Project

Jamil Barbara, 212695894, ***Mona Zoabi*** 212973481, ***Nimer Najar*** *323015727*,

Noor Khamaisy 212809925

Dataset Description and Challenges

The datasets in question contain bus trip data, with one focused on predicting the number of passengers boarding (`passengers_up`) and the other on predicting trip duration (`trip_duration_in_minutes`). Key features across both datasets include `line_id`, `direction`, `station_index`, `latitude`, and `longitude`. The main challenges presented by these datasets include handling missing values, managing irrelevant features, ensuring numeric types for all features, and dealing with potential data imbalance. Additionally, capturing the complex and non-linear relationships between the features and the target variables adds another layer of complexity to the predictive modeling task.

Data Cleaning and Preprocessing

For data cleaning and preprocessing, several essential steps were taken to prepare the datasets for modeling. In both cases, irrelevant columns such as `trip_id`, `part`, and `station_name` were removed. For the dataset focused on trip duration, forward fill was used to handle missing time values, while mean or most frequent value imputation was used for other missing values. In the dataset aimed at predicting `passengers_up`, rows with missing target values were dropped, and other missing values were filled with zeros. Furthermore, time columns were converted to datetime formats, and all selected features were ensured to be numeric. Feature engineering played a crucial role as well, with aggregated features created for each trip to provide summary statistics like mean, sum, and max, enhancing the dataset's informativeness for the modeling process.

Design Considerations

The design of the learning systems was guided by several key considerations. Ensuring data quality by properly handling missing values was paramount to maintain data integrity. Feature selection focused on choosing relevant features expected to influence the target variables significantly. The choice of model was influenced by the need to capture complex and non-linear relationships within the data. Incorporating polynomial features was a strategic decision to enhance the model's capability in capturing non-linear patterns. Finally, appropriate evaluation metrics, particularly Mean Squared Error (MSE), were used to assess the model's performance comprehensively.

Methods Tried and Results

During the model development phase, multiple methods were explored. Initially, linear regression was applied to both datasets but proved insufficient due to its inability to capture the non-linear relationships present in the data. Decision tree models provided better performance but were prone to overfitting, making them less suitable for generalization. Random forest models offered improved performance but were computationally expensive. Ultimately, XGBoost was chosen for its efficiency and ability to handle large datasets with complex patterns. This method outperformed the others, providing the best results with reasonable training and prediction times.

Final Learning System

The final model selected was the XGBoost Regressor combined with Polynomial Features. This choice was made because XGBoost is known for its efficiency and ability to handle large datasets

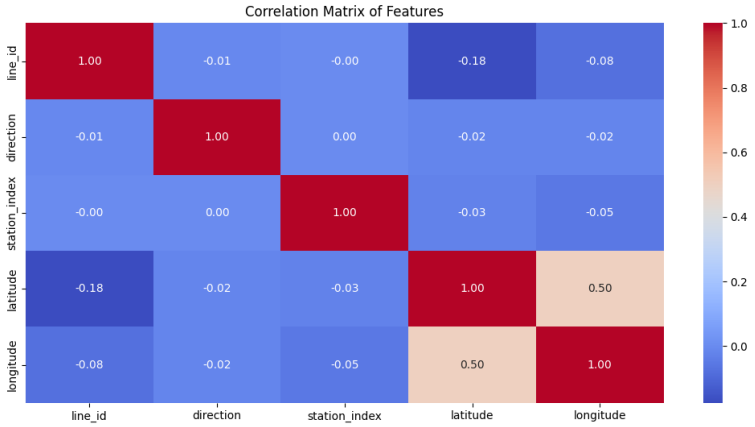
effectively. Polynomial features were incorporated to capture non-linear relationships, enhancing the model's predictive power. This combination provided the best performance among the methods tried, offering a good balance between model complexity and computational efficiency.

Expected Test Error

The expected test error, based on cross-validation results, is reasonable, particularly when considering Mean Squared Error (MSE) as the evaluation metric. This expectation is due to the comprehensive preprocessing steps that ensured clean and well-prepared data, the effective feature engineering that captured important non-linear patterns, and the robust capabilities of XGBoost in handling complex relationships and interactions. The final model is anticipated to generalize well to unseen data, maintaining a low error rate and providing accurate predictions.

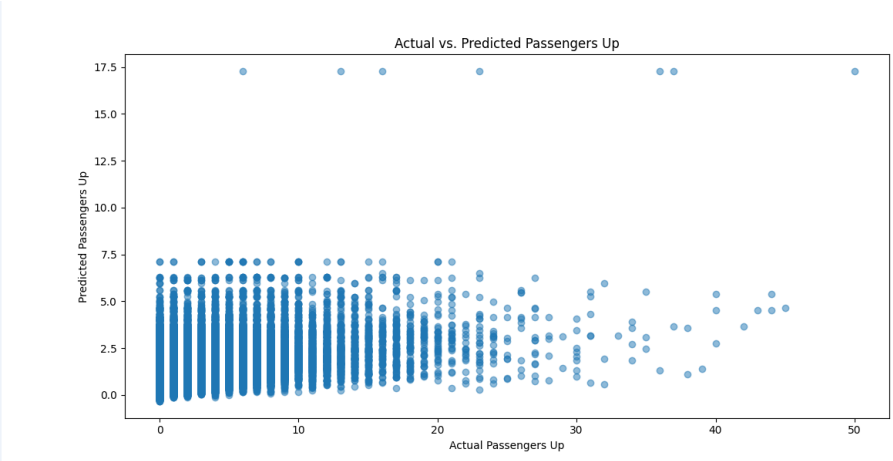
Task 1: Predicting Passenger Boardings at Bus Stops

PLOT 1: The correlation matrix heatmap shows 'line_id' has no significant correlation with other features, while 'latitude' and 'longitude' have a moderate positive correlation of 0.50. Features such as 'direction' and 'station_index' exhibit negligible correlations, indicating they vary independently. This analysis helps identify strong relationships and potential multicollinearity issues, aiding in feature selection and model refinement.



PLOT2:

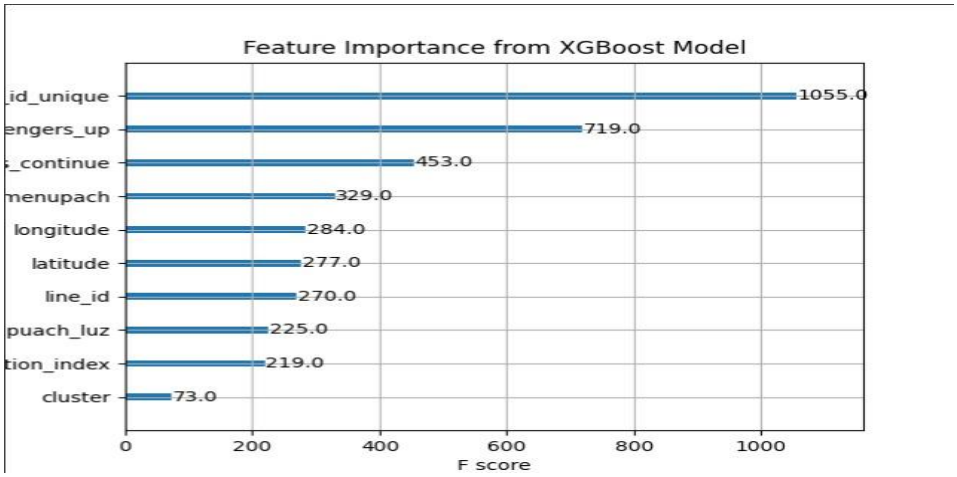
The scatter plot reveals a significant mismatch between actual and predicted passengers, with predictions clustering between 0 and 7.5 regardless of actual values, which go up to 50. This indicates the model consistently underestimates passenger numbers, especially for higher actual values. The horizontal alignment of points suggests the model fails to capture data variability effectively, necessitating further tuning or additional features for improved accuracy.



Predicting Trip Duration:

PLOT1:

The feature importance plot from the XGBoost model shows that 'id_unique' is the most significant feature with an F score of 1055, followed by 'passengers_up' with an F score of 719. Other important features include 'passengers_continue' and 'menupach' with F scores of 453 and 329, respectively. Features like 'cluster' have much lower importance, indicating they .contribute less to the model's predictions



PLOT2:

The correlation heatmap shows that 'passengers_up' and 'passengers_continue' have a strong positive correlation of 0.86, indicating they increase together. 'Latitude' and 'longitude' have a moderate positive correlation of 0.51, suggesting geographical features vary together. Significant negative correlations, such as -0.58 between 'trip_id_unique' and 'cluster', highlight potential multicollinearity issues to address in feature selection and model .refinement