Data Science Practices in Modern Banking: From Transactional Data to Customer Insights

Jingzhe Gao
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
pp23467@bristol.ac.uk

Taiyu Zhao
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
hk23190@bristol.ac.uk

Bojing Hou
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
yl22179@bristol.ac.uk

Xiaotong Jin
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
my23936@bristol.ac.uk

***Abstract*—In today's banking industry, the effective use of customer transaction data is one of the keys to preventing financial fraud and improving user satisfaction. The simulated transaction data provided by Lloyd's Banking Group (LBG) provides an opportunity for this study to explore the possibilities of data science techniques to help banks optimize their decisions and services. This study uses the data set obtained from LBG, including transaction time, amount, account information, etc., through data pre-processing and visual analysis, to predict the customer's transaction amount and provide the basis for banks to judge the transaction risk. In addition, based on the classification of user transaction behavior, this research develops a personalized recommendation system, which uses unsupervised machine learning technology to further refine the customer portrait. According to the historical transaction mode of the user, two-way recommendations are made to merchants and cardholders to create greater profit space for the bank. Through practical case studies, this paper demonstrates the practicability and innovation of data science in the modern banking business, which is expected to provide new ideas and methods for customer data analysis in the industry.**

I. Introduction

In modern life, credit card consumption and online shopping have become more and more frequent, and a large amount of daily transaction data has also been generated. For the banking industry, proper management and rational use of this data is one of the keys to preventing financial risks and improving customer service quality [1]. Although the potential of data is huge, it is not easy to extract useful information accurately from the huge and chaotic transaction data and use it as the basis for business improvement.

Lloyd's Banking Group (LBG), one of the largest financial services organizations in the UK, generates a large amount of customer transaction data every day. These data contain extremely rich transaction details, if they can be properly analyzed and used, it will not only greatly improve customer satisfaction, but also gain more benefits for the bank [2]. Based on this need, this research project was born to demonstrate how machine learning techniques can be used to conduct in-depth analysis of simulated transaction data provided by LBG through a comprehensive data analysis framework. Through this project, we will conduct data preprocessing, visual analysis, transaction amount prediction, and build a personalized recommendation system based on user consumption behavior.

The first step of this research is to conduct accurate pre-processing of the transaction data. Since the data source contains some missing values, which is also a possible problem in the actual transaction or data preservation, we converted the time stamp and explored the method of data repair, so as to provide more accurate and complete data for the subsequent analysis. Next, the project uses data visualization techniques to detail time series analysis of transaction activity to help banks identify and understand patterns and trends in consumer behavior. In terms of transaction amount prediction, this project compares a variety of machine learning models, such as support vector machines, decision trees, random forests, linear regression, and multi-layer perceptrons, in order to select the optimal model to predict the future transaction amount, which supports the transaction risk prediction of banks. In addition, the classification-based recommendation system developed in this research can recommend potential users for merchants by analyzing users' consumption behaviors and recommending merchants that cardholders may be interested in, so as to improve customers' personalized experience and satisfaction.

II. Literature Review

At the beginning of the study, we reviewed the previous literature on the use of machine learning techniques in banking. Previous studies have shown that complex and variable financial data sets can be well processed and analyzed using machine learning techniques. For example, algorithms such as random forests and support vector machines have been widely used in areas such as credit scoring, fraud detection, and market trend prediction. Some studies have shown that these

models can identify trends in trading data to predict future trading behavior [3]. In addition, for real-world transaction data, the pre-processing step before analysis is very important. Various approaches have been used to ensure data integrity and accuracy, such as one study that describes the use of data interpolation and anomaly detection methods to clean up financial transaction data [4].

In terms of transaction amount prediction, deep learning models such as multi-layer perceptrons (MLP) have received special attention due to their efficiency in dealing with non-linear relationships. These models have been successfully applied to predict stock prices and trading volumes, showing better performance than traditional statistical methods [5]. Finally, regarding the recommendation system based on user consumption behavior, there have been many successful cases in the literature. These systems often combine cluster analysis and classification algorithms to identify different groups of consumers and provide personalized product recommendations based on their purchase history and behavior patterns [6] [7]. This not only enhances the customer's consumption experience but also significantly improves the sales efficiency and user loyalty of merchants.

III. Methodology

A. Time Series Analysis

In the data preprocessing phase, we convert the time stamps in the data set into a date format that is easy to use for analysis. Then in the data visualization stage, we use the method of time series analysis to initially explore the relationship between customer trading behavior and time. By plotting customer transactions in different time dimensions, we can observe the overall trading trend of all accounts and discover possible cyclical changes.

B. One-Hot Encoding

In this study, in order to properly handle categorical variables and make them acceptable to machine learning models, the One-Hot Encoding method is used. This method transforms classification characteristics by generating a binary column for each category, where 1 indicates that a particular record belongs to that category and 0 indicates that it does not. This approach ensures that the model correctly understands the independent contributions of each category without introducing possible numerical misinterpretations [8]. In the prediction of transaction amount, we did not carry out unique thermal coding for all variables, but carried out selectively. We tried many times and finally chose the combination that performed well.

C. Transaction Amount Prediction

In this section, we used multiple machine-learning models to predict transaction amounts.

Support Vector Regression (SVR) is the application of Support Vector Machine (SVM) to regression problems. The goal of SVR is to find a function that approximates the target values as closely as possible while allowing for some deviations for particular data points. The optimization problem of SVR can be expressed as:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{1}$$

Here, C is the penalty parameter, ξ_i, ξ_i^* are slack variables that measure the degree of violation for the data points that do not fall within the established margin.

Decision Tree Model constructs predictive models by recursively partitioning the data into increasingly smaller subsets. **Random Forest Model** improves prediction accuracy and control overfitting by combining the predictions of multiple decision trees. They train numerous decision trees on different subsets of data, where the prediction of each tree is averaged or otherwise synthesized into the final outcome. Fig.1 provides an intuitive illustration of how a Random Forest Regression Model works.

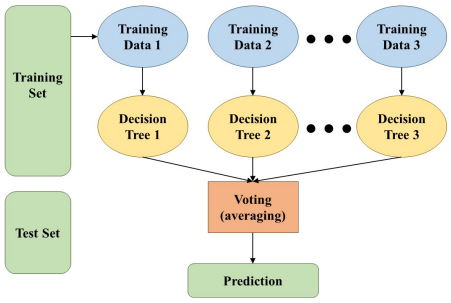

Fig. 1. Working of Random Forest Regression Mode

Linear Regression is a method for predicting a numerical response variable by fitting a linear equation that minimizes the prediction error. The model form is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \tag{2}$$

where β_i are coefficients, x_i are feature variables, and ϵ is the error term.

Multilayer Perceptron (MLP) is a type of feedforward artificial neural network that transforms and combines input features through one or more intermediary layers (hidden layers). It is capable of capturing nonlinear relationships between input variables and is suitable for complex regression tasks.

The performance of each model is evaluated using the Mean Squared Error (MSE) and the Coefficient of Determination (R2) to determine which model is the best fit for our data.

D. User Behavior Analysis

In our project, to analyze consumer behavior, we utilized the K-means clustering algorithm to group user data. K-means is an unsupervised learning method that classifies data through the following steps:

1) Initialization: Choose the number of clusters K and randomly establish K cluster centers.

2) Assignment: Assign each data point to the nearest cluster center based on the Euclidean distance between the point and the cluster center.
3) Update: Update the center of each cluster to be the mean of all points assigned to the cluster.

This process is repeated until the cluster centers stabilize or a predetermined number of iterations is reached. By analyzing features such as transaction amounts, frequency, and account balances, we employ K-means to identify different user groups. This approach helps us understand the consumption patterns and preferences of various groups [9].

IV. Data Description / Preparation

A. Data Description

The dataset used in this project comes from Dr Marie Anderson, Machine Learning Engineer, Chief Data and Analytics Office, Lloyds Banking Group, Bristol. Raw transaction data provides a record of customer interactions, including date, time, amount, pattern, and relative location. For confidentiality reasons, the raw transaction data were subjected to agent-based simulations to generate individual-level transaction data, which make up the dataset we use in this project.

In the simulated transaction dataset, there are 230,596 rows of transaction data and seven variables: transaction date (Date), transaction time (Timestamp), account number (Account No), account balance (Balance), transaction amount (Amount), third party account number (Third Party Account No), and third party name (Third Party Name).

B. Data Quality and Preparation

1) Timestamp: To facilitate analysis, we first merge the 'Date' and 'Timestamp', and convert them into 'datetime' format to support subsequent time series analysis.

2) Third Party Accounts: We have classified third-party accounts based on their business content, with specific rules as follows:

- Class 0: Personal Account or Unknown Account
- Class 1: Technology & Cultural Development
- Class 2: Fashion Trend
- Class 3: Lifestyle & Entertainment Crafts
- Class 4: Health & Living Services
- Class 5: Dining & Leisure
- Class 6: Comprehensive Retail Market
- Class 7: Active Lifestyle & Fitness
- Class 8: Financial Services & Accommodation

Among them, third party accounts with only 'Account No' but no 'Name' is classified as 'Personal Account', and third party accounts without any valid information is classified as 'Unknown Account'.

3) Missing Values and Outliers: Considering the actual significance of transaction data, it is unreasonable to delete or simply fill in missing values and outliers. Since the transaction account is unique, and there is a certain relationship between the transaction amount and the account balance (the account balance equals the last balance of this account plus the amount of this transaction), we decided to use the following method to repair the missing values and outliers. In the process, we set the minimum vacancy threshold (*min_vacancy*) because the balance in the original dataset has inconsistent decimal point retention digits.

a) Timestamp: For the case where the timestamp is empty, we use interpolation filling, using the average of the last and next transaction times of this transaction record.

b) Balance and Amount:

- We process each user's transaction records in reverse order, eliminating anomalies in individual transaction amounts by calculating the differences in balances before and after transactions.
- We then attempt to fix any missing Balance or Amount data. If the absolute difference in balance or amount discovered during the calculation exceeds the set minimum vacancy threshold (min_vacancy), we try to fill these gaps using data from unassigned accounts.
- If the filling fails, we record the failed attempt and consider further steps (such as adding new rows) in the final output.

c) Third Party Account:

- If 'Account No' is missing, it will be filled with 0.
- If 'Name' is missing, it will be filled according to its classification.

V. Results and Discussion

The analyses of the following results are all based on the second dataset provided.

A. Data visualization and analysis of user consumption habits

1) Daily consumption patterns and cyclical variations: To understand the annual changes in the consumption of the users, a graph of the annual changes in the different types of consumption and income was generated as in Fig. 2. The two curves with the largest amount of change can be clearly observed from the graph: Income and Financialservice&Accommodation. The Income curve shows a triangle-like cyclical variation and peaks at the beginning of each month, which may reflect the timing of payroll. The Financialservice&Accommodation curve, on the other hand, is in a steady state most of the time, while it reaches a trough near June, August, and November, most likely related to the end of the school holidays, the end of the summer season and the beginning of the winter season, which may be during the low season for traveling [10] and consumers are more inclined to save. Banks can use this data to adjust the timing of credit products, such as offering overdrafts before a user's salary is paid, or offering special travel loan deals during the low travel season.

In more detail, the categories Personal and Technology&CulturalDevelopment are also cyclical, and the curve of the trend is very similar to income, a consistency that suggests that users tend to spend in these categories as soon as their income arrives. Therefore, banks can introduce reward points

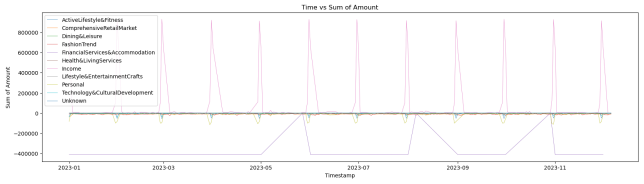

Fig. 2. Changes in Total Expenditure (Income) by Category Over the Year

or cashback programs related to these spending categories to encourage users to spend through their cards.

Fig. 3. Change in Total ComprehensiveRetailMarket Consumption Annually

Fig. 4. Change in Total Dining&Leisure Consumption Annually

Further, there are some expenditures that do not show significant cyclical changes over time, such as the Comprehensive Retail Market, which fluctuates within the range of 0-15,000 throughout the year, as shown in Fig. 3. Whereas Figure 4 shows that the Dining&Leisure category rises gradually over time until it reaches its peak at the end of the year, a change that suggests that the increase in savings over time promotes people's spending on dining and leisure. In summary, ComprehensiveRetailMarket's steady spending pattern and the growing trend of the Dining&Leisure category over time offer potential marketing opportunities for banks. For example, banks can promote their partnerships with the RetailMarket and F&B, such as co-branded credit cards or loyalty programmes, to increase user stickiness and meet their long-term needs [11]. The most unusual category is FashionTrend, which, as shown in Figure 5, has an almost irregular curve, with a high probability of guessing that it is related to the aesthetics of the moment.

Fig. 5. Change in Total FashionTrend Consumption Annually

In addition to looking at consumption patterns on a yearly macro basis, weekdays versus weekends are crucial. Figure 6 illustrates the comparison of total spending on weekdays versus weekends under different categories, with most categories not differing much during the week and on weekends. It is worth noting that more salaries are being paid on weekdays, and banks could develop flexible payroll solutions, such as offering more customised payment day options, including early payment services to suit different customers' cash flow needs [12]. Also, people tend to spend more on fashion on weekends than during the week. Based on this, banks can partner with fashion retailers to offer weekend-exclusive discounts or cashback campaigns [13] via credit card or mobile payment methods to entice consumers to spend on weekends and increase the bank's transaction volume. A step further, Figure 7 shows the comparison between the number of times spent on weekdays and weekends under different categories, and this image can also further confirm that people are more inclined to spend on fashion on weekends.

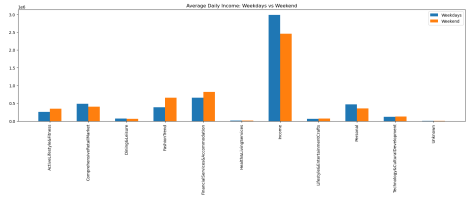

Fig. 6. Averange Daily Consume: Weekday vs Weekand

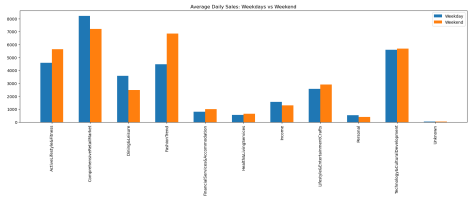

Fig. 7. Averange Daily Sales: Weekday vs Weekand

Overall, these data visualizations and analyses provide banks with important clues to understand their users' spending habits and develop marketing strategies accordingly.

2) Monthly and Annual Consumption Trend Analysis: For monthly income and expenditure totals, frequency, and individual transaction amounts, Figure 8 visualizes some trends. Vertically, the three graphs on the left represent spending, with January having the highest totals and frequency, a time when banks can help customers better manage their finances by providing budget planning tools and savings accounts. While December was the lowest of the year in terms of both total spending and frequency, but with the highest single transaction value, almost double the usual amount, suggesting that people would only be doing a small, but high-value, amount of shopping over the holidays. These year-end spending habits suggest that customers may need year-end financial planning services, and banks can offer investment advice and tax planning services at this time of year [14]. The three graphs on the right represent earnings, with December being the lowest of the year by almost a third, December has the highest single earnings, suggesting that people hardly ever choose to work over the holidays unless their company is willing to pay a higher salary. This provides banks with important clues about the movement of their customers' money.

Figure 9 visualizes the change in balances over the course of a year, where the red dotted line represents the fact that all users arriving at this time have either earned or spent money.

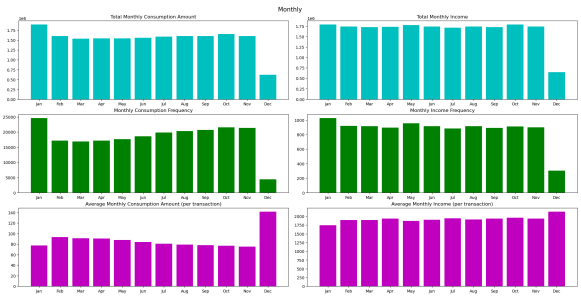

Fig. 8. Monthly Expenditure(Income) Comparison

Fig. 9. Trends in Overall Annual Balances

It is clear from the graph that the curve is changing cyclically, basically at the junction of every two months there will be a steep rise, and then fall with a certain slope, indicating that people like to spend money after payday, while the macro view of the trend of the whole graph, gradually increasing deposits can be observed, indicating that in general, each month's income is still greater than the expenditure and that the population's deposits are gradually increasing. Fluctuations in annual balances provide banks with valuable information on the flow of customer funds. Banks can develop data-based financial counseling services to help customers manage their funds more effectively [15], such as automatic transfers to high-interest savings accounts.

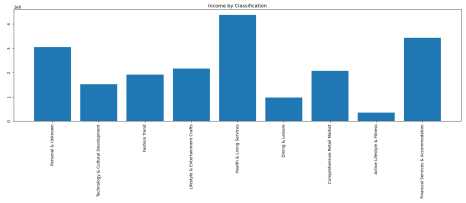

Fig. 10. Income by Classifiction

3) Consumption Category Distribution and Account-Specific Case Studies: Leaving aside the variation over time, Figure 10 shows a comparison of the total amount of different types of spending, with Health&LivingService at the top of the list, suggesting that there is likely to be a greater market demand for health insurance and healthcare financial products [16] and that banks could develop appropriate financial products to address this. Financialservice&Accommodation and personal also ranked high, indicating that demand for these categories of consumption is also high among the population.

One of the users with the highest transaction volume (858989281) was selected for comprehensive analysis in Figure 11. By analyzing the largest transaction volume user in detail, banks can identify potential high-net-worth individ-

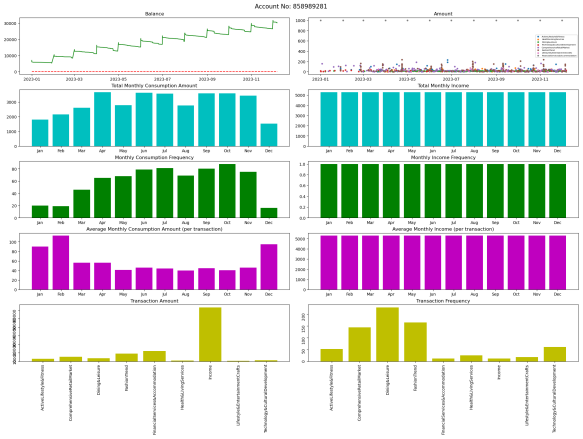

Fig. 11. Full Analysis of Account 858989281

ual customers and provide them with customized financial management services [17] such as wealth management and investment advice. Throughout the year, this user's balance gradually increased from about 7,000 to 30,000, and his frugal spending habits can be seen in the barely decreasing curve. His spending habits are focusing on multiple small purchases in March-November and some large and fewer purchases in January, February, and December. Although the transaction amount is small, the number of transactions in the categories Dining&Leisure, FashionTrend, and ComprehensiveRetailMarket is high, suggesting that these three categories have a small single transaction amount. From the above analysis, it is also possible to make a rough portrait of the user: he is a person who has a regular high monthly income and has no large monthly expenditure other than Financialservice&Accommodation, which indicates that he is relatively regular in his daily life and does not have the habit of squandering. Observing the user's spending pattern after payday, the bank can provide automatic bill payment and budget tracking features to help the customer manage his monthly expenses.

B. Comparison of the predictive performance of different models

In this study, by comparing the performance of different machine learning models in predicting the amount of bank transactions as shown in Table I, we find that various models are suitable for different banking needs. The following is a further refinement and detailed description of the performance analysis of the above models, with special emphasis on how they can help banks improve their business efficiency and customer satisfaction.

The linear regression model excels with its highest r2 score (0.909) and lowest RMSE (36.5648), which indicates that it has the highest predictive accuracy among all the models tested. The strength of the linear model lies in its simplicity and high interpretability, which makes it well-suited for banking operations where the modelling decision-making process needs to be clearly explained to regulators or

TABLE I
MODEL PERFORMANCE COMPARISON

Model	r2 score	RMSE
SVR	0.8855	41.0213
Decision Tree	0.839	48.6369
RandomForest Regressor	0.886	40.9374
Linear Regression	**0.909**	**36.5648**
Neural Network	0.8962	39.0538

customers [18]. For example, in credit approval or financial counselling services, clearly explaining to customers the basis for loan limits or investment recommendations can enhance customer trust and transparency in customer service.

Although SVR and random forests perform slightly less well than linear regression, they demonstrate the ability to capture complex variability in data. Random forests are particularly suitable for dealing with complex data sets that contain a large number of non-linear relationships, and their application is especially important in financial market analysis and asset management. By analyzing and predicting market trends or fluctuations in demand for financial products, Random Forests can help banks develop more accurate investment strategies and risk management measures [19].

With its r2 score of 0.8962 and RMSE of 39.0538, the neural network proves its strong ability to model complex non-linear relationships. For high net-worth customer management, neural networks are able to provide personalized asset management and investment advice by learning in-depth about the customer's transaction patterns, thus helping banks improve service quality and customer loyalty. In addition, this ability of neural networks makes them a powerful tool for preventing financial fraud [20], identifying abnormal transaction behaviors, and supporting the safe operation of banks.

Although the performance of decision trees is not as good as other models, their model structure is intuitive and easy to understand and implement, which makes them particularly suitable for application scenarios with high requirements for real-time trade monitoring and immediate response. However, decision tree models are prone to overfitting when faced with complex or large datasets, and thus need to be used with appropriately adjusted model parameters and pruning strategies to avoid overfitting the data.

C. Bidirectional Interactive Personalised Recommendation Analytics

Better matching of merchants and consumers is also a way to boost consumption.

Figure 12 shows the results obtained after analyzing the user's transaction behavior and account characteristics through clustering, thus providing the bank with a deeper understanding of the user group. With this graph, the distribution of different user groups in terms of income, expenditure, and account balance can be visualized, allowing for a better understanding of the characteristics and needs of different user groups in order to recommend potential consumers to merchants. Banks can carry out more precise marketing activities and

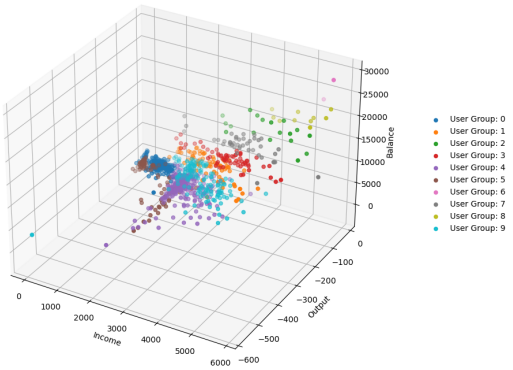

Fig. 12. Distribution of Different User Groups in Terms of Income, Expenditure and Account Balance

product positioning based on their in-depth understanding of different user groups. For example, for high-income and high-expenditure user groups with low account balances, banks can launch high-end financial products or credit cards to meet their investment and consumption needs; for low-income user groups with high expenditures, banks can launch flexible loan products or consumer installment services to help them better manage their funds.

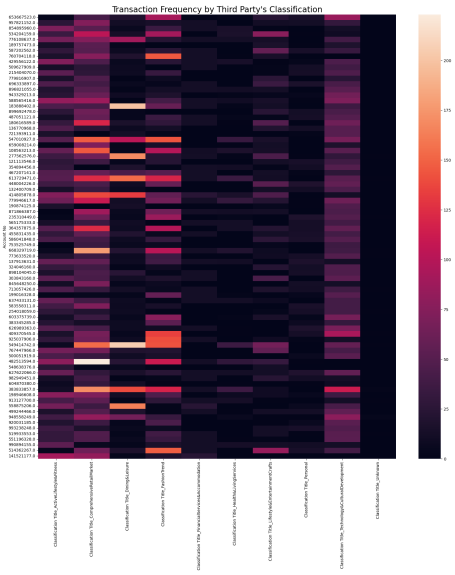

Fig. 13. Transaction Frequency by Third Party's Classification

Meanwhile, generating a heat map of transaction frequency as shown in Figure 13 can also help merchants explore potential users. In this heat map, brighter colors mean that consumers are more likely to make purchases in the corresponding merchant category, so merchants can recommend that category of service to potential users. Banks can use this graph to provide recommendation strategies to merchants to help them attract more potential customers.

The same heat map can be used to recommend merchants

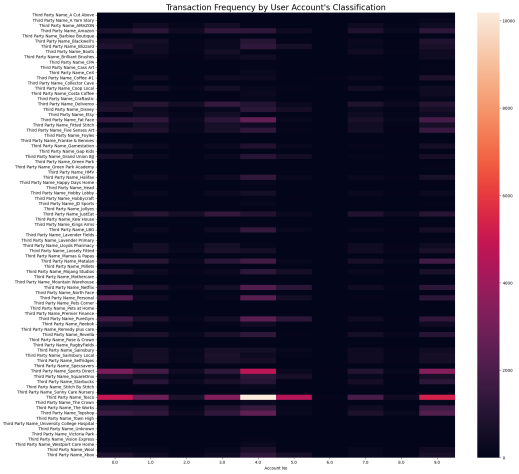

Fig. 14. Transaction Frequency by User Account's Classification

to users, except that the horizontal axis is changed to users, and the vertical axis is changed to merchants, as in Figure 14. Similarly, the brightness of the colors indicates the frequency of transactions between the user group and the merchant, with brighter colors indicating a higher frequency of transactions. Through this graph, we can clearly see the attractiveness of different merchants to different user groups and the frequency of transactions, providing users with the basis for personalized recommendations.

With the above analyses, banks can provide value-added services to their users, such as recommending special offers, new products, or merchant discounts. By recommending merchants and products to users that match their consumption preferences and lifestyle habits, banks can increase user satisfaction and loyalty, and increase the frequency and amount of user transactions, thereby increasing the bank's revenue and profitability [21].

VI. Further Work and Improvement

A. Short-term improvement plan

While current consumption forecasting models have provided effective analysis, further refinement of user profiles is needed, especially for users in different life stages and occupations. By collecting more granular consumption data, such as the user's occupational background, family structure, and lifestyle habits, the consumption behavior of various market segments can be predicted more accurately [22]. It also helps to identify atypical consumption patterns and increase anti-fraud warning mechanisms.

What's more, the current model relies on the manual entry of large amounts of data, which can be solved by integrating the bank's internal systems and automating external data sources. For example, collecting data on users' consumer behavior directly from e-commerce platforms and social media through API interfaces will improve data processing efficiency and reduce errors. This automation will also support real-time monitoring, increasing the speed and accuracy of detection of fraudulent activity [23].

Finally, based on the consumption prediction model, a real-time consumption warning system is developed, which can instantly notify users that their consumption exceeds the budget or abnormal consumption behavior. The system will help users gain better control of their finances, while also providing banks with the opportunity to intervene in a timely manner to prevent potential credit risks, including timely identification of and response to possible fraud.

B. Long-term development plans

On the basis of current consumer data analytics, it would be beneficial to develop a multi-dimensional consumer analytics platform that would integrate more types of data, including but not limited to social media trends, economic indicators, and geographic information, to enable us to understand consumer behavior from a wider perspective and capture subtle market movements, thereby providing the Bank with the ability to adjust its strategy in different economic scenarios.

Meanwhile, an intelligent financial product customization and recommendation system can be developed using deep learning and big data technologies [24]. The system will provide personalized financial product recommendations based on the user's consumption behavior, financial situation, and life events (e.g. home purchase, marriage, children's education). This approach not only improves the attractiveness and applicability of financial products but also enhances customer loyalty by providing help when users need it most.

Moreover, if a global consumption database is established, combined with advanced prediction models, such as the neural network that has been successfully applied before, the bank will be able to predict consumption trends in different countries and regions. This will provide data support for the bank's international business expansion, and help the bank make more accurate market entry and product positioning decisions on a global scale.

Finally, in order to cope with rapidly changing market conditions and consumer behavior, the system of the future will have the ability to learn and adapt itself. By constantly collecting and analyzing new consumer data, the model is able to automatically adjust its algorithms and update its predictions in real time, ensuring that banks are able to respond quickly to changes in the market. At the same time, this technology will also be used to improve the efficiency of anti-fraud systems, enabling them to instantly recognize and respond to emerging fraud patterns [25].

VII. Conclusion

In this work, based on the second dataset, we first visualized the user's consumption patterns, cycle variations, and consumption category distributions in detail, and selected a user with a high transaction volume for a case study. The analysis shows that consumption patterns are closely linked to specific months and events (e.g. holidays), providing an important basis for banks to adjust the timing of their products and services. In the meantime, consumers tend to spend on personal and techno-cultural categories immediately after receiving their

salaries, which provides direction for the design of relevant financial products. There are also, through detailed case studies of the largest volume users, we are able to gain a deeper understanding of the spending patterns and financial needs of HNW individual customers, which has important implications for banks in terms of providing personalized services and wealth management advice.

Next, we evaluated the efficacy of various machine learning models such as linear regression, SVR, random forest, decision tree, and neural networks in predicting the amount of bank transactions. The linear regression model was found to exhibit the highest prediction accuracy among all models and is suitable for application scenarios that require high interpretability. Neural networks and random forests, on an alternative note, perform better in handling large amounts of non-linear data and are suitable for complex market trend analyses. Overall, each model has its unique strengths and limitations, and banks can choose the right model by considering its predictive performance, operational complexity, and fit with their business strategy. By implementing these advanced analytics models, banks can ensure compliance while also improving decision-making efficiency, optimizing the customer experience, and enhancing risk control.

Finally, through cluster analysis and heat maps of transaction frequency, we gained insights into the distribution of income, expenditure, and account balances of the user group. This analytics approach not only improves the effectiveness of marketing campaigns but also facilitates collaboration between banks and merchants, enhancing user satisfaction and loyalty. At the same time, banks can also obtain a share of merchants' revenue through co-op promotions and transaction sharing, adding a new source of profitability to their business.

In general, we use comprehensive data analytics and advanced machine learning techniques to explore and parse the consumption patterns of bank users in a comprehensive manner. Through these in-depth insights, banks will be able to better understand customer behavior, thereby optimizing product design, making timely adjustments to service strategies, and ultimately significantly improving service quality and market responsiveness. Going forward, we plan to continue to expand these analytics to utilize a wider range of datasets and more sophisticated algorithmic models to further enhance forecasting accuracy and operational efficiency, ensuring the Bank's continued leadership in the highly competitive financial market.

REFERENCES

[1] Nesrin Ozatac, Tulen Saner, and Zeynep Suzmen Sen. Customer satisfaction in the banking sector: The case of north cyprus. *Procedia Economics and Finance*, 39:870–878, 2016. 3rd GLOBAL CONFERENCE on BUSINESS, ECONOMICS, MANAGEMENT and TOURISM.

[2] Hossein Hassani, Xu Huang, and Emmanuel Silva. Digitalisation and big data mining in banking. *Big Data and Cognitive Computing*, 2(3), 2018.

[3] Martin Leo, Suneel Sharma, and K. Maddulety. Machine learning in banking risk management: A literature review. *Risks*, 7(1), 2019.

[4] Xi Wang and Chen Wang. Time series data cleaning: A survey. *IEEE Access*, 8:1866–1881, 2020.

[5] Rodolfo C. Cavalcante, Rodrigo C. Brasileiro, Victor L.F. Souza, Jarley P. Nobrega, and Adriano L.I. Oliveira. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55:194–211, 2016.

[6] Kai Wang, Tiantian Zhang, Tianqiao Xue, Yu Lu, and Sang-Gyun Na. E-commerce personalized recommendation analysis by deeply-learned clustering. *Journal of Visual Communication and Image Representation*, 71:102735, 2020.

[7] Soma Bandyopadhyay, S. S. Thakur, and J. K. Mandal. Product recommendation for e-commerce business by applying principal component analysis (pca) and k-means clustering: benefit for the society. *Innovations in Systems and Software Engineering*, 17(1):45–52, 03 2021.

[8] A. Zheng and A. Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, 2018.

[9] Kristina P. Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727, 2020.

[10] Juan Antonio Duro and Judith Turrión-Prats. Tourism seasonality worldwide. *Tourism Management Perspectives*, 31:38–53, 2019.

[11] Matthew Tingchi Liu, Rongwei Chu, IpKin Anthony Wong, Miguel Angel Zúñiga, Yan Meng, and Chuan Pang. Exploring the relationship among affective loyalty, perceived benefits, attitude, and intention to use co-branded products. *Asia Pacific Journal of Marketing and Logistics*, 24(4):561–582, 2012.

[12] David Brown. How flexible pay can help employees in 2023. *Strategic HR Review*, 22(2):47–51, 2023.

[13] Patrícia Saltorato, Larissa Cecília Domingues, Júlio César Donadone, and Márcia Regina Neves Guimarães. From stores to banks: The financialization of the retail trade in brazil. *Latin American Perspectives*, 41(5):110–128, 2014.

[14] Sherman D Hanna. The demand for financial planning services. *Journal of Personal Finance*, 10(1):36–62, 2011.

[15] M N A Shahani, Usman Maharoof, V M Senananda, R T S Rajapaksha, Dasuni Nawinna, and Buddhima Attanayaka. Investopal - smart financial investment advisory system. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–6, 2023.

[16] Emily Gustafsson-Wright, Abay Asfaw, and Jacques van der Gaag. Willingness to pay for health insurance: An analysis of the potential market for new low-cost health insurance products in namibia. *Social Science Medicine*, 69(9):1351–1359, 2009.

[17] Joseph Bamidele Awotunde, Emmanuel Abidemi Adeniyi, Roseline Oluwaseun Ogundokun, and Femi Emmanuel Ayo. Application of big data with fintech in financial services. In *Fintech with artificial intelligence, big data, and blockchain*, pages 107–132. Springer, 2021.

[18] Akber Rajwani, Tahir Syed, Behraj Khan, and Sadaf Behlim. Regression analysis for atm cash flow prediction. In *2017 International Conference on Frontiers of Information Technology (FIT)*, pages 212–217, 2017.

[19] Martin Leo, Suneel Sharma, and Koilakuntla Maddulety. Machine learning in banking risk management: A literature review. *Risks*, 7(1):29, 2019.

[20] John Zhong Lei and Ali A. Ghorbani. Improved competitive learning neural networks for network intrusion and fraud detection. *Neurocomputing*, 75(1):135–145, 2012. Brazilian Symposium on Neural Networks (SBRN 2010) International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010).

[21] Sankar Krishnan. *The power of mobile banking: how to profit from the revolution in retail financial services*. John Wiley & Sons, 2014.

[22] Dimitris C. Gkikas and Prokopis K. Theodoridis. *AI in Consumer Behavior*, pages 147–176. Springer International Publishing, Cham, 2022.

[23] Iqbal H Sarker. Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects. *Annals of Data Science*, 10(6):1473–1498, 2023.

[24] Elena Hernández-Nieves, Guillermo Hernández, Ana-Belén Gil-González, Sara Rodríguez-González, and Juan M. Corchado. Fog computing architecture for personalized recommendation of banking products. *Expert Systems with Applications*, 140:112900, 2020.

[25] Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi, et al. A survey of credit card fraud detection techniques: data and technique oriented perspective. *arXiv preprint arXiv:1611.06439*, 2016.

APPENDIX

All relevant code and scripts of this project are hosted on GitHub and can be accessed via the following link: https://github.com/UoB-DSMP-2023-24/dsmp-2024-group-29.git