Python: Customer Spending Segmentation and Predicting Campaign Offer Acceptance

Goals

Understanding customers is crucial for creating effective marketing strategies by predicting high-potential customers who are most likely to respond to marketing campaigns while optimizing limited resources. I aim to improve campaign efficiency by classifying customers into distinct spending groups and accurately predicting customer responses to the most recent marketing campaign using historical marketing campaign response data and customer demographic information. To achieve this goal, I separated data analysis into 2 different parts. First, I used CART regression and K-means clustering for customer segmentation depending on their spending on products. I analyzed customer data from the retail industry to segment customers on their spending behavior. This segmentation identifies which spending groups are more likely to respond to marketing campaigns and highlights effective marketing channels for targeting specific groups. Second, I used several machine learning models to predict customer responses to the last marketing campaign. I divided the entire dataset into a train and test dataset to evaluate each model performance and select the best-performing model based on several metrics such as accuracy, precision and recall.

Dataset Review

The dataset contains 2,240 unique customer records. The customers are distinguished by customer ID and each ID number is not repeated. There 24 null values in the 'Income' variable among the total 2240 rows. The dataset consists of 5 distinct components. First, it includes **customers' demographics**, such as educational level, income and the number of children. Specifically, there are separate columns for 'Kids at home' and 'Teenagers at home'. I decided to combine them into a single categorical variable by assigning '1' if customers said 'yes' in either column and '0' if they answered 'no' in both columns to avoid overlaps that could occur when customers check both columns, even with only one kid or teenager at home, as no age range is specified in the dataset. Second, there are columns indicating the **amount spent on products** such as meat, fish, wines and so on. These products are primarily necessities but wines and gold products are categorized as luxury items. I summed all spending columns and created a new 'Spending' column to make it as the dependent variable for CART regression to segment customers into distinct spending groups to identify target spending groups for future marketing campaigns. Third, there are categorical variables related to **marketing campaign acceptance** by customers encoded by '1' (Response Yes) and '0' (No response). There are 6 variables and they indicate the customer acceptance for campaigns from the 1st to the 6th (the last) campaign. Customer responses for the most recent campaign is included in the 'Response' variable which serves as the dependent variable for classifying customers into responders and non-responders during predictive modeling. Fourth, the dataset lists the **purchasing channels** chosen by customers. The number of purchases through different purchasing channels such as in-store, online or other purchasing methods are specified. Lastly, there are some variables related to this company such as customer enrollment date (Dt_customer), the number of days since their last purchase (recency) and whether customers complained within the last 2 years (Complaints).

Data Pre-processing

(1) Outliers

The dataset contains 2,240 unique customer records defined by a unique customer ID. There are 24 null values only in the 'Income' variable. I removed 24 null values as the proportion of null values is relatively small so elimination of null values does not have a significant negative impact on the data accuracy. I found outliers in the 'age' variable which was obtained by subtracting the current year (2024) and the year of birthday for each customer. Referring to the box plot of the 'age' variable, there are 3 outliers exceeding the upper whisker which were removed from the dataset. In contrast, there are many outliers in columns related to spending on products and 'Income' variable as the distribution of those spending columns are right-skewed. I decided not to remove outliers as most distributions for spending and income typically are right-skewed.

(2) Pre-processing

There are some categorical variables such as marital status, education level, and the number of kids. For marital status, I dropped 'Absurd', 'Alone', 'YOLO', 'Widow' and left 'Married', 'Together', 'Single', 'Divorced' as the total proportion of dropped variables are relatively small and after dropping categories, the distribution became more balanced which is necessary for more robust analysis. Due to the same reason to make the analysis more robust by balancing the distribution for the education level, I combined PhD and Master into an Advanced degree and 2nd cycle and Basic into Other. For the number of kids related variables such as ‘Kidhome’ and ‘Teenhome’, I combined these variables and created a new variable called as ‘Has_kids_or_Teenagers’ because there is no pre-defined age range in the original dataset which could overlap counting the number of kids or teenagers from the same household occurred by checking both columns even if they have only one child. If either column was yes for having kids or teenagers, 1 is assigned to the new variable whereas if neither column was yes for having kids or teenagers, 0 is assigned to the new variable. Finally, there are customer response variables from the 1st to 5th campaigns and Response variable is assigned to the most recent campaign which is dependent variable for machine learning models. There are only a few number of customers responding ‘Yes’ to each marketing campaign which could affect the robust analysis due to their imbalance. Typically, most customers do not frequently say ‘yes’ to marketing campaigns which was also exhibited in this dataset. I combined all customer acceptance variables and created a ‘AnyAccepetedCmp’ to assign 1 (yes) if customers respond at least once to any of the marketing campaigns in order to increase the proportion of ‘yes’ response of the new variable for enhancing the analysis power. I draw the distribution of Recency variable but there seems no specific distribution. I converted ‘Dt_Customer’ into a new variable ‘length’ by subtracting 2024 and the year of customer registration date. There are only 3 categories which are 10, 11, and 12 years. From the correlation matrix with numerical variables, there wasn’t any strong correlation between variables. The correlation between amount spent on meat products and the new total spending variable is 0.85 which is quite strong correlation. However, the data seems quite dispersed which reduces the reliability of the correlation value and possible outliers seems to lift the correlation upward based on the scatter plot.

(3) Summary table, Non-parametric test and Chi-squared test of Independence

I implemented a non-parametric test for non-normal numerical data such as income, recency, spend, length, age, amount spent on product and the number of purchases through different purchasing channels to analyze the population distributions between responders and non-responders. For the recency, the population distribution of responders is to the left of the population distribution of non-responders which indicates that responders more frequently shop than non-responders. For the income, the population distribution of income for responders is to the right of the population distribution of income for non-responders, which means that responders have higher income than non-responders. For the spending variable, responders spend more than non-responders. Finally, for purchasing methods, the distribution of the number of web purchases and catalog purchases of responders is to the right of the distribution for non-responders which indicates that responders purchase more than non-responders online and through catalog. However, there are no distribution differences between them in offline stores, the number of web visits and the number of purchases through discounts. In addition, I implemented Chi-squared test of independence for categorical variables to identify if there is a difference in response rate among different categories per categorical variable. For the education level, customers with advanced degree respond more often than customers with low education level. For the marital_status, customers living with partners respond less than customers living alone. For the number of kids or teenagers, customers with kids respond less than customers without kids. For the customer response variable, customers who respond to at least once marketing campaign are more likely to respond than customer who have never respond to one of the marketing campaigns.

Explaining Results

(1) Customer Segmentation

I segmented customers using a CART regression and K-mean clustering. I tried to implement multiple linear regression using the total spending variable as the dependent variable and identified independent variables using stepwise variable selection. However, there were severe multicollinearity and heteroscedasticity. Multicollinearity does not negatively impact predictive accuracy but violating homoscedasticity assumption could adversely affects the accuracy of the linear regression model. I decided not to use multiple linear regression and instead opted for the CART regression model which can be used without satisfying assumptions mandatory for linear regression such as homoscedasticity. I used the 'spend' variable as the dependent variable and selected independent variables using important feature values obtained from the CART regression. I removed all independent variables which did not contribute to improving the predictive power of the regression model and regressed the dependent variable on selected independent variables based on feature importance. The final selected independent variables are the number of purchases through catalogues, income, the number of purchases through a store, customer campaign responses ('AnyAcceptedCmp'), length and the number of purchases through website. I divided customers into 'Low Spender', 'Medium Spender', 'High Spender', and 'Very High Spender' using K-means clustering. The segmented customer groups exhibited a significant difference in the number of purchases through catalogs. In addition, customers in the medium spender exhibited the highest amount of predicted spending, followed by those in the high spender. In conclusion, future marketing campaigns could be effective when targeting customers in the medium spender and high spender through catalog-related campaigns to increase conversions.

(2) Classification

I implemented machine learning models such as logistics regression, decision tree, random forest, KNN to predict customer campaign responses. The objective aims to identify the best predictive model and accurately classify customers as non-responders or responders, particularly improving the accuracy of identifying responders. I used a confusion matrix to evaluate the classification performance. I divided the entire dataset into training and testing sets to use the test set for the model evaluation. My primary focus was on improving the accuracy of correctly predicting responders rather than non-responders. The dependent variable is 'Response' variable which is related to customer responses to the most recent campaign. However, there are a significant larger number of non-responders compared to the limited number of responders in the dataset, which decreased the predictive power of correctly identifying responders. I implemented several metrics for the model evaluation such as Accuracy, F1 score and AUC. Accuracy measures how well each model predicts both responders and non-responders. It is related to overall model accuracy. F1 score is a harmonic mean between Precision and Recall. Precision is also related to the model predictive accuracy. It compared predicted values with actual values to evaluate the proportion of correctly predicted values among all predicted values. In contrast, Recall is related to how predictive model is sensitive to identify the actual responders from the dataset. AUC score is related to the model performance to distinguish true responders from true non-responders.

Final model selection

I finally selected Random Forest as the best predictive model for customer campaign responses by comparing it with other machine learning models based on Accuracy, F1 score and AUC score. Random forest achieved the highest score in each metric, particularly I focused on F1 score for responders as the objective aims to accurately predict the number of customers who actually responded to the most recent marketing campaign. However, due to the imbalance between the number of responders and non-responders, F1-score for responders is relatively low compared to that for non-responders.