CREDIT ANALYSIS

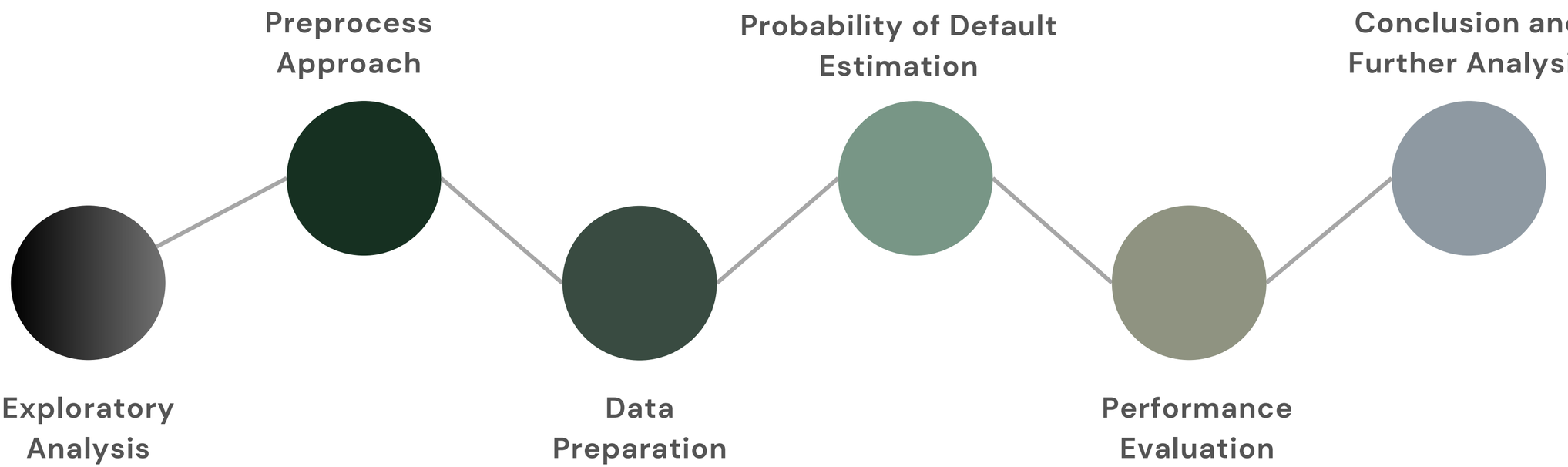
Fabiana Arca Cruz Tortorelli

INTRODUCTION

Credit analysis is a import process used by financial institutions and lenders to evaluate the creditworthiness of individuals or businesses applying for loans or credit. It involves assessing the applicant's ability to repay the borrowed funds by analyzing various financial indicators, such as income, debt levels, credit history, and financial stability. The goal of credit analysis is to minimize the risk of default and ensure that loans are extended to reliable borrowers.

The first stage in the credit lifecycle is the credit granting process, where credit is granted based on a thorough assessment of the applicant’s creditworthiness. This involves evaluating the individual's or company's financial situation, including income, debt, credit score, and overall financial health. Lenders typically use credit scoring models, to predict the likelihood of default. Therefore, the **goal** of this work is to build a Credit Scoring using logistic regression.

PROJECT TIMELINE



EXPLORATORY ANALYSIS

The client dataset contains **10,738** observations and **81** features (78 features anonymized).

It contains **72** (88.89%) features with missing values.

✦ INT TYPES

- **id:** unique key
- **safra:** the month and year in which the credit was granted
- **y:** target variable (1: bad, 0: good)
- **VAR_20:**
  - values from 3 to 12, no missing values
  - mean 10.435, median 12
- **VAR_57:**
  - values from 18 to 78, no missing values
  - mean 45.963, median 46
- **VAR_64:**
  - 0 and 1, no missing values
  - 60.7% contains 1 and 39.3% contains 0

✦FLOAT TYPES

›› 3 with no missing values:

- 'VAR_9', 'VAR_32', and 'VAR_60'

›› 72 with missing values:

- 'VAR_1', 'VAR_2', 'VAR_3', 'VAR_4', 'VAR_5', 'VAR_6', 'VAR_7', 'VAR_8', 'VAR_10', 'VAR_11', 'VAR_12', 'VAR_13', 'VAR_14', 'VAR_15', 'VAR_16', 'VAR_17', 'VAR_18', 'VAR_21', 'VAR_22', 'VAR_23', 'VAR_24', 'VAR_25', 'VAR_26', 'VAR_27', 'VAR_28', 'VAR_29', 'VAR_30', 'VAR_31', 'VAR_33', 'VAR_34', 'VAR_35', 'VAR_36', 'VAR_37', 'VAR_38', 'VAR_39', 'VAR_40', 'VAR_41', 'VAR_42', 'VAR_43', 'VAR_44', 'VAR_45', 'VAR_46', 'VAR_47', 'VAR_48', 'VAR_49', 'VAR_50', 'VAR_51', 'VAR_52', 'VAR_53', 'VAR_54', 'VAR_55', 'VAR_56', 'VAR_58', 'VAR_59', 'VAR_61', 'VAR_62', 'VAR_63', 'VAR_65', 'VAR_66', 'VAR_67', 'VAR_68', 'VAR_69', 'VAR_70', 'VAR_71', 'VAR_72', 'VAR_73', 'VAR_74', 'VAR_75', 'VAR_76', 'VAR_77', and 'VAR_78'

PREPROCESS APPROACH

Missing values

From exploratory analysis, 75% of the features have more than 37.81% of missing values.

As we don't know the meaning of each feature, we set the following rule:

- For features with 50% or less missing values, we assigned them to a "Missing" category during the Fine (reduce feature to discrete ranges) and Coarse Classing process
- For features with more than 50% missing values, we dropped them (38 features)

Date time feature

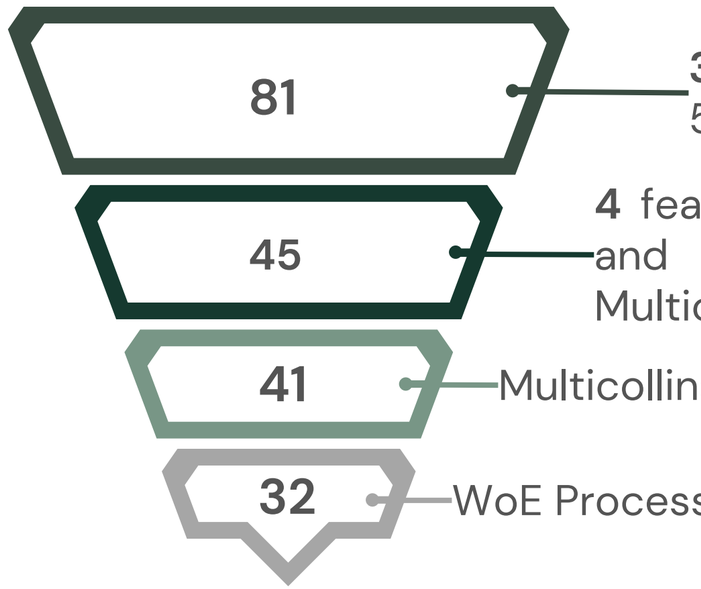
For safra feature, we transformed it to date time and calculated the number of months since the credit line, considering as reference date 2017-12-01

DATA PREPARATION

For this approach, the model needs to be interpretable, and one way to achieve this is by converting all features into dummy variables.

Therefore, we preprocessed the data to transform the features into dummy variables using Weight of Evidence (WoE) - applied to discrete predictors, whether nominal, ordinal, or numeric.

To apply WoE process we considered the features that have 50% or less of missing values and don't have high correlation with another feature (threshold = 0.8).



38 features with more than 50% of missing values

4 features (id, safra, safra_dt, and y) not include on Multicolinearity process

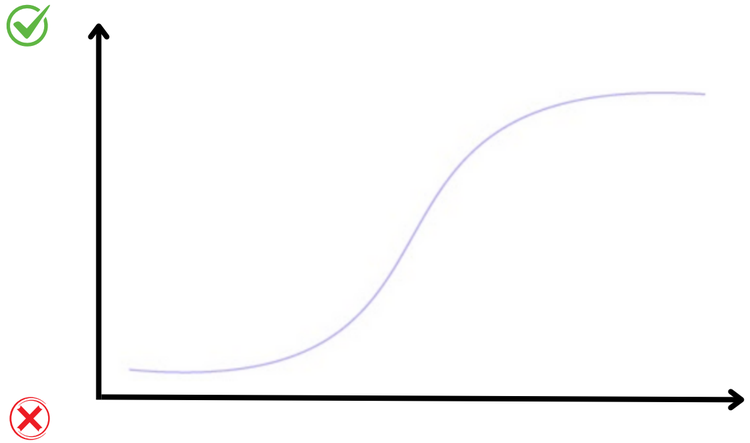
Multicollinearity Process

WoE Process

22 features (62 dummy features):

'VAR_2', 'VAR_3', 'VAR_4', 'VAR_20', 'VAR_28', 'VAR_32'
'VAR_33', 'VAR_60', 'VAR_64', 'months_since_cr_line'
'VAR_1', 'VAR_6', 'VAR_7', 'VAR_9', 'VAR_11', 'VAR_17'
'VAR_22', 'VAR_35', 'VAR_53', 'VAR_54', 'VAR_65', 'VAR_72'

PD ESTIMATION | PERFORMANCE EVALUATION



$$\ln\left(\frac{1-p}{p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Final Features:

VAR_2	VAR_1
VAR_20	VAR_7
VAR_32	VAR_9
VAR_60	VAR_27
months_since_cr_line	VAR_65

Threshold = 0.5

Actual \ Predicted	Good	Bad
Good	1,391	138
Bad	415	204

- **Sensitivity (Good):** 0.90
- **Specificity (Bad):** 0.32
- **Accuracy (Good + Bad):** 0.74
- **Balanced accuracy:** 0.61

AUC	Gini	KS
0.72	0.45	0.34

- The area under the ROC Curve **(AUC)** shows a good but not excellent performance
- The **Gini coefficient** of 0.45 shows that logistic regression model has moderate discriminatory power in distinguishing between class 1 (good) and class 0 (bad)
- The **Kolmogorov-Smirnov statistics** (KS) of 0.34 indicates that the model is performing well in distinguishing between the two classes

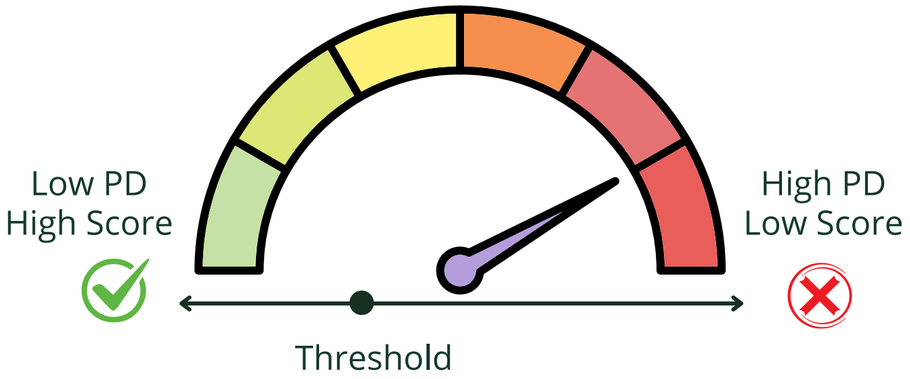
PD AND SCORE

$$\ln\left(\frac{1-PD}{PD}\right) = -1.9851$$

Intercept

$$\begin{aligned}
&+ 0 \times \text{VAR_2:1-14_or_missing} \\
&+ 0 \times \text{VAR_20:11-12} \\
&+ 0.7098 \times \text{VAR_32:>=0.13} \\
&+ 0.5384 \times \text{VAR_60:>-0.004} \\
&+ 0.6936 \times \text{months_since_cr_line:41-46} \\
&+ 0.2957 \times \text{VAR_1:66.56-128.96} \\
&- 0.0441 \times \text{VAR_7:<=62.138} \\
&+ 0.6443 \times \text{VAR_9:<=606.0} \\
&+ 0.6854 \times \text{VAR_17:>2833.422} \\
&- 0.3678 \times \text{VAR_65:<=1120.334}
\end{aligned}$$

Estimated coefficients for each dummy feature



$$\left(\frac{1-PD}{PD}\right) = e^{1.1702} = 1.6919 \Rightarrow (1-PD) = \frac{3.2226}{(3.2226+1)} = 0.7631$$

The probability of not default for this person is 76.31%.
Hence, the probability of default is 23,69%

PD AND SCORE

- We transformed PD into a scorecard, which can facilitate lending decisions by systematically assessing credit risk
- From the table below, we noted that for the worst score range (0 to 200), the percentage of bad loans is 86.71%, whereas for the best score range (800 to 1000), this percentage is only 4.14%

Credit Scoring

Feature	Category	Coeff	Score
VAR_32	VAR_32:>=0.13	0.7098	109
VAR_60	VAR_60:>-0.004	0.5384	83
VAR_7	VAR_7:<=62.138	-0.0441	-7

• • •

Score Range	Pct Clients	Pct Bad
0-200	1.66%	86.71%
200-400	15.87%	58.91%
400-600	40.90%	32.22%
600-800	36.52%	13.77%
800-1000	5.05%	4.14%

FUTURE ANALYSIS OPPORTUNITIES

- **Outliers:** Through interquartile range (IQR) method, we observed that 71 of 78 (91.02%) independent features present outliers, and as an outlier is an observation that is abnormally distant from other values it is crucial to understand them to handle the best treatment. In our dataset, the features are anonymized, and the maximum percentage of outliers is 16.86%, therefore, we considered them as normal values, and we didn't set any treatment or remove them during data preprocessing

- **Missing Values:** We identified 75% of the features have more than 37.81% of missing values. Understanding the nature of missing values is very important, in many cases, missing data may carry intrinsic meaning. However, we don't know the meaning of each feature because they are anonymized, and we followed the rules:

  - For features with 50% or less missing values, we assigned them to a "Missing" category during the Fine (reduce feature to discrete ranges) and Coarse Classing process
  - For features with more than 50% missing values, we removed them (38 features)

- **Gini coefficient and KS statistics:** both metrics showed that the model is performing well, but it could be refined. It can be done by adding new features, transforming existing ones, or trying different algorithms such as random forests or gradient boosting

- **Loss Given Default (LGD), Exposure at Default (EAD), and Expected Loss (EL):** as we don't know the meaning of the features it wasn't possible to calculate LGD, EAD and EL