A Comprehensive Analysis Of Bank Transaction Data: Exploring The Value Behind The Data

Data Science Mini Project
Problem C - Group 25
https://github.com/UoB-DSMP-2023-24/dsmp-2024-group-25

Bohan Liu
Data Science MSc
University Of Bristol
mj20403@bristol.ac.uk

Haolong Li
Data Science MSc
University Of Bristol
bm23475@bristol.ac.uk

Shuyao Yi
Data Science MSc
University Of Bristol
kr23140@bristol.ac.uk

***Abstract*—This research deeply explored Lloyds Bank's simulated transaction data, provided a deep understanding of the transaction data through visualisation, used machine learning and deep learning models to analyse customer transactional data's potential value. We identify key consumption trends and high-value customer segments through RFM and CLV analysis. Implementing LSTM and Decision Tree improves the accuracy of predicting customer spending patterns and helps provide targeted services. In addition, the detection of fraudulent transfers was enhanced using the Isolated Forest algorithm, improving account security. These insights provide strategic recommendations for banks' customer service and improve bank account security. Based on these findings, we also suggest possible directions for future research to optimise banking services further and enhance customer satisfaction and loyalty.**

I. Introduction

In the rapidly developing financial industry, there are both opportunities and challenges. Banks have to conduct more business, provide better services to their customers, and be alert to the dangers in the financial industry. This report demonstrates the analysis of a simulated bank transaction dataset and provides insights and interpretations of the transaction data through visualisation. For instance, we discovered that many of our customers tend to make large transactions on weekends, indicating potential investment opportunities. Based on the analysis results, several models from a range of machine learning and deep learning, such as Long Short Term Memory (LSTM), are used to explore the potential role of the dataset. In addition, to better fit the context of the analysis, several business models, such as Recency, Frequency, Monetary (RFM) analysis and Customer Lifetime Value (CLV) models, were applied to segment customers into more specialised groups based on their transactional behaviour to target business. These analyses reveal the potential value of customers over time and provide constructive insights into the bank's customer service operations.

With the above techniques, we focus on analysing the following three directions:

- **Customer Value:** We use RFM analysis to analyse customer transactional behaviour and combine it with CLV to predict lifetime value.
- **Customer Spending Habits:** We use a Decision Tree algorithm to analyse customer spending habits in the first dataset and process the second dataset using a LSTM network to more accurately predict customer spending patterns for targeted services.
- **Customer Account Security:** We use isolation forest algorithms, a type of unsupervised learning algorithm for anomaly detection that works on the principle of isolating anomalies to detect unusual transaction patterns, as well as unusual income and expense relationships, and factors that may result in revenues that do not support current spending habits.

II. Literature Review

The increasingly widespread application of big data technology in the banking industry provides banks new opportunities to improve their service quality and operational efficiency. The following content mainly discusses the help of data analysis in the banking industry and its specific applications.

A. Roles

H.Nobanee et al. [1] review the application of big data in the banking industry from 2012 to 2020 from the perspective of bibliometric analysis, emphasising its impact on risk management, customer relationship management and investment Impact on banking and other fields. The study shows that through data analysis, banks can better assess credit risks,

improve investment profits, and enhance customer interaction, thereby improving competitiveness.

S.Sun et al. [2] point out that extensive data analysis can help banks gain a deeper understanding of customer needs, optimise products and services, and improve the scientific management level of marketing strategies. This insight enables banks to maintain an edge in a competitive market.

B. Methods

Big data analytics can help banks by using user profiling. According to C.Huan [3], it is mentioned that the logs of commercial banks can be analysed to build customer portraits. U. Srivastava and S. Gopalkrishna [4] state that credit card transactional data can be used to draw user profiles to help banks provide customised financial services to customers.

C. Main Effects

Big data analytics can play a crucial role in assessing customer value. S. Y. Kim et al. [5] provide a method to develop corresponding marketing strategies by understanding and measuring the actual value of customers. M. S. Chang and H. J. Kim [6] use machine learning methods, combined with demographic information and customer transaction behaviour data, to effectively segment bank clients. These methods help banks identify high-value customers and develop targeted marketing strategies.

Data analysis is also used to study customers' spending habits. U. Srivastava and S. Gopalkrishna [4] introduce how Indian banks identify customers' consumption trends and cyclical changes by analysing big data technology, allowing banks to understand customers' needs better and thus provide more consistent services—products and services in demand.

Big data analytics can also contribute to customer account security and fraud detection. M. Soltani Delgosha et al. [7] emphasise that one of the essential applications of big data in the banking industry is fraud detection and credit risk analysis. This technology can quickly identify abnormal transaction patterns and improve banks' ability to prevent and control fraud.

Based on the abovementioned literature, the application of big data in the banking industry covers critical areas such as customer value, customer consumption habits, and customer account security. These research results show that through data analysis, banks can improve their risk management capabilities, develop more effective marketing strategies, and improve customer relationships. For banks, the research and application of these directions help improve operational efficiency and enhance their competitiveness in the market, proving the practical significance and feasibility of these research directions.

III. Methodology

A. Base

1) EDA: Exploratory data analysis (EDA) is a fundamental process that visually and statistically examines a data set's main characteristics. It summarises key characteristics of data through charts, plots, and statistics to identify patterns, anomalies, and relationships to gain insight into the underlying structure and quality of the data. This critical step helps us understand the structure and quality of the data to guide further analysis.

B. Customer Value Analysis

1) RFM Analysis: Segment your customers by examining their recent transaction frequency, transaction frequency, and amount spent to identify your most valuable customers.

2) Customer Lifetime Value (CLV): CLV analysis uses RFM data to use the results from the BG/NBD and Gamma-Gamma models. The two models are used to predict the probability of a customer's future purchasing behaviour and predict the average profit of a customer who has already made a repeat purchase.

The BG/NBD model is based on two assumptions: that the time intervals between customers' purchasing behaviours follow a negative binomial distribution and that the probability of churning a customer after no purchasing activity increases with time according to an exponential distribution. Specifically, the model predicts the number of transactions each customer will make at a given time in the future.

$$P(\text{alive}) = \frac{1}{1 + \frac{frequency \cdot (\beta + T)}{\alpha + \beta + T}} \tag{1}$$

The Gamma-Gamma model assumes that the value of an individual customer's transactions is independent of the frequency of their purchases and that the value of a customer's transactions follows a gamma distribution. Based on the outputs of the two models, the number of trades and the average profit per trade for each customer over a certain period in the future (e.g., 12 months) are predicted as the customer's lifecycle value.

$$E(M|F, M) = \frac{(p + F \cdot x)}{(q + F)} \tag{2}$$

C. Customer Spending Habits

1) Tableau Visualisation: Tableau provides powerful visualization capabilities that transform complex transactional data into intuitive images. This tool can initially analyze the trends and patterns of customer consumption habits and help determine the direction of further analysis.

2) K-means Clustering: K-means Clustering is a widely used unsupervised learning algorithm for dividing a data set into groups or "clusters" such that data points within the same cluster are very similar. In contrast, data points within different clusters are relatively dissimilar. Customers can, therefore, be categorised according to the amount they spend. Different groups of consumers are identified, e.g. distinguishing between high and low-spending customers. This analysis results can be used to position marketing strategies, optimise resource allocation and improve customer service, among other things.

3) Decision Tree: Decision Tree is a supervised learning algorithm for classification and regression tasks. It simulates the decision-making process through a series of questions, thus partitioning the dataset into smaller and smaller parts, ultimately leading to effective classification or prediction. So, in this project, we focus on classifying the customers through the type of purchase through a decision tree.

4) Long Short Term Memory: Long Short Term Memory (LSTM): LSTMs are neural networks applied to sequence prediction problems by analysing data sequences to predict future behaviour. They can help us understand patterns in transaction sequences and thus identify customer spending habits to speculate on the customer's future spending tendencies.

D. Customer Account Security

1) Isolation Forest: This method isolates anomalies, making it effective in identifying fraudulent transactions in banking data.

2) Neural Networks: Enhance real-time fraud detection capabilities using advanced pattern recognition to detect deviations in established spending behaviour that may indicate fraud.

By applying these methods, banks can gain a deeper understanding of consumer spending habits, more accurately assess customer value, and improve fraud detection, thereby increasing customer satisfaction and account security.

IV. Data Sources and Data Processing

A. Data Sources

The data of this project is divided into two stages. The first stage data set is named *"fake transactional data 24.csv"* and contains four data features as follow.

TABLE I: First Dataset

Features	Description
from totally fake account	virtually generated account that initiates the transaction
monopoly money amount	the virtually generated transaction amount
to randomly generated account	the receiving account in the virtually generated transaction
not happened yet date	the virtually generated transaction date

The name of the data set in the second stage is *"simulated transaction 2024.csv"*, which contains 7 data features. The details are shown below.

The first-stage data set focuses more on simulated transactions between accounts. In contrast, the second-stage data set provides more detailed transaction records, including timestamps, balances, and third-party transaction participants.

B. Data Preprocessing

1) Base: When performing data preprocessing, because there are no clear rules for reference in the account transaction data, it is impossible to choose the average value to supplement the missing values in the data set, and the missing values account for a small proportion of the data set, so we decided to delete the missing values. In order to meet different analysis needs, we divided the processed data set into several sub-data sets.

TABLE II: Second Dataset

Features	Description
Date	the date when the transaction occurred
Timestamp	the specific time when the transaction occurred
Account No	the account number that initiated the transaction
Balance	the account balance after the transaction
Amount	the amount of a single transaction
Third Party Account No	the recipient of this transaction
Third Party Name	the name of the recipient merchant

For the first data set, we distinguish between two types of data: account transfer records and transaction records with merchants firstly. Then we divide the merchants into six categories when conducting customer segmentation research.

There is a slight difference for the second data set because of the difference in the features it contains. First, we distinguish between two types of data: account transfer records and transaction records with merchants similarly. Next, we extract the positive part (representing salary income) from the merchant transaction records and form a separate sub-dataset. The remaining transaction records, the consumption expenditure portion, are stored as another independent sub-dataset. Merchants were divided into ten categories based on merchant type when conducting customer segmentation research.

2) Data Preprocessing For Models: For some algorithm implementations, the dataset requires further preprocessing.

RFM Analysis:

- First dataset:
  At the beginning of the project, an initial attempt was made to use the first dataset for RFM analysis. However, it quickly became apparent that the features of the dataset could not demonstrate results worth analysing. Therefore, it was decided not to use the first dataset for this analysis.
- Second dataset:
  To be more relevant for our research purposes, all merchant names in this dataset were replaced with the names of the categories to which they belonged, a step taken to simplify the model's categorisation and make the results more general and easy to interpret. All irrelevant or unnecessary features were then removed to focus on three dimensions that are crucial for RFM analysis: time of last purchase (Recency), frequency of purchase (Frequency) and amount spent (Monetary).

CLV: The content of the dataset required for this analysis is the same as for the RFM analysis, so the data processing is

the same as for the RFM analysis.

LSTM:

- First dataset:
  The features of the first dataset could not demonstrate results worth analysing. Alternatively, it is hard to make an effective LSTM model.
- Second dataset:
  In order to effectively utilise LSTM, the consumer sub-dataset first needs to be segmented into a training set and a validation set so that the model can learn in a controlled environment and verify the accuracy of its predictions. Then, we used the PyTorch framework to preprocess the data carefully. First, we assign labels to merchant types (ranging from 0 to 9) and then divide the time into two-hour intervals (sliding window) to capture changes in consumer behaviour at different times of the day. In addition, to ensure robustness, we removed some variables that may lead to overfitting, such as date, timestamp, merchant type name, and account ID. Finally, to ensure that the data can be processed efficiently in the LSTM network, we converted all data lists into Tensors with a consistent structure for easy training. Through such processing, LSTM can more accurately capture and learn the time series characteristics of consumer behaviour.

Fraud Detection for Customer Account Security:

- First dataset:
  Generate account balances from the original data set that has been processed for missing values. Set thresholds for detecting abnormal behavior: High-frequency transfers: The account transaction frequency is 2 times the average Large transfer: The amount of a single transaction exceeds 80% of the account balance Low-frequency large-amount transfers: transaction frequency is less than the median and transaction volume exceeds 80%
- Second dataset
  Through the experience of analyzing the first data set, the features and thresholds used for detection were optimized in this analysis. We use the partitioned account transfer record dataset and split it into either an income or expense dataset based on transfer type and analyze the two new datasets separately. Set thresholds for detecting abnormal behavior: Transfers at abnormal times: transfers before 7 a.m. and after 22 p.m. Large transfer: For income transfers, if it exceeds 1.5 times the average income, it is abnormal. For expenditure transfers, exceeding 50% of the account balance is an exception. High-frequency transfers: If the number of account transactions exceeds twice the average number, it is considered abnormal.

V. Results

A. Customer Value Analysis

1) EDA: Because the information contained in the features in the first dataset was not sufficiently informative about customer value, no valid analyses were produced in the first dataset.

Through classification and visual analysis of bank customers with different deposit levels in the second dataset, we found that the spending of low-deposit customers is mainly concentrated in Halifax and LBG, which may be mainly loan repayments or essential financial services, indicating that these customers have low spending power. There may be higher financial stress. High-deposit and high-asset customers' primary income comes from financial services, advanced medical care, and education industries, which show high economic activity and spending power.

2) RFM Analysis: Based on the data preprocessing for RFM analysis described earlier, we use the quartile method to grade the values of each dimension of RFM, thus classifying each RFM value into four levels, with level 1 being the best and level 4 being the worst. This rating method helps to distinguish different types of customer groups clearly. Each customer's RFM category is then based on the combination of their scores on the three dimensions, e.g. a customer's RFM rating might be "1-1-1", indicating that the customer is in the best category on all three dimensions.

In order to better illustrate the results of the grading of the customers in each merchant type, An image has been provided below as an example an image that shows the grading of the customers in the merchant category "Large Retailers". According to the colour bar in the graph, the position of the colour lower in the colour bar, the higher the customer's RFM rating, meaning the higher the value of this customer.

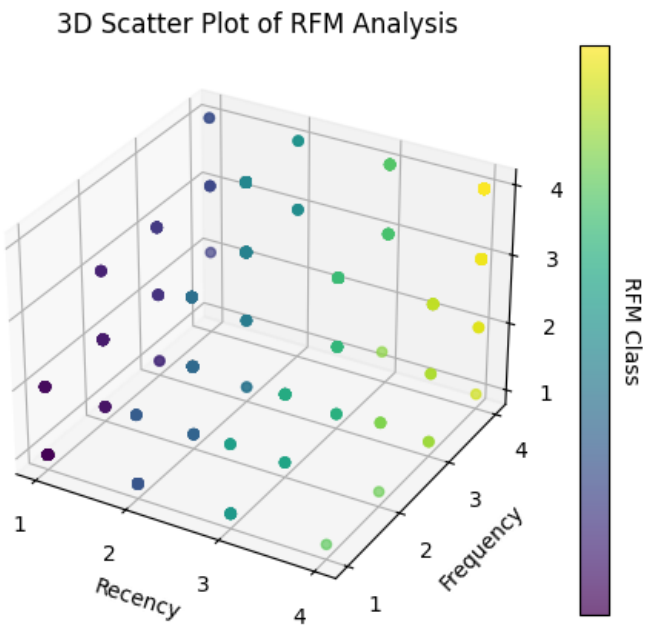

Fig. 1: RFM Analysis Example

3) CLV: Implement a BG/NBD model based on prior data preprocessing to predict future purchasing behaviour based on historical data. The Gamma-Gamma model is then used to estimate the monetary value of future transactions for each customer. Finally, the CLV of each customer is calculated by

combining the outputs of the BG/NBD and Gamma-Gamma models.

In order to show the results of CLV more intuitively, the table below shows the lifetime value of each customer for the merchant category "Large Retailers", where the higher the value, the greater the customer's future value.

In addition, there are some customers in the graph that do not have a CLV value, which means that the customer has never spent money in this area and there is no value to be analysed.

TABLE III: CLV Results

Account No	CLV
104832000	0
105375973	2405.004113743131
106601471	748.7939643284171
108481285	2135.1585605758746
108563213	1834.9326859990094
108812033	131.4811602647312
⋮	⋮
999752672	333.99685259479963

B. Customer Spending Habits

1) EDA: We used Tableau visualisation tools to initially analyse bank transaction data, revealing trends in consumer behaviour, transfer patterns, and customer segmentation based on deposit levels, providing suggestions to improve bank operations and enhance customer experience. The results are as follows.

Time Analysis: The first data set shows that weekend consumption is significantly higher than weekdays, especially in the first three weekends of December, when consumption reaches an annual high, which may be related to the holiday shopping season.

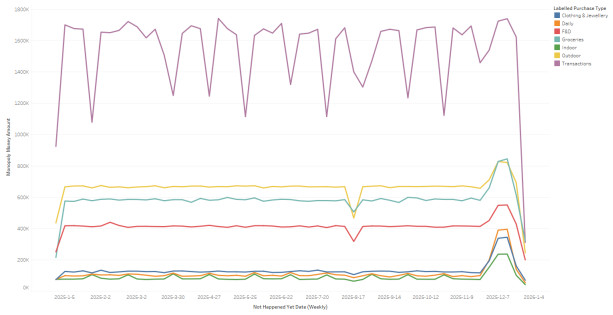

Fig. 2: First Dataset Time Analysis

The second data set shows that consumption peaks on each month's first and last days, possibly due to cyclical large transactions such as loan repayments and payroll payments. Non-transfer consumption is mainly concentrated between 8:00 am and 8:00 pm. The peak consumption on weekdays is 9:00 am and 5:00 pm, and the peak on Saturdays and Sundays is 10:00 am.

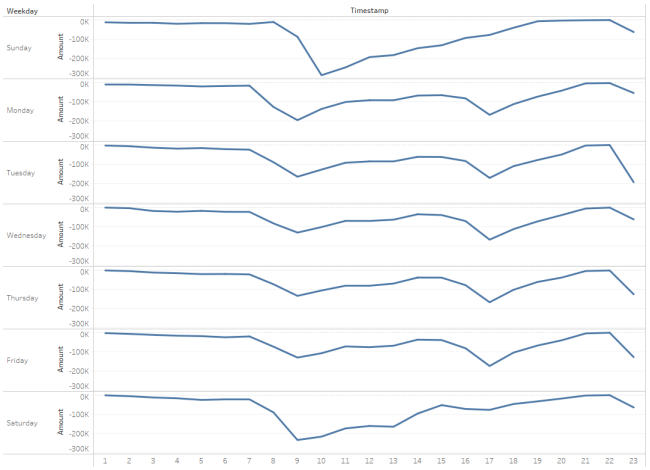

Fig. 3: Second Dataset Time Analysis

Merchant Analysis: We divided merchants in the first dataset into six different types. We found that people spend more on coffee to work efficiently on weekdays, while on weekends, they spend more time at bars and restaurants as people relax and catch up with family and friends, which shows in figure 4.

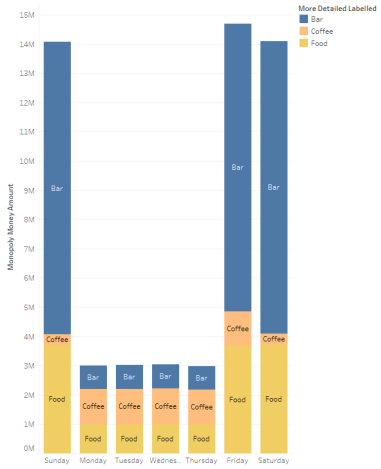

Fig. 4: Preferable Merchant Type

In the second dataset, there were too many merchants, and the names of the merchants did not show enough visualisation of their types to analyse them clearly, so there was no valid analysis of the results.

Transaction Analysis: We can also see from the first dataset that the total amount of transfers on weekends is about 1.5 times that of weekdays. However, the frequency is ten times higher, indicating that people may make large transactions on weekdays and many small transfers on weekends. This result indicates that people make large work transactions during the week and split the bill with family and friends while hanging out on the weekends. This situation may suggest that people split the bill with their family and friends while hanging out

on weekends.

In the second dataset, customers' spending habits with different incomes and assets vary considerably and are therefore analysed separately. Taking the account average balance of 5,000 pounds as a threshold, most of the low-income group's central transfers are between banks (loan repayments), daily necessities, and other merchants (income, expenses). Higher-income groups' central transfers are primarily from hospitals, care, and finance.

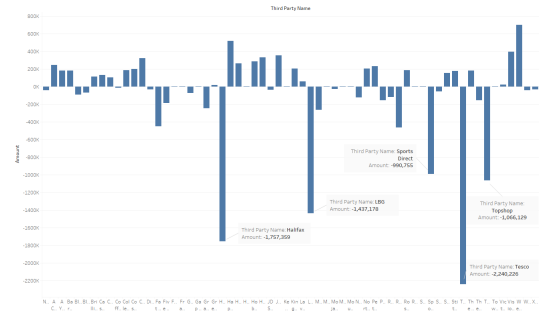

Fig. 5: Lower Income

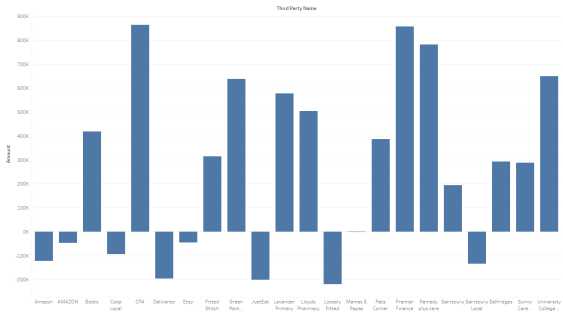

Fig. 6: Higher Income

2) Decision Tree: Due to the low number of features in the first dataset, we decided to use a decision tree for customer type classification, so this algorithm was not used in the second dataset. Based on the preprocessing results of the first dataset, we obtained the results shown in the table below by using a decision tree, demonstrating the group of customers who mainly spend in a certain category. We consider a customer who spends more than twice as much on a merchant type as the total spend on that type to be a customer who spends primarily on that type.

TABLE IV: Classification By Decision Tree

Type	Sum	Percentage
Daily Purchase	1562	19.2%
Clothing & Jewellery	1556	19.1%
Indoor Entertainment	1441	17.7%
Groceries & Department Store	593	7.3%
Food & Drink	538	6.6%
Outdoor Entertainment	180	2.2%
No Type	3691	45.3%

3) LSTM: We use LSTM networks to predict and understand the dynamics of consumer behaviour. According to the requirements of the model, we set the input and output value sizes to 2 and 10, respectively, corresponding to the input values (amount and tag) and the output values (probability of consuming in the ten types indicated by the tag). Then, follow the normal deep learning model training process and tune the parameters (learning rate, hidden size, number of layers) based on the results. We ended up with a Loss of about 1.93, which is not a relatively good result, but an acceptable one for a multi-category classification task on a dataset with up to ten types and less than well-defined classification criteria. Afterwards, the parameters were adjusted, such as shortening the sliding window size to one hour, increasing the hidden size, the learning rate, the number of training layers. The final result became 1.75, which is an improved result, but because of the risk of overfitting and the problem of insufficient arithmetic power, further parameter tuning was abandoned.

C. Customer Account Security

1) Fraud Detection: In this section, the results of the previously mentioned methodology on security are mixed and shown below.

- First dataset:
  We classify abnormal behaviours into three categories: high-value transfers, high-frequency transfers, and low-frequency high-value transfers. A total of 34,672 abnormal transfer records were detected. The detailed classification is shown in table V.

TABLE V: Classification for abnormal transaction

Type	Percentage
High-value Transfers	63.63%
High-frequency Transfers	7.87%
Low-frequency & High-value Transfers	28.5%

- Second dataset:
  We discovered a certain number of suspicious income transfers by detecting the income and expense data set using the three characteristics of value, frequency and transaction time. These transactions include high-frequency transactions, high-value transactions, and high-value and high-frequency abnormal transfers. Only two types of abnormal transfers, high frequency and high value, were found among the expenditure transfers. The details are shown as below.
  Abnormal transfer percentage:

TABLE VI: Abnormal Income Transfer Percentage

Type	Percentage
Suspicious Transfers	13%
High-frequency Transfers	9%
Large-amount Transfers	3%
High-frequency & Large-amount Transfers	1%

TABLE VII: Abnormal Expenditure Transfer Percentage

Type	Percentage
Suspicious Transfers	81%
High-frequency Transfers	16%
Large-amount Transfers	3%

Abnormal detection of revenue and expenditure relationship:
After calculating the average revenue for each account, we merge it with the spending data. Calculating the proportion of each expenditure relative to average income is used as a benchmark for analysing transfer behaviour. A total of 1,001 transfers triggered the reaction strategy; the details of the reaction strategy are shown below.

TABLE VIII: Reaction Strategy Trigger Percentage

Type	Percentage
Freeze Account	81%
Send Warning Messages	13%
Send Reminders to confirm operation	7%

VI. Discussion

A. Customer Value

In response to the results of RFM analysis and CLV, banks can offer customers with high RFM ratings or CLV (at least two dimensions of 1) on a particular spending type preferential activities on the corresponding spending type, as well as can co-operate with a particular merchant in the corresponding category and offer more preferential activities of that merchant to promote customers' spending at that merchant, thus achieving a win-win situation to some extent.

B. Customer Spending Habits

1) EDA: Based on the results of the analyses, these services can be offered in response to consumption habits as follows:

- Peak Spending Service:
  Provide customised credit card offers and reward point programmes with additional benefits for peak shopping periods on weekends or specific festivals.
- Online Banking Service:
  Develop account management tools on mobile banking apps to help customers track spending and budgeting, especially during high-spending periods.
- Customised Services For Different Income Groups:
  Design essential banking services with lower fees and easier approval for low-income customers. Provide more detailed financial management and investment services for high-income customers, including personal wealth management and customised financial advice.

2) K-means Clustering: In our dataset, K-means clustering methods performed poorly, mainly because the requirements of this method on the dataset needed to match the parameters we had. K-means clustering requires rich feature dimensions to distinguish different customer groups effectively. However, our data mainly includes basic transaction information, such as transaction amount and date, and needs more dimensions of customer behaviour characteristics. In addition, K-means is more sensitive to outliers, and extreme transaction values common in financial data may lead to unsatisfactory clustering results. Hierarchical clustering provides a detailed view of the data structure. However, its computational complexity is high and time-consuming for large-scale data processing, and it is also sensitive to noise and outliers. These factors make these two methods unsuitable for direct application to our simple dimensional data set without optimisation to accommodate the specificities of financial data.

3) Decision Tree: As can be seen from the table in the above result part, the data embodied in this dataset, the decision tree, is challenging to demonstrate a better classification, and many customers cannot be classified. Nevertheless, we can still apply the current results to offer services to each type of customer, such as coupons on weekends for those who tend to spend more on groceries or bank points (which can be exchanged for gifts) for those who swipe their bank card.

4) LSTM: The results of the LSTM analysis mean that, to some extent, we can predict the consumer behaviour of this user, but only up to two or three types. This situation results from the dataset features being less amenable to accurate multi-category classification tasks using LSTM. If the merchant types in the dataset could be so accurate that they could be classified into less than or equal to three categories, this prediction would be much more effective. Even if the prediction is inaccurate, we can still use the prediction to serve our customers at the appropriate time. For example, we predict that the user will order food or make online purchases between 10:00 am and 11:00 am. We can push some relevant coupons or bonus activities during this time.

C. Customer Account Security

Abnormal transfers may be caused by fraud, and malicious accounts may exploit vulnerabilities or steal other people's account information to conduct illegal transactions. For example, transfers at unusual times may be to avoid detection, high-frequency transfers may be to obfuscate funds or transfer transaction records, and large transfers may be irrational behaviour caused by victims being coaxed by fraudsters. These unusual transactions may harm account security and liquidity. High-value transactions may result in the loss of funds, while high-frequency transactions may imply the risk of account compromise. In order to deal with these abnormal transactions, it is recommended that account security measures, such as user identity verification, be strengthened. In addition, fraud detection should be applied more to daily life, and a real-time monitoring system should be established for abnormal transfer behaviours to detect and prevent abnormal transactions promptly.

Some things could be improved in the analysis process. Average salary or personal investment income may be misjudged as abnormal transactions or detected as abnormal trading behaviour due to some manual errors. In order to mitigate this

misjudgment, it is necessary to use more features and have a deeper understanding of the difference between everyday and abnormal transactions to better distinguish between normal and abnormal transactions. Because different accounts' income and spending patterns can vary significantly, static thresholds may not be suitable for detecting all accounts. Secondly, relying solely on the ratio of expenditure to average income to assess risk may ignore the actual situation behind the transaction. To make a comprehensive judgment, more contextual information should be combined with the analysis, such as the user's credit rating and economic status.

VII. Further Work and Improvement

A. Customer Value

Based on the analytics, we can use deep learning models to predict the risk of customer churn and improve customer retention. However, the remaining time does not support us to do this step. In addition, for these two business models, the bank can periodically assess the RFM level of the customer as well as the CLV in the long term to adjust the targeted service to a single customer to improve customer loyalty.

B. Customer Spending Habits

In order to analyse user behaviour more effectively by using the methodologies we mentioned above and improve service quality, banks can consider legally collecting the following additional information: transaction type (such as deposits, withdrawals, and payments), customer's age, gender, occupation and education level. In addition, it is helpful to include residential address information to analyse geographic spending patterns, transaction frequency and size, and the device through which transactions were completed (mobile banking, online banking, or traditional methods). Understanding a customer's credit history and the specific location of transactions can also help banks assess credit risk and prevent fraud. When collecting this data, banks must ensure compliance with relevant data protection regulations, ensure transparency in processing data and obtain customer consent.

In addition, we need better merchant-type classification algorithms for better data prediction, so merchant information-related datasets are also necessary. For example, there is a merchant dataset that stores information about the items consumed, and the type of items consumed by the customer is used to be more specific about what merchant type the merchant can be classified.

C. Customer Account Security

Future research can develop more accurate anomaly detection algorithms based on the patterns and characteristics of abnormal transactions in practice. In addition, the long-term impact of abnormal transactions on accounts and users can be studied, and corresponding response strategies can be proposed.

In addition, Static thresholds may not be appropriate for testing all accounts, as income and expenditure patterns may vary significantly from account to account. Secondly, relying on more than just the ratio of expenditure to average income to assess risk may ignore the reality behind the transaction. It should be analysed in conjunction with more contextual information, such as the user's credit rating and financial status.

VIII. Conclusion

By analysing bank-simulated transaction data, we used RFM and CLV to classify customers and predict potential values. We also used decision trees and LSTM to analyse consumption habits, providing a basis for customised services. In addition, techniques such as the isolation forest algorithm are used to detect fraudulent transfers. These explorations can improve the service level and competitiveness of banks and enhance user dependence and trust. In the future, banks should continue exploring deep learning models to predict customer churn risks and regularly adjust service strategies based on customers' RFM and CLV indicators to improve user loyalty. At the same time, banks should legally collect more comprehensive customer and merchant information to more accurately analyse consumer behaviour and provide better financial service. Regarding account security, banks should develop more sophisticated detection algorithms based on more comprehensive characteristics of actual transactions to formulate more effective strategies. Through these approaches, banks can better manage risk and provide more personalised and secure customer service, thereby staying ahead of the competition.

References

[1] H. Nobanee, M. N. Dilshad, M. Al Dhanhani, M. Al Neyadi, S. Al Qubaisi, and S. Al Shamsi, "Big Data applications in the banking sector: A bibliometric analysis approach," Sage Open, vol. 11, no. 4, Art. no. 21582440211067234, 2021.

[2] S. Sun, D. Hu, Z. Zhou, X. Hu, and Q. Shao, "Application and research of big data analysis in commercial banks," in Proceedings of the 2020 International Conference on Big Data and Social Sciences (ICBDSS), IEEE, 2020, pp. 76-79.

[3] C. Huan, "Design and application research of bank customer portrait system based on big data technology," in Proceedings of the 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), IEEE, 2023, pp. 1323-1327.

[4] U. Srivastava and S. Gopalkrishnan, "Impact of big data analytics on banking sector: Learning for Indian banks," Procedia Computer Science, vol. 50, pp. 643-652, 2015.

[5] S. Y. Kim, T. S. Jung, E. H. Suh, and H. S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," Expert Systems with Applications, vol. 31, no. 1, pp. 101-107, 2006.

[6] M. S. Chang and H. J. Kim, "A customer segmentation scheme based on big data in a bank," Journal of Digital Contents Society, vol. 19, no. 1, pp. 85-91, 2018.

[7] M. Soltani Delgosha, N. Hajiheydari, and S. M. Fahimi, "Elucidation of big data analytics in banking: a four-stage Delphi study," Journal of Enterprise Information Management, vol. 34, no. 6, pp. 1577-1596, 2021.