HU.BER– Optimizing Public Transportation Routes

Overall Data Analysis

Some features of the dataset were hard to understand, such as the difference between `trip_id` and `trip_id_unique`, making it difficult to determine their relevance in different contexts. Additionally, features that seemed straightforward had confusing data. For example, the column `alternative` appeared simple but contained a mix of numbers and `#`, and its definition was unclear. We were uncertain whether it referred to buses traveling from the same departure point to the same destination or those with the exact same stations.

Despite these challenges, the extensive dataset allowed us to extract interesting information. For instance, we determined the number of stations for a particular line by extracting the maximal `station_index` for that line. This insight helped us understand the line's location within a trip, as a station at the beginning typically had more passengers boarding than one at the end. We assumed that different regions experience varying levels of traffic congestion, significantly affecting bus travel time. Additionally, regional socio-economic status likely influenced the number of people using public transportation. The time of day also impacted travel time and passenger numbers, with morning and afternoon hours being peak times due to work commutes, leading to more congested roads and increased public transportation usage.

Furthermore, we hypothesized that stations with numerous bus lines would have higher boarding numbers, as these stations provide diverse travel options, reducing dependence on a single bus.

Task One: Predicting Passenger Boardings

Extracting Insights from the Dataset

Despite these challenges, the extensive dataset allowed us to extract interesting information. For instance, we determined the number of stations for a particular line by extracting the maximal `station_index` for that line. This insight helped us understand the line's location within a trip, as a station at the beginning typically had more passengers boarding than one at the end.

We assumed that different regions experience varying levels of traffic congestion, significantly affecting bus travel time. Additionally, regional socio-economic status likely influenced the number of people using public transportation. The time of day also impacted travel time and

passenger numbers, with morning and afternoon hours being peak times due to work commutes, leading to more congested roads and increased public transportation usage.

Furthermore, we hypothesized that stations with numerous bus lines would have higher boarding numbers, as these stations provide diverse travel options, reducing dependence on a single bus.

Model Selection and Performance

Initially, we opted for Linear Regression due to its simplicity and transparency, making it ideal for the initial exploration of our dataset and baseline predictions. It provided a clear understanding of how individual features influenced the target variable, "Passenger Up," which was crucial for interpretability. However, after evaluating the model's performance and finding a Mean Squared Error (MSE) resulting in a loss of 3.5, we recognized its limitations in capturing complex interactions and nonlinear patterns present in the data.
This realization prompted us to pivot to the Random Forest Regressor, known for its capability to handle nonlinearity and capture intricate relationships through an ensemble of decision trees. Leveraging insights from multiple trees and optimizing parameters, we achieved a significant reduction in MSE to 2.2, underscoring the model's enhanced predictive power. This shift highlighted the importance of selecting models that align with the inherent complexities of the dataset, ensuring more accurate and reliable predictions in data-driven endeavors.

Task Two: Predicting Trip Duration

Dataset Challenges and Insights

For Task Two, the output format was drastically different from the input. While the input contained information on each station, the output required data relative to each trip individually. Preprocessing involved a deep understanding of pandas functions to transform the input from a station-based to a trip-based format. Many features seemed irrelevant, while the relevant ones required additional processing.

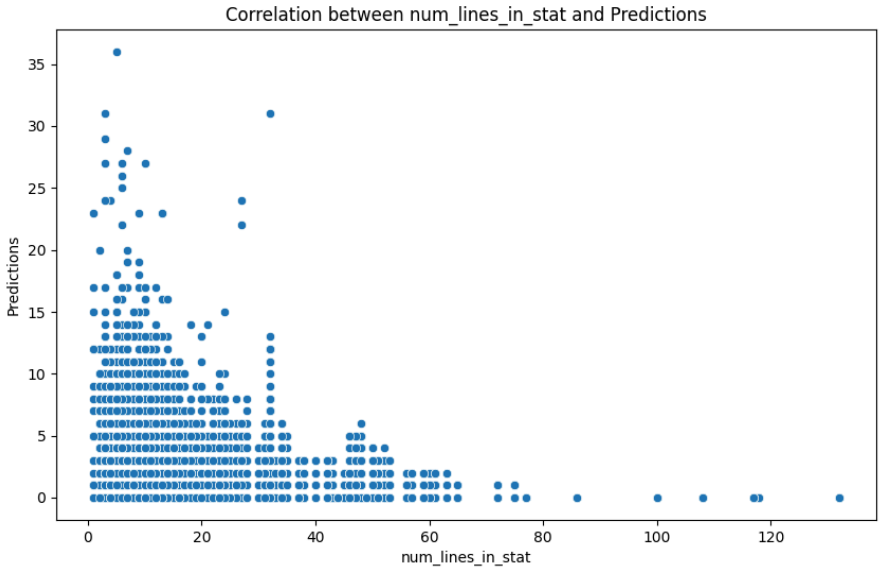
Enhancing Feature Engineering

For this task, we added new columns for total distance and average distance between stations for each trip. However, this distance was calculated as the aerial distance, which might be significantly shorter than the actual travel distance. This feature was assumed to be fairly accurate for short trips within the same town but less useful for longer distances.

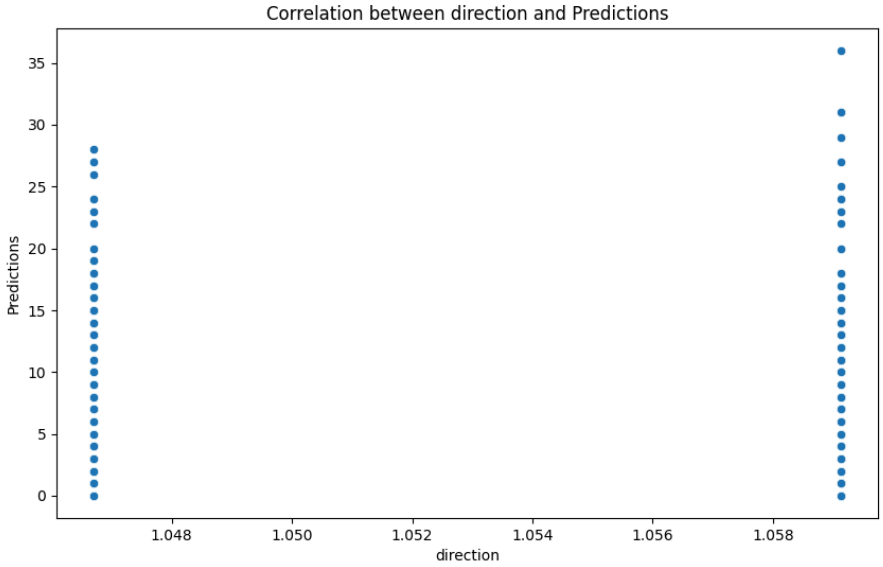
Model Selection and Performance

Similar to Task One, we initially used Linear Regression to explore the dataset and make baseline predictions. However, after finding its limitations in capturing the complex relationships within the data, we shifted to the Random Forest Regressor. This model effectively handled

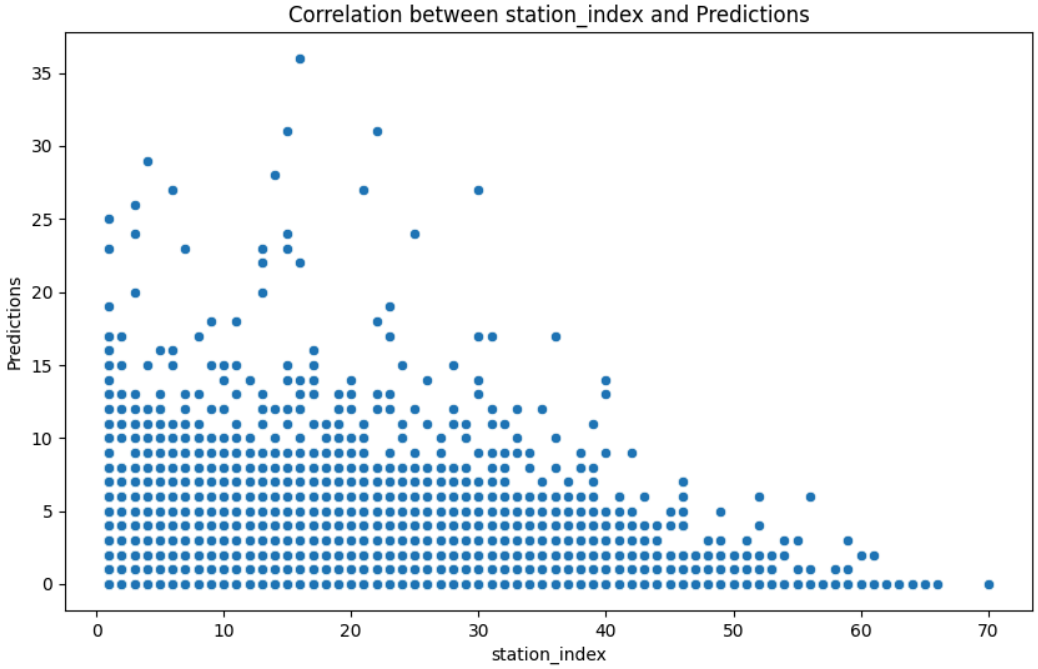
nonlinearity and provided better predictive performance. By optimizing parameters and leveraging the strengths of the Random Forest Regressor, we improved the accuracy of our predictions, ensuring more reliable outcomes for predicting trip duration.



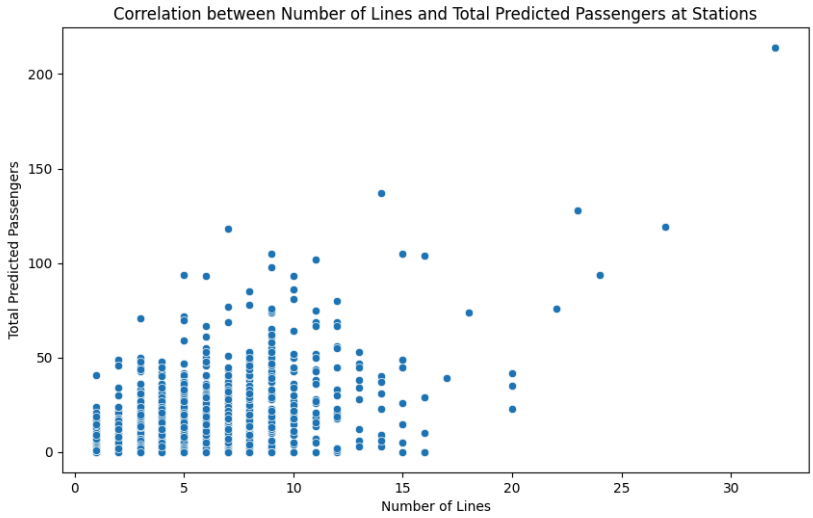
Correlation between number of lines in a given station and number of passengers up. We can see that when the number of lines stopping at a station goes up, less passengers go on the bus, probably because there are more lines to choose from.



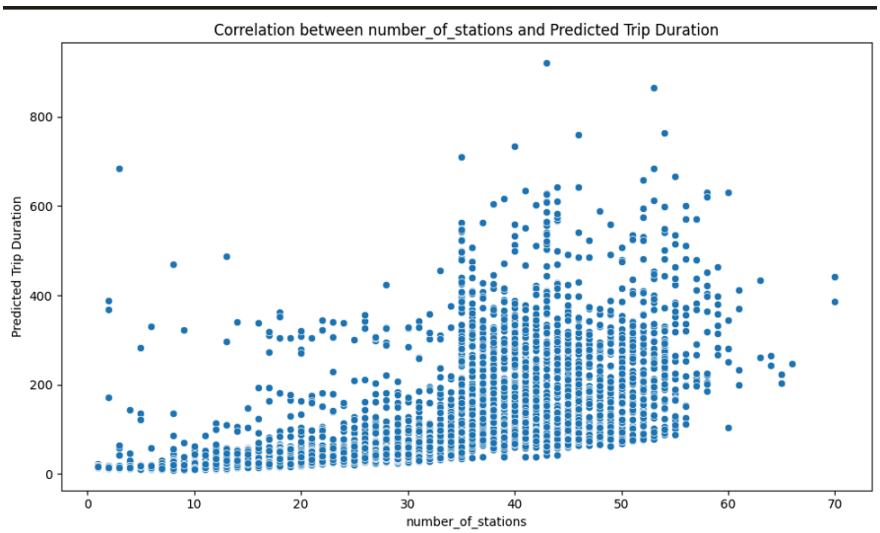
The direction doesn't seem to have an impact on the number of passengers boarding a bus.



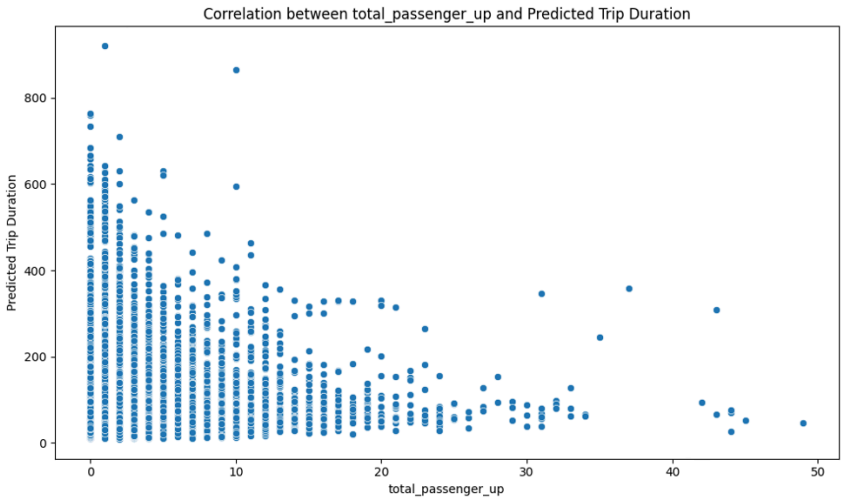
When the station_index is higher, that is the station is located more towards the end of the trip, less passengers board the bus, because there are less stations to go off at.



This graph shows the correlation between the number of lines at a given station and the sum of the predicted passengers for this station. We can see that up to about seven lines per station, there aren't a lot of passengers and they come in frequent waves. From ten lines and above, more passengers are boarding lines each time but in less frequent waves.



From this plot we can see from the graph that as the number of stations increases, the length of the route also increases. This indicates that each station adds total travel time to the bus, both because it takes time to stop and because people get on and off at each station, which delays the journey.



The plot demonstrates a negative correlation between the number of passengers getting on the bus and the predicted trip duration. As the number of passengers increases, the predicted trip duration generally decreases. There is significant variability in trip durations for routes with fewer passengers, whereas routes with more passengers show shorter and more consistent trip durations. This suggests that bus routes with higher passenger volumes might be more optimized or experience fewer delays. Additionally, it indicates that people may prefer not to board buses that have longer predicted trip durations.