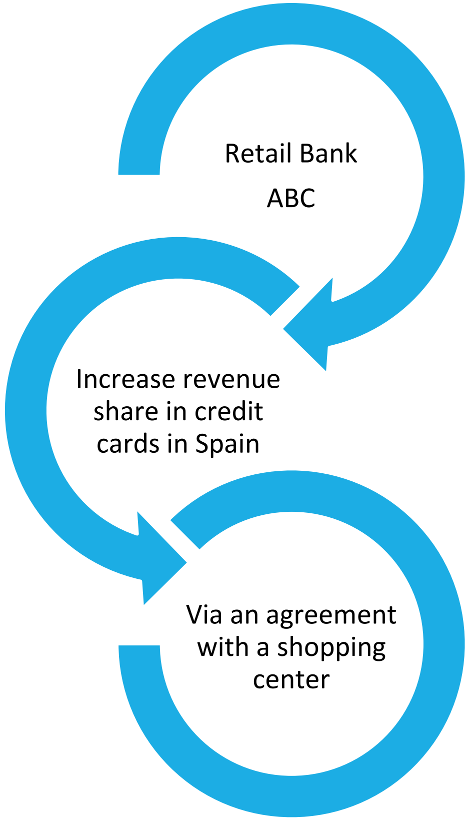
Knowledge Discovery Process

MODELLING CREDIT SCORING

PROFESSOR: ROBERTO COSTUMERO MORENO
Student: Amanda Marques

Business Problem


Retail Bank
ABC
Increase revenue share in credit cards in Spain
Via an agreement with a shopping center

The retail bank ABC wants to increase their revenue share in credit cards in Spain, via an agreement with a shopping center to sell credit cards via different channels.

The major business problems related to this goal is correctly scoring potential customers in order to set the correct risk and interest rate applied.

The reason why scoring is important relates to credit risk: the risk that a borrower (user of the credit card) defaults and does not honor its obligation to service debt. It can occur when the counterpart is unable to pay or cannot pay on time.

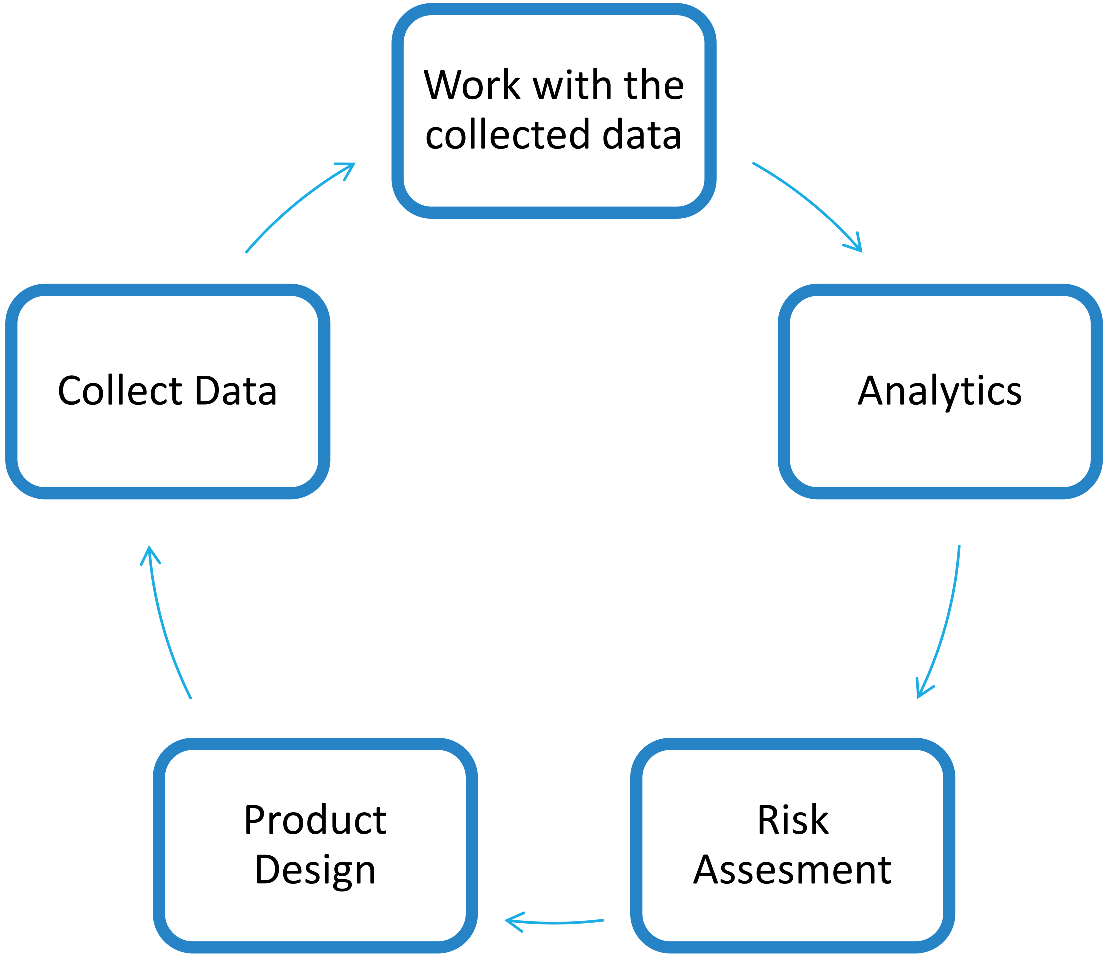
Credit risk is the most obvious risk of a bank by the nature of its activity and, in terms of potential losses, it is typically the largest type of risk

Having a poor scoring model for credit increase the risk of default.

CREDIT

How can it be addressed?

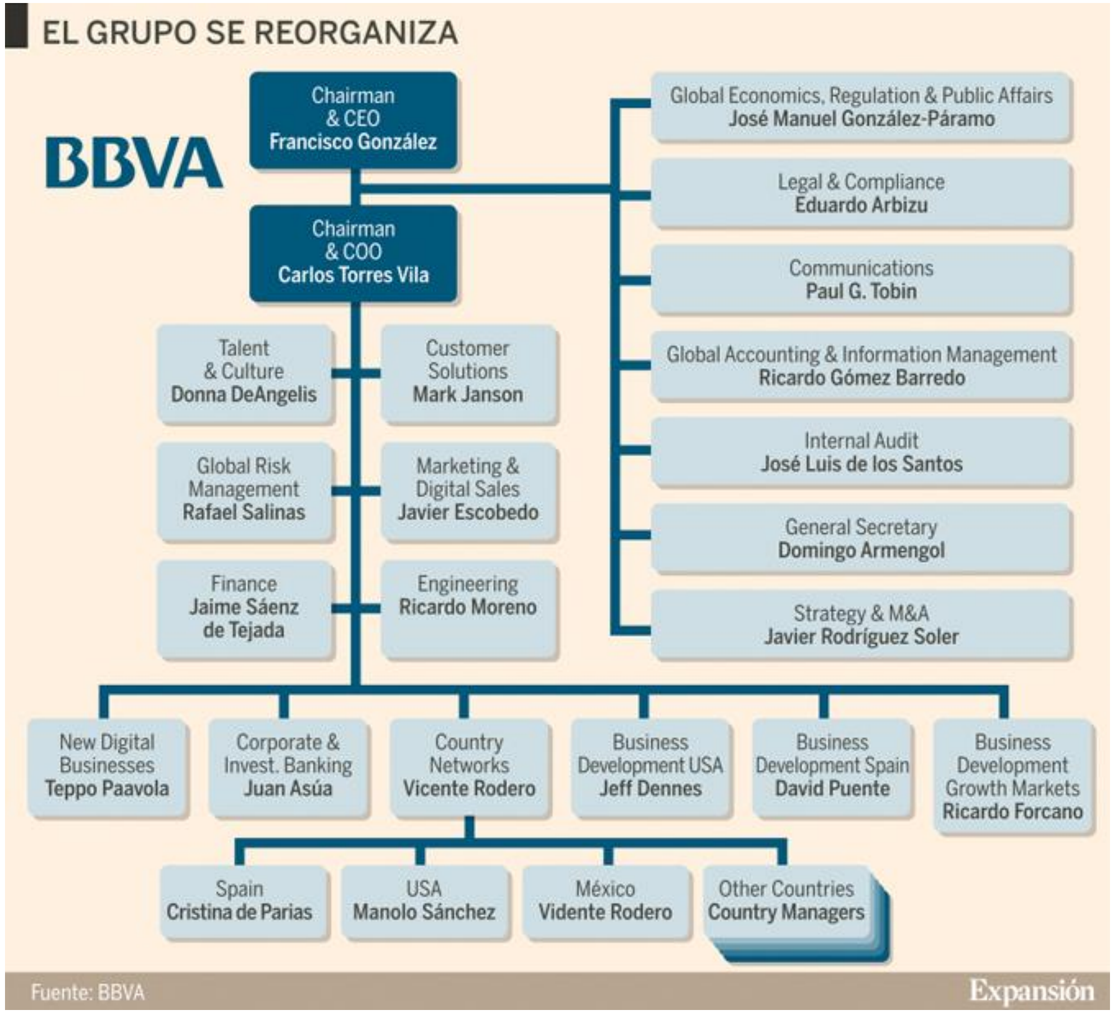
Addressing the risk credit should follow a process:



Start by analyzing the historical data already collected by cleaning and taking important insights out of it;
Later on, the department of Risk Assessment and Finance should work on designing a model that would be acceptable for the bank to offer the credit card;
Once there is an approved model by Finance, the department of Product Design and Manageement of the bank should be involved and work on a product that fits the bank's philosophy and strategy;
Finally, after the product is modeled and out in the market, it is crucial to keep collecting consistent data so the financial model can keep on being more precise and accurate.

Involved areas of the bank

Addressing the risk credit should follow a process:



EL GRUPO SE REORGANIZA

BBVA

Chairman & CEO Francisco González

Chairman & COO Carlos Torres Vila

Talent & Culture Donna DeAngelis

Customer Solutions Mark Janson

Global Risk Management Rafael Salinas

Marketing & Digital Sales Javier Escobedo

Finance Jaime Sáenz de Tejada

Engineering Ricardo Moreno

Global Economics, Regulation & Public Affairs José Manuel González-Páramo

Legal & Compliance Eduardo Arbizu

Communications Paul G. Tobin

Global Accounting & Information Management Ricardo Gómez Barredo

Internal Audit José Luis de los Santos

General Secretary Domingo Armengol

Strategy & M&A Javier Rodríguez Soler

New Digital Businesses Teppo Paavola

Corporate & Invest. Banking Juan Asúa

Country Networks Vicente Rodero

Business Development USA Jeff Dennes

Business Development Spain David Puente

Business Development Growth Markets Ricardo Forcano

Spain Cristina de Parias

USA Manolo Sánchez

México Vidente Rodero

Other Countries Country Managers

Fuente: BBVA

Expansión

BBVA organigram

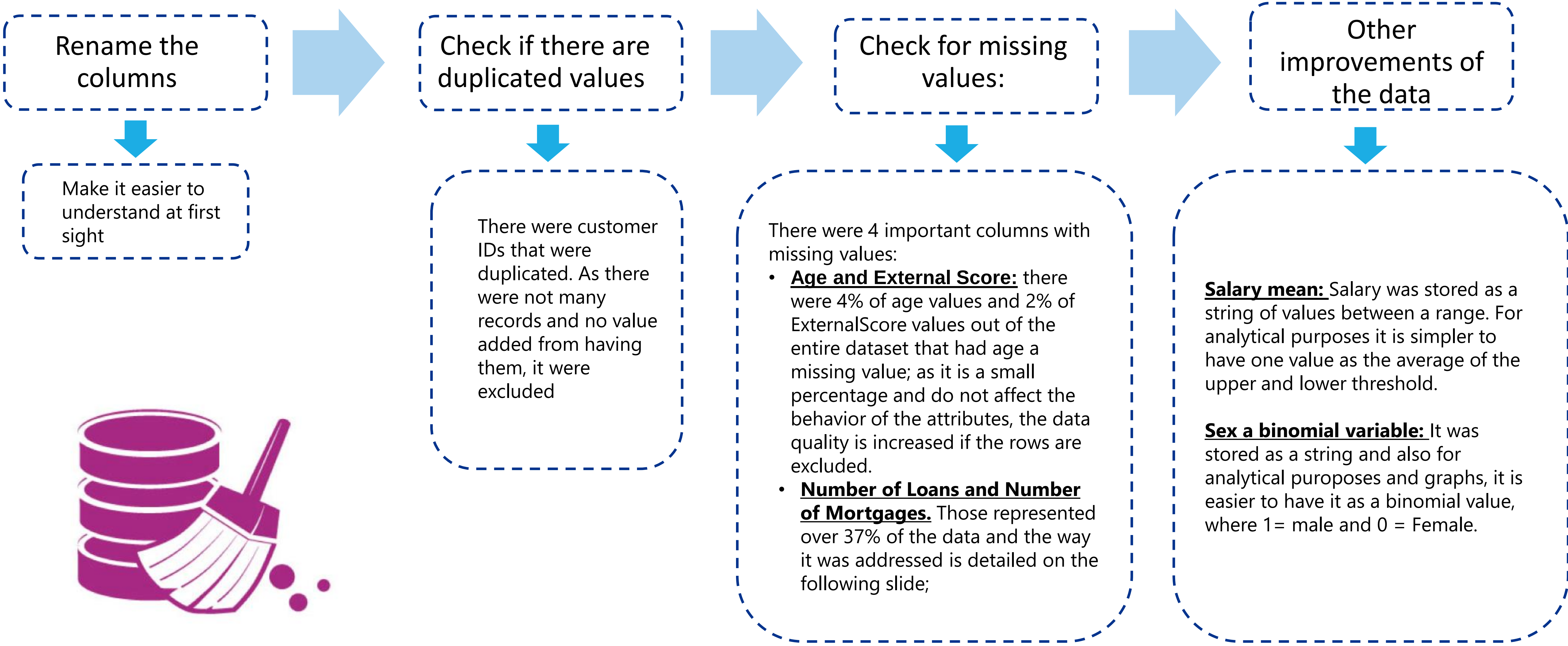
The areas of the bank that should be involved are Finance, Technology of Information and Product Designers.

The principle is to have a the most efficient and precise financial model that suits the market needs and can be applied to a product in accordance to the bank’s strategy.

In case of the Spanish bank BBVA represented on the organigram, the areas should be Finance, Global Risk Management and Business Development in different countries and markets.

Data cleaning and preparation

The historical data provided had some inconsistencies and it required to be cleaned and prepared to be able to input it on the final model:



Rename the columns

Make it easier to understand at first sight

Check if there are duplicated values

There were customer IDs that were duplicated. As there were not many records and no value added from having them, it were excluded

Check for missing values:

There were 4 important columns with missing values:

- **Age and External Score:** there were 4% of age values and 2% of ExternalScore values out of the entire dataset that had age a missing value; as it is a small percentage and do not affect the behavior of the attributes, the data quality is increased if the rows are excluded.
- **Number of Loans and Number of Mortgages.** Those represented over 37% of the data and the way it was addressed is detailed on the following slide;

Other improvements of the data

Salary mean: Salary was stored as a string of values between a range. For analytical purposes it is simpler to have one value as the average of the upper and lower threshold.

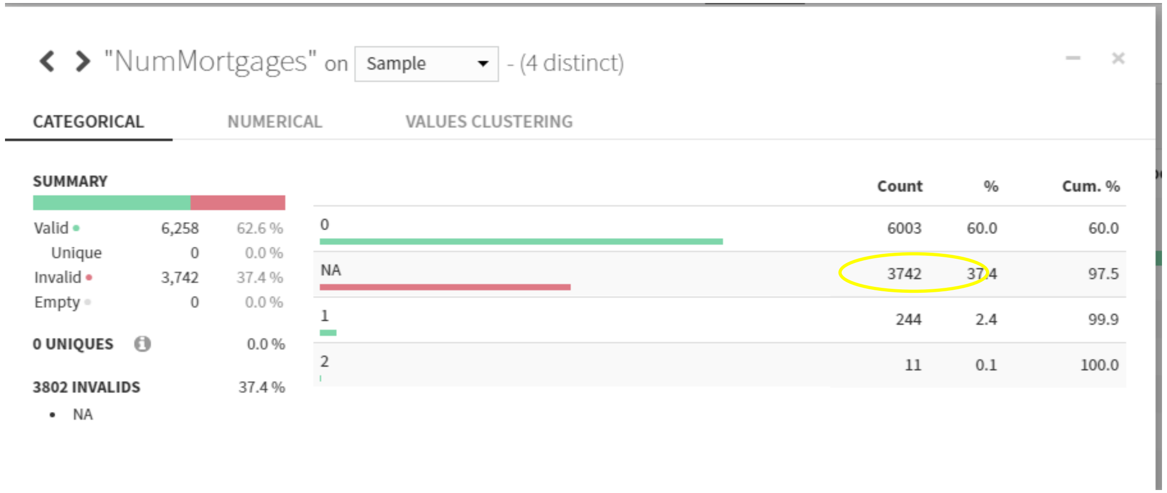
Sex a binomial variable: It was stored as a string and also for analytical puroposes and graphs, it is easier to have it as a binomial value, where 1= male and 0 = Female.

Missing Values:

The biggest inconsistency of this dataset were the missing values in two fairly significant variables: number of mortgages and number of loans per client.

Number of Mortgages:

The data were split in two: One of them had all the known values; the other one had all the missing values. One logistic model were trained to predict if the customer had a mortgage or not based on the other significant variables of the data set. The model was then applied to the data seta with missing values and both data sets were merged.



SUMMARY

Valid	6,258	62.6 %
Unique	0	0.0 %
Invalid	3,742	37.4 %
Empty	0	0.0 %
0 UNIQUES		0.0 %
3802 INVALIDS		37.4 %

- NA

	Count	%	Cum. %
0	6003	60.0	60.0
NA	3742	37.4	97.5
1	244	2.4	99.9
2	11	0.1	100.0

Number of Loans:

Number of Loans: Once the data set had only the number of loans missing, a clustering model was applied. The data was divided in 5 clusters and the missing values of number of loans were applied the average.

The clusters were defined on:

Cluster_0:76% has a Refinancing for Previous Debt, against 14% globally;

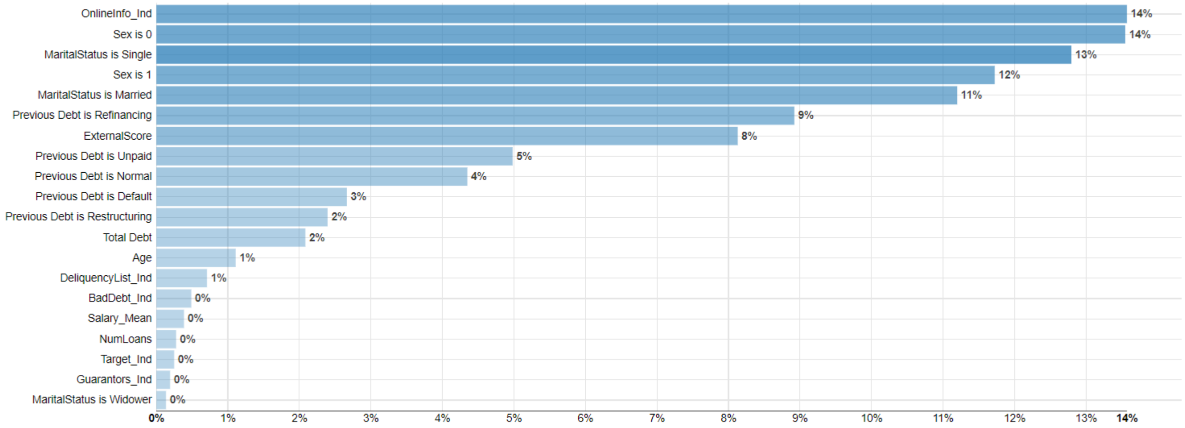
Cluster_1: Has an External Score 27% greater than the average

Cluster_2: 98% of the customers in this cluster are married

Cluster_3: 100% of the customers in this cluster has unpaid its previous Debt

Cluster_4: 99% of this cluster is Female

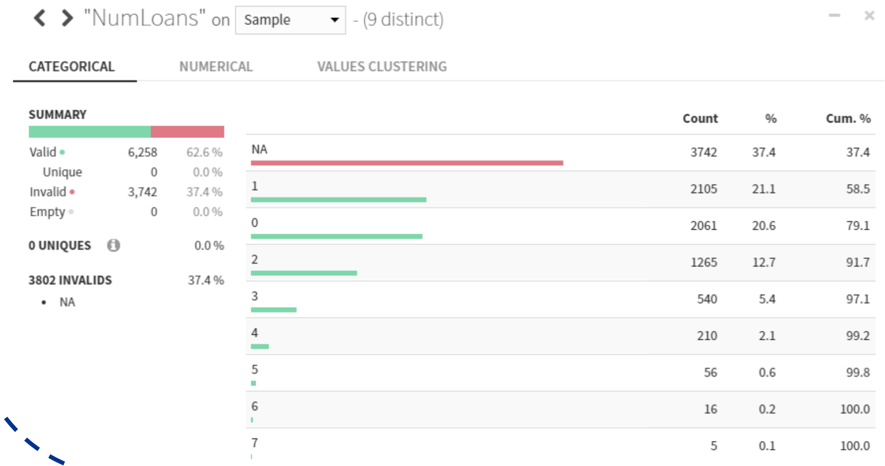
Cluster_5: 42% has its PublicDebt g12% greater than the average;



Exploratory analysis

The first analysis of the data was to see if there were any inconsistencies such as duplicated or missing values. In Dataiku, it is simpler to check for those and one of the biggest issues were related to missing values in Number of Loans and Mortgages.

The first solution proposed was to erase the rows with no values for those variables. However, for it to be possible a correlation matrix was needed: if there were another variable highly correlated with the latter, it was acceptable for the data quality.



SUMMARY

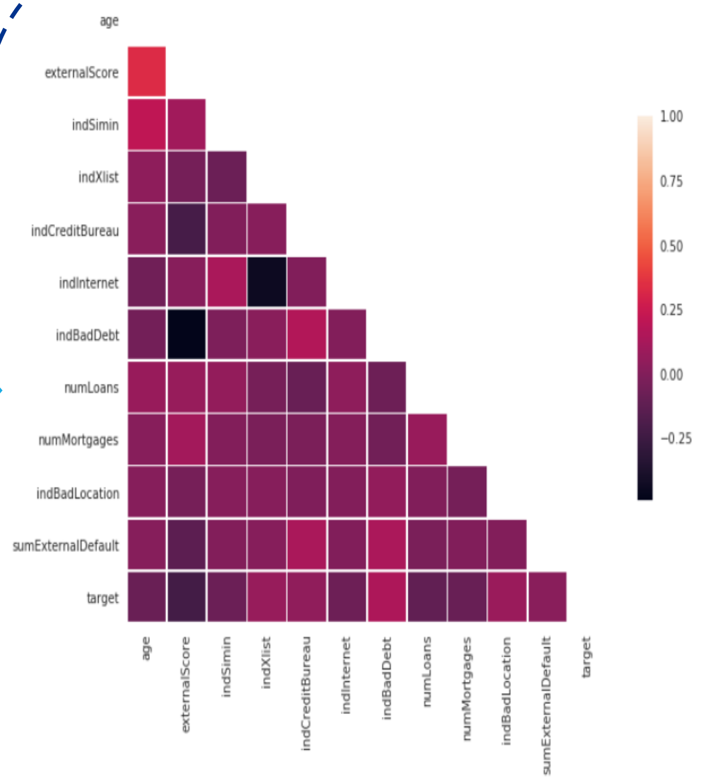
Valid	6,258	62.6 %
Unique	0	0.0 %
Invalid	3,742	37.4 %
Empty	0	0.0 %

0 UNIQUES 0.0 %

3802 INVALIDS 37.4 %

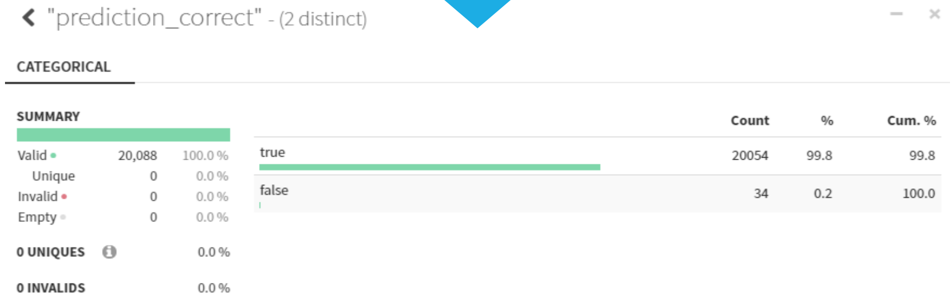
- NA

	Count	%	Cum. %
NA	3742	37.4	37.4
1	2105	21.1	58.5
0	2061	20.6	79.1
2	1265	12.7	91.7
3	540	5.4	97.1
4	210	2.1	99.2
5	56	0.6	99.8
6	16	0.2	100.0
7	5	0.1	100.0



It was created a notebook with a predefined code in R where it was possible to see the mentioned correlated matrix. In order to see if it was possible to eliminate any variable, in case they were highly correlated. As observed in the matrix , there was not a strong correlation and therefore no motivation to delete any variable.

Once it was concluded that the missing values in Number of Loans and Number of Mortgages were not explained by any other variable, a logistic regression to predict if the missing values of mortgages were 1 (have mortgage) or 0 (don't have mortgage). The predictive data were 99.8% correctly predicted.



SUMMARY

Valid	20,088	100.0 %
Unique	0	0.0 %
Invalid	0	0.0 %
Empty	0	0.0 %

0 UNIQUES 0.0 %

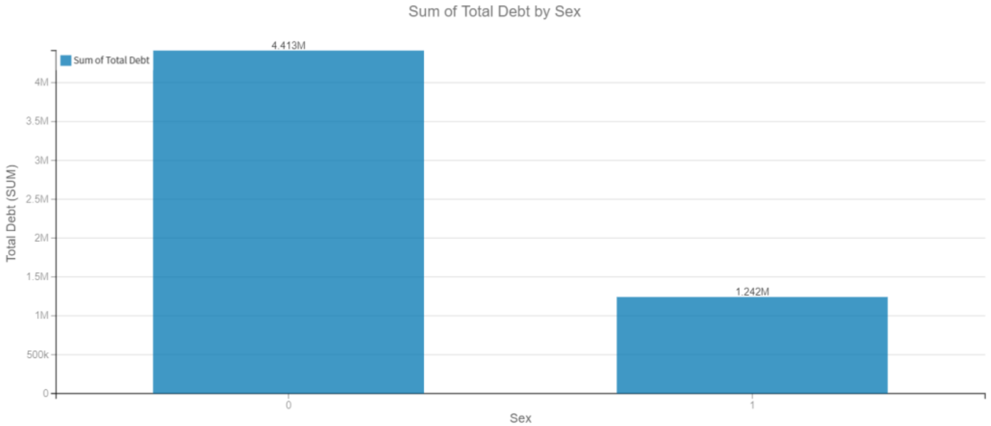
0 INVALIDS 0.0 %

	Count	%	Cum. %
true	20054	99.8	99.8
false	34	0.2	100.0

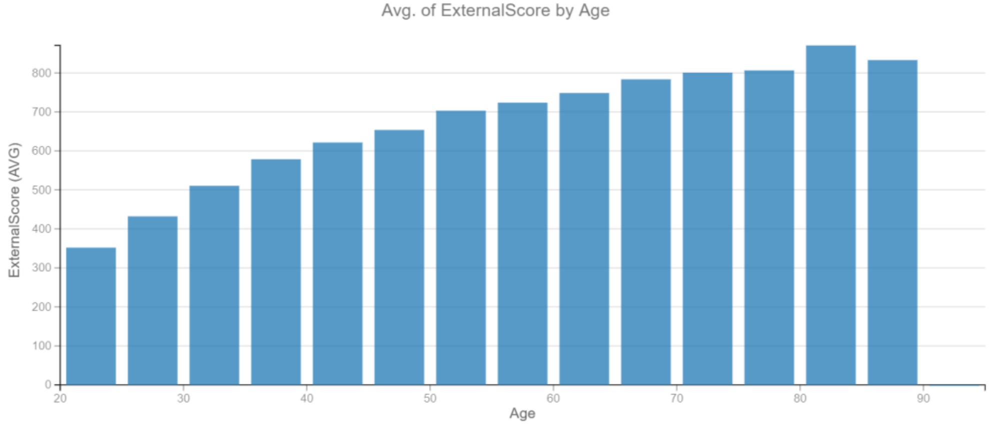
Insights

Some interesting insights for the business problem are:

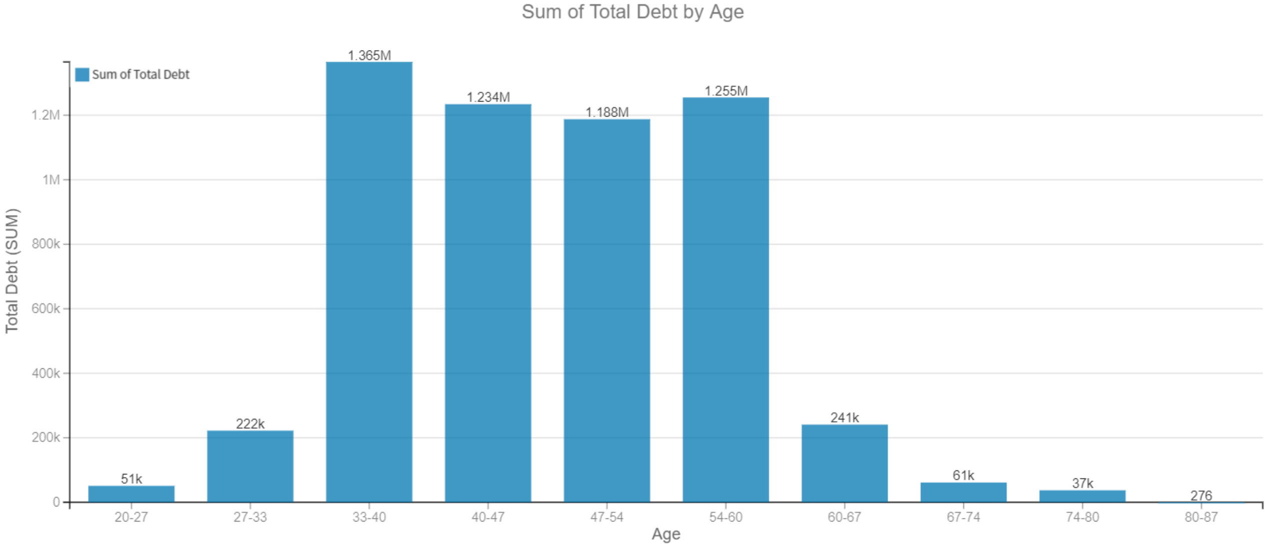
1. Women have 3,5 times the amount of total debt then men:



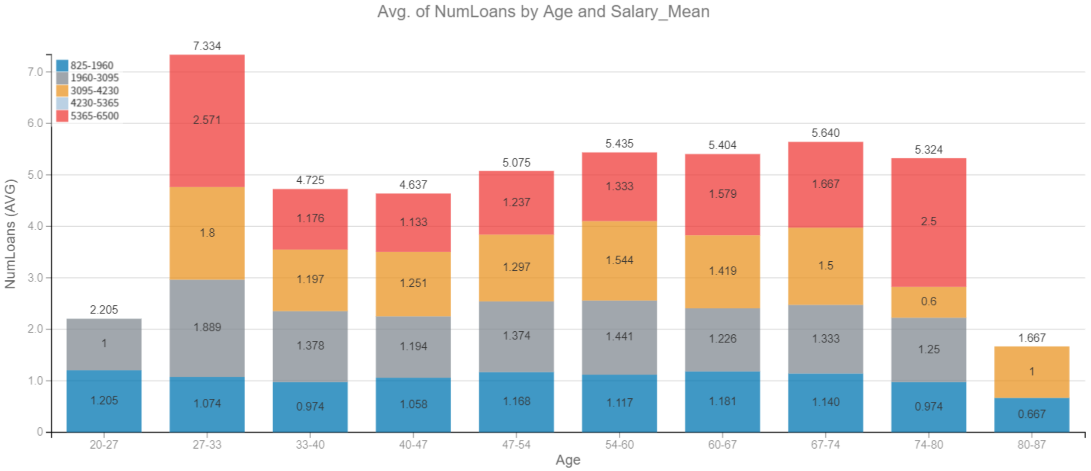
2. The External Score increases with age:



3. Customers between 50 and 60 years have the largest amount of Debt



4. Clients between 27 and 33 years have highest number of loans and also earn a higher salary in average.



SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Thank you

MODELLING CREDIT SCORING

PROFESSOR: ROBERTO
COSTUMERO MORENO
Student:
Amanda Marques