# Customer Personality Modeling and Evaluation

Jin-Zhou Li, Wei-Zhi Hou
Department of Applied Mathematics
National Dong Hwa University

June 12, 2024

**Abstract** This study will model customer personalities and address common issues in financial data, such as multicollinearity and outliers, by introducing robust logistic regression analysis to predict whether a customer will participate in an event. Due to the imbalance in the predicted target, the commonly used accuracy metric is abandoned in favor of developing a profit-based model evaluation metric to maximize profit as the model selection criterion. Finally, based on principal component selection of variables, the study analyzes and explores the market positioning behind the event and identifies the characteristics of potential customers.

**Keywords** Customer personality analysis, machine learning, robustness.

## 1 Introduction

Customer personality modeling can help businesses identify potential customers to increase revenue and implement targeted marketing to reduce expenses. In event promotion and product marketing, not all customers will participate, and executing targeted marketing can effectively reduce marketing costs. By predicting customer characteristics, businesses can identify potential customers. This study utilizes data from `kaggle.com`, with details in the appendix (on page 5). The dataset includes information on 2,000 customers and 19 customer features. The goal is to predict whether a customer will participate in a marketing activity based on their characteristics, and the data has the following attributes:

1. Several variables have outliers, such as age.
2. Several variables exhibit high multicollinearity, such as age, income, and the number of children in the household.
3. The target prediction is highly imbalanced, with only 7% of participants in the activity.
4. Due to the anonymized nature of the data, the theme of the marketing activity is unknown.

Due to the presence of numerous outliers, traditional linear models may produce significant bias. This study introduces robust modeling; principal component analysis (PCA) is used to eliminate multicollinearity and create several simple indicators. Logistic regression (LR) is employed to build a model with strong explanatory power. Finally, due to the imbalance in the prediction target, accuracy is abandoned in favor of developing a profit-based evaluation metric. This approach answers key questions for businesses, such as potential activity revenue, targeting potential customer groups, and assisting in the creation of precise marketing strategies.

## 2 Modeling

Given that the data exhibits strong multicollinearity and requires an interpretable model, principal component analysis (PCA) is first performed, followed by logistic regression (LR) for modeling. Due to the presence of numerous outliers across multiple dimensions of the data, traditional linear modeling would lead to significant estimation errors. Therefore, robust estimators are needed to mitigate the impact of outliers. The study introduces robust PCA (RPCA) as proposed by Hubert et al. (2005) and robust LR (RLR) as suggested by Feng et al. (2014). For model implementation, the R language's `rrcov` package is used, with the `PcaHubert` function for RPCA, and the `robustbase` package is used, with the `glmrob` function for RLR.

## 2.1 Robust Principal Component Analysis

Let the original data matrix be $M \in \mathbb{R}^{n \times p}$. PCA projects the data points onto a hyperplane. Let $x_i \in \mathbb{R}^p$ be the coordinates of the $i$-th data point, and $p_{x_i}$ be the coordinates of the $i$-th data point after projection. The PCA minimization objective is to minimize the difference in distance between the original data points and their projections, i.e.,

$$\|M\|_* = \min \sum_{i=1}^{n} \|x_i - p_{x_i}\|^2. \tag{1}$$

However, PCA is sensitive to outliers. Robust PCA (RPCA) assumes that most data have high collinearity, forming a low-rank matrix $L$, with a small number of outliers forming a sparse matrix $S$, such that $M = L + S$. The minimization objective is then revised to:

$$\min_{L+S=M} \|L\|_* + \lambda \|S\|_1, \tag{2}$$

where $\|S\|_1 = \sum_{ij} |S_{ij}|$, and $\lambda > 0$ is a tuning parameter that determines the influence of outliers. A smaller $\lambda$ will tend to classify more data as outliers. If the term $\|S\|_1$ is removed, the problem degenerates into PCA.

In Equation (2), if outliers can be identified and treated as points in $S$, the PCA projection of the majority data points in $L$ can be better fitted, significantly reducing the value of $\|L\|_*$. However, the growth rate of $\|S\|_1$ is faster than that of $\|L\|_*$, allowing only a small number of points to be moved into $S$ to achieve balance.

## 2.2 Robust Logistic Regression

Let the prediction target for activity participation be $y$, where $y_i = 1$ if the $i$-th customer participates in the activity, and $y_i = 0$ otherwise. Logistic Regression (LR) uses the sigmoid ($\sigma$) function to map the linear combination of the data into the $(0, 1)$ interval. The prediction indicator for whether a customer participates in the activity is:

$$\hat{y} = \sigma(x^T \beta) = \frac{\exp(x^T \beta)}{1 + \exp(x^T \beta)}, \tag{3}$$

where the choice of $\beta$ depends on minimizing the Cross-Entropy Loss:

$$\varepsilon_i = \begin{cases} -\ln(\hat{y}_i) & \text{if } y_i = 1, \\ -\ln(1 - \hat{y}_i) & \text{if } y_i = 0, \end{cases} \tag{4}$$

$$= -y_i \ln(\hat{y}_i) - (1 - y_i) \ln(1 - \hat{y}_i). \tag{5}$$

As a generalized linear model, LR is also sensitive to outliers. Common robust approaches include:

1. **Weighted Residuals:** Transform the original residuals, for example, by assigning weights to each data point as the inverse of their residuals and iterating several times; or by using the density function of the normal distribution:

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon^2}{2}\right), \tag{6}$$

   such that smaller residuals are assigned larger weights.

2. **Robust Loss Functions:** Set an upper limit for the loss function, e.g., by defining a threshold $t$ and reducing residuals exceeding the threshold to $t$:

$$f(\varepsilon) = \begin{cases} \varepsilon & \text{if } \varepsilon \le t, \\ t & \text{if } \varepsilon > t, \end{cases} \tag{7}$$

$$= \varepsilon I\{\varepsilon \le t\} + t I(\varepsilon > t). \tag{8}$$

3. **Random Sample Consensus (RANSAC):** Similar to RPCA, the goal is to identify outliers. By randomly sampling, certain points may cause a significant decrease in testing accuracy during model training. These points can be considered outliers, with their weights reduced or removed before retraining the model.

# 3 Evaluation

Using RLR, each customer is assigned a predicted value $\hat{y}$ between $(0, 1)$, where a larger value indicates a higher likelihood of participating in the activity. At this point, a threshold value $t$ must be chosen, and $\tilde{y} = I\{\hat{y} > t\}$ is used to determine the predicted class. However, due to the imbalanced nature of the data, setting $t = 1$ results in all predicted values being 0, yielding 93% accuracy. Nevertheless, this is not a reasonable choice, and maximizing accuracy is not an appropriate criterion for selecting $t$.

Inspired by the ROC curve, this study uses profit as the evaluation metric:

$$\text{Profit}_{t,g} = \text{TP}_t(g - 1) - \text{FP}_t, \tag{9}$$

where $g$ is defined as the profit margin ratio, i.e., the percentage profit obtained by selling one unit of a product. Both true positives (TP) and false positives (FP) are affected by the choice of $t$:

$$\text{TP}_t = \sum_{i=1}^{n} I\{y_i = 1, \tilde{y}_i = 1\} = \sum_{i=1}^{n} I\{y_i = 1\} I\{\hat{y}_i > t\}, \tag{10}$$

$$\text{FP}_t = \sum_{i=1}^{n} I\{y_i = 0, \tilde{y}_i = 1\} = \sum_{i=1}^{n} I\{y_i = 0\} I\{\hat{y}_i > t\}. \tag{11}$$

Given $g$, the threshold $t$ is chosen to maximize the profit:

$$t = \underset{t \in [0,1]}{\arg\max}\, \text{Profit}_{t,g} = \underset{t \in [0,1]}{\arg\max} [\text{TP}_t(g - 1) - \text{FP}_t]. \tag{12}$$

This threshold is used to determine whether to send an invitation to a customer. If $\hat{y}_i > t$, an invitation is sent, and the potential profit of the activity is reported as $\max_{t \in [0,1]} \text{Profit}_{t,g}$. Subsequent research confirms that the choice of $g$ does not significantly affect the selection of $t$.

Using a model that implements RPCA and RLR as an example, Figure 1 (left) shows the ROC curve for 30% of the test dataset. Given $g = 10$, the transformation using Equation (9) yields the plot on the right. At this point, selecting $t$ such that FP is around 0.2 results in the maximum estimated profit of $230.

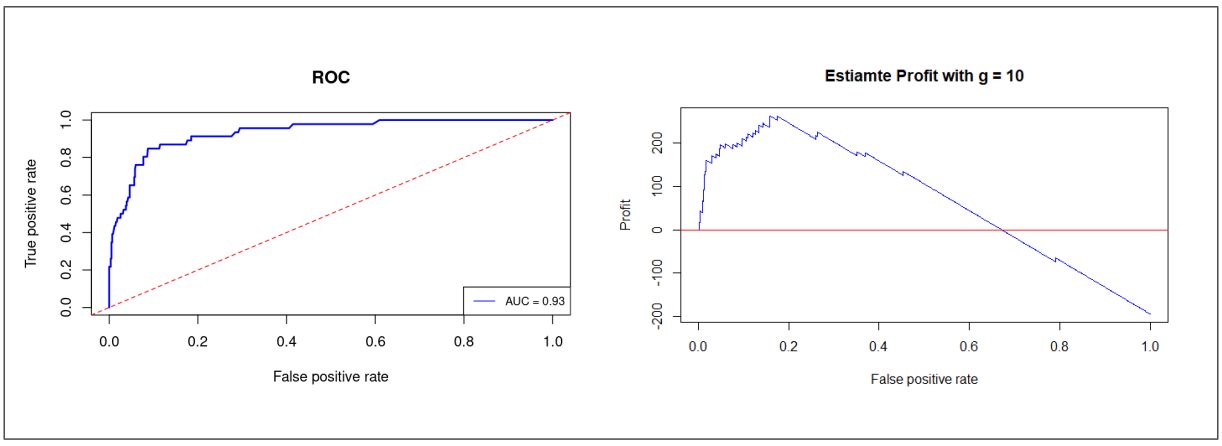


Figure 1: Transformation from ROC to Profit

When $g$ is selected within the range of 7 to 20, the corresponding impact on $t$ is minimal, with FP consistently around 0.2. Therefore, the choice of $g$ can be based on past activity experience or product positioning in the market. Selecting a reasonable profit margin ratio will not significantly alter the prediction results. Using profit as the evaluation criterion, this study also compares the performance of various machine learning models, such as linear discriminant analysis (LDA), random forest (RF), and classification and regression tree (CART). Observing the profit-maximizing models under different machine learning methods shows that RLR with RPCA variables performs similarly to other machine learning models. Thus, RLR is chosen for further investigation. Changes in $t$ with varying $g$ and comparisons across models can be found at `https://imgur.com/a/Lz5IrTY`.

# 4 Conclusion

When dealing with multidimensional data containing outliers, the estimators of linear models often suffer from severe bias. While it is feasible to examine each dimension individually, this approach is inefficient. Robust modeling provides an effective way to mitigate the impact of outliers while retaining as much data as possible.

Using accuracy as the model evaluation criterion implies equal weights for different groups, which is equivalent to assuming identical costs for misclassification errors across groups. However, in this dataset, the costs for different groups are clearly unequal. Predicting participants as non-participants results in a loss of profit, while predicting non-participants as participants merely increases marketing expenses. Developing a profit-maximizing evaluation method not only balances the error costs between groups but also addresses the potential revenue of the activity.

Analyzing the RLR results, four key indicators are derived from RPCA:

| | Estimate | S.E. | Z-value | P-value |
|---|---|---|---|---|
| Intercept | -1.69 | 0.09 | -17.92 | 2e-16 |
| Income Indicator | 0.75 | 0.03 | 22.29 | 2e-16 |
| Youth Group Indicator | 0.48 | 0.05 | 10.19 | 2e-16 |
| Consumption Frequency Indicator | 0.24 | 0.05 | 4.53 | 5.81e-06 |
| Spending Power Indicator | -0.50 | 0.05 | -10.06 | 2e-16 |

Table 1: RLR Coefficients

The RPCA results are interpreted based on variable coefficients with absolute values greater than 0.2 as the significance threshold.

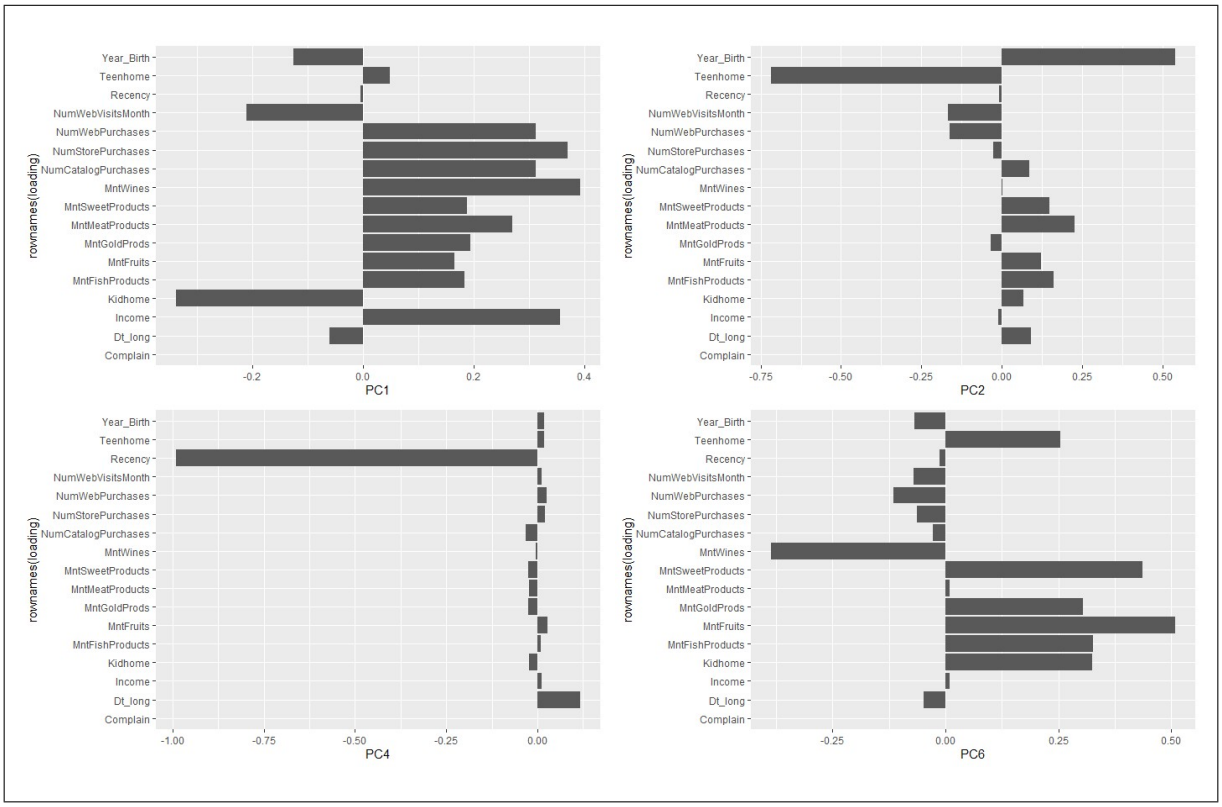


Figure 2: RPCA Coefficients

- PC1, Income Indicator: Negatively correlated with the number of children at home, and positively influenced by preferred shopping platforms, wine sales, and income. PC1 is therefore identified as an income indicator. Considering income as a crucial factor, this suggests that vendors can participate in price-setting and trading models. Hence, the market is inferred to be monopolistic competition. Based on variable design in the dataset, this market is likely branded retail or large-scale supermarkets.

- PC2, Youth Group Indicator: Negatively correlated with the number of young people at home and positively correlated with year of birth. PC2 is identified as a youth group indicator. We infer that the target audience for this activity may primarily consist of young members who are more willing to participate.
- PC4, Consumption Frequency Indicator: Negatively correlated with the number of days since the last purchase. PC4 is identified as a consumption frequency indicator, which, when combined with PC1, increases the likelihood of the market being a large-scale supermarket. After all, supermarket visits are not typically frequent; customers rarely shop daily and might instead make large purchases weekly or monthly.
- PC6, Spending Power Indicator: Positively correlated with wine sales and negatively correlated with some essential food items, the number of children, and the number of young people at home. PC6 is identified as a spending power indicator. We infer that the related variables represent a significant proportion of a household's major expenses, particularly necessities.

Finally, based on the inferred market characteristics and the study on family resource allocation by Hong (2017), we observe the density distribution of wine consumption (Figure 3, left) and the number of children in the household (Figure 3, right) for the two groups. This analysis suggests that the activity might be a supermarket promotion for wine. Participants are likely to be younger groups or families with fewer children, who are more inclined to engage in this activity.

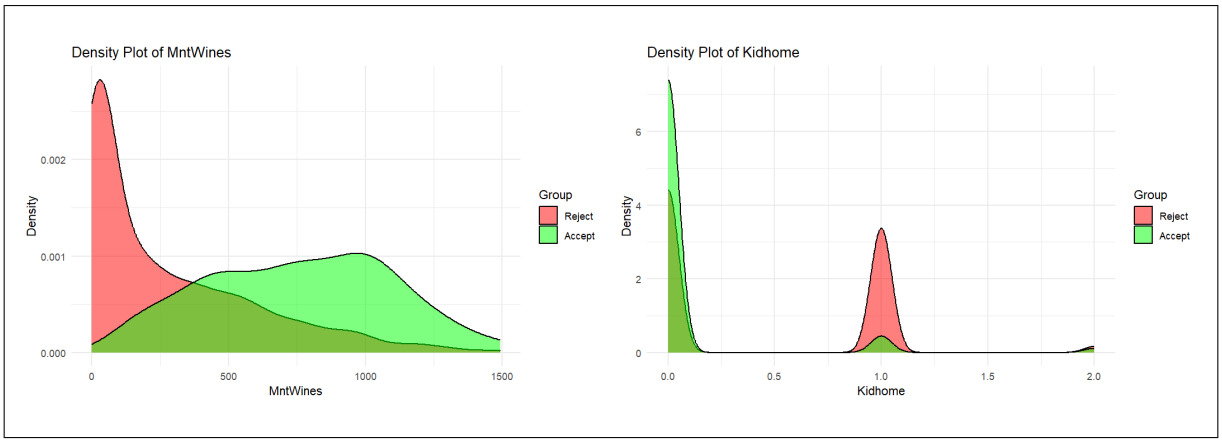


Figure 3: Density Distribution of Two Groups

# 5 References


Feng, J., Xu, H., Mannor, S. and Yan, S. (2014), Robust Logistic Regression and Classification, *in* 'Advances in Neural Information Processing Systems', Vol. 27, Curran Associates, Inc.

Hong, Y.-F. (2017), The Measurement of Child Costs and Interhousehold Resource Allocation—Evidence from Taiwan, 國立中央大學.

Hubert, M., Rousseeuw, P. J. and Vanden Branden, K. (2005), 'ROBPCA: A New Approach to Robust Principal Component Analysis', *Technometrics* **47**(1), 64–79.


# 6 Appendix

The data is sourced from `https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis`, presenting 10 records with information on the year of birth (birth year), education level (education), income (income), number of children in the household (children), wine consumption quantity (alcohol), and participation in the activity (participate).

| ID | birth of birth | education | income | children | alcohol | participate |
|---|---|---|---|---|---|---|
| 2114 | 1946 | PhD | 82800 | 0 | 1006 | 1 |
| 2174 | 1954 | Graduation | 46344 | 1 | 11 | 0 |
| 2225 | 1977 | Graduation | 82582 | 0 | 54 | 1 |
| 4141 | 1965 | Graduation | 71613 | 0 | 426 | 0 |
| 4855 | 1974 | PhD | 30351 | 1 | 14 | 1 |
| 5324 | 1981 | PhD | 58293 | 1 | 173 | 0 |
| 5524 | 1957 | Graduation | 58138 | 0 | 635 | 0 |
| 6182 | 1984 | Graduation | 26646 | 1 | 11 | 0 |
| 7373 | 1952 | PhD | 46610 | 0 | 8 | 1 |
| 9909 | 1996 | 2n Cycle | 7500 | 0 | 24 | 1 |

Table 2: Partial data set