Introduction to Machine Learning (67577)
Hackathon 2024 - Challenge 1:
HU.BER– Optimizing Public Transportation Routes

Authors :

Zenab Waked - 323000026
Adan Irshed - 211508353
Haya Natsheh - 325098440
Adan Mahameed - 214310666

Preprocessing :

1. Rows with Invalid Values:
    - Rows without arrival time or with invalid arrival time format:nAny row where the `arrival_time` is missing or cannot be parsed correctly.
    - Rows with negative passenger counts: Any row where the number of passengers boarding (`passengers_up`) or continuing (`passengers_continue`) is negative.
    - Rows with invalid location data (latitude and longitude): If the standard deviation of the location data is too high, indicating that the location is far from the average area, those rows are removed.
2. Dropping Irrelevant Columns:
    - Identifiers without quantitative value: Columns like `trip_id`, `trip_id_unique`, `trip_id_unique_station` are dropped because they are identifiers and do not have quantitative value for prediction.
    - Undefined values: Columns like `part` and `alternative` are dropped because they are not well-defined.
    - Non-numeric station names: The `station_name` column is dropped because it is non-numeric. While it can be converted to a categorical variable, the station information is also present in the location data (latitude and longitude), so it is omitted.
3. Modifying Existing Columns, Creating New Columns, and Filling Values:
    - Converting `direction` to categorical values (0 or 1): The `direction` column is transformed into binary values (0 or 1) using `LabelEncoder.`
    - Converting `cluster` to categorical variables using `LabelEncoder`: The `cluster` column is transformed into categorical variables using `LabelEncoder` from `sklearn`.
    - Converting `arrival_is_estimated` to categorical values (0 or 1): The `arrival_is_estimated` column is converted to binary values (0 or 1).
    - Creating `time_in_station`: A new column `time_in_station` is created, which is calculated as the difference between `door_closing_time` and `arrival_time`. This column contains information that may affect the number of passengers boarding.
4. Filling Predictable Missing Values:
    - Filling missing values in `time_in_station`: If there are no data on door closing time (e.g., when the arrival time was estimated), the average time the bus waited at the station is filled in.

Model Selection :

To select an appropriate model for predicting the number of passengers boarding the bus at each stop, we first examined the data to understand its characteristics and behavior. Our initial analysis involved checking the linearity of the relationships between various features and the target variable (`passengers_up`). By calculating the correlation coefficients between each feature and the target, we identified that some features exhibited a linear relationship with `passengers_up`, suggesting that a linear regression model might be suitable for these cases.

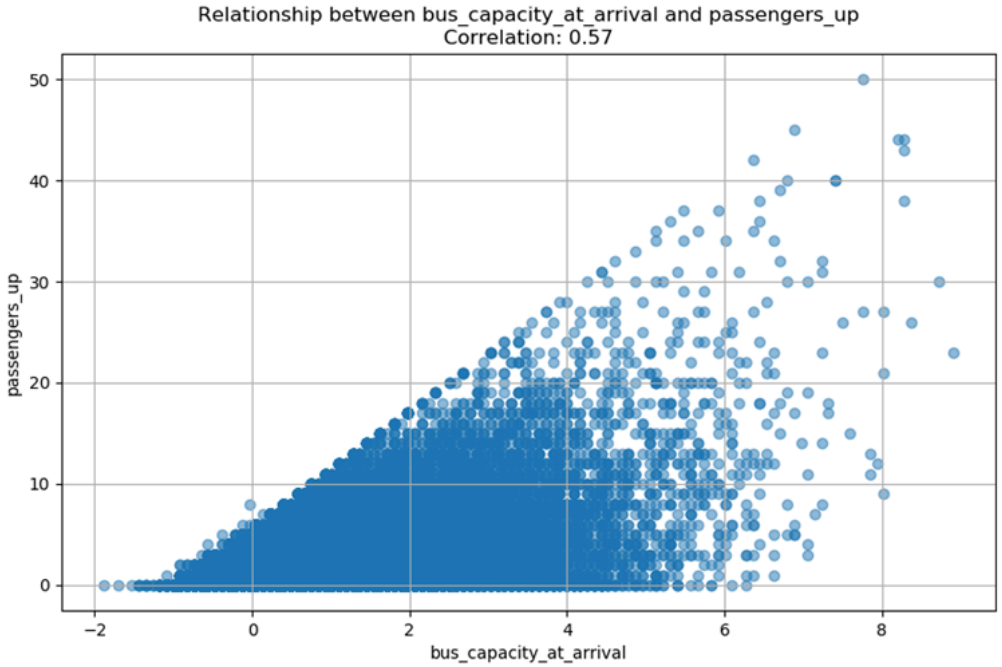
Given the presence of linear correlations for specific features, we decided to employ a Linear Regression model as our primary approach. Linear regression is straightforward and effective for modeling relationships where a linear trend exists between the independent variables and the dependent variable. This model allows us to understand the direct impact of each feature on the number of passengers boarding, providing interpretable coefficients that highlight the significance of each predictor.

However, we also recognized that not all features had a strong linear correlation with `passengers_up`. In some instances, the relationships were more complex or non-linear, which could limit the effectiveness of a linear model. To address this, we explored other modeling techniques such as Decision Tree Regressors and Random Forests. Decision trees can capture non-linear relationships and interactions between features, making them suitable for data with complex patterns. Random forests, being an ensemble method of decision trees, offer improved accuracy and robustness by averaging the predictions of multiple trees, reducing overfitting and variance.

For the second part of our task, predicting the duration of bus trips, we opted to use decision trees instead of linear regression. The rationale behind this choice lies in the nature of the data and the problem. Bus trip durations can be influenced by various factors such as traffic conditions, number of stops, and time of day, which introduce non-linearities and categorical variables into the dataset. Decision trees are well-suited for handling such complexities as they can effectively manage non-linear relationships and categorical features. Moreover, decision trees can handle missing values more gracefully and provide clear insights into the factors affecting trip duration.

In conclusion, our model selection process was driven by the characteristics of the data and the specific requirements of each prediction task. By using a combination of linear regression and decision tree-based models, we aimed to leverage the strengths of each method to achieve accurate and interpretable predictions. This approach ensures that we can effectively model both linear and non-linear relationships, improving our ability to predict the number of passengers boarding at each stop and the duration of bus trips in a comprehensive manner.

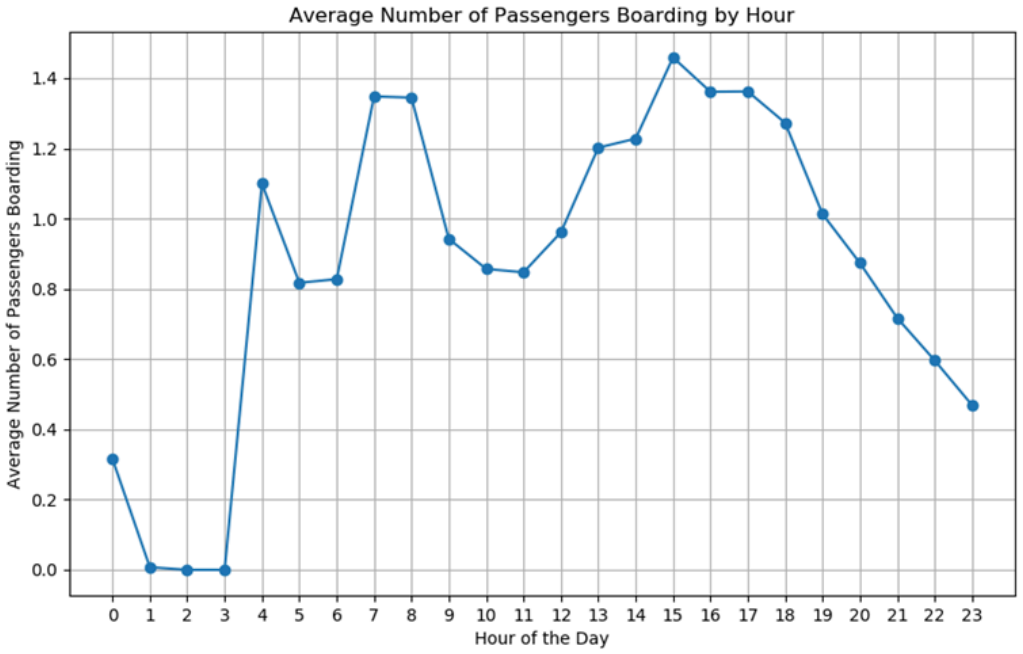
plots :

1. 

There is a clear positive correlation (0.57) between `bus_capacity_at_arrival` and `passengers_up`.

Explanation: As the bus capacity at arrival increases, the number of passengers boarding the bus (`passengers_up`) also tends to increase. This indicates that buses with more passengers already on board tend to pick up more passengers at subsequent stops.

2.



The plot titled "Average Number of Passengers Boarding by Hour" illustrates the average number of passengers boarding the bus at each hour of the day. Here's a detailed explanation of the observed results:

1. Early Morning (0:00 - 4:00)

The average number of passengers boarding the bus is very low, with some hours having close to zero passengers.

Reason: This period typically represents late night to early morning hours when bus services are either limited or not operating, and fewer people are traveling.

2. Morning Peak (5:00 - 9:00)

There is a significant increase in the number of passengers boarding the bus starting around 5:00, peaking between 7:00 and 9:00.

Reason: This peak corresponds to the morning rush hour when people are commuting to work, school, or other morning activities.

3. Midday (10:00 - 14:00)

The average number of passengers boarding stabilizes and shows moderate activity during these hours.

Reason: This period includes a mix of late morning commuters, lunchtime travelers, and possibly non-commuter passengers engaging in midday activities.

4. Afternoon Peak (15:00 - 18:00)

There is another noticeable peak in the number of passengers boarding the bus, with the highest point around 16:00.

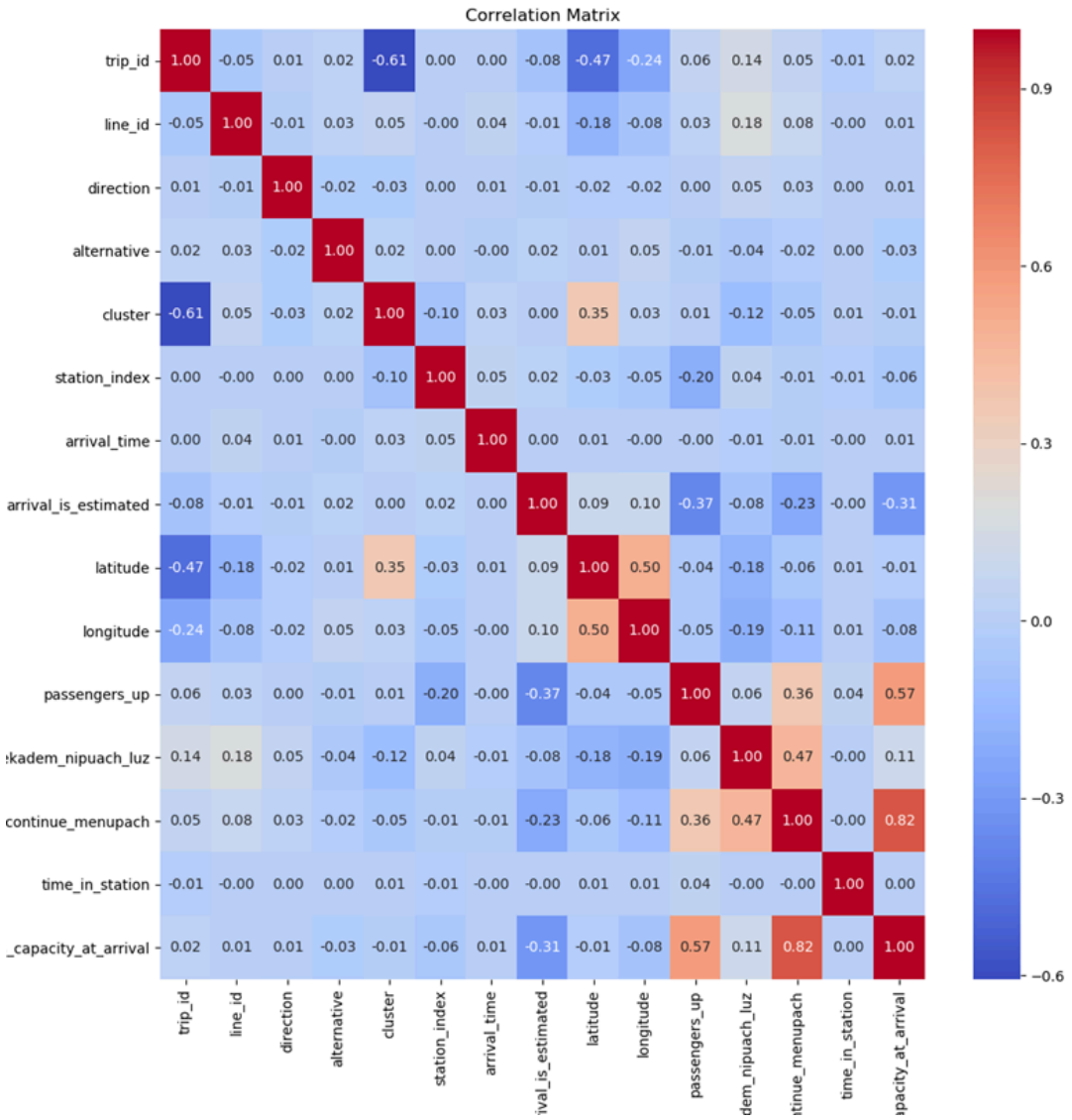
Reason: This second peak corresponds to the afternoon rush hour when people are returning from work, school, and other afternoon engagements.

5. Evening (19:00 - 23:00)

The number of passengers boarding gradually decreases as the evening progresses, reaching lower numbers by 23:00.

Reason: Fewer people travel during late evening hours, leading to a decline in bus usage.

3.



The correlation matrix is a visual representation of the Pearson correlation coefficients between different features in the dataset. The values range from -1 to 1, where:

- 1 indicates a perfect positive correlation.
- -1 indicates a perfect negative correlation.
- 0 indicates no correlation.

Here's an explanation of the key observations from the correlation matrix:

1. Passengers Up (`passengers_up`) Correlations:

Positive Correlations:

`bus_capacity_at_arrival` (0.57): This indicates a strong positive correlation, meaning as the bus capacity at arrival increases, the number of passengers boarding tends to increase.

`mekadem_nipuach_luz` (0.36): This feature also shows a moderate positive correlation with the number of passengers boarding.

Negative Correlations:

`arrival_is_estimated` (-0.37): This negative correlation suggests that when the arrival time is estimated, fewer passengers tend to board.

2. Feature Inter-Correlations:

`latitude` and `longitude` (0.50): There is a strong positive correlation between these two features, which is expected as they are geographical coordinates.

`mekadem_nipuach_luz` and `bus_capacity_at_arrival` (0.82): This high positive correlation suggests that these features are closely related. It might be useful to consider if one of them can be derived from the other or if they provide redundant information.

3. Other Notable Correlations:

`arrival_is_estimated` and `bus_capacity_at_arrival` (-0.31): This negative correlation implies that when arrival times are estimated, the bus capacity at arrival is generally lower.