**PRODUCT RECOMMENDATION SYSTEM**

OYTUN GÜNGÖR

# TABLE OF CONTENTS

[ABSTRACT i](#_Toc157518660)

[ÖZET ii](#_Toc157518661)

[TABLE OF CONTENTS iii](#_Toc157518662)

[LIST OF TABLES iv](#_Toc157518663)

[LIST OF FIGURES v](#_Toc157518664)

[INTRODUCTION 1](#_Toc157518665)

[1. LITERATURE REVIEW 2](#_Toc157518666)

[1.1. Personalized Marketing 2](#_Toc157518667)

[1.2. Predicting customer behaviour 3](#_Toc157518668)

[2. ABOUT THE DATA AND THE COMPANY 5](#_Toc157518669)

[2.1. About Instacart 5](#_Toc157518670)

[2.2. Product Related Datasets 6](#_Toc157518671)

[2.2.1. Products Dataset 6](#_Toc157518672)

[2.2.2. Aisles Dataset 6](#_Toc157518673)

[2.2.3. Department Dataset 6](#_Toc157518674)

[2.3. User Related Datasets 6](#_Toc157518675)

[2.3.1. Orders Dataset 6](#_Toc157518676)

[2.3.2. Order Products Prior and Train Datasets 7](#_Toc157518677)

[2.3. Project Definition and Methodology 7](#_Toc157518678)

[2.3.1. Machine Learning Algorithms 8](#_Toc157518679)

[2.3.2. Hyperparameter Optimization 8](#_Toc157518680)

[2.3.3. Tools and Libraries 8](#_Toc157518681)

[3. EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING 10](#_Toc157518682)

[3.1. Exploratory Data Analysis: 10](#_Toc157518683)

[3.2. Feature Engineering 14](#_Toc157518684)

[3.2.1. User-Based Features 15](#_Toc157518685)

[3.2.2. Product-Based Features 15](#_Toc157518686)

[3.2.3. User-Product Based Features 16](#_Toc157518687)

[RESULTS 17](#_Toc157518688)

[a. Algorithm Performance 17](#_Toc157518689)

[b. Data Handling and Model Development 17](#_Toc157518690)

[c. Model Consistency and Selection 18](#_Toc157518691)

[d. Comparison with Existing Studies 18](#_Toc157518692)

[REFERENCES 20](#_Toc157518693)

# LIST OF TABLES

[Table 3.1: User based features 15](#_Toc157372135)

[Table 3.2: Product based features 16](#_Toc157372136)

[Table 3.3: User-product based features 16](#_Toc157372137)

[Table 4.1: The results of the models employed 17](#_Toc157372141)

# LIST OF FIGURES

[Figure 3.1: Distribution of the evaluation set 10](#_Toc157372264)

[Figure 3.2: Order countsdistribution by days 11](#_Toc157372265)

[Figure 3.3: Order countsdistribution by hours 11](#_Toc157372266)

[Figure 3.4: User distribution according to the order counts 12](#_Toc157372267)

[Figure 3.5: Product distribution department by department 13](#_Toc157372268)

[Figure 3.6: Reorder status 13](#_Toc157372269)

[Figure 3.7: Top 20 most ordered products and top 5 departments 14](#_Toc157372270)

# INTRODUCTION

The integration of information technologies and the decrease in data collection expenses has had a significant impact on how individuals and organizations generate and store data. In today's world, the disclosure of information is closely intertwined with our daily lives. Our smartphones and applications constantly provide us with information about our location, supermarkets gather insights about our preferences through every transaction, and social media platforms offer us a platform to express our thoughts and emotions. As a result, companies are increasingly motivated to harness the vast amount of available data to gain a competitive edge (Lammervo, 2021).

The collection and study of retail transaction data, known as market basket analysis, have become increasingly prevalent in recent decades (Mauri, 2003). Many supermarkets, for example, issue loyalty cards. These cards not only provide discounts to customers but also allow retailers to develop a better understanding of individuals' purchasing habits by associating customers with transactions. The uses of this information vary and may include informing product placement decisions, designing personalized marketing campaigns, and determining the timing and extent of product promotions, etc. (Adomavicius & Tuzhilin, 1999; Agrawal & Srikant, 1994).

In this study, we use the 3m+ dataset of Instacart, an online grocery shopping application, to predict the products that users will buy in their next order. In addition to the orders data set, products, aisles, departments, and prior orders are combined with order data to prepare features.

The organization of this report is as follows. Section 1 provides a brief summary of the relevant literature, section 2 summarizes the dataset employed in this study and company, section 3 explains the exploratory data analysis and feature engineering, section 4 presents the results and section 5 is the conclusion.

# 1. LITERATURE REVIEW

## 1.1. Personalized Marketing

In the evolving digital landscape, there's a noticeable shift towards personalized technology. This trend empowers individuals to tailor their devices to meet their specific needs. This inclination towards customization is also prominent in marketing content creation, with a notable surge in the production of personalized digital content in recent years (Lammervo, 2021)

Personalized marketing strategies can be broadly categorized into three domains: content, data, and experience. A study by the Forbes Insights Team underscores the interconnection between personalization and customer experience. The study highlights that effective personalization is heavily dependent on data availability. It reveals that approximately 70 percent of consumers are likely to make purchases based on their experience alone. Furthermore, nearly 80 percent of consumers rate customer experience as equally significant as the quality of products or services (Morgan, 2019).

The traditional one-size-fits-all approach, while occasionally effective, finds limited applicability in digital marketing (Nair & Gupta, 2021). With the help of the data massively generated and collected by companies and the companies’ abilities of personalization, personalized marketing approaches are increasingly preferred over generic campaigns in the current digital era. (Chaffey & Chahdwick, 2019).

There has been extensive research aimed at understanding and predicting consumer decision-making. This has led to the concept of a multi-category shopping basket, representing a comprehensive collection of items that a consumer purchases together, commonly known as the market basket (Rauber et al., 2019). Both online and offline merchants have shown keen interest in understanding the contents of customers' market baskets. This interest is rooted in the potential to derive insights for informing personalized marketing strategies and targeted cross-selling initiatives (Kraus & Feuerriegel, 2019)

## 1.2. Predicting customer behaviour

The importance of predicting customer behavior has been known for decades, although it has started gaining popularity in recent years due to the aforementioned reasons (Miglautsch, 2000). For instance, Chen et al. (2005) extracted and analyzed the data to predict the customer’s behaviour using some then novel data mining techniques. However, in recent years, data generation has exploded, which allows us to use more sophisticated techniques about predicting customer behaviours.

Recently, Tahiri et al. (2019) analyzed the customer behaviour using data from the grocery stores in Canada. They employed two different neural network methods to predict the products that a customer will purchase and the store at which the customer will purchase. They utilize the transactional history along with other parameters such as promotion offers or distance of the grocery stores to the customer’s address. They noted that the success of the model, which is evaluated with the F1-score of 49% and 0.37, can be improved using more data.

Dou (2020) developed a model to estimate the customer’s consumption of certain products using a gradient lifting method. The dataset involves 17 unique features. They employed the CatBoost algorithm (Dorogush et al., 2018). The model achieved the accuracy of 88.51% with a sensitivity rate of 0.84 and an F1-score of 0.90. The author noted that the model performs well in cases with common overfitting problems and decreases the information loss in model establishment.

Kaneko et al. (2016) employed deep learning techniques that considered L1 regularization. They employed data that were collected at various supermarkets in Japan over the duration of three years as 29 months for training and 7 years for verification. Their accuracy varied from 75% to 86% based on the product attributes that they chose. The authors noted that deep learning models, even the ones not considering L1 regularization, were considerably better than logistic regression.

Sharma and Shafiq (2020) analyzed the historical data of Instacart to predict the possibility of retail items being purchased. They employed various techniques including Random Forests, Extreme Gradient Boosting (XGBoost) and Convolution Neural Networks along with a voting system for the final output. The model’s success was evaluated using accuracy, precision, F1 score and some other parameters. Their method that combines various techniques has shown similar success when compared with the other relevant studies.

# 2. ABOUT THE DATA AND THE COMPANY

## 2.1. About Instacart

Instacart is a technology company founded in 2012 in San Francisco representing an online grocery shopping platform operating in the United States and Canada. With over 80,000 physical stores and 600,000 shoppers, the platform brings together more than 750,000 daily users, offering them a convenient and fast way to do their grocery shopping.

The size and popularity of Instacart pose significant challenges in maintaining its success and increasing profitability. The complexity of its operations highlights the importance of critical processes such as demand planning, operational excellence, personalized product recommendations, and future order forecasting. Accurately predicting customer demands is vital for improving operational efficiency and enhancing user experience. Additionally, anticipating customers' future orders and planning personalized campaigns are crucial for sustaining Instacart's success.

In this context, Instacart gained attention in 2017 by organizing a significant data science competition on the Kaggle platform. The competition provided participants with anonymized data from over 200,000 users and more than 3,000,000 orders. Participants were tasked with developing data science models capable of predicting users' future orders. The dataset included a minimum of 4 and a maximum of 100 orders per user, along with information on the day and time of each order and the number of days since the previous order. This competition aimed to contribute to Instacart's goals of improving future order forecasting and enhancing user experience. As a result, large-scale online grocery shopping platforms like Instacart play a crucial role in accurately predicting customer demands and optimizing operational processes. These platforms leverage technologies such as data science and artificial intelligence to improve user experience and sustainably grow their businesses using various sources of data. The 2017 Kaggle competition organized by Instacart can be considered a significant step that underscores the importance of data science and analytics in shaping the future success of such platforms.

## 2.2. Product Related Datasets

In this capstone project, we have six data files: Department, Aisles, Products, Orders, Order Products Prior, and Order Products Train. Each dataset provides valuable insights into different aspects of the e-commerce platform. Let's take a closer look at each dataset and its contents.

### 2.2.1. Products Dataset

The data set contains detailed information about products.

* **product\_id:** Unique identifier for each product.
* **product\_name:** The name of the product.
* **aisle\_id:** Indicates the aisle in which the product is located.
* **department\_id:** Indicates the department to which the product belongs.

### 2.2.2. Aisles Dataset

This dataset provides information about the various aisles in the store.

* **aisle\_id:** Unique identifier for each aisle.
* **aisle:** The name of the aisle.

### 2.2.3. Department Dataset

This dataset includes details about different departments in the store.

* **department\_id:** Identifier for each department.
* **department:** The name of the department.

## 2.3. User Related Datasets

### 2.3.1. Orders Dataset

This dataset contains information about orders placed by users. It consists of 3.2 million orders placed by 206,209 unique users.

* **order\_id:** Unique identifier for each order.
* **user\_id:** Unique identifier for each customer.
* **eval\_set:** Indica tes which evaluation set the order belongs to (prior, train, or test).
* **order\_number:** Represents the order sequence number for each user.
* **order\_dow:** The day of the week when the order was placed (0 = Sunday, 3 = Wednesday).
* **order\_hour:** The hour of the day when the order was placed.
* **Days\_since\_pri or\_order:** Number of days between the current order and the previous order by the same user.

### 2.3.2. Order Products Prior and Train Datasets

These two datasets share the same structure and provide information about the products ordered by users.

* **order\_id:** Foreign key linking to the order in which the product was purchased.
* **product\_id:** Foreign key linking to the product details.
* **add\_to\_cart\_order:** The order in which each product was added to the cart.
* **reordered:** Indicates whether the user has ordered this product in the past (1 if reordered, 0 otherwise).

These datasets collectively offer valuable insights into product details, aisle information, department categorization, user behavior, and ordering patterns. By analyzing these datasets, we can gain a comprehensive understanding of the e-commerce platform and make data-driven decisions to improve customer experience and drive business growth.

## 2.3. Project Definition and Methodology

This capstone project aims to leverage the publicly available dataset released by Instacart to predict future customer orders using machine learning techniques. The core of this project is framed as a classification problem, where the model predicts whether a user will repurchase a product. Given the inherent class imbalance in the dataset, an oversampling strategy was employed to mitigate this imbalance, thereby reducing the potential bias in the model.

### 2.3.1. Machine Learning Algorithms

* A series of machine learning algorithms were systematically applied and evaluated. The algorithms included:
* Logistic Regression: A foundational approach for binary classification problems.
* XGBoost (Extreme Gradient Boosting): An efficient and scalable implementation of gradient boosting.
* CatBoost: An algorithm specializing in handling categorical data.
* LightGBM (Light Gradient Boosting Machine): A fast, distributed, and high-performance gradient boosting framework.
* Random Forest Classifier: An ensemble method known for its robustness and accuracy.
* Decision Trees: A simple yet effective classification and regression technique.

### 2.3.2. Hyperparameter Optimization

For optimizing the hyperparameters of these models, Randomized Search Cross-Validation (Randomized Search CV) was utilized. This method efficiently searches the hyperparameter space and enhances the performance of the machine learning models. [ref]

### 2.3.3. Tools and Libraries

The project was developed using Python version 3.12.1, which is known for its powerful data handling and machine learning capabilities. Jupyter Notebook served as the primary development environment, offering an interactive and user-friendly interface for coding, data analysis, and visualization. Additionally, Google Colab Pro+ was utilized for running one of the models, leveraging its advanced computational resources to handle the intensive processing requirements of the machine learning algorithms.

This methodology section outlines the systematic approach adopted for predicting user behavior in the context of online grocery shopping, emphasizing the use of advanced machine learning techniques and tools to address a practical and relevant problem in the realm of e-commerce analytics.

The key libraries we use are Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, LightGBM, CatBoost, XGBoost, Joblib, Pickle, Astropy, PIL, SQLite3, Scipy.

# 3. EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING

## 3.1. Exploratory Data Analysis:

In the initial phase of the exploratory data analysis process, we started by examining the summary information of our available datasets. By looking at the distributions of classes within these datasets, we gained insights into their characteristics. For instance, analyzing the percentage distribution of the 'eval set' column in the 'orders' dataset provided a foundational understanding of its structure.

One notable observation across all datasets was the presence of a single NA value, which we decided to remove for the integrity of our analysis.

Figure 3.1 illustrates the distribution of the 'evaluation set' column in the 'orders' dataset. This significant proportion marked as 'prior' indicates a substantial dataset available for feature creation. It reveals a predominant percentage of 94% marked as 'prior', followed by 4% as 'train' and 2% as 'test'.

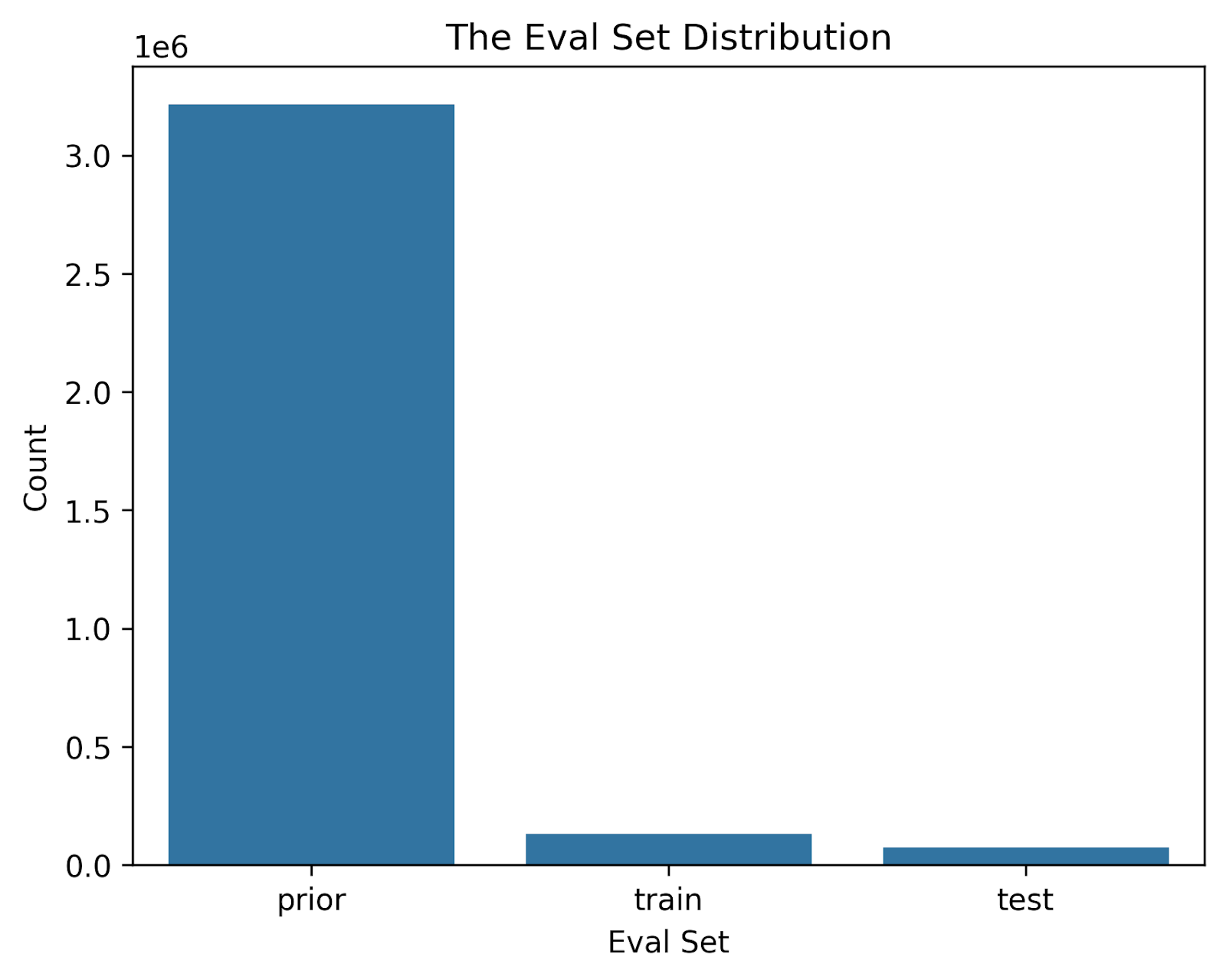


Figure 3.1: Distribution of the evaluation set

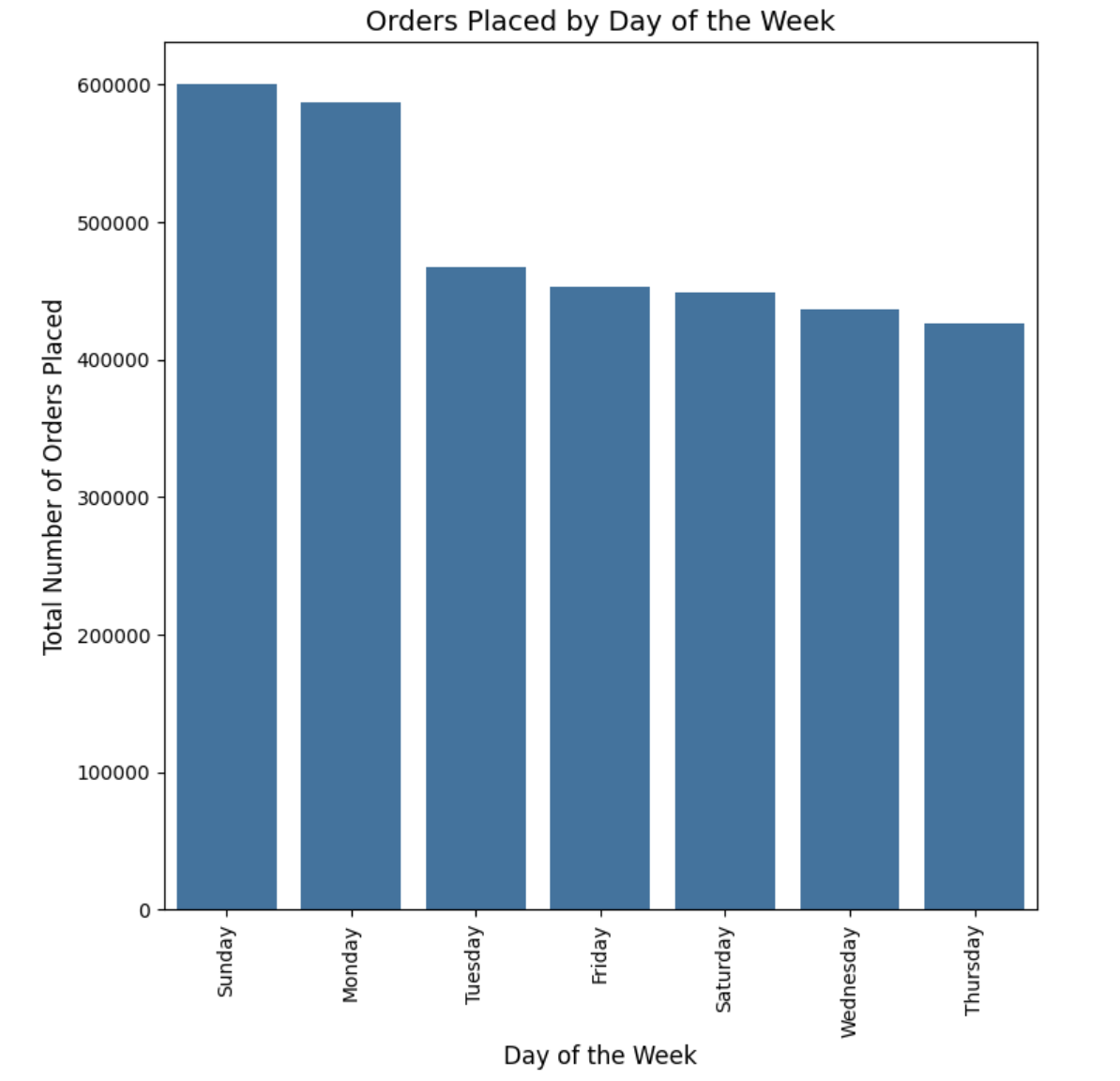
****

Figure 3.2: Order countsdistribution by days

Figure 3.2 displays the distribution of order counts across different days. It highlights that Sundays and Mondays are notably busier compared to other days. The highest number of orders is seen on Sundays, exceeding 600,000 orders, closely followed by Mondays with a roughly 2.5% difference. The least busy day for orders turns out to be Thursday.

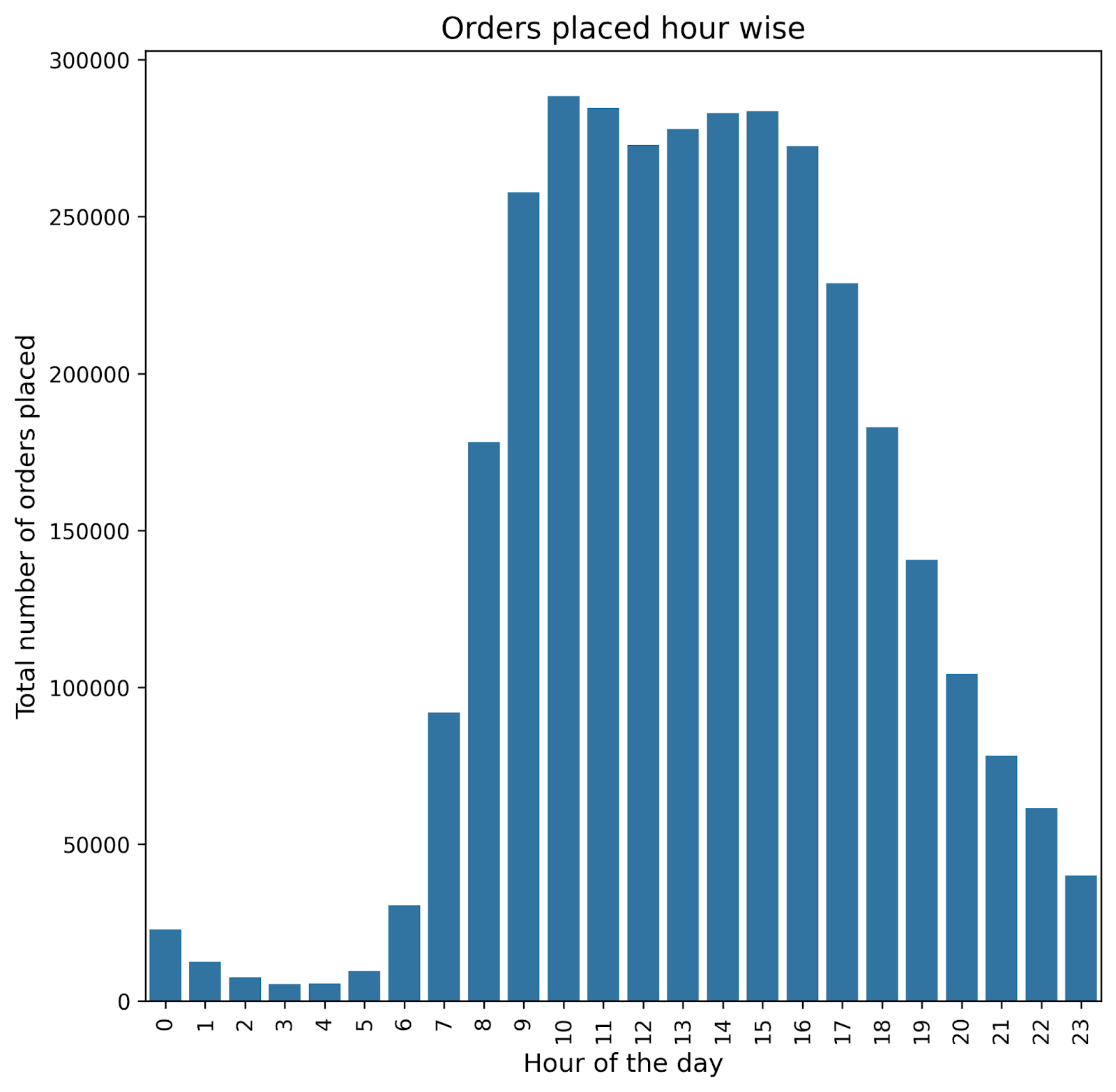


Figure 3.3: Order countsdistribution by hours

Figure 3.3, same as figure 2, shows the total number of orders distributed across various hours of the day. From the graph, we observe a sharp increase in order counts starting at 7 AM, peaking at around 10 AM with over 25,000 orders. This high level of activity remains consistent until 5 PM, after which there is a gradual decline, with the hourly order count dropping below 5,000 past midnight.

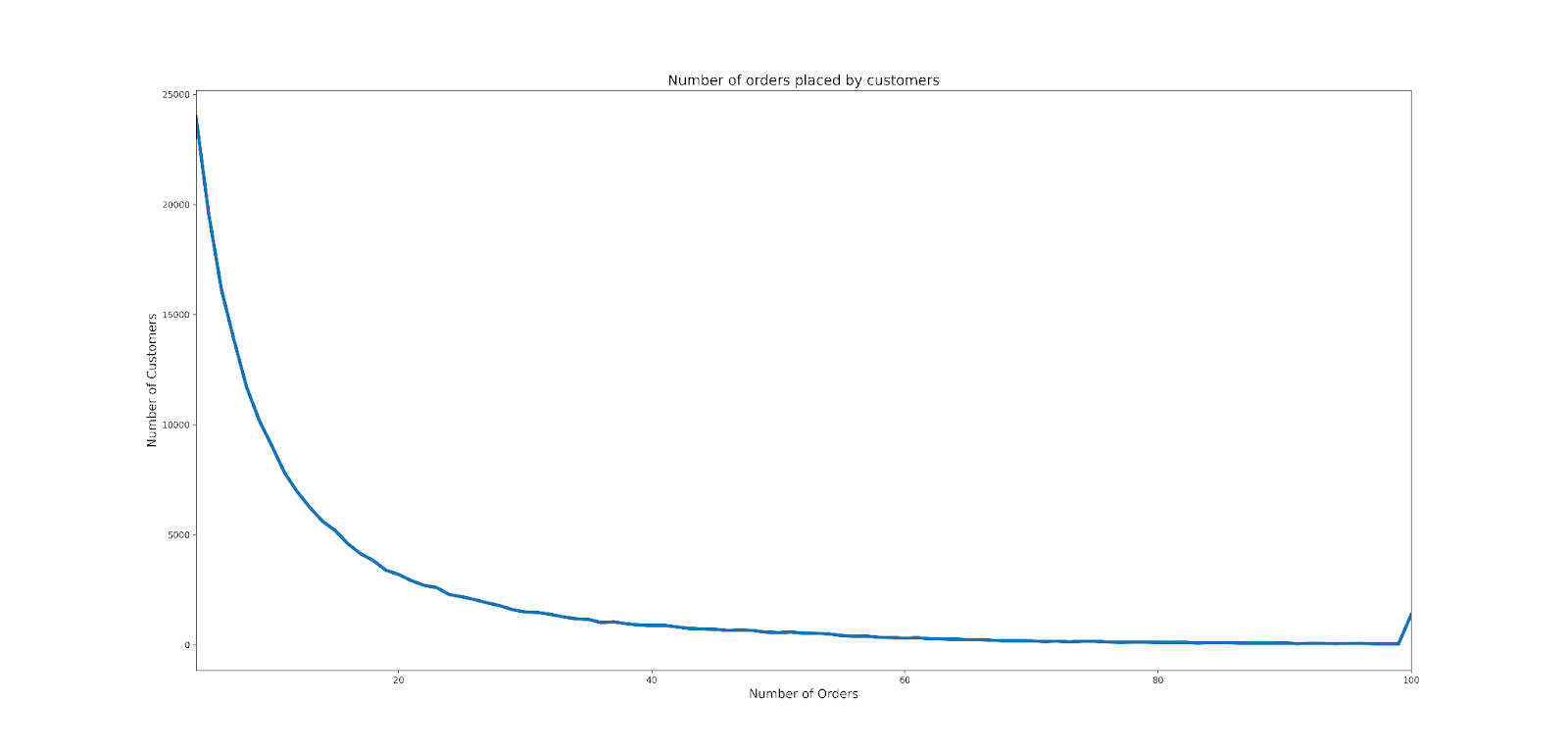
****

Figure 3.4: User distribution according to the order counts

Figure 3.4 illustrates the distribution of users based on their order counts. According to the dataset description, there are a total of 206,209 users whose order data is included. The minimum number of orders made by a user is 4, while the maximum is 100. The number of users decreases as the number of orders increases, indicating an inverse relationship. However, there is a slight increase in the number of users at 100 orders, which can be attributed to the fact that users with more than 100 orders are also categorized under this group. Notably, users with 10 or fewer total orders make up a significant portion of the user base, totaling 104,513 users. This group accounts for over 50% of all users and exhibits a right-skewed distribution.

Figure 3.5 shows the distribution of distinct products across different departments, offering a categorical representation of the product variety. The department with the highest number of unique items is personal care, boasting over 6,000 products, followed closely by snacks, also surpassing the 6,000. On the other hand, the department with the lowest number of unique products is bulk.

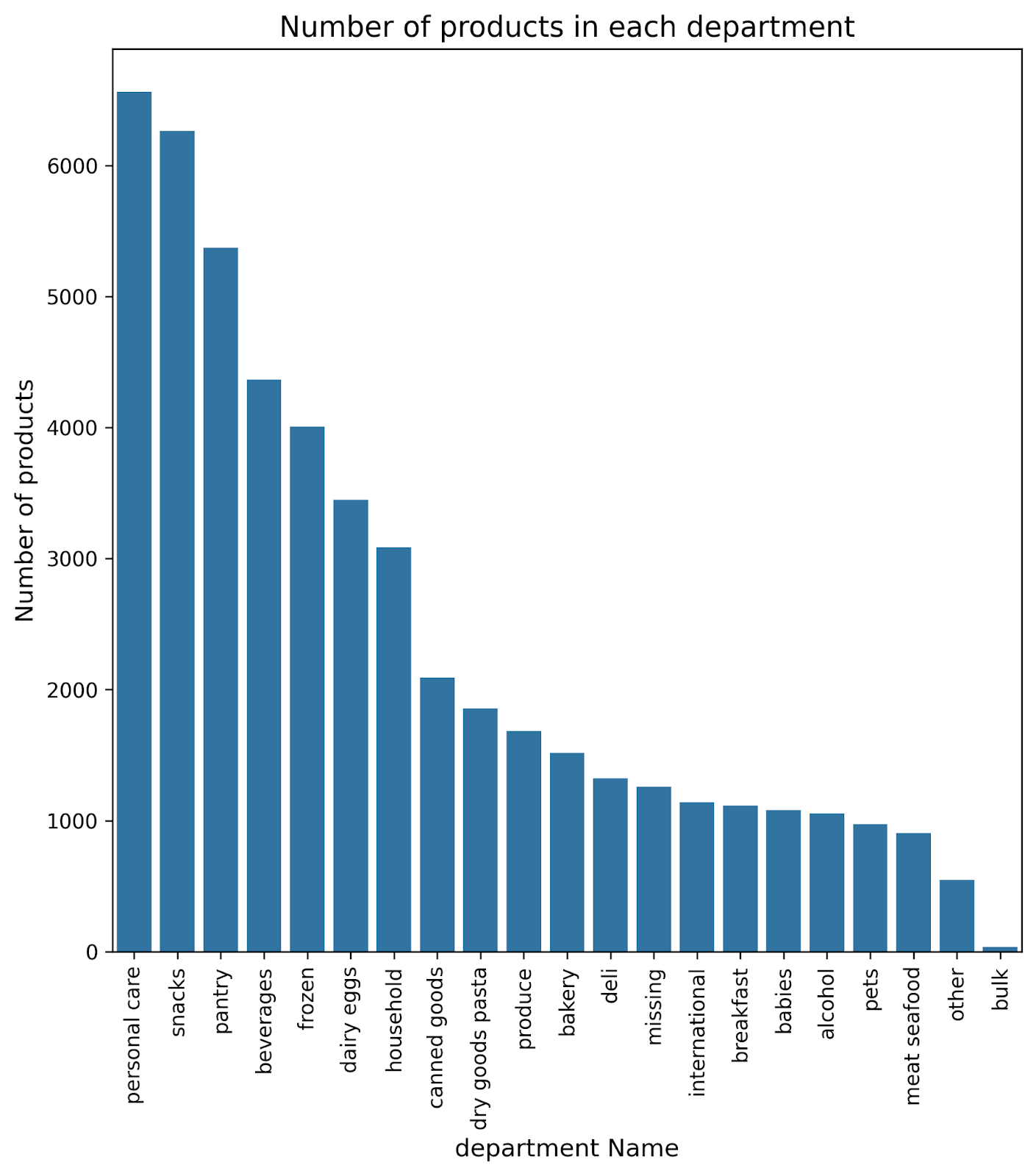


Figure 3.5: Product distribution department by department

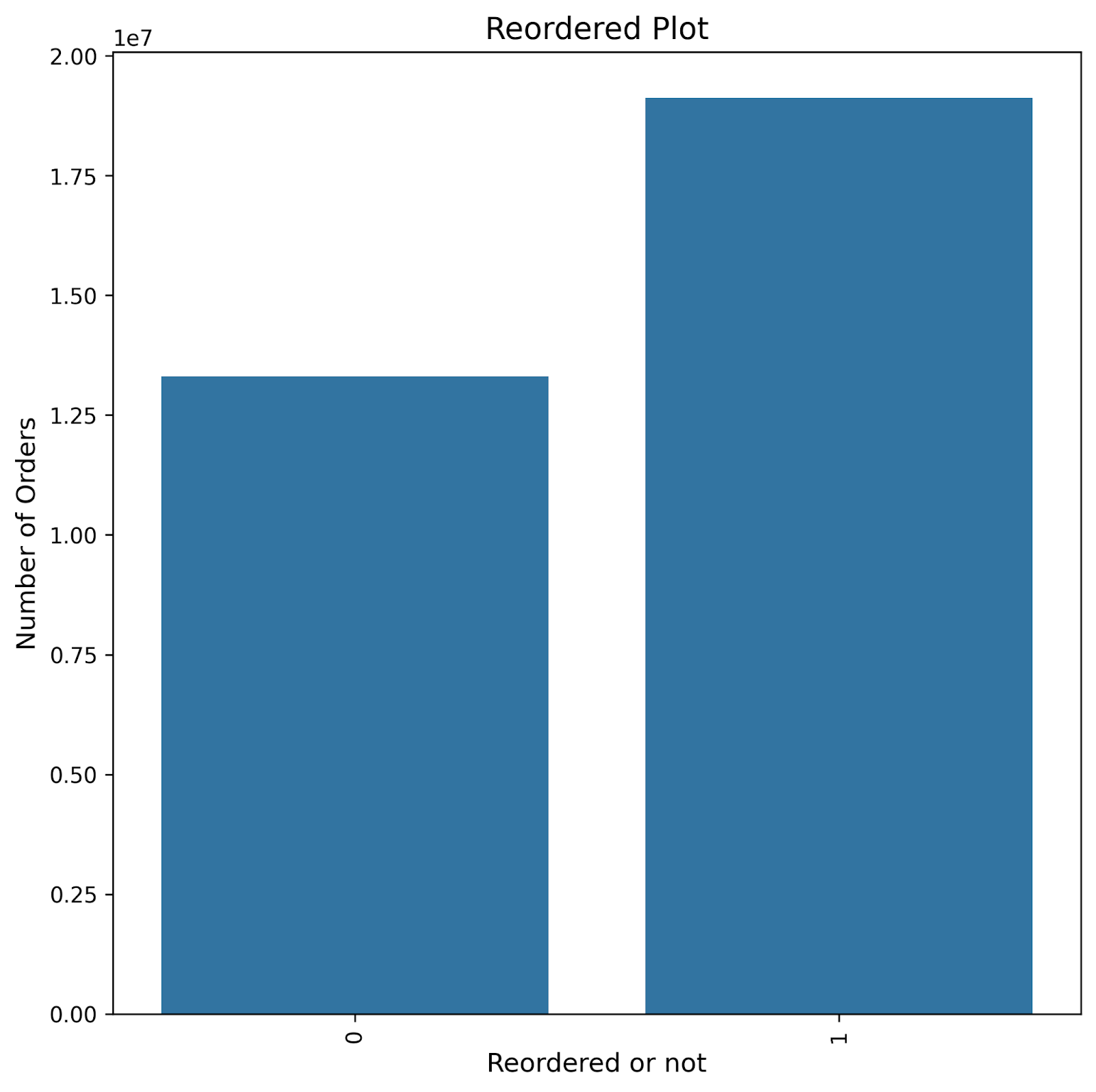


Figure 3.6: Reorder status

Figure 3.6 presents the reorder status of products, with a categorization of '0' indicating products that have not been reordered, and '1' indicating products that have been reordered. The analysis reveals a noteworthy finding: within the dataset of 'order products prior', there exists an approximate 60% reorder rate. This finding suggests a substantial inclination towards repeat purchases among the products under consideration.

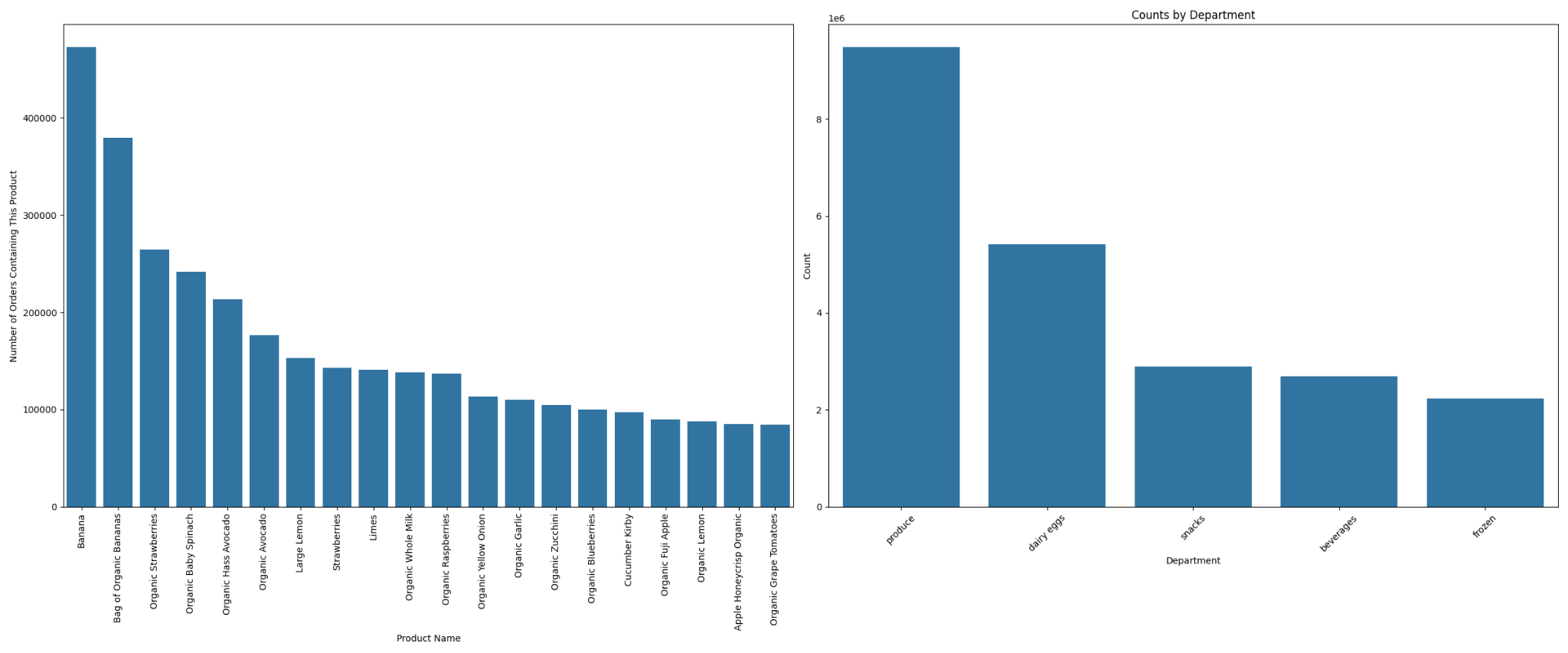


Figure 3.7: Top 20 most ordered products and top 5 departments

Figure 3.7 presents a comprehensive analysis of consumer preferences and market trends by showcasing the top 20 most frequently ordered products and the top 5 departments. This visual depiction provides valuable insights into the popularity of specific products among users and the corresponding departments to which these products belong. By succinctly representing this information, the figure facilitates a deeper understanding of key areas of consumer demand and the strategic significance of different departments, thereby aiding in inventory management and strategic marketing planning.

## 3.2. Feature Engineering

In the field of retail analytics, particularly when working with datasets such as Instacart's, the phase of feature engineering is of crucial importance. Feature engineering can be broadly categorized into three segments: user-based features, product-based features, and user-product features. This process heavily relies on the group-by functionality provided by the pandas library, in addition to basic mathematical operations performed on each column. In this discussion, we will focus on the development and significance of user-based features.

### 3.2.1. User-Based Features

The primary objective of this study is to develop a forecasting model for predicting future orders of Instacart users. This objective emphasizes the significance of creating user-centric features that can provide valuable insights into consumer purchasing behavior.

The dataset used in this study consists of 206,209 unique users. By leveraging the 'prior orders' dataset, we have generated a comprehensive set of user-based features, which have been integrated into a unified dataset. This dataset, referred to as the User-Based Features dataset, encompasses various attributes related to users. The final dataset comprises 206,209 rows and 13 columns, including features such as Unique Product Count, Total Product Count, Average Basket Size, User Reorder Count, as well as additional attributes related to order timings and frequencies (refer to Table 1 for a detailed list).

Table 3.1: User based features

|  |  |
| --- | --- |
| **Feature Name** | **Feature Name** |
| Unique Product Count | User Most Order Count |
| Total Product Count | User Most Order Hod |
| Total Orders Count | User Most Order Count in Most Ordered Hour |
| Average Basket Size | User Most Order Count in Most Ordered Day |
| User Order Hours of the Day Count | User Reorder Count |
| User Order Day of the Week Count | User Reorder Ratio |
| User Most Order Day of the Week | User Most Ordered Department |
| User Most Ordered Aisle |  |

### 3.2.2. Product-Based Features

In the context of Product-Based Features, the dataset comprises 49,677 distinct products. Similar to the User-Based Features, these features were generated using pandas groupby and mathematical operations. Each feature was carefully prepared at the product level and subsequently merged based on the product ID. As a result, the Product-Based Features dataset contains 49,677 rows and 8 columns.

Table 3.2: Product based features

|  |  |
| --- | --- |
| **Feature Name** | **Feature Name** |
| Product Order Count | Product Most Ordered Day |
| Product Reorder Count | Product Most Ordered Hour |
| Product Reorder Rate | Product Average Order Hour |
| Product Average Added to the Cart | Product Average Order Day |

### 3.2.3. User-Product Based Features

This category involves merging user ID and product ID pairs to create user-product-based features. By grouping the prior orders dataset by these IDs, we can generate relevant features. This process results in a dataset containing 13,307,953 unique user-product combinations, which translates to 13,307,953 rows and 9 columns. Once these sets of features have been developed, they are combined into a comprehensive dataset, using appropriate foreign keys for association. The subsequent steps include standardizing the feature dataset using StandardScaler and implementing a train-test split for model validation. These methods demonstrate the sophisticated process of feature engineering in large-scale retail data analysis, establishing a solid foundation for building more accurate predictive models in consumer behavior.

Table 3.3: User-product based features

|  |  |
| --- | --- |
| **Feature Name** | **Feature Name** |
| User Product Order Count | User Product average added to cart |
| User Product Reorder Count | User Product First Order Number |
| User Product Most Ordered Day | User Product Last Order Number |
| User Product Most Ordered Hour | User Product Reorder Ratio |
| User Product Most Added to Cart |  |



# RESULTS

In this capstone project, we aimed to predict the future orders of Instacart users by utilizing the Instacart dataset. This involved setting up a classification problem on an imbalanced dataset to forecast which products users would order in their subsequent purchases.

## a. Algorithm Performance

We employed various algorithms and assessed their performance based on F1-Scores across training, validation, and test datasets, along with private and public leaderboards in a Kaggle competition context. The performance of the models was as follows:

Table 4.1: The results of the models employed

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODELS** | **Train F1 Score** | **Validation F1 Score** | **TEST** | |
| **Private F1 Score** | **Public F1 Score** |
| Random Forest | 0,426 | 0,421 | 0,367 | 0,368 |
| Light GBM | 0,445 | 0,436 | 0,364 | 0,366 |
| XG Boost | 0,443 | 0,4350 | 0,377 | 0,377 |
| Decision Trees | 0,417 | 0,416 | 0,364 | 0,366 |

Among these, the XG Boost algorithm exhibited the best performance, although the results across all models were closely aligned.

## b. Data Handling and Model Development

To address the imbalance in the dataset, the 'stratify' parameter was used while splitting the dataset. Additionally, we experimented with oversampling techniques during the model development phase. However, as oversampling did not lead to significant improvements in model performance, it was not adopted in the final models.

The hyperparameters for the models were fine-tuned using Randomized Search CV followed by Grid Search CV, ensuring the identification of the most optimal parameters for training.

## c. Model Consistency and Selection

The consistency across training, validation, and test scores suggests the absence of overfitting or underfitting in our models. This reliability across different data segments reinforced the robustness of our modeling approach.

Ultimately, based on the comprehensive evaluation of performance metrics, the XG Boost algorithm was selected as the most effective model. This model demonstrated superior accuracy in predicting future purchasing preferences of Instacart users.

## d. Comparison with Existing Studies

In the present study, a total of 32 features were incorporated. Similar to other approaches in the field, this research utilized user-based, product-based, and user-product interaction features. A range of algorithms, including XGBoost, Light Gradient Boosting, Decision Trees, and Random Forest Classifier were employed. Among these, XGBoost achieved the highest F1 score of 0.377, indicating its effectiveness in this context.

Parmar (2020) studied the same dataset and utilized 78 features. Parmar's approach to feature engineering shares similarities with the current study. Notably, Parmar employed ensemble methods in the modeling process leveraging XGBoost and Light Gradient Boosting algorithms to achieve a notable F1 score of 40%. Conversely, Agrawal(2020)’s study, which also analyzed the same dataset, utilized a different set of 46 features. Agrawal's study employed Light Gradient Boosting and Cat Boost algorithms, culminating in an F1 score of 0.37. Both studies obtain F1 scores that are close to the present study. A critical examination of both models reveals a common pattern: more successful predictive outcomes were obtained for class 0, while the results for class 1 were comparatively lower. This trend underscores the challenges in accurately predicting class 1. This is also an issue in our study.

These observations indicate that despite variations in the number of features and the algorithms used, Parmar (2020)’s and Agrawal (2020)'s studies and the current study achieved similarly effective results. However, the better F1 score obtained by Parmar underscores the impact of number of features and algorithm choice on the performance of classification models. The utilization of robust algorithms like XGBoost and Light Gradient Boosting are very useful for this kind of studies. Observing lower success rates for class 1 predictions suggests that there is room for improvement for the model's ability to accurately predict class 1.

# REFERENCES

Adomavicius, G., & Tuzhilin, A. (1999, August). User profiling in personalization applications through rule discovery and validation. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 377-381).

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Chaffey, D., & Ellis-Chadwick, F. (2019). Digital marketing. Pearson uk.

Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4), 773-781.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

Dou, X. (2020, April). Online purchase behavior prediction and analysis using ensemble learning. In *2020 IEEE 5th International conference on cloud computing and big data analytics (ICCCBDA*) (pp. 532-536). IEEE.

Kaneko, Y., & Yada, K. (2016, December). A deep learning approach for the prediction of retail store sales. In *2016 IEEE 16th International conference on data mining workshops (ICDMW)* (pp. 531-537). IEEE.

Kraus, M., & Feuerriegel, S. (2019, July). Personalized purchase prediction of market baskets with Wasserstein-based sequence matching. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2643-2652).

Lammervo, S. (2021). Towards personalization of content marketing through data-driven customer experience.

Mauri, C. (2003). Card loyalty. A new emerging issue in grocery retailing. *Journal of Retailing and Consumer Services,* 10(1), 13-25.

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management*, 8, 67-72.

Morgan, B. (2021). Three lasting changes to grocery shopping after covid-19. *Forbes*. Retrieved November, 25, 2023.

Nair, K., & Gupta, R. (2021). Application of AI technology in modern digital marketing environment. *World Journal of Entrepreneurship, Management and Sustainable Development*, 17(3), 318-328.

Rauber, A., Trasarti, R., & Giannotti, F. (2019). Transparency in algorithmic decision making. *ERCIM News*, 116, 10-11.

Sharma, A., & Shafiq, M. O. (2020, December). Predicting purchase probability of retail items using an ensemble learning approach and historical data. In 2020 19th IEEE International *Conference on Machine Learning and Applications (ICMLA)* (pp. 723-728). IEEE.

Tahiri, N., Mazoure, B., & Makarenkov, V. (2019). An intelligent shopping list based on the application of partitioning and machine learning algorithms. In *Proceedings of the 18th Python in Science Conference (SCIPY 2019*).