Chicago Divvy Bike-sharing Program Bike Usage Prediction

IST.718.M001.FALL24.Big Data Analytics

Team Members

Sai Mani Kiran Chatrathi

Hang Tian

Goutham Sri Vishwesh Bikkumalla

Instructor's Name

Christopher Dunham

December 13, 2024

Introduction :

The Chicago Divvy Bike-sharing program, managed by the Chicago Department of Transportation (CDOT), is a popular transportation option for both residents and visitors. Understanding the factors that drive bike usage is crucial for stakeholders who are tasked with making decisions about station placement, bike fleet management, and service optimization. By analyzing historical data from 2022 to 2023, this project aims to develop a predictive model to forecast Divvy bike usage across Chicago, providing actionable insights into fleet distribution and operational efficiency.

The primary stakeholder for this project is the Chicago Department of Transportation (CDOT). Our predictive insights into bike usage patterns can guide CDOT in planning and allocating resources for the bike-sharing program, enhancing service availability and reliability. Additionally, this information supports strategic decisions regarding the expansion of services and integration with other transportation modes, aiming to optimize the city's bike-sharing infrastructure and improve urban mobility.

In this project, we utilized a comprehensive dataset that includes millions of anonymized bike trips, geographical boundaries of Chicago neighborhoods, and environmental factors like weather conditions, combined with public holiday data. By applying machine learning models to these datasets, we have developed predictive tools that can forecast daily bike usage demand with a high degree of accuracy.

Data :

1. Bike-sharing trip data:

Link: *https://divvy-tripdata.s3.amazonaws.com/index.html*

Since the Divvy bike-sharing program is ran by CDOT (Chicago Department of Transportation), this dataset is highly reliable. We used bike trip data from 2022 to 2023 for this project.

2. Holiday information:

According the calendar, we've compiled all the weekdays and holidays from 2022 to 2023.

3. Daily weather data:

Weather data for 2022 to 2023 collected from the Visual Crossing website.

For the final data we used for modeling, we have total 472885 rows and 21 columns. Including our dependent variable aggregated daily counts of bike usage; weather data: temperature, snow and visibility; date, holiday and weekday information.

Data PreProcessing :

Step 1: Data Processing

To get the dataset ready for analysis, a few key steps were taken. First, the columns end_station_name, end_station_id, end_lat, and end_lng were removed because they weren't relevant to the focus of this part of the project. By dropping these columns, the dataset became simpler and more focused on the essential information.

After that, rows with missing values in the start_station_id column were removed. Since having a valid starting station is crucial for the analysis, this step ensured that only complete and reliable data was kept. These changes made the dataset cleaner and more usable for the next stages of the project.

Step 2: Data Processing

To facilitate consistent reference to each station, two new columns were added to the dataset: mean_latitude and mean_longitude. These columns were created by calculating the average latitude and longitude for each station, grouped by start_station_id. The mean latitude was stored in mean_latitude, while the mean longitude was stored in mean_longitude.

By including these aggregated values, the dataset now provides a single, representative set of coordinates for each station, simplifying the process of identifying and working with station locations in the analysis.

Step 3: Data Aggregation

To analyze the dataset at an aggregated level, the data was grouped by mean_latitude, mean_longitude, and date. Within each group, the total number of rides was calculated by counting the occurrences of rideable_type, representing the ride count for that specific combination of location and date.

The resulting summary was stored in a new DataFrame, with a column named ride_count that contains the ride totals for each grouping. This aggregation provides a clear overview of ride activity based on location and date, making it easier to identify patterns and trends.

Step 4: Feature Engineering for Weekdays and Holidays

To enrich the dataset with additional context, two new features were introduced. First, the is_weekday column was created by checking the day of the week for each date. If the day was a weekday (Monday to Friday), the value was set to 1; otherwise, it was set to 0 to indicate a weekend.

Next, a list of U.S. holidays for 2022 and 2023 was defined and converted into a datetime format for comparison. The is_holiday column was then added by checking whether each date in the dataset matched any of the holiday dates. If it was a holiday, the column was set to True; otherwise, it was set to False.

These features provide insights into how ride activity varies across weekdays, weekends, and holidays, offering valuable context for temporal analysis.

Step 5: Merging Ride Data with Weather Data

To incorporate weather information into the analysis, the ride data was merged with a weather dataset. The merge operation was performed by aligning the date column from the ride data (grouped) with the datetime column from the weather data.

The resulting DataFrame, merged_df, combines ride activity details with corresponding weather conditions for each date. This integration enables a more comprehensive analysis of how weather influences ride behavior and trends.

Step 6: Handling Missing Weather Data

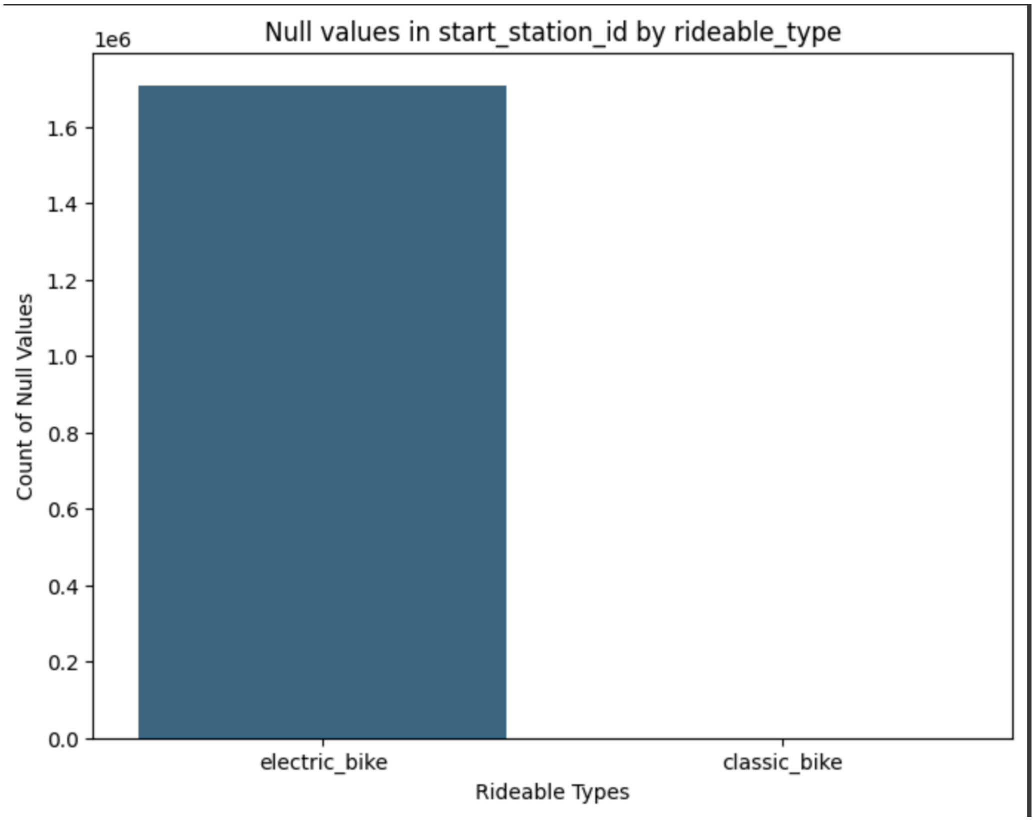
To handle missing values in the weather-related data, the severerisk and windgust columns were filled with their respective mean values. This approach ensures that all rows have complete data without introducing significant bias. By replacing missing values with the average, the dataset remains consistent and reliable for analysis while preserving the overall distribution of these features.

Step 7: Feature Selection Based on Correlation Analysis

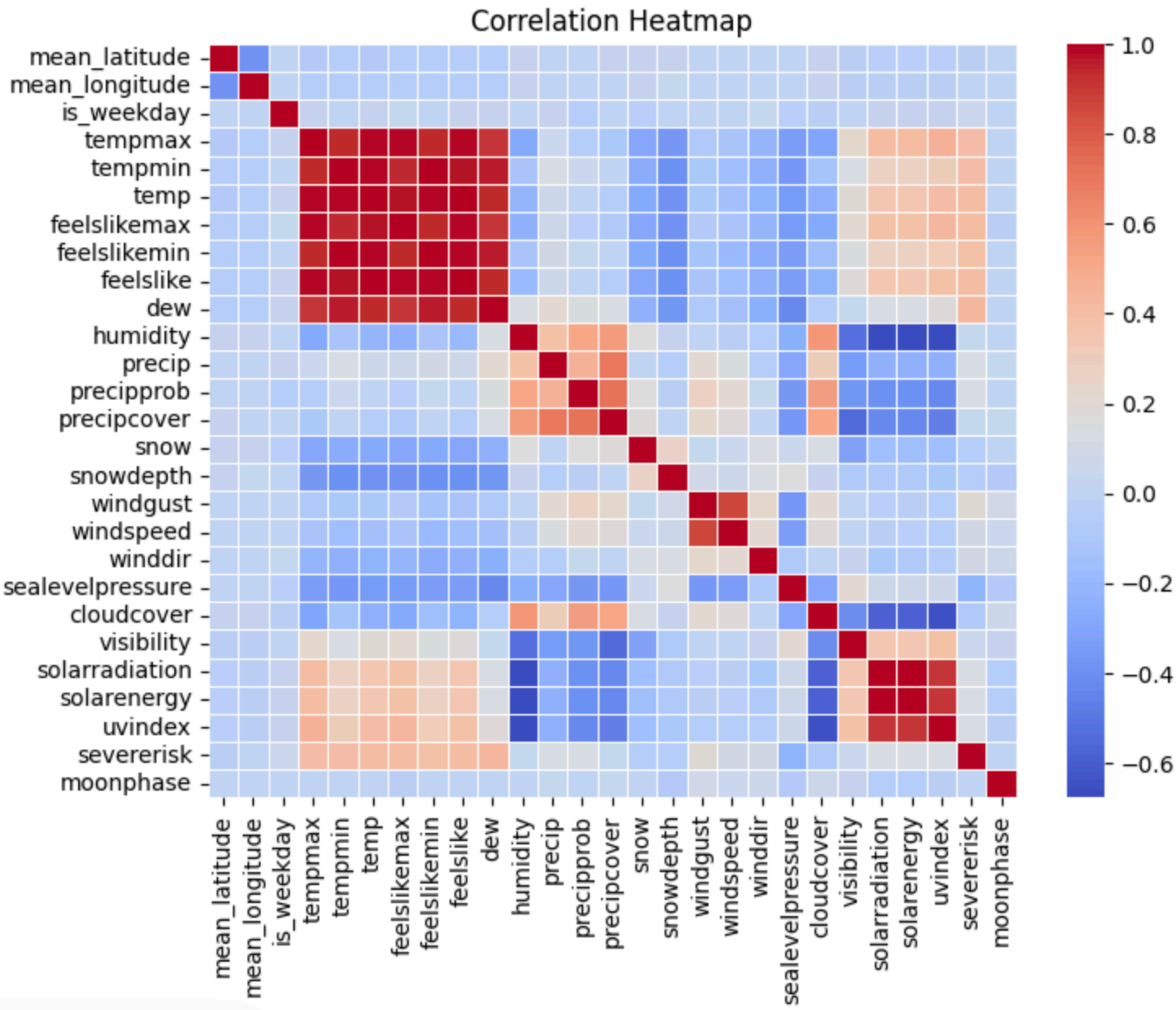
Through performing correlations between the various variables in the dataset, we identified which features have the strongest relationships with the target variable. Based on this analysis, we selected certain columns as the most important features for the model. These features are expected to provide the most valuable insights and contribute significantly to the model’s performance, helping to make accurate predictions.

Insights :

Not-at-station records : The majority of missing station IDs are linked to electric bikes, while classic bikes have significantly fewer missing values. This indicates that electric bikes are more likely to have missing station data, possibly due to issues like unrecorded station locations or incomplete data entry. Understanding this trend is important, as it helps us focus on why this is happening and find ways to address these missing values in future analyses or models.



Inter-Related Columns : weather-related factors, like temperature and wind speed, are interrelated. For instance, temperature variables such as tempmax and tempmin are strongly correlated, which makes sense as they reflect similar data. While there isn't a strong direct correlation with rideable_count, weather conditions like temperature and wind speed could still influence bike usage. This suggests that weather features might play a role in predicting bike usage, especially extreme conditions.


Correlation Heatmap
mean_latitude
mean_longitude
is_weekday
tempmax
tempmin
temp
feelslikemax
feelslikemin
feelslike
dew
humidity
precip
precipprob
precipcover
snow
snowdepth
windgust
windspeed
winddir
sealevelpressure
cloudcover
visibility
solarradiation
solarenergy
uvindex
severerisk
moonphase
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Models Chosen and Reasons:

Linear Regression:

Linear Regression was used for its simplicity and ability to capture linear relationships between features and daily demand. It provides a quick baseline, helping us understand basic trends in the data and offering a starting point for comparison with more complex models.

Random Forest:

Random Forest was chosen for its ability to handle complex, non-linear relationships. It's ideal for our problem, where demand can be influenced by many interacting factors. Its multiple decision trees help reduce overfitting and provide a robust prediction, especially with large datasets.

Gradient Boosted Trees:

Gradient Boosted Trees were selected for their high accuracy in capturing subtle patterns in the data. Their iterative approach allows them to improve upon errors, making them powerful for predicting demand in complex scenarios, though they require careful tuning

Decision Tree:

A Decision Tree was used for its interpretability. It provides a clear view of how different factors, like weather or time, influence demand. While it can overfit, it's helpful for understanding feature importance and decision-making processes.

Results :

Algorithm	R squared	RMSE (usage counts)

Random Forest	0.459	22.73
GBT	0.735	15.91
Decision Tree	0.516	21.5
Linear Regression	0.217	27.4

For the daily usage count prediction (demand prediction), we worked on these four models and their evaluation metrics are shown in the table above. Unlike the best result in Sci-kit learn package (which we tried before running the pyspark machine learning one), the best R squared in pyspark is just 0.735.

Discussions & Limitations :

Data Limitations:

Forecast accuracy depends to a large extent on the quality of the data used. Problems such as incomplete datasets and missing or inaccurate values (especially for external data such as weather reports) can lead to less reliable forecasts. Prioritizing improvements in data collection, validation and pre-processing is essential to maintain forecast accuracy. Besides, the bike usage data has its intrinsic flaws such as spatial and temporal imbalance. For prediction on a smaller scale, more records are needed.

Station Limitations:

In our analysis, the predictions are limited to estimating the usage of bikes that are taken from the stations. We did not include data or predictions related to bikes that are taken from other locations, such as road sides or footpaths. This limitation means that the model only considers a subset of all potential bike usage, excluding important variables that may affect the overall demand for bikes in the area. As a result, the predictions might not fully reflect the true demand for bike rentals, as they do not account for bikes that are used outside of station-based pick-up points.

References :

[1]: Ashqar H I, Elhenawy M, Almannaa M H, et al. Modeling bike availability in a bike-sharing system using machine learning[C]//2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, 2017: 374-378.

[2]: Yang H, Xie K, Ozbay K, et al. Use of deep learning to predict daily usage of bike sharing systems[J]. Transportation research record, 2018, 2672(36): 92-102.

[3]: Mehdizadeh Dastjerdi A, Morency C. Bike-sharing demand prediction at community level under COVID-19 using deep learning[J]. Sensors, 2022, 22(3): 1060.