Financial Distress Prediction in Colombian Infrastructure Firms Using Logistic Regression and Support Vector Machines

Jacobo Arango and Carlos A. Caro[a]

[a]University of los Andes, Department of Industrial Engineering, Cra 1 Este No 19A-40, Bogota, Colombia, South America

Abstract

Bankruptcy describes the condition in which a business cannot repay their outstanding debts, which forces them to follow legal and financial liquidation processes where many of the company´s assets are used to repay a portion of their liabilities. Bankruptcies incur severe consequences to shareholders, creditors, and employees. Advanced statistics and machine learning techniques have been used in the past years to predict many business failure cases. Such models have been of great use for investors, creditors, auditors, banks and government policymakers. In this study, logistic regression and support vector machine models have been carried out with the aim of predicting the financial distress risk of firms belonging to the construction industry in Colombia, one-year prior of its occurrence.

Keywords

Financial Distress; Machine Learning; Logistic Regression; Support Vector Machines

Introduction

Bankruptcy refers to the condition in which a company is unable to meet its obligations and pay liabilities that have become due, and thus forces them to undergo a debt reorganization or assets liquidation. Bankruptcies affect shareholders, creditors, and employees, and in extreme cases, they affect a country´s economy and society as a whole.

Financial distress is a state prior to bankruptcy in which a company cannot generate revenue because it is unable to pay its creditors and lenders. Ignoring the signs of financial distress can be catastrophic as there is one point where financial distress cannot be remedied because there is not enough revenue to offset the debt. If this occurs, bankruptcy is the only way out. For corporate and commercial matters, the general legal framework in Colombia is the Colombian Commercial Code (Decree 410/1971). As reported by this outline, one of the causes for declaring financial distress in a company is the reduction of its total equity below 50% of the paid-in capital. If financial distress cannot be relieved within the next 18 months since the remark was made to the board of directors, legal bankruptcy must be declared (Ministry of Commerce, Industry and Tourism, 2019).

With the recent ambitious infrastructure investment plan, the 4G Toll Road Concession Program, the largest of its kind in Latin America today, the Colombian government is seeking to complete 47 projects spanning 8,000 kilometers of roadway, 3,500 kilometers of four-lane highways as well as an expansion of ports and railways (The Worldfolio, 2016). In addition, the National Infrastructure Agency (ANI) is working on the expansion and modernization of nine airports in the country, having the opportunity to build an additional one for Bogotá and Cartagena (International Trade Administration, 2018). In short, the construction industry is one of the biggest branches of the Colombian economy, contributing 6.8% to the country's GDP, which brings attention to the importance of looking after and being aware of the companies belonging to this sector.

Being able to predict whether a firm could go bankrupt or not based on information contained in the financial statements has several advantages. First, it alerts managers for potential problems that seek attention. Second, it serves as a metric for investors to evaluate a firm. Last, it has a significant impact on lending decisions and profitability of financial institutions by helping them decide whether to grant a credit or not to a specific company (Kirkos, 2012).

The first studies made in bankruptcy analyzed whether financial ratios provide useful information (Beaver, 1966), (Altman, 1968). Since then, many different studies decided to use financial ratios for bankruptcy prediction, but instead of using univariate analysis as Beaver did, they started using discriminant analysis (Altman, 1968), (Deakin, 1972), (Edmister, 1972), and (Blum, 1974). However, these conventional statistical methods were valid only under certain restrictive assumptions, like linearity, normality, and independence between input variables, restrictions that led to the blossom of many other techniques by late '70s and early '80s such as the Linear Probability and Multivariate Conditional Probability Models (Logit & Probit) (Ohlson, 1980), (Zmijewski, 1984). Over time, other methods were also developed to estimate the bankruptcy risk of a particular firm, among those, Recursive Partitioning Algorithm in the 80's, Artificial Neural Networks in the '90s, Support Vector Machines in the 2000s, etc.

The purpose of this study is to evaluate and compare the performance of a Logistic Regression and a Support Vector Machine in predicting financial distress on companies of the construction industry in Colombia. In the remainder of this paper, the methodology will be presented where the research data will be described and the SVM and Logistic Regression models will be explained, demonstrating its several superior points over other algorithms. The next section will show the results obtained and the last part of the text will discuss the conclusions and future research issues.

Methodology

Data Collection and Preprocessing

The Database used in the study was obtained from the Colombian Superintendency of Corporations, a regulatory agency of the Ministry of Commerce, Industry, and Tourism. Its main function is to inspect, supervise and control corporations within the country. Data was collected from the financial statements (balance sheet, income statement and cash flow statement) of the different companies and the sample consisted of a total of 1678 construction firms, where 105 went financially distressed during 2017 and 1573 were not financially fragile during the same period.

The input variable selection process consisted of two stages. At the first stage, 9 financial ratios were selected given the advice of an expert and considering that these were the most frequent features used in forty-two different studies made in the topic (Kirkos, 2012). Features were calculated using information from the financial statements of the year 2016, year before the analyzed-predicted period.

In the second stage, we applied stepwise regression to determine statistically significant variables (Table 1). The backward (step-down) selection was carried out to eliminate statistically insignificant predictors. At this approach, the model starts with all candidate variables (x_1 to x_9) and at each step, feature that is least significant is removed. The process continues until no nonsignificant variables remain. The significance level for this study was set to 5%.

Table 1. Variables-Features

Variable	Definition	P-Value	Statistically Significant
x_1	Return on Total Assets (ROTA)	0.00	YES
x_2	Return on Equity (ROE)	0.01	YES
x_3	Current Ratio	0.48	NO
x_4	Debt Ratio	0.00	YES
x_5	Working Capital to Total Assets Ratio	0.00	YES
x_6	Equity to Debt Ratio	0.00	YES
x_7	Assets Turnover Ratio	0.01	YES
x_8	Cash Flow to Debt Ratio	0.08	NO
x_9	Cash Ratio	0.38	NO

Given that one class dominates the other (93.69% vs 6.31%), that is to say, the sample was certainly imbalanced, SMOTE (Synthetic Minority Over-Sampling Technique) algorithm was run to reduce majority class and oversample minority class. The algorithm selects k similar examples, in this case, 5 neighbors, and by randomly shuffling one attribute at a time within the range of that attribute in the selected neighbors, a new instance is created (Lu, et al., 2018). As a result, the group of examples turned into a perfectly balanced sample consisting of 1104 financially distressed companies and 1104 healthy firms. After that, two subsamples were developed. The first (training) subsample made of 80% of the total sample was used as the training set for both Machine Learning techniques. The remaining 20% of the sample was used as the validation set in order to measure the results of our algorithms.

In order to visualize the performance of both of the classification models, the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve was calculated (Figure 5) and (Figure 8). The AUC-ROC measures how much the model is capable of distinguishing one class from the other. The higher the AUC, the better the model is at predicting healthy firms as healthy firms and financially distressed firms as financially distressed firms. The ROC curve is plotted with True Positive Rate (TPR) on the y-axis and False Positive Rate (FPR) on the x-axis, where TPR and FPR are defined by Equation 1 and Equation 2. By decreasing the threshold value at which data points are being classified, we get more positive values thus we increase TPR and FPR; whilst if we increase threshold value, we get more negative values thus we lower TPR and FPR (Kleinbaum & Klein, 2002).

$$TPR = \frac{TP}{TP + FN}$$

Equation 1. True Positive Rate

$$FPR = \frac{FP}{TN + FP}$$

Equation 2. False Positive Rate

An ideal model has AUC equal to 1 which means it has a perfect measure of separability. A poor model has AUC close to 0 which means it is reciprocating the results. It is predicting healthy firms as financially distressed firms and risky firms as healthy ones.

Logistic Regression

History of Logistic Regression goes back to the year 1838 when mathematician Pierre-François Verhulst invented the Logistic Function to describe the growth in populations and the evolution of autocatalytic chemical reactions. Today, Logistic Regression is widely used in statistics, in the field of medicine and social sciences. For example, Logistic Regression might be used to predict the risk of a patient developing a particular disease, given its characteristics (gender, age, weight, blood type, etc.) or it might be used to predict whether a certain person will vote for a political candidate or not based on its age, income, race, etc (Cramer, 2002).

Unlike linear regression, logistic regression does not predict the value of a numeric variable given a set of entries. Instead, it estimates the probability that the given input belongs to a certain class using a logistic/sigmoid function. Once the probability is calculated, the algorithm can be used as a binary classifier, by choosing a threshold value and consequently classifying inputs with probability greater than the threshold as one class and below the threshold as the other (Hosmer, Lemeshow, & Sturdivant, 2013).

Logistic Regression uses a linear equation (Equation 3) with independent variables to predict a value. Nonetheless, because of the nature of the equation, this value can be anywhere between negative infinity and positive infinity. With the aim of getting as an output, a class variable (i.e. 0-healthy firm, 1-financially distressed company), the outcome of the linear equation needs to be converted into a range of [0,1]. For doing this, the result of Equation 3 is inserted into the sigmoid function (Equation 4).

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots$$

Equation 3. Linear Equation

$$g(z) = \frac{1}{1 + e^{-z}}$$

Equation 4. Sigmoid Function

To understand how Equation 4 transforms the values within the range [0,1], Figure 1 shows the graph of the function. Sigmoid function becomes asymptote to y=1 for positive values of z and becomes asymptote to y=0 for negative values of z.

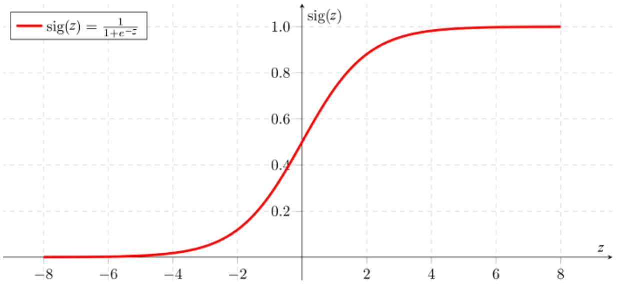

Figure 1. Sigmoid Function

One of the biggest assumptions in which Logistic Regressions rely on is that features should be independent of each other. That is, the model should have little or no multicollinearity (Kleinbaum & Klein, 2002). In order to prove this, correlations among pairs of financial ratios were calculated through the correlation matrix (Table 2).

Correlations among pairs of predictors are limiting, as it is likely that pairwise correlations are insignificant, and yet a linear dependence exists among three or more variables. Therefore, in an effort of further investigating multicollinearity, we proceeded to calculate Variance Inflation Factors (VIF) on financial ratios (Table 3). The VIF measures how much the variance of an estimated jth regression coefficient increases by the existence of correlation with the remaining predictor variables in the model. VIF´s can be calculated by Equation 5, where R_j^2 is the R^2-value obtained by regressing the jth feature on the remaining features (Olusegun Akinwande, Garba Dikko, & Samson, 2014).

$$VIF_j = \frac{1}{1 - R_j^2}$$

Equation 5. Variance Inflation Factor

Besides the AUC, model's performance was measured by its accuracy conducting k-fold cross-validation, a scheme that reduces the variance in the reported accuracy of the model and minimizes the underfitting and overfitting, problems which machine learning models are subject to. The procedure consists of partitioning the database into k equally sized segments (5 in this case). One-fold is held out for validation while the other k-1 folds are used to train the model. Finally, the performance of each of the k models is registered and averaged at the end (Jung & Hu, 2015).

Logistic Regression has been one of the most used statistical methods throughout history because of its many benefits. For instance, it is very efficient in terms of computational resources, it is highly interpretable, it does not require input features to be scaled and it doesn't require any tuning besides the adjustment of the regularization parameter. Because of its simplicity and easy implementation, Logistic Regression is commonly used as a baseline and benchmark for other more complex algorithms (Kleinbaum & Klein, 2002).

Support Vector Machines

Birth of SVM´s goes back to the year 1995 when Statistician Vladimir Vapnik first introduced the basics of the algorithm. Support Vector Machines have several applications, among them, hand-writing recognition, credit rating, time series prediction, fraud detection and so on. SVM´s have been gaining popularity in the last few years due to many provocative advantages and encouraging empirical performances (Gangsar & Tiwari, 2016).

The algorithm is based on the concept of hyperplanes that define separation boundaries between classes. The SVM creates a linear boundary (or hyperplane) between positive instances and negative instances so that it maximizes the distance between the hyperplane and the nearest data points of both classes (Figure 2). The training examples that are closest to the optimal separating hyperplane are called support vectors. These vectors contain all the relevant information required to identify new data points and any other training examples are not essential for defining the binary class boundaries (Figure 3) (Gangsar & Tiwari, 2019).

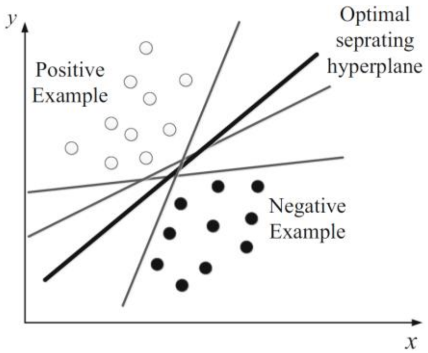

Figure 2. Optimal Separating Hyperplane (OSH)

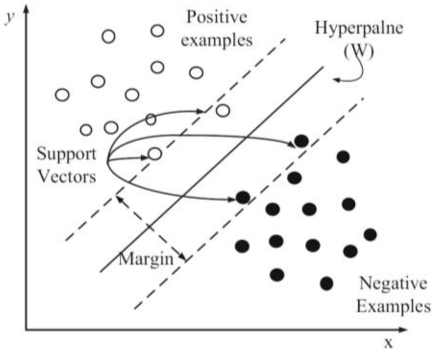

Figure 3. Separation of Binary Classes by OSH

Figure 4 shows the case when a more complex structure is needed in order to make an optimal separation. The original data points, shown on the left side of the figure, are transformed using a set of mathematical functions, known as kernels. In the new setting, the rearranged examples are linearly separable and the SVM is now able to find the optimal line which can separate the positive from the negative objects (TIBCO Software Inc., 2019).

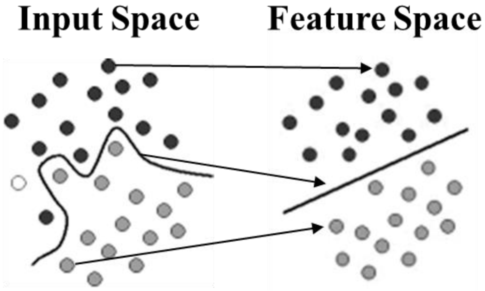

Figure 4. Kernel Transformation

The major advantage of SVM´s over other Artificial Intelligence techniques is its few parameters; namely, regularization parameter “C” and kernel parameter. On the other side, algorithms such as neural networks have a large number of controlling parameters (# of hidden layers, # of hidden nodes, learning rate, epochs, transfer function, etc.…), which prevents them to obtain a unique and global optimum, thing that the SVM´s are capable of finding. In addition, SVM´s captures geometric characteristics of feature space without deriving weights of networks from training data, allowing them to extract the optimal solution with a small training set size (Shin, Lee, & Kim, 2005). In short, SVM techniques have been introduced to many financial applications and have proved superiority over other classifiers including Artificial Neural Networks, Case-Based Reasoning, Multiple Discriminant Analysis, and Logit. Given that, we applied SVM in this study to test its better predictive performance over Logistic Regression in the domain of financial distress risk prediction.

To start with the problem, we standardized original data by removing the mean and scaling it to unit variance. Centering and scaling ensure larger value input attributes do not overcome smaller value inputs. In other words, if a feature has a variance which orders of magnitude are larger than the others, it might dominate the cost function and will impede the machine learning algorithm to learn from other features correctly as expected (Scikit-Learn, 2019).

Once data was ready to train the models, we needed to start choosing kernel parameters. Nevertheless, this is not a simple and trivial task; reason why guideline by (Min & Lee, 2005) was implemented to use grid-search and cross-validation in the search of the parameters. The goal was to identify the optimal choice of C, γ, and d in the case of the polynomial kernel so that the classifiers could accurately predict unknown data from the test set. Cross-validation, in contrast, helped to prevent overfitting.

In this study, grid-search on C, γ, d and kernel type was used using 5-fold cross-validation. Basically, all combinations of C, γ, d, and types of kernels were tried and the one with the best cross-validation accuracy was selected. As suggested by (Min & Lee, 2005), exponentially growing sequences of C and γ were implemented as a practical method to identify optimal parameters. The reason why grid-search was used in the study was because of its simplicity as an exhaustive method. Besides, the computational time required to search for the optimal parameter values is not much more than those needed by any other advanced method or heuristic. Furthermore, grid-search was easily parallelized with the help of an Intel Xeon E5-2630 v3 @ 2.4 GHz x 16 processor, as pairs of C and γ were independent of each other (Hsu, Chang, & Lin, 2003).

Results

Correlation Matrix (Table 2) was calculated to check multicollinearity. Results showed that the only interaction we needed to monitor was the one between variables x3 and x9 referring to the current ratio and the cash ratio respectively.

Table 2. Correlation Matrix

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	1	0.01	-0.01	-0.41	0.19	0.02	0.24	0.04	0.00
x_2	0.01	1	0.00	-0.02	-0.00	0.00	0.02	0.00	0.00
x_3	-0.01	0.00	1	0.02	0.13	0.13	-0.05	0.00	0.44
x_4	-0.41	-0.02	0.02	1	-0.26	-0.08	-0.04	-0.18	-0.07
x_5	0.19	-0.00	0.13	-0.26	1	0.02	0.06	0.14	0.16
x_6	0.02	0.00	0.13	-0.08	0.02	1	-0.02	0.05	0.01
x_7	0.24	0.02	-0.05	-0.04	0.06	-0.01	1	0.08	-0.03
x_8	0.04	0.00	0.00	-0.18	0.14	0.05	0.08	1	0.27
x_9	0.00	0.00	0.44	-0.07	0.16	0.01	-0.03	0.27	1

Moreover, Variance Inflation Factors were estimated for each one of the features (Table 3).

Table 3. Variance Inflation Factors

Features	VIF Factor
Return on Total Assets (ROTA)	1.25
Return on Equity (ROE)	1.00
Current Ratio	1.31
Debt Ratio	1.77
Working Capital to Total Assets Ratio	1.62
Equity to Debt Ratio	1.03
Assets Turnover Ratio	1.73
Cash Flow to Debt Ratio	1.17
Cash Ratio	1.41

A VIF of 1 would mean that there is absolutely no correlation among the jth predictor and the remaining variables. Literature suggests that VIFs exceeding 5 require additional investigation, while VIFs exceeding 10 is clear evidence of serious multicollinearity and requires correction. Considering that, we could rule out the possible existence of multicollinearity among variables, seeing that the highest VIF factor was 1.62 for the Working Capital to Total Assets Ratio.

Accuracy of the Logistic Regression classifier was of 83.51%, having a precision and recall for the healthy and financially distressed firms of (87%, 79%) and (81%, 88%) correspondingly. Precision (Equation 6) talks about how many of the firms that were classified as financially distressed are at risk, while recall (Equation 7) calculates how many of the actual financially distressed firms our model captures by labeling it as a risky firm.

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{Total\ Predicted\ Positive}$$

Equation 6. Precision

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{Total\ Actual\ Positive}$$

Equation 7. Recall

As said before, one of the main advantages of Logistic Regression models is its interpretability, among which odds ratios (OR) are one of its most remarkable readings. The OR represents the odds that an outcome (becoming financially distressed next year) will occur given a particular exposure (an increase of one-unit value in a financial ratio), compared to the odds of that same outcome occurring in the absence of that exposure. An odds ratio of 1 indicates that becoming financially distressed on the next fiscal year is equally likely to occur regardless of the value of that financial ratio. On the contrary, an OR of 1.4 means that the odds of becoming financially distressed is 1.4 times higher if that variable suffers a one unit-value increase, while an odds ratio of 0.8 means that the odds of becoming healthy (opposite of financially distressed) is 1.25 (1/0.8) times higher if that feature is increased by one-unit value (Andrade, 2015). The Odds ratio for the implemented model on firms belonging to the construction industry in Colombia is summarized in Table 4.

Table 4. Odds Ratio

Features	**Odds**
ROTA	0.0000
ROE	0.9503
Debt Ratio	2.6382
Working Capital to Total Assets Ratio	0.4765
Equity to Debt Ratio	0.2049
Assets Turnover Ratio	1.2393

AUC obtained in the implemented model was equal to 0.9315 (Figure 5), a value that reflects the good ability of the algorithm to detect between healthy and risky firms.

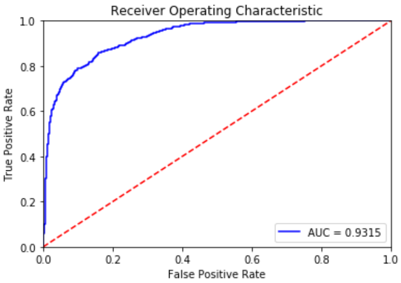

Figure 5. AUC - ROC Curve for Logistic Regression

Moving over to the SVM, we conducted grid-search for the training data as suggested by (Min & Lee, 2005), and we found out that the optimal combination of parameters and kernel, was using a Radial Basis Function with C and γ of 2^{13} and 2^{-1} respectively with a cross-validation rate of 92.6%. After optimal (C, γ) was found, the whole training data was trained again to produce the final model. The prediction accuracy of the validation set turned out to be 93.89% with a precision and recall for the healthy and financially distressed firms of (97%, 91%) and (91%, 97%) correspondingly. Table 5 summarizes the results of the grid-search process using 5-fold cross-validation. Moreover, we proved that even though polynomial functions take longer time in the training stage than the RBF, it provides significantly worse results than the latter. (Kim, 2003).

Table 5. Results of Grid-Search

Kernel	C	γ	d	Accuracy (%)
Linear	2	N/A	N/A	84.39
RBF	2^{13}	2^{-1}	N/A	93.89
Polynomial	2	8	3	88.46
Sigmoid	2^{13}	2^{-13}	N/A	84.16

With the aim of building a measure analogous to the odds ratio in the Logistic Regression, coefficients of the SVM classifier were extracted. Coefficients represent the weights of the input “dimensions”; however, this attribute only works for the linear kernel as the coefficients associated with the separating plane in the feature space are not directly related to the input space in other kernels. In fact, for the RBF kernel, the transformed space is infinite-dimensional (Eigensatz & Giesen, 2006). The coefficients or weights of the hyperplane denote the coordinates of a vector which is orthogonal to the boundary that separates the classes as best as possible. Figure 6 illustrates the importance of each one of the features obtained for the classification task in the Linear SVM.

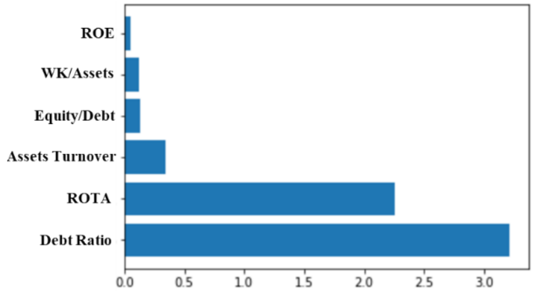

Figure 6. Feature Importance - Linear SVM

Alternatively, Permutation Feature Importance was applied to measure the impact of the different variables in the selected model (RBF Support Vector Machine, Penalty Parameter C: 2^{13} and γ: 2^{-1}). This procedure consists of shuffling feature values, one attribute at a time, and measuring the performance of the model before and after the permutation. Important features are usually more sensitive to the shuffling process and will thus result in higher importance scores (Altmann, Toloşi, Sander, & Lengauer, 2010). Figure 7 shows scores obtained from the permutation feature importance process applied to the Radial Basis Function (RBF) SVM.

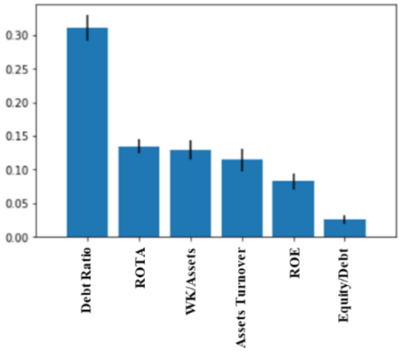

Figure 7. Feature Importance - RBF SVM

As expected, the early detection of bankruptcy or any other kind of financial risk would be explained mainly by profitability and leverage ratios as in this case, where the features that aided both algorithms on predicting correctly next year financial situation were the Debt Ratio and the Return Over Total Assets.

AUC of the RBF Support Vector Machine with optimal choice of kernel function parameters was of 0.9864 (Figure 8). Clearly, models designed to increase the performance (e.g. SVM, Random Forest, etc.) have higher AUC than those who have a less flexible (linear) relation between the predictors and the response (e.g. Logistic Regression, Linear Discriminant Analysis, etc.).

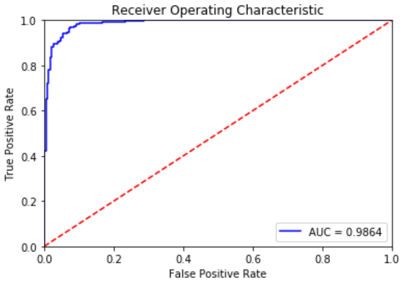

Figure 8. AUC - ROC Curve for RBF SVM

Conclusions

In this study, SVM and Logistic Regression models were applied to the problem of financially distress risk prediction for companies belonging to the construction industry in Colombia. Results obtained in both methods proved to be better than what can be expected by pure chance as the predictive power accomplished by each one of them was higher than the proportion of each one of the classes (50%).

The optimal choice of kernel function parameters for the SVM could be achieved using grid search and 5-fold cross-validation. Selecting optimal parameter values through these techniques, allowed us to build a prediction model with high stability and predictive power, outperforming Logistic Regression by 10% in accuracy. This major difference obtained in both models might suggest relation between predictor variables and response (0-healthy firm, 1-financially distressed company) is likely to be complex enough to be expressed as a linear combination as in the Logistic Regression.

Permutation Feature Importance approach enabled us to close the gap between interpretation and prediction in black box predictive algorithms, a fact that has prevented the widespread use of these techniques in the field of artificial intelligence. This method helped us explain to what extent do the different financial ratios affect the problem of financial distress.

Main influential variables which helped both algorithms determine whether a company would be on financial risk next year or not were primarily Debt Ratio and Return Over Total Assets, indicators that measure the extent of a company's leverage and the efficiency of a company's management at using its assets to generate earnings.

Financial distress prediction studies in Colombia have been limited and oriented to identify relevant factors that explain this phenomenon using multivariate conditional probability models (Logit & Probit), and discriminant analysis techniques (Pérez, González C., & Lopera C., 2013) (Calderon C., 2016) (Berrio Guzmán & Cabeza de Vergara, 2003). In the present study, we provided benchmark values to the predictive performance that could be achieved with different methods besides the above mentioned.

However, this study limited itself to include solely financial ratios, so it would be of great interest to further investigate the impact of incorporating other variables such as the size of the firm, time-series data (more than just one year's previous financial data), macroeconomic variables, and many other. Furthermore, it would be of great use to direct future research towards cost-sensitive classification, as bankruptcy prediction is a topic where misclassification costs are not equal considering that the costs of lending to a defaulter are far greater than the lost-business cost of refusing a loan to a healthy firm.

References

Alpaydin, E. (2016). *Introduction to Machine Learning*. Boston: MIT Press.

Altman, E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 589-609.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation Importance: a Corrected Feature Importance Measure. *Bioinformatics*, 1340–1347.

Andrade, C. (2015). Understanding Relative Risk, Odds Ratio, and Related Terms: As Simple as It Can Get. *The Journal of Clinical Psychiatry*, 1-5.

Beaver, W. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, 71-111.

Berrio Guzmán, D., & Cabeza de Vergara, L. (2003). Verificación y Adaptación del Modelo de Altman a la Superintendencia de Sociedades de Colombia. *Pensamiento y Gestión - Universidad del Norte*, 26-51.

Blum, M. (1974). Failing Company Discriminant Analysis. *Journal of Accounting Research*, 1-25.

Calderon C., E. (2016). *Evaluación de los Modelos de Predicción de Fracaso Empresarial en el Sector Manufacturero Colombiano en los Años 2010-2014*. Bogota: Universidad Nacional de Colombia.

Cramer, J. (2002). *The Origins of Logistic Regression*. Amsterdam: Tinbergen Institute - University of Amsterdam.

Deakin, E. (1972). A Discriminant Analysis of Predictors of Business Failure. *Journal of Accounting Research*, 167-179.

Edmister, R. (1972). An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. *Journal of Financial and Quantitative Analysis*, 1477-1493.

Eigensatz, M., & Giesen, J. (2006). *Insights into the Geometry of the Gaussian Kernel*. Zürich: Swiss Federal Institute of Technology Zürich.

Gangsar, P., & Tiwari, R. (2016). Taxonomy of Induction Motor Mechanical Fault Based on Time Domain Vibration Signals by Multiclass SVM Classifiers. *Intelligent Industrial Systems*, 269-281.

Gangsar, P., & Tiwari, R. (2019). A Support Vector Machine Based Fault Diagnostics of Induction Motors for Practical Situation of Multi Sensor Limited Data Case. *Measurement*, 694-711.

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression*. Hoboken - New Jersey: John Wiley & Sons Inc.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. *Department of Computer Science and Information Engineering, University of National Taiwan*, 1-12.

International Trade Administration. (2018, 08 17). *Colombia - Infrastructure*. Retrieved from Export.gov: https://www.export.gov/article?id=Colombia-infrastructure

Jung, Y., & Hu, J. (2015). A K-fold Averaging Cross-Validation Procedure. *Journal of Nonparametric Statistics*, 167-179.

Kim, K. (2003). Financial Time Series Forecasting Using Support Vector Machines. *Neurocomputing*, 307-319.

Kirkos, E. (2012). Assessing Methodologies for Intelligent Bankruptcy. *Springer*.

Kleinbaum, D., & Klein, M. (2002). *Logistic Regression. A Self-Learning Text*. New York: Springer Publishers.

Lu, P., Ericson, G., Takaki, J., Petersen, T., Sharkey, K., & Martens, J. (2018, 01 09). *SMOTE*. Retrieved from Microsoft Azure: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

Min, J., & Lee, Y.-C. (2005). Bankruptcy Prediction Using Support Vector Machine With Optimal Choice Of Kernel Function Parameters. *Expert Systems With Applications*, 603-614.

Ministry of Commerce, Industry and Tourism. (2019). Decree 410/1971. In *Colombian Comercial Code*.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 109-131.

Olusegun Akinwande, M., Garba Dikko, H., & Samson, A. (2014). *Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis*. Zaria: Open Journal of Statistics.

Pérez, J. I., González C., K. L., & Lopera C., M. (2013). Modelos de Predicción de la Fragilidad Empresarial: Aplicación al Caso Colombiano para el Año 2011. *Perfil de Coyuntura Económica - Universidad de Antioquia*, 205-228.

Scikit-Learn. (2019). *Standard Scaler*. Retrieved from Scikit-Learn. Machine Learning in Python: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Shin, K., Lee, T., & Kim, H. (2005). An Application Of Support Vector Machines In Bankruptcy Prediction Model. *Expert Systems With Applications*, *28*(2), 189-394.

The Worldfolio. (2016, 03 09). Colombia Thinks Big With $70 billion Infrastructure Program. *The Worldfolio*.

TIBCO Software Inc. (2019). *Support Vector Machines (SVM)*. Retrieved from Stat Soft: http://www.statsoft.com/Textbook/Support-Vector-Machines#overview

Zmijewski, M. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, 59-82.