Analytical Framework

TRANSFORMING TECHNICAL DATA INTO BUSINESS VALUE

WHY CHURN IN BANKING

- Customer churn is a critical business problem in banking.
- Retaining customers is cheaper and more profitable than acquiring new ones.
- Machine Learning enables accurate prediction of churn by analysing large volumes of historical data.
- It supports proactive retention strategies and improves decision-making.

GETTY

🎯 RESEARCH QUESTIONS

1. Which specific customers and strategic segments are most likely to churn?

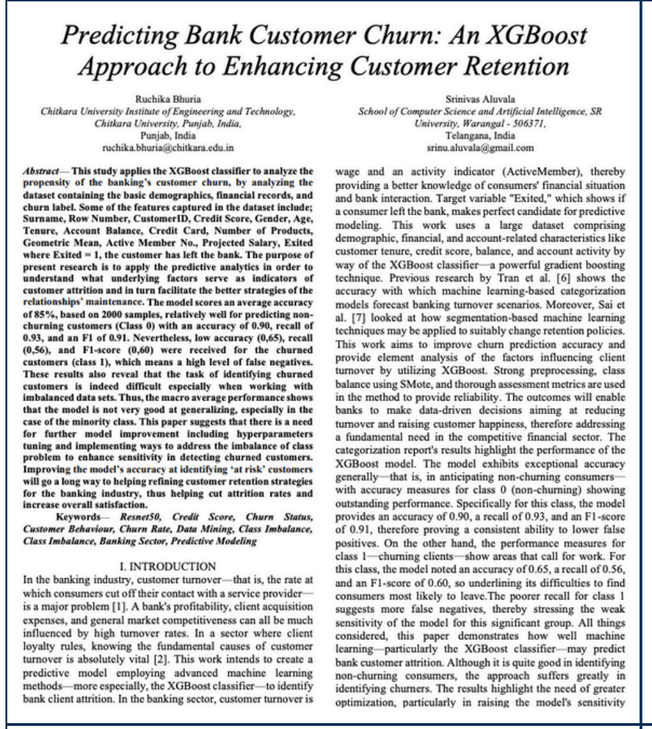
2. What are the key factors that contribute to their churn prediction in each segment?

LITERATURE REVIEW

Different studies support the use of machine learning techniques in the banking sector, especially for customer churn prediction, fraud detection, credit scoring and customer segmentation.

While existing research covers these areas independently, there is limited evidence of end to end projects that integrate segmentation, churn prediction, explainability and business intelligence in a unified workflow.

However, all point in the same direction which is use the data and analytics to improve business performance in terms of security and/or commercial performance.



Predicting Bank Customer Churn: An XGBoost Approach to Enhancing Customer Retention

Ruchika Bhuria
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India,
Punjab, India
ruchika.bhuria@chitkara.edu.in

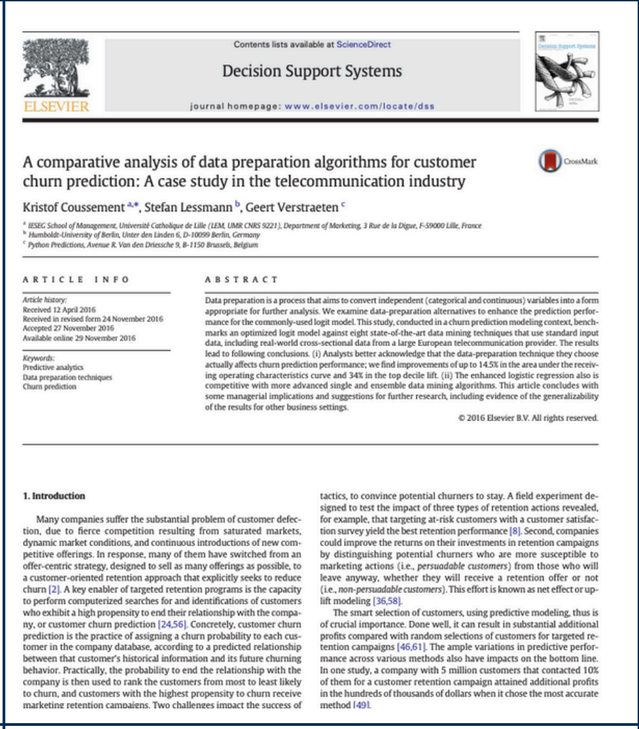
Srinivas Aluvala
School of Computer Science and Artificial Intelligence, SR University, Warangal - 506371,
Telangana, India
srinu.aluvala@gmail.com

***Abstract*— This study applies the XGBoost classifier to analyze the propensity of the banking's customer churn, by analyzing the dataset containing the basic demographics, financial records, and churn label. Some of the features captured in the dataset include; Surname, Row Number, CustomerID, Credit Score, Gender, Age, Tenure, Account Balance, Credit Card, Number of Products, Geometric Mean, Active Member No., Projected Salary, Exited where Exited = 1, the customer has left the bank. The purpose of present research is to apply the predictive analytics in order to understand what underlying factors serve as indicators of customer attrition and in turn facilitate the better strategies of the relationships' maintenance. The model scores an average accuracy of 85%, based on 2000 samples, relatively well for predicting non-churning customers (Class 0) with an accuracy of 0.90, recall of 0.93, and an F1 of 0.91. Nevertheless, low accuracy (0,65), recall (0,56), and F1-score (0,60) were received for the churned customers (class 1), which means a high level of false negatives. These results also reveal that the task of identifying churned customers is indeed difficult especially when working with imbalanced data sets. Thus, the macro average performance shows that the model is not very good at generalizing, especially in the case of the minority class. This paper suggests that there is a need for further model improvement including hyperparameters tuning and implementing ways to address the imbalance of class problem to enhance sensitivity in detecting churned customers. Improving the model's accuracy at identifying 'at risk' customers will go a long way to helping refining customer retention strategies for the banking industry, thus helping cut attrition rates and increase overall satisfaction.**

Keywords*— *Resnet50, Credit Score, Churn Status, Customer Behaviour, Churn Rate, Data Mining, Class Imbalance, Class Imbalance, Banking Sector, Predictive Modeling

I. INTRODUCTION

In the banking industry, customer turnover—that is, the rate at which consumers cut off their contact with a service provider—is a major problem [1]. A bank's profitability, client acquisition expenses, and general market competitiveness can all be much influenced by high turnover rates. In a sector where client loyalty rules, knowing the fundamental causes of customer turnover is absolutely vital [2]. This work intends to create a predictive model employing advanced machine learning methods—more especially, the XGBoost classifier—to identify bank client attrition. In the banking sector, customer turnover is wage and an activity indicator (ActiveMember), thereby providing a better knowledge of consumers' financial situation and bank interaction. Target variable "Exited," which shows if a consumer left the bank, makes perfect candidate for predictive modeling. This work uses a large dataset comprising demographic, financial, and account-related characteristics like customer tenure, credit score, balance, and account activity by way of the XGBoost classifier—a powerful gradient boosting technique. Previous research by Tran et al. [6] shows the accuracy with which machine learning-based categorization models forecast banking turnover scenarios. Moreover, Sai et al. [7] looked at how segmentation-based machine learning techniques may be applied to suitably change retention policies. This work aims to improve churn prediction accuracy and provide element analysis of the factors influencing client turnover by utilizing XGBoost. Strong preprocessing, class balance using SMote, and thorough assessment metrics are used in the method to provide reliability. The outcomes will enable banks to make data-driven decisions aiming at reducing turnover and raising customer happiness, therefore addressing a fundamental need in the competitive financial sector. The categorization report's results highlight the performance of the XGBoost model. The model exhibits exceptional accuracy generally—that is, in anticipating non-churning consumers—with accuracy measures for class 0 (non-churning) showing outstanding performance. Specifically for this class, the model provides an accuracy of 0.90, a recall of 0.93, and an F1-score of 0.91, therefore proving a consistent ability to lower false positives. On the other hand, the performance measures for class 1—churning clients—show areas that call for work. For this class, the model noted an accuracy of 0.65, a recall of 0.56, and an F1-score of 0.60, so underlining its difficulties to find consumers most likely to leave.The poorer recall for class 1 suggests more false negatives, thereby stressing the weak sensitivity of the model for this significant group. All things considered, this paper demonstrates how well machine learning—particularly the XGBoost classifier—may predict bank customer attrition. Although it is quite good in identifying non-churning consumers, the approach suffers greatly in identifying churners. The results highlight the need of greater optimization, particularly in raising the model's sensitivity



Contents lists available at ScienceDirect
Decision Support Systems
journal homepage: www.elsevier.com/locate/dss

A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry

Kristof Coussement [a,*], Stefan Lessmann [b], Geert Verstraeten [c]
[a] IESEG School of Management, Université Catholique de Lille (LEM, UMR CNRS 9221), Department of Marketing, 3 Rue de la Digue, F-59000 Lille, France
[b] Humboldt-University of Berlin, Unter den Linden 6, D-10099 Berlin, Germany
[c] Python Predictions, Avenue R. Van den Driessche 9, B-1150 Brussels, Belgium

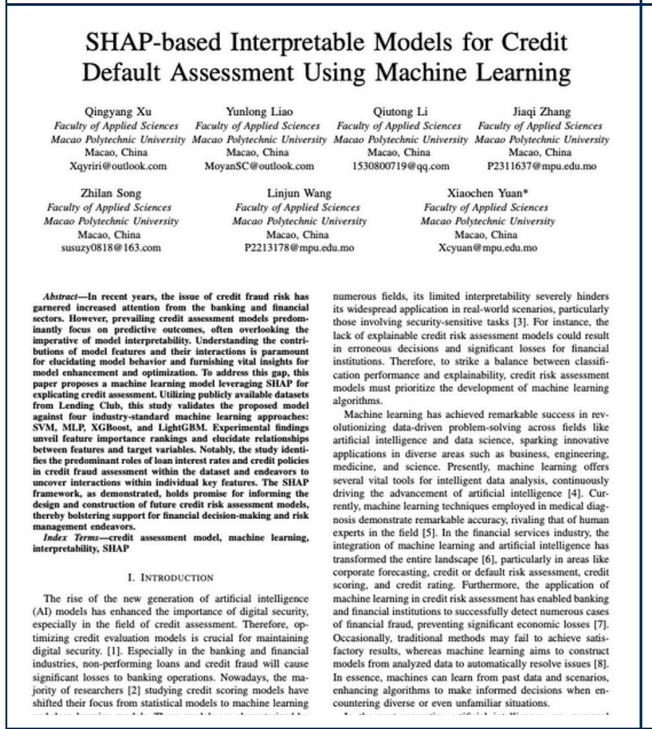
ARTICLE INFO
Article history:
Received 12 April 2016
Received in revised form 24 November 2016
Accepted 27 November 2016
Available online 29 November 2016
Keywords:
Predictive analytics
Data preparation techniques
Churn prediction

ABSTRACT
Data preparation is a process that aims to convert independent (categorical and continuous) variables into a form appropriate for further analysis. We examine data-preparation alternatives to enhance the prediction performance for the commonly-used logit model. This study, conducted in a churn prediction modeling context, benchmarks an optimized logit model against eight state-of-the-art data mining techniques that use standard input data, including real-world cross-sectional data from a large European telecommunication provider. The results lead to following conclusions. (i) Analysts better acknowledge that the data-preparation technique they choose actually affects churn prediction performance; we find improvements of up to 14.5% in the area under the receiving operating characteristics curve and 34% in the top decile lift. (ii) The enhanced logistic regression also is competitive with more advanced single and ensemble data mining algorithms. This article concludes with some managerial implications and suggestions for further research, including evidence of the generalizability of the results for other business settings.

1. Introduction

Many companies suffer the substantial problem of customer defection, due to fierce competition resulting from saturated markets, dynamic market conditions, and continuous introductions of new competitive offerings. In response, many of them have switched from an offer-centric strategy, designed to sell as many offerings as possible, to a customer-oriented retention approach that explicitly seeks to reduce churn [2]. A key enabler of targeted retention programs is the capacity to perform computerized searches for and identifications of customers who exhibit a high propensity to end their relationship with the company, or customer churn prediction [24,56]. Concretely, customer churn prediction is the practice of assigning a churn probability to each customer in the company database, according to a predicted relationship between that customer's historical information and its future churning behavior. Practically, the probability to end the relationship with the company is then used to rank the customers from most to least likely to churn, and customers with the highest propensity to churn receive marketing retention campaigns. Two challenges impact the success of tactics, to convince potential churners to stay. A field experiment designed to test the impact of three types of retention actions revealed, for example, that targeting at-risk customers with a customer satisfaction survey yield the best retention performance [8]. Second, companies could improve the returns on their investments in retention campaigns by distinguishing potential churners who are more susceptible to marketing actions (i.e., *persuadable customers*) from those who will leave anyway, whether they will receive a retention offer or not (i.e., *non-persuadable customers*). This effort is known as net effect or uplift modeling [36,58].

The smart selection of customers, using predictive modeling, thus is of crucial importance. Done well, it can result in substantial additional profits compared with random selections of customers for targeted retention campaigns [46,61]. The ample variations in predictive performance across various methods also have impacts on the bottom line. In one study, a company with 5 million customers that contacted 10% of them for a customer retention campaign attained additional profits in the hundreds of thousands of dollars when it chose the most accurate method [49].



SHAP-based Interpretable Models for Credit Default Assessment Using Machine Learning

Qingyang Xu
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
Xqyriri@outlook.com

Yunlong Liao
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
MoyanSC@outlook.com

Qiutong Li
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
1530800719@qq.com

Jiaqi Zhang
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
P2311637@mpu.edu.mo

Zhilan Song
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
susuzy0818@163.com

Linjun Wang
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
P2213178@mpu.edu.mo

Xiaochen Yuan*
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
Xcyuan@mpu.edu.mo

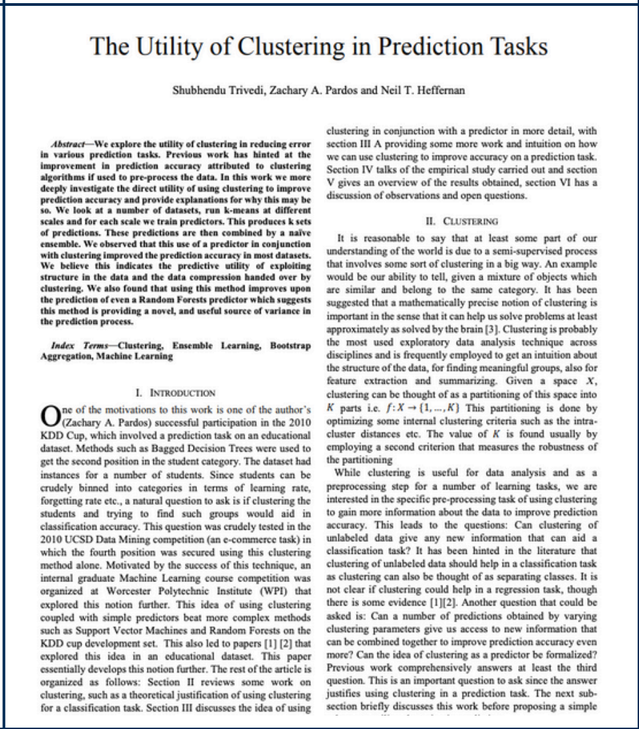
***Abstract*—In recent years, the issue of credit fraud risk has garnered increased attention from the banking and financial sectors. However, prevailing credit assessment models predominantly focus on predictive outcomes, often overlooking the imperative of model interpretability. Understanding the contributions of model features and their interactions is paramount for elucidating model behavior and furnishing vital insights for model enhancement and optimization. To address this gap, this paper proposes a machine learning model leveraging SHAP for explicating credit assessment. Utilizing publicly available datasets from Lending Club, this study validates the proposed model against four industry-standard machine learning approaches: SVM, MLP, XGBoost, and LightGBM. Experimental findings unveil feature importance rankings and elucidate relationships between features and target variables. Notably, the study identifies the predominant roles of loan interest rates and credit policies in credit fraud assessment within the dataset and endeavors to uncover interactions within individual key features. The SHAP framework, as demonstrated, holds promise for informing the design and construction of future credit risk assessment models, thereby bolstering support for financial decision-making and risk management endeavors.**

***Index Terms*—credit assessment model, machine learning, interpretability, SHAP**

I. Introduction

The rise of the new generation of artificial intelligence (AI) models has enhanced the importance of digital security, especially in the field of credit assessment. Therefore, optimizing credit evaluation models is crucial for maintaining digital security. [1]. Especially in the banking and financial industries, non-performing loans and credit fraud will cause significant losses to banking operations. Nowadays, the majority of researchers [2] studying credit scoring models have shifted their focus from statistical models to machine learning numerous fields, its limited interpretability severely hinders its widespread application in real-world scenarios, particularly those involving security-sensitive tasks [3]. For instance, the lack of explainable credit risk assessment models could result in erroneous decisions and significant losses for financial institutions. Therefore, to strike a balance between classification performance and explainability, credit risk assessment models must prioritize the development of machine learning algorithms.

Machine learning has achieved remarkable success in revolutionizing data-driven problem-solving across fields like artificial intelligence and data science, sparking innovative applications in diverse areas such as business, engineering, medicine, and science. Presently, machine learning offers several vital tools for intelligent data analysis, continuously driving the advancement of artificial intelligence [4]. Currently, machine learning techniques employed in medical diagnosis demonstrate remarkable accuracy, rivaling that of human experts in the field [5]. In the financial services industry, the integration of machine learning and artificial intelligence has transformed the entire landscape [6], particularly in areas like corporate forecasting, credit or default risk assessment, credit scoring, and credit rating. Furthermore, the application of machine learning in credit risk assessment has enabled banking and financial institutions to successfully detect numerous cases of financial fraud, preventing significant economic losses [7]. Occasionally, traditional methods may fail to achieve satisfactory results, whereas machine learning aims to construct models from analyzed data to automatically resolve issues [8]. In essence, machines can learn from past data and scenarios, enhancing algorithms to make informed decisions when encountering diverse or even unfamiliar situations.



The Utility of Clustering in Prediction Tasks

Shubhendu Trivedi, Zachary A. Pardos and Neil T. Heffernan

***Abstract*—We explore the utility of clustering in reducing error in various prediction tasks. Previous work has hinted at the improvement in prediction accuracy attributed to clustering algorithms if used to pre-process the data. In this work we more deeply investigate the direct utility of using clustering to improve prediction accuracy and provide explanations for why this may be so. We look at a number of datasets, run k-means at different scales and for each scale we train predictors. This produces k sets of predictions. These predictions are then combined by a naive ensemble. We observed that this use of a predictor in conjunction with clustering improved the prediction accuracy in most datasets. We believe this indicates the predictive utility of exploiting structure in the data and the data compression handed over by clustering. We also found that using this method improves upon the prediction of even a Random Forests predictor which suggests this method is providing a novel, and useful source of variance in the prediction process.**

***Index Terms*—Clustering, Ensemble Learning, Bootstrap Aggregation, Machine Learning**

I. Introduction

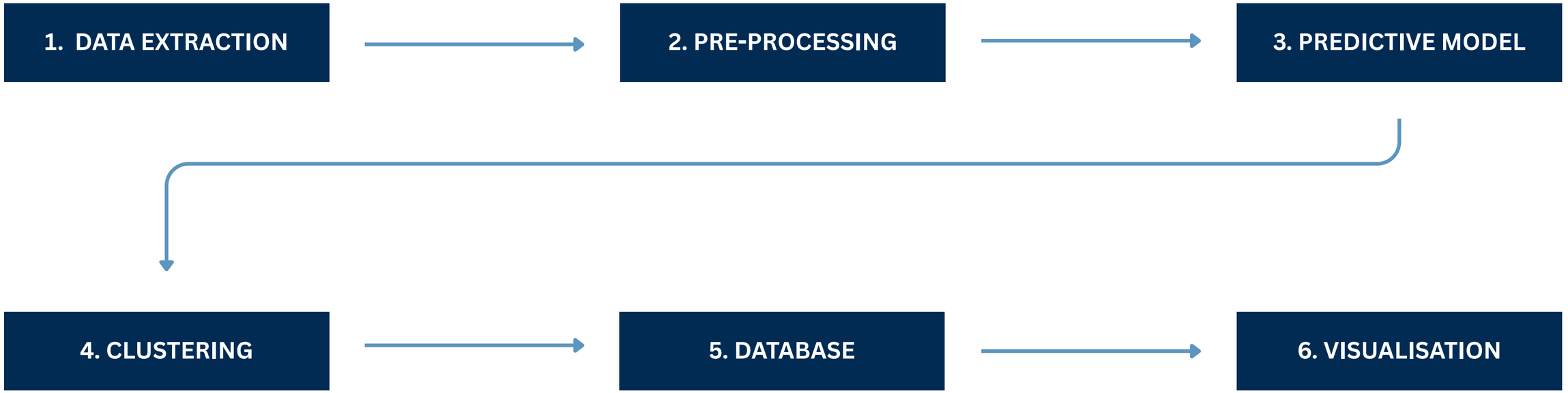
One of the motivations to this work is one of the author's (Zachary A. Pardos) successful participation in the 2010 KDD Cup, which involved a prediction task on an educational dataset. Methods such as Bagged Decision Trees were used to get the second position in the student category. The dataset had instances for a number of students. Since students can be crudely binned into categories in terms of learning rate, forgetting rate etc., a natural question to ask is if clustering the students and trying to find such groups would aid in classification accuracy. This question was crudely tested in the 2010 UCSD Data Mining competition (an e-commerce task) in which the fourth position was secured using this clustering method alone. Motivated by the success of this technique, an internal graduate Machine Learning course competition was organized at Worcester Polytechnic Institute (WPI) that explored this notion further. This idea of using clustering coupled with simple predictors beat more complex methods such as Support Vector Machines and Random Forests on the KDD cup development set. This also led to papers [1] [2] that explored this idea in an educational dataset. This paper essentially develops this notion further. The rest of the article is organized as follows: Section II reviews some work on clustering, such as a theoretical justification of using clustering for a classification task. Section III discusses the idea of using clustering in conjunction with a predictor in more detail, with section III A providing some more work and intuition on how we can use clustering to improve accuracy on a prediction task. Section IV talks of the empirical study carried out and section V gives an overview of the results obtained, section VI has a discussion of observations and open questions.

II. Clustering

It is reasonable to say that at least some part of our understanding of the world is due to a semi-supervised process that involves some sort of clustering in a big way. An example would be our ability to tell, given a mixture of objects which are similar and belong to the same category. It has been suggested that a mathematically precise notion of clustering is important in the sense that it can help us solve problems at least approximately as solved by the brain [3]. Clustering is probably the most used exploratory data analysis technique across disciplines and is frequently employed to get an intuition about the structure of the data, for finding meaningful groups, also for feature extraction and summarizing. Given a space X, clustering can be thought of as a partitioning of this space into K parts i.e. $f: X \rightarrow \{1, ..., K\}$ This partitioning is done by optimizing some internal clustering criteria such as the intra-cluster distances etc. The value of K is found usually by employing a second criterion that measures the robustness of the partitioning

While clustering is useful for data analysis and as a preprocessing step for a number of learning tasks, we are interested in the specific pre-processing task of using clustering to gain more information about the data to improve prediction accuracy. This leads to the questions: Can clustering of unlabeled data give any new information that can aid a classification task? It has been hinted in the literature that clustering of unlabeled data should help in a classification task as clustering can also be thought of as separating classes. It is not clear if clustering could help in a regression task, though there is some evidence [1][2]. Another question that could be asked is: Can a number of predictions obtained by varying clustering parameters give us access to new information that can be combined together to improve prediction accuracy even more? Can the idea of clustering as a predictor be formalized? Previous work comprehensively answers at least the third question. This is an important question to ask since the answer justifies using clustering in a prediction task. The next subsection briefly discusses this work before proposing a simple

METHODOLOGY OVERVIEW



DATASET

- Source: Kaggle
- 10,128 records and 23 variables
- Anonymous bank customer data
- Demographic, financial, and behavioural variables
- Purpose: to study customer churn

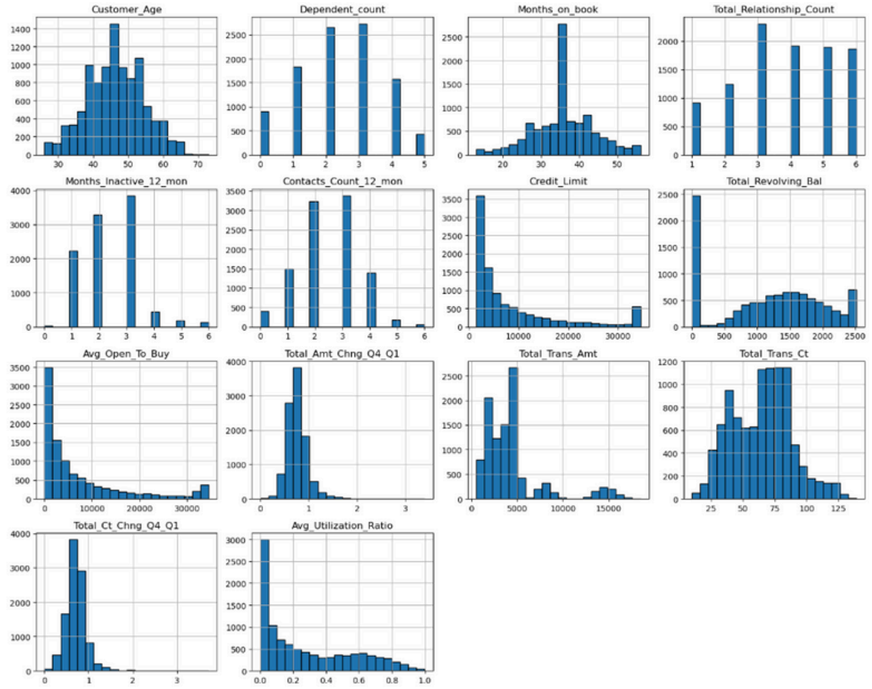
kaggle

# CLIENTNUM Client number. Unique identifier for the customer holding the account	Attrition_Flag Internal event (customer activity) variable - if the account is closed then 1 else 0	# Customer_Age Demographic variable - Customer's Age in Years	Gender Demographic variable - M=Male, F=Female	# Dependent_count Demographic variable - Number of dependents
708m 828m	Existing Customer 84% Attrited Customer 16%	26 73	F 53% M 47%	0 5
768805383	Existing Customer	45	M	3
818770008	Existing Customer	49	F	5
713982108	Existing Customer	51	M	3
769911858	Existing Customer	40	F	4
709106358	Existing Customer	40	M	3
713061558	Existing Customer	44	M	2
810347208	Existing Customer	51	M	4
818906208	Existing Customer	32	M	0
710930508	Existing Customer	37	M	3
719661558	Existing Customer	48	M	2
708790833	Existing Customer	42	M	5
710821833	Existing Customer	65	M	1
710599683	Existing Customer	56	M	1
816082233	Existing Customer	35	M	3
712396908	Existing Customer	57	F	2

Head of dataset

EDA & DATA VALIDATION

- Verification of null and duplicate values (none found)
- Exploratory analysis: distributions, outliers, correlation



Numerical Distribution Charts

Checking missing values

```
# Check for missing values
df.isnull().sum()
```

```
CLIENTNUM                 0
Attrition_Flag            0
Customer_Age              0
Gender                    0
Dependent_count           0
Education_Level           0
Marital_Status            0
Income_Category           0
Card_Category             0
Months_on_book            0
Total_Relationship_Count  0
Months_Inactive_12_mon    0
Contacts_Count_12_mon     0
Credit_Limit              0
Total_Revolving_Bal       0
Avg_Open_To_Buy           0
Total_Amt_Chng_Q4_Q1      0
Total_Trans_Amt           0
Total_Trans_Ct            0
Total_Ct_Chng_Q4_Q1       0
Avg_Utilization_Ratio     0
dtype: int64
```

Checking duplicated values

```
# Check duplicated rows
df.duplicated().sum()
```

```
0
```

```
# Check duplicated IDs
df['CLIENTNUM'].duplicated().any()
```

```
False
```

BOXPLOTS - OUTLIERS

Outliers were found in variables that could affect the model:

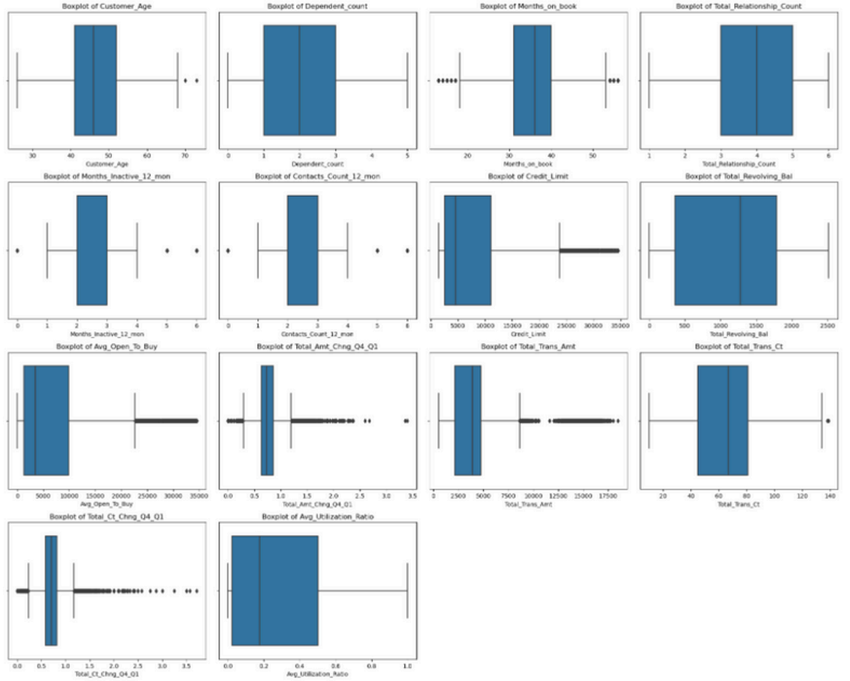
- credit_limit
- avg_open_to_buy
- total_amt_change_q4_q1
- total_trans_amt
- total_ct_chang_q4_q1

Winsorising was applied to limit these extreme values without deleting them.

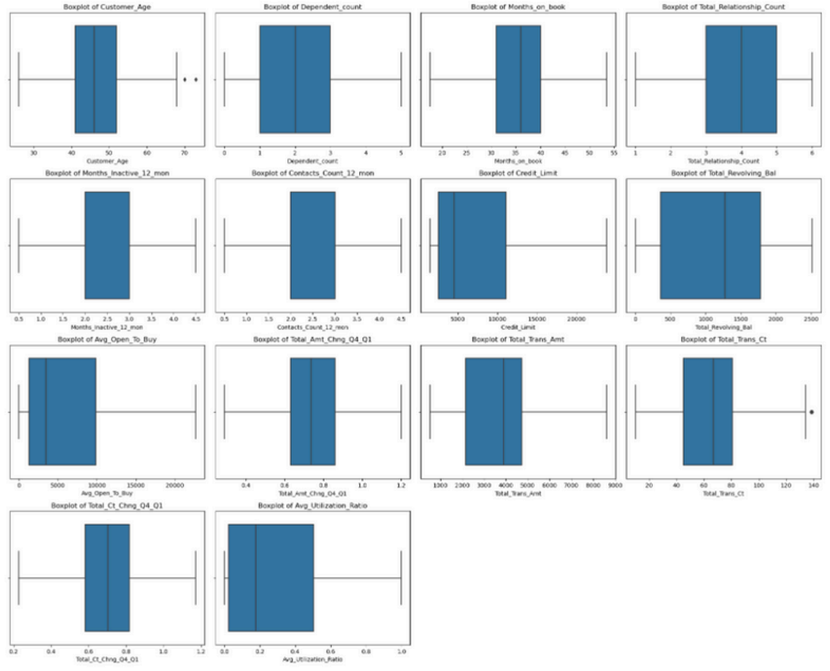
- Lower limit = Q1 – 1.5 × IQR
- Upper limit = Q3 + 1.5 × IQR

The result was a reduction in all outliers found.

Before →



After →



CORRELATION

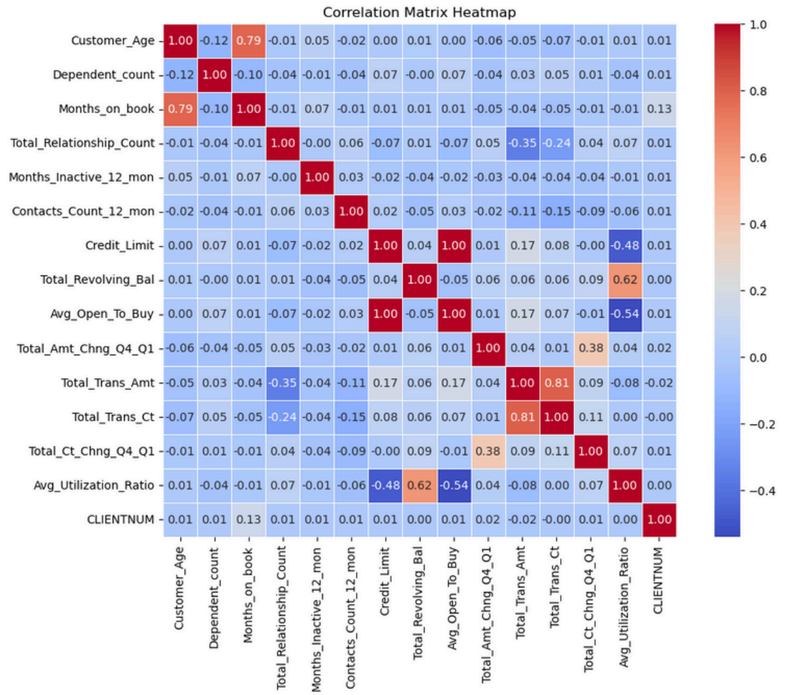
A correlation matrix was performed to detect linear relationships between numerical variables.

Highly correlated pairs were identified:

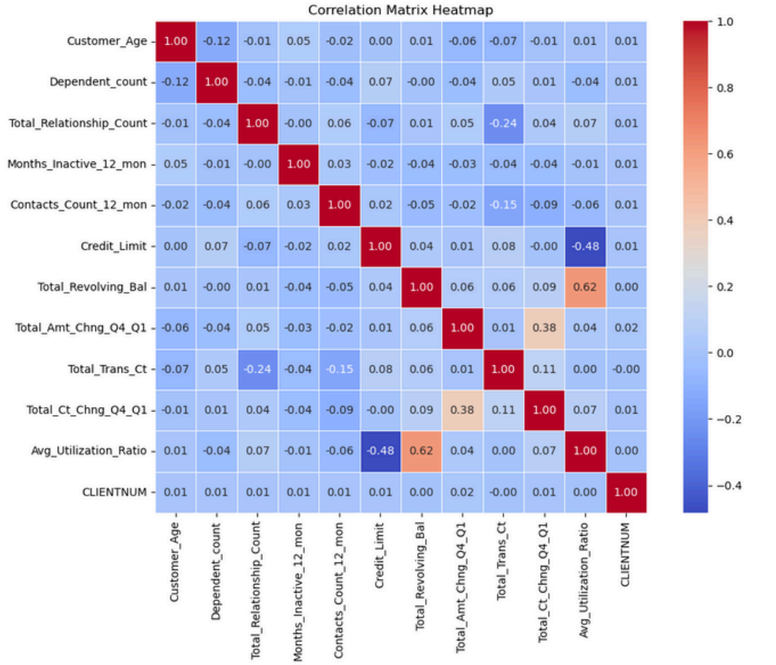
- Avg_Open_To_Buy and Credit_Limit (1.00)
- Total_Trans_Amt and Total_Trans_Ct (0.81)
- Months_on_book and Customer_Age (0.79)

Redundant variables were dropped to avoid multicollinearity

Before →



After →



ENCODING AND SCALING

- One-hot encoding was applied to categorical variables to convert them into binary variables (excluding CLIENTNUM).
- StandardScaler was applied to numerical variables to normalise values
- The data was stored in independent datasets to perform the models.

```
df_final.head()
```

Revolving_Bal	Total_Amt_Chng_Q4_Q1	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	...	Income_Category_ $40K-60K$	Income_Category_ $60K-80K$	Income_Category_ $80K-120K$	Income_Category_Less than $40K
-0.473422	2.623494	-0.973895	3.834003	...	0.0	1.0	0.0	0.0
-0.366667	3.563293	-1.357340	12.608573	...	0.0	0.0	0.0	1.0
-1.426858	8.367214	-1.911206	6.807864	...	0.0	0.0	1.0	0.0
1.661686	2.942843	-1.911206	6.807864	...	0.0	0.0	0.0	1.0
-1.426858	6.455682	-1.570365	7.509325	...	0.0	1.0	0.0	0.0

Example of dataset for scoring y shap

CHURN SCORE

The dataset was divided into 80% for training and 20% for testing.

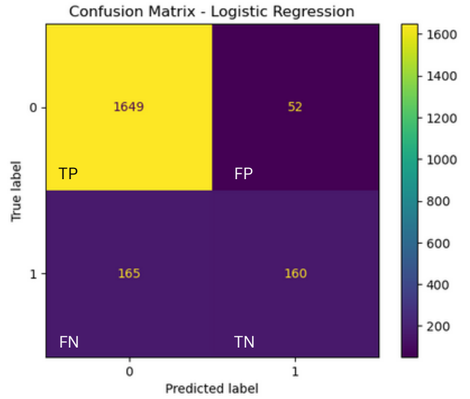
Three classification models were used:

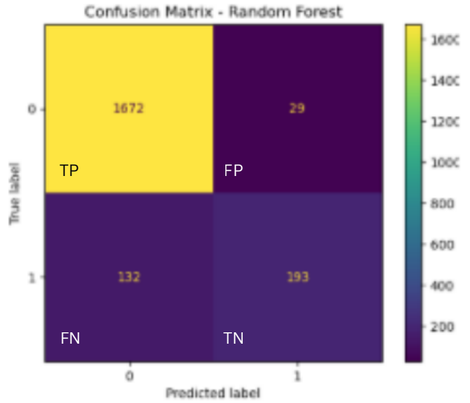
- Logistic Regression
- Random Forest
- XGBoost

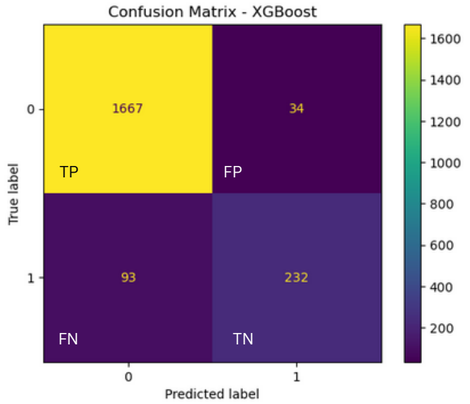
The metrics **Accuracy, Precision, Recall, and F1-Score** were used to evaluate the performance of the models.

Based on the churn identification results, it was observed that XGBoost is the best model for identifying churn customers.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.89	0.75	0.49	0.60
Random Forest	0.93	0.87	0.64	0.74
XGBoost	0.94	0.87	0.71	0.79








```
results_existing_customers.head()
```

	CLIENTNUM	Churn_Score
0	768805383	0.000025
1	818770008	0.000027
2	713982108	0.000348
3	769911858	0.000612
4	709106358	0.000104

XGBoost output for active customers

EXPLANATORY SHAP

An global explanatory SHAP model was conducted to understand which variables contribute most to the prediction of churn.

The most influential variables are to predict churn:

- Total_Trans_Ct (count of total transactions)
- Total_Amt_Chng_Q4_Q1 (Changes in quarterly activity)
- Contacts_Count_12_mon (contacts made in the last 12 months)

Variables contributing to retention:

- Total_Amt_Chng_Q4_Q1 (Changes in quarterly activity)
- Total_Revolving_Bal (revolving balance)
- Total_Relationship_Count (number of relations with the bank)

CLUSTERING

PCA was applied with 2 PCs on the dataset to achieve clear visualisation, but the K-Mean model was trained on the dataset prepared for clusters.

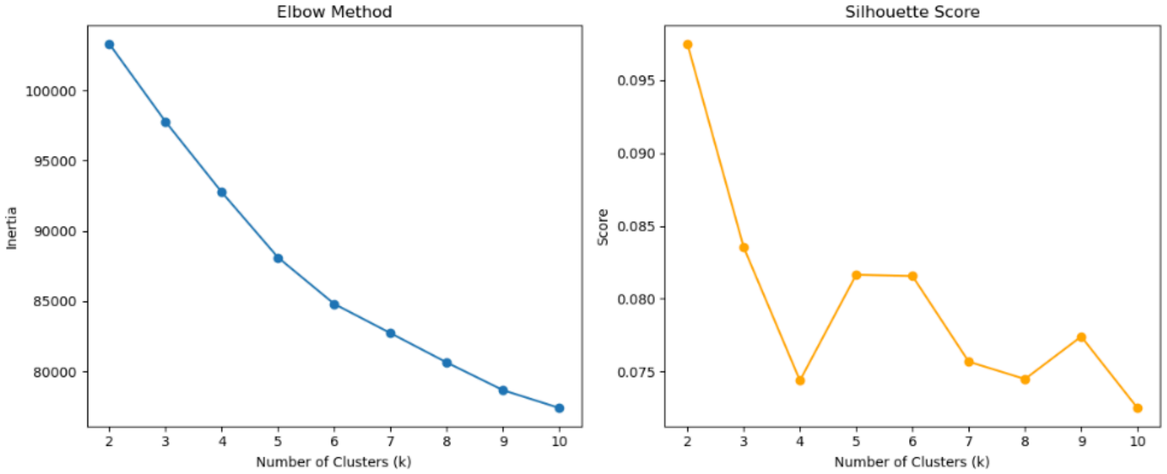
The number of clusters was evaluated through Silhouette Score, where although it shows that the optimal number is 2, it was decided to choose 3 clusters in order to have more diversity of clusters.

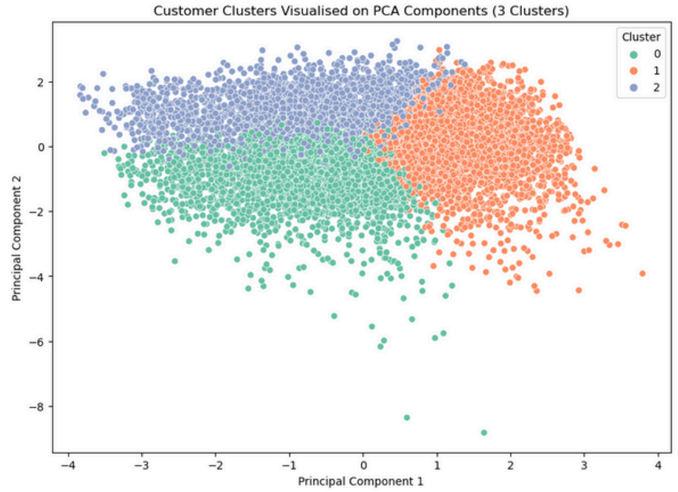
Three different clusters were identified:

- **(Cluster 0)** Medium Income and Low Usage Customers
- **(Cluster 1)** Low Income and High Usage Customers
- **(Cluster 2)** Medium-High Income and Active Usage

The most relevant variables for the segmentation were:

- Number of products contracted
- Contacts with the bank
- Use of credit
- Volume and frequency of transactions





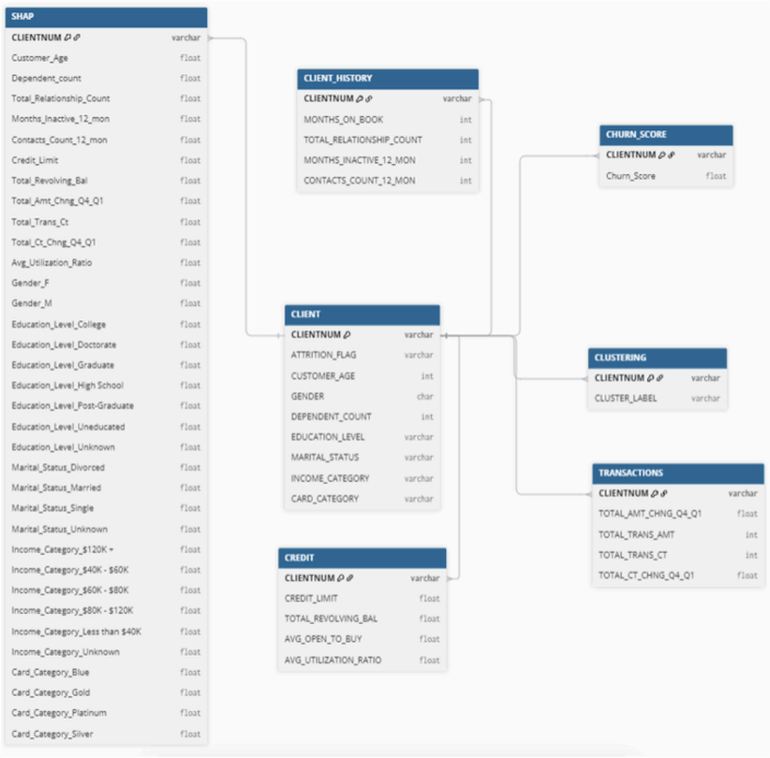
Income_Category	$120K +	$40K – 60K	$60K – 80K	$80K – 120K	Less than $40K	Unknown
Cluster						
0	0.11	0.15	0.19	0.23	0.19	0.12
1	0.02	0.21	0.08	0.06	0.54	0.09
2	0.10	0.18	0.16	0.18	0.26	0.12

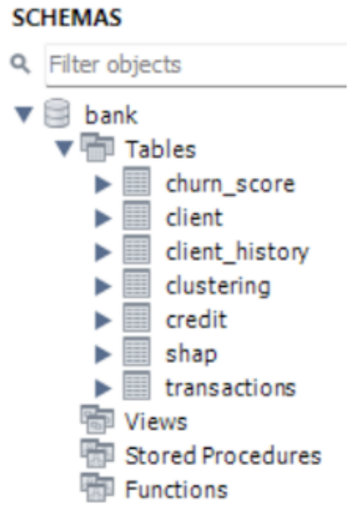
DATABASE

A relational database was built to consolidate the processed information and connect it with visualisation tools.

The structure includes 7 tables, all related by the primary key CLIENTNUM.

- Each table focuses on different information
  - CLIENT: demographic information
  - TRANSACTIONS: financial behaviour
  - CLIENT_HISTORY: contact history
  - CLUSTERING: cluster labels
  - CHURN_SCORE: churn scores
  - SHAP: Interpretability of the prediction by variable





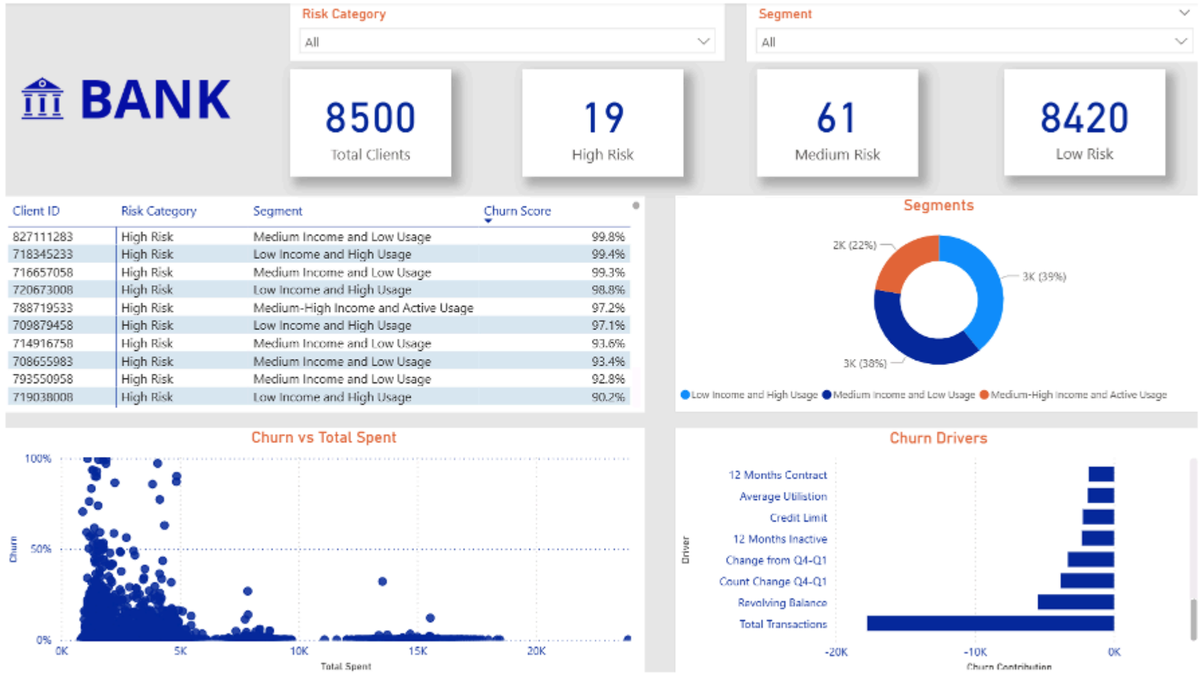
POWER BI DASHBOARD

An interactive dashboard was created in Power BI and connected to the database.

A new measure was developed to classify customers into three risk segments based on their churn probability scores.

- Low Risk: 0% to30%
- Medium Risk: 31% to70%
- High Risk: 71% to 100%

The information as a whole was explored and the research questions were answered.



RESEARCH QUESTION 1

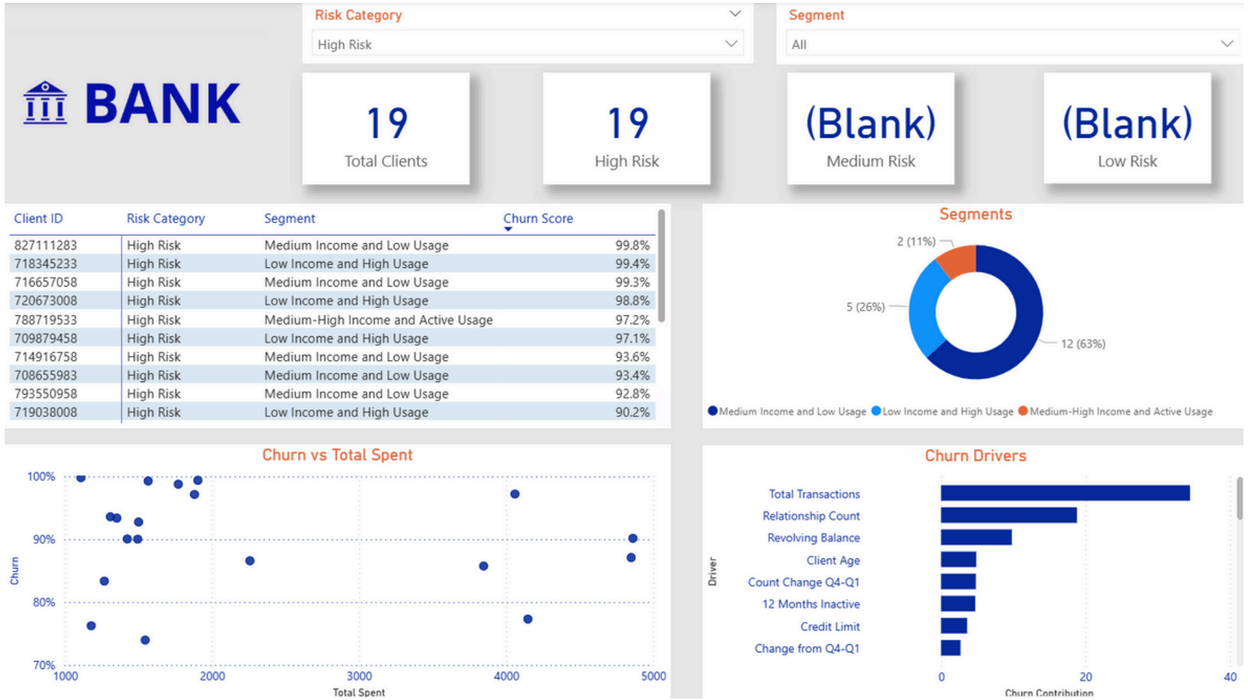
1. WHICH SPECIFIC CUSTOMERS AND STRATEGIC SEGMENTS ARE MOST LIKELY TO CHURN?

Higher risk segments:

- [Cluster 0] Medium Income and Low Usage
- [Cluster 1] Low Income and High Usage

Together they represent more than 85% of medium and high churn risk customers.

Although [Cluster 3] Medium High Income and Active Usage has less volume of clients at high risk it is key because of its high spending.



RESEARCH QUESTION 2

2. WHAT ARE THE KEY FACTORS THAT CONTRIBUTE TO THEIR CHURN PREDICTION IN EACH SEGMENT?

[Cluster 1] Low Income and High Usage

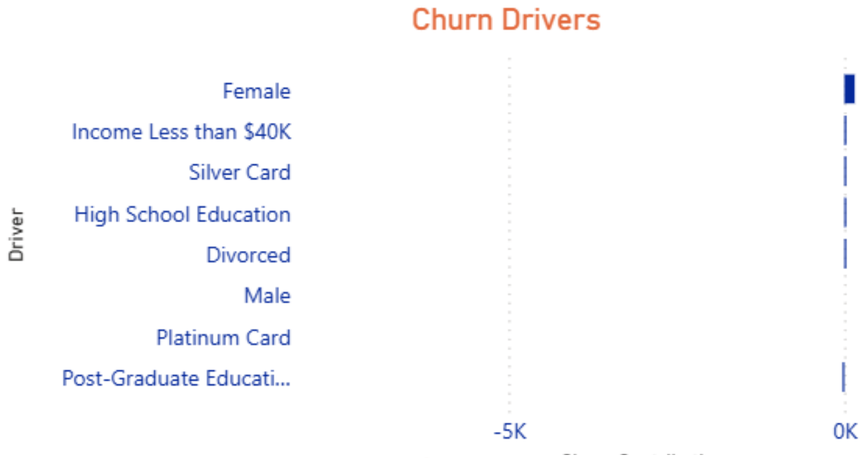
- Factors that increase churn: Female, income <40K, Silver card, high school education.

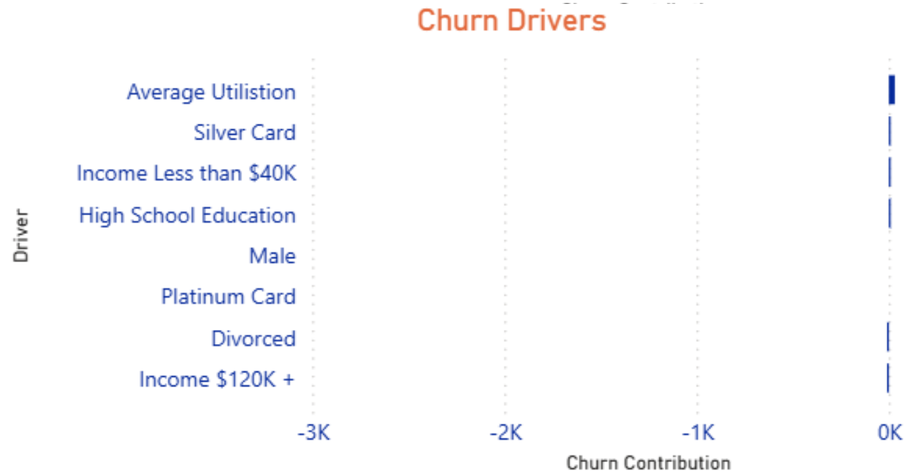
[Cluster 0] Medium Income and Low Usage

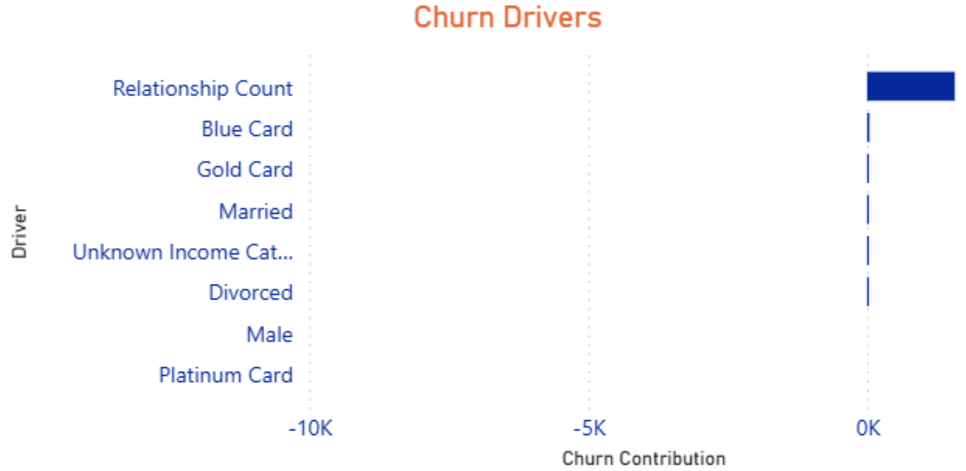
- Factors that increase churn: Average credit utilisation, Silver card, Income less than 40K, high school education.

[Cluster 3] Medium High Income and Active Usage:

- Factors that increase churn: Relationship Count, Blue Card, Gold Card and Married.







FUTURE WORK

- Merge the dataset with other sources, such as CRM.
- Improve dashboard design and visualisations.
- Improving prediction models by adjusting hyperparameters.

THANK YOU