University of Manouba
National School of Computer Sciences

Summer immersion internship report

Customer segmentation in banking

Realized by :
Olfa MESSAOUD

Organism : ProxymIT
Supervised by : Tahar Jarbaoui
Address : Technopole of Sousse BP 184 Khezama TN 4051, Avenue Khezama, Sousse
TEL : 73 821 222
eMail : tahar.jarboui@proxym-group.com

University year :
2019-2020

Contents

List of Figures

List of Tables

General introduction

Over these few years artificial intelligence and machine learning have become in almost every field which leads us to cope with these new technologies and tools in order to use them to perform a desired task to test their validity and to decide about their performances. Talking about machine learning that was the result of a really hard work to integrate artificial intelligence into computer science to give the possibility of making smart systems and helping people to predict and decide about a needed task.

Researches have shown that a lot of algorithms and methods were adopted by machine learning to fulfill decision making tasks offering quasi exact results what makes us excited to use these predefined algorithms ans methods to make predictions, proceed clustering, and having an idea toward the desired results. Specifically in the case of customer segmentation which is really delicate because the society need to understand their clients and know them for more in their habits and behaviours.

Dive deeper in the topic of clustering customers of a given enterprise will help the marketing manager to embrace his consumers and know them for more to enhance the customer relationship management and to be aware and take all the convenient precautions by making offers and promotions to retain his loyal customers and attract new ones. More over than that, we aim to reveal how customers could be partitioned into clusters of segments by the specified characteristics.

Hence, We will help the marketing manager to carry out a segmentation in a retail banking sector. Having the possibility of clustering consumers by their geography, demography, psychography, and also behaviour so we will give the possibility to visualize every segment of the customer in the given kind of clustering besides we could even reveal and figure out the customer lifetime value in the company and his loyalty grade using recency, frequency, and monetary scores ending by a set of plots and figures.

Chapter 1

State of art

Introduction

In these days, the main necessity of the industry in any field consists on classifying their customers and segmenting them in order to more embrace their profiles and improve the targeted customer relationship management.

In this first chapter, we will introduce and define the most used keywords in our project to make it clearer then we will talk about the existing solutions and their limitations ending by an introduction to our proposed solution with the to do tasks.

1.1 Machine learning and customer segmentation bases

1.1.1 Machine learning

Machine Learning [**?**] (ML) is a set of algorithms and statistical models implemented in order to emulate and compute the human intelligence opting for fast learning and offering quasi exact outputs such as human results. ML has got supervised learning algorithms that include the presence of a data scientist, unsupervised learning algorithms in which data don't need to be trained, and reinforcement learning algorithms that does not need inputs and outputs.

1.1.2 Customer Relationship Management

Customer Relationship Management[**?**] (CRM) is a set of strategies, practices and technologies that the companies adopt to analyze and manage customer interactions, transactions and data throughout its life cycle. CRM aims to catch customers and enhance customer retention for driving sales growth. CRM systems provide either a detailed descriptions about customers personal information, purchases, buying preferences, and concerns.

1.1.3 Customer segmentation

Customer segmentation[**?**] or clustering is the process of dividing the customers of a given company into groups of individuals based on their similarities such as age, gender, location, interests, spending habits, etc. Clustering customers is a major need of enterprises to differentiate between their customers and have a deeper understanding of them and even their needs and their intentions with the want of discovering the characteristics of each segment.

1.1.4 Customer lifetime value

Customer Lifetime Value [**?**] (CLV) is a measurement value that represents the total amount the customer is expected to spend on the society during his lifetime or presence and it is a

really important figure to decide the money to invest in acquiring new customers or retaining old ones. Besides, CLV is used to calculate churn probability for customers which is directly relied to CRM to improve customer profitability and loyalty of course.

1.1.5 Segmentation types

Segmentation types [**?**] are mainly geographic classification by country, region, city, or other geographic basis. Then, demographic classification by population characteristics such as age, gender, marital status, occupation, ans so on. Also, psychographic segmentation relied directly to customer lifetime value and churn probability. And, behaviour segmentation calculated through the recency, frequency, and monetary (RFM) scores.

1.1.6 Feature engineering process

Feature engineering [**?**] is a process of converting the given data into a form that is easier to interpret and work with. The new data might be calculated from already existing data or predicted after performing of a machine learning algorithm. So, feature engineering is fundamental for applying machine learning algorithms because it provides better results than using the principle data because the programmer could definitely know the desired kind, shape, or type of data that will enhance results.

1.1.7 RFM scores

RFM [**?**] (Recency, Frequency, Monetary) attributes are calculated by the existing purchase and invoice data of the customers to precise more the loyalty of the customer toward the company. SO, recency represents how recently the customer has purchased in days, frequency means how often customer has purchased, and monetary determine how much the customer spends on purchases. Then, RFM values are converted into RFM scores where every score is between 1 and 5 where 1 is the lowest score.

1.2 General presentation of the project

1.2.1 Problem settings

Based on the issue of classifying customers and labelling them to help companies know the behaviour of their clients and be able to understand them. And because of the potential use of machine learning algorithms to make patterns recognition with getting interesting results which let us demand about the power of machine learning clustering algorithms and how accurate are they ?

1.2.2 Required work

Therefore, we opt to help the companies find a solution while treating their customers and trying to know them more and more by clustering them into groups and put a description to each one of them. Otherwise, we are going to make a segmentation for the customers based on their psychographic or loyalty, geographic, demographic, and behaviour or churn. By making a comparison about the belonging the a specific customer to the cluster number in each side.

1.3 Overview of existing solutions

1.3.1 Existing solutions

- Clustering customers by their loyalty by calculating the Recency, Frequency, and Monetary (RFM) derived values for every customer to describe his or her loyalty toward the company without the use of any machine learning algorithm what has not been needed recently.
- Use of k-means, fuzzy c-means, genetic algorithms, Gaussian mixture algorithms, expectation maximization, mean shift algorithm, etc... With RFM attributes for segmentation which give a great view about customer loyalty ignoring the other kinds of segmentation.
- Segmentation methods were often used for performing only one clustering part on customer loyalty, customer geography, customer demography, or customer behaviour when making the four segmentation parts will be more and more interesting to discover all the details about the customer.

1.3.2 Balance sheet

Existing cases	Limits
Clustering customers by their loyalty after calculating the Recency, Frequency, and Monetary.	Recency, Frequency, and Monetary attributes could not give good results for segmentation without the use of a machine learning algorithm for clustering.
Applying machine learning clustering algorithm with RFM attributes for segmentation give a great results about customer loyalty.	Making segmentation with RFM scores and a clustering algorithm will figure out only the loyalty side of the customer which could not be enough to embrace the customers while segmenting.
Segmentation is performed to classify clients only by customer loyalty or customer geography or customer demography or customer behaviour when making the four segmentation parts.	Proceeding many segmentation types will be more and more interesting and exact while proceeding clustering to make the client discover all the details about his customers.

Table 1.1 – Existing solutions and their limitations

1.3.3 Proposed solution

Therefore, we proposed to make a segmentation model for customers using a retail banking dataset that highlights the results of clustering models used. The models are loyalty, demography, geography, and behaviour of the customer that's why we are going to classify the clients of the bank following these four kinds of segmentation and then we will make a find out the best number of clusters thus the belonging of every customer to each segment of cluster.

Conclusion

In this chapter we emphasized the key words of our project defining them, we presented our work, and we put on an overview about the existing solutions for segmentation with their limits and we finished by presenting our proposed solution.

Chapter 2

Analysis and design

Introduction

As we presented customer segmentation basis and process reminding the existing solutions and their limitations and ending by proposing our proper method. In this chapter we will give a detailed analysis and modelling for the project.

2.1 Requirements analysis

2.1.1 Actors

In our project, we have only a main actor who is the marketing manager that will handle all the segmentation sides and will find the plots of clusters and the classification of customers by every category what makes him able to estimate the partition of his customers in every type of segmentation.

2.1.2 Functional requirements

- ☐ The marketing manager has the capability of making a segmentation based on loyalty of customer in terms of invoice, demographic, geographic, and behavioural or his lifetime value side to cluster consumers who are a bank customers in our case and the segmentation results will be showed in figures.
- ☐ Our main actor will figure out the dashboards of the clustering algorithms results after applying them to be ready to classify customers and reveal the cluster of every one of the client. So, he will be able to know the repartition of the clusters of every segmentation base and the densities of every cluster.
- ☐ After segmenting and classifying consumers, the marketing manager could probably make a comparison between the used clustering algorithms to find out the best one for our dataset and case of study then he would also compare the belonging of every client to each segmentation side .

2.1.3 Non functional requirements

2.1.3.1 Performance

The response time should be very fast in terms of few seconds at maximum in every running because we are using a machine learning algorithm for a normal amount of data.

2.1.3.2 Extensibility

The platform should be open and ready for future growth when the client needs to extend or to add a new functionality or even to modify an existing functionality to enhance the work.

2.1.3.3 Maintainability

The realized work should be well commented and documented to make it clear to read and easier to understand which make its maintenance easy and convenient.

2.1.3.4 Ergonomics

The plots must be attractive and well structured with comfortable colors and beautiful dashboards to ensure that the actor will be pleasant each time he uses the platform.

2.2 Modelling

2.2.1 Use case diagram

The use case diagram describes the behaviour of the system and the designed actions of each actor to show how an actor interact with the system to achieve a task and it contains the most important functionalities that actors are going to realize.

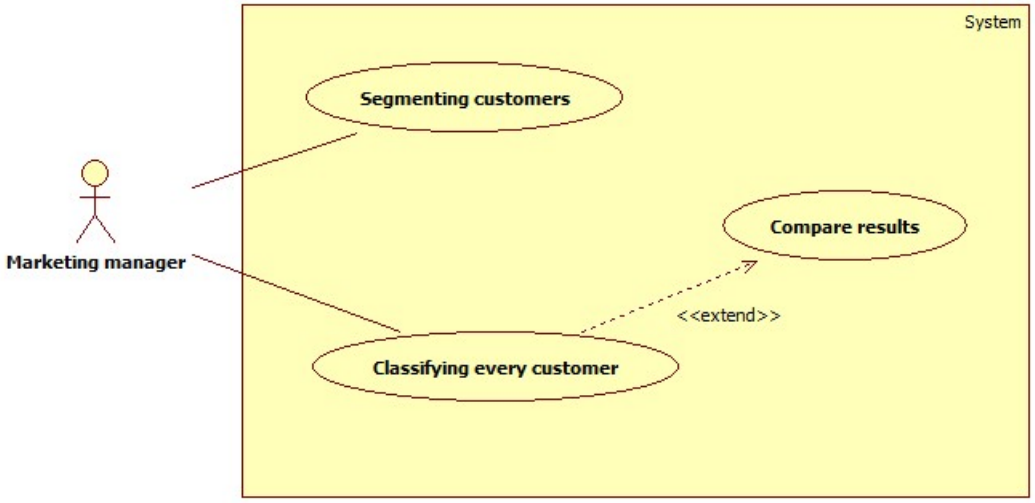

Figure 2.1 – Use case diagram

The figure 2.1 represents the use case diagram for our project in where we find our marketing manager actor that will interact with the platform to make segmentation then to classify every client in the corresponding cluster. Also, the actor could eventually compare clustering results and even clustering algorithms used.

2.2.2 Sequence diagram

The general sequence diagram shows the interaction between every actor and the system which is considered as a black bottle. Because we have an only actor in our case we will make a sequence diagram for him to describe his tasks through the time.

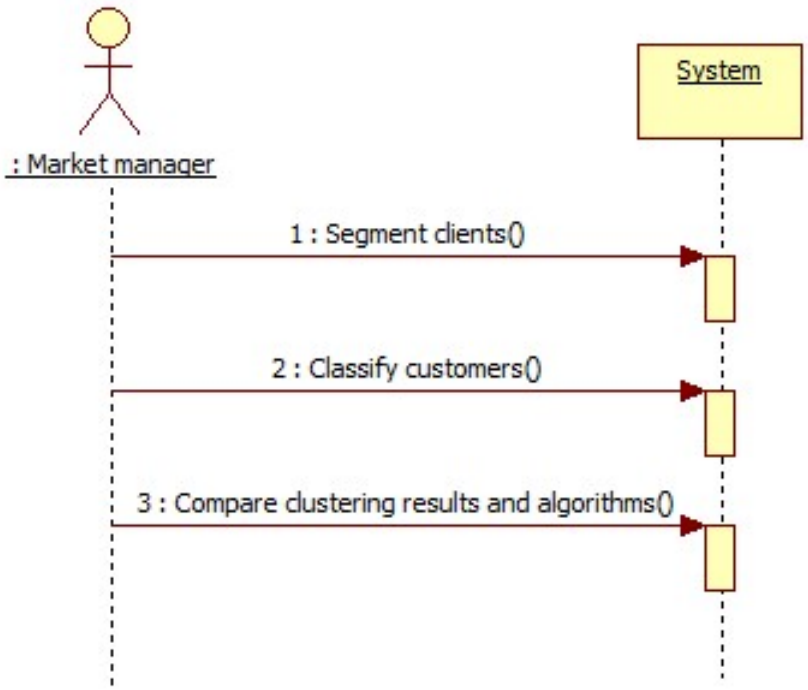

Figure 2.2 – Sequence diagram

So, the sequence diagram corresponds to the marketing manager and how he communicates with the system to realize different tasks. The figure 2.2 shows this interaction and the corresponding scenario.

2.3 Design

The design phase consists on justifying the models used for both design pattern and architect pattern as well as presenting diagrams of the detailed design for front end and for back end to give a global view of the main components, packages, and classes employed in the desktop application from coding until deployment.

2.3.1 Design architecture

Software architecture or design architecture defines the structure of hardware and software in terms of components and interactions between them. We note that architecture design has two types, to know, logical architecture directly relied to the software and physical architecture relied to the technical components used.

2.3.2 Physical architecture

The physical architecture fits more with our project is the 2-tiers architecture precisely client-server architecture where the server tier contains both database and treatments while the client who is a thin client corresponding to a fat server having the hole application and the client navigates to the server to make a request and get the results.

Else, we find a thick client or also called a fat client who gets the graphical interface of the application in his local machine, executes some treatments, and even manages some data from

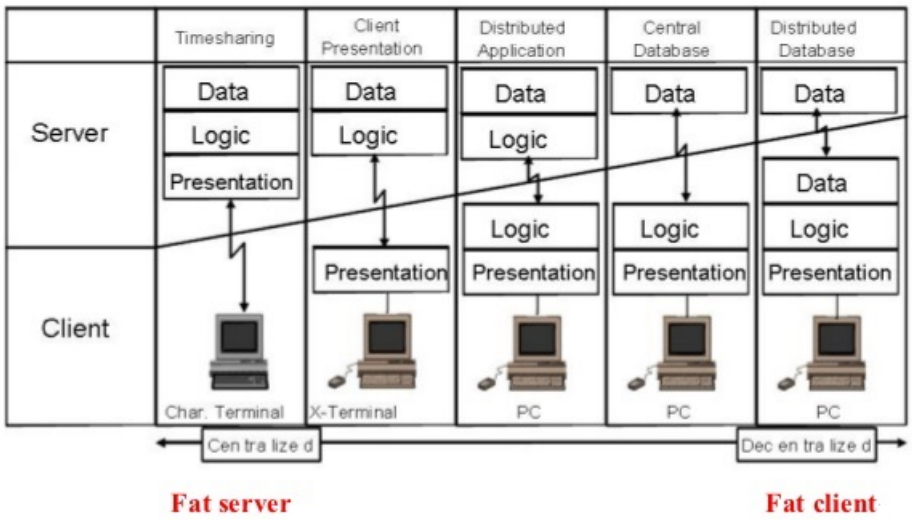

Figure 2.3 – Fat client vs thin client architecture

his local host. The figure 2.3 explains for more the difference between a fat client and a fat server.

2.3.3 Logical architecture

Logical architecture is a structured design detailing the application without constraining the architecture to a particular technology or environment. In our case, we could use any of the logical multi layer architectures such as graphical layer and business or logical layer.

The graphical layer have the interface of the application that contain all the needed functionalities with their results and plots the actor want to perform. While the business layer have the hole code processing the required tasks and the given clustering results.

2.4 Detailed design

As we presented the architectural design of the application, we are now going to detail our design by detailed design diagrams to show how we did divide classes and interfaces in different packages to improve almost modularity, coupling, and cohesion.

2.4.1 Class diagram

The class diagram contains the objects of the application that are related to each other so in our class we have the marketing manager, the segmentation, the classification, and the interface of the visualization which is shown in the figure 2.4.

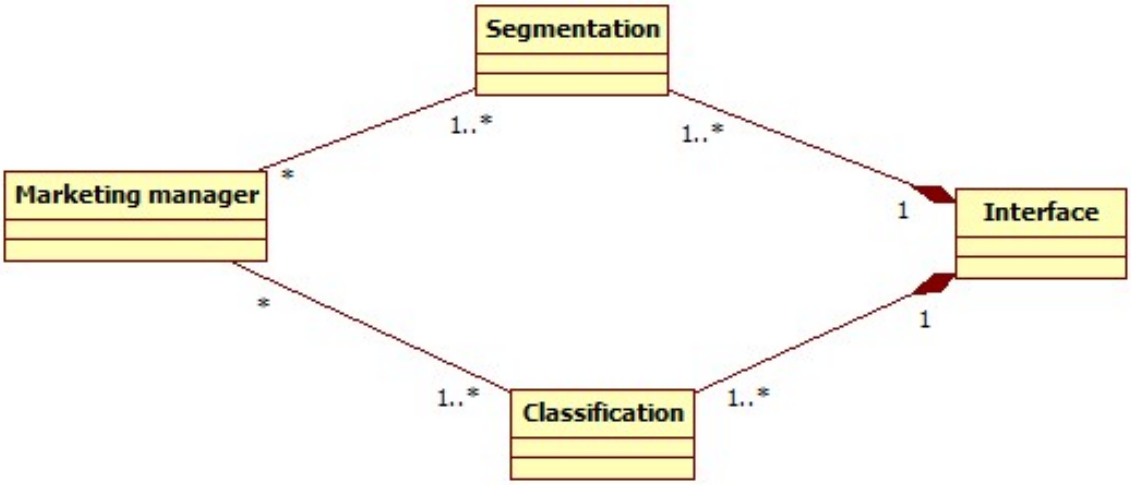

Figure 2.4 – Class diagram

2.4.2 Activity diagram

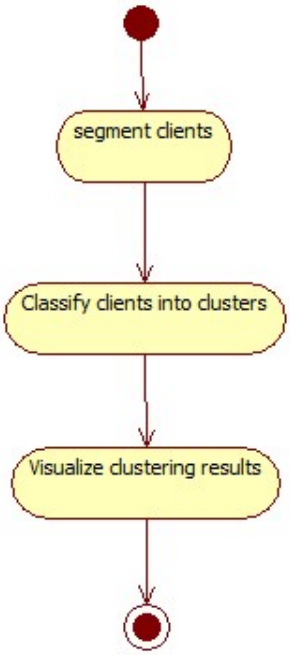

Figure 2.5 – Activity diagram

Activity diagram describes dynamic aspects or behaviours of the system. The figure 2.5 shows that the marketing manager will perform the segmentation, then he will make customers classification ending by a simple comparison with the accuracy.

2.5 Data science workflow

This project belongs to the field of data science because we are clearly going to work at the back to carry out the desired tasks that the marketing manager has posed using all of data mining, feature engineering, data analytics, and machine learning algorithms.

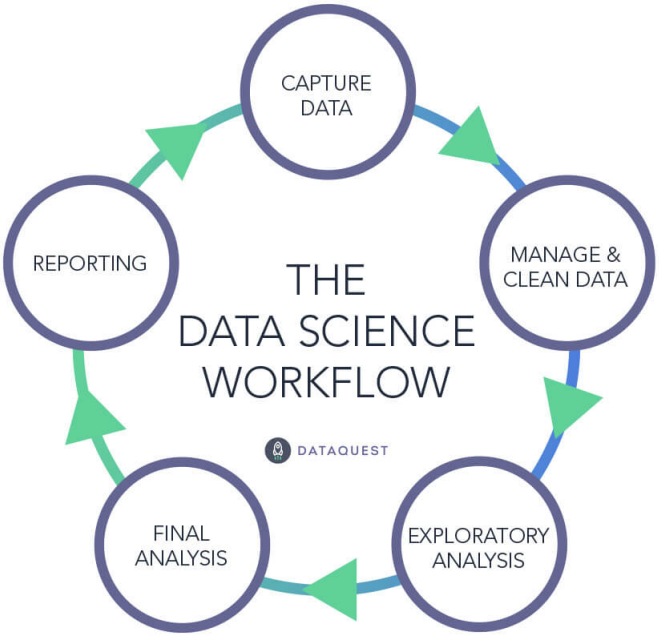

Figure 2.6 – Data science workflow

The process of a data scientist consists on capturing data from a company or scraping them from a website. Then, managing data by storing them and cleaning them, following by an exploratory analysis to study the data and a final analysis to explore them completely. And, a reporting phase to represent results be accessed easily by the user as shown in the figure 2.6.

Conclusion

In this chapter, we focused on the analysis and specification requirements followed by the modelling part and we turned to the design part to produce class diagram and activity diagram. Finally, we introduced the lifetime cycle of a data science project with the diagram of machine learning.

In the next chapter, we will introduce all environments used with the technologies, tools, and libraries used in order to achieve this project. Also, a complete representation of the released work and the graphical interfaces for each task with the desired results accomplished by the application.

Chapter 3

Achievement

Introduction

In the first chapter, we presented the most important terms of our project following by studying existing applications and their lacks, and proposing our proper solution to achieve customers segmentation. While the second chapter contained the overall requirements, modelling, and design diagrams.

In this last chapter we will define all the environments and components we have used in order to help the marketing manager fulfill his customers clustering. Then, we will put on the dashboards and the figures of the customers clusters and their distribution by every segmentation kind.

3.1 Work environment

The work environment is composed of two sides, the hardware environment which presents the machines or components we used while running our code and the software environment which represents the advanced programming interfaces, integrated development environments, editors, technologies and tools we used.

3.1.1 Hardware environment

The table 3.1 describes the used computer used for the related work noting that was no configuration or specification were needed.

PC	Brand	Processor	RAM	HD	OS	Graphics Card
1	ASUS	2.2 GHz	8 GB	1 TB	Windows 10	NVIDIA GeForce

Table 3.1 – Characteristics of used computer

3.1.2 Software environment

In this section, we will give an overview about all the tools and products we used in our application and we will mention the main role of each one of them with justifying why we used them. To clarify some points, we note that there are a lot software products we could have used them instead of those ones.

3.1.2.1 Anaconda

Anaconda is a free and open source distribution of python and R programming languages for scientific computing such as data science, machine learning applications, large scale data processing, predictive analytics, etc. It integrates the desktop graphical user interface anaconda navigator that allows launching applications.

3.1.2.2 Jupyter notebook

Jupyter notebook is a client server application that permits editing and running documents written in python via a web browser on a local desktop without internet access or on a remote server accessed through the internet. It gives the possibility to write the code with some headings and downmarks to make the code clear and easy.

3.1.2.3 Elasticsearch

Elasticsearch is a search engine developed in java and capable of addressing a growing number of use cases. It is an oriented document database for storing, retrieving, and managing data. It is a distributed engine used commonly for log analytics, full text search, security intelligence, business analytics, and operational intelligence use cases.

3.1.2.4 Kibana

Kibana is an open source analytics and visualization platform designed to work with Elasticsearch to search, view, and interact with data stored in Elasticsearch and it allows perform advanced data analysis and visualize your data in a variety of charts, tables, and maps easily and it enables create and share dynamic plots causing changes to Elasticsearch in real time.

3.1.2.5 Star UML

Star UML is a UML sophisticated software modeller for concise and agile modelling diagrams. It is a a cross platform written in java that supports most of the diagram types for presenting the design of an application. Other than that, it gives the possibility of reversing engineer which means translating generated drawn diagrams into a test script.

3.1.2.6 Overleaf

Overleaf is a free, modern, and online editor to develop documents in LaTex. It includes unicode support, spell checking, auto completion, code folding and built in the pdf. Also, it is easy to use and configure just like Microsoft word office. And, it locates errors and warnings by logging them.

3.2 Technologies

3.2.1 Python programming languages

Python is a managed or interpreted programming language released in 1991 to emphasize the clarity and the readability of the code. Python is dynamically typed and garbage collected supporting procedural, functional, and even object oriented programming and provides different constructs.

Python interpreters are available for many operating systems and it is also an open source software. The main force of python is using a very simple syntax close to natural thoughts of

a developer which make resolving problems much easier and faster than other programming languages.

As a result, python is the most used programming language in iRobot machines, You Tube, Bit Torrent, video games, Google Search, and machine learning. In addition, python uses many libraries in its back end while accomplishing a task and abstracts details.

3.2.2 Libraries and predefined modules

3.2.2.1 PySpark

PySpark is the collaboration of Apache Spark and python when Apache Spark is an open source cluster computing framework that guarantee speed, ease of use, and streaming analytics and python. It is a python advanced programming language for Spark that offers the simplicity of working with data and the fastness while taming a big amount of data.

3.2.2.2 Sckilit learn

Sckilit learn or also sklearn simply is a free machine learning library for python with a lot of features for classification, regression, or clustering including support vector machine, random forests, gradient boosting, and k neighbours supporting other libraries.

3.2.2.3 Pandas

Pandas is an open source library providing high performance and simple way to use and analyze data by python. It is the most used library for analyzing data offering a set of methods for the dataframe type to read even csv files.

3.2.2.4 Matplotlib

Matplotlib is a library dedicated to plotting figures with python providing an object oriented API and may make plots in many other toolkits. It gives a lot of methods to customize the plots and change them from a figure to another.

3.3 Achieved work and results

In this stadium after detailing all the ambiguous points of the project and showing the overview of it, we are now supposed to introduce the segmentation results and to try to deduce the profile of every customer in the bank.

3.3.1 Customer segmentation process

For the total process of segmentation, we will start by downloading the Berka dataset which is free on the internet and have a set of 7 tables related through primary keys and foreign keys though it is a bit old since 1998.

Then, we managed to join all the tables to have the complete information about the customer and his transactions. Hence, we cleaned the data and save the necessary to proceed with. And, we started by calculating RFM values and the RFM scores to produce the loyalty of every customer.

The recency is the number of days since the last purchase, the frequency is the number of transactions, and the monetary is the total amount of money spent. And, the scores between 1 and 5 with 1 is the lowest score.

Next, by applying k-means and calculating the RFM scores for every cluster found we could have segmented our customers through their loyalty. And of course we tested many algorithms to figure out the best one for our case.

And by calculating accuracy score for tested algorithm and highlighting the best one, we used it for the rest of segmentation criteria whatever it was demographic segmentation, geographic segmentation, and customer lifetime or behaviour segmentation.

3.3.2 tested algorithms for clustering

To cluster consumers we plotted the elbow graph and the curve of the best k was 3 so we are going to choose 3 as the number of clusters in the tested algorithms.

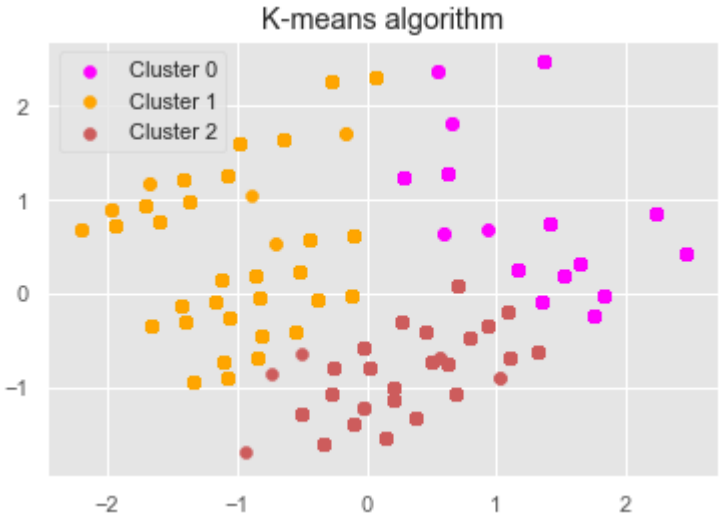

Figure 3.1 – K-means algorithm

The figure 3.1 showed the partition of the 3 clusters, the clusters are clearly separated and the execution time for k-means is very fast.

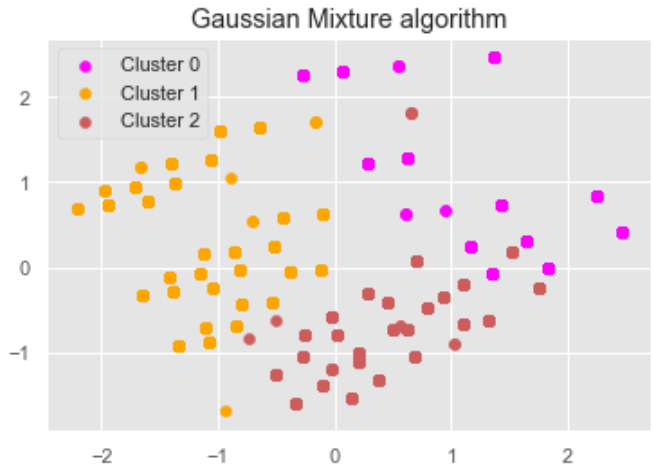

Figure 3.2 – Gaussian Mixture algorithm

The figure 3.2 of the Gaussian Mixture Model (GMM) showed a beautiful partition of the 3 clusters such as k-means and we note that GMM calculate centroids using euclidean distance and probabilities.

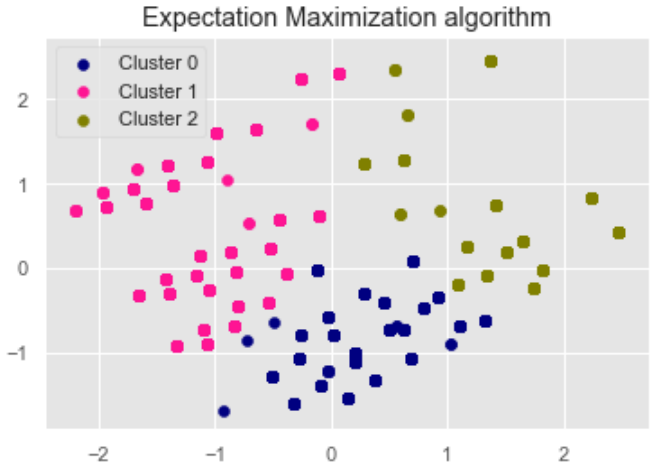

Figure 3.3 – Expectation Maximization algorithm

The figure 3.3 shows the expectation maximization implemented algorithm which is similar to GMM in terms of computations and gives interesting results results but it works only with a converted data of ours.

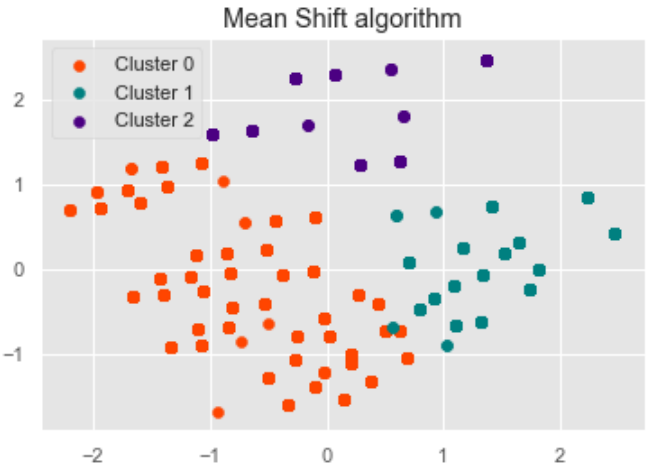

Figure 3.4 – Mean Shift algorithm

The figure 3.4 of the tested Mean Shift clustering algorithm shows that clusters are partitioned well though some points of the cluster 0 and 1 are very close which makes clustering not very precise.

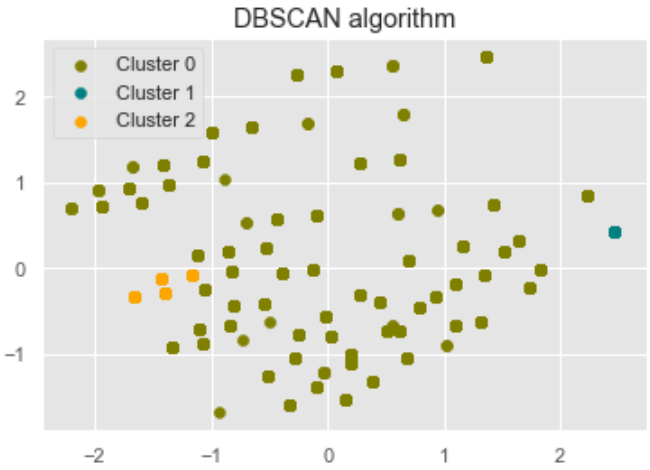

Figure 3.5 – DBSCAN algorithm

The figure 3.5 of the last tested algorithm DBSCAN which is implemented through a neural network inside to tackle density based data and auto estimate number of clusters though its epsilon parameter is very critic.

3.3.3 Accuracy results and most convenient algorithm

In terms of evaluating accuracy for clustering algorithms such as k-means, Gaussian mixture, and mean shift which are not real classification tools and they do not use labels to train. Whereas, we could deal with their results to find out the most precise and accurate clustering algorithm to use.

Expectation Maximization algorithm is not predefined so it has not a predict method and hence we could not apply the accuracy_score method to calculate its accuracy but we could confirm its rapidity and the interesting portioning it gave us.

DBSCAN also has not a predict method that's why we can not calculate its accuracy nevertheless it is important while dealing with real attributes and considering densities variation of labels though its latency and we can not determine the optimal epsilon value for clustering.

Clustering algorithm	Accuracy score
K-means	0.39
Gaussian mixture	0.32
Mean Shift	0.006

Table 3.2 – Accuracy results for clustering algorithms

And finally the table 3.2 showed us the accuracy scores for the algorithms we used. As a result we will keep on using K-means clustering algorithm because it gave us the best performance results in terms of execution time and accuracy score.

3.3.4 Customer loyalty

The average RFM score for every cluster makes us determine the loyalty level of every customer. Noting that we have 3 clusters named Best, Risky, and About to sleep where Best

customer is the one having the highest frequency and monetary with the lowest recency.

The Risky customer has a high frequency and monetary with a high recency and the about to sleep is the cluster having the lowest frequency and monetary with high recency.

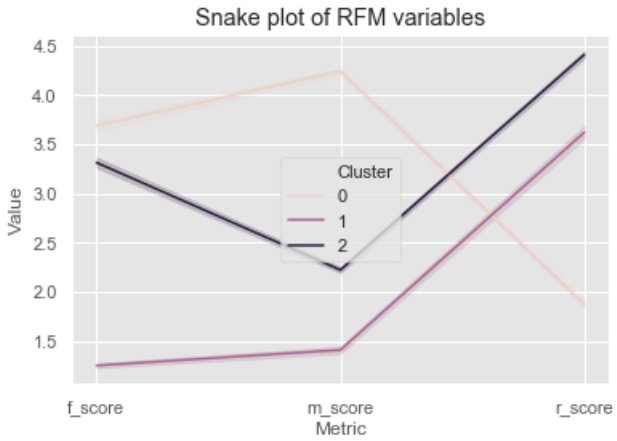

Figure 3.6 – Average RFM score per cluster

The figure 3.6 shows that the cluster number 0 is the Best, the cluster number 2 is the risky, and the cluster number 1 is the about to sleep one as we deduced from the RFM average score for every cluster.

3.3.5 Customer behaviour

Customer behaviour could be issued from customer lifetime value (CLV) in the bank which is also issued from RFM values to emphasize the churn decisions of every customer and so to make the bank think about their customer retention.

The clustering was based on customer lifetime value by applying k-means clustering algorithm with 3 as number of clusters and by describing every cluster we could classify them into short living, well living, or long living customer. The figure 3.7 shows how the clusters are partitioned.

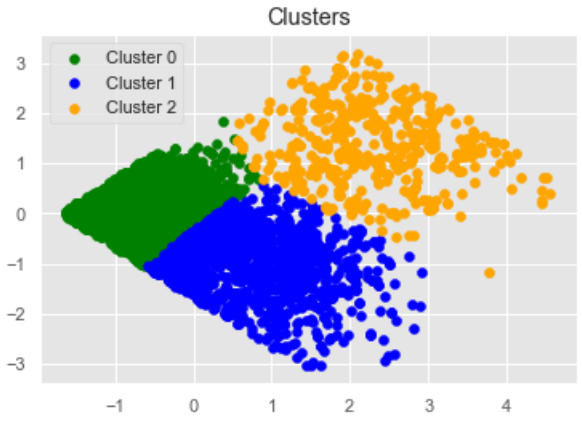

Figure 3.7 – Customer behavioral segmentation

Cluster 0 have an average CLV of 672786 with values range between 184560 and 1507055 so this is the cluster having the minimum CLV that could be named short living cluster.

Cluster 1 have an average CLV of 2251916 with values range between 1414514 and 4608918 so this is the most long living cluster that will stay and keep being loyal to the bank.

Cluster 2 have an average CLV of 1800602 with values range between 340377 and 4326604 so this is the well living cluster that has an important CLV.

3.3.6 Customer geographic

To cluster customers of Berka bank by geographics, we noticed first the presence of 8 regions for all the customers which are the Bohemian region in north, south, east, west, and center with the Prague which is its capital. And the north and south Moravia.

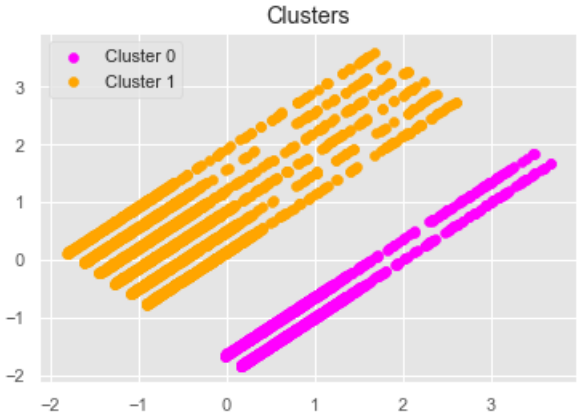

Figure 3.8 – Customer geographic segmentation

The figure 3.8 makes it clear that we are optimally 2 clusters which are Bohemian customer and Moravian customer and after describing the data of every cluster we knew that the cluster 0 is the Bohemian and the cluster 1 is the Moravian.

3.3.7 Customer demographic

In this kind of segmentation, we used the the age and the gender of the customer to cluster customers by their demographics and because that's what we have about the customers only as informations. Then, we plotted the description of every cluster to label them.

The figure 3.9 shows the results of clustering k-mans algorithm with k = 3 and how clusters are partitioned.

Figure 3.9 – Customer demographic segmentation

For the cluster 0, we found an average age up to 73 from 50 to 107 years-old and all the customers of the cluster 0 are females. Therefore, we could name it ancient females.

For the cluster 1, we found an average age up to 60 from 33 to 93 years-old with a female majority customers. So, we could name it mixed juniors.

For the cluster 2, we found an average age up to 43 from 31 to 50 years-old with a fully male cluster customers that's why we could name it senior males.

Now by installing kibana and uploading the csv files that we need to perform visualizations of customers clusters per category and whatever we want, we could get displays of all the customers and for everyone the cluster name by every kind of segmentation or we could even make a filter for a needed customer or a set of customers.

Conclusion

In this chapter, we presented first the necessary software, tools, technologies, programming languages, and libraries we used to develop our project. Then, we detailed all the work precising the algorithms used and the best algorithm for our case. We displayed our results by plots or kibana visualizations.

The project emphasizes the use of machine learning algorithms to determine the profile of a bank customer throughout his loyalty, behaviour, geographic, and even demographic information to help the marketing manager ensure best understanding his consumers and aim to retain them and enhance customer relationship management.

Bibliography