Lloyds Transactional Data Analysis

Lakshmi Pranatharthy Haran
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom
ci22246@bristol.ac.uk

Swaminathan Ganapathy
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom
mf22210@bristol.ac.uk

Suriyaprakash Vadivelu
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom
ty22216@bristol.ac.uk

Rahul Krishnamurthy
Department of Engineering Mathematics
University Of Bristol
Bristol, United Kingdom
zd22604@bristol.ac.uk

***Abstract*— This report comprehensively analyses two transactional datasets using machine learning and exploratory data analysis techniques. The primary objective is to provide the bank and its clients with valuable insights and applications. Exploratory Data Analysis (EDA), Machine Learning Analysis, and Use Case Analysis are the three major sections of the analysis. By identifying patterns, anomalies, and associations, EDA provided a deeper understanding of the datasets. Key findings include higher transaction volumes on weekends, transaction frequency distribution, and spending patterns in different transaction types. Machine Learning Analysis includes K-means clustering for customer segmentation, Isolation Forest for anomaly detection, and K-Nearest Neighbours for sector classification. In addition, the creditworthiness was evaluated using Logistic Regression and RandomForest Classifier, with the latter showing superior performance. Use Case Analysis focused on churn prediction, categorizing customers into different risk categories based on their transaction frequency. Further, business accounts were manually classified into industries to comprehend the transaction density and expenditure patterns of various industries. The report emphasises the potential for data-driven approaches to enhance banking services, personalize consumer experiences, and optimise risk management strategies.**

I. Introduction

The banking industry is currently experiencing a paradigm shift driven by rapidly evolving customer expectations, technological advancements, and increasing competition. In this dynamic environment, understanding customer behaviour and preferences has become more critical than ever for banks to stay relevant, deliver personalized services, and enhance customer satisfaction. This report presents a comprehensive analysis of transaction data provided by Lloyds Bank, focusing on customer-to-customer (C2C) and customer-to-business (C2B) transactions.

The primary motivation behind this study is to gain insights into the transaction patterns and preferences of Lloyds Bank customers. By doing so, the bank will be better equipped to tailor its products and services to meet the needs of its diverse clientele. Furthermore, understanding the factors that influence customer behaviour can help the bank develop targeted strategies for customer engagement, retention, and satisfaction.

The problem context revolves around the challenges faced by traditional banks in the age of digital transformation. With the emergence of innovative financial service providers, such as fintech companies and digital banks, competition for customers has intensified. This development has heightened the importance of understanding customer transaction behaviour as a means to differentiate and deliver value-added services.

This report aims to provide Lloyds Bank with in-depth insights into customer transaction behaviour, enabling the bank to make well-informed decisions and execute data-driven strategic initiatives. By capitalizing on these insights, the bank will be better equipped to strengthen its market position, fine-tune marketing strategies, elevate customer experiences, and uncover potential avenues for expansion and innovation.

Utilizing customers spending habits and transaction frequency, the bank can create targeted marketing campaigns that align with their unique needs and preferences. This data also allows the bank to offer tailored business and personal loans. Also, on analyzing transaction patterns provides insights into customers' investment preferences, risk tolerance, and financial objectives, enabling Lloyds Bank to deliver personalized wealth management services, including customized investment portfolios, financial planning, and tax-efficient strategies. By identifying customers' cash flow trends and instances of financial stress, the bank can provide arranged overdraft facilities with flexible terms, addressing specific needs during periods of financial difficulty. This comprehensive approach enhances the bank's offerings and supports customer satisfaction.

Moreover, these insights will facilitate the bank in optimizing its marketing efforts, ensuring that the right message reaches the right audience at the right time. This targeted approach will not only yield a higher return on investment but also contribute to a more personalized and engaging customer experience.

The report is organized into the following sections: Section 1 provides a detailed description of the datasets and the data pre-processing steps; Section 2 presents the methodology employed in the analysis; and finally, Section 3 concludes the report with key findings and recommendations for Lloyds Bank.

II. Literature Review

Numerous studies have focused on customer segmentation and profiling based on transaction data. Clustering algorithm like KMeans is frequently employed in these studies to categorize customers exhibiting comparable transaction behaviors. This segmentation allows banks to customize marketing campaigns and product offerings, effectively addressing the distinct needs and preferences of each customer segment [1].

In another research paper, the authors review various data mining techniques used for financial-accounting fraud detection in the banking sector. Investigations in this area gave a clear idea on how the transactional data plays a major role in

classifying customer's wealth management and how these could be used for the application of transaction data to provide customized wealth management services [2].

Research has probed the utilization of transaction data to discern customers' cash flow trends and instances of financial stress [3]. The focus here is on the neural network approach for credit risk evaluation. The basic ideology and the approach used suggest that banks can extend adaptable overdraft facilities based on customers' particular requirements, thus supplying assistance during periods of financial hardship and promoting enduring customer loyalty.

The importance of transactional data for credit risk analysis was explained in detail in this research paper [4] wherein the results suggest that banks can extend adaptable overdraft facilities based on customers' particular requirements, thus supplying assistance during periods of financial hardship and promoting enduring customer loyalty.

Th KNN algorithm showed promising results in various bank related contexts. This was well articulated in this paper [5], where the authors propose a KNN-based model for bank credit risk assessment. The proposed approach utilizes transactional data and other relevant information to evaluate the creditworthiness of customers. The study demonstrates that the KNN-based model outperforms other traditional credit assessment methods, showcasing its potential for enhancing loan decision-making processes in the banking industry.

III. METHODOLOGY

A. Data Extraction

The raw dataset was loaded into a pandas Data Frame. The shape of the dataset was examined to find the total number of rows and columns, which revealed that the dataset 1 contains 12004116 rows (transactions) and 4 columns (features) and dataset 2 contains 174601 rows (transactions) and 7 columns (features). Some of these features include transaction amount, senders account details, recipient account details and transaction dates. After going through the data by using the info() function which provided an overview of the data types, we came across the number of non-null values for each feature.

B. Missing Values Identification

To determine the number of missing values in each column, the 'isnull()' function was used in combination with the 'sum()' function. This analysis revealed missing values in the following columns: 'to_randomly_generated_account', 'not_happened_yet_date', and 'from_totally_fake_account' for dataset 1 and 'Third_Party_Account_Number' in dataset 2. A data frame was created to clearly display the number of missing values in each column.

C. Data Cleaning

The data cleaning process involved addressing the missing values identified in the dataset. For each of the columns with missing values, the mode (most frequent value) was used as an imputation method to fill in the missing data points. The following steps were taken for each column in dataset 1:

- 'to_randomly_generated_account': The mode of this column was calculated and used to fill in the missing values. The number of missing values was then re-evaluated to ensure that all missing values had been addressed.
- 'not_happened_yet_date': Similarly, the mode of this column was calculated and used to fill in the missing values. The number of missing values was re-evaluated to confirm that all missing values had been addressed.
- 'from_totally_fake_account': The mode of this column was calculated and used to fill in the missing values. The number of missing values was re-evaluated to ensure that all missing values had been addressed.

The following steps were taken for each column in the 2nd dataset :

- 'Third_Party_Account_Number': This column had null values for two business accounts namely 'Deliveroo' and 'Halifax' which was replaced by dummy account numbers of '1' and '2' respectively.

After completing the data cleaning process for each column, the final data frame 'data' was verified to have the same shape as the original data frame, confirming that no rows or columns were removed during the cleaning process. The cleaned dataset is now ready for further analysis, ensuring a higher level of data quality and reliability for any subsequent insights and modeling.

D. Data Preparation

During data preparation, the senders account number column's data type was converted to 'object' for better understanding and manipulation. The 'Date' column was split into three separate columns: 'day', 'month', and 'year', and its data type was converted to datetime using the pd.to_datetime() function. A new column, 'day of the week', was extracted from the 'not_happened_yet_date' column using the dt.day_name() function for dataset 1. A list called 'order' was defined to represent the days of the week in their proper order (Monday through Sunday). For the dataset 2 a new column 'Type_of_Transaction' was added which shows whether the transaction is debit or credit. Finally, the prepared dataset was saved as a new CSV file in the S3 bucket, which can be used for further analysis or loaded into a database.

E. Data Engineering

a) S3: Using a file storage will make the files accessible across different platforms seamlessly. The transactional data provided is in CSV formats for which we require a file storage to store the files efficiently. Also, the split and cleaned data is also written to a CSV file.

Amazon Simple Storage Service is an object storage service with highest scalability, durability, availability, performance, and security [6]. It also offers different classes of storage which makes it more cost effective [6].

AWS S3 buckets are used as a file storage service to store the transactional data in the CSV files. Two buckets were created and used such as 'lloydsbanking' and 'lloydscleandata'. The 'lloydsbanking' bucket is used to store the raw data provided by Lloyds. The 'lloydscleandata' bucket will be used by python to store the data in CSV format after data cleaning and after splitting the data.

b) Python: The data provided is in raw format which may have unwanted columns or null values which may impact the analysis process. These values need to be removed or replaced to perform the tasks efficiently.

Python is a programming language which provides multiple libraries like pandas and numpy to perform multiple operations on the data before analyzing the data [7].

Python has been used to clean the data which is in this case to remove the null values from the raw data. Then the cleaned data is split into customer-to-customer transactions and customer-to-business transactions. This data is written to a csv file and uploaded to the 'lloydscleandata' S3 bucket.

c) Talend: ETL is the process of extracting the raw data and transforming it as required for the use and storing the data to a data lake or a database. In a real-world scenario streamlining this process by building a pipeline is required to store the data effectively in the expected format.

Talend open studio is an open-source tool which offers drag and drop methodology to build ETL and ELT pipeline for streamlining data processing [8]. The tool offers integration with different third-party services like AWS, Azure, etc. [8].

Talend open studio has been used to create ETL jobs to streamline the process of loading the data from the csv files to the database. Also, the Talend jobs are used to split the data as required and load it to the database tables.

The data has been grouped based on number of transactions, total amount of the transaction for each month and for the whole year appropriately and stored to different tables for customer-to-customer transactions and customer-to-business transactions. Multiple jobs were created to perform the data splits and group the data as needed.

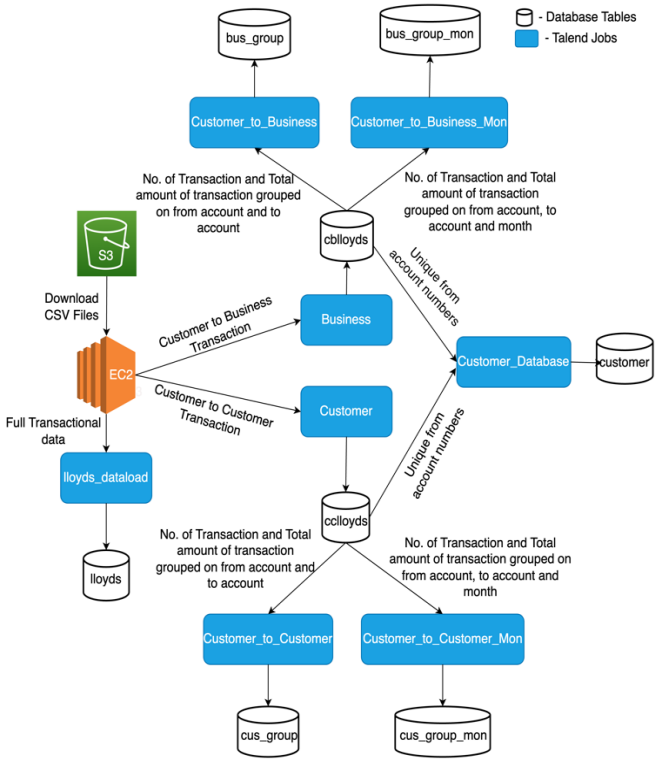

Fig 1. ETL job and Database table integration diagram

d) EC2: A server is required to implement the proposed system to execute the ETL jobs. Using a cloud server will be more efficient in this case as it can be accessed remotely.

Amazon EC2 offers remote compute platform with multiple options for storage, compute power, networking, and operating system. It is highly reliable, scalable with high availability [9].

EC2 has been used to execute the talend ETL jobs to streamline the data. The build of the talend jobs is uploaded to a Linux EC2 instance. Shell scripts are used to download the files from the S3 buckets and to trigger the talend jobs.

e) RDS: The transactional data provided is a structured data which required a relational database storage. Storing the data in a database makes it more efficient and easily accessible. Storing the data in cloud makes it more remote.

Amazon RDS provides a collection of databases in the cloud where the infrastructure is maintained by AWS [10]. It is a highly scalable and highly available service [10].

Amazon RDS MySQL database has been used as the database for the proposed system. Talend ETL jobs extract the data from the CSV files and load them to the database tables.

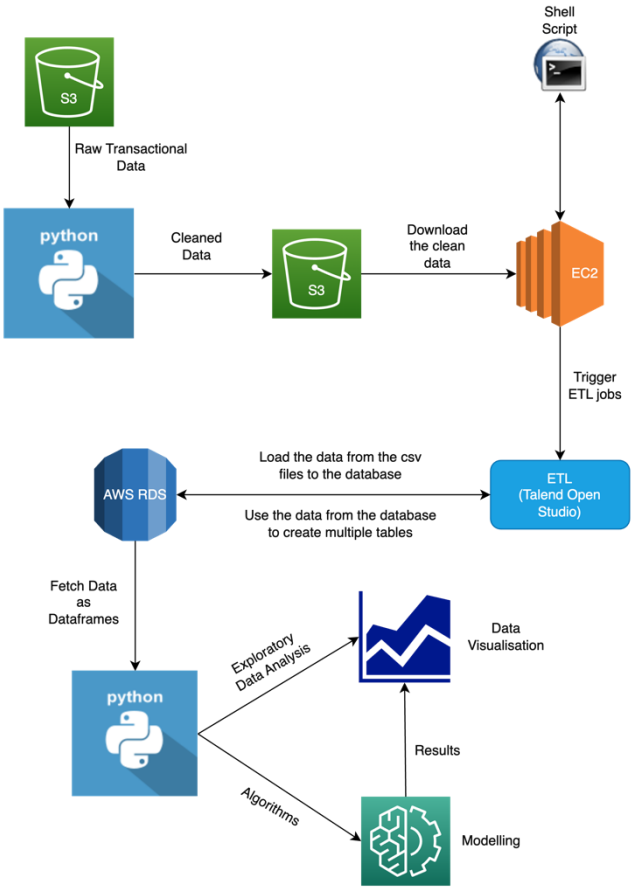

Fig 2. Overall architecture of the proposed system

F. Investigative Methods

The goal of this project is to analyse the customer data to understand customer behaviour, spend patterns and customer credibility. Various investigative methods such as statistics, data visualisations, exploratory data analysis and machine learning has been used to achieve this. Using descriptive statistics a summary of the data was obtained which indicates measures such as central tendency, measure of variability. Data visualisation techniques such as scatterplots, bar plots, histogram has been used to understand relationship between various features. Machine learning methods are used to find and track customer behaviour and categorise them accordingly. Various use cases deem useful for such tasks and various methods have been deployed.

a) Anomaly Detection: It is a technique used to identify datapoints which are significantly deviated from the majority of observations [11]. In such transactional data, detecting anomaly helps in isolating fraud transactions. Isolation Forest, which is an ensemble learning method, is a ML algorithm used for anomaly detection. The algorithm works by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of that feature [12]. This process is repeated recursively, resulting in a binary tree structure. The anomalies are identified as the data points that require fewer splits to be isolated, meaning that they are different from the rest of the data and can be easily separated [12]. An equation is used to determine the anomaly score through Isolation Forest which is given below[13]. The anomaly score (1) is a measure of how anomalous a data point is relative to the other points in the dataset.

$$s(x,m) = 2^{\frac{-E(h(x))}{c(m)}} \qquad (1)$$

Using this measure, we can detect the fraud transactions in the given dataset. With the given data, fraud transactions include, unusual amounts spent at unusual times and location. This can be used by the bank to protect their customers and improve their risk management capability.

b) Customer Segmentation: This is a critical tool for banks to understand their clientele and provide personalised services. With its implementation, banks can improve customer experience, satisfaction and loyalty. K-means Clustering, which is a popular unsupervised ML algorithm, has been used to deploy segmentation [14]. The distance between two datapoints are used to categorise them into groups. The formula for Euclidean distance (2)[15] between two points (x1, y1) and (x2, y2) is:

$$d = \ sqrt((x2 - x1)^2 - (y2 - y1)^2) \qquad (2)$$

Identifying relevant features to be used to group customers is a crucial step. The amount of each transaction and the overall number of transactions has been used to group similar customers together. Banks analyse and understand the inherent characteristics of these groups. One group may be high-spending luxury shoppers, while another may be budget-conscious debit card users. This data may help banks focus marketing efforts, promote products, and optimise pricing.

c) Sector Classification: KNN, a supervised machine learning algorithm, is used to classify the customer transactions into various categories such as food/drink, fashion, academic/leisure and so on, by leveraging the input vectors provided in the dataset [16]. It operates by identifying the K nearest data points in the training set to the target data point, and then classifying the target data point based on the majority class of these K neighbours [16]. Performing sector classification especially on customer to business transactions can be helpful for banks to know the common categories and improve their services or offer new products. For instance, a bank may decide to offer a credit card with discounts on fashion purchases if they observe that a significant number of their consumers are making fashion-related transactions.

G. Customer Credibility

This refers to the extent to which a bank can rely on its customers to repay their loans and fulfil their financial obligations. This is an important measure as it allows the banks to assess the risks involved in providing customers with credit cards and personal loans [17]. Financial behaviour is a key factor in determining customer credibility. It is important to analyse transactional data since it illustrates spending behaviours, balance and much more.

a) Credit card/Personal Loan Eligibility: The user's account balance and transactional data patterns are significant factors that credit card companies consider when determining if an individual is eligible for a credit card. The bank can determine the user's creditworthiness and eligibility for a credit card by analysing their historical account balance and transactional data patterns. The user's account balance may be checked to determine this. The user may be qualified for a credit card if they have a big account balance and have kept it for a long time, which indicates good consistency and will also in turn reflect as a good credit score. However, if the user has a low account balance or is repeatedly overdrawn, it may disqualify them as they are deemed unfit to be trusted with the bank's money. Both Logistic Regression and RandomForest Classifier has been used to determine eligibility majorly based on balance of the customer.

Credit cards are beneficial to both customers and banks. It enables customers to receive rewards on making purchases. These incentives might take the shape of money, travel credits, points redeemable for purchase, or other benefits. This encourages customers to use credit cards instead of cash or debit cards, which helps build credit history and improve their credit score.

Credit card income helps banks in various ways. First, they earn interest on client amounts carried over from billing cycle to billing cycle. Second, they levy yearly, late, cash advance, and balance transfer fees. Thirdly, interchange fees, which retailers pay to accept credit cards, generate income. Finally, they may sell co-branded credit cards with shops or other businesses that give incentives or discounts on purchases.

b) Churn Prediction: This is a technique used by banks to predict whether a customer is likely to stop using their services and close the bank account in the near future [18]. In

the context of transactional data, churn can be defined as those customers who barley have any transactions compared to the mean number of transactions of the general customer base. By implementing churn prediction, the bank can take proactive measure to retain customers and avoid risks associated with losing potentially valuable customers.

For the given dataset, number of transactions has been used as a base to determine churn prediction. This prediction can be used by the bank in effective ways. For instance, banks might provide frequent clients cashback or reward points. Banks can also contact clients who have lowered their transaction frequency with personalised offers based on their transaction history. If many clients leave the bank at the same time, this is also an indication of poor customer service. Early detection allows them to prevent churn. Banks can increase their profits by cross-selling and up-selling additional financial products by retaining customers.

IV. INFERENCE AND RESULTS

Multiple experiments have been carried out on both the datasets provided. These experiments were carried out in a manner that will benefit both the customer and the bank. On a high-level the analysis is divided into three namely- 'Exploratory Data Analysis', 'Machine Learning Analysis' and 'Use Case Analysis'. These experiments were carried out for both the datasets independently where few experiments may be the same for both. More experiments were carried out on the 2nd dataset because it contained more features compared to 1st dataset.

A. Exploratory Data Analysis:

EDA provides a deeper understanding of the dataset by identifying patterns, anomalies and relationships that exist. Different EDA steps have been carried out for each dataset. These experiments are carried out for both customer-to-customer and customer-to-business transactions but either one of the results is displayed in this report as appropriate.

Below is the inference and result of the experiments done on the 1st dataset:

Days of the week Transactions: Bar graph has been plotted to depict the number of transactions occurred on different days of the week. It is clearly visible from Fig 3 that the number of transactions is more on the weekends compared to the ones on weekdays. Below is the plot of customer-to-business transactions,

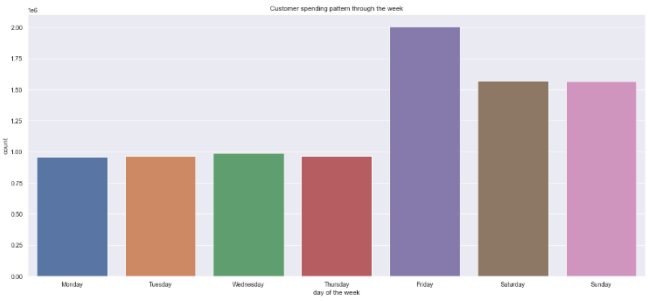

Fig 3. Number of transactions for each day of the week

Transaction Frequency: Histogram, in fig 4, is generated to present the frequency of the number of transactions. The below plot shows that a greater number of customers have transacted 0-5 times to the same recipient in the customer-to-customer transactions.

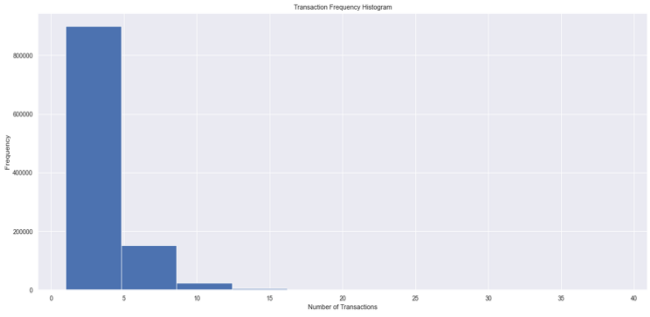

Fig 4. Frequency of number of transactions

Customer Transactions: The histogram, in Fig 5, depicts the amount of money a customer spends in a single transaction. The plot clearly states that most people transact around 200 to 500 pounds in a single transaction and very less people have transacted above 100 pounds while considering customer-to-business transactions.

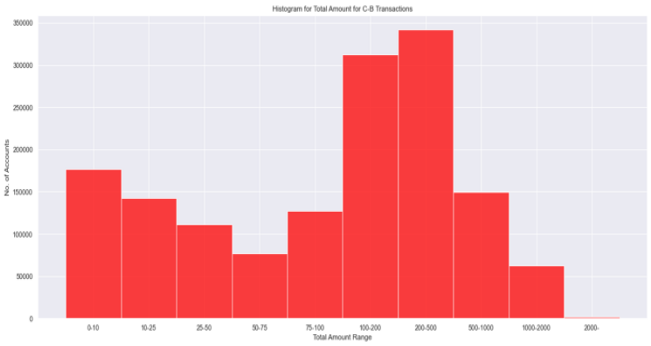

Fig 5. Money spent in a single transaction by customers

Below is the inference and result of the experiments done on the 2nd dataset:

Amount-Frequency Analysis: Scatter plot in Fig 6 is plotted to depict number of transactions and the total amount sent. This shows that more transactions are done with amount less than 10000 pounds and a smaller number of transactions if the amount of the transaction is high.

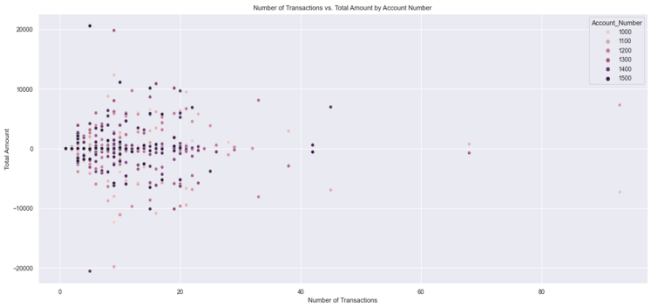

Fig 6. Total Amount sent and Number of transactions

Popular Businesses: A bar graph in Fig 7 was plotted which depicts the number of transactions for different business accounts. The graph illustrates the top 10 business accounts for customer-to-business transactions.

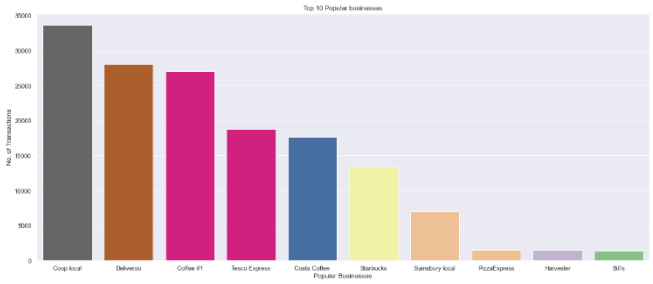

Fig 7. Top 10 business accounts based on number of transactions

Customer Expenditure: The scatter plot in Fig 8 shows an in-depth analysis of the spending patterns of customers with respect to their balance. Customers with a low balance and high spending are the problematic ones that the bank must focus on to prevent any risks.

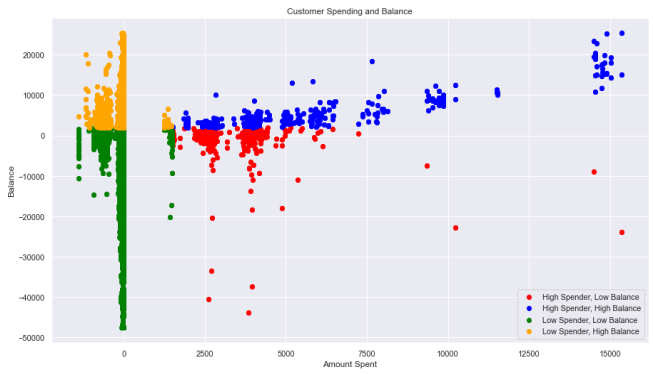

Fig 8. Customer segmentation based on balance and amount spent on single transaction

Customer Income: Salary has been paid by few business accounts to the customers which has been identified using the type of transaction column and plotted the different business accounts which paid salary with the help of a bar chart in Fig 9.

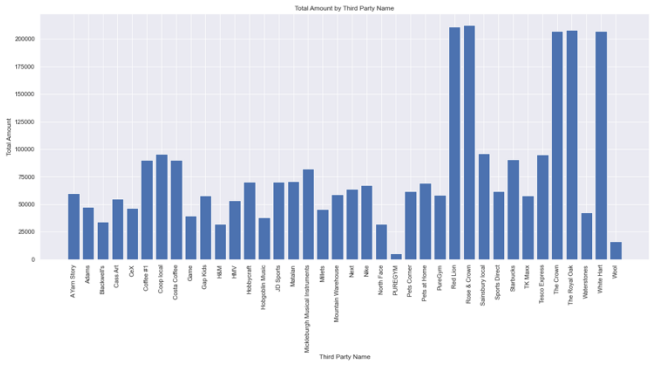

Fig 9. Business accounts paying salary to customers

B. Machine Learning Analysis:

Customer segmentation: This was carried out with the help of K-means clustering algorithm for both the datasets separately and below are the results of the experiment.

Four clusters of customers can be found for the 1st dataset as shown in the scatter plot, Fig 10. The fourth cluster is the outlier which is not visible in the graph. The anomaly detection for these outliers is carried out further in the report.

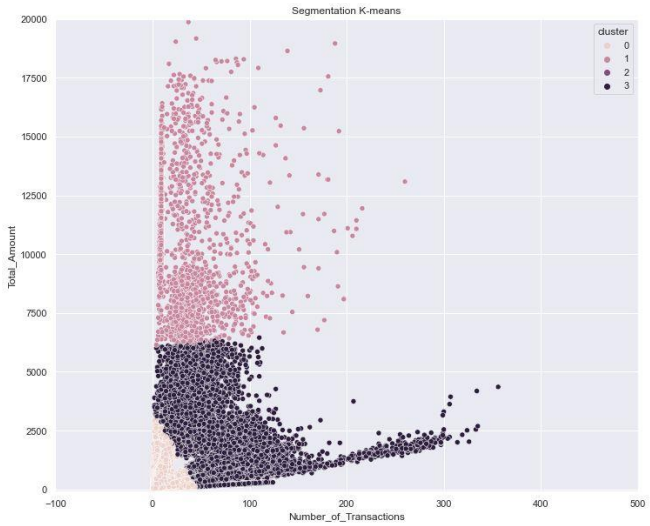

Fig 10. Customer Segmentation for dataset 1

Four clusters of customers can be found for the 2nd dataset as shown in the scatter plot, Fig 11

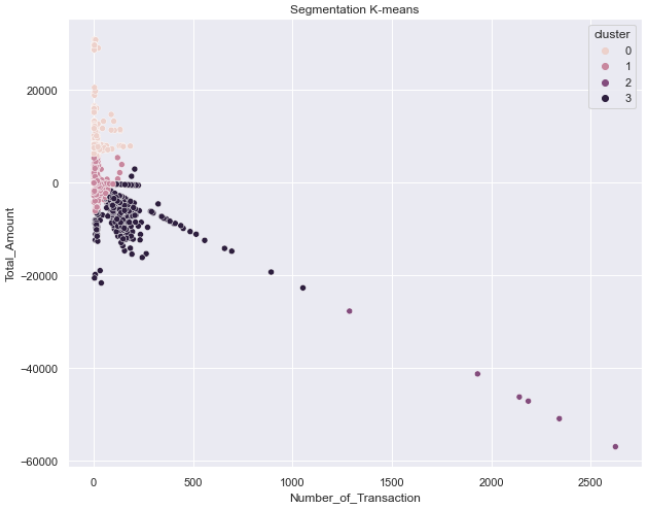

Fig 11. Customer Segmentation for dataset 2

The following experiments are carried out only for the 2nd dataset as the 1st dataset didn't have the columns to carry out the experiments.

Anomaly detection: This has been implemented for the 2nd dataset using the 'Number_of_Transaction' and the 'Total_Amount' columns with the help of the 'Isolation Forest' algorithm. Based on the model's output, it is evident from the scatter plot, Fig 11, that there are quite a few anomalous transactions (red data points) compared to normal transactions (blue datapoints). Using this anomaly predictor, the bank can take proactive measure to improve customer protection.

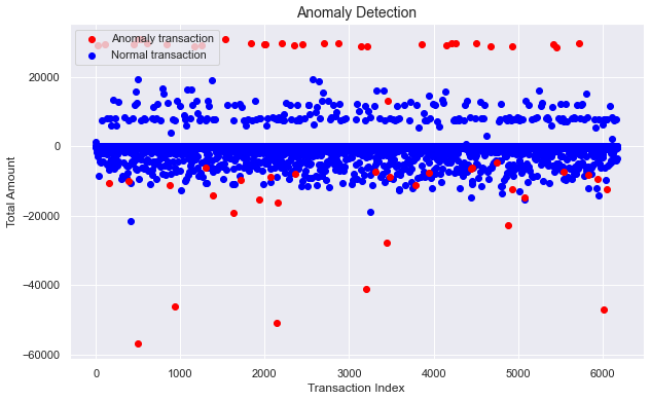

Fig 12. Anomaly transactions and Normal transactions

Credit card eligibility: To identify worthy customers for the bank to provide credit card or loans 2 machine learning models have been trained and compared to classify a customer as 'Creditworthy' or 'Not Creditworthy' based on the recurring balance of the customer. 'Accuracy Score' metric as in TABLE I, has been used the assess performance of the model and below mentioned scores were obtained.

TABLE I. Model Evaluation

Model	Accuracy
Logistic Regression	0.82
Random Forest	0.95

Sector Segmentation: A machine learning model was designed using the K-Nearest Neighbors classifier to classify the sector on which the transaction has been carried out. The model presents with 93% f1-score, as seen in TABLE II, in predicting the sector given the senders account number, recipient account number and the amount of the transaction.

TABLE II. KNN Model performance metrics

	precision	recall	f1-score
0	0.326531	0.271186	0.296296
1	0.564356	0.457219	0.505170
2	0.948503	0.944104	0.946298
3	0.912355	0.933392	0.922754
4	0.799729	0.623418	0.700652
accuracy	0.928381	0.928381	0.928381
macro avg	0.710295	0.645864	0.674234
weighted avg	0.927153	0.928381	0.927486

C. Use Case Analysis:

Churn Prediction: This has been done for the data using the number of transactions carried out by each customer. Based on a threshold each account has been termed as 'High Churn Risk', 'Medium Churn Risk' and 'Low Churn Risk'. The scatter plot, in fig 12, shows all the accounts in different churn risk category.

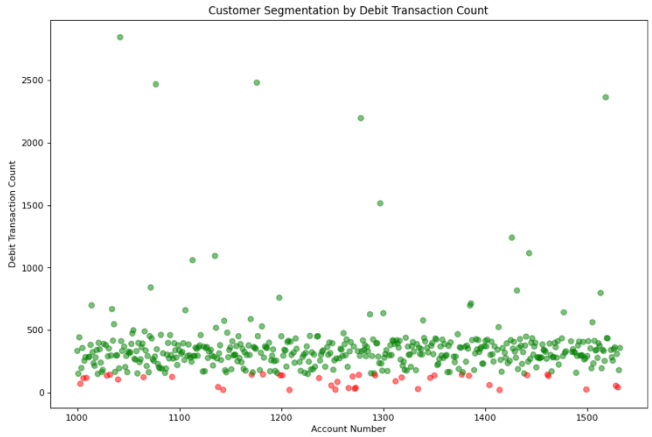

Fig 13. Churn Prediction

Customer Sector-wise Segmentation: The business accounts were categorized on different sectors manually to identify the density of the transaction on different sectors. The scatter plot in fig 13 depicts the sector-based expenditure pattern. Below are the inferences from this experiment,

- Minimal amount is being spent on miscellaneous sector.
- Most of the amount is spent on food and drinks.
- Customers with a negative balance rarely go out for a drink or to a restaurant and rarely spend money on clothes.

Fig 14. Amount spent on different sectors

V. FURTHER WORK AND IMPROVEMENTS

The proposed system can be improved in many areas. Several machine learning algorithms have been used to analyze and gain insights on the data. The performance of these algorithms can be improved further by implementing hyper parameter tuning, cross validation and much more. Also, we can use the data to predict customer expenditure patterns for the upcoming years with the help of 'Time Series Analysis'. The analysis itself can be improved by using a different split and group of the data. For credit card or loan worthiness of a customer only one aspect of the criteria that is the balance of the account has been considered in the experiments. This can be improved if there is more data available such as entire transaction data of the customer and credit score of the customer. The implemented architecture of the system is a basic data processing pipeline which can be improved by using different services and different configurations. Instead of using AWS RDS as a data store, AWS Redshift can be used which is a data warehouse and would be more suitable for this use case. It is easy to get insights into the data that is stored in the Redshift tables with

high availability. Basic configurations are used for the database server in the current architecture which can be improved higher configurations to make the read write operations more efficient and quicker. The EC2 server used to run the talend jobs can also be configured with high compute power which makes the jobs run in a more efficient way. Another suggestion to improve the efficiency of the data engineering system is to break the python code into multiple chunks and move it to AWS Lambda. Lambda is a serverless architecture that can be used to compute small and quick workloads. The ETL architecture built using Talend open studio can be moved to AWS Glue to implement an end-to-end cloud system for the transaction data pre-processing.

VI. CONCLUSION

Throughout the course of this project the main goal has been to derive in-depth analysis of the dataset which will be beneficial for both the bank and the customers. It is observed that most of the expenses occur during the weekends for customer-to-business transactions and the most popular businesses belong to the food and grocery sectors. The different types of analysis which includes statistics, visual and machine learning analysis proved useful in its own way. Anomaly detection can be used to find out fraud transactions and help the banks to proactively improve their risk management capability. This in turn helps the customers to trust their banks more and continue utilising their services for a prolonged time. Customer segmentation and credit card eligibility enables the service providers to get a detail profiling about their clientele by looking into their spending patterns, balance and salary. Churn prediction is used to predict whether a customer is likely to stop using the bank's services, allowing measures to be taken to retain valuable customers. Overall, these methods help banks to provide personalized services, and assess the risks involved in providing credit cards and personal loans and keep their customers happy.

VII. REFERENCES

[1] Bansal, K. and Bohra, A. (n.d.). *K-Mean Clustering Algorithm Implemented To E-Banking*. [online] Available at: https://www.ijert.org/research/k-mean-clustering-algorithm-implemented-to-e-banking-IJERTV2IS4761.pdf.

[2] Sharma, A. and Kumar Panigrahi, P. (2012). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications*, [online] 39(1), pp.37–47. doi:https://doi.org/10.5120/4787-7016.

[3] Angelini, E., di Tollo, G. and Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, [online] 48(4), pp.733–755. doi:https://doi.org/10.1016/j.qref.2007.04.001.

[4] Credit AnAlysis And lending MAnAgeMent. (n.d.). Available at: https://mirabelpublishing.com/wp-content/uploads/2021/07/r8Vr21s6rp66lMTydbQAJZjGBZQ4LlUZ.CALM_booklet.pdf.

[5] Hemachandran, K., George, P. mary, Rodriguez, R.V., Roy, S. and Kulkarni, R. (2021). Performance Analysis of K-Nearest Neighbor Classification Algorithms for Bank Loan Sectors. [online] doi:https://doi.org/10.3233/APC210004.

[6] AWS (2018). Cloud Object Storage | Store & Retrieve Data Anywhere | Amazon Simple Storage Service. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/s3/.

[7] Codecademy. (n.d.). Introduction to Pandas and NumPy. [online] Available at: https://www.codecademy.com/article/introduction-to-numpy-and-pandas.

[8] Talend - A Leader in Data Integration & Data Integrity. (n.d.). Talend Data Integration — Software to Connect, Access, and Transform Data. [online] Available at: https://www.talend.com/uk/products/integrate-data/

[9] AWS (2019). Amazon EC2. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/ec2/.

[10] AWS (2019). Amazon Relational Database Service (RDS) – AWS. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/rds/.

[11] discover.strongdm.com. (n.d.). What Is Anomaly Detection? Methods, Examples, and More | StrongDM. [online] Available at: https://www.strongdm.com/blog/anomaly-detection.

[12] Singh, S. (2020). Anomaly Detection Using Isolation Forest Algorithm. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/anomaly-detection-using-isolation-forest-algorithm-8cf36c38d6f7

[13] Cuemath. (n.d.). Euclidean Distance Formula - Derivation, Examples. [online] Available at: https://www.cuemath.com/euclidean-distance-formula/

[14] Sharma, N. (2021). K-Means Clustering Explained. [online] neptune.ai. Available at: https://neptune.ai/blog/k-means-clustering.

[15] Mavuduru, Amol. "How to Perform Anomaly Detection with the Isolation Forest Algorithm." *Medium*, 8 Apr. 2022, towardsdatascience.com/how-to-perform-anomaly-detection-with-the-isolation-forest-algorithm-e8c8372520bc.

[16] Christopher, A. (2021). K-Nearest Neighbor. [online] Medium. Available at: https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4.

[17] Financial Credibility and Liquidity Scores. [online] Available at: https://docs.planky.com/product-guides/financial-credibility-and-liquidity-score.

[18] Predicting & Preventing Churn: Building a Churn Prediction Model | Mode. [online] Available at: https://mode.com/blog/predicting-and-preventing-churn/#:~:text=Download%20now-.

APPENDIX

GitHub Repository Link:
https://github.com/SuriyaVadivelu/LLOYDS_BANKING

AWS Learner Lab Credentials:
Username - ci22246@bristol.ac.uk
Password - Seetha#181627