NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Implementing a Bank Sales Analytics Solution and a Predictive model for the Next Best Offer

Internship Report

Ziad El Abbass

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

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

2018

Title: **Implementing a Bank Sales Analytics Solution and a Predictive Model for the Next Best Offer**

Subtitle: Internship Report

Student
full name Ziad El Abbass

MAA

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

IMPLEMENTING A BANK SALES ANALYTICS SOLUTION AND A PREDICTIVE MODEL FOR THE NEXT BEST OFFER

by

Ziad El Abbass

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor: Mauro Castelli

November 2018

DEDICATION

“This document is dedicated to my expired dear Father Abdelhai El Abbass who, despite his illness, always craved to support me endlessly so he can achieve his dream of seeing me shine of success.”

ABSTRACT

In the banking industry, the quantity of information that is processed is huge. Knowing also that clients are doted with changing needs every time, companies must adapt their approaches to attract clients with the best offers. That can be done by various machine learning and data mining techniques that enable them to understand better the clients. Also, internally, banks should be equipped with fast and efficient processes that enable them to take quickly the best decision. That is why real-time reporting tools should be implemented as an upper layer of the data sources. In this optic, this internship report is presenting 2 ambitious projects that aim to leverage Millennium BCP bank to a greater level in Analytics and Data Science. The first one is about building a Sales Analytics Solution to track weekly sales of retail products in the bank. The second one is about building a mechanism that will help reach to each client's best adequate product to recommend.

KEYWORDS

Sales; Marketing Campaign; Recommendation System; Reporting; Next Best Offer; Recommendation System

INDEX

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

NBO: Next Best Offer

KDD: Knowledge Discovery in Databases

ABT: Analytical Base Table

RS: Recommendation System

MBCP: Millennium BCP

1. INTRODUCTION

Nowadays in modern companies, it is widely recognized that data has taken an important twist in the recent decade across all economical fields, making it mandatory for the technological infrastructure inside companies to be strongly adapted to support its non-stopping growth along business timeline. However most of the modern companies, specifically the small medium-sized companies were only interested in having powerful storage capacities for statistical purposes. In the other hand, few companies not only they have reached that level, but they are also exploiting this historical accumulation of data using prescription methods to discover underlaying patterns hidden from the actual and previous checkpoints of their business. Based on that, companies constitute a clear roadmap on how to orientate their business to their most profitable clients, thus accomplishing a higher turnover than doing it without data exploitation. In this optic, Millennium BCP which is the largest private bank in Portugal and where the internship has been carried out, is one of the few banks in Portugal, that has succeeded in the field of Big Data and Machine Learning and has been an example of a leader company that exploits Big Data and does prescription studies aiming for a better business. This know-how on both theoretical and practical levels, has been applied in various facets of banking such as retail, funds and investments, insurances, loans, real-estate and private-banking. Millennium BCP, given its important base of clients, a rigorous attention is permanently given to the client's interactions with his patrimony inside the bank. Each client has a determined profile regarding his purchases which, on a large scale, would help Customer Relationship Management teams inside the bank communicate more adequately with the client about the products and offers he is more likely to have. Overall, this internship which was mainly conducted inside the CRM team of the retail marketing department, was mostly based in data collection, cleaning, analysis, predicting and communicating. There will be launched various trade-offs between a panoply of data mining projects methods and processes based on the specificities of each conducted project. On the other hand, Data visualization also will take a very essential role in understanding the evolution of some indicators quickly avoiding time consuming querying, to get the comparison baselines and thus making the data much more understandable through graphical figures.

2. LITERATURE REVIEW

In a world in which internet and technology does not stop from moving forward, Companies must adopt their ways of dealing with massive data in terms of storage and structure. But there is a big part of the companies who behave in a report centric fashion (Kirpes) which leads them take very poor decisions because they haven't drawn attention to the other facet of the medal. Most of the companies are adopting very powerful containers of data which are called Data Warehouses that contain all the summarized information that users need across various axes of analyses called dimensions. Still, there are some data warehouses who have been built very successfully but still lack in terms of efficiency in providing the information. The reason for this issue is that most Data warehouses are built without giving an importance to the conceptual or reference models that they must follow (Matthias Goeken, 2007). The starting point for building a data warehouse is to put into a list all the eventual questions to which this solution is going to come up with answers. The most widely used reference model that Data warehouses rely on is the Entity Relationship Model proposed by (Chen, 1976). This model represents the different stakeholders in a company's daily activity as well as the actions involved between each of these participants. Later, these models can easily translate to Relational Models that is a model widely followed by most of the Online Transactions Systems (OLTP) (K.J Raiha, 1992). Few companies have a long-term vision and opted instead for implementing a Lake. Contrarily to a Data warehouse, a Data Lake contains all types of data. Structured, Semi-Structured, and unstructured. It is a repository that has basically no respect for the shape of the data flow. According to (Llave, 2018) There are 3 purposes behind implementing a data lake which are: Multiple Data sources, an experimentation platform for data scientists and data analysts, and the application of self-service Business Intelligence. The main process behind building a Data Warehouse is the ETL (Extract, Transform, Load) process which is according to (Davenport, 2008), is considered the traditional and widely accepted approach. The ELT (Extract, Load, Transform) Processes has also emerged recently which loads the data from the data sources before cleansing and transforming the data. The ELT has been proven to be less risky than the traditional ETL Process in a way that the transformation is isolated which enables easy maintenance and management. However, it lacks a lot of technological tools in comparison with the ETL that has many in the market.

Business Intelligence is known to be a strong technique that helps companies through their business decisions. However, it is now considered very traditional with the presence of Data Science and Machine learning techniques that help not only in taking decisions but also to anticipate them. Companies nowadays have a great opportunity to extend their traditional Business Intelligence platform to a more powerful one since they already have massive data in their servers. It is in this optic that (Fayyad, 1996) are inciting on extracting knowledge from the cumulated data using processes of Knowledge Discovery Databases. In brief, the KDD Process aims first to create a targeted data from which one wants to extract useful information. Afterwards, a series of data pre-processing and transformation methods are applied, followed by data mining techniques to extract value and make interpretations. Data mining also has been a field to which research has been giving a lot. (Mark Brown) from SAS suggested the SEMMA Method that is widely used in SAS technological tools. (Shearer, 2000) as well in the Journal of Data Warehousing described a very simple and concise data mining method that starts particularly with business understanding followed by data understanding. These data mining techniques help build very powerful predictive models that can be used to forecast some rare events and consequently act efficiently based on these predictions. In the

Marketing and Retail banking, for example, some Machine Learning techniques have proved to have given very good results in terms of making offers to clients for products that might interest them. Recommendation Systems are powerful models that are built based on the behavior of a lot of contributors. The best example would be (FERNÁNDEZ, 2018) with her Movie Recommendation System built for Netflix which was based on over 17000 movies and 500000 Netflix users. (Robin Burke, 2008) distinguishes between 5 different types of Recommendation Systems. collaborative, content-based, utility-based, demographic, and knowledge-based. But the most used ones are collaborative-based which relies on the ratings of each user towards the item, and content-based which considers the features presented by each item. Few companies have used so many data mining techniques and reached a high level of understanding each client's needs. This made it easy for them to suggest the best deal that they can come up with for each client. This concept is called the Next Best Offer (or Action). (Andrea Fabrizi, 2014) have built a very robust solution for predicting the Next Best Offer stressing out also other companies in the commercial field to transform their old marketing strategy of being product-centric into a new one of being client-centric. Their Next Best Offer model relies on getting all the products' profiles and the clients' profile in an information module. This module pushes the information to the Recommendation Engine that comes up with the best offer. The recommendation engine is also trained with the feedback of the clients following this offer. This makes (Andrea Fabrizi, 2014) solution acts always in an efficient manner.

It is with the light of all these research contributions stated above that the following work will be presented.

3. THE CRM SALES ENGINE IN MILLENNIUM BCP

3.1. OVERVIEW

Millennium BCP retail marketing department is composed of several teams that try to find approaches of conveying the sales of specific products. This department has a very compound sales infrastructure that contains various processes, from marketing campaigns to purchased products by the client. Given that the marketing department is composed by teams that are responsible of managing the business of each product category (which mainly are Investment and saving funds team, Solutions team, Bank accounts team, Credit and Debit Cards team, Insurances team, Personal loans and other loans team, and real-estate loans team), a permanent cooperation is held between each of these teams and the CRM team. Core entity that launches the starting point of a marketing campaign chain through which a specific product, service or promotion is being communicated to the client. Clients subject to campaigns are usually chosen by either applying determined criteria that are dealt with between the CRM team and the team concerned by branding its product(s), or by the outcome of a prediction model built also by other team members of the CRM Team.

Contacting clients for branding and selling Millennium BCP products is made through various means of communication called channels. Here below there are stated the most used ones.

- MBCP bank branch (Sucursal): Which are simply small bank branches located all over the country that offer banking face-to-face services to the clients. Moreover, they receive a list of clients by product from the CRM team who should be contacted for eventual offers and services or promotions according to the marketing campaign related to the purpose of the contact.

- Email: Considered as one an important canal due to its free cost. It is usually used to send temporary promotions of discounts to clients.

- Outbound Channels: Contacting approaches launched from the initiative of the bank to the client.

- Inbound Channels: Cases in which the client, by his initiative, reaches the bank directly or indirectly seeking for information or advice about a product.

The following figure illustrates the main channels through which communications between the bank and the client occurs

		OUTBOUND Proactive Approach	**INBOUND** Proactive Approach during a Customer Contact	**INBOUND CAPTURE** Commercial Opportunity
NOT HUMAN	**AUTOMATIC** "One Shot"	MAILING EXTERNAL E-MAIL SMS SECURE E-MAIL (BANCOMAIL)		
	By View Available On Customer View	CAT'S INTERNET – CUSTOMIZED BANNER		CAT'S – CONTACT REQUEST INTERNET - SIMULATIONS INTERNET - CLICKS STREAM INTERNET – CONTACT REQUEST
HUMAN	**BRANCHE/PHONE** Human Contact	BRANCH & ACCOUNT MANAGER CALL CENTER OUTBOUND MORTGAGE SPECIALISTS INVESTIMENT CONSULTANTS TELEMARKETING	BRANCH & ACCOUNT MANAGER CALL CENTER INBOUND	BRANCH –SIMULATIONS REQUEST PHONE – CONTACT REQUEST

Figure 1 : Table of contact channels

One of the most profitable channels among the ones cited above is obviously the bank branches since they establish close contact with the client. MBCP Branches are equipped with softwares created by the IT team of the bank that are used to track their sales weekly according to their objectives that they must fulfill. Each director of the branch has its own list of clients with whom he is considered the direct contact for any purpose that has to do with any service of pre-sale proposition. This direct contact with the client is the reason behind the success of this marketing channel in comparison to the others cited above

3.2. COMERCIAL TIMEFRAME

From the time perspective, Millennium BCP organizes its commercial balance of retail banking by dividing the year from January to December into 4 quarters which are called “commercial cycles”. Each cycle contains 13 up to 14 weeks. The division might differ from a year to another so as the number of weeks by cycle. It is made taking into consideration a lot of criteria that have to do with holidays and special events of the bank.

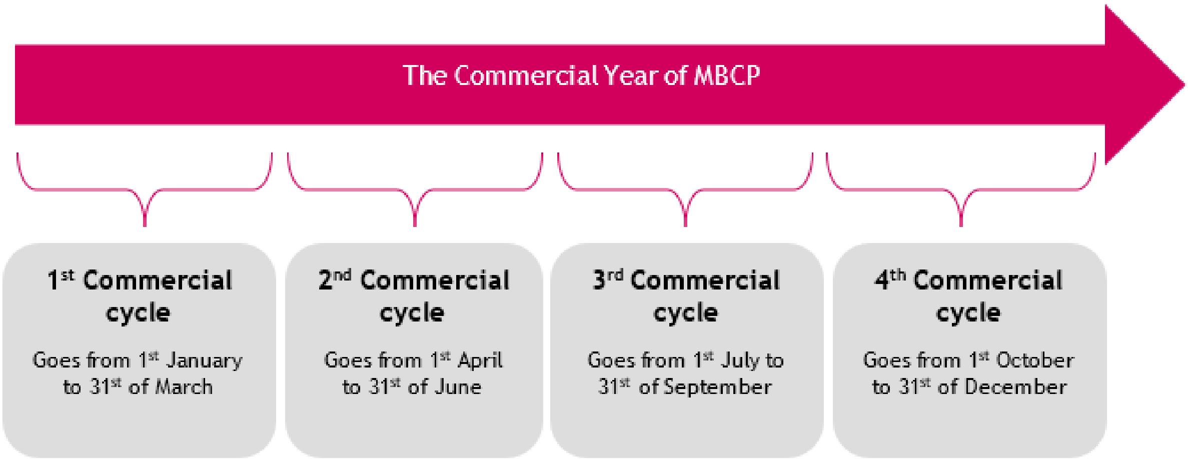

Figure 2 The Comercial year of Millennium BCP

The 1st commercial cycle is considered a very active cycle. As it’s the beginning of the year, it has been noticed that MBCP Clients have more availability to interact with their affiliated branches. Moreover, it is the period known by holidays which make it easy for the bank to target promotion lovers specially during new year’s and the Carnival.

The 2nd commercial cycle is also a good cycle in terms of sales. During this cycle, the CRM team usually focuses on creating marketing campaigns that will retain new clients that have entered during the first commercial cycle. It is also a very active cycle specially in March and April

The 3rd commercial cycle, since it somehow overlaps with summer holidays, attracting new clients becomes a very challenging task during this period. Thus, the sales objectives for this cycle are usually low in comparison to others

The 4th commercial cycle is a period in which business starts to increase from the summer stagnation that is usually occurred in the 3rd commercial cycle. Essentially due to periods such as the school entrance or the new academic year that starts in September.

3.3. The CRM Team and Big Data

The CRM team is playing a very important role in dealing with big data. It is always very challenging to orientate the most adequate product to a small group of clients among the huge clients' base in Millennium BCP. For that purpose, the CRM team is dealing with 3 main types of functions which are the following

- Managing and scheduling the marketing campaigns for client contacting
- Building predictive models for product acquisition using SAS Technologies
- Managing and updating the data from the information systems and store it into the data mart

This figure will illustrate the CRM Team composition and how each of the cells coordinate efficiently to execute their main functions

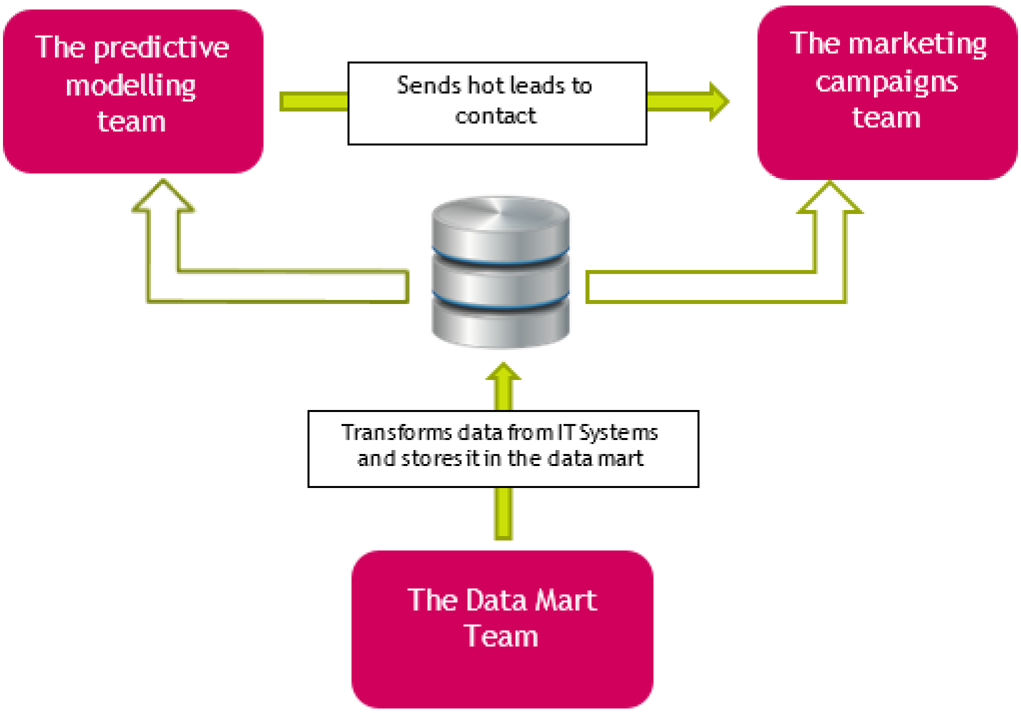

Figure 3 : Mechanism of the CRM Team

The data mart in this use case means only a library or a repository of SAS tables that contain structured data according to different business logics with basic transformations and aggregations of the transactional systems conceived by the IT department (Process that is not covered in the current report). The predictive models team use this information to look at the historical data needed, to build accurate predictive models for specific products. After determining the perfect leads, they communicate them to the marketing campaign that, based on a previous demand from the responsible team of the product, will also create the details of the marketing campaign that will appear in the corresponding channels. The big data part is mainly assured by the data mart team that also deals with storage capacities for the other teams to perform their tasks.

The marketing campaigns that are destined to clients who should be contacted by branches (Outbound) and other channels as well are created by the marketing campaigns team using a platform that acts as an automatic software scheduler for these campaigns to their respective channels. As stated previously, bank branches have **weekly** objectives that should be fulfilled. Those objectives are defined by a department that assesses their performance based on sales made over contacts. The major challenge presented in this workflow is sales tracking. How could the CRM team track their sales on a weekly basis given the challenges this team faces regarding data complexity and the non-stopping growing data size?

4. THE CRM SALES ANALYTICS AS A NEW SOLUTION FOR SALES TRACKING

In this part, the sales process will be discussed in much more depth. There will be presented the main stakeholders that participate in the sales process and across which data analysis was applied. There will also be presented the conceptual model that has been used as an input for the final Sales Analytics solution with a panoply of dashboards implemented for sales tracking making the sales process more flexible and optimized for all concerned parts.

4.1. TECHNOLOGICAL PRESENTATION

Millennium BCP possesses a strong technological infrastructure to support the growing amount of data generated by daily banking operations from its clients. As most of the big companies, Millennium BCP since has always opted for the use of the SAS softwares alongside its panoply of powerful components.

4.1.1. SAS® Software

SAS®, which stands for Statistical Analysis System is a software suite developed by the SAS Institute for purposes of Data Science, Statistics, Multivariate data Analysis, Predictive Analysis and Data Mining. Although the SAS Suite contains more than 200 tools that fulfill different business needs in an automated manner, the main function of the SAS Suite is basically to unite data from different data sources and perform statistical studies, predictive, descriptive and prescriptive analytics for a better understanding of the business. SAS Suite provides a wide variety of technologies with a high level of technicity for users who have tendency to program from scratch, as well as other tools which are based on graphics and drag-and-drop actions that are considered more adapted for users with business knowledge than technical skills.

The main component that is going to be used frequently is the SAS Base. SAS Base or SAS Console is an interface in which there can be written SAS code and executed. The SAS Base is a component that relates to the server of the bank which contains different kinds of banking data. Therefore, data retrieving, and data cleaning is going to take major role in this stage, to come up with the best structured data that will serve as the main input for building the conceptual model.

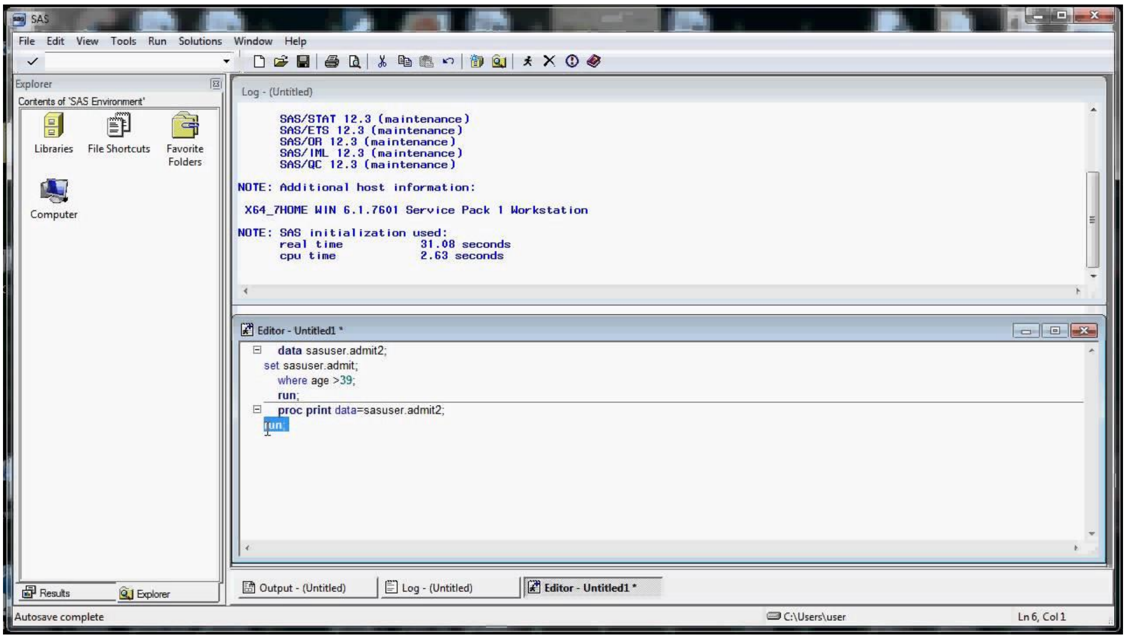

Figure 4 : SAS Base - Left pane: Environment, Upper pane: Logs and outputs of statements, Lower part: Scripts

4.1.2. Microsoft® Power BI

Microsoft Power BI is a Business Intelligence and Business Analytics suite that provides tools for reporting and building dashboards guaranteeing a very simple user-friendly atmosphere for their users aiming for independence from Database Administrators or technical staff.

Power BI contains two main components necessary for supporting, building and sharing reports and dashboards which are the following :

- ***Power BI Desktop***: The only free of cost component in the Power BI suite that enables its users to mainly connect the data sources and make them ready to populate reports and dashboards. Power BI Desktop also offers intelligent capabilities of data editing and data cleansing. It can also be connected to a variety of data sources citing among them excel files, data base management systems (DBMS) as well as other external sources.

- ***Power BI Services***: is a component that is, contrarily to Power BI Desktop, is considered as the web version of Power BI Desktop. It is a cloud-based platform in which dashboards and reports are shared among the rest of the end-users. It has limited editing capabilities as it is much more dedicated for final visualization of the reports and dashboards. However, being connected to a third-party component which is Power BI Gateway, there can be scheduled automatic updates daily, weekly or monthly from the latest version of the data to update the published reports and dashboards without requiring any use of Power BI Desktop.

We can look at these two components in a very simplistic perspective. Power BI Desktop is usually used as a development environment (Building phase), while Power BI Services is used as a Production environment (Tests and deployment phase). Power BI Services' automatic refresh is in fact the main feature that will optimize and automatize the sales process in the CRM Department. It has been 1 year and a half that Millennium BCP has acquired full license for the use of Power BI aiming for flexible data flow between different entities inside the bank as well as for simple data sharing and easy access to data presented in the best adequate graphic visualization. Here below is an example of a power BI Dashboard

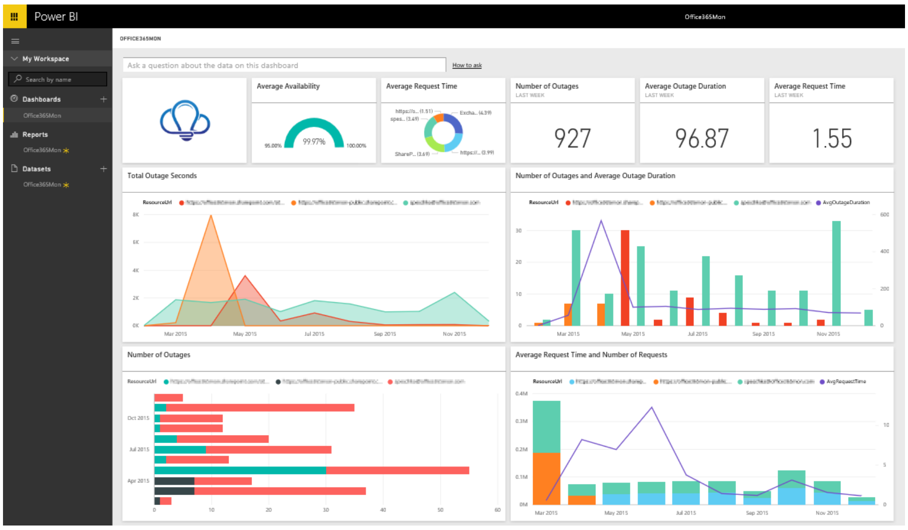

Figure 5 : Example of a Dashboard in Power BI

4.1.3. Microsoft® SQL Server

4.1.3.1. Why a DBMS (Data Base Management Server) ?

The expected scenario that has been planned for the optimized process of sales tracking was connecting SAS datasets and tables directly to Power BI Desktop after preparing the necessary SAS datasets. Unfortunately, there was no possible way to link the data from SAS to Power BI. The main probable reason is the structure with which SAS was built. In fact, this is due to a SAS concept which is "SAS Library". A SAS Library is simply a SAS repository of datasets that is linked to a physical repository in which there can be mapped various data sources. These libraries refer to physical folders in the server. As it is illustrated here below.

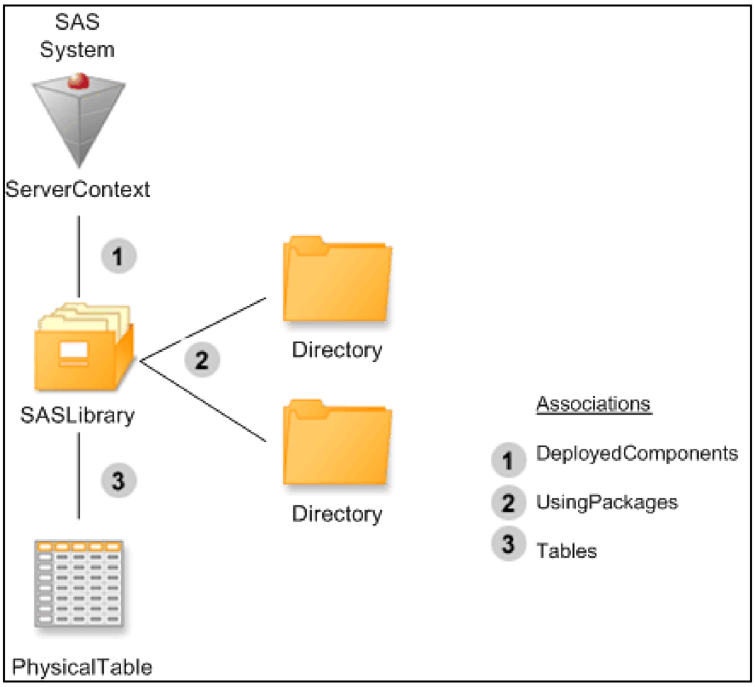

Figure 6 : Structure of The SAS System

Hence, this structure hides the availability of these datasets to other applications including Power BI. Those libraries can contain various datasets which are physically SAS tables having the ".sas7bdat extension". This framework doesn't make of SAS a DBMS with which it can be associated to Power BI. It is in this optic that a DBMS is needed to ensure this connectivity between both sides.

4.1.3.2. SQL Server as a Gatewaty

SQL Server is a Data Base Management System used to store data in databases with high security levels. It enables also a very simple interaction for data selection using SQL queries. In this scenario, SQL Server will serve as a gateway between SAS Datasets and Power BI as illustrated below.

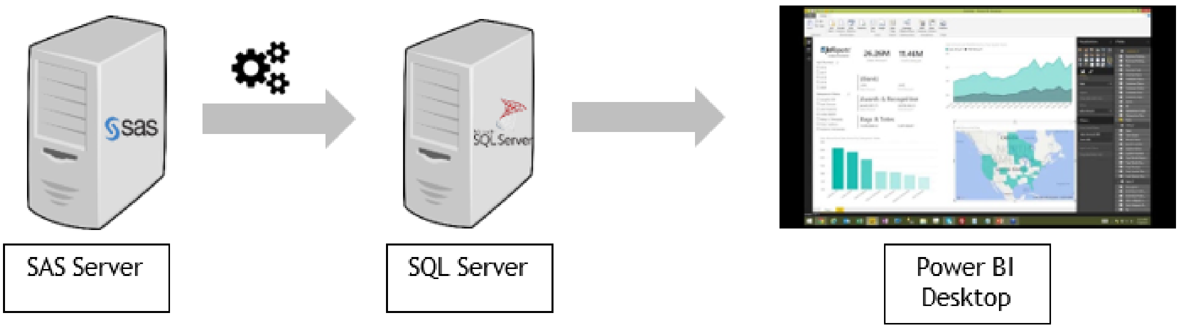

Figure 7 : Mecanism used to build the data model in Power BI Desktop

With this structure, an important part of the work is going to be focused on building a proper data source that is going to gather all the necessary components from the SAS tables required. It has been considered as a best practice, performing the data cleaning, transformation and preparation at the level of SAS Server using multiple tables with which the CRM team works daily. These tables that exist in the Datamart SAS library are updated on a daily, weekly, cyclic (Commercial Cycle) or monthly basis. Some of these tables are going to be used in the process of building the so-called "core table" that will serve as an input in the SQL Server. Later in PowerBI Desktop, the core table will be imported to build the data model

4.2. Data sources and elements required for the CRM Sales Analytics Solution

In this section, there will be presented the main components from the business logic perspective, that will be inputted in the Core-Table. As a reminder, the process of building the Core-Table in the SAS System is not going to be covered since it was not a part of the internship main functions. Therefore, this section will provide detailed information on the CRM Sales process between main communication channels, Marketing campaigns and client contact schedules

To begin with, the figure below is a general overview of the Sales CRM process Management that shows how the CRM Team is playing a major role between different units responsible for product marketing and the different channels for contacting clients.

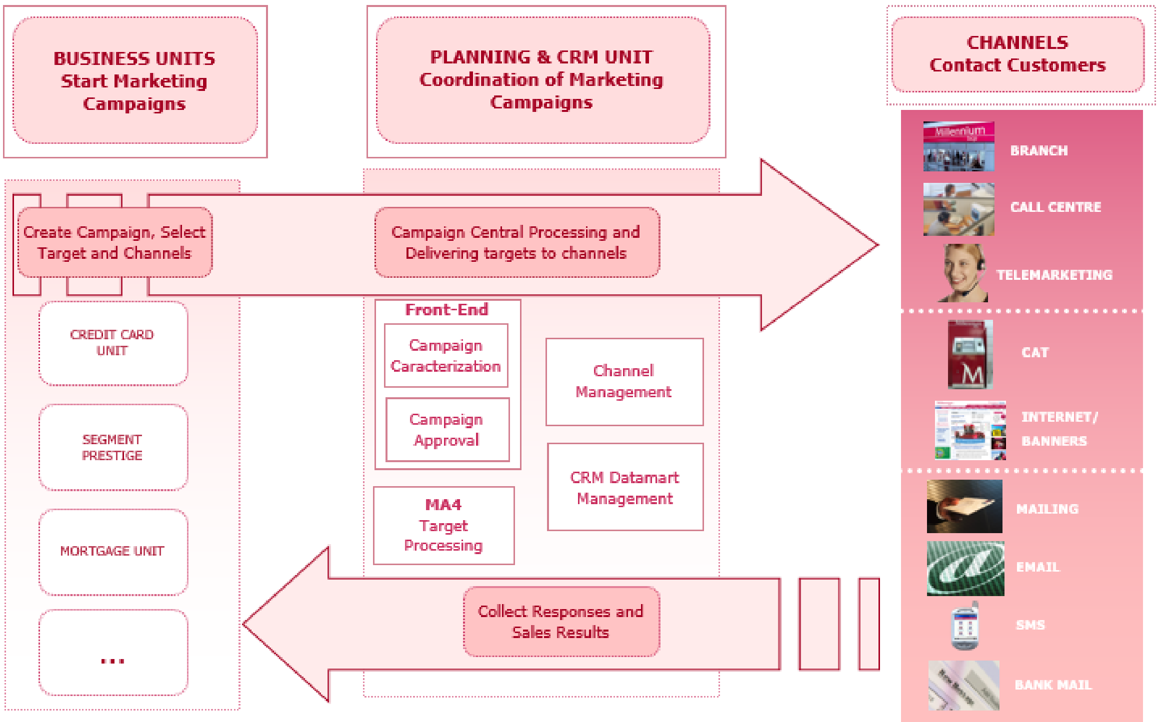

Figure 8 : The CRM Sales Process

It is clear how the CRM Team is positioned as a gateway between product units in the aim of managing and preparing the necessary marketing campaigns to increase sales in a way to guarantee the customer experience in profiting from the wide category of MBCP products. The Sales tracking is made **weekly** as stated before.

Thus, to constitute the core-table, Information is needed about the marketing campaigns, the commercial zone in which the clients are registered, the purpose of contact and the quantity of listed and contacted clients.

4.2.1. MBCP Marketing Campaigns

A marketing campaign is an action that is organized by the bank in the aim of branding a new or an existing product, service or promotion. It can also be made for purposes that are not related to cross-selling such as informative marketing campaigns or other campaigns used to increase client's loyalty and prevent product attrition (Retention campaigns). Usually marketing campaigns target a big number of clients that are considered accurate clients in tems of their profile. Their goal can be to brand one or multiple products. A marketing campaign can have one or more than one contact channel depending on the nature of the product.

In Millennium BCP, as there have been implemented various marketing campaigns. in the core table there will be needed the code of each campaign, its main title, the channel through which it will be launched and the reason or purpose of this marketing campaign.

4.2.2. The Commercial Structure

In the expected Sales Analytics solution, analyzing the sales across all the country to view the sales performance from a geographical perspective would be a strong feature of the solution. That would enable the commercial responsibles for the sales done by all the branches and centers of each locality in the country to track their own sales. To achieve this goal, Millennium BCP adopted a very detailed division of its branches and call centers according to their location. Every entity charged of reaching out the client belongs to a single point in this division. Usually, a marketing campaign is destined to a contacting area (Balcão). This area is the name of the locality in which this entity exists. For example, if the contacting entity existed in Chiado (a very popular neighborhood in Lisbon) the contacting area will be called Chiado. This contacting area hierarchically belongs to a square (Praça). The square itself belongs to another classification of the national level which is the following.

- Northern Part
- Central part
- Southern part and Islands

The following figure shows this national classification on the map of Portugal :

Figure 9 : The Geographical division of Portugal

4.2.3. Time Line and Contacting Process

As it was stated before in several parts, the marketing campaigns are directed to its concerned clients on a frequent basis (weekly, daily, monthly, cyclic). For this matter, the time dimension will be needed. However, this time dimension will be different than the regular time dimension (Year, quarter, month, day) due to the unique timeframe that Millennium BCP uses to quantify its activity (4 Commercial cycles per year with 12 to 13 weeks each). For each marketing campaign, target clients are selected to contact (using automated processes). Next in the contacting process, usually not all the targeted clients are contacted. Only those who present a high propensity of welcoming the offer are subject of campaign. In the next step, from those contacted clients and based on their answers regarding the marketing campaign, clients with a positive intension or expressed an optimistic feedback regarding the campaign are collected. Finally, from this category of people, clients who bought the product or answered positively to the subject of the campaign are also distinguished. The following figure explains this funnel on which the Sales analytics solution will be based for sales tracking

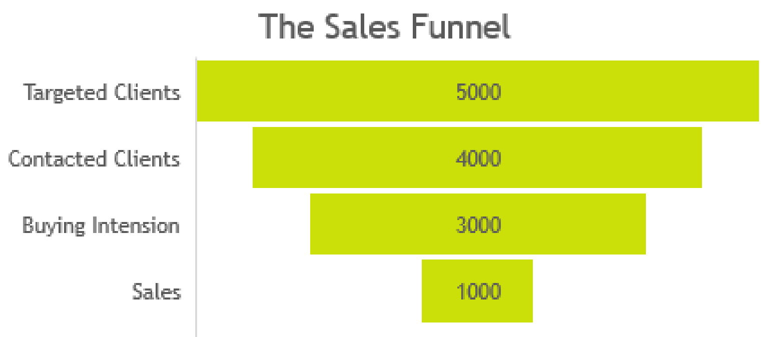

Figure 10 : The Sales Funnel

4.2.4. The final Core-Table metadata

In most cases, the number of targeted clients per marketing campaign is very huge. Therefore, if the core-table was made on the client level, the resulted table would contain millions of rows knowing that in average, one marketing campaign can target up to 10000 clients. As a result, the core-table would be very heavy and difficult to process in PowerBI. For this reason, it has been decided from the CRM data mart team that the core-table ought to be built on the level of each marketing campaign in the respective week of the commercial cycle in which it has been launched. Regarding the number of clients, it will be summarized from targeted clients to clients who accepted the offer. In this figure, there will be shown the metadata of the table core-table.

Column Name	Type	Length	Format	Informat	Label	Transcode
campcode	Text	12	$12.	$12.	campcode	Yes
commcode	Text	16	$16.	$16.	commcode	Yes
campnome	Text	70	$70.	$70.	campnome	Yes
CodBalcao	Num...	8			CodBalcao	No
NomeBalcao	Text	200	$200.	$200.	NomeBalcao	Yes
MacroSegme...	Text	14	$14.	$14.	MacroSegmento	Yes
Cod_Zona	Num...	8			Cod_Zona	No
NomeZona	Text	40	$40.	$40.	NomeZona	Yes
NomeDirCom...	Text	200	$200.	$200.	NomeDirComercial	Yes
Praca	Text	200	$200.	$200.	Praca	Yes
Tipologia	Text	200	$200.	$200.	Tipologia	Yes
CCustoGestor	Num...	8			CCustoGestor	No
NomeGestor	Text	40	$40.	$40.	NomeGestor	Yes
Canal	Text	23	$23.	$23.	Canal	Yes
RazaoContacto	Text	19	$19.	$19.	RazaoContacto	Yes
RazaoContact...	Text	31	$31.	$31.	RazaoContactoOld	Yes
Produto	Text	26	$26.	$26.	Produto	Yes
Ciclo	Num...	8			Ciclo	No
SemCiclo	Num...	8			SemCiclo	No
ListadosCiclo	Num...	8			ListadosCiclo	No
ListadosSema...	Num...	8			ListadosSemana	No
PorTratar	Num...	8			PorTratar	No
Contactado	Num...	8			Contactado	No
Insucesso	Num...	8			Insucesso	No
Sucesso	Num...	8			Sucesso	No
Aberto	Num...	8			Aberto	No
NaoContactado	Num...	8			NaoContactado	No
Expirado	Num...	8			Expirado	No
QtdVendas	Num...	8			QtdVendas	No
MontanteVen...	Num...	8			MontanteVenda	No

Figure 11 : Metadata of the core-table

Campcode : code of the marketing campaign

Commcode : Specifies the segment or category of clinets to whom the marketing campaign is gonna be added

Campnome : Name of the marketing campaign

CodBalcao : code of the contacting area (see page 12)

NomeBalcao : name of the contacting area

Cod_Zona : Code of the zone

NomeZona : Name of the Zone

Praça : Square to which the zone belongs

Tipologia : is Basically the segment of clients to whom the marketing campaign is addressed. It can be mass market, Prestige or Mass plus for individuals

CCustoGestor : It's the contacting center in which the client is registered to

NomeGestor : The name of the manager responsible to call and manage the client's matters directly

RazaoDeContacto: The purpose behind the marketing campaign (Acquisition, Fidelity, Inquiry). There are 2 types of contact purposes used both in different types of marketing campaigns

Produto: the family of products to which the product belongs to (Credit Cards, Debit Cards, Mortgage loans…)

Ciclo: Commercial Cycle

Semana: week of the corresponding cycle

ListadosCiclo: The selected contacts for the cycle

ListadosSemana: The selected contacts for the week

PorTratar: The clients to be contacted

Contactado: The contacted clients

Insucesso : Clients who weren't interested by the offer

Sucesso : Clients who expressed an interest about the offer

Aberto : Clients who's contact plan has been scheduled for next week

NaoContactado : Clients who were listed to be contacted in the week but were not

Expirado: Clients who haven't been contacted in the necessary time in which it should be

QtdVendas: Sales that were realized

MontanteVendas: Sales amount of the sales

4.3. BUILDING THE DATA MODEL IN POWER BI DESKTOP

4.3.1. The main fact tables

As the core table above contains all the information required for the needed solution, it can be imported to Power BI. There are 2 data connection modes in Power BI Desktop. The "Direct Query" mode and the import mode. The Direct Query Connection mode enables the Power BI engine to send queries directly to the data source and brings back the results. Manual data refreshes are not required in this mode. The import connection mode imports a copy of the whole data source in the Power BI file (Available only if the Power BI file doesn't exceed 1GB of memory). As it imports a copy of the data source, manual refreshes are required in this mode. Knowing that the core table is going to increase weekly, using the Direct Query mode on a table that would probably have more than 10 million rows is extremely time consuming at each interaction with the report. Thus, it has been decided to opt for the import mode to connect the core table.

As the table has been imported to the power BI desktop file, it has been built a star schema from it to add more flexibility to our model, especially in terms of interactive filtering. The following figure shows the final data model that was built inside the Power BI Desktop.

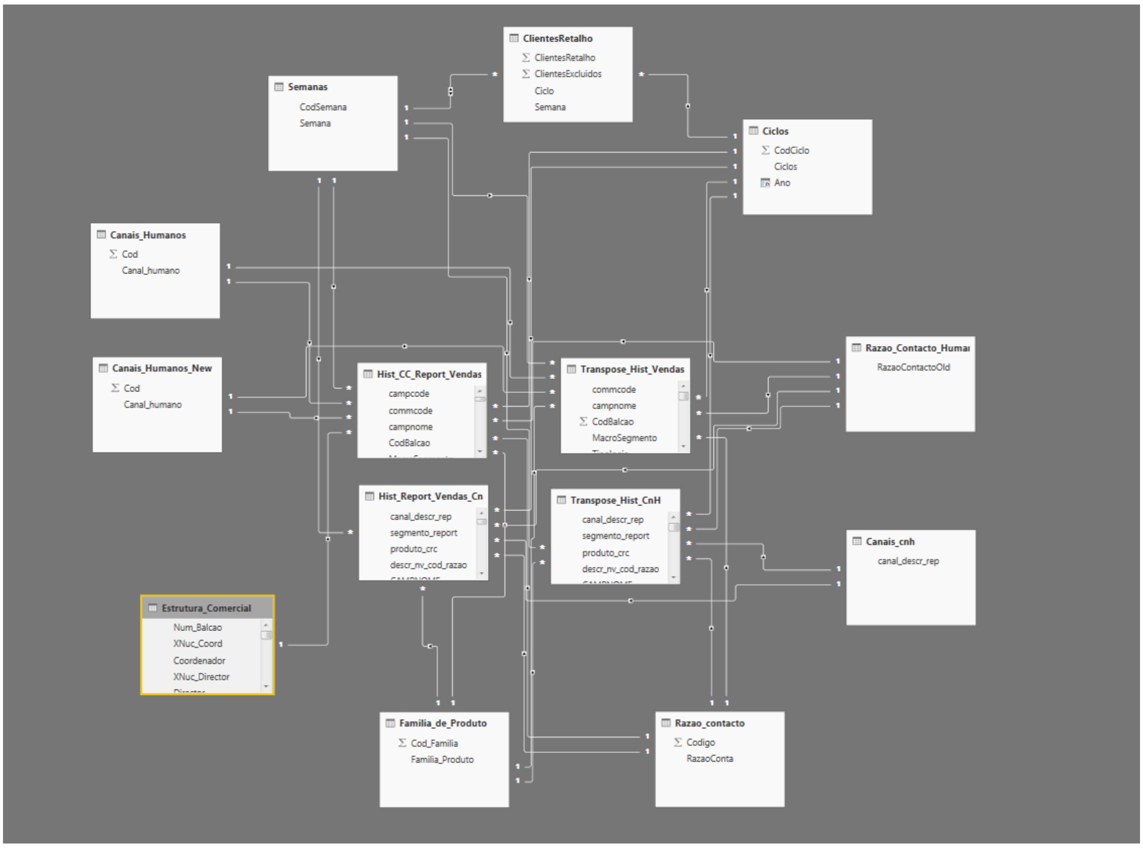

Figure 12 : The data model of the CRM Sales Analytics solution

In this model, the core tables (can be called as fact tables) is the Hist_cc_report_vendas and the Hist_report_vendas_cnh. Both are almost the same. What differs between the two tables is that the first one gathers the marketing campaigns communicated through the human channels and the other contains the marketing campaigns communicated through the non-human channels. From these two core tables, there has been generated two other tables which are the corresponding transposed tables of the two core tables.

The purpose of doing so is to generate the sales funnel (Page 14). The power BI desktop visuals are sophisticated items that take one or various variables from the data source and automatically generate the graph. For generating the funnel graph, the input of its correspondent visual should be a variable containing the values of ("Listed", "Contacted", "Not contacted", "Success", "Fail", "Sale") as many times as the number of these clients.

Below is a small example of a dummy marketing campaign of credit cards that will illustrate the transpose operation applied for both tables to include the funnel graphs for both human channels sales and non-human channels sales in the Sales Analytics Solution.

This is one row from the core table (Either ways which one)

Marketing Campaign	Contacting area	Contacting zone	Square	Channel	Contacting Purpose	Listed	Contacted	Interested	Sale
DCAR-O716	Baixa	Lisboa	Lisboa	Branch	Aquisition	6	5	3	1

The output of the transposed table would then be the following

Marketing Campaign	Contacting area	Contacting zone	Square	Channel	Contacting Purpose	Category	Value
DCAR-O716	Baixa	Lisboa	Lisboa	Branch	Aquisition	Listed	6
DCAR-O716	Baixa	Lisboa	Lisboa	Branch	Aquisition	Contacted	5
DCAR-O716	Baixa	Lisboa	Lisboa	Branch	Aquisition	Interested	3
DCAR-O716	Baixa	Lisboa	Lisboa	Branch	Aquisition	Sale	1

One row of the core table would basically correspond to 4 rows in the transposed tables. The category variable will then contain the values wanted in the input of the funnel sales. And their sum would be the values in the "value" variable. From there the necessity of building two transpose corresponding fact tables.

4.3.2. The Dimension Tables

Knowing that the core table is the only data source provided by the data mart team, the dimension tables that will complete the data model were created and deducted from the values of the corresponding variables in the fact tables. In fact, when building a model in Power BI desktop, relationships creation between tables requires that one of the tables' field should have unique values. This is not a limitation of the software but rather a principle in relational databases (Primary and foreign keys). There will be guaranteed a fast search or filtering when interacting with the report. Therefore, the main dimension tables would contain unique values that would easily ensure the relationship with the fact tables.

These are figures about the main dimension tables that were needed for the Sales Analytics Solution.

Canal_humano
UVD Outbound
CC Outbound
Sucursal Inbound
Sucursal

Figure 13 : Human channels dimension table

canal_descr_rep
Email Web Empresas
SMS
Banner Web Particulares
Email
Banner Web Empresas

Figure 14 : Non-human channels dimension table

Concerning the time dimension, Commercial managers in branches of the bank are most of the times interested by the **weekly** sales. Whereas the commercial directors of the entire zone in which branches operate are more interested to analyze the entire commercial cycle. Thus, the Sales Analytics Solution would not drill deeper than the week hierarchy. As each commercial cycle has between 12 and 13 weeks, a cycle table alongside a week's table have been created with respectively unique values to the periods of the marketing campaigns.

CodCiclo	Ciclos	Ano
1	201504	2015
2	201601	2016
3	201602	2016
4	201603	2016
5	201604	2016
6	201701	2017
7	201702	2017
8	201703	2017
9	201704	2017
10	201801	2018
11	201802	2018
12	201803	2018
13	201804	2018
14	201901	2019

Figure 15 : The commercial cycles dimension table

As the marketing campaigns were promoting a different variety of products, analyzing the sales at the product level is really relevant to the commercial directors and branch managers. Millennium BCP has a fixed classification of all the products. They are grouped into families of products "Familias de produto" or product categories. It is through this classification that commercial directors need to have the information per each product's family.

Familia_Produto
1. Clientes
2. Ordenados
3. Investimento e Poupança
4. Vencimentos
5. Crédito Habitação
5. Crédito Negócios
5. Crédito Pessoal
6. Soluções
7. Seguros
8. Cartão Débito e Crédito
9. Serviços e Digital
99. Outros

Figure 16 : The products' families Dimension Tables

Sometimes it might be important to closely analyse the behavior of the clients regarding the marketing campaigns. Knowing that any marketing campaign must have one commercial purpose (Acquisition, Fidelity, Data update…), It is important to track the marketing campaign from various purposes. The hierarchy of the contacting purposes can differ thus it can always be relevant to analyze the sales through this dimension.

RazaoConta
Fidelização
Fidelização Cliente
Informativo
Inquérito
Newsletter
Onboarding
Retenção
Venda

Figure 17 : Contacting purposes for Non-Human channels

RazaoContactoOld
Divulgação / Esclarecimento
Upgrade
Alerta Produto
Ativação Cartão
Estimulação
Renovação/Vencimentos
Atualização de Dados
Aquisição
Inquérito Momentos Verdade
Fidelização Produto
Acuso Pin
Fidelização Cliente
Substituição de Produto/Serviço
Alerta Cliente

Figure 18 : contacting purposes for Human channels

4.4. Results and Reports

There have been implemented various KPIs for the CRM Analytics Solution. KPIs are indicators with which it can be assessed the performance of the CRM Marketing campaigns not only sales wise but also to understand the global interest of the clients of millennium BCP towards each product category. These KPIs were developed using the DAX Language that is integrated in Power BI Desktop. Most of the indicators that were implemented in the solution concerned mainly Sales and Contacted Clients. The ideal situation would always be a greater percentage of sales over contacted clients. Even better if there were few contacted clients which gives an idea about the good accuracy of listed clients.

Here below is a list of the main KPI used for the CRM Sales Analytics Solution (For formulas look at the annexes)

KPI Code	KPI Name	Description
M_CicloAnterior	Previous commercial cycle	KPI that is used by other KPIs for comparing the performance of the previous commercial cycle
M_UltimaPassadaSemana	Previous week	Used to get retrieve KPIs of last week
M_Contactados	Total of Contacted Clients	which is the sum of all the contacted clients
M_ListadosCiclo	Total of Listed Clients (Cycle)	Listed clients for the current cycle
M_ListadosSemana	Total of Listed Clients (Week)	Sum of the clients to be contacted in the week (Portratar) and the clients which contact has been scheduled for next week (Abertos)
M_Sucesso	Total of Success Clients	Total of the clients with a positive intension towards the marketing campaign

M_Vendas	Total of sales	Sum of the clients who concretized the sale through the marketing campaign
M_tx.Contacto	Ratio of Contacted Clients	Total of clients contacted divided by the total of listed Clients
M_Tx.IntCompra	Ratio of Success	Sum of the success clients / Sum of contacted clients
M_Tx.Venda	Ratio of Sale	Sum of sales / Sum of Contacted Clients
M_Vendas_CicloAnterior	Total of Sales in the previous Cycle	Total of the sales in the previous Cycle positioned in the last week of the Cycle
M_Vendas_SemanaPassada	Sales of Last week	Last week sales in comparison to the current week
Prev_M_Tx.Contacto	Contacting Ratio of previous Cycle	
Prev_M_Tx.Venda	Sales Ratio of the previous Cycle	

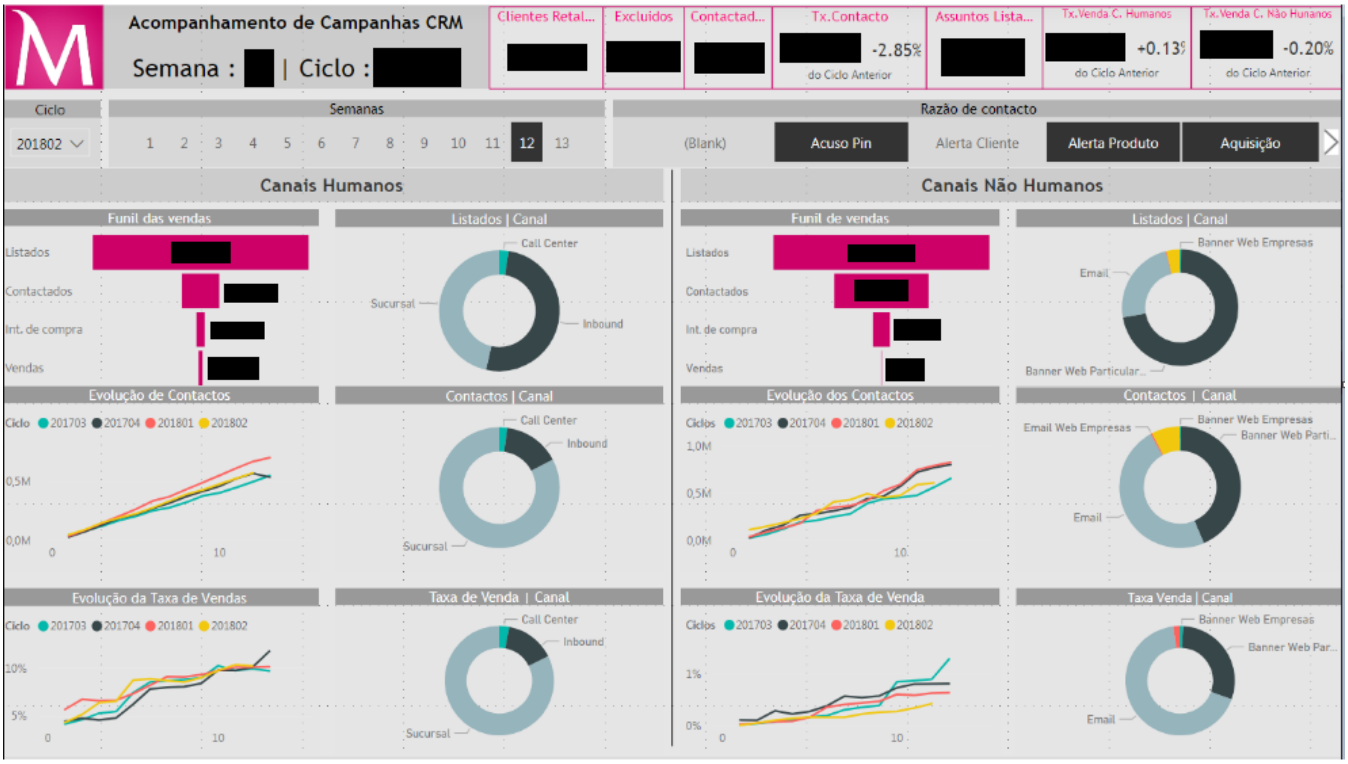

Figure 19 : Global View of the CRM Sales

The upper dashboard presents a global view of the sales in both human canals (Right panel) and non-human canals (Left panel). In both panels, we can see the funnel graph from listed clients to sales that was created using the transpose version of the core-table, the distribution of Listed and contacted clients per channel, the evolution of the contact acceptance rate and the sales rate per cycle and week (each line represents a cycle). On top of the Dashboard there are presented Indicators for the current cycle among them there are, all Clients listed from the Retail Segment, the excluded ones from being listed from marketing campaigns due to high risk or imparity signs, contacted clients, Contacting rate and a comparison of this value with the previous cycle (2.85% less than the previous cycle) and finally both sales rate for both human and non-human canals alongside their difference between the previous cycle.

5. THE NEXT BEST OFFER: A GATEWAY TO EFFICIENT CUSTOMER-CENTRIC RELATIONSHIPS IN THE CRM DEPARTMENT

Nowadays, in a world where clients are becoming more exigent about their needs and expectations from a product, companies must adapt to this sudden twist by investing massively for a marketing with efficient strategies and powerful longstanding customer relationships. It is in this perspective that some companies took advantage and coupled the traditional marketing with Analytics. This component made marketing very precise. However, the predictive models for product acquisition that were designed to give a strong facet to a modern marketing were found to behave very poorly and inadequately in some situations despite the use of significant, powerful and accurate predictive models. At this level, companies started to understand that an attention should be paid more at "the customer level". Let's say for example that clients who are more likely to get a mortgage-loan are the ones who have more logins to the banking phone application. There might be an important number of cases in which young clients who access the app various times to only check their balance receive as well offers for acquiring a mortgage-loan. These situations might be rare to happen. But on a larger scale, especially for companies with a clients' base of more than one million clients, it is having a negative impact on the relationship bank-customer inducing him to decrease his satisfaction due to offers he is not interested in. Therefore, offers should be sent with the certainty that it's going to be an offer the client is less likely to refuse. Consequently, the transition to an analytical product-centric marketing happened to be inefficient though the successful use of strong machine learning techniques in the implementation of the necessary predictive models.

Being aware of this actual situation, Millennium BCP is intending to take a further to adopt a strategy that is more client-centric. The bank launched the challenging project of The Next Best Offer (or action) that has as main objective determine the most adequate product, service, promotion, or product upgrade for the client at the right moment, the right time and throughout the right channel. This project will be the fruit of various machine learning and data science techniques that will be applied on different sides of the project as the major challenge is to have a 360 degrees client vision from a big amount of data.

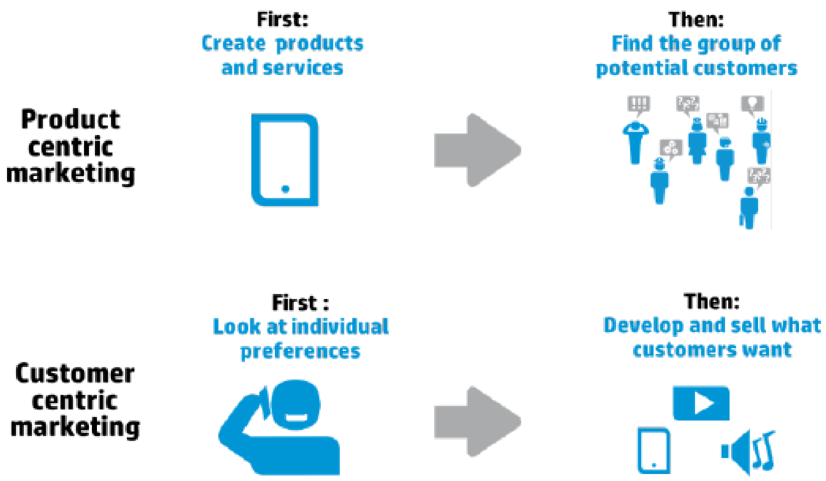

Figure 20 : The switch from product-centric marketing to client-centric marketing (Andrea Fabrizi, 2014)

In this present chapter, there will be presented a brief overview of the initial simple solution of the Next Best Offer, under the name of NBO1.0 (Next Best Offer 1.0), followed by the methodology adopted to Build a more robust solution that considers other aspects of the clients that are not related to his purchasing profile. Also, there will be explained more in-depth each component of the NBO 2.0 analytics engine as well as the rating grid that enables it to reach the best offer for the client.

5.1. The existing solution of the NBO 1.0

5.1.1. Predictive models of the CRM Team

As stated previously, the CRM has 3 teams. The Data Mart team, The Marketing Campaign team and the predictive models' team. The predictive models' team has implemented a panoply of powerful predictive models. Most of them are generally used to study the outcome of clients acquiring a new product. Others are used to calculate the churn rate (Ex Churn rate from becoming a client of Millennium BCP, Churn from acquiring the banking solution "Programa Prestige"). Sometimes during the building of these models, the universe of clients might be restricted only to a segment of clients. (Mass Market clients, Plus Clients, or Prestige Clients which are the 3 categories of retail individual clients) to increase diversification of products possession across other types of clients. Here is a list of the predictive models that were built in the predictive models' team

- Personal Loans (3 models one for each segment)
- Habitation Loans (3 models one for each segment)
- Credit Cards (3 models one for each segment)
- Millennium GO! (All segments included)
- Programa Prestige (All segments included)
- Bank Attrition Model (All segments included)
- Cliente Frequente (All segments included)
- Mais Portugal + Programa Prestige (Hybrid model for the segment of clients living outside Portugal)
- Móbis
- Cliente Frequente (Segment of Negocios)
- Other models for the companies' segment

5.1.2. From predictive models to the Next Best Offer 1.0 (NBO 1.0)

These models provide scores (from 0 to 1) which are themselves classified by the Score category (Low, Medium low, Medium High, High, Superior). For each client, the scores from the corresponding models in which he has been subject to analysis are calculated monthly. Therefore, the approach of the Next Best Offer 1.0 was the following: **The next best offer is simply going to be the product which model gives the highest score across all the other models**. Consequently, each month with new scores coming, there is a monthly Next Best Offer. There were a lot of marketing campaigns that were created by the Marketing Campaigns team to establish contacts to these clients proposing them monthly offers of their most suited product. Unfortunately, most of them didn't have success and that's due to several reasons. The existing predictive models do not cover all the products and the upgrades of each product. Moreover, it doesn't include clients' data that would give an overview of how strong the client's relationship with the bank is, even outside of the commercial perspective. This approach of the Next Best Offer is very limited and was proved that it has poor results judging from the Marketing campaigns. Instead, this version of the Next Best Offer 1.0 will serve as an input for the second version of the Next Best Offer 2.0 alongside other components related to Predictive Analytics.

5.2. THE NEXT BEST OFFER 2.0 (A THEORETICAL ROADMAP)

Having a 360 degrees view requires an onerous effort of data collection about all the client's information. It would have also been interesting to have real time data or at least fast refreshable client's data (Pseudo Real Time). This would be the starting point to a new innovative solution that would maximize the sales of the bank since the offer is built from that client's portfolio. Then this portfolio would be the input of an engine that will be implemented, composed by different machine learning algorithms that would intelligently detect the best suited product to offer for the client.

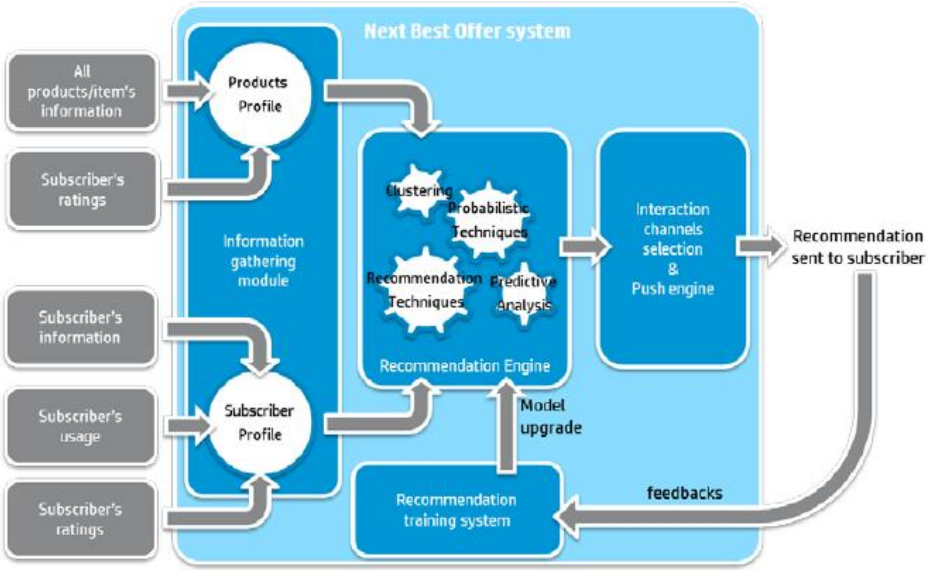

Figure 21 : Conception of the NBO 2.0 architecture (Andrea Fabrizi, 2014)

The upper architecture in Fig 20 from (Andrea Fabrizi, 2014) presents a robust version of the expected NBO 2.0. It is composed mainly of 3 components.

- ***The Information Gathering module***: Which is all the part of data gathering. It will itself contain mainly information about the products catalog and the subscriber's (client) preferences, products that he doesn't own yet, the history of his purchases, his reaction towards bank offers and his previous transactions

- ***The NBO Engine:*** Considered as the heart beating of the whole solution. It will contain a panoply of data science and machine learning techniques that will take the input data provided by the gathering module, then find clear patterns upon which it will be determined the product that is more likely to be purchased by the customers

- ***The Interaction Channels:*** Considered as the client's most preferred communication channels as well as the best timing for contacting him about the offer. These 2 points are considered the most important factors for increasing the probability of success of the offer.

The final feedback issued from the client following the predicted best offer is also participating in empowering the NBO Engine. This is how the NBO engine learns the recommendation training system

5.3. The Millennium BCP Roadmap for the implementation of the Next Best Offer 2.0

(Andrea Fabrizi, 2014) in their technical HP Paper of "Next Best Offer: How to re-think your marketing" have defined a strong customer-centric platform that predicts the best suited product for each of their client. Their solution relies strongly on the information gathering module and the NBO engine. This solution is usually the recommended one for companies wishing to switch their traditional marketing. Therefore, since Millennium BCP is a bank that offers a big number of products and services to a wide universe of clients and companies, the architecture from (Andrea Fabrizi, 2014) cannot be followed nor applied to the letter due to the lack of some important aspects. For instance, the bank does not possess a fixed rating system that can be easily applied following the usage of a given product by a given client. There exist some reports about customer satisfaction only for the most purchased products thus it is needless to draw a rating from these reports. Moreover, it would be time consuming to collect ratings from clients upon basic services or products they might even be neutral towards them. Another reason is that more than 90% of the products that a bank offers are products that if purchased by clients might present a percentage of risk. Risky products for example such as a habitation loan or a personal loan cannot be recommended to a client that presents imparity signs in the bank

Consequently, only some components of the (Andrea Fabrizi, 2014)'s model would be considered as sources of inspiration to reach an adequate roadmap for an implementable solution. Hence, The Millennium BCP roadmap for the implementation of the Next Best Offer 2.0 would mainly rely on 5 pillars which are the following.

- The list of the recommendable products
- The communication profile of each client
- Recommendation System through the similarity Matrix
- The propensity models
- Visits and simulations in the MBCP Website

The output of these 5 pillars will be gathered to a scoring system that will attribute a score to each pillar, then the next best offer will be deduced from the pillar with the highest score. A consideration would also be given to the pillar that has the second score. The following figure explains how the new architecture of the Next Best Offer would work.

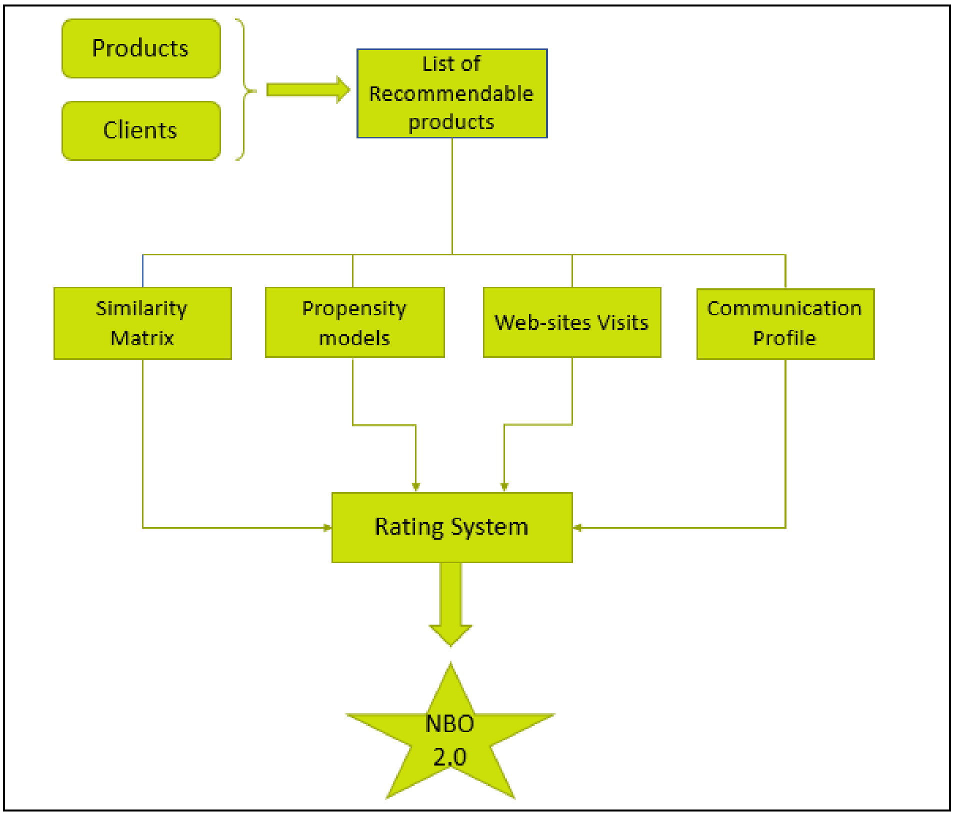

Figure 22 : Millennium BCP roadmap to the Next Best Offer 2.0

The list of recommendable products englobes all the products that can be recommended to clients who haven't own them yet. From this list that also has information about the clients who possess those products or not, the communication profile will be drawn on each of those clients as well as their website visits about each product page, the scores provided by the predictive models (NBO1.0) and the recommended products given by the similarity matrix. After that, these results are assessed by a rating system that attributes coefficients to the components that provide a product to recommend

5.3.1. The list of the recommendable products

The bank provides a wide variety of products with many customized functionalities oriented to clients from different segments other than mass-market. Thus, to reach the most suitable product to recommend to the client, a list should be established for products that can possibly be recommended to the client. To know that, the clients must check all the eligibility rules that this product has. On a large scale, the initial base of clients to analyze must check all the conditions for all the products in the list of the recommendable products. By adopting this strategy, there will be avoided cases in which the next best product to recommend for a client would be a product that he doesn't have the right to acquire. Therefore, the starting point for building this list would be defining the list of the products with respect to their product categories. Then the clients would be deducted by gathering all the clients that satisfy all the eligibility rules stated by each product category.

Figure 23 : Building the List of recommendable products

This process can be repeated a lot of times as long as the definition of this list depends on the products that were chosen before. The choice of the products can be done by 3 different ways: The first one is using all the products of a product category. In this case, there won't be a lot of eligibility rules to verify which means clients with interest over one product category would be more likely to be targeted and their eventual next best offer would be more of an upgrade or a substitution of a product. Another way is to guarantee that the products chosen would be from different products categories. In this case, the next best offer would be a product to increase diversity possession. Then the other way is more oriented to the business logic in a way that some products are chosen for analysis due to some pre-defined rules by other bank divisions.

5.3.2. The Communication Profile of the client

Contrarily to the other pillars, this aspect of the analysis doesn't relate strongly to the next best offer. However, as it is considered a behavioral study for each client. It will help give a general idea of how the client reacts to the previous marketing campaigns communicated to them. Therefore, it has been selected only the communications of marketing campaigns with acquisition or sale purposes to make sure we get the proper behavior from the client towards this type of marketing campaigns. In the DataMart SAS library, there is a daily updated table that gathers all the marketing campaigns that the client received. Depending on the type of response the client will give following the campaign, a rate of positive reception of the offer suggested by the campaign is created to assess the openness of he clients to the acquisition/Sale marketing campaigns. This analysis has been done following two approaches. Analyzing all the communications through human and non-human channels to have an idea of what is the favorite means of communication of the client. The second approach is looking at the communications received by the clients and group them by product category. Therefore, the output will be a personal profile per client for each product category. That will help to determine if the client has a preference or interest for some types of products over the others.

	Num_cli	COMUNICACOES_H	CONT_COM_SUCES	CONT_SEM_SU	CONT_POR_DECISAO	CONT_EXPIRADO_	TX_SUCESSO_H	PERFIL_COM_H
1	68	2	1	1	0	0	50.00%	3-MODERADO
2	72							
3	82	2	1	0	1	0	50.00%	3-MODERADO
4	85							
5	93	1	1	0	0	0	100.0%	5-ALTAMENTE RECEPTIVO
6	129	2	1	1	0	0	50.00%	3-MODERADO
7	133	4	2	2	0	0	50.00%	3-MODERADO
8	207	10	0	10	0	0	0.00%	1-FECHADO
9	258	5	0	4	1	0	0.00%	1-FECHADO
10	260	1	1	0	0	0	100.0%	5-ALTAMENTE RECEPTIVO
11	299							
12	324							
13	325	3	0	3	0	0	0.00%	1-FECHADO
14	334	1	0	1	0	0	0.00%	1-FECHADO

Figure 24 : Table of the clients and their communication profile

There has also been created categories based on the values of the success rate of the past communications to make it easier for other future analysis. The categories have been classified as the following rules shown in this piece of code

```
RSUBMIT;
DATA ZIAD.PERFIL_COMUNICACAO_CLI;
SET SUMMARY_CLI;
TX_SUCESSO = CONT_COM_SUCESSO / COMMUNICACOES;
FORMAT TX_SUCESSO PERCENT8.2;
LENGTH PERFIL_COM $22.;
IF 0 <= TX_SUCESSO <0.2 THEN PERFIL_COM = "1-FRIO";
ELSE IF 0.2 <= TX_SUCESSO <0.4 THEN PERFIL_COM = "2-POUCO FRIO";
ELSE IF 0.4 <= TX_SUCESSO <0.6 THEN PERFIL_COM = "3-MODERADO";
ELSE IF 0.6 <= TX_SUCESSO <0.8 THEN PERFIL_COM = "4-RECEPTIVO";
ELSE IF 0.8 <= TX_SUCESSO <=1 THEN PERFIL_COM = "5-ALTAMENTE RECEPTIVO";
RUN;
ENDRSUBMIT;
```

Figure 25 : Classes of the communication profile

The classes are translated here:

- 1-Frio = Cold (Success rate between 0 and 20%)
- 2-Pouco Frio = Less cold (Success rate between 20 and 40%)
- 3-Moderado = Moderate (Success rate between 40 and 60%)
- 4-Receptivo = Reptive (Success rate between 60 and 80%)
- 5-Altamente Receptivo = Highly Receptive (Success rate between 80 and 100%)

Looking at both approaches, it has been opted to adopt the second one that includes the communication profile per product family, maintaining the same classes explained above.

5.3.3. The Recommendation System using a similarity matrix between products

5.3.3.1. Definitions and theoretical concepts

Following the same logic of the Next Best Offer, there exist many products purchased by a big number of clients. One reason behind may be the positive satisfaction of the clients or simply because they are the most sold products in the bank. One good idea is indeed try to identify the group of clients that do not possess the most sold products in the aim of increasing upselling. This is the goal of the Recommendation Systems that are going to be implemented in the bank. Among the main pillars of the NBO Engine, it will be granted a special attention to its output since it focuses on the products that the client is more likely to buy according to (Miao Nie & Shanshan Cong, 2016). According to the same reference, A Recommendation System is a machine learning technique that consists of suggesting one or several items to be obtained by a user (Client in this case). There are various approaches that can be used to recommend products to customers. (Robin Burke, 2008) distinguishes between 5 different ones. Only the main 3 mostly used will be cited.

- ***Content-Based:*** This type of recommendation systems recommend products to clients based on the positive feedback that they had upon their possession of these products. It also calculates the degree of similarity between products based on the characteristics that each product contains.
- ***Collaborative filtering:*** This recommendation system finds patterns between products liked by a big number of clients. It calculates the similarity based on the rating given to each of the products by the clients. Therefore, it is used to recommend the similar products not possessed by other clients to themselves.
- ***Demographic:*** This type recommends items to clients with a particular demographic background (Age,Country, Income etc…). It is usually applied on clusters of clients that have different demographic profiles to understand the preferences of each cluster that are assumed to be different.

5.3.3.2. The Recommendation System practically in the CRM Team

A content-based recommendation would be the ideal type of recommendation systems to build since it takes into consideration not only the reviews of the client but also the features of each product. Thus, the outcome of this approach will be based on both clients and items. But given the fact that the bank has 10 product categories with hierarchies that can go up to the 5th level of granularity which would result into more than 800 different products, collecting reviews on each product from each client from the monthly customer satisfaction reports would be a very heavy task. Thus, it has been opted to build the CRM Recommendation System using the collaborative filtering method without considering the reviews or ratings form the customer. (Rajendra LVN, 2014) have drawn a roadmap of building a user-based recommendation system. In other words, it is based on whether the client read or not read a book for example. Thus, instead of liking the product which is the main event on which most recommendation systems rely on, there will be considered the event of possessing a given product in the bank or not.

Given the massive number of single products in the bank, computational resources will be highly needed to perform the process of computing similarity matrices. Thus, to avoid this problem, it makes sense to proceed first by doing some pre-prediction on the category of the Next Best Product that the client is more likely to buy. Then, build the recommendation system based on the products of this product category. Some product categories like debit cards or credit cards count more than 70 different cards. Computing a recommendation system with 70 products is also considered as resource consuming.

Therefore, it has been chosen to compute the recommendation system between the first 8 more detained products within this category. In the other hand, if the category has less than 8 products, all the products will be used to build the recommendation system.

5.3.3.3. Roadmap for building the CRM Recommendation System

Many approaches can be explored for determining the NBO's product category of a given client. The most straightforward candidate families of product would be the ones whose clients do not own yet any product from that category. Consequently, the easiest product category to sell from a marketing point of view would be the one candidate. The problem can also be tackled analytically using two different methods.

- Using what would be called as an "outer-recommendation system" (Applied on product categories)
- Using a predictive model with a multi-nominal target
- Using a hybrid method of both above

Building the CRM Outer-Recommendation system

As stated before, this model will serve to predict the product category that the client is more likely to purchase. In other words, there will be depicted the different behaviors of the clients towards the panoply of products offered by millennium BCP. This overview would come with a great help towards the probable Next Best Offer of each client.

The starting point of building a recommendation system is what is called A sparse matrix.

A sparse matrix is a table which columns are the different product categories and rows are the clients. Each client would have a 1 in the corresponding product category if he possesses a product in that category, 0 in the contrary case. Therefore, the Sparse Matrix A would be the following

$$A(i,j) = \{ a_{ij} \} \mid a_{ij} = \begin{cases} 1 \ if\ the\ client\ i\ has\ a\ product\ in\ category\ j \\ 0\ otherwise \end{cases}$$

Based on this matrix, it will have to be assessed the degree of similarity between all pairwise categories. In other words, it will be measured the ability of how 2 items are possessed together throughout the whole universe of clients. The more this measure is, the more similar are the categories. There are a lot of similarity measures but there will be used the most common one used across recommendation systems which is the cosine similarity. The cosine similarity is expressed as it follows with n the number of clients analyzed in the sparse matrix, x and y are two products:

$$s(x,y) = \frac{x.y}{\| x \| * \| y \|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 * \sum_{i=1}^{n} y_i^2}}$$

With analogy to the cosine function it is expressed by the scalar product of both product categories divided by the norm of each product category. This measure is based on calculating the ratio of matching ones in the Sparse Matrix. In other words, it is a measure based on which one can have an idea about the likelihood of two products to be possessed by the clients. It is also similar to the logic of an association rule that can be built from frequent item sets. The more the similarity is close to one the more similar are the products.

One efficient way to calculate the similarity matrix is to use the matricial representation. The similarity matrix would then be the multiplication of the transpose of the standardized sparse matrix and itself.

$$S = \|A\|^t \, \|A\| = \{s_{ij}\}$$

It has been necessary to standardize the matrix and its transpose using the Euclidean norm. The expression of a standardized vector v in the sparse matrix is the following:

The similarity matrix is a symmetric matrix. That is due to the pairwise calculations of the scalar product similarities that are present in the terms of the matrix

Thus, the similarity matrix would be represented as the following with $s_{ij} = s_{ji}$

$$S = \begin{pmatrix} 1 & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & 1 \end{pmatrix}$$

This is the theoretical starting point that is going to be applied on the information of client's possession of product families to build the outer-recommendation system. They are going to be considered 10 different product categories for our analysis which are the following:

Category	Name in the SAS Sparse Matrix
Savings Products	IND_X01_PO_EMEA
Personal loans	IND_X04_CP_EMEA
Habitation loans	IND_X06_CI_EMEA
Insurances	IND_X08_SR_EMEA
Integrated Solutions	IND_PACOTE
Investment products	IND_X01_IV_EMEA
Mobis (Partership product)	IND_MOBIS
Médis (Partnership product)	IND_MEDIS
Homing (Partner Products)	IND_HOMING
Credit Cards	IND_X03_CC_EMEA

The first step is obviously to build the sparse matrix. The sparse matrix was built using possession tables of the DataMart SAS Library. It has been chosen to opt for the client's possession rather than sales. The difference between sales and possession tables is that sales tables register consumed loans by clients whereas the possession tables do not keep this information. Thus, people who have pending loans will appear in possession tables. This approach has been adopted assuming that the behavior of some clients might change after finishing their pending loans. Thus, the sparse matrix will be built with ongoing habitation and personal loans.

The following figure represents the top 50 rows of the sparse matrix

IND_X01_PO_EMEA	IND_X01_IV_EMEA	IND_X03_CC_EMEA	IND_X04_CP_EMEA	IND_X06_CI_EMEA	IND_X08_SR_EMEA	IND_PACOTE	IND_MOBIS	IND_MEDIS	IND_HOMIN
0	0	1	0	0	0	1	0	0	0
0	0	0	1	0	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0
0	1	1	1	0	0	0	0	0	0
0	1	1	1	0	0	1	0	0	0
1	1	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	0	0
1	1	1	0	0	1	1	1	0	0
0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0	0	0
0	1	1	0	0	0	1	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0
1	1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0	0	0
0	0	0	1	1	1	1	0	0	0
0	1	1	0	0	1	1	0	1	0
0	0	0	0	0	1	0	0	1	0
0	1	1	1	0	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	0	0	0
0	0	1	1	0	0	1	0	0	0
0	0	1	1	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0	0
0	1	0	0	0	1	0	0	1	0
0	1	0	0	0	0	0	0	0	0
0	0	1	1	0	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	1	0

Figure 26 : Sparse Matrix

In addition to that, each client has been identified with his client number and segment. It has also been decided to execute the expected recommendation system considering each segment of the client. As stated previously, There are mainly 3 categories of clients in Millennium BCP's retail banking. The Mass Market segment, The Mass Market + segment and the Prestige respectively with higher bank patrimony and higher purchasing power. The recommendation system might also be able to discover patterns about possession behaviors across the 3 segments. In the sparse matrix, there is also information about the client ID and the client's segment.

	NUM_CLI	COD_MACRO_SEG_CLI
1	50	MS
2	68	UU
3	71	MS
4	72	PP
5	80	PP
6	82	PP
7	85	UU
8	93	MS

Figure 27 : Clients Information (MS: Mass | UU: Mass Market+ | PP: Prestige)

The second step of building the outer recommendation system is to compute the similarity matrix using the formulas explained above. The standardization has been made using the PROC STDIZE procedure in SAS using the Euclidean Norm for each column of the matrix. The similarity matrix computed between different product categories per each clients' segment is shown here below.

MASSE PLUS

	IND_X01_PO_EMEA	IND_X01_IV_EMEA	IND_X03_CC_EMEA	IND_X04_CP_EMEA	IND_X06_CI_EMEA	IND_X08_SR_EMEA	IND_PACOTE	IND_MOBIS	IND_MEDIS	IND_HOMIN
1	0	0.3708710279	0.362356218	0.2458770774	0.1774053361	0.3435994845	0.3475429818	0.1841647485	0.1226417865	0.159235613
2	0.3708710279	0	0.3310058495	0.2389425717	0.169719145	0.3310147426	0.2713321268	0.1723048293	0.1328153884	0.1369203924
3	0.362356218	0.3310058495	0	0.7092272907	0.3809616938	0.5291228921	0.5819605993	0.2156741427	0.2115704728	0.2167630468
4	0.2458770774	0.2389425717	0.7092272907	0	0.3792164765	0.5338565073	0.5190130297	0.187006877	0.2119062583	0.2027462432
5	0.1774053361	0.169719145	0.3809616938	0.3792164765	0	0.668398828	0.4968792728	0.1494459459	0.1466276765	0.3471787835
6	0.3435994845	0.3310147426	0.5291228921	0.5338565073	0.668398828	0	0.6108684634	0.3734817382	0.3133793064	0.4174835704
7	0.3475429818	0.2713321268	0.5819605993	0.5190130297	0.4968792728	0.6108684634	0	0.2113749134	0.2025785157	0.2748636344
8	0.1841647485	0.1723048293	0.2156741427	0.187006877	0.1494459459	0.3734817382	0.2113749134	0	0.107802749	0.1867811554
9	0.1226417865	0.1328153884	0.2115704728	0.2119062583	0.1466276765	0.3133793064	0.2025785157	0.107802749	0	0.100743414
10	0.159235613	0.1369203924	0.2167630468	0.2027462432	0.3471787835	0.4174835704	0.2748636344	0.1867811554	0.100743414	0

PRESTIGE

	IND_X01_PO_EMEA	IND_X01_IV_EMEA	IND_X03_CC_EMEA	IND_X04_CP_EMEA	IND_X06_CI_EMEA	IND_X08_SR_EMEA	IND_PACOTE	IND_MOBIS	IND_MEDIS	IND_HOMIN
1	0	0.554924159	0.4975737442	0.3374900928	0.2126011239	0.4346987286	0.5430154275	0.2374273235	0.2010669293	0.2210873048
2	0.554924159	0	0.5647071789	0.3738590749	0.2106387498	0.4814018974	0.6021873168	0.2797536117	0.2464086422	0.2505491411
3	0.4975737442	0.5647071789	0	0.7275557578	0.3654809094	0.5920747503	0.7235141215	0.2833084103	0.3145084565	0.274023906
4	0.3374900928	0.3738590749	0.7275557578	0	0.3807868787	0.5373409242	0.5591095418	0.2171168877	0.2902530203	0.2154006684
5	0.2126011239	0.2106387498	0.3654809094	0.3807868787	0	0.5475195157	0.3639865627	0.1413850051	0.1630011183	0.2805834873
6	0.4346987286	0.4814018974	0.5920747503	0.5373409242	0.5475195157	0	0.6166248094	0.4448842109	0.4548259412	0.4536118073
7	0.5430154275	0.6021873168	0.7235141215	0.5591095418	0.3639865627	0.6166248094	0	0.2819374572	0.310752626	0.2996894299
8	0.2374273235	0.2797536117	0.2833084103	0.2171168877	0.1413850051	0.4448842109	0.2819374572	0	0.1792648951	0.2519883973
9	0.2010669293	0.2464086422	0.3145084565	0.2902530203	0.1630011183	0.4548259412	0.310752626	0.1792648951	0	0.1650725796
10	0.2210873048	0.2505491411	0.274023906	0.2154006684	0.2805834873	0.4536118073	0.2996894299	0.2519883973	0.1650725796	0

MASSE

	IND_X01_PO_EMEA	IND_X01_IV_EMEA	IND_X03_CC_EMEA	IND_X04_CP_EMEA	IND_X06_CI_EMEA	IND_X08_SR_EMEA	IND_PACOTE	IND_MOBIS	IND_MEDIS	IND_HOMIN
1	0	0.1359372799	0.2083729739	0.1441086754	0.064562098	0.1729943396	0.2374302464	0.0816034738	0.0775905656	0.0713861995
2	0.1359372799	0	0.1396418145	0.1051734601	0.0773158694	0.1323791188	0.1233013266	0.0675038831	0.0759948032	0.0516154098
3	0.2083729739	0.1396418145	0	0.6865555198	0.1890641736	0.4270067276	0.5257784002	0.1632451221	0.1971172338	0.1516708156
4	0.1441086754	0.1051734601	0.6865555198	0	0.1511853868	0.5202245349	0.4696797129	0.15314888	0.2091806317	0.1236615603
5	0.064562098	0.0773158694	0.1890641736	0.1511853868	0	0.4325853772	0.2412329103	0.070458139	0.0412734045	0.2917194596
6	0.1729943396	0.1323791188	0.4270067276	0.5202245349	0.4325853772	0	0.4949995982	0.3082428348	0.362458059	0.337911
7	0.2374302464	0.1233013266	0.5257784002	0.4696797129	0.2412329103	0.4949995982	0	0.1588089138	0.1903954951	0.1792490589
8	0.0816034738	0.0675038831	0.1632451221	0.15314888	0.070458139	0.3082428348	0.1588089138	0	0.0896231447	0.1398051534
9	0.0775905656	0.0759948032	0.1971172338	0.2091806317	0.0412734045	0.362458059	0.1903954951	0.0896231447	0	0.063781875
10	0.0713861995	0.0516154098	0.1516708156	0.1236615603	0.2917194596	0.337911	0.1792490589	0.1398051534	0.063781875	0

Figure 28 : Similarity matrices between the 10 different product categories in Millennium BCP per Client's Segment

Notice that the real similarity matrices computed initially contained 1s in the diagonals. The motivation behind altering 1s to 0s in the diagonal will be seen further. Also, as Mass-Market clients represent almost 1.7 billion clients. Computing the similarity matrix with a sparse matrix of this number of rows was computationally heavy. Therefore, the above similarity matrix for Mass Market clients was computed using a random sample of 300.000 clients of this segment. By looking at the values. We clearly see a difference between the range of values obtained. Clearly, Prestige clients have higher numbers of similarity followed by Mass Plus clients and then Mass-Market clients. That is obvious due to the high purchasing power that Prestige clients have.

Depending on the threshold used, there can be determined the most similar products for each product category. For instance, in the Mass-Market segment, it can be noticed the high similarity that exists between credit cards and personal loans (about 0.68). From Mass-Market clients, there is 68% chances to find one client that possesses products from both Credit Cards Category and Personal loans category. The value is also higher for the remaining segments. This to state that the marketing department can find some strategies using this information to brand more their marketing campaigns. There can be targeted clients who have products from either one of the categories to try to brand them the other one they do not own.

However, to get more precise results for each client, there is a third and last step to conduct for building the final outer-recommendation system. There will be built the so called “Recommendation

Strength table". This table will be indeed used to go to the clients' level and determine what is the likelihood of buying a product from each Product Category.

The principle of the recommendation strength is simple. A recommendation strength of a product not possessed yet by the client is simply the sum of the similarity values between this product and the product he possesses. This measure will only be computed for not possessed products per client. The product with higher recommendation strength is the one that is most likely to be bought next by the client. Here below is an illustrative example.

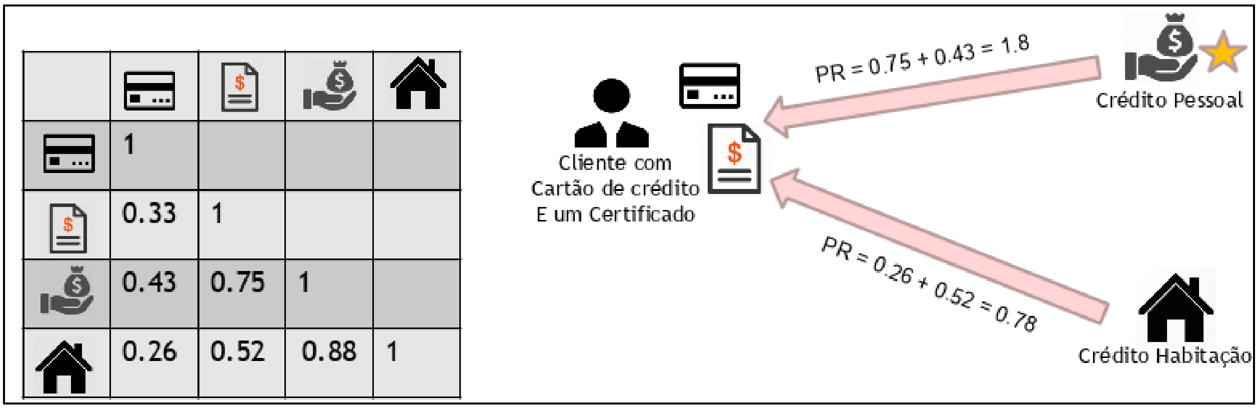

Figure 29 : Illustrative Example of a Recommendation Strength

Above is the similarity matrix between 4 products (Credit Card, A bank Certificate, A Personal loan and a habitation loan). Given a client that possesses a certificate and a credit card. The recommendation strength of the personal loan would be the sum of the similarity between the credit card and the personal loan and the similarity between the certificate and the personal loan (In this case 0.75 + 0.43 = 1.8). Same for the Habitation loan which gives the value 0.26+0.52 = 0.78. Thus, we can conclude that the personal loan is the product that is more likely to be bought by the client given the similarity matrix computed.

This metric should be computed for all the clients whose possession information was used to calculate the similarity matrix. That is why there is an efficient way to calculate the recommendation strength for all the clients for each product category. The resulting table would be called the recommendation strength table which can be expressed by this formula

$$R_{Strength} = A * S$$

With A the initial sparse matrix without standardization and S the similarity matrix with the values on the diagonals to 0. That's is why the similarity matrices in figure 26 have 0s in their diagonals. Therefore, it can be said that the outer-recommendation system that will predict the next product category that the client is more likely to buy, will mainly rely on the recommendation strength tables. It is also important to note that the recommendation strength values can be above one because it is considered as a sum of similarities.

IND_X01_PO_EMEA	IND_X01_IV_EMEA	IND_X03_CC_EMEA	IND_X04_CP_EMEA	IND_X06_CI_EMEA	IND_X08_SR_EMEA	IND_PACOTE	IND_MOBIS	IND_MEDIS	IND_HOMIN
0.2374302464	0.1233013266	0.5257784002	0.4696797129	0.2412329103	0.4949995982	.	0.1588089138	0.1903954951	0.1792490589
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
.	0	0.1396418145	0.1051734601	0.0773158694	0.1323791188	0.1233013266	0.0675038831	0.0759948032	0.0516154098
0.2083729739	0.1396418145	.	0.6865555198	0.1890641736	0.4270067276	0.5257784002	0.1632451221	0.1971172338	0.1516708156
0.4589578791	0.3480157365	.	1.4159606864	0.6629229553	.	1.2111734934	0.5611111016	.	0.5533636905
0.8445097091	0.5679996031	.	.	1.084525987	.	.	.	1.0487745643	0.9322975883
0	0	0	0	0	0	0	0	0	0
.	0.5040538812	.	.	0.7233604382	1.7476043191	.	0.6243102728	0.7502787294	0.5775830442
.	0.1233013266	0.6654202147	0.574853173	0.3185487797	0.627378717	.	0.2263127969	0.2663902983	0.2308644687
0.2374302464	0.1233013266	0.5257784002	0.4696797129	0.2412329103	0.4949995982	.	0.1588089138	0.1903954951	0.1792490589
0	0	0	0	0	0	0	0	0	0
0.5899118957	0.3681166013	.	.	0.5814824708	1.4422308607	.	0.4752029159	0.5966933607	0.4545814349
0.3813673135	0.2720209333	.	1.2067800547	0.6216495509	.	1.0207779984	0.4714879569	0.5595752929	0.4895818156
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0.5899118957	0.3681166013	.	.	0.5814824708	1.4422308607	.	0.4752029159	0.5966933607	0.4545814349

Figure 30 : Recommendation Strength table for Prestige Clients

This snapshot is a representation of how the recommendation strength tables look like. They are tables per client that have information about the possessed product categories as well as the recommendation strengths for the non-possessed product categories. The product category that has the maximum recommendation strength is defined as the next best product category to be acquired by the client. There are various strategies with which the marketing campaign can use this output such as:

- Cross-sell a certain product using leads that will have as the next best product category, the product category of the product in question.
- Create more confident multi-product campaigns for clients that have higher recommendation strengths in 2 or 3 product families
- Create marketing campaigns that involve certain promotions or discounts of products from product families that are highly similar.
- Cross-sell the weakest product family by targeting clients that will have it as the next best product family to acquire

Indeed the “Outer-recommendation system” is a very powerful machine learning solution that helps increase business using the possession profile of the clients. The next step then is to couple this recommendation system with another one that will be called “The Inner-Recommendation system”

Building the CRM Inner-Recommendation System

The outer-recommendation system will help orient better to match the needs of the clients by providing the product category that will most likely be acquired by the client. Departing from this result, there will be applied another recommendation system but between the most owned products in this category. Their number would be fixed at 8 products. In other words, after determining the best product category, the same process of the outer-recommendation system will be applied between the first 8 products in this family to determine which one is the next best product to acquire. However, there are some product families that do not contain a lot of products like Personal and Habitation loans (Due to the definition of the products hierarchy of Millennium BCP). Therefore, it wouldn't make sense to build a recommendation system with only 2 products. In this case, it has been opted to include them with the credit card product category since both products involve credit. This recommendation system, as well as part of the outer-recommendation system, has been created by using a SAS Script that mainly has 3 big parts (3 Macro Functions) which are:

- The Recommend Macro Function: That prepares the Sparse Matrix from the 8 most owned products of the product family
- The Similarity Macro Function: That computes the similarity matrix between the 8 products
- The Recommendation Strength Macro function: That computes the Recommendation strength table between all the clients used to build the sparse matrix.
- The Joining Clients Macro Function that simply joins the client IDs to the resulting recommendation strength tables.

Since the outer-recommendation model will be based on the products that exist in each category, it will be interesting to present an overall picture of the product hierarchy in the retail department of the bank.

The product line of Millennium BCP contains a variety of products following a fixed determined hierarchy which is the following

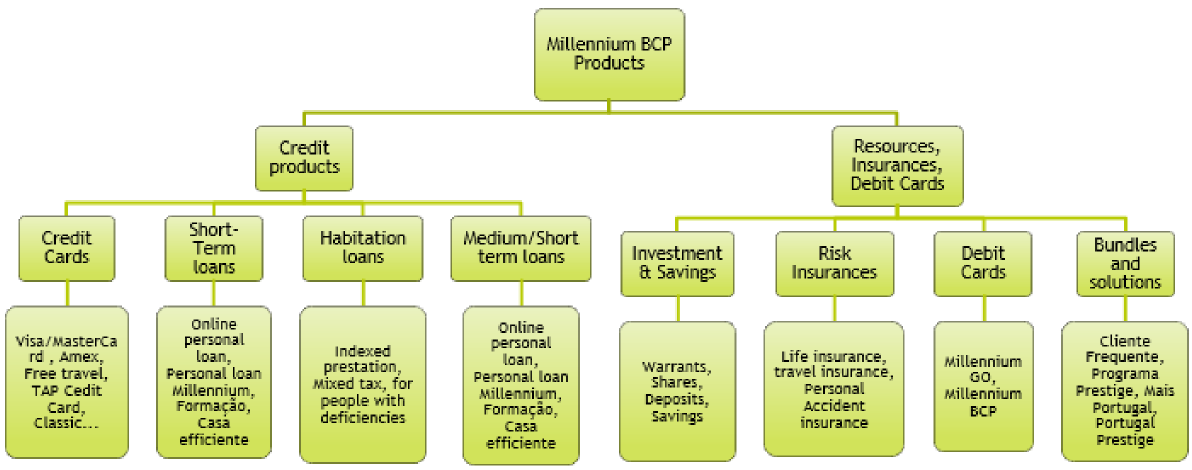

Figure 31 : Snapshot of the retail products hierarchy in Millennium BCP

In general, Millennium BCP retail products are distinguished through 2 big categories which principal discriminator is Credit. In other words, there is a clear dispatch between products that involve credit than others that do not in the first hierarchy. The second hierarchy reveals then the usual products commercialized by the bank which is approximately the equivalent of the concept of the product category. The fourth level represented in the above figure has got also a 5th hierarchy that is not represented.

Having mentioned that and given the fact that the Inner-Recommendation System takes into its argument the product category generated by the Outer-Recommendation system, it has been conducted the 4 steps in the SAS Script stated above. Notice that the loans product category and the habitation loan product category contained few products which were the same, with only some functionalities added in the contract of these products. Therefore, since they are credit products, they will be inputted with the credit cards with respect to the second level of the Hierarchy. Moreover, Debit Cards won't be considered in this analysis since there has been discovered that 99% of the clients in the bank have got debit cards. The Integrated Solutions (Bundles) product category as well won't be considered in the Inner-Recommendation system process since they are products formed by a panoply of already existing products in the bank. And moreover, a segmented client in the bank that has an account in Millennium BCP has the right to have one kind of integrated solution in his account. The other related accounts that belong to his familiars for instance will also follow the same integrated solution from the mother account. Thus, it is irrelevant to conduct a recommendation system between different bundles. As a conclusion, the inner recommendation system will consider the following families.

- Credit Cards/Habitation Loans
- Credit Cards/Personal Loans
- Risk Insurances
- Investment & Savings

Here are snapshots of the recommendation strengths between the most 8 owned products in each product for all clients by each of the four product families stated above

VIEWTABLE: Ziad.Poder_recom_cc_ci_adj

	Num_Cli	P1027	P0121	P1022	P2157	P3285	P1024	P1982	P1026	P0034
1	50	0	0	0	0	0	0	0	0	0
2	72	.	0.0077997462	0.0060930806	0.0255228381	0.0192917458	0.0037747917	0.0005728675	0.0082504948	0.045998517
3	80	.	0.0077997462	0.0060930806	0.0255228381	0.0192917458	0.0037747917	0.0005728675	0.0082504948	0.045998517
4	82	0	0	0	0	0	0	0	0	0
5	85	0.0537982632	.	0.0685440184	0.0995718216	0.0737847162	0.1572235952	0.0295516858	0.042440606	.
6	93	0.0077997462	.	0.0072652579	0.0303421357	0.0079359554	0.0071895523	0.0041739966	0.0094351095	0.3316949003
7	117	0.0060930806	0.0072652579	.	0.0196164713	0.0049552778	0.003102039	0.0003558037	0.0032122605	0.0612787605
8	118	.	0.0077997462	0.0060930806	0.0255228381	0.0192917458	0.0037747917	0.0005728675	0.0082504948	0.045998517
9	133	0.0448145838	0.0382780911	0.0245717491	.	.	0.0281074114	0.0074754347	0.0203934889	0.1350784468
10	199	0.0060930806	0.0072652579	.	0.0196164713	0.0049552778	0.003102039	0.0003558037	0.0032122605	0.0612787605
11	325	0.0253848264	0.0152012133	.	0.0463116924	.	0.0089194492	0.0015356601	0.0107573975	0.1271275214
12	380	0	0	0	0	0	0	0	0	0
13	388	0.0255228381	0.0303421357	0.0196164713	.	0.026695221	0.0222900012	0.0062955783	0.0128483519	0.0692296859
14	415	.	0.0077997462	0.0060930806	0.0255228381	0.0192917458	0.0037747917	0.0005728675	0.0082504948	0.045998517
15	419	0.0192917458	0.0079359554	0.0049552778	0.026695221	.	0.0058174102	0.0011798564	0.007545137	0.0658487608
16	427	0.0037747917	0.0071895523	0.003102039	0.0222900012	0.0058174102	.	0.0012735833	0.0053525831	0.1500340429

Figure 32 : Recommendation strength for Credit Cards & Habitation loans

VIEWTABLE: Ziad.Poder_recom_cc_cp_adj

	Num_Cli	P1027	P0121	P1022	P2157	P3285	P1024	P1982	P1026	P0883	P0138
1	50	0	0	0	0	0	0	0	0	0	0
2	72	.	0.0077997462	0.0060930806	0.0255228381	0.0192917458	0.0037747917	0.0005728675	0.0082504948	0.0051180039	0.0116882449
3	80	.	0.0077997462	0.0060930806	0.0255228381	0.0192917458	0.0037747917	0.0005728675	0.0082504948	0.0051180039	0.0116882449
4	82	0	0	0	0	0	0	0	0	0	0
5	85	0.0077997462	.	0.0072652579	0.0303421357	0.0079359554	0.0071895523	0.0041739966	0.0094351095	0.023285487	0.0038534962
6	93	0.0077997462	.	0.0072652579	0.0303421357	0.0079359554	0.0071895523	0.0041739966	0.0094351095	0.023285487	0.0038534962
7	117	0.0060930806	0.0072652579	.	0.0196164713	0.0049552778	0.003102039	0.0003558037	0.0032122605	0.0070638972	0.001613218
8	118	.	0.0310852332	0.0131569778	0.0355434558	0.0219269579	0.0133874136	0.0129686232	0.0149317566	0.0051180039	0.0154630394
9	133	0.0448145838	0.0382780911	0.0245717491	.	0.026695221	0.0281074114	0.0074754347	0.0203934889	0.0126558299	0.0380987111
10	199	0.0060930806	0.0072652579	.	0.0196164713	0.0049552778	0.003102039	0.0003558037	0.0032122605	0.0070638972	0.001613218
11	325	0.0253848264	0.0152012133	.	0.0463116924	0.0049552778	0.0089194492	0.0015356601	0.0107573975	0.0096991093	0.0256858885
12	380	0	0	0	0	0	0	0	0	0	0
13	388	0.0255228381	0.0303421357	0.0196164713	.	0.026695221	0.0222900012	0.0062955783	0.0128483519	0.0100206177	0.0140260407
14	415	.	0.0077997462	0.0060930806	0.0255228381	0.0192917458	0.0037747917	0.0005728675	0.0082504948	0.0051180039	0.0116882449
15	419	0.0192917458	0.0079359554	0.0049552778	0.026695221	.	0.0058174102	0.0011798564	0.007545137	0.0026352122	0.0240726705

Figure 33 : Recommendation strength for Credit Cards & Personal loans

VIEWTABLE: Ziad.Poder_recom_sr_adj

	Num_Cli	P0784	P1089	P2078	P0829	P1128	P0238	P2863	P1140
1	68	.	.	.	0.2894883702	0.365741987	0.3338991426	0.3679881248	0.8356017521
2	85	0.2302188526	0.3871377701	.	.	.	0.2984211546	0.5234081553	0.2720792629
3	93	0.0724357962	0.104282735	.	0.0902111401	0.1243811257	0.1089598711	0.1708231905	0.0714429677
4	100	0	0	0	0	0	0	0	0
5	118	0.3074986531	.	0.2132426061	0.2178613826	0.254454936	.	0.2072537059	0.6982331134
6	176	0.0899781716	0.1349610999	0.1089598711	0.0860248789	0.1034364045	.	0.0925833302	0.0996643507
7	198	0	0	0	0	0	0	0	0
8	318	0	0	0	0	0	0	0	0
9	322	0	0	0	0	0	0	0	0
10	324	0.1577830564	0.2828550351	0.2145922657	.	.	0.1894612835	0.3525849648	0.2006362953
11	325	0.0899781716	0.1349610999	0.1089598711	0.0860248789	0.1034364045	.	0.0925833302	0.0996643507
12	329	0.0899781716	0.1349610999	0.1089598711	0.0860248789	0.1034364045	.	0.0925833302	0.0996643507
13	334	0	0	0	0	0	0	0	0
14	404	0.1624139677	0.2392438349	.	0.176236019	0.2278175302	.	0.2634065207	0.1711073184
15	422	0.0899781716	0.1349610999	0.1089598711	0.0860248789	0.1034364045	.	0.0925833302	0.0996643507
16	427	0	0	0	0	0	0	0	0
17	437	0.0899781716	0.1349610999	0.1089598711	0.0860248789	0.1034364045	.	0.0925833302	0.0996643507
18	447	0.2175204815	.	0.104282735	0.1318365037	0.1510185315	0.1349610999	0.1146703757	0.5985687627

Figure 34 : Recommendation strength for Insurances

VIEWTABLE: Ziad.Poder_recom_ip_adj

	Num_Cli	P1055	P0957	P0925	P0191	P1343	P2005	P2479	P2477
1	72	0.3664605371	.	0.0896251877	0.0817347161	0.18115254	0.0631276852	0.0493735852	0.0430915807
2	80	.	.	0.14742336	0.1443514674	0.3062742062	0.0999281029	0.0787646133	0.0700579304
3	82	0	0	0	0	0	0	0	0
4	85	0.0577981722	0.0896251877	.	0.0185972698	0.0475066236	0.0646293455	0.0609002137	0.0571038457
5	98	.	0.3664605371	0.0577981722	0.0626167513	0.1251216662	0.0368004177	0.0293910281	0.0269663497
6	130	.	.	0.14742336	0.1443514674	0.3062742062	0.0999281029	0.0787646133	0.0700579304
7	155	.	0.3664605371	0.0577981722	0.0626167513	0.1251216662	0.0368004177	0.0293910281	0.0269663497
8	199	0	0	0	0	0	0	0	0
9	260	0.3664605371	.	0.0896251877	0.0817347161	0.18115254	0.0631276852	0.0493735852	0.0430915807
10	325	0.4915822033	.	0.1371318113	0.1297806121	.	0.08697784	0.0761612408	0.0594404192
11	334	.	.	0.14742336	0.1443514674	0.3062742062	0.0999281029	0.0787646133	0.0700579304
12	360	0.0577981722	0.0896251877	.	0.0185972698	0.0475066236	0.0646293455	0.0609002137	0.0571038457
13	380	0	0	0	0	0	0	0	0
14	388	0.0368004177	0.0631276852	0.0646293455	0.0188185585	0.0238501548	.	0.0415934209	0.0297090074
15	406	.	.	0.14742336	0.1443514674	0.3062742062	0.0999281029	0.0787646133	0.0700579304
16	422	0.3664605371	.	0.0896251877	0.0817347161	0.18115254	0.0631276852	0.0493735852	0.0430915807
17	424	.	0.3664605371	0.0577981722	0.0626167513	0.1251216662	0.0368004177	0.0293910281	0.0269663497

Figure 35 : Recommendation strength for Savings & Investments

The PXXXX is the code of the products belonging to each product category. It also should be noted that the clients possessing already one or various products of the 8 resulting products are marked with a missing value. Their recommendation strength in the matrix has been ignored since it is not needed for the analysis. Following the same approach as the outer-recommendation system, the maximum recommendation strength given by one of the products would indeed be the more likely to be purchased.

It is important to explore some general statistics about the resulting product categories from the outer-recommendation system as well as the best products that scored high recommendation strength values

To sum up, there has been implemented 2 types of recommendation systems. The product category recommendation system denoted as the outer-recommendation system, and the recommendation system that deals with products proper to the best product category denoted as the inner recommendation system. Both parts are complementary to determine the next best offer of our clients based on other clients profiling (collaborative filtering). This component would constitute a very important pillar of the expected recommendation engine in the optic of the Next Best Offer 2.0 Roadmap. It provides very analytical results for the next best offer 2.0. which is very important for the analysis of the Next Best Offer 2.0. The objective is to reach to an adequate offer by combining both analytical machine learning techniques and business data that are available about the bank's clients.

5.3.4. The Websites Visits of the official website of Millennium BCP

While the recommendation system architecture that was built makes predictions about the most probable products to be purchased by the clients using analytical techniques of similarity, this part is going to focus concretely on how clients behave while visiting the website of Millennium BCP. The bank's website is the perfect place to understand the needs of the clients which in most of cases can be totally far from the output of the analytical methods. All Millennium BCP clients have credentials to have not only access to their bank account and conduct various series of operations such as transactions or payments, but also express their needs by checking the panoply of products available in the Millennium BCP website. This need should be caught with great attention and should be translated into an offer that meets best the expectations of the clients at the right time. This approach is going to be valued at a greater point in the Recommendation engine of the Next Best Offer 2.0 since it represents approximately a real-time information about what the client is looking for. Thanks to the datamart team, it has made available a daily-updated SAS table about the visits of all the clients logged in using the information from their web sessions. Therefore, if a client entered to view the Savings & Investments page in the website, the next day the CRM Team will know about this information. Now, the next step would be to prepare an offer concretely by sending an outbound offer through his bank branch to better understand the purpose of his visit.

Figure 36 : Home page of the Millennium BCP website

As shown in the upper figure, the website of Millennium BCP is organized in a way that shows in the header the different product categories that cover the products offered by the bank. The client goes directly to the area which interests him the most. The table to which each visit is registered using the sessions concept in web development is shown here below

VIEWTABLE: Siteinfo.Clicks_novo_001

	num_cli	Data	Hora	SK_Page	SK_Channel	Canal
1	904782	20180922	1727	4370	7	Web - Particulares
2	22552046	20180922	1727	118	7	Web - Particulares
3	22552046	20180922	1727	4370	7	Web - Particulares
4	22552046	20180922	1727	451	7	Web - Particulares
5	22552046	20180922	1727	4370	7	Web - Particulares
6	104900304	20180922	1727	118	7	Web - Particulares
7	104900304	20180922	1727	118	7	Web - Particulares
8	101967688	20180922	1727	4392	7	Web - Particulares
9	106368467	20180922	1727	266924	8	Web - Empresas
10	3090762	20180922	1727	4370	7	Web - Particulares
11	3090762	20180922	1727	4370	7	Web - Particulares
12	3090762	20180922	1727	4370	7	Web - Particulares
13	104744961	20180922	1727	14444	8	Web - Empresas
14	104744961	20180922	1727	14965	8	Web - Empresas
15	1336423	20180922	1727	4360	7	Web - Particulares
16	1336423	20180922	1727	4360	7	Web - Particulares
17	1336423	20180922	1727	4360	7	Web - Particulares
18	2438364	20180922	1727	4377	7	Web - Particulares
19	2438364	20180922	1727	4377	7	Web - Particulares
20	2438364	20180922	1727	4360	7	Web - Particulares
21	2438364	20180922	1727	4360	7	Web - Particulares
22	4506723	20180922	1727	12373	7	Web - Particulares
23	4506723	20180922	1727	461	7	Web - Particulares
24	4506723	20180922	1727	12361	7	Web - Particulares
25	4506723	20180922	1727	118	7	Web - Particulares
26	102635717	20180922	1727	4370	7	Web - Particulares

Figure 37 : Visits and clicks in the Millennium BCP Website

The table records each visit of any page (Login and welcome pages included) after authentication of the client by date and time as well as the type of the client (Individual or Company). As the purpose of the next best offer is only restricted to individual clients, companies will be excluded from this analysis. It should be mentioned that this table has been used very frequently by the marketing campaigns team of the CRM Team for creating follow-up marketing campaigns for clients who went to look for information about a product through which they can assist clients by phone about their internet visits. However, these follow-up marketing campaigns had poor success rates and sales rates and weren't created for all products. Therefore, this information will be used in a very broad manner. in other words, the websites visits will be coupled with the previous pillars that have been discussed and explained previously for an assertive approach of the next best offer. Knowing that the table showed in figure 36 increases daily by 20 million rows (visits), it would be inefficient to work with it knowing that across all the clients that visit the website in one given day, the average pages visited by client are around 12 without including the login page and the automatic refreshes. Consequently, the Data mart Team made available a summarized table by client that visited each page related to a product category with the information of the number of clicks in that given page.

	num_cli	Produto	Familia	UltimaData	QtdClicks
1	2112	Seguros Habitação	SEGH	20180928	1
2	2297	Mobile	MOBILE	20180924	1
3	3434	Cartão Classic	CC	20181010	1
4	3434	Cartão Débito MasterCard	CD	20181010	1
5	4123	Investimento	INV	20181005	1
6	4625	Seguros Viagem	SEGV	20180928	2
7	4854	Poupança	POUP	20181003	6
8	5859	Investimento	INV	20181021	3
9	9345	Cr Habit Indexado (Part)	CH	20180924	1
10	9345	Cr Pessoal Formação	CP	20181001	3
11	9345	Cr Pessoal Millennium	CP	20181019	4
12	9345	Cr Pessoal Online	CP	20181017	7
13	9345	Mobile	MOBILE	20180927	1
14	16607	Cr Habit Indexado (Part)	CH	20181003	1
15	17108	Cr Pessoal Online	CP	20181008	2
16	18072	Reforma	REF	20181001	2
17	19058	Investimento	INV	20181011	1
18	19878	Seguros Habitação	SEGH	20180926	1
19	21159	Investimento	INV	20181014	1
20	21517	Investimento	INV	20181019	1
21	21735	Seguros Auto	SEGA	20180925	1
22	21838	Poupança	POUP	20181008	2
23	22282	Mobile	MOBILE	20181001	1
24	22465	Poupança	POUP	20181019	3
25	23816	Seguros Habitação	SEGH	20180926	1
26	24125	Poupança	POUP	20181008	2
27	25137	Investimento	INV	20180927	4
28	25569	Cr Habit Out16 (Part)	CH	20181007	1
29	28596	Poupança	POUP	20181004	1
30	29636	Mobile	MOBILE	20180928	1

Figure 38 : Summarized website visits table by client

This form of the table is much more in line with the final table that will contain all the information helpful for determining the Next Best Offer. This information aspect of the NBO 2.0 model will have a higher impact in comparison to the other aspects of the analysis due to the possible concrete need that can be caught and give a useful value to the model.

5.3.5. The Propensity Models

The last pillar of the NBO 2.0 engine is indeed the fruit of the main activity of the CRM Team which is developing predictive models of product acquisition. These predictive models are closely linked to marketing campaigns as their output is the ranking of clients by their model scores. The clients with higher scores (also denoted hot leads) are top targets of marketing campaigns that aim the cross-selling of the product to which the predictive model has been developed. It has been built various predictive models that do not limit themselves to the product category but rather go deeper to model events at the level of the final product. In this section, it will be explained the standard process mounted by the CRM Team from the definition of the target variable to the scoring of the clients.

5.3.5.1. Variables, the gold of the CRM Team

The SAS Library of the DataMart mentioned previously in the chapter of the CRM Analytics Solution for sales tracking is a massive library that contains SAS tables that have information about numerous banking information of each client for instance (Offers received, Credit Cards plafonds, Loans, Risk, Rentability, TPA Purchases, Money withdrawals etc.…). This information is represented by a conjunction of various tables and variables. However, as the CRM team's main tasks is to give sense to the behavior of these variables through time to look at the historical data of the clients for making prediction in the future, there has been created various scripts that are executed every month to constitute a series of variables categorized by their banking logic. The output of these scripts are tables saved in a SAS library dedicated for the process of model building. The variables are classified by various and different business logics. The most used ones are the following:

- Transationality: Gathers all the variables that relate to transactions of the clients
- Rentability: Contains the variables that express the turnover that the bank gets from the services offered to the client
- Possession: Variables that are related to product possession
- Realtionship: Variables that sums the communications or offers proposed to the client
- Companies: Are more economic variables that concern companies clients
- Demographics : Demographic variables of the clients

Here below is a figure that illustrated some examples of these variables :

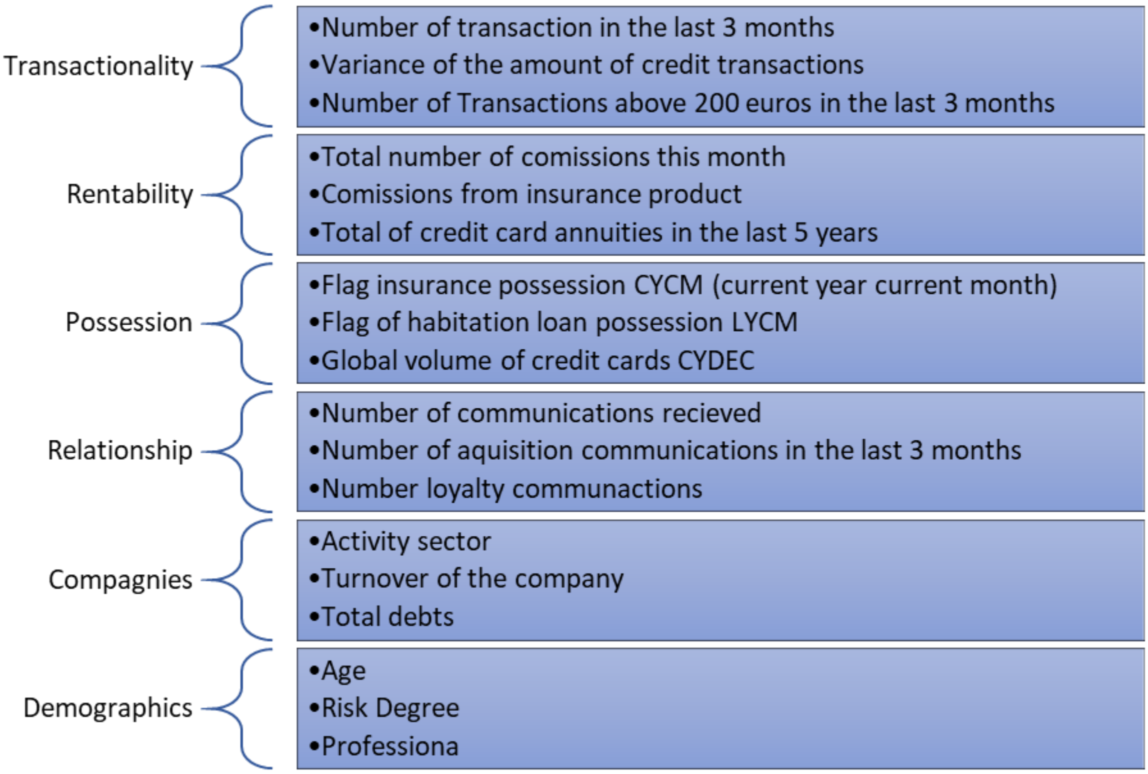

Figure 39 : Example of Variables per each Variable category

Besides this structure upon which variables are organized, this information is available by month from January 2016. In other words, 2 years of historical data is available for all clients of the bank, companies and individuals across all the dimensions showed above. This makes a very comfortable playground for building predictive models. Also, these variables are organized in monthly tables that will serve for the process of building a predictive model

5.3.5.2. Target Definition

This is the first very crucial step with which each project of building a predictive model should start with. A very important step as it is the main step where the goal of the predictive model is determined. In analogy with the CRISP Method for Data Mining (Shearer, 2000), this step enters in the description of the business goals which is the first phase of the CRISP Methodology. In other words, specifically in this scenario, this step aims to define the event as well as the non-event that is tried to be modeled. Since most of the times it is required to look at the historical data to understand this behavior of the event, it is considered monthly checkpoints (windows in timeline) to view how was the event before this checkpoint and after it. For instance, let’s assume that it is wished to build a predictive model for the churn of the Millennium banking solution “Programa Prestige”. The considered target was built as the following:

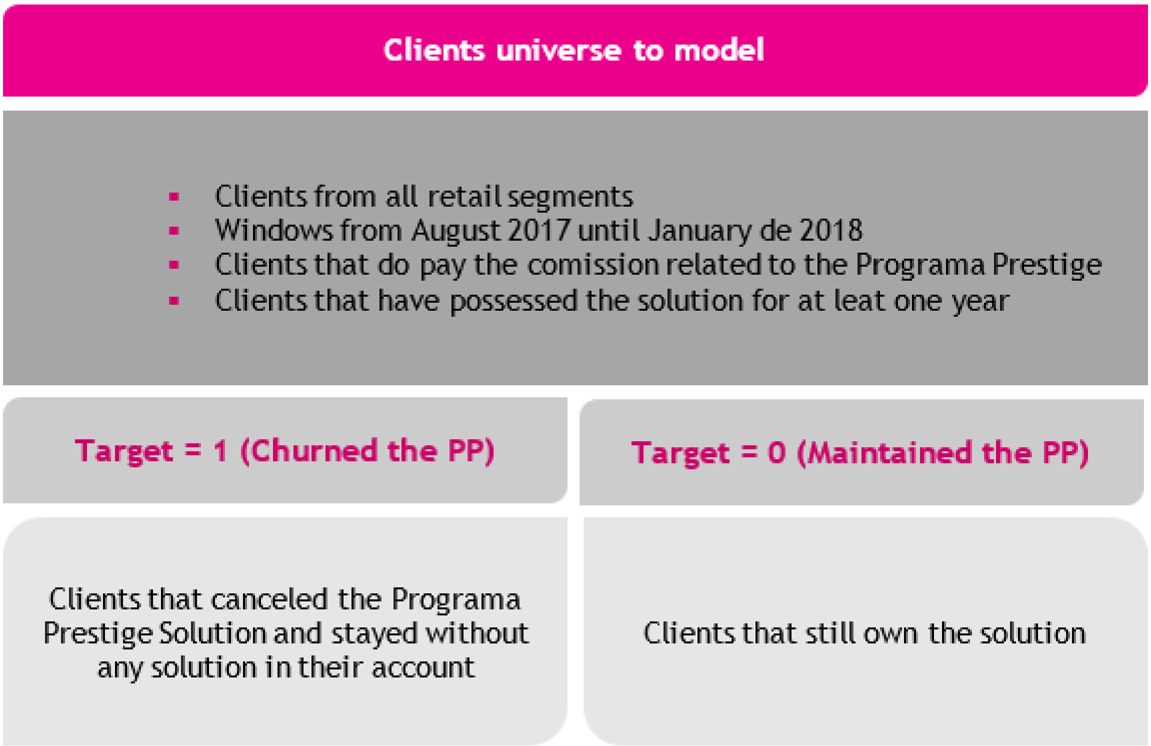

Figure 40 : Example of a Target Definition

It has been chosen to eliminate clients that do not pay the commission from using the solution as it is considered a very specific case that might biase the behavior of the model. Moreover, it also has been considered a minimum period of 1 year of possessing the solution to eliminate the cases of clients that want to beneficiate from the first 3 months that are commission free. Those business decisions are going to contribute for a better understanding of the event of churning from the Programa Prestige. Proceeding as explained above, this is how 1s and 0s are collected from the windows time points.

Figure 41 : Collecting 0s and 1s using the windows timeline

These are considered the reference months of analysis that will serve as a modelling playground for the churn event. The orange brace represents the time a client has got the solution until the reference month that should be for both 0s and 1s at least one year. The purple brace represents the period to be analyzed to see if there was a churn or not. The 0s and 1s generated by each month are gathered in a table called target_ap_MonthOfReference (ap stands for attrition prestige). Afterwards, there are run some automatic SAS Scripts that arrange the target tables all in one table given it as a key the client number concatenated with the reference month.

It is also important to notice that the purpose of the model is going to be the creation of a marketing campaign that targets the clients with high scores, specially the models that are focused on the purchase of one specific product. There is another way to build the target by eliminating clients that already have been contacted or have already bought the product in question and the clients who have been contacted recently (to a maximum of 90 days from the latest contact). That is why there exist other parallel processes that are run monthly such as the list of all contacted clients per month as well as sales (purchases from the client perspective) also per month. For instance, if it is needed to build a predictive model to predict the propensity of getting a personal loan, the target tables will be built by joining the table of contacted clients with the sales. The clients who have been contacted more than 90 days ago and didn't get a personal loan are going to be targeted as 0. The ones who have been contacted and got the personal loan are going to be targeted as 1. The ones who have been contacted in less than 90 days are not going to be subject of the analysis to avoid calling them repeatedly.

5.3.5.3. Pre-Analysis Phase

After building the target tables according to each reference month, these tables are afterwards combined and sorted by the earliest month reference and ascending client code (12 digits code) making what is called the base table. There might be repeated clients depending on each reference month. The pre-analysis phase is very important. It contains 2 essential steps. The first one is the joining of the business variables that have been explained in 5.3.5.1. This step will join automatically the target clients of a given reference month with the corresponding variables of all types of that reference month. Since the purchases or the churn events occur after the reference month, the goal of the predictive model is to try to find patterns in the variables for clients before the event happens 2 months or 3 months after. Thus, the result of this join would be the following.

Reference Month	Client Number	Demographics	Rentability	Possession	Relationship	Target(0/1)

Figure 42 : Analytical Base Table

The second step is to make changes on some specific class variables that have a lot of categories. For example, there is a subsegment of the client that has various categories in each segment (Mass-Market, Mass-Market+, Prestige). Categorical variables with more than 6 levels would be poor variables to target the models. Therefore, it has been discussed with the responsible business areas about finding rules of correctly joining the levels not only for analytical purpose but also in a way that makes sense from the business point of view. After several meetings, the predictive models team has come up with a script with all rules implemented for the adjustment of these categorical variables. In brief, the pre-analysis phase is only about joining the bank variables with the built target table.

5.3.5.4. Columns/Rows Analysis Phase

This phase takes as input the Analytical Base Table and analyses all the variables. In parallel to this analysis, there is a table that is created which has all the statistics and information of all the variables of the analytical base table called the columns trace table. In terms of columns, there are computed informative variables about them such as the percentage of missing values in each variable, the percentage of variables which are equal to the value of the mode, and some statistics such as the mean, standard deviation, skewness and kurtosis. For rows, each variable, the corresponding rows which are considered as outliers are labeled in a table that is dedicated for the rows analysis. However, these outliers are not eliminated from the analytical base table. As most of the cases the number of 1s that are collected in each reference month represents only 13% in comparison of the number of 0s, and also knowing that it is hard to build a predictive model in SAS Enterprise Miner with a table of more than 2 million rows, There is an important step that is run after the outliers part which is building randomly a sample of the analytical base table that has half observations with the target value equal to 1 and the other half equal to 0. By this way, it is given consistency to the event in comparison with the non-event. There will also be captured the maximum number of insights that would distinguish clients who purchase (or churn) than the other that do not. The focus in the upcoming steps will be specifically in this sample

5.3.5.5. Pre-Modeling Phase

This phase is considered the last phase before building the predictive model in SAS Enterprise Miner, it operates in the random sample created at the end of the Rows/Columns analysis phase. It also adds new variables of analytical statistics that explain how well the variables of the sample ABT explain the target variable. These variables are the following:

- ***The worth*** (or the importance) of a variable: is a statistic computed for each interval variable using a decision tree. It evaluates the relevance of the variable in the splitting rules. The higher the value of the worth is, the higher is the importance of the variable meaning that the split of the decision tree would garantee a better classification of ones and zeros

- ***The chi-squared*** statistic is a test of significance that is computed for categorical variable, it evaluates if there is any relationship of dependence between the categorical variable and the target variable. The null hypothesis is that there is no relationship between the two variables. The statistic is simply calculated using a cross-tabulation count between the levels of both variables.

- ***The R-squared***: Calculated for all interval variables, it expresses the amount of variability explained of the target variable

- ***The indicator of correlation***: It is computed on the top of the sample of the ABT table to eliminate inter-correlated variables as well as variables that are correlated with the target variable. This indicator value is set to one if the variable is correlated and set to 0 in the opposite case.

- ***The information value (IV):*** is a very powerful measure for variable selection as it calculates the number of ones over the number of zeros for each level of the categorical variable for which the information value is evaluated. This is the formula how the information value is calculated.

  $$Information\,Value = \sum_{levels}(CountOf1s - CountOf0s) * \ln(CountOf1s/CountOf0s)$$

After computing these statistics. The variables from the business categories seen before and that are present in the ABT sample table are ranked according to the statistics mentioned earlier. Depending on each predictive modeler, the best 300 or 400 ranked variables are going to be used for the process of modelling inside SAS Enterprise Miner.

5.3.5.6. Modeling Phase in SAS Enterprise Miner:

It is in this phase where the modelling part from a data mining point of view begins. Technically, all the previous parts that have been seen are preliminary and mandatory to this phase. The input dataset of this phase is the randomly sampled ABT table that contains the top 300-400 ranked variables and half observations with target equals to 1 (Event) and the other half at target equals to 0 (Non-Event). The modelling phase is conducted in SAS Enterprise Miner that follows the SEMMA Methodology that was developed by the SAS Institute (Mark Brown) .The SEMMA stands for:

Sample: Is designed to create a sample data that is considered to be representative enough of the real population. Following the process from the very beginning, this step is considered done and already applied since the input of the SAS Miner process is going to be the ABT Sampled table which is a well-prepared table for the modelling phase.

Explore: This phase consists of applying statistics and visualization techniques (Histograms, Pie charts…) on the variables to understand better the data. One common practice is also Viewing the distribution of each variable to understand its behavior alongside the target variable.

Modify: Consists in applying the necessary transformations to the variables to increase their prediction power. One common method in this step is to transform continuous variables to categorical variables which levels are simply 2 to 3 bins of the continuous variables. This method is called optimal binning and is very useful specially when the target is binary.

Model: Applying the predictive models(Trees, Regressions, Neural Networks, Ensembles) to the input data set considering a training and validation data set for model evaluation

Assess: consists of assessing the models by applying various statistics such as the model comparison node or the segment profile if the goal was to perform clustering tasks.

For predictive modelling, the predictive modelling team in the CRM team considers this benchmark for model assessment. In this figure, it is shown the final model of predicting the churn from the solution “Programa Prestige” which was an example discussed in the part of the target definition.

Tipo de Medida	Indicador	Benchmark	Ensemble
			Regression (x3) +DMNeural (x1)
Estatística	Kolmogorov-Smirnov Statistic (+)	0.35	0.51
	Average Squared Error (-)	0.2	0.16
	Gini Coefficient (+)	0.5	0.67
Classificação	Roc Index (+)	0.75	0.84
	Misclassification Rate (-)	0.30	0.25
	Cumulative Percent Captured Response (+)*	45	46.24
Data Mining	Lift (+)*	1.5	1.58
	Cumulative Lift (+)*	1.8	1.85

Figure 43 : Benchmark Statistics Versus Model Statistics

One of the important statistics to consider, is the lift which is in this case 1.85. This value means that it is 85% worth using the model that has been built than using the baseline model that uses univariate random probabilities between 0 and 1.

5.3.5.7. Scoring phase

After validating the necessary analytical metrics and before using the model in production mode which is to score clients monthly, it is needed to apply the model on a brand-new data set of which the values of the target are known. This step is called the back-testing. To make it simple, it has been built score ranges for each observation depending on the percentiles of the scores distribution. There are 4 levels that are considered of the score values which are the following:

- Baixo (Low in Portuguese): which corresponds to scores from 0 to the 25th percentile
- Médio Baixo (Average Low in Portuguese): which corresponds to scores from 25th percentile to the 50th percentile
- Médio Alto (Average High in Portuguese): which corresponds to scores from 50th percentile to the 75th percentile
- Alto (High in Portuguese): which corresponds to scores from 75th percentile to the 90th percentile
- Superior (Superior in Portuguese): which corresponds to scores from 90th percentile to the 100th percentile

The following figure shows the distribution of the churn target values of the reference month February 2018 that are equal to 1(Purple) across each class of scores in the graph

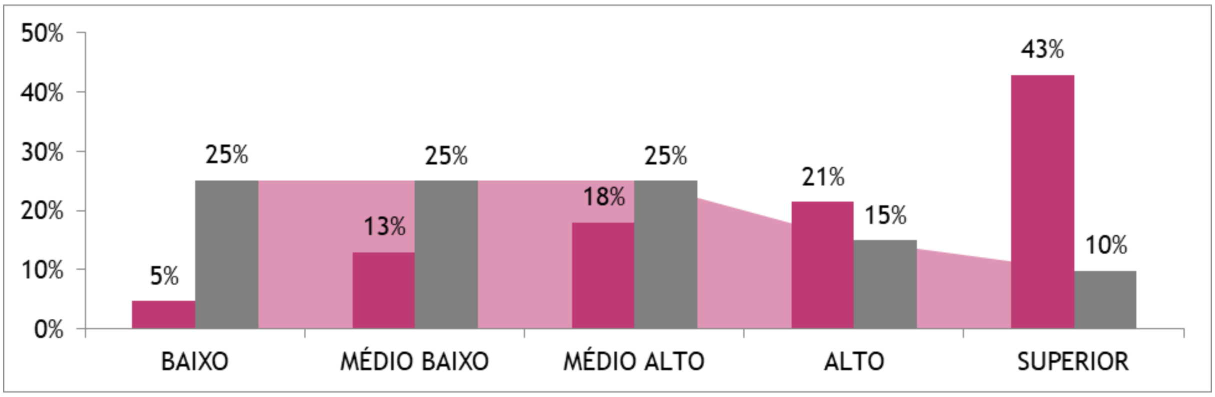

Figure 44 : Back-testing distribution

A predictive model is considered ready for entering in production if the percentage of 1 captured in the "Superior" Class is higher than the percentage captured in the other 4 classes. Usually the clients who present higher scores in the "Alto" and "Superior" are the ones who are considered possible for an eventual contact plan in the marketing campaigns. In this example of this model, in only the higher 2 score classes, it has been possible to catch 64 % of the clients that might churn from the solution of "Programa Prestige", which is considered a good percentage.

Later, the model is directly used in production by providing the "monthly" scores collected from the "Alto" and "Superior" classes to the marketing campaigns team that prepare a marketing campaign contact plan to these clients.

5.3.5.8. The propensity models for the Next Best Offer 2.0

After explaining all the details and steps of the process of building each predictive model, it is important to notice that they contribute with a greater value to the decision of the Next Best Offer 2.0. As mentioned previously, the next best offer 1.0 developed by Millennium BCP is nothing but the gathering of all the built models and detect the one that gives the higher score. The main output of the scoring phase seen previously is a scoring table that contains the client ID, the reference month, and all the scores and the 5 score classes seen previously from all predictive models. The table has got all clients, individuals as well as companies knowing that there have been built some predictive models for companies. A small part of the table is shown below

PROPENSÃO À AQUISIÇÃO DE CONFIRMING	CLASSE DE PROPENSÃO À AQUISIÇÃO DE CONFIRMING	PROPENSÃO À AQUISIÇÃO DE FUNDOS PPR	CLASSE DE PROPENSÃO À AQUISIÇÃO DE FUNDOS PPR	PROPENSÃO AO ABANDONO DO BANCO	CLASSE DE PROPENSÃO AO ABANDONO DO BANCO	PROPENSÃO À AQUISIÇÃO DE CRÉDITO HABITAÇÃO	CLASSE DE PROPENSÃO À AQUISIÇÃO DE CRÉDITO HABITAÇÃO	PROPENSÃO À AQUISIÇÃO DE SOLUÇÃO INTEGRADA - PROGRAMA PRESTIGE	CLASSE DE PROPENSÃO À AQUISIÇÃO DE SOLUÇÃO INTEGRADA - PROGRAMA PRESTIGE
0.00000	00: NULO	0.66334	05: SUPERIOR	0.04145	01: BAIXO	0.30387	03: MEDIO ALTO	0.88331	05: SUPERIOR
0.00000	00: NULO	0.83696	05: SUPERIOR	0.02447	01: BAIXO	0.55436	04: ALTO	0.00000	00: NULO
0.00000	00: NULO	0.26798	01: BAIXO	0.24917	02: MEDIO BAIXO	0.76096	05: SUPERIOR	0.44112	02: MEDIO BAIXO
0.00000	00: NULO	0.27126	01: BAIXO	0.15853	02: MEDIO BAIXO	0.70919	05: SUPERIOR	0.00000	00: NULO
0.00000	00: NULO	0.13379	01: BAIXO	0.58243	04: ALTO	0.65585	05: SUPERIOR	0.86523	05: SUPERIOR
0.00000	00: NULO	0.36630	02: MEDIO BAIXO	0.07492	01: BAIXO	0.46169	04: ALTO	0.41733	02: MEDIO BAIXO
0.00000	00: NULO	0.16621	01: BAIXO	0.11600	01: BAIXO	0.22117	02: MEDIO BAIXO	0.00000	00: NULO
0.00000	00: NULO	0.16256	01: BAIXO	0.04762	01: BAIXO	0.36313	03: MEDIO ALTO	0.42040	02: MEDIO BAIXO
0.00000	00: NULO	0.11210	01: BAIXO	0.04512	01: BAIXO	0.29846	02: MEDIO BAIXO	0.70268	03: MEDIO ALTO
0.00000	00: NULO	0.34217	02: MEDIO BAIXO	0.02995	01: BAIXO	0.40746	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.37405	02: MEDIO BAIXO	0.09015	01: BAIXO	0.40746	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.65389	05: SUPERIOR	0.04029	01: BAIXO	0.57467	04: ALTO	0.00000	00: NULO
0.00000	00: NULO	0.26912	01: BAIXO	0.06435	01: BAIXO	0.40746	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.09954	01: BAIXO	0.45738	03: MEDIO ALTO	0.68651	05: SUPERIOR	0.75579	04: ALTO
0.00000	00: NULO	0.32132	02: MEDIO BAIXO	0.20447	02: MEDIO BAIXO	0.33308	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.30032	02: MEDIO BAIXO	0.07456	01: BAIXO	0.46940	04: ALTO	0.83600	04: ALTO
0.00000	00: NULO	0.25915	01: BAIXO	0.47415	03: MEDIO ALTO	0.15469	01: BAIXO	0.27542	01: BAIXO
0.00000	00: NULO	0.45705	03: MEDIO ALTO	0.00000	00: NULO	0.44780	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.25534	01: BAIXO	0.18621	02: MEDIO BAIXO	0.31560	03: MEDIO ALTO	0.55382	03: MEDIO ALTO
0.00000	00: NULO	0.40027	03: MEDIO ALTO	0.15044	02: MEDIO BAIXO	0.32165	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.60501	05: SUPERIOR	0.02686	01: BAIXO	0.64570	05: SUPERIOR	0.70937	03: MEDIO ALTO
0.00000	00: NULO	0.43432	03: MEDIO ALTO	0.05762	01: BAIXO	0.17372	01: BAIXO	0.45727	02: MEDIO BAIXO
0.00000	00: NULO	0.29411	02: MEDIO BAIXO	0.04039	01: BAIXO	0.35127	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.31090	02: MEDIO BAIXO	0.06743	01: BAIXO	0.38247	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.28895	02: MEDIO BAIXO	0.14438	02: MEDIO BAIXO	0.53852	04: ALTO	0.00000	00: NULO
0.00000	00: NULO	0.70878	05: SUPERIOR	0.10148	01: BAIXO	0.42229	03: MEDIO ALTO	0.00000	00: NULO
0.00000	00: NULO	0.20654	01: BAIXO	0.03496	01: BAIXO	0.34881	03: MEDIO ALTO	0.00000	00: NULO

Figure 45 : The CRM Scoring table

Besides the 5 classes of scores that have been covered, the “00: Nulo” Classe of score is a reference for clients that have not been scored by the predictive model probably because they already have the product for which the model has been built or simply because they are not eligible to own the product. As it is being considered only individual clients for the Next Best Offer 2.0, the models that are going to be used are the following :

Credit Cards	Credit Card Acquisition Model (CC)
Habitation Loan	Habitation Loan Acquisition Model (CI)
Personal Loan	Personal Loan Acquisition Model (CP)
Insurances	Life Insurance Acquisition Model (PS) Personal Accidents Acquisition Model (PA) Travel Insurance Acquisition Model (PV) Medis Acquisition Model (MD) Homing Acquisition Model (HM) Mobis Acquisition Model (MB)
Solutions	Programa Prestige Acquisition Model (PP) Millennium GO! Acquisition Model (MG) Cliente Frequente Acquisition Model (CF)
Investment & Savings	Retirment/Savings Plan Acquisition Model (PPR)

Figure 46 : Table of the used models for the NBO 2.0

5.3.6. The Rating System for the NBO 2.0

In this part, there is going to be the gathering of all the output resulted from the 5.3.2, 5.3.3, 5.3.4 and 5.3.5 parts. As a reminder, it has been performed an analysis of the behavior of clients towards the bank's communication across all channels (Email, Branch, SMS etc.) for each product category. Their receptivity rate has also been calculated and classified per product category. It has also been built two recommendation systems which are the outer (for product families) and the inner (for the first 8 items in the product family) Recommendation System that is also going to be used for the decision of the Next Best Offer. There has also been collected the websites pages visits from the clients using the session from their web session to catch if they are interested by a certain type of product. This information is also by each product category. And finally, it has been stressed out the process of building a predictive model in general and provided the ones who are going to be used for this analysis. The next step is to gather this information based on the importance of each aspect in the contribution to the Next Best Offer.

It has been decided that based on each aspect's results, there will be a maximum of 100 points that will be shared among the elite product category from each pillar as all of the pillars discussed previously decide which is the best product category (Except the Inner Recommendation). Thus, for each client and depending on each aspect, the product category from which the client will probably buy his Next Best Offer is going to be the product category elite by all or most of the aspects.

From the 4 aspects that have been discussed, the closest one that represents an inbound action from the client is the website visits since, as mentioned before, it represents a real opportunity to understand what really the client is looking for. Based on this information, an offer will be adapted to his needs. Therefore, as this aspect is considered by far very important, it will be attributed a total of 50 points to the Elite product category. Afterwards, the predictive models give a very strong analytical strength in a way that each one of them is focused on branding a specific product. A total of 25 points are going to be given by the model (or the models) to the Elite Product category. For the outer-recommendation System, A total of 20 points will be given to the Next Best Product category with 5 points less than the predictive models due to the fact of being based on the possession profile only, whereas a predictive model is a set of variables that can go beyond the possession criteria. The last 5 points are going to be attributed to the product category elite from the communication profile, since the information given for some clients might be biased specially for clients who have been contacted very few times and answered the call.

SCORE_SOLUCOES	SCORE_CARTOES	SCORE_CPESSOAL	SCORE_CIMMOBI	SCORE_SEGUROS	SCORE_INVESTIMENTO	SCORE_POUPANCA
8	20	0	25	8	0	5
0	0	22	0	8	0	0
0	0	25	0	12	20	0
3	45	25	0	4	0	0
20	25	5	0	8	0	0
0	25	0	0	8	20	0
0	25	25	0	24	0	0
0	0	0	0	8	32	12
0	25	25	0	24	0	0
8	45	0	25	8	0	0
0	0	0	0	28	0	0
8	0	0	0	8	20	2
20	0	0	0	4	0	0
0	20	0	0	4	0	0
20	5	25	0	8	0	0
0	25	5	0	24	0	0
20	25	25	0	8	12	12
20	0	0	0	4	0	0
0	25	2	0	8	20	0
0	0	0	0	8	20	0

Figure 47 : Outer R.System Results

5.4. SUMMARY

The Millennium BCP Roadmap for the Next Best Offer 2.0 has been implemented by an inspiration from the structure proposed by the HP Engineers (Andrea Fabrizi, 2014). Their infrastructure of the NBO System has relied on ratings as well as the profiles of the products. Whereas the Millennium BCP infrastructure analyses the general possession profile from all the costumers without taking into consideration the gathering of the products information. Also, as mentioned previously, there has not been used any client's personal feedback or rating towards a given product or product category. Thus, the general structure of the Millennium BCP's Next Best Offer 2.0 can be represented as the following.

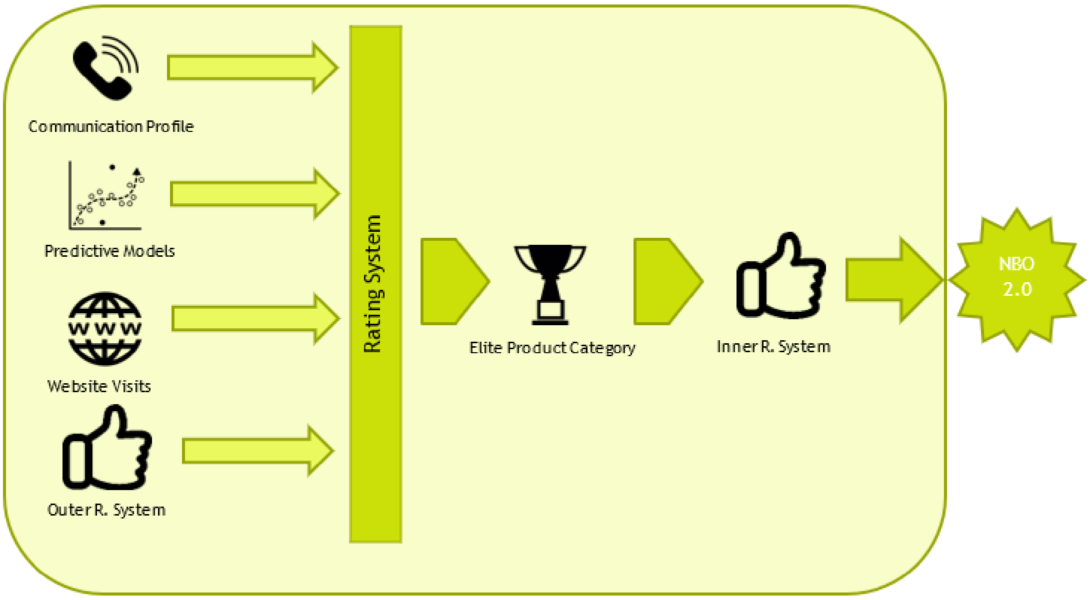

Figure 48 : The Millennium BCP implemented process for the Prediction of the Next Best Offer

Currently, the Model of the Next Best Offer 2.0 is put into a phase test. Various Marketing campaigns are created to test each Elite Product Category coming out from the Rating System.

Future steps that are going to be taken are to try to incorporate in this existing infrastructure, the clients' transactions history. Although transactions are not related to any product category, but they would tell more about the client's intensions in case of sudden high amounts of transactions (High amount purchases, monthly money transfer, dislocation of funds from one account to the other…) There are various possibilities that can be concluded when dealing with higher amounts of money. There is also another possibility that can enable this current engine to be much more efficient which is coupling it with the real time events that can be developed using SAS-Real Time Analytics. As this project is in the test phase using specific marketing campaigns, real results haven't been furnished yet from which the NBO 2.0 would be assessed. However, the feedback of the clients will be incorporated to in order to be learnt from by the model.

6. CONCLUSION

In this document there has been implemented 2 Data Science and Analytics projects. The first one was more related to the Business Intelligence area in a way that it has empowered the CRM Team to have a global vision on the evolution of the Banks retail products performance. The First project has added the CRM team more flexibility and more time saving from product managers to prepare their cyclic meetings for results presentation. With the panoply of dashboards built, many product managers from different areas can know exactly how well their marketing strategies have performed or have been performing. It also enables them to take actions quickly since the information they need is always available with the built CRM Sales Analytics Solution. The Second project was more related to Machine learning techniques and Data Analytics about the historical data and behavior of the client. In the future in Millennium BCP, All the marketing campaigns offers are going to be based on various behavioral, transactional, analytical data of the client collected since day 1 of his relationship with the bank. The Next Best Offer 2.0 that has been presented in only the beginning of a very ambitious project that will continue to include other aspects of analysis such as data that is not structured or not even registered. For example, just a friendly conversation with the branch manager can give a lot of insights and possibilities of cross-selling. It is planned to develop a platform in which there will be understood very well the needs of the clients. This information will contribute alongside the Next Best Offer Model to be more self-confident about the offer that will suit best the needs of the client. The projects discussed in this document might be diverging one another, but both aim to increase value and business in Millennium BCP since they are in line with the digital transformation that all companies are aiming for.

7. REFERENCES

Andrea Fabrizi, T. B. (2014). *HP SPS Next Best Offer, How to re-think your marketing.* HP.

Chen. (1976). *The Entity Relationship Model : Towards a Unified View of Data.* Massachusetts Institute of Technology.

Davenport, R. J. (2008). *ETL vs ELT.* Reading, England.

Fayyad, G. P.-S. (1996). *From Data Mining to Knowledge Discovery in Databases.* AAAI.

FERNÁNDEZ, L. E. (2018). *Recommendation System for Netflix.* Vrije University, Amsterdam.

K.J Raiha, H. M. (1992). *The Design of Relational Databases.*

Kirpes, D. (n.d.). *Dimensional Model Data Warehouse: An Overview (Why).* Novato, California: Fireman's Fund Insurance Company.

Llave, M. R. (2018). *Data lakes in business intelligence: Reporting from the trenches.* Kristiansand, Norway.

Mark Brown, J. B. (n.d.). *Data Mining.* The SAS Institute Inc.

Matthias Goeken, R. K. (2007). *Multidimensional Reference Models for Datawarehouse Development.* Frankfurt.

Miao Nie, A. A., & Shanshan Cong, S. I. (2016). *Build Recommender system with SAS to improve cross-selling etc for online business.* SAS Institute Inc.

Rajendra LVN, Q. W. (2014). *Recommending News Articles using Cosine Similarity Function.* The SAS Institute Inc.

Robin Burke, M. R. (2008). *Matching Recommendation Technologies and Domains.* DePaul University, Chicago, Illinois, USA.

Shearer, C. (2000). *The CRISP-DM Model: The New Blueprint for Data Mining.* The Journal of Data Warehousing.