Ridge Regression Model Report

By- Devendra Garwa

1. Objective

The goal of this project is to predict the 'segment_actual_time'—the time taken to complete a delivery segment—based on operational and location-based data. Ridge Regression was chosen as a regularized linear model to handle feature multicollinearity and serve as a robust, interpretable baseline.

2. Dataset Overview

The dataset used is 'selected_features_dataset.csv' containing 144,867 rows and 11 features. These include numerical fields such as distances, cutoff factors, segment efficiency metrics, and categorical fields like source and destination cities and cutoff flags. The target variable for regression is 'segment_actual_time'.

3. Ridge Regression Pipeline (Selected Model)

- Categorical columns ('source_city', 'destination_city', 'data') were encoded using OneHotEncoding.
- Numerical columns were scaled using StandardScaler to ensure uniform input ranges.
- Ridge Regression was applied with regularization parameter alpha = 1.0.
- The model was evaluated using R^2 score and RMSE.
- Final $R^2 \approx 0.54$ and RMSE ≈ 0.60, indicating moderate predictive performance.

4. Other Methods Tried (Code Not Included in the Final Notebook)

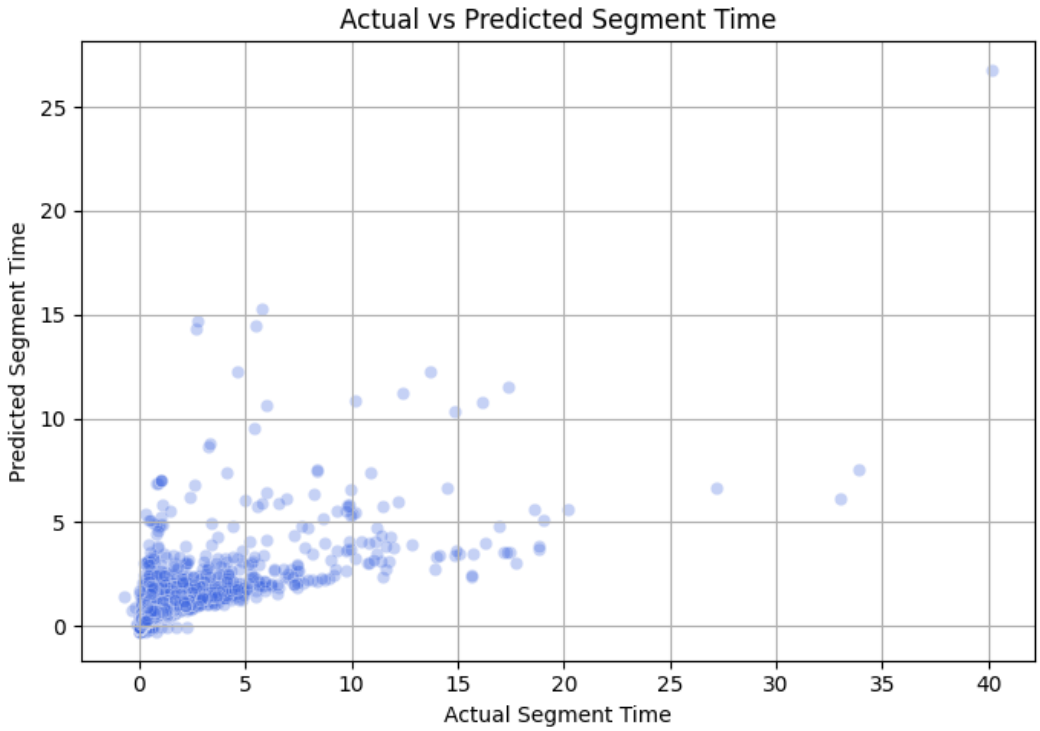
- A Ridge model with GridSearchCV-based alpha tuning was tested but showed only marginal improvements over alpha = 1.0.
- A version of Ridge Regression with log-transformed target (log1p on 'segment_actual_time') was also tested. However, it resulted in a severe performance drop ($R^2 = -90.96$, RMSE > 8), indicating that the log-transform was not suitable for this dataset.

5. Justification for Final Model

The final selected Ridge Regression model provides a solid trade-off between simplicity and effectiveness. It offers reasonable predictive performance without overfitting and is easy to interpret and maintain. Attempts to improve it with log-transformed targets or hyperparameter tuning did not result in substantial performance gains. Hence, the model with alpha = 1.0, OneHotEncoded categorical variables, and scaled numerical inputs was chosen for submission.

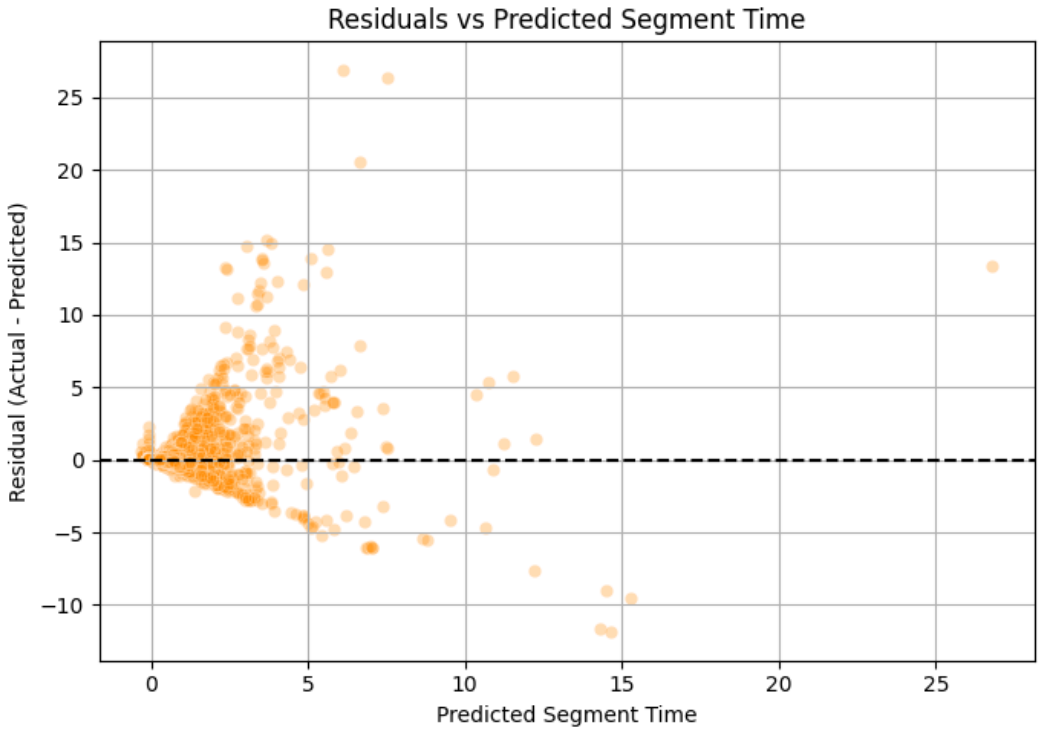
6. Model Visualizations and Interpretations

1. Actual vs Predicted Segment Time



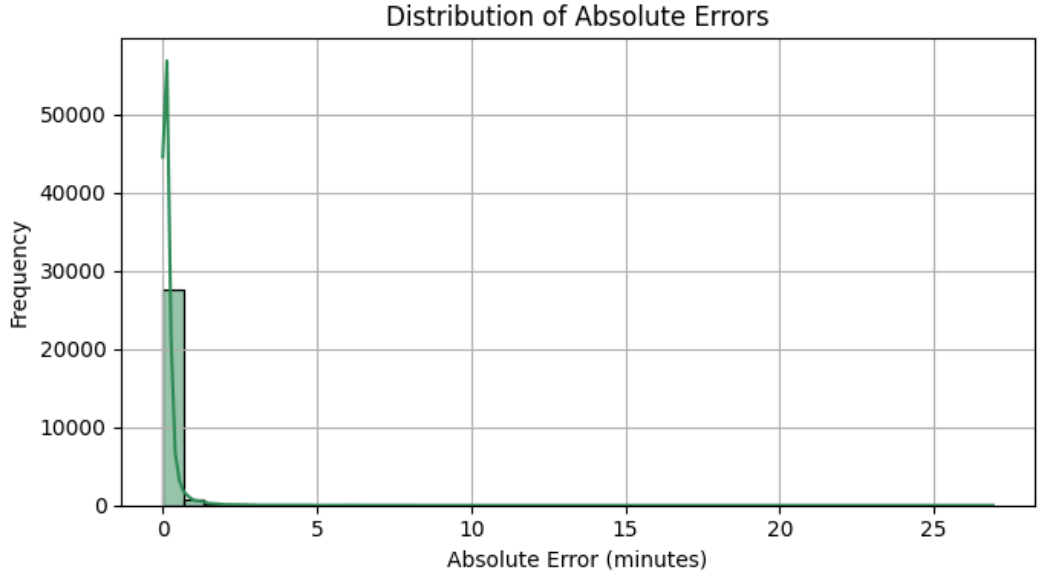
This scatter plot compares actual segment times with predicted values. A tighter cluster around the diagonal line would indicate higher model accuracy. Here, while there's a general upward trend, the spread shows variability—especially for larger actual times, where predictions tend to underestimate the true value.

2. Residuals vs Predicted Segment Time



This residual plot shows the prediction errors (Actual - Predicted) across predicted values. Ideally, residuals should be randomly dispersed around zero. In our model, there's slight funneling which suggests increasing variance at higher prediction levels—an indicator of model bias under higher segment times.

3. Distribution of Absolute Errors



The histogram shows that most absolute errors are within 2 minutes, suggesting the model performs acceptably for the majority of cases. However, a long tail exists, representing a small number of deliveries with high prediction error.