Credit Risk Score
Introduction & Data Cleaning

KNOWLEDGE DISCOVERY PROCESS

Eduardo Cort Pons

What is the business problem of the company & how can it be addressed?

The goal is to give credit cards as quick as possible to economically secure people with the lowest risk of having payment problems.

To achieve this, a credit risk score must be created having as a resource a document with 522.940 rows of client's information. The first step is to clean that document.

What Bank´s areas should participate?

The four main areas that should participate are:

- **The Compliance department** is the one in charge of saying if a client lives legally: it is the department which analyze if a client is corrupt, if he makes money-laundering, etc. So this department will participate to check the information about debts, client's guarantors, etc.

- **The Back Office** is in charge of arranging administrative processes, so it will participate when the credit risk score allows to start the process of giving a credit card, and also when clients ask for specific information, in cooperation with some local offices.

- **The Tech/Big Data department** will be in charge of creating the model to achieve the credit risk score, so it will be the main department which will be in contact with others if questions appeared (for example, with Compliance if they have doubts about the column of "externalScore")

- **The Sales department** will bring information about how credit cards are being sold. They will bring clarity when doubts regarding client's information appears

How the data has been cleaned?

We used Dataiku to clean de data. The transformations that we have done are:

1) Delete the rows with an **age** outside of 1.5 IQR (which is 19). With this, we only keep rows with an age in the **range of 18-75.5** (and we delete rows with an age of, for example, 110).

2) We **removed** the column **“indBadLocation”**

> We know the salary of the client thanks to a range of column “Salary”, and this range gives us more information about the client’s rent that a binary value which reflects if the client's address is located in a low rent per capita zone → “Salary” already reflects the information that interest us from “IndBadLocation”

3) The median of SumExternalDefault is 0 but we have noisy values such us 11496.06, so we remove rows with a **SumExternalDefault** outside of 5 IQR (which means to keep rows with a SumExternalDefault value between **0 and 272.2**

How the data has been cleaned?

4) We **created** column **“LowExtScore”** with a binary value that reflects if the value of “externalScore” is lower than 300, which means that the third party indicated that the client is risky. After this, we **deleted** column **“externalScore”**.

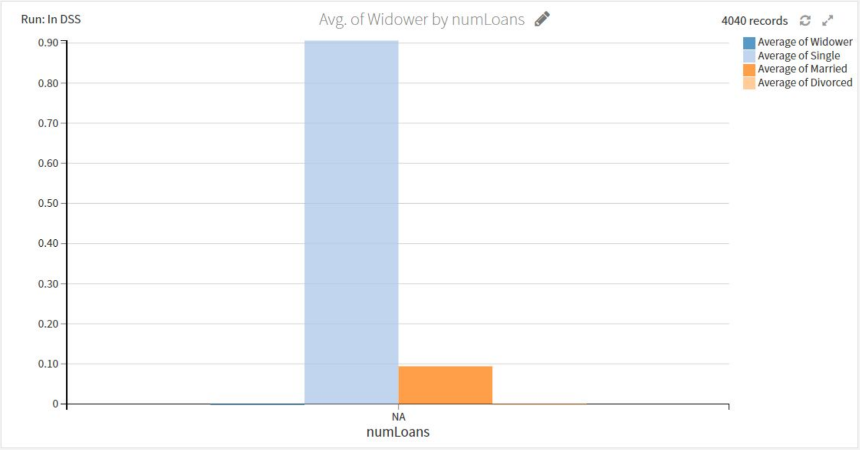
5) We categorized columns “Sex”, “Status”, “Channel” → We **created** columns **Male, Female, Single, Married, Divorced, Widower, ChannelApp, ChannelBranch, ChannelCallCenter, ChannelExternalAgent, ChannelOnline, ChannelRecovery and ChannelUnknown with values 1 or 0.** Then, we deleted columns “Sex”, “Status” and “Channel” to not have columns in excess.

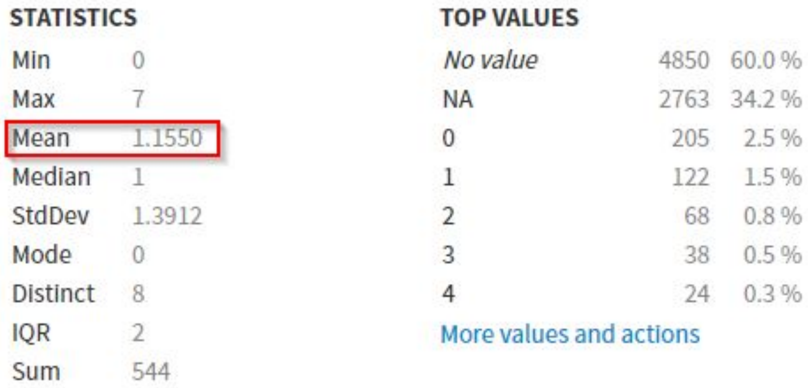
6) We **transformed** the **range of** the **salary into its average** → For example, range [1500-2000) has been transform to 1750. The new value appears in a new column called “**AverageSalary**” and the original “Salary” column has been removed

How the data has been cleaned?

7) Column “numLoans” had ‘NA’ values → 49.9% of the values were ‘NA”. We could delete the rows because even though it will imply deleting half of the dataset, we would still have around 261.000 of rows. Another option would be to put the average value of the column in all rows with ‘NA’ value.

- We analyzed the dataset and we find out that almost 100% of ‘NA’ values are from Single Females and 76% of them have value 1 in “indInternet” (client asked for information about credit cards online), so we chose to put the average value of the column but only based on Single Females with indInternet=1, which is 1.15 (instead of 1.30)





STATISTICS

Min	0
Max	7
Mean	1.1550
Median	1
StdDev	1.3912
Mode	0
Distinct	8
IQR	2
Sum	544

TOP VALUES

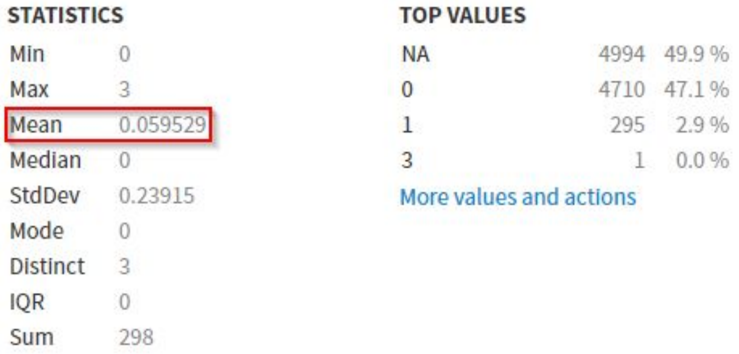
No value	4850	60.0 %
NA	2763	34.2 %
0	205	2.5 %
1	122	1.5 %
2	68	0.8 %
3	38	0.5 %
4	24	0.3 %

More values and actions

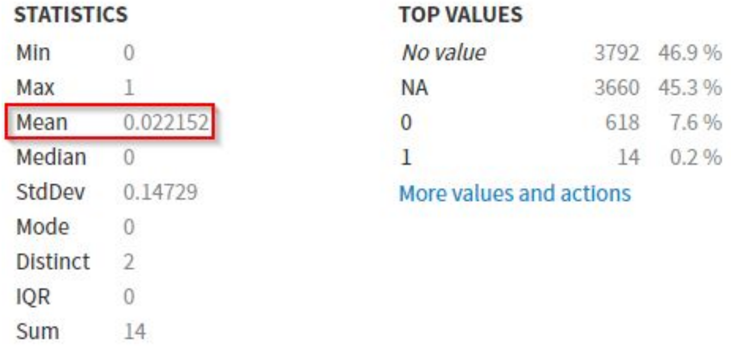
How the data has been cleaned?

8) Column “numMortgages” had ‘NA’ values → 50.1% of the values were ‘NA”. The approach that we chose was just like with “numLoans”:

- We analyzed the dataset and we find out that almost 100% of ‘NA’ values are from Single Females, so we chose to put the average value of the column but only based on Single Females, which is 0.022 (instead of the 0.059 value that we get as the mean if we check all values of numMortages, not only Single Females values)



```
STATISTICS                TOP VALUES
Min       0               NA        4994  49.9 %
Max       3               0         4710  47.1 %
Mean      0.059529        1          295   2.9 %
Median    0               3            1   0.0 %
StdDev    0.23915         More values and actions
Mode      0
Distinct  3
IQR       0
Sum       298
```



```
STATISTICS                TOP VALUES
Min       0               No value  3792  46.9 %
Max       1               NA        3660  45.3 %
Mean      0.022152        0          618   7.6 %
Median    0               1           14   0.2 %
StdDev    0.14729         More values and actions
Mode      0
Distinct  2
IQR       0
Sum       14
```

How the data has been cleaned?

9) Finally, we remove duplicate rows. We defined a duplicate row as: a row that has in every column the same values that another row or a row that has the same value in “customerID” that another row.

We took the assumption that when the customerID appears more than once, we consider that row as a duplicated row and we deleted both, despite the values of the other columns, as we have enough data to do so.

- We realized that the majority of rows with duplicated customerID has only one column different: “previous” column, which show the client debt classification of the previous year. The problem is that we do not know which of the two rows is the most recent one.

- We do not have a column to show us which of the duplicate rows is the one to take into account.

10) We **changed** the next **column names** (to clarify the meaning of the column)): indSimin to hasGuarantors, indXlist to hasDebts, indCreditBureau to isDelinquentCreditBureau, indInternet to askInfoOnline, indBadDebt to riskBadDebt and target to unpaidCreditCardFee

What exploratory analysis have been done?

First of all, I searched information to confirm that all columns were necessary for our goal, and to understand all columns (I did not know the meaning of a “Credit Score” of 300, 500 or 800).

During this process, I found out that in countries like Canada people age 18-22 with a Loan (numLoans>0) is quite normal, in contrary to Spain.

As seen from page 6, using the “Charts” of Dataiku was very important to confirm or reject hypothesis that affects to the process.

What insights did you discover?

Apart from the insights commented in the “data cleaning” section (like the relationship between Female and ‘NA’ values) I discovered that:

- With regard to Canada-Spain comparison, knowing not only the laws of a country but also the way of living of their people is important to know how to treat your data.

- When the channel for which the credit card has been contracted is ‘Branch’ or ‘Online’, we have a considerable percentage of credit card fee which has been unpaid (target=1)

- The lower the salary, the higher the possibility of having an unpaid credit card fee

References

-To create the column “LowExtScore” (explanation in page 5) I support my decision in the next references:

-https://www.consumerfinance.gov/es/obtener-respuestas/que-es-un-puntaje-de-fico-es-1883/

-https://www.experian.com/blogs/ask-experian/credit-education/score-basics/300-credit-score/

-https://www.consumerfinance.gov/es/obtener-respuestas/que-es-un-puntaje-de-credito-es-315/

-https://www.mybanktracker.com/credit-cards/credit-score/how-fix-very-bad-credit-score-254021

Table of Figures

-Salary/Target chart to support the insight reflected in page 9:

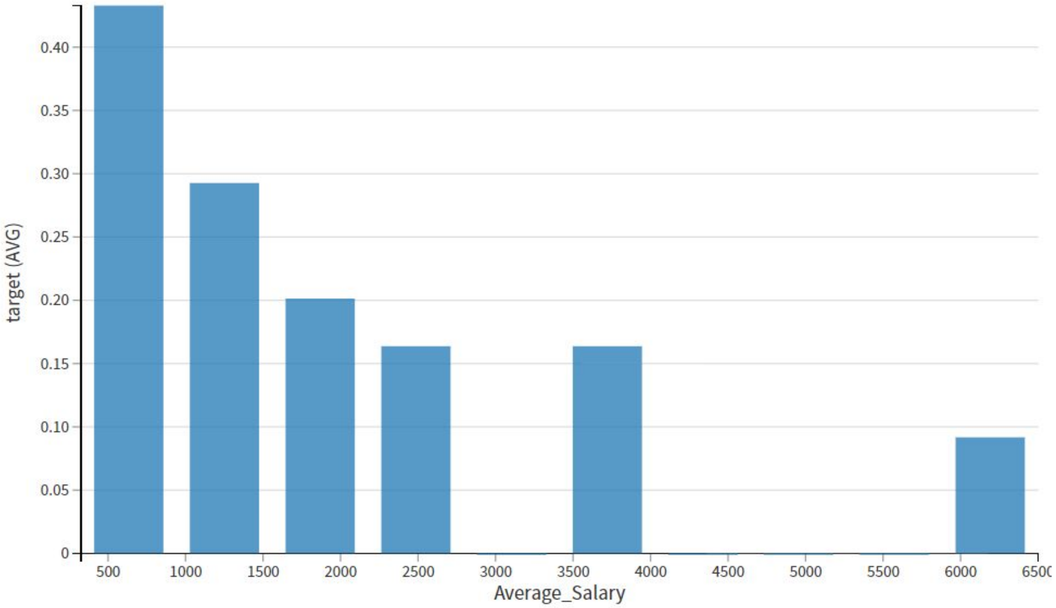
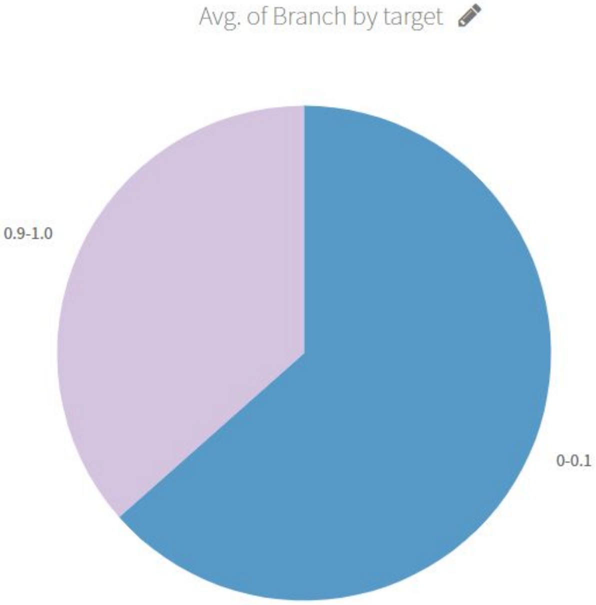
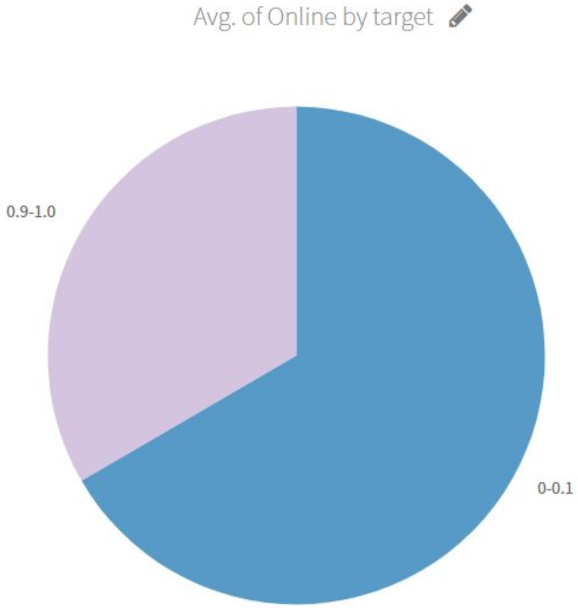


Table of Figures

-Branch channel/Target and Online channel/Target charts to support the insights reflected in page 9:


Avg. of Branch by target


Avg. of Online by target