MERCER UNIVERSITY

A Capstone Project

Presented to

The Academic Faculty

Enhancing DoorDash Customer Satisfaction and Revenue: Optimizing Delivery Efficiency

By

Nguyen Trang

Wang Qi

Advisor: Dr. Rui Sun

In Partial Fulfillment of the Requirements of Master of Science in the Business Analytics Program Offered by Stetson-Hatcher School of Business at Mercer University

May 2024

TABLE OF CONTENTS

1. Introduction

How the way people eat is changing significantly in the modern world. In the past, restaurant-quality meal delivery was limited. Nowadays, with significant advancements in digital technology and shifts in consumer expectations, the food delivery ecosystem has experienced exponential growth during the COVID-19 global pandemic and has expanded steadily, greatly impacting the dine-in restaurant business. DoorDash is one of the most popular online food ordering platforms among the four top-rated food delivery services: Grubhub, Uber Eats, Postmates, and ChowNow. The platform operates in many cities across the United States and in other countries as well, serving as a connection between consumers, technology-driven delivery networks, and local restaurants. It functions as a third-party aggregator, offering customers a convenient way to quickly access menus, prices, order customization, food orders, and delivery tracking [1]. The company generates a substantial portion of its revenue through commission charges from partner restaurants, customer delivery and service fees, subscription plans, advertising, catering services, and customer tips to Dashers [2]. Do you know that the dominance of food delivery services has led to a 27% year-over-year increase in DoorDash's revenue, leading to a market capitalization of $39.37 billion in 2023 [3]? This achievement positions DoorDash as one of the most valuable companies and has been recognized in Fortune's Future 50 Companies in 2023 [4].

To gain deeper insights into DoorDash's operations and performance, a study was conducted by examining a large historical dataset containing all delivery order records received by DoorDash in early 2015. This dataset provides comprehensive information, including timestamps of order submission and completion, store details, order protocol systems, customer order characteristics, and market features. Delivery time is a crucial factor for customer satisfaction in the food delivery business. The primary objective of this study is to assist the business by developing a model to predict the accuracy of total delivery duration. This analysis aims to understand how DoorDash operates efficiently to enhance customer experience, gain insights, and refine strategies to keep DoorDash competitive and successful in the food delivery market.

Main objective:

- Identify Influencing Factors

  Analyze and determine the main factors affecting delivery efficiency and enhance the overall delivery process.

- ❖ Formulate Strategic Recommendations

  Present strategies and approaches to optimize the total delivery time and increase profits.

To achieve these objectives, data mining techniques and machine learning algorithms will be employed to conduct further analysis in this study.

2. Literature Review

Several articles were reviewed to conduct this study, ensuring that the methodologies and models utilized in this analysis were rigorous and aligned with contemporary research. For example, Ahmed Saad (2021) published a paper analyzing online food delivery services in Bangladesh. The purpose of the study is to determine significant factors that impact customers' behavior when ordering food through online food delivery intermediaries. The author utilized statistical analysis, including t-test and factor analysis, to conduct the study. From the findings, the author highlights the most important factors affecting successful online food delivery services, including delivery time, service quality, condition of the food delivered, and price. In addition, the author emphasizes that significant aspects influencing delivery time service are order characteristics and price. Online food ordering has emerged as a rapidly growing sector globally due to the development and widespread availability of the Internet, and the increasing consumer demands in today's fast-paced world. The author recommends that understanding customer needs is an essential factor for business growth and highlights the positive impact of online food delivery services on businesses and the enhancement of people's quality of life [5].

Machine Learning Algorithms

Another study conducted by Farhana (2020), examined factors affecting the accuracy of predicting shipment delivery lead times, which represent the duration between the initiation of an order and the completion of a processed order. In this study, the researcher utilized machine learning algorithms, including Linear Regression, Ridge Regression, Lasso Regression, Random Forest (RF), Support Vector Machine (SVM), Multivariate Adaptive Regression Splines (MARS), K-Nearest Neighbors (KNN), and Artificial Neural Network (ANN) to identify the significant factors influencing the accurate prediction of the delivery times of purchase orders. To evaluate the accuracy performance of the predictive models, five major performance metrics were utilized in this analysis, including Mean Absolute Scaled Error (MASE), Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error

(RMSE), and Normalized Root Mean Square Error (NRMSE). According to the findings, the author recommended that Random Forest (RF) demonstrated the best performance model, as indicated by the Normalized Root Mean Square Error (NRMSE) and the Root Mean Square Error (RMSE). In addition, the author discovered that the accuracy of predicting shipment delivery lead times is influenced by factors such as order part numbers, order types, supplier origin for the purchase part, buyer destination, and requested delivery date[6].

Linear Regression Algorithms

Wahyudi & Arroufu (2022) focused on using linear regression algorithms to predict delivery times, analyzing a dataset comprising 1000 orders and considering various factors, including pickup times, claim delivery time and delivered time and so on. Their study demonstrated the efficacy of linear regression models in analyzing the impact of these factors on delivery times. Through precise data analysis, their model achieved an RMSE value of 0.370%, proving the practicality of linear regression methods in specific contexts[7].

Regression Tree-Based Models and Quantile Regression Forests

Salari, Liu, and Shen (2022) adopted a more complex approach by leveraging machine learning technologies, particularly tree-based models and quantile regression forests, to predict real-time delivery times in online retail. They also introduced a cost-effectiveness decision rule for setting customer-promised delivery times. Tested on the real dataset from JD.com, their model showed superior forecasting performance and demonstrated the potential of precise delivery time predictions to enhance sales volumes and customer satisfaction[8].

Boosting Algorithms

Khiari & Olaverri-Monreal (2022) explored boosting algorithms for delivery time prediction in transportation logistics. Their study highlighted the effectiveness of light gradient boosting and CatBoost over traditional models like linear regression and random forest. Utilizing metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for model evaluation. Their results demonstrated superior accuracy and efficiency of the boosting algorithms, suggesting significant improvements in operational efficiencies and customer satisfaction for postal services. Through these evaluation metrics, the study underscores the potential of advanced machine learning techniques in enhancing the precision of delivery time predictions[9].

In our study, the objective is to improve the accuracy of predicting DoorDash delivery times to enhance customer satisfaction. With guidance from insights in the literature, various types of machine learning algorithms are employed to conduct our study, including the Multilinear Regression model, Classification and Regression Tree models, Random Forests, and advanced machine learning methodologies such as Boosting algorithms. These algorithms are utilized to uncover complex data patterns and relationships that could further refine the accuracy of the delivery time predictions.

3. Data Collection, Cleaning, and Preprocessing

3.1 Data Collection and Description

The study utilized data from StrataScratch, a platform known for gathering a wide range of datasets from different sectors. All the data in this study originated from DoorDash and was acquired through StrataScratch. The dataset describes 197,428 records of customer orders through the DoorDash system. It includes 16 variables outlining DoorDash operation in 6 different locations (Figure 1), timestamps for customer order placement and delivery, 73 restaurant cuisine categories (Figure 2), and customer order descriptions.

3.2 Data Cleaning and Preprocessing

The purpose of this study is to examine factors affecting delivery times and establish a predictive model to forecast delivery times more accurately. To achieve this goal, data cleaning and preprocessing are essential steps before conducting the study and modeling. This process includes identifying and correcting any inaccuracies, inconsistencies, missing values, and outliers in the dataset to ensure the reliability and validity of the analyses and predictions.

3.2.1 Dimension and Characteristics of the Dataset

After conducting summary statistics to provide an initial overview of the dataset, we observed several issues, including inconsistencies in variable types, the presence of missing values, and potential outliers inferred from the maximum and minimum values.

3.2.2 Data Type Check and Conversion

All variables presented in the data were transformed into the correct types to ensure effective analysis. Four variables, including market_id, store_id, order_protocol, and store_primary_category, are correctly classified as categorical variables.

3.2.3 Handling Missing Values

The dataset contains missing values in several variables. There are 987 missing records identified in market_id, and 7 missing records in actual_delivery_time. In addition, 995 missing records are found in order_protocol, and 4760 missing records in store_primary_category. Furthermore, 16262 records are reported as missing in each of the variables of total_onshift_dashers, total_busy_dashers, and total_outstanding_orders. There are 526 records with missing values revealed in the estimated_store_to_consumer_driving_duration. The overall missing values in the dataset total 56,061, spread across 21,651 records, which constitutes approximately 10.97% of the total records. Given that this proportion is relatively small, it has been decided to remove these records with missing values from the dataset.This proportion is relatively small; therefore, these records with missing values are removed from the dataset.

3.2.4 Duplicate Detection

After a thorough review of the dataset, there are no duplicate records present.

3.2.5 Correcting Data Inconsistencies

The dataset had no spelling errors or other common issues, such as inconsistent capitalization or mislabeling. However, semantic inconsistencies were observed in the data, particularly between the store_id and store_primary_category variables, which can be seen in Table 2. Some records had the same store_id but were associated with different primary categories, as seen in Table 2. To address these semantic inconsistencies, a corrective measure was implemented. The store_id and store_primary_category were collected separately within each market_id location. Then, each store_id was mapped to the most frequently occurring cuisine category, which was considered representative of the store's primary category.

Structural inconsistencies were also identified in the variables created_at and actual_delivery_time, where both date and time were included in the same column. To correct these inconsistencies, separate columns were created to indicate the date and time of customer order placement submitted to DoorDash, as well as the date and time of order delivery to consumers.

3.2.6 Feature Engineering: Creating New Variables

The aim of this study is to investigate the factors influencing the total delivery duration and

develop a predictive model to enhance the accuracy of forecasting delivery times to customers. To enhance the predictive modeling performance in the dataset, the target variable total_delivery_duration was created by subtracting the created_at timestamp from the actual_delivery_time. In addition, the dataset is enriched with new variables to capture crucial information:

busy_dashers_ratio: This variable indicates the ratio of busy dashers to the total dashers on shift, where a higher ratio implies a lower availability of dashers, potentially increasing delivery times. This is directly related to the efficiency of deliveries.

distinct_item_of_total: Reflects the diversity of items in an order, highlighting the variety of items purchased.

price_range: The difference between the highest and lowest item prices in an order, indicating the range of item prices.

estimate_non_preparation_duration: A combination of estimated_order_ place_ duration, the estimated time for a restaurant to receive an order from DoorDash, and estimated_store_to_consumer_driving_duration, the estimated travel time from store to consumer. This sum excludes the time taken for food preparation by the restaurant.

weekend and time_peak (Figure 3 & Figure4): Variables based on the analysis of delivery times across various days and times. Delivery durations are found to be longer on Fridays, Saturdays, and Sundays, attributed to increased weekend activities and the transition back to work routines on Monday. Furthermore, peak periods identified as 6 AM to 9 AM, 11 AM to 1 PM, and 5 PM to 8 PM witness extended delivery times, likely due to higher traffic, increased order volumes, and potential decreases in delivery efficiency.

3.2.7 Identify Potential Errors and Inconsistencies in Numerical Variables

Through the Descriptive Statistics summary (Table 3), particularly by examining the maximum and minimum values, we can identify unusual values within the dataset. Some variables, such as total_onshift_dashers, total_busy_dashers, total_ outstanding_orders, and min_item_price, exhibit negative values, which are nonsensical in this context. Consequently, we opted to remove these observations to maintain the integrity of our analysis.

3.2.8 Data Distribution, Outlier Detection and Handling

a) Data Distribution

According to Figure 5, some variables in the data exhibit non-normal distributions. Specifically, the histogram for Total_Outstanding_Orders illustrates a right-skewed

distribution. This means that most of the time, there are relatively few pending orders within a 10-mile radius of the order being processed. However, during peak periods, this number can increase significantly, reaching a maximum value of 285. On average, there were approximately 61 orders within a 10-mile radius of the order being processed.

Similarly, the data distribution of the variable total_items is right-skewed, indicating that some orders contain more items than the median. Most orders consist of a relatively small number of items, with the median order containing 3 items. However, there are some orders with a significantly higher number of items, with a maximum of 411 items in the order.

The distributions of the variables estimate_non_preparation_duration and avg_price_per_item are leptokurtic, indicating heavier tails and a sharper peak, as evidenced by their respective kurtosis values of approximately 2.818701 and 8.383186.

Regarding the distribution of estimate_non_preparation_duration, the shortest estimated non-preparation duration is around 179 seconds (3 minutes), while the longest duration reached 3222 seconds (53 minutes and 42 seconds). The typical estimated non-preparation duration across the dataset was approximately 856.1 seconds (14 minutes).

This examination of skewness offers valuable insights into the data distribution structure. The next steps of this study involve identifying any potential outliers and determining the appropriate strategies for managing them.

b) Outlier Detection and Handling

By employing boxplot visualizations for a more detailed examination of outliers across each continuous variable, we observed that every continuous variable contains outliers, as seen in Figure 6. In our analysis, we chose to remove outliers that lie beyond 3 times the interquartile range (IQR) to enhance the accuracy and clarity of our findings. This method serves to minimize statistical distortions and the potential for misleading conclusions based on extreme data points.

3.2.9 Unit Standardization

To ensure uniformity in units and to facilitate easier interpretation of time and monetary variables, we converted the units of variables related to time and money. Specifically, we

change time from seconds to minutes and monetary values from cents to dollars.

3.2.10 Data Dimension Reduction

Structural inconsistencies were identified in the variables `created_at` and `actual_delivery_time`, where both date and time were included in the same column. The inconsistencies were then corrected. Separate columns were created to indicate the date and time of customer order placement submitted to DoorDash and the date and time of order delivery to consumers.

3.3 Data Partitioning

Data partitioning is an essential step in assessing the stability, performance, and generalization capabilities of data analysis and machine learning (ML) models. Before conducting the study and modeling, the cleaned and preprocessed dataset was partitioned into a 70% training dataset and a 30% validation dataset. There are 81,776 customer records in the training set and 35,048 customer records in the validation set. These datasets include 110 variables outlining DoorDash's various operational locations, timestamps for customer order placement and delivery, primary restaurant cuisine categories, and customer order descriptions. The primary objective of utilizing the training dataset was to train models to discern patterns and relationships within the data. After the models were trained, they underwent testing on the validation dataset. The model's performance on the validation dataset provided an estimation of its efficacy when performing with new and unseen data.

4. Methodology

4. 1 Feature Selection Methodologies

4.1.1 Wrapper method - Backward Elimination

Backward Elimination is one of the wrapper methods for feature selection. A wrapper method is used in feature selection to improve the performance of machine learning models by identifying the best subset of features. This method involves selecting a set of features, training a model using these features, and evaluating the quality of the subset based on the performance of the model. The process is often iterative, seeking to find the optimal combination of features. Wrapper methods include various algorithms like Recursive Feature Elimination (RFE), Forward Selection, and Backward Elimination. We use the Backward elimination method, which starts with all features and then removes the least significant feature at each iteration until reaching an optimal set of features.

In this study, the backward elimination method integrated diagnostic tests, including tolerance and VIF, to evaluate multicollinearity among the predictors in the model and to select the most significant predictors for the models.

To illustrate, through the examination of Pearson correlation coefficients, we discovered moderately positive correlations among various pairs of variables in the dataset (correlation coefficient > 0.8), as seen in Figure 7 and Table 5. The moderate correlation coefficients detected between these variable pairs may suggest the possible presence of multicollinearity, leading to instability in parameter estimates, and inflated standard errors. According to Table 6, the tolerance values for the variables total_busy_dashers, total_onshift_dashers, and total_outstanding_orders are below 0.1, with their corresponding variance inflation factors (VIF) exceeding 10. These findings suggest the presence of moderate multicollinearity among these variables. To mitigate these concerns, redundant variables total_busy_dashers and total_onshift_dashers were removed, and busy_dashers_ratio was introduced. This new variable represents the ratio of total_busy_dashers to total_onshift_dashers, effectively reducing multicollinearity while retaining the information content of the variables.

4.1.2 Embedded Method - LASSO Regression

Embedded methods integrate feature selection directly into the model training process and optimize regularization to enhance prediction accuracy, with techniques like LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge regression. Group LASSO is a variant of LASSO that penalizes groups of coefficients, promoting sparsity not just at the individual coefficient level but also at the group level, effectively performing variable selection within groups of related predictors. This method helps to prevent overfitting by shrinking the coefficients of less important predictors towards zero, thereby reducing model complexity. By selecting the most relevant predictors and shrinking the coefficients of others, Lasso regression is a powerful method for enhancing predictive accuracy and improving the interpretability of statistical models in machine learning and data analysis.

Figure 8, Figure 9 and Figure 10 display significant predictors of the total delivery duration of DoorDash among 109 predictors.

4.2 Model Selection and Evaluation Methods

In our analysis aimed at predicting DoorDash's total delivery duration, we will adopt a range of predictive models, each chosen for its specific strengths and adaptability to our forecasting challenge. The models include Multiple Linear Regression (MLR), Decision Tree, Random Forest, Support Vector Regression (SVR), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM).

Multiple Linear Regression is our foundational statistical tool. Its main benefit is revealing the direct linear relationships between the independent variables and the dependent variable, aiding in pinpointing which factors most influence the total delivery time. Additionally, it provides a clear interpretation of the results, which is crucial for comprehending the mechanics behind delivery efficiency. However, its predictive performance may sometimes fall short when compared to other machine learning algorithms.

Decision Trees offer a clear way to navigate through complex data. They visually break down the decision process, making it easier to understand how various factors affect total delivery times. But they have a downside: they can get too fitted to our specific data set and might not perform well if there's a slight change in data.

To mitigate these potential limitations, we incorporate Ensemble Methods such as Random Forest, Gradient Boosting, XGBoost, and LightGBM. Random Forest combats the overfitting issue inherent in single decision trees by constructing a multitude of trees and integrating their outcomes, either through averaging or majority voting. This method enhances predictive accuracy and model robustness without significantly compromising interpretability.

Support Vector Regression (SVR) stands out in its capacity to manage high-dimensional spaces and nonlinear data relationships via kernel functions. It's exceptionally skilled at capturing complex patterns within delivery data, serving as a powerful enhancer of our prediction accuracy.

Gradient Boosting, XGBoost, and LightGBM are advanced ensemble techniques known for their high efficiency and predictive performance. These methods sequentially build trees, each correcting errors made by the previous ones, leading to progressively improved accuracy. XGBoost and LightGBM, in particular, offer enhancements in speed and efficiency, enabling the handling of large data sets without compromising on model performance.

By combining these models, we can explore and leverage both linear and non-linear relationships in the data, balance the interpretability and predictive performance of the model, as well as computational efficiency and the ability to handle large datasets. This multi-model strategy allows us to understand the data from different angles and depths, enhancing the accuracy and reliability of predictions and thereby providing accurate forecasts for DoorDash's total delivery duration.

Furthermore, for evaluating the predictive capabilities of our models, we have chosen metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error). Both metrics serve a specific purpose in quantifying the accuracy of our predictions. MAE provides an average of the absolute differences, offering a straightforward interpretation of prediction accuracy. RMSE takes the square root of MSE, scaling errors back to their original units, which makes it particularly useful for understanding the magnitude of prediction errors. In addition to these metrics, we also employ lift charts and decile-wise lift charts for visualizing the model's capability in identifying longer delivery duration cases as compared to random selection.

5. Empirical Results

5.1 Model Evaluation

5.1.1 Performance Metrics Evaluation

To evaluate the accuracy prediction of the predictive models, two major performance metrics were used in this analysis, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE). According to our findings, as presented in Table 7, LightGBM emerges as a standout performer, demonstrating the lowest RMSE and MAE values among all tested models. This performance indicates its superior capability in providing accurate and reliable predictions. Following closely behind is XGBoost, which also shows commendable performance metrics but slightly higher in both RMSE and MAE compared to LightGBM.

We also consider the efficiency of these models in terms of training time and prediction speed. Efficiency is a critical factor, especially in environments where time and computational resources are limited. In this context, LightGBM not only leads in accuracy but also excels in operational efficiency, making it the superior choice for scenarios requiring both high performance and fast model training and prediction capabilities.

Overall, the combined analysis of accuracy and efficiency highlights LightGBM's potential as a highly effective tool in predictive modeling, capable of delivering quick and precise predictions while ensuring an optimal fit to complex data sets. This makes it an ideal choice for real-world applications where both performance and speed are crucial.

5.1.2 Model Evaluation Visualization

Lift charts and decile-wise lift charts are effective tools for comparing the performance of different predictive models. The lift chart is a visual tool that compares the cumulative response of the models against a random selection, ordered by the model's estimated probabilities. The distance between the model's curve and the diagonal line (which represents a random guess) indicates the lift the model provides. A curve that remains above this line suggests that the model offers a better predictive capability than random chance. The decile-wise lift chart divides the data into deciles based on the model's predicted delivery duration values. For each decile, it calculates the average actual delivery duration and compares it to a baseline prediction. The y-axis shows the lift, which is the ratio of the average predicted response to the baseline response. A value greater than 1 indicates that the model performs better than random chance in that decile.

The lift chart (Figure 11) demonstrates that all models perform better than random chance, as evidenced by the curves that remain above the diagonal benchmark line throughout the range of cases. Although the curves of the different models are close to each other, LightGBM's curve is the one that most closely approaches the upper left corner, signifying its superior performance in predicting longer delivery durations with higher accuracy over the others. Further insight is provided by the decile-wise lift chart (Figure 12), which is particularly useful for assessing each model's ability to identify orders that will have longer total delivery durations.In this chart, all models demonstrate lift ratios above 1 in the initial deciles, indicating that these models are effective at predicting longer delivery times compared to a random selection. This suggests that the models are accurately identifying orders with longer actual delivery durations than typically expected. LightGBM stands out as it shows the highest lift values in these early segments, signifying its superior precision in detecting orders that are prone to the longest delivery times when compared to the other models.

5.2 Key Features Analysis

Among the 7 models used in this study, the LightGBM model demonstrates superior

performance in predicting longer delivery durations with the highest accuracy. From the LightGBM model, we found that the five most significant factors impacting the total delivery duration are non-preparation duration, busy dashers ratio, total outstanding orders, subtotal, and time peak, as seen in Figure 13.

To assess the impact of factors on total delivery duration, we use SHAP value analysis, a method in machine learning used to analyze how each feature contributes to the prediction. This analysis shows that the average predicted total delivery duration across 81,776 order records is 47.913 minutes. As seen in Figure 14, non-preparation duration is the most significant feature affecting the total delivery duration. The SHAP value of non-preparation duration is +4.26, meaning that for every minute increase in non-preparation duration, the total delivery duration until customers receive their order will be 4.26 minutes longer. Similarly, the following factors affecting the total delivery duration are the busy dasher ratio and total outstanding orders. For every one unit increase in the busy dasher ratio or an additional order in total outstanding orders, the total delivery duration will increase by 3.15 minutes and 3.03 minutes, respectively. Additionally, if customers place orders during peak hours, the total delivery duration will be longer than for customers placing orders during regular hours by 1.6 minutes.

5.3 "Real World" Prediction

In our analysis, we employ the LightGBM model, which has shown exceptional predictive performance in previous evaluations, making it well-suited for estimating the total delivery time of hypothetical orders. We examine three distinct scenarios in Table 8, each reflecting how different factors such as market ID, order size, order value, dasher availability impact the delivery duration.

Case A: Involves a small, simple order from a sandwich shop during a non-peak period. The predicted delivery duration is approximately 25.80 minutes. This quick delivery can be attributed to the small size of the order and the off-peak timing, which typically means less traffic and faster preparation times.

Case B: Describes a medium-sized order from a moderately busy restaurant during lunchtime peak. The estimated delivery time is about 45.35 minutes. This scenario illustrates the impact of a moderately busy period, where, despite the order not being exceptionally large, the peak traffic and potential delays contribute to a longer delivery duration.

Case C: Concerns a large, complex order from a popular restaurant during weekend peak

hours. The model forecasts a delivery duration of 70.02 minutes, primarily due to the complexity of the order and the busy traffic typical of peak periods. In such cases, both preparation and transit delays significantly extend the delivery time.

These detailed examples clearly demonstrate how the LightGBM model adjusts its predictions based on various input variables and accurately reflects the actual impact of these factors on delivery time.

6. Conclusion and Recommendation

It is concluded that LightGBM outperforms Multiple Linear Regression (MLR), Random Forest (RF), Gradient Boosting Machines (GBM), XGBoost, and Support Vector Regression (SVR) with superior performance, providing more accurate predictions across all performance metrics, including RMSE, MAE. The model can be implemented for forecasting total delivery duration and predicting customer behaviors, providing valuable insights into DoorDash delivery operations, and it can also be used in marketing analytics.

Customer service is important across all industries, as it helps businesses retain customers, increase revenue, and maintain a competitive advantage in the market. According to the findings, we have some recommendations to improve the DoorDash customer experience.

a) Optimize Dasher Allocation

To enhance service efficiency, DoorDash should increase the number of Dashers during peak times and in high-demand locations. We also recommend continuous monitoring of outstanding orders to enable real-time adjustment of Dasher distribution and better respond to fluctuations in order volumes.

b) Dasher Rewards Program

DoorDash should implement a rewards program for Dashers working during high-demand periods to enhance delivery efficiency and increase Dasher satisfaction and loyalty.

c) Dedicated Delivery Teams

DoorDash should establish dedicated delivery teams for such orders to ensure efficient and secure delivery, address the unique needs of such orders, and enhance customer satisfaction.

d) Keep Improving Order Dispatch System

DoorDash should establish a more advanced order dispatch system that schedules delivery drivers' pickup times based on the estimated order preparation time at restaurants. For instance, the system can automatically adjust pickup times for known high-value orders or large orders requiring longer preparation times to minimize driver wait times and ensure orders are ready

upon arrival, optimizing efficiency.

e) Multi-Modal Delivery Options

DooDash should introduce multi-modal delivery options, such as combining drones, electric bicycles, motorcycles, and walking, among other delivery methods. This allows for flexible selection of the most suitable delivery method based on different areas of the city and actual traffic conditions, particularly during congested periods, effectively reducing delivery times.

f) Data-Driven Decision Making

DoorDash should use data-driven methods to analyze market and store categories involving collecting detailed information on market size, consumer preferences, and competitor strategies. This data helps in making informed decisions about store location, pricing, and product assortment, optimizing operations to better meet market demands and enhance business performance.

References

[1] Littman, Julie. (2022). DoorDash's expansion beyond the US will come with challenges. *Restaurant Dive: Restaurant News and Trends.*

[2] Reaume, Amanda. (2021). 8 Ways DoorDash Makes Money. *Seeking Alpha | Stock Market Analysis & Tools for Investors.*

[3] GuruFocus Research. (2024). DoorDash Inc (DASH) Reports Strong Q4 and Full Year 2023 Results. *Yahoo Finance.*

[4] Reeves, Martin. (2023). The Future 50: Companies built for growth in uncertain times. *Fortune - Fortune 500 Daily & Breaking Business News.*

[5]. Saad, A.T. (2021). Factors affecting online food delivery service in Bangladesh: an empirical study. British Food Journal, Vol. 123 No. 2, pp. 535-550.

[6]. Farhana, N. (2020).Using Machine Learning Methods to Predict Order Lead Times. International Journal of Sciences: Basic and Applied Research (IJSBAR), Vol. 54 No. 3, pp. 87-96.

[7]. Wahyudi, T., & Arroufu, D. S. (2022). Implementation of Data Mining Prediction Delivery Time Using Linear Regression Algorithm. Journal of Applied Engineering and Technological Science (JAETS), 4(1), 84-92.

[8]. Salari, N., Liu, S., & Shen, Z. J. M. (2022). Real-time delivery time forecasting and promising in online retailing: When will your package arrive?. Manufacturing & Service Operations Management, 24(3), 1421-1436.

[9]. Khiari, J., & Olaverri-Monreal, C. (2020, November). Boosting algorithms for delivery time prediction in transportation logistics. In 2020 International Conference on Data Mining Workshops (ICDMW) (pp. 251-258). IEEE.

Appendix

Table 1. Data dictionary

Variable Name	Variable Type	Unit of Measurement	Description
market_id	Categorical	ID	The unique identifier for the city/region where DoorDash operates.
created_at	Timestamp	UTC	The date and time when the consumer submitted the order.
actual_delivery_time	Timestamp	UTC	The date and time when the order was delivered to the consumer.
store_id	Categorical	ID	The unique identifier representing the restaurant for which the order was submitted.
store_primary_category	Categorical	NA	The cuisine category of the restaurant.
order_protocol	Categorical	ID	The protocol ID through which the store receives orders from DoorDash.
total_items	Numerical	Items	The total number of items in the order.
subtotal	Numerical	Cents	The total value of the order submitted.
num_distinct_items	Numerical	Items	The number of distinct items included in the order.
min_item_price	Numerical	Cents	The price of the item with the least cost in the order.
max_item_price	Numerical	Cents	The price of the item with the highest cost in the order.
total_onshift_dashers	Numerical	Dashers	The number of available dashers who are within 10 miles of the store at the time of order creation.
total_busy_dashers	Numerical	Dashers	The subset of onshift dashers who are currently working on an order.
total_outstanding_orders	Numerical	Orders	The number of orders within 10 miles of this order that are currently being processed.
estimated_order_place_duration	Numerical	Seconds	The estimated time for the restaurant to receive the order

			from DoorDash.
estimated_store_to_consumer_driving_duration	Numerical	Seconds	The estimated travel time between the store and the consumer.

Table 2. Data Inconsistencies

market_id	created_at	actual_delivery_time	store_id	store_primary_category
4	2015-02-04 04:28:35	2015-02-04 05:00:18	77	japanese
4	2015-02-16 02:24:29	2015-02-16 03:12:06	77	japanese
4	2015-02-01 03:14:03	2015-02-01 04:00:36	77	japanese
3	2015-02-04 20:07:09	2015-02-04 20:50:45	77	thai
4	2015-01-26 19:32:23	2015-01-26 20:26:06	77	japanese
4	2015-02-04 01:19:36	2015-02-04 01:59:08	77	burger
4	2015-02-01 23:46:50	2015-02-02 00:18:44	77	japanese
4	2015-01-29 01:12:51	2015-01-29 01:50:37	77	japanese
1	2015-02-09 19:52:17	2015-02-09 20:56:44	77	chinese
4	2015-02-12 00:15:56	2015-02-12 00:45:20	77	japanese
4	2015-01-22 19:53:00	2015-01-22 20:41:29	77	japanese
4	2015-02-10 01:13:49	2015-02-10 01:58:10	77	japanese
4	2015-02-07 19:48:20	2015-02-07 20:23:55	77	japanese
4	2015-02-14 20:16:35	2015-02-14 20:50:11	77	japanese
4	2015-01-24 19:16:03	2015-01-24 19:44:55	77	japanese
1	2015-02-09 00:35:54	2015-02-09 01:13:22	77	american
4	2015-02-06 20:43:19	2015-02-06 21:07:30	77	japanese
4	2015-01-23 02:09:09	2015-01-23 03:09:59	77	japanese
4	2015-02-10 01:01:19	2015-02-10 01:30:14	77	japanese
4	2015-02-07 04:59:25	2015-02-07 05:35:20	77	japanese
4	2015-02-06 03:04:05	2015-02-06 03:46:29	77	japanese

Table 3. Descriptive Statistics Summary

	total_items	subtotal	min_item_price	max_item_price	estimated_order_place_duration	estimated_store_to_consumer_driving_duration
Min	1	100	0	60	0	0
1st Qu	2	1798	250	845	251	385
Median	3	3109	449	1095	251	546
Mean	4	2649	526	1177	308.3	547.8
3rd Qu	5	3885	725	1400	446	704
Max	411	26800	6022	8500	2715	2088

Table 5. Correlation Coefficient and Relationship between Variables

Variable 1	Variable 2	Correlation Coefficient
total_busy_dashers	total_onshift_dashers	0.9538
total_outstanding_orders	total_busy_dashers	0.9469
total_outstanding_orders	total_onshift_dashers	0.9337
estimate_non_preparation_duration	estimated_store_to_consumer_driving_duration	0.9245
total_items	num_distinct_items	0.8647

Table 6. Tolerance and Variance Inflation Factor (VIF)

Coefficients											
		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Zero-orde	Partial	Part	Tolerance	VIF
1	(Constant)	11.629	0.513		22.68	<.001					
	total_busy_dashers	-0.33	0.007	-0.631	-46.098	<.001	0.131	-0.134	-0.115	0.033	30.412
	total_onshift_dashers	-0.152	0.006	-0.317	-24.639	<.001	0.098	-0.072	-0.061	0.037	26.792
	total_outstanding_orders	0.316	0.003	1.014	122.792	<.001	0.216	0.338	0.305	0.09	11.06
	busy_dashers_ratio	13.295	0.351	0.147	37.853	<.001	0.161	0.11	0.094	0.409	2.443
	total_items	0.307	0.042	0.035	7.372	<.001	0.127	0.022	0.018	0.278	3.603
	min_item_price	-0.063	0.016	-0.013	-3.921	<.001	0.047	-0.011	-0.01	0.548	1.824
	max_item_price	0.167	0.014	0.046	11.953	<.001	0.152	0.035	0.03	0.423	2.363
	subtotal	0.117	0.005	0.116	21.564	<.001	0.21	0.063	0.054	0.214	4.669
	estimated_order_place_	0.003	0.041	0	0.078	0.938	0.105	0	0	0.461	2.169
	estimate_non_preparatic	1.143	0.011	0.267	99.975	<.001	0.274	0.281	0.248	0.867	1.154
	time_peak	5.288	0.097	0.15	54.64	<.001	0.214	0.158	0.136	0.818	1.222
	Weekend	2.416	0.085	0.072	28.471	<.001	0.127	0.083	0.071	0.958	1.043
	market_id	-0.429	0.032	-0.034	-13.305	<.001	-0.04	-0.039	-0.033	0.945	1.058
	order_protocol	-0.017	0.039	-0.002	-0.429	0.668	-0.083	-0.001	-0.001	0.501	1.996

a. Dependent Variable: total_delivery_duration

Table 7. Predictive Metrics Across Seven Models

Metrics	MLR	Reg Tree	RF	GBM	XGBoost	LightGBM
RMSE	14.6702	15.4257	14.3540	14.2936	14.1853	14.1674
MAE	11.1544	11.806	10.9245	10.8546	10.7706	10.7548

Table 8. "Real world" Prediction Using LightGBM Model

	Market ID	Store Primary Category	Order Protocol	Total Items	Subtotal	Max Item Price	Total Outstanding Orders	Busy Dashers Ratio	Estimate Non Preparation Duration	Percent Distinct item of Total	Price Range	Time Peak	Weekend
A	3	Sandwich	5	2	12.11	10.42	18	0.75	7.12	50	0.33	0	0
B	2	Mexican	2	3	12.24	11.99	174	0.88	12.85	66.67	7.74	1	0
C	1	American	3	6	52.17	13.18	197	1	19.27	100	8.29	1	1

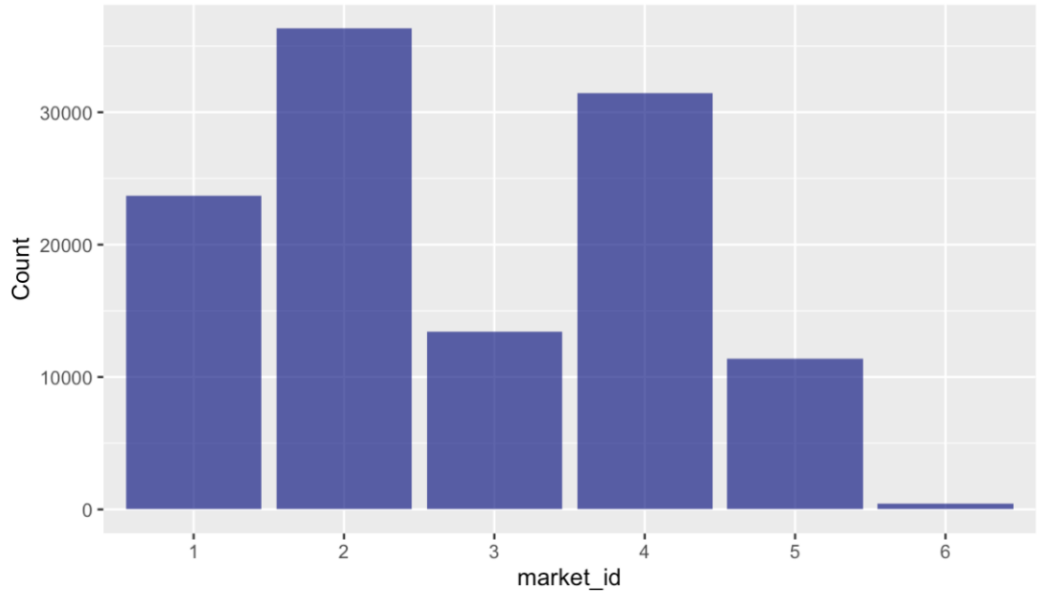

Figure 1. Distribution of City/region DoorDash Operation

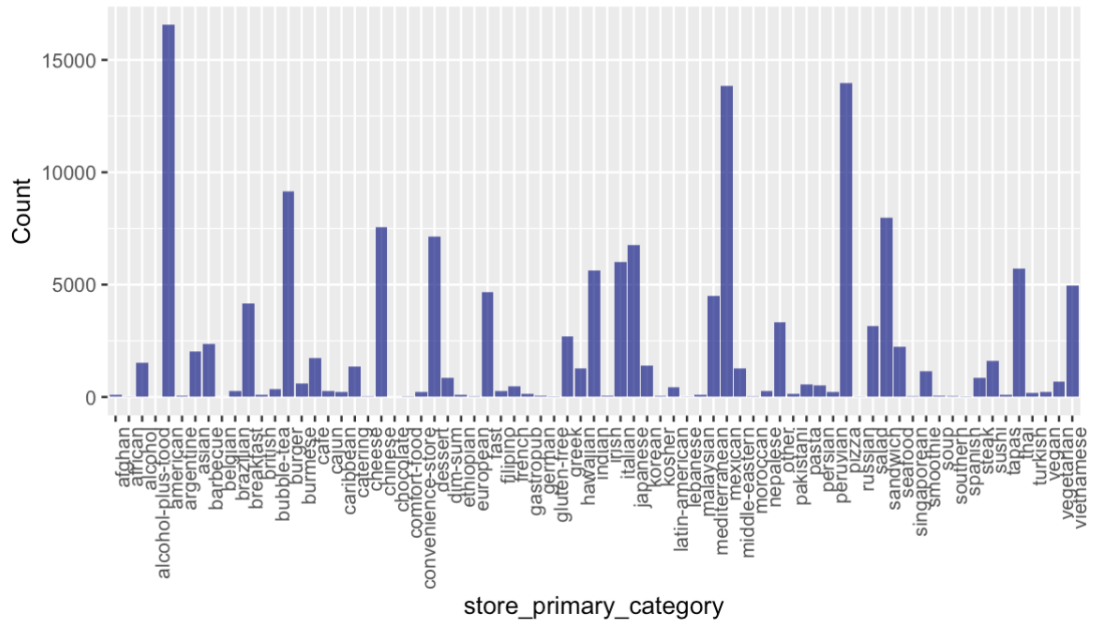

Figure 2. Distribution of Store Primary Category

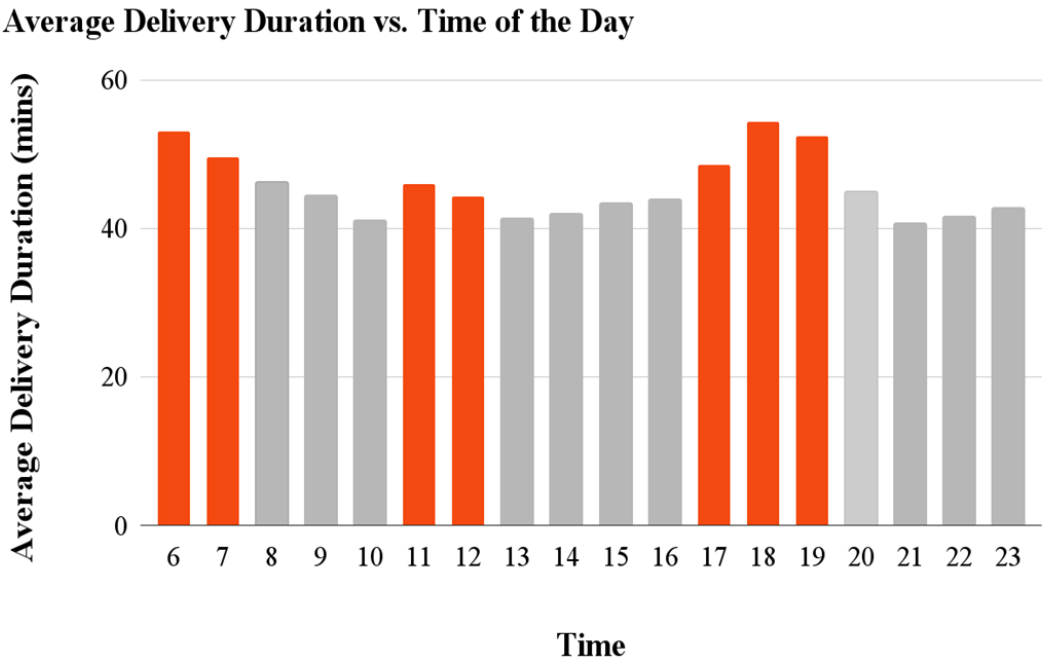

Figure 3. Time Peak

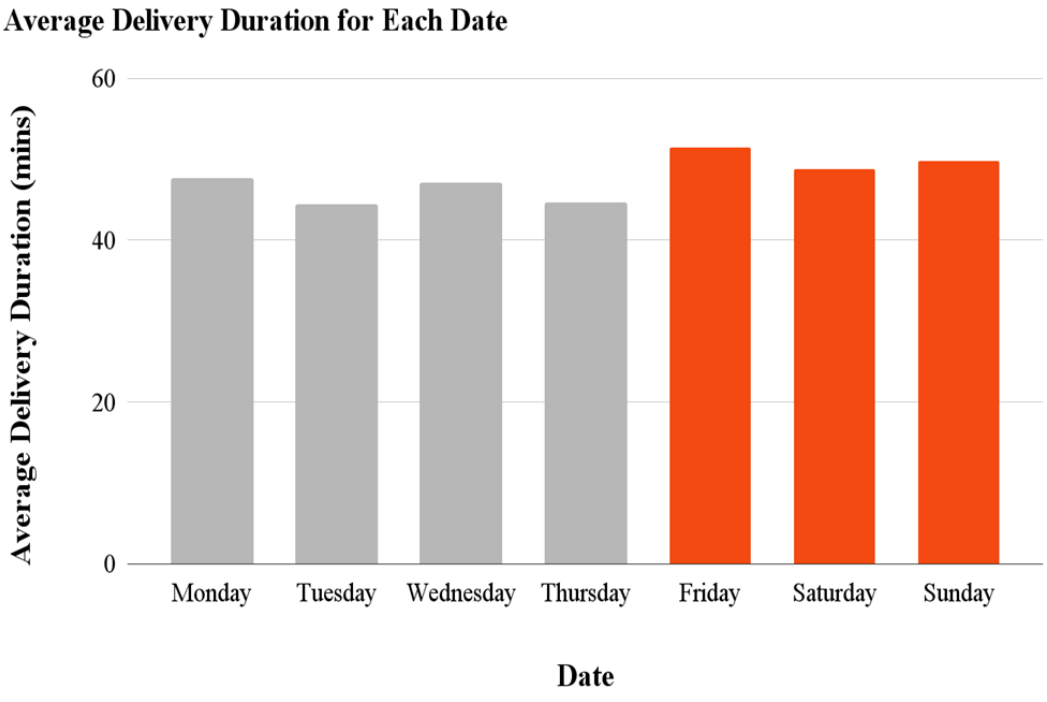

Figure 4. Dates Peak

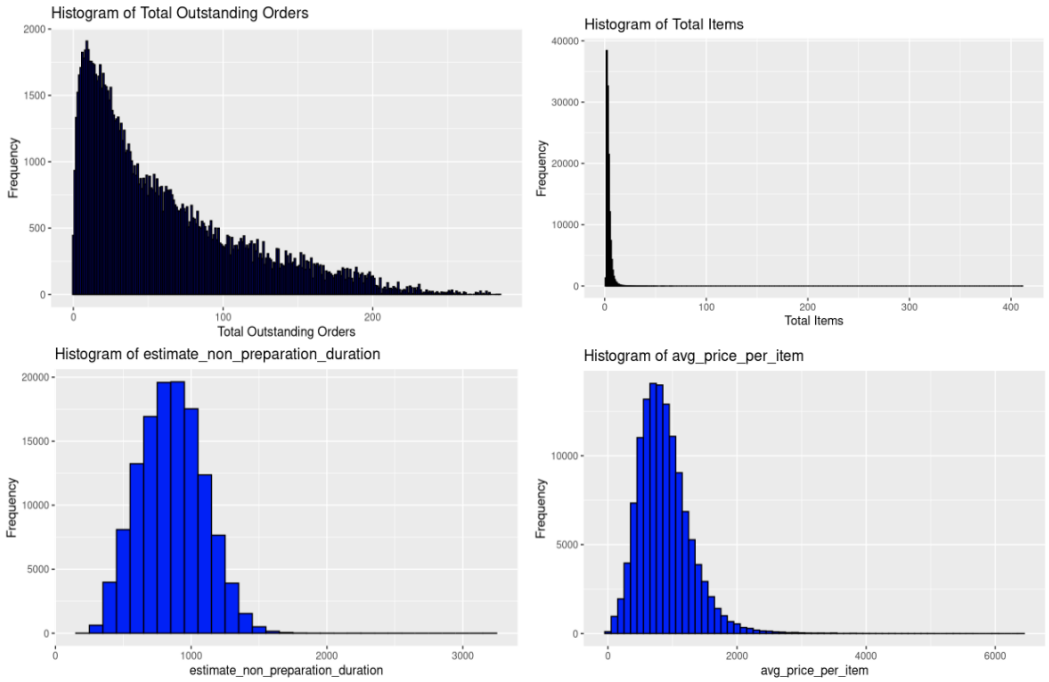

Figure 5. Histograms of Data Distribution

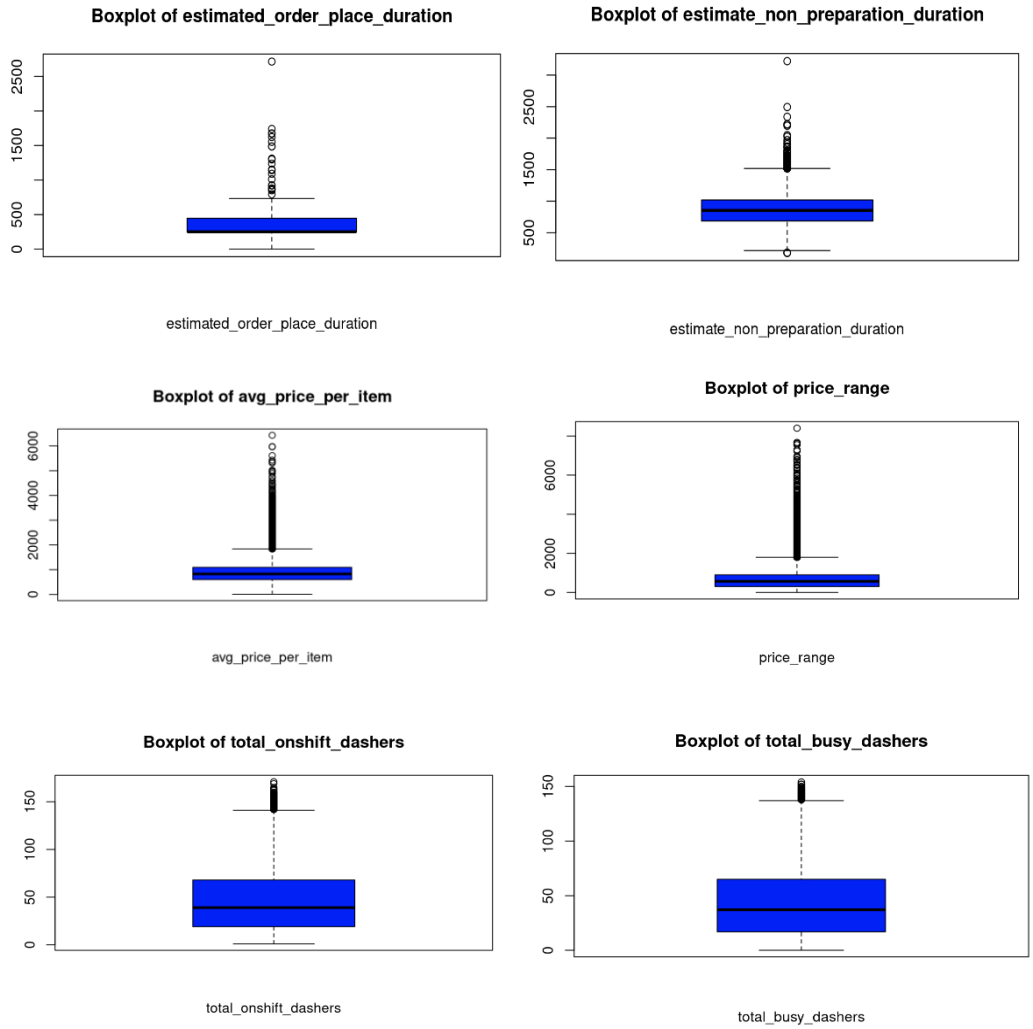

Figure 6. Box Plots of Numerical Variables

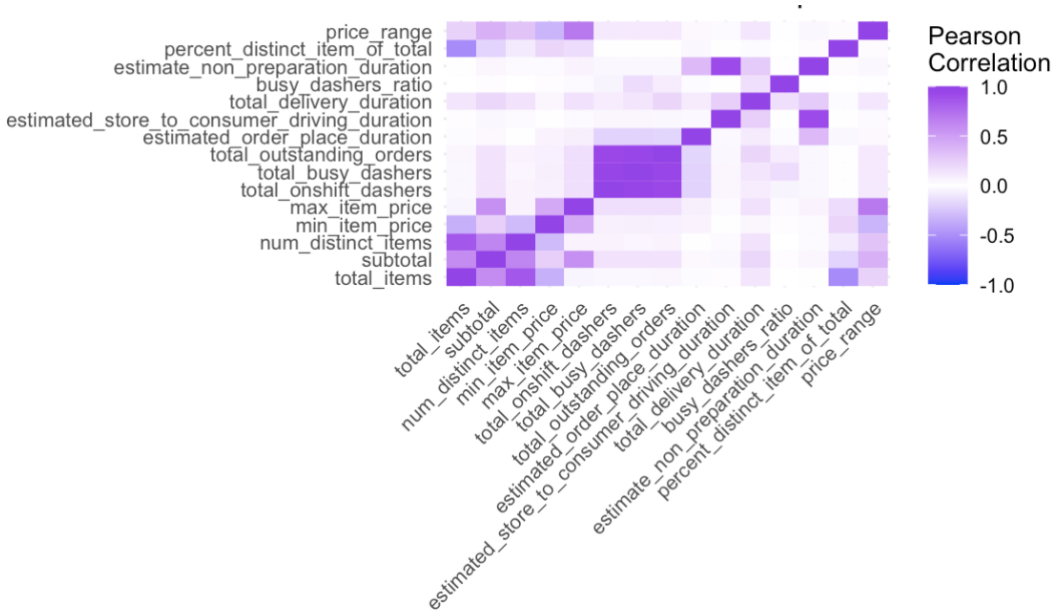

Figure 7. Correlation Matrix Heatmap

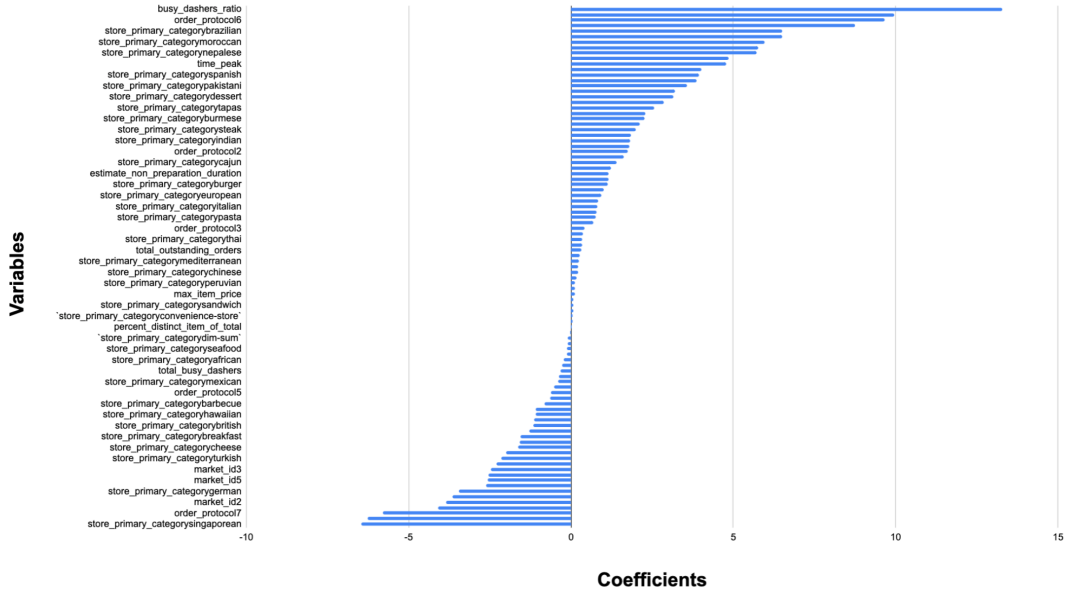

Figure 8. All Significant Variables Selected by LASSO

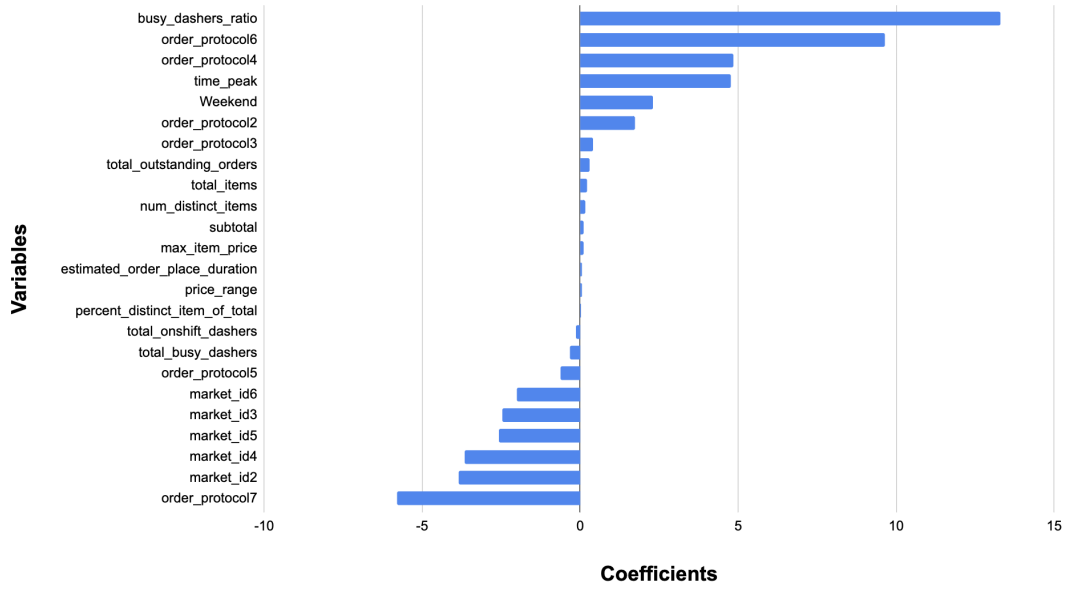

Figure 9. Significant Variables Selected by LASSO Excluded Significant Store Primary Categories

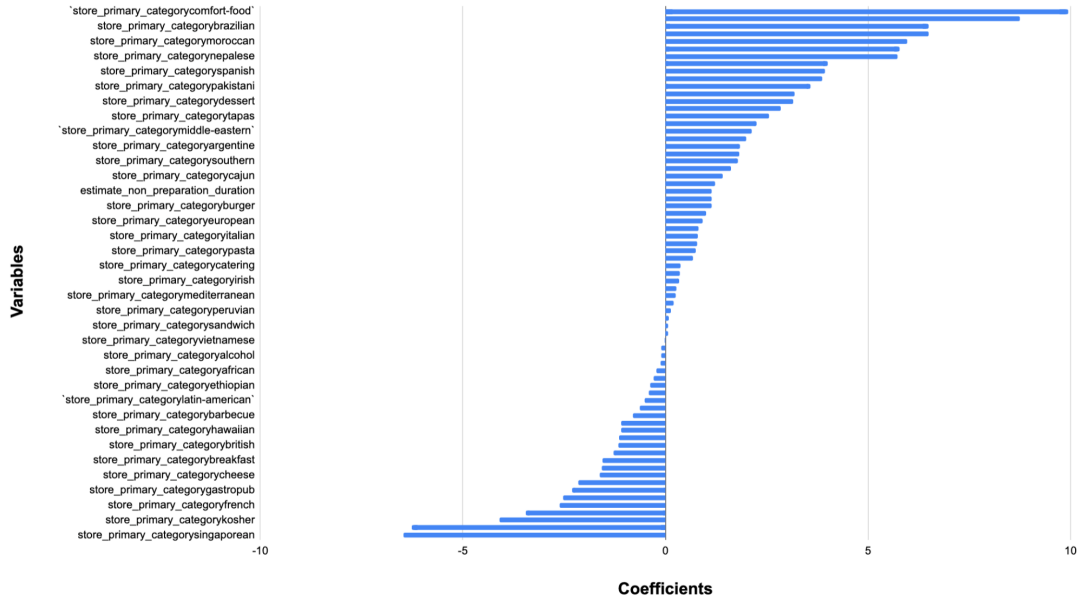

Figure 10. Significant Store Primary Categories Selected by LASSO

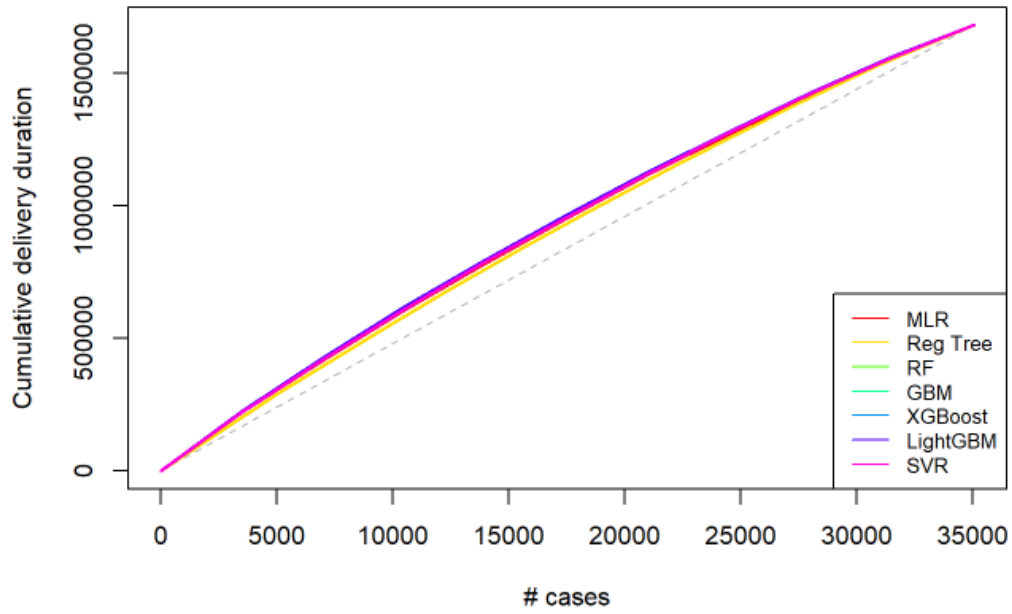

Figure 11. Lift Chart of all Model

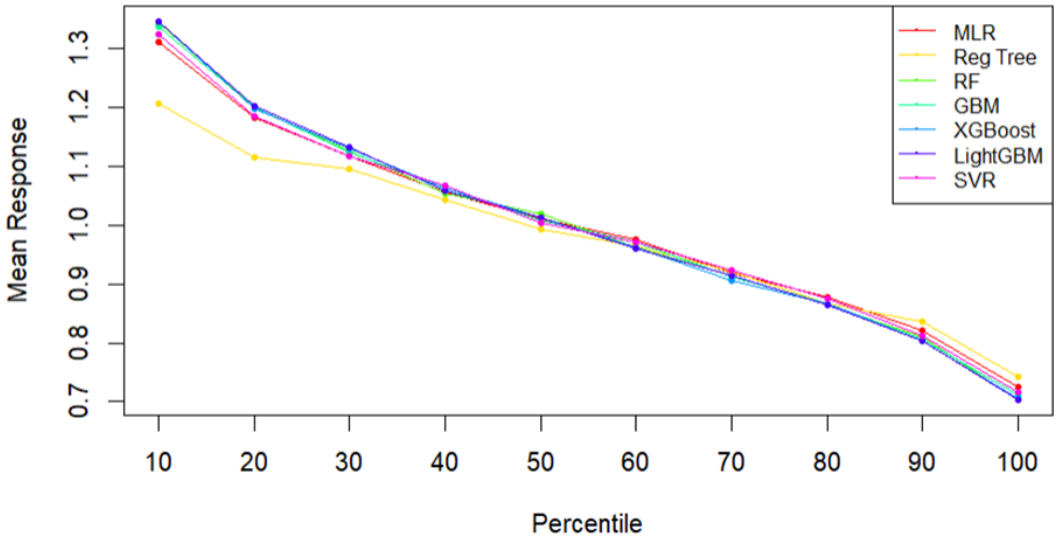

Figure 12. Decile-wise Lift Chart of all Model

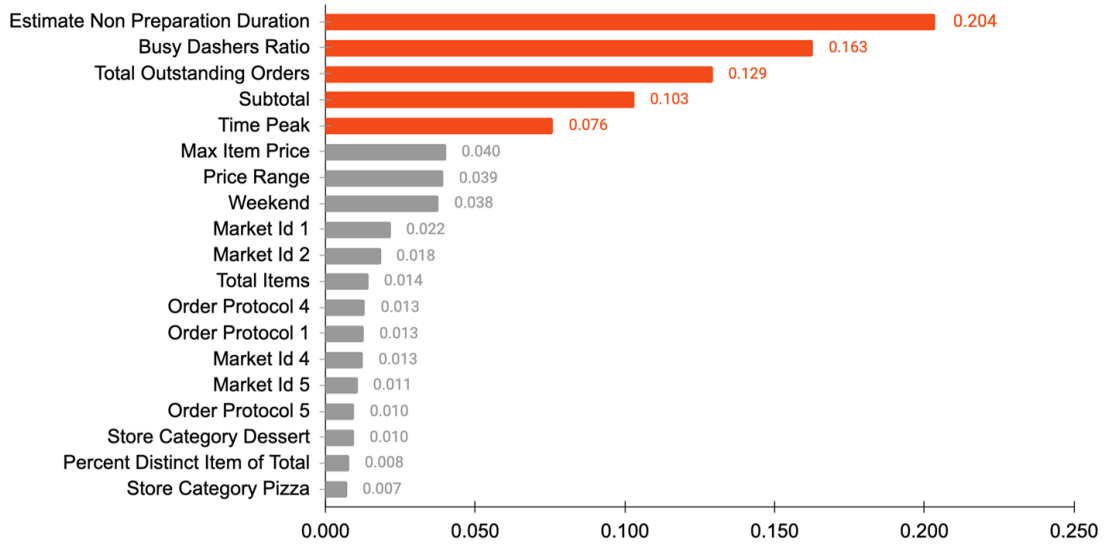

Figure 13. Most significant factors in LightGBM Model

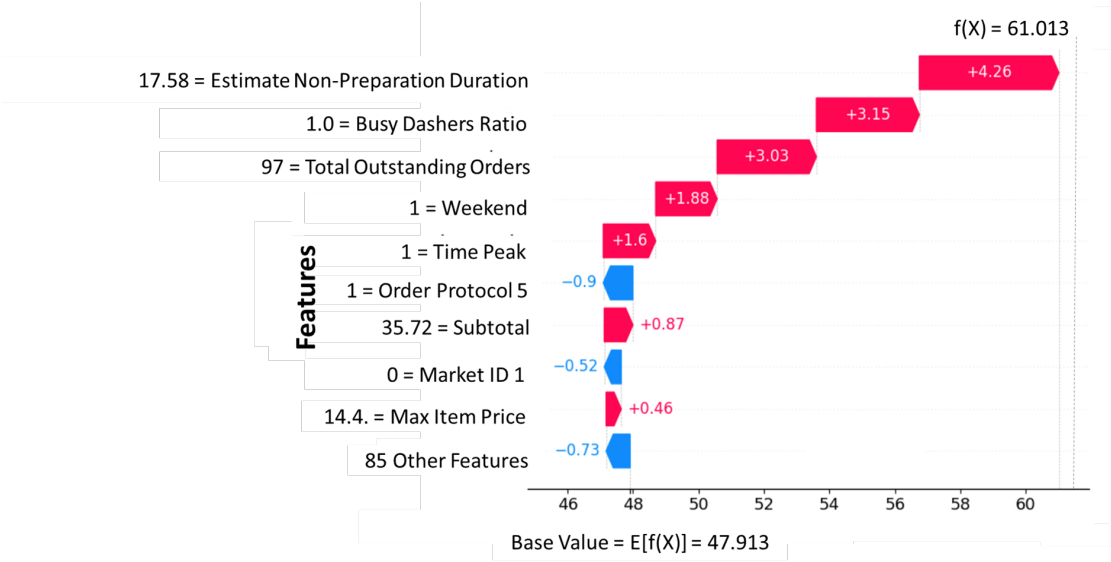

Figure 14. SHAP Value for LightGBM Model

SHAP Values Codes

```
import pandas as pd
import numpy as np
import shap
import lightgbm as lgbm
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt

# Load Data
data = pd.read_excel("/Users/iluv3trang/Desktop/MSBA/SPRING 2024/Capstone Project/CleanedDatasets/DoorDash2.xlsx")
train_data, valid_data = train_test_split( *arrays: data, test_size=0.3, random_state=42)

X_train = train_data.drop('total_delivery_duration', axis=1)
y_train = train_data['total_delivery_duration']
X_valid = valid_data.drop('total_delivery_duration', axis=1)
y_valid = valid_data['total_delivery_duration']

train_set = lgbm.Dataset(X_train, label=y_train)
valid_set = lgbm.Dataset(X_valid, label=y_valid)

```

```
params = {
    'objective': 'regression',
    'metric': 'rmse',
    'num_leaves': 31,
    'learning_rate': 0.05,
    'feature_fraction': 0.9,
    'bagging_fraction': 0.8,
    'bagging_freq': 5,
    'verbose': 1
}

num_round = 100
model = lgbm.train(params, train_set, num_boost_round=num_round, valid_sets=[valid_set])

y_pred = model.predict(X_valid, num_iteration=model.best_iteration)

import shap
explainer = shap.TreeExplainer(model)
shap_values = explainer(X_valid)

np.shape(shap_values.values)
shap.plots.waterfall(shap_values[14766])
```