Order Delivery Time Prediction - Data Analysis and Modeling Report

1. Introduction

This report summarizes the exploratory data analysis (EDA), feature engineering, model building, and evaluation steps undertaken to predict order delivery times for Porter orders. The goal is to identify key factors influencing delivery duration and develop an accurate predictive model.

2. Data Overview

The dataset contains information on Porter orders, including timestamps, restaurant categories, pricing, delivery personnel deployment, and distances. Key fields include created_at, actual_delivery_time, store_primary_category, order_protocol, total_items, and distance.

3. Data Preprocessing and Feature Engineering

3.1 Data Type Conversion

created_at and actual_delivery_time were converted from object to datetime types for temporal analysis.

Categorical fields such as store_primary_category and order_protocol were encoded as categories for efficient processing.

3.2 Feature Extraction

Delivery time (delivery_time_minutes) was calculated as the difference between delivery and order timestamps.

Temporal features: order_hour and isWeekend were extracted to understand order timing.

Unnecessary columns like created_at were dropped post feature creation.

3.3 Handling Outliers

Analysis via boxplots and IQR method identified outliers in delivery times, which were removed to improve model robustness. The number of removed outliers was approximately 1,749, ensuring cleaner data.

4. Exploratory Data Analysis (EDA)

4.1 Distribution of Features

Numerical features such as distance, total_items, and subtotal showed skewed distributions, with some variables like distance having positive correlation with delivery_time_minutes.

Categorical features, especially store_primary_category and order_protocol, displayed diverse counts, indicating varied order types.

4.2 Distribution of Target Variable

The delivery_time_minutes histogram revealed a right-skewed distribution, with most deliveries completed within 30-60 minutes but some long-tail outliers.

4.3 Relationships Between Variables

Scatter plots indicated a positive correlation between distance and delivery time.

Boxplots of delivery_time_minutes by hour showed peak durations during late-night and evening hours, highlighting time-of-day effects.

Correlation heatmap confirmed distance as the most influential numerical predictor.

4.4 Feature Selection

Features with weak correlation (below 0.1 to the target) such as min_item_price, store_primary_category, and actual_delivery_time were dropped, reducing multicollinearity and noise.

5. Model Building and Evaluation

5.1 Data Splitting and Scaling

The data was split into 80% training and 20% testing sets. Numerical features were scaled using StandardScaler, facilitating model convergence and interpretability.

5.2 Linear Regression Model

Initial linear regression achieved an MAE of approximately 2.34 minutes, MSE of 10.39, and RMSE of 3.22 on the test set, indicating reasonable prediction accuracy.

5.3 Feature Importance via Recursive Feature Elimination

Testing models with increasing features, the optimal set was found to be the top 5 features, providing a balance between simplicity and performance.

The most influential features included distance, total_outstanding_orders, and order_protocol.

5.4 Residual Analysis

Residual plots showed random dispersion around zero, suggesting that assumptions such as linearity and homoscedasticity hold. No significant patterns indicated the model's suitability.

6. Insights and Outcomes

The key driver of delivery time is distance, reflecting the importance of proximity on operational efficiency.

Order-related factors like total_outstanding_orders and order_protocol also significantly influence delivery duration.

Outliers in delivery times were effectively handled, improving model stability.

The chosen features and linear regression model provide a transparent, interpretable basis for future operational adjustments.

7. Visualizations

Distribution of delivery times: right-skewed, with most deliveries under 60 minutes.

Correlation heatmap shows distance has the highest positive correlation.

Residual plots confirm the model's assumptions and good fit.

8. Conclusion

The analysis demonstrates that delivery time can be reasonably predicted using factors like distance, order protocol, and current workload indicators. Incorporating these insights allows Porter to optimize delivery operations and enhance customer satisfaction.