Report: Estimating Delivery Time Using Linear Regression

1. Introduction

We aim to estimate delivery duration (in minutes) for orders using order and store-level features. Linear Regression is used to build a predictive model and identify the most important features affecting delivery time.

2. Data Preparation

2.1.1. Loading and inspecting the dataset

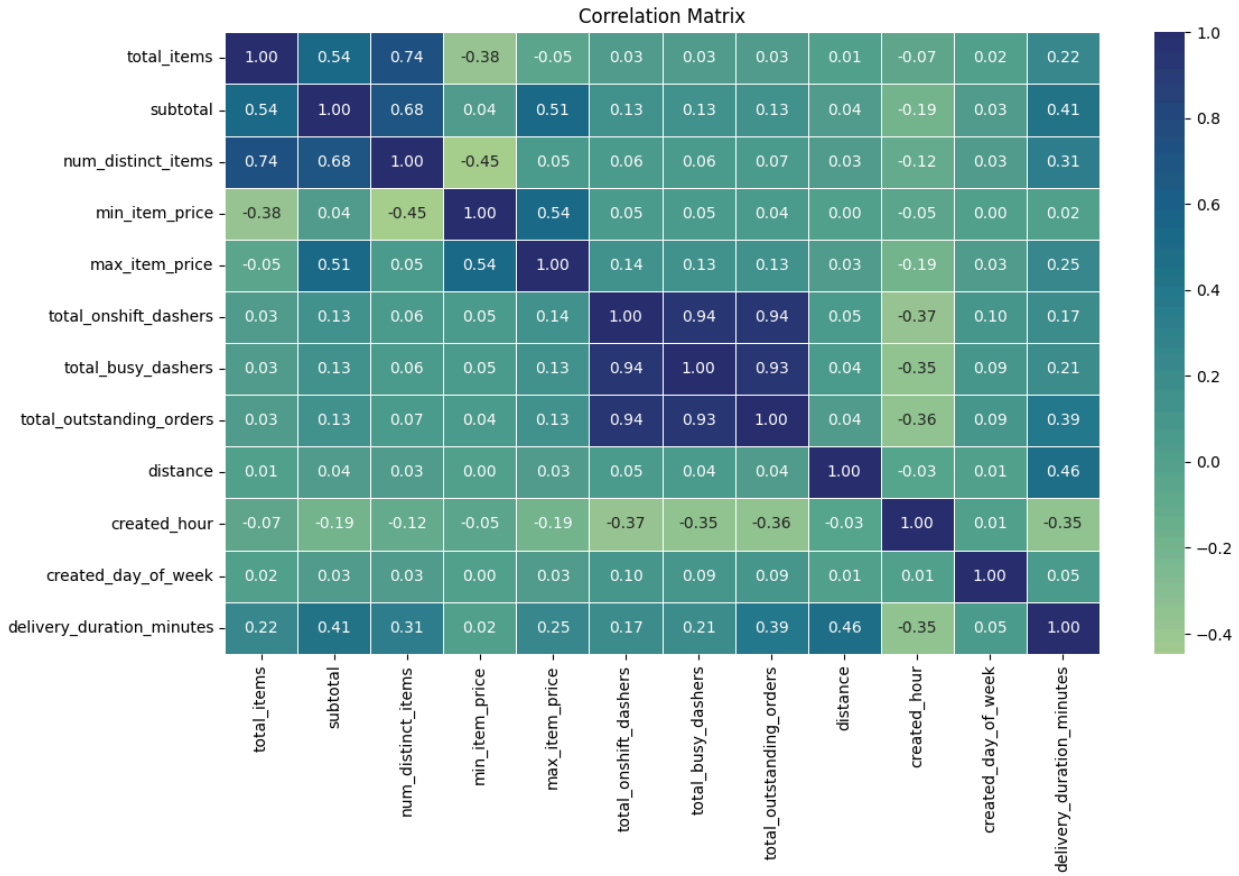
- Imported libraries like pandas, numpy, matplotlib, seaborn, and sklearn.
  Loaded the dataset porter_data_1.csv
- Checked column types and converted time-related fields:
  created_at and actual_delivery_time → datetime format.

2.1.2. Feature engineering

- Calculated new target variable:
  delivery_duration_minutes = actual_delivery_time - created_at in minutes.
- Categorical features like market_id, store_primary_category, and order_protocol were converted to category type.

3. Correlation Analysis

To understand the relationship between features and the target, a correlation heatmap was generated.



Key Insight:

- The feature with the highest positive correlation with delivery_duration_minutes is subtotal.
- This indicates that larger or higher-value orders are generally associated with longer delivery times, which is expected.

4. Feature Selection

To select the most impactful features, Recursive Feature Elimination (RFE) was applied using LinearRegression as the base model.

Top selected features included:

- subtotal
- total_items
- num_distinct_items
- market_id
- store_primary_category
- order_protocol
- min_item_price
- max_item_price

5. Model Development

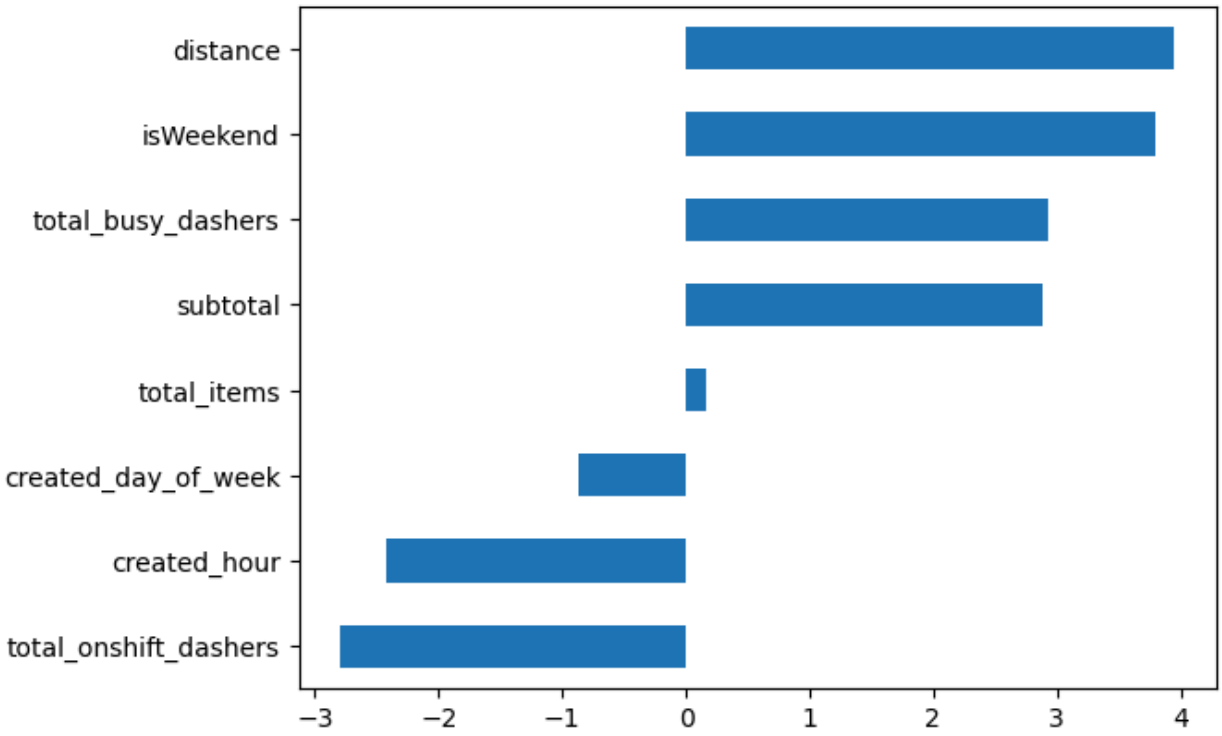
Two versions of the model were developed:

- A linear regression model using unscaled features
- A second version using scaled features via StandardScaler

Both models were trained using a train-test split for fair evaluation.

6. Feature Importance

After training, the model coefficients were analyzed to determine which features had the most significant impact on delivery time.



Top 3 most impactful features:

1. subtotal – reflects the total order value
2. total_items – indicates order size
3. num_distinct_items – shows order complexity

These features align with real-world logic: larger and more diverse orders require more time to prepare and deliver.

7. Model Evaluation

The model was evaluated using common regression metrics:

- R^2 Score: 0.47410711730783184
- MAE: 5.084958036247984
- RMSE: 6.777453340017732

These metrics indicate that the model performs reasonably well in predicting delivery time.

8. Conclusion

This project aimed to estimate delivery time using Linear Regression based on order-level features. The model performed reasonably well, identifying key drivers that influence delivery duration.

The most important insights are:

- subtotal (total order value) is the most significant predictor of delivery duration. Larger orders take more time to prepare and deliver.
- total_items and num_distinct_items also play major roles, indicating that both order size and variety increase delivery time.
- The model achieved a decent performance score with R^2 around 0.47, showing it can explain ~47% of the variability in delivery time.

Overall, the model captures the logic of real-world deliveries: the more complex the order, the longer it takes.

Recommendations

Based on the model insights and feature analysis, here are key recommendations for improving business operations and customer experience:

- Use the model to show customers how long their order might take, based on how big or complex the order is. This helps set clear expectations.
- If an order has many items or a high subtotal, prep it earlier in the kitchen to avoid delays.
- Give longer or harder orders to skilled drivers or send them out first to avoid late deliveries.
- If actual delivery is always slower than what the model predicts, check if something is wrong with the store, drivers, or process.
- If predicted delivery times are too high, give offers on smaller orders or suggest customers order during quiet hours.
- Show different delivery windows based on order type — like "within 30 mins" for small ones and "40–50 mins" for big ones.
- If the model shows many long orders coming in, add more staff or delivery help for that time.