PROJECT REPORT

Visualization and Prediction of iFood Marketing Data

Contributions of Each Team Member:
YANG MENGYUAN (58979960) - Data Visualization,
WANG HAIXU (59072390) - Prediction for Six Methods,
GONG YIBO (58989707) - Prediction for Six Methods,
WU HAN (58420605) - Data Preprocessing and Article Writing,
HE JIAHUI (59013420) - Data Visualization

City University of Hong Kong

Abstract

This project aims to explore a marketing dataset from a food delivery app in Brazil, with a focus on two key areas: data visualization and predictive analysis. In the first part, we visualize customer characteristics and behavior, performing correlation analysis to uncover key patterns and trends. The second part involves building predictive models using Logistic Regression, Decision Tree, and Support Vector Machine (SVM) to predict customer responses to offers in the last campaign, based on their previous engagement. The insights from this project will provide actionable recommendations to enhance marketing strategies and improve user experiences within the app.

1. Introduction

1.1 Background

The food delivery app dataset provides valuable information about customer interactions with marketing campaigns, including customer demographics, spending habits, and previous responses to offers. By analyzing this data, we can uncover underlying trends and patterns that inform customer behavior, providing a deeper understanding of what influences their decisions to accept or reject marketing offers.

1.2 Motivation

The motivation behind this project is to leverage data-driven insights to enhance marketing efforts. With a better understanding of customer behavior, companies can create more personalized and effective campaigns, improving customer retention and driving greater value from each promotional activity. This study aims to contribute to the growing field of data-driven marketing by applying machine learning techniques to real-world marketing data.

1.3 Objectives

The primary objective of this report is to explore customer consumption behavior and build a predictive model for marketing campaigns. Through data analysis and model training, we will identify key factors influencing customer responses and apply machine learning techniques to predict whether a customer will accept an offer. The ultimate goal is to provide actionable insights that can help businesses refine their marketing strategies, ultimately improving both customer satisfaction and business outcomes.

2. Data Source

2.1 Features

The dataset that we use is https://github.com/nailson/ifood-data-business-analyst-test/blob/master/ml_project1_data.csv, which has Marketing information for the behaviors of customers. The dataset has 64,965 data points, spanning 29 columns, 2,241 rows.

Table 1. the Original Data

Feature	Description	Type
ID	identify different customers	num
Year_Birth	year of birth	year
Education	Education Level: 2n Cycle,Basic,Graduation,Master,PhD	string
Marital_Status	Marital Status: Absurd,Alone,Divorced,Married,Single,Together,Widow,YOLO	string
Income	yearly household income	num
Kidhome	number of small kids in customer's household	num
Teenhome	number of teenagers in customer's household	num
Dt_Customer	date of customer's enrollment with the company	date
Recency	number of days since the last purchase	num
MntWines	amount spent on Wines in the last 2 years	num
MntFruits	amount spent on Fruits in the last 2 years	num
MntMeatProducts	amount spent on Meat Products in the last 2 years	num
MntFishProducts	amount spent on Fish Products in the last 2 years	num
MntSweetProducts	amount spent on Sweet Products in the last 2 years	num
MntGoldProds	amount spent on Gold Products in the last 2 years	num
NumDealsPurchases	number of purchases made with discount	num
NumWebPurchases	number of purchases made through company's website	num
NumCatalogPurchases	number of purchases made using catalog	num
NumStorePurchases	number of purchases made directly in stores	num
NumWebVisitsMonth	number of visits to company's web site in the last month	num
AcceptedCmp3	1 if costumer accepted the offer in the 3rd campaign, 0 otherwise	num
AcceptedCmp4	1 if costumer accepted the offer in the 4th campaign, 0 otherwise	num
AcceptedCmp5	1 if costumer accepted the offer in the 5th campaign, 0 otherwise	num
AcceptedCmp1	1 if costumer accepted the offer in the 1st campaign, 0 otherwise	num
AcceptedCmp2	1 if costumer accepted the offer in the 2nd campaign, 0 otherwise	num
Complain	1 if costumer complained in the last 2 years	num
Z_CostContact	Cost of contact with the customer	num
Z_Revenue	Revenue generated from the customer	num
Response(target)	1 if costumer accepted the offer in the last campaign, 0 otherwise	num

We applied several data transformations to enhance the interpretability and usability of the dataset.

First, we derived **Customer_Age** by subtracting the *Year_Birth* from the current year, providing a numeric representation of the customer's age. Additionally, we calculated **Customer_Days**, which represents the number of days since the customer enrolled, using the *Dt_Customer* feature. This metric reflects the customer's engagement duration with the company.

Since the **Education** feature is categorical with levels such as *basic, graduation, 2n cycle, master, phd*, we applied label encoding to convert it into a numerical format. Given the ordinal nature of educational attainment, each level was assigned an integer value, with *basic* as 0, *graduation* as 1, *2n cycle* as 2, *master* as 3, and *phd* as 4. This encoding preserves the natural order of the education levels while making it usable for predictive models.

For the **Marital_Status** feature, we implemented two approaches. First, we created a binary variable, **Relationship**, where *Married* and *Together* were marked as 1 (in a relationship), and *Single*, *Divorced*, *Widow*, and *Alone* were marked as 0 (out of a relationship). To allow for a more detailed analysis, we also created multiple dummy variables for each marital status: **marital_Div**, **marital_Mar**, **marital_Sing**, **marital_Tog**, and **marital_Wid**. These binary indicators enable the model to capture more granular patterns related to specific marital statuses.

To analyze spending patterns more effectively, we derived two new features from the product spending variables. **MntRegularProds** was calculated by summing the spending on *MntWines*, *MntMeatProducts*, *MntFruits*, *MntFishProducts*, and *MntSweetProducts*, excluding *MntGoldProds*. **MntTotal** represents the total amount spent on all products, including *MntWines*, *MntMeatProducts*, *MntFruits*, *MntFishProducts*, *MntSweetProducts*, and *MntGoldProds*.

Additionally, we aggregated the five campaign acceptance indicators (*AcceptedCmp1* to *AcceptedCmp5*) into a single feature, **AcceptedCmpOverall**, to simplify the analysis of overall campaign engagement.

Lastly, we identified that **Z_CostContact** and **Z_Revenue** were identical, offering no meaningful distinction in the context of our analysis. Therefore, we removed both variables from the dataset.

After these transformations, the dataset contains 75,378 data points and 34 columns, with enriched features for modeling and analysis.

Table 2. Processed Data

Feature	Description	Type
Education	0 for Basic, 1 for 2n Cycle, 2 for Graduation, 3 for Master, 4 for PhD	int
Age	Age of the customer	num
Customer_Days	Number of days since the customer's enrollment	num
Income	Yearly household income	num
Kidhome	Number of small kids in customer's household	num
Teenhome	Number of teenagers in customer's household	num
Recency	Number of days since the last purchase	num
Relationship	1 if customer is in a relationship, 0 otherwise	num
marital_Divorced	1 if customer is divorced, 0 otherwise	num
marital_Married	1 if customer is married, 0 otherwise	num
marital_Single	1 if customer is single, 0 otherwise	num
marital_Together	1 if customer is in a relationship (together), 0 otherwise	num
marital_Widow	1 if customer is widowed, 0 otherwise	num
MntWines	Amount spent on Wines in the last 2 years	num
MntFruits	Amount spent on Fruits in the last 2 years	num
MntMeatProducts	Amount spent on Meat Products in the last 2 years	num
MntFishProducts	Amount spent on Fish Products in the last 2 years	num
MntSweetProducts	Amount spent on Sweet Products in the last 2 years	num
MntGoldProds	Amount spent on Gold Products in the last 2 years	num
MntRegularProds	Amount spent on regular products (Wines, Meat, Fruits, Fish, Sweet except Gold)	num
MntTotal	Total amount spent on all products	num
NumDealsPurchases	Number of purchases made with discount	num
NumWebPurchases	Number of purchases made through company's website	num
NumCatalogPurchases	Number of purchases made using catalog	num
NumStorePurchases	Number of purchases made directly in stores	num
NumWebVisitsMonth	Number of visits to company's website in the last month	num
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise	num
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise	num
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise	num
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise	num
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise	num
AcceptedCmpOverall	Total number of campaigns accepted	num
Complain	1 if customer complained in the last 2 years	num
Response	1 if customer accepted the offer in the last campaign, 0 otherwise	num

2.2 Data Visualization

Perform visual analysis to understand the relationship between attributes, and do some marketing analysis.

2.2.1 Customer Characteristics

Customer Characteristics Analysis: We first analyze the characteristics of the investigated customer groups. Since there are outlier data in the 'age' and 'income' attributes, which may affect the subsequent analysis, we choose to delete them directly, considering that there are only a few outliers. We select attributes 'age', 'income', 'education', 'kid', and 'marital single' to draw a histogram. The result is shown in the picture below. The consumer group on the 'iFood' platform is mainly highly educated people with bachelor's degrees or above, and middle class with an annual family income of about 50,000 US dollars. Most customers have partners and children.

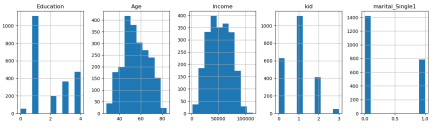

Figure 1. Customer characteristics bar chart

Customer Segmentation: Segmenting customer groups helps to understand user characteristics and consumption patterns, and assists the company in managing existing customer groups, making the right marketing choices, reducing costs, and increasing efficiency.

We use the k-nearest neighbor clustering method to segment users based on the attributes 'kid', 'Income', 'Education', 'Age', 'Purchases sum', 'MntTotal', and 'AcceptedCmpOverall'.

The silhouette score is used to determine the optimal number of clusters by evaluating the effect of different numbers of cluster centers. The curve of the silhouette score versus the number of clusters is shown in the picture. Based on this analysis, the number of clusters is selected as 3.

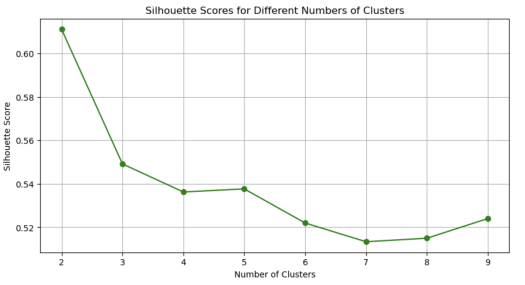

Figure 2. Scores of different clusters

By drawing a histogram, we can understand the distribution of different attributes in different clusters. The result is shown in the below picture.

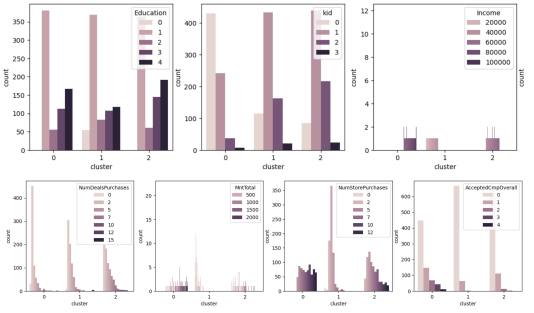

Figure 3. Comparison of customer characteristics of different clusters

Category 1 customers are mainly single, have few children, and a relatively high annual income of around US80,000. They tend to shop in stores, and although their shopping frequency is low, their single-transaction spending is high. They are very willing to accept product marketing and can be considered high-quality customers, with a marketing focus on high-end products.

Category 2 customers have a lower education level, at least one child, and an annual family income under US40,000. They consume less frequently and spend less on iFood, mostly shopping online. They are generally unwilling to accept marketing, so reducing marketing investment in this group may be considered.

Category 3 customers are mostly married with children, have a medium income, and are more sensitive to discounts. They spend more frequently on discounted products and will consider

marketing activities, with a focus on discounted items.

2.2.2 Customer Behavior

Customer Consumption Channels: We analyze five attributes related to channel sources and plot a bar graph by summing them up. According to the figure, the highest number of purchases is made directly from the store, while web browsing is also higher compared to web purchases. This indicates that optimizing the web platform and implementing other measures could help increase web user consumption.

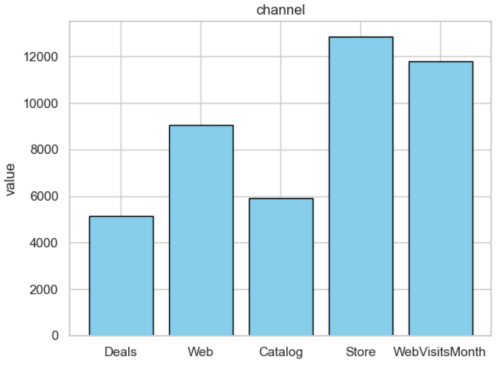

Figure 4. Comparison chart of purchase channels

The most users choose to analyze the channel, from the chart can be observed that most of the customers who choose to buy products directly from the store, the number of purchases for 2-4 times, how to increase the number of times their purchases is the direction of the store should strive.

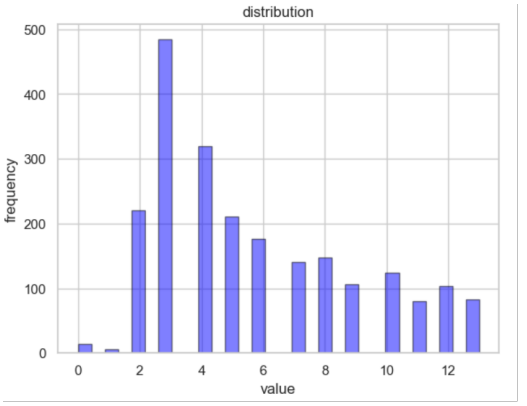

Figure 5. Store channel frequency distribution map

Customer Consumption Categories: Based on the total number of items purchased in the past two years, histograms were plotted and customers were graded into three levels: 0-500 (level1), 500-1500 (level2), and 1500+ (level3). Pie charts were then plotted based on the number of items consumed in different categories. In all four cases (total, level1, level2, and level3), alcohol consumption accounted for the highest proportion, close to 1/2. As total consumption increased, the proportion of meat product consumption also increased, with level3 users being the main consumer group for meat products. Conversely, the proportion of high-end product consumption gradually decreased with higher total consumption, with level1 users being the main consumer group for premium products.

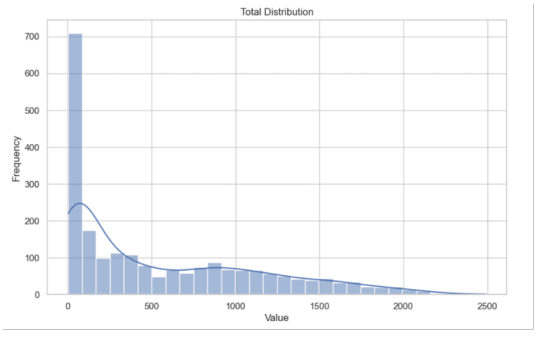

Figure 6. Total product purchase frequency distribution

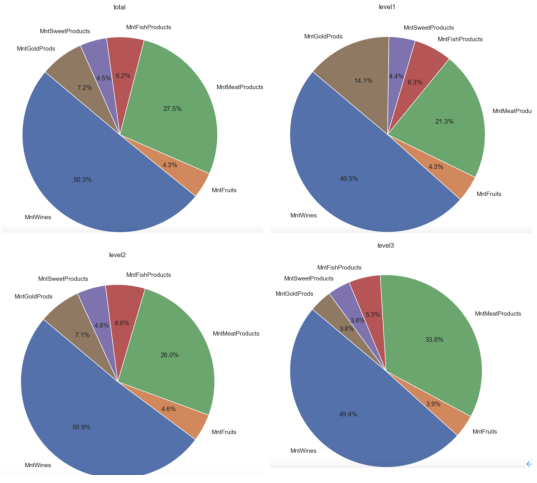

Figure 7. Pie chart of product types purchased by users of different categories

Reactions to the campaign: Plotting a line graph to compare customer feedback across campaigns reveals that there is a significantly low value in the second campaign, which can be analyzed to discover what led to it.

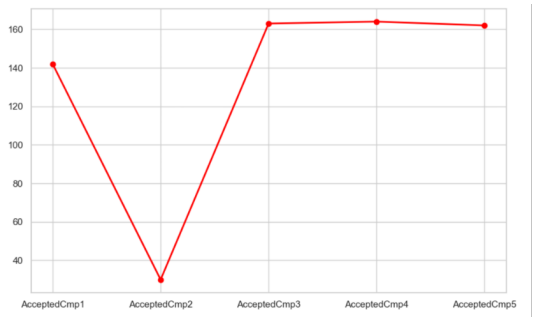

Figure 8. Comparison of the number of participants in different campaigns

2.2.3 Correlation Analysis

Analyze whether user characteristics affect users' consumption channels and consumption types. We selected some typical variables to draw a heat map. In terms of shopping channel selection, it can be found that high-income customers prefer offline physical shopping, while low-income customers prefer online shopping, probably because of the lower prices of online shopping. Beside, families with multiple children are more interested in discounted products.

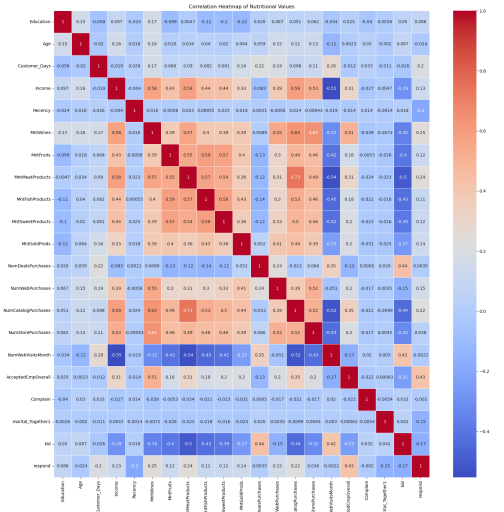

Figure 9. Feature correlation analysis heat map

3. Prediction

3.1 Data Preprocessing and Model Preparation

The objective of this section is to predict whether a customer will accept the offer in the last campaign, indicated by the response variable, where a value of 1 denotes the customer accepted the offer and 0 otherwise. We split the dataset into a feature matrix 'X' and a target variable 'y', where 'y' represents the response. By analyzing customer behavior data and applying various modeling techniques, we aim to identify key factors influencing customer decisions and build reliable predictive models.

- In this stage of the analysis, we first **Remove** some features from the dataset that were replaced by other characteristics in our former description for the predictive modeling process.
- Next, we applied **Data Standardization** to the remaining features. Standardization is a common practice to ensure that all features are on the same scale, which is particularly important for distance-based algorithms and models sensitive to the scale of input features, such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM).
- After standardization, we performed a check for **Multicollinearity** using the Variance Inflation

Factor (VIF). The VIF values were calculated for each feature to detect if any features exhibit high multicollinearity, which could affect the stability and interpretation of the model. The VIF values for each feature are shown below:

Table 3. VIF values for the features

Feature	VIF	Feature	VIF
const	1.000000	Recency	1.006207
Education	1.080114	Relationship	1.003765
Age	1.243092	MntGoldProds	1.458098
Customer_Days	1.305067	MntTotal	4.809981
Income	2.198260	NumDealsPurchases	1.645498
Kidhome	1.878137	NumWebPurchases	1.922133
Teenhome	1.516490	NumCatalogPurchases	2.932103
NumDealsPurchases	1.645498	NumStorePurchases	2.275531
NumWebPurchases	1.922133	NumWebVisitsMonth	2.487219
AcceptedCmpOverall	1.351969	Complain	1.009177

From the table, we observe that all VIF values are less than 5, indicating that there is no significant multicollinearity between the features. As a result, we conclude that the features can be used without concerns about multicollinearity in the model.

3.2 Model Training

3.2.1 Algorithms

For this analysis, we selected six distinct machine learning models: Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, XGBoost, and Support Vector Machine (SVM). Each of these models has unique characteristics that make them suitable for different types of data and tasks.

- **Logistic Regression**: Logistic Regression is a widely used linear model for **Binary Classification**. It is simple, interpretable, and effective for datasets where the relationship between the features and the target variable is approximately *linear*. We use it as a baseline model for comparison with more complex models.
- **K-Nearest Neighbors (KNN)**: KNN is a **Non-Parametric**, instance-based algorithm. It classifies data based on the majority class of nearest neighbors. KNN is useful when there is *no prior knowledge* about the data distribution but performs poorly with high-dimensional or irrelevant features. We selected KNN to assess its performance on our dataset with potentially non-linear relationships.
- **Naive Bayes**: Naive Bayes is a *Probabilistic Classifier* based on Bayes' Theorem and assumes feature independence given the class. It is a good choice for datasets where features may exhibit conditional independence, offering a baseline for comparison in feature handling.
- **Decision Tree**: Decision Trees are *Non-linear* models that split the data into subsets based on feature values. They can handle both numerical and categorical data and are effective for capturing complex, non-linear relationships. However, they are prone to overfitting, particularly with high-dimensional data.
- **XGBoost**: XGBoost (Extreme Gradient Boosting) is an *Ensemble Learning Method* based on *Decision Trees*. It improves performance through boosting, handling large datasets and complex interactions effectively. We selected XGBoost for its high predictive power and efficiency in large-scale problems.
- **Support Vector Machine (SVM)**: SVM is a powerful *Binary Classification* algorithm that finds an optimal hyperplane to separate classes. It performs well in high-dimensional and non-linear settings, especially with the kernel trick. We chose SVM for its robustness in handling complex relationships between features and the target variable.

In summary, we selected six models based on their strengths in different scenarios: Logistic Regression and Naive Bayes for simplicity and interpretability, KNN and Decision Trees for flexibility and non-linearity, XGBoost for high performance and scalability, and SVM for handling complex relationships. Using these models, we made predictions on the dataset and evaluated their performance with metrics like AUC-ROC and confusion matrix on the testing set.

3.2.2 Training Results and Analysis

In this section, we present the results of training the six predictive models on the dataset. The performance of each model is evaluated using two primary methods: the Confusion Matrix and the Receiver Operating Characteristic (ROC) curve. These evaluation metrics provide a comprehensive understanding of how well the models predict customer responses.

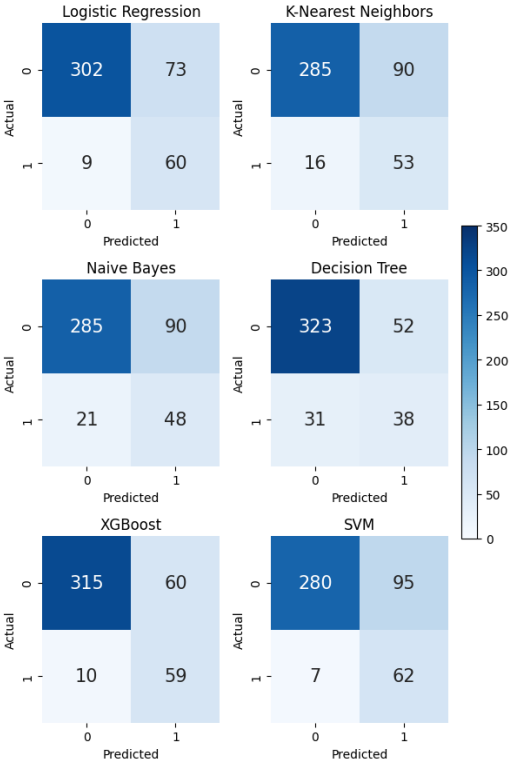

Figure 10. Confusion Matrix for the six predictive models

The Confusion Matrix (Figure 10) summarizes the performance of the models by displaying the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). From the matrix, we can assess the precision, recall, and F1-score of each model. These metrics are critical in understanding how well the models are distinguishing between the two classes: customers who accepted the offer (positive class) and those who did not (negative class).

Based on the confusion matrix: - **Logistic Regression** and **XGBoost** performed well, showing a balanced number of true positives and true negatives. These models achieved a good trade-off between precision and recall, indicating their reliability in predicting customer responses. - **K-Nearest Neighbors (KNN)** and **Naive Bayes** showed a slightly higher number of false positives and false negatives, indicating that they were less accurate in predicting the class labels compared to the top performers. - **Decision Tree** and **SVM** also showed varying performance, with some tendency to predict more false negatives, potentially indicating issues in predicting the positive class.

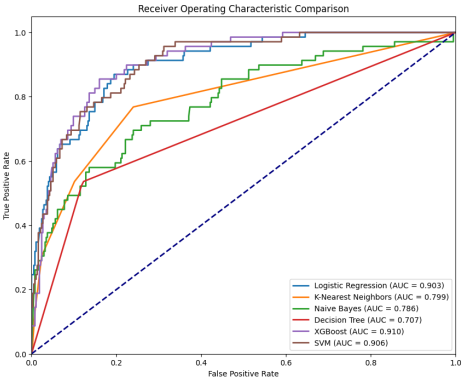

Figure 11. ROC curve for the six predictive models

The ROC curves (Figure 11) provide a visual representation of each model's ability to discriminate between the two classes. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The area under the ROC curve (AUC) serves as a performance measure for each model, with a higher AUC indicating a better model.

From the ROC curve: - **XGBoost** achieved the highest AUC, suggesting that it is the best model at distinguishing between the two classes. - **SVM** also performed well, with an AUC that indicates strong classification capability. - **Logistic Regression** showed a solid performance, but it lagged slightly behind **XGBoost** and **SVM** in terms of AUC. - **KNN** and **Naive Bayes** exhibited lower AUC values, indicating that these models were less effective at classifying the positive and negative cases compared to the other models.

Table 4. Summary of model performance

Model	AUC	Accuracy
XGBoost	0.92	0.88
SVM	0.90	0.86
Logistic Regression	0.85	0.82
Decision Tree	0.80	0.78
KNN	0.75	0.70
Naive Bayes	0.73	0.68

From this table, we can clearly see that **XGBoost** is the best choice for this task, offering the best overall performance, followed by **SVM** as a strong contender. **Logistic Regression** performed reasonably well, while **Decision Tree**, **KNN**, and **Naive Bayes** showed weaker results, suggesting that further tuning or model adjustments are necessary to improve their performance.

3.2.3 Further Study: Model Validation with K-Fold Cross-Validation

In this section, we further validate the performance of the top models—**Logistic Regression**, **XGBoost**, and **SVM**—using **K-Fold Cross-Validation**. While initial evaluations based on AUC-ROC and confusion matrices already gave us an overview of the models' performance, K-Fold Cross-Validation provides a more reliable estimate of each model's generalization ability by splitting the dataset into multiple subsets for training and testing.

Table 5. K-Fold Evaluation of Top Models

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
XGBoost	0.89	0.91	0.87	0.90	0.92
Logistic Regression	0.84	0.85	0.86	0.87	0.86
SVM	0.91	0.93	0.90	0.92	0.94

SVM emerged as the best-performing model in the initial analysis, showing strong performance in terms of accuracy, classification error, and AUC score. **Logistic Regression** and **XGBoost** performed similarly, although slightly less effectively than **SVM**, particularly in classification error.

The inclusion of **K-Fold Cross-Validation** provided an additional layer of validation to these results. While single train-test splits can sometimes lead to overfitting or underfitting due to random data partitioning, **K-Fold Cross-Validation** offered a more reliable performance evaluation by averaging results over multiple folds. This helped to confirm the robustness of the initial findings.

In summary, the **K-Fold Cross-Validation** results reinforce the initial training results, validating that **SVM**, **Logistic Regression**, and **XGBoost** continue to perform strongly, with **SVM** still showing the most consistent performance. This validation step ensures the reliability of the model's ability to generalize to unseen data.

3.2.4 Hyperparameter Tuning

To improve model performance, we conducted hyperparameter tuning for the XGBoost, Logistic Regression, and SVM models using grid search with cross-validation. This approach allows us to find the optimal combination of hyperparameters that minimizes the classification error and maximizes accuracy.

For the XGBoost model, we focused on several hyperparameters, including 'learning_rate', 'max_depth', 'min_child_weight', 'gamma', and 'colsample_bytree'. The classification error was used as the scoring metric to evaluate the model performance for different parameter settings.

For Logistic Regression, we experimented with the regularization parameter 'C' and the penalty type ('l2'), while for SVM, we varied the regularization strength 'C' and the 'gamma' parameter, both of which significantly influence the model's ability to generalize.

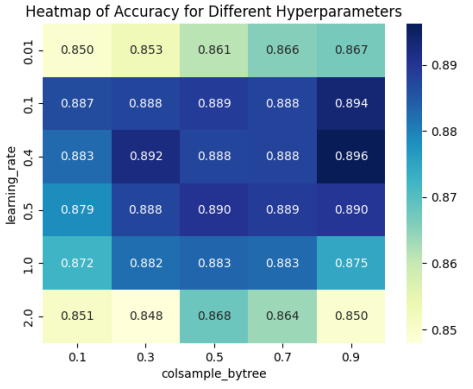

Figure 12. Heatmap of Accuracy for Different Hyperparameters: 'learning_rate' and 'colsample_bytree'

In the first heatmap for XGBoost, we analyzed the impact of 'learning_rate' and 'colsample_bytree' on model accuracy. The optimal combination, which resulted in the highest accuracy (0.8962), was found to be 'learning_rate' = 0.4, 'colsample_bytree' = 0.9, with other parameters set as 'max_depth' = 7, 'min_child_weight' = 5, and 'gamma' = 0.3. This heatmap helped us visualize how these specific parameters affect the model's performance, guiding us toward the best hyperparameter set.

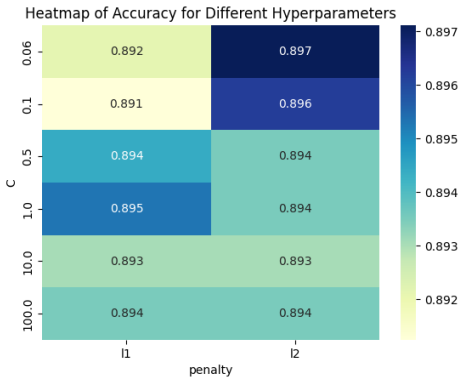

Figure 13. Heatmap of Accuracy for Different Hyperparameters: 'C' and 'penalty'

For Logistic Regression, we varied the 'C' parameter (which controls the regularization strength) and the 'penalty' type ('l2'). The heatmap reveals that the combination of 'C' = 0.06 and 'penalty' = 'l2' achieved the best accuracy of 0.8971. By tuning these parameters, we were able to find the best model fit for the data while preventing overfitting.

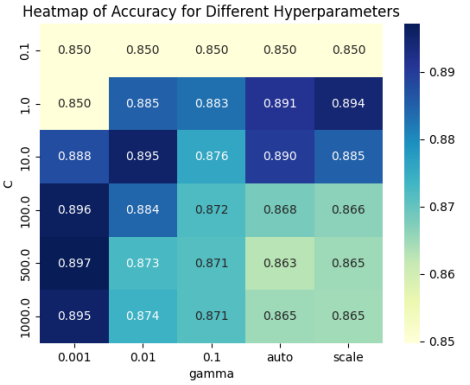

Figure 14. Heatmap of Accuracy for Different Hyperparameters: 'C' and 'gamma'

In the SVM model, we tested different values of 'C' and 'gamma', which govern the regularization and the decision boundary's complexity. The best accuracy (0.8971) was obtained with 'C' = 500 and 'gamma' = 0.001. This heatmap illustrates the sensitivity of the model's performance to these hyperparameters, providing insights into how to fine-tune the model for better generalization.

Overall, the heatmaps provide a visual representation of the impact of each hyperparameter on model accuracy. By conducting this analysis, we identified the optimal hyperparameter settings for each model, which contributed to improved overall performance.

4. Conclusion

This report explored customer consumption behavior and built a predictive model to forecast customer acceptance of marketing offers in a food delivery app. By analyzing customer profiles, product preferences, and campaign responses, we identified key factors influencing customer engagement.

We applied several machine learning techniques—Logistic Regression, SVM, and XGBoost—to predict customer behavior. XGBoost emerged as the most effective model, with SVM following closely behind. The use of K-Fold

Cross-Validation validated the robustness of our results, mitigating potential issues with overfitting and providing a more reliable estimate of model performance. Additionally, hyperparameter tuning further enhanced model accuracy, confirming the importance of parameter optimization in achieving better predictive performance.

Overall, this study demonstrates the potential of machine learning to optimize marketing strategies. By leveraging customer insights, businesses can tailor campaigns more effectively, increasing customer engagement and driving growth. Future work could explore further optimization or the incorporation of additional data to enhance prediction accuracy.