Detailed Summary: Order Delivery Time Prediction

Order Delivery Time Prediction

OBJECTIVES

The primary goal is to build a regression model to predict the delivery time for food orders placed via Porter.

The model should help optimize operational efficiency and uncover key factors influencing delivery time.

KEY OBJECTIVES:

- Predict delivery time using multiple features
- Improve accuracy and efficiency in operations
- Interpret key drivers of delivery delay

DATA PIPELINE:

1. Data Loading
2. Data Preprocessing and Feature Engineering
3. Exploratory Data Analysis (EDA)
4. Model Building & Inference

DATA DESCRIPTION

Key columns include market_id, created_at, actual_delivery_time, store_primary_category, order_protocol, total_items, subtotal, num_distinct_items, min/max item price, total_onshift_dashers, busy dashers, outstanding orders, and distance.

FEATURE ENGINEERING:

- Converted timestamps to datetime format

- Calculated time taken in minutes (actual_delivery_time - created_at)

- Extracted hour and day of week, and created 'isWeekend' flag

- Dropped unnecessary columns to reduce multicollinearity

EDA HIGHLIGHTS:

- Identified feature distributions (numerical and categorical)

- Observed high correlation between 'distance', 'subtotal', and 'total_outstanding_orders' with time_taken

- Categorical effects visualized through countplots and boxplots

- Outliers handled using IQR method and filtered invalid data (e.g., negative distances)

MODEL BUILDING:

- Scaled features using MinMaxScaler

- Performed Recursive Feature Elimination (RFE) to select top features

- Built final linear regression model using statsmodels OLS

- Evaluated with RMSE and R (Test R ~ 0.5455)

TOP FEATURES IMPACTING DELIVERY TIME:

1. Distance (positive correlation)

2. Total Outstanding Orders (positive)

3. Num Distinct Items (positive)

4. Market ID dummies (some had negative correlation)

MODEL PERFORMANCE:

- Train RMSE: 0.160

- Test RMSE: 0.185

- Train R: 0.527

- Test R: 0.546

RESIDUAL ANALYSIS:

- Residual plots showed good spread and no major non-linear pattern

- QQ Plot confirmed normality in residuals

COEFFICIENT ANALYSIS:

- Distance had the highest positive scaled coefficient

- Market ID and Order Protocol dummies had significant negative impact

CONCLUSION:

The linear regression model explains around 54% of the variance in delivery time. Key influencing factors include order distance, outstanding orders, and item diversity. The model is interpretable and performs consistently across training and test sets.

This analysis provides a strong foundation for further optimization and deeper exploration into operational factors.