NOVA IMS Information Management School

MDDM

Master Degree Program in

Data-Driven Marketing

AI-driven Customer Analytics

Implementation of Machine Learning Solutions into bank's CRM

Julia Marianna Trzos

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data-Driven Marketing

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

AI – DRIVEN CUSTOMER ANALYTICS – IMPLEMENTATION OF MACHINE LEARNING SOLUTIONS INTO BANK'S CRM

By

Julia Marianna Trzos

Internship Report presented as partial requirement for obtaining the Master's degree in Data-Driven Marketing, with a specialization in Digital Marketing and Analytics

Supervisor/Orientador(a): Miguel de Castro Simões Ferreira Neto

Or Co-Supervisors/Co-Orientadores*:* João Bruno Morais de Sousa Jardim

November 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 23.11.2023

ACKNOWLEDGEMENTS

I would like to express my gratitude for Asseco PST and its team for accepting my internship and allowing me to develop this project.

I would like to thank Bruno Jardim for his excellent work and support as my co-supervisor.

Finally, I would like to thank professor Miguel Neto for accepting to be my supervisor.

ABSTRACT

In the ever-evolving landscape of modern banking, the incorporation of emerging technologies, specifically Artificial Intelligence and machine learning, holds great significance in order to remain relevant and efficiently address customer needs, having a potential for significant enhancements in customer relationship management practises within the banking industry. The objective of this project, conducted in collaboration with the Asseco PST Data & Analytics team, is to improve their CRM solution by incorporating a comprehensive machine learning framework. This involves utilising machine learning techniques to segment clients, with the goal of optimising customer relationship management (CRM) and providing data-driven campaigns for their bank clients. The project aims to develop a clustering-based Recommendation System that delivers customised product recommendations. Furthermore, the project presents a deployment demonstration involving the creation of apps aimed at achieving a scalable solution for clustering and predictive modelling, hence facilitating the implementation process for new clients. Additionally, this project intends to establish itself as an important component within Asseco PST's comprehensive offering. The integration of this work within their pre-existing CRM development offer serves to underscore its importance and possible influence within the constantly developing world of present-day banking.

KEYWORDS

Business Intelligence, Machine Learning, Banking, Customer Relationship Management, Artificial Intelligence, Customer Segmentation

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
BI	Business Intelligence
CRM	Customer Relationship Management
DT	Decision Tree
DW	Data Warehouse
ETL	Extract, Transform, Load
HR	Human Resources
KNN	K-nearest Neighbours
LTV	Lifetime value
ML	Machine Learning
RF	Random Forest
RFM	Recency, Frequency, Monetary value
RS	Recommender System
SVM	Support-vector Machine
WCSS	Within-cluster Sum of Squares

1. INTRODUCTION

Modern banks operate in a highly dynamic and competitive market with a necessity for control and risk management, where also the customer demands are constantly changing. Business Intelligence solutions for banks need to consider the characteristics of the sector to support better management and decision-making process, covering many banks areas, including Customer Relationship Management (Ubiparipovi & Durkovic, 2011). Customer Relationship Management (CRM) involves the planned deployment of targeted technological solutions and integrated methodologies to comprehensively manage and enhance customer relationships, thereby generating value for shareholders and cultivating customer loyalty while improving organisational profitability (Payne, 2005). Big Data and Artificial Intelligence (AI) are useful in a variety of functional areas of marketing, providing effective assistance in decision-making and reducing the risk of poor marketing actions (D'Arco et al, 2019). Machine learning, as an application of AI, has enormous capacity for lowering product and service costs, speeding up business processes, and providing better customer service. It has been identified as one of the most important application areas in this era of unprecedented technological development, and its adoption is gaining traction across almost all industries (Lee & Shin, 2020).

According to Hlavac and Stefanovic (2020), modern BI is based on the cooperation of BI applications with machine learning processing. Banks have quickly adapted to the digitalization of its services, which resulted in availability of Big Data and the need of data processing, fuelling the adoption of Artificial Intelligence, which has a capacity to transform commercial banks to intelligence.

McKinsey & Company (2021) report describes the need of traditional banks to become "AI – first", adopting AI technologies at the core of new value propositions and customer experiences to remain relevant and competitive. According to the authors, the adoption of AI by banks can help boost revenues, lower costs, and uncover new and previously unrealized opportunities, as well as bring at-scale personalization, omnichannel experiences, and innovation cycles.

The implementation of AI and machine learning by banks has the potential to raise their competitive edge, boost their efficiency, and maximize their profitability. These cutting-edge technologies may be implemented in a variety of contexts, including customer relationship management, risk management, fraud detection, and identity verification, in order to bolster the security of services and prevent fraudulent actions. Nonetheless, due to the novelty of the field and the recent transition from experimental to real-world implementations, there are several chances to extend product offers and understand customer preferences (Antal-Vaida, 2022).

The activities of commercial banks generate large amounts of data, but the banking system was designed with the conventional financial industry in mind, without taking consumer preferences and lifestyle into account (Yu et al., 2018). However, commercial banks are required to promote and offer their products and services directly to consumers. Due to the homogeneity of the goods offered by different commercial banks, customer relationship management can be challenging; effective customer targeting is therefore an important component of the management of client relationships in commercial banking (Königstorfer and Thalmann, 2020). With data mining technology, massive volumes of data may be leveraged to their utmost potential for decision-making and to fuel CRM efforts to increase the effectiveness and efficiency of recruiting and maintaining key clients (Yu et al., 2018).

One of the applications of AI for analytics is Machine Learning, where the model can learn from past experiences or the patterns within the data (Hlavac & Stefanovic, 2020). The use of machine learning for CRM can support the application in expanding customer information to develop customer-centric business solutions (Amnur, 2017). One of the use cases of AI in banking can be found within the personalization of the customer journey, allowing to improve the efficiency of marketing actions and customer satisfaction (LI, 2021; Chen, 2020). However, even though AI can have a revolutionary potential in marketing, many managers do not fully understand the full benefits it can provide or how to adopt it (Campbell et al, 2019).

McKinsey & Company report (2021) expresses the need of banks to become customer-centric through integration of personalization elements across all customer touchpoints to deliver a superior experience and results. This can be accomplished by employing AI to gain a greater comprehension of each customer's context, behaviour, needs, and preferences, enabling the bank to provide customized services.

One of the ways of understanding more in-depth different types of bank's customers is through segmentation (Fares et al., 2022). Through deeper understanding of client profiles, banks can offer more personalized products and services, and inform the design and targeting of marketing strategies and campaign channels, as well as identification of new or previously unknown market sectors (Raiter, 2021). According to Zakrzewska and Murleski (2005), clustering is the methodology that is the most applied in the area of market segmentation for knowledge-based marketing. An unsupervised ML method, clustering, allows for more effective targeting of the most valuable customers, resulting in increased revenue and decrease in the marketing costs (Djurisic et al, 2020). According to Djurisic et al. (2020), studies have shown that customer segmentation plays one of the key roles in the banking

sector, allowing for more effective targeting of the most valuable customers, resulting in the increased revenue and decrease in the marketing costs.

From a business standpoint, the absence of an intelligent decision support system is a problem faced by several financial institutions (Cheng et al., 2009). AI allows for better segmentation, targeting and positioning of bank products and services (Fares et al., 2022). Product and offer recommendations can be achieved through the use of recommender systems, which utilize AI methods (Portugal et al., 2018). Due to its potential business value, the application of recommender systems began to extend to other sectors, including banking. Recommender systems can provide substantial advantages for financial services by supporting the sales representatives or through automating the decision-making process for customers (Zibriczky, 2016).

McKinsey & Company (2021) report emphasizes the significance of deploying AI solutions at scale, where the models are built to be re-used throughout the organization, as well as selecting the use cases that have the most influence on customer experience and provide the most value to the bank.

1.1. COMPANY OVERVIEW

Asseco PST is an information technology company specializing in the development of core banking software, as well as differentiated technological solutions and knowledge for bank clients. Asseco PST is a part of the Asseco Group, one of the largest software suppliers in Europe, with presence in 60 countries. Their systems and solutions are used by banks, energy and telecommunication companies, public sector, and health care.

Asseco PST operates mainly in Portuguese-speaking countries, present in eight markets across three continents. Its main offices are located in Madeira and Lisbon, with additional branches in Angola and Mozambique. The majority of the clients are based in Angola, Mozambique, Cape Verde, Namibia as well as Sao Tome and Principe, East Timor, Malta and Portugal. Asseco PST is a reference for Portuguese-speaking information systems with over 60 bank clients. In 2021, the company acquired a majority stake in the Portuguese capital market solutions firm Finantech in order to diversify its products, clients, and markets. Finantech's primary product is the SIFOX platform.

Its primary business is developing solutions for the financial industry on the Promosoft Financial Suite (PFS) software platform to address the present banking issues. The Comprehensive, Modular, and Scalable Core Banking System (CBS) from Asseco PST provides a bank's core business operations, including retail, corporate, financial markets, and payments. In addition to banking software development, the company offers IT Infrastructure & Security, Consulting, and Development services, as well as Data & Analytics solutions.

The Data & Analytics sector provides a plan for the integration, processing, and accessibility of data, based on a uniform and standardized architecture. This design is based on the three-layer organization of data:

- The structural layer consists of data sources (Staging and DW)
- Access layer - ETL services and procedures that expose data from Staging and DW to applications requiring the data.
- Application layer — the layer where internal and external applications that consume data exist.

The solutions of the Data & Analytics department focus on the data extraction engine, data quality engine and advanced BI, supporting the clients with integration of CRM, HR, and Sales solutions as well as offering management of financial performance, company predictions and indicators, and oversight of noncompliance.

1.2. OBJECTIVES AND METHODOLOGY

The purpose of this project is to support Asseco PST Data & Analytics team in providing a machine learning solution to be implemented as an extension of their current CRM solution that they offer. Although the current CRM system uses automation, the marketing campaigns rely heavily on the staff's expertise, and the campaign response probability is assigned manually to the customers. The aim is to enrich the current CRM solution with ML generated customer segments to optimize the marketing campaigns, as well as offer product recommendation system to offer new products to current and new customers.

The objective of the project is to provide a solution that is readily scalable among Asseco PST clients, taking into account the integration with the existing CRM architecture and the deployment of the project. In addition, the design of the deployment of the machine learning model will include how it can be applied to other clients and how the code can be reused for other applications. In order to do this, the project provides a model deployment architectural design in which data preparation is accomplished through ETL and stored in a data lake, which communicates with the application and returns results that may be moved to the main CRM database. This enables the reuse and application of model code packages to various clients and use cases.

The solution will be developed in a lab environment and use client data of one of the Asseco PST clients, an Angolan bank. The dataset consists of data extracted from the bank core system of 46 008 active customers who have made at least one transaction in the past year.

The project will follow the CRISP-DM data mining approach to apply machine learning methods on the customers dataset. As proposed by Aryuni et al (2018), the clustering will be done by combining movements data, sociodemographic information, and product ownership data to create customer segments, which will be used to create lists of customers for CRM segments to be used in specific campaigns. Additionally, a predictive method will be used to classify new customers into existing segments. This then will be used to create an agnostic solution to be implemented into the system data pipeline to be added into the database. The second proposed part of the project is to develop a Recommendation System to predict which customers will be interested in certain bank products. Finally, PowerBI reports will be used to communicate the findings, by visualizing the results of segmentation and recommendations.

This project is expected to provide the client with a solution to enrich their current CRM platform, which would be ready to implement as a part of their current BI service. The expectation is that the

findings would be easily understood by managers, and it will be beneficial for the end user to create personalized marketing campaigns.

2. LITERATURE REVIEW

2.1. CRM AND AI IN BANKING

According to Ozdemir et al. (2022), customer data is the core of customer relationship management (CRM), and technology assesses it and utilizes it to analyse and to identify and solve problems or improve the current situation. As technology continues to progress, organizations are using it to predict their consumers' future attitudes and actions and to comprehend the motivations behind their existing behaviours and activities. According to the report, investments in CRM technology, CRM strategy, and innovation capacity have a beneficial impact on financial success and customer satisfaction. The study also indicated that organizations with the ability to innovate are more likely to successfully respond to competitors and acquire new capabilities that provide a competitive advantage.

The theoretical study of Gallego-Gomez and De-Pablos-Heredero (2020) demonstrate the capability of AI to promote new customer relationships, detect their needs or experiences, and adapt the services they provide to be more competitive. This study also demonstrates that proper AI implementation allows for a reshaping of traditional banking scenarios. Detection, absorption, integration, and innovation, that AI in banking brings, are capabilities that allow for cost savings, increased efficiency, and increased competitiveness.

According to McKinsey & Company (2021) report, in order to improve customer experiences and outcomes, banks need to effectively customize their interactions with clients. This may be accomplished by analysing client data and gaining a deeper understanding of their requirements and preferences to fuel the development of sophisticated, customized client propositions that go beyond typical financial services and cater to a range of consumer demands. To do so, businesses must take a customer-centric strategy and incorporate AI and analytics capabilities into their systems and platforms. According to the report, banks should prioritize using advanced analytics and machine learning in decisions across the customer life cycle, from customer acquisition, credit decisioning, through monitoring and collections, deepening relationships, and smart servicing.

According to Omoge et al. (2022), the use of AI driven CRM systems has a positive and direct effect on service quality, customer satisfaction, and consumer purchasing behaviour. Nonetheless, in the banking context of Nigeria, the study also establishes that technology downtime has a moderating effect on technology usage, consumer purchasing behaviour, and customer satisfaction. The paper points out the technological downtime, frequently occurring in developing countries, that has yet to be studied on a large scale.

2.2. CUSTOMER SEGMENTATION IN BANKING

Customer segmentation has been widely used in banking for a very long time to identify homogenous groups of clients that are distinctively different from each other, idea presented by Smith (1956) to create alternative marketing strategies. A-priori strategies of the customer segments being chosen by analysts based on their sectors, geographic or demographic criteria have been slowly replaced by post hoc methods, where they are clusters based on homogenous patterns of features within a group that is heterogenous within the population (Green, 1977; Gwin and Lindgren, 1982).

The customer segmentation research in the XXI century moved into the latter approach, using data and distance measures to divide the segments. Such an approach can be seen in Machauer and Morgner (2001) research, which used survey of bank customers towards information and technology attitudes and used the responses to segment customers using agglomerative hierarchical method of clustering. The study showed that the clustering method was superior to the a-priori segmentation defined by simple demographic criteria. With the rapid technology advancement and the digitalization of banking, more data became available for analysis and the computational costs of such analyses decreased, allowing for more computationally demanding algorithms to be used using large databases of customer data (Machauer & Morgner, 2001).

Zakrzewska and Murlewski (2005) compared three clustering algorithms using high dimensional data for bank customer segmentation. They used DBSCAN, k-means and two-phase clustering consisting of modified k-means and hierarchical agglomerative methods, concluding that k-means is very efficient for high dimensional datasets, and offers the best performance in terms of scalability and run times due to its simplicity. Although the two-phase algorithm had a very good performance for noisy data with small number of dimensions, it performed poorly when scaled, due to its complexity as well as the use of hierarchical method that is time consuming. DBSCAN was found to be difficult to use due to its sensitivity to input parameters as well as the detection of too many outliers.

Martens et al. (2016) present a method for targeting existing bank customers by analysing the transactions of 1.2 million customers of a large European bank using an AI with an application of identifying prospective customers for marketing offers in banking. They argue that fine-grained consumer behavioural data can predict which consumers will be good candidates for specific offers. To do that, they created pseudo-social network where the consumers are linked if they shopped at the same merchant, and then calculated their behavioural similarity scores (BeSim). They emphasize that the proposed method was much more accurate than the traditional targeting method in identifying customers who purchased the product, arguing that AI can assist commercial bank employees in approaching customers with offers that are more likely to be accepted, benefiting both the customer

and the bank. The study provides a valuable approach to combining the traditional variables like demographics or RFM scores with the transactions BeSim scores. Even though the approach resulted in more accurate predictions, there is no information whether it could be successful in other banks, especially in other regions where the card transactions are not the dominant form of payment, such as in African countries where the cash payments are more widely used. Moreover, the study focused on just on debit transactions, not analysing the bank transfers and cash withdrawals. Another limitation is that the proposed method assumes the variety of merchants that they shop at, which in developing countries might not be applicable as most bank customers use bank transactions mainly in the large supermarkets, energy providers and petrol stations and use other methods of payment in smaller merchants; thus, by the proposed method of weighing down those popular merchants would be left with not much available data for the accurate analysis.

Aryuni et al. (2018) conducted research on customer segmentation in a bank to compare K-means and K-medoids clustering methods, using customers' RFM scores of their internet banking transactions of an Indonesian bank. The results of the study show that K-means method outperformed the K-medoids methods. The evaluation metrics to rank the clustering techniques consisted of just two metrics: Average Within Cluster and Davies-Bouldin Index, however these are not the only evaluation metrics that can evaluate the clustering model, and moreover none of the visual representations of the clustering were used. Moreover, the study uses just three variables, which are a-priori classifies into scores of 1-5, thus possibly impacting the algorithm that could possibly could have found different patterns within the data, and the authors did not clearly state a reason for that. It would have been interesting to compare the results with method using normalized and without scores transformation variables. The results of the research are proposed to be used to group customers based on their spending behaviour, however, do not propose an example of that. They also state that the clusters formed during the clustering method can be applied to CRM, sales, and marketing for the targeted customers, however without any concrete example. The study doesn't look at the business value of the proposed methods nor shows a real-life implementation. Authors suggest that for future works, clustering should be applied in combination with socio-demographic and product ownership data.

Djurisic et al. (2020) propose an approach to segmentation of credit card users in banking. The method integrates the Recency, Frequency, and Monetary (RFM) method, clustering using the k-means model, and predictive classification with the Support Vector Machine (SVM). They argue that the proposed approach allows marketing managers to more effectively target the most valuable, increasing revenue while also reducing unnecessary costs due to incorrectly targeted valuable clients. The results show the potential of using the unbiased characterization of different client categories to enable marketers to develop different offers for specific segments, as well as to place highly accurate marketing offers

to the right audience. Compared to other methods, they used a combination of unsupervised and supervised methods for clustering, proposing classification of segments to be used as well as the application of DT to explain the feature importance. However, the study puts more focus on the predictive task, not the clustering method that is applied first, as they used just one method for the k means evaluation for the optimal number of clusters and did not extensively evaluate the method used, as the result of segmentation could impact the results of classification models. Additionally, the study focused on just the credit card transactions outside of other features that could impact the results of the model.

The research of Namvar et al. (2010) provides a further illustration of the use of algorithm combinations. They presented a method for customer segmentation using k-means clustering on a dataset of Iranian bank customers, in combination with neural network classification, utilizing RFM (recency, frequency, and monetary value of customer transactions), demographic data, and LTV (lifetime value of the customer). RFM scores and demographic characteristics were utilized to categorize clients into nine groupings. The average LTV for each cluster was obtained using a neural network classification algorithm to anticipate each customer's prospective value in the evaluation of clusters profiles. The resulting clusters can provide the management indicators to boost client retention, loyalty, and profitability. Moreover, it is claimed that the process is more systematic than single-point-of-view methods for client segmentation. According to the study, this strategy can help marketers improve their strategy, boost customer relationship management, customer loyalty, income, and up- and cross-selling prospects.

Mihova & Pavlov (2018) in their paper on customer segmentation approach in commercial banks offer a different view on the evaluation of clustering method that focuses on the goal of the segmentation for marketing strategy. Through comparison of different methods using k-means model, they argue that the clustering method should be chosen to inform the marketing strategies based on the variable they want to play the most important role and the goal of the marketing actions, thus offering a business-based perspective to the evaluation of the clustering performance.

2.3. RECOMMENDATION SYSTEMS

Recommender systems (RSs) are a type of artificial intelligence that give item recommendations to users (Portugal et al., 2018). RSs were introduced for the first time in 1992 when Tapestry, a collaborative filtering recommendation appeared (Goldberg et al., 1992). Since their adoption in the 1990s, they progressed to employ machine learning algorithms to categorize goods by genre and

propose other products to consumers and are categorized as collaborative, content-based, or hybrid filtering (Adomavicius & Tuzhilin, 2005).When processing information for recommendations, collaborative RSs consider user input, content-based RSs base recommendations on item data, whereas hybrid RSs combine the two preceding types and provide suggestions based on both user and item data (Portugal et al., 2018). Fayyaz et al. (2020) extend the types of recommendation system types, including demographic, utility, and knowledge-based methods.

In the systematic review on recommendation systems based on 121 primary studies, Portugal et al. (2018) found that collaborative techniques, particularly those employing neighbourhood-based methodologies, are prevalent in RS development, while hybrid techniques are still an area of research potential. Clustering algorithms, ensemble methods, and support vector machines (SVMs) are often utilized ML approaches in RSs. The study revealed that the main application of using machine learning for product recommendations is found in movies, followed by social and academic domains. Clustering algorithms, ensemble methods, and support vector machines (SVMs) are often utilized ML approaches in RSs. According to Portugal et al. (2018), future research opportunities include studies in the big data technologies as well as studies analyzing early and late stages of RS development.

Fayyaz et al. (2020) describe main challenges in modern applications of recommender systems that can affect its performance. Cold-start can happen when there is not enough initial data such as user interactions, and when a new user does not have any preferences or history available to base recommendations of a product. To tackle the problem of scalability, authors propose using clustering techniques to segment users, which brings two main benefits: alleviates the sparsity of the dataset and divides the dataset into smaller partitions, resulting in a significant reduction of the prediction generation speed. Other challenges mentioned by the authors include data sparsity, diversity, and habituation effect.

2.3.1. Recommendation Systems in Banking

According to Oyebode and Orji (2020), application of recommender systems can be challenging in the banking sector, as there are no product ratings available. Moreover, the cold-start problem can make it difficult to a prospective or new customer who does not have any preferences yet. The study suggests utilization of a hybrid approach to firstly derive the customers preferences from the transaction data, and secondly applying item-based collaborative filtering in combination with demographic-based approach. The results of the study suggest that the hybrid-based approach performed better than a

single method of either item-based or demographic-based approaches. The proposed recommender system can be scalable and practical in a banking environment, as well as in other financial domains.

Gallego and Huecas (2012) present a context-aware mobile recommender system that incorporates the banking data of user profiles, credit card transactions, and the merchants where payments were made. They propose a three-phase solution beginning with Social Context Generation, where user profile segmentation is used to assign clients to social clusters, followed by the computation of the clusters trends map using transaction and place data, and then for new customers, the system assigns them to existing social clusters based on the information profile extracted from their bank account. The location of the mobile device is then used as an input to the recommender system in order to further customize the recommendation based on the user's location. In the third phase, User Context Filtering, a collection of characteristics such as the time of day or current activity, in conjunction with the user's input preferences for place category, are utilized to rank recommended locations. Lastly, the user has the ability to provide comments on the suggested location, which may result in a cluster reassignment. This study provides research on the architecture and implementation of the deployed solution in a real-world banking environment. Moreover, the solution provides transparency through an explanation of how the recommendations were derived and has been recognized by users as trustworthy and efficient. Future expansion of the solution to cross-domain suggestions is mentioned by the authors.

Wang et al. (2018) presented a personalized recommendation method for financial advertising recommendation based on customer segmentation. Through the segmentation of the customer groups, the method can reveal different consumption habits and consumer preferences of the customer groups, which becomes the basis for the association rules mining to provide the targeted customer personalised service based in the customer segment. The proposed method based on segmentation performs better than the traditional recommendation method.

FOBA (Fog Oriented Banking Architecture), a solution based on Fog computing, was proposed by Hernandez-Nieves et al. (2020) as a novel deployment-oriented recommendation system for financial products in banking. FOBA integrates predictive systems and enhances customer support services with increased security, transparency, and agility while reducing management costs. The architecture consists of fog nodes that process data in real-time using light intelligent agents and a hybrid recommendation method that combines collaborative filtering and content-based filtering. The recommendation engine utilizes information from sensory networks, such as static nodes (located in the bank office) and mobile nodes (phones, cards, wristbands), and through development in the fog layer, provides personalized recommendations via a personal agent. The proposal includes the

validation of recommendation success rates via Case Based Reasoning (CBR) hosted in the cloud and permits the sharing of business intelligence and context data in fog nodes.

2.4. LITERATURE REVIEW CONCLUSIONS AND LIMITATIONS

The available literature emphasizes the relevance of using AI solutions in the banking CRM, whereas the majority of research concentrate on specific use-cases. In the context of customer segmentation, there is a lack of study on the implementation and development of the recommended approaches in the CRM system. In addition, the studies do not demonstrate how the offered solutions may be reapplied in other situations, with a few exceptions such as Gallego and Huecas's (2012) recommender system, which has the potential to be used in other fields.

In addition, the majority of the literature primarily describes potential solutions and does not give data on the proposals' actual implementation, adoption, or effectiveness. Most of the literature use cases are based on western banks, which may not be appropriate to developing nations. Omoge et al.'s work (2022) is the only one that discusses the application of AI in CRM systems in Africa. The authors provide insights on developing country difficulties, such as technological downtime, that must be addressed while constructing a data-mining solution for an African bank.

According to the Literature Review, most consumer segmentation studies used the RFM technique for feature transformation. Some of the studies are based on customer profiles, while others are based on transaction data or a combination of the two. K-means is the most common and effective approach for clustering customer data in banking. However, the studies lack in-depth analysis of the feature transformation and its significance. The majority of research assess segmentation outcomes based on performance criteria, such as silhouette score, and not on the commercial value of cluster profiles, as did Mihova and Pavlov (2018).

Banking Recommendation Systems are very new, and as a result, there is little research in this topic. The banking sector's difficulties in processing massive amounts of data is one of the obstacles to the widespread use of this technology in the banking industry. The absence of rankings of bank products and services, as well as the system's inability to scale, are other concerns. The bulk of banking recommendation systems use hybrid techniques and clustering to analyse consumer data. This could imply that the customer segmentation solution can be used in conjunction with the recommender system, as the two machine learning solutions could become integrated and form the basis of a more advanced artificial intelligence system that can feed information from one ML solution to another. Although most applications are based on back-office data, certain suggested applications, such as

FOBA by Hernandez-Nieves et al. (2020), provide a novel approach to environment-based suggestions that may be implemented in the bank's office. However, the cost and complexity of the suggested solutions are not addressed in the research, which might impact the attitude toward its adoption. In addition, none of the research address privacy concerns associated with these solutions.

2.5. NEXT STEPS

Next steps include a deeper understanding of the African banking culture and the business understanding of the Angolan Bank, as well as data understanding of the dataset. For the clustering solution, a sample k-means clustering has been applied, which required feature transformation and data pre-processing. This resulted in a formation of 8 clusters formed from the customer profile data and their transactions. A sample PowerBI has been made for the business understanding of the customer segments. The next step requires designing architecture for the solution, as well as application for model deployment, coded in Python. The design of the application requires for the data preparation and transformation to be done in ETL process, taking the data from core system to a data lake that will feed the results to the CRM database. This approach to design of the python application allows for the code re-usability for other projects and use cases. The next step is to design the recommender system solution which will benefit from the clustering method applied earlier.

3. METHODOLOGY

In today's competitive marketplace, businesses endeavour to provide customized consumer experiences. AI-powered consumer analytics have emerged as a valuable instrument for businesses to comprehend the behaviour, preferences, and requirements of their customers. The Asseco PST Data & Analytics team intends to enhance their current CRM solution by integrating machine learning solutions that can optimize their marketing campaigns and provide customers with a product recommendation system. This project's objective is to assist the Asseco PST Data & Analytics team in implementing a machine learning solution as an extension of their current CRM system. The project concentrates on augmenting the existing CRM solution with customer segments derived by machine learning in order to optimize marketing campaigns and provide a product recommendation system. The goal is to develop an extensible solution for Asseco PST clients, taking into consideration the integration with the existing CRM architecture and the deployment of the project.

3.1. RESEARCH DESIGN

The project uses the CRISP-DM data mining methodology to implement machine learning techniques on the client's dataset. The dataset consists of data extracted from the bank's main system pertaining to 46 008 active consumers who have conducted at least one transaction within the past 12 months. The clustering will be accomplished by combining movement data, sociodemographic information, and product ownership data to generate customer segments, which will be used to generate lists of customers for CRM segments to be utilized in specific campaigns. In addition, a predictive method will be employed to categorize new consumers into pre-existing segments. This will then be used to develop an agnostic solution that will be integrated into the system data pipeline and appended to the database. The second portion of the proposed project is to develop a Recommendation System to predict which consumers will be interested in particular bank products. The findings will be communicated using PowerBI reports, which will visualize the results of segmentation and recommendations. The project will be developed in a laboratory using client data from one of Asseco PST's Angolan bank clients. The project methodology will emphasize deployment and how the project is structured and programmed to ensure that Asseco PST clients can simply implement the solution. The success of the project will be determined by the precision of the customer segments generated, the efficacy of the recommendation system, and the simplicity of implementation and deployment of the solution.

3.1.1. CRISP – DM Methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a popular and highly regarded data mining project methodology (Chapman et al., 2000). It is comprised of six phases, each with its own distinct roles and objectives:

1. Business Understanding

This phase entails comprehending the project's objectives and how it will contribute value to the organization. It entails identifying business objectives, evaluating the circumstance, and defining the project plan.

2. Data Understanding

In this phase, the emphasis is placed on accumulating and comprehending the data that will be used for the endeavor. It includes identifying the data sources, collecting and analyzing the data, and validating the data's integrity.

3. Data Preparation

After collecting and comprehending the data, it must be prepared for analysis. This phase entails cleansing, transforming, and integrating the data to generate an analysis-ready dataset.

4. Modelling

Actual data analysis occurs during the modelling stage. During this phase, various modelling techniques are used to identify patterns and relationships in the data. The models are then evaluated and, if necessary, refined.

5. Evaluation

In the evaluation phase, the efficacy of the models is evaluated. This involves comparing the models to the business objectives and determining whether or not they achieve the intended results.

6. Deployment

The concluding phase of the project involves implementing the project's outcomes within the organization. This may entail integrating the models into existing systems, creating new reports or interfaces, or instructing personnel to use the models effectively.

CRISP-DM is crucial for data mining initiatives because it provides a structured and methodical approach to data analysis. By adhering to this methodology, data scientists and analysts can ensure

that all aspects of the project are taken into account and that the outcomes are in line with business objectives. This reduces the risk of errors, ensures the project is completed on time and within budget, and increases the project's value to the organization. In addition, the standardized approach of CRISP-DM makes it simpler to communicate project results to stakeholders, including non-technical organization personnel.

3.2. BUSINESS UNDERSTANDING

The main objective of the project is to create a sample solution that will function as a framework and proposal to introduce the Machine Learning solutions to the clients. As mentioned earlier, in the company description section, the clients of Asseco PST are mainly traditional banks based in Portuguese-speaking countries. Athough, the objective of the solution is to be easily scalable across all the possible clients, the project will focus on one of them, an Angolan Bank. Currently, there is no advanced analytics or machine learning applications implemented, and therefore, the main objective of the solution is to have a high potential return on investment value to the company.

3.3. DATA UNDERSTANDING

This section describes the methods and techniques utilized during the Data Understanding phase of the study, including the data sources, the collection process, data exploration and visualization techniques, and data quality assessment methods. In addition, the results of this phase are presented, including the main characteristics of the data, the identified patterns and relationships, and the quality issues that must be addressed.

3.3.1. Collection of Initial Data

Asseco PST provides core banking system software, Promosoft. The historical data is extracted and stored in Core Staging SQL server which then is used to build CRM Data Warehouse. The data available for the analysis is within the transactional and customer SQL Server tables. The available tables are presented in the table below.

Table name	Description
CORE_ACTIVE_ENTITIES_CUSTOMERS	Contains information on the activity state of accounts and entities
CORE_CLIENT	Contains client information
CORE_CONTRACT	Contains product contracts for each client
CORE_CREDIT_BALANCE	Contains credit balance and credit type data

CORE_ENTITY	Contains detailed information on each bank entity
CORE_MARITALSTATUS	Description of the marital status code
CORE_MOVEMENT	Collects all the data on bank movements
CORE_PRODUCT	Description of core product codes
CORE_PRODUCTCLASS	Description of product class codes
CORE_QUALIFICATION	Description of qualification codes
CORE_RELATIONSHIP	Contains information on account and entities, whether they are the

Table 1 Core Staging Database Tables

There are 342 059 customers in database, out of which 343 399 are private active entities. After querying the data, the sample dataset was obtained from Core Staging database. The rules for extracting the data were to only get entries from active private customers who has made at least one bank movement in the past 6 months (rules that are used for the current CRM solution to classify a customer as active). The data was pulled out from SQLServer Movements and Customer related tables using SQL query and stored in a table in Machine Learning Data Lake to be used as basis for further analysis. The resulting dataset consists of 48 337 (46 008 unique) entries that were anonymized, using only the Customer Number for identification.

3.3.2. Data Exploration

In order to understand the structure of data, and what variables might be useful for further analysis, a thorough exploration of the data has been made. Since the dataset is obtained directly from the SQL database, it is important to explore its structure and relationships between the tables.

To obtain demographic data of customers, a join needs to be performed on CORE_CLIENT and CORE_ENTITY table (refer to diagram in Appendix A). Clients are the main holders of the account and entities can have a 'Titular 1' or 'Titular 2' when the account is made in the name of the main holder, which are expressed in the Name column of CORE_RELATIONSHIP table. For the purpose of the analysis, it has been decided to focus on the main holders of the accounts ('Titular 1'). Since the aim of the project consists of two separate tasks: gaining insights on customers and product recommendation system, the dataset will be divided into customers data and products recommendation data, which will be stored in separate tables.

3.3.3. Data Quality Verification

After the initial assessment of the data, it appeared there has been data quality issues. The main concern appeared to be with some of the columns in the CORE_ENTITY table. Majority of the values in ANNUALINCOME or JOBTITLE columns are missing values, meaning they cannot be used for the analysis. Moreover, the POSTALCODE column has not been verified for quality when the data was inputted, as more than 50% of the clients have the same Postal Code value, meaning that the location of the customer is not available for the analysis as well. Apart from that, there are very few missing values (less than 1%) in QUALIFICATION, PROFESSIONALACTIVITY and MARITALSTATUS tables. Regarding duplicates, there were 2247 duplicated rows which will need to be dropped.

3.4. DATA PREPARATION

The data preparation phase is essential for assuring the quality and usability of data in the context of data analysis and data mining. The data preparation component of the CRISP-DM methodology includes data collection, data cleaning, data integration, data transformation, and data reduction. All of these stages were completed using VS Integration Services and a SQL query during the ETL phase to create new table in the Machine Learning Data Lake for the clustering purposes. Thus, we are able to perform the ETL processes independently of the Machine Learning Solution, allowing for the scalability and reusability of the code for the future solutions of other clients, as well as the clustering of different datasets. At this stage, the essential feature engineering can be performed, and any variations of the data preparation can be recorded in distinct tables so that the results and feature importance can be compared within new projects utilizing the same solution.

Data selection is the first stage in the data preparation process. As mentioned in the previous section, the data for this project was the historical data was stored on the Core Staging SQL server, which was then utilized to construct the CRM Data Warehouse and was extracted from the bank's core system. The SQL Server tables containing transactional and customer data and related tables are available for analysis.

Data was collected from multiple tables in the CORE STAGING database. CORE_CLIENT, CORE_RELATIONSHIP, CORE_ENTITY, CORE_ACTIVE_ENTITIES_CUSTOMERS, CORE_CREDIT_BALANCE, CORE_MOVEMENT, and CORE_CONTRACT were among the tables included.

Data cleaning is the second stage in the data preparation process. Several initial data cleaning processes were outlined in the Data Collection phase, filtering out non-active customers, as well as customers who did not make a movement in the past 6 months.

The SQL query eliminated customers who were not of type 'P' (individuals) and whose last movement date was unknown. In addition, credit balances with the 'V' situation code were omitted. The query excluded contracts with a status code of 'E' and only included contracts with the 'DO', 'DP', 'CRR', and 'CRF' product classes, which are the main bank products.

For the purposes of the Recommendation System, a separate table was constructed by combining the Clustering result table with product data from the CORE_CONTRACT table, again only including contracts within the 'DO', 'DP', 'CRR', and 'CRF' product classes.

3.4.1. Data Transformation and Integration

Since the purpose of the clustering solution was to enhance the CRM system for Marketing Purposes, necessary features for Marketing need to be obtained. To design effective marketing campaigns, the bank needs to know basic information of the client like age, education, gender, or relationship status.

First, client information was retrieved from the CORE_CLIENT table, and then the DATEDIFF function was used to determine the length of time since account creation for each client. Second, demographic information was extracted from the CORE_ENTITY table, including the client's marital status, education level, occupation, age, and gender. The MARITIALSTATUSCODE column was converted into a binary variable, with a value of 1 indicating a singular status and a value of 0 indicating otherwise. Similarly, the QUALIFICATIONCODE column was transformed into a binary variable, with a value of 1 indicating a higher education level and a value of 0 indicating otherwise. Additionally, the PROFESSIONALACTIVITYCODE column was transformed into a binary variable, with 1 indicating that the client is a student and 0 indicating otherwise. The AGE column was determined by subtracting the client's birthdate from the current date and dividing the resulting number by 10000. Finally, the GENDERCODE column was transformed into a binary variable, where 1 indicates that the client is female and 0 indicates otherwise.

The third step involved retrieving transactional data. As suggested Aryuni et al (2018), Martens et al. (2016) and Namvar et al. (2010), the chosen approach for feature selection was to combine customer characteristics and demographic data with Recency, Frequency and Monetary values commonly used in Marketing Analytics. Those values would need to be calculated based on the data from CORE_MOVEMENTS. Recency is expressed as days since the last movement, Frequency is the number of movements in the past 6 months, and Monetary Value is the sum of all the movements in the past

6 months. The LAST_MOVEMENT_DATE column was converted to a Recency variable by subtracting the date of the most recent transaction from the extraction date. The average available balance for each client was extracted from the CORE_CONTRACT table containing banking data. Finally, credit information was retrieved from the CORE_CREDIT_BALANCE table and a binary variable was constructed to denote whether a client has a credit history.

Using SQL join statements to link tables based on their shared key values, all these steps were performed. The resulting dataset has been cleaned, the missing values and duplicates have been removed, and filtered to include only the necessary variables for analysis. The figure below represents the Clients_Clustering table with the transformed dataset, which is stored in a non-relational database MLDataLake.

Clients_Clustering	
CLIENT_NUMBER	VARCHAR
CLIENT_SINCE	INT
AGE	INT
AVERAGE_AVAILABLE_BALANCE	NUMERIC
IS_SINGLE	INT
HIGHER_EDUCATION	INT
IS_STUDENT	INT
GENDER	INT
RECENCY	INT
MONETARY_VALUE	NUMERIC
FREQUENCY	INT
IS_CREDIT	INT
TS	DATETIME

Figure 1 Clients_Clustering Table and its Variables

The Product Recommendation dataset utilized a comparable ETL process utilizing VS Integration Services. In this instance, the Clients_Clustering table was joined with the pivot counts of each product code, resulting in the creation of a new column for each product code containing the product count as its value. Even though the resulting table contains over 120 columns for each product, its structure is readily accessible for further analysis, eliminating the need for pivoting or encoding the products data within the solution.

3.5. Machine Learning Methods

This section will discuss the machine learning techniques used to enrich bank's CRM and build a recommendation system. This section will provide a detailed explanation of the methods used, including clustering and predictive modelling, as well as their advantages, limitations, and selection for

this project's goals. K-means clustering, a frequently employed unsupervised machine learning algorithm for classifying data into discrete clusters, is the first method discussed. It was selected due to its applicability to business scenarios. The second technique is multi-class classification, which requires the development of a predictive model to designate cluster labels predicted by the K-means algorithm. This will facilitate assignment of cluster labels for new data points. A recommendation system employing clustering-based solutions and item-based collaborative filtering is finally proposed. This system will aid in providing consumers with personalized recommendations based on the customer characteristics or their ownership of bank products.

3.5.1. K-Means Clustering

K-means clustering, also known as Lloyd's algorithm, is a commonly used unsupervised machine learning algorithm for categorizing data into distinct clusters. It functions by partitioning data points iteratively into K clusters based on their proximity to a set of centroids. In customer relationship management (CRM) systems, K-means clustering can be used to segment consumers according to their purchasing behaviours, demographics, and other relevant metrics. This can help businesses tailor their marketing campaigns and customer service strategies to specific customer segments, thereby enhancing customer retention and satisfaction. Due to its applicability to business scenarios, K-means clustering was selected as the segmentation technique for customers.

K-means clustering algorithm has three primary stages. Initially, the initial centroids are selected, typically by selecting k samples from the dataset X at random. The algorithm then iterates between two additional steps: designating each sample to its nearest centroid and creating new centroids by calculating the mean value of all samples assigned to each previous centroid. The algorithm then computes the difference between the old and new centroids and replicates these last two stages until the difference falls below a predetermined threshold. In other words, the algorithm continues until the centroids no longer move considerably, indicating a stable clustering solution. The K-means algorithm aims to minimize the within-cluster sum of squares (WCSS), which is the squared sum of the distances between each data point and its assigned centroid, also known as inertia. The algorithm can converge to a local maximum, which means that the ultimate clustering result may not always be optimal. Therefore, it is suggested to execute the algorithm multiple times with various initial centroids in order to achieve a superior ultimate result (David & Vassilvitskii, 2007).

K-means clustering was applied using an open-source library, scikit-learn. The chosen version of the algorithm was K-means++, which is used to determine initial cluster centroids in the K-means clustering algorithm. Unlike the traditional K-means algorithm, which selects initial centroids at random, K-means++ chooses initial centroids intelligently to increase the likelihood of obtaining a superior final

clustering. The K-means++ algorithm selects the first centroid from the dataset at random. The selection of subsequent centroids is then based on their distance from the previously chosen centroids. The probability of designating a data point as a centroid is proportional to the square of its distance from the nearest centroid. This ensures that the designated centroids are distributed out and not too close to one another. By selecting better initial centroids, K-means++ can converge more quickly and produce a superior final clustering result than traditional K-means. One of the limitations of K-means++ is that it may get stuck in local optima; therefore, the scikit-learn library employs an extension of the algorithm called Greedy K-means++ that addresses this limitation by conducting multiple trials at each sampling phase and selecting the best centroid among them (David & Vassilvitskii, 2007).

One of the benefits of K-means clustering is its scalability and effectiveness. It can manage large datasets with minimal computational expense and is simple to implement. In addition, it does not require prior knowledge of the data, making it a versatile method for exploratory data analysis. K-means clustering has some limitations, however. In real-world datasets, it is not always the case that clusters are spherical and of equal dimensions. In addition, it necessitates the selection of the optimal number of clusters, which can be subjective and influence the quality of the results. K-means clustering remains a popular and useful instrument for data analysis in numerous disciplines, including CRM, despite these limitations.

3.5.2. Multi-Class Classification

The next step of designing a clustering solution involves development of predictive model to assign the cluster labels predicted by the K-means algorithm for a new set of data. In the context of K-means clustering, semi-supervised learning can also refer to a technique involving the development of a predictive model to attribute cluster labels to new unlabeled data points. Using K-means, the initial dataset is clustered to acquire the cluster labels for the labeled data points. The labeled data is then used to train a predictive model.

The appropriate model to be chosen for this task is a classification model that predicts an observation's category or class based on a set of input features. These models are trained on a labeled dataset in which each observation is associated with a known class label, and the model learns to recognize patterns and relationships in the data that can be used to accurately predict the class of new, unlabeled data points. Since, the results of clustering would usually form more than two labels, the chosen classification model would need to support multi-class output. Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were chosen as classification algorithms, due to its support of the multi-class output within the scikit-learn library.

3.5.2.1. K-Nearest Neighbours

K-Nearest Neighbours was selected as the first predictive model due to its similarity to a clustering method, which divides the dataset based on its distance from the centroids. Nearest Neighbour methods are based on locating a predetermined number of training samples that are the closest in distance to the new point and predicting the label based on these. This is accomplished through a simple majority vote of each point's nearest neighbours, where the query point is designated the data class with the most representatives among its nearest neighbours. This method is known as neighbours-based classification, which is a form of instance-based or non-generalizable learning. Instead of attempting to build a generalized internal model, it simply stores instances of the training data (Fix and Hodges, 1952).

K-nearest neighbours (KNN) is a non-parametric classification algorithm that assigns a label based on the majority class of k-nearest neighbours for a given data point. The algorithm employs a constant number of training samples that are closest to the test point, with the number of neighbours k being user-defined. Alternately, radius-based neighbour learning can be used, in which the number of neighbours fluctuates according to the local density of objects within a certain radius. The distance metric can be any metric measure, but the standard Euclidean distance is the most common option. KNN is a neighbours-based method known as a non-generalizing machine learning technique because it stores all of its training data without attempting to build a general internal model. Utilizing quick indexing structures, such as Ball Tree or KD Tree, can enhance the algorithm by optimizing the search for nearest neighbours (Guo et al., 2003)

As a non-parametric method, it makes no assumptions about the underlying distribution of the data, which is one of its strengths. KNN also performs well on datasets with low dimensionality and highly irregular decision boundaries. KNN has some limitations, however. It is a memory-based algorithm that can be computationally costly for large datasets. Choosing the right value for k can also be difficult and is highly dependent on the dataset. In addition, KNN presupposes that all features are of equal importance, which is not always the case. Lastly, KNN is sensitive to the data scaling, and standardizing the features is frequently required for optimal performance.

3.5.2.2. Support Vector Machine

SVMs are a class of supervised learning algorithms used for classification, regression, and outlier detection. SVM seeks to identify a hyperplane that best separates data into distinct classifications. This hyperplane is a line in two dimensions, but it is a hyperplane in higher dimensions. The SVM algorithm operates by mapping the input data to a high-dimensional feature space and then locating the hyperplane that maximally separates the classes. Support vectors are the coordinates closest to the

hyperplane that are used to define the decision boundary. SVM can also use various kernel functions, such as polynomial and radial basis function (RBF) kernels, to convert the data into a higher dimensional space. This allows SVM to identify non-linear decision boundaries, making it an excellent classification tool for post clustering predictions (Chang and Lin, 2022).

SVM used in scikit-learn is called SVC and implements the "one-versus-one" classification method for the multi-class output. This means that N*(N-1)/2 binary classifiers are trained for a problem with N classes, with each classifier separating a pair of classes. During prediction, each classifier votes for their assigned class, and the class with the most votes is predicted to be the final class. This method is computationally more efficient than the "one-versus-all" method, in which N binary classifiers are trained for N classes and each classifier distinguishes one class from all others. Additionally, the "one-versus-one" approach avoids some of the class imbalances that can arise with the "one-versus-all" approach (Crammer and Singer, 2001).

SVMs perform particularly well in high-dimensional spaces, making them suitable for large-featured datasets. Additionally, SVMs perform well when the number of dimensions exceeds the number of samples. SVMs are memory-efficient because they use a subset of training points in the decision function, which is one of their strengths. A further benefit of SVMs is their adaptability. Different kernel functions, including standard kernels and custom kernels, can be specified for the decision function. However, if the number of features is significantly greater than the number of samples, over-fitting can occur, requiring cautious selection of kernel functions and regularization terms. In addition, SVMs do not provide direct probability estimates, and deriving these estimates using time-consuming and costly cross-validation methods can be time-consuming (Chang and Lin, 2022).

3.5.2.3. Decision Trees

Decision Trees (DTs) are a well-known non-parametric supervised learning technique for classification and regression. DTs are especially useful in multi-class classification problems because they enable the separation of multiple classes using basic decision criteria. The fundamental concept underlying decision trees is to partition the feature space into ever-smaller regions by applying simple decision rules at each phase. At each decision point, the algorithm chooses the characteristic that best separates the data based on a predetermined criterion, such as information gain or Gini impurity. After selecting a feature, the data is partitioned into two subsets based on the threshold value. This procedure is then repeated recursively until all of the data has been separated into its respective target classes (Loh, 2011).

The simplicity of decision trees is one of their greatest assets, as they are clear and easy to understand. Decision trees can be readily viewed and comprehended by non-specialists, making them a popular option for applications where transparency and interpretability are crucial. Due to this, Decision Tree was one of the chosen predictive methods. Moreover, decision trees can manage both numeric and categorical data and can be applied to multi-output problems. In addition, they are comparatively insensitive to irrelevant features and able to manage missing data without imputation.

However, there are also limitations to decision trees. A common issue is overfitting, which occurs when the tree becomes excessively complex and begins to capture noise rather than the data's underlying patterns. Various techniques, such as pruning or requiring a minimum number of samples at each leaf node, can be used to address this issue. In addition, decision trees can be erratic, meaning that minor modifications to the data can result in entirely different trees. Using decision trees within an ensemble method, such as random forests or gradient boosting, can mitigate this issue. Lastly, decision trees can be biased towards dominant classes; therefore, it is essential to balance the dataset prior to fitting the model (Loh, 2011).

3.5.2.4. Random Forests

Random Forest (RF) is an ensemble learning technique that incorporates multiple decision trees to enhance the model's precision and robustness. In RF, each ensemble tree is constructed using a bootstrap sample drawn with replacement from the training set, and the optimal split for each node is determined using either all input features or a random subset of size max_features. This incorporates randomness into the trees and aids in reducing their variance, which typically results from overfitting. RF reduces variance and improves generalization performance by combining the predictions of individual trees (Breiman, 2001). This may, however, result in a modest increase in bias. To address this problem, the scikit-learn implementation of RF averages the probabilistic predictions of the individual trees rather than allowing each tree to vote for a specific class. This enables for a model output that is more granular and interpretable. It has been demonstrated that RF performs well on a wide variety of classification and regression tasks, making it a popular choice among practitioners of machine learning.

Real-world applications frequently involve high-dimensional data with a large number of features, which is one of the greatest advantages of random forests. Without extensive pre-processing, they can also manage missing values and outliers in the data. In contrast to many other machine learning algorithms, random forests are resistant to overfitting. This is a prevalent issue with other machine learning algorithms. Randomizing both the samples and the features used to construct each decision tree in the ensemble reduces the variance of the model, thereby making it less susceptible to

overfitting. This results in a model that can perform well on unseen data and is more generalizable. Random forests also provide a measure of feature significance that can be used to identify the most pertinent features for a given task. This can be especially helpful for feature selection and dimension reduction (Breiman, 2001).

There are, however, some limitations to the use of random forests. One limitation is their computational complexity, which becomes problematic when working with large datasets or a large number of features. The training period of a random forest can be lengthy, particularly if the number of trees in the ensemble or the depth of the trees is large. Interpretability is another limitation of random forests. Even though they can provide feature importance measures, the algorithm's inner workings can be difficult to interpret, particularly when a large number of trees are involved. This can make it difficult to interpret how the model arrived at a particular prediction.

3.5.3. Recommendation System

Traditional recommendation methods, as introduced in the Literature Review section, rely on user ratings and product characteristics, which are inaccessible in this particular use case. A recommendation system utilizing clustering-based solutions and item-based collaborative filtering is proposed to address these issues. Our method combines user characteristics with product ownership information to aggregate users with similar preferences into clusters and recommend products based on these clusters. Two methods were used to build the proposed system of recommendations: clustering- based solution and item-based collaborative filtering.

3.5.3.1. Clustering-based Recommendation System

K-Means clustering was used in this proposed method to group users based on their characteristics and product ownership information. The algorithm divides individuals into distinct categories or clusters based on consumer characteristics and product data. By categorizing users with comparable characteristics and product ownership, it is presumed they have similar preferences, and products are recommended based on their collective ownership of the same products.

To calculate the recommendation score for each product within a cluster, products with ownership values below 100 across all customers were removed as this data was considered insufficient for a reliable recommendation. In order to account for differences in cluster sizes, a cluster weight was assigned, which was calculated as the total product quantity in all clusters divided by the total number of products in a cluster, expressed in the following formula:

$$Cluster\ Weight\ =\ total\ product\ quantity\ in\ all\ clusters / total\ products\ in\ a\ cluster$$

Product popularity was accounted for by dividing the product quantity by the average product quantity in a cluster to determine the popularity weight:

$$Popularity\ Weight\ =\ product\ quantity\ /\ average\ product\ quantity\ in\ a\ cluster$$

To derive an initial score that captures differences in product ownership distribution within clusters, the product quantity in a cluster was divided by the mean product quantity across all clusters:

Score = product quantity in a cluster / average of the product quantity for all clusters

The final score for each product was obtained by multiplying the initial score by the cluster and popularity weights.

$$Final\ Score\ =\ Score\ x\ Cluster\ Weight\ x\ Popularity\ Weight$$

The final step is to normalize the scores across the vertical axis of clusters so that there is a maximal score product within each cluster. Min-max standardization was chosen as the normalization method to ensure that the score values are scaled to a range of 0 to 1, which is beneficial for comparing the values across clusters. Normalization was computed with the following formula:

$$X\ normalized\ =\ (X - X\ min)\ /\ (X\ max\ -\ X\ min)$$

Where X is the original score value, X min is the minimum score value in the cluster column, X max is the maximum score value in the column, and X normalized is the normalized value of the data point in range of 0 to 1.

The resulting normalized scores are stored in a table of product scores for each cluster and are ready to be used as part of the solution, where, based on the given cluster label, the user is provided with a list of the most highly recommended products based on their scores for the given cluster.

Utilizing both user characteristics and product ownership information, the clustering-based recommendation system provides a personalized and robust approach to product recommendations. Using K-Means clustering to group users into clusters with similar preferences can potentially allow for more accurate and relevant product recommendations. In addition, the scoring and normalization procedure ensures that recommendations across clusters are uniform and comparable.

3.5.3.2. Item-based Collaborative Filtering

Item-based collaborative filtering is a popular method for creating personalized suggestions based on user preferences. This method employs historical user behavior data to identify items similar to those in which a user has previously expressed interest, and then recommends these items to the user (Ricci et al., 2011).

The first step in implementing item-based collaborative filtering is to construct an item-item similarity matrix. This matrix calculates the cosine similarity between the feature vectors of two objects to determine their similarity. The cosine similarity between these vectors is defined as the cosine of the angle between them, which can be expressed as:

$$Cosine\ Similarity\ (A, B)\ =\ (A\ \cdot\ B)\ /\ (||A||\ ||B||)$$

Where A · B is the dot product of vectors A and B, ||A|| is the Euclidean length (magnitude) of vector A, and ||B|| is the Euclidean length (magnitude) of vector B. The cosine similarity score is a value between -1 and 1, with values closer to 1 indicating that the two vectors are highly similar and values closer to -1 indicating that they are dissimilar. A score of 0 for cosine similarity indicates that two vectors are orthogonal, or completely dissimilar (Manning et al., 2009).

The next step, following the construction of the item-item similarity matrix, is to identify the items with which a user has already interacted, such as items they have purchased or rated. The similarity ratings between these items and all other items are then calculated using the item-item similarity matrix. Then, the user is recommended the products with the highest similarity scores.

Item-based collaborative filtering has the ability to generate recommendations for new users with no interaction history with the system. This is due to the fact that the system can recommend products that are comparable to those that have been popular with users who share similar interests or preferences. However, item-based collaborative filtering may be unable to provide recommendations for new items that have not yet been interacted with by a user. Insufficient data on user preferences or a poorly constructed item-item similarity matrix can also negatively impact the quality of recommendations. Therefore, the item-based collaborative filtering is proposed to work in conjunction with the proposed clustering-based recommendation system (Ricci et al., 2011).

Overall, item-based collaborative filtering is a scalable and efficient method for generating personalized user-preference-based recommendations. By leveraging historical user behavior data

and item-item similarity scores, the system can provide users with relevant product recommendations and a customized purchasing experience.

3.6. Evaluation Metrics

In this section, the metrics used to evaluate the model performance will be discussed.

3.6.1. K-means Clustering

In the context of K-means clustering, the evaluation metrics can not only be used to evaluate the performance of the model, but also to choose the optimal K (the number of clusters). To assess the effectiveness of a clustering algorithm, it is essential to use evaluation metrics. Clustering algorithm performance evaluation is not as straightforward as that for supervised classification algorithms. Any evaluation metric for clustering should not take into account the absolute values of the cluster labels, but rather evaluate whether the clustering defines separations of the data similar to some ground truth set of classes or satisfying some assumption, such that members of the same class are more similar than members of different classes according to some similarity metric. If the labels of the ground truth are unknown, evaluation must be conducted using the model itself (Halkidi et al., 2001).

Since the ground truth was unknown in this study, the following metrics have been used to assess the clustering performance:

3.6.1.1. Inertia or Within-Cluster Squared Sum (WCSS)

The WCSS metric gauges the sum of squared distances between each data point and its assigned cluster centroid. This metric assesses the density of the clusters. A lesser WCSS indicates that the data points are closer to their designated centroids, indicating superior efficacy in clustering. Inertia can also be used to choose the optimal number of clusters, using elbow method. The elbow method is a popular technique for determining the optimal number of clusters for clustering algorithms in a dataset. It works by plotting the relationship between the number of clusters and the cluster inertia or sum of squared distances between each data point and its assigned cluster centroid. The inertia decreases as the number of clusters increases, but at some point the rate of decrease diminishes, resulting in an elbow-shaped parabola. Typically, the number of clusters at this elbow point represents the optimal number of clusters for the provided dataset. It is essential to note, however, that the elbow method is a heuristic method and may not always yield the finest results. To determine the optimal number of clusters, it is recommended to use other clustering evaluation metrics in conjunction with the elbow method (Edwards& Cavalli-Sforza, 1965).

3.6.1.2. Silhouette Score

The silhouette score indicates the degree to which each data point is analogous to its own cluster relative to other clusters. Silhouette Coefficient, which is defined for each sample, consists of two scores: the mean distance between a sample and all other points in the same class, and the mean distance between a sample and all other points in the next nearest cluster. The Silhouette Coefficient for a collection of samples is the average of the Silhouette Coefficient for each sample. A model with better-defined clusters corresponds to a greater Silhouette Coefficient score. The score is between -1 and 1 for erroneous clustering and dense clustering, respectively. Scores near zero indicate cluster overlap (Rousseeuw, 1987).

3.6.1.3. Davies-Bouldin Index (DBI)

DBI calculates the average similarity between each cluster and its most similar cluster, considering each cluster's size and density into consideration. The Davies-Bouldin index is another metric that can be used to evaluate the model if the ground truth labels are unknown. This index represents the average "similarity" between clusters, where "similarity" is a measure that contrasts the distance between clusters to their own size. A model with a lesser Davies-Bouldin index has greater separation between clusters. The calculation of Davies-Bouldin scores is more straightforward than that of Silhouette scores, and the index is dependent solely on quantities and characteristics inherent to the dataset, as its calculation only employs point-wise distances (Davis & Bouldin, 1979).

3.6.1.4. Calinski-Harabasz Index (CHI)

If the ground truth labels are unknown, the Calinski-Harabasz index, also known as the Variance Ratio Criterion, can be used to evaluate the model. CHI is a metric that evaluates the ratio of the variance between clusters to the variance within clusters. The index is the ratio between the sum of dispersion between clusters and dispersion within clusters for all clusters. A greater Calinski-Harabasz score corresponds to a model with more precisely defined clusters. The score can be calculated more quickly than the Silhouette Coefficient (Calinski & Harabasz, 1974).

3.6.2. Classification

The objective of multiclass classification is to place each observation into one of multiple classifications. Several evaluation metrics can be used to determine the precision and effectiveness of a multi-class classification model when assessing its performance. Accuracy, precision, recall, the F1 score, and the confusion matrix are the most commonly employed metrics (Grandini et al., 2020).

3.6.2.1. Accuracy

A common metric used to evaluate the performance of a multi-class classification model is the accuracy of a classifier. Commonly, the accuracy score is used to compute the classifier's accuracy as the fraction or number of true predictions. Accuracy of a classification model is defined as the proportion of instances correctly classified relative to the total number of instances:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

where TP stands for a true positive, TN for a true negative, FP for a false positive, and FN for a false negative. In the presence of imbalanced datasets or misclassification errors, accuracy alone may not provide a complete picture of the classifier's performance (Grandini et al., 2020).

3.6.2.2. Confusion Matrix

Another crucial evaluation metric for multi-class classification is the confusion matrix. It computes the confusion matrix, with each row corresponding to the true class, in order to evaluate classification accuracy. The confusion matrix is a table that summarizes classification results, with each row representing the actual class and each column representing the predicted class. For each class, the matrix displays the number of true positives, true negatives, false positives, and false negatives. Other evaluation metrics, such as precision and recall, can be derived from the confusion matrix (Vujović, 2021).

3.6.2.3. Precision, Recall and F-1 Score

Other evaluation metrics used in multi-class classification include precision and recall. Precision is the classifier's inability to incorrectly designate a sample as positive when it is actually negative. Precision is defined as:

$$Precision = TP / (TP + FP)$$

Recall, on the other hand, is the classifier's ability to locate all positive samples, and it is defined as:

$$Recall = TP / (TP + FN)$$

The F-1 score is used to evaluate the tradeoff between precision and recall. The F1 score is a weighted average of precision and recall, with 1 representing the greatest performance and 0 the worst:

$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall)$$

When evaluating the performance of a multi-class classification model, it is crucial to consider multiple evaluation metrics in order to gain a comprehensive understanding of the model's efficacy. By combining metrics such as accuracy, precision, recall, F1 score, and confusion matrix, one can gain a deeper understanding of the model's strengths and weaknesses and make more informed decisions about how to enhance its performance (Grandini et al., 2020).

4. RESULTS AND DISCUSSION

The following section provides an in-depth overview of the study's extensive results, which were intended to assist the Asseco PST Data & Analytics team in integrating a machine learning solution into their existing CRM system. The stated objectives encompassed the segmentation of customers, integration of CRM for the optimization of campaigns, creation of a Recommendation System, and providing the solution's scalability. The subsequent subsections provide a comprehensive summary of the results obtained.

4.1. OVERVIEW

The dataset used in the study was extracted from the main system of an Angolan bank and consisted of 46008 unique active customers who had conducted at least one transaction in the previous 6 months. Several pre-processing processes were performed to guarantee data quality and suitability for analysis. These steps included data cleaning, missing value management, and data normalization. The resulting dataset was deemed suitable for analysis, fulfilling the project's requirements.

4.2. CUSTOMER SEGMENTATION

In this section, all the results of all the processes needed for customer segmentation will be discussed, starting from the k-means clustering, through predictive modelling of the resultant segments to the PowerBI reports and proposed CRM integration.

4.2.1. K-means clustering

For the customer segmentation objective, metrics including silhouette score, Calculated Calinski-Harabasz Index, Davis Bouldin Index, and inertia were used to evaluate the quality and distinctiveness of the generated segments. These metrics provided insight into the effectiveness of the clustering algorithms in developing meaningful and well-defined customer segments. Presented in the table below are the results of the k-means evaluation metrics.

Number of K	Silhouette Score	Calculated Calinski-Harabasz Index	Davis Bouldin Index	Inertia
2	0.3595	17367.8792	1.3278	100377.4734
3	0.4153	17886.0931	1.2192	83082.9195
4	0.4570	22892.6754	0.9910	68495.7178
5	0.4967	21430.6497	0.8778	59648.4372
6	0.5424	28130.4270	0.9007	52209.9947
7	0.5571	37259.7640	0.9076	45637.3703
8	0.5992	45248.3672	0.7415	39574.6226
9	0.5198	44794.5232	0.8300	35042.1147
10	0.6334	76886.1426	0.7040	30756.8592
11	0.5356	64759.1539	0.7149	28032.4986
12	0.6568	97051.4918	0.5808	25551.7391

Table 2 Clustering Performance Metrics for Customer Segmentation

As seen in the table 5.1, the greatest Silhouette Score was observed for k = 12, followed by k = 10 and k = 8. The greatest Calculated Calinski-Harabasz index was for k = 12 (97052.4918), followed by k = 10 and k =11. The lowest Davis Bouldin Index was observed in k = 12 (0.5808), and k = 10 (0.7040). The inertia kept on decreasing with the increase of number of clusters. According to the evaluation metrics the overall best number of clusters is 12, followed by 10.

To further asses the optimal number of clusters, it is valuable to look at the mean values of each variable for each cluster and assess the distinctiveness of the clusters for the business purpose. The table presented below shows the cluster centroids for k = 10 and k =12.

As observed in the table of centroids, k = 12 does not introduce new distinctive clusters that would be relevant for the design of marketing actions. Moreover, for the business purposes a smaller number of clusters would be easier to understand by the business users.

4.2.1.1. Description of clusters

To fully understand the segmentation of customers by the clustering algorithm and to evaluate the final clusters, it is necessary to define the clusters based on the values of the characteristics of their centroids. By doing so, descriptive names can be assigned to clusters and their characteristics can be described so that business users can better comprehend customer partitioning.

- Cluster 0 – “Young Females”

  This cluster represents female clients who are on average 26 years old. 100% of them are students, with 9% already having higher education. 97% of them are single. They are on average clients of the bank for 44 months and 2% have credit or overdraft.

- Cluster 1 – “Most Active”

  This segment represents clients who have best average values in terms of Recency, Frequency and Monetary Value. They are the largest segment and represent over 16 thousand clients. They are single males, without higher education who are on average 37 years old. They are on average with the bank for more than a year, and do not have any credit.
- Cluster 2 – “Female savers”

  This segment represents females who are single without higher education who are on average 35 years old. They have the greatest total transactions value out of all female segments, meaning that they spend the least. They are the second largest segment. They have the lowest average of total transactions value, meaning that they spend a lot.
- Cluster 3 – “Savers”

  This segment consists of 62 % males and 38% females who have the greatest average of total transactions value. They are in relationships and do not have higher education. They have the second greatest average available balance and the greatest average of total transactions value, meaning that they spend the least compared to other segments. They are also second when it comes to being the longest with the bank.
- Cluster 4 – “Young Males”

  They are the shortest with the bank, on average 42 months. They are male students who on average are 26 years old. They have second lowest average available balance.
- Cluster 5 – “Best Clients”

  They are smallest segment and represent clients who have the greatest average available balance. These are clients who have the highest average number of transactions and made their last transaction most recently. They are single males, who are on average 50 years old and have higher education. 15% of them have credit.
- Cluster 6 – “Least Active”

  They are the third largest segment. These are the clients who have made their last transaction on average 119 days ago. They also have the lowest average number of transactions (11). On average, they are with the bank for just over a year and their mean age is 36. They are single males, who do not have higher education. Although they do not have any credit, they display

the third lowest average of transaction value indicating that they could have unauthorized debt.

- Cluster 7 – “Medium Value Females”

  These are females who are on average 41 years old and have higher education. 66% of them are single. They tend to spend a lot (second lowest average of total transactions value) and have close to overall values for the frequency and recency of transactions, as well as average available balance.

- Cluster 8 – “Medium Value Males”

  This cluster represents males who are on average 39 years old, are single and have higher education. They have the second greatest total monetary transaction value but average available balance. They make frequent transactions but average when it comes to recency.

- Cluster 9 – “Most Loyal”

  These are the clients who have been the longest with the bank on average (over 2 years). All of them have credit and therefore have the lowest average available balance. Their last transaction was on average most recent (6 days ago) and their transaction frequency is close to the overall average. 11.4 % of them are in relationships, and 13.5 % are females. They are on average 44 years old.

4.2.2. PowerBI reports

The purpose of using PowerBI reports is to visualize the clustering process's outcomes. This enables business users to quickly and easily comprehend the customer segments. It facilitates the comparison of clusters, an understanding of segments in relation to all customers, and the examination of each segment in greater detail.

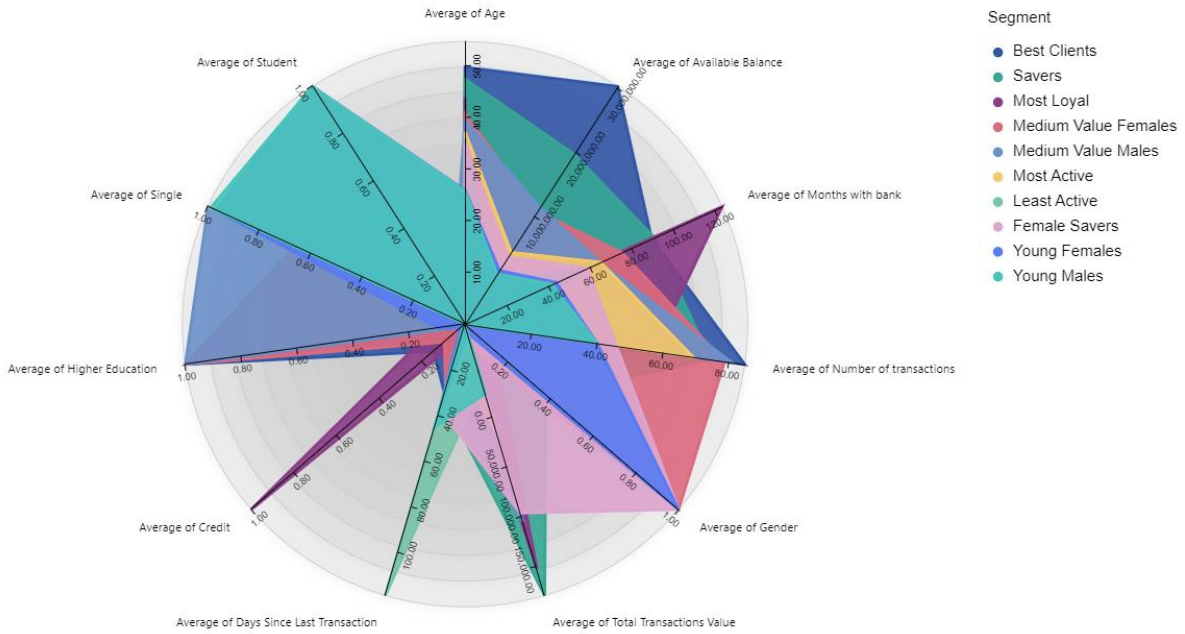

Figure 2 Radar Chart

The initial visualization implemented for the PowerBI reports is a radar chart that provides a visual overview of the distinctions between the client features segments. It enables rapid comprehension of segment characteristics and the ability to filter through each segment.

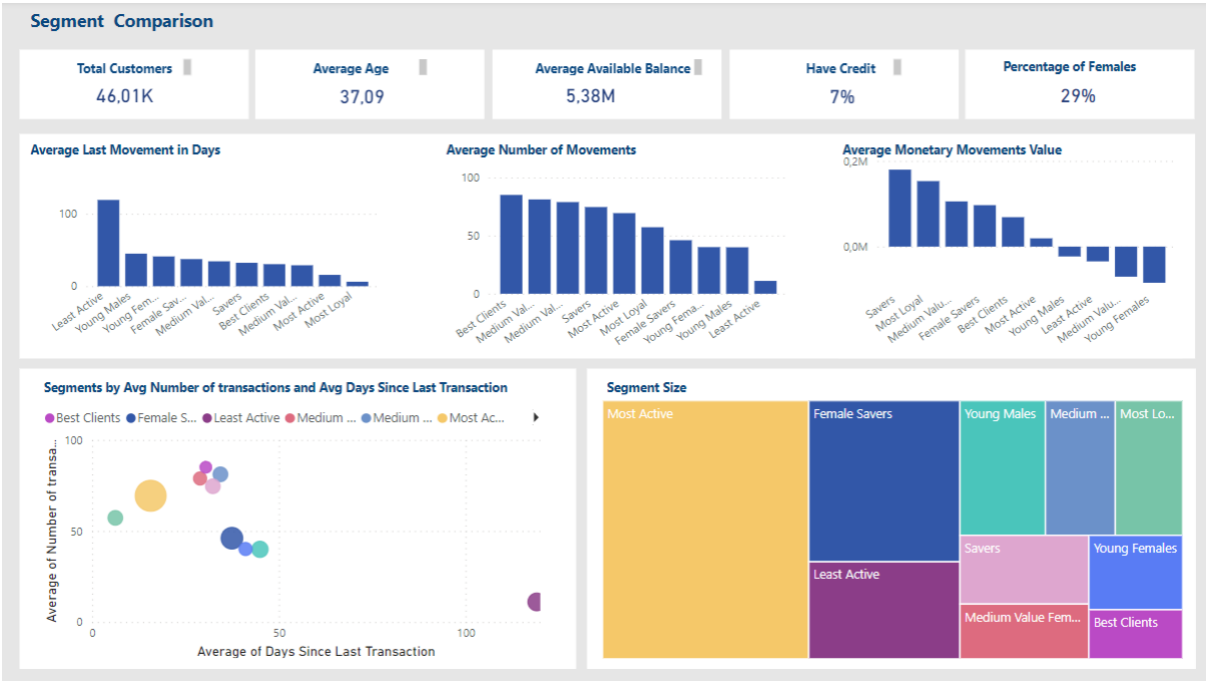

Figure 3 Segment Comparison

As its name suggests, the segment comparison report focuses on the differences between the segments. It displays an overview of the bank's customers, which can be filtered by selecting the segment name in any of the other visuals on the page. It compares Recency, Frequency, and Monetary values across segments, as Last Movement in Days, Number of Movements, and Value of Monetary Movements. The scatter plot displays the average Number of transactions by Days since the last transaction for each segment, with the size of the bubbles representing the cluster size. And in the bottom right corner, a treemap was used to quickly visualize the differences in segment proportions, as well as to enable fast segment selection for report page filtering.

Figure 4 Segment in Detail

Lastly, the Segments in Detail page provides a deeper understanding of each segment. The slicer enables selection of the segment name that filters out all of the page's visuals. The leftmost section displays the segment's average values for all numerical features. Pie charts illustrate the distribution of binary variables. The remaining visuals aid in a more thorough comprehension of the distribution of variables related to client value for the bank.

4.2.3. Predictive modeling

In order to avoid repeating clustering solution and continuously redefining the clusters, a semi-supervised learning approach was taken. Thus, the cluster membership can be continuously updated and monitored, as well as cluster labels can be easily predicted for new customers. The output data of the clustering solution with cluster labels was used to train a multi-label classification algorithm.

The proposed models were, as mentioned in section 3.5.2 were Support Vector Machine (SVM), K-nearest Neighbors (KNN), Decision Tree (DT) and Random Forest (RF).

To choose the best model hyperparameters a 10 k-fold random search validation technique was used, which resulted in the following hyperparameters being used to train the selected models shown in the table below:

Model	Best Hyperparameters
SVM	tol: 0.001, shrinking: False, random_state: 42, probability: True, max_iter: -1, kernel: linear, gamma: scale, degree: 2, coef0: 1, class_weight: balanced, C: 100
KNN	weights: distance, p: 1, n_neighbors: 5, metric: minkowski, leaf_size: 16, algorithm: kd_tree
DT	Splitter: best, random_state: 42, min_samples_split: 5, min_samples_leaf: 1, max_features: sqrt, max_depth: 50, criterion: entropy
RF	Random_state: 42, n_estimators: 200, min_samples_split: 2, min_samples_leaf: 2, max_features: sqrt, max_depth: 50, criterion: gini, bootstrap: True

Table 3 Best hyperparameters for the chosen models

4.2.4. Evaluation of model performance

The models were trained on 70 % of the dataset and tested on the remaining 30%. Below are presented the results of evaluation of each model based on the test dataset:

Model	Accuracy	Precision	Recall	F1-score
SVM	0.999493	0.999298	0.999498	0.999493
KNN	0.996233	0.995841	0.997415	0.996614
DT	0.990002	0.986592	0.985958	0.986236
RF	0.997839	0.997839	0.998229	0.998033

Table 4 Evaluation Metrics for each model

As shown in table 5.3, all of the models perform exceedingly well on the test dataset, with above 99% accuracy, thus the f1-score was selected as the primary metric for evaluation. Compared to other models, SVM has the highest f1-score (0.999493), but the distinctions between them are minimal. Decision tree was the model with the lowest f1 score. Since the purpose of the model is to assign cluster labels, model explainability is not of the utmost significance; therefore, the best performing model is selected. SVM has the highest values across all evaluation metrics and was therefore chosen as the best model for future predictions.

For a better understanding of the predictions made, the confusion matrix of the chosen best model, Support Vector Machine, is shown below; it reveals that only cluster labels 0, 1, and 6 exhibited prediction errors. The most significant error occurred when the true label was 0 and the predicted label was 7. However, the errors are not particularly significant, as they occurred in only 0.47 percent of instances.

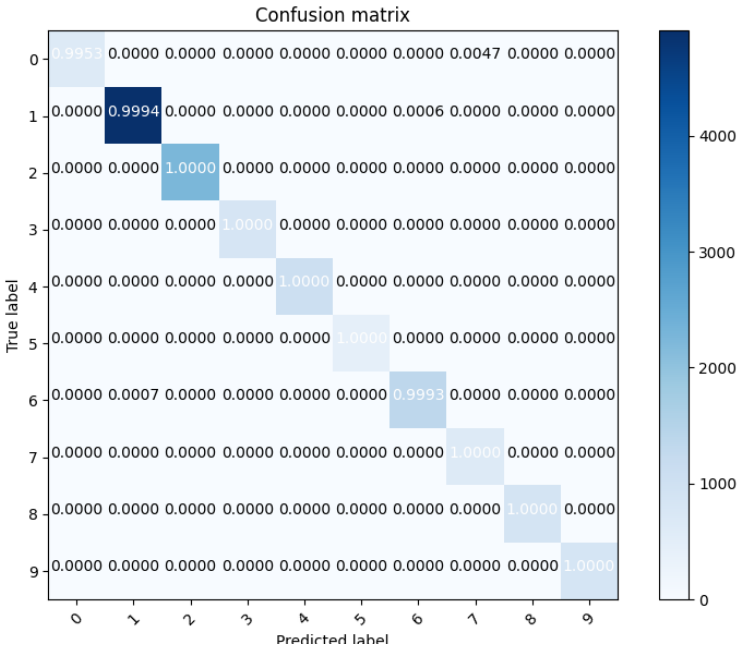

Figure 5 Confusion Matrix or Support Vector Machine model

Moreover, based on all of the evaluation metrics of all of the models used, it appears that the clustering was effective in separating the data points into distinct clusters, and that the models were able to recognize the pattern that the clustering algorithm used to segment the consumers.

4.3. RECOMMENDATION SYSTEM

As described in section 3.5, it was proposed for the recommendation system to employ a clustering-based strategy, based on the premise that similar clients would select similar products. To test this, a clustering procedure was conducted using the same dataset in conjunction with data on product ownership count, where the columns represent the product, and the values represent the product count. Then, each product is assigned a score for each customer cluster based on the clustering results. To further evaluate the approach, a classification model that utilizes only customer characteristics and excludes product ownership data has been developed to assist in the recommendation of products to new customers.

4.3.1. K-means clustering

The first step of the development of the recommendation system was to conduct K-means clustering on the client features that were previously used for customer segmentation as well as the product ownership values. To evaluate the optimal number of clusters (k), visual assessment was performed first. The data was visualized in a 2-dimensional space using PCA method for dimensionality reduction. This was done to observe how the data points naturally group. As seen in fig 5, the data points tend to group themselves into 16 clusters, with a clear separation in the middle.

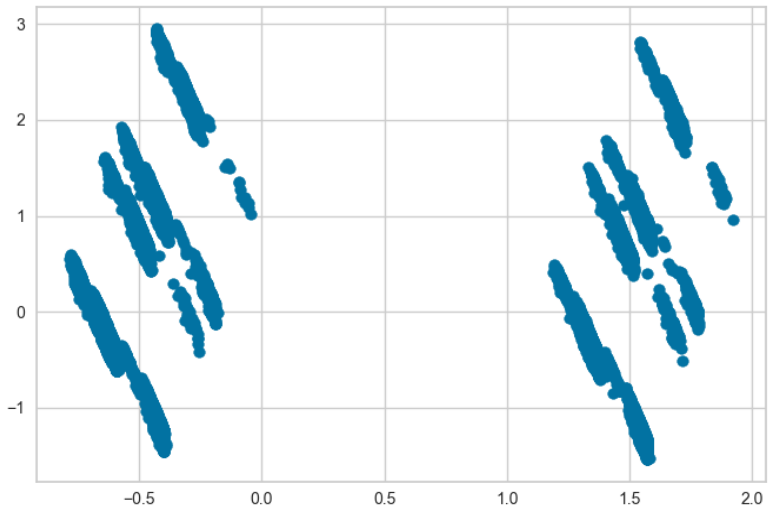

Figure 6 Visualization of 2-D Space after PCA dimensionality reduction

Another visual method for assessment of the optimal number of k was performed using elbow method, using distortion score. The results suggest an optimal elbow at k =6.

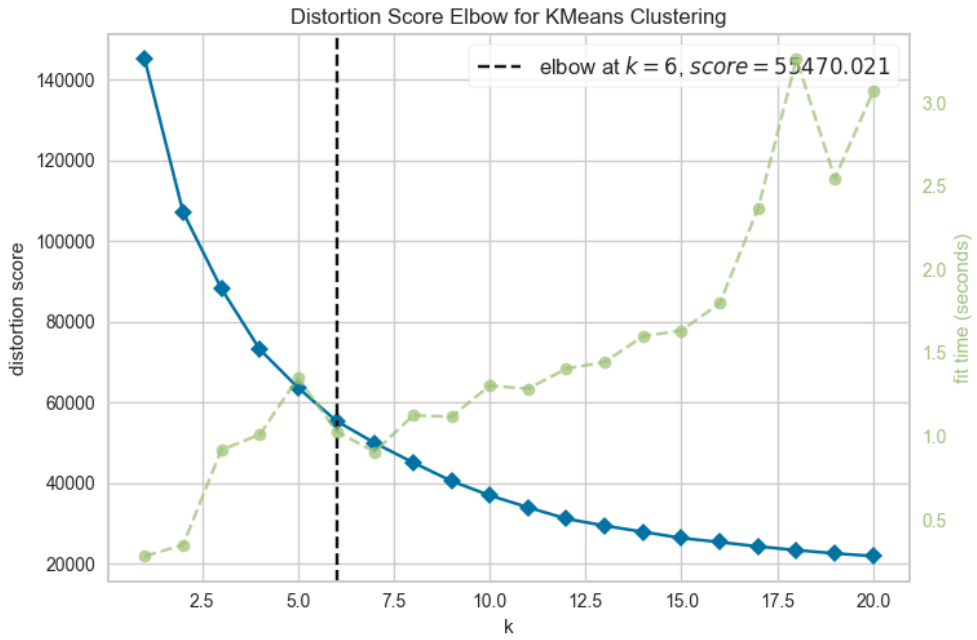

Figure 7 Elbow method graph

As done previously, the clustering performance was evaluated using the Silhouette Score, Calculated Calinski-Harabasz Index, Davis Bouldin Index and Inertia metrics.

Number of K	Silhouette Score	Calculated Calinski-Harabasz Index	Davis Bouldin Index	Inertia
2	0.3355	16261.3475	1.3909	107255.4535
3	0.3871	16579.7617	1.2684	89831.3065
4	0.4274	18027.1979	1.0498	74767.2700
5	0.4728	20631.6976	0.9747	64909.5577
6	0.5047	31873.2142	0.8849	58051.8035
7	0.5159	23723.8839	0.9501	51962.1352
8	0.5398	31815.7208	0.8001	46920.0678
9	0.4603	35379.6711	0.8211	42341.1166
10	0.5894	45532.4880	0.7142	38857.1720
11	0.5425	55009.2698	0.7264	35186.9485
12	0.6340	107893.5779	0.6747	33119.6038
13	0.5047	73718.5880	0.6739	30383.5140
14	0.6495	92285.9891	0.7266	28515.4207
15	0.5093	64981.1985	0.8160	26869.6980

16	0.6775	111117.1058	0.5931	25679.5209
17	0.6493	90701.4262	0.6266	25907.9833
18	0.6337	106190.2001	0.6959	25158.7823
19	0.6758	112393.6620	0.5978	23765.9925
20	0.6464	151717.1926	0.5503	21670.3437

Table 5 Clustering Evaluation Metrics

According to table 5.4, k = 16 had the highest Silhouette Score, followed closely by k = 19. The highest calculated value of the Calinski Harabasz Index was k = 20, followed by k =19 and k =16. The lowest calculated Davis Bouldin Index was k = 20, followed by k = 16. The decrease in inertia was proportional to the increase in the number of k. Across all clustering performance metrics, k = 16 clustering was the most optimal.

Even though the elbow method suggested k = 6 as the optimal number of clusters, for this segmentation, k = 16 was chosen as the optimal number of clusters due to its superior performance across all of the previously mentioned metrics and the visual analysis of data points after PCA reduction suggesting 16 clusters.

To reaffirm the choice of k, the centroids table for k = 16 was exhaustively investigated and showed that the values of cluster centroids across the customer features and product counts were distinct.

4.3.1.1. Description of clusters

After the clustering was performed, it was necessary to evaluate the results across the distinctiveness in client features. In Appendix B can be found the centroid values of each cluster.

Based on their characteristics, the clusters were then given names according to their characteristics to be more easily understood by the business users:

- Cluster 0 – “Best Customers”
- Cluster 1 – “New Customers”
- Cluster 2 “Least Active Females”
- Cluster 3 – “Male Savers”
- Cluster 4 “Most Loyal”
- Cluster 5 “Female Students”
- Cluster 6 “Least Active Males”
- Cluster 7 “Female Big Spenders”

- Cluster 8 – “Female Savers”
- Cluster 9 – “Moderately Active”
- Cluster 10 – “Active Male Students”
- Cluster 11 – “Higher Education Single Females”
- Cluster 12 – “Least Active Big Spenders”
- Cluster 13 – “Greatest Balance Females”
- Cluster 14 – “Future Best Male Clients”
- Cluster 15 – “Future Best Female Clients”

The customer segments which were results of the clustering show distinctiveness across the client characteristics as well as product ownership. The matrix (see figure 8) below represents the total product counts for each cluster.

Product	Active Male Students	Best Customers	Female Big Spenders	Female 'Savers'	Female Students	Future Best Female Customers	Future Best Male Customers	Greatest Balance Females	Higher Ed. Single Females	Least Active Big Spenders	Least Active Females	Least Active Males	Male 'Savers'	Moderately Active	Most Loyal	New Customers	Total
Conta D/O (AKZ)	712	992	522	2241	574	651	3445	539	781	1460	465	211	1662	1128	1982	2773	20138
Depósito à Ordem BANKITA (AKZ)	2014	31	15	2249	1388	110	1028	32	153	1824	902	671	199	127	74	3925	14742
Conta Salário (AKZ)	109	171	36	1463	78	212	2193	108	294	717	159	16	609	380	135	3271	9951
Conta Plus RE24 (USD)	13	49	30	103	9	46	247	23	35	39	22	1	81	69	127		894
Conta D/O (USD)	8	82	30	48	3	40	219	28	30	34	17	4	72	89	148	2	854
Conta Ordem (AKZ)	5	136	57	43	9	50	88	41	52	13	3	1	80	96	106	10	790
Colateral Cash (AKZ)	3	83	21	42	5	42	65	34	54	8	10		53	59	53		532
Dp. Ordem Funcionário Público (AKZ)	9	17		81	4	12	117	6	20	12	4		47	14	13	21	377
Conta Júnior D/O (AKZ)	49			51	83	8	27			40	25	12	4	7	2	54	362
Conta Colaboradores BNI (AKZ)	3	21	15	9	1	4	18	15	15				23	13	38	9	184
BNI JÚNIOR (AKZ)	23	3		26	14	6	14		2	23	23	8		2		27	171
Conta Ordem (USD)	1	38	6	11		1	18	13	2				15	30	27		162
Conta Ordem Plus (USD)		31	18	6		5	26	8	6	3			14	15	23		155
Conta D/O - Não Residente (AKZ)		22		7		4	1	3	4	4		1	31	12		59	148
Conta D/O (EUR)	1	27	7	12	5	8	17	6	15				16	7	18	2	141
Depósito a Prazo (AKZ)	1	29	10	14	3	5	15	6	3	5	3		9	8	1	9	121
Conta à Ordem Simplificada (AKZ)	31	1		20	11	2			1	2	2	4	9		1	24	108
D/P BNI Sobre Rodas (AKZ)	4	11	2	11	2	4	15	6	12				15	10	4	7	103
D/P Penhor (AKZ)		18	14		1			7					1		45		86
Colateral Cash (USD)		13	2	5		7	9	15	1	1			10	6	10		79
Conta Ordem (EUR)		22	2	3	4	1	7	1	2				10	6	7		65
Crédito Rendas Pessoal - RE24 (AKZ)	1	1	9					1							45		57
D/P Penhor - CARC (AKZ)		4	6	2		1	1						5	1	26	2	48
D/P BNI 16 Anos - Juros Postic(AKZ)	3	3		12		4	6	1	1				5	2		4	41
Conta Plus RE24 (EUR)		4	2	2	2		9	5	2				3	2	9		40
Conta Interna BNI - Subs.Falha(AKZ)		2	7	2			2		3				10	1	7	1	35
Conta Ordem Plus (EUR)		4	1			7	3	3		1			2	2	5		28
DP BNI - NET (AKZ)	1	5		1		4	3	2	1				3	4		1	25
Colateral Cash (EUR)						3	5	1	2	1			2	2	5		21

Figure 8 Product Ownership across clusters

As can be seen, the most common products include Conta D/O (AKZ), Deposito an Ordem Bankita, and Conta Salario. There are however variations in the product quantity distribution across the clusters. Conta D/O (AKZ) is the most popular product among Future Best Male Clients, and the least popular among Least Active Males. New Customers and Active Male Students are the most likely to own a Bankita product, while Female Big Spenders, Best Customers, and Greatest Balance Females are the least likely. As an example, USD products such as Conta Ordem USD or Conta Ordem Plus USD are

primarily selected by Best Customers, Future Best Male Customers, Moderately Active, or Most Loyal customers, whereas they are unpopular among segments that are least active, have students, or New Customers.

4.3.2. Predictive modeling

As was done previously, a predictive model was created to classify the cluster labels for new customers and reassign existing customers to their respective clusters each month. Due to the large number of clusters, it was not possible to train the Support Vector Machine model at this time because it demanded an excessive amount of time and memory available to train. K-nearest neighbours, decision tree, and random forest were selected as potential models. As has been demonstrated beforehand, a random search with 10 k-fold validation was used to identify the optimal hyperparameters for each model (see table 5.5).

Model	Best Hyperparameters
KNN	weights: distance, p: 1, n_neighbors: 13, metric: l2, leaf_size: 40, algorithm: auto
DT	splitter: best, random_state: 123, min_samples_split: 2, min_samples_leaf: 2, max_features: auto, max_depth: 50, criterion: gini
RF	random_state: 42, n_estimators: 100, min_samples_split: 2, min_samples_leaf: 2, max_features: auto, max_depth: 20, criterion: entropy, bootstrap: False

Table 6 Best Hyperparameters of each model

Each model was trained on the training dataset of 70% of the data and tested on the remaining 30%. The evaluation was performed on the test dataset using accuracy, precision, recall and f1-score metrics.

Model	Accuracy	Precision	Recall	F1-score
KNN	0.987960	0.988467	0.993749	0.991025
DT	0.964967	0.957504	0.955259	0.955259
RF	0.990571	0.993677	0.995232	0.994442

Table 7 Evaluation metrics of each model

All models performed remarkably well on the test dataset, indicating that the cluster partitioning was of high quality. Decision Tree was the model with the lowest performance metric values. Random Forest, which performed the best across all evaluation metrics with a 99% accuracy and an average F1 score of over 99%, was selected as the model that will be used to predict cluster labels in the future.

4.3.3. Calculating Recommendation Scores

As described in section 3.5, based on the collective ownership of products within each cluster, recommendations are made. The recommendation score for each product is calculated using variables such as the number of products owned, the cluster weight, and the popularity weight. The scores are then normalized using min-max standardization to ensure comparability across clusters, resulting in product recommendation scores for each cluster.

The figure below shows a matrix of products and scores, where the values represent the recommendation score in the range of 0 to 1 for each cluster.

Product	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
BNI JÚNIOR (AKZ)	0,00	0,01	0,10	0,00		0,03	0,01		0,03	0,00	0,01	0,00	0,04		0,01	0,01
Colateral Cash (AKZ)	0,44	0,00	0,00	0,11	0,09	0,00		0,28	0,03	0,17	0,00	0,23	0,00	0,41	0,03	0,27
Colateral Cash (EUR)	0,05			0,01	0,02	0,00		0,02	0,00	0,02		0,01	0,00	0,01	0,00	0,06
Colateral Cash (USD)	0,15		0,00	0,02	0,02			0,01	0,00	0,06		0,02	0,00	0,20	0,01	0,12
Conta à Ordem Simplificada (AKZ)	0,00	0,01	0,00	0,00	0,00	0,01	0,00		0,01		0,03	0,00	0,00			0,00
Conta Colaboradores BNI (AKZ)	0,09	0,00		0,02	0,13	0,00		0,46	0,01	0,01	0,00	0,04		0,11	0,01	0,02
Conta Colaboradores BNI (USD)	0,03			0,00	0,06	0,00		0,24	0,00	0,00	0,00	0,01		0,04	0,00	0,01
Conta D/O (AKZ)	0,88	0,32	0,21	1,00	1,00	0,12	0,06	0,75	0,80	1,00	0,09	1,00	0,46	1,00	1,00	1,00
Conta D/O (EUR)	0,11	0,00	0,00	0,05	0,03	0,00	0,00	0,06	0,01	0,08	0,00	0,03	0,00	0,07	0,02	0,07
Conta D/O (USD)	0,28	0,00	0,01	0,11	0,23	0,00	0,00	0,12	0,03	0,26	0,00	0,08	0,01	0,19	0,17	0,22
Conta D/O - Não Residente (AKZ)	0,05	0,02		0,04			0,00		0,00	0,01		0,00	0,00	0,00	0,00	0,00
Conta Interna BNI - Subs.Falha(AKZ)	0,00	0,00		0,01	0,05			0,16	0,00	0,00	0,00	0,01		0,03	0,00	0,01
Conta Júnior D/O (AKZ)	0,00	0,01	0,07	0,00	0,00	0,14	0,02		0,03	0,00	0,03	0,00	0,03		0,00	0,01
Conta Ordem (AKZ)	1,00	0,00	0,00	0,08	0,31	0,00	0,00	1,00	0,01	0,20	0,00	0,11	0,00	0,44	0,02	0,12
Conta Ordem (EUR)	0,29			0,01	0,05	0,00		0,06	0,00	0,04		0,01	0,00	0,06	0,00	0,02
Conta Ordem (USD)	0,62			0,02	0,09	0,00		0,14	0,00	0,15	0,00	0,02	0,00	0,09	0,01	0,05
Conta Ordem Plus (USD)	0,18		0,00	0,01	0,04	0,00		0,10	0,00	0,02	0,00	0,01	0,00	0,06	0,01	0,02
Conta Plus RE24 (EUR)	0,02		0,00	0,01	0,01	0,00		0,03	0,01	0,02		0,02	0,00	0,02	0,01	0,00
Conta Plus RE24 (USD)	0,14		0,02	0,12	0,16	0,00	0,00	0,18	0,07	0,18	0,00	0,12	0,01	0,16	0,21	0,17
Conta Salário (AKZ)	0,05	0,98	0,05	0,24	0,01	0,00	0,00	0,01	0,63	0,19	0,00	0,23	0,22	0,06	0,78	0,19
Crédito Rendas Pessoal - RE24 (AKZ)	0,00				0,29			0,18			0,00			0,00		
D/P BNI 16 Anos - Juros Postic(AKZ)	0,01	0,00		0,01	0,01			0,00	0,01	0,01	0,00	0,02		0,01	0,00	0,03
D/P BNI Sobre Rodas (AKZ)	0,04	0,00		0,05	0,01	0,00		0,00	0,02	0,02	0,00	0,06	0,00	0,07	0,01	0,02
D/P Penhor (AKZ)	0,26	0,00		0,00	0,19	0,00		0,40	0,00	0,00		0,00	0,00	0,08	0,00	
D/P Penhor - CARC (AKZ)	0,02	0,00		0,00	0,10			0,06	0,00	0,00	0,00	0,00		0,01	0,00	0,00
Depósito à Ordem BANKITA (AKZ)	0,00	1,00	1,00	0,02	0,00	1,00	1,00	0,00	1,00	0,01	1,00	0,04	1,00	0,00	0,10	0,03
Depósito a Prazo (AKZ)	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,04	0,01	0,02	0,00	0,01	0,00	0,01	0,01	0,04
Dp. Ordem Funcionário Público (AKZ)	0,01	0,00	0,00	0,04	0,00	0,00		0,00	0,05	0,01	0,00	0,03	0,00	0,00	0,06	0,02

Segment Code	Segment Name
0	Best Customers
1	New Customers
2	Least Active Females
3	Male 'Savers'
4	Most Loyal
5	Female Students
6	Least Active Males
7	Female Big Spenders
8	Female Savers
9	Moderately Active
10	Active Male Students
11	Higher Ed. Single Females
12	Least Active Big Spenders
13	Greatest Balance Females
14	Future Best Male Customers
15	Future Best Female Customers

Figure 9 Recommendation scores matrix for each cluster and product

The most recommended products are Deposito a Ordem BANKITA for clusters 1, 2, 5, 6, 8, 10 and 12, and Conta D/O (AKZ) for cluster 3, 4, 9, 11, 13, 14 and 15. The products with the lowest recommendation scores are Conta a Ordem Simplificada (AKZ), Conta Interna, Conta Plus RE24 (EUR), D/P BNI 16 Anos – Jutos Postic (AKZ) and Dp. Ordem Funcionario Publico (AKZ).

4.3.4. PowerBI Reports

The PowerBI Reports were developed to provide an in-depth understanding of the consumer segments' product ownership. It also enables the differentiation of customer groups according to their preferred products. The first report identifies the most popular products by segmenting the products. The second page provides a cluster overview that facilitates comprehension of each segment in relation to its average characteristics and product ownership. The product detail page displays the average features of customers who own the selected product as well as the customer segments who own the product. Additionally, product scores matrix was provided for easy identification of best scored products for each customer segment (see Figure 9).

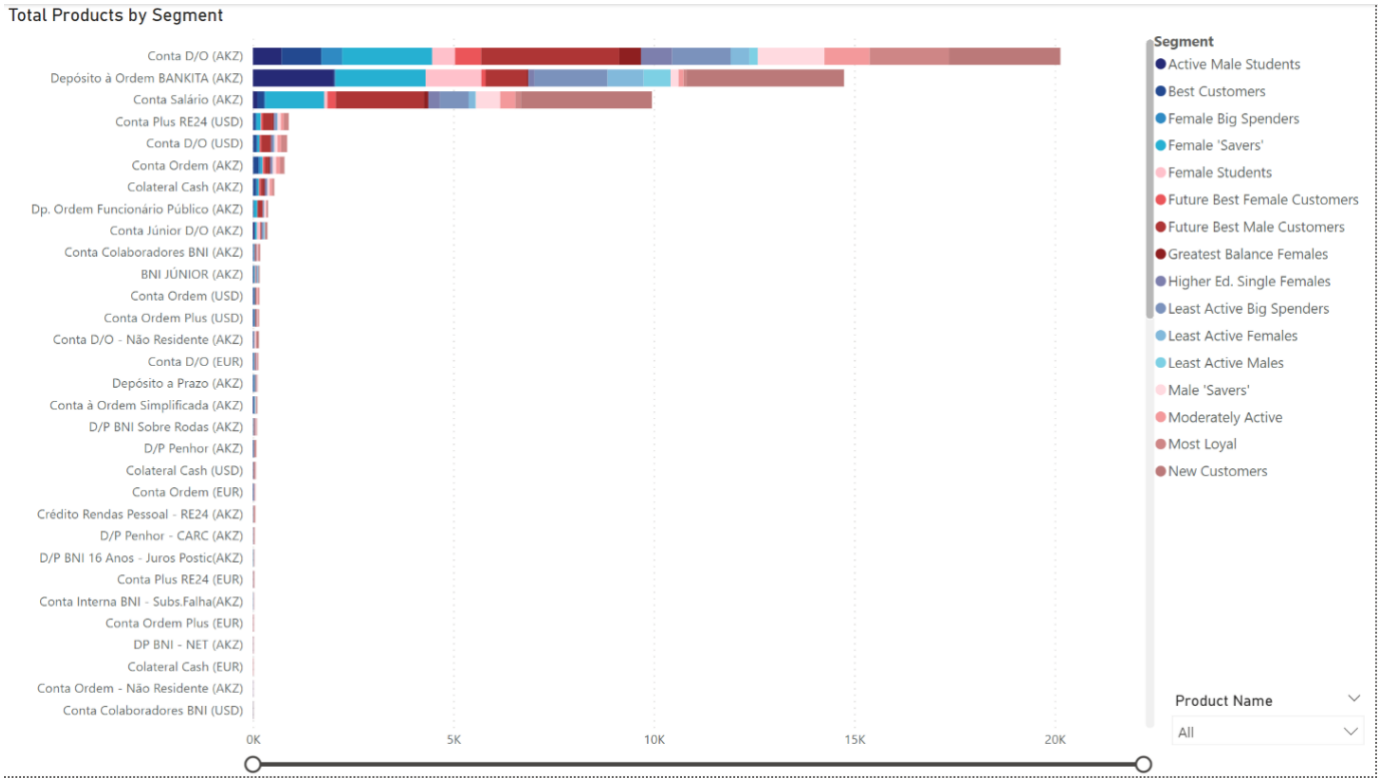

Figure 10 Total Products by Segment PowerBI report

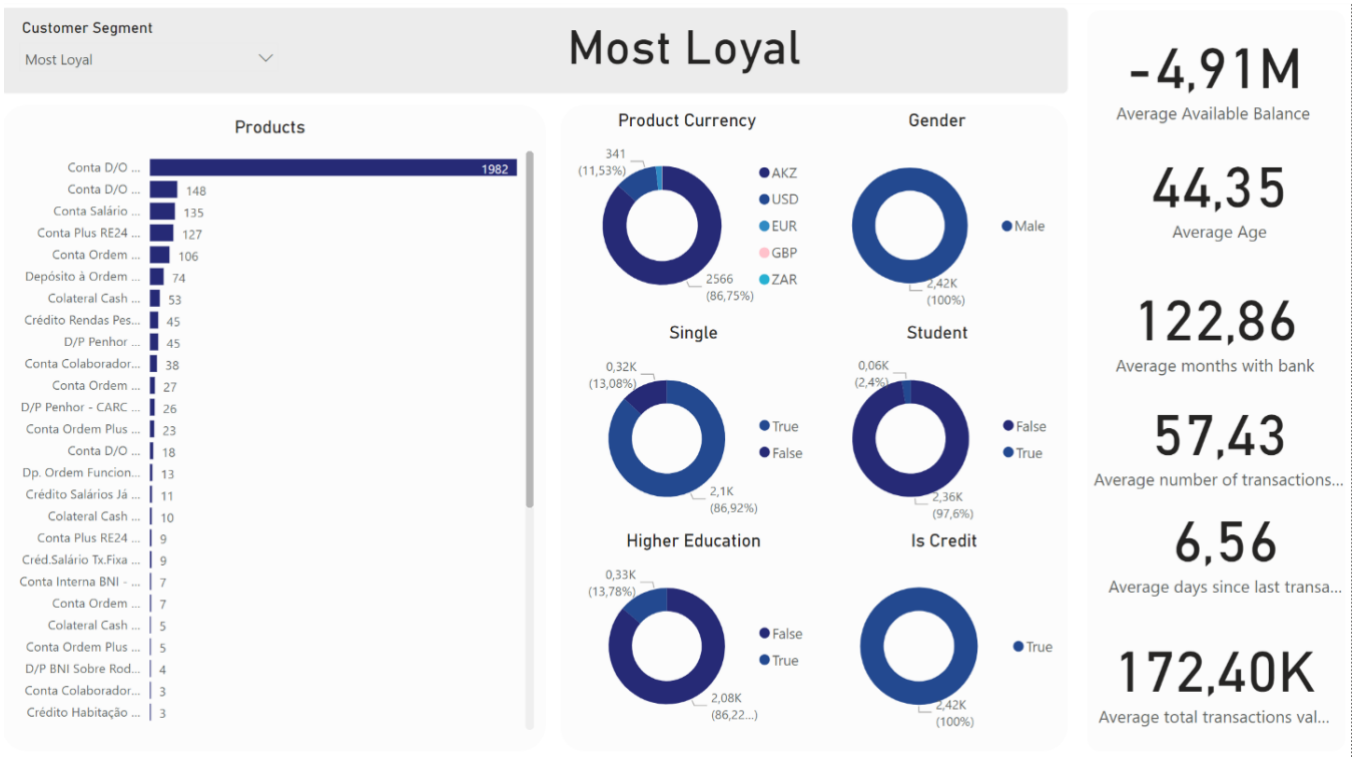

Figure 11 Segments in detail PowerBI report

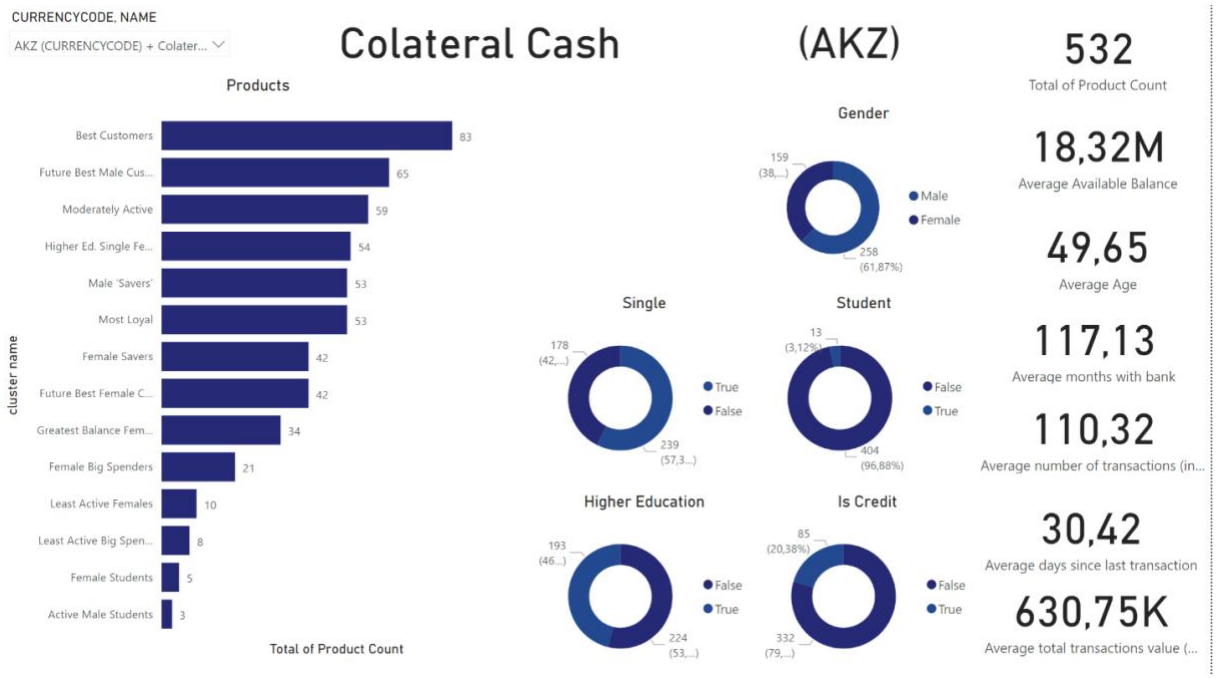

Figure 12 Products in detail PowerBI report

4.3.5. Item-based collaborative filtering

As introduced in section 3.5, item-based collaborative filtering is a method that utilizes historical user behaviour data to recommend items similar to those a user has shown interest in, based on the cosine similarity between their feature vectors.

The product ownership data was used to construct a product-product similarity matrix with calculated cosine similarity scores. The results can be presented in the appendix.

The results indicate that there is not much similarity within the products as there no frequent occurrences of cosine similarity values greater than 0.5. (in the case of Junior DO AKZ and BNI Junior AKZ = 0.56). The results do not meet criteria to be used for product recommendation, however they can be utilized to understand the products better. Moreover, the approach should be tested on a dataset of a different client, as the results might differ.

4.4. ADDITIONAL FINDINGS

At the stage of data understanding, it became clear that only a small percentage of bank consumers have been active within the past six months. This indicates that action is required to prevent loss of clients.

While analysing the data on product ownership, it became apparent that only 25% of the 120 products are owned by more than 100 clients. This suggests that it may be necessary to reduce the number of products available, as they are not popular among customers. Moreover, some products are only owned by a small number of customers or a single customer, and it may be advisable to contact these customers to offer them a more popular product and remove it from the Bank's offer.

4.5. LIMITATIONS

The first limitation of the study is the quality of the available data, where some of the variables that could be significant for marketing purposes (like income) were not appropriate for the analysis. Secondly, the dataset represents just one of the clients of Asseco PST, which does not guarantee the success of the solution for other clients.

Regarding the recommendation system, there was no information provided on the product characteristics and the types of consumers to whom they are directed, which would have enriched the solution and facilitated the testing of the results. Due to the nature of the recommendation system, it is difficult to evaluate the success of the proposed solution using historical data. Therefore, it is

suggested that the Recommendation System's performance is evaluated following its implementation and modified accordingly.

4.6. SUMMARY OF THE RESULTS

In summary, the study successfully achieved its aims. The utilisation of machine learning in consumer segmentation has allowed the identification of distinct segments, hence enabling the implementation of focused marketing strategies. The opportunity for enhancing campaign optimisation lies in the integration of the result segments within the CRM system. The implemented Recommendation System has facilitated the provision of tailored product recommendations, hence offering the opportunity to enhance customer engagement.

The evaluation of model performance included metrics that demonstrated the efficacy of the segmentation and recommendation models, resulting in satisfactory outcomes. The inclusion of additional findings enhanced the understanding of the dataset and consumer behaviour, hence paving the way for future research opportunities.

The study's findings make a significant contribution to the improvement of Asseco PST's CRM offering. This is achieved through the use of a machine learning framework that enables personalised marketing campaigns. Additionally, the study provides important insights that can inform decision-making processes.

5. DEPLOYMENT

This section will focus on the proposed deployment of the solution. It will present solutions developed for K-means clustering, Predictive classification modelling and the proposed recommendation system that were designed to offer scalability and repeatability of them to other clients and projects that could benefit from the same models used. In order to make the solutions scalable, and easy to use, for each machine learning method used, separate applications were developed, using python programming and communicating with the solution through REST APIs. This section will explain the design of the clustering, predictive modelling, and recommendation applications in more detail.

5.1. SEGMENTATION SOLUTION

This solution as presented, can be used to segment customers, but is not limited to it, and can expand to any project that could benefit from data points segmentation.

5.1.1. Solution Architecture Overview

The Solution Architecture consists of 3 layers: Data Layer, Machine Learning Layer and Results Layer, presented in the chart below:

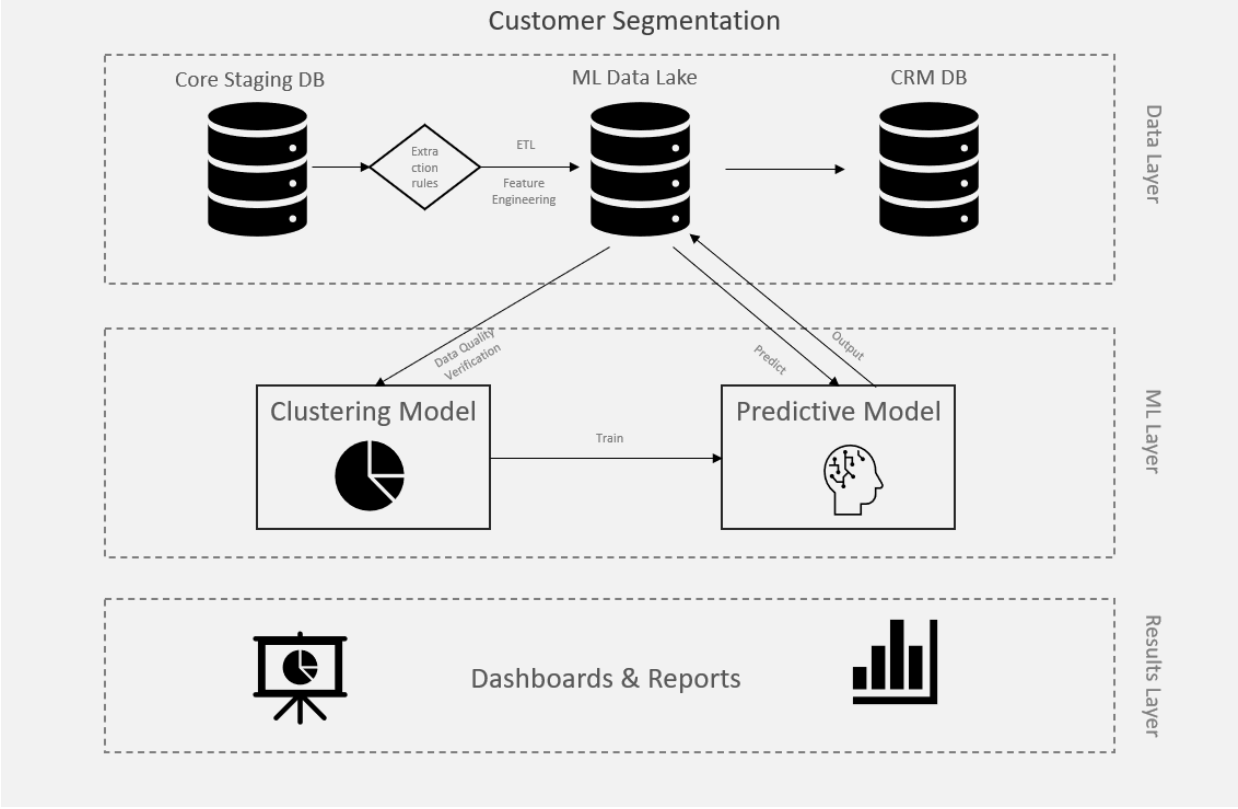

Figure 13 Segmentation Solution Architecture

5.1.2. Data Layer

Data Layer covers the data pipeline for the Machine Learning Models, from the Data Source, Machine Learning Data Store to the CRM Database. In this layer all the necessary processes for data collection and pre-processing as well as feature engineering take place so that the data is already prepared for Machine Learning modelling. In the proposed architecture, the source of the initial data comes from the Core Staging database. The data is then extracted and transformed using an SQL Query utilizing SQL Server Integration Services tools. The transformed data is then loaded into a customer segmentation table within a Machine Learning Data Lake. The Machine Learning Data Lake is an unstructured database created to store the data prepared for machine learning modelling, as well as the outputs of the ML models. It is designed for lab purposes of ML modelling to not interfere with the production environment and serves as an intermediary environment between the ML layer and the CRM database, used for the operation processes in the production environment.

5.1.3. Machine Learning Layer

Clustering Modelling and Predictive Modelling, which can both be used independently of one another, make up the Machine Learning Layer.

The main part of this layer, clustering modelling, needs to be run only once and updated as needed (for example, once a year). To specify the number of clusters and assess the clustering outcomes, the solution needs the assistance of a data analyst. Segment labels for each client are the product of the solution.

5.1.3.1. Clustering Solution

First, the clustering project receives the prepared data from the Machine Learning Data Lake, which is then examined, and its quality verified during the first stage of the project. At this point, any duplicates or missing values will be eliminated. It then performs the clustering model iterations and aids the data analyst in determining the ideal number of clusters. The best clustering model is selected by the data analyst, saved, and a cluster label (new variable with numerical labels) is assigned to each customer. The customer segmentation output table in the ML Data Lake stores the segmentation result. The data analyst's job is to assess the clustering results and then produce reports with names and descriptions of the newly created segments.

Application Development

The Clustering Application is designed for easier and faster application of clustering. The technologies and python packages used include Flask for python web framework, SQLAlchemy to communicate with databases, Scikit-learn for K-means clustering, Yellowbrick, Matplotlib, Seaborn and Dataframe -image for visualization, Pandas for dataframe operations and Pickle to save models.

The project space consists of main file for running the application, configurations file written in yaml, and clustering module that contains the functions and objects and their methods used to perform clustering.

Within the configuration file, the user can specify the connections to the database or choose to connect to a csv file, which columns need to be dropped, where to store visuals and models, as well as where to store the output of the clustering.

Except for pre-processing and analysing of the data, the clustering project follows Object-Oriented Programming approach for k-means clustering. A Parent class object has been created with its own methods, as well as children objects for evaluation of clustering.

The application reads the dataset, then drops duplicates, and normalizes the data. It then follows five steps that need to be performed for clustering, and each one of them has its own route, which will be explained in more detail.

1. **Analyse**

This step is used to re-assess the previously prepared dataset, and it outputs summary statistics table and correlation tables for the user to check if the data has been prepared correctly, and the right variables have been chosen for modelling.

2. **Choose K**

This step takes in request from the user of the maximum number of clusters. It then outputs an elbow graph as well as PCA – reduced graph to assess within what range of K the clustering seems the most optimal, and through the PCA method graph to observe whether the data points form groups naturally within a two-dimensional space.

3. **Metrics**

This step allows for the user to re-submit the maximum number of clusters based on the results provided in the previous step. It calculates metrics used to evaluate the effectiveness of k-means

clustering for each K. It then outputs a table with Silhouette Score, Calculated Calinski-Harabasz Index, Davis Bouldin Index and Inertia for each number of K.

This allows the user together with the results of previous step to make an informed choice in the number of clusters for the modelling to be performed.

4. **Apply K-Means**

In this step, k-means clustering is performed. The route takes in an input from the user of the number of clusters, as well as the number of model iterations. Due to the nature of k-means clustering behaviour with random seeding, each model results in slightly different outcomes. Therefore, this step allows for multiple iterations of the model for the user to choose the best performing one. Each model is saved in a pickle file to be accessed in the next step. The output of this step are inter-cluster distance graphs, that show the clusters and their sizes in a two-dimensional space.

5. **Predict**

This is the final step, that takes on a request of the iteration index of the chosen model from the previous step to perform final modelling. It opens the saved model from the pickle file, predicts the cluster labels and outputs the results to a csv file or sql table.

5.1.3.2. Predictive Modelling Solution

The beginning of the Predictive Modelling process is dependent upon the completion of the Clustering Modelling phase. The objective of the Predictive Modelling Solution is to provide a model that accurately predicts the cluster labels of individual customers on a monthly basis, including those who are new to the company. The output of the clustering solution is sent into the predictive modelling solution, where the model is trained using the clustering variables. Subsequently, the model predicts the cluster label for each individual consumer. In order to accommodate the inclusion of new customers and account for potential shifts in customer behaviour that may lead to migration between segments, the model is trained first and subsequently stored for future predictions. These predictions are recommended to be conducted monthly.

Application Development

The Predictive Modelling Application is designed for almost automatic use of predictive classification modelling. The technologies and python packages used include Flask for python web framework, SQLAlchemy to communicate with databases, Scikit-learn for predictive modelling, Matplotlib,

Seaborn and Dataframe -image for visualization, Pandas for dataframe operations and Pickle to save models.

The project space consists of main file for running the application, configurations file written in yaml, and modelling module that contains the functions and objects and their methods used to perform predictive modelling.

Within the configuration file, the user can specify the connections to the database or choose to connect to a csv file. The application has three modes: train, train and predict and predict, which also can be chosen, allowing for the initial training use and continuing predictions. The user can choose which columns need to be dropped, which one is the ID column, target column and the target names, where to store visuals and models, as well as where to store the output of the predictions.

Except for pre-processing and analysing of the data, the predictive modelling application project space follows Object-Oriented Programming approach for predictive classification modelling. A Parent class object has been created with its own methods, as well as children objects for different classification models: Decision Tree, Random Forest, K-Nearest Neighbours and Support-Vector Machine. Additionally, data preparation module contains pre-processing, oversampling, and splitting functions. Train_test module contains functions that uses the Classify object defined in the modelling module.

The application reads the dataset, then drops the columns specified in the configuration file. It consists of four routes:

1. **Train and test**

In this route, the application trains and tests a single model, specified by the user in the configuration file. It then utilizes the train_test function, which drops the ID column, splits the dataset into train and test sets, normalizes the data, performs oversampling, tunes the hyperparameters of the chosen model. Then it fits, predicts, evaluates, and saves the model. The user receives the best parameters of the model and confusion matrix and classification report.

2. **Best model**

The best model route requests the user to specify the evaluation metric to choose the best model. It uses choose best model function, in which all the specified models are trained and tested and undergo the same preparation and train and test process as specified in the step 1. (Train and test), however, all the models are saved and the best one is chosen, accordingly to the best result from the evaluation metric chosen by the user. The best model is then saved to a pickle file for future use. If the chosen

mode is to predict, it then predicts the labels which are either saved to sql table or csv file. The route gives the user the output of the best model and its parameters.

3. **Predict**

The predict route is intended for continuous use once the model is trained and saved. It takes the dataset, prepares it by dropping the ID column and normalizing the dataset. It then opens the saved model (which path is specified in the configuration file) and predicts the labels, which are then saved to sql table or csv file.

4. **Predict form**

Predict form route is designed for quick, one-off predictions given the values. It requests the values for each variable as a user input in a form request within the flask application and it then normalizes the data, opens the saved model and predicts the label. It then outputs the predicted class and its probability to the user.

5.1.3.3. Results Layer

Following the use of clustering techniques, the Dashboards and Reports layer is a final but crucial part of the customer segmentation process. For stakeholders, analysts, and decision-makers, this layer acts as a comprehensive visualization and reporting tool that enables them to learn substantial information from the segmented customer data. The Dashboards and Reports layer facilitates efficient decision-making and the development of targeted strategies by presenting the findings in an intuitive and user-friendly manner.

The Results layer's main goal is to give users a clear understanding of the customer segmentation findings from clustering analysis. In order to enable users to learn more about the traits, tendencies, and preferences of each customer segment, it aims to transform complex data into information that can be used to make decisions. The dashboards and reports act as an outlet for the findings, ensuring that stakeholders are able to quickly understand the lessons learned from the clustering process.

Additionally, the results of the customer segmentation can be easily implemented into Dynamics 365 Marketing Platform, creating marketing lists for the segments created through clustering. Dynamics 365 allows marketers to personalise their campaigns by tailoring them to the specific characteristics and behaviours of certain segments. Through the utilisation of marketing lists, businesses have the ability to effectively customise their messages, offers, and content in order to cater to the specific

needs and preferences of each group. The automation functionalities offered by Dynamics 365 allow marketers to optimise and automate their marketing processes. The integration of segmentation data into marketing lists enables marketers to automate the distribution of personalised messages, monitor customer interactions, and initiate automatic replies in accordance with consumer behaviour. Through the use of built-in reporting, the value of the ML segmentation process can be re-evaluated. The comprehensive analytics and reporting functionalities offered by Dynamics 365 offer useful insights into the effectiveness of targeted marketing efforts. Marketers have the ability to evaluate the effectiveness of their efforts, track key performance indicators (KPIs), and acquire valuable insights into client responses. This data can also allow for fine-tuning of the Machine Learning processes.

5.2. RECOMMENDATIONS SOLUTION

The recommendations solution follows the clustering and predictive modelling parts described in the Segmentation Solution, with an addition of calculating recommendation scores. Additionally, it contains a web-based application that can take on values of variables of a client provided by a user and predict the customer segment and output the recommended products with their recommendation scores. The architecture of the recommendation's solution can be shown in the presented graph.

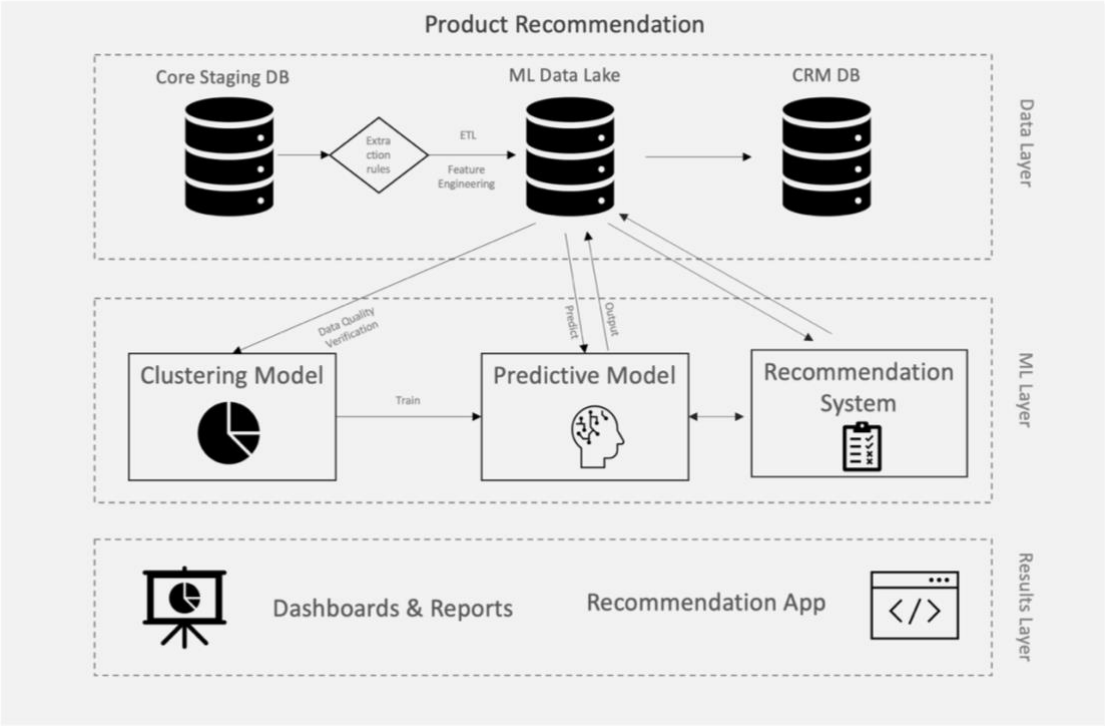

Figure 14 Recommendation Solution Architecture

All the processes, described in the 5.1. Segmentation Solution section, are used for the recommendation solution, with a difference in the addition of a simple script that calculates the scores

accordingly to the formula presented in chapter 3.5.3.1 Clustering-based Recommendation System. The automation of that part needs further development in order to automate the recommendation score threshold input.

In addition to the visual representation of the clustering and recommendation score results, a flask web application was developed to quickly predict the segment membership and provide personalized product recommendations based on the client characteristics given as an input.

Additionally, recommended client lists for each product were uploaded to the CRM Dynamics365 system for easy access when designing marketing campaigns. To achieve this, a recommendation score threshold of 0.2 was applied.

5.3. Deployment Conclusion

The deployment architecture and development of clustering and predictive modelling applications mentioned above were utilised as a proof-of-concept (POC) to showcase the initial prototype of the project. The study effectively presented empirical data supporting the applicability of the concept in practical situations. However, it is worth noting that the creation of the applications was characterised by a rudimentary level of complexity. The apps can be utilised in diverse contexts, extending beyond a singular use case, by leveraging the configuration file. However, additional progress is required to conduct further experimentation on the concept, including the refinement of the front-end user interface and the evaluation of various database connections. In addition, it is necessary to conduct further testing of the suggested solution on other projects and datasets in order to assess its scalability and potential for recurrence.

6. CONCLUSIONS AND FUTURE WORKS

This study introduced two potential machine learning approaches that can be deployed as products that complement the existing customer relationship management (CRM) and database management system provided by Asseco PST. Both solutions have successfully achieved the initial project objectives, which were to provide machine learning solutions that are user-friendly and adaptable. These solutions aim to optimise the marketing campaigns of Asseco's bank clients. One of the project's requirements was to provide a machine learning framework that may seamlessly interact with Asseco PST's existing CRM infrastructure, as provided to its clients. The project successfully achieved its goal of automating and enhancing the marketing endeavours of bank customers by implementing consumer segmentation through machine learning and a product recommendation system.

The proof-of-concept study employed customer relationship management (CRM) data from a bank client of Asseco to investigate the practicality of the suggested solution. The study's findings show that K-means clustering can be used to create meaningful client segmentation. The clusters that have been formed have distinct characteristics and have the potential to offer novel perspectives on customer attributes and behaviours. The effectiveness of segmentation was further validated using the evaluation metrics of K-means clustering and prediction models. The recommendation solution presented in this study offers an alternative approach to address the limitations of accessible data in recommendation systems. The study demonstrated the effectiveness of employing k-means clustering for computing product recommendation scores for customer segments, as opposed to individual consumers. Moreover, it highlighted the potential of this approach in developing personalised marketing strategies aimed at promoting the sale of additional products and services offered by the bank. Upon the presentation of the findings to the Sales team of Asseco PST, they expressed agreement with the results, confirming the effectiveness and practical significance of the machine learning solution that was built. The similarity between the identified consumer categories and what they know about bank’s customers from their own experience, the intuitive cluster names, and the proposed marketing techniques was well received by the Sales team, affirming the strategic importance of implementing AI-driven methods for future offer development. The validation received from the Sales team serves to confirm the rigorous academic methodology of the research and highlights the possibility for smooth integration of machine learning technologies within a dynamic commercial environment.

The next natural step is presenting the successful proof-of-concept to clients, signifying its transformation from an internal innovation to a solution that is accessible to clients. The incorporation of this solution into the organization's range of offerings demonstrates its dedication to remaining at

the forefront of technical progress and providing innovative solutions as a component of its service portfolio.

In conclusion, this study successfully achieves its intended goals and also establishes the groundwork for a transformative journey for Asseco PST's Data & Analytics team. The utilisation of segmentation and recommendation models, coupled with their effective implementation, enables the organisation to offer its bank clients to adopt a customer-centric approach by implementing personalised marketing campaigns and making data-driven decisions. Considering the ongoing evolution of the financial landscape and the rise of Artificial Intelligence applications, the outcomes of this research provide not only valuable insights but also a practical guide for effectively navigating the intricate dynamics of contemporary banking through strategic thinking and inventive approaches.

6.1. Limitations & Future Work

Despite the encouraging results obtained from the research, it is imperative to acknowledge and address the several limitations inherent in this study. Firstly, as indicated in the Results section, the project solely relied on the data provided by one of Asseco's clients, specifically an Angolan bank. Consequently, it is not appropriate to draw conclusions about the project's performance for other clients. Furthermore, it should be noted that the analysis was limited by the quality of the data, resulting in the exclusion of certain variables. The exclusion of specific variables, such as income or postal code, may limit the comprehensiveness of the investigation and its resulting consequences, thus affecting the effectiveness of machine learning models. Additionally, it is important to acknowledge that the dataset utilised in this study only encompasses a specific subset of bank clients, namely those considered active, which represents approximately 15% of the total customer base. Consequently, caution should be exercised when generalising the findings to the wider banking sector.

The project only focused on one customer, thus depending on the client base—especially if it's a foreign client—there may be differences in requirements and behaviour. As a result, the method and results may not be as effective when applied to different clients. According to the insights provided by the Asseco PST team, it is observed that a significant proportion of transactions in African countries are conducted using cash. Consequently, this poses a challenge in evaluating customer behaviour, as individuals infrequently utilise their bank accounts. This limited usage of bank accounts may have implications for understanding customer preferences in a comprehensive manner.

An additional concern associated with the work presented relates to the ethical considerations surrounding the selection of variables for analysis. The utilisation of age, gender, education, marital

status, and bank balance as factors in the research may introduce bias and yield discriminatory effects, hence raising concerns about the results obtained from both proposed solutions and the implementation of personalised marketing tactics for the identified segments.

With additional specific details about the product attributes or the kind of clients they are intended for, the recommendation system's effectiveness could be further increased. Moreover, possessing knowledge about the customer journey and the specific point at which the products were obtained may potentially influence the outcomes of the recommendations. The recommendations' precision and level of information could be affected by the lack of these specifics. Moreover, the current evaluation of the recommendation system is constrained as it solely depends on clustering and predictive model evaluation methodologies. However, it has not undergone real-life testing, particularly in relation to customer engagement and satisfaction. The success of the offered solution necessitates validation upon implementation in the production phase.

The study lacks a clear comparison with alternative approaches or solutions. While the applicability of the proposed machine learning solution is evident, conducting a comparative analysis with conventional methods or alternative AI-based approaches can offer a more comprehensive understanding of its advantages.

The transition from the proof-of-concept phase to the full implementation of a project can encounter operational challenges. The practical viability of the proposed solution may be impacted by several aspects, such as the interaction with existing systems, scalability, and user acceptance. Furthermore, the suggested solution architecture entails the implementation of a data pipeline methodology, which involves the establishment of a data lake specifically designed for the machine learning solution. Additionally, an Extract, Transform, Load (ETL) process is incorporated into the pipeline, necessitating the use of SQL and Integration Services. The exploration of other techniques has been lacking, and it is possible that the proposed architecture may not offer a cost-effective solution for the clients. Furthermore, the software lacks the capability to provide a comparative analysis with a cloud-based alternative or establish connectivity with the data source via REST APIs.

Finally, the banking sector is characterised by its dynamic nature, as it continuously adapts to shifting customer behaviours, regulatory modifications, and technological improvements. The conclusions of the study, although valid throughout the period of research, may require ongoing adjustments to ensure their effectiveness in light of evolving industry environment.

As future work, the project would benefit from utilization of more rich and higher quality data, to do so more emphasis needs to be put on the data quality significance within the Asseco PST clients, which already has been partially addressed through development of products focusing on the assessment of data quality in the database. Additionally, it would be interesting to see how the findings could change when incorporating additional variables or through incorporation of data from different sources, which could potentially contribute to a more comprehensive understanding of machine learning solutions' applicability across various banking contexts.

The customization and adaptation of the developed solution for different clients and their unique business models should be a focus of future research. Investigating the transferability and effectiveness of the solution across diverse banking scenarios would contribute to its broader applicability. Moreover, the proposed deployment and applications serve only as a demonstration and require further programming development and the building of the user-friendly front-end interface. Furthermore, greater automation in the data extraction and transformation rules and processes needs to be further explored. The next step should consist of testing the solution on a dataset from a different client and implementing the project into a production process to assess its real-life value. Further processes and analyses that measure the effectiveness of the proposed solutions need to be developed, working together with the business analysts and marketing teams of the Asseco's PST clients. At that stage, personalized marketing campaigns need to be developed to address the characteristics and needs of each segment and specific metrics need to be established to measure the success of the solutions. Additionally, there needs to be ongoing monitoring of the performance of the models and the solutions.

This project signifies the starting point of a more extensive endeavour, acting as a crucial starting point that lays the groundwork for continuous machine learning advancement within the scope of Asseco PST's Data & Analytics services. The current study effectively fulfils certain objectives related to client segmentation, CRM integration, and the deployment of a recommendation system. However, it is crucial to emphasise that its value surpasses these immediate outcomes. This project establishes a foundation for ongoing innovation and improvement in the field of machine learning applications within the banking sector of Asseco PST clients based mainly in the African region. The symbolic cornerstone serves as a fundamental basis upon which subsequent developments, adaptations, and expansions of the technology will be constructed. This perspective emphasises the project's future-oriented approach, highlighting its role not only as a standalone initiative but as the starting point for an ongoing and developing exploration of the revolutionary capabilities of machine learning in the constantly evolving field of financial services.

BIBLIOGRAPHICAL REFERENCES

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of Recommender Systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. https://doi.org/10.1109/tkde.2005.99

Alexandra, J., & Sinaga, K. P. (2021). Machine learning approaches for marketing campaign in portuguese banks. *2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS)*. https://doi.org/10.1109/icoris52787.2021.9649623

Amnur, H. (2017). Customer relationship management and machine learning technology for identifying the customer. *JOIV : International Journal on Informatics Visualization*, *1*(4), 12. https://doi.org/10.30630/joiv.1.1.10

Antal-Vaida, C. (2022). A review of Artificial Intelligence and machine learning adoption in banks, during the covid-19 Outbreak. *Proceedings of the International Conference on Business Excellence*, *16*(1), 1316–1328. https://doi.org/10.2478/picbe-2022-0120

Aryuni, M., Didik Madyatmadja, E., & Miranda, E. (2018). Customer segmentation in XYZ Bank using K-means and K- medoids clustering.*2018 International Conference on Information Management and Technology (ICIMTech)*. https://doi.org/10.1109/icimtech.2018.8528086

Beheshti, A., Yakhchi, S., Mousaeirad, S., Ghafari, S. M., Goluguri, S. R., & Edrisi, M. A. (2020). Towards Cognitive Recommender Systems. *Algorithms*, *13*(8), 176. https://doi.org/10.3390/a13080176

Bogaert, M., Lootens, J., Van den Poel, D., & Ballings, M. (2019). Evaluating multi-label classifiers and recommender systems in the Financial Service Sector. *European Journal of Operational Research*, *279*(2), 620–634. https://doi.org/10.1016/j.ejor.2019.05.037

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27.

Campbell, C., Sands, S., Ferraro, C., Tsao, H.-Y. (J., & Mavrommatis, A. (2020). From data to action: How marketers can leverage AI. *Business Horizons*, *63*(2), 227–243. https://doi.org/10.1016/j.bushor.2019.12.002

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

Chapman, P. C., Clinton, J., Ncr, R. K., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0.

Chen, T.-H. (2020). Do you know your customer? bank risk assessment based on machine learning. *Applied Soft Computing*, *86*, 105779. https://doi.org/10.1016/j.asoc.2019.105779

Crammer, K., & Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. Journal of machine learning research, 2(Dec), 265-292.

D'Arco, M., Lo Presti, L., Marino, V., & Resciniti, R. (2019). Embracing AI and Big Data in customer journey mapping: From literature review to a theoretical framework. *Innovative Marketing*, *15*(4), 102–115. https://doi.org/10.21511/im.15(4).2019.09

David, A. (2007). Vassilvitskii s.: K-means++: The advantages of careful seeding. In 18th annual ACM-SIAM symposium on Discrete algorithms (SODA), New Orleans, Louisiana (pp. 1027-1035).

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

Djurisic, V., Kascelan, L., Rogic, S., & Melovic, B. (2020). Bank CRM optimization using predictive classification based on the support vector machine method. *Applied Artificial Intelligence*, *34*(12), 941–955. https://doi.org/10.1080/08839514.2020.1790248

Djurisic, V., Kascelan, L., Rogic, S., & Melovic, B. (2020). Bank CRM optimization using predictive classification based on the support vector machine method. *Applied Artificial Intelligence*, *34*(12), 941–955. https://doi.org/10.1080/08839514.2020.1790248

Djurisic, V., Kascelan, L., Rogic, S., & Melovic, B. (2020). Bank CRM optimization using predictive classification based on the support vector machine method. *Applied Artificial Intelligence*, *34*(12), 941–955. https://doi.org/10.1080/08839514.2020.1790248

Edwards, A. W., & Cavalli-Sforza, L. L. (1965). A method for cluster analysis. Biometrics, 362-375.

Fares, O. H., Butt, I., & Lee, S. H. (2022). Utilization of artificial intelligence in the banking sector: A Systematic Literature Review. *Journal of Financial Services Marketing*. https://doi.org/10.1057/s41264-022-00176-7

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Sciences*, *10*(21), 7748. https://doi.org/10.3390/app10217748

Fix, E., & Hodges Jr, J. L. (1952). Discriminatory analysis-nonparametric discrimination: Small sample performance. California Univ Berkeley.

Gallego-Gomez, C., & De-Pablos-Heredero, C. (2020). Artificial Intelligence as an enabling tool for the development of dynamic capabilities in the banking industry. *International Journal of Enterprise Information Systems*, *16*(3), 20–33. https://doi.org/10.4018/ijeis.2020070102

Gallego, D., & Huecas, G. (2012). An empirical case of a context-aware mobile recommender system in a banking environment. *2012 Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing*. https://doi.org/10.1109/music.2012.11

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61–70. https://doi.org/10.1145/138859.138867

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of intelligent information systems, 17, 107-145.

Hentzen, J. K., Hoffmann, A., Dolan, R., & Pala, E. (2021). Artificial Intelligence in customer-facing financial services: A systematic literature review and agenda for future research. *International Journal of Bank Marketing*, *40*(6), 1299–1336. https://doi.org/10.1108/ijbm-09-2021-0417

Hernández-Nieves, E., Hernández, G., Gil-González, A.-B., Rodríguez-González, S., & Corchado, J. M. (2020). Fog computing architecture for personalized recommendation of banking products. *Expert Systems with Applications*, *140*, 112900. https://doi.org/10.1016/j.eswa.2019.112900

Hlavac, J., & Stefanovic, J. (2020). Machine learning and business intelligence or from descriptive analytics to predictive analytics. *2020 Cybernetics & Informatics (K&I)*. https://doi.org/10.1109/ki48306.2020.9039874

KAUR, N. A. V. L. E. E. N., SAHDEV, S. U. P. R. I. Y. A. L. A. M. B. A., SHARMA, M. O. N. I. K. A., & SIDDIQUI, L. A. R. A. I. B. E. (2020). Banking 4.0: "The influence of artificial intelligence on the Banking Industry & How Ai is changing the face of modern day banks." *INTERNATIONAL JOURNAL OF MANAGEMENT*, *11*(6). https://doi.org/10.34218/ijm.11.6.2020.049

Königstorfer, F., & Thalmann, S. (2020). Applications of artificial intelligence in commercial banks – a research agenda for Behavioral Finance. *Journal of Behavioral and Experimental Finance*, *27*, 100352. https://doi.org/10.1016/j.jbef.2020.100352

Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, *63*(2), 157–170. https://doi.org/10.1016/j.bushor.2019.10.005

Li, X. (2021). Application and influence of Artificial Intelligence Technology in commercial banks. *2021 2nd International Conference on Computer Science and Management Technology (ICCSMT)*. https://doi.org/10.1109/iccsmt54525.2021.00089

Lin, R.-H., Chuang, W.-W., Chuang, C.-L., & Chang, W.-S. (2021). Applied Big Data Analysis to build customer product recommendation model. *Sustainability*, *13*(9), 4985. https://doi.org/10.3390/su13094985

Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

Machauer, A., & Morgner, S. (2001). Segmentation of bank customers by expected benefits and attitudes. *International Journal of Bank Marketing*, *19*(1), 6–18. https://doi.org/10.1108/02652320110366472

McKinsey & Company. (2021). AI-bank of the future: Can banks meet the AI challenge. Retrieved from https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/building%20the%20ai%20bank%20of%20the%20future/building-the-ai-bank-of-the-future.pdf

Mihova, V., & Pavlov, V. (2018). A customer segmentation approach in commercial banks. *AIP Conference Proceedings*. https://doi.org/10.1063/1.5064881

Namvar, M., Gholamian, M. R., & KhakAbi, S. (2010). A two phase clustering method for intelligent customer segmentation. 2010 International Conference on Intelligent Systems, Modelling and Simulation. https://doi.org/10.1109/isms.2010.48

Omoge, A. P., Gala, P., & Horky, A. (2022). Disruptive technology and AI in the banking industry of an emerging market. *International Journal of Bank Marketing*, *40*(6), 1217–1247. https://doi.org/10.1108/ijbm-09-2021-0403

Oyebode, O., & Orji, R. (2020). A hybrid recommender system for product sales in a banking environment. Journal of Banking and Financial Technology, 4(1), 15–25. https://doi.org/10.1007/s42786-019-00014-w

Özdemir, S., Sonmez Cakir, F., & Adiguzel, Z. (2022). Examination of customer relations management in banks in terms of strategic, technological and Innovation Capability. *Journal of Contemporary Marketing Science*, *5*(2), 176–195. https://doi.org/10.1108/jcmars-12-2021-0044

Payne, A. (2005). HANDBOOK OF CRM: Achieving Excellence in Customer Management. Elsevier.

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in Recommender Systems: A systematic review. *Expert Systems with Applications*, *97*, 205–227. https://doi.org/10.1016/j.eswa.2017.12.020

Raiter, O. (2021). Segmentation of Bank Consumers for Artificial Intelligence Marketing . *International Journal of Contemporary Financial Issues*, *1*(1), 39–54.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. Recommender systems handbook, 1-34.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Sharifihosseini, A. (2019). A case study for presenting Bank Recommender Systems based on Bon Card Transaction Data. *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*. https://doi.org/10.1109/iccke48569.2019.8964698

Sheth, J. N., Jain, V., Roy, G., & Chakraborty, A. (2022). AI-driven banking services: The Next Frontier for a personalised experience in the emerging market. *International Journal of Bank Marketing*, *40*(6), 1248–1271. https://doi.org/10.1108/ijbm-09-2021-0449

Thisarani, M., & Fernando, S. (2021). Artificial Intelligence for Futuristic Banking. *2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. https://doi.org/10.1109/ice/itmc52061.2021.9570253

Ubiparipovi, B., & Durkovic, E. (2011). Application of Business Intelligence in the Banking Industry. *Management Information Systems*, *6*(4), 023–030.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.

Wang, L., Liu, Y., & Wu, J. (2018). Research on financial advertisement personalised recommendation method based on customer segmentation. *International Journal of Wireless and Mobile Computing*, *14*(1), 97. https://doi.org/10.1504/ijwmc.2018.090005

Yu, Q., Jiang, H., & Ma, X. (2018). The Application of Data Mining Technology in Customer Relationship Management of Commercial Banks. *14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 1368–1373.

Zakrzewska, D., & Murlewski, J. (2005). Clustering algorithms for bank customer segmentation. *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*. https://doi.org/10.1109/isda.2005.33

Zibriczky, D. (2016). Recommender systems meet finance: a literature review. *FINREC*.

APPENDIX A


CORE_ENTITY
NUMBER PK, VARCHAR
GENDERCODE VARCHAR
BIRTHDATE VARCHAR
PLACEOFBIRTH VARCHAR
ANNUALINCOME VARCHAR
JOBTITLE VARCHAR
EMPLOYER VARCHAR
ENTITYTYPECODE VARCHAR
NATURETYPE VARCHAR
POSTALCODE VARCHAR
MARITALSTATUS FK, VARCHAR
NATIONALITYCODE VARCHAR
QUALIFICATIONCODE FK, VARCHAR
PROFESSIONALACTIVITY FK, VARCHAR
SECTORIALCODE VARCHAR
ISRESIDENT VARCHAR
COUNTRYCODE VARCHAR
CLIENTRELATIONSHIP FK, VARCHAR
TS DATETIME
RISKLEVEL VARCHAR
CORE_MARITALSTATUS
CODE PK, VARCHAR
NAME VARCHAR
STATE VARCHAR
TS DATETIME
CORE_QUALIFICATION
CODE PK, VARCHAR
NAME VARCHAR
STATE VARCHAR
TS DATETIME
CORE_RELATIONSHIP
ACCOUNT_CODE PK, VARCHAR
ENTITY_CODE FK, VARCHAR
STATE VARCHAR
NAME INT
TS DATETIME
CORE_ACTIVE_ENTITIES_CUSTOMERS
ACCOUNT_CODE PK, VARCHAR
ENTITY_CODE FK, VARCHAR
STATE VARCHAR
TS DATETIME
CORE_MOVEMENT
MOVEMENY_ID PK, INT
CLIENT_NUMBER FK, VARCHAR
CONTRACT_NUMBER FK, VARCHAR
MOVEMENT_VALLUE NUMERIC
CURRENCY_CODE VARCHAR
MOVEMENT_BALANCE_AFTER NUMERIC
MOVEMENT_RATE NUMERIC
MOVEMENT_DATE VARCHAR
MOVEMENT_DESCRIPTION VARCHAR
OPERATION_CODE VARCHAR
OPERATION_DESCRIPTION VARCHAR
TS DATETIME
MOVEMENT_USER VARCHAR
MOVEMENT_CHANNEL VARCHAR
MOVEMENT_TYPE VARCHAR
MOVEMENT_EXECUTOR VARCHAR
CORE_CLIENT
CLIENT_NUMBER PK, VARCHAR
BRANCH_CODE VARCHAR
COUNTRY_CODE VARCHAR
SEGMENT_CODE VARCHAR
USER_CODE VARCHAR
CLIENT_CREATIONDATE VARCHAR
CLIENT_TYPE VARCHAR
CLIENT_SECTORIAL VARCHAR
CLIENT_ISRESIDENT VARCHAR
CLIENT_RISKCLASS VARCHAR
TS VARCHAR
CLIENT_RISKAML VARCHAR
CORE_CONTRACT
CONTRACT_NUMBER PK, VARCHAR
PRODUCT_CLASS FK, VARCHAR
CONTRACT_OPENINGDATE VARCHAR
CONTRACTSTATUS VARCHAR
CLIENT_NUMBER FK, VARCHAR
PRODUCT_CODE FK, VARCHAR
CURRENCY_CODE VARCHAR
CONTRACT_DUEDATE VARCHAR
ISBLOCK VARCHAR
CONTRACT_BALANCE NUMERIC
CONTRACT_IRREGULARBALANCE NUMERIC
CONTRACT_RATE VARCHAR
OVERDRAFT NUMERIC
OVERDRAFT_UNAUTHORIZED NUMERIC
DAYS_OVERDRAFT VARCHAR
TS DATETIME
STATE VARCHAR
MEDIUMBALANCEAVAILABLE NUMERIC
AVGMEDIUMBALANCEBYDAYS NUMERIC
CORE_PRODUCT
CODE PK, VARCHAR
NAME VARCHAR
STATE VARCHAR
CLASSCODE FK, VARCHAR
SUBSCRIPTION INT
CURRENCYCODE VARCHAR
TS DATETIME
SEGMENTS VARCHAR
CREDITTYPE VARCHAR
CORE_PRODUCTCLASS
CODE PK, VARCHAR
NAME VARCHAR
TS DATETIME
CORE_CREDIT_BALANCE
CREDIT_CLASS VARCHAR
CREDIT_TYPE VARCHAR
CREDIT_SITU VARCHAR
CREDIT_BALANCE VARCHAR
CREDIT_CONTRACT VARCHAR
CREDIT_DATECREATION DATETIME
CREDIT_CLIENTNUMBER FK, VARCHAR
CREDIT_QT INT
CREDIT_CURRENCY VARCHAR
TS DATETIME

APPENDIX B

Table of centroids for k = 16

cluster	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
CLIENT_SINCE	89.7504	32.8689	47.9348	67.9250	122.8647	41.6345	39.8854	120.5487	61.9339	91.5133	41.9714	66.3600	57.3998	84.5527	112.8180	87.6852
AGE	49.7865	33.5383	33.0407	40.5974	44.3526	25.6675	25.6496	41.3726	35.9011	48.8521	25.9514	38.6962	35.4121	44.0711	42.7422	44.5010
AVERAGE_AVAILABLE_BALANCE	30392698.2505	3048311.8641	412581.0206	12456841.6986	-4905535.6112	1552963.1869	63774.7900	-1907330.3754	4555335.9460	20734692.2152	1414536.7357	7750353.1382	523869.4390	16554346.6455	6335341.1140	15312509.2462
IS_SINGLE	0.0000	1.0000	1.0000	1.0000	0.8692	0.9942	0.9944	0.8522	1.0000	0.0000	0.9944	1.0000	1.0000	0.0000	1.0000	0.0000
HIGHER_EDUCATION	1.0000	0.0000	0.0000	1.0000	0.1378	0.0956	0.0779	0.2123	0.0000	0.0000	0.0583	1.0000	0.0000	1.0000	0.0000	0.0000
IS_STUDENT	0.0113	0.0000	0.0000	0.0049	0.0240	1.0000	1.0000	0.0535	0.0000	0.0118	1.0000	0.0000	0.0000	0.0315	0.0000	0.0486
GENDER	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000	0.0000	1.0000	1.0000	0.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
RECENCY	30.3546	18.4014	116.7263	33.0369	6.5608	41.9184	124.7397	6.1336	16.7001	34.1167	20.2972	30.8462	124.4321	29.4774	16.1490	31.4676
MONETARY_VALUE	68908.9969	31378.3002	-11299.6502	129950.6793	172401.3680	2090.5637	-8435.0609	-183537.5734	125032.3302	72544.0197	-24616.9646	-14312.3726	-32363.2426	-125945.9468	-3155.1134	-13418.2160
FREQUENCY	85.2050	56.2495	8.4203	83.6796	57.4267	39.1942	9.4894	84.0912	55.6721	76.6659	49.9951	71.1592	10.4746	79.8057	84.4209	68.9393
IS_CREDIT	0.1475	0.0000	0.0000	0.0053	1.0000	0.0063	0.0011	1.0000	0.0000	0.0000	0.0091	0.0008	0.0000	0.1040	0.0000	0.0000