Part 3.1:

During the preprocessing phase, we strategically removed features with low correlation to enhance the model's predictive performance. To do that we made a heat map of the correlation between the different features. Features such as `trip_id_unique_station`, `trip_id_unique`, `station_name`, `alternative`, `door_closing_time`, `trip_id`, and `station_id` were identified as having low correlation.

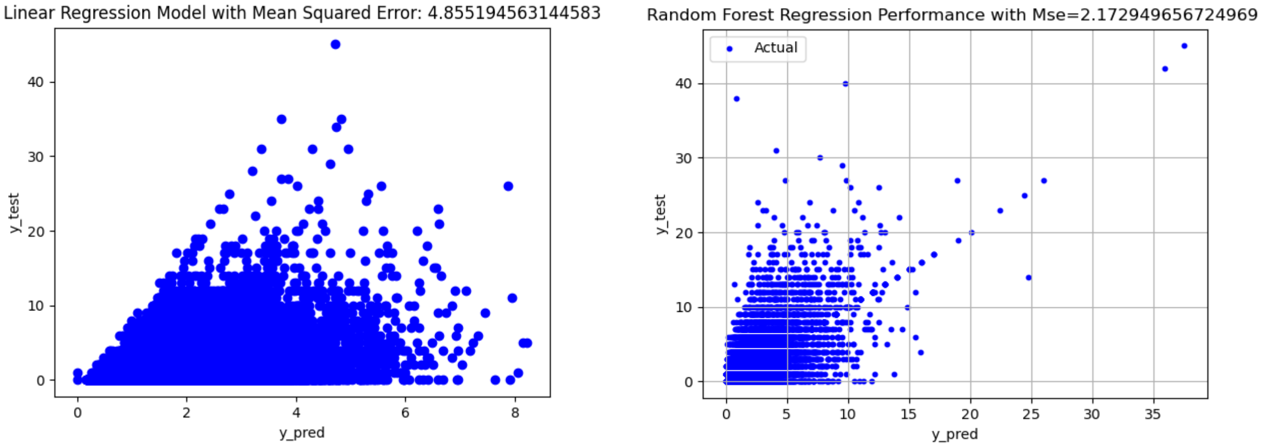


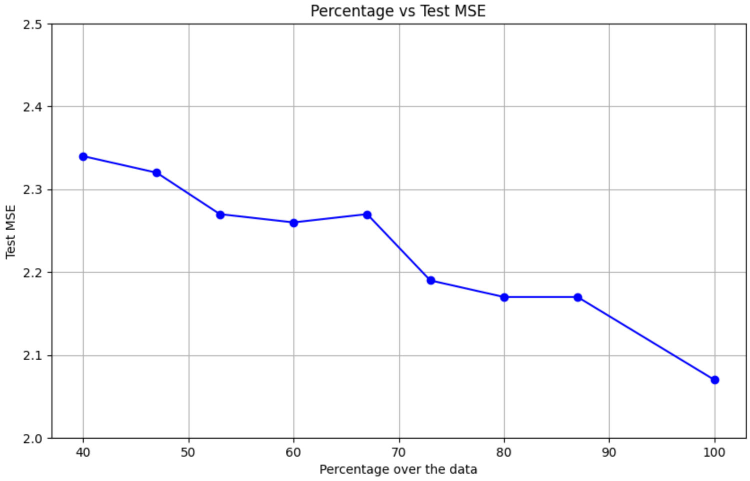
Figure 1. first baseline error and the final model

Our initial baseline model utilized linear regression, which yielded a Mean Squared Error (MSE) of 4.86. However, we observed that this model exhibited high bias.

Moving forward, further preprocessing steps were necessary to refine feature selection and address any remaining biases in the model. By iteratively improving our preprocessing techniques and carefully selecting features based on their correlations and impact on model performance, we aimed to enhance the predictive accuracy of our models in subsequent analyses.

We tried different algorithms until transitioning to Random Forest, which significantly improved our predictive performance to 2.46 MSE. when later we transitioned to Random Forest Regression model. Through iterative refinement and feature selection, we systematically reduced the MSE to 2.31. This improvement underscores the Random Forest's ability to capture non-linear relationships and feature interactions inherent in the dataset.

Further enhancement was achieved through regularization techniques, where we strategically tuned model parameters to control overfitting and achieve a final MSE of 2.07. This approach involved balancing model complexity with generalization, ensuring robust performance on unseen data. The main parameter that we controlled was to limit the depth of the base model trees to 40.



In this graph we can see how we implied are model on different amounts of the total data. In each iteration we split are data to 80% train and 20% test. We see that by using more data we managed to lower are error.

Part 3.2:

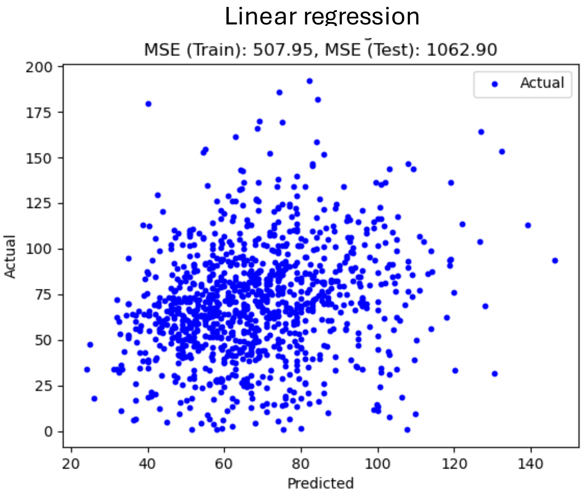
We proceeded to enhance the model further by incorporating additional features.

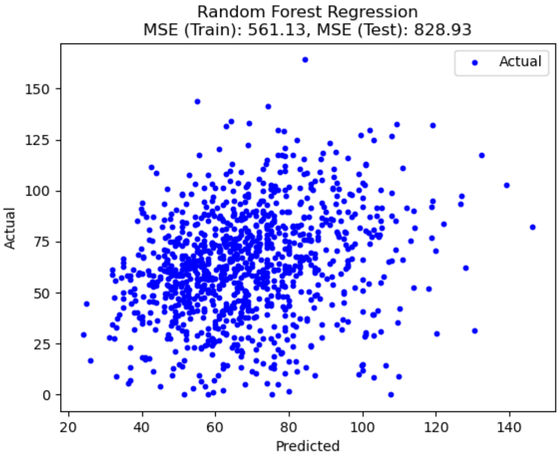
arrival_time_diff: Represents the time it took for a specific trip to move between two stations. This feature demonstrated a high correlation with the overall trip duration, providing valuable insights into transit efficiency.

longitude_diff, latitude_diff: Calculated distances between two stations in a specific trip based on their geographical coordinates. These features were found to correlate strongly with travel times, contributing significantly to the predictive power of our model.

Utilizing Random Forest regression once again, we integrated these new features into our model. By leveraging the ensemble learning capabilities of Random Forests, we effectively captured complex relationships between input features and the target variable.

Through this iterative approach of feature engineering, model selection, and parameter tuning, we continuously refined our predictive model's performance. After many tries, we got MSE 850. We can explain the high error by trying to reserve as much measures as we can. we used every sample since we believed there is correlation between the time it takes to move from one station to the next and the trip duration. Also, we believed shrinking 40~ samples into one will get us a lot of data loss. The "good" part is that our model is not overfitted. It made our MSE a lot higher, but now out of lack of time, that’s what we have.





Part 3.3:

In that part we wanted to extract interesting conclusions from analyzing the data ourselves. We chose to check if we can manage resources (bus drivers and buses) more efficiently during rush hours.