

# **easyMoney - Business Case**

## **End-to-End Machine Learning Project**



**Project Title: easyMoney Data Science Capstone Project**  
**Project Team Members: Gonalo, Seema, Chukwubuikem**  
**Date: 06/04/2024**

# Index

<b>Index.....</b>	<b>2</b>
<b>1. Introduction.....</b>	<b>4</b>
1.1 Company Profile.....	4
1.2 Scope of the Project: easy Money.....	5
<b>2. Task 1 - Data Analysis: Exploratory Data Analysis.....</b>	<b>6</b>
2.1 EDA (Exploratory Data Analysis).....	6
2.1.1 Data Exploration.....	7
2.1.2 Objective of Data Analysis.....	7
2.1.3 Data Understanding:.....	7
2.2 Univariate Analysis.....	7
2.2.1 Number of rows in the dataset.....	7
2.2.2 Age grouping.....	8
2.3 Multivariate Analysis.....	10
2.3.1 Entry Channel Analysis.....	10
2.3.2 Customer Segment Analysis.....	10
2.3.3 Geographic Distribution.....	10
2.3.4 Customer Activity Status.....	10
2.3.5 Demographic Distribution.....	10
2.3.6 Active and In-Active customers.....	10
2.3.7 Definition of Old and New Customers.....	11
2.3.8 Analysis of the products.....	13
2.3.9 Product Analysis by type of product.....	14
2.3.10 Product Analysis by type of product.....	14
2.3.11 Analysis of sales/churn.....	15
2.3.12 Sales in the last month (May 2019).....	17
2.3.13 Pre-analysis for clustering.....	19
<b>3. Task 2 - Customer Segmentation.....</b>	<b>21</b>
3.1 Sales pattern in each segment.....	21
3.2 Splitting datasets.....	21
3.3 K-Means Clustering.....	22
3.3.1 Number of clusters.....	23
3.4 Clusters' analysis.....	24
<b>4. Task 3 - Target customers recommendation.....</b>	<b>29</b>
4.1 Analysis of 2018/2019 revenues.....	29
4.2 Gradient Boosting Classifier.....	31
4.2.1 Target: Saving Products.....	31
4.2.2 Target: Financing Products.....	32
4.2.3 Target: Investment Products.....	33

4.2.4 Target: Account Type Products.....	34
4.3 Expected response rate.....	35
4.4 Customers recommendations.....	36
4.4.1 Linear Programming.....	37
4.5 Incremental revenue.....	38
<b>5. Task 4 - Monitoring the process.....</b>	<b>40</b>
<b>6. Task 5 - Coordination and Planning.....</b>	<b>41</b>
<b>7. Conclusions/Further work.....</b>	<b>44</b>
<b>8. Appendixes.....</b>	<b>45</b>
8.1 Entry dates time series.....	45
8.2 Data dictionary.....	46

# 1. Introduction

## 1.1 Company Profile

**Background:** easyMoney was born almost 4 years ago from the imagination of Carol Denver, an investment banking professional, who after more than 10 years working for large firms, decided to launch her own business project: a multi-channel platform for marketing financial products (savings, investment, financing) with a friendly interface. It would allow customers to find solutions to their financial needs and to acquire them in a simple way. Their first product, the easyMoney piggy bank account (accumulate money in your piggy bank effortlessly and automatically by rounding up your purchases) was a great success, after which they have been expanding the product offering with investment solutions, cards, etc.

### **Challenges easy Money was undergoing :**

- After 4 years of activity, easyMoney is facing some challenges that put its continuity at risk: the addition of products to its catalog due to pressure from its easyBanking partners has changed the initial vision of offering simple products to just respond to customer needs. Furthermore, the funding obtained in the rounds has almost run out, without yet obtaining the expected positive EBITDA that allows them to start walking alone.
- On the other hand, the high turnover in the IT team and the lack of investment in technology is beginning to generate problems in all areas of the company, which complain
- of not having adequate means of work.
- Finally, internal tensions in the company are holding back the agile spirit that characterized the first developments.

**Probable Solution:** easyMoney's management has decided to fill this vacancy with the incorporation of a Data Scientist, who will help in this new stage to increase the profitability of the current client portfolio.

## 1.2 Scope of the Project: easy Money

The goal of this project is to complement programming practices with a scenario close to the reality of the Data Science profession, in which the requirements are poorly (or not at all) defined, and to provide data science solutions in an area of analytical marketing and business development.

**Task 1: Data Analysis:** We start from the dataset provided for the development of data of EasyMoney. The objective is to analyze and explore the data we have from a statistical and business perspective, to obtain an adequate vision of our business and in the future be able to face the tasks entrusted to us.

**Task 1 b : Dashboard:** The direct result of this task is the obtaining of a BI tool (Dashboard) that will be used to interactively explore all the visualizations and information provided below.

**Task 2: Segmentation:** After the analysis, we shall explore the segmentation of EasyMoney customers with the aim of guiding the company's commercial activity and helping to define a marketing plan based on the analysis of existing data.

**Task 3: Recommendation:** Helping the marketing team to understand the “ball of the future” (expected ROIs are of interest) and help them understand the response rate before sending the 10,000 emails.

Proposed idea from marketing team: To be able to send our customers the product that interests them the most (and the one that helps us earn more). Revenue is approximately €10 for each account sold, €40 for savings and investment products (plans, funds, etc.) and €60 for financing products (loans and cards).

**Task 4: Monitoring:** In this task we propose some KPIs to monitor the marketing campaign to follow up the process and understand if the feedback from the customers are according to our expectations.

**Task 5: Coordination:** At last, the team was asked to create a project plan for each phase of the project, and for that purpose, we used the kanban-style, list-making application Trello to coordinate the team as well as the timings and tasks.

## 2. Task 1 - Data Analysis: Exploratory Data Analysis

In this chapter we describe the procedure to perform data analysis. Figure 2.1 shows an overview of the analysis process we used.

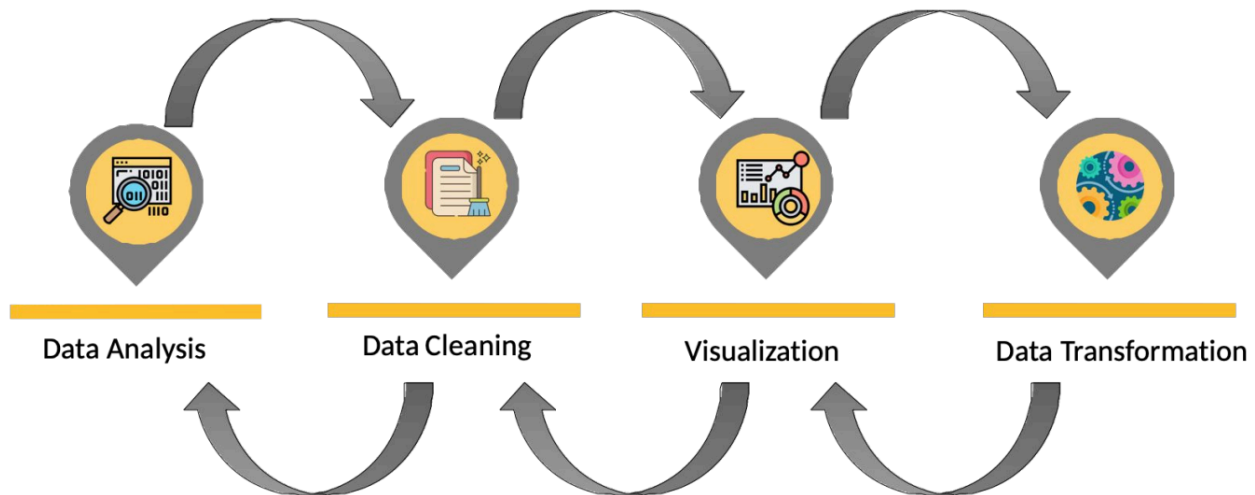


Fig. 2.1 - Task 1 Analysis: easyMoney: Data Cleaning and Data Analysis

Customer Database provided in csv format.

Table 2.1 - List of dataset provided to perform the project.

Socio	Provides information about features related directly with customers.
Commercial Activity	Provides information about the customer segments, entry date and respective channel.
Product	Provides specific information of how many products were sold at each day and customer.

### 2.1 EDA (Exploratory Data Analysis)

We Merged the 3 clean data sets and observed that in this Dataset, we have 2 columns in common with the other datasets: pk\_cid (Customer identifier) and pk\_partition (Data ingestion date). Before merging the cleaned data set, the below mentioned steps of data exploration were carried out on all 3 data sets separately.

### 2.1.1 Data Exploration

Step 1: After uploading the data and importing relevant libraries, we had an initial understanding of the data set: including its size, structure, and features.

Step 2: We understood Summary statistics, distributions, and visualizations of key variables

Step 3: Data pre-processing steps such as cleaning, missing value imputation, and feature engineering clean were done on the provided 3 data frames.

### 2.1.2 Objective of Data Analysis

Once we have merged the 3 cleaned datasets, we got a data set with size: 5962924 rows  $\times$  32 columns

### 2.1.3 Data Understanding:

Before exploring, we need to make sure the problem statement is duly addressed. The management and business expectations are:

- 1) To understand the total number of products we've sold this month / year
- 2) Sales made by customers are new or old
- 3) The Monthly and Annual Sales of the product

## 2.2 Univariate Analysis

### 2.2.1 Number of rows in the dataset

Table 2.2 - Number of rows in the merged dataset.

Total unique customers	unique pk_cid: 456373
Total unique rows	5962924

From this information, we can draw some conclusions:

- ❖ There are more rows in the dataset (5,962,924) than there are unique customers (456,373), suggesting that each customer has multiple records or interactions within the dataset. This could include multiple transactions, interactions with different products or services, or changes over time.
- ❖ The ratio of unique customers to total rows ( $456,373 / 5,962,924 \approx 0.076$ ) indicates that, on average, each customer has approximately 13 rows of data associated with them. This average number of records per customer can provide insight into the level of detail available for analyzing individual customer behavior.

### 2.2.2 Age grouping

We defined the Age Groups as shown in table 2.3.

Table 2.3 - Number of rows in the merged dataset.

Group name	Age Range
Teenager	0-18
Young Adults	18-34
Adults	35-54
Senior	55-64
Old Age	65+

We understood the split of the minimum age is 2 years. The maximum age is 105 years, while the average age in the dataset is 29.76 years.

Above 70% of the age group are 'Young Adults' and are within the range 18 - 34 years of age, as shown fig. 2.2.

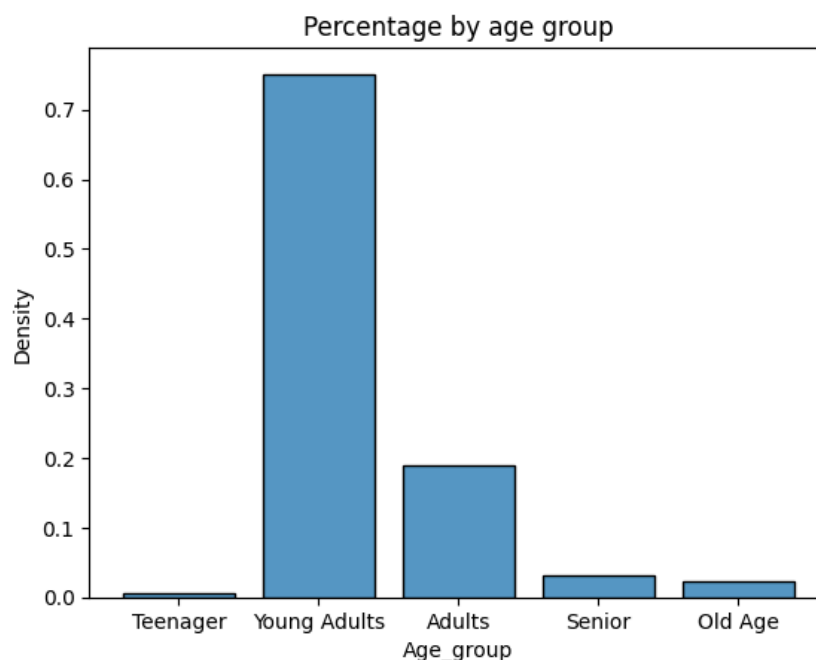


Fig. 2.2 - Percentages of clients by each group age.

Further Analysis of Salary: The minimum salary is €1,202.73 while the maximum salary is €28,894,396, while the average salary is €115816.72.



Average salary with reference to the Age Group is shown in table 2.4.

Table 2.4 - Average salary by age group.

Age Group	Average salary (€)
Teenager	122403.69
Young Adults	85236.22
Adults	85862.04
Senior	98060.71
Old Age	106590.35

Also we observe the salary with reference to the top 3 segments in the data set (table 2.5).

Table 2.5 - Average salary by segment.

Segment	Average salary (€)
Top	114306.27
Particulares	88760.14
Universitario	86603.52

### Conclusions with reference to Feature Salary:

#### Age Group Analysis:

- ❖ Teenagers have the highest average salary among all age groups, followed by Old Age individuals.
- ❖ Young Adults and Adults have similar average salaries, suggesting that there might not be significant differences in earnings between these two age groups.

#### Segment Analysis:

- ❖ The top segment (01 - TOP) has the highest average salary among the segments analyzed
- ❖ The UNIVERSITARIO segment has the lowest average salary among the segments analyzed. This could indicate that individuals in this segment, possibly recent graduates or those in academic or research roles, tend to earn less on average compared to those in other segments.

## 2.3 Multivariate Analysis

### 2.3.1 Entry Channel Analysis

KHE dominates among all the channels, comprising approximately 52% of the total share. Following KHE, KFC holds around 15% of the customer acquisition share.

### 2.3.2 Customer Segment Analysis

The UNIVERSITARIO segment stands out, constituting approximately 65.4% of Easy Money Bank's customer base. PARTICULARS make up around 31% of the total customers.

### 2.3.3 Geographic Distribution

Almost all customers, approximately 99.96%, hail from Spain, specifically from region code 28.

### 2.3.4 Customer Activity Status

On around 60% of the dataset ingestions, the customers were classified as inactive, while the remaining 40% as active.

### 2.3.5 Demographic Distribution

The Young Adults category represents a significant portion, contributing to 75% of the total customer entries on the set. Adults make up around 19% of the customer demographic.

These observations provide valuable insights into the customer acquisition, segmentation, geographic distribution, activity status, and demographic composition of Easy Money Bank's customer base. Further analysis and strategies can be developed based on these findings to optimize marketing efforts, improve customer engagement, and enhance overall business performance.

### 2.3.6 Active and In-Active customers

Considering the data ingestion date from period 28-01-2018 to 28-05-2019, we have observed that the activity of customers at each date/month was according to one shown in table 2.6.

Table 2.6 - Monthly percentage of active/inactive customers.

pk_partition	2018-01-28	2018-02-28	2018-03-28	2018-04-28	2018-05-28	2018-06-28	2018-07-28	2018-08-28	2018-09-28	2018-10-28	2018-11-28	2018-12-28	2019-01-28	2019-02-28	2019-03-28	2019-04-28	2019-05-28
perc_active_customers	45.18	45.8	46.4	47.0	47.64	48.22	38.1	38.63	38.58	37.77	37.54	37.69	38.08	38.23	38.56	38.67	38.73
perc_inactive_customers	54.82	54.2	53.6	53.0	52.36	51.78	61.9	61.37	61.42	62.23	62.46	62.31	61.92	61.77	61.44	61.33	61.27

So, by observing the table 2.6, we conclude to have a drop in Active Customers since 28-07-2018:

- ❖ There has been a noticeable drop in active customers from 28-07-2018 onwards.
- ❖ This drop indicates a decline in customer activity or engagement with Easy Money Bank's services.
- ❖ Active Customers in June 2018:
- ❖ June 2018 is highlighted as having the highest percentage of active customers.
- ❖ This suggests that June 2018 was a particularly strong month in terms of customer engagement or usage of Easy Money Bank's offerings.

#### Comparison between May 2018 and May 2019:

Active customers in May 2018 accounted for 47.64% of the total customer base.

However, in May 2019, the percentage of active customers dropped to 38.73%.

This represents a significant decrease in active customer participation between May 2018 and May 2019.

The percentage drop in active customers from May 2018 to May 2019 highlights a decline in customer activity or engagement over this period.

These observations suggest fluctuations in customer engagement over time, with notable drops in active customer percentages, especially from May 2018 to May 2019. Further analysis may be needed to understand the factors contributing to these fluctuations and to develop strategies to address them.

### **2.3.7 Definition of Old and New Customers**

It was performed some visualizations to understand the amount of old and new customers along the months, and based on the provided information, we obtained the graph shown in fig. 2.3.

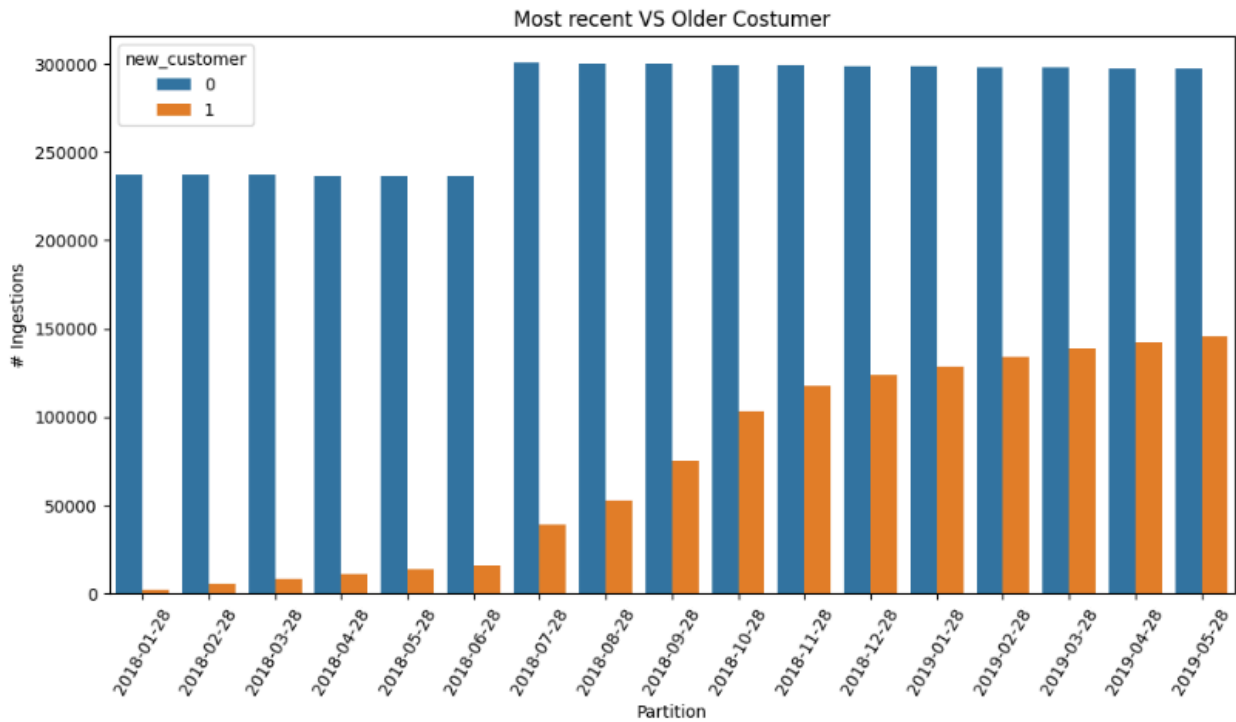


Fig. 2.3 - Number of old/new customers by partition dates.

**Note:** Old customers are defined as those who entered before 01/01/2018. New customers are defined as those who entered after 01/01/2018.

### Trend Analysis

There is a gradual increasing trend in the number of new customers throughout 2018. This indicates that Easy Money Bank was acquiring new customers steadily over the course of 2018, suggesting successful marketing or customer acquisition efforts during that period.

### Stagnation in New Customers in 2019

From January 2019 to May 2019, the number of new customers remains relatively constant. This suggests that there was a slowdown or stagnation in acquiring new customers during this period. The observation implies that Easy Money Bank did not acquire many new customers during the first five months of 2019 compared to the previous year.

### Conclusion

The analysis indicates a shift in the trend of acquiring new customers, with a significant slowdown in 2019 compared to 2018. Further investigation may be necessary to understand the reasons behind this stagnation and to develop strategies to attract new customers effectively in the future.

These observations provide insights into the dynamics of customer acquisition at Easy Money Bank, highlighting changes in the rate of acquiring new customers over time.

### 2.3.8 Analysis of the products

Based on the provided data, table 2.7 shows the **top 5 popular products** offered by Easy Money Bank, along with their respective percentages (in terms of ingestions in the dataset):

Table 2.7 - Top 5 most popular products.

Product	Percentage of ingestions
em_account	70.05%
debit_card	9.00%
payroll_account	5.27%
emc_account	5.23%
pension_plan	3.48%

These percentages represent the relative popularity or usage of each product among Easy Money Bank customers. It's evident that "*em\_account*" is the most popular product by a significant margin, followed by "*debit\_card*," "*payroll\_account*," "*emc\_account*," and "*pension\_plan*" respectively. Understanding the popularity of these products can help Easy Money Bank focus its marketing efforts and product development strategies.

#### Trend of "*em\_count*" Across All Age Groups

Across all age groups, there is a consistent trend of increasing usage of the "*em\_count*" product over time.

This suggests a general preference or adoption of the "*em\_count*" product among Easy Money Bank customers, regardless of age.

#### Unique Pattern for Teenagers

Despite fluctuations in usage over time, the most common product among teenagers is the "*pension\_plan*," not the "*em\_count*" product.

This indicates that teenagers, unlike other age groups, show a distinct preference for the "*pension\_plan*" product over the "*em\_count*" product.

The reasons behind this unique pattern for teenagers, such as specific marketing strategies or demographic characteristics, would require further investigation.

#### Conclusion

Overall, while there is a consistent trend of increasing usage of the "*em\_count*" product across all age groups, teenagers stand out by showing a preference for the "*pension\_plan*" product over "*em\_count*." Understanding these differences can help Easy Money Bank tailor its product offerings and marketing strategies to better meet the needs and preferences of different customer segments.~

### 2.3.9 Product Analysis by type of product

Based on the classification provided, here's how the products offered by EasyMoney are grouped into the four product families:

Table 2.8 - Classification of each product (Savings, Financial, Investment or Account type).

Product type	Products
Savings Products:	<i>'long_term_deposit';</i> <i>'pension_plan';</i> <i>'short_term_deposit'.</i>
Financial Products:	<i>'funds'</i> <i>'securities'</i>
Investment Products:	<i>'credit_card'</i> <i>'loans'</i> <i>'mortgage'</i>
Account Type Products:	<i>'debit_card'</i> <i>'em_acount'</i> <i>'emc_account'</i> <i>'payroll'</i> <i>'payroll_account'</i> <i>'em_account_p'</i> <i>'em_account_pp'</i>

These classifications group the products offered by EasyMoney into four distinct families based on their typology: Savings, Financial, Investment, and Account Type. This categorization can help in analyzing and managing the product portfolio effectively and understanding customer preferences within each product family.

### 2.3.10 Product Analysis by type of product

Another analysis was performed to understand if some products have relation with another, i.e, to understand if when a customer buys some product, he will buy another specific product. It was created a headmap which can be visualized in the coding file.

Additionally some bar plots were created to show what happens to other products when the client buys each one of the products. In other words, when a client has an entry buying one product (value 1), how are, in general, the values for other products (don't buy - 0, or buy - 1).

By the observations, the following conclusions were made:

- ❖ *short\_term\_deposit*, *loans* and *funds* products have some relation with *em\_acount* products (not the opposite);
- ❖ *mortgage* has some relation with *payroll*, *pension\_plan* and *debit\_card* products;

- ❖ *credit\_card* has some relation with *debit\_card* and the trend is *em\_account* to be 1 when the *credit\_card* is 0;
- ❖ all ingestions of *payroll* products have *pension\_plan* equal to 1. Also has strong relation with *payroll\_account* type and *debit\_card*;
- ❖ *pension\_plan* has strong relation with *payroll*, but not all ingestions have *payroll* equal to 1;
- ❖ *payroll* has some relation with *payroll\_plan* and *pension\_plan*, but it is stronger if we look in the perspective of *payroll* and *pension\_plan* then the opposite.

### 2.3.11 Analysis of sales/churn

The merged dataset gives us valuable information about the profile of each customer as well as about the products owned by this client at each ingestion date.

It means that if the customer, for example, has *em\_account* equal to 1 in two consecutive months, it doesn't mean that customer bought another *em\_account*: it means he maintained the product.

So, in that sense we define that the client buys some product when the difference between two consecutive months is 1 and he stops using the product when that difference is -1, ie, here we're referring to customer churn associated with this specific product.

The details of such computations are described in the coding file. Below, in table 2.9, one can see the total sales for each product by month.

Table 2.9 - Monthly total sales for all products.

pk_partition	sale_loans	sale_short_term_deposit	sale_mortgage	sale_funds	sale_securities	sale_long_term_deposit	sale_credit_card	sale_payroll	sale_pension_plan	sale_payroll_account	sale_emc_account	sale_debit_card	sale_em_account_p	sale_em_account
2018-01-31	19	883	15	786	932	4884	3325	8145	8835	13478	15320	24696	2	215293
2018-02-28	0	665	0	97	68	225	510	1966	2006	974	807	3708	0	4156
2018-03-31	4	652	2	119	47	261	735	1718	1734	1015	915	4058	0	4021
2018-04-30	3	419	1	127	45	249	652	1344	1855	1098	840	3500	0	3608
2018-05-31	3	383	1	83	47	355	621	1594	1606	1125	865	3184	0	3754
2018-06-30	0	443	1	53	42	387	749	2160	2694	1119	685	3690	0	5565
2018-07-31	1	451	1	54	59	450	760	2310	2506	1649	552	3969	0	13289
2018-08-31	2	434	0	63	101	345	694	1642	1665	1556	448	3659	0	14888
2018-09-30	2	498	0	38	77	497	749	2146	2190	1075	586	5288	0	18404
2018-10-31	2	412	0	40	179	506	743	2384	2410	1398	569	6319	0	19963
2018-11-30	0	96	0	52	39	366	713	2331	2370	1861	943	5258	0	11358
2018-12-31	1	6	0	60	86	464	664	2589	2683	2294	901	5154	0	8060
2019-01-31	1	0	1	36	254	194	653	1457	1707	1365	1178	4694	0	6288
2019-02-28	1	0	0	25	120	219	708	4037	4691	1379	1235	6138	0	4997
2019-03-31	3	0	0	28	65	76	741	2391	2458	1516	1033	5580	0	5538
2019-04-30	2	0	3	26	44	27	713	2187	2323	1475	1283	5005	0	4231
2019-05-31	0	0	0	18	39	12	761	2764	2819	1531	1607	5140	0	4035

As we can see (similar to the analysis about products' popularity), the most popular products in terms of sales are *em\_account*, followed by *debit\_card* and *pension\_plan*.

Interesting to see that in the first month of ingestion, the numbers are so high. It suggests that, as such day was the first the bank registered the amount of products owned by those customers, there were inserted all the products they had since the first entry date, ie, since 2015.

Concerning the churn, table 2.10 summarizes the number of customers that stopped using each product by month since January 2019.

Table 2.10 - Monthly total churn for all products.

pk_partition	churn_loans	churn_short_term_deposit	churn_mortgage	churn_funds	churn_securities	churn_long_term_deposit	churn_credit_card	churn_payroll	churn_pension_plan	churn_payroll_account	churn_emc_account	churn_debit_card	churn_em_account_p	churn_em_account
2018-01-31	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2018-02-28	0	205	0	11	35	153	553	1217	1340	206	300	2913	0	2351
2018-03-31	0	331	0	22	47	252	506	862	1336	272	314	2432	0	2436
2018-04-30	2	343	1	12	27	107	410	1173	1176	261	321	2742	0	2433
2018-05-31	0	668	0	22	18	106	558	1502	2014	254	243	3013	0	2290
2018-06-30	0	654	0	39	35	155	516	1198	1198	2271	293	2754	0	2466
2018-07-31	0	413	0	41	25	242	672	1397	1393	278	328	3373	0	3386
2018-08-31	1	381	0	16	21	218	625	2142	2313	354	300	4010	0	3222
2018-09-30	0	446	0	20	9	262	666	1679	1692	331	301	2792	0	3158
2018-10-31	0	424	1	28	32	284	647	1776	1807	315	342	3589	0	3454
2018-11-30	1	452	0	19	27	299	585	1749	1769	368	293	4141	0	4137
2018-12-31	3	496	0	20	26	162	640	1166	1173	2372	352	3168	0	4019
2019-01-31	3	403	0	28	19	284	696	3846	4487	400	298	4744	0	3419
2019-02-28	0	114	0	39	28	220	659	1713	1960	439	356	4064	0	3769
2019-03-31	3	9	0	22	43	197	707	1647	1739	390	347	3557	0	3472
2019-04-30	1	0	0	27	31	196	542	2065	2073	363	307	4229	0	3423
2019-05-31	0	0	0	23	32	251	707	1660	1684	501	306	4003	0	3484

Generally speaking, as expected, the churn rate trend is similar to the sales, ie, on the months and products where the number of sales are higher, the churn rate is higher, because to have a churn rate, the clients must buy the products first. Hoewer, at first sight, the balance tends to be positive because we have more sales than churn (in general).

The information shown in table 2.10 is more clear if we look the following visualization:

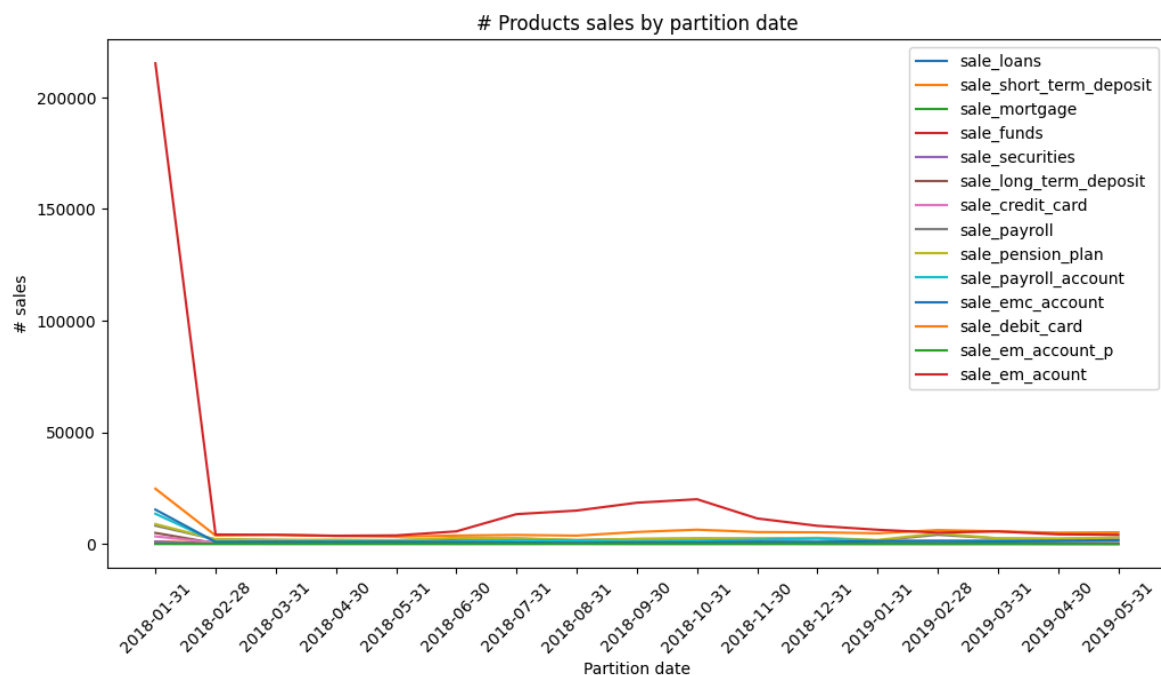


Fig. 2.4 - Monthly sales of each product.

As we can see, the first month amount, especially for em\_account, is high (above 200,000 sales). It's not clear the trend for other products, but we can see that the number of monthly sales of em\_account until September 2018 was increasing, but it started decreasing after around this month.

So, to have a clearer idea about the last 6 months, we created a visualization from December 2018 until the last month (May 2019). It can be observed in fig. 2.5.



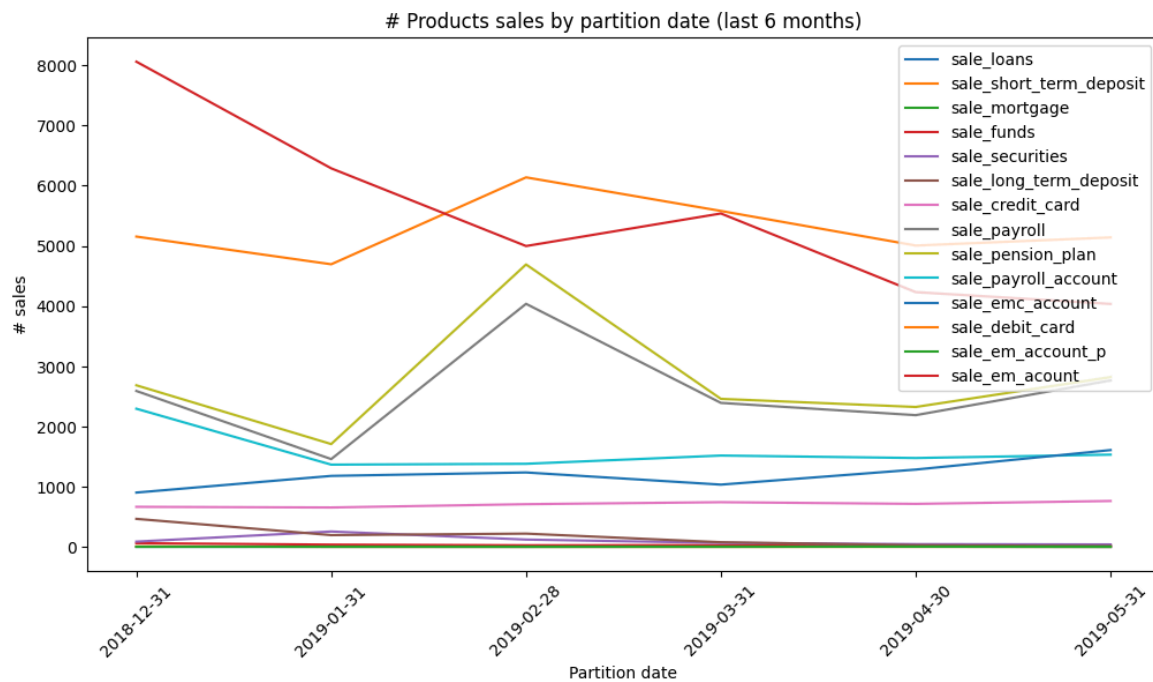


Fig. 2.5 - Monthly sales of each product.

By looking at the line plot above, we can see that for the products with more sales in the last months are *em\_acount* and *debit\_card*, followed by *pension\_plan* and *payroll*. On another hand, since January 2019, the bank hasn't sold any *short\_term\_deposit* and just 4 *mortgages*.

Some visualizations were made to understand if we have a pattern considering the months of the years, in terms of sales/churn, but just comparing the 5 first months of 2019 with the 5 first months of 2018 isn't enough to make reasonable conclusions.

### 2.3.12 Sales in the last month (May 2019)

Regarding the last month, we performed a separated visualization to see the comparison among all products sales, and distinguishing the sales of newer/older customers. Below one can see such visualization.

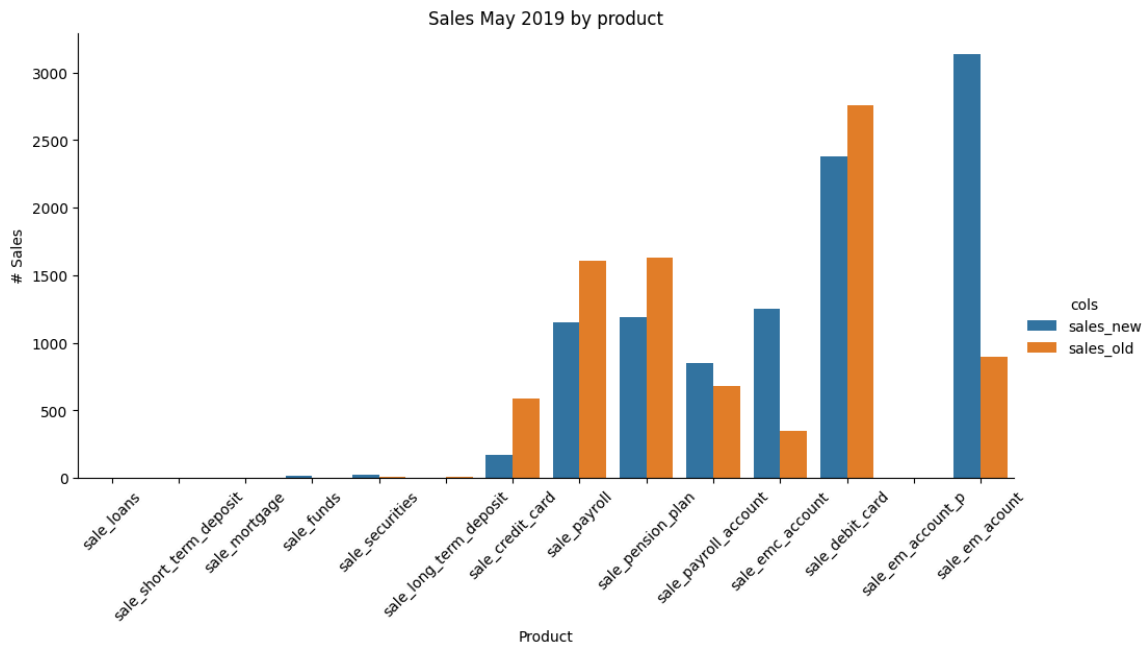


Fig. 2.6 - Sales of each product and new/old clients. (May 2019)

The same approach was made for the churn. See figure 2.7.

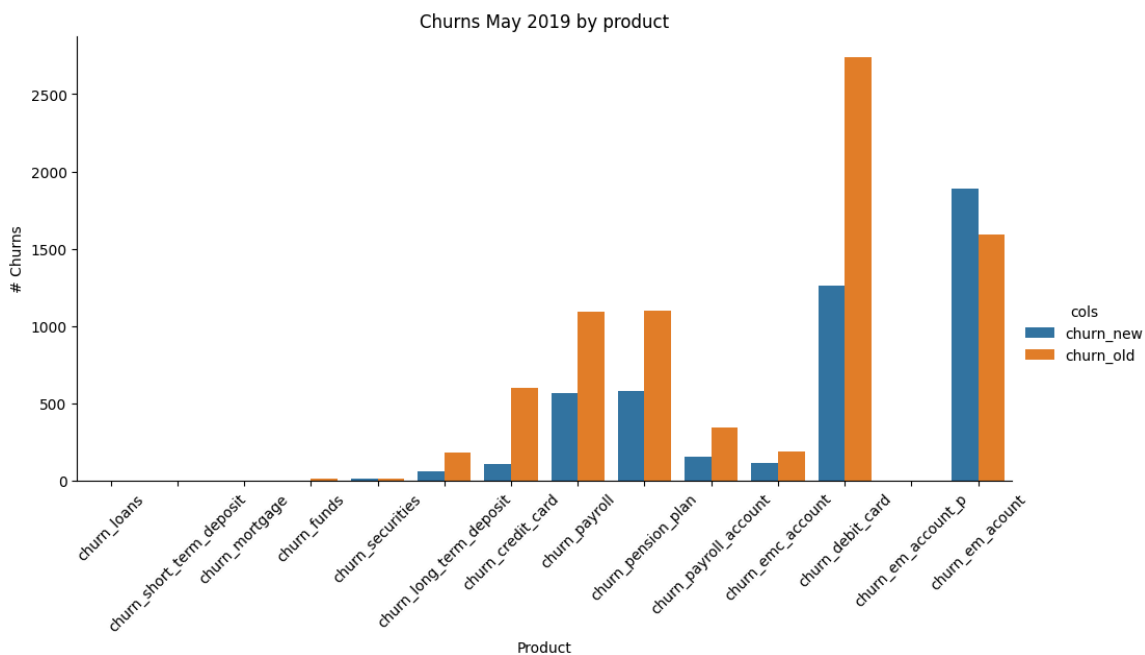


Fig. 2.7 - Churn associated with each product and new/old clients. (May 2019)

Concerning the *em\_acount*, new customers bought more than 3x compared with old customers, which bought just around 1500 *em\_acounts*, having a similar amount of such customers that stopped using the product.

If we talk about *long\_term\_deposits*, *credit\_cards*, and *payroll*, the older customers bought more compared with the newer. However, the difference is not so significant.

In May 2019, the total of 18,726 sales, 54.4% were from new customers and 45.6% from the older ones.

### 2.3.13 Pre-analysis for clustering

The next task we had to perform was to perform the segmentation of customers, so before proceeding to that task, we did a quick analysis to understand a priori some differences or similarities among the customers considering their age/age group, segment, antiquity or salary level.

For purpose of defining the salary level (new feature), we decided to classify each customer according to his salary and the criteria was to assign as 0 all the customers with a salary below the value of the 3rd quartile of salary feature, and to assign as 1 the remaining, ie, the customers with a salary above such value (€ 113.462.20).

With that pre-analysis we made some conclusions:

- ❖ The average age is strongly related with the segment, if we distinguish between Universitario and No Universitario (Top, Unknown and particular) groups;
- ❖ Most of the top/particular clients are fall into the adults group;
- ❖ In general, the segment with more sales for most of the products except em\_acount is the particular group.
- ❖ Practically 100% of the *universitarios* are young adults;
- ❖ Except for mortgage, for most of the sales, the amount of low-salary customers that bought the products is higher (which makes sense because the percentage of such customers is higher too);

Tables 2.11 and 2.12 show some conclusions concerning the segment and the age group, respectively.

Table 2.11 - Conclusions associated with salary level vs segment.

Segment	Comments
<b>Top</b>	0.96% and 0.68% of customers are in the group of lower salaries and higher salaries respectively (59% vs 41% if we consider the proportion on the group);
<b>Particulares</b>	22.81% and 7.90% of customers are in the group of lower salaries and higher salaries respectively (74% vs 26% if we consider the proportion on the group);
<b>Universitarios</b>	49.18% and 16.22% of customers are in the group of lower salaries and higher salaries respectively (75% vs 25% if we consider the proportion on the group);

Table 2.12 - Conclusions associated with salary level vs segment.

Age group	Comments
<b>Teenager</b>	0.32% and 0.29% of customers are in the group of lower salaries and higher salaries respectively (52% vs 48% if we consider the proportion on the group)
<b>Young adults</b>	56.87% and 18.23% of customers are in the group of lower salaries and higher salaries respectively (76% vs 24% if we consider the proportion on the group)
<b>Adults</b>	14.26% and 4.74% of customers are in the group of lower salaries and higher salaries respectively (65% vs 35% if we consider the proportion on the group)
<b>Senior</b>	2.08% and 1.00% of customers are in the group of lower salaries and higher salaries respectively (68% vs 32% if we consider the proportion on the group)
<b>Old age</b>	1.47% and 0.74% of customers are in the group of lower salaries and higher salaries respectively (67% vs 33% if we consider the proportion on the group)

## Conclusions

Considering the previous analysis, we decided that to proceed with the segmentation, we could start by splitting the customer set into 2 groups: **Universitarios** and **No Universitarios**, and after, we considered some features as age, antiquity, salary level or some product type specific sales to create the remaining 5/6 customer clusters.

That decision was made due the fact that it is clear the Universitario segment has significant differences when compared to the other segments, generally speaking.

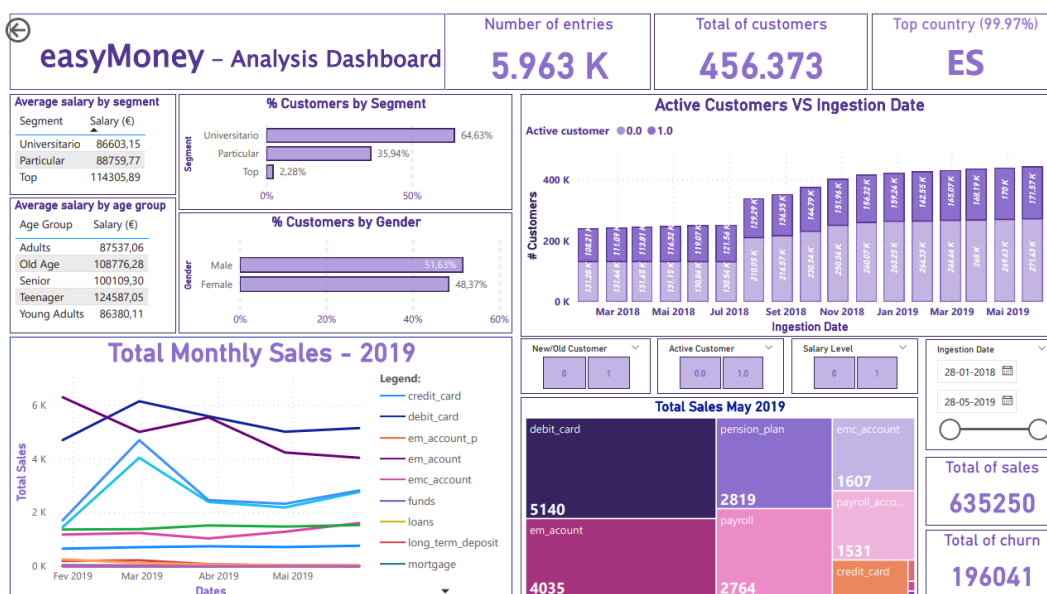


Fig. 2.8 - Dashboard for Analysis Dashboard.

## 3. Task 2 - Customer Segmentation

### 3.1 Sales pattern in each segment

Before performing the PCA and K-Means algorithms to create the customers segments, we created new columns on the dataset to insert the total sales/churn for each product type (by summing the sales of each group, as shown in table 2.8 in page 13).

Detailed calculations were made in the coding file *Task02-Segmentation*, and we present in table 3.1 the general proportion of sales by product type in each segment.

Table 3.1 - Sales' percentage by product type for each segment.

Segment	Savings	Financing	Investment	Account type
Unknown	1.9	0.1	0.0	98.0
Top	26.1	3.7	4.6	65.6
Particulares	14.3	1.0	4.2	80.5
Universitarios	6.0	0.2	0.9	92.9

By analyzing the table above, we can conclude that in the group Universitarios account type products were, by far, the most sold.

For other segments, such products are the most popular, but the distribution among the product types is “more balanced”, ie, 26.1% and 14.3% of the products sold in the segments top and particulares, respectively are saving products and 3.7% are financing products in the top group. This is valuable information in terms of sales customer behavior because we have interest in selling more expensive products, like savings, financing and investment products (revenues will be shown in task 3: Recommendation).

So, it shows that the number of savings, financing and investment products should be an important feature to take into account when performing the segmentation.

Regarding the churn rate, it is information that won't add so much value to proceed with the segmentation, so we decided to not take into account.

### 3.2 Splitting datasets

As mentioned before, the first step to segmenting the customers was to split them into 2 groups: **Universitarios (UNI)** and **No Universitarios (NUNI)**. The first group was composed of 207,383 customers and the second by 248.990.

The features we considered to segment the 2 groups were *salary\_level*, *savings\_sold*, *financing\_sold*, *investment\_sold* and *account\_type\_sold* for Universitarios and *salary\_level*, *new\_customer*, *savings\_sold*, *financing\_sold*, *investment\_sold* and *account\_type\_sold*. The

antiquity for the Universitarios doesn't have importance in distinguishing them because the majority of them are older customers.

### 3.3 K-Means Clustering

Before proceeding to the segmentation, we performed PCA to know the importance of each feature in the dataframe and how much each feature explains our data.

The first step was to standardize the features by removing the mean and scaling to unit variance.

Next we obtained the explained variance ratio for each feature and we created a linear plot for UNI and NUNI groups, which can be seen in figures 3.1 (a) and (b), respectively.

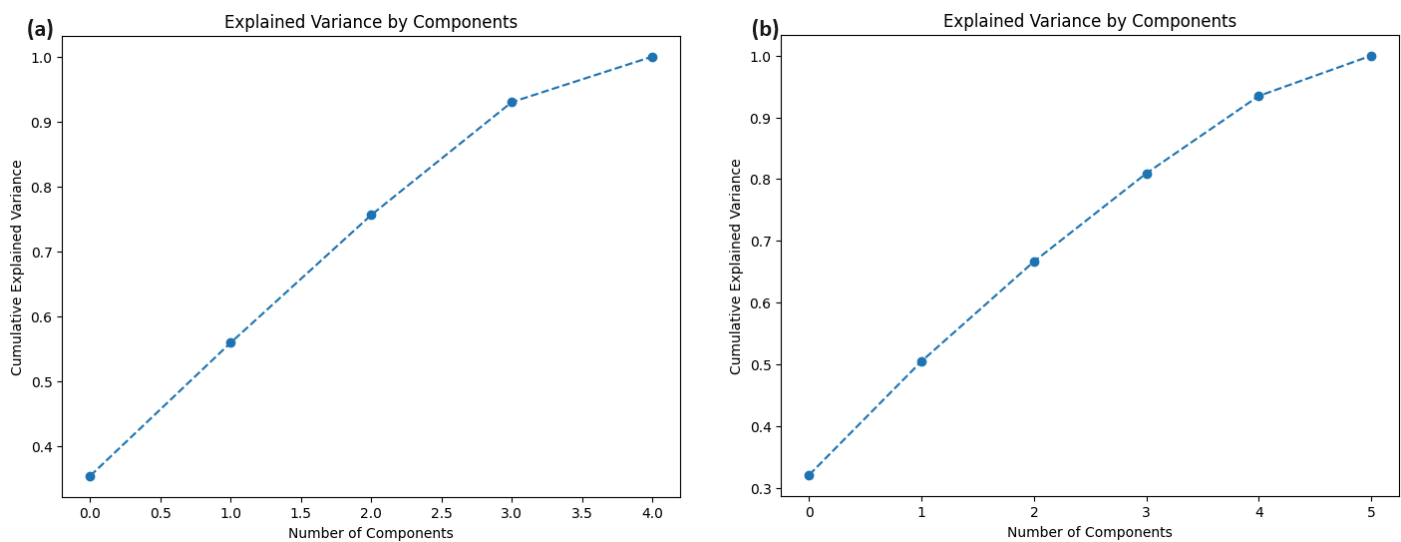


Fig. 3.1 - Cumulative Explained Variance vs number of components for UNI customers (a) and NUNI customers (b).

Looking at the UNI plot, we can see that the first 3 components explain more than 90% of the data and the NUNI plot tells us the first 4 components explain more than 90% of the data.

In that sense, for the UNI and NUNI group we reduced our features to three and four components, respectively, from the original five values that explain the shape the values themselves show the so-called loadings, which are correlations between an original variable and the component.

For instance, the first value of the array shows the loading of the first feature on the first component.

Next we did the visualization on the “weight” of each component for each original feature by creating a heatmap Components VS Original features.

Figure 3.2 (a) and (b) shows the heatmap for UNI and NUNI groups respectively.

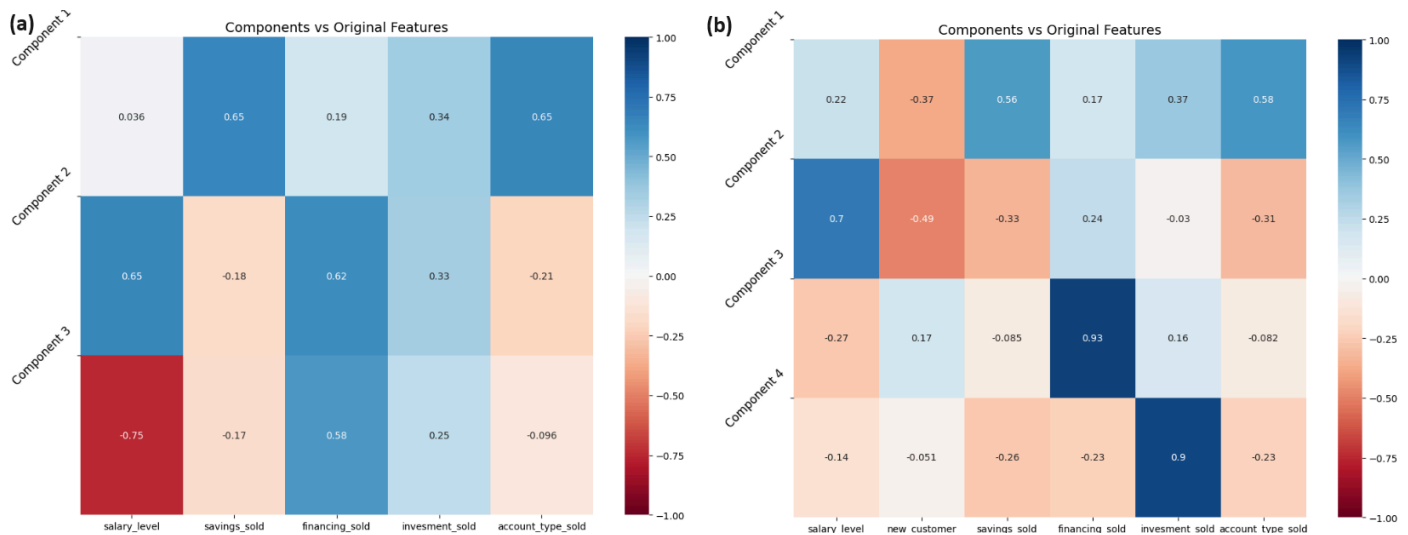


Fig. 3.2 - Components vs original features for UNI customers (a) and NUNI customers (b).

### Concerning the UNI group:

We see that there is a positive correlation between Component 1, amount of savings sold and account type sold. This shows that this customer tends to buy more saving products and account types compared to the others.

For the second component, financing sales and salary level are the most prominent determinants. Such customers tend to buy more financing products and those are the ones with a higher salary level.

For the final component salary is again by far the determinant, but contrary to the second, the relation is negative, meaning that such clients have a lower salary level.

### Concerning the NUNI group:

We see that there is a positive correlation between Component 1, amount of savings sold and account type sold. This shows that this customer tends to buy more saving products and account types compared to the others.

For the second component, new\_customer and salary level are the most prominent determinants. Such customers have higher salaries and most of them are old customers.

For the third component, financing products sold is the most prominent determinant. Such customers tend to buy more financing products compared to the others.

For the final component investment products sold is by far the determinant, meaning that such customers are investment buyers.

Now, we have an idea about our new variables (components). We can clearly see the relationship between components and variables.

### 3.3.1 Number of clusters

Regarding the number of clusters to split the 2 groups, the coding file has more detailed information and we decided to split the UNI group into 3 segments and the NUNI group into 4 distinct groups, which mean features can be observed in tables 3.2 and 3.3.

Table 3.2 - Mean for each feature used to segment the UNI group.

	salary_level	savings_sold	financing_sold	invesment_sold	account_type_sold
<b>Segment K-means</b>					
0	0.000000	0.008176	0.000275	0.000783	0.987805
1	0.360368	1.709169	0.047938	0.280936	5.235925
2	1.000000	0.014343	0.000695	0.001207	0.900990

From the clusters defined above, one can comment them as:

- ❖ 0: Low salary customers: have low salaries;
- ❖ 1: Buyer customers: 64% have low salary, but are saving and account type buyers;
- ❖ 2; High salary customers: have high salaries.

Table 3.3 - Mean for each feature used to segment the NUNI group.

	salary_level	new_customer	savings_sold	financing_sold	invesment_sold	account_type_sold
<b>Segment K-means</b>						
0	0.000000	0.705795	0.026028	0.000000	0.005733	0.889123
1	0.157244	0.327295	1.437406	0.000000	0.385127	4.440729
2	0.367954	0.256175	0.596263	1.116846	0.287524	2.402470
3	1.000000	0.400451	0.130129	0.000000	0.023280	1.169288

From the clusters defined above, one can comment them as:

- ❖ 0: Low Salary No Universitario: low salary customer, 70% are new customers and buy too little products;
- ❖ 1: Saving Buyer No Universitario: buy saving and account\_type products;
- ❖ 2: Financing Buyer No Universitario: buy financing and account\_type products;
- ❖ 3: High Salary No Universitario: high salary customers, 40% are new customers and buy a small amount of products.

### 3.4 Clusters' analysis

In the end, we made a quick analysis, especially on the remaining features of the clients that didn't contribute to the clustering.

A new feature was created to the name of each customer segment: *Segment Customer*, and the names were inserted as shown in table 3.4.



Table 3.4 - Segment Customer names.

Group	Segment K-means	Segment Customer name
Universitarios	0	Low Salary Universitario
	1	Buyer Universitario
	2	High Salary Universitario
No Universitarios	0	Low Salary No Universitario
	1	Saving Buyer No Universitario
	2	Financing Buyer No Universitario
	3	High Salary No Universitario

Concerning the percentage of each segment we can say that:

- ❖ **Low Salary Universitario** corresponds to 70.1% of total Universitario customers;
- ❖ **Buyer Universitario** corresponds to 3.5% of total Universitario customers;
- ❖ **High Salary Universitario** corresponds to 26.4% of total Universitario customers;
- ❖ **Low Salary No Universitario** corresponds to 73.1% of total No Universitario customers;
- ❖ **Saving Buyer No Universitario** corresponds to 9.9% of total No Universitario customers;
- ❖ **Financing Buyer No Universitario** corresponds to 1.3% of total No Universitario customers;
- ❖ **High Salary No Universitario** corresponds to 15.7% of total No Universitario customers;

Concerning the general statistics (detailed in the coding file) of each *Segment Customer*, one can make the conclusions shown in table 3.5.

Table 3.5 - Analysis of some features for all the segments created.

Segment Customer	Comments
<b>Low Salary Universitario</b>	As expected, the level of salaries of most of such customers tends to be low. The average age is low, most of them are no active customers (in mean), and most of them are old clients.
<b>Buyer Universitario</b>	The average age is quite higher than the other 2 groups, most of them are active customers (in mean), and 36% have a high level salary.
<b>High Salary Universitario</b>	As expected, the level of salaries of most of such customers tends to be high. The average age is low, most of them are no active customers (in mean), and most of them are old clients.
<b>Low Salary No Universitario</b>	The salary level tends to be low, as expected. The percentage of new customers is around 70% and the mean age is 34 (higher than universitarios)
<b>Saving Buyer No Universitario</b>	16% have a high salary level. The percentage of new customers is around 32%, almost 100% are very active customers and the mean age is 41.
<b>Financing Buyer No Universitario</b>	37% have a high salary level. The percentage of new customers is around 26% and the mean age is 49 (higher than low and high salary no universitarios)
<b>High Salary No Universitario</b>	The salary level tends to be high, as expected. The percentage of new customers is around 40% and the mean age is 38 (higher than low salary no universitarios)

Further analysis related with the sales and churn behavior of each segment and the visualizations can be seen in the coding file.

Generally speaking, we can say that for most of the products, the Buyer Universitario has a higher percentage of customers that bought more than one product.

It's very clear for *debit\_cards*, *pension\_plan* and *payroll* that the percentage of Buyer Universitarios that bought more than one of such products is higher when compared to the other segments.

The behavior for Low and High Salary Universitario is not so different, but, in general, the percentage of customers with a high salary that bought more than 1 product (as *payroll\_account*, *payroll*...) is strictly higher compared to the low salary.

Concerning the 4 segments of NUNI, Comparing between the low and high salary customers groups, one can say that the trend is similar to the observed in the groups of Low/High Universitarios.

Regarding the Saving Buyer No Universitarios, as expected, the percentage of customers who buy saving products is high, especially for *pension\_plan*, with more than 40% buying 1 *pension\_plan* and almost 20% buying 2 of such products. Those customers also tend to buy *payroll\_accounts* and *payrolls*.

Regarding the Financing Buyer No Universitarios, as expected, the percentage of customer buying *funds* and *securities* is high compared with other groups. Those customers also tend to buy *payroll\_accounts*, *payrolls* and *emc\_accounts*.

## Conclusions

After segmenting the customers in 7 distinct groups according to their segment, antiquity, salary level and sales behavior in terms of product type, we can proceed to the next task. The main objective is to analyze those groups and understand what groups of interest are associated with each product type, so we can have a clearer idea about the response rate of each segment when they receive the advertisement email, and based on that, choose the right customers for the right products.

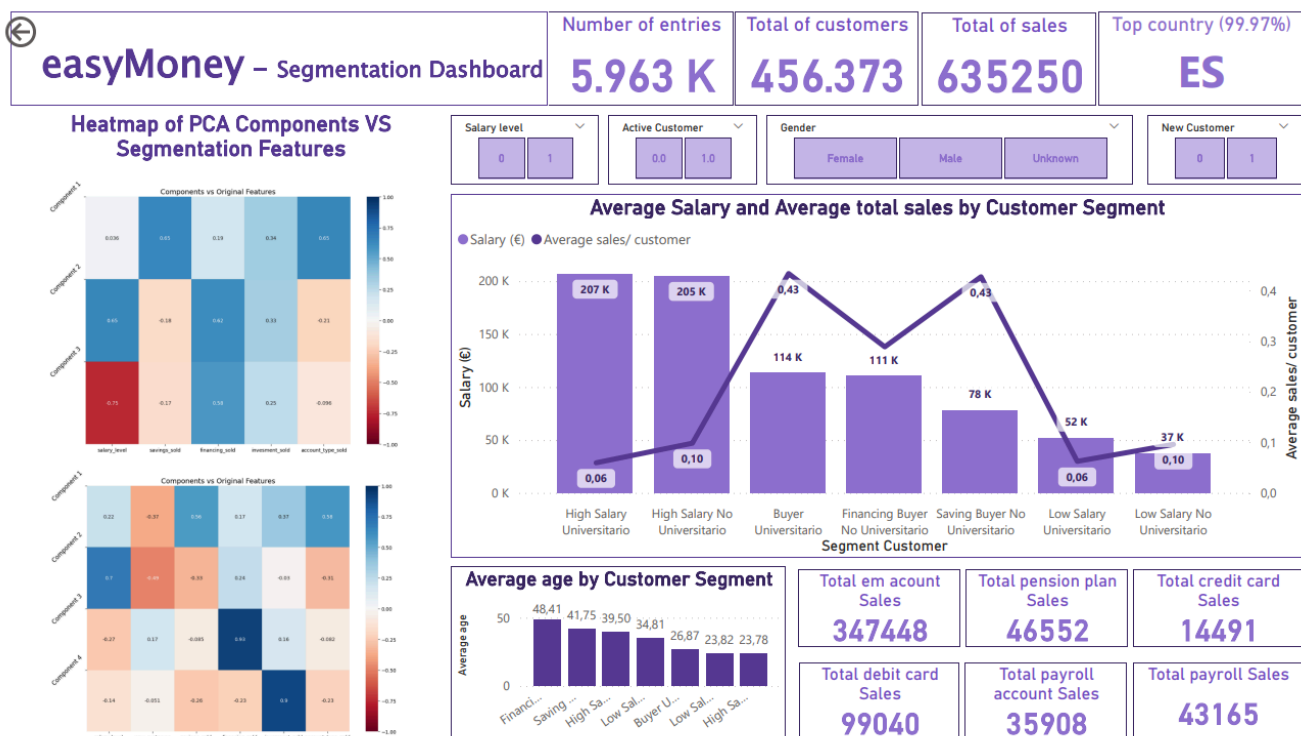


Fig. 3.3 - Dashboard for Segmentation.

Below the names and description of each segment to consider:

- ❖ **Low Salary Universitario:** customers from segment UNIVERSITARIO whose salary is classified as low level salary;
- ❖ **High Salary Universitario:** customers from segment UNIVERSITARIO whose salary is classified as high level salary;
- ❖ **Buyer Universitario:** customers from segment UNIVERSITARIO with the highest amount of sales;
- ❖ **Low Salary No Universitario:** customers from segments TOP, PARTICULAR and UNKNOWN (all except UNIVERSITARIO) whose salary is classified as low level salary;
- ❖ **High Salary No Universitario:** customers from segments TOP, PARTICULAR and UNKNOWN (all except UNIVERSITARIO) whose salary is classified as high level salary;
- ❖ **Savings Buyer No Universitario:** customers from segments TOP, PARTICULAR and UNKNOWN (all except UNIVERSITARIO) with the highest amount of saving products sales;
- ❖ **Financing Buyer No Universitario:** customers from segments TOP, PARTICULAR and UNKNOWN (all except UNIVERSITARIO) with the highest amount of financing products sales.

## 4. Task 3 - Target customers recommendation

### 4.1 Analysis of 2018/2019 revenues

To perform the analysis of the revenue from January 2018 to May 2019, we considered that the revenue of each product type is the same as mentioned by the Direct Marketing Director. The revenue, in Euros (€) for each sold product is as shown in table 4.1.

Table 4.1 - Revenue, in Euros (€), for each sold product.

Revenue	
Savings	40
Financing	60
Investment	40
Account Type	10

Table 4.2 shows the percentage of the number of customers in each segment compared with the total customers.

Table 4.2 - Analysis of some features for all the segments created.

Customer Segment	Percentage
Low Salary Universitario	38.9
Buyer Universitario	2.0
High Salary Universitario	14.1
Low Salary No Universitario	29.3
Saving Buyer No Universitario	6.1
Financing Buyer No Universitario	0.8
High Salary No Universitario	8.8

Table 4.3 shows the overall of the total revenue percentages for each product type that was associated with each one of the segments.

Table 4.3 - Overall of the total revenue percentages for each product type that was associated with each one of the segments.

	High Univ	Low Univ	Buyer Univ	Low NUniv	High NUniv	Saving NUniv	Financing NUniv	Total
Savings	1.3	1.9	20.0	7.7	8.3	57.8	3.1	100.1
Financing	1.0	1.0	8.7	0.0	0.0	0.0	89.3	100.0
Investment	0.5	0.8	13.8	7.2	6.2	65.3	6.2	100.0
Account Type	8.9	25.9	6.8	29.2	8.2	19.7	1.4	100.1
All Products	6.0	16.9	11.0	21.0	7.9	32.8	4.5	100.1
% Customers	14.1	38.9	2.0	29.3	8.8	6.1	0.8	100.0

More detailed information concerning each individual segment can be seen in the coding file, but the data we have and obtained, so far, one can say:

- ❖ Regarding the **High Salary Universitarios**, 6% of the total revenue came from this group, and 93.1% of those 6% came from Account type products. For that segment, the remaining products don't have importance in the total revenue. That segment has 14% of the total customers.
- ❖ Regarding the **Low Salary Universitarios**, 16.9% of the total revenue came from this group, and 96.3% of those 16.9% came from Account type products. The remaining products don't have importance in the total revenue, especially if we think that segment has 39% of the total customers (*easymoney* doesn't earn so much per customer)
- ❖ Regarding the **Buyer Universitarios**, 11% of the total revenue came from this group, and 50.7% of those 11% came from Saving products. Account type products also have some importance in the total revenue. The most interesting part related with those customers is that those 11% revenue is made by just 2% of total customers.
- ❖ Regarding the **Low Salary No Universitarios**, 21% of the total revenue came from this group, and 87.5% of those 21% came from Account type products. 7.7% of the total savings revenues came from that segment and it's the main reason for the percentage of total revenue being higher than the Low Salary Universitario segment.
- ❖ Regarding the **High Salary No Universitarios**, 8% of the total revenue came from this group, and 65.6% of those 8% came from Account type products. Compared to the High Salary Universitarios, the percentage of revenue is high if we think that just 8.8% of customers belong to such a segment, while 14.1% belong to the Universitario.
- ❖ Regarding the **Saving Buyer No Universitarios**, 33% of the total revenue came from this group, and 49% of those 33% came from Savings products. This is an interesting group because only 6.1% of total customers generated 33% of the revenue. This is clearly because those clients bought expensive products (40€ each).
- ❖ Regarding the **Financing Buyer No Universitarios**, 4.5% of the total revenue came from this group, and 53% of those 4.5% came from Financing products. This is also an interesting group because only 0.8% of total customers generated 4.5% of the revenue. This is clearly because those clients bought the most expensive products (60€ each), even if the group is composed of just 0.8% of total customers. Each sale on this segment means 6x more revenue compared to the groups that buy account products.

## 4.2 Gradient Boosting Classifier

Regarding the recommendation task, we decided to analyze the problem by looking for the sales for each client segment and to perform a Gradient Boosting Classifier to perform predictions related with the decision of each client to buy/not buy the product.

So, to perform this approach we decided to target the sales of each type of product on the model and predict whose clients will or won't buy the product.

Regarding the data transformation, we decide to create a target column in each model `{*product_type*}_buy` (ie, *savings\_buy*, *financing\_buy*, etc) which is 0 if the client never bought that type of product and 1 if the client bought at least one product.

Then there were created dummies to replace the *Customer Segment* feature.

The model was trained with 70% of the data and 30% was saved to perform the test evaluation, and the implementation can be seen in the coding file.

Below we can see the metrics obtained after each model.

### 4.2.1 Target: Saving Products

Table 4.4 - Confusion matrix for training (a) and test datasets.

Train Dataset		Test Dataset	
289629	1647	124074	659
10550	17635	4474	7705

Table 4.5 - Accuracy, precision and recall for train and test datasets.

Dataset	Accuracy	Precision	Recall
Train	0.96	0.91	0.63
Test	0.96	0.92	0.63

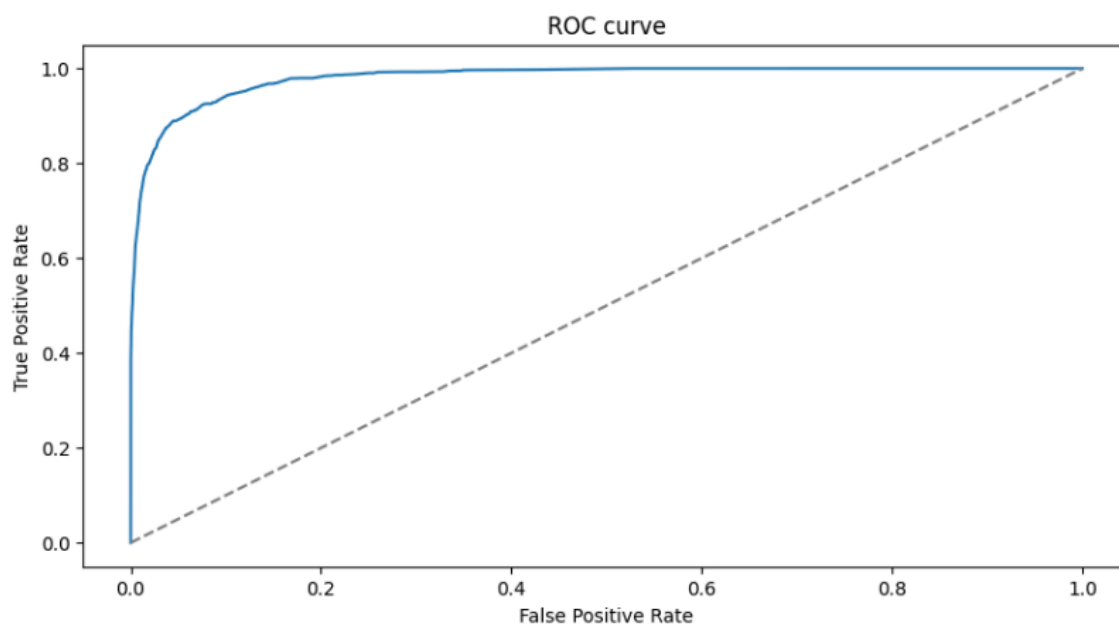


Fig. 4.1 - ROC curve of Savings Model (test dataset)

### 4.2.2 Target: Financing Products

Table 4.6 - Confusion matrix for training (a) and test datasets.

Train Dataset		Test Dataset	
316963	29	135828	17
277	2192	108	959

Table 4.7 - Accuracy, precision and recall for train and test datasets.

Dataset	Accuracy	Precision	Recall
Train	0.99	0.98	0.89
Test	0.99	0.98	0.90



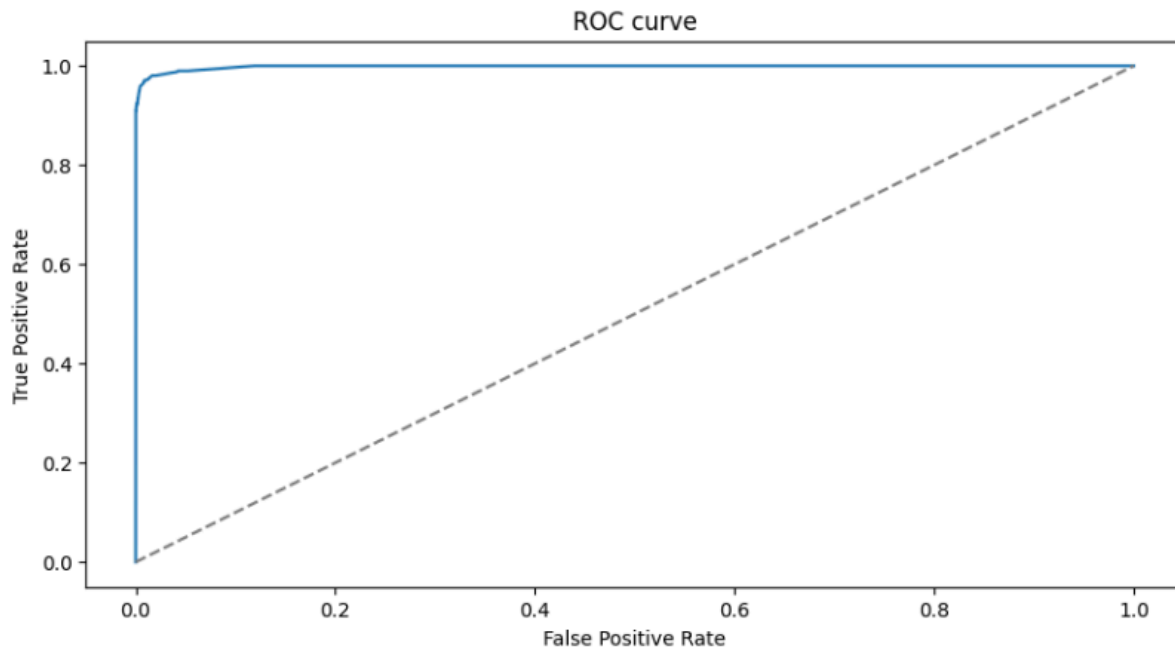


Fig. 4.2 - ROC curve of Financing Model (test dataset)

### 4.2.3 Target: Investment Products

Table 4.8 - Confusion matrix for training (a) and test (b) datasets.

Train Dataset		Test Dataset	
313386	127	134245	52
5388	560	2369	246

Table 4.9 - Accuracy, precision and recall for train and test datasets.

Dataset	Accuracy	Precision	Recall
Train	0.98	0.81	0.09
Test	0.98	0.83	0.09

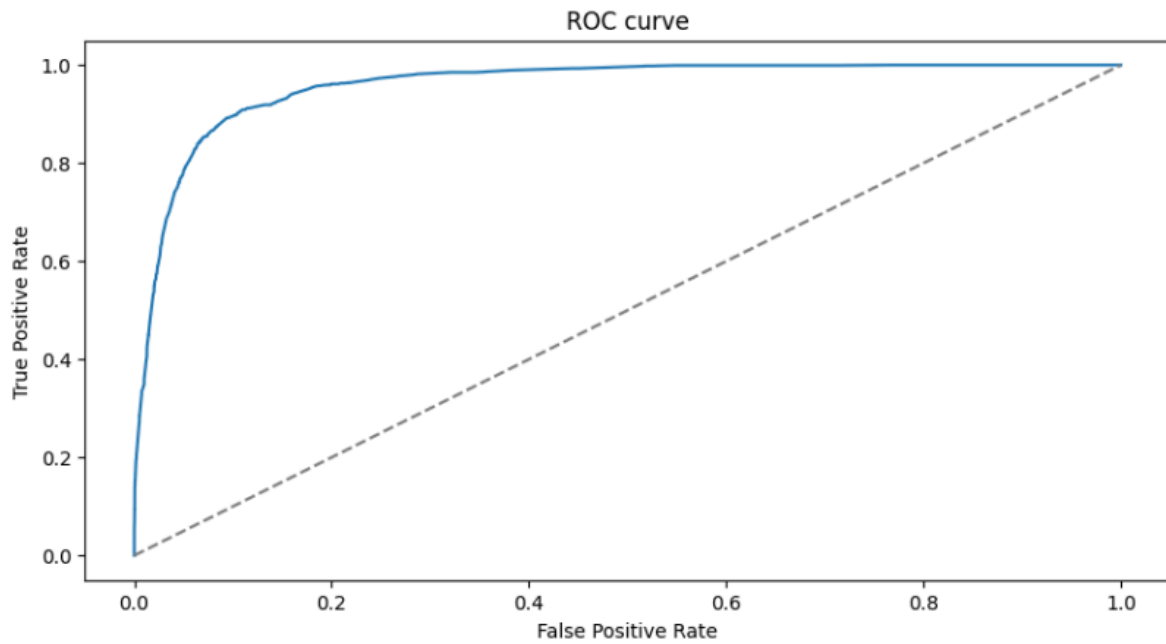


Fig. 4.3 - ROC curve of Investment Model (test dataset)

#### 4.2.4 Target: Account Type Products

Table 4.10 - Confusion matrix for training (a) and test datasets.

Train Dataset		Test Dataset	
23030	52274	9883	22372
7330	236727	3183	101474

Table 4.11 - Accuracy, precision and recall for train and test datasets.

Dataset	Accuracy	Precision	Recall
Train	0.81	0.82	0.97
Test	0.81	0.82	0.97

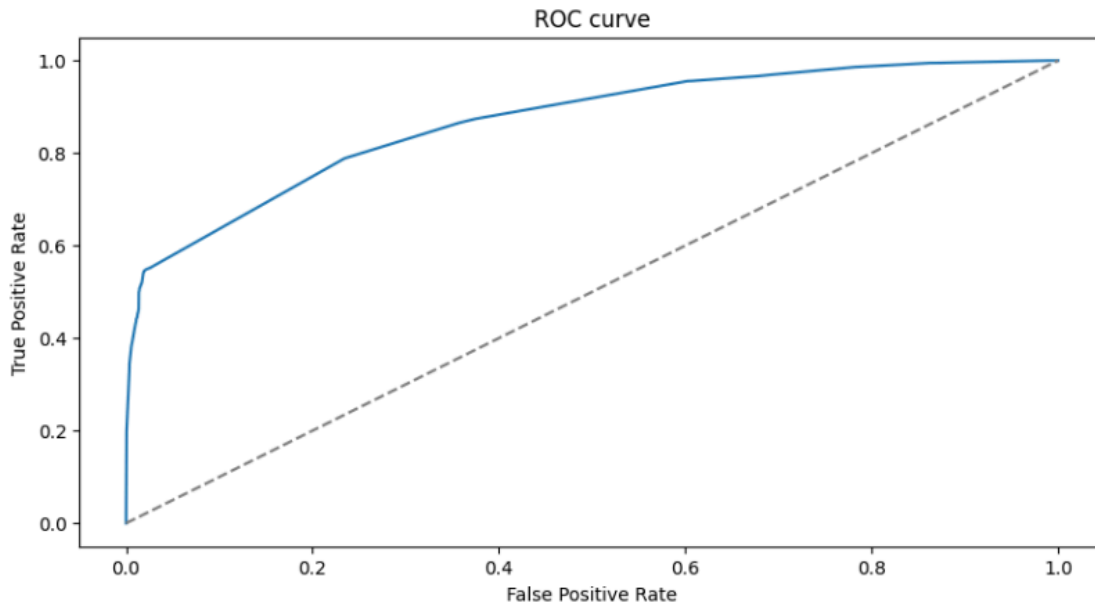


Fig. 4.4 - ROC curve of Account type Model (test dataset)

As we can observe before, the overall performance of all models is good. The worst accuracy was obtained in the Account type model and the Investment model didn't achieve a good recall value.

### 4.3 Expected response rate

In the end, we made an estimation of sales for each segment based on the 4 models performed, which can be seen in table 4.5.

Table 4.5 - Expected positive responses for each segment (if the customer buys or doesn't buy).

	Low Salary Universitario	High Salary Universitario	Buyer Universitario	Low Salary No Universitario	High Salary No Universitario	Saving Buyer No Universitario	Financing Buyer No Universitario
savings_sold	0.0	0.0	6862.0	6.0	43.0	20652.0	83.0
financing_sold	3.0	1.0	24.0	0.0	0.0	21.0	3148.0
investment_sold	0.0	0.0	19.0	0.0	0.0	966.0	0.0
account_type_sold	145546.0	54661.0	7176.0	140783.0	36953.0	24670.0	3158.0

By looking at table 4.4, and according to our models, one can conclude that, if we send an email advertising each of the 4 product types for each segment, just 3 Low Salary Universitarios will have a positive response to financing products.

That response is, by far, higher if we refer to account type products, ie, we expect that around 145 thousand customers will respond positively.

Some interesting information is related to the fact that, according to the savings model, we expect that 6,862 Buyer Universitarios will have a positive response and 20,652 Saving Buyer No Universitario will respond positively to the campaign.

The same conclusions can be taken if we talk about Financing Buyer No Universitarios, a segment with almost 3,500 customers that are expected to have a positive response.

Below we have table 4.5 in terms of expected positive response rate expected to the email campaign for each type of product.

Table 4.5 - Expected response rate for each segment.

	Low Salary Universitario	High Salary Universitario	Buyer Universitario	Low Salary No Universitario	High Salary No Universitario	Saving Buyer No Universitario	Financing Buyer No Universitario
savings_sold	0.0	0.0	0.956243	0.000033	0.001101	0.83696	0.026282
financing_sold	0.000021	0.000018	0.003344	0.0	0.0	0.000851	0.996833
investment_sold	0.0	0.0	0.002648	0.0	0.0	0.039149	0.0
account_type_sold	1.0	1.0	1.0	0.773061	0.946397	0.999797	1.0

According to our model, one can say that practically all the segments have a 100% positive response for all customers. The only exception is the *Low Salary No Universitario* Segment. Regarding the remaining products, *Low* and *High Salary Universitario* segments won't have a positive response, as well as the *Low* and *High Salary No Universitario* segments. Also 96% of *Buyer Universitario* and 84% of *Saving Buyer No Universitario* are expected to have a positive response to saving products. The high rate for financing products is also in that last segment.

The *Financing Buyer No Universitario* are expected to have almost 100% of positive responses.

## 4.4 Customers recommendations

Marketing team will send 10000 emails to the customers, so they are expecting to have a recommendation with respect to the customers segments that we should prioritize to send the emails.

In that sense, our team decided to “translate” the response rates to revenues per sale expected in each segment, because all the products have distinct revenues.

In that sense, according to the values in table 4.5, it is expected to have the revenues per email sent as shown in table 4.6.

Table 4.6 - Expected revenues per email sent considering the different segments.

Segment	Expected revenue per sent email (€)
<b>Low Salary Univ</b>	$0 \times 40 + 2 \times 10^{-6} \times 60 + 0 \times 40 + 1 \times 10 \approx 10$
<b>High Salary Univ</b>	$0 \times 40 + 1.8 \times 10^{-6} \times 60 + 0 \times 40 + 1 \times 10 \approx 10$
<b>Buyer Univ</b>	$0.956 \times 40 + 3 \times 10^{-3} \times 60 + 2.6 \times 10^{-3} \times 40 + 1 \times 10 \approx 48.56$
<b>Low Salary No Univ</b>	$3 \times 10^{-5} \times 40 + 0 \times 60 + 0 \times 40 + 0.773 \times 10 \approx 7.73$
<b>High Salary No Univ</b>	$10^{-3} \times 40 + 0 \times 60 + 0 \times 40 + 0.946 \times 10 \approx 9.46$
<b>Savings Buyer No Univ</b>	$0.837 \times 40 + 8.5 \times 10^{-4} \times 60 + 4 \times 10^{-2} \times 40 + 1 \times 10 \approx 45.09$
<b>Financing Buyer No Univ</b>	$2.6 \times 10^{-3} \times 40 + 0.997 \times 60 + 0 \times 40 + 1 \times 10 \approx 70.87$

So, it is easy to understand that we can rank the top most rentable segments per email sent as shown in table 4.7.

Table 4.7 - Top 3 most rentable customer segments (by order from top to bottom).

<b>Financing Buyer No Universitario</b>
<b>Buyer Universitario</b>
<b>Savings Buyer No Univ</b>
<b>Low Salary Univ</b>
<b>High Salary Univ</b>
<b>High Salary No Univ</b>
<b>Low Salary No Univ</b>

So, we should prioritize those segments when the time comes to send the 10000 emails. To confirm our results, our team implemented a simple Linear Programming to maximize the expected revenue.

Below one can see the definition of the Problem:

#### 4.4.1 Linear Programming

$$\max z = 10a + 10b + 48.56c + 7.73d + 9.46e + 45.04f + 70.81g$$

$$a + b + c + d + e + f + g = 10000$$

$$a, b, c, d, e, f, g \geq 0$$

$$a \leq 145546, b \leq 54661, c \leq 7176, d \leq 182111, e \leq 39046, f \leq 24675, g \leq 3158$$

$a \rightarrow$  number of low salary universitario

$b \rightarrow$  number of high salary universitario

$c \rightarrow$  number of buyer universitario

$d \rightarrow$  number of low salary no universitario

$e \rightarrow$  number of high salary no universitario

$f \rightarrow$  number of saving buyer no universitario

$g \rightarrow$  number of financing buyer no universitario

After running the Linear Programming, the solutions just confirmed what we had already concluded: we should send the email to the 3158 *Financing Buyer No Universitario* customers and the remaining 6842 to the *Buyer Universitario* customers. The revenue expected would be 556002.20 €.

## 4.5 Incremental revenue

Above we estimated expected revenue if the business decides to follow our recommendations, which were based on the results of a trained Data Science Model, so it's important to show the importance of considering "smart" models to perform predictions.

In that sense, and using a weighted mean customers response estimation for each product type, we pretend to show the *incremental revenue* comparing the six scenarios shown in table 4.8, with the scenario in which our model is used.

*Incremental revenue* is the profit a business gains from an increase in sales. It can be used to determine the additional revenue generated by a certain product, investment or direct sale from a marketing campaign when the quantity of sales has grown.

Table 4.8 - Alternative random scenario of choosing the customers to send the emails .

Scenario	Estimated response rate (%)	Expected revenue/ ER (€)	Incremental revenue/ IR (€)*
1: Advertising randomly saving products	7.04	$ER = 7.04\% \times 10000 \times 40 = 28,160$	$IR \approx 527842.20$ $(\frac{ER^{ML\ Model}}{ER} \approx 20)$
2: Advertising randomly financing products	0.8	$ER = 0.8\% \times 10000 \times 60 = 4800$	$IR \approx 551400.20$ $(\frac{ER^{ML\ Model}}{ER} \approx 116)$
3: Advertising randomly investment products	0.2	$ER = 0.2\% \times 10000 \times 40 = 800$	$IR \approx 555202.20$ $(\frac{ER^{ML\ Model}}{ER} \approx 695)$
4: Advertising randomly account types products	92.8	$ER = 92.8\% \times 10000 \times 10 = 92800$	$IR \approx 463202.20$ $(\frac{ER^{ML\ Model}}{ER} \approx 6)$
5: Advertising randomly all products (25% each)		$ER^{sav} = 7.04\% \times 2500 \times 40 = 7040$ $ER^{fin} = 0.8\% \times 2500 \times 60 = 1200$ $ER^{inv} = 0.2\% \times 2500 \times 40 = 200$ $ER^{act} = 92.8\% \times 2500 \times 10 = 23200$ $ER^{all} = 7040 + 1200 + 200 + 23200 = 31640$	$IR \approx 524362.20$ $(\frac{ER^{ML\ Model}}{ER} \approx 18)$
6: Advertising specific products for the top rentable segments (Savings for Buyer Univ ( $ER^{sav}$ ) and Financing for Financing Buyer No Univ ( $ER^{fin}$ ))		$ER^{sav} = 95.62\% \times 6842 \times 40 = 261692.82$ $ER^{fin} = 99.68\% \times 3158 \times 60 = 188873.66$ $ER^{all} = 261692.82 + 188873.66$	$IR \approx 450566.48$ $(\frac{ER^{ML\ Model}}{ER} \approx 1.2)$

\*  $IR = ER^{ML\ Model} - ER^{sav/fin/inv/act/all}$  (€)

## Conclusions

By looking for the Incremental revenues in percentage, it's clear that we have a significant increment on the expected revenues if we consider the scenario 1 to 5 and if we consider the recommendations we made before. Even in the scenario with the best expected revenue (forth one), the revenue is practically 6 times lower compared with the expected revenue of using Machine Learning.

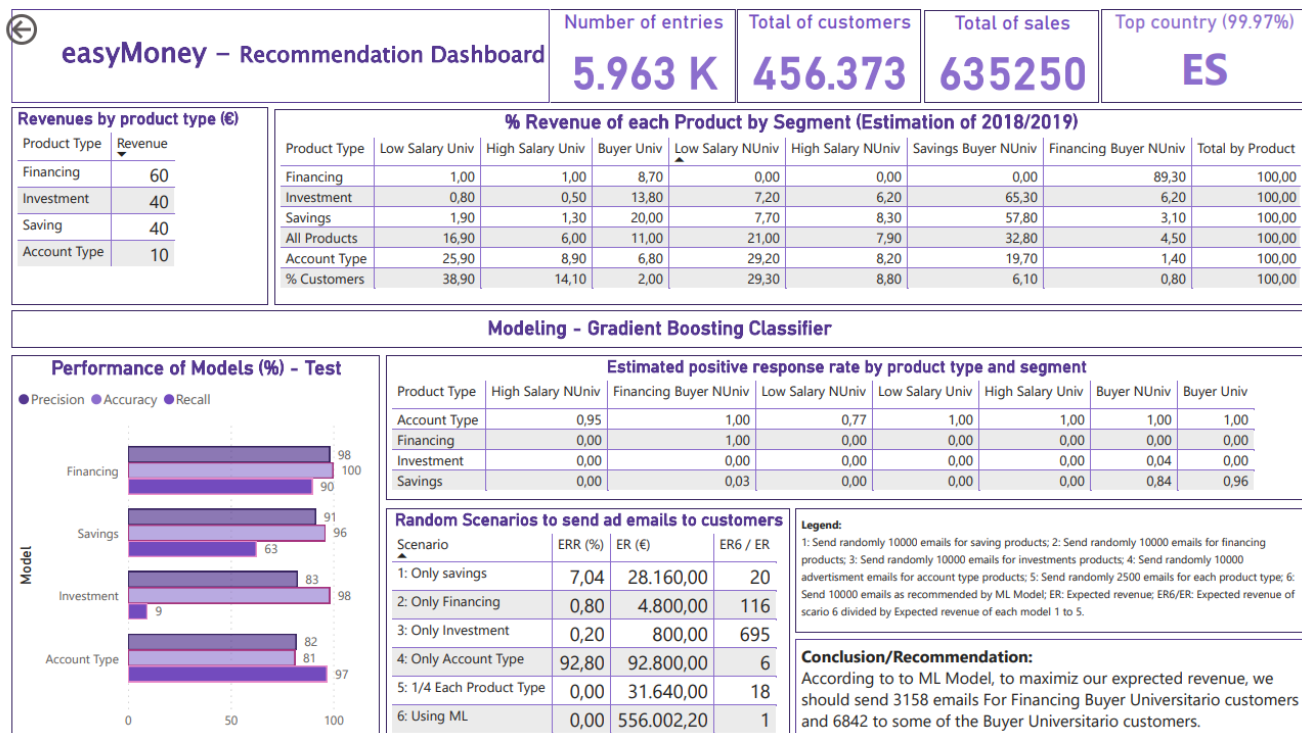


Fig. 4.5 - Dashboard for Recommendation.

## 5. Task 4 - Monitoring the process

Regarding the Monitoring of the process, the team were asked to define some KPIs to perform the task. After discussing, we decided that it should be interesting to look for the KPIs below (formulas in the appendix 8.2):

- ❖ **Open Rate:** The percentage of emails opened relative to the total number of emails sent, indicating the initial level of interest from customers in the advertising messages.
- ❖ **Click-through Rate (CTR):** The percentage of customers who click on the links included in the emails relative to the total number of emails sent, indicating the level of engagement with the email content.
- ❖ **Conversion Rate:** The percentage of customers who take a desired action after clicking on the email link, such as making a purchase or filling out a form, relative to the total number of emails sent, indicating the success in converting interest into action.
- ❖ **Response Rate:** The percentage of customers who directly respond to the advertising emails, for example, by sending a reply email or contacting the company in another way, indicating the level of engagement and interest from customers.
- ❖ **Specific Product Conversion Rate:** The percentage of customers who take a desired action related to a specific product after clicking on the email link, indicating the specific interest of customers in certain products and the effectiveness of the advertising campaign segmentation.


 <b>easyMoney – Monitoring Dashboard</b>	<b>Number of entries</b> <b>5.963 K</b>	<b>Total of customers</b> <b>456.373</b>	<b>Total of sales</b> <b>635250</b>	<b>Top country (99.97%)</b> <b>ES</b>
<b>Suggested KPIs to monitor the project</b>				
$\text{Open rate} = \frac{\text{Total No. opened emails}}{\text{Total No. delivered emails}} \times 100 \quad \textbf{(a)}$				
$\text{Click through rate} = \frac{\text{Total No. clicks}}{\text{Total No. impressions}} \times 100 \quad \textbf{(b)}$				
$\text{Conversion rate} = \frac{\text{Total No. conversions}}{\text{Total No. visitors}} \times 100 \quad \textbf{(c)}$				
$\text{Response rate} = \frac{\text{Total No. email direct response}}{\text{Total No. delivered emails}} \times 100 \quad \textbf{(d)}$				
$\text{Specific product conversion rate} = \frac{\text{Total No. conversions of a specific product}}{\text{Total No. visitors/ delivered emails for a specific product}} \times 100 \quad \textbf{(e)}$				

Fig. 5.1 - Dashboard for Monitoring.



## 6. Task 5 - Coordination and Planning

Our team was asked to create a plan to coordinate the team and tasks. After analyzing the possibilities, we decided to use Trello, a web-based project management tool that utilizes a system of boards, lists, and cards to help us organize and prioritize tasks in a visual and collaborative way.

Each board represents a project, and within each board, users can create lists to represent different stages or categories of tasks. Cards, which represent individual tasks or items, can then be moved between lists to indicate progress or changes in status.

One of the main advantages of Trello is its simplicity and ease of use, allowing us to quickly set up and customize boards to suit their specific workflow. Additionally, Trello promotes transparency and collaboration by providing a real-time overview of tasks and progress, facilitating communication and alignment among team members. Its flexibility, accessibility across devices, and integrations with other tools make Trello a popular choice for teams of all sizes looking to streamline project management processes.

Due to all those advantages, we found it to be a great option to perform that task.

The base for the planning was to create the a list of cards organized by the 7 Stages of a Data Science Project: **Problem Definition and Problem, Data Collection, Data Preparation, Data Analysis, Model Building, Model Evaluation and Model Deployment**, plus one list related to **Monitoring**, one to related to the **Report and Dashboards**, and a last one to reflect **POS (Point of Situations)** which can be used each time an element of the groups thinks the team needs to reunite and discuss some topic/issue.

Next we show some examples of the timeline, board card and Dashboard of our planning.

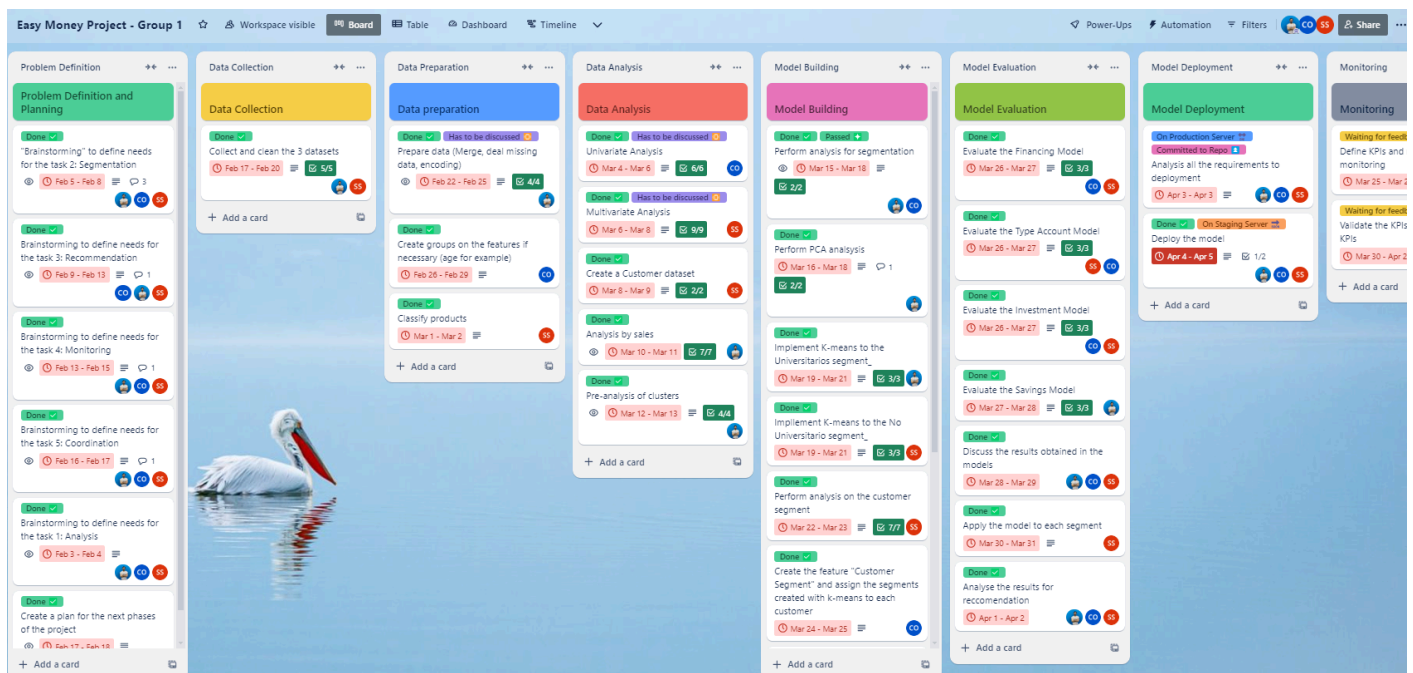


Fig. 6.1 - Card board showing all the cards and tasks to be performed by the team.

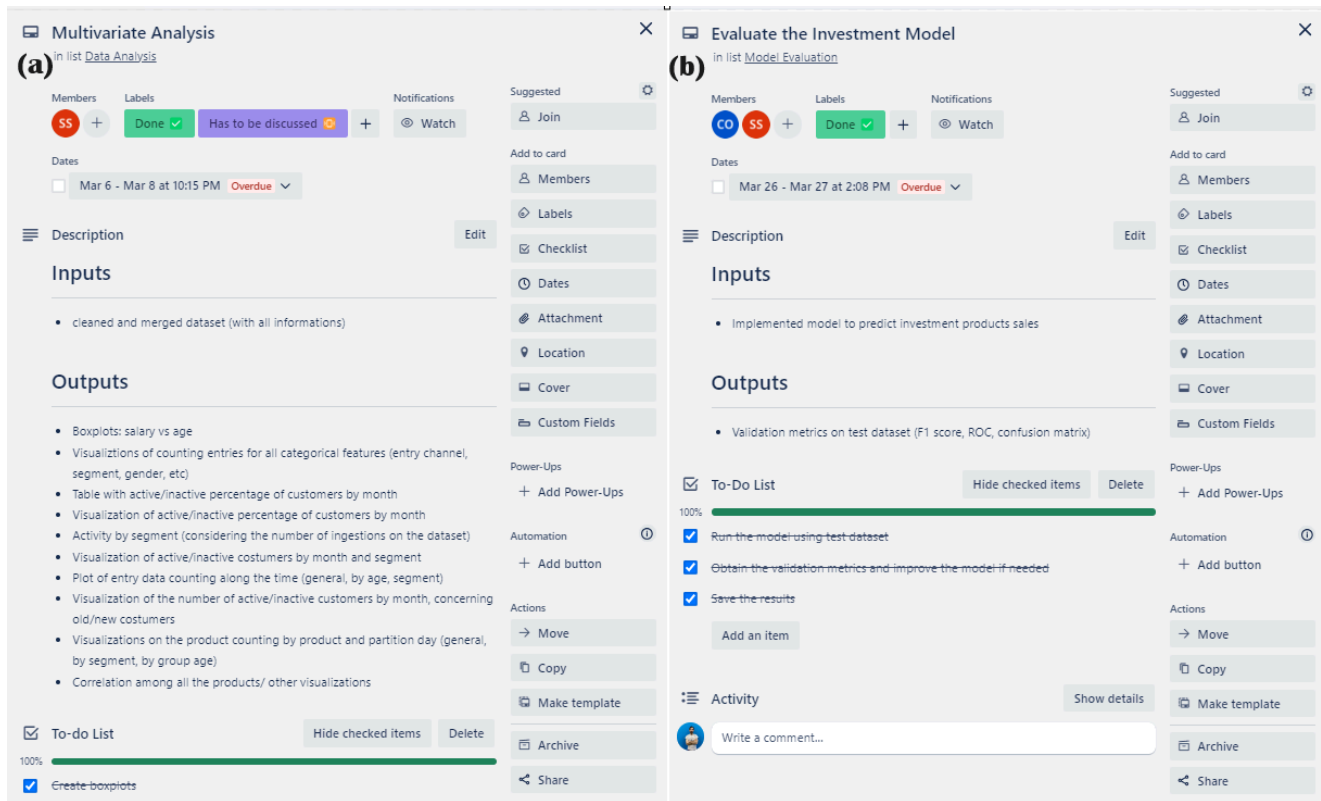


Fig 6.2 - Example of the detail for 2 cards: (a) Multivariate Analysis; (b) Evaluate The Investment Model.

Card	List	Labels	Members	Due date
Problem Definition and Planning	Problem Definition	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 17
"Brainstorming" to define needs for the task 2: Segmentation	Problem Definition	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 8
Brainstorming to define needs for the task 3: Recommendation	Problem Definition	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 13
Brainstorming to define needs for the task 4: Monitoring	Problem Definition	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 15
Brainstorming to define needs for the task 5: Coordination	Problem Definition	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 17
Brainstorming to define needs for the task 1: Analysis	Problem Definition	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 4
Create a plan for the next phases of the project	Problem Definition	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 18
Data Collection	Data Collection	Done ✓	SS, CO, PS	<input type="checkbox"/> Feb 21
Collect and clean the 3 datasets	Data Collection	Done ✓	SS, PS	<input type="checkbox"/> Feb 20
Data preparation	Data Preparation	Done ✓	SS, CO, PS	<input type="checkbox"/> Mar 2
Prepare data (Merge, deal missing data, encoding)	Data Preparation	Done ✓ Has to be discussed	PS	<input type="checkbox"/> Feb 25
Create groups on the features if necessary (age for example)	Data Preparation	Done ✓	CO	<input type="checkbox"/> Feb 29
Classify products	Data Preparation	Done ✓	SS	<input type="checkbox"/> Mar 2
Data Analysis	Data Analysis	Done ✓	SS, CO, PS	<input type="checkbox"/> Mar 13
Univariate Analysis	Data Analysis	Done ✓ Has to be discussed	CO	<input type="checkbox"/> Mar 6
Multivariate Analysis	Data Analysis	Done ✓ Has to be discussed	SS	<input type="checkbox"/> Mar 8
Create a Customer dataset	Data Analysis	Done ✓	SS	<input type="checkbox"/> Mar 9
Analysis by sales	Data Analysis	Done ✓	PS	<input type="checkbox"/> Mar 11
Pre-analysis of clusters	Data Analysis	Done ✓	PS	<input type="checkbox"/> Mar 13
Model Building	Model Building	Done ✓	SS, CO, PS	<input type="checkbox"/> Mar 29
Perform analysis for segmentation	Model Building	Done ✓ Passed	CO, PS	<input type="checkbox"/> Mar 18

Fig 6.3 - Part of the table that contains all the information related with cards, list, members assigned, status and due dates.



Fig 6.4 - Dashboard to manage the distribution of (from top to right) cards per list, the due dates, assignment among the team members and the status of each card.

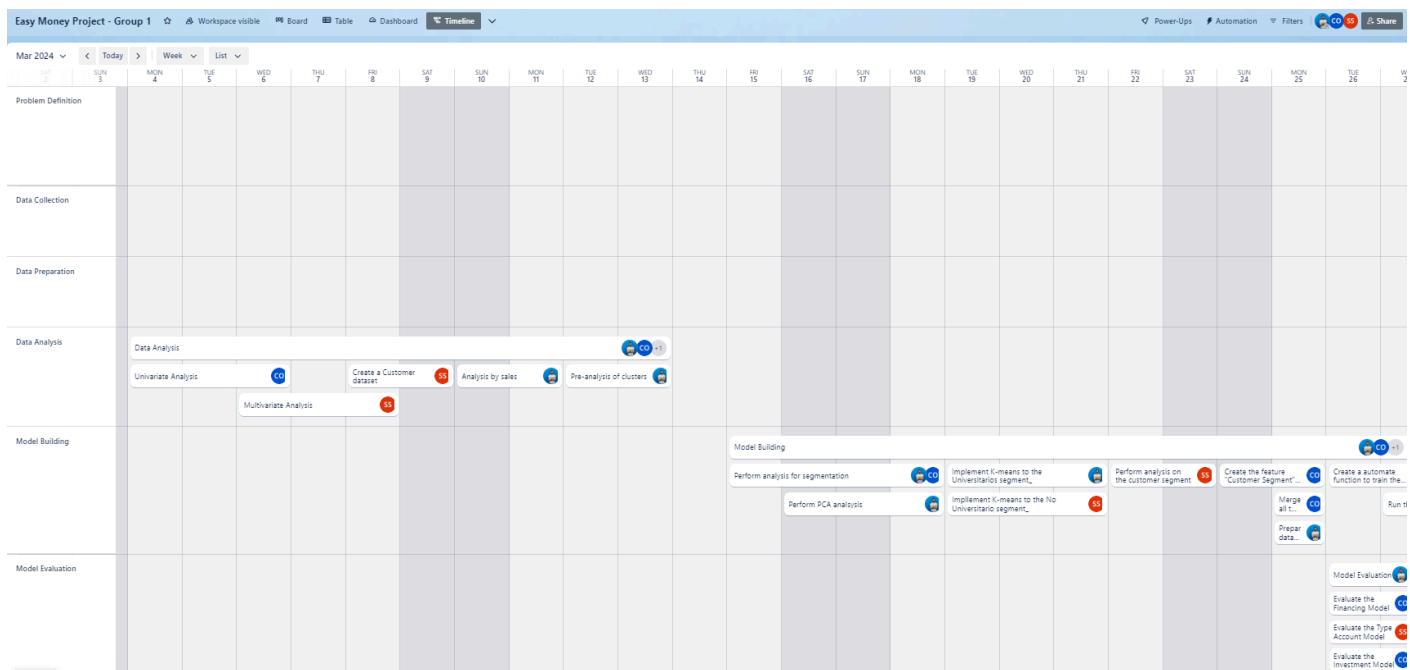


Fig 6.5 - Part of the timeline used to visually manage the due dates.

## 7. Conclusions/Further work

Performing the data analysis as explained before, allied to the implementation of K-Means to segmentation purposes and using a Gradient Boosting Classifier to determine specific response rates of customers related with an advertisement, have shown leading to a notable incremental revenue comparing to a few scenarios without consistent criteria for choosing the target customers.

In that sense, one of the main conclusions we can take from this project is that, with a proper exploration, analysis and treatment of data we can simulate and predict customers sales behaviors, and it would lead to a better planning when comes to choose the right targets, ie, the target that would be expected to have more probabilities to respond positively to some specific types of products and lead to an incremental revenue, which is some cases almost 700 times (see scenario 3, table 4.8).

Considering the dataset, one can say that, besides having some missing data, we believe that the information we have is enough to create a viables models with accuracies above 90%.

Regarding further work related with this project, apart from following the monitoring and measuring the KPIs mentioned in chapter 5, it would be interesting to perform some studies on the entry dates of the customers because we observed that the number of customers that signed their first contract follows a clear pattern, specially from segment Universitario. That pattern seems to coincide with the academic schedules. (We attached at the end of the report some time plot showing the pattern.)

That study could give us, for example, what would be the best days or months to contact the customers because if they tend to sign the contract on those days, they would be more available to respond positively to a campaign.

## 8. Appendixes

### 8.1 Entry dates time series

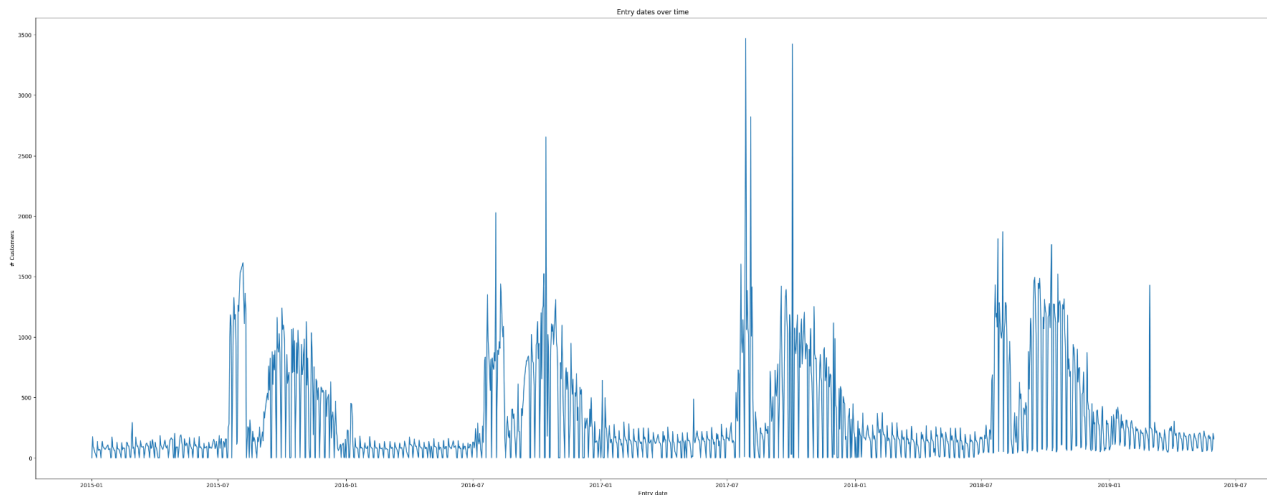


Fig 8.1 - Timeseries: Entry date vs total customers that signed first contract.

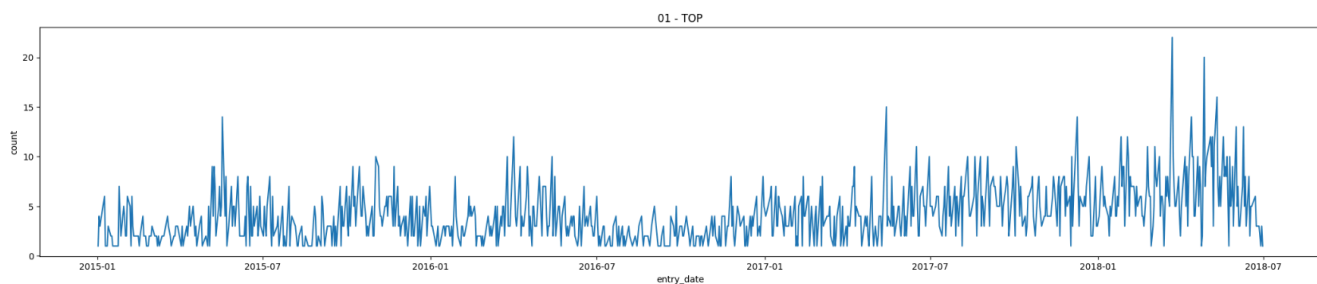


Fig 8.2 - Timeseries: Entry date vs total customers that signed first contract. (TOP segment)

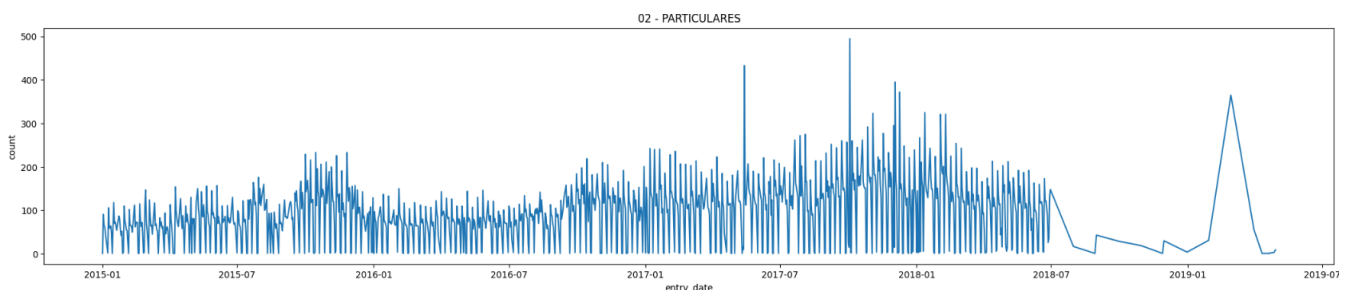


Fig 8.3 - Timeseries: Entry date vs total customers that signed first contract. (PARTICULARES segment)

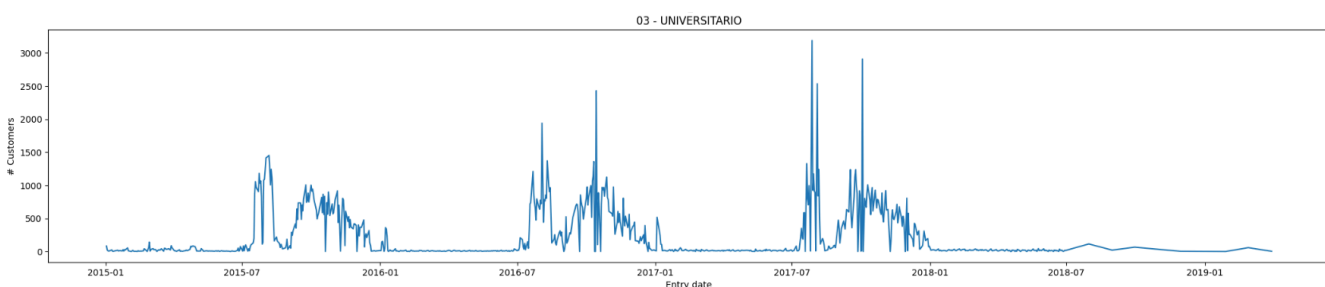


Fig 8.4 - Timeseries: Entry date vs total customers that signed first contract. (UNIVERSITARIO segment)

## 8.2 Data dictionary

### Data dictionary

Name	Table	Description
active_customer	commercial_activity	Client activity indicator in our application
age	sociodemographic	Customer age
country_id	sociodemographic	Country of residence of the client
credit_card	products	Credit cards
debit_card	products	Debit card
deceased	sociodemographic	Deceased index. N/S
em_account_p	products	easyMoney+ account
em_account_pp	products	easyMoney++ account
em_account	products	easyMoney account
emc_account	products	easyMoney Crypto account
entry_channel	commercial_activity	Customer acquisition channel
entry_date	commercial_activity	Date on which first easyMoney contract was signed
funds	products	Investment funds
gender	sociodemographic	Gender
loans	products	Loans
long_term_deposit	products	Long term deposits
mortgage	products	Mortgage
payroll	products	payroll
payroll_account	products	Account awarded with a bonus due to payroll
pension_plan	products	Pension plan
pk_cid	pk	Customer identifier
pk_partition	pk	Data ingestion date
region_code	sociodemographic	Customer's province of residence (for ES)
salary	sociodemographic	Household gross income
securities	products	Securities
segment	commercial_activity	Customer business segment
short_term_deposit	products	Short-term deposits

Fig 8.5 - Data dictionary.