Porter Delivery Time Estimation

Abstract:

In this article, we cover the construction of a model to predict the relationship between order information and delivery time using the Porter Delivery Time Estimation dataset. To accomplish this, we employed two approaches.

Firstly, we opted for a regression model to predict the continuous target variable (delivery time).

Secondly, we chose a classification model to predict the categorical target variable (fast, moderate, slightly slow, slow) after converting it.

In this article, we will cover the second approach, which involves handling the Porter Delivery Time Estimation dataset using a classification model.

To achieve this, we created a new target variable for delivery time ("How_Long") by utilizing "created_at" and "actual_delivery_time" and established the following criteria to convert this variable into categorical data.

1 to 30 minutes: Fast

30 to 60 minutes: Average.

60 to 90 minutes: Slightly Slow.

More than 90 minutes: Slow.

By categorically transforming the target variable in this manner, a classification model was constructed. This Model is expected to provide customers with intuitive delivery time information for new orders and contribute to efficiently managing the delivery process.

1.1 Objective:

Porter has a number of delivery partners available for delivering the food, from various restaurants and wants to get an estimated delivery time that it can provide the customers on the basis of what they are ordering, from where and also the delivery partners.

This dataset has the required data to train a regression model that will do the delivery time esimation, based on all those features

1.2 Challenges:

□ **Missing Values**:

- The store_primary_category column has missing values.
- It's important to check other columns for missing values as well, as they can affect analysis and model performance.

□ **Data Types**:

- Ensure that columns like created_at and actual_delivery_time are properly parsed as datetime objects.
- Other columns might need type conversion for accurate analysis.

□ **Consistency**:

- Check for inconsistencies in categorical data such as store_primary_category.
- Ensure that numerical values are within expected ranges (e.g., no negative prices).

□ **Outliers**:

- Identify and handle outliers in columns like subtotal, min_item_price, max_item_price, etc.

□ **Feature Engineering**:

- Calculate new features like delivery time (actual_delivery_time - created_at).
- Aggregate data by categories, time periods, etc., for more insightful analysis.

□ **Duplicates**:

- Check for duplicate entries that might skew the results.

□ **Distribution of Data**:

- Analyze the distribution of data to ensure that there are no biases or skewed distributions which could affect modeling.

□ **Correlation**:

- Investigate the correlation between different features to identify multicollinearity which might affect model performance.

1.3 Real World Scenario:

□ **Improved Operational Efficiency**:

- By analyzing delivery times and identifying patterns, companies can optimize delivery routes and schedules, reducing delivery times and operational costs.
- Understanding peak times for orders can help in better workforce management, ensuring the right number of partners are on shift.

□ **Enhanced Customer Experience**:

- Faster and more reliable deliveries can lead to increased customer satisfaction.
- Identifying and addressing issues with specific stores or categories can improve overall service quality.

□ **Strategic Decision Making**:

- Insights into popular categories and order patterns can inform inventory and menu decisions for stores.

- Data-driven decisions can be made about expanding to new markets or adjusting offerings based on customer preferences.

□ **Revenue Growth**:

- By optimizing pricing strategies and promotions based on order data, businesses can potentially increase their revenue.
- Identifying high-value customers and targeting them with personalized offers can enhance customer loyalty and drive sales.

□ **Risk Management**:

- Analyzing data can help in identifying potential risks such as late deliveries or operational bottlenecks.
- Proactively addressing these risks can prevent loss of customers and reputational damage.

□ **Sustainability**:

- Optimizing delivery routes and reducing the number of trips can contribute to lower fuel consumption and reduced carbon emissions.
- Efficient resource management leads to sustainable business practices.

□ **Competitive Advantage**:

- Companies that leverage data analysis effectively can gain a competitive edge by providing better services and making more informed business decisions.
- Staying ahead of market trends and customer expectations can set a business apart from its competitors.

□ **Policy and Planning**:

- For urban planners and policymakers, data on delivery patterns can provide insights into traffic management and urban infrastructure needs.
- Informing policies related to local business regulations and support.

2.Data Fields:

market_id : integer id for the market where the restaurant lies

created_at : the timestamp at which the order was placed

actual_delivery_time : the timestamp when the order was delivered

store_primary_category: category for the restaurant

order_protocol : integer code value for order protocol(how the order was placed ie: through porter, call to restaurant, prebooked, third part etc)

total_items : no. of items

subtotal : final price of the order

num_distinct_items : the number of distinct items in the order

min_item_price : price of the cheapest item in the order

max_item_price : price of the costliest item in order

total_onshift_dashers : number of delivery partners on duty at the time order was placed

total_busy_dashers : number of delivery partners attending to other tasks

total_outstanding_orders : total number of orders to be fulfilled at the moment

estimated_store_to_consumer_driving_duration : approximate travel time from restaurant to customer

2.1 **Data Understanding & Tools:**

Data comes from a Kaggle competition so it can be downloaded directly for the solution but if we want to productionize the live data we might have to make a data pipeline for the same. Cloud solutions and SQL queries for data pipelines are very commonly seen in companies which can be used effectively. For this particular instance we can use Pandas and Numpy libraries to process the data as we have data in CSV format. As the data is company specific additional data can be acquired by having business understanding of the same.

3.Real world challenges and constraints:

1. Data Quality Issues:

- **Incomplete Data**: Missing values in critical columns can lead to incomplete analysis and inaccurate predictions.
- **Inconsistent Data**: Variations in data entry (e.g., different naming conventions for the same category) can complicate analysis.
- **Outliers**: Extreme values can skew analysis results and lead to incorrect conclusions.

2. Data Integration:

- **Multiple Data Sources**: Combining data from different sources can be complex, especially if the data is not standardized.
- **Data Silos**: Relevant data might be stored in isolated systems, making it difficult to get a comprehensive view.

3. Data Privacy and Security:

- **Regulatory Compliance**: Adhering to data protection regulations (e.g., GDPR, CCPA) is crucial when handling customer data.
- **Data Security**: Ensuring that sensitive information is protected against breaches is essential for maintaining customer trust and compliance.

4. Scalability:

- **Data Volume**: Large datasets require efficient storage, processing power, and sophisticated analysis techniques.

- **Real-Time Analysis**: For some applications, such as dynamic pricing or real-time delivery optimization, the ability to process and analyze data in real time is critical.

5. Interpretability:

- **Complex Models**: Advanced analytical models (e.g., machine learning) can be difficult to interpret and explain to non-technical stakeholders.
- **Actionable Insights**: Translating data analysis into actionable business insights requires a deep understanding of both the data and the business context.

6. Resource Constraints:

- **Technical Expertise**: Skilled data scientists, analysts, and engineers are required to perform high-quality analysis, and they may be in short supply.
- **Financial Resources**: Investing in necessary technology, software, and personnel can be costly.

7. Operational Constraints:

- **Integration with Existing Systems**: Implementing insights from data analysis into existing operational systems can be challenging.
- **Change Management**: Adapting to new data-driven processes and decision-making frameworks can meet resistance within the organization.

8. Data Interpretation Challenges:

- **Biases in Data**: Historical data might contain biases that, if not addressed, can lead to biased insights and decisions.
- **Dynamic Environments**: The business environment is constantly changing, and historical data might not always be a reliable predictor of future trends.

9. Ethical Considerations:

- **Fairness and Equity**: Ensuring that data analysis does not lead to unfair treatment of any customer segment is crucial.

- **Transparency**: Maintaining transparency about how data is used and how decisions are made is important for building trust.

10. External Factors:

- **Market Conditions**: Changes in market conditions (e.g., economic downturns, new competitors) can affect the relevance of the insights derived from the data.
- **Regulatory Changes**: New regulations can impact how data can be used and what types of analyses are permissible.

4. Solutions to similar problems:

4.1 Solution Approach and Problem Type:

1. Logistic Regression
2. Decision Trees
3. Random Forest Classifier
4. Gradient Boosting Machines(GBM)
5. XGBoost
6. Light GBM
7. Ensemble – Hard Voting

5. References:

www.kaggle.com

Google

ChatGPT

BIA recording

Self prepared notes

Online possible references.

Phase 2 : EDA and Feature Extraction

EDA and Feature extraction to understand the intricacies of the Data.

We will explore the data using this steps :-

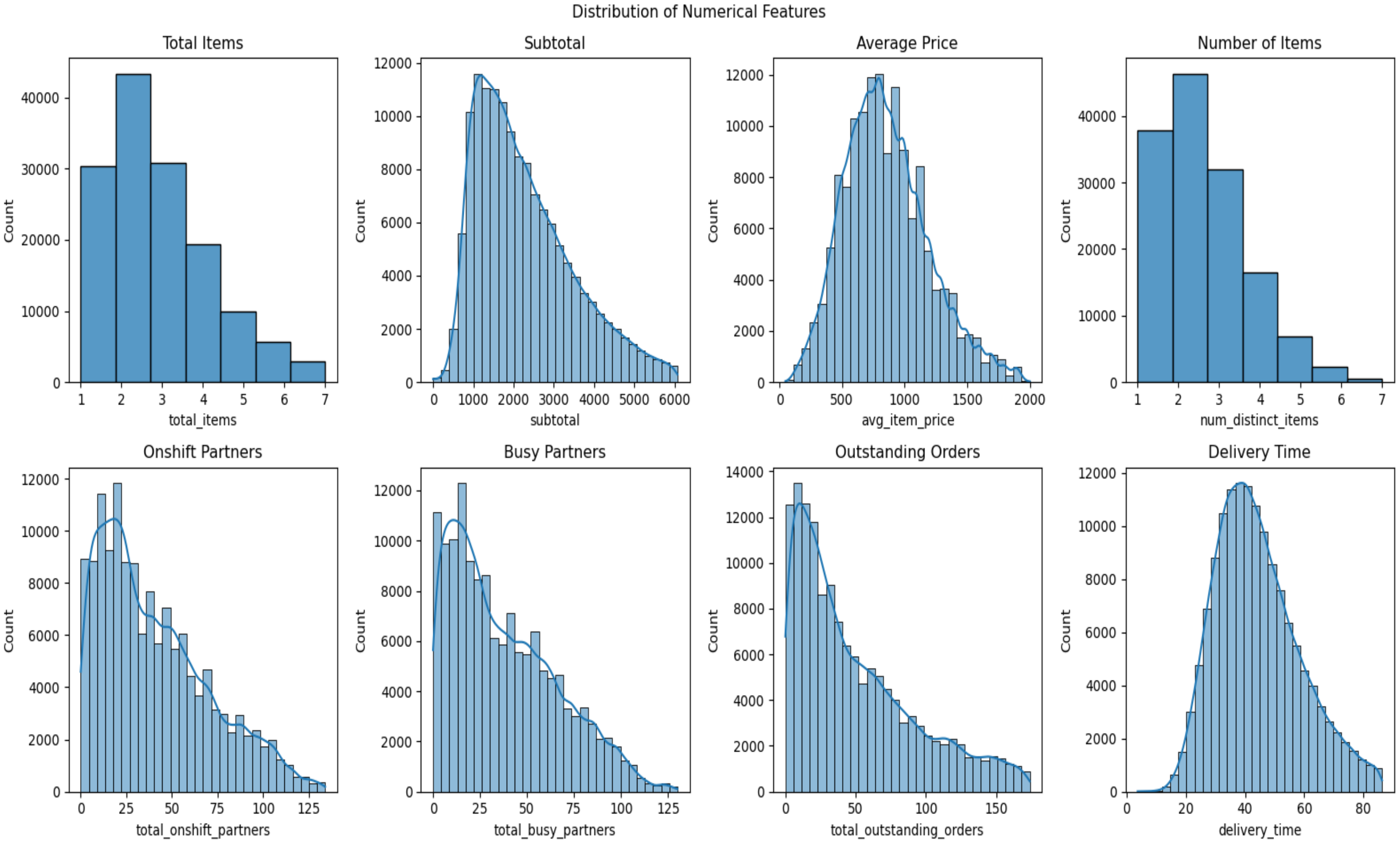
1. Porter Data understanding and insights

2. Data Cleaning or Manipulation

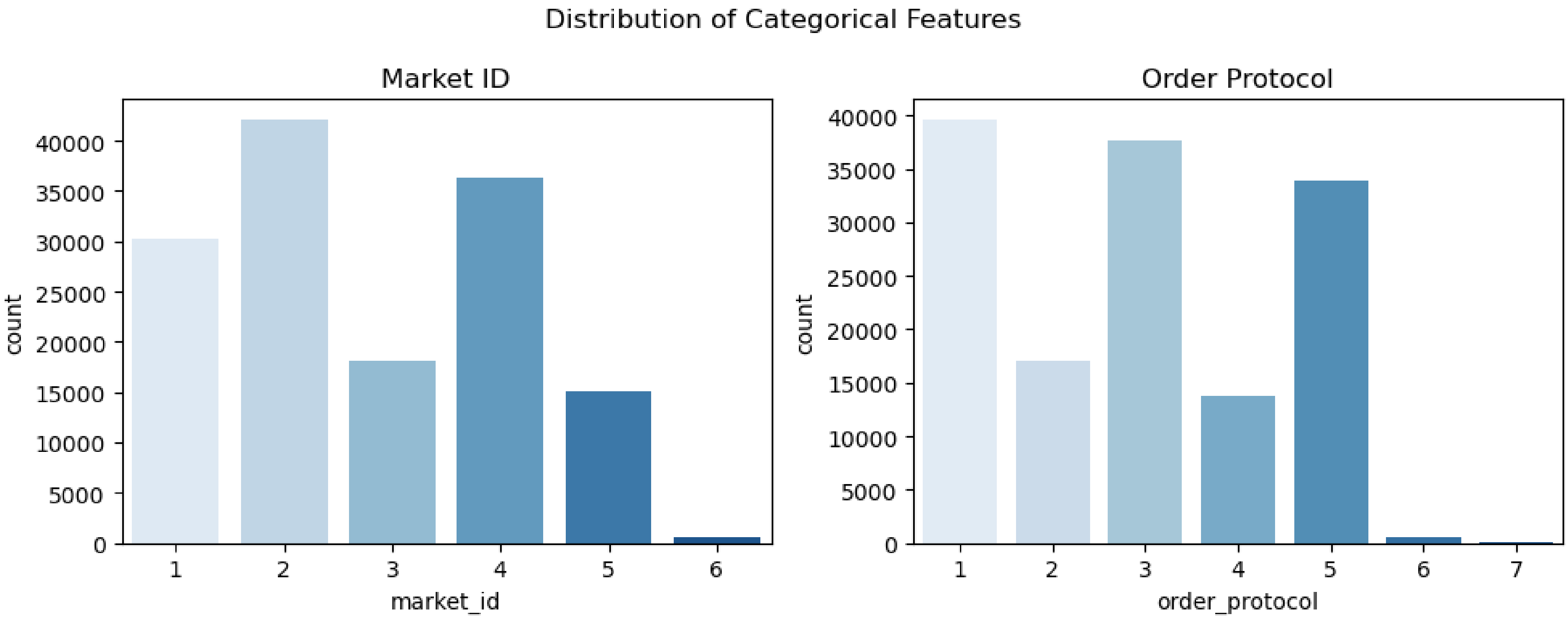
3. Data PreProcessing

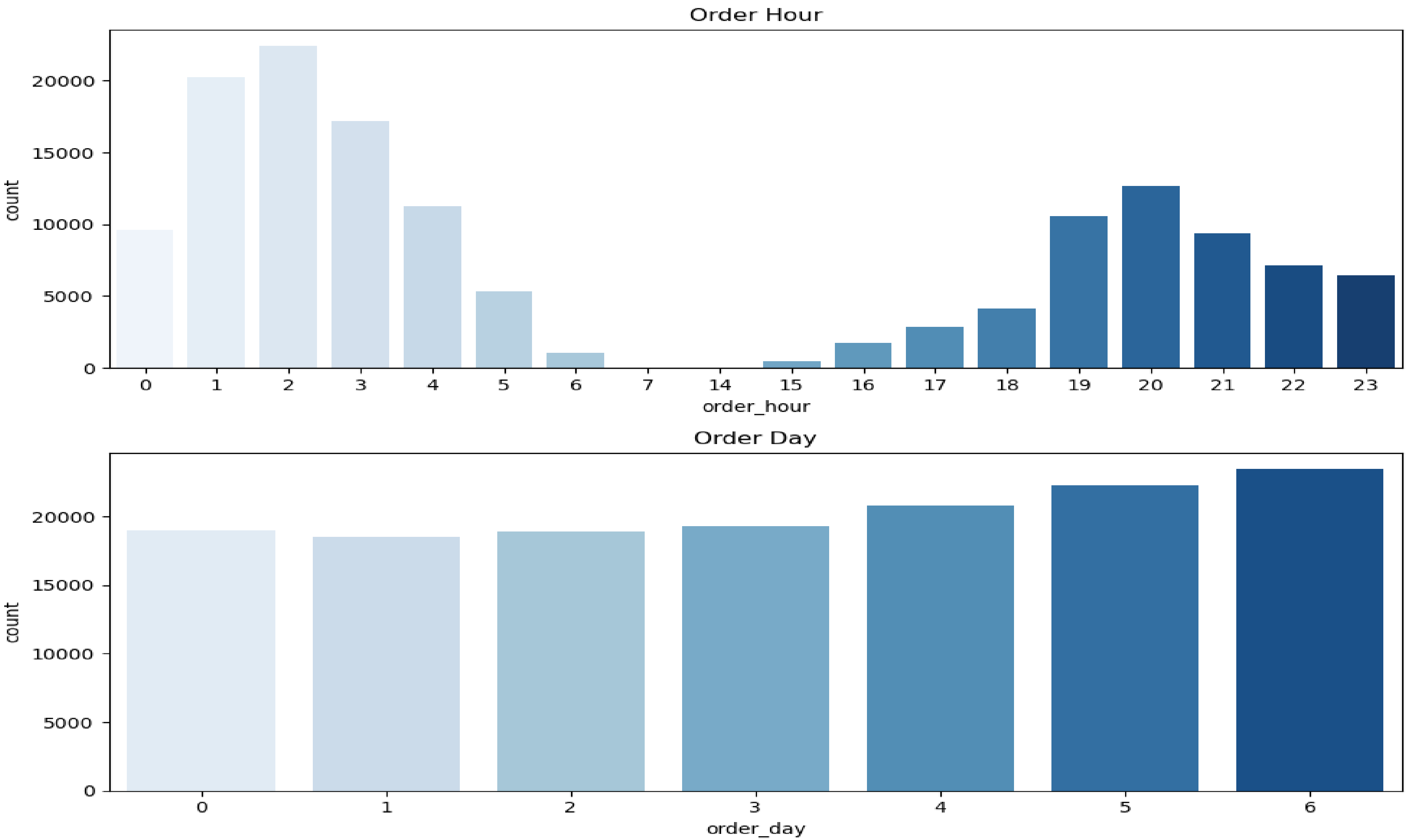
6.Feature Engineering

7.Checking the accuracy_score

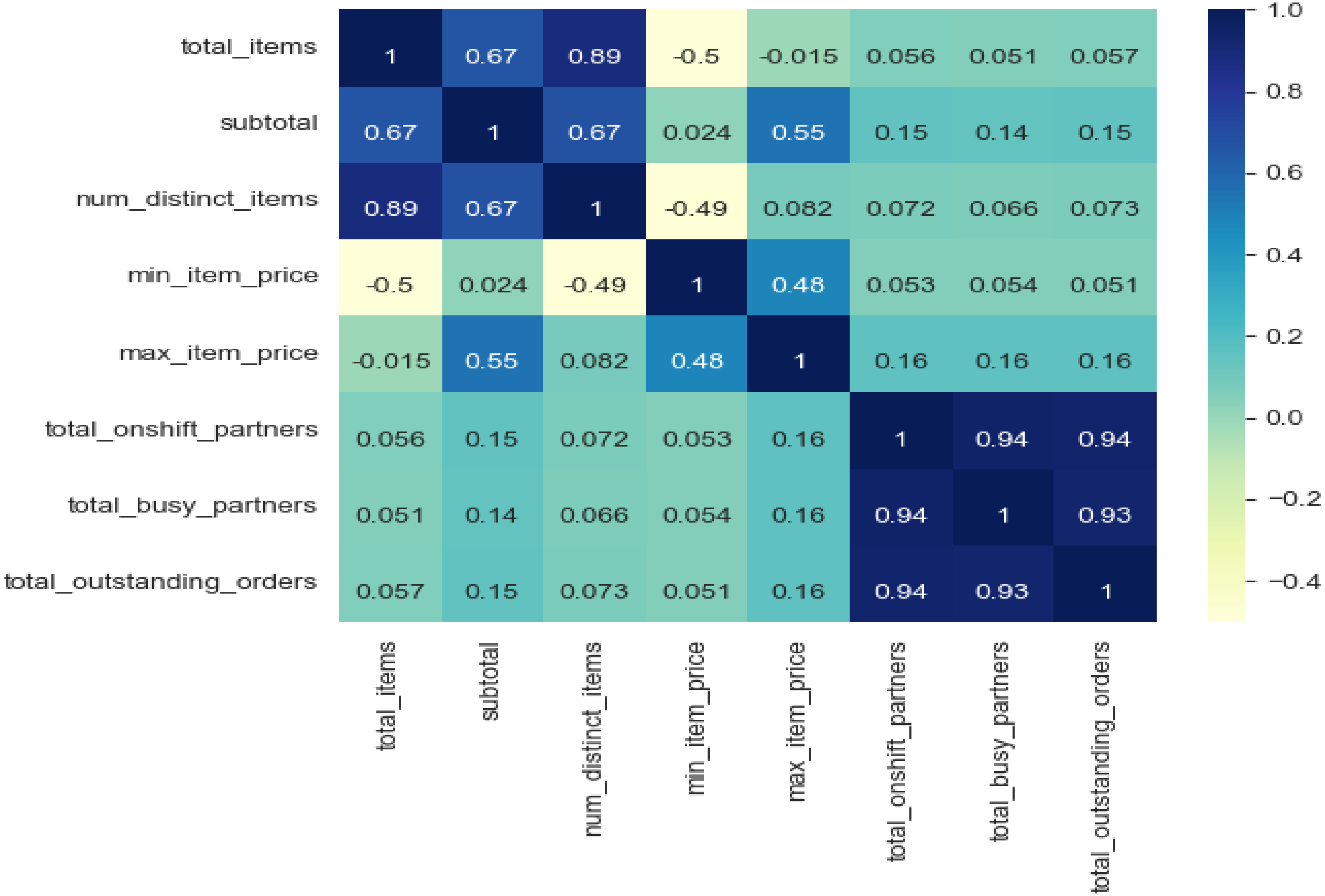


It can be observed that all of the numerical features have a right skew due to presence of outliers


Distribution of Categorical Features
Market ID
count
40000
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
market_id
Order Protocol
count
40000
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
order_protocol



Saturdays and Sundays have the highest order volume
Highest volume of orders is between 1am-4am. There are no orders between 8 am to 2 pm



	total_items	subtotal	num_distinct_items	min_item_price	max_item_price	total_onshift_partners	total_busy_partners	total_outstanding_orders
total_items	1	0.67	0.89	-0.5	-0.015	0.056	0.051	0.057
subtotal	0.67	1	0.67	0.024	0.55	0.15	0.14	0.15
num_distinct_items	0.89	0.67	1	-0.49	0.082	0.072	0.066	0.073
min_item_price	-0.5	0.024	-0.49	1	0.48	0.053	0.054	0.051
max_item_price	-0.015	0.55	0.082	0.48	1	0.16	0.16	0.16
total_onshift_partners	0.056	0.15	0.072	0.053	0.16	1	0.94	0.94
total_busy_partners	0.051	0.14	0.066	0.054	0.16	0.94	1	0.93
total_outstanding_orders	0.057	0.15	0.073	0.051	0.16	0.94	0.93	1

total_onshift_partners, total_busy_partners and total_outstanding_orders have a high correlation.

Similarly, total_items, subtotal and num_distinct_items also have a high correlation

Conclusion

We created a target variable 'How Long' and converted it into a categorical variable.

After running predictions on the classification models, we observed that the Light GBM Classifier yielded the highest accuracy among the models.

Following that, the XGB Classifier and Random Forest Classifier achieved around 50% accuracy, while other models notably exhibited lower accuracy.

Based on these results, we concluded that applying the Light GBM Classifier model to the Porter Delivery Time Estimation dataset is the most suitable choice.

Furthermore, comparing the results of predicting the 'How Long' variable using the first method as a continuous variable, we found that the performance was even better when treating it as a categorical variable.

This suggests that predicting with categorical values yields higher accuracy than predicting precise continuous values, thus confirming our hypothesis.