Fall 2024 STAT650 - Customer Personality Analysis

Hwiyoon Kim

Nov 26, 2024

1 Introduction

The project focuses on analyzing customer behavior using a dataset derived from a marketing campaign.[1] The goal is to examine various customer characteristics (such as income, education, spending habits, and web visits) and determine how these influence their response to marketing offers. This analysis compose various regression models and exploratory data analysis (EDA) techniques to predict and interpret customer behavior.

The objectives of this project:

- To explore and preprocess customer data for insights into their behavior.
- To understand how various customer features, such as income and spending habits, relate to their response to marketing campaigns.
- To build and evaluate different regression models (simple, multiple, logistic, polynomial) to predict customer behavior.

Research questions:

- How do customer characteristics relate to the likelihood of accepting marketing offers?
- Which features are most predictive of customer responses to campaigns?
- How can regression models be applied to predict customer behavior and campaign success?

2 Dataset Description

The dataset used for this analysis is sourced from Kaggle, titled Customer Personality Analysis formed with 2240 data observations along with 29 variables. It contains information about customers' demographics, purchasing behavior, and responses to marketing campaigns.

Quantitative Variables:

1. Income: Measures the annual income of the customer in monetary units.
2. Recency: Measures the number of days since the customer's last purchase.
3. MntWines: Measures the amount spent on wines.
4. MntMeatProducts: Measures the amount spent on meat products.
5. NumWebVisitsMonth: Measures the number of visits to the company's web page per month.
6. NumStorePurchases: Measures the number of purchases made in stores.

Qualitative Variables:

1. ID: A unique identifier for each customer.

2. YearBirth: The birth year of the customer.
3. Education: The level of education attained by the customer.
4. MaritalStatus: The marital status of the customer.
5. DtCustomer: The date when the customer registered.
6. Kidhome: Counts the number of children.
7. Teenhome: Counts the number of teenagers.
8. AcceptedCmp: Customer respond to each of five different marketing campaigns (yes/no).
9. Complain: Indicates whether the customer has complained (yes/no).
10. Response: Indicates whether the customer responded to the last marketing campaign (yes/no).

3 Data Pre-processing

Missing values in the dataset were filled by getting the median value of the respective columns. Specifically, the Income column, which had the most missing columns, was imputed with the median income value to avoid data loss and maintain consistency. For data cleaning and preparation, The MaritalStatus variable underwent label encoding to convert it into a numerical form, making it suitable for inclusion in analytical models. Additionally, a selection of variables was carefully chosen based on their relevance to the target variable, which in this case is the response to marketing campaigns. These variables include Income, Recency, MntWines, MntMeatProducts, NumWebVisitsMonth, and NumStorePurchases. Each of these variables was selected for their potential impact on understanding and predicting campaign responses. Also, Recency and Income were transformed into numerical formats for consistency.

4 Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

The dataset consists of 2240 rows and 29 columns. For the Income variable, the average income was 52,247, with a median of 51,251. The standard deviation was 24,247, indicating a wide spread in customer incomes. The minimum income observed was 1,730, and the maximum was 666,666. A histogram of the Income variable revealed a right-skewed distribution, suggesting that most customers had incomes below the mean. Similarly, the YearBirth feature showed a population concentration around customers born between 1893 and 1996. A boxplot of Income identified several outliers, primarily at the upper end of the distribution.

4.2 Bivariate Analysis

The correlation matrix revealed several notable relationships between the variables. There is a strong positive correlation between Income and MntWines (correlation: 0.579), indicating that higher-income customers tend to spend more on wine. Similarly, Income and MntMeatProducts showed a correlation of 0.585, suggesting that higher-income customers also spend more on meat products. Another strong correlation was observed between MntWines and NumStorePurchases (0.642), which highlights that customers who spend more on wine are also likely to make frequent purchases in physical stores. A moderate correlation of 0.542 between MntWines and NumWebPurchases indicates that wine buyers also make a notable number of online purchases, though the preference is stronger for physical stores.

Weak or insignificant correlations were also observed. For instance, the correlation between Recency and Income was -0.004, suggesting no meaningful relationship between how recently a customer made a purchase and their income level. Similarly, the correlation between Recency and MntWines was 0.016, indicating that recency has little to no effect on wine spending.

Negative correlations were identified between YearBirth and spending behaviors. YearBirth and Income had a correlation of -0.162, meaning younger customers (those with a higher birth year) generally have lower incomes. Likewise, the correlation between YearBirth and MntWines was -0.158, showing that younger customers tend to spend slightly less on wine compared to older customers.

4.3 Multivariate Analysis

The recency quartile distribution divides customers into four groups based on how recently they made a purchase. The groups were fairly balanced, with 567 customers in Q1 (most recent), 555 in Q2, 567 in Q3, and 551 in Q4 (least recent). This even distribution ensures that each quartile provides meaningful insights.

The mean wine spending (MntWines) and income for each quartile highlight interesting trends. Customers in Q1 (most recent) have the highest average income (52,976.74) but spend moderately on wine (290.66). In contrast, customers in Q3 (moderately recent) spend the most on wine (333.30), despite having a slightly lower average income (52,952.95) compared to Q1. Customers in Q2 have the lowest average income (51,359.82) but spend slightly more on wine (298.30) than Q1. Lastly, customers in Q4 (least recent) have an average income of 51,660.19 and spend 293.06 on wine, which is slightly higher than Q1 but less than Q3.

These results shows that customers in Q3 represent a prime segment for wine-related promotions, as they have the highest spending on wine. Customers in Q1, who have the highest incomes, might be better suited for premium or exclusive product offerings. The weak correlation between recency and spending behaviors (e.g., wine and income) suggests that targeting strategies should focus on high-value customers rather than recency alone. Furthermore, the strong correlation between wine spending and store purchases (0.642) indicates that physical store promotions may yield better results for wine buyers compared to online channels.

In summary, the demonstration shows the importance of segmenting customers by both income and spending behavior to design effective marketing strategies. While recent customers with high incomes (Q1) provide opportunities for premium campaigns, moderately recent customers (Q3) stand out for their spending patterns and responsiveness to wine promotions.

5 Regression Analysis

5.1 Simple Linear Regression

A simple linear regression model was fitted using Income as the predictor and MntWines as the target variable. The regression equation derived was:

$$MntWines = -208.5 + 0.013 \times Income.$$

This model explained 41% of the variance in MntWines ($R^2 = 0.41$), with a Mean Squared Error (MSE) of 62351.62.

5.2 Multiple Linear Regression

A multiple linear regression model, incorporating Income, Recency, and Kidhome as predictors, slightly improved the fit. The adjusted $\mathrm{R}^{\mathbf{2}}$ increased to 0.57, and the MSE decreased to 46693.88. Variance Inflation Factor (VIF) analysis confirmed no significant multicollinearity among the predictors.

5.3 Polynomial Regression

Polynomial regression was used to capture non-linear relationships, including quadratic terms for Income. This model further improved the fit, achieving an $\mathrm{R}^{\mathbf{2}}$ of 0.45 and reducing the MSE to 54651.0. This indicates that the polynomial model captured additional complexities in the data compared to linear models.

5.4 Logistic Regression

Logistic regression was applied to predict the Response variable (acceptance of a marketing campaign). Coefficients revealed that Income and Recency significantly influenced the likelihood of a positive response. The model achieved an accuracy of 91% , precision of 91% , and recall of 91% , suggesting a well-balanced performance for binary classification.

5.5 Regularization Techniques

Regularization methods, including LASSO, Ridge, and Elastic Net, were implemented to reduce overfitting and improve generalizability. LASSO effectively shrank non-significant coefficients to zero, while Ridge regression provided consistent shrinkage across coefficients. Elastic Net balanced these approaches. MSE values for these methods were:

- LASSO: 46,692.85
- Ridge: 46,693.84
- Elastic Net: 46,661.18

R^2 values followed a similar trend, with Elastic Net achieving the highest value of 0.5657.

6 Model Evaluation and Comparison

Across all models, polynomial regression exhibited the best fit for regression tasks, with the lowest MSE and highest R^2. For binary classification, logistic regression performed well, balancing precision and recall effectively. Regularized regression methods demonstrated better generalizability by reducing overfitting, with Elastic Net offering the best trade-off between feature selection and predictive power.

7 Model Evaluation and Comparison

In conclusion, income was a significant predictor of spending across all models. Polynomial regression proved effective in capturing non-linear relationships, while logistic regression highlighted key factors driving campaign acceptance. Regularization techniques improved model interpretability and reduced overfitting. These findings suggest that marketing strategies should focus on higher-income customers for product promotion and campaign targeting, improving the insights provided by these models.

References

[1] Kaggle. Customer personality analysis dataset. *Kaggle Dataset*, 2024. Retrieved from https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis on 2024-11-26.