Transactional Data Analysis using Machine Learning in the Banking Sector

Kishan Shetty, Amir Sahil Shailk

Abstract—The use of transactional data analysis in the banking sector has become increasingly important in recent years. With the rise of digitization, banks have access to vast amounts of data that can be used to improve their operations and make informed business decisions. Transactional data analysis allows banks to assess the creditworthiness of loan applicants, determine interest rates, and target marketing efforts to retain loyal customers. In this study, we reviewed the literature on the use of statistical and machine-learning models for transactional data analysis in the banking sector. Our findings show that these models can be used to create insights into various aspects of customer behaviour and the performance of banking operations. The results of this study demonstrate the potential for transactional data analysis to provide valuable information to the banking sector and help drive growth and profitability.

Index Terms—Transactional Data, Analytics, Machine Learning, Banking

1 INTRODUCTION

As a part of the post-graduation programme, this project was undertaken with Lloyd's Bank which supplied the synthetic data for the analysis. The banking sector generates a vast amount of data every day, including information on account transactions, loans, and customer demographics. This data has the potential to provide valuable insights into customer spending patterns and capitalization opportunities. Transactional data analysis is the process of using data analytics techniques to extract meaningful information from transactional data in the banking sector. This information can then be used to inform business decisions and improve performance. The Lloyds Banking Group generates vast amounts of daily data, including account information, transactions, loans, and demographic data. Our goal is to analyze and visualize this transactional data in order to uncover any spending patterns and discover new capitalization opportunities. The purpose of our analysis is to use advanced data analytics techniques to gain insights into customer behaviour and generate business growth opportunities. The analysis process involves data preparation, manipulation, and analysis through various methods such as machine learning and statistical analysis. The results of this work will provide insights into unusual spending patterns and other opportunities, including salary analysis, loan repayments, and individual expenditure. The analysis is divided into three main areas: statistical analysis, visual analysis, and machine learning. Statistical analysis includes the investigation of correlations, variance, mean, and other data dimensions. The use of Tableau software allows us to create high-quality plots to identify patterns and dependencies. Machine learning models were constructed using Python and libraries to achieve the objectives.

• *Kishan Shetty, MSc Data Science, University of Bristol, Bristol, UK, Corresponding author E-mail: shkishan98@gmail.com*
• *Amir Sahil Shailk, MSc Data Science, University of Bristol, Bristol, UK, E-mail: skasahil026@gmail.com*

2 LITERATURE SURVEY

The banking sector has become one of the most data-intensive industries, generating millions of pieces of data every day from account information, transactions, loans, and other financial demographics. This large amount of

data creates opportunities for banks to improve their operations and make better business decisions by identifying patterns, trends, and individual statistics of customer spending. In recent years, the use of transactional data analysis in the banking sector has increased significantly. Transactional data analysis refers to the process of collecting, cleaning, transforming, and modelling customer transaction data to uncover insights and identify business opportunities. These insights are used to support decision-making, optimize marketing efforts, detect fraud, and enhance customer experience. [1]

One of the key benefits of transactional data analysis in the banking sector is the ability to detect fraud. Fraudulent activities can cause significant losses for banks, and traditional fraud detection methods such as manual reviews and rulebased systems are often slow and ineffective. Transactional data analysis allows banks to identify unusual patterns in customer spending, which can indicate potential fraud. Machine learning algorithms can be used to classify transactions as normal or suspicious, and the results can be used to enhance fraud detection systems. Another benefit of transactional data analysis in the banking sector is customer segmentation. Customer segmentation refers to the process of dividing a large customer base into smaller groups based on common characteristics such as spending habits, demographics, and behaviour.

By analyzing transactional data, banks can identify different segments of customers with different spending patterns and target their marketing efforts accordingly. This leads to a more personalized and relevant customer experience, which can increase customer loyalty and drive business growth. [2]

In addition to customer segmentation, transactional data analysis can also be used for customer profiling. Customer profiling refers to the process of creating a detailed picture of a customer's behaviour, preferences, and demographics based on their transactional data. By analyzing transactional data, banks can gain a deeper understanding of their customers and develop more effective marketing strategies. For example, banks can analyze spending patterns to identify when and where customers are most likely to make purchases and target their marketing efforts accordingly. [2] According to Agrawal, data analytics in finance offers a new perspective on the interpretation of data. It holds a vital position in predicting future outcomes based on past statistics, improving performance, and utilizing current data. This field has significant importance in financial decision-making. In a financial case study, Ouahilal et al. utilized predictive analysis, which involved multiple regression models, including decision tree regression and multiple linear regression. The aim was to predict future spending over a specified time frame and to compare the performance of the models. [3]

Cai et al. studied the use of clustering techniques in financial data analysis, focusing on k-means clustering. This algorithm partitions the data into k clusters and calculates the mean as the centroid of all data within each cluster. The advantage of this method is that it treats all data equally, allowing it to be visualized in multiple clusters and allowing for the examination of specific patterns or trends within each cluster. Yeomen delved into the application of random forest methods in finance. Random forest is beneficial as each tree is independent, mimicking different personality types in terms of financial spending patterns, and eliminates bias within the model. [4] While classification and regression models are the primary models in financial data analysis, Yeomen highlighted the shift in perception from asking "if X increases, how much will Y increase?" to "is value X going to change?" which can have a significant impact on the overall understanding of the problem being addressed. Cankurt et al. compared the performance of a multi-layer perceptron model and a linear regression model in forecasting tourist arrivals. The performance of the MLP model was evaluated based on mean squared and absolute errors, and it learns by adjusting the strength of connections through the weights in the input data. This model could be useful for predicting individual spending based on a training dataset and detecting outliers to identify ab-

normal spending patterns. [4]

Despite the numerous benefits of transactional data analysis in the banking sector, there are also some challenges associated with it. One of the main challenges is the quality and accuracy of the data. Transactional data can be messy and contain errors, which can affect the accuracy of the results. Therefore, it is important for banks to have robust data cleaning and preprocessing procedures in place to ensure that the data is of high quality and ready for analysis. Another challenge is privacy and security. Transactional data often contains sensitive information about customers, such as their spending habits and financial information. Banks must take steps to protect this information and ensure that it is not misused. This includes implementing strict security measures and following privacy regulations, such as the General Data Protection Regulation (GDPR) in Europe. [5]

Finally, the choice of methods and tools for transactional data analysis can also be a challenge. There are many different methods and tools available, ranging from basic statistical techniques to advanced machine learning algorithms. Banks must choose the right methods and tools for their specific needs and goals, taking into account factors such as the size of the data, the complexity of the analysis, and the available resources. [5]

In conclusion, transactional data analysis is a valuable tool for the banking sector, providing opportunities for banks to improve their operations and make better business decisions. Despite the challenges associated with it, the benefits of transactional data analysis in terms of fraud detection, customer segmentation, and customer profiling make it an essential part of modern banking practices. As the amount of data generated by the banking sector continues to grow, the importance of transactional data analysis will only continue to increase, making it a crucial tool for success in the industry. [6]

3 METHODOLOGY

Lloyds Banking Group (LBG) set the task of finding novel ways to use daily transactional data. The vagueness of the problem statement left the direction of the project open-ended; our chosen goal was to discover new ways of generating revenue for LBG through our analysis. Naturally, genuine transaction data is treated with absolute confidentiality, therefore all data supplied to us is entirely generated by simulation. Without knowing the model used to simulate said data, we must assume that it accurately mimics authentic transaction data – allowing our research to have real-world applications. All data was supplied directly by LBG, the initial dataset contained over 9,000,000 transactions. Each transaction can be seen as a vector X_{tr} which typically contains information on the personal account number N_0 , destination account number N_1 , amount a and date d of transaction. The secondary expanded dataset contained only 300,000 observations. The secondary dataset gives X_{tr} two additional characteristics: account balance b and time t of transaction.

3.1 How do banks make money?

After the relaxation of Covid-19 restrictions, there has been a surge in economic activity, especially in the mortgage/credit lending sector. To leverage this growth, we explored ways to use transaction data to market credit cards or mortgages effectively. By analyzing the transaction data, we could identify potential customers and offer them better rates to attract more business and open more accounts. We also analyzed the commercial industries that are experiencing significant growth/spending to identify profitable investment opportunities or business loans. [7]

3.2 Investigative Methods

The project involved analyzing two datasets using different techniques. The first technique used was statistical analysis, which was conducted through Python and Tableau.

This provided some preliminary insights on the structure of the data and the spending patterns. The second technique was machine learning, which involved using numerous models.

One interesting observation from the statistical analysis was that the initial dataset was about 30 times larger than the second dataset. Additionally, the first dataset contained more high-value transactions, with an average expense of £21.96, compared to £1.39 for the second dataset. The exploratory analysis also looked at transactions spread over the day of the week. The first dataset had a steady increase of transactions across the week, with a spike on Friday, while the second dataset had alternating days of high to low counts of transactions, and the cause of this pattern is unknown. Data visualization was a crucial part of the project, and we utilized Tableau software to create compelling visualizations that would help us better understand the data. Tableau allowed us to create a variety of charts and graphs that were both informative and easy to read. Additionally, we were able to manipulate the data within Tableau before creating visualizations, which saved us a lot of time and effort.

Our main goal was to analyze the time and amount spent by specific individuals, and we used different types of charts to showcase this data. For instance, bar charts were ideal for displaying values, such as expenditure or turnover, for a particular user or organization at a given time. On the other hand, line graphs were effective in representing trends or comparing data over time. For example, we could use a

line chart to track how much an individual spends each day of the month, with each point on the line representing their spending on a particular day. This approach helped us to identify insights and make informed decisions.

1) Category Classification: In order to identify promising investment prospects for the bank, we employed a KNN classification technique that leveraged all the available characteristics in vector X_{tr}. We first partitioned N_1 into five distinct sectors, based on their respective business types, which included Grocery, Food/Drink, Fashion, Misc., and Academic/Leisure. We intentionally excluded any personal transactions (i.e., those between personal accounts) from the dataset. We then proceeded to train the model using a standard 80:20 split, and carefully fine-tuned various hyperparameters, including n neighbor, metric, and weights, via a grid search. [8] The central equation driving this approach is chiefly based on calculating the distances between various data points and is given by,

$$dist(x,z) = \left(\sum_{r=1}^{d} (x_r - z_r|^p\right)^{1/p} \quad (1)$$

2) Customer Segmentation: We explored a technique to categorize customers based on their spending habits, to enhance the marketing of interest rates and assess creditworthiness. We adopted the K-means clustering method and standardized scaler for normalizing feature values to divide customers into groups based on their spending habits and month of spending. We utilized the elbow method to determine the number of clusters. The customers were separated into four clusters, signifying their eligibility for various bank promotions. We reduced dimensionality and improved clustering outcomes using principal component analysis after the initial clustering round. Transaction details such as the month of the transaction, account number, spent amount, and spending category were taken into consideration for creating a basic segmentation model. [9] The algorithm's general equation is based on distance and feature values which is given by,

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left\| x^i - \mu_k \right\|^2 \quad (2)$$

3) Forecasting: Forecasting is a process of predicting future values based on past observations. In this project, we attempted to forecast individual account spending using KNN regression model. By comparing predicted and actual spending amounts, we can identify any anomalies or fraudulent transactions. We also considered the possibility that an individual may have multiple bank accounts, which could affect the accuracy of our predictions. Our goal was to identify accounts with larger errors and investigate them

further. However, it's important to note that outliers must be carefully examined before taking any action, as they may not necessarily indicate fraudulent behaviour. Overall, the project aimed to improve the accuracy of banking information forecasting and prevent financial fraud. [7]

4) Anomaly Detection: Fraudulent transactions pose a great threat to the integrity of financial institutions and the trust of their customers. Anomaly detection is a powerful tool that can help banks identify and flag suspicious transactions by analyzing patterns and previous spending history. By doing so, banks can significantly reduce the risk of fraud and prevent customers from losing their hard-earned money. To implement this, we had to create a new feature in the dataset called "Outlier," which would train the model to identify potentially fraudulent transactions. This column is marked as 1 or 0 based on a set of rules that help the model distinguish between normal and suspicious activity. It's crucial for banks to stay vigilant and proactive in preventing fraud, as it not only protects their customers but also maintains the integrity of the financial system as a whole. [10]

For every customer, the mean of their transaction amount was taken throughout the year of the data, and then a permissible constant C was decided (in our case, 1000). So, for every transaction of that customer if the amount spent was greater than the sum of the mean and the constant then it was flagged as fraud (1) else normal (0)

$$\text{Outlier } (\mathrm{X_i}) = \begin{cases} 1 & \text{if } \mathrm{Amount(Xi)} >= (\mathrm{Mean(Xi)} + \mathrm{C} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

3.3 Evaluation Metrics and Justification

When building the KNN regression model, we selected metrics that were appropriate for a non-linear model. Although the R2 metric is acceptable for non-linear models, it is mainly used for linear regression. Hence, we chose to use mean squared error and mean absolute error as our metrics. While we still output the R2 score to gauge model performance, our primary focus is on mean absolute error as it shows us the difference between predicted and actual values for each data point.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \qquad MAE = \frac{\sum_{i=1}^{n}(y_i - x_i|}{n} \tag{4}$$

To determine the accuracy of each class, we examined precision and recall statistics in our classification analysis. Additionally, we employed the weighted f1 score to determine the overall accuracy, taking into consideration the amount of data in each class. Given the limited amount of outlier data throughout the entire dataset, this was particularly important.

$$\begin{aligned} \text{Precision } &= \frac{\text{True Positive}}{\text{True Positive } + \text{ False Positive}} \\ \text{Recall } &= \frac{\text{True Positive}}{\text{True Positive } + \text{ False Negative}} \end{aligned} \tag{5}$$

The choice of the algorithm in machine learning is critical to achieving the desired outcome, and the justification for the chosen algorithm is crucial to the success of the project. The K-Nearest Neighbor (KNN) algorithm was chosen due to its ability to save the whole training dataset, which makes it easy to represent itself and calculate the similarity between the input sample and each training instance. This results in faster prediction times and a wide range of distance measures available to match the input structure.

K-Means is another algorithm that is simple to build, scalable and adaptable, and guarantees convergence. The Expectation-Maximization technique is used to allocate data points to the closest cluster and calculate the centroid of each cluster. This algorithm is suitable for large datasets, and it performs well when the data points are evenly distributed. Principal Component Analysis (PCA) is used to reduce the dimensions of the data and remove correlated features, which can improve clustering performance and reduce overfitting. This algorithm is useful when dealing with high-dimensional datasets that have a large number of features. Isolation Forest is an ensemble algorithm that employs Isolation Trees to identify anomalies in the data. Anomalies are observations that are infrequent and dis-

tinct, making them easier to detect. This algorithm works by generating partitions on the dataset in a recursive manner, randomly choosing a feature and selecting a split value for the feature. If the anomalies require fewer random partitions to be isolated than ”normal” points in the dataset, then the anomalies will be the points in the tree with the shortest path length.

In conclusion, the choice of the algorithm in machine learning should be justified based on the requirements of the project. Each algorithm has unique features, advantages, and disadvantages, and it is essential to choose the one that best suits the project's needs.

4 DATA DESCRIPTION AND PREPARATION

The data comes from the UK's largest retail bank as well as one of the UK's major retail financial service companies – Lloyds Banking Group. During the entire process of the project, data was released in two individual sets. The first data set comes from the data synthesizer developed by LBG, which can be used to generate some content that is very similar to the actual data. The second data set is richer and more realistic than the first data set and includes income and transactions to banking groups themselves.

The first data set is a simulation of multiple individuals account spending to business accounts and other banking accounts. The data consists of four features: the date of the transaction, account number/name to and from and the amount spent. There are 9,629,737 instances within the dataset. There were multiple null values within the dataset which during the process of this project was either imputed or removed for specific methods. The second data set contains a simulation of expenditure and income for individual accounts to which timestamps are added which could allow us to look more closely at subscriptions, bank repayments and salaries as these mainly occur at 00:00. The the second data set also includes two different columns for an account name or number, which is slightly different to the first data set as it only contained one column. The second set also includes more detail in the business account names, compared to the first simulation where simplistic names such as “COFFEE ” was used, this uses actual store names such as “Tkmaxx” and “Co-op” which could allow us to group more efficiently.

To prepare the data in the first data set for model implementation, column names were changed from e.g., “from totally fake account” to “account from” for simplicity reasons. Dates were divided into days, months and years for reasons used for regression, classification, and clustering. Dates are necessary for statistical analysis so in some situations this column was removed, and, in some situations, this column was used. Business accounts were classified into larger groups for heatmap data visualization. Due to many unique business names, business names were classified into 11 different groups so that the visualization was not overcomplex. Kishan used imputation for null values, where the most occurring value per account replaced that specific null value.

The second dataset did not need a lot of preparation, in some scenarios the time stamps were changed to minutes and the dates were replaced with individual day, month, and year columns. Null data was removed and business names and account numbers that the transaction was going to be merged into one column. Filtering took place to look at salaries and loan repayments back to the bank by taking the statistics from the timestamp at 0:00 and filtering through income/expenditure for accounts at that time. Members of the groups looked at null data individually to find any specific patterns within that mini-data set.

5 RESULTS AND DISCUSSION

While keeping the larger goals of the business in mind, we conducted various experiments on multiple datasets and discovered valuable insights. Our aim was to find ideas to help the business increase its sales and revenue, and we followed sound data analytics practices. We broke down the

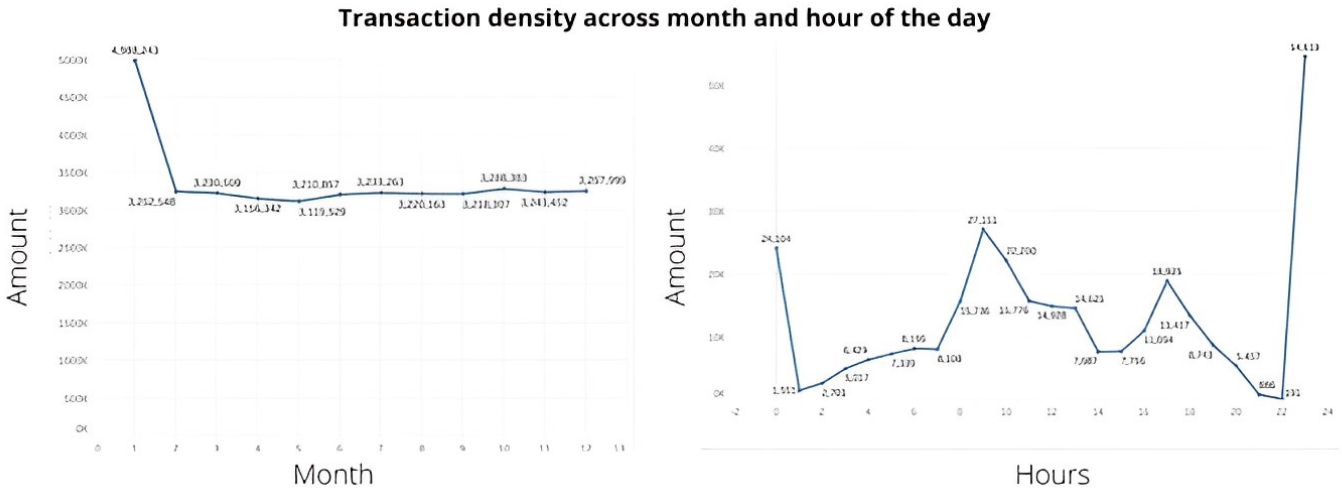

Fig. 1. Customer Transaction Frequency

task of experimentation into three main categories: Visual Analysis, Statistical Analysis, and Machine Learning. The first two categories helped us scope for insights, while the latter focused more on the technical aspects, such as classifying transactions into organizational categories, predicting future customer spending, segmenting customers based on their spending locations, and identifying potentially fraudulent transactions.

5.1 Statistical and Visual Analysis

Heatmap was generated that identified which organization the sector had the most spending spread throughout the months of the year portrayed in Fig. 2. This helped in identifying the sector which had the highest transactions and a huge customer base. Analysing the years' worth of data and mapping the date of the transactions to the day of the week and further taking the average amount that was transacted on that day it was found that surprisingly Friday had the lowest amount spent despite being a weekend.

data set and 10-11 am on the second data set was the golden hour for transactions as depicted in Fig. 1.

The graph in Fig. 3, showcases the typical annual spending of customers across three categories: entertainment,

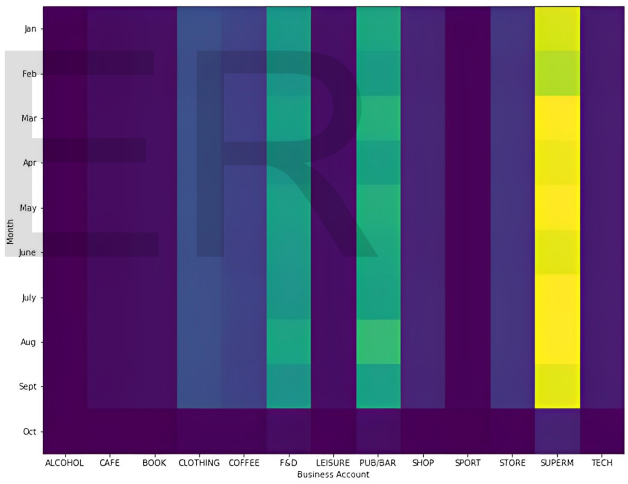

Fig. 2. Heatmap of sector revenue month-wise

food, and retail. Interestingly, entertainment is the least popular category with a maximum average spend of £3,500, while food and retail have significantly higher averages of £11,800 and £13,000, respectively. These findings suggest that the entertainment industry has a narrower scope of expenditures than the food and retail industries. This knowledge could be beneficial for banks when deciding on loan approvals for businesses within these categories.

The visual representation in Fig. 4 displays the top tencompanies that receive the most significant yearly cus-

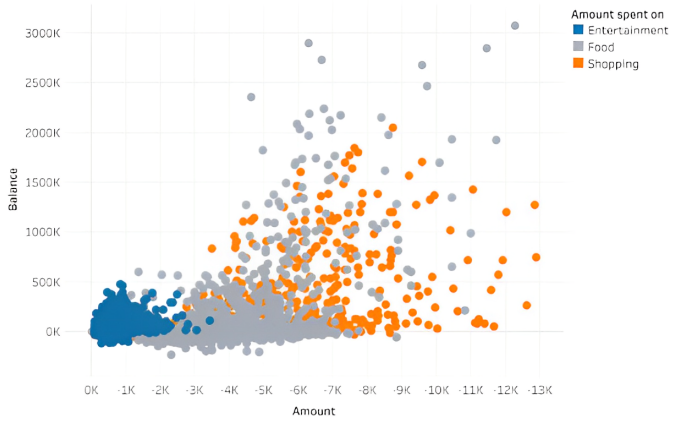

Fig. 3. Customers spent on entertainment, food and shopping

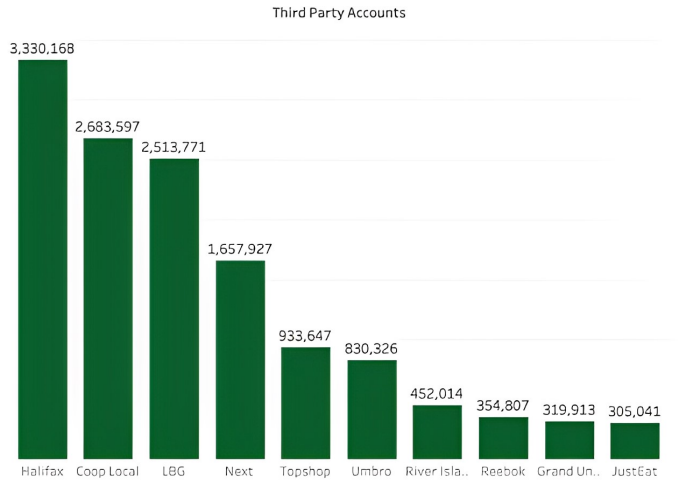

Fig. 4. Top 10 profiting businesses

tomerpayments. Among the top three amounts received, the first and third are credited to banks, namely Halifax and LBG. Thisindicates that the transferred amount could be a repaymentof a loan or credit card payment by the customers to thebanks. As per this estimate, customers repay around 6 millionpounds annually to Halifax and LBG. Apart from banks,Coop Local, a supermarket, receives approximately 2.6 millionpounds yearly, which reinforces the previous visualization'sobservation that food and shopping businesses have the highestcustomer spending. The remaining brands listed in the visual-ization also belong to the retail and food sectors.

During the second half of the data, January was the monthwith the most transactions, followed by a sharp decline inFebruary, which remained constant until the end of the year.However, in the first half, there was a high volume of transac-tions over the first three quarters, but it dropped significantlyin the last quarter. An analysis was conducted to determinewhich business organizations customers were spending moreor less on, using a ratio plot. It was observed that successfullarger businesses generated a lot of revenue in the first quarterand it declined over the subsequent quarters. A subset ofcustomers who made recurring payments to banks such asHalifax and LBG was analyzed in detail to identify mortgage,credit card bills, or EMI payments. Monthly income andspending habits were also analyzed to identify customers whosave significantly more money for investment opportunities,and customers who spend more to offer them credit cardswere targeted. Fig. 5, shows the graph of all the customersof Lloyd's Bank with their salaries and how much money arethey paying back to the bank in total every year in the formof mortgages or personal loans or credit.

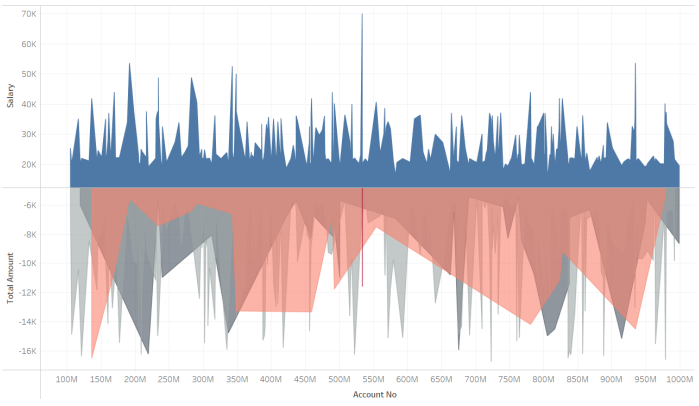

Fig. 5. Salary vs Repayments to LBG

Banks typically earn revenue from credit card interest on late bill repayments. One of the primary considerations for issuing credit cards is a customer's annual income. We analyzed customer income to create a credit card eligibility

plot that divides them into four groups. The first three groups have good annual incomes and can receive cards with varying credit limits. The fourth group has average income and may include students or part-time workers, so it is best to avoid offering them credit cards to prevent them from falling into unnecessary debt. See Fig. 6 for the visualization of customer segmentation. B. Machine Learning Analysis Several machine learning algorithms were utilized to accomplish specific objectives and deliver corresponding outcomes, with a focus on creating business value at the conclusion of each task.

Fig. 6. Credit Card Eligibility

5.2 Machine Learning Analysis

Several machine learning algorithms were utilized to accom-plish specific objectives and deliver corresponding outcomes,with a focus on creating business value at the conclusion ofeach task.

1) Sector Classification: A classification algorithm was implemented. The algorithm we chose was K-Nearest Neighbour and the aim of this task was to classify every transaction that was fed into the model to a particular sector of the organization. There were 4 distinct sectors we considered: Supermarket Grocery, Food Drinks, Fashion, Miscellaneous and Stationary Academic. We were able to classify each transaction into a particular sector based on account number, the amount spent and the date of the transaction. Our overall accuracy of this model was 55%.

2) Amount Forecasting: To predict each customer's transaction amount a regression algorithm was trained and fit. Features like account number, date of transaction and category of spending were considered, and a forecast was made to accurately predict the amount that was likely to be spent by the customer. This is helpful for the business to avoid surprises and be well prepared to manage any crisis that may or may not happen. K-NN regression was provisioned for this task and multiple metrics were used including an R2 which achieved an average of 0.44 and a mean absolute error (MAE) score which varied between a range of 5.00-25.00 for individual accounts. The regression model predictions were used against the true values within a threshold to find predictions that were not within the range of the threshold +/-. This allowed us to highlight outlier data. However, there is a clear emphasis that certain accounts did not contain a lot of instances, this could be due to the individual owning multiple bank accounts. Therefore, further investigation into specific individual pieces of data that contain limited transactions for the account is needed to propose that as fraud or an uncommon payment.

3) Customer Segmentation: t is crucial to know what type of customers a business is dealing with and with that in mind a clustering model was built to segment the customer corpus based on their spending habits. This includes the amount that was spent, the sector of the organization they spent on, and the month of the transaction. To achieve this goal easily, KMeans clustering was employed, and we were able to make 4 clusters of customers as shown in Fig. 7.

4) Anomaly Detection: It is of absolute importance that business-like banks be ready to tackle unforeseen situations like fraudulent transactions, and it is better to be pre-

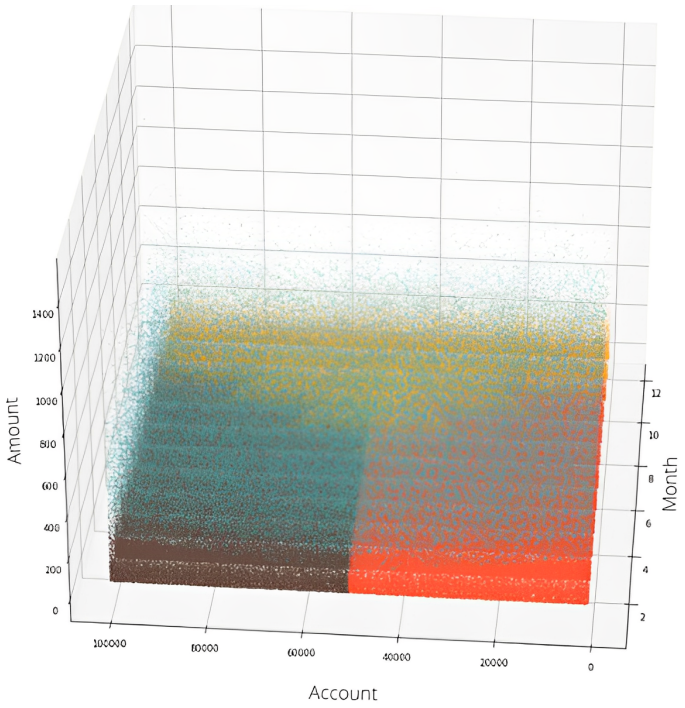

Fig. 7. Customer segmentation

pared to handle them. To attain this task, we have resourced a classification algorithm to classify all the potentially fraudulent transactions based on other features like date, amount, account, and sector of the transaction, we use Isolation Forest to attain the goal. The algorithm works so well and avoids overfitting as much as possible and it generates the results with greater accuracy without hyperparameter tuning itself. By doing so, we have successfully trained and tested the model to get an accuracy of 80%. With the isolation forest algorithm, we were able to deduce the correlation between individual variables in the data set.

Table 1 briefly summarizes the performance metric that was used for the machine learning model and gives thethe metric value for that chosen parameter.

TABLE 1
PERFORMANCE REPORTOFTHE MODELS

Algorithm	Performancre Metric	Value
KNN Classifier	Accuracy	85%
Isolation Forest	Accuracy	80%
KNN Regressor	R2	0.5
	Mean Absolute Error	5-25
K-Means Clustering	Silhouette Score	0.65

6 FURTHER WORK AND IMPROVEMENTS

The implemented work can be extended and improved to achieve better efficiency and move the advanced model to benefit the business. As a part of evolution, it is our responsibility to make use of the work that has been built upon and there are a few areas where this can happen. To begin with, an advanced technique of behaviour analysis can be done based on customer transactions. When we say behaviour analysis, we mean to identify the personality traits of customers and then segment customers to target each segment with its unique offers and benefits. There are traits like materialism and new experience which if predicted properly for customers then banks can target specific credit cards or loans for customers.

Our model's classification, regression and fraud detection can also be further improvised where if a person has a spending record in some sectors, and if a new sector is identified in a truncation, then the transaction could be a fraudulent one and immediately necessary action could take place such as blocking the credit card. This approach can also be merged with the amount spent so that the area spent, and amount spent variables are constantly monitored to avoid fraud.

Personality identification was a key aspect that we tried to implement using random forest. However, this was unsuccessful. To be able to get a model to infer personalities within the data, individuals within the bank would have to take a personality survey to deduce the big five personality traits (extraversion, agreeableness, openness, conscientiousness, and neuroticism) and infer that with the data provided. Due to the lack of this information, this was an incomplete task, and with this information, it would have given us further insight into individuals' personalities to contrast their spending against fraud detection and irregu-

lar spending etc

7 CONCLUSION

Throughout this project, three distinct types of analysis were employed - statistical, visual, and machine learning - each of which proved to be effective in its own way. By examining spending patterns, it was discovered that customers tend to spend more during specific times of the day and on certain days of the week. Furthermore, it was observed that January had the highest overall expenditure across all accounts while other months had lower expenditure. By comparing these findings with industry statistics, it was determined that retail supermarkets and food and drink establishments in the hospitality sector generate the most revenue. Additionally, anomaly detection and customer segmentation were implemented to prepare for unexpected events and to target specific customer groups, respectively. In the end, a wealth of insights were uncovered while also considering the business framework. Furthermore, several ideas were brainstormed to help the business generate more revenue and attract new customers while retaining existing ones.

REFERENCES

[1] T. Comp, "COMP TIA. [no date]. How is Data Analytics Used in Finance? [online]. Available from: https://www.comptia.org/content/articles/how-is-data- analytics-used-in-finance: :text=Data%20analytics%20helps%20finance %20teams,detect%20fraud%20in%20revenue%20turnover."

[2] V. Agrawal, "AGRAWAL, V. [8th July 2021]. Finance Data Analyst: 7 Critical Aspects. [online]. Available at: https://hevodata.com/learn/finance- data- analyst/: :text=Financial%20Data%20Analytics%20is %20a,perspective%20to%20the%20financial%20data," 2021.

[3] F. Giannotti, C. Gozzi, and G. Manco, "Giannotti, Fosca, Cristian Gozzi, and Giuseppe Manco. "Clustering transactional data." Principles of Data Mining and Knowledge Discov- ery: 6th European Conference, PKDD 2002 Helsinki, Finland, August 19–23, 2002 Proceedings 6. Springer Berlin Heidelberg, 2002.," 2002.

[4] Y. Yang, X. Guan, and J. You, "CLOPE: a fast and effective clustering algorithm for transactional data," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[5] M. Ouahilal, M.E. Mohajir, M.E.C. B, and E. Moha, "A comparative study of predictive algorithms for business analytics and decision support systems: Finance as a case stud," *International Conference on Information Technology for Organizations Development (IT4OD). 2016. pp. 1-6.*

[6] F. Cai, N.L.K. T, and Kechadi, "Clustering Approaches for Financial Data Analysis: A Survey," in *Clustering Approaches for Financial Data Analysis: A Survey*, (University College Dublin, Ireland), 2016.

[7] A. Yeoman, "YEOMAN, A. [9th February 2022]. Random Decision Forest in Finance: Preparing for The Unexpected. [online]. Available from: https:// www.bairesdev.com/blog/random-decision- forests-in-finance/.," 2022.

[8] S.A. Cankut and Subasi, "CANKUT, S. A. SUBASI. (2012) "Comparison on linear regression and neural network models forecasting tourist arrivals to Turkey". ISSD 2012. 2012.," 2012.

[9] D. Subramanian, "SUBRAMANIAN, D. [8th June 2019] A Simple Introduction to K-Nearest Neighbors Algorithm [online]. Available from: https://towards datascience.com/a- simple-introduction-to-k-nearest-neighbors-algorithm- b3519ed98e.," 2019.

[10] B.R. Mihaela, "Statistical methods applied to the financial analysis of a publicly funded investment

project," *7th International Conference on Applied Statistics*, pp. 304–313, 2014.