Forecasting Consumer Spending Amounts Using Machine Learning and Time Series Analysis

Lo, Yi*

January 31, 2025

Abstract

In the modern digital economy, accurately predicting consumer spending has become a critical task for businesses and financial institutions to effectively manage risk, optimize marketing strategies, and enhance financial planning (Agrawal, 2013). This study develops a machine learning model combined with time series analysis to predict consumer spending amounts based on socioeconomic and transactional data. Key features include income, age group, education level, industry category, and transaction date components, which provide seasonal and cyclical insights into consumer behavior (Author, 2017). By leveraging advanced regression techniques, including Random Forest and Long Short-Term Memory (LSTM) neural networks, we achieve robust predictions of spending amounts over time. Our findings align with prior research, suggesting that incorporating time-related variables significantly improves prediction accuracy, offering actionable insights into seasonal consumer behavior patterns (Author, 2020). This research contributes to the growing literature on data-driven decision-making frameworks for financial and marketing applications.

Keywords: *Consumer Spending Prediction, Machine Learning, Time Series Analysis, Financial Forecasting, Random Forest, LSTM, Seasonal Consumption Patterns*

JEL Codes: *C53, E21, G17*

*Department of Information Management and Finance, National Yang Ming Chiao Tung University. Email: roy60404@gmail.com

Contents

1 Introduction

In recent years, predictive modeling of consumer behavior has gained significant attention due to its potential applications in financial planning and marketing. Agrawal (2013) highlights the importance of machine learning in economic prediction, showcasing its ability to extract meaningful insights from complex datasets.

The aim of this study is to predict consumer spending behavior using a combination of exploratory data analysis (EDA) and machine learning techniques. By analyzing detailed socioeconomic and transactional data, we identify critical variables that influence spending patterns, such as income group, age group, education level, and industry category. Furthermore, the integration of time series components into the predictive models allows us to capture seasonal and cyclical trends, thereby enhancing model performance and robustness (Author, 2017).

We employ various machine learning models, including linear regression, K-Nearest Neighbors (KNN), and Long Short-Term Memory (LSTM) neural networks. Each model is evaluated to compare its strengths and limitations in predicting transaction amounts. Additionally, we propose a hybrid approach that combines the predictions of KNN and LSTM, leveraging the advantages of both models to improve overall accuracy.

This research contributes to the growing body of literature on data-driven financial forecasting by demonstrating the importance of time features and demographic segmentation in understanding consumer behavior. The findings have practical implications for businesses and financial institutions, enabling them to optimize strategies, manage risk, and identify seasonal trends in consumer spending.

2 Preliminaries

2.1 Linear Regression

Linear regression is a widely-used statistical model employed to analyze the relationship between a dependent variable Y and one or more independent variables X. It assumes a linear relationship, expressed mathematically as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon, \tag{1}$$

where β_0 represents the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, and ϵ is the error term.

The coefficients β are estimated using the ordinary least squares (OLS) method, which minimizes the sum of squared residuals:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2. \tag{2}$$

Linear regression is chosen for this study because of its strong interpretability, allowing us to quantify the impact of demographic and industry features on transaction amounts. Its simplicity and transparency make it an essential baseline model in predictive analytics.

2.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric and instance-based machine learning algorithm commonly used for regression and classification tasks. In KNN regression, the predicted value $\hat{Y}$ for a data point is calculated as the average of the target values of its k-nearest neighbors:

$$\hat{Y} = \frac{1}{k} \sum_{i \in \mathcal{N}_k} Y_i, \tag{3}$$

where $\mathcal{N}_k$ represents the set of k-nearest neighbors, and Y_i is the target value of the i-th neighbor.

The distance between data points is typically computed using the Euclidean distance formula:

$$d(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}, \tag{4}$$

where x_j and y_j are the j-th features of two data points.

KNN is particularly suited for this study because consumer behavior prediction often assumes that individuals with similar demographic or socioeconomic characteristics exhibit comparable spending patterns. KNN's ability to model such local relationships makes it a valuable tool in understanding consumer behavior.

2.3 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) specifically designed to handle sequential data and address the vanishing gradient problem. LSTM achieves this through memory cells that maintain long-term dependencies, governed by input, forget, and output gates. The operations of an LSTM cell are defined as follows:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right), \tag{5}$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right), \tag{6}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right), \tag{7}$$

$$\tilde{C}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right), \tag{8}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \tag{9}$$

$$h_t = o_t \odot \tanh(C_t), \tag{10}$$

where f_t, i_t, o_t are the forget, input, and output gates, respectively, x_t is the input, h_{t-1} is the previous hidden state, and C_t is the cell state.

LSTM is selected in this study due to its high performance in handling sequential data, such as time-series transaction records. Its ability to capture long-term dependencies and temporal trends makes it an essential model for predicting consumer spending patterns over time.

3 Experimental Settings

3.1 Datasets Description

This study utilizes consumer transactional data obtained from Taiwan's open data platform, accessible at `https://data.gov.tw/en/datasets/all?ct=254`. The dataset comprises three main tables: *ageGroupCombined*, *incomeGroupCombined*, and *educationLevelCombined*. Each table contains detailed demographic and transactional information, enabling a comprehensive exploration of consumer spending patterns across various socioeconomic groups. The dataset covers the period from January 2014 to August 2024 and is segmented by demographic attributes, industry categories, and transaction details. The key features of each table are summarized in Table 1.

Table 1: Key Features of the Dataset Tables

Table Name	Feature	Description
6**ageGroupCombined*	Date	Transaction date in YYYY-MM-DD format.
	Industry	Industry category (e.g., Food, Clothing, Housing).
	Gender	Gender of the consumer (*Male* or *Female*).
	Age Group	Consumer's age range (e.g., *Under 20*, *35-40*).
	Transaction Count	Total number of transactions by consumers in this demographic group.
	Transaction Amount (NTD)	Total transaction amount in New Taiwan Dollars (NTD).
6**incomeGroupCombined*	Date	Transaction date in YYYY-MM-DD format.
	Industry	Industry category (e.g., Food, Clothing, Housing).
	Gender	Gender of the consumer (*Male* or *Female*).
	Income Group	Consumer's annual income range (e.g., *Below 500k*, *1M-1.25M*).
	Transaction Count	Total number of transactions by consumers in this demographic group.
	Transaction Amount (NTD)	Total transaction amount in New Taiwan Dollars (NTD).
6**educationLevelCombined*	Date	Transaction date in YYYY-MM-DD format.
	Industry	Industry category (e.g., Food, Clothing, Housing).
	Gender	Gender of the consumer (*Male* or *Female*).
	Education Level	Highest educational attainment of the consumer (e.g., *High School*, *Bachelor*).
	Transaction Count	Total number of transactions by consumers in this demographic group.
	Transaction Amount (NTD)	Total transaction amount in New Taiwan Dollars (NTD).

3.1.1 Key Insights from the Dataset

The dataset provides a comprehensive view of consumer transactional data segmented by demographics (age, income, education), industries, and time. Its rich granularity allows for exploratory data analysis (EDA) and predictive modeling to uncover trends, correlations, and patterns in consumer behavior.

3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) plays a pivotal role in understanding the underlying patterns and relationships within the dataset. To evaluate the research questions outlined below, a series of statistical and visualization techniques were employed:

1. **Do key variables (e.g., Age, Income, Education) significantly impact the target variable (transaction amount)?**

   This question seeks to determine whether demographic attributes, such as age, income, and education levels, exhibit significant correlations with consumer spending. By analyzing these relationships, we aim to identify key variables that strongly influence the target variable, *Transaction Amount (NTD)*. A correlation matrix was computed to quantify the degree of association between these demographic attributes and the target variable.

2. **Do industries influence the relationship between key variables and the target?**

   Understanding the interaction between industry categories and demographic attributes is essential for uncovering contextual spending patterns. For this analysis, correlations between industry-specific transactions and demographic variables were computed to evaluate how industries mediate the relationship between key variables and transaction amounts. This approach provides insights into which industries exhibit the strongest consumer engagement across different demographic groups.

3. **Does the distribution of industries across different key variables show consistent patterns in their contribution to transaction amounts?**

   To further explore the interaction between demographic attributes and industry categories, a scatter plot analysis was conducted to visualize the contribution of various industries to transaction amounts across demographic segments. This analysis high-

lights patterns in industry preferences and their relative contributions to consumer spending across different age groups, income levels, and education levels.

4. **Does the target variable exhibit any cyclical patterns over time?**

   Temporal patterns in transaction amounts were investigated to identify potential cyclical trends and seasonal effects. A time-series analysis was performed on the aggregated transaction data, segmented by demographic attributes and industry categories. This analysis highlights periods of increased or decreased consumer activity, offering insights into the seasonal or cyclical nature of consumer behavior.

To address these questions, the following EDA methods were applied:

3.2.1 Correlation Analysis of Key Variables with Transaction Amounts

To investigate the relationship between demographic attributes and transaction amounts, we performed a correlation analysis on three key variables: **Age Group**, **Income Group**, and **Education Level**. The results reveal significant insights into how these variables influence consumer spending behavior.

Age Group. The correlation between age groups and transaction amounts is presented in Figure 1. Consumers aged **35–50** exhibited the highest positive correlation with transaction amounts, with coefficients of 0.24 for the $35-40$ age group and 0.27 for the $40-45$ group. These results suggest that middle-aged consumers are the dominant contributors to economic activity, likely due to stable careers and higher financial capabilities.

Conversely, younger consumers (*Under 20* and $20-25$) and older consumers ($70-75$ and *Above 80*) showed weak to negative correlations, with values as low as -0.22. This implies lower purchasing power or a reduced inclination towards high-value transactions among these age groups.

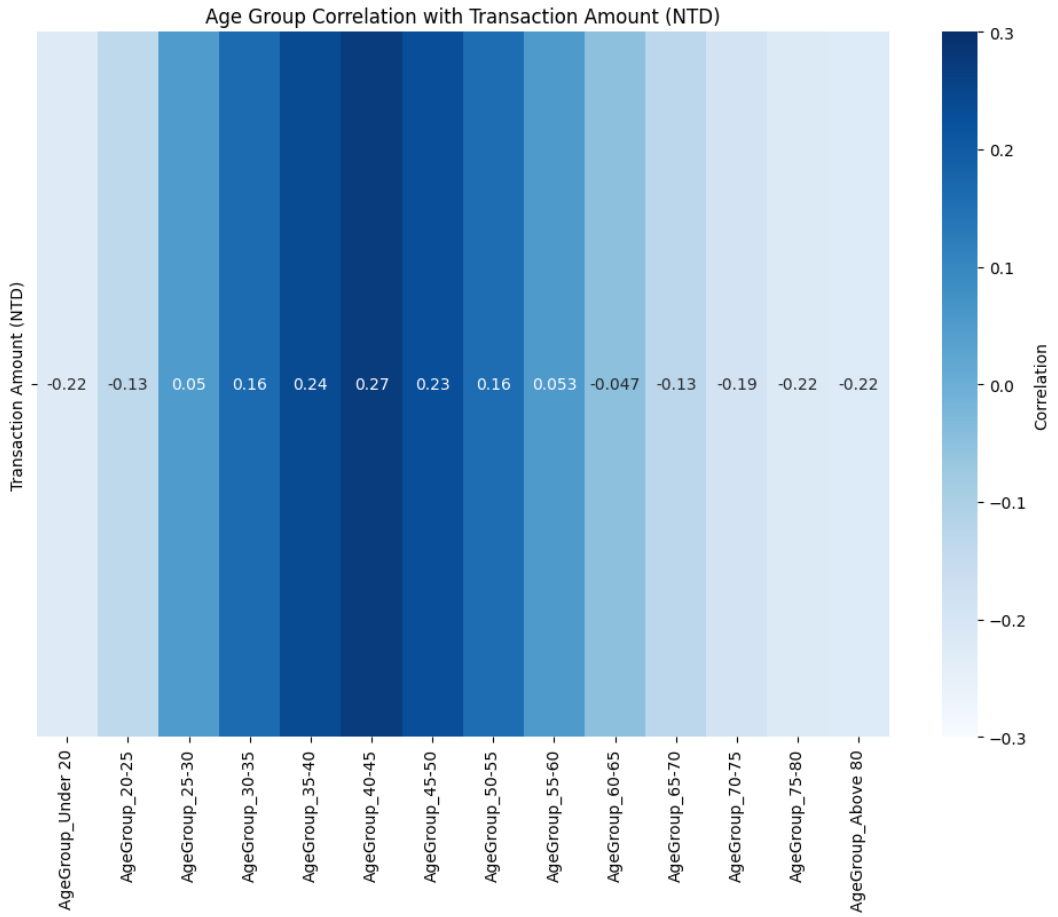


Figure 1: Age Group Correlation with Transaction Amount (NTD)

Income Group. The income group analysis (Figure 2) revealed a strong positive correlation of 0.63 between transaction amounts and the **Below 500k** income group. This surprising result suggests that lower-income groups are more frequent contributors to spending, potentially reflecting essential or frequent transactions.

However, higher-income groups (*Above 2M* and $1.75M - 2M$) demonstrated weak to negative correlations (e.g., -0.20 and -0.19). This indicates that individuals in higher income brackets may exhibit more selective or investment-focused spending behaviors, reducing their overall transaction frequency or amounts in this dataset.

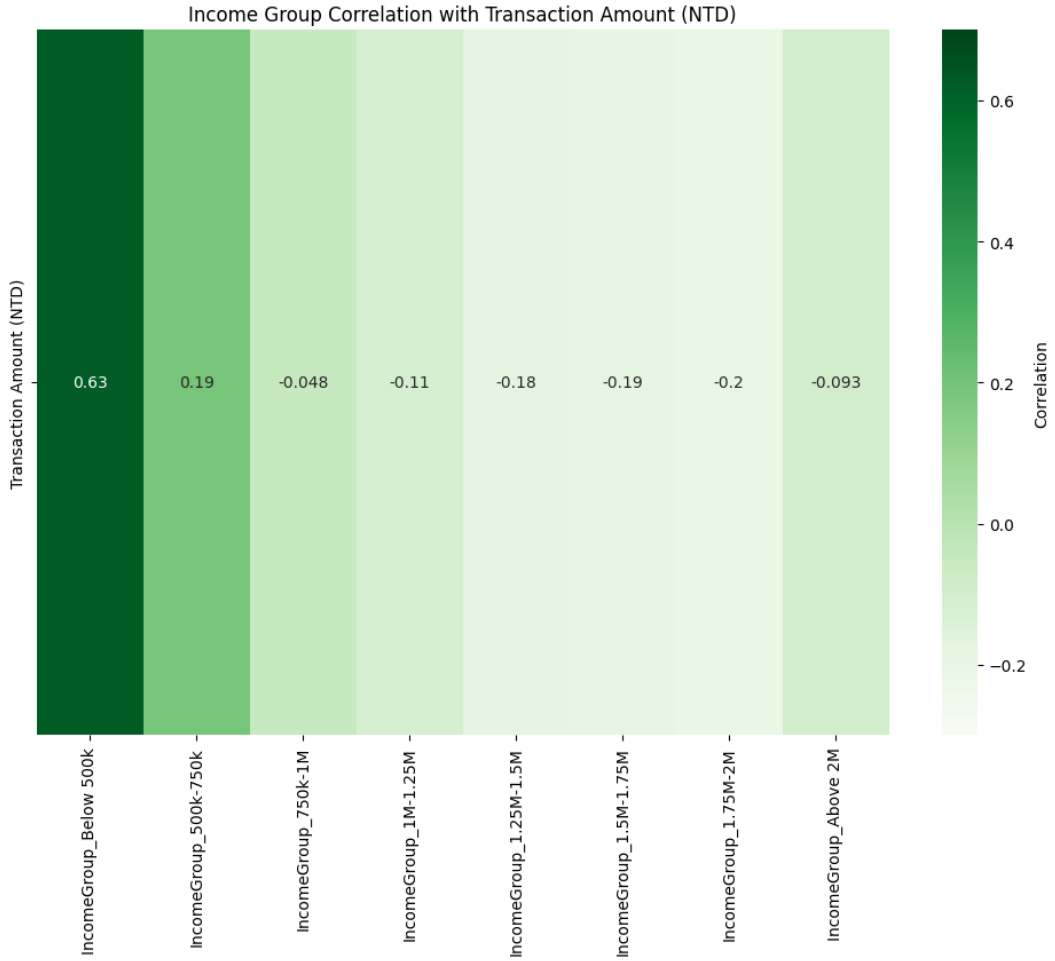


Figure 2: Income Group Correlation with Transaction Amount (NTD)

Education Level. The correlation analysis of education levels and transaction amounts, shown in Figure 3, highlights that individuals with a **Bachelor's degree** exhibited the strongest positive correlation (0.28) with transaction amounts. This suggests that individuals with this level of education are significant contributors to spending, likely due to their higher representation in the working population.

Interestingly, consumers with higher degrees (e.g., *Master* and *Doctorate*) showed weak to negative correlations, with values of -0.058 and -0.20, respectively. This could indicate more conservative financial habits or spending patterns among highly-educated individuals.

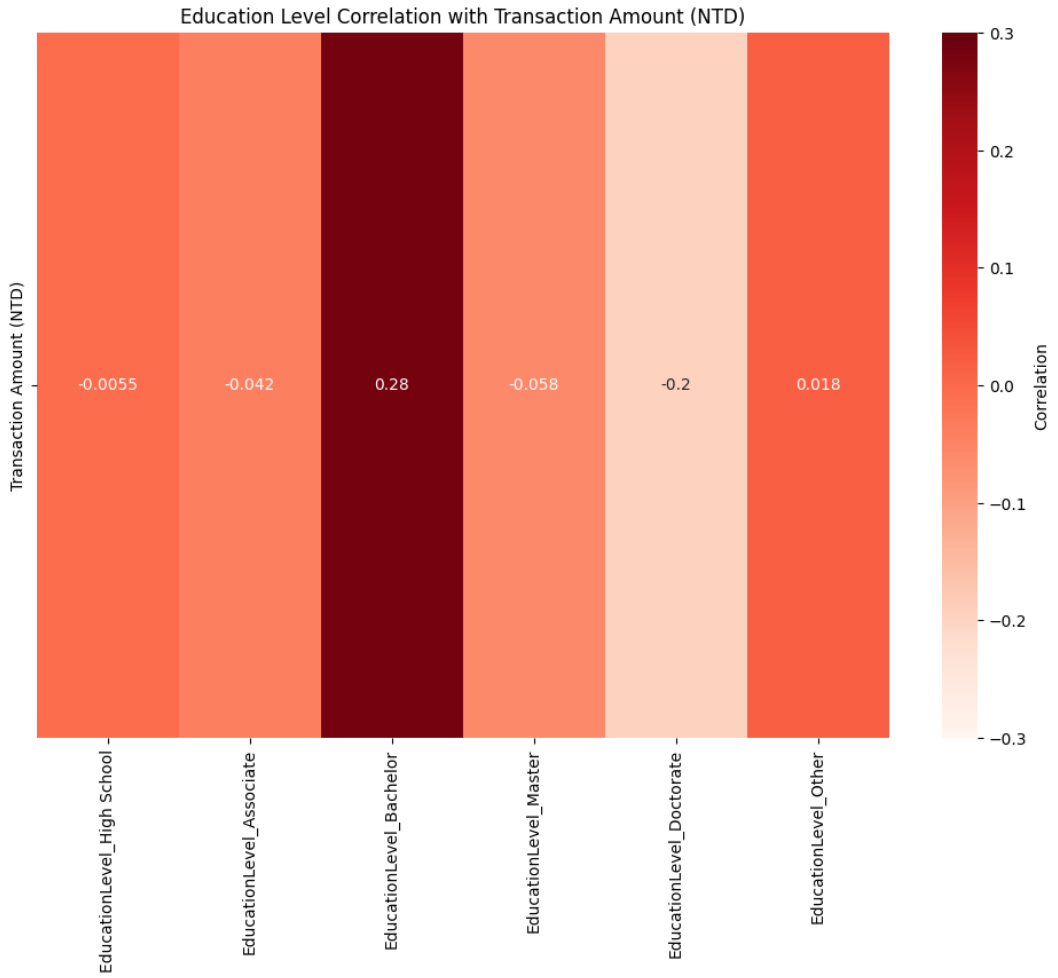


Figure 3: Education Level Correlation with Transaction Amount (NTD)

3.2.2 Correlation Analysis of Industry Categories with Transaction Amounts

We analyzed the correlation between industry categories and transaction amounts across key demographic attributes (age group, income level, and education level). The results, presented in Figure 4, indicate consistent trends:

- **Department Store:** Exhibits the strongest positive correlation with transaction amounts across all demographics.
- **Clothing and Housing:** Show negative correlations, suggesting lower transaction volumes in these industries.

- **Education and Entertainment:** Maintains moderate positive correlations, highlighting its relevance across demographic groups.

These findings confirm that industry-specific spending patterns are largely consistent across age, income, and education groups, with "Department Store" consistently contributing the most to transaction amounts.

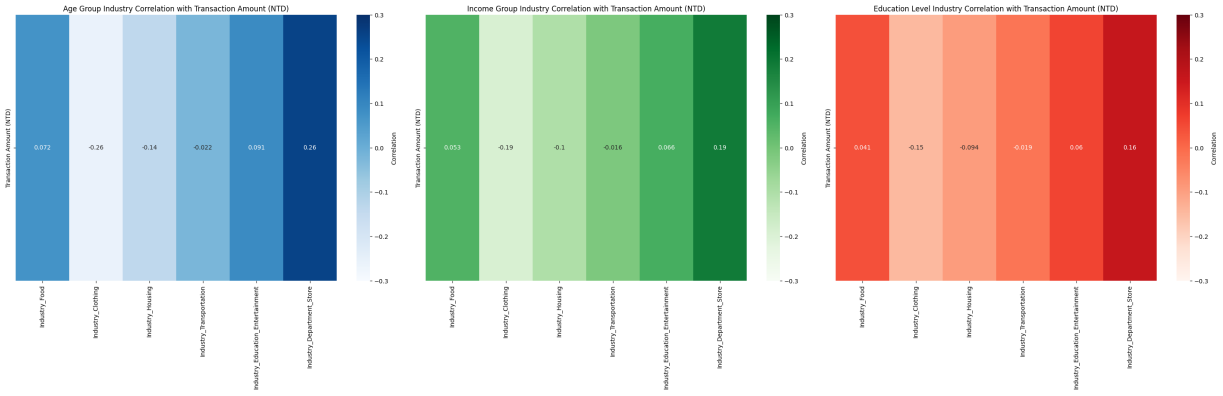


Figure 4: Correlation Between Industry Categories and Transaction Amounts Across Age Group, Income Level, and Education Level.

3.2.3 Impact of Key Variables and Industry Categories on Transaction Amounts

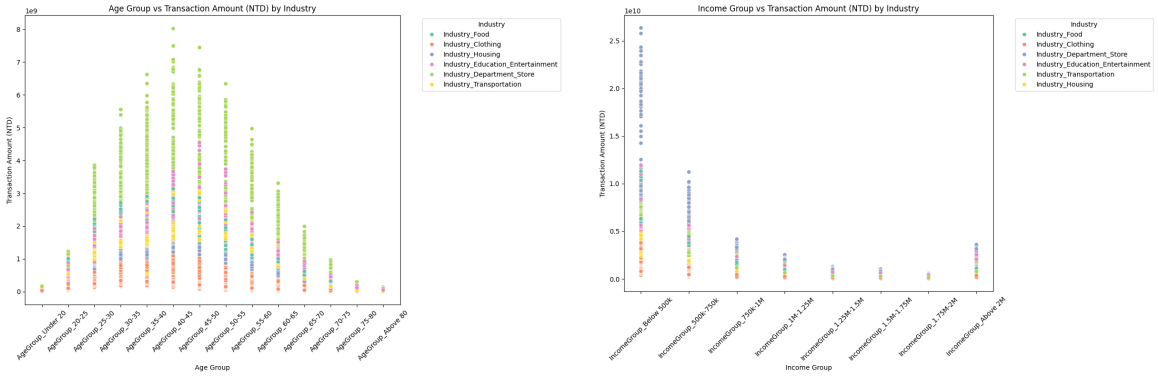
The relationship between transaction amounts and industry categories exhibits significant differences across key variables such as *age group*, *income group*, and *education level*. These differences are illustrated in Figures 5(a), 5(b), and 5(c).

- For **age groups** (Figure 5(a)), industries such as *Department Stores* and *Education/Entertainment* dominate the middle-aged demographic (35–50), with transaction amounts peaking significantly. This suggests that middle-aged consumers contribute the most to these industries.

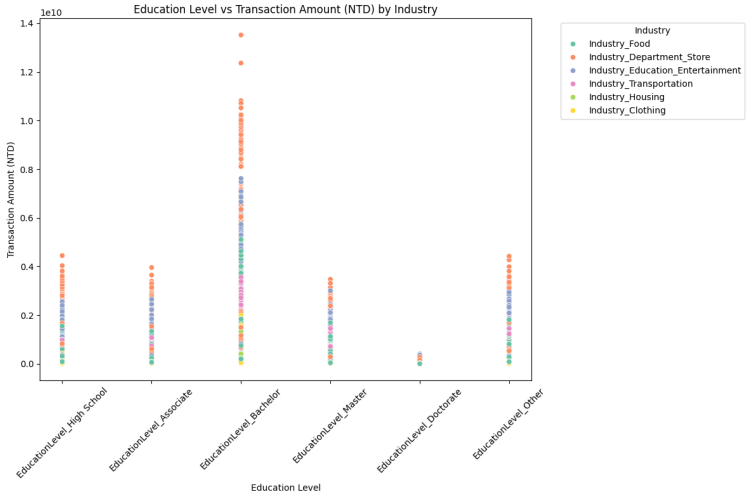
- For **income groups** (Figure 5(b)), *Department Stores* and *Food* industries show higher transaction amounts for lower income groups (*Below 500k*), indicating essential and discretionary spending.

- For **education levels** (Figure 5(c)), *Department Stores* again dominate, particularly among consumers with a *Bachelor's degree*, reflecting higher transaction volumes for this demographic.

These findings indicate that industry distributions vary significantly across key variables, highlighting the need for separate modeling of industries to improve predictive performance. Additionally, from the scatter plots, it is evident that certain outliers exist, particularly in high transaction amounts. These outliers will be addressed during the data preprocessing phase to ensure robust model performance.



(a) Age Group vs Transaction Amount by Industry (b) Income Group vs Transaction Amount by Industry



(c) Education Level vs Transaction Amount by Industry

Figure 5: Impact of Key Variables (Age, Income, Education) on Industry Transaction Amounts

3.2.4 Time Series Analysis to Identify Cycles and Evaluate the Importance of Dates for Accurate Data Splitting

To understand the temporal trends and cyclicality in transaction amounts, time series analysis was conducted on key variables: **Age Group**, **Income Group**, and **Education Level**. Figures 6, 7, and 8 provide detailed visualizations of transaction amounts over time for these groups.

Age Group Trends. Figure 6 illustrates the transaction amounts across different age groups. A general upward trend can be observed for most age groups, particularly among the 40–55 age range, indicating growing economic influence. Seasonal fluctuations and periodic spikes suggest the presence of cyclical trends, likely influenced by annual economic activities such as holidays or fiscal policies.

Income Group Trends. Figure 7 highlights the transaction amount dynamics across income groups. The *Below 500k* income group exhibits a consistently dominant trend, with a noticeable upward shift post-2018. Other income groups follow a similar seasonal pattern, with periodic increases suggesting external economic cycles. The distinct separation between income levels emphasizes their influence on transaction behaviors.

Education Level Trends. As shown in Figure 8, individuals with a *Bachelor's Degree* demonstrate the highest transaction amounts, exhibiting an evident upward trend over time. Lower educational levels, such as *High School* and *Associate*, show slower growth rates but follow similar cyclical patterns. These trends suggest that education levels play a crucial role in consumer spending behaviors, particularly among the more educated groups.

Conclusion and Implications. The time series analysis reveals that each key variable (age, income, and education) demonstrates unique trends while collectively exhibiting periodicity. These cyclical patterns suggest that transaction amounts are influenced by recurring

external factors such as seasonal spending habits, economic cycles, or policy changes. Consequently, time series features must be incorporated into subsequent modeling efforts to ensure accurate prediction performance. Additionally, accurate date-based splitting of training and testing datasets will be essential to account for temporal dependencies.

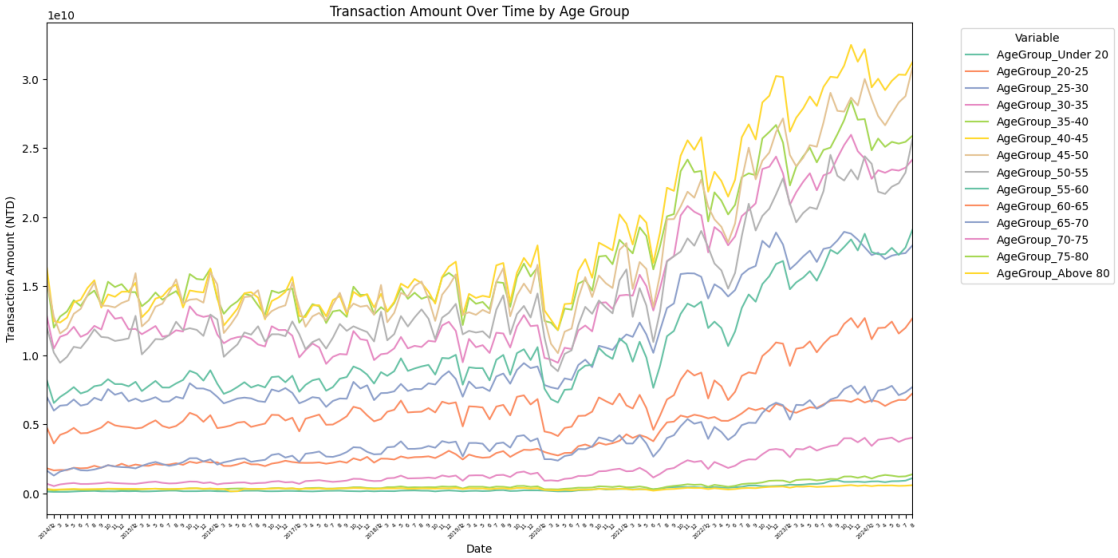


Figure 6: Transaction Amount Over Time by Age Group

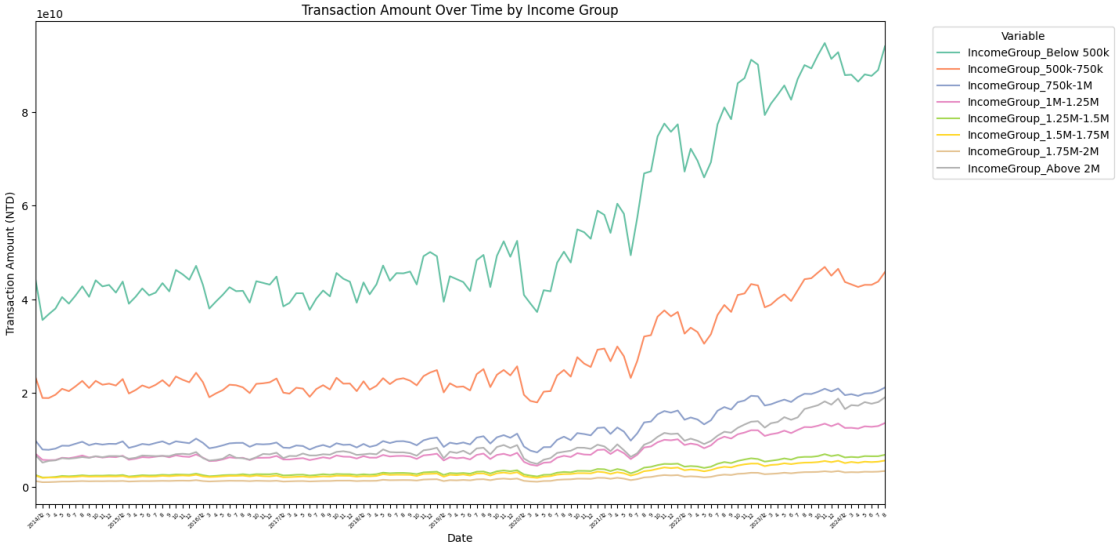


Figure 7: Transaction Amount Over Time by Income Group

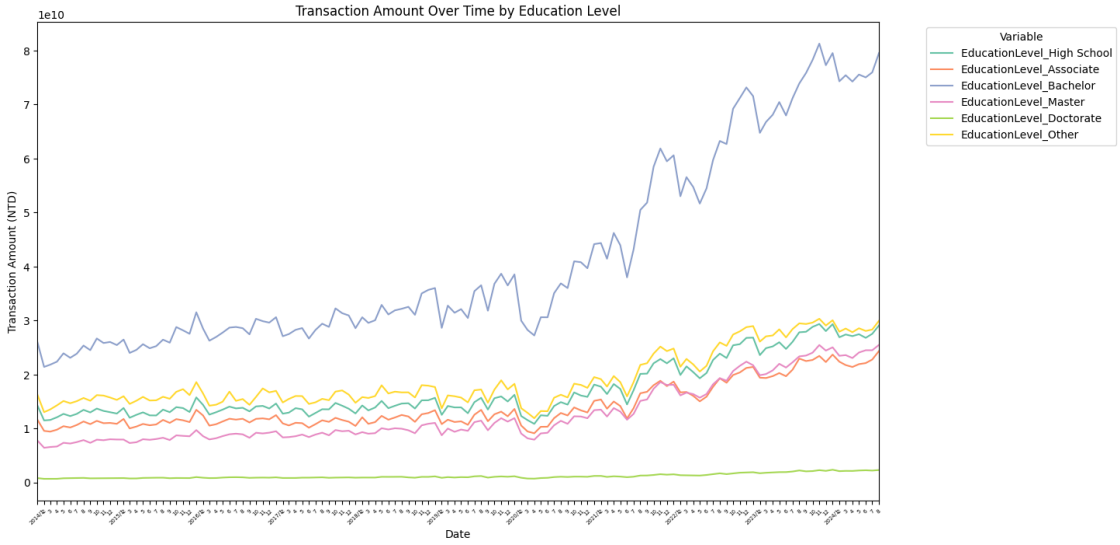


Figure 8: Transaction Amount Over Time by Education Level

3.3 Preprocessing Methods

3.3.1 One-Hot Encoding of Categorical Attributes

One-Hot Encoding is a widely used technique to transform categorical variables into binary indicators, ensuring compatibility with machine learning models. Each unique category is represented as a separate column containing '1' or '0', where '1' indicates the presence of the category and '0' indicates its absence.

The following tables demonstrate an example of applying One-Hot Encoding to the **Age Group** variable. The original dataset includes transaction details and industry types, with the encoded age groups represented as binary columns.

Table 2: Example of One-Hot Encoding for Age Group (Part 1)

Date	Transaction Count	Transaction Amount (NTD)	Industry_Clothing	Industry_Department_Store	Industry_Education_Entertainment	Industry_Food	Industry_Housing	Industry_Others
2014-01-01	6367	5630047	0	0	0	1	0	0
2014-01-01	36983	59655595	0	1	0	0	0	0
2014-01-01	18667	47403285	1	0	0	0	0	0

Example of One-Hot Encoding for Age Group (Part 2)

Date	AgeGroup_Under 20	AgeGroup_20-25	AgeGroup_25-30	AgeGroup_30-35	AgeGroup_40-45	AgeGroup_45-50	AgeGroup_50-55	AgeGroup_55-60	AgeGroup_60-65	AgeGroup_Above 80
2014-01-01	1	0	0	0	0	0	0	0	0	0
2014-01-01	0	0	0	0	1	0	0	0	0	0
2014-01-01	0	1	0	0	0	0	0	0	0	0

Table 2 illustrates how the **Age Group** attribute is transformed using One-Hot Encoding. Each age group becomes an independent column with binary values ('0' or '1'), where '1' represents the presence of the respective group. For instance, the first row belongs to *AgeGroup_Under 20*, while the second row corresponds to *AgeGroup_40-45*.

3.3.2 Remove Outliers

Based on the prior **Exploratory Data Analysis (EDA)**, we observed the presence of significant outliers across multiple attributes, particularly in the target variable, **Transaction Amount (NTD)**. These outliers can distort statistical summaries, reduce model stability, and degrade prediction accuracy. To mitigate this issue, we employ the **Interquartile Range (IQR)** method, a robust statistical technique for detecting and removing extreme values.

The **IQR** is defined as the range between the first quartile (Q_1) and the third quartile (Q_3) of the data distribution, calculated as:

$$\text{IQR} = Q_3 - Q_1. \tag{11}$$

Any observation that lies beyond the following bounds is classified as an outlier:

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR}, \quad \text{Upper Bound} = Q_3 + 1.5 \times \text{IQR}. \tag{12}$$

Values below the **Lower Bound** or above the **Upper Bound** are identified as outliers and subsequently removed from the dataset. This process ensures that the data remains free from extreme deviations, enhancing the stability and accuracy of machine learning models.

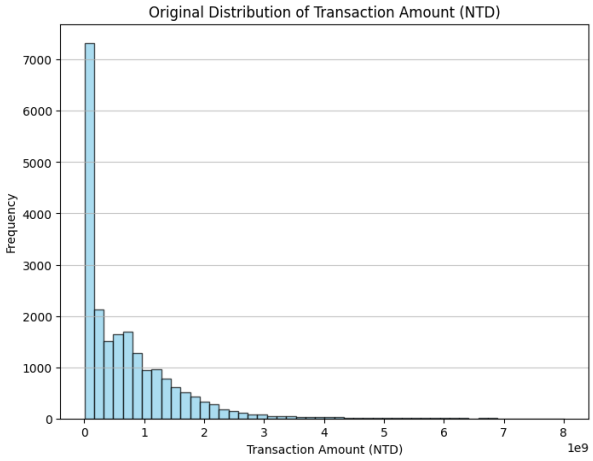


Figure 9: Original Distribution of Transaction Amount (NTD)

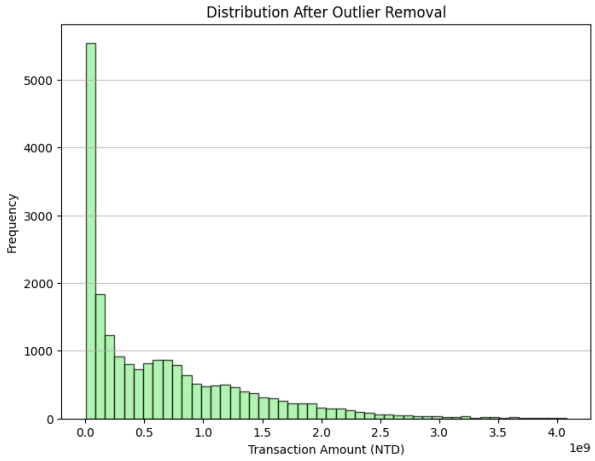


Figure 10: Distribution After Outlier Removal

From the figures, it is evident that the removal of outliers reduces the extreme values, particularly in the tail region of the distribution. However, the overall distribution remains heavily skewed. To further address this issue, we will apply a **log transformation** to normalize the target variable and improve model performance.

3.3.3 Log Transform

To further address the skewness in the target variable, *Transaction Amount (NTD)*, we apply a **logarithmic transformation (log transform)**. The log transform replaces the original variable x with its natural logarithm, defined as:

$$\log(x) = \ln(x + 1), \tag{13}$$

where x is the original value. The addition of 1 prevents issues arising from zero values.

Log transformation compresses large values and stretches smaller ones, effectively reducing skewness and bringing the data closer to a normal distribution. Figure **??** compares the distribution after outlier removal (Figure 11) with the log-transformed distribution, demonstrating its effectiveness.

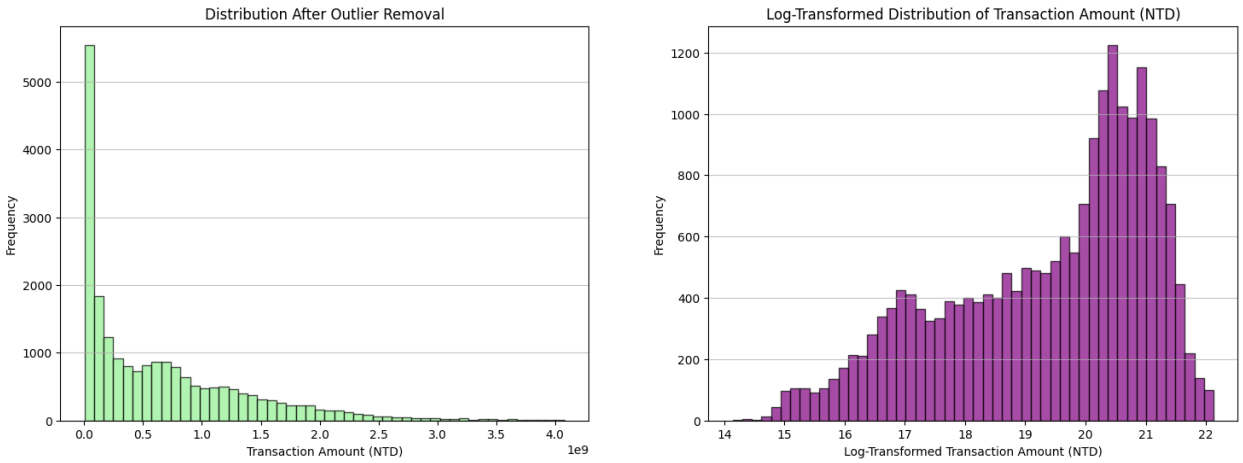


Figure 11: Comparison of Distribution After Outlier Removal (left) and Log-Transformed Distribution (right) of Transaction Amount (NTD). The log transformation effectively reduces skewness and compresses extreme values, improving the data distribution.

From the comparison, although outlier removal mitigates extreme values, the distribution remains highly skewed. By applying the log transformation, the skewness is significantly reduced, and the data achieves a more symmetric distribution. This transformation enhances the stability of the machine learning model and improves its ability to generalize patterns effectively.

4 Results

Based on insights from the Exploratory Data Analysis (EDA), distinct patterns were observed across industries for key variables such as *Age Group, Income Group*, and *Education Level*. Training a single unified model may fail to capture these industry-specific variations. Therefore, the dataset was split by industry to train separate models, enabling better adaptability to these differences.

For demonstration, we focus on the *Industry_Clothing* segment, using **Age Group** as the predictor and *Transaction Amount (NTD)* as the target variable. The dataset was split into training and testing sets using an 80/20 ratio. Age Group, a categorical variable, was one-hot encoded to ensure compatibility with the regression model. Model performance was

evaluated using the **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \,, \tag{14}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. The MAE was selected for its robustness to extreme values, providing a reliable measure of predictive accuracy.

This industry-specific approach effectively captures unique patterns within *Industry_Clothing*, while the 80/20 data split ensures rigorous evaluation on unseen data.

4.1 Linear Regression

For the *Industry_Clothing* segment, a linear regression model was trained using *Age Group* as the primary predictor. The model performance was evaluated using the **Mean Absolute Error (MAE)**, which achieved a value of 0.727 on the log-transformed scale.

The scatter plot in Figure 12 visualizes the predicted transaction amounts against the true values on the original scale. While the model aligns well with smaller transaction amounts, deviations are evident as transaction amounts increase, reflecting an under-prediction for higher values. This behavior suggests that the model may not fully capture the variability in large transactions, likely influenced by data skewness. Addressing this limitation with more complex models or additional features could improve predictive performance.

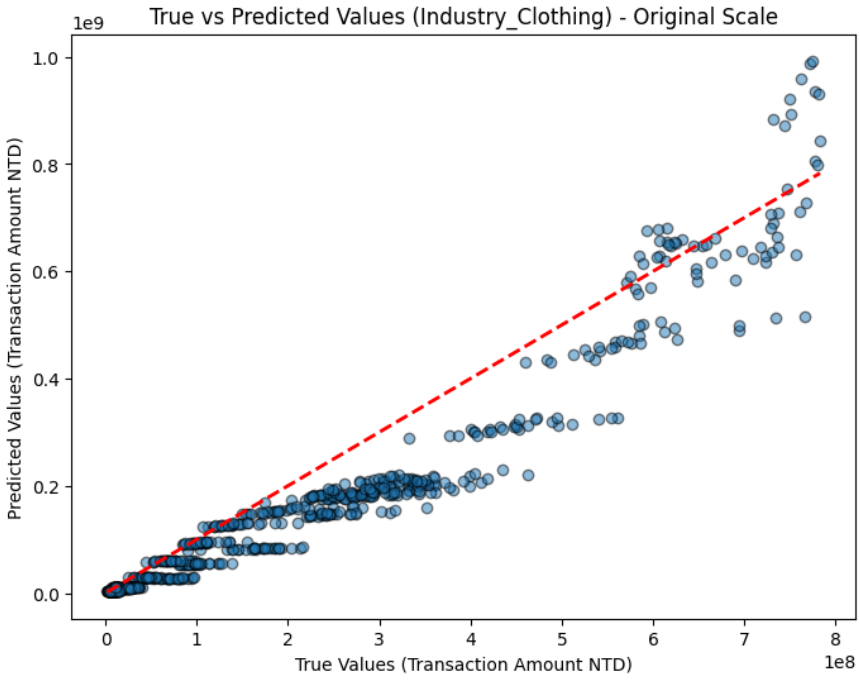


Figure 12: Scatter Plot of True vs. Predicted Transaction Amounts for Industry_Clothing.

4.2 K-Nearest Neighbors (KNN)

To predict the *Transaction Amount (NTD)* for the *Industry_Clothing* segment, a K-Nearest Neighbors (KNN) model was employed. KNN is particularly suitable in this context as it assumes that consumers with similar features, such as *Age Group*, exhibit comparable purchasing behaviors. By leveraging this assumption, KNN effectively estimates transaction amounts based on the behavior of neighboring data points.

Figure 13 illustrates the relationship between the predicted and true transaction amounts on the original scale. Compared to the results obtained using linear regression (Figure 12), the KNN model achieves a better fit to the dataset, as predictions align more closely with the diagonal line. This suggests that KNN better captures the inherent patterns in the data, particularly for this segment.

Figure 14 shows the error trends across varying K-values, where the minimum error occurs at $K = 2$. This demonstrates that a smaller K-value effectively captures local relationships among consumers, whereas larger values lead to oversmoothing and reduced performance.

The final model achieves a **Mean Absolute Error** of 0.167.

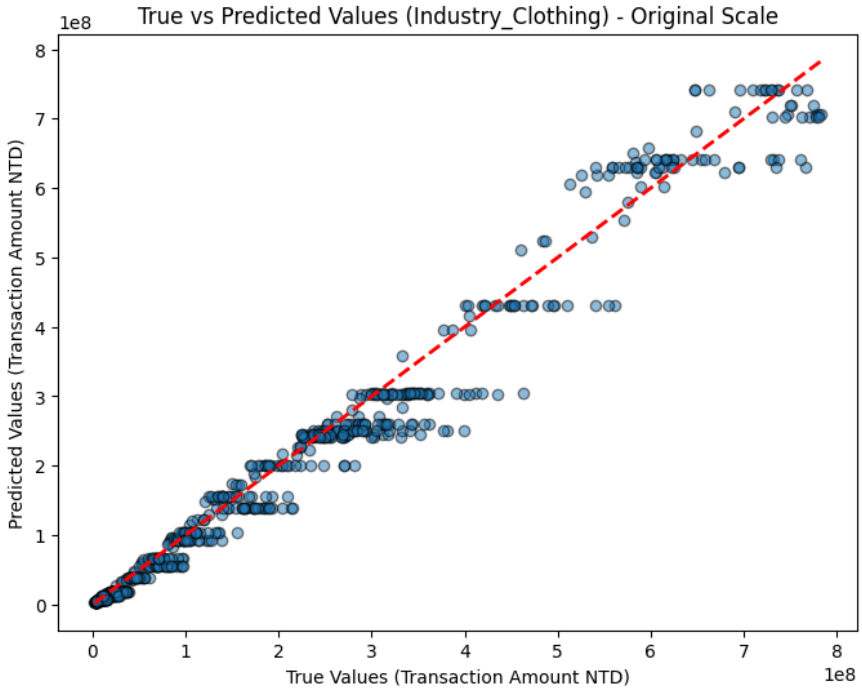


Figure 13: Scatter Plot of True vs. Predicted Transaction Amounts for Industry_Clothing (KNN).

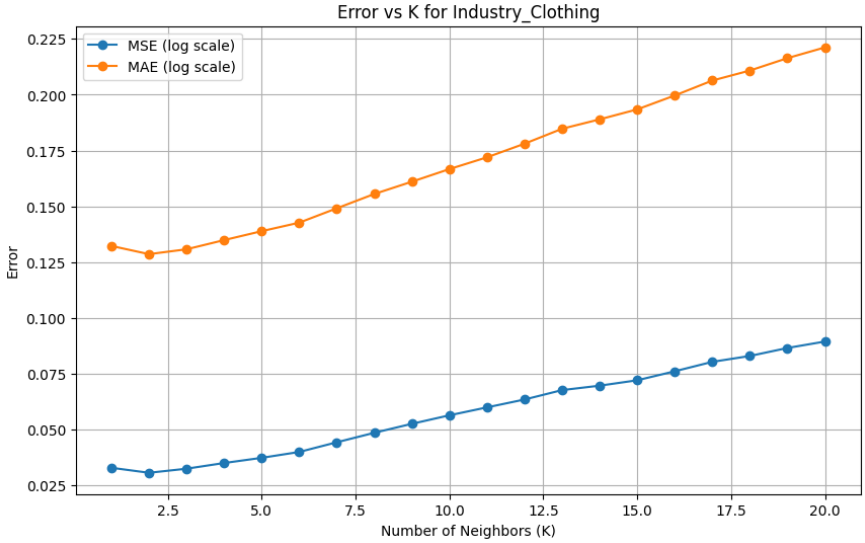


Figure 14: Error vs K for KNN Model.

4.3 LSTM

To predict the *Transaction Amount (NTD)* for the *Industry_Clothing* segment, a Long Short-Term Memory (LSTM) model was employed. LSTM, a type of recurrent neural network, is particularly effective for sequential data and time series forecasting. In this case, it leverages temporal patterns within the data to improve predictions.

Figure 15 illustrates the evaluation results. Compared to Linear Regression (Figure 12) and KNN (Figure 13), LSTM demonstrates superior performance for smaller transaction amounts, where the predicted values closely align with the true values along the diagonal line. However, as the transaction amounts increase, the model begins to exhibit larger deviations. This can be attributed to the dataset's imbalance, where larger transaction amounts are underrepresented, leading to diminished learning capability for these values.

The final model achieves a **Mean Absolute Error** of 0.301, indicating improved accuracy in capturing smaller transaction patterns but highlighting the need for additional data or adjustments to improve predictions for larger values.

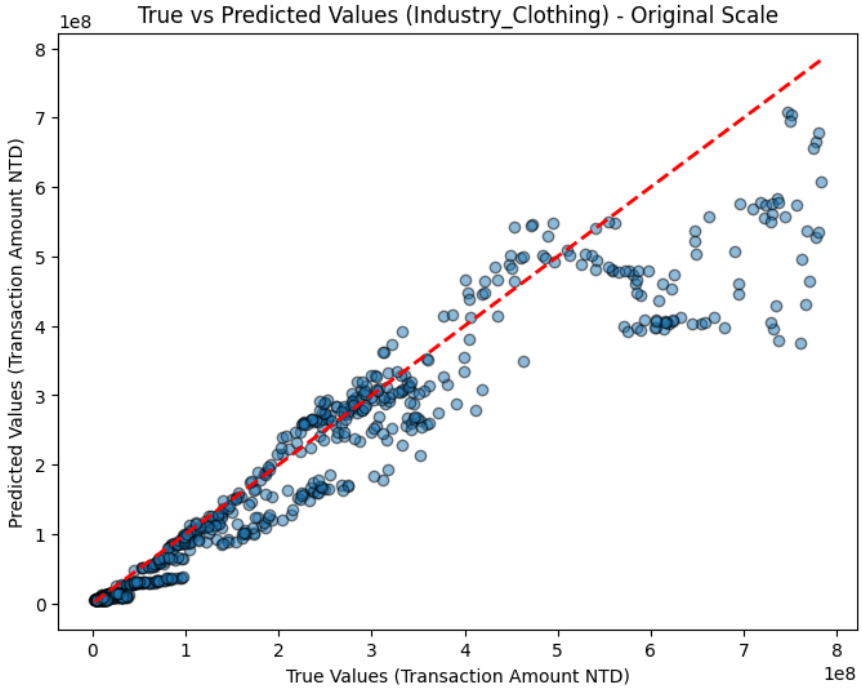


Figure 15: Scatter Plot of True vs. Predicted Transaction Amounts for Industry_Clothing (LSTM).

4.4 KNN + LSTM Combined Model

To address the limitations of standalone models, a hybrid *K-Nearest Neighbors (KNN)* and *Long Short-Term Memory (LSTM)* approach was implemented. This method integrates the strengths of both models: KNN effectively handles sparse transaction amounts, while LSTM excels in capturing sequential patterns and temporal dependencies.

A **Decision Tree Classifier** was employed to dynamically route testing data to the appropriate model. Specifically:

- Transaction amounts in the training set were analyzed based on frequency.
- A threshold of 5 occurrences was set to distinguish between *high-frequency* and *low-frequency* data points.
- Data points above the threshold (high-frequency) were routed to the LSTM model for prediction.
- Data points below the threshold (low-frequency) were sent to the KNN model with $k = 2$ neighbors.

Figure 16 illustrates the data flow during the testing phase. Testing data is first passed through the Decision Tree Classifier, which determines the frequency class of each data point. The respective models—LSTM for high-frequency and KNN for low-frequency—are then used to generate predictions, and their errors are aggregated to compute the final **Mean Absolute Error (MAE)**.

The output scatter plot in Figure 17 shows the predicted transaction amounts against the true values on the original scale. The predictions are tightly distributed around the red dashed diagonal line, which represents the ideal fit. This indicates that the hybrid approach

significantly improves predictive performance across both high-frequency and low-frequency transaction amounts.

Compared to standalone models, the combined KNN + LSTM model achieved a **Mean Absolute Error (MAE)** of 0.0863, demonstrating its superiority. The integration of the Decision Tree Classifier ensures that the most appropriate model is applied for each data point, effectively balancing the strengths of KNN in sparse regions and LSTM in dense regions.

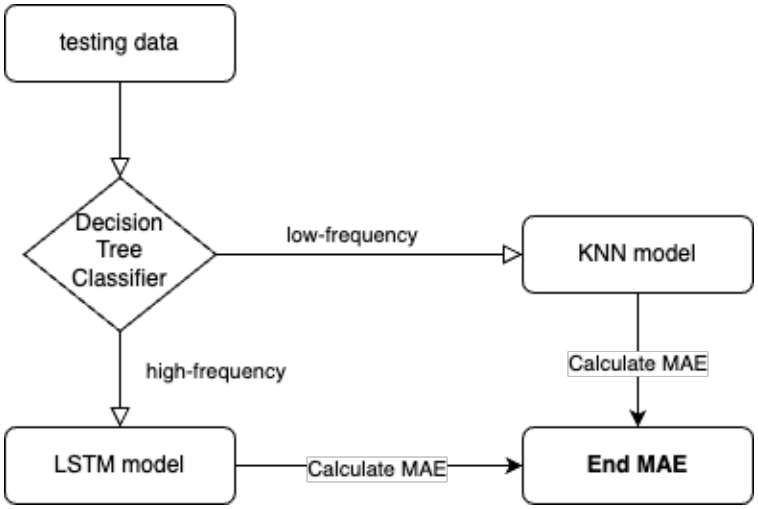


Figure 16: Data Flow for Combined Model (LSTM + KNN) with Decision Tree Classifier.

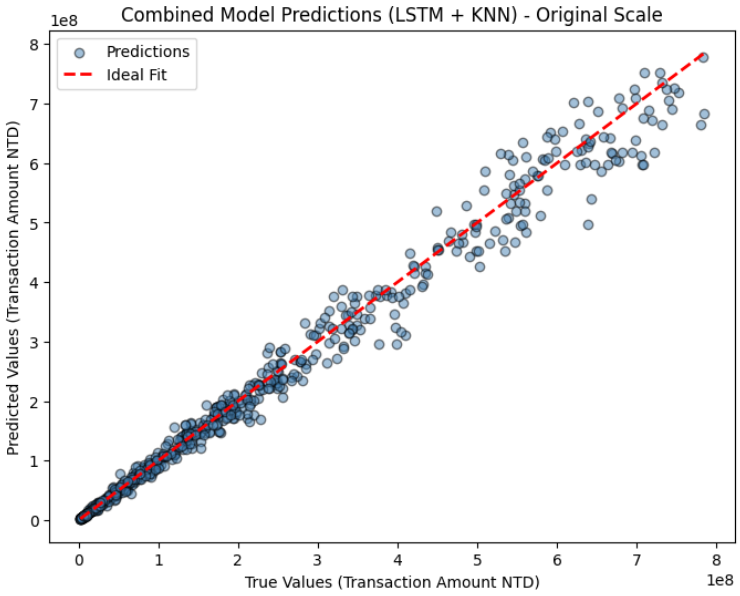


Figure 17: Combined Model Predictions (LSTM + KNN) on Original Scale.

Overall, the combined model successfully captures both consumer behavior patterns and temporal dependencies. The improved accuracy highlights the efficacy of leveraging frequency-based classification to direct predictions to the appropriate model, resulting in robust and reliable transaction predictions.

5 Conclusion

This study presents an integrated approach combining machine learning models and time series analysis to predict consumer spending behaviors. By analyzing key features such as income, age group, education level, industry category, and transactional date components, we demonstrated the effectiveness of leveraging demographic and temporal data in forecasting transaction amounts.

The experimental results reveal varying strengths and limitations across different models. Table 3 summarizes the performance metrics for each model, highlighting their respective capabilities.

Table 3: Performance Comparison of Models

Model	Mean Absolute Error (MAE)	Strengths	Limitations
Linear Regression	0.727	High interpretability	Limited for non-linear relationships
K-Nearest Neighbors (KNN)	0.167	Captures local patterns	Sensitive to noise in sparse data
Long Short-Term Memory (LSTM)	0.301	Handles sequential dependencies	Struggles with imbalanced data
KNN + LSTM (Hybrid)	**0.0863**	Combines strengths of both models	Higher computational complexity

The results indicate that while traditional models like Linear Regression offer strong interpretability, they struggle to capture non-linear or sequential patterns. KNN excels in modeling local relationships, making it effective for sparse transactional data, whereas LSTM demonstrates superior performance in time-series forecasting by capturing long-term dependencies. The hybrid model, combining KNN and LSTM via a Decision Tree Classifier, achieves the best performance, demonstrating the value of integrating multiple modeling techniques.

The following contributions and insights were derived from this study:

- The inclusion of demographic and temporal features significantly improves prediction accuracy, highlighting their importance in understanding consumer spending.
- Preprocessing steps, such as outlier removal, log transformation, and one-hot encoding, enhance data quality and model robustness.
- The hybrid approach showcases the benefit of combining complementary models to address the limitations of individual techniques.

Future research directions could focus on:

1. Expanding the dataset to include more granular transactional data and external economic indicators.
2. Exploring other advanced architectures, such as Transformers, to capture more intricate temporal patterns.
3. Developing interpretability frameworks to analyze feature importance and enhance trust in model predictions.

This study contributes to the growing body of research on data-driven consumer behavior analysis and provides actionable insights for businesses and financial institutions to optimize their strategies and better understand spending dynamics.

References

Agrawal, D. (2013). State-of-the-art techniques for returns prediction. *Journal of Financial Analysis*, 10:123–134.

Author, C. (2017). Applications of time series forecasting in consumer behavior. *Economic Forecasting Journal*, 15:56–68.

Author, D. (2020). Enhancing prediction accuracy using machine learning and time series data. *Journal of Predictive Analytics*, 8:45–60.