FINAL-TERM REPORT

PREDICT CUSTOMER CHURN WITH CLASSIFICATION ALGORITHM THROUGH DATA MINING

Supervisor: Thon-Da Nguyen Ph.D

Group: 1

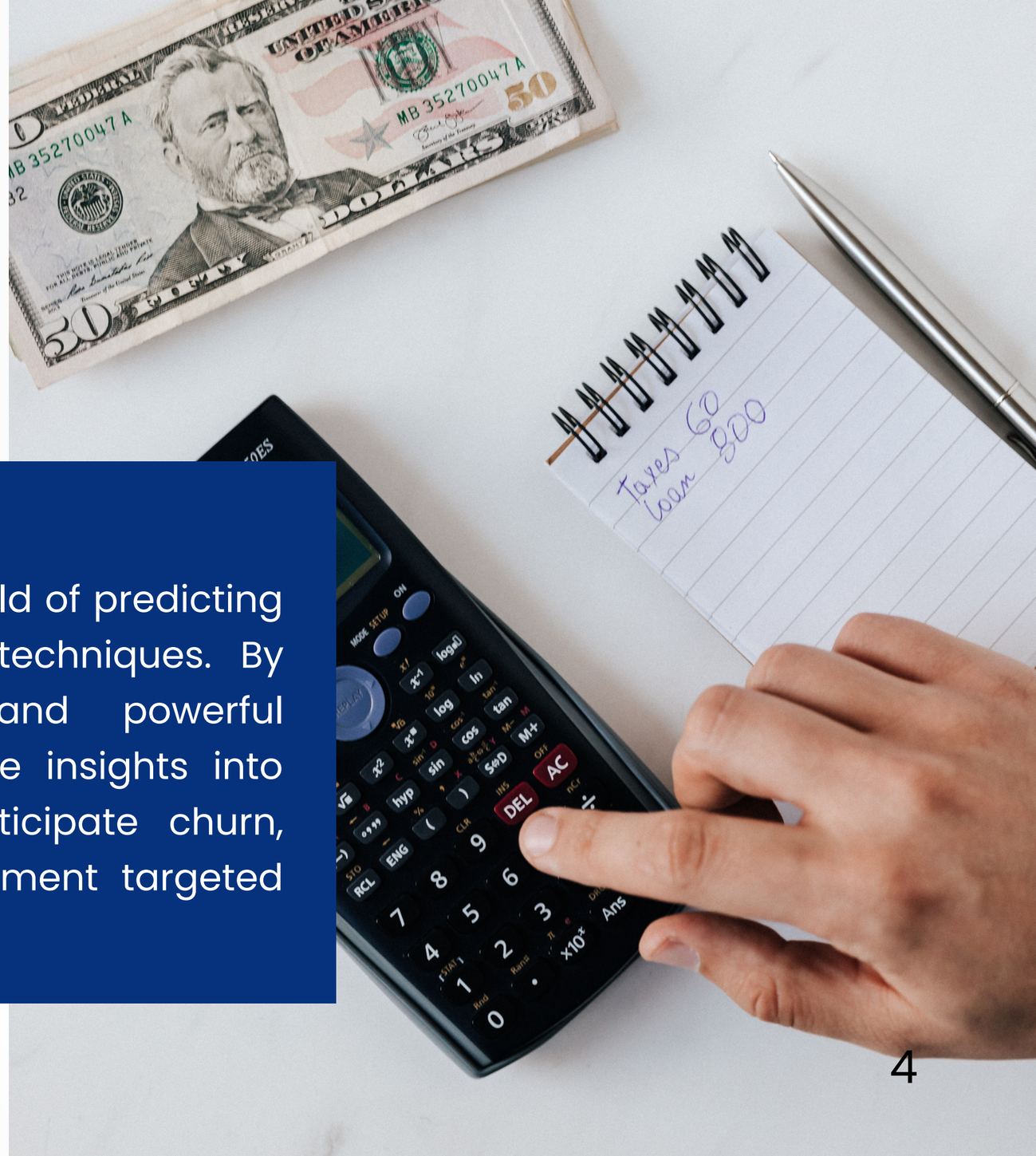
Hoang Thi Thanh Phuong - K204061445
Nguyen Tran Thuy Quynh - K204060307
Nguyen Thi Huyen Thuong - K204061450
Nguyen Hoang Tinh - K204061451

TABLE OF CONTENT

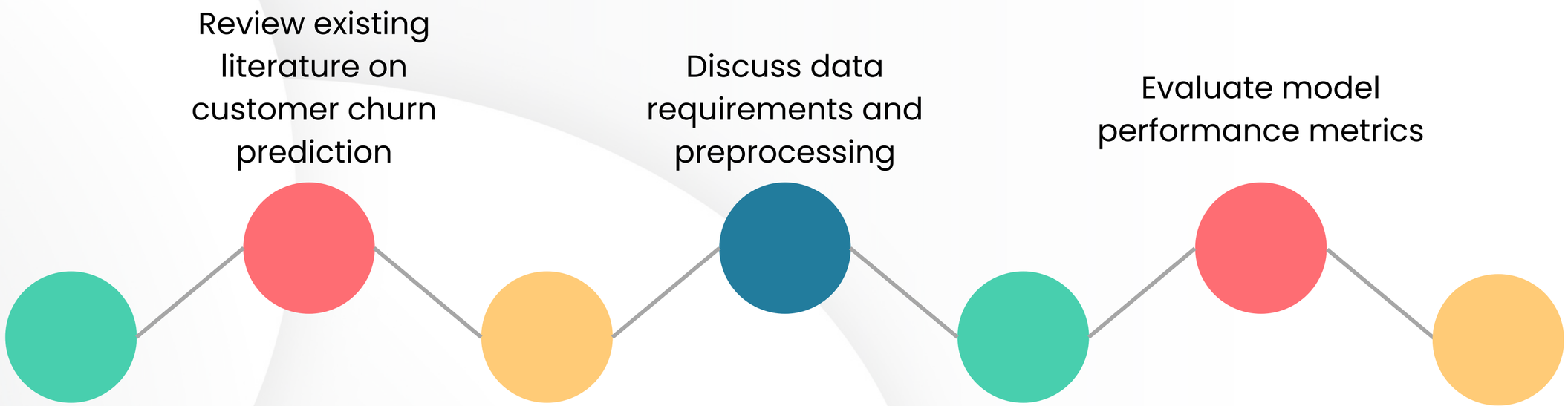
1.
INTRODUCTION

1. INTRODUCTION

This subject report aims to delve into the world of predicting customer churn using machine learning techniques. By leveraging historical customer data and powerful algorithms, organizations can gain valuable insights into customer behavior, enabling them to anticipate churn, identify potential warning signs, and implement targeted retention strategies.



1. INTRODUCTION



Key Objectives

2.
THEORETICAL BASIS

2. THEORETICAL BASIS

DATA MINING

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data

Data mining is often used in fields such as finance, healthcare, and marketing to help businesses make more informed decisions based on their data

2. THEORETICAL BASIS

CHURN RATE

According to Arun Velu (2021), the situation where customers stop using the service or switch to a competitor is called churning.

Voluntary and involuntary churn are the two main categories of churn.

Businesses need to focus on voluntary churn and come up with strategies to meet customer needs and satisfaction.

2. THEORETICAL BASIS

ASSOCIATION DECISION TREE (ADT)

ADT is a machine learning technique that combines decision tree algorithms with association rule mining.

ADT is useful for analyzing data where there may be complex relationships and dependencies between attributes

2. THEORETICAL BASIS

MULTICLASS ASSOCIATIVE CLASSIFICATION

MAC is a type of classification algorithm that can handle multiple classes of data by leveraging association rules. MAC builds on the concept of Associative Classification (AC), which uses association rule mining to generate a set of rules that predict the class label of a new instance based on its attribute values.

In MAC, the class label is not binary, but rather a set of possible classes.

2. THEORETICAL BASIS

CLASSIFICATION BASED ON MULTIPLE ASSOCIATION RULES

This is another type of classification algorithm that uses association rule mining to generate a set of rules that can be used to classify new instances. CMAR is similar to MAC, but it differs in the way that it handles missing values and noisy data.

2. THEORETICAL BASIS

CLASSIFICATION BASED ON MULTIPLE ASSOCIATION RULES

This is another type of classification algorithm that uses association rule mining to generate a set of rules that can be used to classify new instances. CMAR is similar to MAC, but it differs in the way that it handles missing values and noisy data.

2. THEORETICAL BASIS

THE K-NEAREST NEIGHBOR (KNN)

The k-nearest neighbor (k-NN) algorithm is a type of supervised learning algorithm used for classification and regression tasks.

The k-NN algorithm is a non-parametric algorithm, which means that it does not make any assumptions about the distribution of the data. Instead, it directly uses the training data to make predictions.

2. THEORETICAL BASIS

THE K-NEAREST NEIGHBOR (KNN)

To determine the closest groups or the nearest points for a query point, there are 3 commonly used calculation methods:

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k}|x_i - y_i|$$

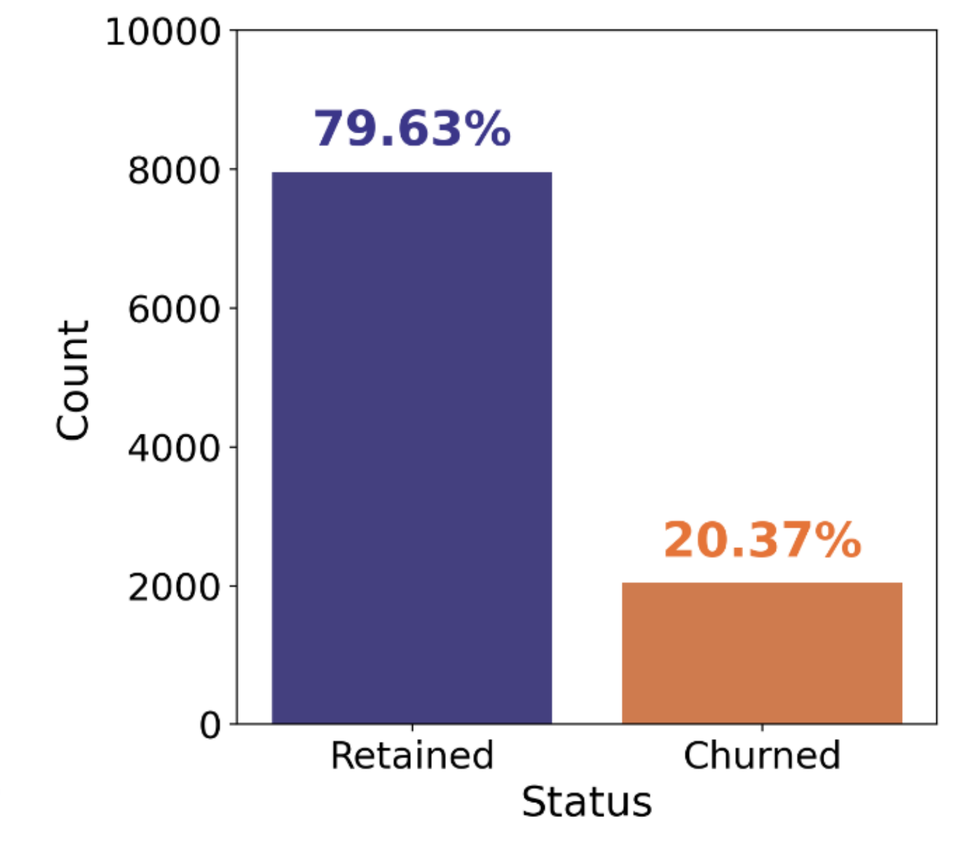
Minkowski

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

3.
DATA EXPLORATION

3. DATA EXPLORATION

CATEGORIES VARIABLES

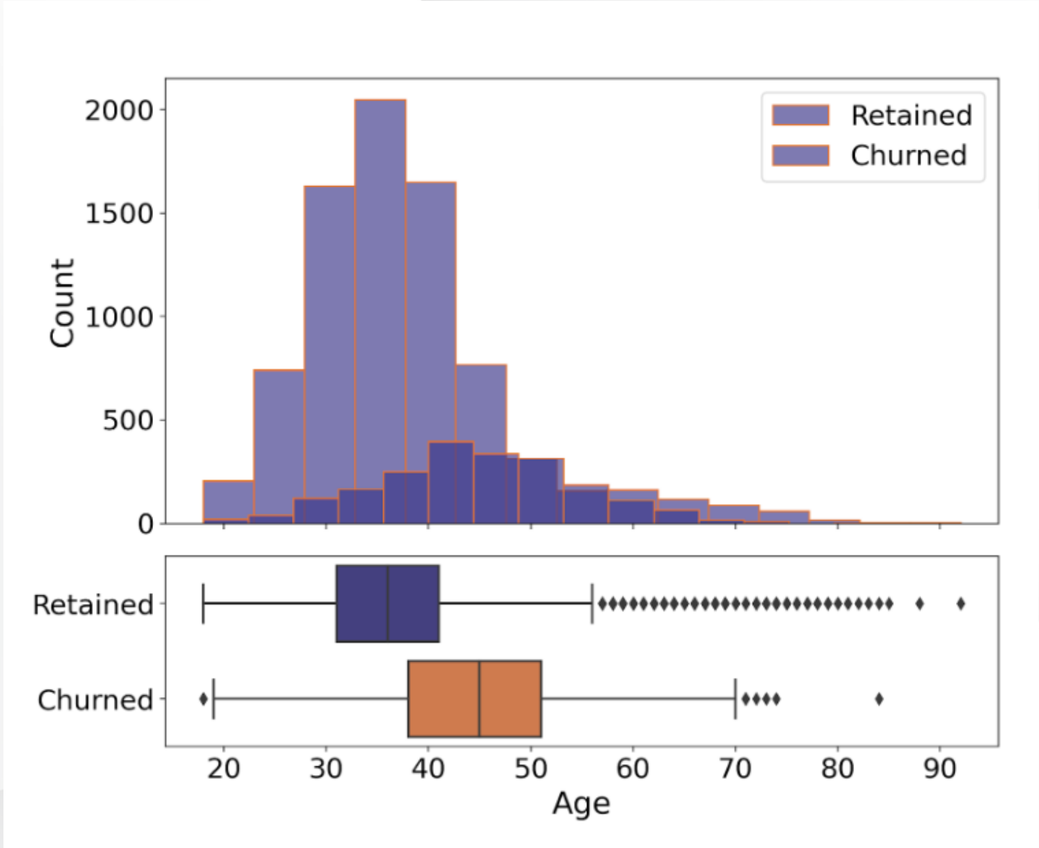


Zero (0) for a customer who hasn't left and one (1) for a customer who has left.

Low churn rate is not necessarily a good thing because the bank may grow slowly and find it difficult to replace new customers, if the churn rate is high, the bank may have quality and service problems. It is necessary to consider the bank's services and strategy.

3. DATA EXPLORATION

RELATIONSHIP OF TARGET VARIABLE

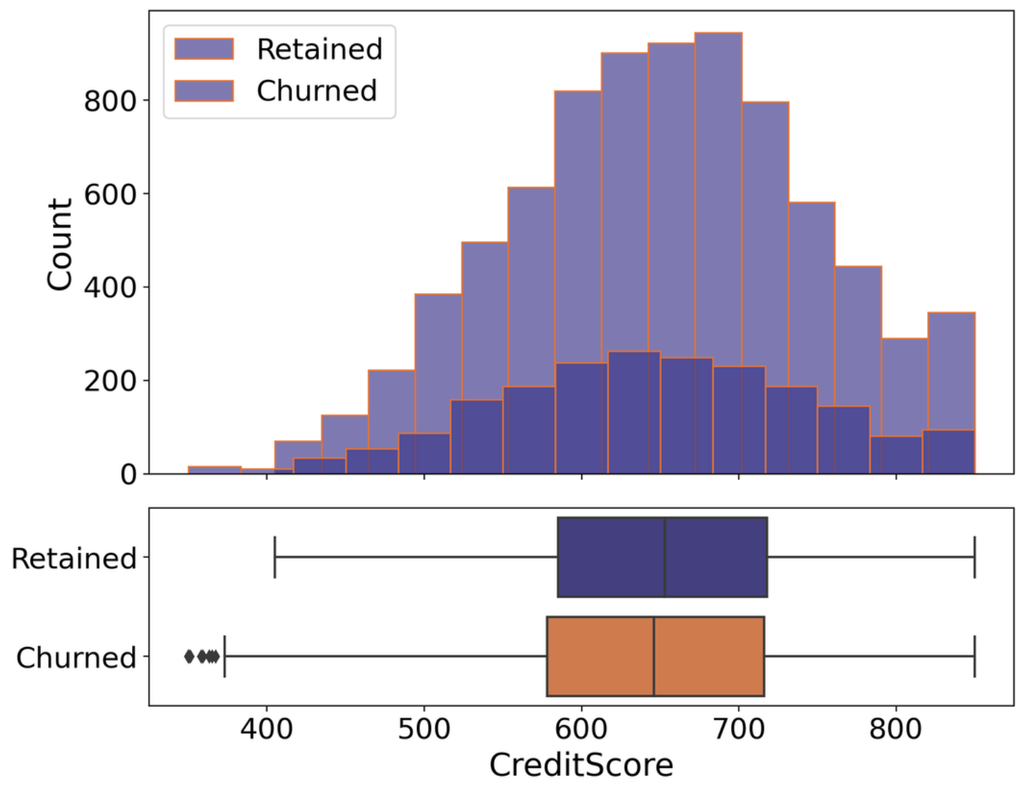


The older the customer file, the more likely to leave

the bank probably did not have many surveys on middle-aged people to understand their needs and preferences

3. DATA EXPLORATION

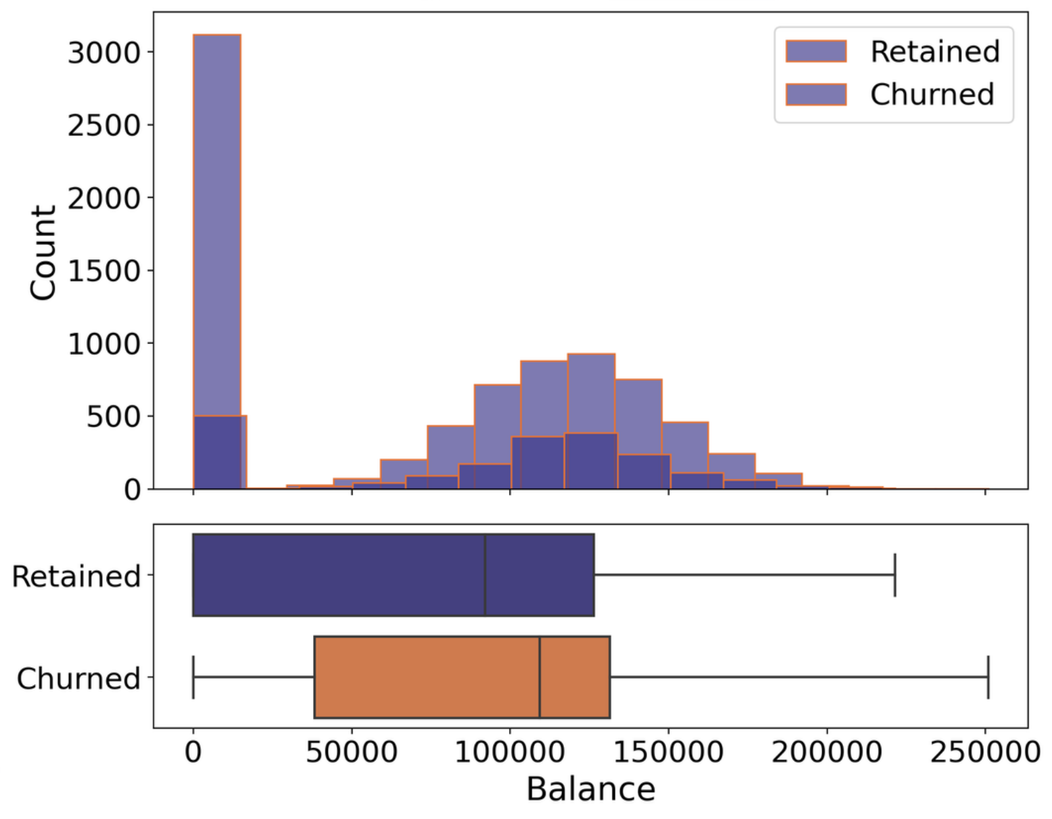
RELATIONSHIP OF TARGET VARIABLE



The credit score of customers who leave and customers who continue to use the service at the bank does not have much difference as well as a significant difference

3. DATA EXPLORATION

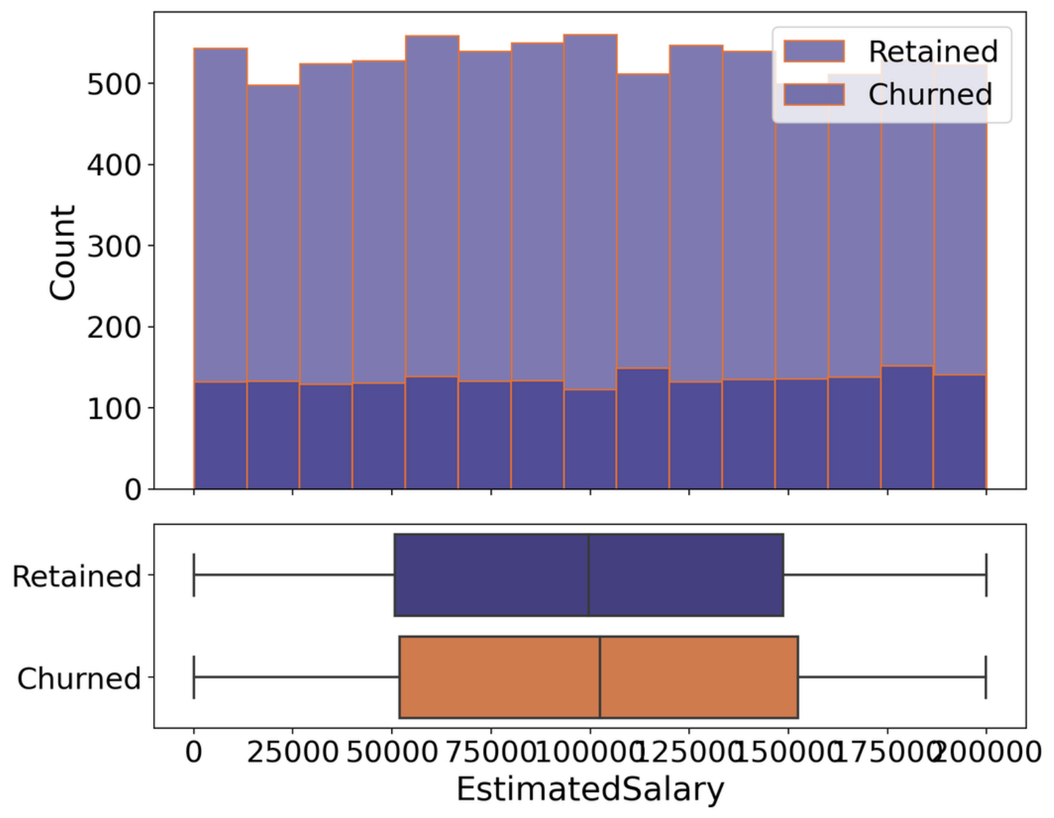
RELATIONSHIP OF TARGET VARIABLE



It can be seen that the number of customers who do not leave with a low account balance is almost zero, which is understandable because the customer has just opened a bank account, so there is not much account balance

3. DATA EXPLORATION

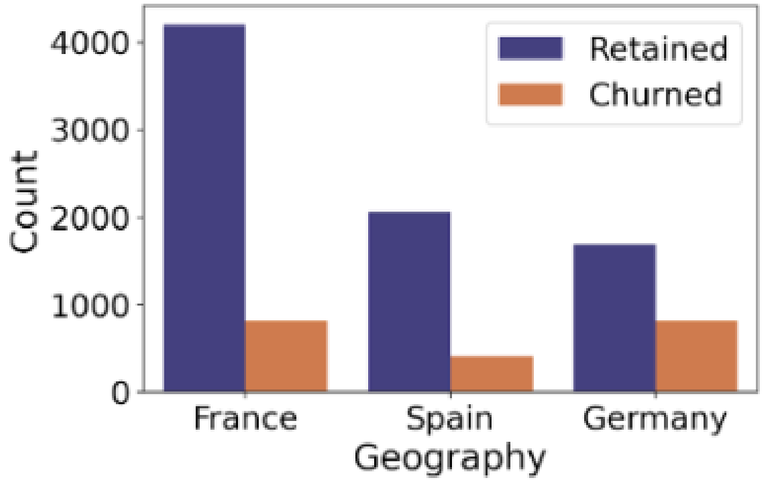
RELATIONSHIP OF TARGET VARIABLE

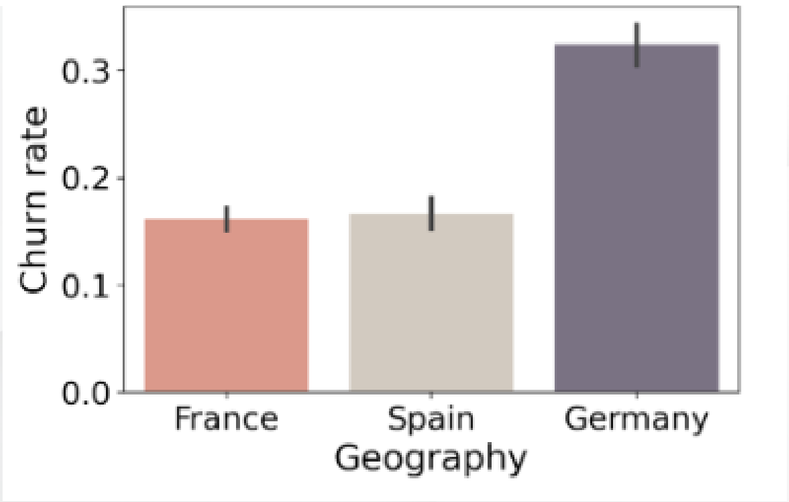


The salary distribution of customers who stay and leave the bank with almost no difference, which shows the uniformity of their salaries.

3. DATA EXPLORATION

RELATIONSHIP OF TARGET VARIABLE

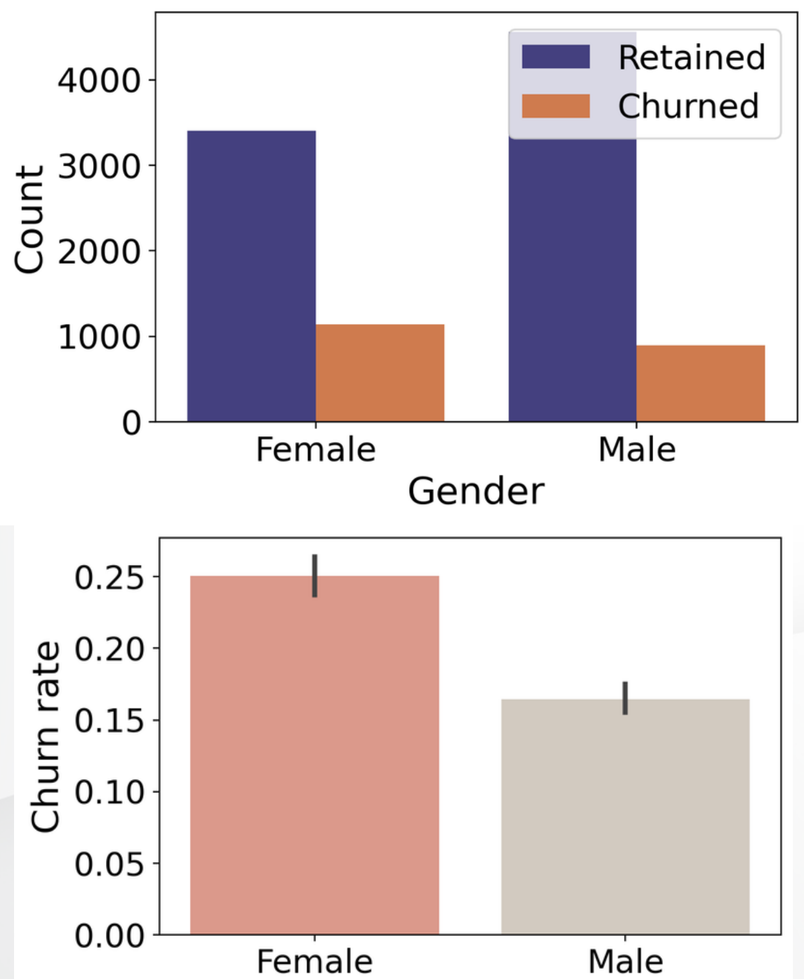




Customers in Germany are more likely to leave than customers in the other two countries (the churn rate is almost twice that of Spain and France) . Many reasons could explain this finding, such as higher competition or different preferences of German customers.

3. DATA EXPLORATION

RELATIONSHIP OF TARGET VARIABLE



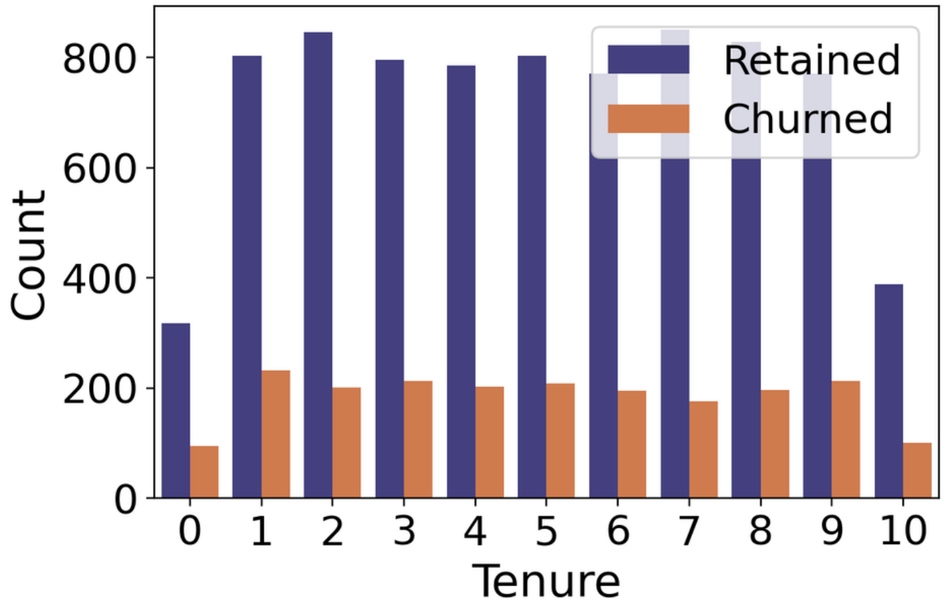
Female customers are more likely to leave than men.

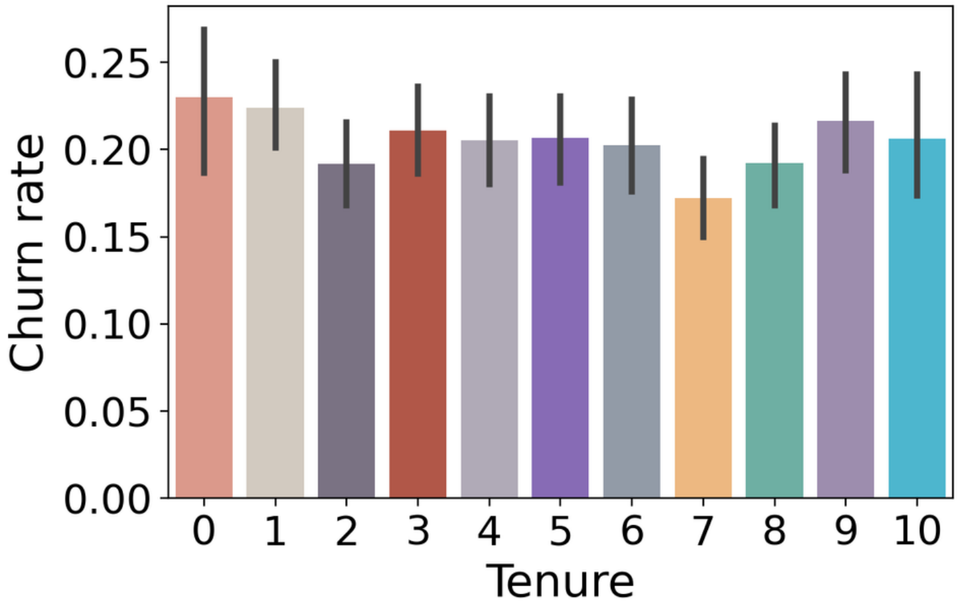
We can however explain that female customers have high requirements for service quality as well as purchase decisions based on emotions.

For men, they are more interested in financial factors, interest rates, risk or safety are also carefully analyzed by them, so they may be less likely to leave than women.

3. DATA EXPLORATION

RELATIONSHIP OF TARGET VARIABLE

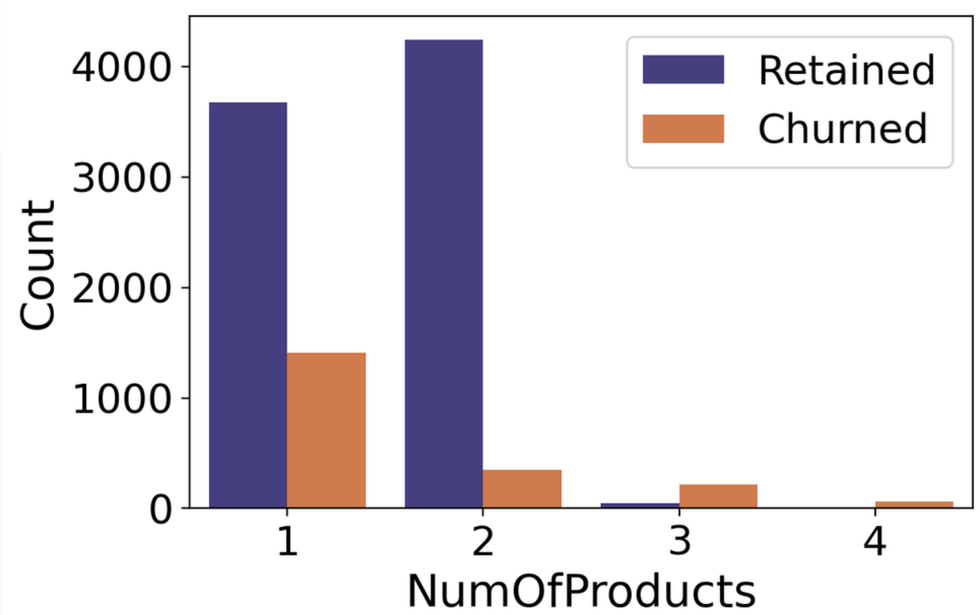


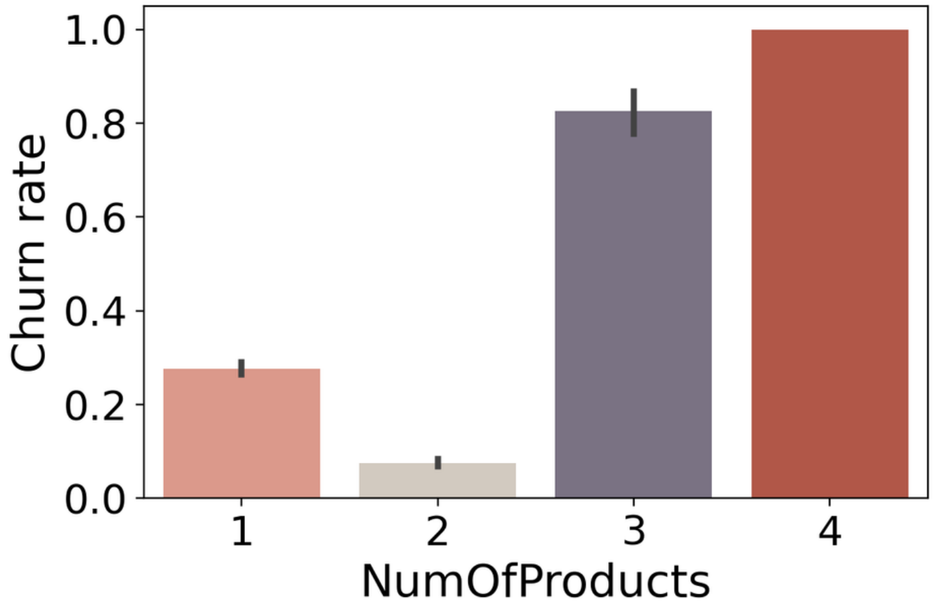


The number of years of banking service has almost no effect on the churn rate. However, we can clearly see that the customer churn rate from the first year is higher than the remaining years.

3. DATA EXPLORATION

RELATIONSHIP OF TARGET VARIABLE

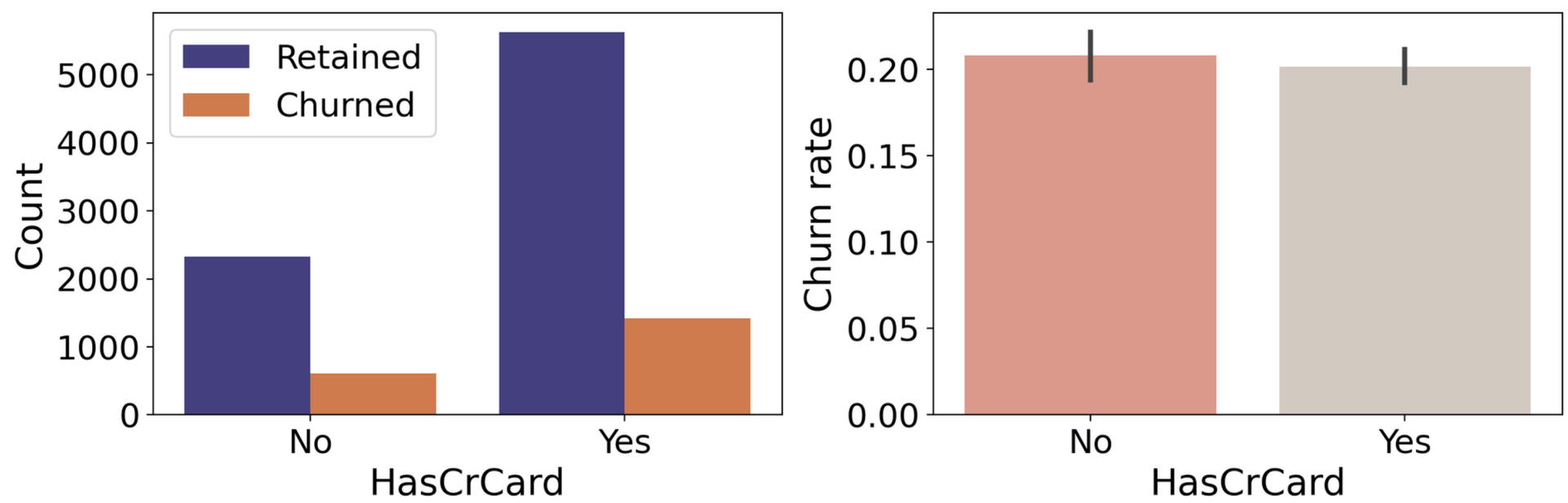




Many customers who only use 1-2 services. This shows that the bank does not do well in providing many of their services to customers

3. DATA EXPLORATION

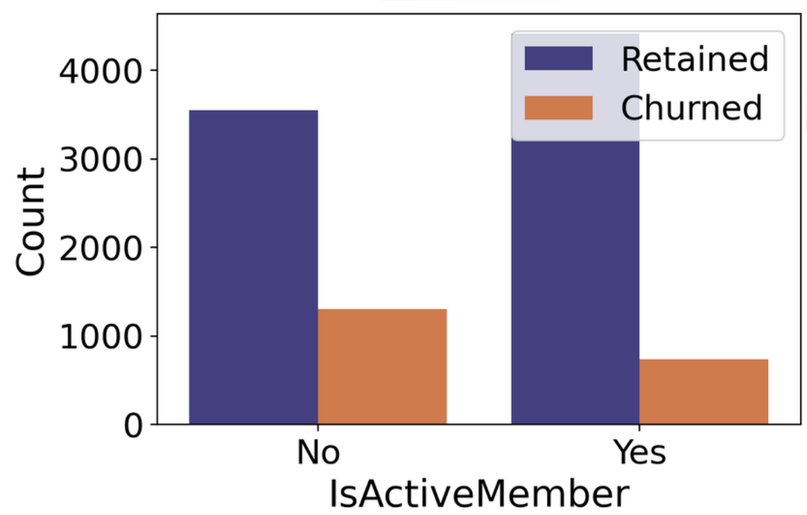
RELATIONSHIP OF TARGET VARIABLE

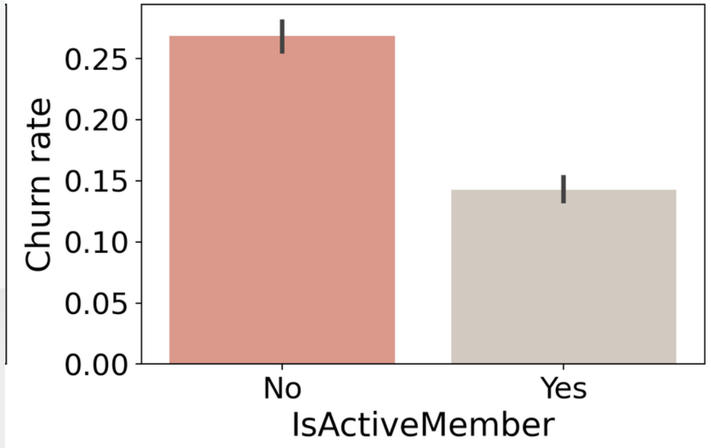


That whether or not a customer has a credit card doesn't seem to affect customer churn

3. DATA EXPLORATION

RELATIONSHIP OF TARGET VARIABLE





We see that inactive customers account for a much higher abandonment rate, which is understandable when these two values are closely related

This is a really important factor and greatly affects the churn rate of customers, once it is no longer active, most customers will not want to continue using the bank's services

3. DATA EXPLORATION

CORRELATIONS MATRIX

We can easily see that there is no significant correlation between the variables. This reassures us about the reduction of multicollinearity

4. PROPOSED MODEL

	CBA	CBA2	ADT	ACAC	ACCF	ACN	L3	MAC	CMAR	KNN	ID3
Accuracy Score	0.7902	0.8062	0.7963	0.7963	0.7963	0.7963	0.7963	0.7963	0.7963	0.9898	0.1636

From the result we can see the logistic outperformance (highest) so the rest of the algorithm and all the accuracy from the result is more than 80%. So, the team will start with **K-Near Neighbors** regulation with hyperparameter adjustments.

5. EXPERIMENTAL RESULT AND ANALYSIS

COMPARE RESULTS

#NAME:	ADT	CBA	CBA2	MAC	L3	CMAR	KNN
#ACCURACY:	0.7963	0.7902	0.8062	0.7963	0.7963	0.7963	0.9898
#RECALL:	0.5	0.6397	0.5281	0.5	0.5	0.5	0.9749
#PRECISION:	0.3982	0.6659	0.828	0.3982	0.3982	0.3982	0.9937
#KAPPA:	0	0.3017	0.0861	0	0	0	0.9679
#FMICRO:	0.7963	0.7902	0.8062	0.7963	0.7963	0.7963	0.9898
#FMACRO:	0.4433	0.6499	0.5007	0.4433	0.4433	0.4433	0.9839
#TIMEms:	4	10	3	15	3	12	1594
#MEMORYmb:	39.0993	40.6013	42.0993	49.0993	51.6013	54.6013	45.6663
#NOPREDICTION:	0.0	0.0	0.0	0.0	0.0	0.0	0.0

We have hidden the highest accuracy measurement values to easily consider the algorithms that get good results in the training set. We can clearly see that the ratio of KNN stands out from the rest. However, the execution time is much longer

5. EXPERIMENTAL RESULT AND ANALYSIS

COMPARE RESULTS

#NAME:	ADT	CBA	CBA2	MAC	L3	CMAR	KNN
#ACCURACY:	0.796	0.802	0.816	0.796	0.796	0.796	0.996
#RECALL:	0.5	0.6733	0.5508	0.5	0.5	0.5	0.9902
#PRECISION:	0.398	0.6907	0.8837	0.398	0.398	0.398	0.9975
#KAPPA:	0	0.3624	0.1522	0	0	0	0.9876
#FMICRO:	0.796	0.802	0.816	0.796	0.796	0.796	0.996
#FMACRO:	0.4432	0.6809	0.5411	0.4432	0.4432	0.4432	0.9938
#TIMEms:	0	0	0	2	1	1	161
#MEMORYmb:	39.0993	40.6013	42.0993	49.6013	51.6013	54.6013	5.67
#NOPREDICTION:	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The value at the test set is the one we are most interested in. Here, the KNN algorithm is still very prominent with rates from 0.99 and up. However, memory and execution time still consume too much.

5. EXPERIMENTAL RESULT AND ANALYSIS

EVALUATION

There is a clear difference in the execution time of the algorithms. With a data set of 10,000 lines, timing really doesn't matter, but with large amounts of data, we are forced to consider and choose optimal algorithms in terms of time and memory.

6. CONCLUSION

RESULT

Throughout the study, we analyzed the data comprehensively, extracting meaningful features that shed light on factors influencing customer churn.

By using the KNN classifier, we have proposed a highly reliable predictive model that accurately identifies potential leavers as well as the factors and characteristics affecting the abandonment rate. Thereby, helping the bank understand customer behavior better and take proactive measures in retaining valuable customers.

6. CONCLUSION

LIMITATION

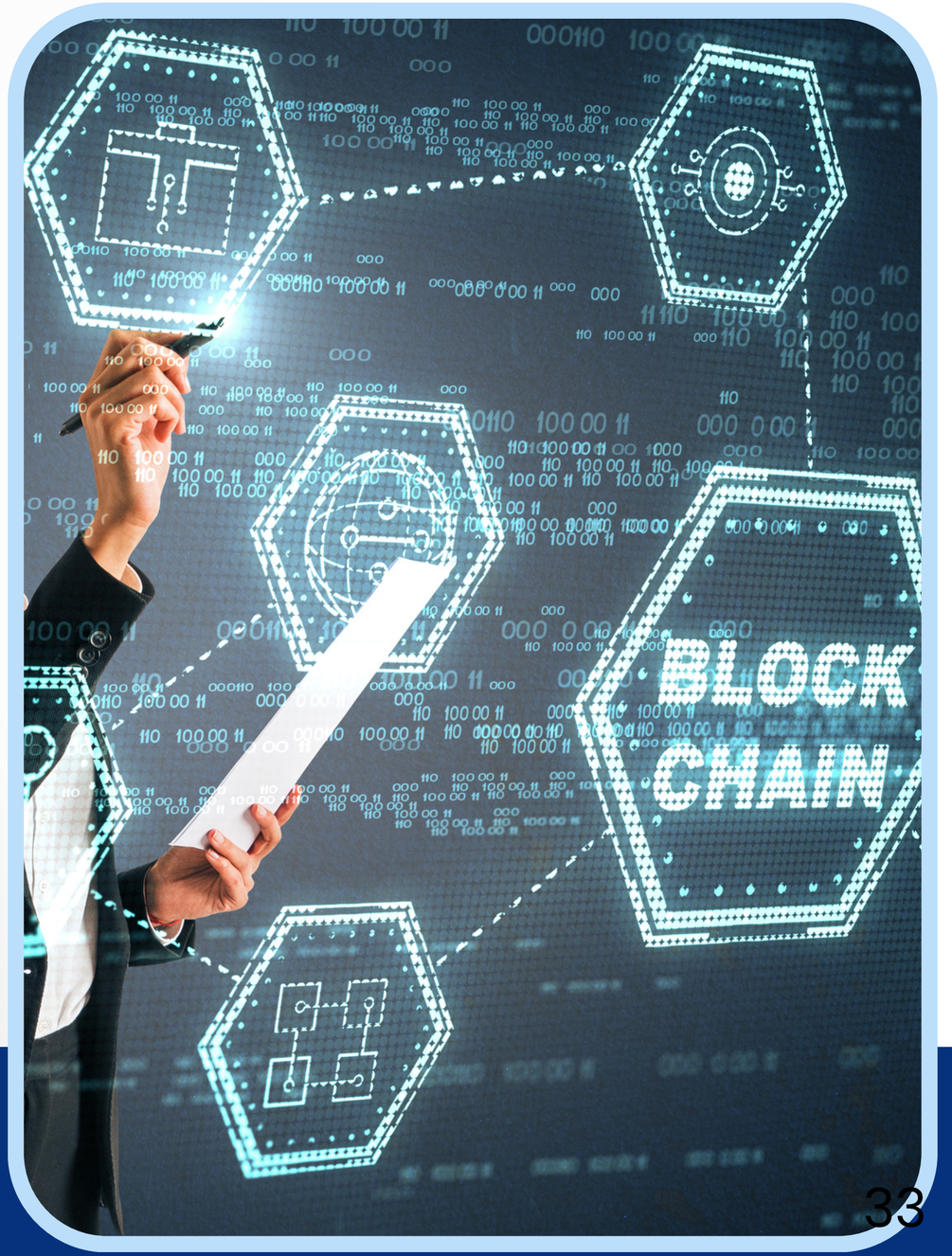
Data preprocessing and feature engineering play a key role in enhancing model performance, but the quality and completeness of the original dataset pose certain barriers

The research and analysis time is limited, so the team has not studied deeply during the implementation of the project.

6. CONCLUSION

DEVELOPMENT ORIENTATION

Future research should incorporate additional data sources and explore more complex models like ensemble methods, deep learning, or support vector machines to potentially yield more refined churn prediction models. Investigating the impact of personalized marketing strategies and customer retention initiatives based on the model's predictions would also be a valuable area of exploration



Thank You