Research Article

Exploring Gender-Based Differences in Banking Transactions: A Data-Driven Analysis Using Multiple Linear Models and Permutation Feature Importance

Nathan Ma

Senior, Diamond Bar High School, 21400 Pathfinder Rd, Diamond Bar, CA 91765, USA

ABSTRACT

This study examines gender-based differences in financial behavior, focusing on various bank transactions and demographic factors using a large dataset from Kaggle, consisting of over one million entries. By applying multiple linear regression models and permutation feature importance rankings, the study explores how different variables, such as age, geographic location, and transactional timing, influence banking patterns across genders. The findings reveal that age is a significant predictor of transaction behavior for males, while it is less impactful for females. Geographic location, particularly "Location_West" and "Location_Other," plays a crucial role for males but has minimal influence on females. For females, transactional timing, indicated by "TransactionHour," shows more importance in predicting banking behaviors. In contrast, gender itself does not significantly affect transaction outcomes when controlling for other variables. Overall, the study highlights the importance of demographic and contextual factors, such as age and geography, over inherent gender-based differences. These insights provide valuable guidance for financial institutions aiming to tailor their services more effectively to meet the needs of diverse customer segments. The results emphasize the need for data-driven approaches to better understand gender-specific financial behavior and improve service offerings.

Keywords: Gender; Transaction Patterns; Financial Behavior; Multivariate Linear Models; Customer Satisfaction

INTRODUCTION

Bank transactions are fundamental to financial activities, involving a wide range of types, purposes, and frequencies (1). These transactions can be categorized into deposits, withdrawals, transfers, and payments, each serving distinct financial purposes such as saving, investing, daily expenditures, or bill payments (2). The frequency of transactions varies greatly depending on individual financial behaviors, economic conditions, and technological advancements. For instance, daily transactions often involve retail purchases or ATM withdrawals, while less frequent transactions may relate to larger sums, such as investments or loan repayments.

Among the many factors influencing bank transactions, gender has emerged as a significant determinant of

Corresponding author: Nathan Ma, E-mail: nnathanma@gmail.com.

Received September 25, 2024**; Accepted** October 10, 2024
https://doi.org/10.70251/HYJR2348.23145154

financial behavior. Research suggests that men and women exhibit distinct patterns in how they engage with financial services. For example, women are generally more inclined to prioritize saving and budgeting, whereas men may demonstrate a greater propensity for risk-taking and investments (3). Understanding these gender-based differences is crucial for financial institutions aiming to tailor their services and products to meet the needs of diverse customer groups.

Several studies have previously explored the relationship between gender and banking behavior. Brown and Green (4) examined banking penetration among genders and the subsequent usage patterns. Jones (5) focused on how gender influences investment choices and risk-taking behaviors through a global survey. Clark, Doe, and White (6) investigated gender disparities in the adoption and use of online banking platforms, while Smith (7) studied how financial literacy varies between genders and its impact on financial decision-making.

While these studies have enhanced the understanding of gender-based differences in financial behavior, they are often limited by relatively small sample sizes. To address this limitation, the current study utilizes a large dataset from Kaggle, comprising over one million entries from various financial institutions, and includes a diverse set of variables. This study applies three multiple linear models and ranking tool based on permutation feature importance to analyze gender-based differences in bank transactions across various scenarios.

The primary contribution of this research is to provide a deeper understanding of gender-specific patterns in banking transactions, offering actionable insights that can help financial institutions improve their service offerings and more effectively cater to the diverse needs of their customers.

DATA DESCRIPTION

The bank customer segmentation dataset utilized in this research comprises over 1,000,000 entries, each providing detailed financial and demographic information essential for understanding customer behavior and financial trends within the banking sector. The dataset includes key variables such as age, gender, geographic location, account balances, transaction amounts, and other relevant financial metrics, offering a comprehensive foundation for segmenting banking customers and analyzing their financial activities.

Prior to conducting the analysis, a rigorous data cleaning process was implemented to ensure the accuracy, reliability, and integrity of the dataset. This involved several critical steps:

- Removal of Entries with Missing Age Values: Since age is a pivotal variable for customer segmentation, entries lacking age data were excluded to avoid skewing age-based analyses and ensure meaningful segmentation results.
- Exclusion of Entries with Zero Balances: Accounts with zero balances were removed from the dataset. These entries could distort percentage-based calculations, such as the account balance ratio, and may not reflect active customer behavior, thus diminishing the relevance of their inclusion.
- Validation of Transaction Data: Transaction amounts and other financial details were checked for inconsistencies or outliers to maintain a high level of data reliability and prevent erroneous results in predictive modeling and segmentation.

The resulting refined dataset is well-suited for identifying distinct customer groups within the banking population, supporting the development of targeted marketing strategies and personalized customer service approaches. By leveraging detailed financial information, this dataset enables the identification of key customer segments based on behaviors such as spending patterns, savings habits, and transactional preferences.

Furthermore, the dataset supports advanced predictive modeling techniques aimed at anticipating future customer needs and trends. This includes forecasting potential financial product uptake, identifying customers at risk of attrition, and predicting changes in transactional behavior. As such, it serves as a critical resource for banks aiming to enhance customer retention and service personalization.

The dataset is publicly available on Kaggle (8), where it has gained significant traction in the research community. With over 110,000 views and more than 12,000 downloads, the dataset's popularity attests to its relevance and utility in various financial and academic contexts. Its large volume and extensive variables align well with the objectives of this research, which seeks to analyze broad banking customer trends with statistical rigor.

In conclusion, the bank customer segmentation dataset provides a rich and reliable foundation for this study. By utilizing over one million entries and a refined, clean dataset, this research is well-positioned to yield meaningful insights into customer segmentation, financial behavior, and predictive modeling in the banking industry. Table 1 provides more detailed information on the key variables and characteristics of the dataset.

Table 1. Variable Definitions and Summary Statistics for Bank Transaction Data

Variables	Type	Definition	Descriptive statistics
TransactionID	Categorical	A unique identifier for each transaction in the dataset.	**Count**: 1,048,567
CustomerID	Categorical	A unique identifier assigned to each customer in the dataset.	**Unique**: 884,265
CustomerAge	Numerical	Age of the customer, derived from the date of birth.	**Mean**: 35; **SD**: 12; **Min**: 18; **Max**: 67
CustGender	Categorical	The gender of the customer.	**Male**: 48%; **Female**: 50%; **Other**: 2%
TransactionAmount	Numerical	The amount of money transacted, denoted in INR.	**Mean**: 1,574 INR; **SD**: 3,000; **Min**: 10; **Max**: 15,000
TimeCategory	Categorical	The category of time the transaction (Morning, Afternoon, Evening).	**Morning**: 30%; **Afternoon**: 50%; **Evening**: 20%
LocationCategory	Categorical	The geographical location category of the transaction.	**North**: 25%; **South**: 25%; **East**: 20%; **West**: 20%; **Other**: 10%
Weekday_weekend	Categorical	Indicates whether the transaction occurred on a weekday or a weekend.	**Weekday**: 75%; **Weekend**: 25%

Notes: 1. SD represents standard deviation; 2. INR represents Indian Rupees.

METHODOLOGIES

Pearson correlation Matrix

A correlation matrix is a table that displays the Pearson correlation coefficients between multiple variables, allowing us to assess the linear relationships between predictors and the target variable. The Pearson correlation coefficient between two variables, X and Y, is calculated as (9):

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (1)$$

where $\overline{X}$ and $\overline{Y}$ and are the means of variables X and Y, and n is the number of observations. This coefficient r_{XY} ranges from -1 to 1, where values close to 1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship, and values close to 0 suggest no linear relationship. A correlation matrix provides a compact summary of these relationships, allowing us to easily spot which predictors are most strongly correlated with the target variable.

In the context of this analysis, the correlation matrix helps identify which predictors have the strongest linear associations with the target variable, thereby guiding feature selection and model interpretation. A high positive correlation (close to 1) between a predictor and the target indicates that as the predictor increases, the target also tends to increase, which could suggest that this predictor is important for predictive modeling. On the other hand, a strong negative correlation (close to -1) implies an inverse relationship. Variables that show very low or no correlation with the target may contribute little to the model's predictive power, unless there are complex non-linear interactions, which would not be captured by the correlation coefficient. Therefore, in the paper, constructing and analyzing the correlation matrix before building models allow us to understand the underlying data relationships and can serve as an initial feature screening step before moving into more advanced modeling techniques.

Multiple Linear Regression Models

A multiple linear regression model is used to analyze the relationship between a dependent variable, in this case, the account-to-balance ratio, and several independent predictors, such as gender, location, transaction hour, and other related factors. The general form of the multiple linear regression equation can be written as (10):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \quad (2)$$

where Y is the dependent variable (account-to-balance ratio), β_0 is the intercept, β_1, β_2, …, β_p are the coefficients for each predictor X_1, X_2, …, X_p and ϵ represents the error term. Each coefficient β_i represents the change in Y for a one-unit change in X_i, assuming all other predictors remain constant. For example, the coefficient for gender will indicate the difference in the account-to-balance ratio between genders, controlling for other variables such as location or transaction hour.

In evaluating the multiple regression model, information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) provide a robust way to assess model fit and compare models with different sets of predictors. Both AIC and BIC penalize the inclusion of additional variables to discourage overfitting. AIC is calculated as:

$$AIC = 2k - 2\ln(L) \quad (3)$$

where k is the number of parameters and L is the likelihood of the model. BIC, on the other hand, introduces a stronger penalty for models with more parameters and is defined as:

$$BIC = \ln(n)\, k - 2\ln(L) \quad (4)$$

where n is the number of observations. Lower AIC and BIC values indicate better model performance. The choice between AIC and BIC depends on the trade-off between goodness of fit and model complexity, with BIC being more conservative as it penalizes additional parameters more strongly. In the case of modeling the account-to-balance ratio, comparing the AIC and BIC values across different models will help determine the most parsimonious model—one that balances explanatory power with simplicity. Furthermore, significance testing through t-tests can help assess whether variables like location or transaction hour contribute significantly to the model, while AIC and BIC provide guidance on the overall performance of the model.

Permutation Feature Importance

There are a wide range of methods for evaluation the ranking of variable importance. One of the most popular ones, Permutation feature importance (11), provides an intuitive way to measure the contribution of each variable in a predictive model by observing how the model's performance changes when the values of a specific feature are randomly shuffled. The underlying principle is that if permuting the values of a feature results in a significant drop in model accuracy, then that feature plays a critical role in the model's predictions. Mathematically, let f(x) represent the model's prediction, and R_i be the feature importance score for feature i. The importance is quantified by measuring the change in a chosen performance metric (such as accuracy or mean squared error) before and after permuting the feature:

$$R_i = \frac{1}{n} * \sum_{k=1}^{n} \left(L(y_k, f(x_k)) - L(y_k, f(x_k^{perm,i}))\right) \quad (5)$$

Where $L(y_k, f(x_k))$ is the loss function for observation k, and $x_k^{perm,i}$ is the model's prediction after feature i has been permuted. A larger R_i indicates higher importance, implying that the feature is more critical in determining the outcome.

In the present paper, Permutation feature importance was selected over other variable importance ranking methods, such as coefficient-based or tree-based importance, because of its intuitive nature and model-agnostic approach. Unlike coefficient-based methods, which are limited to linear models and rely on assumptions about the linearity of relationships between variables, permutation importance can be applied to any model, including complex machine learning models like decision trees, neural networks, or ensemble methods. Furthermore, tree-based importance methods, such as those derived from decision trees or random forests, can sometimes be biased toward variables with more categories or higher cardinality, leading to misleading importance rankings. Permutation importance directly assesses the impact of each variable by observing how randomizing its values affects model performance, offering a more robust and interpretable measure of feature contribution. This method provides a more accurate representation of how each feature influences the predictions, making it particularly suitable when working with a variety of model types and feature structures.

RESULTS

The detailed research results are presented in order in the following subsections to demonstrate the transactions among different genders and other predictors.

Correlation Matrix Results

The correlation matrices for both genders reveal similar patterns (as shown in Figures 1 and 2, with weak relationships across most variables. For both males and females, the Amount_balance_ratio shows no significant correlation with any other variables. Age exhibits weak

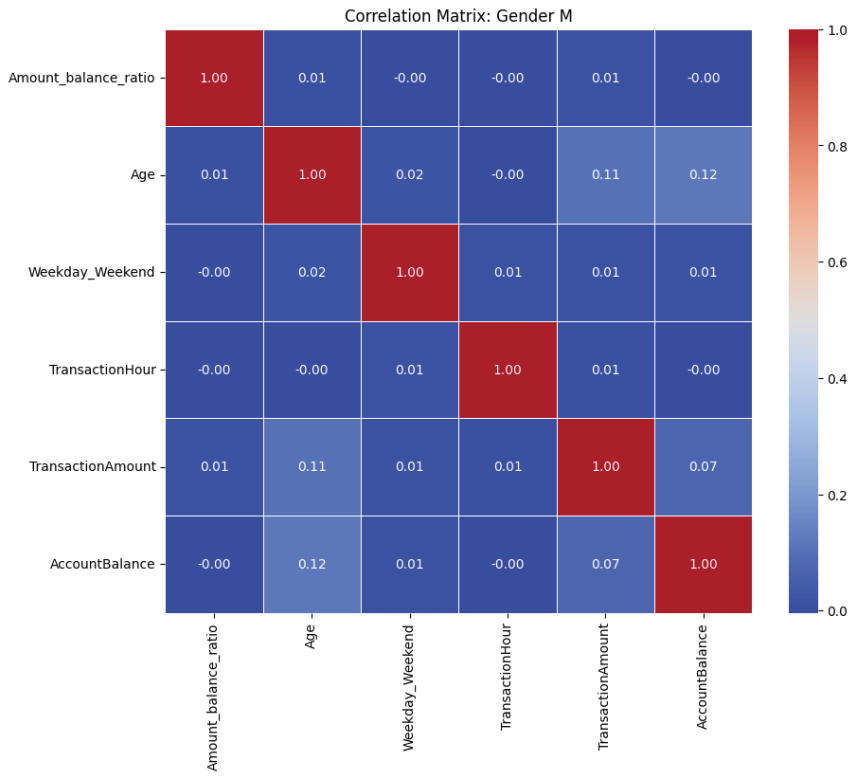

Figure 1. Correlation Matrix Results for Males Only.

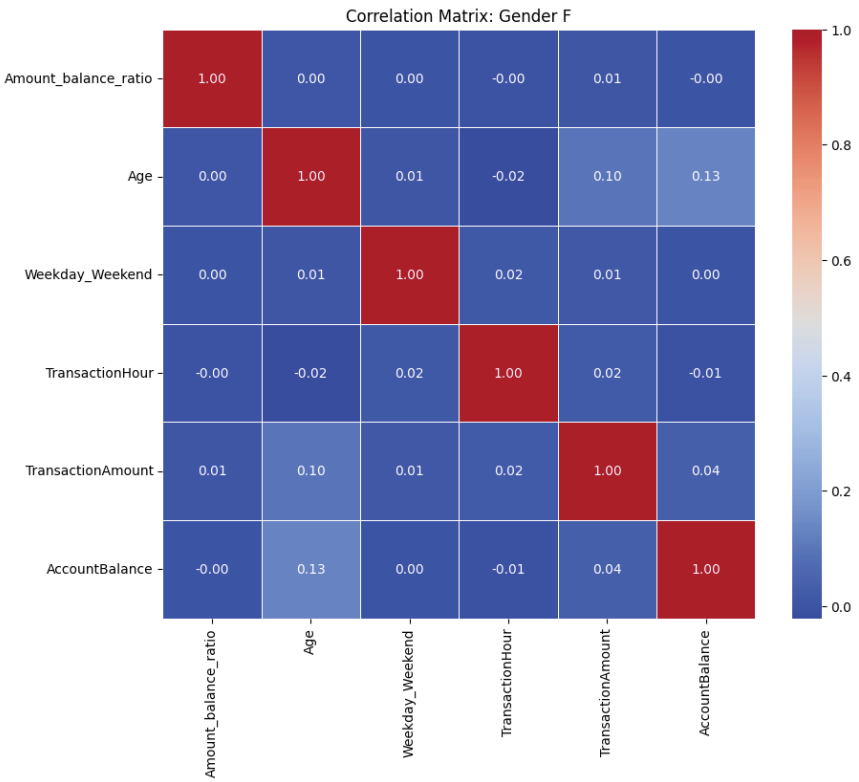

Figure 2. Correlation Matrix Results for Females Only.

positive correlations with TransactionAmount (0.11 for males and 0.10 for females) and AccountBalance (0.12 for males and 0.13 for females), indicating a slight tendency for older individuals to have higher transaction amounts and account balances, with this effect being marginally stronger for females. TransactionAmount also shows a weak correlation with AccountBalance for both genders, though this relationship is somewhat stronger for males (0.07) compared to females (0.04).

Other variables such as Weekday_Weekend and TransactionHour show no meaningful correlations with the other variables for either gender. Overall, while there are minor differences in the strength of correlations between certain variables (e.g., the relationship between TransactionAmount and AccountBalance), the general patterns of correlation are consistent across both genders, with no strong relationships observed. These findings suggest that demographic features such as Age have only a modest influence on financial behavior, regardless of gender.

Correlation Matrix Results

The results of the three models—Females Only (Table 4), Males Only (Table 3), and All Genders (Table 2)—highlight differences in how various factors impact the dependent variable across gender segments. Age emerges as a significant predictor in the models for males and all genders, with a stronger effect in the male-only model (Coef = 486.44, P = 0) compared to the female-only model, where it is not statistically significant (Coef = 196.48, P = 0.171). Interestingly, while the constant term and several variables such as Weekday_Weekend and TransactionHour show no significant impact across any of the models, the coefficients for location variables vary between genders. However, none of the locations demonstrate statistical

Table 2. Multiple Linear Model Results based on All Genders

Variables	coef	std err	t	P>|t|
const	-4728.55	7179.72	-0.66	0.51
Age	401.12	74.53	5.38	0.00
Weekday_Weekend	-70.26	1399.88	-0.05	0.96
TransactionHour	-87.79	134.36	-0.65	0.51
location_East	-4174.64	7249.11	-0.58	0.57
location_North	8074.39	6473.25	1.25	0.21
location_Other	-834.63	6140.91	-0.14	0.89
location_South	-2958.24	6339.15	-0.47	0.64
location_West	1992.82	6303.56	0.32	0.75
Gender_M	-1376.87	1446.41	-0.95	0.34
Log-Likelihood: -1.4535e+07				
No. Observations: 982828; AIC: 2.907e+07				
Df Residuals: 982818; BIC: 2.907e+07				

Table 3. Multiple Linear Model Results based on Males Only

Variables	coef	std err	t	P>|t|
const	-6517.71	8103.03	-0.80	0.42
Age	486.44	87.16	5.58	0.00
Weekday_Weekend	-1426.16	1624.30	-0.88	0.38
TransactionHour	-31.22	151.78	-0.21	0.84
location_East	-5787.74	8338.17	-0.69	0.49
location_North	11710.00	7388.56	1.59	0.11
location_Other	-2329.96	6971.07	-0.33	0.74
location_South	-4810.52	7208.65	-0.67	0.51
location_West	-1158.67	7168.32	-0.16	0.87
Log-Likelihood: -1.0498e+07				
No. Observations: 710557 AIC: 2.100e+07				
Df Residuals: 710548 BIC: 2.100e+07				

Table 4. Multiple Linear Model Results based on Females Only

Variables	coef	std err	t	P>|t|
const	-4949.05	14600.00	-0.34	0.74
Age	196.48	143.65	1.37	0.17
Weekday_Weekend	3510.93	2749.72	1.28	0.20
TransactionHour	-268.92	284.94	-0.94	0.35
location_East	956.25	14700.00	0.07	0.95
location_North	1832.74	13400.00	0.14	0.89
location_Other	4217.57	12800.00	0.33	0.74
location_South	2783.83	13200.00	0.21	0.83
location_West	10990.00	13100.00	0.84	0.40
Log-Likelihood: -4.0370e+06				
No. Observations: 272271; AIC: 8.074e+06				
Df Residuals: 272262; BIC: 8.074e+06				

significance, suggesting that geographic differences do not play a critical role in predicting the outcome. Moreover, in the all-genders model, gender itself is not a significant predictor, indicating that controlling for other factors minimizes the differences in the dependent variable based on gender.

The overall model fit, indicated by Log-Likelihood, AIC, and BIC values, shows that the model for males has better fit metrics compared to the female-only model, with the all-genders model fitting the largest dataset. This reflects better predictive power for the male model, driven mainly by the influence of age. Across all models, the impact of transactional timing and geographic location appears minimal. These results suggest that while age is a strong determinant, other variables such as location and time of transaction do not significantly influence the predictions, and the effect of gender diminishes when controlling for other factors.

Variable Importance Ranking Results

In comparing the variable importance across the three models (Females Only, Males Only, and All Genders), distinct differences emerge in terms of how features impact predictions. In the female-only model (Figure 5), Age and TransactionHour are the most important predictors, indicating that age and the time of transaction play a significant role in influencing the outcomes for females. Geographic variables like Location_West and Weekday_Weekend also contribute to the predictions but with less importance than age. In contrast, Location_East, Location_South, and Location_North show very little influence, suggesting that geographic differences are less critical in this model.

For males (Figure 4), the most important variables are Location_Other and Location_West, indicating that geographic location has a stronger influence on predictions for males than it does for females. Age plays a smaller role compared to the female-only model, while TransactionHour and Weekday_Weekend have minimal or even negative importance. In the all-genders model (Figure 3), geographic variables like Location_West and Location_Other remain the most influential, similar to the male-only model, while TransactionHour shows negative importance. Age has a moderate influence, and Gender_M contributes minimally, aligning with the regression results that suggest gender alone does not significantly impact predictions. These findings illustrate how age and transactional timing are more influential for females, while geographic location plays a more significant role for males and in the combined model.

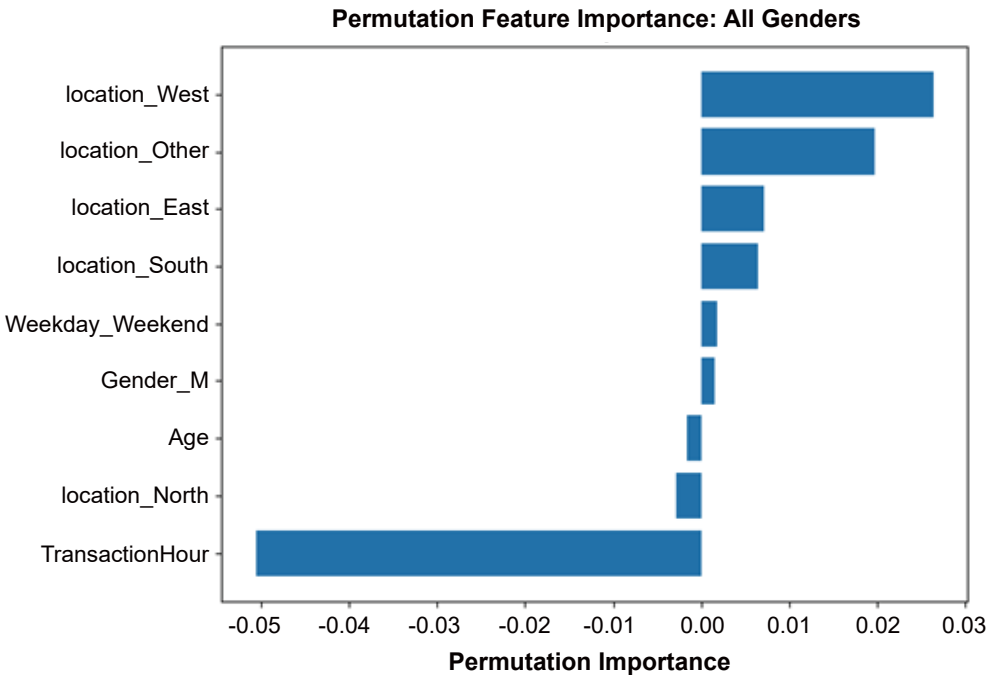

Figure 3. Permutation Feature Importance Results for All Genders.

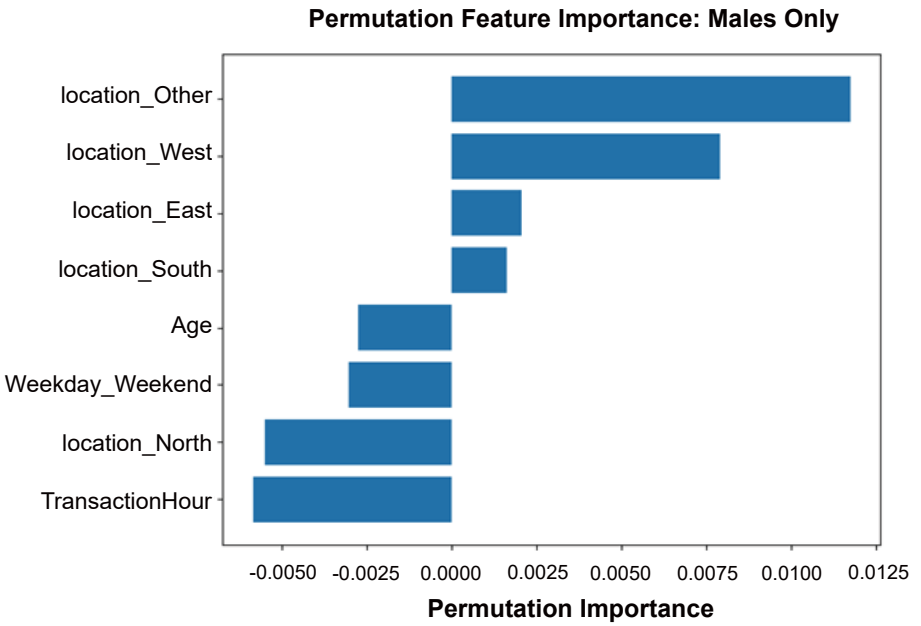

Figure 4. Permutation Feature Importance Results for Males Only.

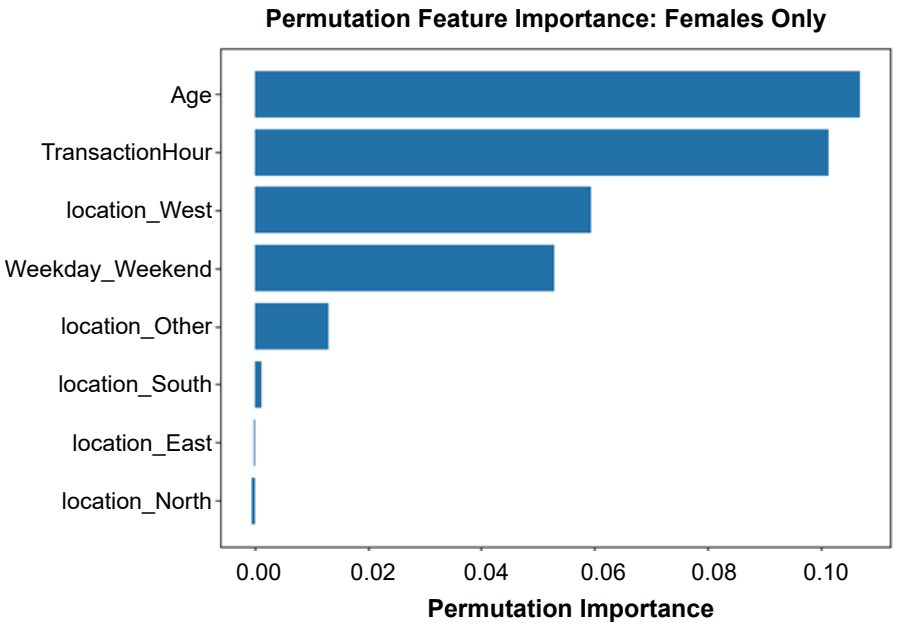

Figure 5. Permutation Feature Importance Results for Females Only.

CONCLUSION AND RECOMMENDATIONS

This study explored the gender-based differences in financial behavior, focusing on various bank transaction types, demographic factors, and geographic variables. By utilizing a large dataset from Kaggle and employing multiple statistical models, including three linear regression models and variable rankings through permutation feature importance, the author sought to uncover patterns and predictors of banking behavior across gender segments. The findings provide insights into the distinct factors that influence transaction behaviors for males, females, and all genders combined.

One key finding is the consistent but weak correlation between Age and financial behavior across both genders, with slightly stronger correlations observed in females. Age was a significant predictor in the models, particularly in the male-only model, indicating that older individuals tend to have higher transaction amounts and account balances. Interestingly, while age was influential in the male and all-genders models, it was not significant for females. This difference suggests that financial behaviors associated with age may be more prominent among males than females, highlighting the importance of age in gender-specific financial modeling.

Geographic location emerged as an important predictor, particularly for males. Location variables such as "Location_West" and "Location_Other" were significant in the male-only and all-genders models, while they were less influential for females. This suggests that geography plays a more significant role in determining financial behavior for men, potentially reflecting regional differences in banking services or access. In contrast, transactional timing variables like "TransactionHour" were more important for females, indicating that the time of the transaction may play a greater role in financial decision-making for women.

The analysis also revealed that when other factors such as age and geographic location are controlled, gender itself does not appear to be a significant predictor of banking behavior. This suggests that differences in financial transactions attributed to gender may be largely influenced by demographic and contextual factors rather than inherent gender-based differences in financial behavior. This finding underscores the importance of considering multiple dimensions of customer behavior when analyzing financial data, rather than focusing solely on gender.

With the above conclusive information, the following recommendations are proposed for the future practice or research:

1. Tailored Financial Services: Financial institutions should consider offering tailored services that reflect the differences in financial behavior across genders. For example, since geographic location plays a more critical role in predicting behavior for males, banks could explore region-specific services or marketing strategies that cater to local needs, particularly in areas like "Location_West" and "Location_Other." For females, focusing on transactional timing and encouraging usage during specific hours may enhance engagement and service utilization.

2. Focus on Age-Based Financial Products: Given the significant role of age in influencing financial behavior, especially for males, banks should develop age-targeted financial products and services. These could include savings plans, investment options, or financial literacy programs aimed at older individuals, who tend to have higher transaction amounts and account balances.

3. Further Research on Geographic Influence: The findings show that geographic location is a stronger predictor for males than females. Financial institutions should conduct further research into why location influences male banking behavior more significantly. This could include studies on regional access to financial services, local economic conditions, or cultural factors affecting financial decisions.

4. Enhanced Data-Driven Decision Making: The study highlights the importance of data-driven decision-making in financial services. Financial institutions should continue to leverage large datasets and advanced statistical methods like random forests to identify and rank key factors influencing customer behavior. This approach will allow for more precise and effective service offerings that cater to diverse customer segments.

In conclusion, while gender may not independently drive significant differences in financial behavior, demographic factors such as age and geographic location play substantial roles. By focusing on these factors and tailoring services accordingly, financial institutions can enhance their offerings and better serve the diverse needs of their customer base.

ACKNOWLEDGMENT

I extend my deepest gratitude to the contributors of the dataset used in this study, which enabled the in-depth analysis of over 1 million bank transactions. The richness of the data provided a strong foundation for uncovering key insights into gender disparities in financial behavior. I

am also grateful to the anonymous reviewers and editors for their constructive feedback, which greatly enhanced the quality of this work.

REFERENCES

1. Merton RC & Bodie Z. A conceptual framework for analyzing the financial system. The global financial system: A functional perspective. 1995; 3-31.
2. Nofal S. Identifying highly-valued bank customers with current accounts based on the frequency and amount of transactions. Heliyon. 2024; 10 (13). https://doi.org/10.1016/j.heliyon.2024.e33490
3. Smith JA. Gender differences in financial literacy and usage of banking services. Journal of Financial Studies, 2020; 15 (3): 123-145.
4. Brown PL & Green MS. Exploring gender differences in online banking adoption. Journal of Digital Finance. 2017; 10 (1): 34-56.
5. Jones, M. B. (2018). The influence of gender on investment decisions: Evidence from a global survey. International Journal of Finance, 22(2), 98-115.
6. Clark RE, Doe JL & White ST. Gender and financial inclusion: An analysis. In P. Black (Ed.), Proceedings of the International Conference on Financial Inclusion. 2019; pp. 45-67. Conference Press.
7. Rencher AC. *Methods of Multivariate Analysis*. Wiley. 2002. https://doi.org/10.1002/0471271357
8. Strang G. *Introduction to Linear Algebra*. Wellesley-Cambridge Press. 2009.
9. Aduda J & Kingoo N. The relationship between electronic banking and financial performance among commercial banks in Kenya. Journal of finance and investment analysis. 2012; 1 (3): 99-118.
10. Tranmer M & Elliot M. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR). 2008; 5 (5): 1-5.
11. Altmann A, Toloşi L, Sander O & Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010; 26 (10): 1340-1347. https://doi.org/10.1093/bioinformatics/btq134