Hackathon 2024 - Project Report

Introduction

In this project, we aimed to predict the number of passengers boarding the bus at each stop using a dataset provided for the hackathon. The dataset contains records of bus routes and various features collected at each stop. We were tasked with two main objectives: predicting the number of passengers boarding at each stop and predicting the total trip duration.

Model Selection

To select an appropriate model for our tasks, we first examined the data to understand its characteristics. We calculated the correlation coefficients between each feature and the target variable (passengers_up). Some features exhibited a linear relationship with the target, leading us to initially choose a Linear Regression model. However, due to the presence of non-linear relationships, we also explored Decision Tree Regressors and Random Forests.

For predicting trip duration, we opted for Decision Trees as they can handle non-linearities and categorical variables effectively. This choice was driven by the nature of the data and the problem.

Plots

Conclusion

In conclusion, we found that combining different models allowed us to capture both linear and non-linear relationships in the data. Our analysis provided valuable insights into passenger boarding patterns, highlighting the importance of considering multiple factors such as time of day and bus capacity.

Hackathon 2024 - Project Report

For future improvements, we recommend investigating outliers, enhancing feature engineering, and using regularization techniques to build more robust models. These steps will help improve prediction accuracy and provide better insights into public transportation usage.