Customer Spending Behavior Investigation*

*Supervised Leaning

1st Yoon Kim
Georgia Institute of Technology
Atlanta, USA
blue1357a@gmail.com

***Abstract*—The "Customer Personality" dataset provides a comprehensive analysis of customer demographics and spending behavior, offering valuable insights into a company's ideal customer base. This study aims to investigate the dataset to better understand customer spending patterns based on demographic information, particularly focusing on family size and income levels. By analyzing these attributes, we aim to identify patterns in spending behavior across different product categories. Our goal is to predict which categories of products individuals or families are most likely to spend on, based on their demographic and financial characteristics. This analysis will help businesses tailor their products and marketing strategies to meet the specific needs and preferences of various customer groups.**

I. Hypothesis

Based on family size and per-family income, people spend differently on various categories of products (e.g., wine, fruits, meat, fish, sweets, and gold). These spending patterns are influenced by family dynamics (e.g., single, with children, etc.) and financial capacity, leading to distinct preferences across categories.

II. Description of Classification

A. Objective

The primary objective of this analysis is to predict consumer spending preferences across various product categories, including wine, fruits, meat, fish, sweets, and gold, by utilizing demographic and family-related attributes. By examining factors such as family size, income, and other demographic information, the study aims to identify and rank spending patterns of individuals and families. This ranking will help determine which product categories receive the most spending, providing insights into consumer behavior and preferences.

B. Non-Triviality

The relationship between family size, income, and spending patterns is inherently complex and influenced by various factors such as lifestyle, culture, and financial priorities. For instance, families with children may prioritize spending on essential items like meat and fruits, while single individuals are more likely to allocate resources to luxury items such as wine or gold. These interactions are far from linear or straightforward, making the problem a non-trivial challenge for machine learning models. The complexity of these dependencies requires sophisticated modeling to capture the nuanced relationships between attributes and spending preferences.

C. Algorithmic Comparisons

To address this challenge, several machine learning models, including neural networks, support vector machines (SVMs), and k-nearest neighbors (KNN), are used to predict spending behavior and evaluate their performance. These algorithms are well-suited to handle the mix of categorical and numerical data present in the dataset. By applying these models, the study explores the extent to which demographic and family-related characteristics influence consumer spending and compares the effectiveness of each algorithm in capturing these intricate relationships. The results of these comparisons will provide valuable insight into the most appropriate techniques to analyze consumer spending behavior.

III. Initial Data Investigation

The primary objective of this analysis is to examine the across different product categories. The spending rates for various product types are presented below:

Category	Average Spending Rate
MntWines_rate	0.4585
MntFruits_rate	0.0495
MntMeatProducts_rate	0.2495
MntFishProducts_rate	0.0715
MntSweetProducts_rate	0.0508
MntGoldProds_rate	0.1202

TABLE I
Average Spending Rates per Category

As observed in Table I, spending on **MntWines** is significantly higher compared to other product categories, nearly double that of other spending rates. This indicates that wine purchases constitute the most substantial portion of spending behavior in this market.

A. Comparison Based on Family Size

To further analyze consumer spending behavior, the average spending rates are evaluated based on family size.

From Figure 1, it is evident that as **family size increases, the spending rate on wines (MntWines_rate) also increases**, while spending on fruits and fish products tends to decline. This trend suggests that larger households allocate a higher percentage of their spending towards wine products.

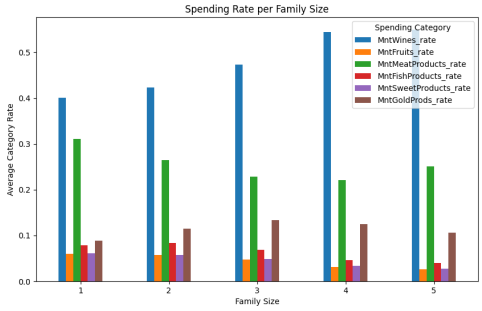

Fig. 1. Spending rate distribution across different family sizes.

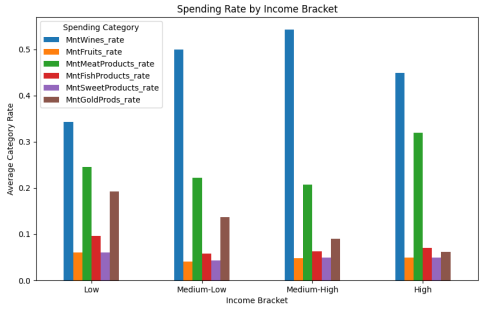

Fig. 2. Spending rate distribution across different family sizes.

B. Comparison Based on Income Bracket

Additionally, the spending rates are compared across different income brackets, as presented in Figure 2. Spending rates on wines and gold products tend to increase with income level, while meat product spending varies significantly across brackets. Notably, individuals in the highest income bracket allocate a larger portion of spending towards meat products (MntMeatProducts_rate = 0.3203compared to other income groups.

C. Correlation Explanation

The correlation heatmap in Fig. 3 provides insights into the relationships between various spending categories and demographic factors. A strong positive correlation between variables indicates that they tend to increase together, whereas a negative correlation suggests an inverse relationship. As you can see, the data shows that larger family sizes have a weak positive correlation with wine purchases but a negative correlation with other product categories such as fruits and fish. Additionally, income per family member exhibits a moderate negative correlation with gold spending, implying that higher-income individuals allocate a smaller proportion of their spending to luxury items. Understanding these correlations helps in identifying key factors influencing consumer purchasing behavior.

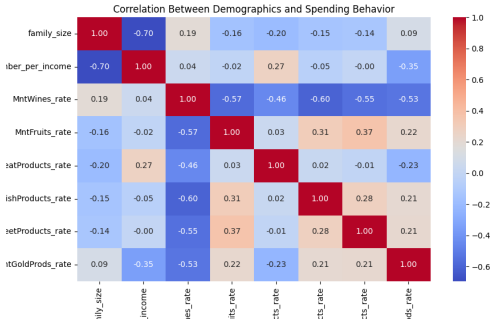

Fig. 3. Correlation heatmap of spending behavior.

MODELING

After conducting this investigation, we developed models to predict the percentage of spending in various categories. Initially, we created neural networks, with plans to include Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) in future sections. These models were designed to predict spending based on household demographics, including family income and the number of family members. The goal was to estimate how much each household spends on different categories. We evaluated the models using Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2 scores, and accuracy.

D. Neural Networks

The neural networks were implemented using TensorFlow. We utilized activation functions such as ReLU and Sigmoid, both of which are suitable for predicting family spending categories. To improve the accuracy of spending behavior predictions, we experimented with different learning rates.

Accuracy Calculation:

TABLE II
SVM PERFORMANCE METRICS FOR LINEAR AND POLYNOMIAL KERNELS

Kernel	Category	MSE	R^2 Score	Accuracy (%)
Linear	MntWines	0.04	0.24	17.12
	MntFruits	0.01	-1.04	5.18
	MntMeatProducts	0.01	0.14	14.41
	MntFishProducts	0.01	-0.02	7.21
	MntSweetProducts	0.01	-0.64	4.73
	MntGoldProds	0.01	0.17	8.78
Polynomial	MntWines	0.04	0.31	18.69
	MntFruits	0.01	-0.99	5.41
	MntMeatProducts	0.01	0.15	17.79
	MntFishProducts	0.01	-0.01	6.31
	MntSweetProducts	0.01	-0.58	5.18
	MntGoldProds	0.01	0.26	9.01

The accuracy was calculated based on the comparison between predicted values and actual test results (y_{test} vs. y_{pred}). For the ReLU activation function, we tested learning rates of 0.01, 0.05, 0.2, and 0.5. As the learning rate increased, both the Mean Squared Error (MSE) and Mean Absolute Error (MAE) also increased, indicating a decline in model performance. Additionally, the R^2 score decreased with higher learning rates, reflecting less reliable predictions.

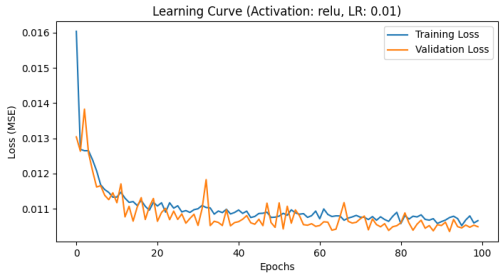

Fig. 4. ReLU Activation with Learning Rate = 0.01

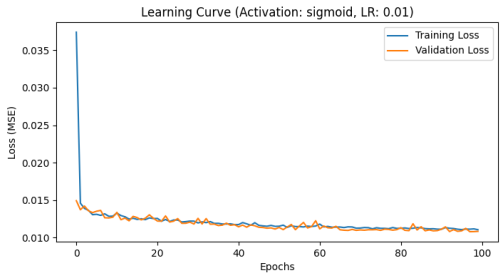

Fig. 5. Sigmoid Activation with Learning Rate = 0.01

Visualization of Model Performance: The figures below illustrate the learning rate performance and prediction accuracy for both ReLU and Sigmoid activation functions.

Furthermore, the predicted vs. actual spending for both activation functions is visualized below:

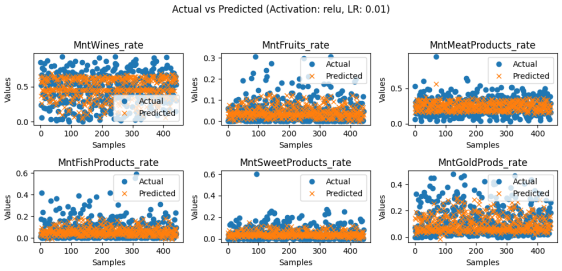

Fig. 6. ReLU Predicted vs. Actual Spending

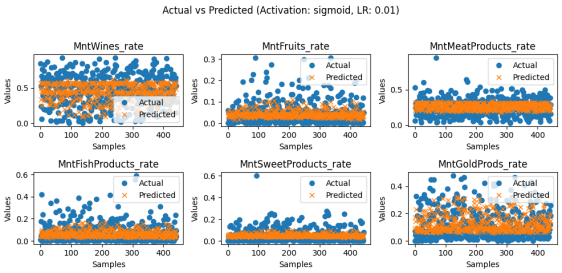

Fig. 7. Sigmoid Predicted vs. Actual Spending

Overall Model Comparison: Based on the results, both ReLU and Sigmoid activation functions demonstrate strong predictive performance, with ReLU showing slightly more stable accuracy across different learning rates. However, Sigmoid achieved a higher peak accuracy at a learning rate of 0.01. These findings suggest that while both functions are effective, tuning the learning rate is crucial for optimizing model performance.

E. SVM Model

In this section, Support Vector Machine (SVM) models were applied using two different kernels: Linear and Polynomial (Poly) The results from both models are presented and compared to evaluate their performance in predicting spending patterns across various categories.

TABLE III
SVM PERFORMANCE METRICS FOR LINEAR AND POLYNOMIAL KERNELS

Kernel	Category	MSE	R^2 Score
Linear	MntWines_rate	0.04	0.24
	MntFruits_rate	0.01	-1.04
	MntMeatProducts_rate	0.01	0.14
	MntFishProducts_rate	0.01	-0.02
	MntSweetProducts_rate	0.01	-0.64
	MntGoldProds_rate	0.01	0.17
Polynomial	MntWines_rate	0.04	0.31
	MntFruits_rate	0.01	-0.99
	MntMeatProducts_rate	0.01	0.15
	MntFishProducts_rate	0.01	-0.01
	MntSweetProducts_rate	0.01	-0.58
	MntGoldProds_rate	0.01	0.26

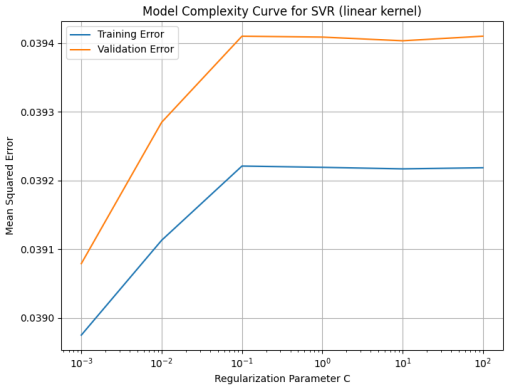

Fig. 8. Model Complexity Curve for SVR (Linear Kernel)

Linear Kernel Analysis: The training error slightly increases as the regularization parameter C increases. This is unusual because typically, a more complex model (higher C) fits the training data better. This trend suggests that the model may be overfitting noise in the data. Initially, the validation error increases as C increases, then stabilizes and slightly decreases. This indicates that after reaching a certain model complexity, further increasing C does not significantly improve performance.

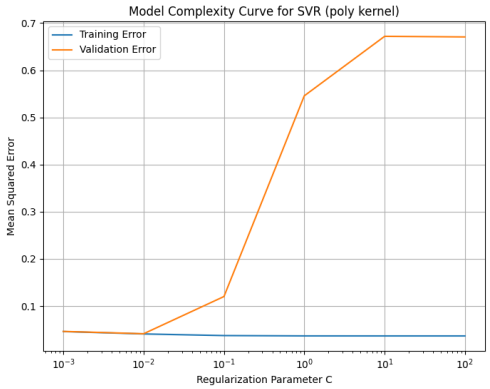

Fig. 9. Model Complexity Curve for SVM (Polynomial Kernel)

k	Category	MSE	R^2	Accuracy
1	MntWines	0.05	-0.01	26.80%
	MntFruits	0.00	-0.67	21.62%
	MntMeatProducts	0.02	-0.23	26.35%
	MntFishProducts	0.01	-0.31	20.95%
	MntSweetProducts	0.01	-0.47	20.50%
	MntGoldProds	0.01	-0.04	18.02%
3	MntWines	0.04	0.16	18.92%
	MntFruits	0.00	-0.15	7.88%
	MntMeatProducts	0.01	0.02	16.44%
	MntFishProducts	0.01	-0.02	6.31%
	MntSweetProducts	0.00	-0.07	6.53%
	MntGoldProds	0.01	0.10	8.56%
5	MntWines	0.04	0.25	16.22%
	MntFruits	0.00	0.03	7.21%
	MntMeatProducts	0.01	0.09	17.34%
	MntFishProducts	0.01	0.01	5.18%
	MntSweetProducts	0.00	0.01	6.31%
	MntGoldProds	0.01	0.21	9.01%

TABLE IV
PERFORMANCE METRICS FOR DIFFERENT CATEGORIES AND VALUES OF K.

Polynomial Kernel Analysis: The training error remains stable, indicating that the model is well-fitted to the training data. However, the validation error increases dramatically as C increases, highlighting severe overfitting. When C is small, the model is well-regularized, showing better generalization to unseen data. However, as C increases, the model loses its regularization effect, leading to poor validation performance.

Overall Model Comparison: Based on the results, both linear and polynomial kernels demonstrate limited predictive performance. The linear kernel shows relatively stable behavior across different regularization parameters, while the polynomial kernel exhibits signs of overfitting as the model complexity increases. These findings highlight the importance of careful hyperparameter tuning and suggest that additional kernel functions or regularization strategies may improve model performance.

F. KNN Models

In this section, I present the analysis of the K-Nearest Neighbors (KNN) models. The models were tested with various kernel functions and different values of k to evaluate their performance. The experiments involved changing clusters and testing on $k = 1, 3, 5$ with different kernel ranges.

1) Model Performance Results: The table below summarizes the MSE and R^2 scores for each category with different k values:

2) Model Performance Visualization:

3) Model Complexity Analysis: Based on the KNN model complexity curve, several key observations emerge. As the training error increases, it gradually rises because the model becomes less flexible, considering more neighbors when making predictions. In contrast, the validation error starts relatively high when k is small, indicating overfitting to the training data, which does not generalize well to new data. As the k value increases, the validation error decreases initially, reflecting improved generalization. Around $k = 10 - 12$, the validation error stabilizes, showing diminishing returns with further increases in k. This stabilization suggests that the model has reached an optimal level of complexity, balancing bias and variance effectively.

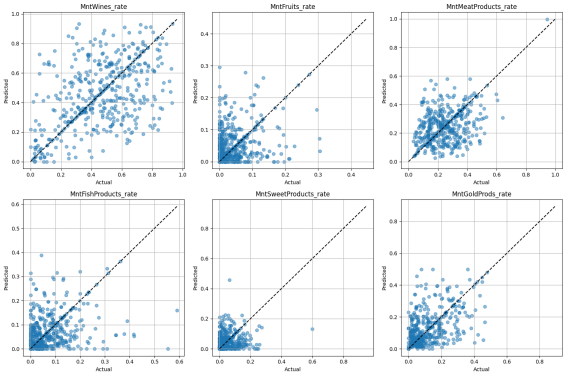

Fig. 10. Cross-Validation Results for KNN with $k = 1$

KNN with Cross-Validation (k = 1, 3, 5): The predicted points are more tightly clustered around the diagonal line ($y = x$), particularly for $k = 3$ and $k = 5$. This indicates better predictive performance. Cross-validation reduces the variance in predictions, leading to more stable and reliable outcomes.

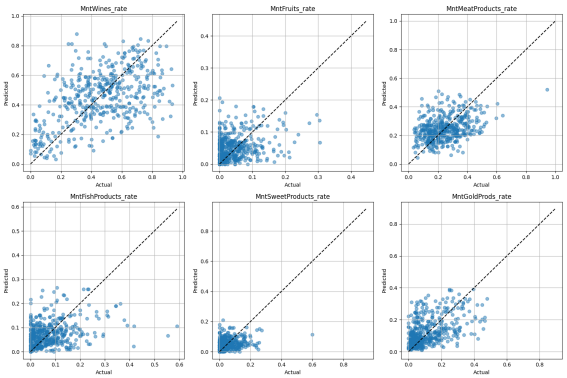

Fig. 11. Cross-Validation Results for KNN with $k = 1$

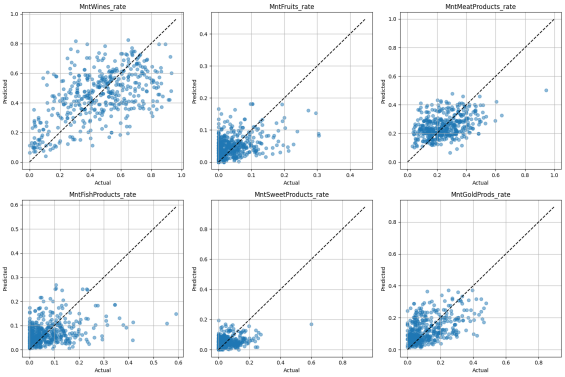

Fig. 12. Cross-Validation Results for KNN with $k = 1$

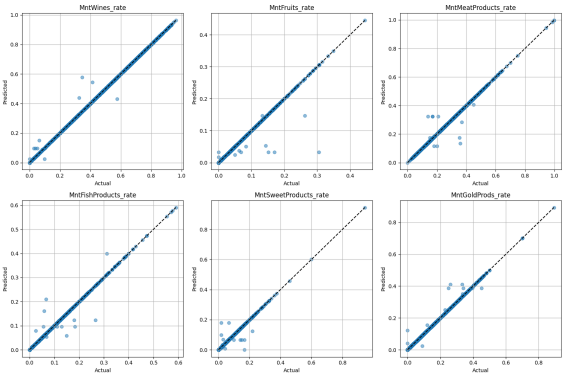

Fig. 13. Cross-Validation Results for KNN with $k = 1$

The use of cross-validation helps reduce overfitting, resulting in smoother predictions that generalize better to unseen data.

IV. Overall Analysis and Discussion

Overall, the models that I have created Neural, SVM, KNN, do not exhibit a highly accurate performance. Both the Mean Squared Error (MSE) and values lack significant results, indicating that the models struggle to generalize well to unseen data. However, it is noteworthy that the predictions for Wine and Fruits spending rates are relatively more accurate

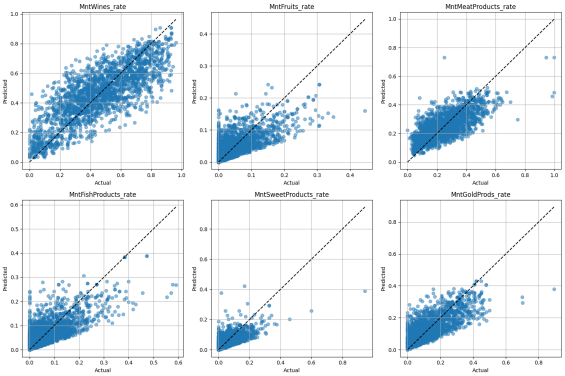

Fig. 14. Cross-Validation Results for KNN with $k = 3$

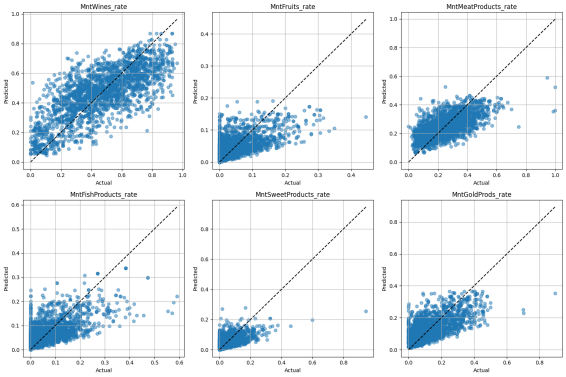

Fig. 15. Cross-Validation Results for KNN with $k = 5$

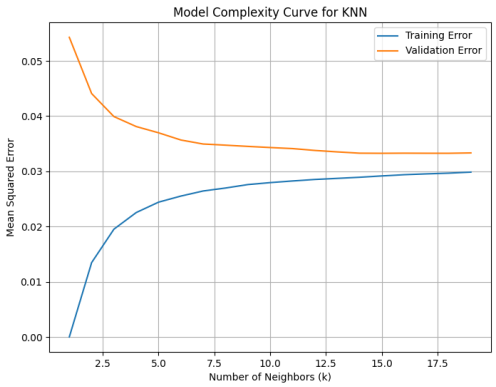

Fig. 16. KNN Model Complexity

compared to other categories. The data conversion process involved calculating spending rates based on total spending, aiming to understand consumer spending behaviors more effectively. Despite this approach, the models showed limitations due to several factors discussed below.

A. Comparative Analysis of Algorithms

Neural networks provided moderate accuracy, especially when using ReLU and Sigmoid activation functions. Performance was sensitive to learning rates. Higher learning rates resulted in increased MSE and decreased scores, reflecting less stable predictions. To improve performance, implementing regularization techniques such as dropout could prevent overfitting, and fine-tuning learning rates along with network architecture adjustments.

Support Vector Machines (SVM) with linear and polynomial kernels displayed limited predictive capabilities. Polynomial kernels tended to be overfitting, especially with high regularization parameters, while linear kernels performed slightly better in terms of stability. To improve performance, experimenting with RBF kernels for capturing non-linear patterns and applying grid search for hyper-parmeters tuning could be beneficial.

K-Nearest Neighbors (KNN) models showed inconsistent accuracy across categories. Predictions for wine and fruit spending were relatively accurate, whereas other categories performed poorly. Lower values resulted in overfitting, while higher values led to underfitting. Optimizing the value of through cross-validation and introducing distance-weighted KNN to give more importance to closer neighbors might improve results.

B. Hypothesis vs. Results

The original hypothesis suggested that family size and income would significantly influence spending behavior. However, the results showed only moderate correlations. Wine and fruits spending align well with the hypothesis, as income positively affects these categories. In contrast, gold and meat products did not show strong predictable patterns based on family size or income, indicating the presence of more complex underlying factors.

C. Performance Considerations

Neural networks required to take long time, especially with deeper architectures, However, SVMs were faster with linear kernels but slower with polynomial kernels due to their complexity. KNN had moderate computation times, heavily dependent on dataset size. Regarding iterations, neural networks needed multiple epochs to converge, while SVMs performed well with fewer iterations due to their optimization nature.

D. Best Performing Algorithm

Among the models, neural networks showed the best overall performance due to their flexibility. The criteria for defining the best-performing algorithm include high accuracy, stable scores, and adaptability to different data patterns. In conclusion, while the models show potential, further optimization is needed to achieve significant predictive power. The results highlight the complexity of consumer spending behavior, emphasizing the need for more sophisticated modeling techniques.

V. Conclusion

This study explored customer spending behavior using demographic and financial data to uncover patterns and trends across various product categories. By applying a range of machine learning models—neural networks, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN)—we aimed to predict consumer preferences and spending tendencies with a focus on factors like family size and income levels.

The results revealed nuanced insights into consumer behavior. Neural networks demonstrated the most promising predictive capabilities, offering stable performance with fine-tuned learning rates and activation functions such as ReLU and Sigmoid. Despite this, model accuracy varied across categories, with spending on wine and fruits being more predictable compared to other products like gold and meat. This suggests that consumer behavior in certain categories may be influenced by external factors beyond basic demographics, such as cultural preferences, lifestyle choices, and economic conditions.

SVM models showed moderate performance, with linear kernels providing more stability compared to polynomial kernels, which tended to overfit the data. Meanwhile, KNN models highlighted the importance of parameter tuning, as variations in k significantly affected model accuracy. Although KNN achieved decent results in specific categories, it struggled to generalize across the dataset, especially for products with more complex purchasing patterns.

While our initial hypothesis posited strong correlations between family size, income, and spending habits, the analysis revealed only moderate associations. This highlights the multifaceted nature of consumer behavior, where psychological, social, and environmental factors interplay with economic variables.

For future work, incorporating additional data sources, such as behavioral metrics or transaction histories, could enhance predictive accuracy. Additionally, leveraging advanced modeling techniques like ensemble learning or deep learning architectures may offer deeper insights. Continuous model optimization, coupled with robust feature engineering, will be key to refining predictions and driving more actionable business insights.

Ultimately, this investigation underscores the potential of data-driven approaches in understanding consumer behavior, providing valuable guidance for businesses aiming to tailor their marketing strategies and product offerings to meet diverse customer needs.

References

[1] Customer Personality Analysis Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

[2] Overleaf Project. [Online]. Available: https://www.overleaf.com/read/grnxnvpshskk#1c368c