CS Get-Skilled Academy

Model Performance Report

New York City Trip Duration Prediction Project

Dr.Prof. Mostafa Saad

By: Sawsan Abdulbari

June 2024

1. Introduction

This report presents the findings of a comprehensive analysis and modeling exercise aimed at predicting NYC taxi trip durations. Key features were extracted through Exploratory Data Analysis (EDA) and various modeling approaches were compared to identify the optimal configuration. An ablation study was conducted to evaluate the impact of each feature on model performance.

2. Exploratory Data Analysis (EDA) Summary

EDA revealed several critical patterns and insights, which guided the feature engineering process. The most impactful features identified were:

- **Log Transformation:** Applied to trip duration to reduce skewness and stabilize variance, improving model robustness.
- **Time-Based Features:** Extracted hour, day of the week, month, and day from the pickup datetime to capture time-dependent patterns.
- **Geographical Features:** Calculated distances (Haversine and Manhattan) and directions between pickup and drop-off locations to capture spatial dependencies.
- **Trip Speed:** Calculated as the distance divided by the trip duration in hours, providing a measure of trip efficiency.
- **Airport Proximity:** Created binary features indicating proximity to major airports (JFK, LGA, EWR), which are significant trip destinations.

3. Model Training and Performance

Three primary modeling approaches were explored, culminating in an enhanced Ridge Regression model with additional features and hyperparameter tuning.

Approach 1: Basic Ridge Regression with Direct Features

Enhanced with **Time extraction** and **Log Transformation** for **Duration** (Target)

Train RMSE: 0.6233 Train R^2: 0.1202

Test RMSE: 0.5018 Test R^2: 0.4951

Cross-Validation RMSE: 0.6234 Cross-Validation R^2: 0.1200

Approach 2: One-Hot Encoding and Ridge Regression

Enhanced with **Distance** and **Direction**

Train RMSE: 0.3442 Train R^2: 0.7317

Test RMSE: 0.3440 Test R^2: 0.7343

Cross-Validation RMSE: 0.3549 Cross-Validation R^2: 0.7150

Approach 3: Enhanced Ridge Regression

with **Additional Features** such as Airport, speed and Log Transformation for Distance and Direction

Train RMSE: 0.1868 Train R^2: 0.9210

Test RMSE: 0.1861 Test R^2: 0.9223

Cross-Validation RMSE: 0.1868 Cross-Validation R^2: 0.9209

4. Ablation experiments

Ablation experiments performed to assess the contribution of each feature to the model's overall performance involved systematically removing each feature and evaluating the resulting impact on the model's metrics.

Impact of Key Enhancements:

Enhancement Feature Approach 3	**Train R^2**	**Test R^2**
Without Log Transformation and speed	0.5271	0.5301
Without Log Transformation and airport proximity	0.7018	0.7039
Without Log Transformation	0.7325	0.7351
Without Log Transformation airport proximity	0.9205	0.9218
Full Model	**0.9210**	**0.9223**

Note:

Test (unseen*) Data:

RMSE: 0.2019 **R^2: 0.8934**

*Unseen data used: split_sample/test.csv

5. Conclusion

The enhanced Ridge Regression model significantly outperformed simpler models by incorporating key features such as One-Hot Encoding, Log Transformation, Airport Proximity, and Trip Speed Calculation. Each enhancement was shown to contribute substantially to the model's performance, as evidenced by the ablation study.

The final model achieved high predictive accuracy with:

Train RMSE: 0.1868 **Train R^2: 0.9210**

Test RMSE: 0.1861 **Test R^2: 0.9223**

Cross-Validation RMSE: 0.1868 **Cross-Validation R^2: 0.9209**

By leveraging these insights, the final model is well-equipped for predicting NYC taxi trip durations with a high degree of accuracy.