How and To what extent do costumers' characteristics and sellers' strategies influence costumers spending behaviors?

By YUXI LI, KEXIN CHEN

1. Introduction

Costumers' personality is an essential aspect for firms to consider and analyze, providing a thorough understanding of the ideal customer profiles that a business should target. It helps a company have a better grasp of its customers, making it easier to tailor goods to the unique needs, tastes, and behaviors of different consumer segments. As a result, conducting a personality analysis of a firm's ideal customers makes firms easier to identify prospective customers and formulate more rational marketing strategies. However, precisely analyzing the personalities of these ideal customers and understanding how these characteristics correlate with their spending behaviors present a challenging but crucial task. Previous paper like *Consumer Analysis of Commercial Plant-Based Jerky* indicates that competition between companies in the food sector is becoming increasingly tight, analysis of consumer satisfaction and preference contributes a lot to determine firms' strategy.

Consumer behavior research has identified internal and external factors as the key influences of consumers' purchasing behavior. Internal factors include attitude, beliefs, feelings, lifestyle, motivation, and personality traits, whereas external factors include culture, locality, and the reference group (Sandhusen, 2000). Many previous studies have researched personality in consumer behavior, like brand preferences (Banerjee, 2016), online purchase intentions (Iqbal et al., 2021). But this article aims to explore consumer purchasing behavior from multiple perspectives, examining internal factors such as income, number of children, and purchasing methods, as well as the variety of products (such as sweet and meat) and sellers' promotional tactics.

Our research focus on "How and To what extent do costumers' characteristics and sellers' strategies influence costumers spending behaviors". The purpose of this paper is to analyze the Customer Personality Analysis data by using econometric models to quantify how different consumer characteristics affect their spending behavior and ultimately to obtain some uniform patterns and norms. Based on previous research

papers and common sense, people always assume that households with more incomes will have higher purchasing power, so they should have higher spending on all normal goods, and with more people in the household, the total consumption will be higher in order to satisfy more people's survival needs. In addition, we believe that families with more kids may spend more on sweets, and families with more teenagers may spend more on meat.

For specification check, we use white test to detect heteroscedasticity. Also, for robustness check, we choose an alternative model called ridge regression to compare with the Lasso model to check multicollinearity.

2. Data Description

2.1 Data Overview

The dataset we use in this study is provided by a company named ifood, which is a leading online food ordering and delivery platform based in Brazil. The original source of the dataset can be found on GitHub at https://github.com/nailson/ifood-data-business-analyst-test.The data was collected from a consumer interview in 2020, varying in terms of age, income, and educational background, and the data was not collected by controlling for specific groups of variables which ensures the randomness and authenticity.

This dataset "marketing_campaign" is a sample including 2240 randomly choosing individuals from all the costumers of ifood. It generally includes 26 variables, which can separate into 3 parts: costumers characteristics, company promotion, and costumers' spending on different kinds of products. To analyze how the costumers' characteristics and sellers' behaviors influence costumers spending behaviors, the response variables are costumers' spending behaviors, including the spending on Wines, Fruit, Meat, Fish, Sweet and Gold in the last 2 years. In addition, explanatory variables also include costumers' characteristics and company promotions which imply the sellers' behaviors. The customers can order and acquire products through 3 sales channels: physical stores, catalogs and company's website. Several strategic initiatives are being considered to invert this situation. One is to improve the performance of

marketing activities, with a special focus on marketing campaigns.

People	
ID	Customer's unique identifier
Year_Birth	Customer's birth year
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
Dt_Customer	Date of customer's 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
NumWebPurchases	Number of purchases made through the company's website
NumCatalogPurchases	Number of purchases made using a catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company's website in the last month
Products	
MntWines	Amount spent on wine in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years
Promotion	
NumDealsPurchases	Number of purchases made with a discount
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response	1 if customer accepted the offer in the last campaign, 0 otherwise

TABLE 1: MEANING OF ALL VARIABLES

The dataset features a mix of quantitative and qualitative variables, offering a detailed perspective on consumer behavior. It includes two categorical variables, 'Education' and 'Marital_Status', which denote the educational and marital statuses of consumers, respectively, with 'Education' having five categories and 'Marital_Status' seven. This categorization provides insight into the dataset's demographic diversity. Income emerges as a crucial explanatory variable, directly influencing spending on normal goods; higher income levels correlate with increased purchasing power.

Additionally, family size, indicated by variables such as 'Marital_Status', 'Kidhome', and 'Teenhome', significantly affects spending habits. Larger families require more necessities, altering spending behavior based on family composition and varying product demands. Therefore, the influence of different explanatory variables varies across product types, necessitating further detailed econometric analysis to understand these dynamics better.

2.2 Data Cleaning

In the preliminary phase of our study, we refined the dataset to align with our research goals, removing irrelevant variables such as 'Z_CostContact', 'Z_Revenue', 'ID', 'Year_Birth', and 'Dt_Customer'. To tackle missing data in the 'Income' variable, we adopted a categorical grouping-based imputation approach. By leveraging the categories within 'Education' and 'Marital_Status', we filled missing values with the subgroup's average income, ensuring the imputations were accurate and representative of the potential impacts of educational attainment and marital status on income levels.

Further optimization of our dataset involved the transformation of categorical data from 'Education' and 'Marital_Status' into numerical values, utilizing their relative positions to facilitate their inclusion in quantitative analysis. Through comparative study, we decided to employ lasso's inherent one-hot encoding process for handling categorical variables, maintaining the integrity and independence of the categorical data, which is one of the reasons for our choice of the lasso model. Additionally, after evaluating the correlation matrix, we removed the 'Recency' column due to its minimal correlation with other key variables, thereby focusing our analysis on variables with more significant associations. Moreover, we introduced an aggregated column summarizing responses to five different marketing campaigns, providing a comprehensive view of consumer engagement with these initiatives, as we believe too many dummy variables could impact the model's fit. Lastly, We ultimately selected the sales data of three representative items for our analysis—wine,meat and sweet to simplify the analysis and explore consumer spending behaviors across a wider range of product domains, thereby enhancing the robustness and depth of our analytical insights.

2.3 Summary statistics

	count	mean	std	min	25%	50%	75%	max
Income	2240.0	52248.747768	25039.981164	1730.0	35538.75	51381.5	68289.75	666666.0
Kidhome	2240.0	0.444196	0.538398	0.0	0.00	0.0	1.00	2.0
Teenhome	2240.0	0.506250	0.544538	0.0	0.00	0.0	1.00	2.0
MntWines	2240.0	303.935714	336.597393	0.0	23.75	173.5	504.25	1493.0
MntFruits	2240.0	26.302232	39.773434	0.0	1.00	8.0	33.00	199.0
MntMeatProducts	2240.0	166.950000	225.715373	0.0	16.00	67.0	232.00	1725.0
MntFishProducts	2240.0	37.525446	54.628979	0.0	3.00	12.0	50.00	259.0
MntSweetProducts	2240.0	27.062946	41.280498	0.0	1.00	8.0	33.00	263.0
MntGoldProds	2240.0	44.021875	52.167439	0.0	9.00	24.0	56.00	362.0
NumDealsPurchases	2240.0	2.325000	1.932238	0.0	1.00	2.0	3.00	15.0
NumWebPurchases	2240.0	4.084821	2.778714	0.0	2.00	4.0	6.00	27.0
NumCatalogPurchases	2240.0	2.662054	2.923101	0.0	0.00	2.0	4.00	28.0
NumStorePurchases	2240.0	5.790179	3.250958	0.0	3.00	5.0	8.00	13.0
NumWebVisitsMonth	2240.0	5.316518	2.426645	0.0	3.00	6.0	7.00	20.0
Complain	2240.0	0.009375	0.096391	0.0	0.00	0.0	0.00	1.0
Response	2240.0	0.149107	0.356274	0.0	0.00	0.0	0.00	1.0
TotalAcceptedCmps	2240.0	0.297768	0.678381	0.0	0.00	0.0	0.00	4.0

FIGURE 1: SUMMARY OF THE STATISTICS

Figure 1 above is the statistical summary of the marketing_campaign raw dataset. Taking the variable Income which implies the yearly household income as an example, we lost the data of 24 individuals, so it just counts for 2216; it has the mean of 52,247, with the highest value reaching 66,666, but the lowest value just 1730; it is noticeable that there is a large variation since the standard deviation is 25173, which is half of the mean. Figure 1 also show that, from an average point of view, spending on alcohol and meat are major part of costumers' spending at this company because the average spending on meat and alcohol is significantly higher than the expenditure on other categories.

3. Method and Model

After comparing various models, we chose the lasso model for our dataset, which contains not only numerical variables but also categorical variables such as education and marital status. The lasso model is better suited for handling categorical variables and selecting relevant features based on these variables, making it more appropriate for our dataset than linear regression. LASSO regression model has the function of

determine whether the explanatory variable is correlated or uncorrelated with the independent variables. The following equation show the LASSO regression equation used in this article.

$$\boldsymbol{Y_n = \beta_{0i} + \beta_{1i} D_{1i} + \beta_{2i} D_{2i} + \cdots + \beta_{ki} X_{ki} + \lambda * (|\beta_{1i}| + |\beta_{2i}| + \cdots + |\beta_{ki}|) + \varepsilon_i}$$

Compared to general OSL Model, LASSO (Least Absolute Shrinkage and Selection Operator) regression, a shrinkage and variable selection method for regression models, is an attractive option as it addresses both the overfitting and overestimation of how well the model performs in terms of using the included variables to explain the observed variability ('optimism bias') problems. LASSO regression aims to identify the variables and corresponding regression coefficients that lead to a model that minimizes the prediction error. This is achieved by imposing a constraint on the model parameters, which 'shrinks' the regression coefficients towards zero, that is by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value (λ)(Robert,1996).

By our LASSO regression equation above, $\boldsymbol{Y_n}$ is the dependent variable (target) including spending on Wines, Meat, Sweet. $\boldsymbol{\beta_{0i}, \beta_{1i},, \beta_{2i}, \ldots \beta_{ki}}$ are the coefficients (parameters) to be estimated $\boldsymbol{X_{1i}, X_{2n}, \ldots, X_{ki}}$ are the explanatory variables (features). $\boldsymbol{\varepsilon_i}$ represents the residual or we say the error term for this model. $\boldsymbol{\lambda}$ is the regularization parameter that controls the amount of regularization applied. LASSO regression model is quite suitable for our dataset as it includes ample prospective explanatory variables since too many variables will cause a high error. LASSO regression model will fit the whole dataset including the testing set well, especially when we considering a high variance dataset. It will help us to eliminate the effect of the unrelated response variables for different kinds of products as well since it can make the coefficient of those variables equal to zero. After applying the LASSO regression model, we will find the coefficients for each explanatory variable to different response variables. The coefficient $\boldsymbol{\beta_{0i}, \beta_{1i},, \beta_{2i}, \ldots \beta_{ki}}$ of the explanatory variables quantify the effect of them to the spending on different kinds of goods. The higher absolute value the coefficient is, the stronger effect the corresponding explanatory variables has on the spendings. The sign of the coefficients implies the negative and positive correlation

between the explanatory and independent variables. If the coefficient automatically turns to zero by our LASSO regression model, this means the explanatory and independent variables are uncorrelated. Thus, by analyzing the values of those coefficients, we can have several insights of how the costumers' characteristics and sellers' behaviors will influence the costumers' spending behaviors.

4. Results

Lasso regression for wine:

The first Lasso regression measures the linear relationship between all the explanatory variables ($\boldsymbol{X_{ki}}$ and $\boldsymbol{D_{ji}}$) and the spending in wine ($\boldsymbol{Y_1}$).

$$\begin{aligned} \boldsymbol{Y_1} = \mathbf{260.532} &- \mathbf{34.314}\boldsymbol{D_{1i}} + \mathbf{61.080}\boldsymbol{D_{2i}} + \mathbf{80.476}\boldsymbol{D_{3i}} - \mathbf{0.284}\boldsymbol{D_{4i}} + \mathbf{8.052}\boldsymbol{D_{5i}} + \mathbf{9.797}\boldsymbol{D_{6i}} \\ &+ \mathbf{36.954}\boldsymbol{X_{1i}} - \mathbf{27.326}\boldsymbol{X_{2i}} - \mathbf{9.877}\boldsymbol{X_{3i}} + \mathbf{50.371}\boldsymbol{X_{4i}} + \mathbf{89.819}\boldsymbol{X_{5i}} \\ &+ \mathbf{106.915}\boldsymbol{X_{6i}} + \mathbf{39.224}\boldsymbol{X_{7i}} + \mathbf{100.585}\boldsymbol{X_{8i}} + \mathbf{0.707} \\ &* (|\boldsymbol{\beta_{1i}}| + |\boldsymbol{\beta_{2i}}| + \cdots + |\boldsymbol{\beta_{ki}}|) \end{aligned}$$

Table 2 presents the results of a Lasso regression model analyzing wine spending. This model reduces coefficients of uncorrelated dummy variables such as Edu_graduation, Mari_alone, Teenhome, etc. to zero, indicating their non-impact on wine purchases. Additionally, we exclude complain_1 and response_1 from our models, utilizing complain_0 and response_0 instead to avoid redundancy, which represent whether a customer complained or accepted the last campaign offer, respectively.

Dummy variables in the model show the influence of customer characteristics on wine spending. For instance, customers with 2n cycle education level spend on average $34.314 less on wine, whereas those with a PhD education level spend $84.314 more. Quantitative variables with positive coefficients suggest a positive impact on wine spending; each unit increase in income and in-store purchases increases spending by $36.954 and $106.915, respectively. The higher absolute coefficients the explanatory variables have, the stronger impact they have to the spending on wine. Conversely, with each additional child reducing spending by $27.327, this is consistent with our previous prediction that number of kids home has a negative correlation with the spending on wine.

The model's R-squared value of approximately 0.668 indicates that 66.8% of the variation in wine spending can be explained by the analyzed characteristics. This aligns with the hypothesis that higher income positively correlates with spending on normal goods, and that larger families spend less on wine. The significance of the retained coefficients in Lasso regression negates the need for measuring p-values and F-statistics, as the model inherently dismisses non-significant variables.

Lasso regression for meat

$$Y_2 = 186.500 + 9.408D_{1i} + 5.990D_{2i} + 13.557D_{3i} - 29.783D_{4i} + 35.056X_{1i} - 9.352X_{2i} - 43.090X_{3i} + 3.115X_{4i} + 4.117X_{5i} + 108.282X_{6i} + 13.682X_{7i} - 31.076X_{8i} + 4.299X_{9i} + 1.818 * (|\beta_{1i}| + |\beta_{2i}| + \cdots + |\beta_{ki}|)$$

The second Lasso regression measures the linear relationship between all the explanatory variables (X_{ki} and D_{ji}) and the spending on meat (Y_2).

In general, the lasso regression makes the coefficients of the dummies other than Edu_graduation, Mari_single and Response_0 to zero which means they are uncorrelated with the spending on meat. For the dummy variables, compare with the first Lasso regression model, the PhD and 2nd cycle education level have no effect on the spending on meat, but the graduation education level play a role. Customers with the marital status of Together consume more wine, while those with the marital status of Single consume more on meat. Very similar to the first model is the effect of income on spending, but very different in that number of purchase costumers made in store is not as strongly positively correlated with customers' spending on meat, one unit increase on number of purchase costumers made in store, the average spending on meat only increase by $13.682. One contradiction between the result and our initial expectation is the coefficient of the explanatory variable Teenhome. We usually think that families with more teenagers may buy more meat to meet the growing needs of teenagers, but our results show that number of teenagers at home has a strong negative relationship with spending on meat, one unit increase in number of teenagers, the average spending on meat decrease by $43.090. The R square for this model is around 0.604 which means 60.4% of the variation in the spending on meat can be explained by

customers' characteristics and sellers' strategies. This is a little lower than the first Lasso regression.

Lasso regression for sweet

$$Y_3 = 33.286 + 3.763D_{1i} - 9.445D_{2i} - 12.830D_{3i} + 0.842D_{4i} - 0.059D_{5i} - 1.479D_{6i} + 5.285D_{7i} - 2.538D_{8i} + 3.241X_{1i} - 1.335X_{2i} - 4.794X_{3i} - 1.636X_{4i} + 8.595X_{5i} + 7.504X_{6i} + 5.705X_{7i} - 6.965X_{8i} - 1.017X_{9i} + 0.101 * (|\beta_{1i}| + |\beta_{2i}| + \cdots + |\beta_{ki}|)$$

The last Lasso regression measures the linear relationship between all the explanatory variables (X_{ki} and D_{ji}) and the spending on sweet (Y_3). In general, the lasso regression makes the coefficients of Edu_graduation, Edu_basic, Mari_absurd, Mari_alone, Mari_single, Mari_Yolo, complain_0 which means they are uncorrelated with the spending on sweet.

By comparing the above three regression results, we find that education and marital status don't seem to have a very stable relationship to spending on different goods, it's just that different groups may react differently to different goods. We also find that half of the customers' spending is negatively related to the number of website visits, but wine is an exception. We previously thought that the more kids a household has, the more it spends on sweet, but our regression result tells us otherwise. Kidhome actually has a negative correlation with spending on sweet, one unit increase in the number of kids at home, the average spending on sweet decrease by $1.335. In general, the coefficient on this regression model is smaller than the previous two, showing that people's total spending on sweet is the smallest of the three products. The R square for this model is around 0.394 which means 39.4% of the variation in the spending on sweet can be explained by customers' characteristics and sellers' strategies. This is much lower than the first and second regression model.

5. Discussion

5.1 specification check

For specification check, we use white test. The white test is a statistical test used to detect the presence of heteroscedasticity in a linear regression model. The linear

regression model generally assumes that variance of residuals remains constant across different levels of explanatory variables. However, not all datasets have this property, so we need the white test to check. We take the lasso regression model for wine as an example. The equation goes:

$$\boldsymbol{\varepsilon_i}^2 = \boldsymbol{\theta_0} + \boldsymbol{\theta_1}\widehat{\boldsymbol{Y}}_\mathbf{1} + \boldsymbol{\theta_2}\widehat{\boldsymbol{Y}}_\mathbf{1}^{\,2} + \boldsymbol{\eta_i}$$

We use the white test to see whether the coefficients $\boldsymbol{\theta_1}$and $\boldsymbol{\theta_2}$are significant. By looking at the p-value and comparing to the significant level, we are able to determine the heteroscedasticity. We find that the p-value are about 6.855e-11, 6.848e-08 and 0.0031. These are much smaller than the significant level 0.05 we chosen, so we have enough evidence to say there is a strong relation between the fitted values and the residuals. Therefore, there is evidence of heteroscedasticity in these three lasso regression model.

5.2 Robustness analysis

For robustness checks, we opted to compare the Ridge regression model with the Lasso regression. Given the presence of categorical variables in our data, we decided against using linear regression, despite it being one of the most commonly employed regression techniques. This approach allows us to explore the effectiveness and performance of different regularization methods when dealing with complex datasets that include categorical features.

Lasso and Ridge regression both enhance ordinary least squares regression (OLS) through regularization techniques, improving the model's ability to handle high-dimensional data. By incorporating regularization terms, they limit model complexity, thus reducing the risk of overfitting. In scenarios with highly correlated predictors, both Lasso and Ridge regression offer more robust performance. However, a key difference is that Lasso can perform variable selection, automatically excluding insignificant variables, making it particularly suitable for situations where the number of parameters exceeds the number of samples. In contrast, the Ridge model does not have variable selection capabilities; it retains all variables in the model but reduces their influence. Our analysis revealed that the results from Ridge Regression are essentially similar to

those obtained from the Lasso model in terms of MSE and R-squared values. Therefore, we selected representative wine data for comparison.

Ridge regression for wine

$$\begin{aligned} Y_3 = {} & 256.443 + -60.017D_{1i} - 29.131D_{2i} - 20.304D_{3i} + 45.5202D_{4i} + 63.931D_{5i} \\ & - 55.649D_{6i} + 2.083D_{7i} + 2.083D_{8i} + 17.585D_{9i} + 8.714X_{2i} + 10.591X_{3i} \\ & + 21.326X_{4i} + 4.594X_{5i} - 9.243X_{6i} + 14.518X_{7i} - 14.518X_{7i} + 9.347X_{9i} \\ & - 9.347D_{9i} + 37.277X_{2i} - 27.694X_{2i} - 0.435X_{2i} - 11.232X_{2i} \\ & + 50.715X_{2i} + 91.060X_{2i} + 107.111X_{2i} + 41.875X_{2i} + 101.956X_{2i} + \\ & 4.18 * (|\beta_{1i}| + |\beta_{2i}| + \cdots + |\beta_{ki}|) \end{aligned}$$

Based on the results from the Ridge regression, the R-squared value is essentially the same as that obtained from the Lasso regression (66.8%). This indicates that reducing the coefficients of less important variables to zero in the Lasso model is well-suited for our dataset, reflecting the appropriateness of the model selection. Additionally, the results from the Ridge regression suggest that there is no very strong multicollinearity among the variables.

6. Evaluation

Although we use the LASSO regression model to automatically exclude some uncorrelated explanatory variables to lower the error of our prediction, after applying the data to the model, we find out that the mean squared error is still very high. This probably mean that there is something wrong with our previous model choice or assumption design. By analyzing the original data, we hypothesize that the large mean squared error for wine is due to the extensive variability in wine consumption data. The average value of wine consumption reaches $303, with a standard deviation of 225, which is significantly higher than that for other items.

Our study also has some limitations, particularly in the selection of independent variables, where multicollinearity exists. For instance, the level of education can influence income to some extent, and marital status may affect the number of children. Moreover, at the outset we assumed that customer characteristics and the firm's marketing strategy have a linear relationship with the customer's spending behaviors.

However, nonlinearity could occur, choosing a nonlinear regression model may give us better results. In addition, omitted variables bias could happen, variables such as customers' eating habits, culture and traditions which not included in this dataset can influence customers' consumption on different products.

7. Conclusion

Our project is aimed to find out how and to what extent do costumers' characteristics and sellers' strategies influence costumers spending behaviors. We analyze a sales data for one company. We analyze 3 typical products: wine meat and sweet and build Lasso regression model to help us automatically exclude some of the uncorrelated variables. The results of the three Lasso regression models show the impact of different costumers' behaviors and marketing strategies on the spending on different goods. In conclusion, the negative coefficients show the negative correlations and the larger coefficients in absolute value have stronger effect on the spendings. The results of these analyses are very helpful in helping companies identify target customers and develop effective marketing strategies. Through analysis, we found that some results indeed align with common sense, such as a positive correlation between income and consumption. Our study also uncovered some interesting findings that deviate from common sense, such as consumers who respond to marketing campaigns not necessarily spending more, and families with more children buying less wine.

While our research has made significant strides in applying advanced regression techniques to understand consumer behavior in consumption, the findings also highlight the intricate nature of real-world data and the necessity for continuous model evaluation and adaptation. Further research is encouraged to expand on these preliminary findings and to explore alternative modeling strategies that could address the identified limitations.

REFERENCES

Veniranda, V., & Surya, R. (2022, February 1). Consumer Analysis of Commercial Plant-Based Jerky. IOP Conference Series: Earth and Environmental Science, 998(1), 012059.

Khatri, J., Marín-Morales, J., Moghaddasi, M., Guixeres, J., Giglioli, I. A. C., & Alcañiz, M. (2022). Recognizing Personality Traits Using Consumer Behavior Patterns in a Virtual Retail Store. Frontiers in psychology, 13, 752073.

Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B 1996; 58: 267–288.

Sandhusen, R. L. (2000). Marketing Básico-Série Essencial. Saraiva Educação SA.

Banerjee, S. (2016). Influence of consumer personality, brand personality, and corporate personality on brand preference: An empirical investigation of interaction effect. Asia Pacific Journal of Marketing and Logistics, 28.

Iqbal, M. K., Raza, A., Ahmed, F., Faraz, N. A., & Bhutta, U. S. (2021). Research on influencing mechanism of big five personality traits on customers' online purchase intention: A mediating role of trust. International Journal of Electronic Business, 52–76.

Attribution

For the coding part, Kexin Chen cleaned the data, summarized the statistics and done the Lasso and ridge regression model. Yuxi Li was responsible for the cross validation in choosing lambda and the white test.

For the writing part, in the introduction section, Yuxi Li finished the main part, Kexin Chen made some additions and changes.

In the data description section, Yuxi Li wrote 2.1, the data overview section, Kexin Chen completed the tables and wrote 2.2 2.3.

In the model part, Yuxi Li was the main editor who summarize the Lasso regression model used in this paper. Kexin Chen made some improvements.

In the result section, Yuxi Li summarized the regression results and made the results tables. Kexin Chen helped with the wording and the grammar.

In the discussion and evaluation section, Yuxi Li was responsible for writing the specification (the white test). Kexin Chen was responsible to the robustness analysis (the ridge regression) and the evaluation part. Yuxi Li helped to add more points in the evaluation.

In the conclusion section, Kexin Chen and Yuxi Li clearly summarized the previous sections and answered the initial research question.

Kexin Chen and Yuxi Li both made the final check of essay formatting, grammar and writing.

Appendix

Lasso regression for wine

Dummy	*coefficient*	*Quantitative*	*coefficient*	*R square*	*MSE*	*Lambada (alpha)*
Edu_2n cycle	-34.314	**Income**	36.954	0.6688047	38820.217	0.707
Edu_basic	0	**Kidhome**	-27.326			
Edu_graduation	0	**Teenhome**	0			
Edu_master	61.080	**NumDealsPurch**	-9.877			
Edu_PhD	80.476	**NumWebPurch**	50.371			
Mari_absurd	0	**NumCataogPurch**	89.819			
Mari_alone	0	**NumStorePurch**	106.915			
Mari_divorced	0	**NumWebVisit**	39.224			
Mari_married	-0.284	**TotalAcceptedCmp**	100.585			
Mari_single	0					
Mari_together	8.052	**Constant**	260.532			
Mari_window	0					
Mari_Yolo	0					
Complain_0	0					
Complain_1	0					
Response_0	9.797					
Response_1	-6.847e-14					

Lasso regression for meat

Dummy	*coefficient*	*Quantitative*	*coefficient*	*R square*	*MSE*	*Lambda (alpha)*
Edu_2n cycle	0	**Income**	35.056	0.6041565	17186.436	1.818
Edu_basic	0	**Kidhome**	-9.352			
Edu_graduate	9.408	**Teenhome**	-43.090			
Edu_master	0	**NumDealsPurch**	3.115			
Edu_PhD	0	**NumWebPurch**	4.117			
Mari_absurd	0	**NumCataogPurch**	108.282			
Mari_alone	0	**NumStorePurch**	13.682			
Mari_divorced	0	**NumWebVisit**	-31.076			
Mari_married	0	**TotalAcceptedCmp**	4.299			
Mari_single	5.990					
Mari_together	0	**Constant**	186.500			
Mari_window	0					
Mari_Yolo	0					
Complain_0	0					
Complain_1	0					
Response_0	-29.783					
Response_1	2.819e-14					

Lasso regression for sweet

Dummy	*coefficient*	*Quantitative*	*coefficient*	*R square*	*MSE*	*Lambda (alpha)*
Edu_2 cycle	3.763	**Income**	3.241	0.3936200	1036.220	0.101
Edu_basic	0	**Kidhome**	-1.335			
Edu_graduate	0	**Teenhome**	-4.794			
Edu_master	-9.445	**NumDealsPurch**	-1.636			
Edu_PhD	-12.830	**NumWebPurch**	8.595			
Mari_absurd	0	**NumCataogPurch**	7.504			
Mari_alone	0	**NumStorePurch**	5.705			
Mari_divorced	0.842	**NumWebVisit**	-6.965			
Mari_married	-0.059	**TotalAcceptedCmp**	-1.017			
Mari_single	0					
Mari_together	-1.479	**Constant**	33.286			
Mari_window	5.285					
Mari_Yolo	0					
Complain_0	0					
Complain_1	0					
Response_0	-2.538					
Response_1	5.035e-15					

White test

	Wine	*Meat*	*Sweet*
test statistic	309.605	279.164	219.026
P value	6.855e-11	6.848e-08	0.0031