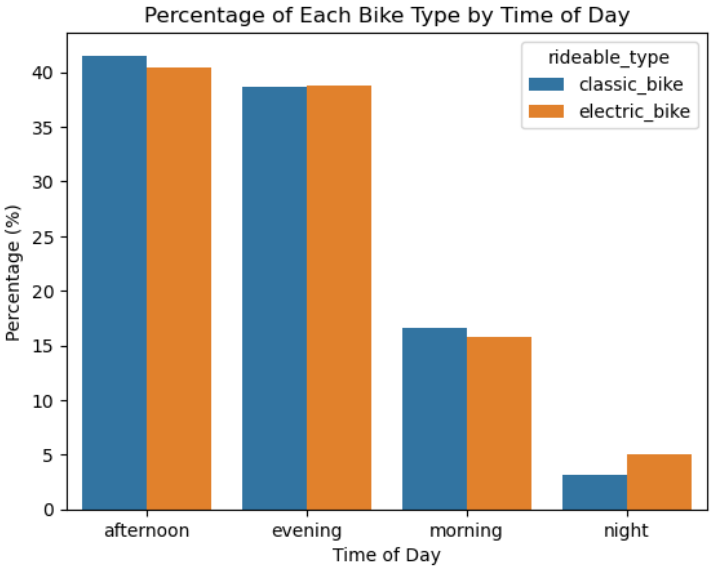
Electric bikes generally have longer rides (both in terms of median and range), which might suggest they are chosen for longer distances. Classic bikes have shorter and more consistent ride distances.



The chart displays the percentage of bike types used within each membership category. Among casual users, 60.6% use classic bikes, while 39.4% prefer ebikes. For members, the distribution is closer, with 58.9% using classic bikes and 41.1% using ebikes. This indicates that while both user groups show a preference for classic bikes, members have a slightly higher tendency to use ebikes compared to casual users.



The distribution between the bike types in the evening is about even. There is slight preference for classic bikes in the morning and afternoons, and vice versa at night. We also see that most rides are taken in the afternoon and evening, while far fewer are in the morning and night.



To summarize, it seems like miles ridden may provide us with the most information on predicting which bike type is chosen. Time of day may also provide some insights, while distance and membership type may provide the least information.

Modeling & Interpretations

To predict type of bike, I am planning on using multiple different classification models to see which performs the best in predicting the bike type chosen. For each of these models, I will use a 80-20 train-test split.

Logistic Regression

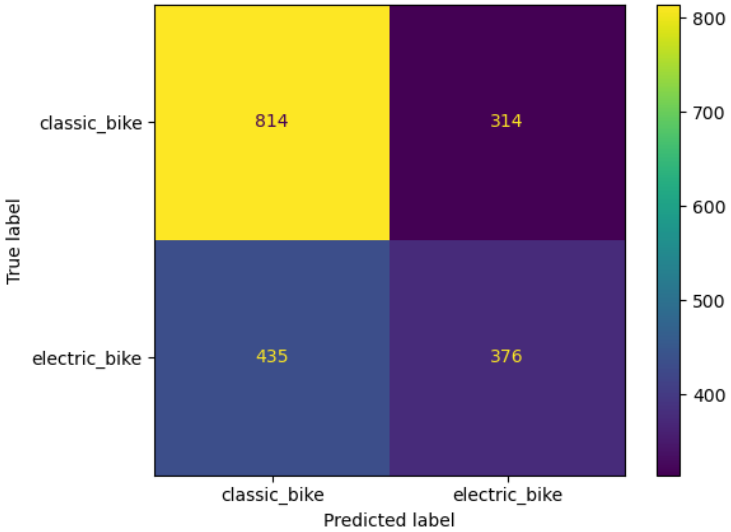
Starting with the Logistic Regression Model, I chose this model as a "baseline" since it is a simple and reliable method for classification problems. It helps show how each feature influences the prediction, making it easy to understand the results. Logistic regression works well when the data is straightforward and there are clear relationships between variables and the target.

```
Train Score: 0.6029411764705882
Test Score: 0.6137184115523465
Precision Score: 0.6070541233907966
Cross-Validation Scores: [0.59445519 0.61637653 0.60709677 0.60064516 0.59612903]
Mean Cross-Validation Accuracy: 0.602940537842391
Classification Report:
              precision    recall  f1-score   support

 classic_bike       0.65      0.72      0.68      1128
electric_bike       0.54      0.46      0.50       811

     accuracy                           0.61      1939
    macro avg       0.60      0.59      0.59      1939
 weighted avg       0.61      0.61      0.61      1939
```

This output shows the performance of a logistic regression model for predicting bike type. The precision and recall results suggest that the model is better at identifying classic bikes compared to electric bikes. The model struggles with correctly classifying electric bikes, as shown by the confusion matrix. There are more electric bikes accidentally classified as classic bikes than those correctly classified, and many fewer vice versa. Overall, the test score is relatively low and we can likely find a more accurate model.



KNearest-Neighbors

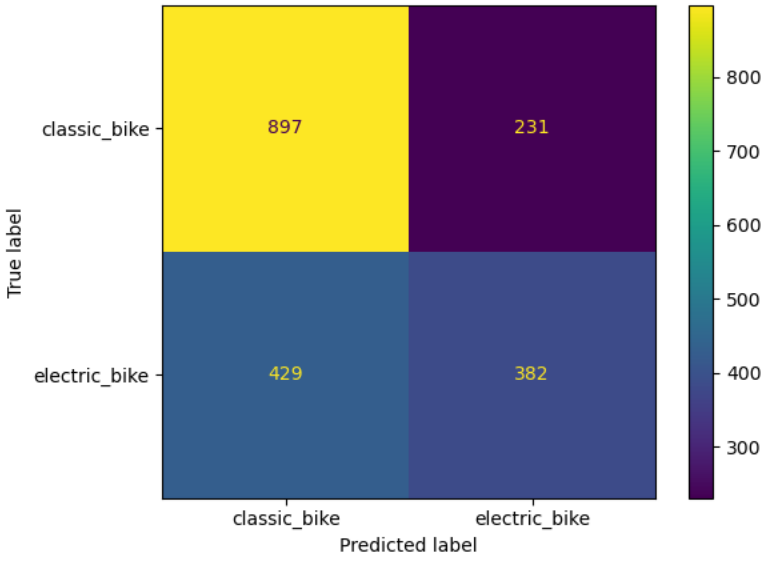
Next I chose to try the KNearest-Neighbors because it is simple and makes predictions based on the similarity between data points. This makes it easy to understand local patterns in the data. KNN works well when the relationship between the variables and the target may be non-linear, which I believe is the case because the relationships and interactions between variables may be more complex. KNN doesn't rely on a fixed model structure, which allows it to adapt to the dataset and identify patterns.

```
Train Score: 1.0
Test Score: 0.6596183599793708
Precision Score: 0.6541750627613074
Cross-Validation Scores: [0.67891683 0.67633785 0.66967742 0.66       0.6683871 ]
Mean Cross-Validation Accuracy: 0.6706638381065286
Classification Report:
              precision    recall  f1-score   support

 classic_bike       0.68      0.80      0.73      1128
electric_bike       0.62      0.47      0.54       811

     accuracy                           0.66      1939
    macro avg       0.65      0.63      0.63      1939
 weighted avg       0.65      0.66      0.65      1939
```

The testing score in my KNN model reflects better performance than logistic regression likely because of multiple factors, such as how it can adapt better to the structure of the dataset, it an capture non-linear patterns, and also how GridSearchCV fine-tunes parameters to find the most accurate model. However, it still seems to misclassify many of the electric bikes as classic bikes, though to a slightly lesser degree. The confusion matrix shows that it correctly classifies more classic bikes than the logistic regression model does.



Decision Tree Model

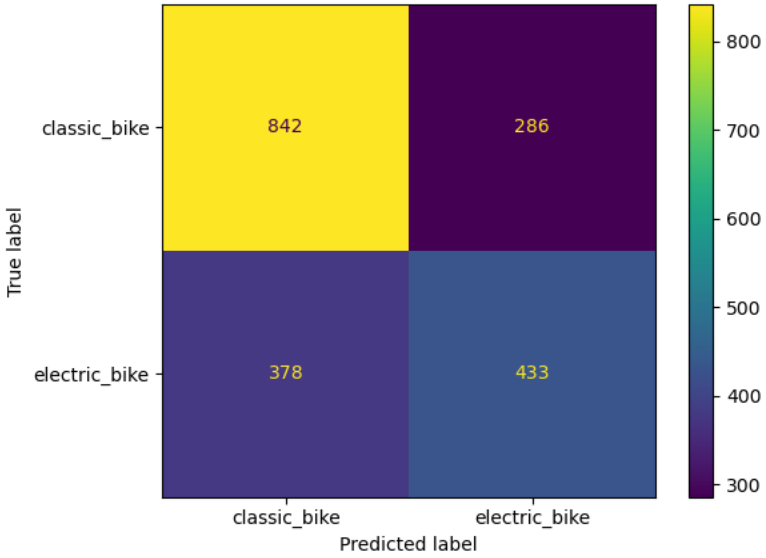
My next model is a Decision Tree Classifier model. Similar to how KNN works, decision trees also capture non-linear relationships. Unlike with KNN, decision trees may be better with unseen relationships since it does not depend on proximity of data points and are better at splitting the data into groups. This may be good for my bike dataset since it provides clear, interpretable rules that show how features influence my target.

```
Train Score: 0.7127192982456141
Test Score: 0.6575554409489428
Precision Score: 0.6533830050658963
Cross-Validation Scores: [0.66795616 0.65764023 0.65548387 0.66967742 0.66645161]
Mean Cross-Validation Accuracy: 0.6634418585303966
Classification Report:
              precision    recall  f1-score   support

 classic_bike       0.69      0.75      0.72      1128
electric_bike       0.60      0.53      0.57       811

     accuracy                           0.66      1939
    macro avg       0.65      0.64      0.64      1939
 weighted avg       0.65      0.66      0.65      1939
```

The Decision Tree model has a higher precision score and mean cross-validation accuracy than both the KNN and Logistic Regression models. At a max_depth of 7 (as found through GridSearchCV), this model more correctly predicts electric bikes correctly. However, it seems to perform worse with correctly predicting classic bikes. This may be because the model focuses on optimizing overall precision and accuracy, which can cause it to focus on classifying electric bikes if it is harder to classify. The tree may be more constrained at a max depth of 7 and it cannot fully split the data into smaller, more specific groups that may accurately predict classic bikes.



Random Forest Classifier Model (best model)

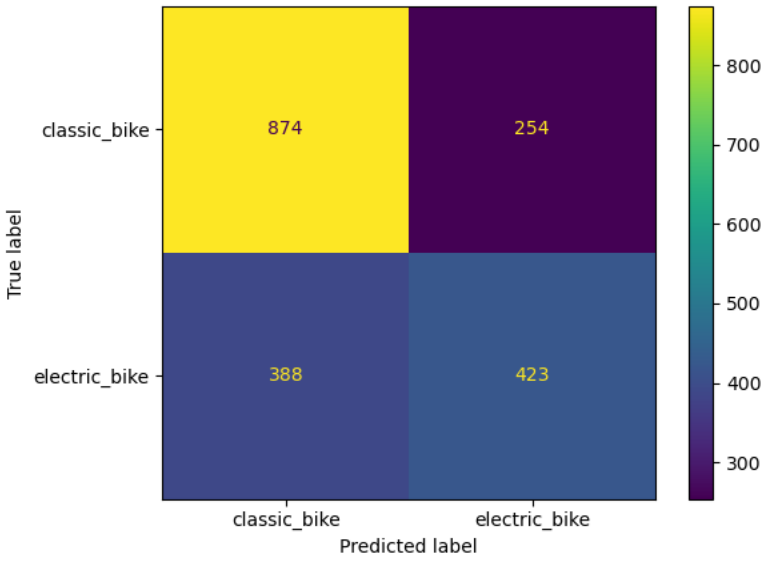
I chose to use a Random Forest Classifier model to potentially handle some class imbalances in my data. Similar to decision trees, random forests can capture non-linear relationships and handle both categorical and numerical data. Unlike a single decision tree, random forests combine multiple trees to reduce overfitting and improve performance on unseen data. This may be good for my bike dataset since it adds robustness to predictions and can capture more complex patterns in the features that may influence bike type chosen.

```
Train Score: 0.7839267285861713
Test Score: 0.6689014956162971
Precision Score: 0.6642204005931347
Cross-Validation Scores: [0.68858801 0.68665377 0.69032258 0.67612903 0.67096774]
Mean Cross-Validation Accuracy: 0.6825322268671618
Classification Report:
              precision    recall  f1-score   support

 classic_bike       0.69      0.77      0.73      1128
electric_bike       0.62      0.52      0.57       811

     accuracy                           0.67      1939
    macro avg       0.66      0.65      0.65      1939
 weighted avg       0.66      0.67      0.66      1939
```

The accuracy seems to be more balanced than that of the Decision Tree - the Random Forest predicts more classic bikes correctly but at the expense of correctly predicting electric bikes. This model also has the highest scores among the ones I ran. While the recall for electric bikes is lower, the precision score is better and more consistent across both classes. Additionally, we find that the most important features are how many miles ridden and how many minutes ridden, while the other features have importances close to 0. This suggests that there is minimal influence from features like time of day or membership type.



LDA Model

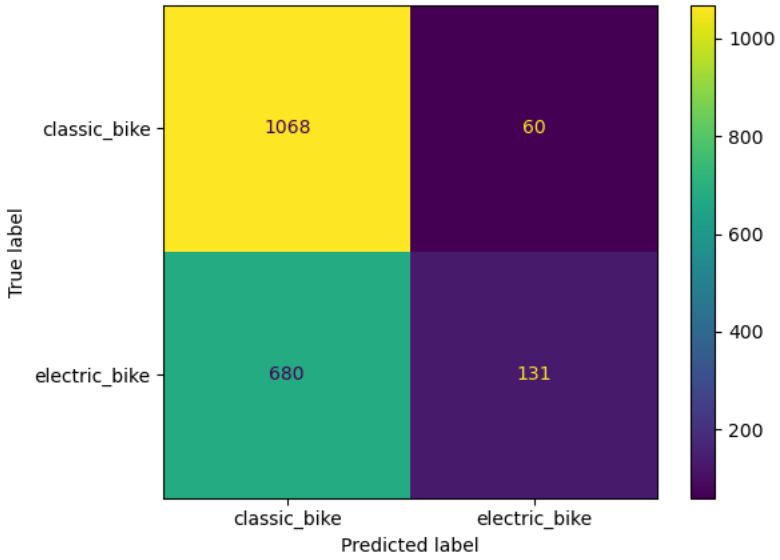
I picked a LDA model as a new one to explore for this project because it can perform dimensionality reduction-unlike the other models ran-which helps simplify the data while retaining the most important features. LDA assumes a linear relationship between the features and the target, which can make it less complex but easier to interpret. Additionally, it is a simpler model that can work like a baseline to compare against more complex models like Random Forest.

```
Train Score: 0.6233230134158927
Test Score: 0.6183599793708097
Precision Score: 0.6423030084808433
Cross-Validation Scores: [0.62282398 0.62346873 0.62129032 0.62258065 0.61870968]
Mean Cross-Validation Accuracy: 0.6217746719078222
Classification Report:
              precision    recall  f1-score   support

 classic_bike       0.61      0.95      0.74      1128
electric_bike       0.69      0.16      0.26       811

     accuracy                           0.62      1939
    macro avg       0.65      0.55      0.50      1939
 weighted avg       0.64      0.62      0.54      1939
```

The LDA model did a very good job of correctly identifying classic_bikes, but did a pretty horrible job of identifying electric bikes. This suggests that the linear boundaries LDA relied on were not sufficient to separate the two classes effectively, which is supported by my findings in our EDA at the beginning of this project. The overlap in feature distributions likely made it difficult for the model to maximize class separation. This indicates the features that influence bike type may have more complex or nonlinear interactions, where LDA is not the best to use.



Next Steps and Discussion

Summary of Findings

In this analysis of CitiBike data to predict bike type chosen, the models I constructed and tested demonstrated varying levels of accuracy in predicting bike type, with overall low accuracy. The models ranked in terms of performance are as follows: Random Forest Classifier, Decision Tree Classifier, K-Nearest Neighbors, Logistic Regression, and Linear Discriminant Analysis.

1) Success of Random Forest Model: The Random Forest model was the most effective, with the highest accuracy and a more balanced performance between classes as compared to other models. This model reduced overfitting and captured more complex patterns, making it better at distinguishing between classic and electric bikes.

2) Important Features: Across all models, the distance (ride_miles) and duration (ride_minutes) of rides were the most important predictors of bike type. Other features like membership type and time of day had minimal influence on predictions.

3) Challenges with Linear Models: The Logistic Regression and LDA models performed worse due to their reliance on linear relationships between variables. This led to weak separation between classes, especially when classifying electric bikes. These models highlighted the need for more complex models that could capture non-linear relationships in the data.

4) Trade-Offs in Model Performance: While the Decision Tree and KNN models showed improved accuracy compared to linear models, KNN still misclassified a significant number of electric bike rides and Decision Tree perfomed better with electric bikes but at the expense of worse accuracy with classic bikes. The Decision Tree's constrained depth and KNN's reliance on local proximity limited their ability to fully capture the complexities.

The Random Forest model proved to be the most accurate among the models tested, but its precision score of 0.66 highlights significant room for improvement. The relatively low accuracy suggests that the features in the dataset may not fully capture the relationships influencing bike type. We need further data exploration or additional features to better model the underlying patterns. These findings emphasize the limitations of the current dataset CitiBike provides, and gives us a starting point for future predictions.

Next Steps/Improvements

To make the predictive capabilities of the models better and gain deeper insights into bike type prediction for CitiBike, I would want to incorporate these additional features into my models:

- Weather Data: Including weather conditions during the rides, such as temperature, precipitation, and wind speed, could provide valuable insights into how external factors influence bike type usage. For example, there may be a greater preference for electric bikes on cold or rainy days.

- User Demographics: Data on user demographics, such as age, gender, or residential neighborhood, could help us find patterns in preferences based on user characteristics. This could reveal how different user groups may have varying CitiBike preferences.
- Traffic and Infrastructure: Data on traffic and proximity to subway stations or bus stops could point to external factors that may influence the type of bike chosen.
- Station-Specific Data: Information about if the bike was picked up at a station that is more often a corporate hub versus residential area or tourist location may provide insight into which bike chosen. For example, if picked up from residential area and dropped at a corporate hub, that rider was potentially on the way to work and may pick an electric bike out of convenience.

By incorporating some additional features into the analysis, the models could capture a broader range of potentially more important factors that influence bike type usage, leading to better accuracy and more actionable insights for CitiBike and other bike ridesharers.