Modeling Customer Responsiveness to Marketing Campaign using Machine Learning

Tu Tong
tongtu@dickinson.edu

Anisha Choudhury
choudhua@dickinson.edu

1. Introduction

With rapid technological advancements and easy access to information, customers today have become increasingly knowledgeable about their preferences. They are more likely to resist standardization, prefer differentiated products and prefer companies that directly cater to their individual needs. This has created a challenge for companies as customers are less likely to pay attention to marketing messages that do not target their needs. Additionally, consumers are characterized by changing buying preferences making it necessary for ‘identification of the ideal customer and understanding their purchasing patterns’ (Dias & Rosario, 2023).

In research by Braverman (2015), 80% of worldwide panelists affirmed the importance of data in deploying their marketing and advertising efforts. Additionally, 77.4% reported that they were confident in data-driven marketing and its prospects for the future. These statistics show the enthusiasm towards data-driven marketing practices and their potential to improve company growth and performance.

By analyzing the factors influencing customer response to companies' campaigns we can understand the characteristics of those customers. Using machine learning, we aimed to create models that would help us learn and understand marketing strategies. This will not only optimize the use of customer information in marketing but also help companies with resource allocation by focusing on their campaigns that are the most effective.

2. Related work

Demographic and purchasing history data are widely recognized as valuable predictors of customer decision-making in many business contexts. Numerous studies have explored the use of regression models and machine learning techniques to forecast customer behavior. For example, Islam et al. (2022) investigated the predictive power of using only socio-demographic variables in purchase decisions by implementing a range of methods and evaluating their performance to determine which yielded the most accurate results. Similarly, Yin (2022) discussed how machine learning can address key marketing challenges, such as market segmentation, recommendation systems, and customer behavior prediction. Yin's work also reviewed various techniques, including random forests, logistic regression, and artificial neural networks, highlighting their respective strengths and limitations in marketing applications.

Inspired by these studies, our project seeks to apply both traditional approaches, like logistic regression, and advanced machine learning models, such as random forests and XGBoost, to better understand customer behavior using our dataset. Specifically, we aim to examine how different customer characteristics or subgroups influence the likelihood of responding to targeted marketing campaigns.

3. Data and features

We use the Kaggle dataset on customer personality analysis which includes a detailed analysis of a company’s ideal customers. Initially, the dataset contains 29 columns and 2240 rows which consist of 28 features and 1 response column.

Some of the important features in the dataset:

People

- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Recency: Number of days since customer's last purchase

Products

- MntWines/Fruits/Meat/Fish/Sweet/Gold: Amount spent on wine, fruits, meat, fish, sweet, gold in last 2 years (6 seperate categories).

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1/2/3/4/5: 1 if customer accepted the offer in the n^{th} campaigns, 0 otherwise (5 seperate categories).
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

We perform the following data pre-processing steps:

i. We remove rows with missing values and construct one-hot encoding for categorical feature: *marital_status* and label encoding for ordinal feature: *education*.

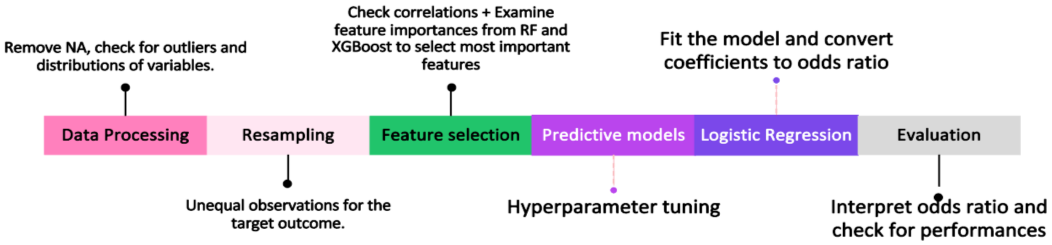
ii. Normalize *income* by scaling to convert a right-skewed income to approximately normal.

iii. Drop numerical and categorical features with significantly lower correlation.

iv. Resample response variable using the method of oversampling.

4. Methods

The following outline steps was used for our project:


Remove NA, check for outliers and distributions of variables.
Check correlations + Examine feature importances from RF and XGBoost to select most important features
Fit the model and convert coefficients to odds ratio
Data Processing
Resampling
Feature selection
Predictive models
Logistic Regression
Evaluation
Unequal observations for the target outcome.
Hyperparameter tuning
Interpret odds ratio and check for performances

20% of the final dataset was set aside to test accuracy and performance for all models. Three models, Random Forests, XGBoost, and Logistic Regression, were used to understand feature importance and interpret features within the model. In this case, we aimed to interpret how each characteristic (feature) impacts the response to our most current campaign (response). Due to unbalanced classes (1400/300), we first use the oversampling method on the initial training set to create a new balanced set that was used for all our models. Before running all models, we also checked for correlation between all features and the response variable and removed all features that have little correlation.

Random Forest

In this project, we used the Random Forest algorithm to assist with feature selection. Random Forest is a robust and versatile machine learning method that improves accuracy by aggregating the results of multiple decision trees built on different subsets of data and features. At each split in a tree, it considers a random subset of features, which allows the model to estimate the relative importance of each feature based on how much it contribute in each split. Given the large number of features in our dataset, Random Forest is particularly useful for identifying the most influential variables in predicting customer response. To ensure accurate feature importance rankings, we trained the model using all features and applied hyperparameter tuning using grid search with cross-validation, adjusting parameters such as number of estimators (*n_estimators*), maximum tree depth (*max_depth*), minimum samples required to split an internal node (*min_samples_split*), minimum samples required at a leaf node (*min_samples_leaf*), and the number of features to consider when looking for the best split (*max_features*). This process ensures that we obtain the most reliable feature rankings under optimal model settings, which helps streamline subsequent model training and improve interpretability.

XGBoost

In addition to Random Forest, we used XGBoost (Extreme Gradient Boosting) as a second model for feature selection and to improve overall predictive performance. XGBoost is a powerful ensemble learning method based on gradient boosting, known for its high efficiency, scalability, and accuracy on structured data. The XGBoost model can learn from previous iterations, which makes it typically reach more accurate results than other models, although it takes much longer time to run. For this reason, we wanted to see how features are ranked in XGBoost and use it as a comparison to Random Forest rankings. Like Random Forest, we also selected the most relevant features based on the correlations for XGBoost in order to help identify the most relevant predictors for the Customer Response variable. To ensure reliable results, we applied hyperparameter tuning using grid search with cross-validation, adjusting parameters for number of estimators *(n_estimators*), maximum tree depth (*max_depth*), learning rate (*learning_rate*), subsample ratio of training instances (*subsample*), subsample ratio of features per tree (*colsample_bytree*), minimum loss reduction for further partitioning (*gamma*), L1 regularization term (*reg_alpha*), and L2 regularization term (*reg_lambda*).

After running feature importances on both Random Forests and XGBoost, we compared to see the overlapping important features and use them in our Logistic Regression. This narrowed down the total number of features and obtained a more robust understanding of which features consistently contribute to predicting customer response.

Logistic Regression

To analyze customer behavior and interpret the influence of various customer attributes on campaign response, we employed a logistic regression model. Logistic regression was chosen due to its suitability for binary classification problems and its interpretability through odds ratios, which allowed us to understand both the direction and strength of the relationship between predictors and the outcome variable.

Our dependent variable was a binary indicator of whether the customer accepted the most recent marketing campaign. The independent variables were selected based on feature importance rankings obtained from Random Forest and XGBoost models, ensuring that only the most predictive features were retained for logistic regression analysis. The final set of predictors used in the logistic regression model included:

- Recency: # of days since customer's last purchase

- MntWines: Amt spent on wine in last 2 years
- MntMeatProducts: Amt spent on meat in last 2yrs
- MntGoldProds: Amt spent on gold in last 2yrs
- Income: Customer's yearly household income
- Marital_Single: Customer's marital status
- Kidhome: # of children in customer's household
- Teenhome: # of teenagers in customer's household
- AcceptedCmp1/2/3/4/5: 1 if customer accepted the offer in the nth campaign, 0 otherwise

The model was implemented using Python's scikit-learn library. Logistic regression was trained on a balanced version of the dataset to address class imbalance and improve generalizability. We increased the maximum number of iterations (*max_iter=2000*) to ensure model convergence and suppressed convergence warnings for cleaner output.

After fitting the model, we extracted the coefficients and transformed them into odds ratios using the exponential function. This transformation allows for a more intuitive interpretation of each predictor's impact on the likelihood of a campaign response. Specifically, odds ratios greater than 1 indicate a positive association with the outcome (i.e., higher odds of accepting the offer), whereas odds ratios less than 1 suggest a negative association. These odds ratios allowed us to identify key behavioral and demographic patterns in customer responses.

5. Results

The logistic regression model was fitted to examine the association between customer attributes and the likelihood of accepting a marketing offer. The exponentiated coefficients (i.e., odds ratios) are presented in Figure 1. The results are interpreted to understand both the magnitude and direction of these relationships.

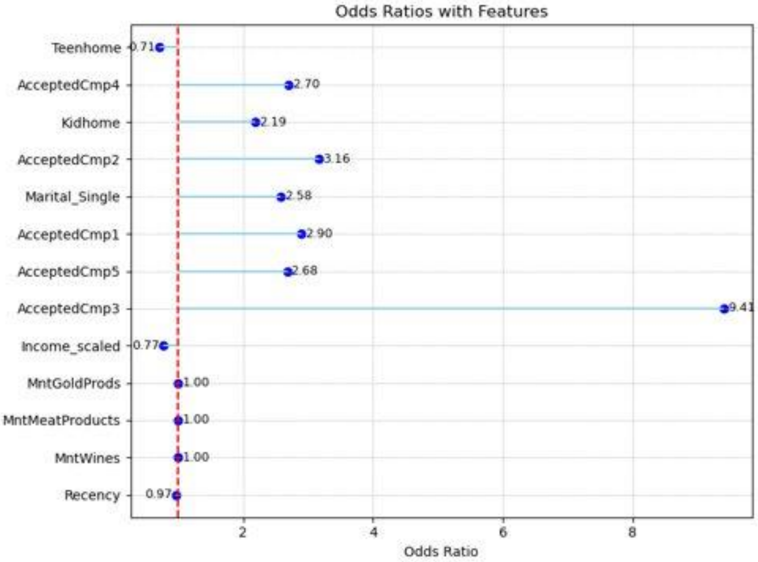


Figure 1

All predictors were found to have statistically significant associations with offer acceptance. The number of teenagers in the household had a negative association with response likelihood; each additional teenager reduced the odds of acceptance by approximately 29%. Similarly, higher income levels were associated with a decreased likelihood of response, with each unit increase in scaled income lowering the odds by 23%. A small but consistent negative effect was also observed for *Recency* (i.e., days since last purchase) - each additional day corresponded to a 2.7% reduction in odds.

In contrast, several variables showed positive associations. For example, each additional child in the household more than doubled the odds of acceptance, increasing them to approximately 2x. Customers who were single had around 2.5x higher odds of accepting compared to those who were married or in other relationship categories. Prior campaign behavior was also highly predictive: customers who had previously accepted offers were significantly more likely to respond again. Specifically, accepting Campaign 1, 2, 4, or 5 increased the odds of acceptance by roughly 3x, 3x, 2.5x, and 2.5x, respectively. Notably, customers who accepted Campaign 3 were over 9x more likely to accept the current offer, underscoring a strong effect of past engagement.

Spending behaviors also had minor but directionally consistent effects. Each additional unit spent on wine, meat, or gold products was associated with a marginal increase in odds - 0.13%, 0.19%, and 0.32%, respectively - indicating a weak but positive correlation between spending and offer receptiveness.

6. Conclusion

With the help of machine learning classifiers like Random Forest and XGBoost, we narrowed down our important features to 13 predictors that are used in logistic regression for classification. Our approach helps skim over all available features in the initial dataset and remove irrelevant ones before putting them in the final regression. By utilizing machine learning for feature selection, we provide a more robust check of feature relevance and importance, ensuring that we do not miss any significant variables.

At the end, we reached an accuracy score of about 0.72 for our regression, which indicates a generally good fit for our dataset. By converting the coefficients into odds ratios, we were able to quantify and compare the likelihood of customer response based on their characteristics.

The results from our project not only help us understand the demographics of customer engagement for our most current marketing campaign but also allow the business to evaluate the performance of its latest marketing campaign. Particularly, seeing whether this campaign reached its initial target audience and using it as a reference for future campaigns.

7. References

- Islam, T., Meade, N., Carson, R. T., Louviere, J. J., & Wang, J. (2022). The usefulness of socio-demographic variables in predicting purchase decisions: Evidence from machine learning procedures. Journal of Business Research, 151, 324–338. https://doi.org/10.1016/j.jbusres.2022.07.004
- Braverman, S. (2015). Global review of data-driven marketing and advertising. Journal of Direct, Data and Digital Marketing Practice, 16, 181–183. https://doi.org/10.1057/dddmp.2015.7
- Albérico Travassos Rosário, Joana Carmo Dias. (2023). How has data-driven marketing evolved: Challenges and opportunities with emerging technologies. https://doi.org/10.1016/j.jjimei.2023.100203
- Yin, Yuanzheng. (2022). Marketing Strategies Based on Machine Learning Approaches. https://doi.org/10.2991/978-94-6463-030-5_42