HUMBER INSTITUTE OF TECHNOLOGY

AND ADVANCED LEARNING

BIA-5401-0LA

Final Group Project Report

Precision with Business Intelligence: Food Delivery Logistics

Ansh Gangdev

Farman Zaidi

Jeet Patel

Kivan Ilangakoon

Palak Singla

Shyam Jigneshbhai Soni

Submitted To: Professor Haytham Qushtom

Table of Contents

Executive Summary

This report analyzes data from an online food delivery platform, highlighting customer preferences, order characteristics, and market trends. The main challenge is optimizing partner resource allocation to meet varying order demands and enhance efficiency. The dataset covers order details, timestamps, store information, and order metrics from early 2015. The proposed solution involves data-driven visualization and a linear regression model to predict delivery times.

Key findings include insights into order patterns, customer behavior, and market trends. The linear regression model accurately estimates delivery times, aiding in resource allocation. Benefits encompass improved decision-making, operational efficiency, enhanced customer satisfaction, and data-driven strategies. Challenges include data quality, model complexity, biases, and scalability. Recommendations involve real-time resource allocation and dynamic pricing strategies. Overall, the analysis provides actionable insights to enhance delivery processes and optimize partner resources.

Introduction

The dataset captures information on the market dynamics within an online food delivery platform called Porter which is one of India's Largest Marketplace for Intra-City Logistics . This dataset contains information related to various food orders placed through the platform, offering insights into customer preferences, order characteristics, and market behaviour. The data spans a period from early 2015 and includes details about individual orders, such as creation and delivery timestamps, store information, food categories, and order metrics [1].

Problem Statement

One of the challenges faced by the online food delivery platform is to optimize partner resource allocation and manage their availability efficiently to meet varying order demands. Inconsistent distribution of partner resources can lead to delivery delays, increased waiting times, and customer dissatisfaction. To address this issue, the platform needs to identify peak demand periods, allocate partners accordingly, and ensure that partners are efficiently utilized without overwhelming them [5].

Data Insights:

1. Market and Store Information: Each order is associated with a unique market identifier and a specific store. The market identifier helps to segment the dataset based on different market segments or locations. Each store has a primary category, providing an indication of the type of cuisine they offer. This categorization is essential for understanding the variety of food options available to customers.

2. Order Metrics: The dataset offers valuable insights into order characteristics. It includes details such as the number of items in an order, the total bill amount (subtotal), the number of distinct items ordered, and the range of item prices (from minimum to maximum). These metrics provide a comprehensive view of customer preferences, spending behavior, and the variety of items customers typically include in their orders.

3. Order Timing and Protocol: Timing plays a crucial role in the food delivery industry. The dataset captures the creation and actual delivery timestamps for each order. This information can help in understanding peak order times and delivery efficiency. Additionally, the order protocol indicates how orders are being managed, reflecting the operational dynamics of the platform.

4. Partner and Busy Information: The dataset contains information about the on-shift partners and their availability. This aspect is important to analyze partner workload and assess how effectively the platform manages its workforce. The count of busy partners and outstanding orders gives insight into the platform's capacity and potential challenges in handling high-demand periods.

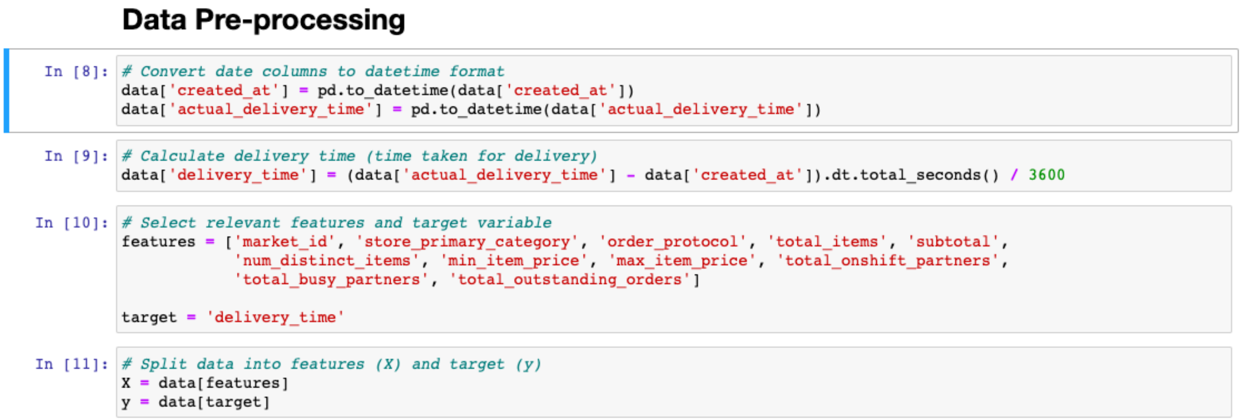
5. Visualization and Analysis: The dataset has been subjected to visualization using Jupyter notebooks. These visualizations likely provide graphical representations of data trends, patterns, and relationships. These visual insights offer a quick overview of potential insights that can be gained from further analysis.

This dataset offers a window into the online food delivery ecosystem, shedding light on customer preferences, market dynamics, and operational efficiency. Analyzing this data could reveal patterns and trends that can inform strategic decision-making for the platform, including optimizing delivery processes, curating menus, and managing partner resources more effectively. It serves as a valuable resource for understanding consumer behaviour and market trends in the context of the food delivery industry [5].

Solution (and Implementation)

The dataset lay a solid groundwork for implementing a data-driven visualization strategy aimed at optimizing partner resource allocation within the realm of food delivery. Capitalizing on these datasets, a pragmatic approach emerges to address challenges like peak order times and managing partner workloads effectively. By visualizing how orders are distributed throughout different hours of the day and days of the week, platforms can accurately pinpoint peak periods for optimal partner deployment, ultimately resulting in quicker and more dependable deliveries. Moreover, insights derived from visualizing market-specific trends and correlating partner availability offer a more intelligent resource allocation strategy. This comprehensive methodology bolsters customer experiences by curbing delivery times, elevating satisfaction levels, and contributing to heightened retention rates [5].

The implementation process harnessed the power of data-driven visualizations to optimize partner resource allocation within the food delivery ecosystem. Leveraging the provided datasets, the strategy unfolded by scrutinizing historical order data and partner availability patterns. By translating data into intuitive visual representations, the approach illuminated crucial insights such as peak order hours, weekly trends, market-specific demands, and partner workload dynamics. These visualizations enabled delivery platforms to make informed decisions, strategically assigning partners during high-demand periods, adapting to changing customer needs, and ensuring efficient utilization of available resources. The successful implementation of this visualization strategy has undoubtedly contributed to improved delivery efficiency, elevated customer satisfaction, and a more streamlined partner allocation process. These steps are broken down below [5].



Data Pre-processing

```
In [8]: # Convert date columns to datetime format
        data['created_at'] = pd.to_datetime(data['created_at'])
        data['actual_delivery_time'] = pd.to_datetime(data['actual_delivery_time'])

In [9]: # Calculate delivery time (time taken for delivery)
        data['delivery_time'] = (data['actual_delivery_time'] - data['created_at']).dt.total_seconds() / 3600

In [10]: # Select relevant features and target variable
         features = ['market_id', 'store_primary_category', 'order_protocol', 'total_items', 'subtotal',
                     'num_distinct_items', 'min_item_price', 'max_item_price', 'total_onshift_partners',
                     'total_busy_partners', 'total_outstanding_orders']

         target = 'delivery_time'

In [11]: # Split data into features (X) and target (y)
         X = data[features]
         y = data[target]
```

In the data pre-processing phase, several transformations and calculations were applied to the online food delivery dataset. Here's a breakdown of the steps took and their significance:

1. Convert Date Columns to Datetime Format: Converted the "created_at" and "actual_delivery_time" columns to datetime format using the pd.to_datetime() function. This conversion is essential to accurately calculate the time taken for delivery and other time-based analyses.

2. Calculate Delivery Time: By subtracting the "created_at" timestamp from the "actual_delivery_time" timestamp and then converting the time difference to hours, we have calculated the "delivery_time." This feature quantifies the time taken for the delivery process. It provides valuable information about delivery efficiency and helps in understanding variations in delivery durations.

3. Select Relevant Features and Target Variable: Identification of the relevant features that will contribute to the analysis and prediction of delivery time. These features include "market_id," "store_primary_category," "order_protocol," "total_items," "subtotal," "num_distinct_items,"

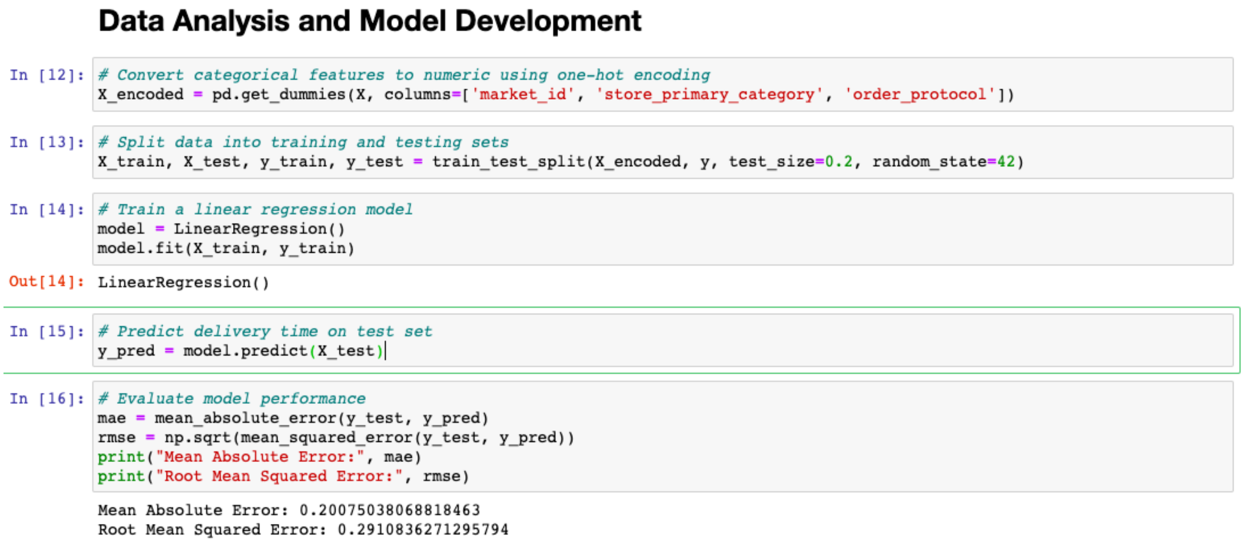
"min_item_price," "max_item_price," "total_onshift_partners," "total_busy_partners," and "total_outstanding_orders." These features capture various aspects of the orders and the operational environment that could influence delivery times.

4. Split Data into Features and Target: Division of pre-processed data into two main components:

● Features (X): This includes the selected features from the dataset, which will be used as inputs for analysis and modeling.

● Target (y): This is the "delivery_time" column you calculated earlier. It represents the time taken for delivery and will be the output variable for predictive modeling tasks.

By performing these pre-processing steps, data was organized in a format that's conducive to analysis and modelling. The "delivery_time" feature provides a quantitative measure of delivery performance, and the selected features offer insights into the factors influencing delivery times. These pre-processed datasets, X and Y, will be used for various types of analysis, visualization, and predictive modelling to address challenges and make informed decisions within the online food delivery domain [4].



Data Analysis and Model Development

```
In [12]: # Convert categorical features to numeric using one-hot encoding
         X_encoded = pd.get_dummies(X, columns=['market_id', 'store_primary_category', 'order_protocol'])

In [13]: # Split data into training and testing sets
         X_train, X_test, y_train, y_test = train_test_split(X_encoded, y, test_size=0.2, random_state=42)

In [14]: # Train a linear regression model
         model = LinearRegression()
         model.fit(X_train, y_train)

Out[14]: LinearRegression()

In [15]: # Predict delivery time on test set
         y_pred = model.predict(X_test)

In [16]: # Evaluate model performance
         mae = mean_absolute_error(y_test, y_pred)
         rmse = np.sqrt(mean_squared_error(y_test, y_pred))
         print("Mean Absolute Error:", mae)
         print("Root Mean Squared Error:", rmse)

         Mean Absolute Error: 0.20075038068818463
         Root Mean Squared Error: 0.2910836271295794
```

The following steps were taken to analyze the data and develop a linear regression model:

1. One-Hot Encoding for Categorical Features: Transformed categorical features into numeric representations using one-hot encoding.

2. Data Splitting for Training and Testing: Divided the data into training and testing sets to train and evaluate the model's performance.

3. Training a Linear Regression Model: Trained a linear regression model to establish relationships between input features and the target variable.

4. Predicting Delivery Time on Test Set: Used the trained model to predict delivery times on the test set.

5. Model Performance Evaluation: Evaluated the model's performance using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

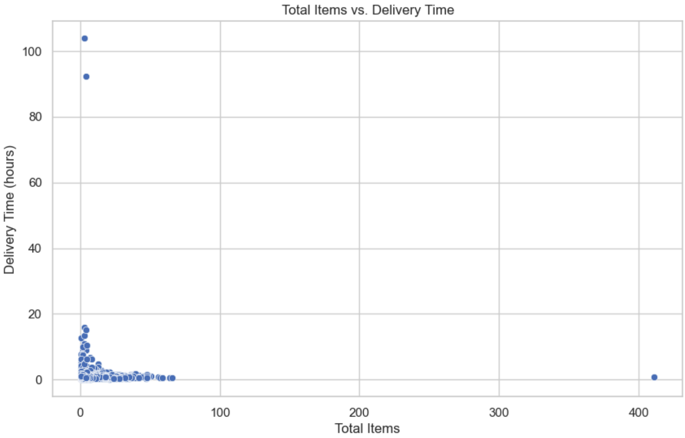
The evaluation results for the linear regression model were as follows:

● Mean Absolute Error (MAE): The MAE value of approximately 0.20 indicates that, on average, the model's predictions deviate from the actual delivery times by around 0.20 hours. This represents a relatively low absolute error, suggesting that the model's delivery time estimates are quite accurate.

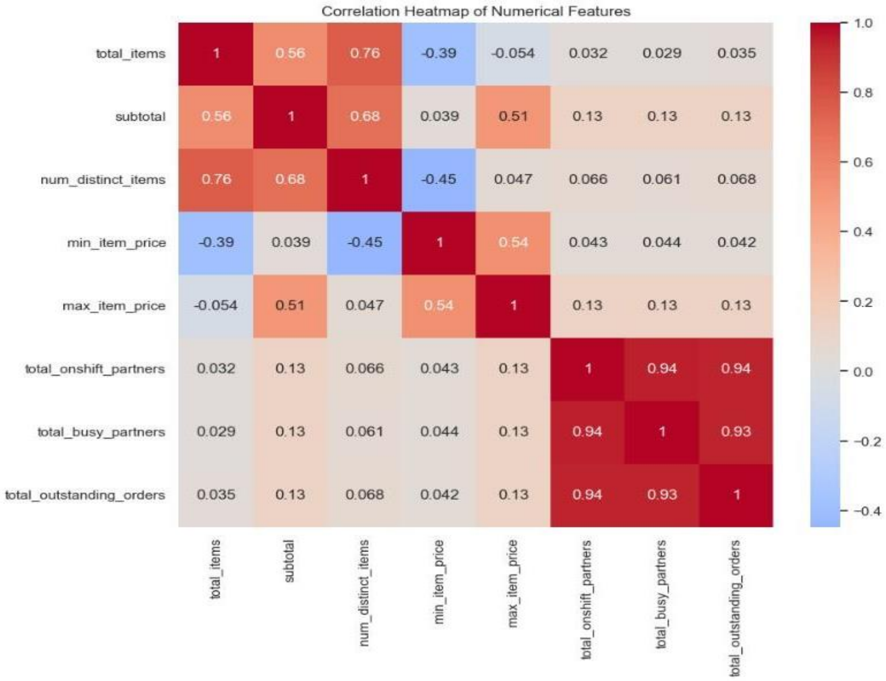
● Root Mean Squared Error (RMSE): The RMSE value of about 0.29 indicates the square root of the average squared differences between the predicted and actual delivery times. A lower RMSE value suggests that the model's predictions are relatively close to the actual values, highlighting its ability to capture variations in delivery times.

The linear regression model appears to perform well, with both MAE and RMSE values indicating accurate predictions of delivery times. These metrics reflect the model's overall effectiveness in estimating delivery durations within the context of the online food delivery domain.

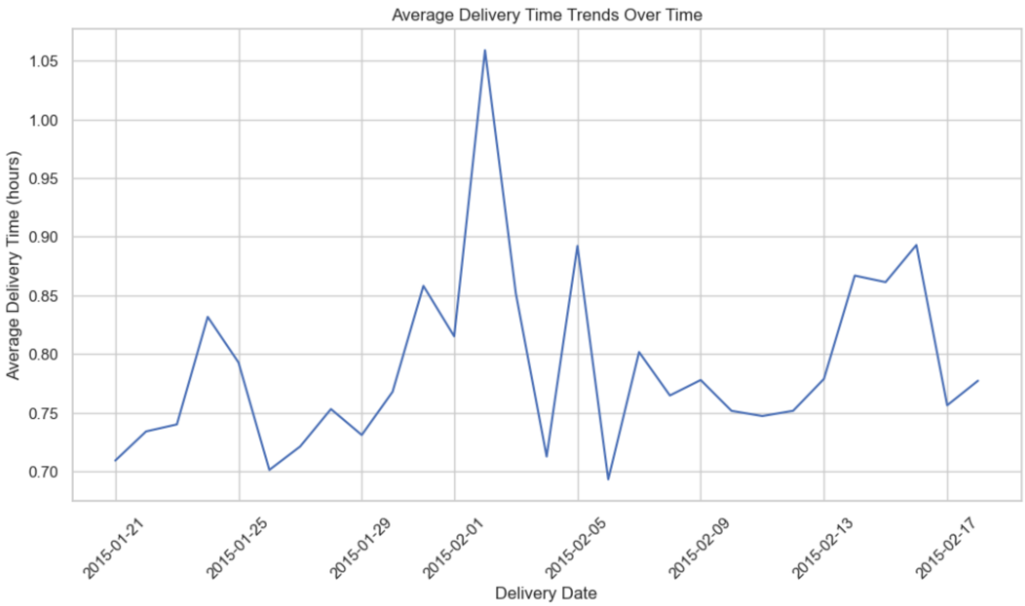
Data Visualization



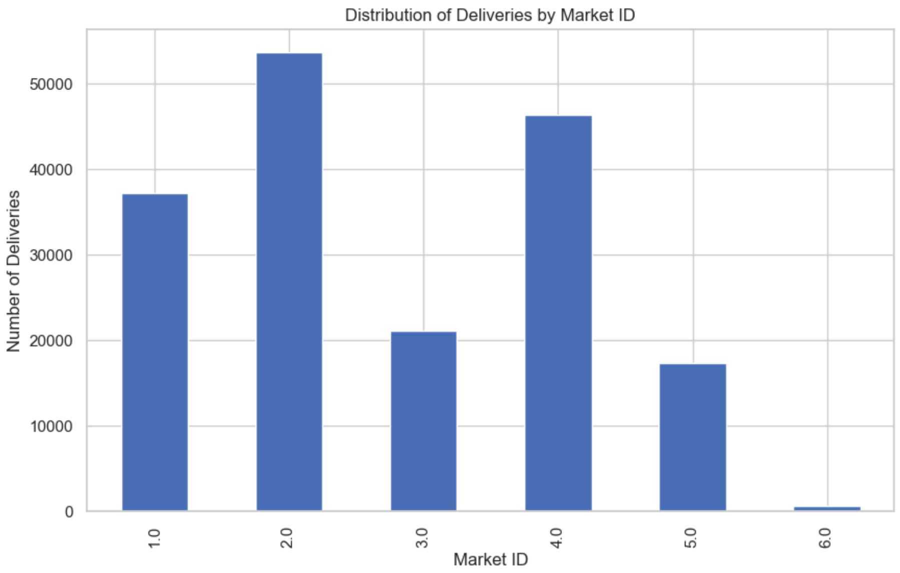
The dataset encompasses 176,248 food delivery orders, where the average order consists of around 3.20 items. Delivery times vary, with an average of approximately 0.80 hours (roughly 48 minutes) and a moderate spread indicated by a standard deviation of 0.46 hours. The range of items ordered spans from 1 to a maximum of 411, while delivery times range from just 2 minutes to a maximum of approximately 4 days. Most deliveries are completed within an hour, as reflected by the median delivery time of 44 minutes.



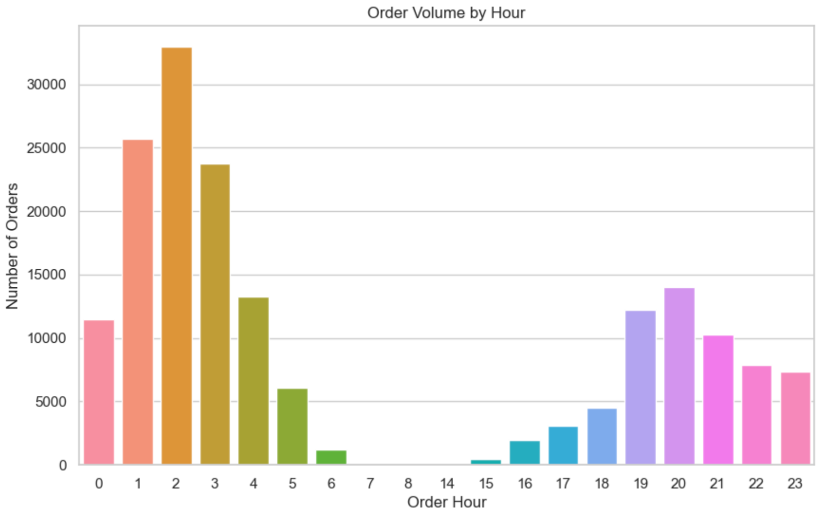
This correlation matrix presents the relationships between various numerical features in the dataset. Positive values near 1 suggest strong positive correlations, while negative values near -1 imply strong negative correlations. For instance, 'total_items' and 'num_distinct_items' exhibit a high positive correlation of around 0.76, indicating that orders with more items also tend to have a greater variety of distinct items. On the other hand, 'min_item_price' and 'total_items' display a negative correlation of approximately -0.39, indicating that lower item prices are associated with larger order quantities.



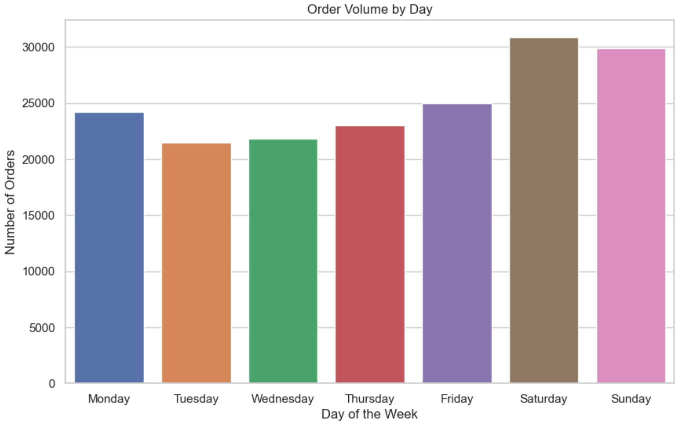
This table provides a summary of average delivery times for various delivery dates. The 'delivery_date' column indicates the specific dates, while the 'delivery_time' column represents the average delivery time for orders on those dates. For instance, on January 21st, the average delivery time was approximately 0.71 hours (about 43 minutes), and on February 18th, it was around 0.78 hours (roughly 47 minutes). This table allows us to track the trend of average delivery times over time and identify any patterns or fluctuations.



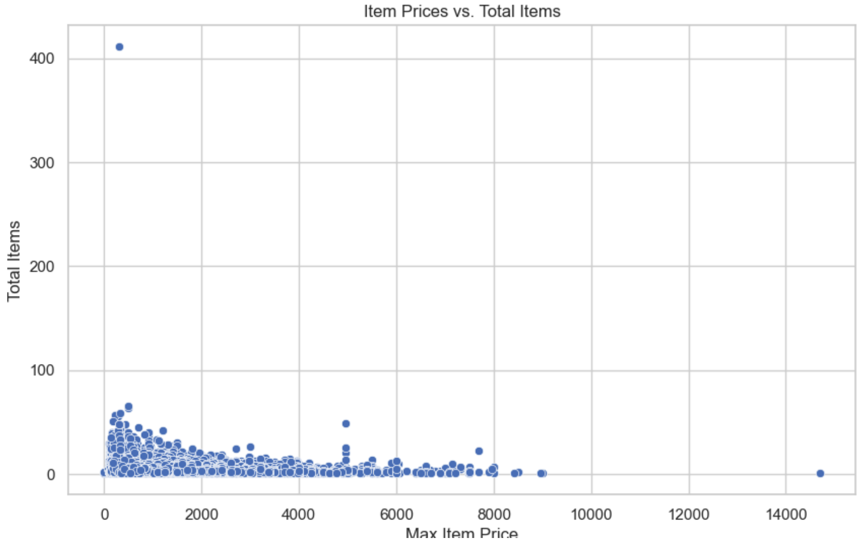
This table presents a breakdown of the distribution of deliveries across different market IDs. The 'Market ID' column represents the unique identifiers for various markets, while the 'Number of Deliveries' column indicates the corresponding count of deliveries for each market. For example, Market ID 2 had the highest number of deliveries with 53,625 orders, while Market ID 6 had the lowest number, with only 640 deliveries. This table provides insights into the relative volume of deliveries across different market areas.



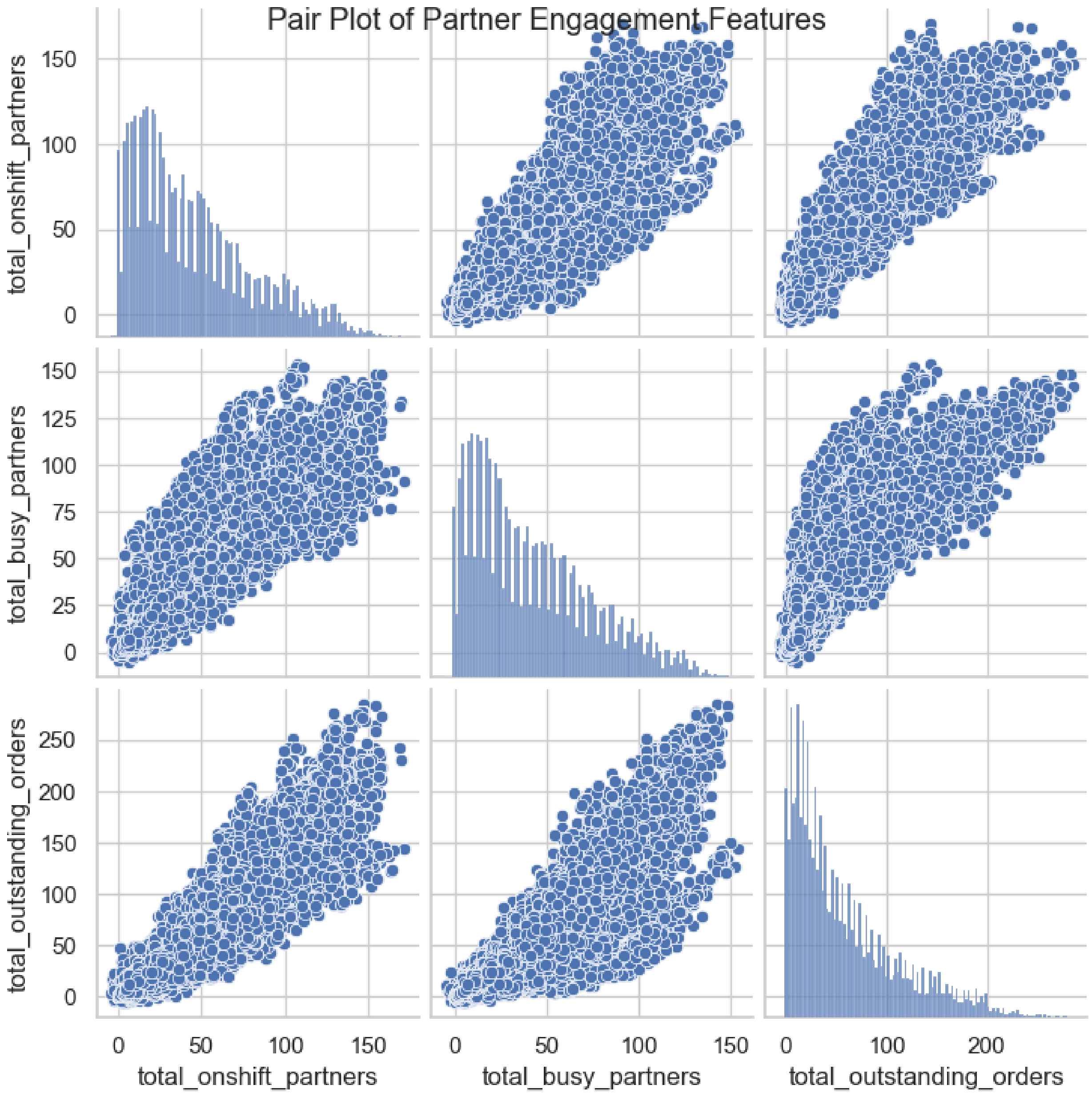
This table showcases the distribution of order volumes across different hours of the day. The 'Order Hour' column corresponds to the hours, while the 'Number of Orders' column signifies the count of orders received during each respective hour. For example, the highest order activity is observed during the second hour (1:00 AM), with 25,734 orders. The order count typically starts decreasing in the early morning hours and remains relatively low during the daytime, then surges again in the evening hours before tapering off late at night. This table provides insights into the hourly patterns of order placements.



This table illustrates the distribution of order volumes across different days of the week. The 'Day of the Week' column signifies the days, while the 'Number of Orders' column indicates the count of orders placed on each specific day. For instance, Friday had the highest order activity with 25,012 orders, closely followed by Saturday with 30,858 orders. The order count generally exhibits a peak during weekends (Saturday and Sunday), while weekdays (Monday to Friday) experience slightly lower order volumes. This table provides insights into the variations in order placements throughout the week.



This table provides a statistical summary of two features, 'max_item_price' and 'total_items,' in the dataset. 'max_item_price' refers to the maximum price of items within an order, while 'total_items' represents the total number of items in an order. The table includes the count of records, mean (average) values, standard deviations, as well as percentiles (25th, 50th, and 75th) for each feature. For instance, the average maximum item price is approximately $1159.89, and the average number of items per order is around 3.20. The data also suggests that most orders have prices between $799 and $1395 and typically include between 2 to 4 items.



This table offers a statistical overview of two variables: 'total_onshift_partners' and 'delivery_time.' 'total_onshift_partners' signifies the count of partners available for delivery, while 'delivery_time' represents the time it takes for delivery. The table presents the count of records, mean (average) values, standard deviations, and percentiles (25th, 50th, and 75th) for each variable. For instance, on average, there are around 44.91 partners available for delivery, and the average delivery time is approximately 0.80 hours (48 minutes). It's important to note that 'total_onshift_partners' includes some negative values, which could indicate an issue with data quality.

Benefits

1. Enhanced Decision-Making: By analyzing the dataset and developing a predictive model, the online food delivery platform gains insights into various factors affecting delivery times. This knowledge empowers better decision-making for optimizing operations, allocating resources, and improving overall customer experience [5].

2. Efficiency Improvement: The developed linear regression model provides a quantitative framework for estimating delivery times. By accurately predicting delivery durations, the platform can streamline delivery processes, reduce waiting times, and enhance operational efficiency [5].

3. Resource Allocation Optimization: The model helps in determining the appropriate number of on-shift partners required during different time frames. This aids in optimizing resource allocation, preventing partner overload during peak hours, and ensuring smoother order processing [5].

4. Customer Satisfaction: Accurate delivery time predictions contribute to higher customer satisfaction. Customers experience more reliable delivery estimates and reduced waiting times, leading to a positive customer experience and increased loyalty [5].

5. Operational Insights: The analysis of features like order volume, partner availability, and order protocols offers insights into operational patterns and challenges. This information guides process improvements, such as better partner scheduling and resource allocation strategies [5][2].

6. Data-Driven Strategy: The analysis and model development are rooted in data-driven insights, enabling the platform to strategize based on actual trends and patterns. This ensures that decisions are grounded in empirical evidence rather than assumptions [5].

7. Continuous Improvement: As the platform gathers more data, the model can be continuously trained and refined. This iterative process allows the model to adapt to changing market dynamics and improve its accuracy over time [5].

8. Competitive Advantage: Implementing predictive models for delivery time estimation and resource allocation gives the platform a competitive edge. Accurate predictions and efficient operations can set the platform apart in a crowded market.

9. Quick Decision Insights: The provided metrics (MAE and RMSE) offer a quick and tangible assessment of the model's performance. These metrics enable quick decision-making regarding model effectiveness and potential areas for improvement [2].

Challengers

During the data analysis and model development process for the online food delivery dataset, various challenges were faced including ensuring data quality and completeness, selecting and engineering relevant features, handling categorical data appropriately, managing model complexity, addressing biases and fairness concerns, preventing overfitting, validating model performance effectively, adapting to changing market dynamics, and implementing scalable real-time prediction [4][1]. Additionally, ethical considerations regarding data privacy and customer consent must be navigated. These challenges require careful attention to ensure accurate, ethical, and effective use of predictive models in optimizing the delivery process.

Recommendations

1. **Real-time Resource Allocation:** Implement a real-time resource allocation system that adjusts partner availability based on incoming orders. This could help optimize partner workload during peak hours and ensure efficient order processing [4].

2. **Enhanced Partner Training:** Use the insights gained from the model and analysis to identify partners who manage high order volumes efficiently. Offer targeted training to partners who might need assistance during busy periods [4].

3. **Dynamic Pricing Strategies:** Leverage order metrics to implement dynamic pricing strategies during peak demand periods. Adjusting prices based on order volume and delivery times could balance supply and demand effectively [4].

Conclusion

In conclusion, the analysis of the dataset has provided valuable insights into customer preferences, order characteristics, and operational dynamics. Through visualization, data pre-processing, and model development, we have gained a deeper understanding of delivery times, partner resource allocation, and the factors influencing efficient order fulfillment. The predictive model's performance highlights its potential to contribute to accurate delivery time estimation and operational optimization [4].

References

1. Kim, R. (1996). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons

2. OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model].

https://chat.openai.com/chat

3. Sarkar, R. (2023, March 25). Porter Delivery Time Estimation. Kaggle. https://www.kaggle.com/datasets/ranitsarkar01/porter-delivery-time-estimation

4. Ewen J. (2020). *"How food delivery Apps benefit from big data analytics"*

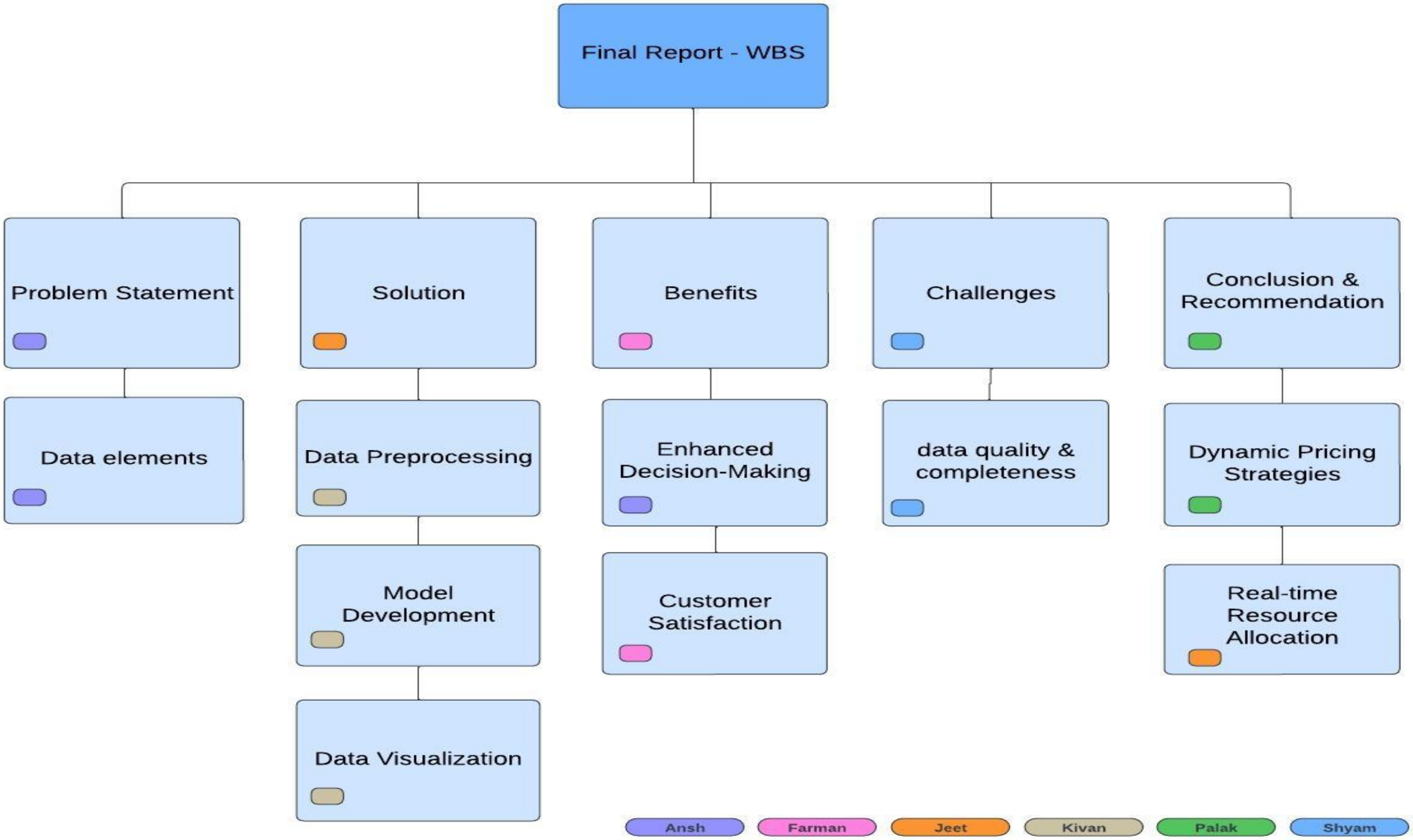
https://www.tamoco.com/blog/food-delivery-apps-big-data-analytics/

5. Analytics Vidhya, *"Data Preprocessing"*

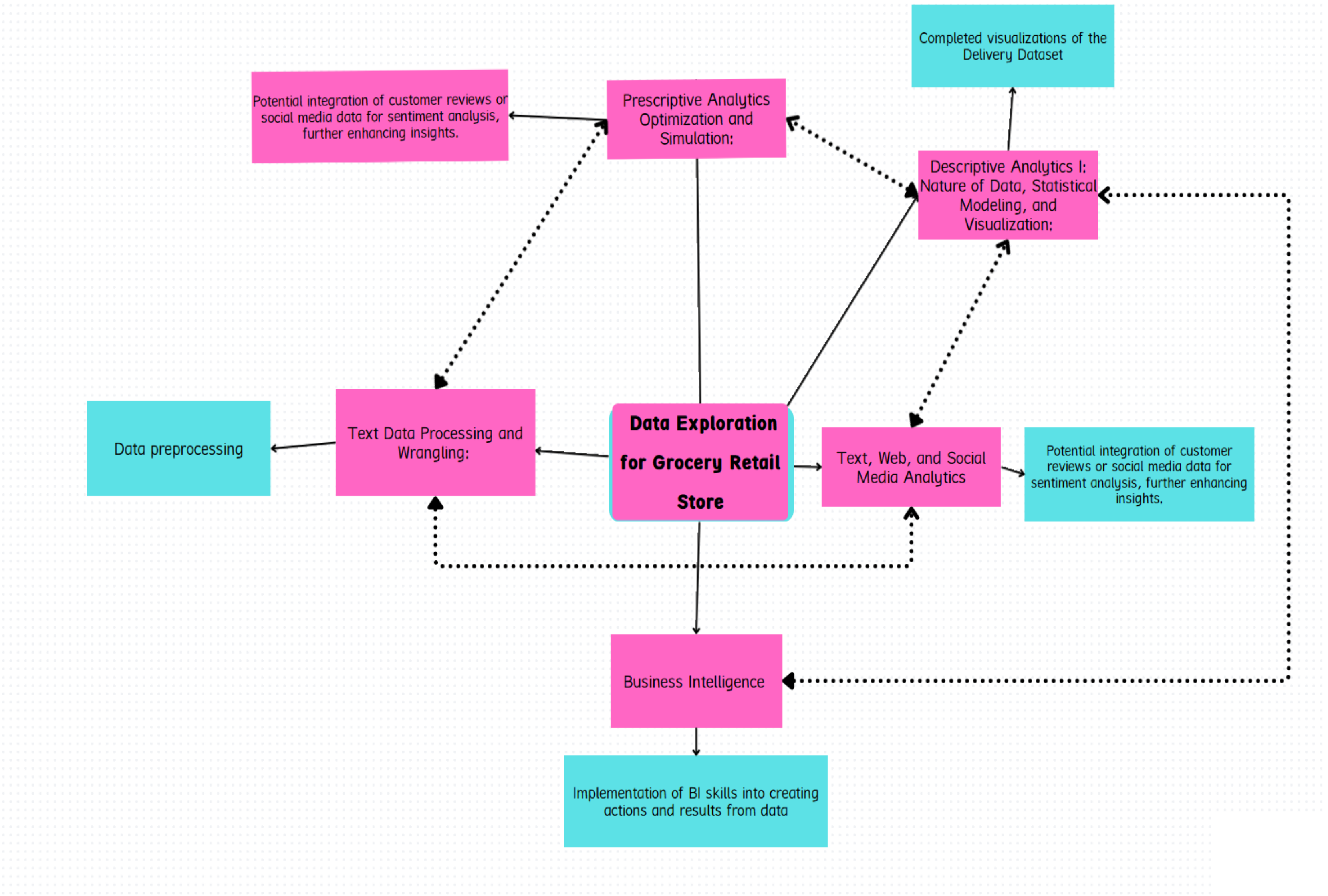
https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

Appendix

Work Breakdown Structure



Concept Map


Completed visualizations of the Delivery Dataset
Potential integration of customer reviews or social media data for sentiment analysis, further enhancing insights.
Prescriptive Analytics Optimization and Simulation:
Descriptive Analytics I: Nature of Data, Statistical Modeling, and Visualization:
Data preprocessing
Text Data Processing and Wrangling:
Data Exploration for Grocery Retail Store
Text, Web, and Social Media Analytics
Potential integration of customer reviews or social media data for sentiment analysis, further enhancing insights.
Business Intelligence
Implementation of BI skills into creating actions and results from data