30416 Big Data and Databases Project

Cl.25

Analysis Report

"*Should This Loan be Approved or Denied?* "

Group 2:

Mert Tekdemir - 3069602

Davide Drago - 3058517

Chiara D'Ignazio - 3065411

Nunzio Fallico - 3072744

Giacomo Bugli - 3058679

Luigi Noto - 3075187

BEMACS

Università Commerciale "Luigi Bocconi"

Italy, Milan

11 December 2020

CONTENTS:

1. Goal of The Analysis

The goal of the analysis is to predict whether or not a loan should be approved based on archives provided by the U.S. Small Business Administration (SBA). In order to do so the probability of default of a loan is estimated to understand which variables are most likely to influence this process.

This dataset has been chosen because the SBA is a reliable government organization, founded in 1953, that fosters small business formation and growth, which have considerable social benefits by creating job opportunities and reducing unemployment in the United States of America. In particular, the SBA pursues its goal through a loan guarantee program which is designed to encourage banks to grant loans to small businesses. The SBA acts much like an insurance provider to reduce the risk for a bank by taking on some of the risk through guaranteeing up to 85% of the loan amount, that is the portion paid by the SBA in case the loan goes into default.

Since SBA loans only guarantee a portion of the entire loan balance, banks will incur some losses if a small business defaults on its SBA-guaranteed loan. Due to the nature of the small businesses involved, banks are still faced with a difficult choice as to whether they should grant such a loan because of the high risk of default associated with them. One way to inform their decision making and to avoid incurring losses, is through analyzing relevant historical data to identify trends to guide managerial decision making.
The analysis aims to provide a useful tool or ruleset for bank officers to assess if a loan should be approved or not based on the probability of default estimated by the characteristics of the small business in question.

2. Data Preprocessing

2.1. Dataset Introduction and Data Preparation

The dataset consists of 899.132 observations and 27 variables (text, date/time, number and currency types) ranging from 1987 to 2014. The variables are listed in the following table:

Table 1: Dataset Overview

Name	Description	Data type
LoanNr_ChkDgt	Identifier – Primary key	Text
Name	Borrower name	Text
City	Borrower city	Text
State	Borrower state	Text
Zip	Borrower zip code	Text
Bank	Bank name	Text
BankState	Bank state	Text
NAICS	North American industry classification system code	Text
ApprovalDate	Date SBA commitment issued	Date/Time
ApprovalFY	Fiscal year of commitment	Text
Term	Loan term in months	Number
NoEmp	Number of business employees	Number
NewExist	1=Existing business, 2=New business	Text
CreateJob	Number of jobs created	Number
RetainedJob	Number of jobs retained	Number
FranchiseCode	Franchise code, (00000 or 00001)=No franchise	Text
UrbanRural	1=Urban, 2=rural, 0=undefined	Text
RevLineCr	Revolving line of credit: Y=Yes, N=No	Text
LowDoc	LowDoc Loan Program: Y=Yes, N=No	Text
ChgOffDate	The date when a loan is declared to be in default	Date/Time
DisbursementDate	Disbursement date	Date/Time
DisbursementGross	Amount disbursed	Currency
BalanceGross	Gross amount outstanding	Currency
MIS_Status	Loan status charged off=CHGOFF, Paid in full=PIF	Text
ChgOffPrinGr	Charged-off amount	Currency
GrAppv	Gross amount of loan approved by bank	Currency
SBA_Appv	SBA's guaranteed amount of approved loan	Currency

The dataset import presented two main bugs, one related to the wrong format of the primary key column and the other related to the wrong datatype of some variables, that have been solved quickly before proceeding with the analysis. Then the currency variables mentioned in the table above needed to be manipulated by removing the dollar sign and converted to the right data type, double. Furthermore, since the date format was not the one used by Knime to manage date variables, we designed a process with Python nodes in order to change the format according to the requirements of the software (and be able to use them in the feature engineering part).

To capture the essence of the *GrAppv* and *DisbursementGross* variables, we calculated the difference between the two. This allowed us to find that ⅔ of the observations had the same values while for the remaining part the difference was negligible. Likely these differences are due to circumstantial factors that vary on a case by case basis. Therefore only the variable *GrAppv* is considered as it ensures greater interpretability in the analysis and applicability from the perspective of a bank, as banks will not be able to rely on the gross disbursement as a data point in the decision making process.

For what concerns the *UrbanRural* variable, more than 300k observations were categorized as undefined even though the *Zip* code variable was available for most of them. Since it is a valuable dummy characteristic an additional database of RUCA codes (Rural-Urban Commuting Area codes) was cross checked with the *Zip* variable to handle missing *UrbanRural* values. Namely, zip codes that had a corresponding RUCA code of 10 were classified as Rural while for values less than 10 it was classified as Urban. After all the process only 6543 observations remained undefined due to wrong Zip Codes which were dropped.

Next the *LowDoc* variable was handled. It is a variable that indicates if a loan can be processed using a one-page application (to speed up the process) applicable only if the amount of the loan request is lower than 150.000$. Examining the data, many observations reported wrong *LowDoc* values, hence the column was reclassified according to *GrAppv* column, if *GrAppv* was lower or equal to 150.000$ it was classified as *LowDoc*=1 and *LowDoc*=0 otherwise.

As far as *FranchiseCode* is concerned, it is more applicable to consider if a particular business is part of a franchising or not, rather than use their specific franchise code. So this column was dummified assigning “1” if the *FranchiseCode* is greater than 1 and “0” in the other cases.

Regarding the *NAICS* variable, it is a 6-digits code that classifies the specific sector where a firm is operating according to the US Census Bureau. Only the first two digits were kept since they represent the larger economic sector and allow a more relevant grouping of the most significant industries.

Further, examining the dataset it could be seen that the *MIS_status* variable was mispecified in many observations (meaning that for example *MIS_status*=0 even though the charged off amount was greater than 0). In these cases *MIS_status* was imputed from the *ChgOffPrinGr* variable, correcting the wrong *MIS_status* values. At the same time the values of the dummies *NewExist* and *UrbanRural* were modified as they ranged from 1 to 2 instead of the traditional 0 or 1. Finally, some observations had missing values in the *State* column and since they were fewer than 100, they were dropped.

As regards to the feature engineering 3 additional variables were created that can provide interesting insights regarding a loans default behavior:

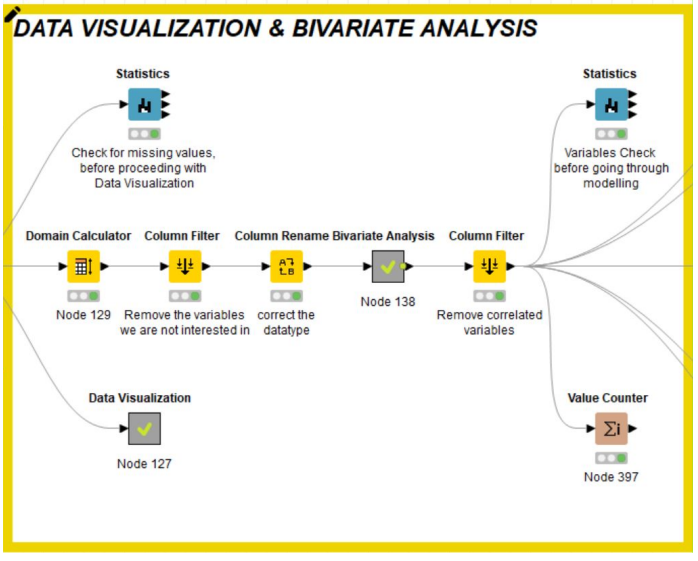
1. *Portion* = *SBA_Appv/GrAppv* ⇒ This feature represents the percentage of the loan that is guaranteed by SBA, this can be a risk indicator for the bank highlighting the level of exposure to possible losses. For this reason may be noticeable including this measure within the analysis.
2. *DaysToDisbursment* = *ApprovalDate* - *DisbursementDate* ⇒ This feature represents the days elapsed between the approval date of the loan and the day the firm receives the funds. It may be an interesting insight since some firms may need the money as soon as possible, thus a longer disbursement times of the funds could result in a higher probability of default of the loan.
3. *Recession* ⇒ This feature represents whether a loan has been active during the Great Recession period or not. This dummy variable has been calculated considering the loan disbursement date and the end of loan date (or defaulted date for defaulted loans), taking value 1 if the loan was active during the 2008 crisis (started in 01/12/2007 and ended in 30/06/2009) and 0 otherwise. Understanding how the economic cycles affect the loan evaluation is fundamental to precisely estimate the default probability, especially in regards to a recession period. It is expected that a loan active during that period should have a higher default probability due to the financial stress that the firm had suffered.

Table 2: Variables Created

Name	Description	Data type
Portion	Portion of the loan guaranteed by SBA	Number
DaysToDisbursment	Days between the approval and the disbursement date of the loan	Number
Recession	If the loan was active during the 2008 Great Recession: 1=Yes, 0=No	Dummy

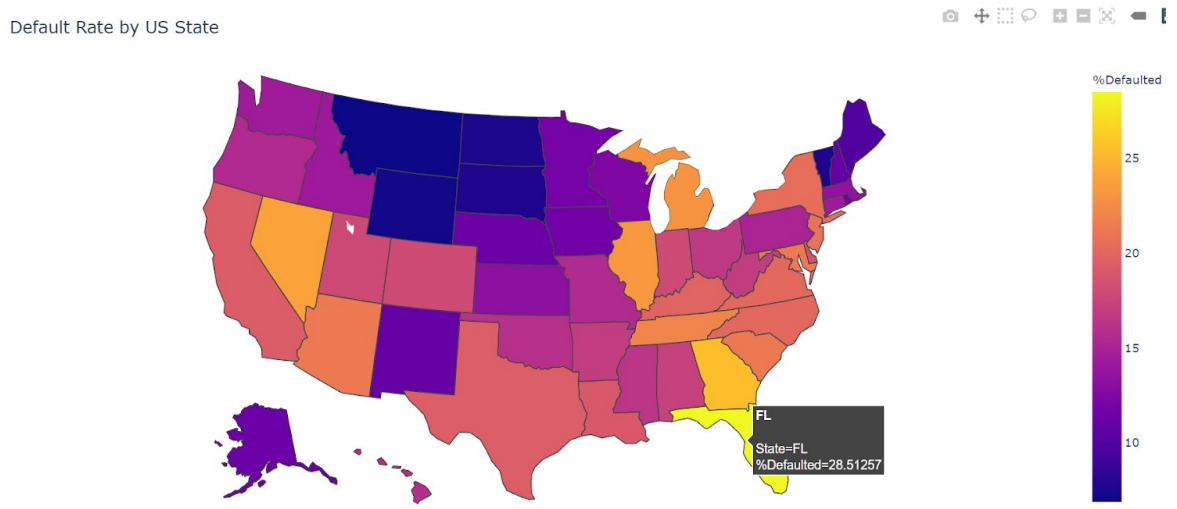
Following the data preparation only less than 14000 observations were dropped (accounting for less than 2% of the total observations), thanks to a meticulous work of imputation of missing or misspecified values. This enables greater confidence in the model's performance.

2.2. Dataset Visualization

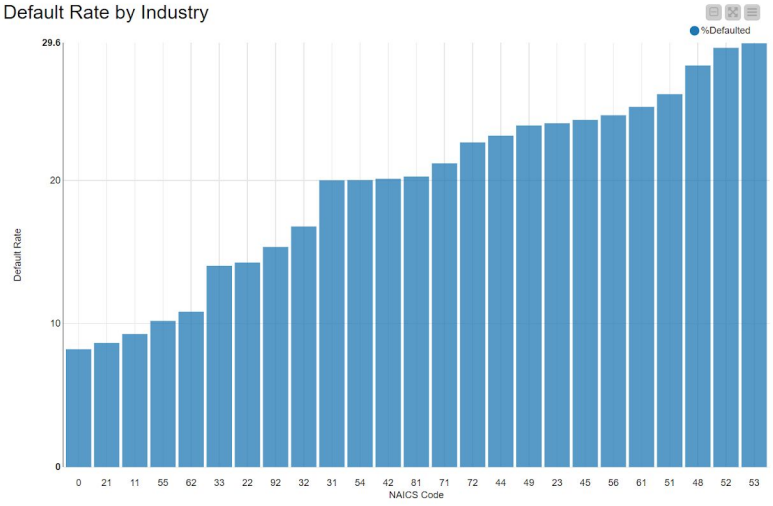


The data visualization focuses on some key explanatory variables which may be good indicators or predictors of the potential risk of a loan. In particular the variables that consistently emerge as indicators of risk in the analysis that could explain the variation of loan's default rate are *Location(State), Industry, New vs Established Businesses, Economic Recession, Term of the loan* and *Urban vs Rural.*

Starting from the *State,* looking at the heatmap it is noticeable that the different economic environments in which companies reside result in different default rates. In particular Florida has the highest default rate, maybe because it houses a significant housing market which suffered major declines in real estate prices during the Great Recession. On the contrary, states such as Wyoming and North Dakota had stronger economies, due to their reliance on minerals,oil and agriculture which are industries with relatively low default rates.

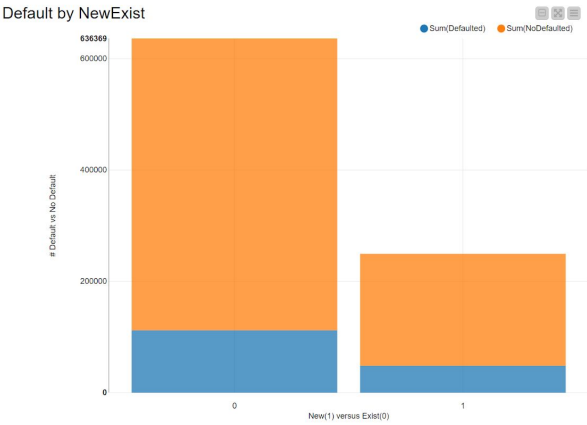


The variable *NAICS* is another key indicator of risk. The sectors with the lowest default rate (8-12%)are mining, oil and gas exploration (21), agriculture (11), management companies (55) and Healthcare (62). At the opposite side of the spectrum, the industries with higher probability of default (28-29%) are financial institutions (52) and real estate agencies (53).
Regarding financial institutions, the sector's propensity to work with large amounts of leverage make it susceptible to defaults. The large default rate for the real estate industry is almost certainly due to the great recession which saw housing prices and subsequently real estate margins plumet. These results also support the previous findings related to the Default Rate variation by State.



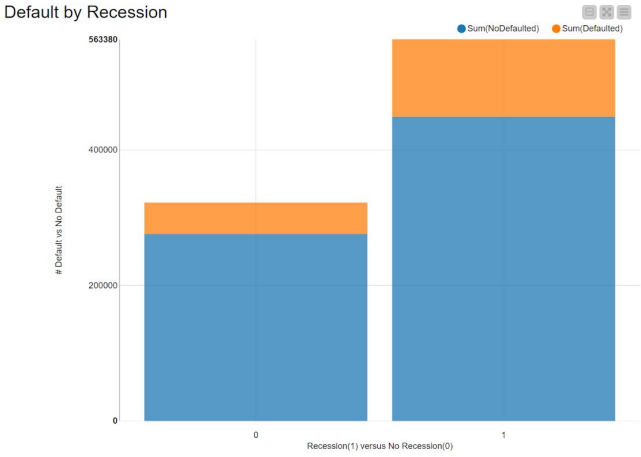
2 digit code	Description
21	Mining, quarrying, and oil and gas extraction
11	Agriculture, forestry, fishing and hunting
55	Management of companies and enterprises
62	Health care and social assistance
22	Utilities
92	Public administration
54	Professional, scientific, and technical services
42	Wholesale trade
31–33	Manufacturing
81	Other services (except public administration)
71	Arts, entertainment, and recreation
72	Accommodation and food services
44–45	Retail trade
23	Construction
56	Administrative/support & waste management/remediation Service
61	Educational services
51	Information
48–49	Transportation and warehousing
52	Finance and insurance
53	Real estate and rental and leasing

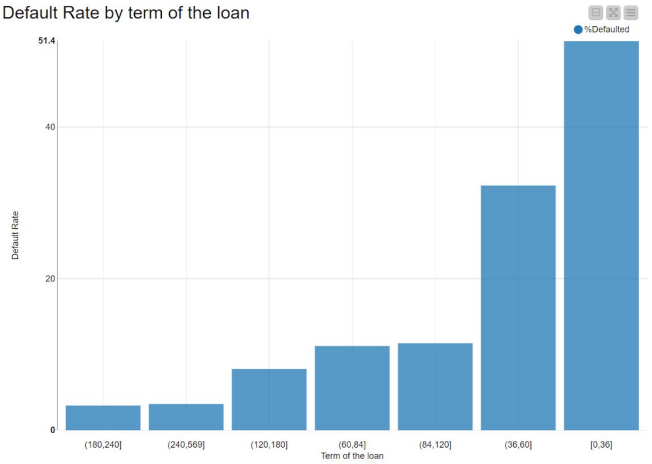
A counter intuitive result is given by the *New vs Established* variable. While it can be generally thought that new businesses may experience a higher default rate if compared to already existing ones, in the dataset there is a negligible difference between them. The default rate is 17.62% for established firms, while 19% for the new ones. Generally, newly established firms are considered likely to default due to their small size and lack of experience. However since all observations in the dataset are consider “small” businesses, the result may imply that the role of a firm's size is much more significant than its experience in the marketplace when it comes to its likelihood of default.



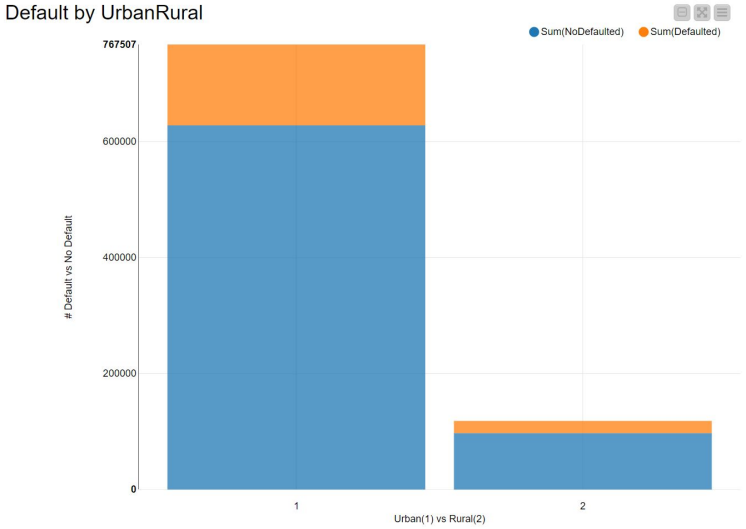
Another important risk indicator is the economic situation in which the firms are active. Indeed as seen from the graph, the default rate is higher for those companies that were active during the Recession period (20.3%), compared to those that were not (14.3%). Banks should therefore be more conservative during economic downturns, while the result may seem intuitive is not always the case due to the structure of managerial incentive programs.

It is interesting to consider also the term of the loan, looking at the variable *Term*. The result is unexpected since usually higher loan terms are associated with a higher risk of default. On the contrary, from the graph an opposite trend is evident, lower length loan terms correspond to a much higher default rate. A possible rationale of lower risk associated with terms above 240 months can be that those loans are necessarily backed by the real estate. As a consequence the possession of land is often large enough to cover the amount of any principal outstanding, thus reducing the probability of default.





Finally, a somewhat counterintuitive result is regarding the default rate of urban businesses compared to rural businesses, which is captured by the variable *UrbanRural*. Interestingly, the percentage of default within the two categories is nearly identical, 18.163% and 17.822% respectively. This suggests that this characteristic cannot be a basis for decision making on its own. It may be more insightful to consider it in combination however with other characteristics, for example the industry.



2.3. Bivariate Analysis & Feature Selection

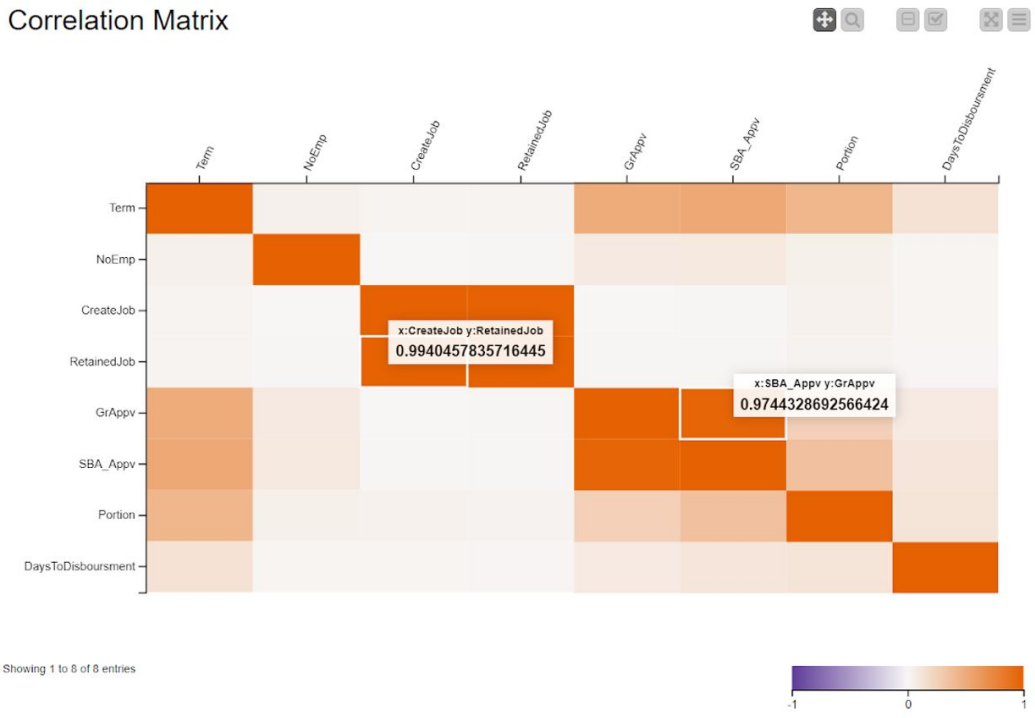
In addition to the univariate analysis and data visualization, it is important to perform also a multivariate analysis on the variables to be included in the model to determine their suitability.

The first step for the variable selection is removing the variables that are redundant or that do not provide interesting information for the final scope of the analysis. This will allow a reduction of the overall noise in the dataset and eventually increase the performance of the models. As a result, the following variables have been removed:

- *Name, ID*, *Bank, Bank State* because they do not provide relevant information for assessing the risk of default of a company;
- *City, Zip* because they provide the similar geographic information as the *State* variable which has been chosen because due to its more relevant clustering;
- *ChgOffDate, ChgOffPrincGr* because they contain additional information of defaulted companies, thus cannot be used as predictors;
- *ApprovalDate, ApprovalFY, DisbursementDate* because the relevant information of these variables is captured by the created variables *DaysToDisboursment* and *Recession;*
- *DisbursementGross* because it contains the same information as the variables *GrAppr* and *SBA_Portion* that have been chosen over the first for providing more relevant insights;
- *RevLinCr* because it contains many missing values and wrong entries.

Moreover a bivariate analysis is performed on the chosen variables to check for multicollinearity issues. For the numerical variables, the linear correlation coefficient has been computed and summarized in the correlation matrix. *RetainedJob* and *SBA_Appv* have been removed because they are highly correlated with *CreateJob* (0.99) and *GrAppv* (0.97) respectively.

For the categorical variables, the cross tabulation and Cramer's V measures have been computed and they do not show any statistical significant correlation among them.
Finally, dealing with the collinearity among categorical and numerical variables, the use of the ANOVA test does not provide reliable results because the low p-values are greatly affected by the large number of observations present in the dataset.



Regarding the outlier detection, given the nature of the dataset, univariate outliers do not constitute multivariate outliers when combined with the other relevant variables, thus they do not need to be removed.

The final result of the whole preprocessing steps is a dataset of 885.887 records and 14 variables including the target one, which are summarized in the following table:

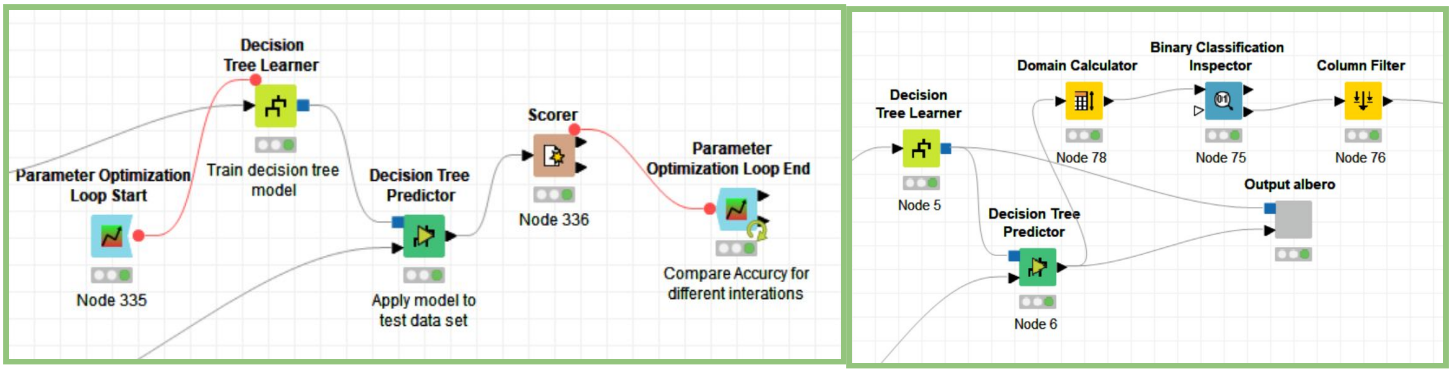
Name	Description	Data Type
MIS_Status	**Target variable:** "0" if defaulted, "1" otherwise	String
State	Borrower State	String
NAICS	North American industry classification code (first 2 digits)	String
Term	Loan term in months	Number
NoEmp	Number of business employees	Number
NewExist	"0" if Existing, "1" if New business	String
CreateJob	Number of Jobs created	Number
FrinchiseCode	"0" if no franchise, "1" if franchise	String
UrbanRural	"1" if urban, "2" if rural	String
LowDoc	"1" if LowDoc loan program, "0" otherwise	String
GrAppv	Gross Amount of the loan approved by the bank	Number
Portion	Portion of the loan guaranteed by SBA	Number
DaysToDisbursement	Days between the approval and the disbursement of the loan	Number
Recession	"1" if the loan was active during the 2008 Great Recession, "0" otherwise	String

4. Data Modelling and Model Evaluation

For the train test splitting, a stratified sampling procedure was used with a 80 - 20 ratio given the large amount of data at disposal.

Although the proportion of the positive class in the data set is 18%, the performance of the models was considered both on the original train set and on a balanced one. The balanced set was created using an *Equal Size Sampling* procedure on the target variable *MIS_Status*.

4.1. Decision Tree



As regard the hyperparameters of the Classification Tree the following have been used:

- Gain Ratio as the measure to optimize for the quality of split, which allows to reduce overfitting by preferring predictors with a few distinct categories;
- Both modes of pruning available to reduce as much as possible the complexity of the final solution.
-

Finally for the minimum number of records for the leaf node and maximum nominal splits, an optimization loop has been run. The following are the resulted optimized parameters:

- Minimum number of cases per node = 10 (Balanced); 15 (Unbalanced)

- Maximum number of nominal splits = 1 (Unbalanced and Balanced)

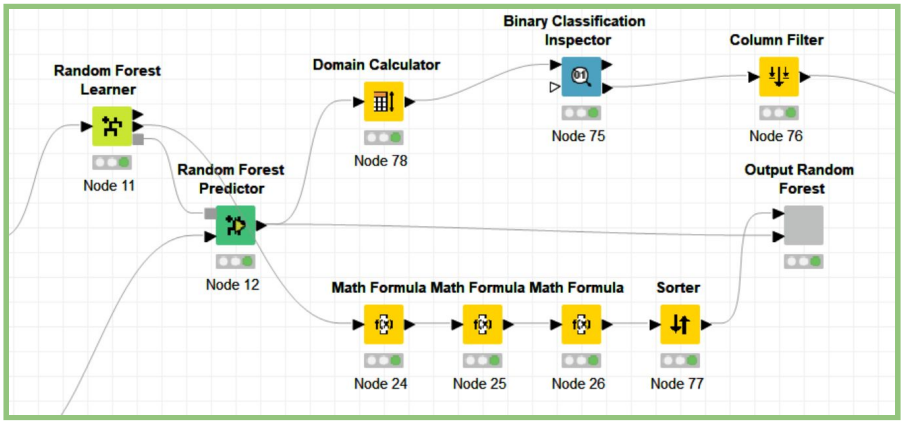
To reduce the possible overfitting of the single tree and thus increase the overall performance of the model, also two ensemble methods were considered: Random Forest and Gradient Boosting Tree. The Random Forest usually performs better on multiclass classification problems which tend to have a lot of statistical noise, on the contrary Gradient Boosting performs better when dealing with unbalanced data and real time risk assessment. Thus, given the structure of the dataset, it is expected that the latter model achieves a higher performance.

4.2. Random Forest

The hyperparameters chosen for the Random Forest Learner, both for the balanced and unbalanced train set are:

- Gain Ratio as the measure to optimize for the quality of split;
- Number of trees = 100;
- Minimum number of cases per node = 10.

This choice is based on the combination of efficiency and performance.



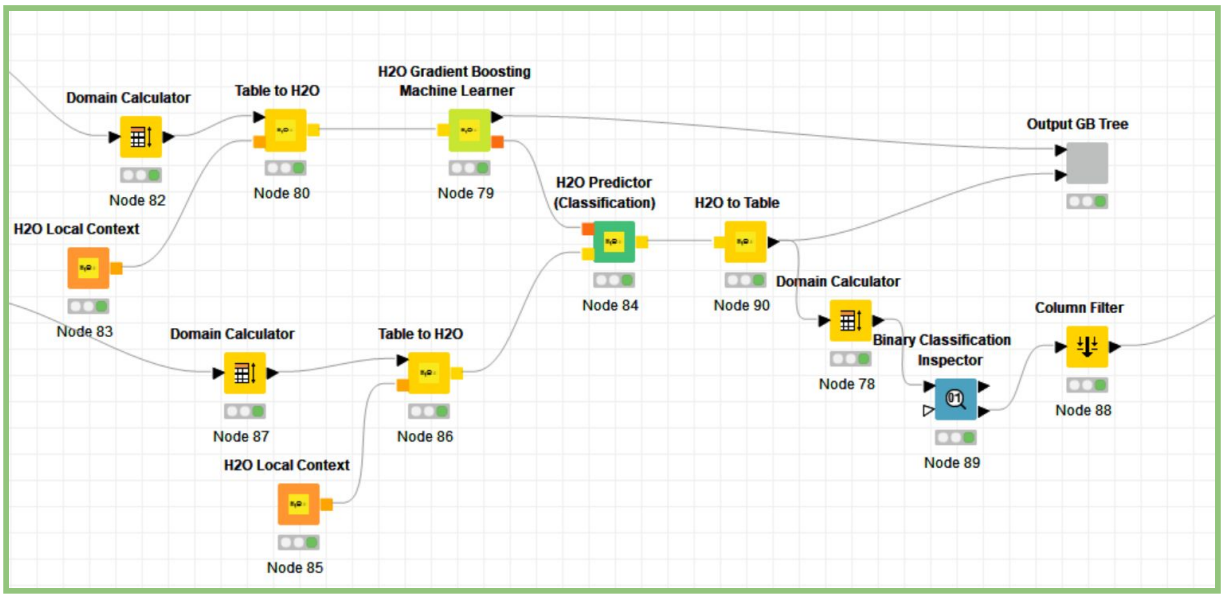
4.3. Gradient Boosting

For the Gradient Boosting the H2O extension was used as it is more efficient and provides as output also the relative importance of the features, which is essential to interpret the model and derive managerial implications.

The hyperparameters chosen for this model are:

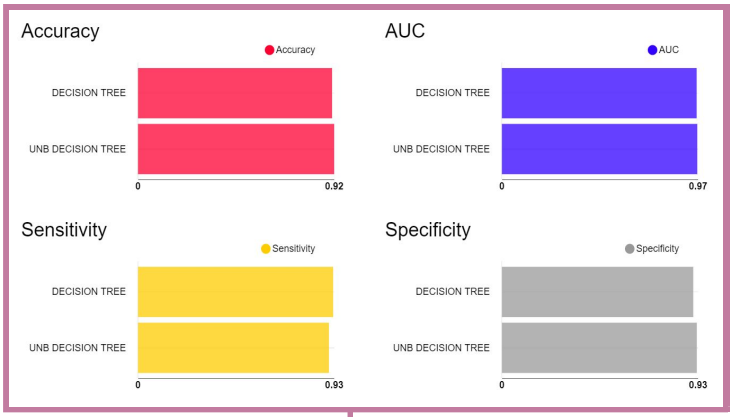
- Tree depth = 10;
- Number of models = 100;
- Learning rate = 0.1.

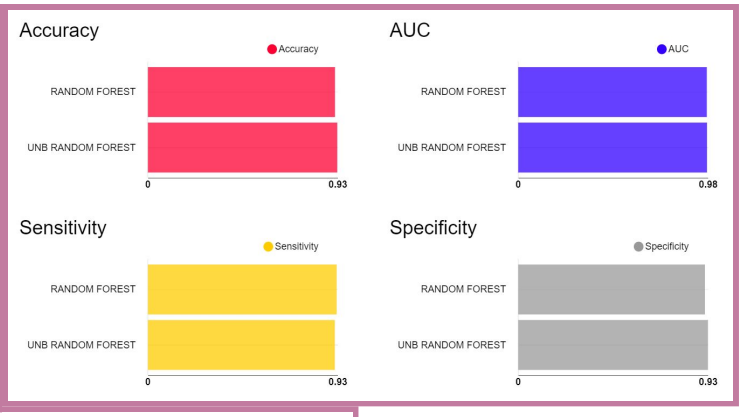
This choice again is based on the combination of efficiency and performance.

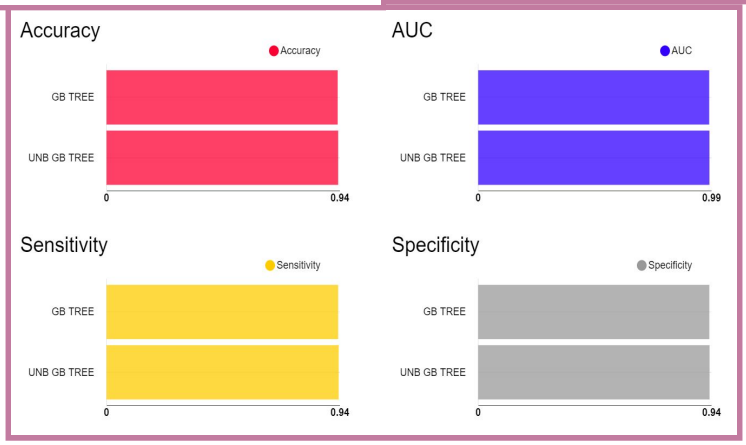


For the model evaluation the measures derived from the confusion matrix are considered. In particular, given the nature of the analysis, the performance has been compared mainly on the sensitivity. Indeed, from the perspective of a bank, the most threatening error is the false negative one, which implies giving a loan to a company that will eventually default. At the same time, even if with a lower weight, also the specificity must be considered in order not to harm the profitability of the bank.

Firstly, these are the results of the models taken individually and trained on the balanced and unbalanced train set:







As for the Decision Tree, there are not significant changes in the measures, except for 1% decrease in precision if trained in the balanced set. On the contrary the Random Forest performs better in terms of sensitivity (+1%) if trained in the balanced set. Finally the Gradient Boosting does not show significant changes in any of the measures. Given these results, it is preferred to proceed with the balanced set given the improved efficiency of the training process due to the lower amount of data.

Comparing the different models trained on the balanced set, the one that outperforms the others considering all the measures is the Gradient Boosting as expected. It has a sensitivity of 93% and an AUC of 98%.

4.4. Logistic Regression

Logistic regression is often used when a relation between a binary categorical variable and explanatory variables needs to be established, modelling the log odds as a linear function of the explanatory variables.

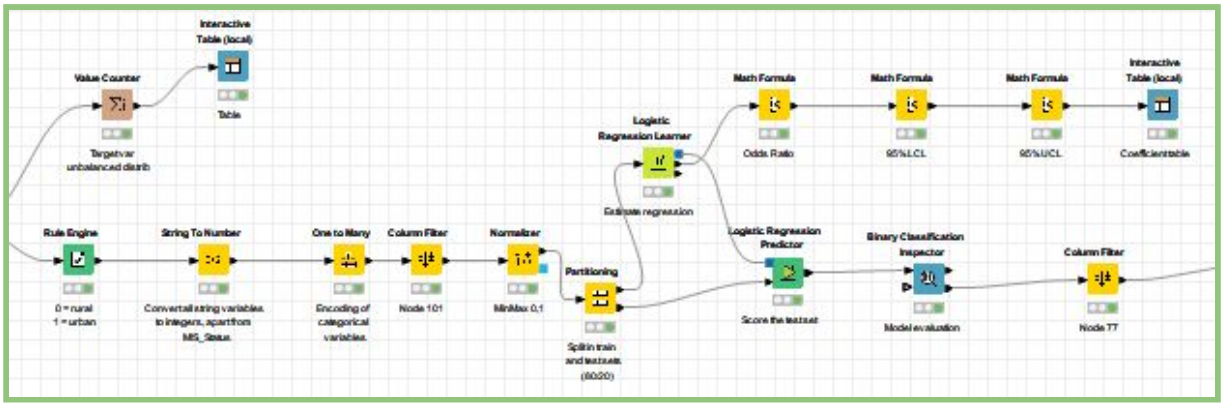
In the analysis the binary dependent variable is “MIS_Status” which is a dummy variable, thus the aim is to predict the probability of a loan defaulting, i.e. the probability of the event “MIS_Status” = 1. The logistic regression is used over a linear regression since when estimating probabilities through the OLS, its assumptions would have been violated, in particular the fitted values would not be constrained in the interval [0, 1],

To implement the regression in Knime the first step is to dummify the categorical variables *State* and *NAICS* and then normalize all the variables in the transformed dataset using the MinMax Normalizer (because it is common practice to do it, since variables with very different scales may have a different importance when optimizing the training function). The target column as mentioned before is *MIS_Status* with 0 as reference category and 1 as target category and all the features mentioned in the previous section were included.

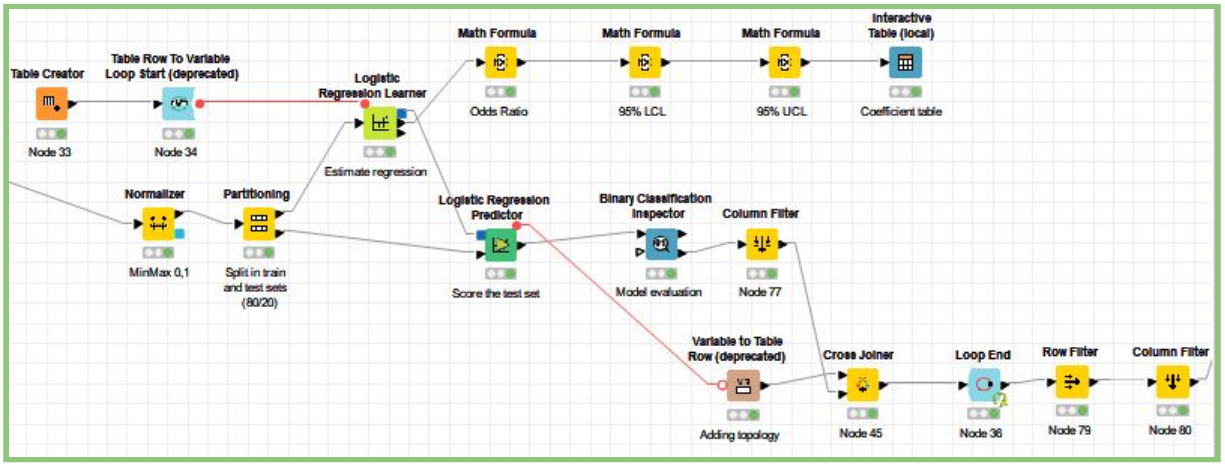
As regard to the hyperparameters of the Logistic Regression, the solver selected was the “iteratively reweighted least squares” that takes only the “Uniform” prior. Thus, the non-regularized version of the logistic regression was implemented, meaning no additional term was inserted in the training function to optimize that control for issues that might negatively affect prediction.

$$logit\,(\pi) = \beta_0 + \sum_{i=1}^{50} \beta_i\, State_i + \sum_{i=51}^{74} \beta_i\, NAICS_i + \beta_{75} Term + \beta_{76} NoEmp + \beta_{77} CreateJob + \beta_{78} GrAppv + \beta_{79} Portion$$
$$+ \beta_{80} DaysToDisbursment + \beta_{81} NewExist + \beta_{82} FranchiseCode + \beta_{83} UrbanRural + \beta_{84} LowDoc + \beta_{85} Recession$$

The indices corresponding to “DC” for the State dummies and “0” for the NAICS dummies were excluded to avoid multicollinearity issues, in fact in the equation there are 50 States and 24 NAICS codes instead of 51 and 25 respectively.



Despite having a sensitivity of 0.724 and ROC value of 0.85 in the naive model, it was preferred to investigate other techniques in order to mitigate the risk of negative effects of the standard techniques and parameters. The first problem of One Hot Encoding is related to the high number of categories of *NAICS* and *State* variables, which was not reduced since single categories had relevantly high frequencies. For this reason it was decided to investigate Frequency and Weight of Evidence encodings, in order to better handle this problem by generating fewer variables (even though limitations arise also in this case due to the higher possibility of overfitting). Moreover, it may be important to control the optimization of the training function in order to avoid the eventual convergence to coefficients that might negatively affect the prediction.



Thus the following combinations of configurations were evaluated:

1. Encoding (methods suitable for nominal variables like *States* and *NAICS*)
    a. One-hot Encoding. As previously anticipated above, it consists of encoding each categorical variable with different Boolean variables (also called dummy variables) which take values 0 or 1, indicating if a category is present in an observation.
    The advantages of One-hot Encoding are the following.
    - It does not assume the distribution of categories of the categorical variable.
    - It keeps all the information of the categorical variable.
    - It is suitable for linear models.
    The limitations of one-hot encoding are the following.
    - It expands the feature space.
    - It does not add extra information while encoding.

- Many dummy variables may be identical, and this can introduce redundant information.

b. Frequency Encoding. It consists of replacing the categories of each categorical variable with its frequency of observations in the dataset.

$$logit(\pi) = \beta_0 + \beta_1 Term + \beta_2 NoEmp + \beta_3 CreateJob + \beta_4 GrAppv + \beta_5 Portion + \beta_6 DaysToDisboursment + \beta_7 State + \beta_8 NAICS + \beta_9 NewExist + \beta_{10} FranchiseCode + \beta_{11} UrbanRural + \beta_{12} LowDoc + \beta_{13} Recession$$

In this way the categorical variables can be treated as numerical one in the interpretation of the coefficients.

The advantages of Frequency Encoding are the following.

- It is straightforward to implement.
- It does not expand the feature space.

The main limitation of frequency encoding is that we can lose valuable information if there are two different categories with the same amount of observations count. This is because we replace them with the same number.

c. Weight of Evidence Encoding. For each variable, we start by grouping each category alone, and for each group, we calculate the ratio between the number of observations having target variable equal to 1 divided by the total number of observations in the group, defined by p(1), and we compute p(0), defined as 1-p(1). Then, we replace each category with the natural logarithm of [p(1)/p(0)].

$$logit(\pi) = \beta_0 + \beta_1 Term + \beta_2 NoEmp + \beta_3 CreateJob + \beta_4 GrAppv + \beta_5 Portion + \beta_6 DaysToDisboursment + \beta_7 State + \beta_8 NAICS + \beta_9 NewExist + \beta_{10} FranchiseCode + \beta_{11} UrbanRural + \beta_{12} LowDoc + \beta_{13} Recession$$

In this way the categorical variables can be treated as numerical one in the interpretation of the coefficients.

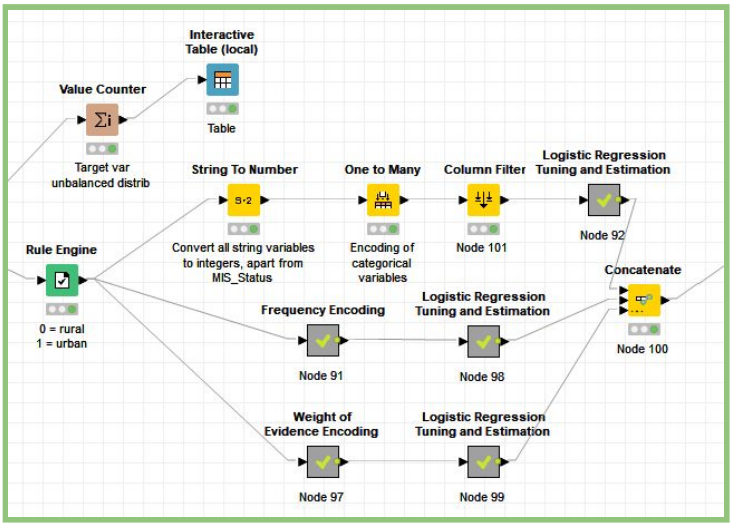
The advantages of Weight of Evidence Encoding are the following.

- It creates a monotonic relationship between the target and the variables.
- It orders the categories on a “logistic” scale, which is natural for logistic regression.

The limitations of Weight of Evidence Encoding are the following.

- It may lead to overfitting.
- It is not defined when the denominator is 0. However, we did not have this problem in our dataset, i.e. for each category of the encoded categorical variables there was at least one observation with the 0 category in the target variable.

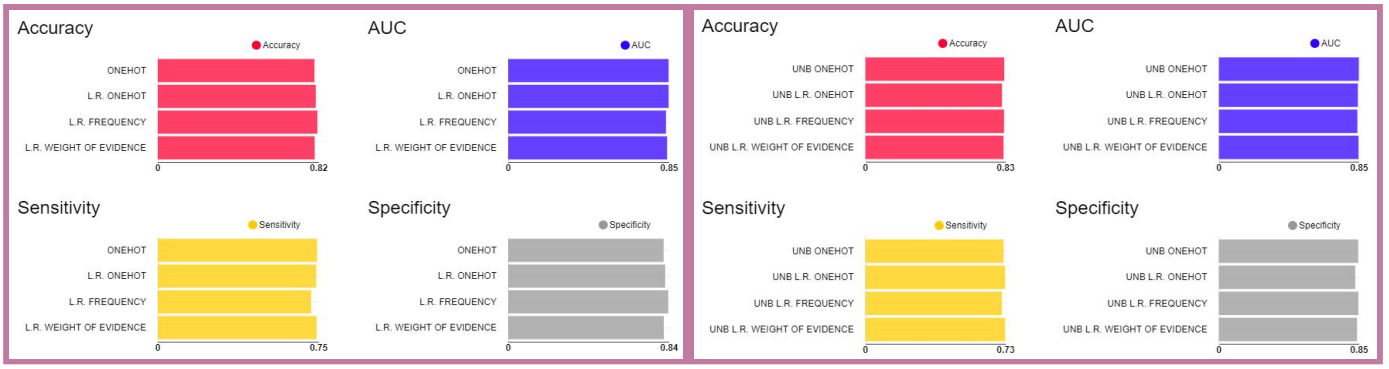
2. Regularization (hyperparameters tuning)
   a. Non-Regularization
   b. L1 Regularization (Lasso Regularization) - Laplace Prior: it forces unimportant coefficients to be zero.
   c. L2 Regularization (Ridge Regression) - Gauss Prior: it is less strict than L1 Regularization, because it keeps the coefficients from becoming too large but does not force them to be zero.

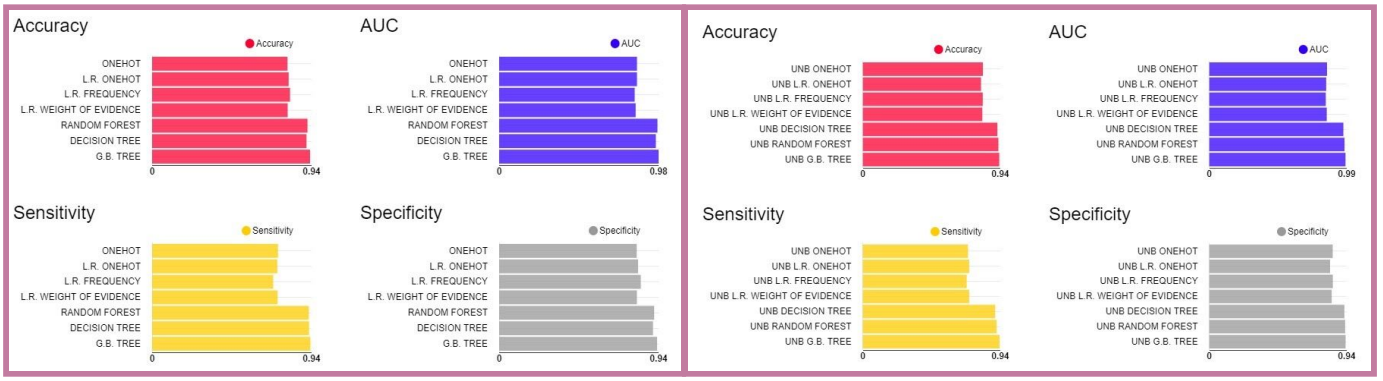


The approach used to choose the best model between all the logistic regressions was about evaluating the model with the highest sensitivity, as deeply explained before, after having maximized the Youden's Index for choosing the threshold of each model.

In particular one can note that in the equal size situation, the model which performs best for the given goals is the one Not Regularized with One Hot Encoding, which reaches a sensitivity of 0.742 and ROC value of 0.847. Whereas in the unbalanced dataset, Weight of Evidence Encoding is the best solution, together with IRLS Solver with Uniform Prior reaching a sensitivity of 0.727 and ROC value of 0.846.

The following are the results for the balanced and unbalanced datasets.





5. Managerial Implications

As has been said, the algorithm with the best performance is the GBTree. For the interpretation of this model, it is useful to look at the splits measures in order to draw managerial implications. In particular, the relative importance is computed in the case of the GBTree algorithm. This is computed by calculating the relative influence of each variable: whether that variable was selected to split on during the tree building process, and how much the squared error (over all trees) improved (decreased) as a result.

The variable *Term* seems to be very influential, having a relative importance of 66.5%. This variable is calculated as a function of the expected lifetime of the assets of the firm, thus we suggest to deeper analyse how the components of the function may impact the prediction of Defaulting companies.

RowID	Relative Importance	Scaled Importance	Percentage
Term	221479.859375	1	0.7409080052961651
Recession	23138.125	0.10447056028161703	0.07740307443042564
State	18804.962890625	0.08490597268614507	0.06290751486062242
NAICS	11985.169921875	0.05411403978536141	0.04009352527588944
Portion	10388.6279296875	0.046905519802132124	0.034752675113977484
DaysToDisboursment	4471.50341796875	0.020189210118640163	0.014958347397506833
GrAppv	4295.1787109375	0.01939308938997063	0.014368495176453363

Moreover, as expected, also the variable *Recession* must be taken into consideration for the final decision. This captures the fact that companies' performance is highly related to worldwide economic factors, which as a result are important to take into consideration when assessing the firm's risk to Default.

It is useful to know that the variables *State, NAICS, Portion, DaystoDisboursement* and *GrAppv* all have almost the same relative importance, following in importance the two variables already described.

It is important to understand how these variables influence the prediction in the Tree-based algorithms. Therefore, it was decided to go through the Random Forest and Decision Tree outcomes as well. In the Random Forest, the feature's ordinal importance is the same as the GBTree one, whereas the Decision Tree output allows to capture more information about the models.

The pattern followed by the algorithm is clear. The Gain Ratio is first reduced by massively splitting the data through the variable *Term*. Then, the macrosets created are classified by slowly selecting little sets of homogeneous elements sharing the same class. This is done mainly through the use of the 5 variables *State, NAICS, Portion, DaystoDisboursement* and *GrAppv* and in some cases also considering *LowDoc* and the other firm specific measures.
The variable *Recession* has been encountered in both the cases, to create the macrosets and to redefine them.

More specifically, by the Tree analysis one can draw the following suggestions:

- The small loans (low *GrApprv<50k*) seem to be riskier and to default more.
- Most of the loans covered for more than 90% by SBA tend to not Default. A *Portion* higher than 70% is still significant. These results may be linked to a reliable SBA evaluation process, applied to ensure a safe investment for such high *Portion*.
- Short Loans should be better analysed as 80% of the ones with *LowDoc* and a *Term* lower than 1 year Defaulted. The percentage lowers to 70% at 3 years and to 55% at 4.
- The variable *Term* does not have a linear relationship with the dependent variable. In fact, the model has shown that there exists a pattern such that the loans with *Term* equal to any number of years (12 months, 24 months, 36 etc.) are mostly classified as non Defaulting. This may be due to the function used to evaluate the *Term* of the loan. As a result we suggest to provide new data and include as additional predictors the variables inside that function so that better insights can be captured.

It is also worth to notice that the cut off value for the Binary Classification probability may be adapted to the risk attitude of the bank so to balance accordingly the trade off between minimizing the mistakes in evaluating as non risky the companies that will eventually default, and not being too much conservative in providing the loans .

The categorical variables such as *NAICS* and *State* are difficult to analyse through the Decision Tree, thus they are covered by looking at the Logistic Regression outputs.

In the previous section it was seen that the best model (highest sensitivity) was the one with the Weight of Evidence encoding. However, as it is very difficult to interpret the coefficients of the categorical variables with this type of encoding, for the managerial implications the coefficients of the logistic regression model with One-hot encoding (dummies) will be interpreted, despite being outperformed by the model with Weight of Evidence encoding.

The dummy variables facilitate the interpretation of the impact on the default probability from each category of the qualitative variables. It is important to point out that in the case of L1 regularization, i.e. Lasso Regression (Laplace prior), that implements *de facto* a feature selection, many dummies of the variable *State* are enforced to be non-significant, for the unbalanced dataset. Thus, one might infer that the impact of this variable on the default probability is not very relevant. Still, the coefficients of the L2 regularization version are analyses because it achieves higher sensitivity and so it is the preferred model with One-hot encoding.

Row ID	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
Row1	1	Term	-17.156	0.067	-257.084	0
Row2	1	NoEmp	-3.586	0.41	-8.735	0
Row3	1	NewExist	0.15	0.008	18.741	0
Row4	1	CreateJob	3.357	0.147	22.855	0
Row5	1	Franchise...	-0.044	0.017	-2.615	0.009
Row6	1	UrbanRural	0.091	0.011	8.43	0
Row7	1	LowDoc	-0.786	0.014	-57.828	0
Row8	1	GrAppv	2.374	0.101	23.595	0
Row9	1	Portion	1.385	0.026	52.879	0
Row10	1	DaysToDi...	-12.454	0.394	-31.607	0
Row11	1	Recession	0.989	0.009	108.311	0
Row12	1	NY	-0.066	0.049	-1.324	0.185
Row13	1	OH	-0.183	0.051	-3.58	0
Row14	1	MO	-0.172	0.053	-3.216	0.001
Row15	1	ME	-0.772	0.071	-10.934	0
Row16	1	TX	0.139	0.049	2.815	0.005
Row17	1	WA	-0.193	0.053	-3.651	0
Row18	1	CO	0.203	0.053	3.839	0
Row19	1	NV	0.522	0.059	8.898	0
Row20	1	CA	0.362	0.049	7.429	0
Row21	1	AZ	0.466	0.053	8.716	0
Row22	1	SC	0.3	0.063	4.755	0
Row23	1	FL	0.68	0.05	13.64	0
Row24	1	MI	0.158	0.052	3.029	0.002
Row25	1	MD	-0.086	0.055	-1.577	0.115
Row26	1	MS	0.042	0.062	0.671	0.502
Row27	1	UT	-0.005	0.053	-0.094	0.925
Row28	1	ND	-1.04	0.078	-13.411	0
Row29	1	IL	0.244	0.051	4.796	0
Row30	1	GA	0.73	0.052	14.001	0
Row31	1	MA	-0.629	0.053	-11.93	0
Row32	1	MN	-0.481	0.054	-8.987	0
Row33	1	NJ	0.073	0.052	1.412	0.158
Row34	1	NH	-0.735	0.059	-12.403	0
Row35	1	NC	0.044	0.054	0.816	0.415
Row36	1	RI	-0.723	0.062	-11.723	0
Row37	1	CT	-0.388	0.057	-6.801	0
Row38	1	VA	0.001	0.055	0.01	0.992
Row39	1	VT	-0.989	0.077	-12.857	0
Row40	1	WV	0.092	0.075	1.229	0.219
Row41	1	ID	-0.152	0.06	-2.529	0.011
Row42	1	AK	-0.385	0.091	-4.219	0
Row43	1	TN	0.393	0.057	6.84	0

Row ID	Logit	Variable	Coeff.	Std. Err.	z-score	P>|z|
Row44	1	AL	0.245	0.061	4.027	0
Row45	1	WI	-0.292	0.054	-5.388	0
Row46	1	PA	-0.491	0.051	-9.616	0
Row47	1	IA	-0.354	0.059	-5.995	0
Row48	1	OK	-0.106	0.059	-1.802	0.072
Row49	1	HI	-0.692	0.071	-9.693	0
Row50	1	KS	-0.326	0.058	-5.607	0
Row51	1	LA	0.009	0.058	0.155	0.877
Row52	1	AR	0.215	0.064	3.379	0.001
Row53	1	IN	-0.149	0.055	-2.703	0.007
Row54	1	KY	0.112	0.06	1.876	0.061
Row55	1	OR	-0.23	0.058	-3.973	0
Row56	1	NM	-0.411	0.069	-5.953	0
Row57	1	SD	-0.639	0.081	-7.859	0
Row58	1	NE	-0.613	0.067	-9.108	0
Row59	1	DE	-0.212	0.082	-2.597	0.009
Row60	1	MT	-1.031	0.069	-14.978	0
Row61	1	WY	-0.776	0.097	-7.988	0
Row62	1	44	0.401	0.016	25.625	0
Row63	1	81	0.306	0.017	18.373	0
Row64	1	45	0.38	0.018	20.819	0
Row65	1	32	-0.012	0.028	-0.417	0.677
Row66	1	31	0.155	0.031	4.969	0
Row67	1	42	-0.038	0.018	-2.029	0.042
Row68	1	23	0.202	0.016	12.273	0
Row69	1	56	0.142	0.02	7.134	0
Row70	1	33	-0.326	0.022	-15.067	0
Row71	1	92	-0.047	0.191	-0.244	0.807
Row72	1	72	0.518	0.017	30.599	0
Row73	1	48	0.228	0.023	9.979	0
Row74	1	54	0.013	0.017	0.76	0.447
Row75	1	53	0.572	0.027	21.203	0
Row76	1	61	0.253	0.038	6.725	0
Row77	1	62	-0.351	0.02	-17.196	0
Row78	1	51	0.232	0.029	7.952	0
Row79	1	71	0.384	0.028	13.514	0
Row80	1	52	0.494	0.031	15.808	0
Row81	1	49	0.178	0.062	2.851	0.004
Row82	1	11	-0.355	0.047	-7.572	0
Row83	1	22	-0.363	0.134	-2.711	0.007
Row84	1	21	-0.856	0.095	-9.033	0
Row85	1	55	0.129	0.24	0.54	0.589
Row86	1	Constant	8.492	0.304	27.931	0

Interpreting the coefficients for the unbalanced dataset, the significant variables are *Term***, NoEmp***, NewExist***, CreateJob***, FranchiseCode***, UrbanRural***, LowDoc***, GrAppv***, Portion***, DaysToDisboursement***, Recession****, all the *State* dummy variables except the ones for NY, MD, MS, UT, NJ, NC, VA, WV and LA (at 0.01, 0.05 or 0.1 significance level), and all the *NAICS* dummy variables except the ones for the codes 32, 92, 54 and 55 (at 0.01 or 0.05 significance level), and the intercept***.

All the other variables do not have a relevant effect on the log-odds, i.e. the ratio of the default probability to the non-default probability, and therefore also on the default probability. We then expect that by plotting these variables against the default probability we obtain almost flat curves.

Among the significant variables, the ones with negative coefficients, thereby indicating a decreasing relation between each of these predictors and the estimate of the default probability, are:

- *Term*
- *NoEmp*
- *FranchiseCode*
- *LowDoc*
- *DaysToDisboursement*
- The *State* dummy variables for OH, MO, ME, WA, ND, MA, MN, NH, RI, CT, VT, ID, AK, WI, PA, IA, OK, HI, KS, IN, OR, NM, SD, NE, DE, MT and WY
- The *NAICS* dummy variables for 42, 33, 62, 11, 22 and 21

whereas the ones with positive coefficients, thereby indicating a increasing relation between each of these predictors and the estimate of the default probability, are

- *NewExist*
- *CreateJob*
- *UrbanRural*

- *GrAppv*
- *Portion*
- *Recession*
- All the other *State* dummy variables
- All the other *NAICS* dummy variables.

From these results, the following implications can be drawn that might be useful for a bank when deciding whether to approve a loan or not.

- Ceteris paribus, the higher the *Term*, the lower the probability of default. This matches the results obtained from the Tree models, but giving a more precise explanation to it may not be easy. In fact, when the *Term* is higher than 240 the lower number of Defaulting companies may be due to the presence of the Real Estate collateral. The rationale for this indicator is that the value of the land is often large enough to cover the amount of any principal outstanding, thereby reducing the probability of default. Though, looking at the data, a significant decrease in the Defaulted companies occurs already for loans with *Term* greater than 96 months.
- Ceteris paribus, issuing a one-page application through the “LowDoc Loan” program rather than a more detailed application might reduce the probability of default. This might be due to the fact that companies’ appreciation of the opportunity to declare less information to the bank might lead them to prioritize the payment of such loans with respect to other loans. However, one must take into account that it is possible to follow the “LowDoc Loan” program only for loans under $150,000. Therefore the decreasing impact on the default probability might capture in part the fact that loans with lower principal are normally less risky, considering that, despite some outliers, most firms in the dataset are of the same size, measured with the NoEmp variable (NoEmp < 50 for approximately 96.5% of the observations).
- Ceteris paribus, for most of the States in the Northern and Northeastern United States, companies based in one of these States generally present a lower likelihood of default. This in part confirms what was seen in the US heatmap by default rate in the data visualization part. It is not easy to identify a possible reason for this pattern. For example, further investigation is needed to check whether the different tax policies implemented over time in the different States might have contributed to such a pattern.
- Ceteris paribus, companies operating in the sectors “Wholesale trade”, “Manufacturing”, “Health care and social assistance”, “Agriculture, forestry, fishing and hunting”, “Utilities” and “Mining, quarrying, and oil and gas extraction” are significantly less risky. The same effect is shown for firms with relatively more employees and longer waiting times for the disbursement to happen. These coefficients are hard to interpret, but show important results. Indeed, the interpretations could be many. For instance, companies with more employees could signal more established companies, with greater expertise which decreases the probability of default. One could also think that companies that are able to wait longer for the Disbursement to occur while remaining healthy and not Defaulting possess greater financial stability. From a social responsibility perspective, one could infer that companies with a higher number of employees are more reluctant to Default to protect the financial well-being of employees.
- Ceteris paribus, new companies and companies active in the recession period are riskier, as was trivally expected.
- Ceteris paribus, the higher the *GrAppv* variable, the higher the default probability. This may be due the fact that higher principals in general lead to riskier loans for

firms of similar size. Higher principals mean higher monthly payments which clearly make each payment more difficult to fulfil.

- Ceteris paribus, the higher the *Portion* variable, the riskier the company. This is in contrast with what was expected and noticed in the data visualization part, therefore it is possible that this variable actually captures also the effect of some omitted variable or interaction term. For instance the repercussions of defaulting on high portion loans may be less severe than low portion loans as less money is owed to banks which are private institutions as opposed to the SBA which is a government institution built on supporting small business and is not as actively pursuing a profit.

6. Limitations

A limitation of the analysis is relative to the lack of the credit risk of the borrowers. Within the past few years, SBA has collected and evaluated Fair Issac (FICO) credit scoring of guarantors and borrowers. If a borrower or guarantor is not a person, then a Dun and Bradstreet score is obtained. Many financial institutions now rely upon credit scores when making smaller loans. If this information was presented, the analysis could have been more accurate as it presents an aggregate measure of trustworthiness.