Customer Segmentation – A Classification Problem with Logistic Regression, Support Vector Machine (SVM) and K-Means Clustering

Peiying Hu
phu46@wisc.edu

1. Introduction

Customer personality analysis involves a detailed examination of a company's ideal customers, enabling businesses to understand their clientele better and tailor products to meet the specific needs, behaviors, and concerns of various customer segments. This targeted approach allows companies to allocate marketing resources more efficiently by focusing on customer segments most likely to purchase a product, rather than marketing indiscriminately to all customers. Such analysis is crucial for developing effective marketing strategies and enhancing customer satisfaction. For instance, research has shown that incorporating personality traits into user-centered modeling can provide valuable insights for improving user experiences in complex systems (Mahmood et al., 2016). Our study, aiming to identify the target customer segment for a specific product based on customer characteristics, used Logistic Regression, Support Vector Machine (SVM) and K-Means Clustering to solve this classification problem. Figure 1 is an overview of the project logic.

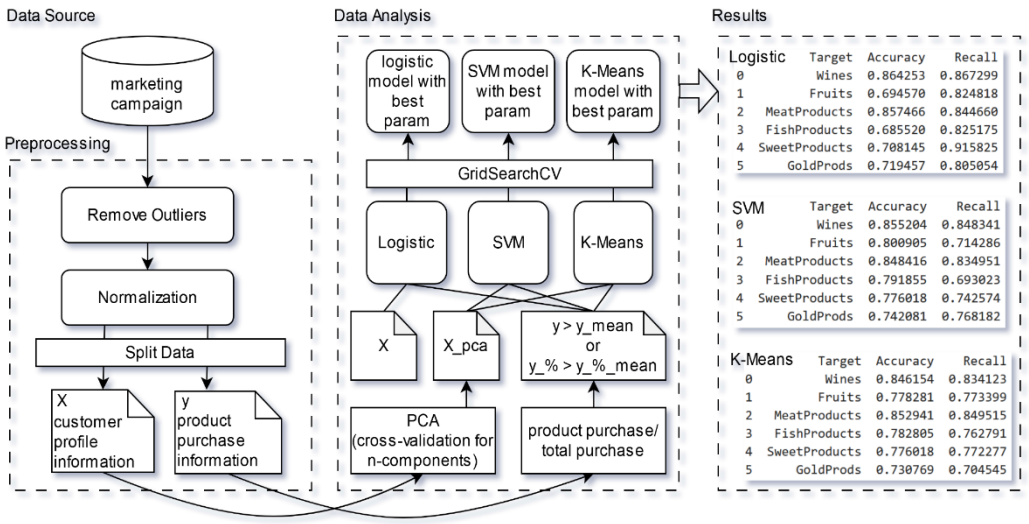

Figure 1: Overview of Project Logic

2. Data and Preprocessing

2.1 Data Cleaning and Organizing

The original dataset we used in this project is Customer Personality Analysis Dataset from a marketing campaign on Kaggle, provided by Dr. Omar Romero-Hernandez. It is a real-world cross-sectional medium-size dataset with 29 features and 2240 observations. The transactions come from a franchise network, uploaded in 2021 covering information from July 2012 to November 2014. The dataset includes information about customer profile, product purchase, promotion and place. Place was later integrated into the customer profile since it indicates the purchasing methods of the customer. After merging, the customer profile category contains basic personal information like year of birth, education level, marital status, income, purchase method, etc. Product purchase category includes the amount of money spent on wines, fruits, meat, fish, sweet products and gold products in the last 2 years. Promotion category includes information about whether the customer accepted a certain discount offer or not. Table 1 shows the specific meanings of each feature.

Customer Profile Information	Product Purchase Information
ID: Customer's unique identifier	**MntWines**: Amount spent on wine in last 2 years
Year_Birth: Customer's birth year	**MntFruits**: Amount spent on fruits in last 2 years
Education: Customer's education level	**MntMeatProducts**: Amount spent on meat in 2 years
Marital_Status: Marital status	**MntFishProducts**: Amount spent on fish in 2 years
Income: Yearly household income	**MntSweetProducts**: Amount spent on sweets in 2 years
Kidhome: Number of children in household	**MntGoldProds**: Amount spent on gold in 2 years
Teenhome: Number of teenagers in household	**NumDealsPurchases**: Purchases made with discount
Dt_Customer: Enrollment date	**Promotion Data**
Recency: Days since last purchase	**AcceptedCmp1**: 1 if accepted offer in 1st campaign
Complain: 1 if complained in 2 years	**AcceptedCmp2**: 1 if accepted offer in 2nd campaign
NumWebPurchases: Purchases via website	**AcceptedCmp3**: 1 if accepted offer in 3rd campaign
NumCatalogPurchases: Purchases via catalogue	**AcceptedCmp4**: 1 if accepted offer in 4th campaign
NumStorePurchases: Purchases in stores	**AcceptedCmp5**: 1 if accepted offer in 5th campaign
NumWebVisitsMonth: Visits to website last month	**Response**: 1 if accepted offer in last campaign

Table 1: Categories and Specific Meaning of Features

We only used the first two categories of features in our analysis because our main interest is studying customers' preference for products based on customer profile information. The promotion data, however, reveals more about preferences for discount which is less of our concern.

The next thing we did was processing the data to a more useful form. We converted year of birth into age; label encoded education into continuous levels from basic, graduation, 2nd cycle /Master to PhD; turned marital status into three new categories, current marital status, divorce history and romantic history; added kid and teen numbers together; changed date of enrollment into days as a customer; transferred numbers of purchases through different methods into proportion of total purchase number.

When looking at the sns scatter plot for data distribution in Figure 2, we noticed some outliers. For example, Age and Income seemed to present normal distributions, but there were some obvious

outliers like age 128 and income 666666 which needed to be dealt with. The plot looked more reasonable after the abnormal values were filtered.

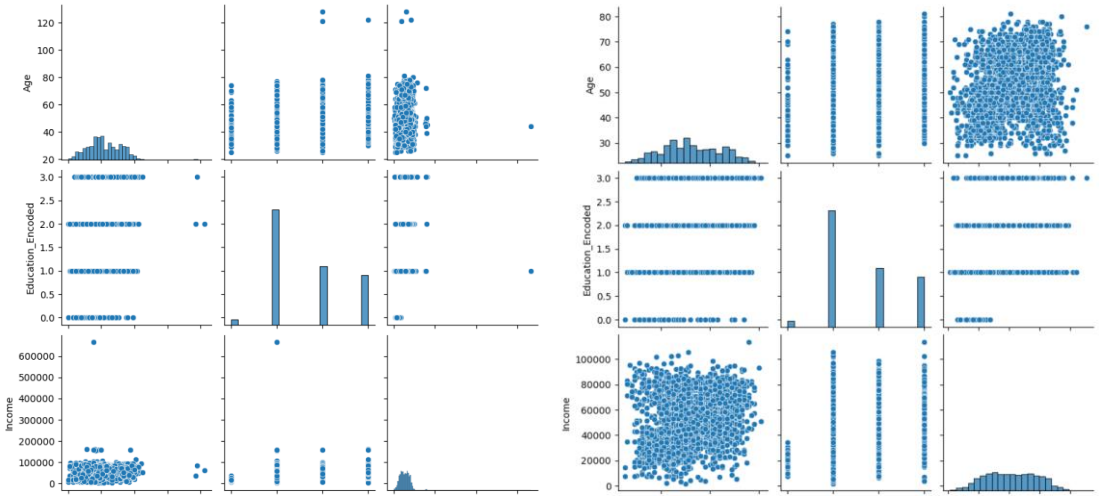

Figure 2: Before and After Outlier Filtering

2.2 Normalization

To better understand customer purchase intent, it is important to recognize the context in which the original data was collected. The dataset records the monetary expenditure of customers on specific product categories, such as “MntWines”. However, since the data reflects purchases made within the sales network of a single store brand, it is not necessarily indicative of the customer’s overall consumption preference for these products. The customer’s spending at this store brand may represent only a portion of their total consumption behavior. To address this limitation, we transformed the absolute expenditure on specific products into the proportion of spending relative to the customer’s total expenditure at the store. This transformation allows for a more accurate representation of the customer’s preference for particular product categories, as customers with a stronger preference for a specific type of product are expected to allocate a higher share of their total expenditure toward that product. This proportional measure captures relative consumer

preference, mitigating biases that may arise from differences in customers' total purchasing power or overall spending habits.

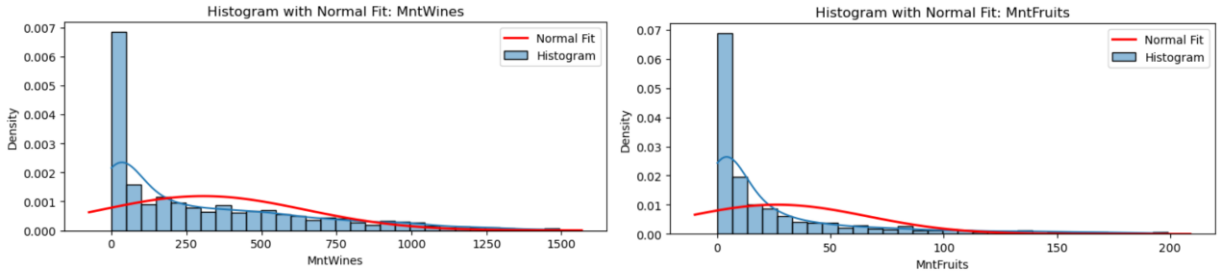

Figure 3: Absolute Expenditure on Specific Products

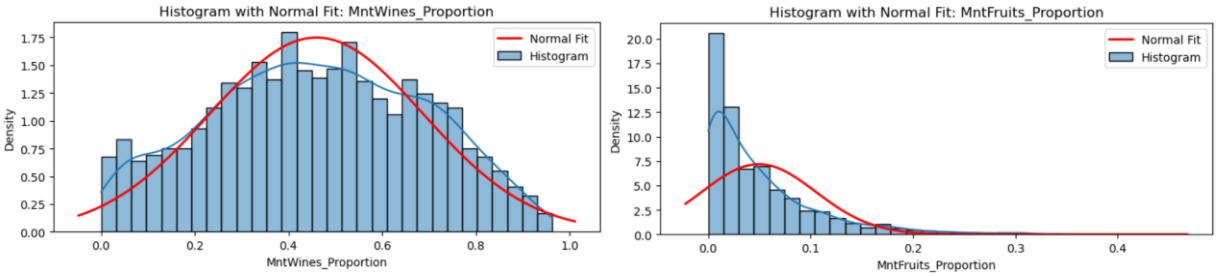

Figure 4: The Proportion of Spending Relative to the Customer's Total Expenditure

As illustrated in Figure 3, the majority of customer's absolute expenditure within this store network is relatively low. This observation supports our earlier hypothesis that customers do not necessarily rely exclusively on this store for their purchases, and as a result, their absolute spending may not fully reflect their preferences for certain product categories. After converting absolute spending into proportional expenditure relative to total store spending, we observed distinct distributional patterns across product categories. As shown in Figure 4, for wine and meat products, customer spending proportions exhibit characteristics of a normal distribution. In contrast, the proportional spending on fruit products, fish products, sweets, and gold products displays a distribution pattern that more closely resembles a skewed or chi-squared distribution, with a large concentration of customers in the lower expenditure proportion range.

This observation suggests that on one hand, customer demand for wine and meat products is relatively stable, with customers maintaining a consistent proportional expenditure on these items, regardless of their overall spending capacity. On the other hand, purchases of fruits, fish, sweets, and gold products appear to be driven more by personal preferences, leading to a broader range of spending behaviors. Additionally, we found that the correlation coefficient between customer's income and proportional spending on wine and meat products is significantly higher than that of other products, as shown in Figure 5. This finding further supports the notion that wine and meat products correspond to more stable consumer demand, while purchases of other product categories exhibit greater variability due to individual preference.

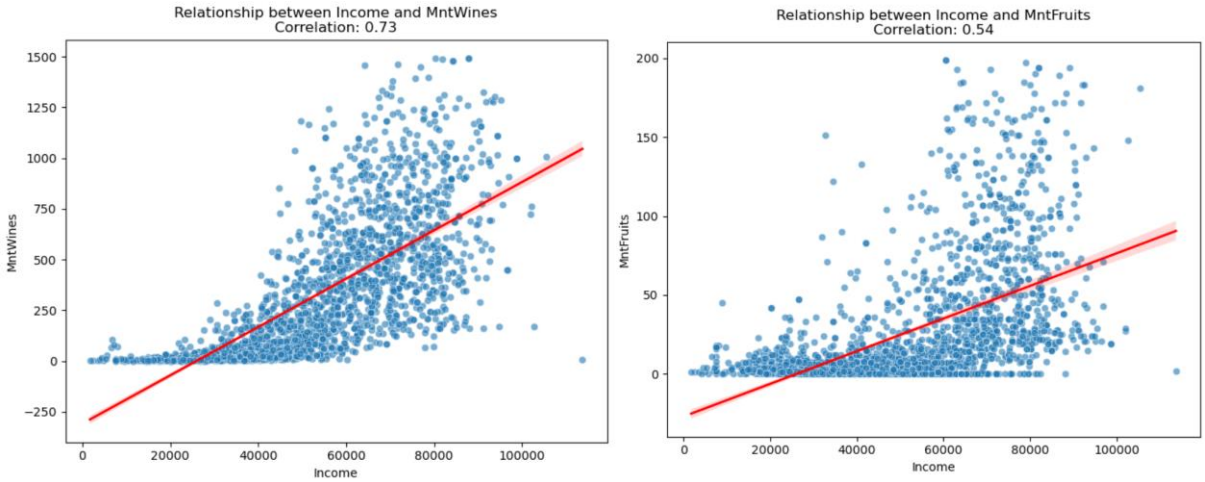

Figure 5: Correlation between Customer's Income and Proportional Spending

Therefore, a higher expenditure on a particular product category alone is insufficient to conclude that a customer has a stronger purchase intent for that category. For instance, customers who spend more on meat products may simply have higher incomes, enabling them to allocate more funds toward essential goods like meat. However, from a business perspective, identifying customers with greater purchasing power is also valuable, as targeted promotions aimed at these customers may yield better marketing outcomes.

To address this dual objective, we incorporated both purchase intent and purchasing power into our classification approach. We defined the target customers for a specific product as those who either 1) have an expenditure on the product that exceeds the population mean, or 2) have a proportional expenditure on the product that surpasses the peak of the chi-squared distribution for that product. This dual criterion ensures that the model identifies both high-spending customers and those with a clear preference for the product, enabling more effective customer segmentation and personalized marketing strategies.

2.3 Principal Component Analysis (PCA)

High-dimensional customer profile data contains a significant amount of noise, which can adversely impact the predictive capabilities of machine learning models. Eliminating irrelevant features can help models focus more on the most critical information, thereby improving prediction accuracy. However, different models require distinct approaches to noise reduction and dimensionality reduction. In our case, the models we used are Logistic Regression, SVM, and K-Means Clustering.

For Logistic Regression, Ridge and Lasso regularization (L2 and L1 norms, respectively) automatically identify and exclude irrelevant features. Since Logistic Regression assigns a weight coefficient to each feature, training directly on the original user profile data is a reasonable choice. This approach maintains the interpretability of the model while still achieving strong predictive performance. The resulting model not only remains interpretable but also effectively handles feature selection as part of the training process.

For SVM, dimensionality reduction using PCA is recommended before model training. SVM defines classification boundaries using support vectors, and PCA removes low-variance directions in the data, leading to smoother, more stable decision boundaries. This dimensionality reduction reduces computational cost, accelerates training, and decreases the risk of overfitting. By focusing on the principal components, SVM can form more robust classification rules.

Similarly, for K-Means Clustering, it is beneficial to apply PCA to reduce dimensionality before clustering. K-Means Clustering partitions data by minimizing the distance between samples and their assigned centroids. By projecting the data onto its principal components, PCA aligns the data with its main variance directions. This step reduces redundancy, improves the quality of clustering, and speeds up the iterative computation of centroids. Reducing dimensionality ensures that the most significant patterns in the data are captured, while computational efficiency is improved.

Given the considerations for each model, we proposed using the original customer profile data for Logistic Regression model training to preserve feature interpretability and leverage automatic feature selection through regularization. For SVM and K-Means Clustering, we applied PCA to reduce dimensionality and noise before training. This approach optimizes computational efficiency, stabilizes decision boundaries in SVM, and improves clustering quality in K-Means Clustering. Next, we combined the characteristics of PCA itself and its reflection in model performance to determine the number of components of PCA.

When selecting the appropriate number of dimensions to retain in PCA, the proportion of variance explained by the retained components is a critical consideration. For the current customer profile

data, retaining more than 7 dimensions captures over 90% of the original data's variance. However, this does not imply that 7 dimensions is the optimal choice. As the dimensionality of the data increases, SVM's decision boundaries can become less robust, and K-Means Clustering may suffer from the “distance concentration phenomenon”, where distances between points tend to become similar, reducing the effectiveness of clustering. Consequently, selecting a smaller number of dimensions may enhance the performance of SVM and K-Means Clustering.

Therefore, in order to identify the optimal dimensionality, we employed a cross-validation approach. We evaluated the classification performance of SVM and K-Means Clustering for PCA-retained dimensions ranging from 1 to 7. The performance results are illustrated in Figure 6. The analysis reveals that when the number of retained dimensions is between 3 and 7, the classification performance of SVM and K-Means Clustering is relatively stable and similar. Considering that a 3D representation allows for more intuitive visualization and interpretability, we selected 3 as the optimal number of dimensions to retain in PCA for downstream modeling with SVM and K-Means Clustering.

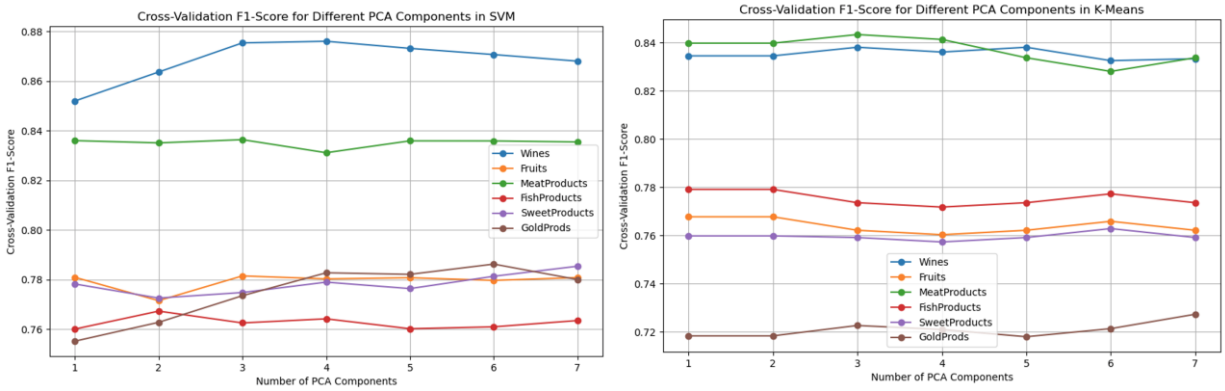

Figure 6: Cross-Validation F-1 Score for Different PCA Components in SVM and K-Means Clustering

3. Models and Analysis

The modeling process was designed to classify consumer spending behavior across six key product spending data: Wines, fruits, meat products, fish products, sweet products, and gold products. The spending data was standardized to ensure uniform feature scaling, which is crucial for models to operate effectively. For each target product, a binary classification label was defined, indicating high or low spending behavior. The labeling was determined using two criteria as mentioned in session 2.2: a threshold based on spending proportion and a threshold based on total spending.

3.1 Logistic Regression

In Logistic Regression model training, a grid search with cross-validation was used to identify the optimal hyperparameters. Hyperparameters in Table 2 were considered:

Hyperparameter	Values	Description
C	[0.1, 1, 10, 100]	Regularization strength
Penalty	['l1', 'l2', 'elasticnet']	Regularization type
Solver	['lbfgs', 'liblinear', 'newton-cg', 'newton-cholesky']	Optimization algorithm

Table 2: Logistic Regression Hyperparameters

Each model was trained and validated using 5-fold cross-validation to identify the optimal combination of hyperparameters for each target product category. After training, the Logistic Regression model's performance was assessed using classification metrics including accuracy, recall, precision, and F1-score. Confusion matrices were plotted to visualize classification errors.

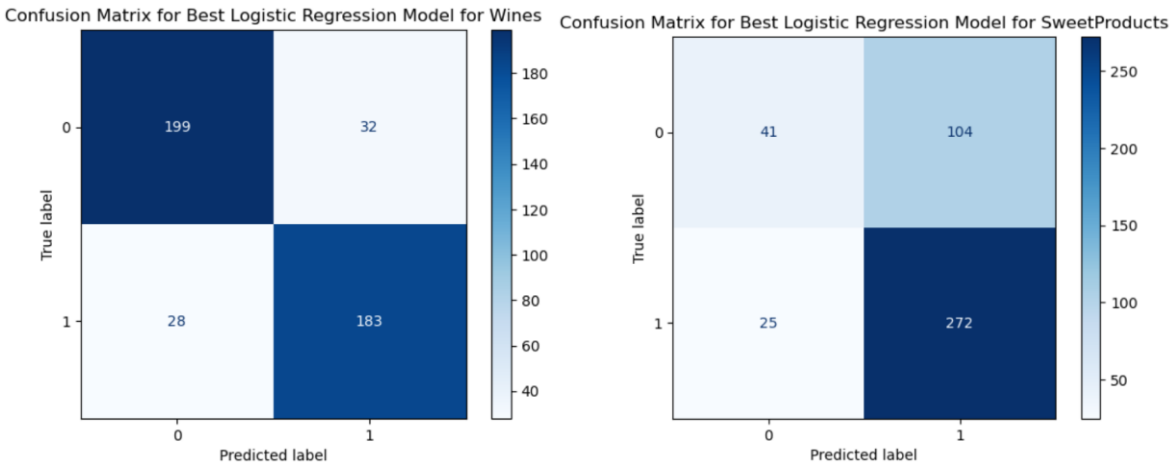

Figure 7: Confusion Matrix for Logistic Classification for Wines and Sweet Products

Target	Accuracy	Recall	Best Parameters
Wines	0.8643	0.8673	C=0.1, penalty='l1', solver='liblinear'
Fruits	0.6946	0.8248	C=0.1, penalty='l1', solver='liblinear'
MeatProducts	0.8575	0.8447	C=0.1, penalty='l1', solver='liblinear'
FishProducts	0.6855	0.8252	C=1, penalty='l1', solver='liblinear'
SweetProducts	0.7081	0.9158	C=0.1, penalty='l1', solver='liblinear'
GoldProds	0.7195	0.8051	C=0.1, penalty='l1', solver='liblinear'

Table 3: Logistic Regression Classification Results

The results indicate that Logistic Regression demonstrates good classification performance for wines and meat products. However, for fruits, sweets, and gold products, Logistic Regression tends to misclassify a larger portion of the population as having high purchase intent. This may be attributed to the limited ability of Logistic Regression to handle high-dimensional data effectively.

In earlier analysis, it was observed that the distributions of wines and meat products approximate normal distributions, and their consumption levels are strongly correlated with income. In such cases, Logistic Regression performs well by assigning higher weights to income-related factors. Conversely, the consumption of other products does not exhibit a clear correlation with any single customer profile factor. Consequently, Logistic Regression performs poorly for these products.

Therefore, alternative classification models that can effectively handle high-dimensional data are required to improve classification performance in such cases.

3.2 Support Vector Machine (SVM)

SVM uses the data after PCA dimensionality reduction (components = 3) for training. A grid search with cross-validation was used to identify the optimal hyperparameters for SVM. Hyperparameters in Table 2 were considered:

Hyperparameter	Values	Description
C	[0.1, 1, 10, 100]	The trade-off between margin and errors
Kernel	['linear', 'rbf']	Type of hyperplane to separate classes
Gamma	['scale', 'auto']	Influence of individual data points

Table 4: SVM Hyperparameters

Each model was trained and validated using 5-fold cross-validation to identify the optimal combination of hyperparameters for each target product category. After training, the SVM model's performance was assessed using classification metrics including accuracy, recall, precision, and F1-score. Confusion matrices were plotted to visualize classification errors, and 3D scatter plots were used to visualize classification boundaries for each product category.

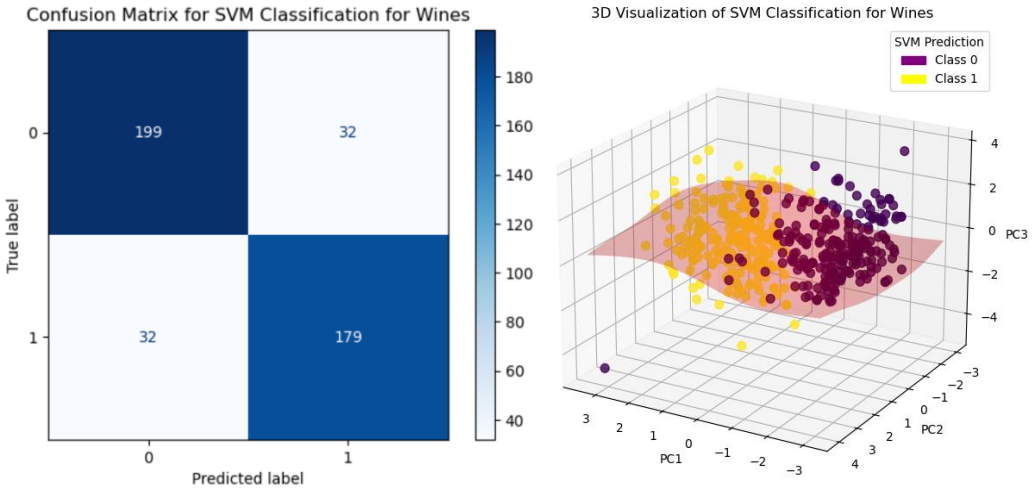

Figure 8: Confusion Matrix and 3D Visualization for SVM Classification for Wines

The SVM model calculated the classification accuracy and recall rate for the high consumption willingness purchase groups for each product under the best hyperparameter selection. The best hyperparameters for each product were also reported to provide insight into the most effective model configurations.

Target	Accuracy	Recall	Best Parameters
Wines	0.8552	0.8483	C=1, kernel='rbf', gamma='scale'
Fruits	0.8009	0.7143	C=1, kernel='rbf', gamma='auto'
MeatProducts	0.8484	0.8350	C=0.1, kernel='rbf', gamma='scale'
FishProducts	0.7919	0.6930	C=0.1, kernel='rbf', gamma='auto'
SweetProducts	0.7760	0.7426	C=0.1, kernel='linear', gamma='scale'
GoldProds	0.7421	0.7682	C=1, kernel='rbf', gamma='auto'

Table 5: SVM Classification Results

Compared to Logistic Regression, SVM demonstrates superior performance on data such as fruits, which do not follow a normal distribution and exhibit a low correlation with income. This improvement may be attributed to SVM's ability to handle non-linear decision boundaries and its reliance on support vectors, which enables it to effectively capture complex patterns in the data without being constrained by assumptions of normality or linearity.

3.3 K-Means Clustering

Upon analyzing the results of the SVM model, it is observed that the two classified groups exhibit a certain degree of cohesion within the groups and separation between the groups in the feature space. Motivated by this observation, we attempted to apply K-Means Clustering to classify the customer profile data based on its inherent characteristics. Similar to SVM, K-Means Clustering uses the data after PCA dimensionality reduction (components=3) for training. K-Means

Clustering was implemented with a fixed number of clusters set to 2 for each product category. The Hungarian algorithm is used to re-align cluster labels with ground-truth labels, ensuring consistency in the evaluation.

After training, visual analysis reveals that the classification results of K-Means Clustering are notably similar to those of SVM, demonstrating strong classification performance on the test data. This indicates that the inherent characteristics of customer profiles are indeed intrinsically related to customers' purchase intention features.

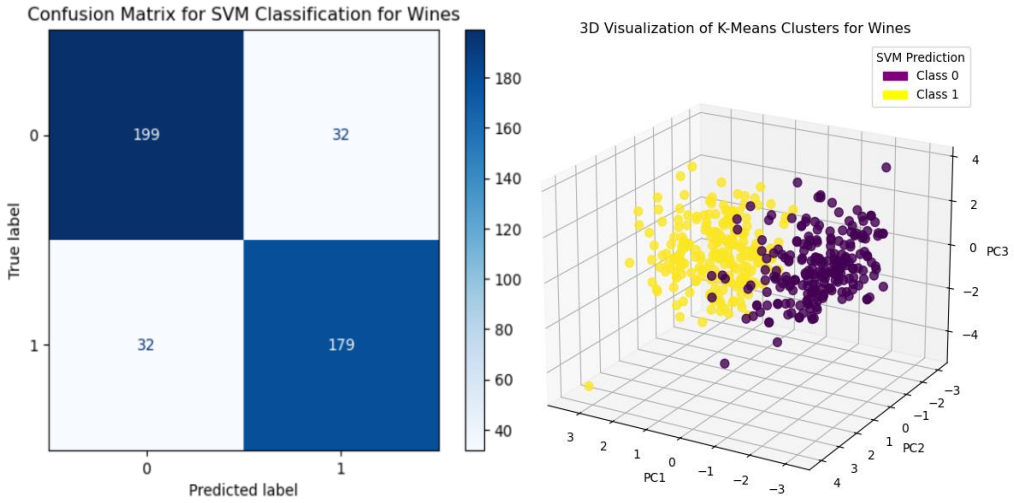

Figure 9: Confusion Matrix and 3D Visualization for K-Means Classification for Wines

Target	Accuracy	Recall
Wines	0.8462	0.8341
Fruits	0.7783	0.7734
MeatProducts	0.8529	0.8495
FishProducts	0.7828	0.7628
SweetProducts	0.7760	0.7723
GoldProds	0.7308	0.7045

Table 6: K-Means Clustering Classification Results

Both the visualized images and results demonstrate that the classifications produced by K-Means are highly similar to those generated by SVM. Moreover, despite having slightly lower accuracy, K-Means achieves better recall across a broader range of product categories. This can be explained by the fact that K-Means relies on the inherent cohesion and separation within the user profile data to form clusters. Unlike SVM, which depends on labeled data, K-Means is less affected by outliers in the training data and does not sacrifice accuracy due to edge cases when defining decision boundaries.

4. Discussion

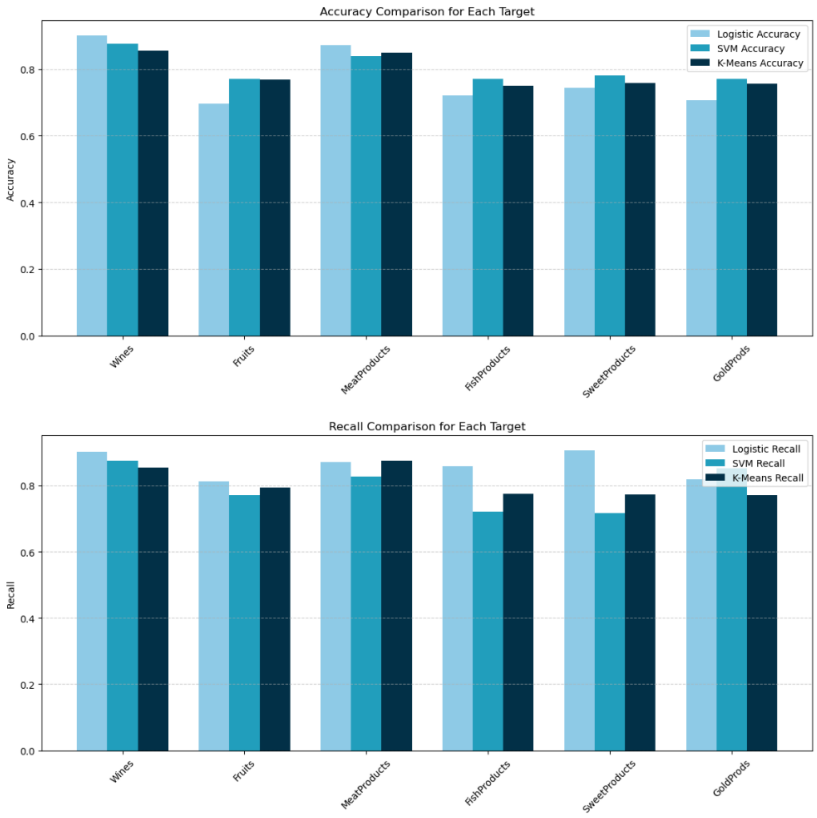

Figure 10: Accuracy and Recall Comparison among Three Models on Each Product

An analysis of the prediction results from the three classification models reveals distinct characteristics in their performance across different product categories. The logistic regression

model performs well for products where purchase behavior is strongly associated with a single factor, but it struggles with products influenced by more complex consumer habits. This limitation reflects the inability of logistic regression to capture nonlinear decision boundaries in complex classification tasks.

SVM achieves the highest overall accuracy, highlighting its strength in handling complex decision boundaries. However, due to the limited size of the dataset, SVM exhibits signs of overfitting, even with hyperparameter tuning for kernel type and distance functions. This overfitting reduces the recall rate for certain product categories, as the model becomes too sensitive to small variations in the training data.

In contrast, the K-means clustering method offers a different perspective. Unlike supervised models, K-means does not require labeled data, allowing it to capture the inherent structure of the data based on feature cohesion. This property is valuable for exploratory analysis and increases the likelihood of capturing potential users for target product categories, thereby improving recall. However, since K-means does not directly consider class labels, its effectiveness depends on the assumption that the natural clusters in the data align with the class boundaries. Additionally, K-means has limited interpretability compared to supervised models, which can hinder decision-making when clear classification logic is required.

We also observed that customer consumption patterns for wine and meat products differ significantly from other product categories. Customers exhibit relatively stable spending proportions on these two categories, suggesting that wine and meat products may have a certain

degree of necessity in consumer behavior. Consequently, while our models perform well in identifying users with high purchase intent for wine and meat products, further analysis may require the integration of additional data. For instance, regional consumption data for wine and meat products could provide valuable context for understanding customer behavior more comprehensively. Such supplementary data would allow for a more nuanced assessment of customer preferences, distinguishing between essential consumption driven by necessity and preference-driven purchasing behavior.

5. Conclusions

Our project implemented customer segmentation through a comprehensive analysis. Starting with data preprocessing and feature engineering, we transformed raw customer data into meaningful metrics, including both absolute spending and proportional expenditure patterns. We then applied three distinct approaches: Logistic Regression on mostly original features, and both SVM and K-Means Clustering on PCA-reduced dimensions. Each method demonstrated unique strengths in classifying customer purchase intent across different product categories. Overall, this project underpins data-driven decisions that enhance economic efficiency and consumer engagement. For businesses, accurately identifying high and low-consumption customers allows for more effective marketing strategies, and better resource allocation. Customer segmentation supports optimizing product offerings, which can potentially drive revenue growth.

One limitation of our analysis lies in its reliance on monetary expenditure as the primary indicator of consumer preference, which may not fully capture the underlying utility maximization behavior of customers. According to consumer choice theory, utility is derived from the consumption of

goods rather than the monetary value spent. Our current metrics - absolute spending and spending proportions - may introduce bias due to price heterogeneity across product categories, particularly for luxury items like gold products which have substantially higher unit prices. Future study might use purchase quantity and frequency if possible as they can serve as more reliable indicators of consumer preferences and better reflect consumption patterns and sustained category engagement, independent of price effects described by Engel's law (Engel, 1857).

6. References

Engel, E. (1857). Die Productions-und Consumtionsverhältnisse des Königreichs Sachsen. *Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministeriums des Innern*, *8*, 1-54.

Mahmood, N. H., Rauf, S., Javed, A., & Mehmood, A. (2016). User-centered modeling incorporating personality traits for improving user experience in complex systems. 10.1007/978-3-319-47175-4_29.