




ISSN: 2617-6548

URL: [www.ijirss.com](http://www.ijirss.com)



## Leveraging predictive intelligence to understand banking customer behavior

 Nihel Ziadi Ben Fadhel

*Labortory ERMA, FSEGT University of Tunis Elmanar, Tunisia.*

(Email: [Nihel.ziadi@fst.utm.tn](mailto:Nihel.ziadi@fst.utm.tn))

### Abstract

This study aims to address the challenge of optimizing customer retention and marketing effectiveness in the banking sector, particularly considering rapidly evolving customer behaviors driven by digitalization and the increasing demand for personalized services. Focusing on a Tunisian financial institution, the research evaluates the combined use of predictive churn modeling, customer segmentation through K-means clustering, and association rule-based recommendation systems to enhance strategic decision-making and revenue growth. Real-world transactional and behavioral data from bank customers were analyzed using machine learning algorithms to develop predictive models for churn, identify distinct customer segments, and uncover affinity patterns for targeted marketing. The findings demonstrate that integrating these analytical techniques significantly improves marketing performance by enabling more precise targeting, tailoring personalized campaigns, and reducing customer attrition rates. The results reveal that predictive intelligence tools contribute to a better understanding and influence of customer behavior, ultimately driving profitability in a competitive banking environment. The study concludes that the strategic application of advanced data analytics is crucial for banks seeking to maintain customer loyalty and optimize marketing investments. These insights underscore the practical importance of financial institutions adopting integrated analytics frameworks, as such approaches support more effective targeting, personalization, and customer retention, thereby ensuring sustained competitiveness and profitability in the digital era.

**Keywords:** Banking customer behavior, Customer churn prediction, Predictive intelligence.

**DOI:** 10.53894/ijirss.v8i5.9273

**Funding:** This study received no specific financial support.

**History: Received:** 16 June 2025 / **Revised:** 16 July 2025 / **Accepted:** 18 July 2025 / **Published:** 13 August 2025

**Copyright:** © 2025 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Competing Interests:** The authors declare that they have no competing interests.

**Authors' Contributions:** All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

**Transparency:** The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

**Publisher:** Innovative Research Publishing

### 1. Introduction

The banking sector is undergoing a profound transformation driven by digitalization and evolving customer expectations. As customers increasingly demand personalized and seamless experiences, banks face the critical challenge

of optimizing customer retention and enhancing marketing effectiveness to maintain competitiveness. Despite the availability of vast amounts of transactional and behavioral data, many financial institutions struggle to fully leverage advanced analytics to anticipate customer churn and tailor marketing strategies effectively.

This study aims to address this gap by focusing on a Tunisian financial institution and exploring how the integration of predictive churn modeling, customer segmentation through K-means clustering, and association rule-based recommendation systems can improve strategic decision-making and revenue growth. While prior research has examined these analytical techniques individually, few studies have investigated their combined application within the banking context, particularly in emerging markets such as Tunisia.

The main research questions guiding this study are:

1. How can predictive churn models accurately identify customers at risk of attrition in the Tunisian banking sector?
2. What distinct customer segments can be identified using clustering techniques, and how do these segments differ in behavior and value?
3. How can association rule mining uncover affinity patterns that enhance targeted marketing campaigns?

To answer these questions, the research analyzes real-world transactional and behavioral data from bank customers using machine learning algorithms. The study develops predictive models to forecast churn, segments customers into meaningful groups, and extracts association rules to inform personalized marketing recommendations. This integrated analytical approach is then evaluated for its impact on marketing performance and customer retention.

In summary, this research contributes to both theory and practice by demonstrating the value of combining multiple data analytics techniques to optimize customer relationship management in a competitive banking environment. The findings provide actionable insights for financial institutions seeking to leverage data-driven strategies to enhance customer loyalty and profitability in the digital era.

## **2. Literature Review**

### **2.1. Churn Prediction in Banking**

According to Nettleton et al. [1], customer churn analysis can be defined as analytical work carried out on the possibility of a customer leaving a product or service. In its simplest definition, it means that customers abandon the company because of competition.

Predictive intelligence uses prediction models that approximate the risk of customer attrition. In fact, predicting customer attrition can help banks to plan suitable marketing campaigns to convince clients who are potentially candidates for leaving. Overall, the demand for customer attrition analysis is increasing. The study of the characteristics related to the customers' profile and behavior, by consulting their transaction history, remains the most widely used approach in research works related to this domain. Most of these are statistical learning methods. This raises the following question: which learning model can best predict customer churn?

By examining methods used in the literature, we find that popular methods for predicting churn likelihood are logistic regression (LR), K-Nearest Neighbors (KNN), decision trees (DT), and SVM. However, Elyusufi and Ait Kbir [2] stated that these methods have reached their limits and have practical difficulties, resulting in the emergence of a new generation of tree-based ensemble algorithms such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosted Machine (LightGBM) models. Based on this, we will apply such algorithms in this project.

### **2.2. Data Segmentation**

In this rapidly evolving environment, understanding customer segmentation within banks has become increasingly imperative. According to Essayem et al. [3], by segmenting customers based on their behaviors, preferences, and needs, banks can tailor their offerings and strategies to better serve their diverse clientele and gain a competitive edge in the market.

This is proven by Azad [4]: « Segmentation processes can help businesses and companies understand their customer groups, target the right groups, and develop effective marketing strategies for different targeted groups. Clustering techniques are the most appropriate methods that enable businesses and companies to identify segments or groups of customers to target the potential user base. Customer segmentation can be performed using a variety of different customer characteristics.

Along with customer segmentation, we also find Marketing segmentation, based on this article by Raiter [5]. « Market segmentation, at its most basic level, pushes financial businesses to assess where they are now and where they aspire to go in the future. As a result, businesses are forced to consider what they are especially strong at in comparison to rivals, as well as trying to understand what customers want. Market segmentation allows for vital new insights and views, as well as the chance to analyze and reconsider, and that, taking market segmentation to its logical conclusion, would entail being able to provide a tailored product or service to very small groups of account holders.

Based on this, the idea of implementing a two-tiered segmentation, also known as Hierarchical Segmentation, was developed. The first tier is broader and more general, representing Customer segmentation. The second tier is more refined, focusing on each customer type; thus, this nested segment will represent Marketing segmentation. According to Chéron et al. [6] « By employing a two-stage segmentation approach, banks can more accurately identify and target specific subgroups within the commercial market. This targeted approach allows for more effective marketing strategies tailored to the unique needs of each segment, leading to improved client engagement and satisfaction.»

Based on Gopalakrishnan [7] the K-Means clustering algorithm stands out as a sophisticated analytical tool that offers a powerful method for segmenting customers based on multivariate data. The application of K-Means clustering to

customer segmentation in the banking sector demonstrates the value of machine learning in identifying distinct customer profiles. By leveraging these insights, banks can implement targeted marketing strategies that cater to the specific needs and behaviors of different customer groups. This targeted approach not only enhances customer engagement but also fosters loyalty and drives profitability by ensuring that customers receive relevant offers and services that resonate with their financial habits and needs. Thus, we will use the K-Means clustering technique to accomplish data segmentation in this project.

### 2.3. Association Rules

Among the key objectives of the bank's strategic marketing initiatives is maximizing the customer value delivered to the organization, primarily through cross-selling and up-selling a diverse range of the bank's products and services to its existing customer base.

This, in turn, is expected to contribute significantly to the bank's revenue targets by improving the efficiency and effectiveness of its sales initiatives, which focus on maximizing customer transaction intensity and value by identifying patterns in their purchasing behavior. To improve the success rate of current targeted marketing campaigns, market basket analysis was employed as a customer analytics technique. According to Fadhel [8] the different types of relationships impact the performance of banks and advocate for the complementarity between transactional and relational approaches to improve perceived performance.

According to this article by Mansingh et al. [9], Market Basket Analysis uses association rule mining to identify the products that customers currently purchase together and can help to identify those products that go well together (in terms of bundles) and therefore should be marketed accordingly. Furthermore, since a customer does not buy a set of products at one time, but rather the basket contains products bought over time, it is conducive to sequential rule mining, which not only shows which products were bought together but also the sequence in which they were bought.

Both association rules and sequential mining will help to increase the effectiveness of sales campaign management and targeting processes. Based on this article by Desai and Kaiwade [10] a priori algorithm is an influential algorithm for mining frequent item sets for boolean association rules. Various data mining techniques were used earlier for pattern analysis. However, for finding locally frequent items, it is most suitable, especially for transactional databases. This has led to various improvisations of the core approach. Thus, in banking, data mining plays a vital role in handling transaction data and customer profiles. From that, using data mining techniques, a user can make effective decisions. Finally, it can be concluded that banks will obtain massive profits if they implement data mining in their process of data and decisions.

### 3. Research question

- How accurately can churn be predicted within the bank's customer base using predictive modeling techniques?
- What distinct segments can be identified within the bank's customer base using K-Means clustering, and how can these segments be leveraged to tailor marketing campaigns?
- How can association rules and recommendation systems be used to identify cross-selling and upselling opportunities to drive revenue growth and improve customer engagement at the bank?

### 4. Research Hypotheses

Based on the literature review, the following research hypotheses are proposed:

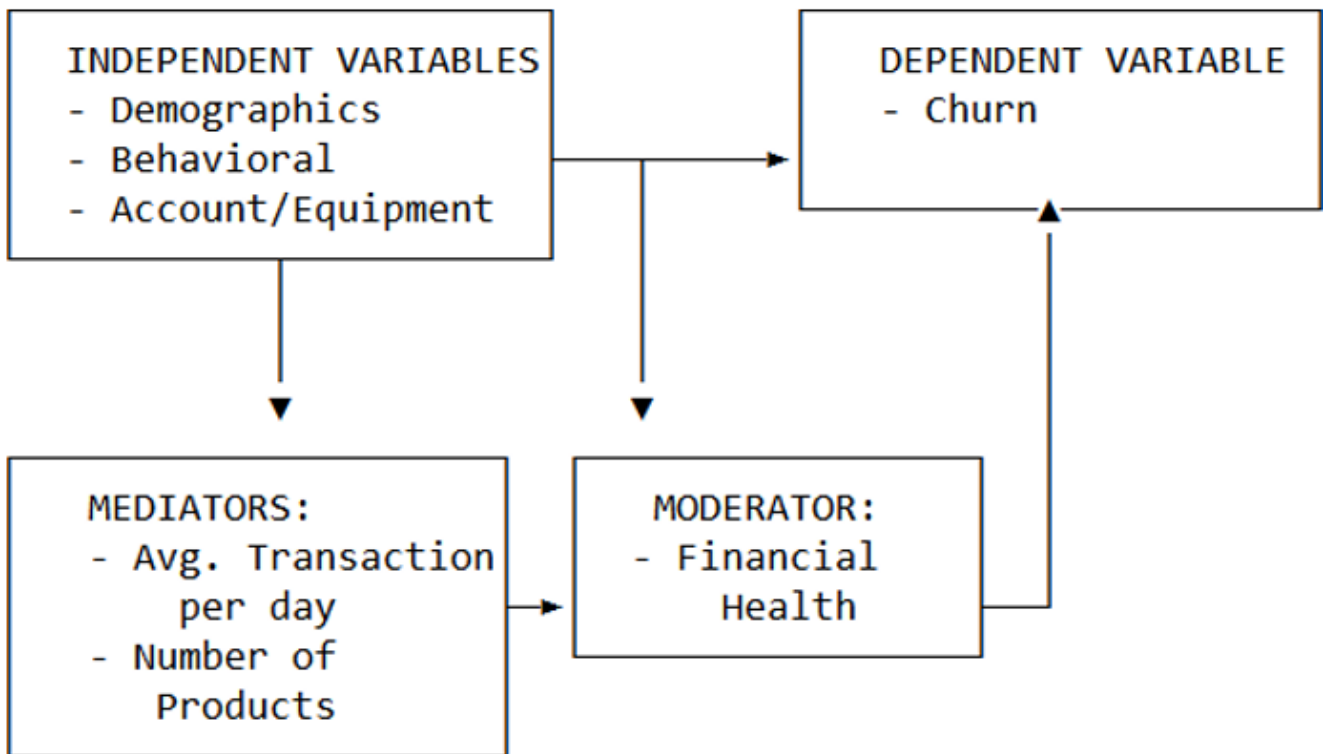
*H<sub>1</sub>: A predictive model incorporating socio-demographic and behavioral characteristics can accurately predict customer churn at a bank*

*H<sub>2</sub>: Hierarchical segmentation using k-means clustering will identify statistically significant segments within the bank's customer base, differentiated by their financial product needs.*

*H<sub>3</sub>: Customers are more likely to purchase products related to those they already use, suggesting that implementing an association rule-based recommendation system has a significant impact on cross-selling and up-selling rates within banks.*

## 5. Theoretical Models

### 5.1. Theoretical Model for Hypothesis 1:



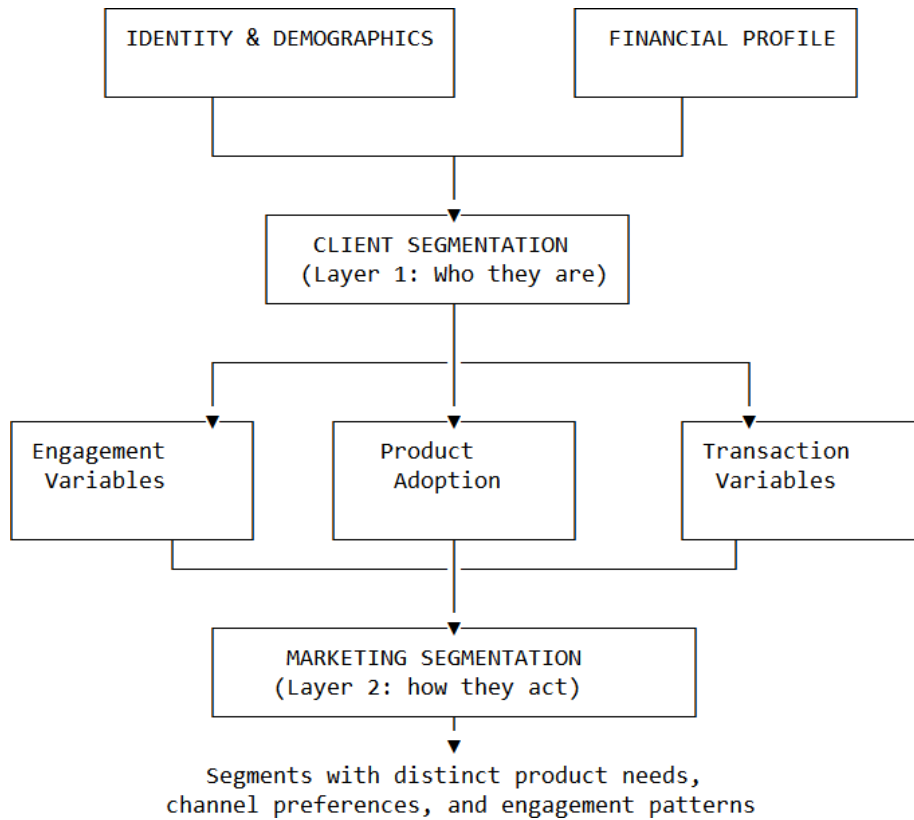
**Figure 1.**  
H1 theoretical model.

This theoretical model highlights the variables used in the predictive model; we have 4 types of variables:

- Independent variables: Also known as "Predictors," these are input variables that help explain or predict outcomes, such as demographics, behavioral variables, and variables that are account- and equipment-related.
- Dependent variable: Also known as "Outcome variable." This variable represents the result of our predictive modeling.
- Moderators: These variables are essential; they help influence the strength of the relationship between independent and dependent variables. They modify how strongly one variable affects another, such as financial health.
- Mediators: These variables help explain the process through which independent variables influence churn. For example, average transactions per day and the number of products help us understand how client behavior leads to churn.

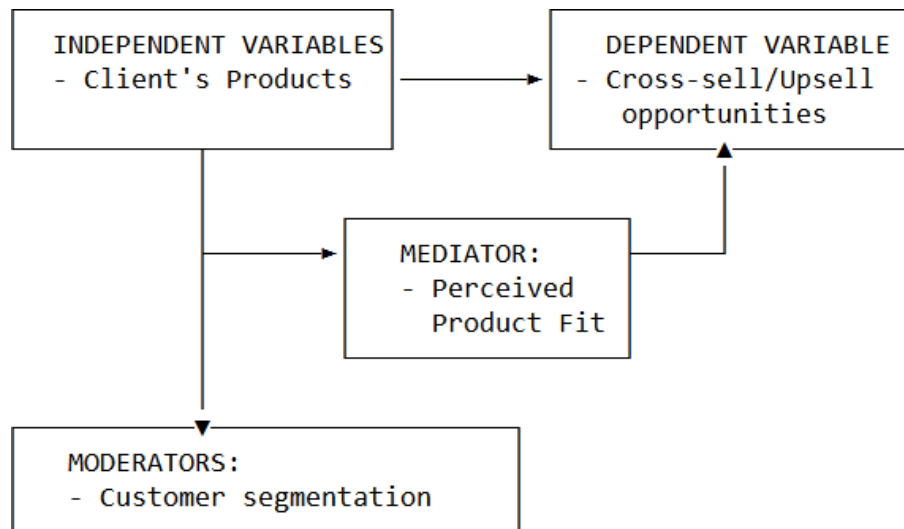
### 5.2. Theoretical Model for Hypothesis 2

This theoretical model (see Figure 2, page 16) highlights a hierarchical approach that creates sub-segments, first by using these independent variables: 'identification and demographic' and 'financial profile variables' to apply k-means clustering to create the first layer of the segmentation: Client segmentation. Then, by using this independent variable: 'Engagement variables,' 'Product Adoption,' and 'Transaction variables' to apply k-means and create the second and final layer representing Marketing segmentation.



**Figure 2.**  
H2 theoretical model.

### 5.3. Theoretical Model for Hypothesis 3



**Figure 3.**  
H3 theoretical model.

This model highlights the variables used in the association rule-based recommendation system. For the independent variable, we used only product (equipment)-related variables, with the help of customer segmentation and the use of perceived product fit (the implied likelihood that a client will see a recommended product as relevant or useful, based on their existing product usage patterns). The system output (dependent variable) is to identify cross-sell/up-sell opportunities.

## 6. Adopted Methodology

For the empirical validation of our hypotheses, the methodology for this project is the Cross-Industry Standard Process for Data Mining, known as CRISP-DM, which is an open standard process model that describes common approaches used by data mining experts. It is the most widely used analytics model. CRISP-DM breaks the process of data mining into six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

In this article, Python was the primary programming language used due to its versatility and extensive libraries suited for data science and machine learning tasks. The development and experimentation were conducted within Jupyter

Notebook, which facilitated interactive coding, visualization, and documentation of the workflow. For data visualization and business intelligence, Microsoft Power BI was employed to transform diverse data sources into insightful static and interactive visualizations. Together, these tools provided a comprehensive and efficient environment for implementing, analyzing, and presenting the machine learning models discussed.

## 6.1. Data Preparation

### 6.1.1. Creating the TARGET Feature

We will begin this phase by creating a new feature named 'TARGET'. This feature is binary, returning 1 if the client exists in the attrition table (indicating that the client has left the bank), and 0 if the client is still with the bank. This feature will be created in all tables except for the attrition table, which will no longer be used.

Defining the TARGET column is crucial, as it serves not only to identify targeted clients but also to guide the selection of key predictor variables for optimized churn analysis by applying statistical hypothesis tests.

### 6.1.2. Statistical Hypothesis Test

Statistical hypothesis testing is a method of statistical inference used to determine whether the data provides sufficient evidence to reject a specific hypothesis. These hypotheses include  $H_0$ , which represents the null hypothesis, and  $H_1$ , which represents the alternative hypothesis. The null hypothesis is always tested using a hypothesis test. The purpose of the test is to assess whether there is no difference or no relationship between variables, based on the data analyzed.

$H_0$ : column  $i$  and TARGET are independent.

$H_1$ : column  $i$  and TARGET are dependent.

To determine whether to reject the null hypothesis, we rely on a value called the p-value. If the p-value is less than the level of significance, the null hypothesis  $H_0$  is rejected; if the p-value is greater than the level of significance, it is not rejected.

### 6.1.3. Level of Significance Funding

A hypothesis test can never reject the null hypothesis with absolute certainty. There is always a certain probability of error that the null hypothesis is rejected even though it is true. This probability of error is called the significance level or  $\alpha$ .

If we set a significance level of 5% ( $\alpha = 0.05$ ), we are accepting a 5% risk of incorrectly rejecting the null hypothesis. This risk exists because we are drawing conclusions about a larger population based on a smaller sample. Even if there is no real difference in the population, random variation in our sample might lead us to observe a difference between the means of two groups. The greater the difference we observe between sample means, the less likely it is that both samples originated from the same underlying population.

The key question is "how big does the difference in means between two groups need to be to convince us it's not just random chance and therefore reject the null hypothesis in favor of the alternative hypothesis?" Our chosen significance level (5%) answers this question. If the p-value is less than or equal to our significance level, it means our results are unlikely enough if the null hypothesis were true. In that case, we reject the null hypothesis.

It is important to note that the significance level is always set before the test and may not be changed afterward to obtain the "desired" statement after all.

### 6.1.4. Types of Errors

A hypothesis can only be rejected with a certain level of probability, leading to different types of errors. One possibility is that the null hypothesis may be rejected by chance due to sample selection, even when it is valid and no difference in means between the two groups exists. Conversely, it is also possible for the null hypothesis not to be rejected when a difference does exist, meaning the alternative hypothesis is true.

Accordingly, there are two types of errors in hypothesis testing:

- Type 1 error: If the alternative hypothesis is accepted, although the null hypothesis is valid.
- Type 2 error: If the null hypothesis is not rejected, although the alternative hypothesis applies.
- Overall, the following cases arise:

	Decision	
	for $H_0$	against $H_0$
$H_0$ true	Correct	Type 1 error
$H_0$ false	Type 2 error	Correct

Figure 4.  
Types of errors.

To test hypotheses, various test procedures are available. On the one hand, these are divided according to the levels of measurement of the sample.

- Nominal variable: is a categorical variable that does not have any intrinsic ordering or ranking (A, B, C, D).

- Ordinal variable: is a categorical variable that has natural, ordered categories (A < B < C < D).
- Metric variables: are variables on which calculations are meaningful, and on the other hand, how many samples are present and how the samples are related to each other. After carefully reviewing the cleaned and preprocessed dataset, I noticed that most variables are categorical (including the binary ones), with just one numerical column: 'Operation\_number'. To analyze the dependence on TARGET, we used the Chi-squared test for categorical variables and a T-test for Operation\_number. Chi-squared test.

The Chi-square test of independence,  $\chi^2$  is a statistical hypothesis test used to determine whether two categorical or nominal variables are likely to be related or not.

The first step is to select the categorical variables in each table and store them in a new data frame called df\_cat. Next, we will apply the Chi-square test.

Using the chi2\_contingency ()function of scipy.stats module, we got these results:

	Variable	Chi2_p-value	degree of freedom	Chi-squared
	MMM_BIN	3.257408e-48	4	228.177714
	AGE_BINNED	2.316892e-20	4	98.252031
	avg_monthly_balance_binned	4.400546e-04	4	20.278141
	ENGAGEMENT_Binned	1.320943e-02	6	16.103434
	Active_Investment_binned	6.518932e-02	4	8.841362
	PNB_BINNED	3.576230e-01	3	3.229171

**Figure 5.**  
Chi-2 result sample.

After this, we calculated the probability of Type 2 Error ( $\beta$ ), which measures the probability of failing to detect a true association (false negative):

	Variable	Chi2_p-value	Type_2_Error
4	MMM_BIN	3.257408e-48	0.000000e+00
0	AGE_BINNED	2.316892e-20	1.479927e-13
5	avg_monthly_balance_binned	4.400546e-04	2.323627e-02
7	ENGAGEMENT_Binned	1.320943e-02	1.075909e-01
1	Active_Investment_binned	6.518932e-02	2.973168e-01
6	PNB_BINNED	3.576230e-01	6.566183e-01
8	DNR   ΔSTVFAR Binned	6.624974e-01	8.135999e-01

**Figure 6.**  
Chi2 type 2 error.

Based on the outputs, we only select features that have a Chi2 p-value lower than  $\alpha=0.05$  and, at the same time, a Type II Error lower than 0.2 (the test had a power higher than 80% to detect the effect, which indicates a low risk of false negatives). Thus, we are selecting features that are dependent on TARGET.

#### 6.1.5. T-Test

T-test, also known as Student's t-test, is a statistical test used to compare the means of two groups. It is often employed in hypothesis testing to determine whether a process or treatment influences the population of interest (Operation\_number and TARGET are dependent), or whether two groups are different from each other (Operation\_number and TARGET are independent).

After importing the test function of scipy.stats module, created two groups: group\_0, which represents all Operation\_number values for clients who stayed (TARGET = 0), and group\_1, which represents all Operation\_number values for clients who churned (TARGET = 1). This separation is to test whether the mean Operation\_number differs between the two groups.

Based on this code, we obtained a T-test p-value of 0.1540, which indicates that TARGET and Operation\_number are dependent. After successfully identifying all features that are dependent on TARGET, we will add them to a new data frame called DF\_Modeling. This data frame will be used in churn prediction after the encoding of the variables.

```
]: print(DF_Modeling.info())

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1500 entries, 0 to 1499
Data columns (total 24 columns):
 #   Column                                  Non-Null Count  Dtype
---  -
 0   CLIENT_ID                             1500 non-null   int64
 1   TARGET                                1500 non-null   int64
 2   Market                                1500 non-null   object
 3   deposit_type                           1500 non-null   object
 4   CLIENT_STATUS                          1500 non-null   object
 5   final_segment                          1500 non-null   object
 6   Pre_segment_final                      1500 non-null   object
 7   AGE_BINNED                             1500 non-null   object
 8   MMM_BIN                                1500 non-null   category
 9   ENGAGEMENT_Binned                     1500 non-null   category
10   avg_monthly_balance_binned             1500 non-null   category
11   TX_TPE                                 1500 non-null   int64
12   TX_BIATNET                             1500 non-null   int64
13   TX_PACK_PLATINUM                       1500 non-null   int64
14   TX_PACK_SILVER                         1500 non-null   int64
15   TX_PACK_BUSINESS                       1500 non-null   int64
16   TX_CREDIT_CONSO                        1500 non-null   int64
17   TX_CREDIT_IMMO                         1500 non-null   int64
18   TX_CREDIT_INV                          1500 non-null   int64
19   NUM_PACKAGES                           1500 non-null   int64
20   NUM_SERVICES                           1500 non-null   int64
21   account_age_bin                        1500 non-null   category
22   Inactivity_months_BIN                  1500 non-null   category
23   account_status                         1500 non-null   int64
dtypes: category(5), int64(13), object(6)
memory usage: 242.9+ KB
None
```

**Figure 7.**  
DF\_Modeling columns.

#### 6.1.6. Feature Encoding

Encoding is the final step in data preparation for churn prediction. Encoding refers to the process of converting categorical or textual data into a numerical format so that it can be used as input for algorithms to process. The reason for encoding is that most machine learning algorithms work with numbers and not with text or categorical variables.

Based on the output above (Figure 6: DF\_Modeling columns), we have 11 categorical variables where we have:

- 4 Nominal variables: 'Market', 'deposit\_type', 'CLIENT\_STATUS', 'AGE\_BINNED'
- 7 Ordinal variables: 'final\_segment', 'Pre\_segment\_final', 'MMM\_BIN', 'ENGAGEMENT\_Binned', 'account\_age\_bin', 'Inactivity\_months\_BIN', 'avg\_monthly\_balance\_binned'

For nominal features, we used label encoding, which converts categorical columns into numerical ones, instead of one-hot encoding, and that's due to its computational efficiency and compatibility with tree-based algorithms. Since tree-based models make splits based on feature thresholds rather than Euclidean distances, they do not misinterpret the integer labels assigned by label encoding as ordinal relationships. Additionally, with 22 features, one-hot encoding could unnecessarily expand the feature space, leading to increased memory usage and training time without significant performance benefits for ensemble methods. For ordinal features, we use ordinal encoding, which converts categories to ordered integers based on their specified order.

After the application of the encoding process using the sklearn.preprocessing library of Python and removing CLIENT\_ID from DF\_Modeling, the following sections will be devoted to the modeling, evaluation, and comparison of machine learning model performances.

#### 6.2. Model Development

In this research, we will use tree-based algorithms to predict BANK's customer churn by training and evaluating three popular models: Random Forest, XGBoost, and LightGBM. Then, we will compare their performance using precision, recall, F1-score, and accuracy metrics.

##### 6.2.1. Model Optimization with Hyperparameter Tuning

Hyperparameter tuning is the process of selecting the optimal values for a machine learning model's hyperparameters. Hyperparameters are configuration settings that control the learning process of the model. The quality of a predictive model mostly depends on the configuration of its hyperparameters, but it is often difficult to know how these hyperparameters interact with each other to affect the results of the model. To determine accuracy and make a comparison between the models, it is always better to perform this tuning when the hyperparameters are optimized [9]. In this part, we will present the results of the tested algorithms and the chosen model. We are now in the evaluation step of CRISP-DM.



### 6.2.2. Evaluation and Results

**Table 1.**

Evaluation table.

Model	Precision	recall	F1-score	Accuracy	AUC-ROC
RF	0.82	0.83	0.83	0.83	0.85
Hyper tuned RF	0.82	0.84	0.83	0.863	0.94
XGBoost	0.81	0.82	0.81	0.85	0.941
Hyper tuned XGBoost	0.8	0.85	0.82	0.85	0.945
LightGBM	0.82	0.89	0.85	0.876	0.945
Hyper-tuned LightGBM	0.8	0.89	0.84	0.866	0.951

Based on the results shown above, both LightGBM and hyper-tuned LightGBM achieved better performance compared to the other models. Globally, hyper-tuned LightGBM shows a near-perfect performance (AUC-ROC = 0.951). However, the default LightGBM model shows slightly higher precision and accuracy compared to the hyper-tuned version. Both these models show excellent results overall; their performances are quite similar. Therefore, to choose the best model between LightGBM and the hyper-tuned LightGBM, we will perform cross-validation to further evaluate their performance.

### 6.2.3. K-Fold Cross Validation

Cross-validation (CV) is one of the most used techniques in measuring the optimal model in machine learning. It is typically used where ML tasks involve prediction, and one seeks to evaluate how accurately a predictive model will perform in training. The goal of CV is to evaluate the model's ability to classify new data that was not employed in estimating the model, to flag problems like selection bias or overfitting. In addition, CV offers in-depth insights into how a trained model will generalize to unseen datasets or real-world problems. Among the various types of CV, k-fold cross-validation is the most common among machine learning practitioners [11].

In K-Fold Cross Validation, we split the dataset into k subsets, known as folds. Then, we perform training on all but one of the subsets, leaving one subset for evaluation of the trained model. In this method, we iterate k times, with a different subset reserved for testing purposes each time. In this project, we chose K=5 since it is the most commonly used value of k [11], then we calculated the average values of precision, recall and F1 score, we found:

**Table 2.**

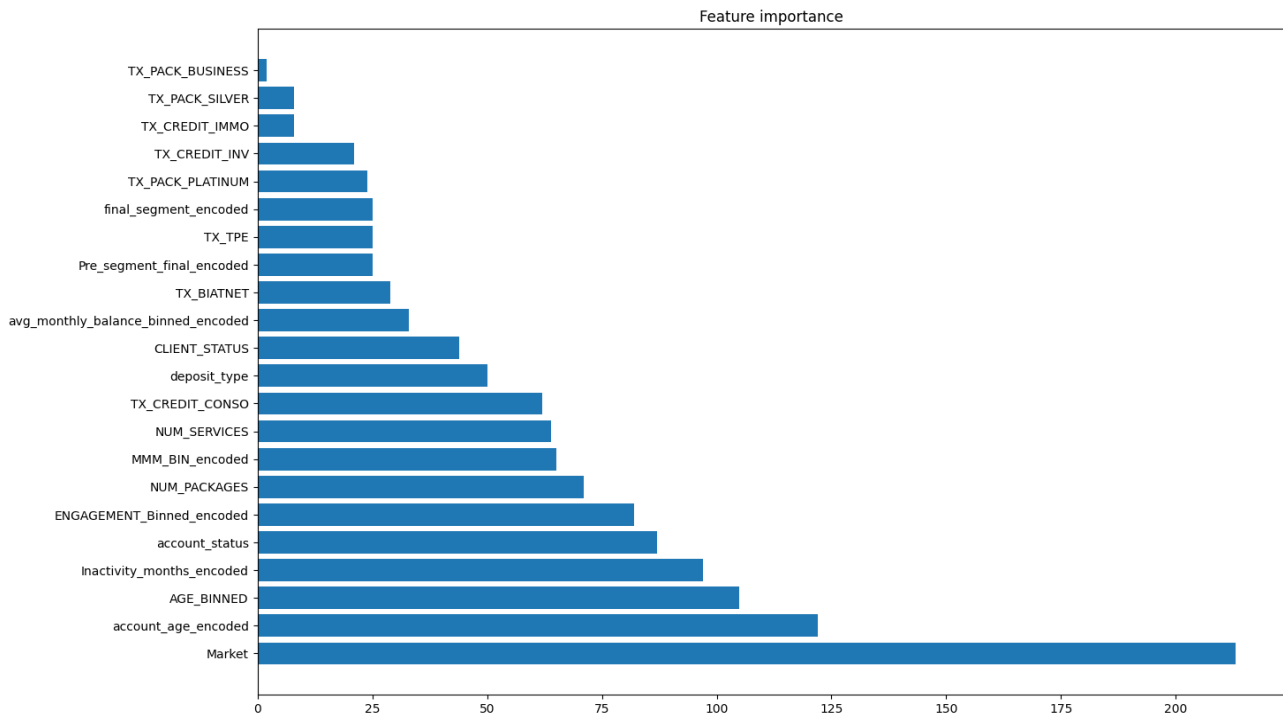
Cross-validation table.

Model	Average Precision	Average Recall	Average F1-score
LightGBM	0.836	0.847	0.841
Hyper-tuned LightGBM	0.832	0.903	0.866

The Cross Validation table, Hyper-tuned LightGBM shows higher recall and F1-score, with a slightly lower precision (-0.004), which can be considered negligible. Based on this, our best model is Hyper-tuned LightGBM, as it identifies 90% of all churners, correctly predicts 83% of actual churners out of all predicted churners, and has a mean precision and recall of 0.866, indicating a high balance between these metrics.

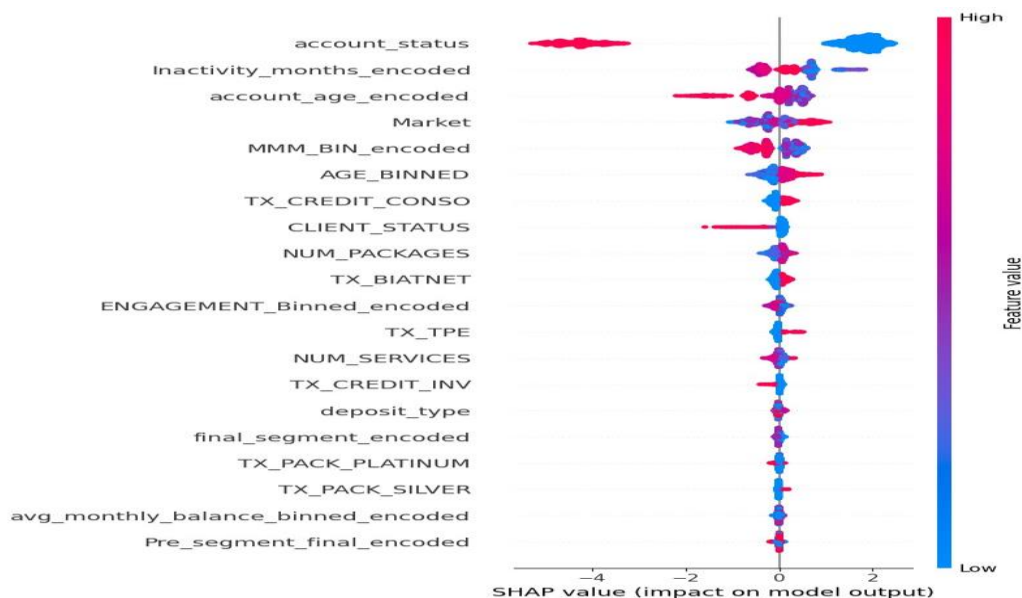
### 6.3. Feature Importance

Feature importance scores are used to determine the relative importance of each feature in a dataset when building a predictive model. They identify which features in a dataset contribute most to the final prediction and which features are less important.



**Figure 8.**  
Feature importance.

This plot provides a general and comprehensive view of the features most utilized by the Hyper Tuned LightGBM model for prediction. It clearly indicates that the Market is the most influential feature, suggesting that the client's employment or business category is the strongest predictor of whether they are likely to churn. In second place is account age\_encoded, which reflects the client's account age, and in third place is the client's age. The least influential features are the bank's products and services. To better understand how these features influence churn prediction, we will apply SHAP. SHAP, which stands for "SHapley Additive exPlanations," is used to explain the output of machine learning models. It is based on Shapley values, which utilize game theory to assign credit for a model's prediction to each feature or feature value. Unlike feature importance, which answers the question "What does the model use most?", SHAP addresses the question "What actually drives predictions up and down?"



**Figure 9.**  
SHAP values.

The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature, and on the x-axis by the Shapley value. The color represents the value of the feature from low to high. Overlapping points are jittered in the y-axis direction, so we get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance.

More specifically, red dots indicate high feature values, blue dots indicate low feature values, and *negative values (left of 0) decrease the churn probability and positive values (right of 0) increase the churn probability.*

**Table 3.**  
SHAP Feature Interpretation Summary

Feature	Description	Interpretation
Account_status	Client account status (1: Active / 0: Inactive)	Blue dots (low feature values) clustered on the right (positive) side indicates customers with inactive accounts are MORE likely to churn. Red dots (HIGH feature values) clustered on the left (negative) side indicate active customers are less likely to churn.
Account_inactivity_months	Number of months the client's account stayed inactive	Shows variant results with moderate impact. Higher inactivity months increase the churn probability.
Account_age_encoded	indicates the client's account age	We see red dots clustered on the left side, which means clients with older accounts are less likely to churn
Market	Client's employment or business category	shows some variation, we see more red dots (High values), more on the positive side, this is interrupted based on the encoding where we assigned, which means Salary (3) and Startups (4) segments tend to increase churn probability, Corporates (0) on the other hand decrease churn probability
MMM_BIN_encoded	the client's average monthly movement.	We see a cluster of red dots (high values) on the negative side, which indicates that high average monthly movements decrease churn prediction.
AGE_BINNED	the client's age.	The plot shows a high concentration of blue dots on the negative side, which indicates that younger clients tend to churn more compared to older ones.

To summarize, our hyper-tuned LightGBM predictive model demonstrated excellent performance in predicting customer churn at BANK. It also successfully highlights the features that lead to customer churn, which is beneficial and important to the bank as it indicates at-risk clients. This also confirms our hypothesis H1 that a predictive model incorporating socio-demographic and behavioral characteristics can accurately predict customer churn.

## 7. Data Segmentation

Banks can offer more personalized products and services by using segmentation solutions. By gaining a deeper understanding of client characteristics, marketers can choose the appropriate promotional content to deliver, select the right marketing channels for the target market, identify new and profitable market sectors, and introduce new products and services.

As mentioned in the literature review, we will implement a two-tiered hierarchical segmentation approach for the data provided by BANK. In this section, we will apply these techniques using k-means clustering at both layers and demonstrate how this approach functions as a predictive intelligence model.

### 7.1. Data Preparation

Just like churn prediction models, K-means also requires encoding. To prepare the data, we will start by dropping all churned clients who have already left the bank because customer segmentation only focuses on existing clients. This will leave us with 900 clients. The second step is to drop all attrition-related columns, which will leave us with a total of 66 columns.

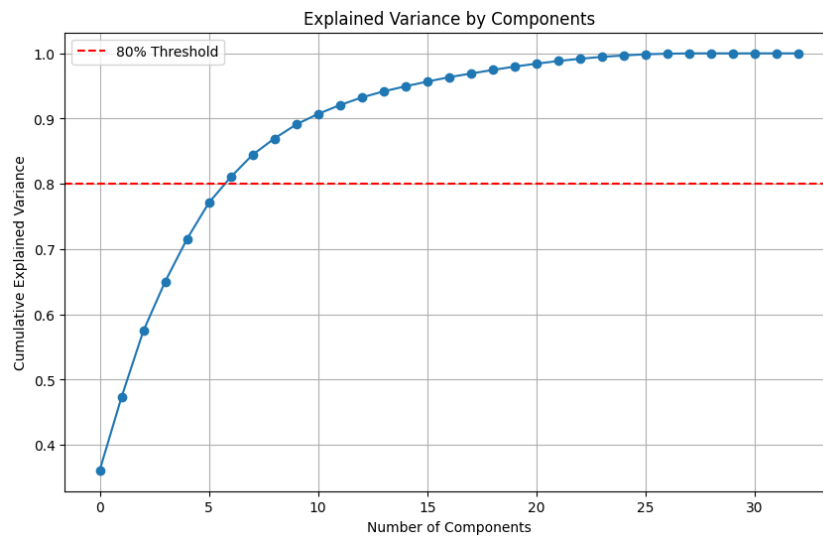
To further reduce the number of columns, we will drop the highly correlated ones. The results showed that MONTHS\_INACTIVE and account\_status are the only columns with high correlation between each other (0.872124). Since MONTHS\_INACTIVE provides more granularity and information, we will drop account\_status. We now have a total of 65 columns.

### 7.2. Segmentation of Bank Dataset

#### 7.2.1. Customer Segmentation

Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction. Based on this article by Azad [4] PCA can analyze data to identify patterns and reduce the dimensions of the dataset with minimal loss of information. High dimensionality means that the dataset has many features, which could lead to overfitting.

By reducing the dimensions of datasets, PCA provides an effective and efficient method for data description and visualization. PCA produces a low-dimensional representation of the feature space by finding a sequence of linear combinations of the features that have maximal variance and are mutually uncorrelated. To do so, we used a **Scree plot** to find the optimal number of PCs with a common threshold of 80%.



**Figure 10.**  
Scree plot.

This plot helps determine how many principal components to retain. The red line at 0.8 is the common threshold; its purpose is to identify how many components are needed to reach 80% of explained variance (information). The intersection between the blue line and the red line indicates the minimum number of PCs required to retain at least 80% of the information (variance) in our data.

The plot shows an intersection at 6 principal components, which means the first 6 principal components together explain about 80% of the total variance in the data. Therefore, if we use only 6 principal components, it will cover 80% of the information in 33 variables, and it is sufficient to produce more reliable clustering results.

### 7.3. Selecting The Number of Clusters $K$

To select the number of clusters we calculated the *Silhouette Score* for  $k \in [2, 12]$ :

```
k=2: Score = 0.271
k=3: Score = 0.262
k=4: Score = 0.222
k=5: Score = 0.224
k=6: Score = 0.205
k=7: Score = 0.207
k=8: Score = 0.187
k=9: Score = 0.192
k=10: Score = 0.190
```

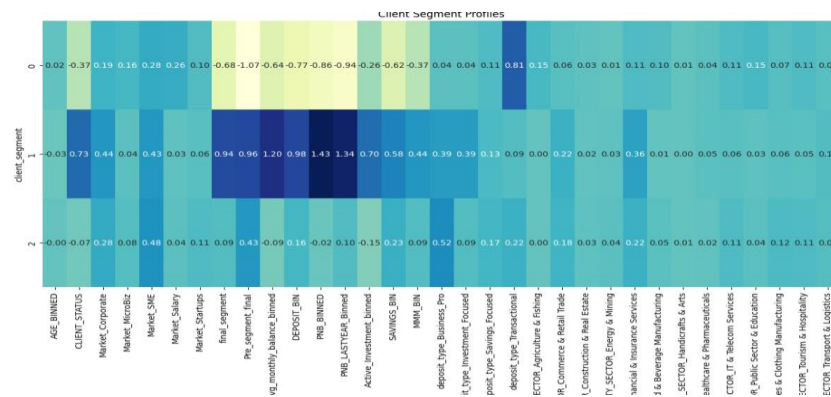
**Figure 11.**  
Silhouette Score.

Based on these results, we can clearly see that  $k=2$  and  $k=3$  have the best scores. We chose  $k=3$ . This choice is not random; k-means with  $k=3$  showed balanced clusters and a more refined segmentation.

```
client_segment
0    330
1    203
2    367
Name: count, dtype: int64
```

**Figure 12.**  
Clusters distribution.

To identify and profile these clusters, we created a heatmap of the original variables (instead of the PCA components) to better understand how each cluster differs in terms of the actual features.



**Figure 13.**  
First tier heat map

Based on this heatmap we distinguished what each cluster represents:

- Client\_segment 0: Lower financial values indicating less engaged and less profitable clients. Thus, this segment's profile is mass-market, transactional clients.
- Client\_segment 1: Higher financial values indicating top engagement and profitable clients. Thus, this segment's profile is affluent/wealthy.
- Client\_segment 2: Moderate financial values indicating Average and basic clients.

#### 7.4. Business, Mid-Market Clients

##### 7.4.1. Marketing segmentation

Marketing segmentation for Mass-market, transactional clients:

- Optimal number of PCA components (80% variance): 19
- Optimal number of marketing clusters: 2
- \* Marketing Segment 0: 113 customers (34.2%)
- \* Marketing Segment 1: 217 customers (65.8%)
- Marketing segmentation for Affluent, wealthy clients:
- Optimal number of PCA components (80% variance): 18
- Optimal number of marketing clusters: 2
- \* Marketing Segment 0: 137 customers (67.5%)
- \* Marketing Segment 1: 66 customers (32.5%)
- Marketing segmentation for Business, Mid-market clients:
- Optimal number of PCA components (80% variance): 19
- Optimal number of marketing clusters: 2
- \* Marketing Segment 0: 260 customers (70.8%)
- \* Marketing Segment 1: 107 customers (29.2%)

These results highlight how each customer segment is composed of two clusters that further define the customer base of BANK. To better understand these results, we selected the top five distinguishing features that show what makes this segment unique compared to the overall population.

##### 7.4.2. Mass-market, transactional clients: Marketing Segment 0

###### Top distinguishing features:

- Amount\_LC\_Binned: 1.02 (higher than segment average of -0.03)
- operation\_family\_Credit: 0.91 (higher than segment average of 0.33)
- operation\_description\_deposit: 0.88 (higher than segment average of 0.32)
- operation\_family\_Debit: 0.09 (lower than segment average of 0.62)
- operation\_description\_withdrawal: 0.06 (lower than segment average of 0.59)

**Figure 14.**  
Mass market: Marketing Segment 0.

This output demonstrates how this sub-segment is made up of customers who are more financially active: they deposit more (operation\_description\_deposit 0.88), make larger transactions (Amount\_LC\_Binned 1.02), and perform more credit-type operations (operation\_family\_Credit 0.91). They are highly engaged with BANK's services, show potential for

increased financial activity, and exhibit lower rates of account withdrawals and spending. This reveals that this sub-segment represents Mass Market Growing Savers.

#### 7.4.3. Mass-Market, Transactional Clients: Marketing Segment 1

Top distinguishing features:

- Amount\_LC\_Binned: -0.58 (lower than segment average of -0.03)
- operation\_familly\_Credit: 0.03 (lower than segment average of 0.33)
- operation\_description\_deposit: 0.03 (lower than segment average of 0.32)
- operation\_familly\_Debit: 0.89 (higher than segment average of 0.62)
- operation\_description\_withdrawal: 0.87 (higher than segment average of 0.59)

**Figure 15.**

Mass market: Marketing Segment 1.

This output demonstrates how this sub-segment has lower variables compared to the Marketing segment 0. It is less financially active, with more spending and withdrawals (operation\_description\_withdrawal 0.87) and less saving or depositing (operation\_description\_deposit 0.03). Likely paycheck-to-paycheck or transactional users. This reveals that this sub-segment represents Mass Market Everyday Spenders.

#### 7.4.4. Affluent, Wealthy Clients: Marketing Segment 0

Top distinguishing features:

- Amount\_LC\_Binned: -0.43 (lower than segment average of 0.15)
- operation\_description\_withdrawal: 0.93 (higher than segment average of 0.66)
- operation\_familly\_Credit: 0.02 (lower than segment average of 0.29)
- operation\_familly\_Debit: 0.95 (higher than segment average of 0.69)
- operation\_description\_deposit: 0.02 (lower than segment average of 0.28)

**Figure 16.**

Wealthy clients: Marketing Segment 0.

This output demonstrates how this sub-segment represents affluent customers who are frequent spenders. They withdraw (operation\_description\_withdrawal 0.93) and spend more (operation\_family\_Debit 0.95), but deposit or save less (operation\_description\_deposit 0.02). They may use their accounts mainly for everyday transactions and cash access, not for growing wealth. Thus, this sub-segment represents Affluent Spenders.

#### 7.4.5. Affluent, wealthy clients: Marketing Segment 1

Top distinguishing features:

- Amount\_LC\_Binned: 1.35 (higher than segment average of 0.15)
- operation\_description\_withdrawal: 0.11 (lower than segment average of 0.66)
- operation\_familly\_Credit: 0.83 (higher than segment average of 0.29)
- operation\_familly\_Debit: 0.15 (lower than segment average of 0.69)
- operation\_description\_deposit: 0.80 (higher than segment average of 0.28)

**Figure 17.**

Wealthy clients: Marketing Segment 1.

This output demonstrates how this sub-segment represents affluent high-value clients who deposit more (operation\_description\_deposit 0.80), transact in larger amounts (Amount\_LC\_Binned 1.35), and use credit products (operation\_family\_Credit 0.83). Compared to segment 1, these clients rarely withdraw or spend. They are more likely to be Affluent Accumulators.

#### 7.4.6. Mid-market clients: Marketing Segment 0

Top distinguishing features:

- Amount\_LC\_Binned: -0.48 (lower than segment average of -0.05)
- operation\_familly\_Credit: 0.02 (lower than segment average of 0.29)
- operation\_description\_deposit: 0.02 (lower than segment average of 0.28)
- operation\_familly\_Debit: 0.87 (higher than segment average of 0.63)
- operation\_description\_withdrawal: 0.84 (higher than segment average of 0.60)

**Figure 18.**

Mid-market clients: Marketing Segment 0.

This output demonstrates how this mid-market segment is transactionally focused and cash-flow oriented. They make

frequent withdrawals (operation\_description withdrawal: 0.84) and debit transactions (operation\_family\_Debit: 0.87) but rarely deposit or use credit. Therefore, this segment represents a specific customer profile characterized by high transaction activity and limited credit usage.

#### 7.4.7. Mid-Market Operational Spenders

##### 7.4.7.1. Mid-market clients: Marketing Segment 1

Top distinguishing features:

- Amount\_LC\_Binned: 0.98 (higher than segment average of -0.05)
- operation\_family\_Credit: 0.93 (higher than segment average of 0.29)
- operation\_description\_deposit: 0.91 (higher than segment average of 0.28)
- operation\_family\_Debit: 0.05 (lower than segment average of 0.63)
- operation\_description\_withdrawal: 0.03 (lower than segment average of 0.60)

**Figure 19.**

Mid-market clients: Marketing Segment 1.

This output demonstrates that this segment of mid-market clients is accumulating and leveraging funds; they deposit more (operation\_description\_deposit 0.91), make larger transactions (Amount\_LC\_Binned 0.98), use credit products (operation\_family\_Credit 0.93), and rarely withdraw (operation\_description\_withdrawal 0.03). Therefore, this segment presents Mid-Market Business Builders.

#### 7.5. Link To Marketing Strategies Funding

By devising these newly identified customer segments, the bank can now implement targeted cross-selling and up-selling strategies customized to each client category. For instance, the bank can set strategic plans for Mass Market Growing Savers by promoting premium savings accounts with better interest rates, promoting Digital Banking (BANK-NET, MYBANK), offering Micro-Investing accounts, etc.

Based on these findings, we conclude that the hierarchical K-means clustering methodology effectively identified distinct customer segments with specific needs and preferences, enabling banks to develop targeted cross-selling and up-selling programs. This successful segmentation validates our initial hypothesis that a multilayered approach can improve customer understanding and inform more effective marketing strategies.

## 8. Association Rule Recommendation System

Association rule mining is a data-mining technique used to identify associations between various item combinations. The concept of association rules originated from market basket analysis, which aims to identify frequent item sets in supermarket shopping. The main idea is to find products that are frequently purchased together to enhance marketing strategies.

#### 8.1. Application of A priori

Following this logic, the bank can now target growing mass-market savers who have a "Future project contract" and offer them deposit accounts and online services if they do not already have them.

- *Mass-market Everyday Spenders clients*

Top 5 association rules:

```
TX_DEBIT & TX_Investment → TX_Future_Project
  Support: 0.060, Confidence: 0.481, Lift: 2.750
TX_Future_Project → TX_DEBIT & TX_Investment
  Support: 0.060, Confidence: 0.342, Lift: 2.750
TX_MULTIVIR → TX_Saving_Account & TX_BIATNET
  Support: 0.055, Confidence: 0.343, Lift: 2.325
TX_Saving_Account & TX_BIATNET → TX_MULTIVIR
  Support: 0.055, Confidence: 0.375, Lift: 2.325
TX_Future_Project → TX_BIAT_DD & TX_Investment
  Support: 0.097, Confidence: 0.553, Lift: 2.221
```

**Figure 20.**

Top 5 association rules for segment 2.



- *Affluent Spenders clients:*

Top 5 association rules:

```
TX_PACK_BUSINESS → TX_PACK_PLATINUM & TX_CREDIT_IMMO
  Support: 0.051, Confidence: 0.318, Lift: 3.633
TX_PACK_PLATINUM & TX_CREDIT_IMMO → TX_PACK_BUSINESS
  Support: 0.051, Confidence: 0.583, Lift: 3.633
TX_BIATNET → TX_MESSAGIS & TX_TPE
  Support: 0.051, Confidence: 0.167, Lift: 3.262
TX_MESSAGIS & TX_TPE → TX_BIATNET
  Support: 0.051, Confidence: 1.000, Lift: 3.262
TX_CREDIT_IMMO & TX_PACK_BUSINESS → TX_PACK_PLATINUM
  Support: 0.051, Confidence: 0.778, Lift: 3.229
```

**Figure 21.**

Top 5 association rules for segment 3.

- *Affluent Accumulators clients:*

Top 5 association rules:

```
TX_PACK_ELITE_PRO → TX_MESSAGIS & TX_PACK_FIRST
  Support: 0.030, Confidence: 0.400, Lift: 13.200
TX_PACK_FIRST → TX_MESSAGIS & TX_PACK_ELITE_PRO
  Support: 0.030, Confidence: 0.400, Lift: 13.200
TX_MESSAGIS & TX_PACK_ELITE_PRO → TX_PACK_FIRST
  Support: 0.030, Confidence: 1.000, Lift: 13.200
TX_MESSAGIS & TX_PACK_FIRST → TX_PACK_ELITE_PRO
  Support: 0.030, Confidence: 1.000, Lift: 13.200
TX_MULTIVIR & TX_PACK_ELITE → TX_PACK_TOUNESNA
  Support: 0.030, Confidence: 1.000, Lift: 11.000
```

**Figure 22.**

Top 5 association rules for segment 4.

- *Mid-Market Operational Spenders:*

Top 5 association rules:

```
TX_Future_Project → TX_CREDIT_CONSO & TX_FAMILIA
  Support: 0.065, Confidence: 0.270, Lift: 2.192
TX_CREDIT_CONSO & TX_FAMILIA → TX_Future_Project
  Support: 0.065, Confidence: 0.531, Lift: 2.192
TX_CREDIT_CONSO → TX_Future_Project & TX_FAMILIA
  Support: 0.065, Confidence: 0.258, Lift: 1.970
TX_Future_Project & TX_FAMILIA → TX_CREDIT_CONSO
  Support: 0.065, Confidence: 0.500, Lift: 1.970
TX_BIAT_DD & TX_PACK_SAFIR_PRO → TX_BIATNET
  Support: 0.065, Confidence: 0.850, Lift: 1.873
```

**Figure 23.**

Top 5 association rules for segment 5.



- *Mid-Market Business Builders:*

Top 5 association rules:

TX\_PACK\_PLATINUM & TX\_DEBIT → TX\_PACK\_SAFIR

Support: 0.065, Confidence: 0.500, Lift: 2.816

TX\_PACK\_SAFIR → TX\_PACK\_PLATINUM & TX\_DEBIT

Support: 0.065, Confidence: 0.368, Lift: 2.816

TX\_PACK\_SILVER & TX\_Investment → TX\_MULTIVIR

Support: 0.065, Confidence: 0.636, Lift: 2.724

TX\_MULTIVIR → TX\_PACK\_SILVER & TX\_Investment

Support: 0.065, Confidence: 0.280, Lift: 2.724

TX\_BIATNET & TX\_FAMILIA → TX\_PACK\_FIRST

Support: 0.056, Confidence: 0.353, Lift: 2.697

**Figure 24.**

Top 5 association rules for segment 6.

These association rules represent a significant advantage for BANK by offering a multitude of benefits across various business functions. These association rules allow the creation of customized offers that target individual customer needs and preferences, thereby increasing the likelihood of successful cross-selling and up-selling. Additionally, the discovered product associations provide valuable insights into customer purchasing behavior, revealing previously unknown relationships between BANK's products. This knowledge empowers the bank to optimize product bundling, refine marketing campaigns, and proactively identify emerging trends in customer demand, leading to improved customer satisfaction and better financial performance.

To summarize, an association rules-based recommendation system can be of immense help to the bank in solving business problems by finding hidden patterns within the database and providing the bank with insightful information to utilize in cross-selling and up-selling their products, thereby confirming our hypothesis.

## 9. Conclusion

This article demonstrates the significant potential of predictive intelligence in understanding and managing banking customer behavior, ultimately driving both financial and marketing growth. By leveraging tree-based ensemble algorithms and selecting hyper-tuned LightGBM as the best model, we accurately predicted customer churn and identified its key drivers through feature importance and SHAP values. This enables the bank to address these drivers' features and avoid potential losses. The study extends beyond mere churn prediction by implementing a two-tiered segmentation strategy. First, customer segmentation identified three distinct customer categories. Second, marketing segmentation aligns marketing efforts with specific customer needs and preferences. These segments, combined with churn prediction insights, provide a foundation for highly effective marketing campaigns and customer retention strategies. Finally, a rule-based recommendation system built using the Apriori algorithm revealed valuable product associations. These associations enable the bank to strategically bundle products and services, cross-sell and up-sell effectively, and create targeted campaigns that anticipate customer needs. In conclusion, this study offers a strategy for the bank to transform its approach to customer management. The bank can:

- Reduce churn: By addressing the identified churn key drivers and retaining valuable customers.
- Optimize marketing strategies: By targeting specific customer segments with relevant and customized offers.
- Increase revenue: By promoting and creating new product bundles and by cross-selling and up-selling opportunities identified by the association rules.
- Enhance customer satisfaction: By delivering personalized experiences and anticipating customer needs through data-driven insights.
- The implementation of these data-driven strategies will position the bank for sustained financial growth and a stronger competitive advantage in an increasingly dynamic market.

## References

- [1] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," *Artificial Intelligence Review*, vol. 33, no. 4, pp. 275-306, 2010. <https://doi.org/10.1007/s10462-010-9156-z>
- [2] Y. Elyusufi and M. Ait Kbir, "Churn prediction analysis by combining machine learning algorithms and best features exploration," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, pp. 615-622, 2022. <https://doi.org/10.14569/IJACSA.2022.0130773>

- [3] A. Essayem, S. Gormus, and M. Guven, "The GCC's regional roller coaster: Do regional factors affect stock market dynamics in the GCC Region? Evidence from non-parametric quantile regression," *Borsa Istanbul Review*, vol. 23, no. 2, pp. 473-494, 2023. <https://doi.org/10.1016/j.bir.2022.11.018>
- [4] A. Azad, "Incorporating k-means, hierarchical clustering and pca in customer segmentation," *Journal of City and Development*, vol. 3, no. 1, pp. 12-30, 2021.
- [5] O. Raiter, "Segmentation of bank consumers for artificial intelligence marketing," *International Journal of Contemporary Financial Issues*, vol. 1, no. 1, pp. 39–54, 2021. <https://doi.org/10.17613/q0h8-m266>
- [6] E. J. Chéron, R. McTavish, and J. Perrien, "Segmentation of bank commercial markets," *International Journal of Bank Marketing*, vol. 7, no. 6, pp. 25-30, 1989. <https://doi.org/10.1108/eum0000000001458>
- [7] K. Gopalakrishnan, "Customer segmentation using k-means clustering for targeted marketing in banking," *International Journal of Artificial Intelligence & Machine Learning*, vol. 3, no. 2, pp. 89–94, 2024. <https://doi.org/10.5281/zenodo.13627403>
- [8] N. Z. B. Fadhel, "From the transactional to the relational in the customer-supplier relationship: Case of the Tunisian banks," *International Journal of Marketing Studies*, vol. 17, no. 1, pp. 44–56, 2025. <https://doi.org/10.5539/ijms.v17n1p44>
- [9] G. Mansingh, K.-M. Osei-Bryson, L. Rao, and M. McNaughton, "Market basket analysis in the financial sector: A customer centric approach," in *Proceedings of the 2016 Pre-ICIS SIGDSA/I-FIP WG8.3 Symposium: Innovations in Data Analytics*, 2016.
- [10] D. B. Desai and A. Kaiwade, "Application of Apriori algorithm for analyzing customer behavior to improve deposits in banks," in *Proceedings of the International Conference on Advances in Computer Technology and Management (ICACTM)* (pp. 188–191). Novateur Publications, 2018.
- [11] M. R. Hossain and D. Timmer, "Machine learning model optimization with hyper-parameter tuning approach," *Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence*, vol. 21, no. 2, pp. 7–13, 2021.
- [12] Ž. Vujović, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599-606, 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0120670>