Customer Personality Analysis using SVM and Random Forest

STATS/CSE 790 Project

2024-04-04

Pao Zhu Vivian Hsu (400547994)

Introduction

Understanding the characteristics of customers and adjusting one's products and services to cater to these characteristics is crucial to running a successful business. As such, building models to accurately analyze customer personality can serve as a powerful tool for businesses. In this paper, we use Support Vector Machines (SVMs) and random forest methods to predict customer responses to a marketing campaign by analyzing customer personality.

The data in this study comes from an open sourced website called Kaggle (Patel, 2021). It contains 2240 observations and 29 variables describing the customer, their buying habits, and their interactions with previous campaigns. The response variable is a binary categorical variable indicating whether the customer accepted the offer in the latest marketing campaign or not. The remaining variables are listed below in Table 1:

Table 1: Variable description

Category	Variables
People	ID, birth year, education level, marital status, income, number of children in household, number of teenagers in household, enrolment date with company, number of days since last purchase, whether the customer complained over the last two years
Products	Amount spent on wine, fruits, meat, fish, sweets, gold over the last 2 years
Promotion	Number of purchases made with discount, whether the customer

Category	Variables
	accepted the offer for the past 5 campaigns
Place	Number of website visits in the last month, number of purchases made on the website, catalog, and in-store

Methods

We began the study by cleansing the data to handle any missing and non-interpretable values. Missing values in the income column were imputed using the mean for the associated education level. Rows containing non-interpretable marital statuses were removed.

After cleansing the data, we then performed data visualization using a pairs plot and correlation plot to check for any patterns in the data. Figure 1 shows a subset of the pairs plot, where blue indicates the case where a customer accepts the campaign offer and red indicates the customer does not accept it. There is an imbalance in the outcome. However, we have chosen not to alter this imbalance since this imbalance is often found in real business settings.

Figure 1: Pairs plot of a selection of the variables

Figure 2 shows the correlation plot. We define a strong correlation as those with a correlation coefficient larger than 0.7. Based on the plot, all values are 0.7 or lower so there is no evidence of strong correlations.

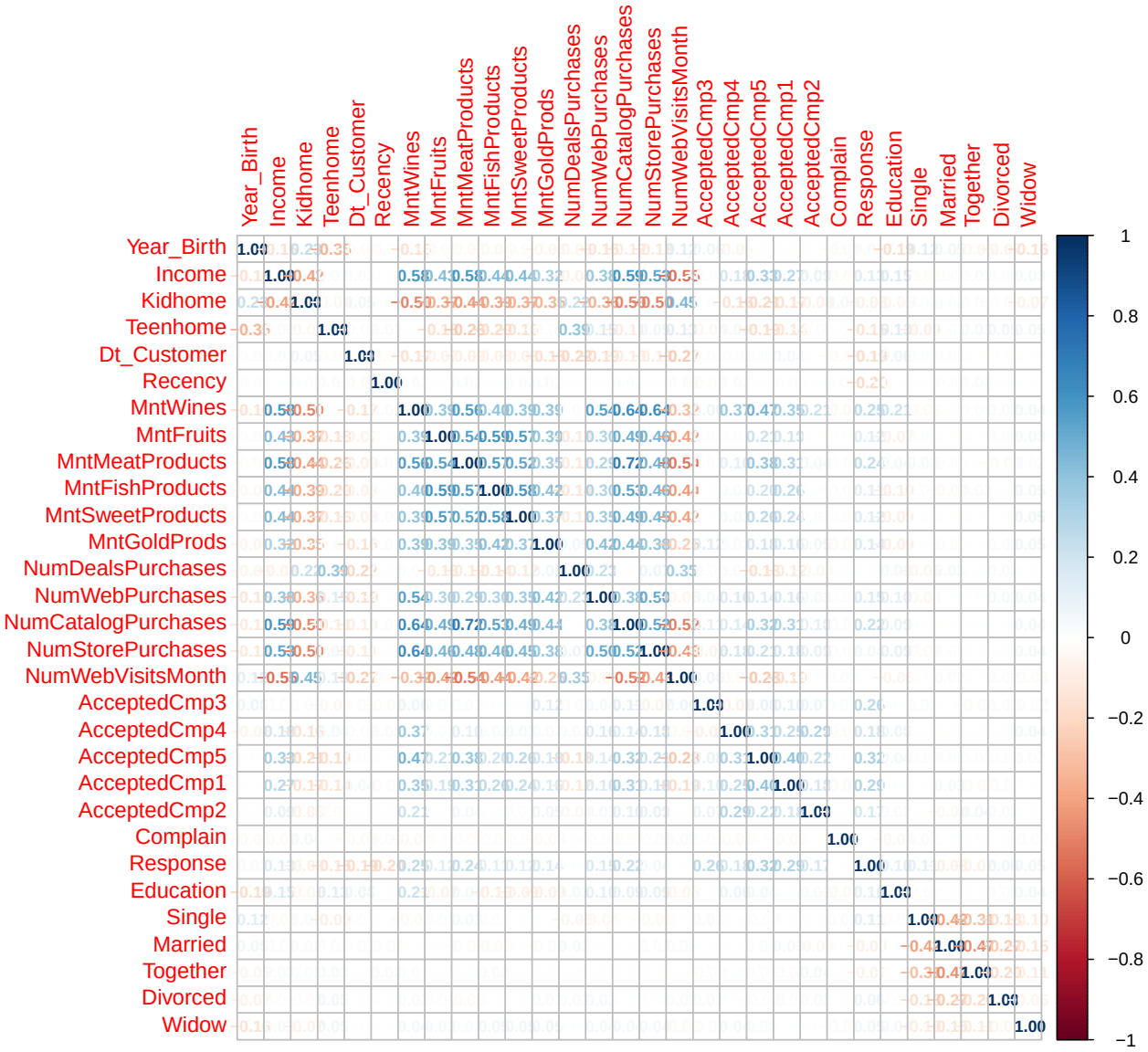

Figure 2: Correlation plot

Next, we split the data into two equal parts to form training and testing sets and applied SVM and random forest methods to the data.

The first method we applied was SVM, which is a supervised learning method that uses decision boundaries to perform classification. These decision boundaries are determined by a kernel, which is a function that measures the similarity between observations and can be used to increase the feature space to accommodate non-linear boundaries between classes (King-Yu, 2024b).

In this study, we used linear, polynomial, and radial kernels to perform SVM. The linear kernel can be written as:

$$K(x_i, x_i{}') = \sum_{j=1}^{p} x_{ij}, x_i{}'_j$$

where i represents the ith observation. The polynomial kernel can be written as:

$$K(x_i, x_i{}') = (1 + \sum_{j=1}^{p} x_{ij}, x_i{}'_j)^d$$

where i represents the ith observation and d represents the degree of the polynomial. The radial kernel can be written as:

$$K(x_i, x_i{}') = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_i{}'_j)^2)$$

where i represents the ith observation and γ is a positive constant (King-Yu, 2024b).

Each of the models were also tuned for the cost parameter C and the gamma constant γ through cross-validation. The cost parameter controls the model's tolerance level towards margin violations. It measures the number of observations that are on the wrong side of the hyperplane by putting a limitation on the sum of slack variables ϵ_i permitted by the model. ϵ_i indicates whether the ith observation is on the wrong side of the hyperplane ($\epsilon_i > 1$), violates the margin ($\epsilon_i > 0$), or is on the correct side of the margin ($\epsilon_i = 0$). The cost parameter C can also be interpreted as a parameter that controls the bias and variance in the model. When it is large, there are wide margins and the model has a high tolerance for margin violations. When it is small, there are narrow margins and the model has lower tolerance towards margin violations. Cross-validation aims to balance the bias-variance trade off involved (King-Yu, 2024a).

In this study, we tested a variety of values for each parameter to tune them. Table 2 provides a summary of the optimal values for each model. For the linear model, the optimal value for the cost parameter was found to be $C = 0.01$. For the polynomial model, the optimal value

obtained for the cost parameter was $C = 0.4$. Finally, for the radial kernel, we the optimal parameters were $C = 2.5$ for the cost parameter and $\gamma = 0.03$ for the gamma parameter. After performing cross-validation, we used the optimal values to fit the models with the test set.

Table 2: Optimal values from cross-validation

Kernel	Optimal Cost	Optimal Gamma
Linear	0.01	N/A
Polynomial	0.40	N/A
Radial	2.50	0.03

The second method we used was random forest, an ensemble learning technique that involves splitting the predictor variables into subsets to form multiple trees. The final model from a random forest would average out the results from each of the individual trees produced and develop a more reproducible and reliable result (King-Yu, 2024c). In this study, we performed 5-fold cross validation to tune the number of trees and variables randomly sampled at each split. We tested tree sizes of 1 to 80 and obtained 43 as the optimal tree size. We also tested variable subset sizes of 1 to 10 and obtained 10 as the optimal size.

Results

Table 3 below summarizes the results of the SVM and random forest. The adjusted Rand index (ARI) is somewhat low for all four models. The SVM model with the radial kernel produced the highest ARI followed by the random forest, polynomial kernel, and linear kernel models.

The misclassification error rate follows a somewhat similar pattern. As with the ARI, the SVM model with the radial kernel produced the best results among the four models. The error rate for the random forest is slightly worse than the polynomial kernel model. Once again, the linear kernel model has the poorest performance.

Table 3: Performance comparison

Method	Adjusted Rand Index	Misclassification Error Rate
SVM Linear Kernel	0.05051	0.14222
SVM Polynomial Kernel	0.24689	0.12165
SVM Radial Kernel	0.32621	0.11628
Random Forest	0.28880	0.13417

Based on these measures, the radial kernel SVM model has the best performance among the four models produced in this study. However, the random forest model is also valuable as it provides insight on the most important factors associated with a customer accepting an offer from a marketing campaign. As shown in Figure 3, the most important factors are the number of days since the last purchase, amount spent on meat products in the last two years, date the customer enrolled with the company, and the amount spent on wine in the last two years.

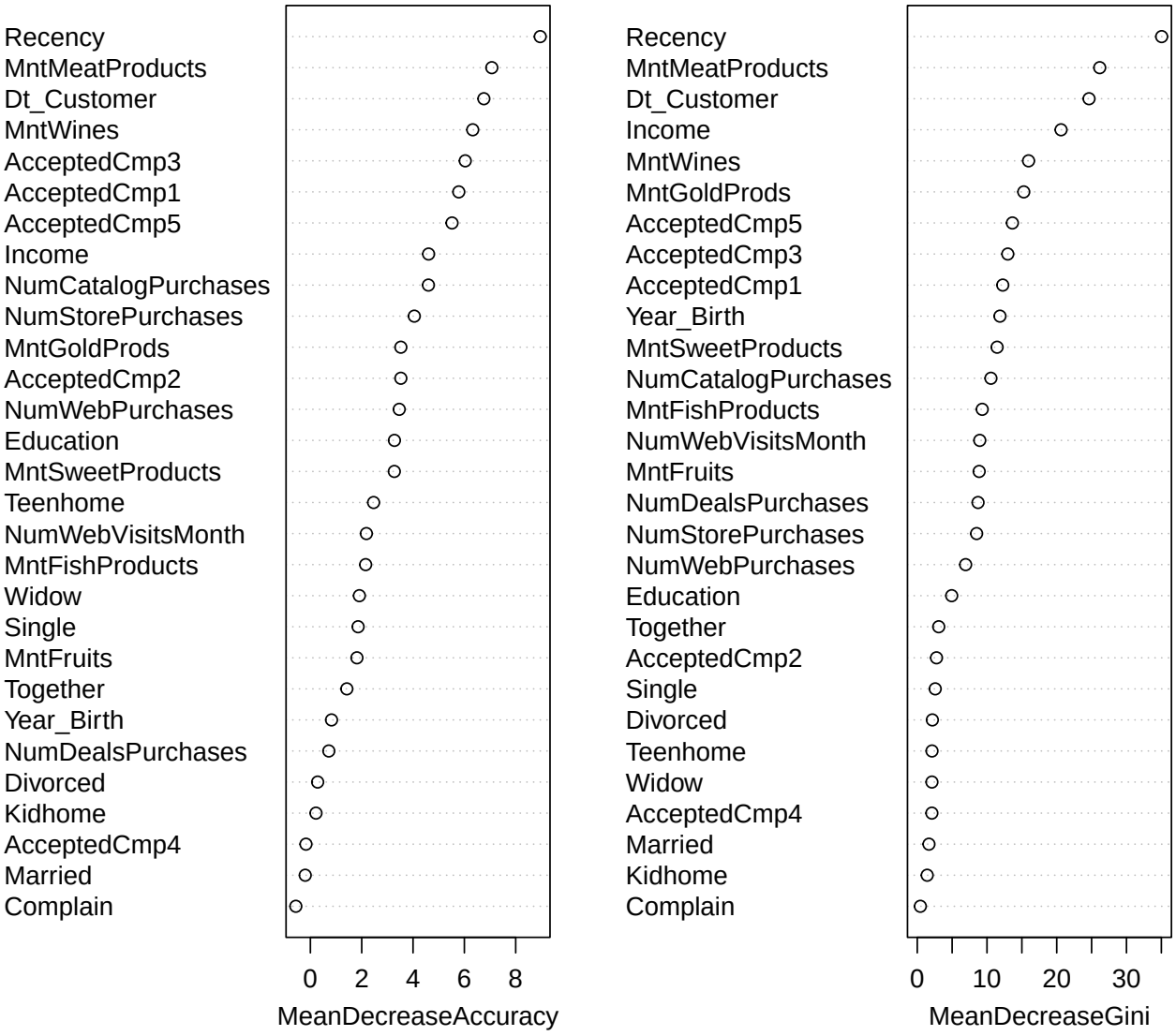

Figure 3: Important factors associated with accepting an offer from a marketing campaign

Conclusion

Overall, the results of our study show four different ways to model the relationship between customer personality and their response to marketing campaigns. Based on the ARIs and misclassification error rates, all of the models have a decent performance. Since the radial kernel model performed the best, we would recommend using this model to help businesses better understand their customers and make changes to their products to increase campaign results.

To improve the accuracy and interpretability of these models, future studies may want to consult subject matter experts to better understand the ARI rates for this type of data. In addition, cleaning the data more extensively, carrying out feature engineering, and performing more fine-tuning may help improve model performance.

References

King-Yu, S. (2024a). *Statistical learning lecture 13: Support vector machines i.*

King-Yu, S. (2024b). *Statistical learning lecture 14: Support vector machines II.*

King-Yu, S. (2024c). *Statistical learning lecture 8: Ensemble learning part III random forests.*

Patel, A. (2021). *Customer personality analysis.* Kaggle. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data