Parcel Delivery Time Estimation using Linear Regression

Name: Prajith Jayarajan
Assignment 2 : LR

---

1. Problem Statement

The objective of this assignment is to build a regression model that predicts the delivery time for orders placed through Porter. The model will use various features such as the items ordered, the restaurant location, the order protocol, and the availability of delivery partners.

The key goals are:

- Predict the delivery time for an order based on multiple input features

- Improve delivery time predictions to optimiae operational efficiency

- Understand the key factors influencing delivery time to enhance the model's accuracy

---

2. Data Pipeline

The data pipeline for this assignment will involve the following steps:

1. **Data Loading**
2. **Data Preprocessing and Feature Engineering**
3. **Exploratory Data Analysis**
4. **Model Building**
5. **Model Inference**

---

3. Data Preprocessing & Feature Engineering

3.1 Fixing Data Types

- Converted timestamps to datetime format
- Transformed categorical variables: market_id, order_protocol, store_primary_category

3.2 Feature Engineering

- Extracted hour & day of week from created_at
- Created isWeekend feature
- Derived 4 categories morning, afternoon, evening and night from the order_hour column
- Calculated time_taken in minutes

3.3 Handling Missing Values

- Dropped rows with missing delivery timestamps
- Dropped order_hour, day_of_week feauture as they are now redundant

3.4 One-Hot Encoding

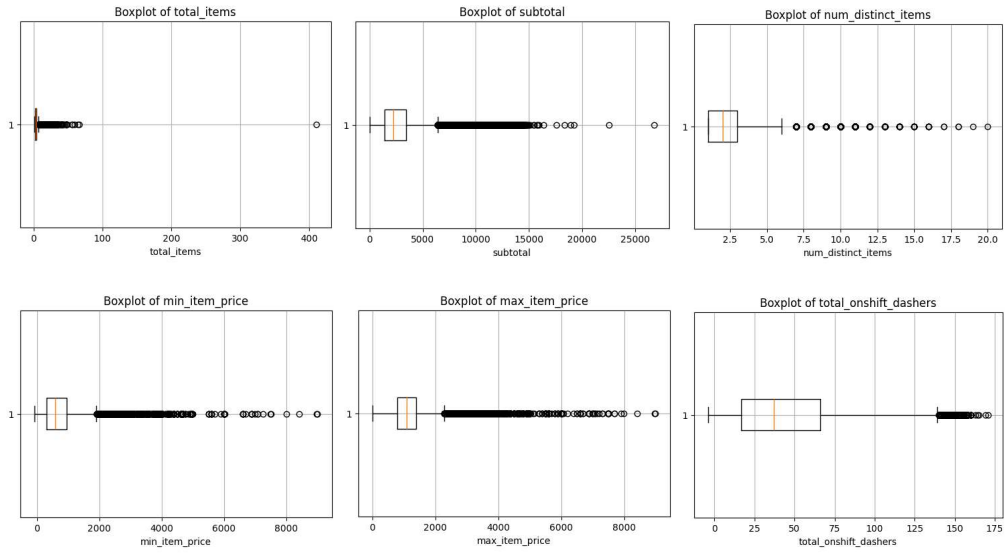
- Applied to categorical features
- Combined rare store categories under "Other" since there were 72 store_primary categories and so many dummy columns will lead to poor performance.

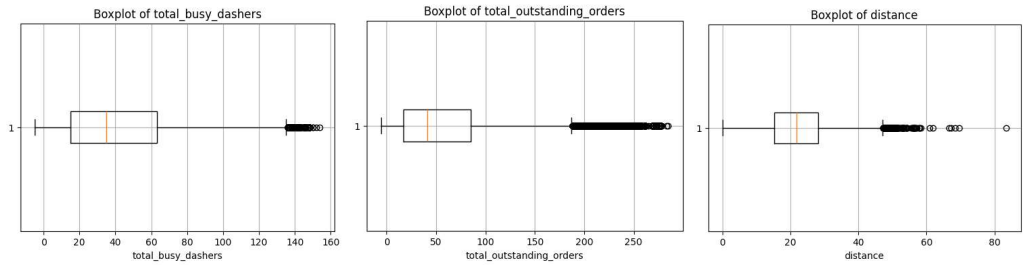
---

4. Exploratory Data Analysis (EDA)

4.1 Distributions

- Numerical columns visualized using boxplots

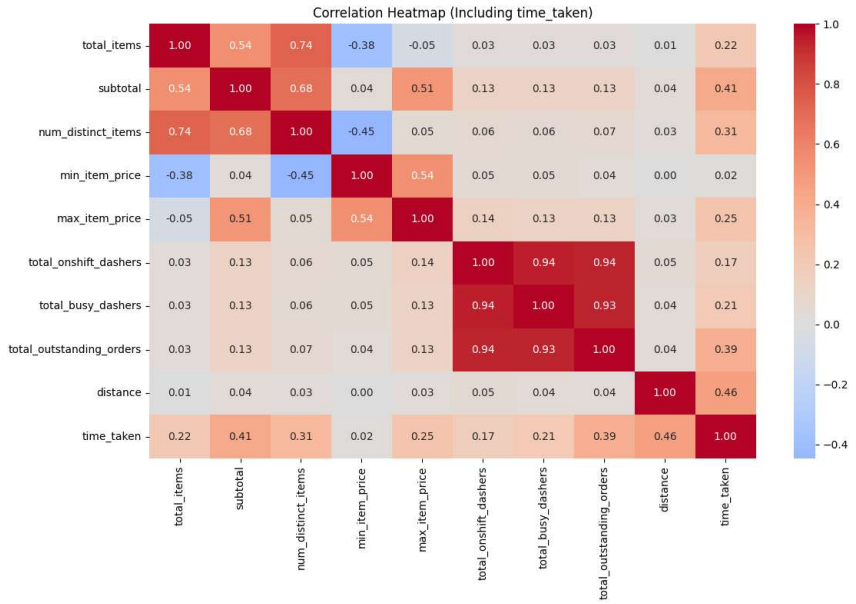




- time_taken mostly ranged between 10-60 mins with right skew

4.2 Relationships

- Moderate positive correlation of distance and subtotal with time_taken



4.3 Correlation Analysis

- Dropped features with correlation < 0.05 which was min_item_price
- Top features: total_outstanding_orders, distance, subtotal

4.4 Outlier Handling

- Applied IQR filtering to trim extreme values in numeric features

---

5. Model Building

5.1 Feature Scaling

Used StandardScaler on numerical columns

5.2 Model Training

- **Algorithm:** Linear Regression (from scikit-learn)
- **Train-Test Split:** 70:30

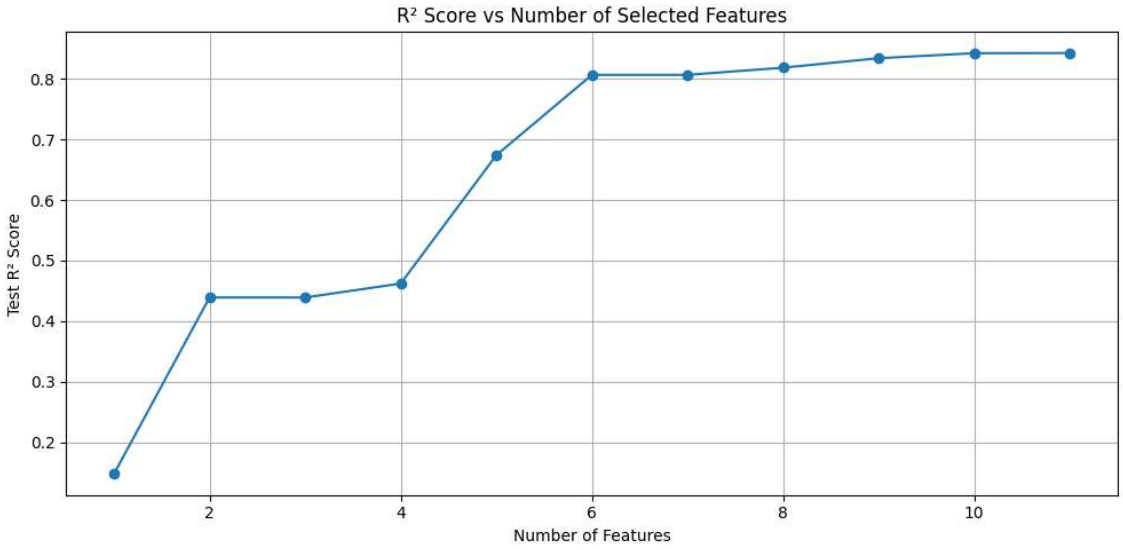
5.3 Model Performance

- **Train R^2:** 0.88.29
- **Test R^2:** 0.88.23
- **Test RMSE:** ~3.21 minutes

---

6. Recursive Feature Elimination (RFE)

- Used RFE to reduce features to most important subset
- Best performing model had 6 features, as awe can see in the graph below, additional features don't improve the model significantly.

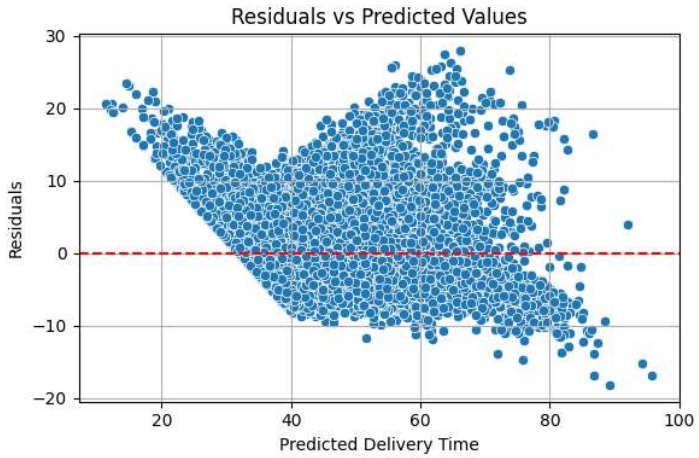


- However, performance slightly dropped (R^2 ~0.81),

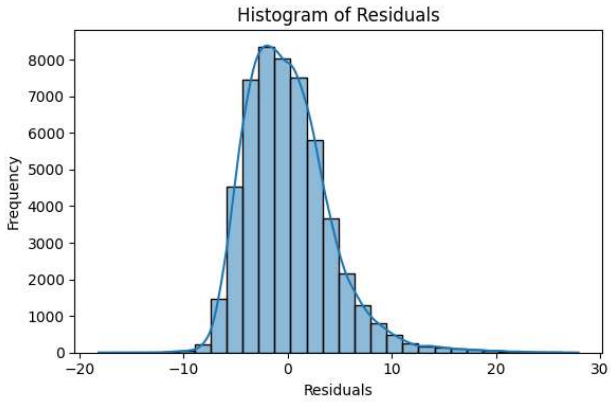
7. Residual & Coefficient Analysis

7.1 Residual Plots

- The residuals are not evenly distributed around zero, except there is a funnel shape - that is narrow for lower delivery times and it gets wider later, This is a sign of heteroscedasticity — the variance of errors increases with prediction magnitude.



- The histogram plot shows that the residuals are normally distributed with peak a bit left of zero and also it has a slight right skew.



7.2 Coefficients

- Coefficient of 6 most important features selected using RFE:

	Feature	Coefficient (scaled)
0	subtotal	3.438563
1	total_onshift_dashers	-12.109860
2	total_busy_dashers	-4.756545
3	total_outstanding_orders	18.751106
4	distance	4.178832
5	store_primary_category_3	7.293929

- Original std deviation of subtotal: 1832.80
- A unit increase in subtotal by std (1832.80) increases delivery time by 3.438 mins (highlighting outlier influence)

---

8. Categorical Variables Insight

- Categorical variables were one-hot encoded
- Some market regions and store categories significantly impacted delivery time

---

9. Summary of Insights

- Delivery time is impacted significantly by operational backlog and distance
- Subtotal affects delivery time but has high variance
- Categorical regions influence model and should not be excluded

---

10. Assumptions

- Outliers were treated using IQR method
- One-hot encoding used for categorical variables
- StandardScaler used only on numeric features

---

11. Recommendations

- Prioritize reducing outstanding orders to improve efficiency
- Use distance as a key predictor in route planning
- Regularly update the model with new data to capture operational changes

SUBJECTIVE QUESTIONS:

Q1) Are there any categorical variables in the data? From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Yes, There are several categorical features, But one of them made it to selected feautures, the store_primary_category_3, the store_primary_category 3 takes longer delivery time according to analysis.

Also It is important to note that categorical variables weren't scaled or standardized.

Q2) What does `test_size = 0.2` refer to during splitting the data into training and test sets?

Answer: If test_size = 0.2, that means 20 percent of data will be used as a validation set, this will not be used for standardizing, EDA or model training

Q3) Looking at the heatmap, which one has the highest correlation with the target variable?

Answer: 'Distance' has highest corr with time_taken

Q4) What was your approach to detect the outliers? How did you address them?

Answer: we did a boxplot to first visualize the outliers, then we drop all the outliers which is 1.5 times std dev away from IQR

Q5) Based on the final model, which are the top 3 features significantly affecting the delivery time?

Answer: Based on the model coefficients, the top three features affecting delivery time are: subtotal, total_onshift_dashers, and total_busy_dashers are the most important 3 features

Q6) Explain the linear regression algorithm in detail

Answer: Linear Regression is a supervised learning algorithm used for predicting a continuous output based on one or more input features.

It models the relationship between the independent variables (X) and the dependent variable (y) by fitting a linear equation to the data.

A best fit line is to be found, that is the coefficients of features need to be found, this is done using gradient descent.

The cost-function used in LR is mean of square of residues, This function tells us how well the line fits the data points, minimizing this cost-function will give the best-fit line. To minimize it computationally, gradient decent algorithm is used, which works as follows:

1)The weights are initialized with random value

2)The cost function is calculated, Also we find the the derivatives of cost functions with respect to the weights

3)The weights are updated as follows w' = w - alpha* d where d is the deivative of cost function wrt w and alpha is learning rate

Q7) Explain the difference between simple linear regression and multiple linear regression

Answer: Simple LR has only one independent feature while multiple LR can have many.

The underlying assumptions remain same, except a few additional assumptions are required for multiple LR, that there should be no multicollinearity between features.

Q8) What is the role of the cost function in linear regression, and how is it minimized?

Answer: The cost-function used in LR is mean of square of residues, This function tells us how well the line fits the data points, minimizing this cost-function will give the best-fit line. To minimize it computationally, gradient decent algorithm is used, which works as follows:

The weights are initialized with random value

The cost function is calculated, Also we find the the derivatives of cost functions with respect to the weights

the weights are updated as follows w' = w - alpha* d where d is the deivative of cost function wrt w and alpha is learning rate

Q9) Explain the difference between overfitting and underfitting.

Answer: Overfitting is when a model memorizes the training data, In technical terms, it takes advantage of increased complexity to forcefully reduce the cost function, this

results in model not generalizing well, and the acuracy drops while trying to predict unseen data.

Underfitting is when the model is to simple to capture the pattern.

Q10) How do residual plots help in diagnosing a linear regression model?

Answer: Linear regression needs certain assumptions to be valid, They are:

1) The residuals should add upto zero, that is if we plot the residuals, they should be approximately balanced around the zero line

2) The variance should not increase or decrease with the features; this also can be confirmed using a plot