Customer Responses to FMCG Retail Marketing Campaigns using Demographic and Shopping Behavior Features: An Imbalanced Binary Classification Problem

Introduction and Problem Description

The retail industry is a highly competitive and dynamic field, where companies must constantly strive to improve customer experience in order to retain existing customers and attract new ones. One of the key strategies employed by retailers is the use of marketing campaigns, which are designed to increase sales, promote brand awareness, and encourage customer loyalty.

With the advent of data science and machine learning techniques, retailers now have the ability to gain a deeper understanding of customer behavior and preferences, enabling a more personalized approach to campaign strategy. This study aims to leverage this capability by predicting the effectiveness of marketing campaigns utilizing demographic and historical spending data of customers within a specified campaign period.

The goal is to classify customers into two categories: those who are likely to respond positively to the campaign, and those who are likely to respond negatively. The target response variable is binary, indicating whether a customer responded positively (1) or negatively (0) to the marketing campaign. However, the dataset presents an imbalanced target response variable with a low ratio of positive responses (approximately 0.16%), this requires special handling to address this issue during model training and evaluation. This study aims to develop a model that can accurately predict customer response to marketing campaigns and tackle the imbalanced data problem, in order to improve the campaign's effectiveness and increase revenue for the retail industry.

Data Description

The dataset used in this study is composed of seven tables, namely customer, customer_response, customer_account, product_groups, categories, transaction_header, and transaction_sale. Each table represents a specific aspect of customer behavior and demographic information, as well as details of the marketing campaign. The data was collected between the time range of 2020-12-01 and 2021-12-01.

The customer table contains demographic information on 28,593 customers, including their individual number, gender, city code, and date of birth. The city code variable in the customer table has a 23% missing value rate. Since it is a categorical variable, null values were not imputed, and instead, these observations were removed from the analysis and considered as out of scope.

The customer_response table provides campaign-specific information, including the individual number of the customer, the category number, the minimum amount of spending required to benefit from the campaign, the reward amount, and the customer's response to the campaign. This table contains 13,115 observations.

The customer_account table includes card information on 35,159 customers, including the card number and individual number. The product_groups table provides category breakdowns of the campaign categories, including the category number, and the category levels 1, 2, 3, and 4. This table contains 3,913 observations. The categories table includes general category names and category numbers, with 50 observations, and five distinct observations (Food, Personal Care, Other, Beverage and Hygiene) for the category column.

The transaction_header table contains receipt information for 1,124,673 transactions, including the basket id, card number, date of transaction, and whether the transaction was virtual or in-store. The transaction_sale table includes product level information for each receipt, including the basket id, category levels 1-4, the amount (measured in Turkish Lira), quantity, and discount types. This table contains 6,537,881 observations.

The distribution of the target variable, customer response to the marketing campaign, is imbalanced with a low ratio of positive responses (approximately 0.16%). To handle this imbalance problem, the study employs special handling methods during model training and evaluation.

The evaluation metrics used in this study are K-fold (K=5) cross-validation and F1 score. The K-fold cross-validation method was chosen to ensure that the model is robust and generalizes well to unseen data, while the F1 score was chosen as it is a suitable metric for imbalanced datasets, taking into account both precision and recall.

As the data set is composed of seven tables, the relationships between them and how the data is structured can be better understood by consulting the Entity Relationship Diagram (ERD) provided as **figure 1** in the report. The ERD illustrates the foreign key-primary key relationships among the tables.

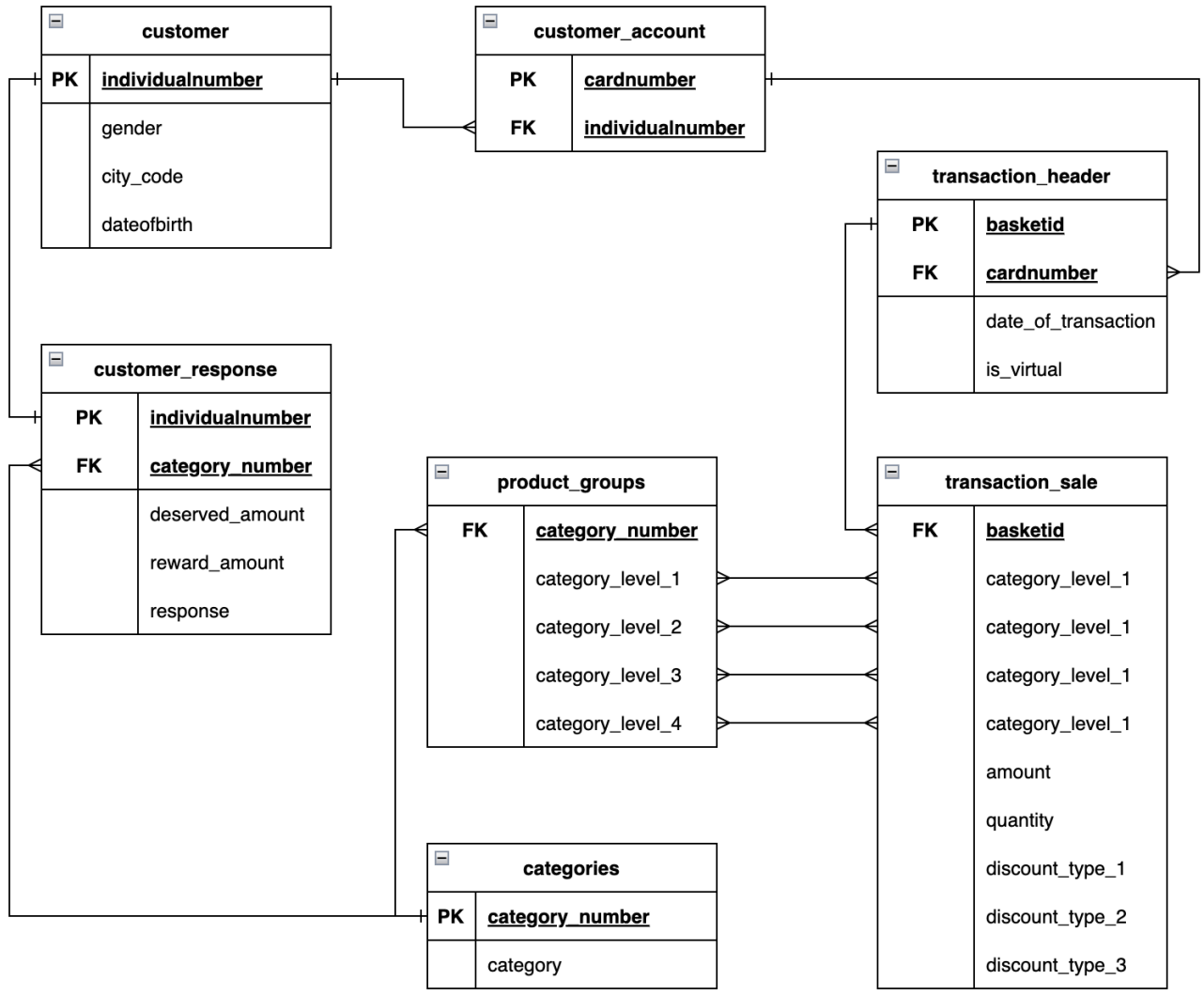

Figure 1: *Entity Relationship Diagram*

Exploratory Data Analysis (EDA)

Demographic analysis of the customer dataset reveals important insights regarding the gender and age distribution of customers and their responses to the marketing campaign. The dataset has a slightly higher representation of male customers (53.14%) compared to female customers (46.85%). A further analysis of customer response to the campaign indicates that the positive response ratio is slightly higher for male customers (0.17%) than female customers (0.14%).

The city code variable in the customer table has a missing value rate of 23%, which indicates that a significant portion of the customers do not have information about their city of residence. As it

is a categorical variable, null values were not imputed and instead, these observations were removed from the analysis and considered as out of scope.

A breakdown of the date of birth variable in the customer table shows that the majority of customers are born in the 1980s and 1990s, with 28% born between 1980-1989 and 23% born between 1990-1999 (***Figure 2: Date of Birth Distribution***) Other age groups include 21% born between 1970-1979, 19% born between 1950-1969, and 9% belong to other dates. There were also 15 null values and 12 observations of customers born after 2020, which were removed from the dataset.

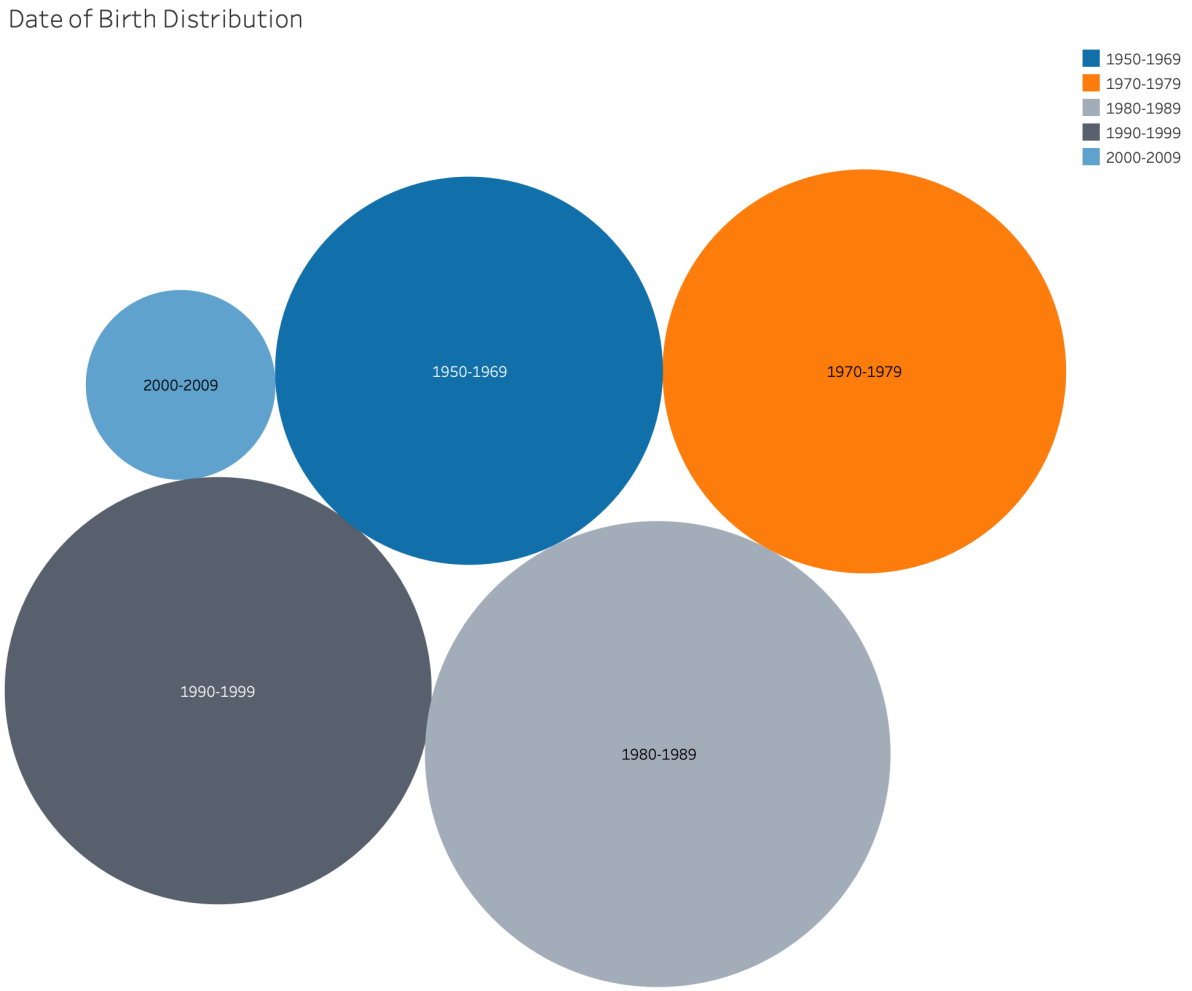

Figure 2: *Date of Birth Distribution*

An analysis of the customer account data shows that a significant portion of customers have multiple cards. 4315 out of 28,578 customers have more than one card and 1055 of them have more than two cards.

An analysis of the transaction data within the 1-year range revealed that a clear pattern of seasonality was not observed when weekly and monthly sales were analyzed (***Figure 3: Shopping Volume by Date***). Despite a thorough examination of the data, no discernible trend was identified that could be attributed to specific days of the week or months. As a result of this lack of seasonality, the sales dates will not be taken into consideration during the model setup. It should be noted that this may be due to the limited time range of the data or other factors not captured in the dataset, and further research or analysis may be required to fully understand the underlying trends in the data. Additionally, it is worth mentioning that the majority of the transactions, approximately 91%, were conducted in physical stores rather than online. This may indicate that the majority of customers prefer in-store shopping experiences.

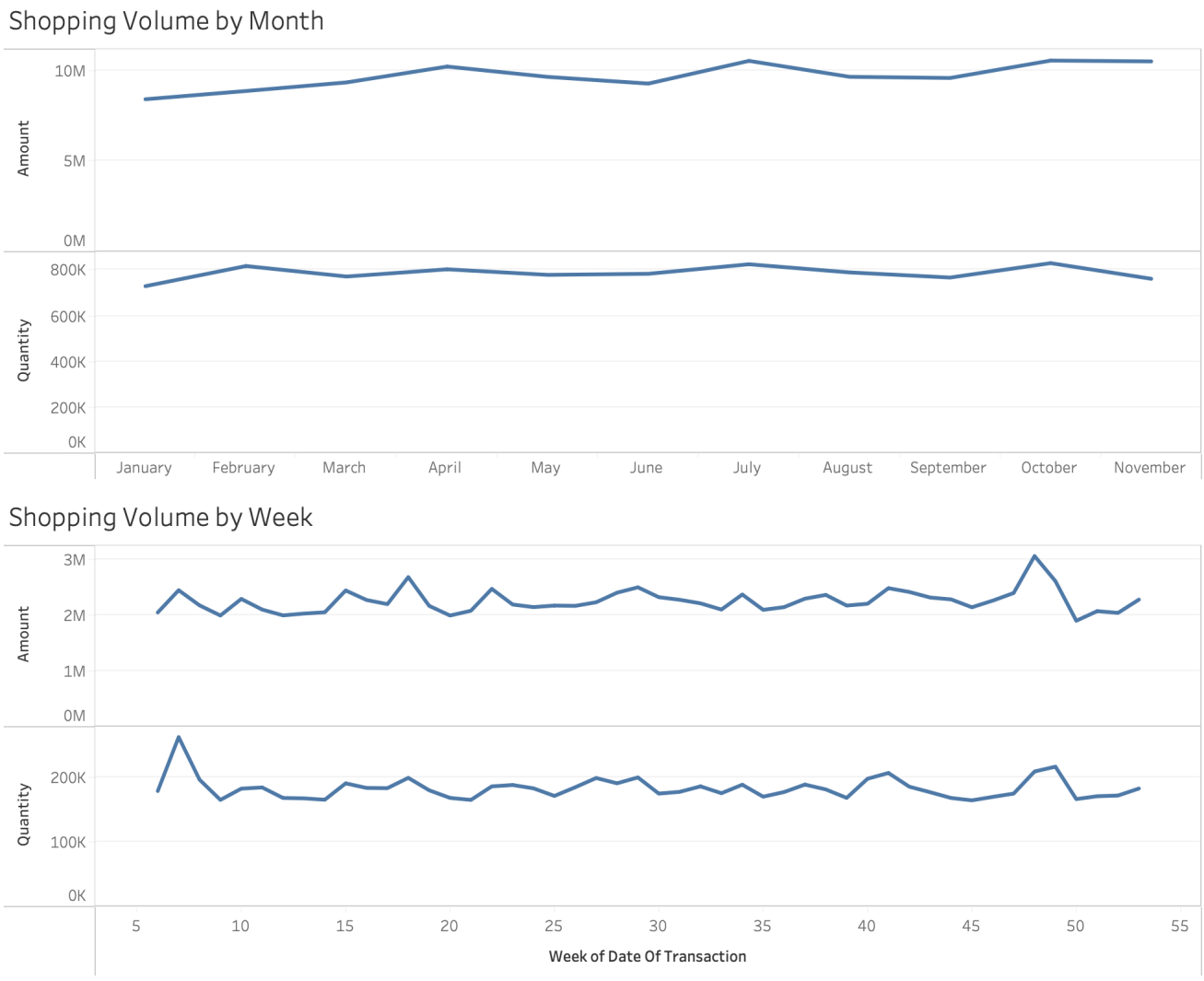

Figure 3: *Shopping Volume by Date*

An analysis of the transaction data by category reveals that the "Food" category has the highest volume of purchases, both in terms of quantity and amount. Specifically, the "Food" category accounts for 65.54% of purchases by quantity and 64.14% of purchases by amount. The

"Beverage" category is the second highest by quantity, representing 18.91% of purchases by quantity and 10.91% of purchases by amount. These findings provide insight into customer purchasing habits and preferences and can inform product and marketing strategies.

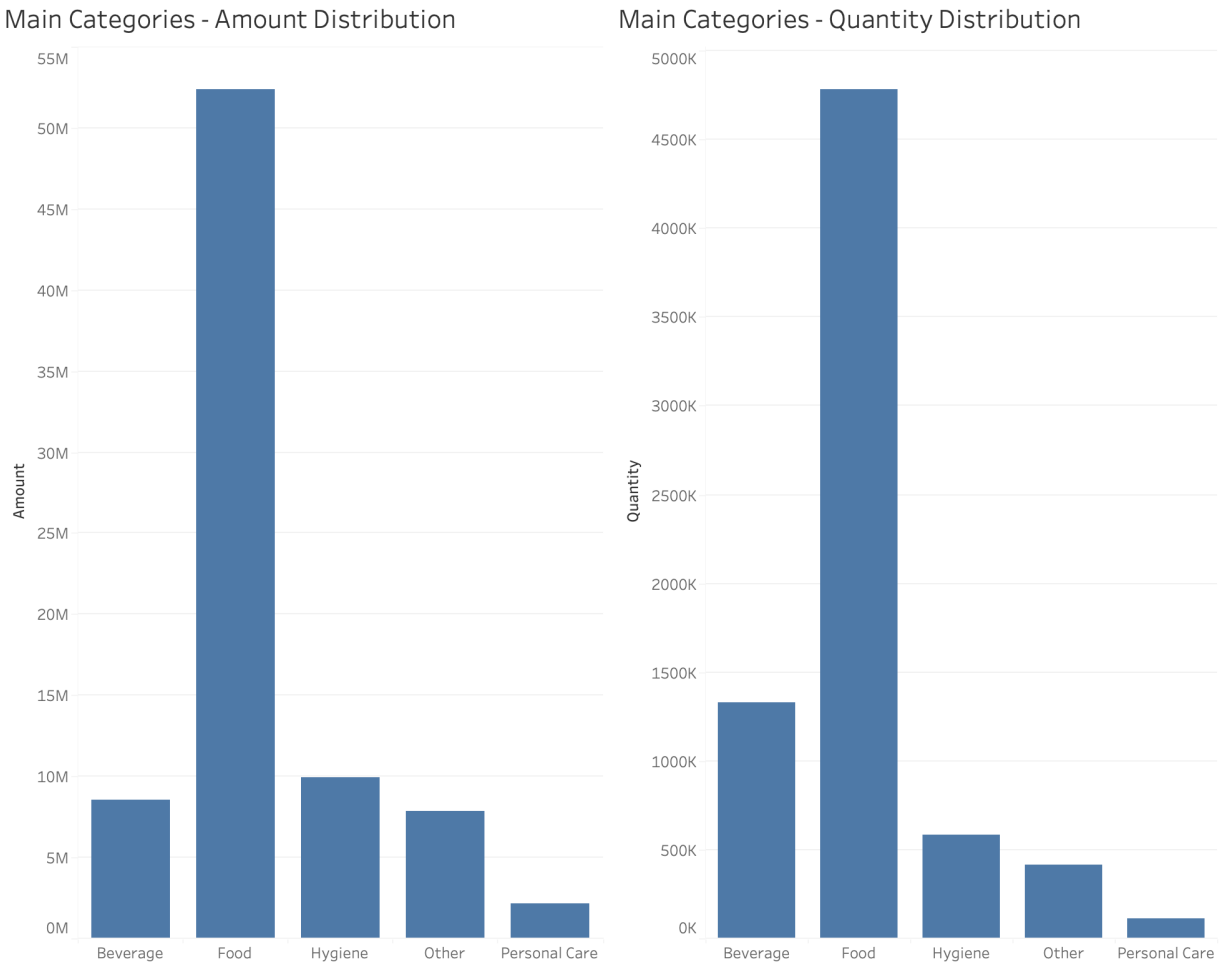

Figure 4: *Main Category Distribution*

A correlation analysis was performed to examine the relationship between the amount of spending required to benefit from a campaign (deserved_amount) and the amount of reward points earned (reward_amount) with the positive response status of customers. The results of this analysis are presented in **Figure 5** as a correlation graph. This analysis provides insight into the relationship between the campaign's requirements and rewards and the likelihood of customers responding positively to the campaign. This information can be used to inform the design and implementation of future marketing campaigns and to optimize the use of incentives to drive customer engagement.

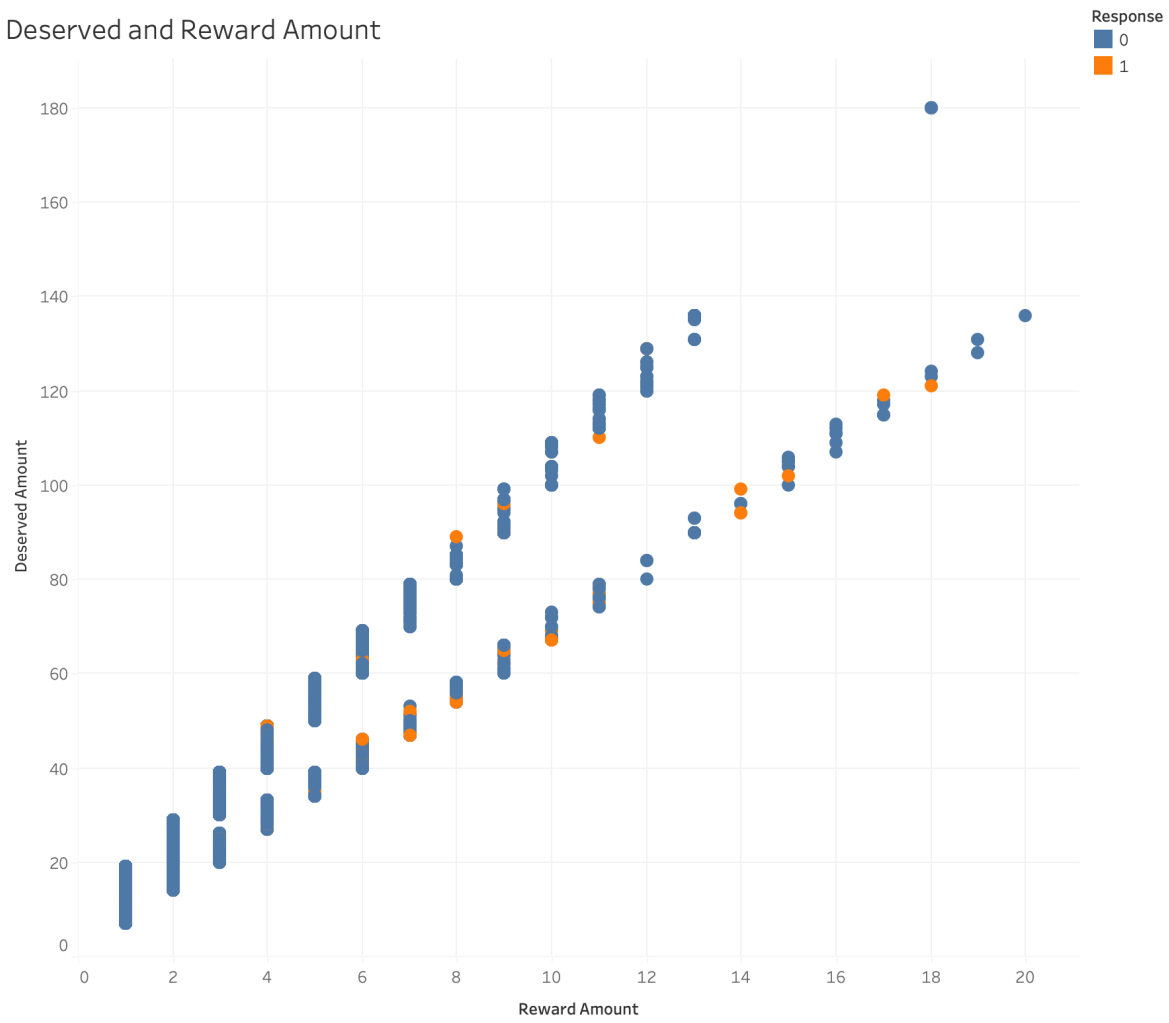

Figure 5: *Deserved & Reward Amount*

A detailed analysis of the discounts applied by the company to the products purchased was conducted, and the results are presented in **Figure 6**. The figure illustrates the number of occurrences of three different discount types applied to products, as well as the correlation of these discount types with each other and with the positive response status(Orange Fill). This analysis provides insight into the company's discount strategies and the effectiveness of these strategies in terms of customer response. It can be used to inform future pricing and promotion decisions

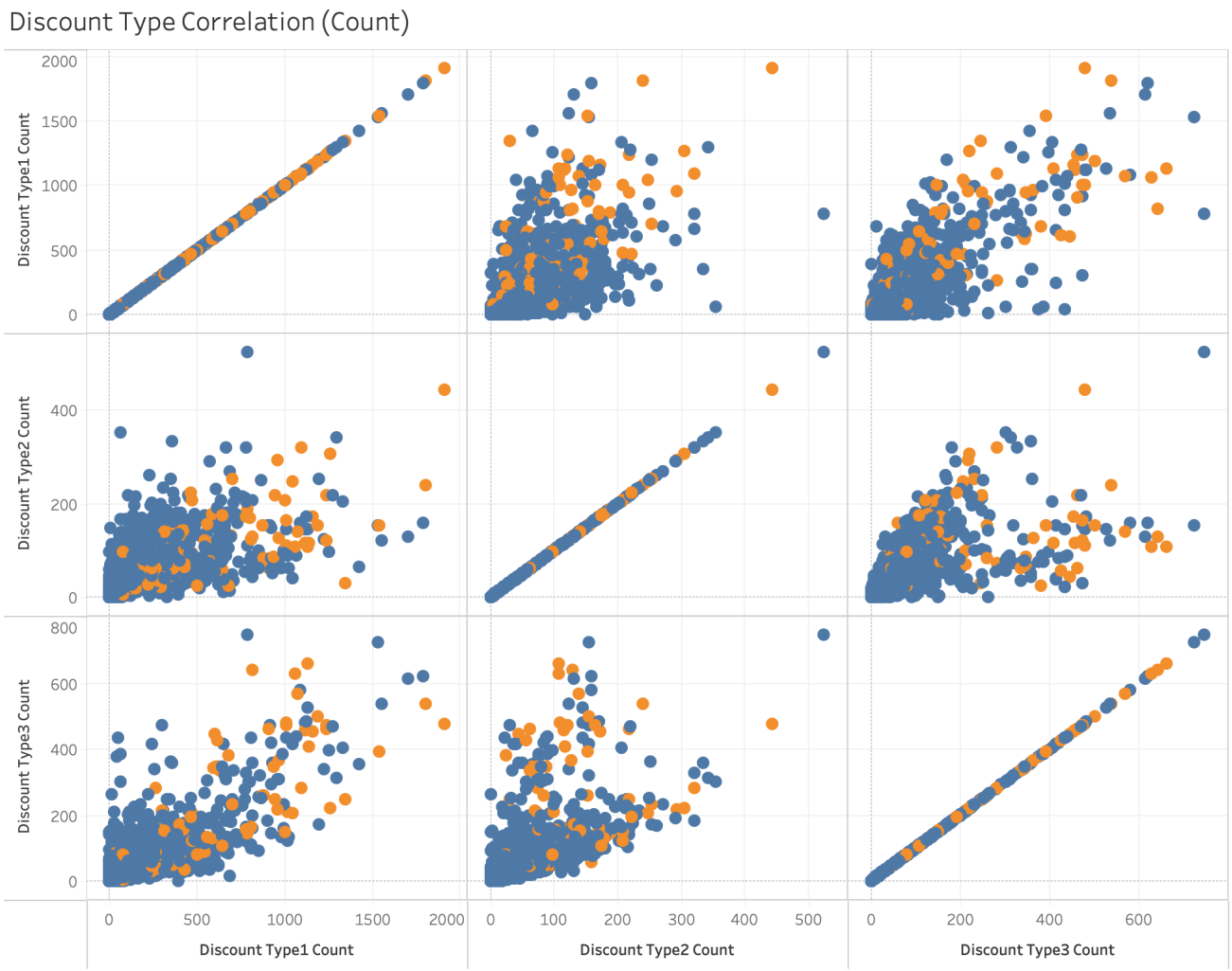

Figure 6: *Discount Type Correlation (Count)*

A further analysis was conducted to understand the relationship between customer spending on the main categories and their response to the marketing campaign. The results of this analysis are presented in **Figure 7**, which displays the relationship between customer spending on the main categories and their likelihood of responding positively to the marketing campaign. This analysis provides insight into the impact of customer spending patterns on their likelihood of responding positively to a marketing campaign. This information can be used to inform targeted marketing strategies and to understand which customer segments may be most responsive to specific types of campaigns.

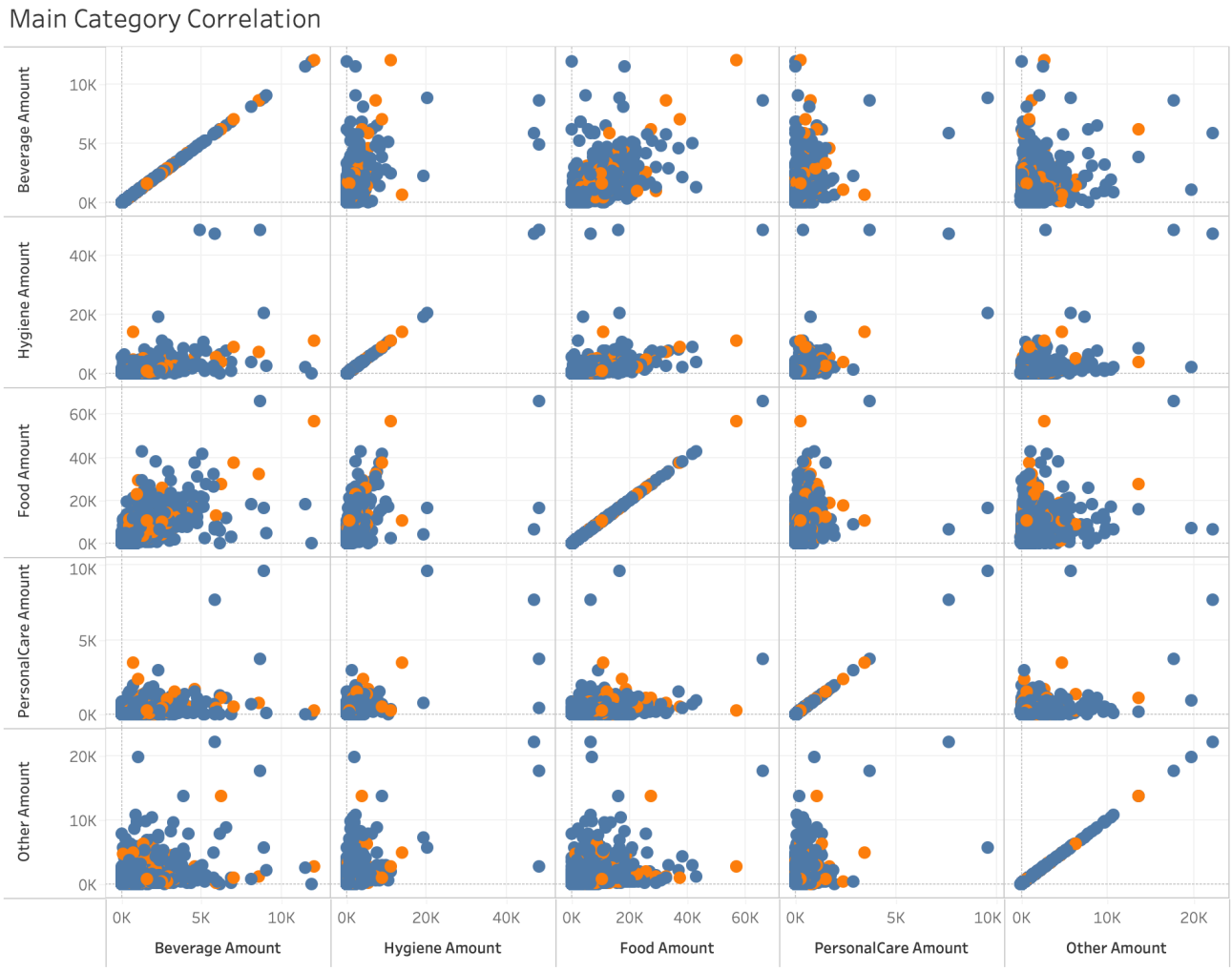

Figure 7: *Main Category Correlation*

Methodology:

- ### Model Selection:

The methodology employed in this study involves the use of three machine learning algorithms for binary classification, including logistic regression, gradient boosting classifier, and XGBoost Classifier. Hyperparameter optimization is performed using the sensitivity analysis technique and the GridSearchCV package. Model evaluation is conducted using the F1 score metric and each model's output is evaluated using StratifiedKFold cross-validation technique, since it is an imbalanced binary classification problem.

- ## Feature Selection:

In order to select the most relevant features for the model, a comprehensive analysis is conducted to identify the variables that have the most significant impact on the model's performance. The following features are selected:

1. The one-year sales performance of customers, which is taken as 10 different variables based on the total number and total amount in 5 main categories

2. The deserved_amount values determined for the customer's reward_amount and customer/category_number

3. The discount_type variables applied for the customers, which are taken as 2 different variables, the number of discount_type and the average of discount_type. The reason why the total discount_type value applied is not taken, instead it is taken as two different variables as average and count, is to take into account the frequency of discount_type applied for the customer within a one-year period.

4. Demographic information of customers, such as date of birth and gender, are excluded from the scope because the average F-score value is lowered in the tests.

- ## Outliers and Normalization:

To ensure that the model is not affected by outliers and to preserve the distribution of the variables, Z-score normalization is applied to a total of 12 variables, including the amount and number information of 5 main categories, plus the reward_amount and deserved_amount variables. Since the problem is a binary classification and decision tree algorithms are not sensitive to outliers, outliers are not excluded from the dataset.

It is worth noting that this normalization technique standardizes the data by subtracting the mean and dividing by the standard deviation, thus transforming the data to have a mean of 0 and a standard deviation of 1. This method of normalization allows the model to make fair comparisons between the different variables and to handle any variability in the data.

In summary, the methodology employed in this study includes the use of three machine learning algorithms for binary classification, feature selection based on the impact on model performance, and the use of Z-score normalization to handle outliers and preserve the distribution of variables. Hyperparameter optimization is performed using the sensitivity analysis technique and the GridSearchCV package. Model evaluation is conducted using the F1 score metric and each model's output is evaluated using StratifiedKFold cross-validation technique, to handle the imbalanced nature of the dataset.

Results:

The results of the model evaluation are summarized in Table 1, showing the mean F1 score for each of the three machine learning algorithms used in this study. It is observed that the XGBoost Classifier has an F1 score of 0.4494, the Logistic Regression has an F1 score of 0.1631, and the Gradient Boosting Classifier has an F1 score of 0.6411.

ML Algorithm	**F1 Score**
XGBoost Classifier	0.4494
Logistic Regression	0.1631
Gradient Boosting Classifier	0.6411

Table 1: *Model Evaluation Summary*

The highest F1 score of 0.6411 is achieved by the Gradient Boosting Classifier, indicating that it has the best performance among the three models in terms of precision and recall. The F1 score is a measure of the trade-off between precision and recall, where a higher score indicates a better balance between these two metrics. Therefore, the Gradient Boosting Classifier can be considered as the best model for this problem.

Confusion matrix, ROC Curve and Feature Importance graph of the gradient boosting algorithm can be examined from Table 2, Figure 8 and Figure 9, respectively.

	True	**False**
Positive	127	2
Negative	12,774	82

Table 2: Confusion Matrix

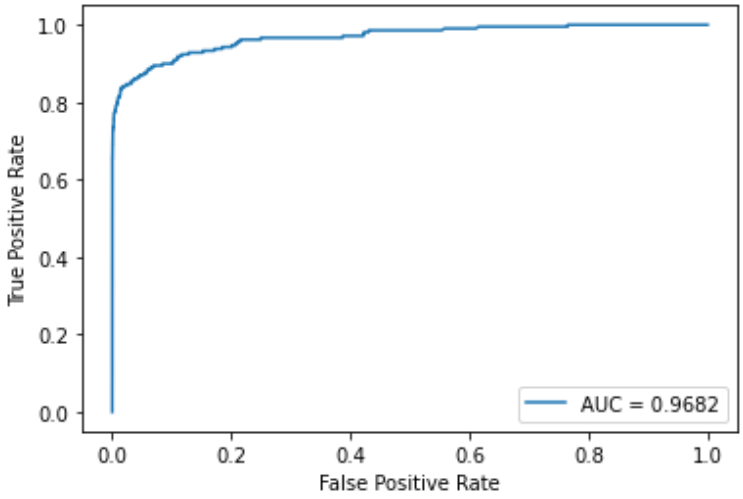

Figure 8: *ROC Curve*

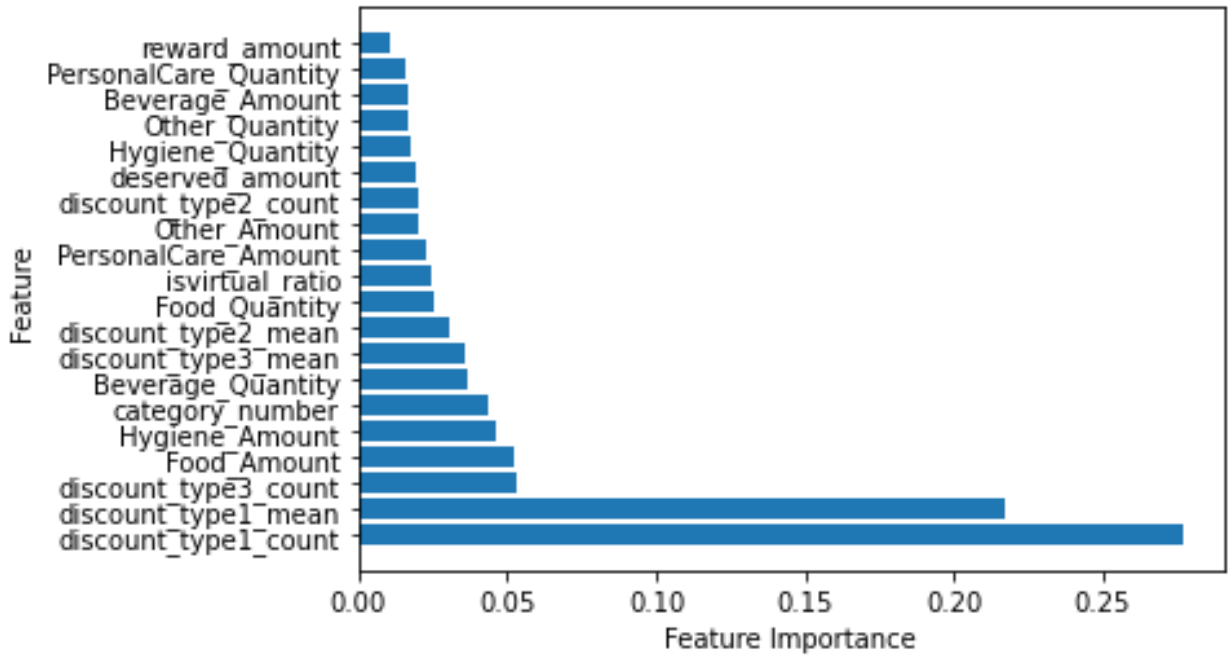

Figure 9: *Feature Importance*

Conclusion

In conclusion, the goal of this study was to predict customer response to a marketing campaign in the Fast-Moving Consumer Goods (FMCG) retail sector using demographic and shopping behavior features. The dataset was found to be imbalanced with a low ratio of positive responses, requiring special handling during the training and evaluation of the model. Through the use of logistic regression, gradient boosting classifier, and XGBoost classifier algorithms, hyperparameter optimization, and feature selection techniques, a model was developed with a mean F1 score of 0.6411 for the gradient boosting classifier. The performance of the model was further evaluated through the use of cross-validation and the calculation of important metrics such as the AUC value and the confusion matrix. The feature importance plot also provided insight into which variables had the greatest impact on the model's predictions. Overall, the study successfully demonstrated the potential for utilizing demographic and shopping behavior data to predict customer response to marketing campaigns in the FMCG retail sector.