Analyzing The Impact of Customer Demographics on Product Category Preferences

Arafat Okino Sadiq

November 23, 2024

1. Introduction

- **Background**: Understanding the relationship between customer demographics (e.g., age, income, education) and consumer behavior is critical for businesses. Insights gained can enhance marketing strategies, optimize inventory, and improve customer satisfaction and retention.
- **Objective**: This study aims to identify key demographic factors influencing customer preferences across different product categories, using machine learning models to enhance prediction and offer insights for personalized marketing strategies.

2. Data and Methodology

The data for this study was sourced from Kaggle's *Retail Sales Customer Behavior Analysis* dataset, which originally contained 1 million rows and 78 columns. For this analysis, we randomly sampled 1,000 rows, focusing on demographic data to explore how customer characteristics relate to product category preferences. This approach allowed us to simplify the dataset while retaining essential demographic factors to predict the product category a customer will likely purchase. The variables used for this analysis include:

Variable	Description	Variable Type	Factor Levels
age	Age(in years) of the customer	Numerical	
gender	Gender(Sex of the customer)	Categorical	Male, Female, Other
Income_bracket	Income bracket of the customer	Categorical	Low, Medium, High
Loyalty_program	Whether the customer is part of a loyalty program	Categorical	Yes, No
membership_years	Number of years the customer has been a member	Numerical	
marital_status	Marital status of the customer	Categorical	Single, Divorced, Married
number_of_children	Number of children the customer has	Numerical	
education_level	Education level of the customer	Categorical	High School, Bachelor's, Master's, PhD
occupation	Occupation of the customer	Categorical	Employed, Unemployed, Self-Employed, Retired
Product_category	Product category of the customer	Categorical	Groceries, Furniture, toys, Electronics, Clothing

Table 2.0: Variable Description

We employed various modeling techniques, including multinomial logistic regression, Lasso regression, and Random Forest, to identify and assess the most influential demographic variables. Data preprocessing steps included feature transformation, encoding categorical variables, and variable selection, ultimately aiming to reveal the demographic factors most predictive of customer purchasing behavior.

Data Preparation:
Cleaning: The dataset was pre-processed to ensure no missing values or inconsistencies.
Feature Transformation and Encoding:
After checking for normality within the numerical variables using qq-plots, the Shapiro-Wilk Test, and realizing they were skewed, we applied log transformations to skewed numerical features (age, membership years, and number of children).
Transformed the numerical variable age to an ordinal category:
Ages 18–29 -- "Young Adults", Ages 30–44 -- "Adults", Ages 45–59 -- "Middle Age."
Ages 60–79 -- "Seniors."
Used label encoding for ordinal categorical variables like the age_group, income_bracket, education_level, one-hot encoding to handle categorical variables like marital status, gender, and occupation.

Modeling Approaches:
Multinomial Logistic Regression: Initially implemented to classify product categories based on demographic variables. However, accuracy was low, indicating a need for more advanced modeling. For each product category k, the probability of that category being selected given the predictor variables X = (x1, x2, …, xp) is given by:

$$P(Y = k \mid X) = \frac{\exp\,(\beta_{k0} + \beta_{k1x1} + \beta_{k2x2} + \cdots + \beta_{kpxp})}{\sum_{j=1}^{k} \exp(\beta_{j0} + \beta_{j1x1} + \beta_{j2x2} + \cdots + \beta_{jpxp})}$$

Where:
Y is the categorical outcome (product category in this case).
β_{kj} are the coefficients for each predictor x_j in the category k.
K is the number of product categories.

Lasso Regression: Employed to improve feature selection by shrinking irrelevant coefficients to zero, resulting in a more interpretable model and a refined set of influential demographic factors.
Random Forest: Tested as a more robust model, leveraging ensemble learning to capture complex interactions in the data.
Model Evaluation:
Used cross-validation and a test set to evaluate each model’s performance based on metrics like accuracy and feature importance.

3. Results

In the data preparation stage, we cleared the missing values and prepared a subset from the source data including the factors we are interested in this analysis. The response variable is “Product category” which is categorized into four groups which is categorical. In all, 1000 rows and 10 columns were used for this analysis.

3.1 Exploratory Analysis

This phase involves examining the relationship between the response variable and selected explanatory variables through numerical summaries and graphical visualizations.

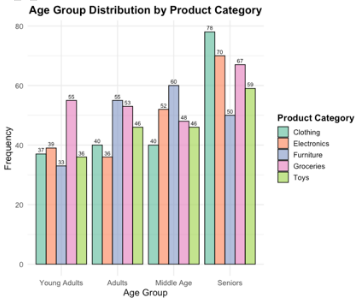
3.1.1 Descriptive Statistics

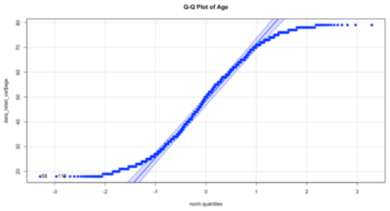
Variable	Mean	Standard Deviation	Min	Max
Age	48.81	18.16	18.00	79.00
Membership Years	4.574	2.84	0.00	9.00
Number of children	2.028	1.39	0.00	4.00

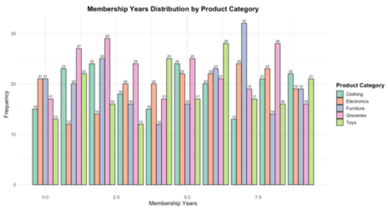
Table 3.0: Summary Statistics

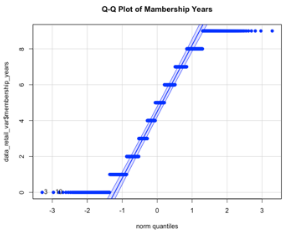
3.1.2 Distribution of Product Category by the numeric demographic variables

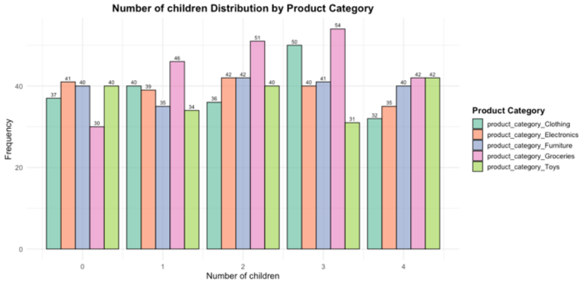
The distributions of product categories based on numeric demographic variables deviate from a normal (bell-shaped) distribution. These variables are inherently discrete and lack the continuous, symmetric shape characteristic of normal distributions. To further analyze these distributions, QQ plots and the Shapiro-Wilk test were utilized. Log transformations were then applied to address skewness and improve the interpretability of the data.

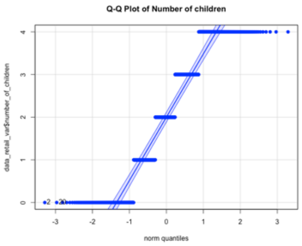










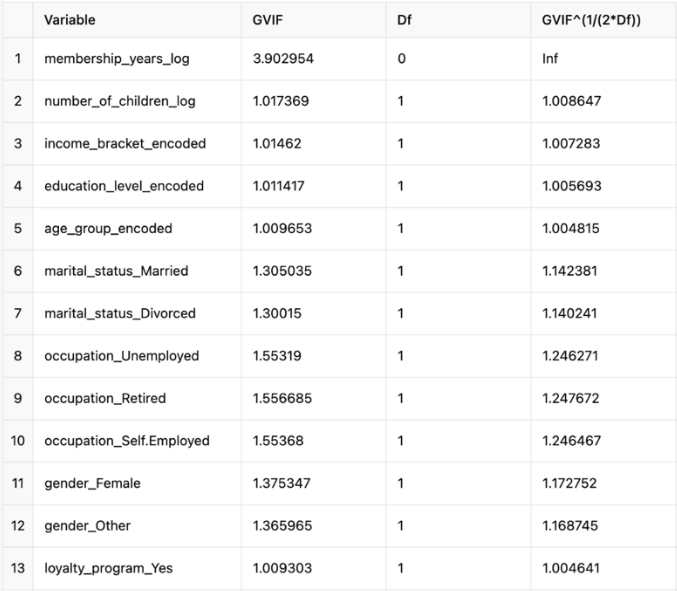


3.1.3 Assumptions and Multicollinearity:

Before proceeding with modeling, key assumptions of linearity, normality, homoscedasticity, and independence were assessed. Normality of continuous predictors was evaluated using statistical tests and visualization techniques such as scatterplots to examine relationships between predictors and product categories.

To assess multicollinearity among the explanatory variables, the Variance Inflation Factor (VIF) test was conducted. As a general rule, VIF values exceeding 4 or 10 indicate problematic multicollinearity. In this analysis, all variables exhibited VIF values below 4, indicating no significant multicollinearity issues.

To enhance model performance, ridge regression was used to address potential multicollinearity, while lasso regression was applied for feature selection by shrinking irrelevant predictors to zero. Cross-validation identified the optimal lambda value for the lasso model, effectively highlighting the most significant demographic predictors influencing product category preferences.



	Variable	GVIF	Df	GVIF^(1/(2*Df))
1	membership_years_log	3.902954	0	Inf
2	number_of_children_log	1.017369	1	1.008647
3	income_bracket_encoded	1.01462	1	1.007283
4	education_level_encoded	1.011417	1	1.005693
5	age_group_encoded	1.009653	1	1.004815
6	marital_status_Married	1.305035	1	1.142381
7	marital_status_Divorced	1.30015	1	1.140241
8	occupation_Unemployed	1.55319	1	1.246271
9	occupation_Retired	1.556685	1	1.247672
10	occupation_Self.Employed	1.55368	1	1.246467
11	gender_Female	1.375347	1	1.172752
12	gender_Other	1.365965	1	1.168745
13	loyalty_program_Yes	1.009303	1	1.004641

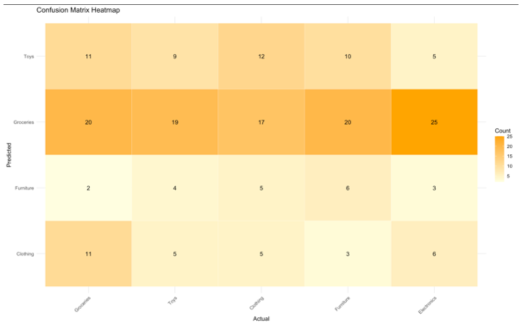
3.1.4 Lasso Regression and Optimal Lambda:

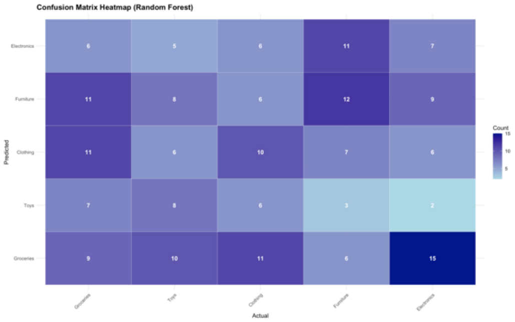
Lasso regression improved feature selection by shrinking less important coefficients to zero, enhancing model interpretability. Cross-validation identified the optimal lambda (λ) value as **0.0217**, balancing model complexity and accuracy. Key demographic predictors selected included income bracket, age group, education level, marital status, employment status, gender, and loyalty program membership. While the model streamlined features, its predictive accuracy was limited, suggesting demographic data alone may not fully capture customer preferences.

3.2 Model Performance

Model Summary	Multinomial Logistic Model Regularized(Lasso)	Random forest
Accuracy	**21%**	**23%**

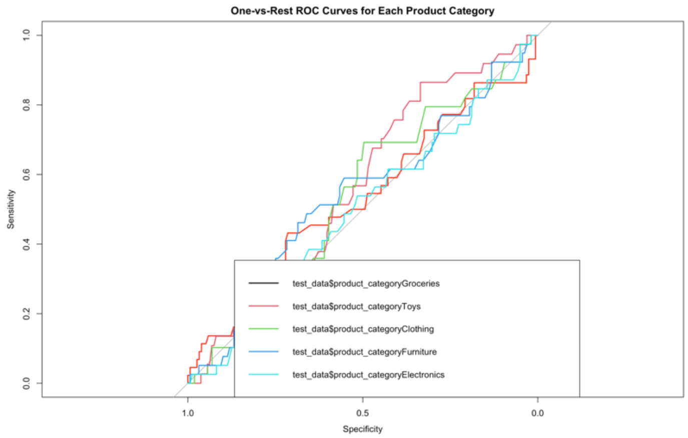
- **Insights**: Demographic variables alone showed limited predictive power.
- **Key Predictors**: Membership years, income bracket, age_group, income_bracket, and education level.





- **Confusion Matrices**: Highlighted misclassifications, particularly between Groceries and Toys.

3.4 ROC Analysis



The one-vs-rest ROC curves showed moderate separability for product categories, with most curves hovering near the diagonal, indicating limited model performance.

4. Conclusion

The analysis highlights the limitations of using demographic data alone for predicting customer product category preferences. While Lasso and Random Forest models identified key demographic predictors, their predictive accuracies remained modest. These findings suggest that businesses should integrate behavioral and transactional data for enhanced predictive performance and customer segmentation.

5. References

1. Kaggle: Retail Sales Customer Behavior Analysis Dataset. https://www.kaggle.com/datasets/utkalk/large-retail-data-set-for-eda
2. Dr. Ruvini Jayamaha. “STAT 7210 /Section 1 - Applied Regression Analysis(Fall 2024)” https://kennesaw.view.usg.edu/d2l/le/content/3287712/viewContent/52032456/View
3. Friedman, Hastie, Tibshirani. **"The Elements of Statistical Learning"** for Lasso and Regularization Techniques.https://link.springer.com/book/10.1007/978-0-387-84858-7
4. Breiman, L. **"Random Forests"** for ensemble learning. https://www.researchgate.net/publication/236952762_Random_Forests
5. Mike Gholi. “Machine Learning: A View of Trends in Customer Behavior” https://www.academia.edu/49350870/Machine_Learning_A_View_of_Trends_in_Customer_Behavior?email_work_card=view-paper