UNIVERSIDADE DE BRASÍLIA
DEPARTAMENTO DE ECONOMIA

GUSTAVO MEDEIROS FERREIRA DOS SANTOS

MACHINE LEARNING MODELS FOR BANK FAILURE CLASSIFICATION USING DIFFERENT TECHNIQUES TO DEAL WITH IMBALANCED DATASET

BRASÍLIA
13 DE DEZEMBRO DE 2021

Gustavo Medeiros Ferreira dos Santos

Machine Learning Models for Bank failure Classification using different techniques to deal with imbalanced dataset

Master thesis presented to obtain Master Degree on Economics Master Program from Universidade de Brasilia.

Supervisor: Daniel Oliveira Cajueiro

Universidade de Brasília – UnB
Departamento de Economia

Brasília
13 de dezembro de 2021

Gustavo Medeiros Ferreira dos Santos

Machine Learning Models for Bank failure Classification using different techniques to deal with imbalanced dataset/ Gustavo Medeiros Ferreira dos Santos. – Brasília, 13 de dezembro de 2021-

37 p. : il. (algumas color.) ; 30 cm.

Supervisor: Daniel Oliveira Cajueiro

Master Thesis – Universidade de Brasília – UnB
Departamento de Economia, 13 de dezembro de 2021.

1. Bank failure. 2. Imbalance data. 3. Machine Learning111 4. Classification models I. Daniel Oliveira Cajueiro. II. Universidade de Brasília. III. Faculdade de Economia, Administração e Contabilidade. IV. Mestre

CDU 02:141:005.7

Gustavo Medeiros Ferreira dos Santos

Machine Learning Models for Bank failure Classification using different techniques to deal with imbalanced dataset

Master thesis presented to obtain Master Degree on Economics Master Program from Universidade de Brasilia.

Trabalho aprovado. Brasília, 13 de dezembro de 2021:

Daniel Oliveira Cajueiro
Orientador

Marina Delmondes de Carvalho Rossi
Membro Interno

Regis Augusto Ely
Membro Externo

Brasília
13 de dezembro de 2021

Resumo

A previsão de falência bancária não é uma tarefa trivial. Não consiste apenas em escolher o melhor modelo, mas também a melhor forma de tratar o conjunto de dados altamente desequilibrado. Dados desequilibrados referem-se a um problema em que o número de observações pertencentes a uma classe é consideravelmente maior do que o das outras classes. É um desafio relativamente novo nos campos industrial e acadêmico porque muitas técnicas de aprendizado de máquina não têm um bom desempenho. Este trabalho tem como objetivo comparar os resultados de diferentes métodos de balanceamento, ou seja, *Random UnderSampling*, *Random OverSampling* e *Synthetic Minority Oversampling Technique* em um problema de classificação de bancos saudáveis e inadimplentes usando um painel de dados filtrado. O painel de dados consiste em vinte anos de instituições financeiras brasileiras e principais características econômicas que vão de 2000 a 2019. Com configurações de validação e classificação adequadas, são treinados modelos Logit com diferentes regularizações e modelos ensemble, como *Random Forest* e *Gradient Boost* em todos as três base de dados balanceados de forma diferente. A principal contribuição deste trabalho é a utilização de um filtro em dados em painel como primeiro passo para reduzir o problema de desequilíbrio.

Key-words: Falência bancária. Dados desbalanceados. Aprendizado de Máquinas. Modelos de classificação.

Abstract

Bank failure prediction is not an easy task. It consists not just in chosing the best model but also the best way to treat the highly imbalanced dataset. Imbalanced data refers to a problem where the number of observations belonging to one class is considerably higher than the other classes. It is a relatively new challenge in both industrial and academic fields because many machine learning techniques do not have a good performance. This work aims to compare the results of different balancing methods namely Random UnderSampling, Random OverSampling and Synthetic Minority Oversampling Technique on a healthy and default banks classification problem using a filtered panel data. The panel data consists in twenty years of Brazilian financial institutions and major economic features ranging from 2000 to 2019. With a proper validation and classification settings it is trained Logit models with different regularizations and ensemble models such as Random Forest and Gradient Boost on all three different balanced datasets. The major contribution of this work uses a filter in panel data as first step to reduce imbalance problem.

Key-words: Bank Failure. Imbalanced data. Machine learning. Classification models.

List of Tables

List of Figures

Summary

1 Introduction

The banking sector is one the most important sectors to create more wealth and it has a major share in the financial field which needs to be tracked constantly. A single bank failure impacts financial stability and harms economic growth in all others sectors (Erdogan and Akyüz 2018). Bank failure prediction is essential to avoid major economic crisis and developing a predictive bankruptcy model have become an important research field of Economics Science. Since bank failure is a very rare event, most studies faces an imbalance data problem. Data imbalance occurs when the proportion of one class is extremely small compared to the other class. Using imbalanced data, in which non-bankruptcy cases dominates bankruptcy cases, to train classification models results in a model predicting every bank as solvent. This leads to a poor performance model and huge bias in classifying new samples.

This work focuses primarily on the misleading proportion between healthy and default Brazilian banks. An approach that filters the extensive panel data as a starting point to deal imbalanced datasets introduces a equal distribution of observation inside each bank even though the classes of banks continues imbalanced. Once we apply this filter, we use the most used balance techniques on training set in order to achieve equal porportion between classes. The final balanced data sets trains a roll of five distinct models. Then, each model trained in different balanced dataset gives scores to compare methods and algorithms.

While most studies regarding failure prediction focuses on creating better classification algorithms, relatively little research explores the imbalance dataset problem and techniques used to balance the dataset. Even though most techniques explored in literature are similar and have very little variation between studies, they can lead to different models and results (Chawla, Lazarevic, et al. 2003; Batista, Prati, and Monard 2004; Drummond, Holte, et al. 2003) . This study offers a first step in data preparation in order to avoid imbalance between each banks observations. Before use any balance specific metodology on banks a filter is applied to capture only the most recent and relevant information from each bank. With a filtered panel we aim the selection of the best suitable technique to deal with imbalanced banks distribution and the choice of the best model to predict bank failure.

Our dataset uses data from financial institutions released quarterly by the Central Bank of Brazil (BCB). We collected asset, accounting information, income statement and capital information data for all banks available on BCB portal from the first quarter of 2000 until the fourth quarter of 2019. This data contains important information about all financial institutions in Brazil. We included interest rate, exchange rate, GDP, among others, 60 economic variables from several different sources in order to control external

market factors. Among the variables are nominal interest rate, inflation indexes, sector GDP variation, exchange rate, investment rate and current transactions.

Since we want to predict the default occurrence in future, we resized the predictors database (x) in relation to the dependent variable (y). The goal is to create a model that predicts, with the information in t, the state of the bank for a period in t + 4 quarters, that is, 1 year ahead. The dependent variable is positioned to be on the same line as the X predictors from a year ago and thus, we eliminated the first 4 lines of the variable Y and the last 4 lines of the variables X for each bank in the base. We selected only those banks that have more than 8 observations, meaning at least 2 years of data, leaving us with 274 banks, in which 16 went bankrupt during the period analyzed and 258 were healthy banks. In order to catch more relevant information preceding the failure event, we kept only the most recent data, using a two year window, totalizing 8 observations for each bank. Thus, we transform a imbalanced data panel into a filtered data panel resulting in a database with 2192 observations and 253 features.

After properly standardize, this final imbalanced data set is transformed over three balance techniques. We present three different methods to deal with imbalanced data: (i) random under sample, which will randomly eliminate samples from the majority class until both classes have similar proportion inside the training set ; (ii) random over sample, that do the opposite and will create multiples observations copies from minority class until minority class reaches a similar observation number as majority class does; (iii) and Synthetic Minority Oversampling Technique (SMOTE), which creates new minority class data based on feature space similarities from the minority samples.

Each balance method results in a different data set which trains a list of statistic models and machine learn algorithms. The models used were Logit Lasso, Logit Rigde, Logit Elastic Net, Gradient Boost and Random Forest. Most previous studies on bank failure prediction have usually adopted random under-sampling or adjustment on weights to the misclassification cost, inversely proportional to class distribution (Chawla, Japkowicz, and Kotcz 2004). This work provides a starting point in imbalanced data problems. A filtered panel data is an alternative approach not explored in most academic papers regarding this kind of problem.

Modeling relies as much on data representation as on the choice of techniques and model structures. Real-world data often contain a large number of features, some of which are either redundant or irrelevant to the given tasks (Fayyad, Piatetsky-Shapiro, Smyth, et al. 1996). The feature construction and selection is an important process that needs knowledge domain, specially when using recent real world data (Zhao, Sinha, and Ge 2009).

Some studies have also attempted to match each failed bank with non-failed banks in several characteristics, such as geographic location and bank size. This method was used by Sinkey (1975) and Tam and Kiang (1992). Lane, Looney, and Wansley (1986) and

Thomson (1992) used most available failed banks during a particular period and sampled a comparable number of surviving banks in the same period. Both techniques are examples of random under sample. He and Garcia (2009) investigates the influence of imbalance ratio on the performance of four re-sampling strategies to deal with imbalanced data sets and found that over-sampling the minority class consistently outperforms under-sampling the majority class when data sets are strongly imbalanced. Zhou (2013) uses japanese banks information with no parameters and no model optimization to test different methods to deal with imbalanced data set and shows that when there are hundreds of bankrupt observations in the dataset, undersampling will perform better than oversampling. However, when there are only dozens of bankrupt cases in the dataset, oversampling method SMOTE is a better choice. For a more extensive review on bank failure prediction models, Balcaen and Ooghe (2006) does a complete review on the classical statistical models and machine learning models used so far for this task.

We consider an approach first focusing on the imbalance problem and then in model selection. This work applies all balance techniques quoted above on our database and train the algorithms mentioned before. The main goal is to provide a model with good failure prediction, that is, focused on prediction of the observations that are minority in a class. With this approach, the model correctly finds the failed banks, but mistakenly labels some banks as "failed". Since the cost of a failed bank incorrectly predicted is way greater than a healthy bank labeled as default, the appropriate metric to evaluate the model is Recall Score.

The outline of the paper is as follows. In Section 2 we detail our dataset and analyze each set of variables that we use. In Section 3 we present the balance techniques, the models, the specifications used for prediction and the procedure adopted to choose the parameters. We then show our general results and examine the best models in Section 4. Lastly, in Section 5, we present a brief conclusion concerning our findings.

2 Data and Features

2.1 Data

In order to build a bankruptcy banks database, we aggregate the data[1] from financial institutions released quarterly by the Central Bank of Brazil (BCB). These data contains information of all institutions BCB authorized to operate and that are in normal operation. The quarterly reports are available 60 days after the end of each quarter or 90 days if is the fourth quarter.

The source contains groups of variables for assets, liabilities, income accounts, capital information and bank data. The institutions that are in the dataset are: (i) Prudential Conglomerates and Independent Institutions; (ii) Financial Conglomerates and Independent Institutions; (iii) Independent Institutions; (iv) Institutions with Foreign Exchange Operations [2].

We collect asset, liability, income statement and capital information data for all banks in the above groups from first quarter of 2000 until fourth quarter of 2019. This dataset mix important information about all financial institutions in Brazil, containing variables such as total assets, total liabilities, net income, shareholders' equity, Basel index and fixed assets index.

After grouping all selected variables horizontally and all bank's quarters vertically, we filter the base to select: (b1) Commercial Bank, Multiple Bank with Commercial Portfolio or Savings Banks; (b2) Multiple Bank without Commercial Portfolio or Foreign Exchange Bank or Investment Bank and (b4) Development Bank. We did not included the consolidated type referring to Credit Cooperative (b3) in our dataset since they are only avaible for their associates and follow a different capital regulamentation. The organized base has 90 variables in it and approximately 350 banks that have gone through the dataset in these 20 years.

Now we analyze the historical financial health of each bank. We did a vast research in all banks that left and joined the database in all 20 years analyzed. We surveyed the situation of each one of the banks at the time. Finally, we discovered 26 banks that had judicial reorganization, declared bankruptcy, had their activities closed or in some cases had a very bad financial situation, and were bought by other banks or changed the type of financial institution, in these 20 years.

[1] https://www3.bcb.gov.br/ifdata/
[2] https://www.bcb.gov.br/estabilidadefinanceira/regprudencialsegmentacao

2.1.1 Features

In order to control external market factors, such as interest rate, exchange rate, GDP, among others, we added 62 economic variables from several different sources, such as: Bacen, IBGE [3], FGV [4] and IPEA [5]. Among the variables are nominal interest rate, inflation indexes, nominal and real GDP, exchange rate, investment rate and current transactions.

We also created categorical year and quarter variables in order to control the date; categorical variables if the Fixed Asset or Basel Ratio were lower than the benchmark for the period; categorical variables of the sector and type of bank; categorical variables of city and federative unit, in order to control by region; and variables that control the segment (s1, s2, s3, s4 and s5) multiplied by variables commonly classified as important, such as: total assets, total liabilities, net income, equity, ROE (Return on Equity), provisioned credit, etc., in an attempt to separate important bases by different groups of banks. Finally, we eliminate rows with little information and unnecessary columns. We define our categorical variable "default" with the value of 1 in case the bank declared bankruptcy and 0 otherwise.

In order to avoid the loss of important information we extrapolated some null values using the move average between 3 growth averages of the last 3 quarters. In this way, high growth rates between quarters are prevented from hindering extrapolation.

With the data set complete, we resized the predictors database (x) in relation to the dependent variable (y). Ergo, the goal is to create a model that can predict, with the information in t, the state of the bank for a period in t + 4 quarters, that is, 1 year ahead. We positioned the dependent variable on the same line as the X predictors from a year ago and thus, we eliminated the first 4 lines of the variable Y and the last 4 lines of the variables X, for each bank in the base. Thus, with the database properly treated, we have a database with just under 8523 observations and 253 features and an imbalanced proportion between classes of 5.71%

2.1.2 Filter

Since healthy banks will have extra observations once they are on operation for the whole data set and default bank will have limited quarter information since they had their operation suspend at some point, we are dealing not only with imbalance between default and healthy banks, but also when take into account samples from each bank. Healthy banks observations subsets will be larger than default banks subsets. So we filter only the most recent samples of each bank, limited to 8 observations, consequently, 8 quarters. We also discard banks that have less then 8 samples. In that way every bank in database have

[3] https://www.ibge.gov.br/en/home-eng.html
[4] https://portal.fgv.br/en
[5] http://www.ipeadata.gov.br/Default.aspx

exactly 8 observation, meaning 2 years of most recent data available for our independent variables, which are displayed 1 year before the event of failure or not, indicating that the first observation is 3 years before bankruptcy.

This whole process transforms an extensive panel data into filtered panel data with 2192 observations and 253 features, where each observation represents a bank x at time t_x where $t_x \in T_x = \{1_x, 2_x, ..., t_x, ..., 8_x\}$. However, banks might have their 8 observations on different time space compared to others bank. The point is that all of them have exactly 8 observations. The proportion between classes is now 11.68%, more than twice the older proportion.

The database is still imbalanced, with 274 Financial Institutions which 258 are classified as healthy banks and 16 as default banks. The default banks are: Banco BRJ SA, Banco Azteca do Brazil SA, Banco BVA SA, Banco Cacique SA, Banco GE Capital SA, Banco Gerador SA, Banco Morada SA, Banco Neon, Banco Sterling SA, Credibel, Cruzeiro do Sul, Intercap, Prosper, Rural, Schahin, Banco Santos.

We separate the variables into five groups: (i) Accounting Information, providing data about balance sheet, financial results and cash flow about the Financial Institution; (ii) Capital Information, which informs how the capital is organized and the bank structure regarding the capital; (iii) Variables interacting Accounting and Capital Information; (iv) Economic Features, representing Macro Economics data, such as prices index, GDP variation by sectors and others features related to the big picture; and (v) Calendar and Location, given geographic and date information about the observations.

Table 1 – Features Groups

Feature Group	Count
Accounting Information	69
Capital Information	23
Accounting and Capital Interaction	61
Economic Features	62
Calendar and Location	38

We use this segregation in order to facilitate further analysis. However we are still able to search for individual relevance inside each group. With the database selected, the next step is to deal with imbalance problem in order to train a model without any bias. The next Section focuses on this matter.

3 Methodology

3.1 Balancing Techniques

After applying two years of information filter on our dataset and achieving almost 12% proportion between classes we continue with a imbalanced dataset, now regarding the financial institution distribution. We present three different methods to deal with imbalanced data: (i) random under sample; (ii) random over sample; and (iii) Synthetic Minority Oversampling Technique (SMOTE). Those techniques are the most used in literature when dealing with imbalanced classification (Sun, Wong, and Kamel 2009). All three techniques must apply only on training set and after any normalization or standardization procedure. Regarding our limitation, since we have many features with different values scales, we standardize all features. In the process, each feature is subtracted by the mean value and then divided by it is standard deviation. The formula is given by

$$z = \frac{x - u}{\sigma}$$

where x is the observation, u is the mean and σ is the standard deviation. After standardizes the data set, we apply all three methods only on our training set, creating three different training sets to train our models.

Random Undersampling consists in randomly selecting examples from the majority class and deleting them from the training dataset until a more balanced distribution is reached. This approach may be more suitable for those data sets where there is a class imbalance and a sufficient number of examples in the minority class in order to fit the model. However, vast quantities of data are discarded, leading to loss of information, making the decision boundary between minority and majority instances harder to learn, resulting in a poor overall accuracy score.

Random Oversampling involves randomly duplicating examples from the minority class and adding them to the training set. This technique influences the training process when we have a misleading distribution that affects machine learning algorithms and multiples duplicate examples enhances the fit process. The random oversampling sometimes increases the likelihood of occurring overfitting, since it makes exact copies of the minority class examples. In this way, a symbolic classifier, for instance, might construct rules that are apparently accurate, but actually cover one replicated example. Oversampling also results in a higher computational cost when fitting the model, especially considering the model is seeing the same examples in the training data set again and again. Moreover, Oversampling sometimes decreases the classifier performance and always increases the computational effort.

SMOTE uses a different approach. Instead of simply duplicating examples in the minority class, we synthesize new examples from the existing ones. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique. We first chose a random example from the minority class. Then selects the k's nearest neighbors for that example, usually five neighbors. Subsequently, it selects one of the neighbors and creates a synthetic example at a randomly selected point between the two examples in feature space. The synthetic instances are generated as a convex combination of the two chosen minority class instances. This procedure can be used to create as many synthetic examples for the minority class as required. The approach is effective because new synthetic examples from the minority class are relatively close in feature space to existing examples from the minority class. A general downside of the approach is that it creates synthetic examples without considering the majority class, possibly resulting in ambiguous examples if there is a strong overlap for the classes (Chawla, Bowyer, et al. 2002).

We take the initial training and test (holdout) sets roughly equal to 75% and 25% of each class samples, which makes training and test set with same proportion of healthy and default banks. However we applied the three different techniques on training set, leaving test set without any interference. That way we achieve a balance data set to train our models and a test set more close to real world. For Random Undersampling we have a training database with 192 observations. When Random Oversampling is applied, we achieve a training set containing 3000 observations. SMOTE process gives us a training set of 2928 observations. All of them nearing 50% balance between classes and 253 features.

To train those dataset, we uses Logit models with different types for regularization and decision tree ensemble models, such as Random Forest and Gradient Boost. In total it will be five models (Logit Lasso, Logit Rigde, Logit Elastic Net, Random Forest and Gradient Boost), all of them set to classification problem. Our estimated equation is given by:

$$y_{i,t+4} = \alpha_0 + \sum_{i=1}^{I}\sum_{t=1}^{T} \beta X_{(it)} + e_{it}$$

in which $i \subset I$ represents the i Financial Institution at time $t \subset T$,α_0 is a constant term, $X_{i,t}$ is a matrix containing all candidate variables, $\boldsymbol{B} = [\beta]$ is a vector with all the linear parameters, and $u_{i,t}$ is an error term. The lag between y and X means that we are trying to predict for $t+4$ periods ahead, using the variables from t to $t-8$. We are using 2 years information to predict one year ahead.

3.2 Models

3.2.1 Logit Models

Logistic Classification models essentially adapt the linear regression formula to allow it to act as a classifier. Ridge and Lasso regularization are also known as 'shrinkage' methods because they reduces or shrinks the coefficients, reducing the variance in the model. In most cases only a small number of variables actually affects the dependent variable. In these cases, a sparse representation is more adequate and shrinkage models become particularly useful since they vanishe coefficients associated with the variables that have little or no effect on the dependent variable (Horowitz 2015).

The least absolute shrinkage and selection operator (Lasso) minimizes the mean squared error (MSE), just as OLS. However, Lasso also has a shrinkage coefficient, λ, that forces irrelevant variables to zero (Tibshirani 1996; Santosa and Symes 1986). The parameters are determined by:

$$\hat{\beta} = \hat{\beta} argmin \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right].$$

Lasso is a generic OLS model because if $\lambda = 0$, then $\hat{\beta}$ and $\hat{\beta}_{OLS}$ are the same. On the other hand, if the shrinkage coefficient is equal to some $\lambda_c > 0$ large enough, then all the variables are considered irrelevant and all β_i's are equal to zero. Generically we have, $\lambda_c \geq \lambda \geq \lambda_{OLS} = 0$.

Ridge model, like Lasso, is a generic version of OLS that also uses a shrinkage coefficient λ (Hoerl and Kennard 1970). However, the difference is that in Ridge the penalization term is squared instead of linear, as in Lasso. Its parameters are determined by:

$$\hat{\beta} = \hat{\beta} argmin \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right].$$

Although this model also reduces overfitting presented in traditional models, it differs from Lasso because the penalization factor uses the Euclidean norm. Therefore, it makes harder to vanish the coefficients in the Ridge algorithm and, consequently, it cannot completely eliminate some irrelevant variables. However, one particular characteristic of this model is that it treats correlated variables in a close way.

Elastic Net is also a regularized regression model that combines the restrictions in Lasso and Ridge Models (Zou and Hastie 2005). The parameters are:

$$\hat{\beta} = \hat{\beta} argmin \left[\|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right].$$

Elastic Net becomes a general case of Lasso and Ridge Regression. Therefore, as in the Ridge model, the Elastic Net method makes the loss function strictly convex, forcing it to have a unique minimum.

3.2.2 Ensemble Models

Random Forest is a machine learning method that combines decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees (Breiman 2001). The technique used is called bagging, also known as bootstrap aggregating, which creates and merges a collection of independent, parallel decision trees using different subsets of the training data.

Gradient Boost is also a tree based ensemble machine learning method. However, instead of building each tree independently, decision trees are trained sequentially and subsequent models try to reduce the errors of the previous model in order to gradually improve the predictive power. This sequential models to create one strong model by focusing on the mistakes in the prior iterations is called boosting (Breiman et al. 2017).

Although models based on conventional decision trees reduce the possibility of overfitting by choosing parameters that controls their depth and number of leaves, they also seek to reduce the overfitting of traditional trees by combining different trees.

3.3 Parameters

Since we are working with models that can be tuned by parameters, we use cross validation during train process. We run several hyper-parameters values for each model, such as the regularization ratio for Logit Models, and leafs and decision tree deep parameters for Ensemble Models. We select the parameters that deliver the highest accuracy score for each step inside the validation process. The balance techniques give us balanced training sets and even tough the validation process uses overall accuracy to train the models, the evaluation of each models on test set uses others metrics besides accuracy.

After we select the hyper-parameters, we run the models for the entire training set and obtain the final trained models. Next, we use these models in our test set, *i.e.*, the models trained on different balanced datasets are evaluated on the same test set five times, each time changing the banks distribution between training and test set. We use several metrics to evaluate our models. The main one is the Recall Score (Hossin and Sulaiman 2015) which focuses on predict the default banks correctly, once the cost of a failed bank incorrectly predicted is way greater than a healthy bank labeled as default. Thus, in order to have a better idea of our models, we use three others metrics: accuracy, precision and F1 score.

As mentioned, Recall Score is focused on minority class accuracy. We measure it by the ratio $\frac{TP}{TP+FN}$. It is also know as Sensitivity of the Model. Accuracy is the overall correct prediction for both classes. It is calculated as $\frac{TP+TN}{TP+TN+FP+FN}$. Precision is intuitively the ability of the classifier not to label as positive a sample that is negative, in our case, a healthy bank as a default bank. The measure is given by the ratio $\frac{TP}{TP+FP}$.

Finally, F1 Score is a ratio between Recall and Precision Scores, given by $\frac{2*Precision*Recall}{Precision+Recall}$ or $\frac{2*TP}{2*TP+FP+FN}$.

All models were developed and trained in Python and we use a 512GB SSD-equipped PC with a 2.8GHz Octa-Core and with 16G of RAM.

4 Results

We run for each model for trained on different balanced data set five random experiments. We use the mean between the five experiments in order to achieve robust results and avoid any randomness in our results. The tables below show the mean metrics scores for every model trained on each different balanced training set and tested on test set.

Table 2 – Undersample models trained scores

Random Undersampling Training Set	Accuracy	Recall	Precision	F1 Score
Logit Lasso	0.59	0.85	0.09	0.17
Logit Rigde	0.66	0.74	0.10	0.18
Logit Elastic Net	0.66	0.74	0.10	0.18
Random Forest	0.68	0.69	0.10	0.18
Gradient Boost	0.88	0.35	0.18	0.24
Mean	0.66	0.74	0.10	0.18

Table 3 – Oversample models trained scores

Random Oversampling Training Set	Accuracy	Recall	Precision	F1 Score
Logit Lasso	0.64	0.90	0.11	0.20
Logit Rigde	0.89	0.52	0.23	0.32
Logit Elastic Net	0.89	0.50	0.23	0.32
Random Forest	0.92	0.25	0.25	0.24
Gradient Boost	0.94	0.01	0.20	0.01
Mean	0.89	0.50	0.23	0.24

Table 4 – SMOTE models trained scores

SMOTE Training Set	Accuracy	Recall	Precision	F1 Score
Logit Lasso	0.66	0.90	0.12	0.21
Logit Rigde	0.89	0.53	0.23	0.32
Logit Elastic Net	0.89	0.50	0.23	0.32
Random Forest	0.92	0.25	0.27	0.25
Gradient Boost	0.94	0.03	0.18	0.06
Mean	0.89	0.50	0.23	0.25

4.1 Best Models and Balance Techniques

The results show that in general, linear models achieved better recall scores when compared to ensemble models. Logit Lasso model provides the best Recall Score in every balanced dataset and the highest values when trained on Oversample and SMOTE

methods, even though the overall recall scores from Undersample trained models were higher. SMOTE and Undersample models deliver better Accuracy, Precision and F1 score than models trained on the Undersample data set.

Table 5 exhibits the best models based on Recall Scores, Accuracy, Precision and F1 Score. Alongside with the model used, we identify the balance technique used. Those results shows Logit Lasso models as the best alternative in every different balanced training set.

Table 5 – Models Rank

Rank	Model	Balance Technique	Recall Score	Accuracy	Precision	F1 Score
1st	Logit Lasso	SMOTE	0.90	0.66	0.12	0.21
2nd	Logit Lasso	Oversampling	0.90	0.64	0.11	0.20
3rd	Logit Lasso	Undersampling	0.85	0.59	0.09	0.17

This table shows the three best models

Logit models perfoms good on Undersampled training set, however it shows poor alternative metrics. Even though the most important metric for this problem is Recall Score, too low accuracy and precision indicate an underfit model. This type of problem is common when Undersample is applied. The lost of information affects the correct fit of the model trained. SMOTE training set showed good Recall Scores and better accuracy scores. Logit models were the best models trained using this balance technique.

4.2 Features Overview

Using the SHAP (SHapley Additive exPlanations) library in Python, which is used to explains the output of machine learning model based on optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (Lundberg and Lee 2017) we can have a better look at the features in the model.

In summary, what SHAP does is calculate what the prediction of the model would be without feature i, calculate the prediction of the model with feature i, and then calculate the difference. The change in the model's prediction is essentially the effect of the feature. However, since the order in which a model fits a features can affect its predictions, this is done in every possible order, so that the features are fairly compared. In the end SHAP value is the average of the marginal contributions across all permutations. SHAP values show how much each predictor contributes to the target variable.

To get an overview of which features are most important for a model we can plot the mean absolute value of the SHAP values for each feature to get a standard bar plot. For the Logit Lasso Model using SMOTE training set we can plot the most correlated features to our target variable.

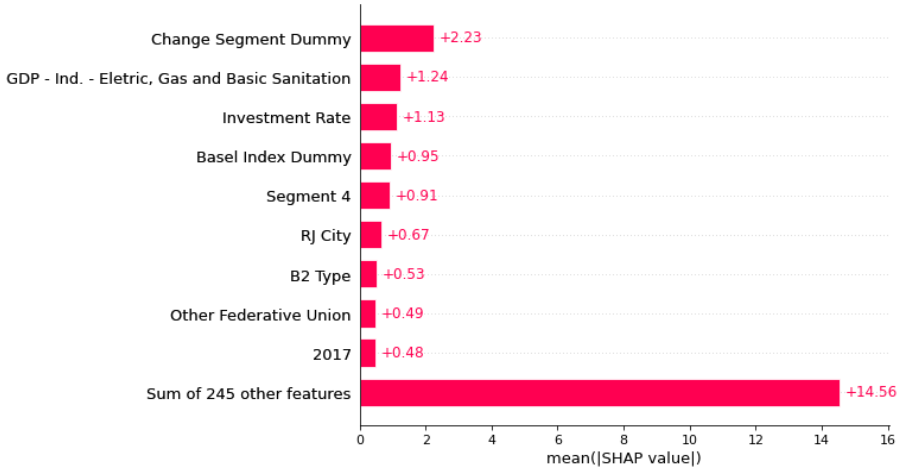

Figure 1 – Most Relevant Features

We can see that different types of variables are relevant to the model. From bank specific characteristics and macro economic features to calendar and geographic information. The Change Segment Dummy allows us to perceive that banks which change segmentation, that is, changes the ratio between assets and GDP are correlated to bank failure. Another feature relevant is RJ City. In our dataset five from sixteen banks were located in Rio de Janeiro, almost one third of all default banks. The Segment 4 dummy shows that small banks are also correlated to failure of a financial institution. Investment Rate and Infrastructure GPD are highly associated with the bank healthy, revealing that Economic features matters. The Basel Index Dummy insight shows us that government regulamentation provides substantial importance on banks failure.

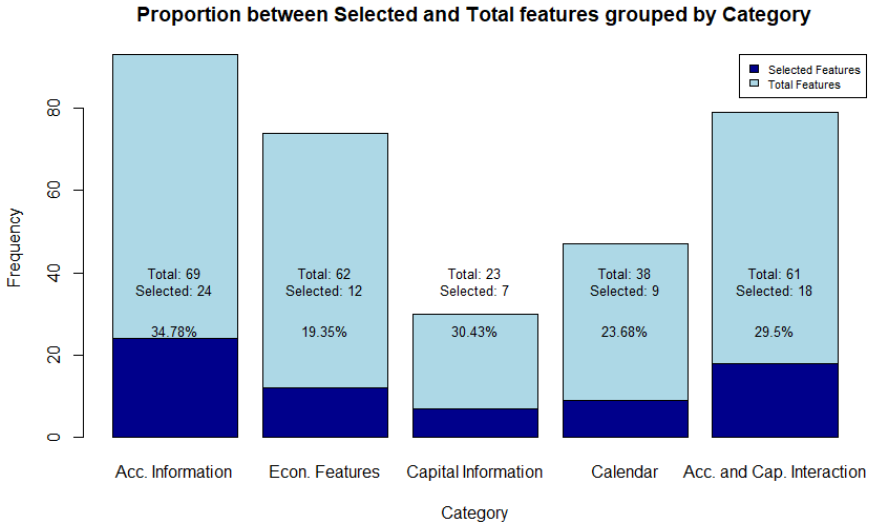

Figure 2 – Features Relavance by Group

We can also take a look at features by group, in order to see relevance of each group. Since the model was trained with Lasso regularization, the least important features will

be forced to zero, leaving us with those that matters.

Even though our model showed high relevance from some Economic Features and Calendar information, these Groups are not relevant as specific information regarding the financial institution. Accounting and Capital information, as well their interactions contributes the most to our classification model. We can conclude that bank failure is more related to administration and governance than general context, excepts for some specific general information about Economic panorama, Region and Date.

5 Conclusion

This work uses imbalanced data to predict bank failure one year ahead of the default event. We show that SMOTE treatment to balance data outperforms simpler ones such as random oversample and random undersample. We tested five different models in three different training set and evaluate them on the same test set. The variables in our database are classified into five groups: Accounting Information, Capital Information, Accounting and Capital interaction, Economic Variables and Calendar and Location features.

Even though all three balance methods achieve good results regarding Recall Scores with Logit Lasso, it is important to emphasize that the balance treatment must take in consideration the specifications of each data set. Since we deal with a small minority class problem, applying undersampling until balance results in too much lost information and low Accuracy and Precision, as our results shows. The alternative is Oversample the minority class. Since random oversample results in an overfitting problem, an alternative is Synthetic Minority Oversampling Technique. Our results shows that in general all models performed better when trained on data treated with SMOTE. Logit models in general performed better than ensemble models. Given the features dimensional size, we can conclude Logit models are enhanced by L1 and L2 regularization.

The most relevant features showed us that every group matters, indicating that bank specification, general economic aspects, region and date have contribution on predicting financial institutions failure. However, there are groups that are more relevant than others. When we group variables into categories, features that brings individual banks information affects the target variable more often compare to general features.

Finally, despite the fact that dealing properly with high imbalanced database is one of the main goals of this work, the approach used on extensive panel data, transforming into a filtered panel data and considering only a gap of most recent observations for each bank showed to be a reliable alternative when the data set is a cross section time series with highly imbalanced distribution observation between classes. Filtering panel data provides a better balance between classes and gives good scores when evaluate models, indicating no loss of crucial information. This method applies not only to imbalanced dataset but general extensive panel data which reduces the size of the dataset in order to train models faster and demand less computional efforts, without losing important information. Another major aspect of this work is to introduce a Brazilian banks dataset. Brazilian banks failure studies and modeling are still in early stages.

However, this work has its own limitations. We chose a class of regularized, easy-to-tune models and with low computational cost of estimation. An alternative approach is to transform panel data into cross section, reducing even more the imbalance problem without apply specific balance techniques and also reducing data size, decreasing the need

for computational cost and leaving space to apply others costly models, as deep learning models. We can explore that approach in a future work.

Bibliography

Balcaen, Sofie and Hubert Ooghe (Mar. 2006). "35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems". In: *The British Accounting Review* 38, pp. 63–93. DOI: 10.1016/j.bar.2005.09.001 (cit. on p. 17).

Batista, Gustavo EAPA, Ronaldo C Prati, and Maria Carolina Monard (2004). "A study of the behavior of several methods for balancing machine learning training data". In: *ACM SIGKDD explorations newsletter* 6.1, pp. 20–29 (cit. on p. 15).

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324 (cit. on p. 26).

Breiman, Leo et al. (2017). *Classification and regression trees*. Routledge (cit. on p. 26).

Chawla, Nitesh V, Kevin W Bowyer, et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357 (cit. on p. 24).

Chawla, Nitesh V, Nathalie Japkowicz, and Aleksander Kotcz (2004). "Special issue on learning from imbalanced data sets". In: *ACM SIGKDD explorations newsletter* 6.1, pp. 1–6 (cit. on p. 16).

Chawla, Nitesh V, Aleksandar Lazarevic, et al. (2003). "SMOTEBoost: Improving prediction of the minority class in boosting". In: *European conference on principles of data mining and knowledge discovery*. Springer, pp. 107–119 (cit. on p. 15).

Drummond, Chris, Robert C Holte, et al. (2003). "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling". In: *Workshop on learning from imbalanced datasets II*. Vol. 11. Citeseer, pp. 1–8 (cit. on p. 15).

Erdogan, Birsen Eygi and Süreyya Özöğür Akyüz (2018). "A weighted ensemble learning by SVM for longitudinal data: Turkish Bank bankruptcy". In: *Trends and perspectives in linear statistical inference*. Springer, pp. 89–103 (cit. on p. 15).

Fayyad, Usama M, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. (1996). "Knowledge Discovery and Data Mining: Towards a Unifying Framework." In: *KDD*. Vol. 96, pp. 82–88 (cit. on p. 16).

He, Haibo and Edwardo A Garcia (2009). "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284 (cit. on p. 17).

Hoerl, A.E. and R. Kennard (1970). "Ridge regression: Biasedestimation for nonorthogonal". In: *Technometrics* 12, pp. 55–67. URL: https://www.jstor.org/stable/1271436?seq=1#page_scan_tab_contents (cit. on p. 25).

Horowitz, J. L. (2015). "Variable selection and estimation in high-dimensional models". In: *Canadian Journal of Economics* 48, pp. 389–407. URL: https://doi.org/10.1111/caje.12130 (cit. on p. 25).

Hossin, Mohammad and Md Nasir Sulaiman (2015). "A review on evaluation metrics for data classification evaluations". In: *International journal of data mining & knowledge management process* 5.2, p. 1 (cit. on p. 26).

Lane, William R, Stephen W Looney, and James W Wansley (1986). "An application of the Cox proportional hazards model to bank failure". In: *Journal of Banking & Finance* 10.4, pp. 511–531 (cit. on p. 16).

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (cit. on p. 30).

Santosa, F. and W. W. Symes (1986). "Linear inversion of band-limited reflection seismograms". In: *SIAM Journal on Scientific and Statistical Computing* 7, pp. 1307–1330. URL: https://epubs.siam.org/doi/abs/10.1137/0907087 (cit. on p. 25).

Sinkey Joseph F, Jr (Mar. 1975). "A Multivariate Statistical Analysis of the Characteristics of Problem Banks". In: *Journal of Finance* 30.1, pp. 21–36. URL: https://ideas.repec.org/a/bla/jfinan/v30y1975i1p21-36.html (cit. on p. 16).

Sun, YANMIN, ANDREW K. C. Wong, and MOHAMED S. Kamel (2009). "CLASSIFICATION OF IMBALANCED DATA: A REVIEW". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.04, pp. 687–719. DOI: 10.1142/S0218001409007326. eprint: https://doi.org/10.1142/S0218001409007326. URL: https://doi.org/10.1142/S0218001409007326 (cit. on p. 23).

Tam, Kar Yan and Melody Y. Kiang (1992). "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions". In: *Management Science* 38.7, pp. 926–947. URL: https://EconPapers.repec.org/RePEc:inm:ormnsc:v:38:y:1992:i:7:p:926-947 (cit. on p. 16).

Thomson, James B (1992). "Modeling the bank regulator's closure option: a two-step logit regression approach". In: *Journal of Financial Services Research* 6.1, pp. 5–23 (cit. on p. 17).

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (methodological)* 58, pp. 267–288. URL: https://www.jstor.org/stable/2346178?seq=1#page_scan_tab_contents (cit. on p. 25).

Zhao, Huimin, Atish P. Sinha, and Wei Ge (2009). "Effects of feature construction on classification performance: An empirical study in bank failure prediction". In: *Expert Systems with Applications* 36.2, Part 2, pp. 2633–2644. DOI: https://doi.org/10.

1016/j.eswa.2008.01.053. URL: https://www.sciencedirect.com/science/article/pii/S0957417408000444 (cit. on p. 16).

Zhou, Ligang (2013). "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods". In: *Knowledge-Based Systems* 41, pp. 16–25 (cit. on p. 17).

Zou, H. and T. Hastie (2005). "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society, Series B* 67, pp. 301–320. URL: https://www.jstor.org/stable/3647580?seq=1#page_scan_tab_contents (cit. on p. 25).