Customer Personality Data Analysis

Team: BigFour
Jingyue Huang (jh8522) Meiyu Li (ml8457)
Yuwen Shen (ys3344) Zehui Gu (zg745)

Introduction

Nowadays, due to the more and more fierce competition in the market, it is increasingly important for companies to analyze the personality and behavior of their customers so that they can modify the products according to the market needs and target the customers that will bring the most benefits to them. For instance, it is costly to advertise a new product to everyone, so the company can use the analysis of customers to advertise only to groups that are more likely to make a purchase or accept offers in a market campaign.

In this project, we aim to analyze the factors that influence customer behaviors and predict the customers' total amount of spending and whether they will accept discount offers. The dataset we use is obtained from Kaggle (Customer Personality Analysis) [1] with 2240 rows and 28 columns. Each row shows the information of one particular customer and the first column represents a unique customer ID. The rest columns contain specific customer features like education, income, the amount spent on several categories of products, and whether they accepted promotions in six campaigns.

Data Preparation
In the original dataset, we find that there are only 24 missing values in the income column, which accounts for just around 1% of the total customers, so we just drop those customers from the dataset. Then we add several columns using the existing ones in the raw data to make the data more informative and interpretable, which would be shown below.

After plotting the boxplots of the features, we find that the 'Age' and 'Income' columns have few large outliers. There are three customers more than 120 years old, but all others are less than 85 years old. Besides, one customer has a yearly income of more than 600,000 but all others earn below 200,000. Since outliers may have a negative impact on our analysis, we remove these outliers and only keep data of customers younger than 85 and earning less than 200,000.

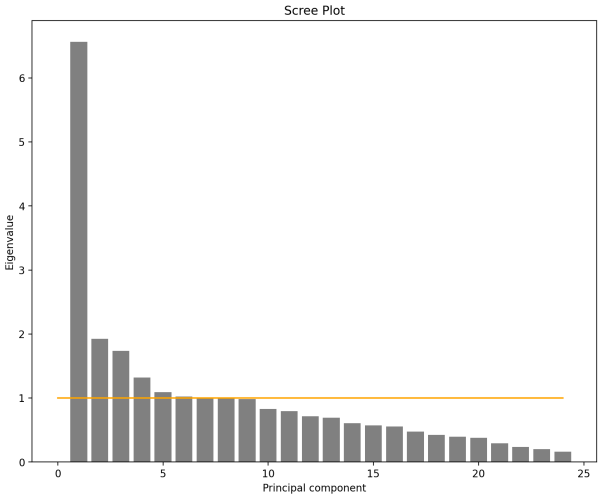


Figure 1: Scree Plot from PCA

Figure 3: Income vs Total Spending after Clustering

Moreover, some features in this dataset are correlated, like the amount spent on different products, so we decide to perform dimension reduction on selected columns of the raw data before doing our analysis. We choose to apply PCA with Kaiser Criterion and the resulting Scree plot (Figure 1) indicates that the first nine principal components play a major role, which explains 69.36% of total variances. Since they are representative of the entire dataset, we use them to divide the customers by the K-means clustering algorithm into two clusters, which is the optimal number of clusters indicated by the highest silhouette score. Also, seeing from the heatmap (Figure 2) that the first principal component is highly correlated with income and total spending, we use the result from clustering to label these features. The scatter plot (Figure 3) clearly shows two groups of customers, and we call them high income with high total spending and low income with low total spending.

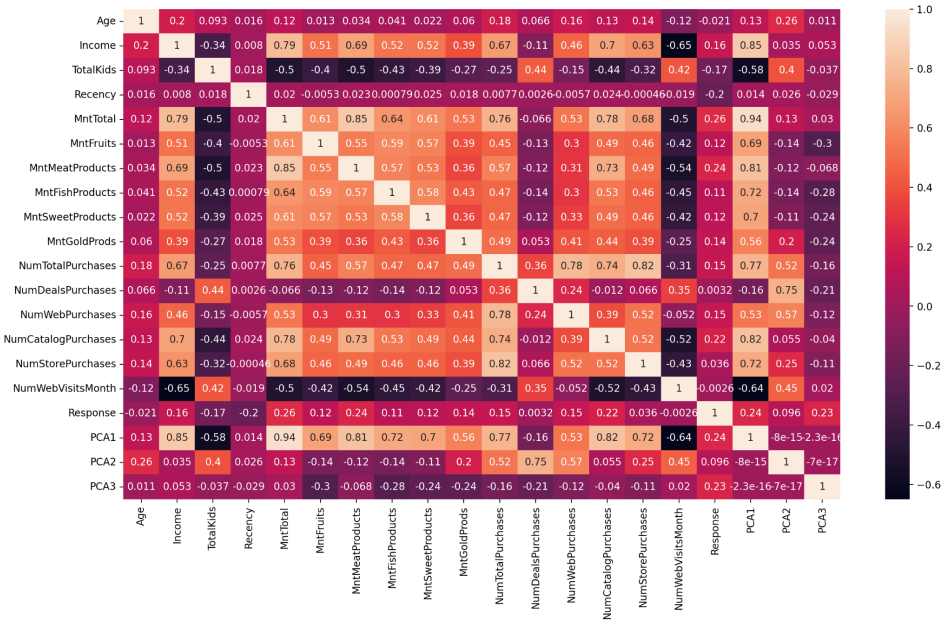


Figure 2: Heatmap of correlation between features and PCA factors

Therefore, after data preparation, we have the following columns:

- 'Age': subtract the 'Year_Birth' from the current year to get the customer's age
- 'Income': customer's yearly household income
- 'Single': obtained from 'Marital_Status', 0 means the customer is together or married, 1 otherwise
- 'Education': 1 if the customer has graduated, 0 otherwise
- 'TotalKids': the total number of kids and teenagers in the customer's home
- 'EnrollmentDuration': number of months from the customer's date of enrollment with the company
- 'Recency': number of days since customer's last purchase
- 'Complain': 1 if the customer complained in the last 2 years, 0 otherwise
- 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds': amount spent on specific categories of products in last 2 years
- 'MntTotal': the total spending by summing up amount spent on different products
- 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases': number of purchases made through the company's website / by catalog / directly at store
- 'NumDealsPurchases': number of purchases made with a discount
- 'NumTotalPurchases': the total number of purchases made from different places
- 'NumWebVisitsMonth': number of visits to company's website in the last month
- 'AcceptedCmp1-5': 1 if customer accepted the offer in each campaign 1-5, 0 otherwise
- 'AcceptNum': the total number of accepted discount offers in the five campaigns
- 'Response': 1 if customer accepted the offer in the last campaign, 0 otherwise
- 'IncomeSpending': 1 means high income and high spending, 0 means low income and low spending

Inference

In this section, we will conduct statistical analysis to explore the impact of both continuous and categorical features on the amount of total spending as well as whether customers accept discount offers in the last campaign.

Question 1. *Is there a difference in the means of continuous features (age, income…) between the group of customers accepting the discount offer and those not accepting, and between the groups of high and low spending customers?*

First, we do the **power analysis** and find the minimum sample size of 40 would result in a power of 0.869. Then, it is easy to split the group of accepting discounts by looking at the Response (0 or 1) variable. To label higher or lower for the total spending, we do a median split. Then, we conduct **Two-Sample T-Tests** for the continuous features, because (1) the data has an approximately normal distribution, (2) each feature's ratio of the variance between each group is less than 1, and (3) customers are independent of each other.

Null Hypothesis: There is no difference in means of continuous features (age, income,...) between different groups

Alternative Hypothesis: There is a difference in means of continuous features (age, income,...) between different groups

	Not Accepted Discount *mean(SD)*	Accepted Discount *mean(SD)*	P-value	Higher Spending *mean(SD)*	Lower Spending *mean(SD)*	P-value
Age	53.2 (11.6)	52.5 (12.3)	**0.346**	55.0 (12.0)	51.2 (11.1)	<0.001
Income	50496.6 (20887.4)	60209.7 (23194.1)	<0.001	67841.2 (14198.3)	36047.6 (14826.2)	<0.001
# Kids	0.5 (0.5)	0.3 (0.5)	<0.001	0.1 (0.4)	0.7 (0.5)	<0.001
# Teens	0.5 (0.5)	0.3 (0.5)	<0.001	0.5 (0.6)	0.5 (0.5)	**0.435**
Total Spending	540.2 (553.2)	985.7 (719.4)	<0.001	-	-	-
Total Purchases	14.4 (7.7)	17.7 (6.9)	<0.001	21.3 (4.6)	8.5 (3.8)	<0.001

Table 1: Means of continuous features comparing in discount accepted or not groups and higher/lower spending groups

Question 2. *Is there a difference in the means of categorical features (income and spending situation, graduated or not, discount campaign 1-5 accepted or not) between the group of customers accepting the discount offer and those not accepting, and between the groups of higher and lower spending customers?*

First, we do the same algorithm as *Question 1* to split the groups. Then, we construct **Chi-Square Tests** for all features when comparing higher spending and lower spending groups. Furthermore, we construct **Fisher's Exact Tests** for discount campaign 2 (whether customers accept 2nd campaign discount) for groups that accepted the final discount offer or not because the frequency of accepting 2nd campaign discount is smaller than 40.

Null Hypothesis: The features (whether graduated, whether accepting 1st campaign offer, …) are independent of discount acceptance groups and higher/lower spending groups.
Alternative Hypothesis: The features (whether graduated, whether accepting 1st campaign offer, …) are not independent of discount acceptance groups or higher/lower spending groups.

		Total	Not Accepted Discount *n(%)*	Accepted Discount n(%)	P-value	Higher Spending *n(%)*	Lower Spending *n(%)*	P-value
Total		2212	1879	333		1107	1105	
High Income High Spending (IncomeSpending)	**Yes**	838 (37.9)	644 (34.3)	194 (58.3)	<0.001	836 (75.5)	2 (0.2)	<0.001
	No	1374 (62.1)	1235 (65.7)	139 (41.7)		271 (24.5)	1103 (99.8)	
Graduated	**Yes**	1115 (50.4)	963 (51.3)	152 (45.6)	**0.112**	566 (51.1)	549 (49.7)	**0.524**
	No	1097 (49.6)	916 (48.7)	181 (54.4)		541 (48.9)	556 (50.3)	
Campaign 1	**Yes**	142 (6.4)	63 (3.4)	79 (23.7)	<0.001	139 (12.6)	3 (0.3)	<0.001
	No	2070 (93.6)	1816 (96.6)	254 (76.3)		968 (87.4)	1102 (99.7)	
Campaign 2	**Yes**	30 (1.4)	10 (0.5)	20 (6.0)	<0.001	28 (2.5)	2 (0.2)	<0.001
	No	2182 (98.6)	1869 (99.5)	313 (94.0)		1079 (97.5)	1103 (99.8)	
Campaign 3	**Yes**	163 (7.4)	86 (4.6)	77 (23.1)	<0.001	89 (8.0)	74 (6.7)	**0.260**
	No	2049 (92.6)	1793 (95.4)	256 (76.9)		1018 (92.0)	1031 (93.3)	
Campaign 4	**Yes**	164 (7.4)	102 (5.4)	62 (18.6)	<0.001	145 (13.1)	19 (1.7)	<0.001
	No	2048 (92.6)	1777 (94.6)	271 (81.4)		962 (86.9)	1086 (98.3)	
Campaign 5	**Yes**	161 (7.3)	70 (3.7)	91 (27.3)	<0.001	161 (14.5)	0	<0.001
	No	2051 (92.7)	1809 (96.3)	242 (72.7)		946 (85.5)	1105 (100.0)	

Table 2: Means of categorical features comparing in discount accepted or not groups and higher/lower spending groups

To sum up, Table 1 and Table 2 indicate income, the number of kids in the customer's family, and total purchases show a significant difference both among the final discount offer acceptance groups and among the higher/lower spending groups. Furthermore, the mean age also shows a difference in higher/lower spending groups. The 5 discount campaigns and whether the customer is lower-income-and-lower-spending impose an impact on customers' spending on goods and their final decision about whether to accept the final discount offer.

Prediction

In this section, we will use regularized regression techniques to predict the total amount of spending based on a set of features we've explored in the Inference Section.

Question 3: *To what extent can we use income and the number of total purchases of a customer to predict the total amount of spending?*

According to the correlation heatmap from data preparation and hypothesis testing results from the Inference Section, we select a set of features that could serve as potential predictors of total amount of spending, build simple linear regression models for each of them and output the following table with each one of the potential predictors with their associated COD values and beta coefficients. We also use a random forest regressor with the same set of predictors to generate a feature importance plot, as shown in the right figure below.

Predictor	COD	Beta1	Beta0
Income	0.6284	0.0222	-545.5676
NumTotalPurchases	0.5729	59.4434	-277.9339
TotalKids	0.2499	-401.9061	988.0977
AcceptNum	0.2084	404.6984	486.5172
Response	0.0699	445.4520	540.2086
EnrollmentDuration	0.0190	10.7016	-491.6583
Age	0.0134	5.9677	290.4639
Complain	0.0012	-217.2322	609.2322
Education	0.0006	30.0853	592.1030
Single	0.0004	26.3567	597.9265
Recency	0.0004	0.4263	586.3701

Figure 4: COD and Betas of Features From Simple Linear Regression

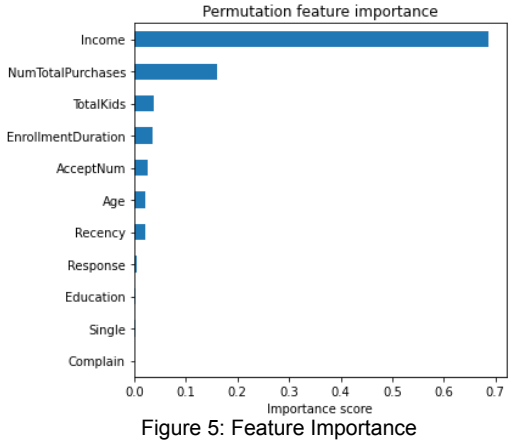


Figure 5: Feature Importance

The feature with the highest COD value (0.628) is Income, indicating that this feature alone could account for 62.8% of the variance in the response variable. Notice that COD values have a distinct cutoff around 0.50, and two features above the cutoff are income and total number of purchases. The feature importance of these two variables can also be shown in the right figure above, and it's consistent with what we find in the Inference Section that they both show a significant difference among higher/lower spending groups. Therefore, we choose these two features, income and the number of total purchases, to predict the total amount of spending.

After we partition our data using an 80/20 split, we build a multiple regression model and obtain the following scatterplot of predictions and actual values of spending with an orange identity line. Its R^2 is 0.691, with a RMSE of 332.654. Beta coefficients of income and total purchases are 0.0149, 31.4983, with an intercept of -632.6630.

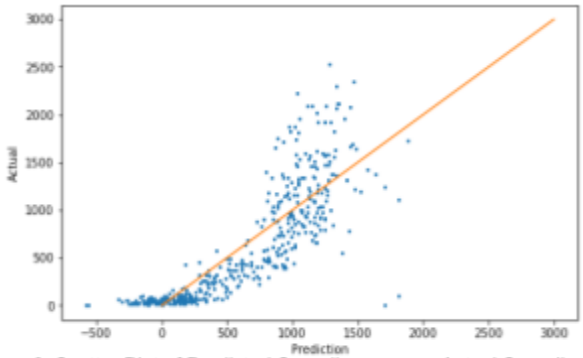


Figure 6: Scatter Plot of Predicted Spendings versus Actual Spendings

We also try ridge and lasso regularized regression to see whether there's any performance improvement of our model. For both Ridge and Lasso, we use grid search with cross validation to find the best hyperparameter and obtain the following results:

Ridge: best α=199, RMSE = 332.692, R^2 = 0.691, betas = [0.0149, 31.387], with intercept = -632.422
Lasso: best α=2,RMSE= 332.676, R^2 = 0.691, betas = [0.0149, 31.435], with intercept = -632.527

Notice that there's no huge difference between RMSE and R^2 values among models, and it makes sense because we don't have too many independent variables in our model. To conclude, beta coefficients of 0.0149, 31.4983 can be interpreted as:
1) The total spending will increase by 1.49 when income increases by 100 with other factors controlled, i.e. for an additional income of 100, the customer is willing to use 1.49% of them as budget to purchase goods at this store.
2) The total spending will increase by 31.50 when an additional purchase is made with other factors controlled, i.e. average additional transaction value is 31.50.

Classification

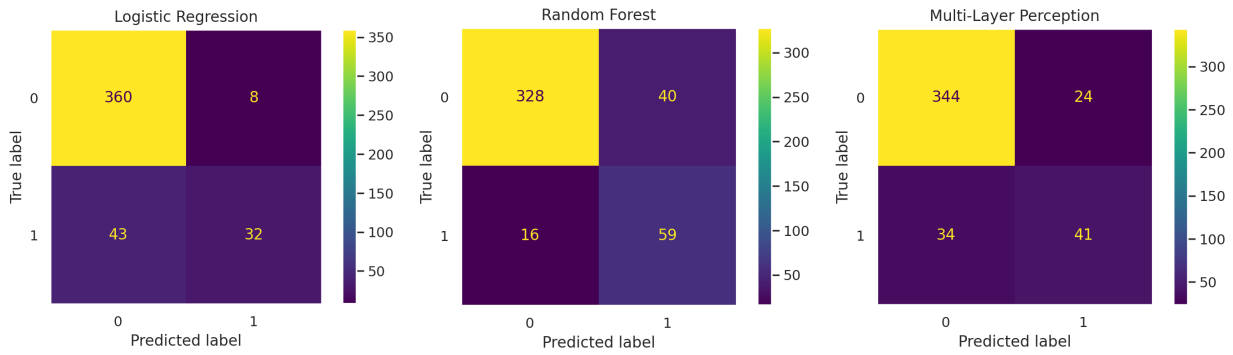
In this section, we will explore several classification methods to classify customers into two categories according to the Response variable, using other variables as predictors.

Question 4. *What is the best fit to predict whether customers accept discount offers in the last campaign?*

It's common that customers who accept offers in a market campaign are in the minority. In our dataset, only 15.1% of customers responded to the last campaign, which means that this is an imbalanced classification problem. A suitable model that could effectively detect these small proportions of customers who are more likely to accept offers would be beneficial to marketing strategy design.

Based on the analysis in the inference section, there exist strong correlations between customer personality and their behavior of accepting discounts. Their interaction records 'AcceptedCmp1-5' also play a role in predicting their future purchases. Therefore, we consider all the other variables described above to predict the 'Response' variable. To reduce the high dimensionality of features, we choose the variable 'IncomeSpending' through dimension reduction and clustering to represent income-spending situations of customers, instead of using raw variables related to income and spending. All of the available data (18 independent variables and 1 dependent variable) is split as 80% training data and 20% test data to make classification. Moreover, besides Accuracy, we use F1 Score and AUC as additional evaluation metrics for this imbalanced problem.

To further avoid over-fitting, we apply Lasso Regression to select significant features. Specifically, we fit a logistic regression model with L1 penalty to find a sparse solution of coefficients while making classification. The best hyperparameter 'C' (inverse of regularization strength) is found by cross-validation. The best F1-Score and AUC are achieved with C = 0.91 and 2 variables with zero coefficients are removed, which results in 16 independent variables. Figure 7(a) shows the confusion matrix of the Logistic Regression model. The high accuracy of 0.885 is achieved by classifying most data into the majority class, i.e., no response.



(a) (b) (c)

Figure 7: Confusion matrix of three classification models

To improve the classification performance, we try two more kinds of models that are robust to the imbalance problem, Random Forest and Multi-Layer Perceptron (MLP). Similarly, we use the 16 variables obtained from feature selection as predictors and perform hyperparameter tuning by cross-validation. The classification results are shown in Table 3.

Model	Hyperparameter	Accuracy	F1-Score	AUC
Logistic Regression	C = 0.97	**0.885**	0.557	0.882
Random Forest	n_estimators = 150 class_weight = balanced_subsample	0.874	**0.678**	**0.904**
Multi-Layer Perceptron	hidden_layer_sizes=(10, 2) alpha = 0.218	0.869	0.586	0.853

Table 3: Classification performance of three models

From the results above, although the accuracy of Logistic Regression is higher than Random Forest and MLP, the F1-Score and AUC of the two latter models are much better, which indicates their capacity of detecting True Response. Moreover, with "balanced_subsample" mode, Random Forest reaches the best F1-Score of 0.678 among three models by adjusting weights to balance the data. The reason why MLP does not achieve comparable results may be that there is not enough data. The best overall performance is achieved by Random Forest, so it is the best fit to predict customers' responses in future campaigns. Specifically, the feature importance evaluated by mean decrease in impurity (MDI) shows that 'Recency' and 'EnrollmentDuration' are the most important factors for response prediction, which suggests that loyal customers should get more attention at the campaign events.

Conclusion

In summary, data analysis on customer personality allows the company to identify specific groups of people they should target in order to increase sales and promote their products.

- To establish the user persona and user profile, it is significant to consider each customer's total spending, purchases, income, and behavior in the first 5 discount campaigns. For people in different groups, we can use different discount methods by comparing the P-values in the inference table. For example, the company could offer more frequent discount campaigns for people with high incomes.
- Both income and number of total purchases have a positive relationship with total spending for a customer: an additional 100 of income would bring up 1.49 spending and an additional purchase would increase spending by 31.5. So the company could potentially increase customers' spending and thereby its profit by implementing strategies such as upselling / cross-selling, offering bundle deals and tiered rewards, etc.
- Moreover, to improve the effectiveness of future campaigns, companies are suggested to apply the Random Forest to find target customers who are more likely to accept the offers. Especially customers who made a purchase recently and those who enrolled with the company for a long time deserve higher attention.

Limitation
One of the limitations of our prediction analysis is that multiple regression models may not be able to capture complex relationships between predictors and outcomes, particularly if the relationships are nonlinear. What's more, our dataset only records the amount of spending in a fixed interval (last 2 years). In order to more accurately predict the spending, we could incorporate time variables while gathering the dataset, such as collecting weekly/monthly spending for a given interval, as well as external factors like economic condition, which helps to capture the dynamics of a customer's spending behaviors and to predict future spendings.

Extra Credit
One interesting finding is that we observe a negative relationship between the number of kids in a household and total spending, with a coefficient of around -400. It seems counterintuitive as we might have expected that with more children in a household, more money will be spent, which suggests that it may be worth further investigating the validity of our dataset and figuring out the reasons behind this.

References

[1] Customer Personality Analysis. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis.

Student Contribution

PCA and Clustering: Yuwen Shen
Inference: Meiyu Li
Prediction: Zehui Gu
Classification: Jingyue Huang