

Understanding bank failure: A close examination of rules created by Genetic Programming.*

Alma Lilia García-Almanza
Banco de México

Biliana Alexandrova-Kabadjova
Banco de México

Serafín Martínez-Jaramillo
Banco de México

Abstract

This paper presents a novel method to predict bankruptcy, using a Genetic Programming (GP) based approach called Evolving Decision Rules (EDR). In order to obtain the optimum parameters of the classifying mechanism, we use a data set, obtained from the US Federal Deposit Insurance Corporation (FDIC). The set consists of limited financial institutions' data, presented as variables widely used to detect bank failure. The outcome is a set of comprehensible decision rules, which allows to identify cases of bankruptcy. Further, the reliability of those rules is measured in terms of the true and false positive rate, calculated over the whole data set and plot over the Receiving Operating Characteristic (ROC) space. In order to test the accuracy performance of the mechanism, we elaborate two experiments: the first, aimed to test the degree of the variables' usefulness, provides a quantitative and a qualitative analysis. The second experiment completed over 1000 different re-sampled cases is used to measure the performance of the approach. To our knowledge this is the first computational technique in this field able to give useful insights of the method's predictive structure. The main contributions of this work are three: first, we want to bring to the arena of bankruptcy prediction a competitive novel method which in pure performance terms is comparable to state of the art methods recently proposed in similar works; second, this method provides the additional advantage of transparency as the generated rules are fully interpretable in terms of simple financial ratios; third and final, the proposed method includes cutting edge techniques to handle highly unbalanced samples, something that is very common in bankruptcy applications.

1 Introduction

Corporate bankruptcy has been an active area of research in finance. Furthermore, in the aftermath of the Lehman Brothers default and its consequences on the global financial system, this topic has attracted even more attention from regulators and researchers. This unprecedented even in recent times

*The authors would like to thank Bolanle Shadia Abudu and the Center for Computational Finance and Economic Agents (CCFEA) for allowing us to use the data.

has brought to the attention of governmental and financial authorities the need to strengthen the regulatory framework. Despite that the principles of the Basel Committee on Banking Supervision or Basel II [1], has been revised just few years ago, today regulators' agenda points toward what has been informally known as Basel III. Consequently, the need of timely signals for supervisory action and the development of tools that helps to determine which financial information is more relevant to predict distress is gaining popularity.

In the modern interconnected economies, the default probability of individual financial institutions is not an independent event, nevertheless an early warning system that could predict bankruptcy with relatively good accuracy could be a crucial element in preventing contagion. To that end, it is important to have a reliable classification method to assist financial distress. In this context, it is worth mention that Breiman et al. [2] pointed out that the main goals of classification are: 1) to generate an accurate classification model able to predict unseen cases and 2) discover the predictive structure of the problem. The last goal means that the method should provide an understanding of the variables and conditions that control or are involved in the event. In many cases both objectives are important, for that reason a good classifier has to be able to produce accurate classifications, with the limitation of the data and provide understanding of the predictive structure of the data. This idea is extended even further in [3], which discusses the distinction between models designed for prediction and models designed for understanding. The authors remark that sometimes models, which provide a poor representation of the solution, have better predictive power than methods that provide comprehensible models, e.g. the Artificial Neural Networks (ANN) [4].

It has been a commonplace for a while, the criticism to classification and prediction models which are opaque and do not provide interpretability of such models. ANNs are the typical example of such good performance poor interpretability models. Although, it is not possible to classify such type of approach as black boxes any more, there is a strong sense that having models which allows a better understanding of how the prediction models work is preferable than having a competitive model only in performance terms which provides very little information about how the model works.

In this study our objective is to achieve both requirements:

the proposed method should be accurate in predicting and it should be able to give insights in the way financial institution bankruptcy is detected. To that end, we propose a machine learning approach called Evolving Decision Rules (EDR) [5], based on Genetic Programming (GP) [6]. This method generate a set of comprehensible decision rules, aimed to detect bank failure, from a data set, obtained by the US Federal Deposit Insurance Corporation (FDIC). The data set consists of limited financial institutions' variables widely used to predict bankruptcy. This paper presents two experiments, the first one makes a quantitative and a qualitative analysis of the results with the aim to rate the predictive ability of the variables in the data set. The second experiment shows the performance of our approach using the Receiving Operating Characteristic (ROC) [7] curve and exhibits a set of decision rules demonstrating the interpretability of the solution.

The remainder of this paper is organized as follows: section 2 contains an overview of the literature related to bankruptcy prediction for banks, section 3 presents a brief description of the Evolving Decision Rules (EDR) approach, while section 4 describes the experiments to test our method. Finally, we present our conclusions in section 5.

2 Bankruptcy prediction for banks

Given the actual state of the global economy and having banks at the centre of a worldwide debate it is more important today than ever to have models which can be used to predict bankruptcy for banks as they are tremendously important institutions in our globalized financial system. We believe that as important as having models that predict bankruptcy accurately is to have models that shed some light on how such bankruptcy can be predicted. More importantly, it is crucial to be able to create simple rules which can warn us before bankruptcy actually happens.

The present study is very much related to [8] crucially in terms of the data set used to generate the rules and on the re-sampling method used in order to make the performance comparison with other related techniques fairer. The information used to generate the decision rules for the bankruptcy prediction for banks consists of simple financial ratios which are very common and well understood within the finance community. Nevertheless, there is nothing that prevent us from using information of different nature to the financial ratios. In fact it would not require a change in paradigm and an exploratory study can be conducted in order to cope with the size of the possible extended input information for the rule generation mechanism.

The first work which used financial ratios to assess the financial health of companies is [9]. After such seminal work, the works by Altman ([10]) and the Z-score became the reference works in this field. Moreover, as it has been proved in [8] the Fisher's Discriminant ([11]) is a quite competent tool in this precise prediction task.

3 Evolving Decision Rules

This section briefly describes the Evolving Decision Rules approach (EDR). It starts by describing the idea behind this approach and finally a brief description of each step is explained. For more detailed information the reader is referred to [12], [5].

Evolving Decision Rules is an evolutionary technique. The main objective of this approach is to generate a set of understandable rules whose purpose is to classify a single class, let us call this set of rules "repository", which is represented as R . Different classifications can be obtained by using distinct subsets of rules from R , these subsets are conformed by rules whose precision reaches a specific threshold. This feature is very useful for users because they can choose between a conservative or a liberal prediction (high and low precision respectively), according to their necessities.

Let us list the main steps of the EDR approach.

1. - Creation of the initial population. Create a population of random decision trees
2. - Extraction of rules. Analyze every decision tree to delimit their rules and select useful patterns
3. - Rule simplification. This step analyses the new rule to remove the redundant and vacuous conditions
4. - Adding new rules to the repository. Detect new rules by comparing the new one with the existing rules in the repository
5. - Creation of a new population. Create the new population taking as parents the rules that are stored in the repository. The next generation will be created using the mutation operator and hill-climbing.
6. - Testing EDR. Once the evolutionary process has finished, EDR is tested by using another data set. It is evaluated by using sub-collections of R . Those rules are grouped according to their precision.

4 Experimental section

To achieve our goals the EDR approach was applied for generating rules to predict bank failure by means of two series of experiments. The first experiment pursues the objective of analyzing the variables in the data set for choosing the most relevant features to generate the model. The second experiment measures the performance of the EDR system in terms of the Area Under the ROC curve (AUC) curve and shows the most successful rules. More detailed description of the experiments is given in the following sections.

4.1 Training data description

The data set to train and test our approach is composed of financial ratios, these are listed in Table 1. The data comes

from the Federal Deposit Insurance Corporation (FDIC) and it was kindly provided by the Centre for Computational Finance and Economic Agents (CCFEA), University of Essex.

To tune the parameters for running the EDR system a series of preliminary experiments was performed. The experiment consisted of fifty running using different values for the parameters. The best results were obtained using the values described in table 2.

4.2 Experiment I

The purpose of the first experiment is to select the most suitable variables to create the model. We devote substantial time to this work because removing irrelevant features from the data set helps to avoid confusion or noise trying to integrate no-relevant information into the model. Additionally the reduction of variables speeds the evolutionary process. To determine the predictive power of each variable a quantitative and a qualitative analysis were carried on.

The quantitative analysis pursues to measure the frequency and the impact of each variable in the resultant rules. Thus, a data set D_0 was randomly re-sampling ten times to create ten different data sets D_1, D_2, \dots, D_{10} , this technique is well known as bootstrapping [13]. Given that, the EDR method is a stochastic process, this was performed ten times per each data set D_i where $i = 1, 2, \dots, 10$. In total EDR was performed 100 times producing the same number of rules' repositories, let us call them R_{ij} where i indicates the number of the data set and j the number of running. Each repository R_{ij} contains 50 rules, let us define $Rep = \bigcup R_{ij} \forall i, j = 1, 2, \dots, 10$, such repository contains in total 5,000 rules. An index to measure the importance of each variable was created, notice that the index involves the frequency as well as the precision of the rule. Let us define the usefulness index as follows: $Index(var_k) = \frac{1}{n} \sum x_{ik} \times p_i \forall i = 1, 2, \dots, n$ n is the total number of analyzed rules r_i is a decision rule where $i = 1, 2, \dots, n$ p_i is the precision of the rule

Table 1. List of variables

Var. Num.	Name	Description
1	NIM	Net interest margin
2	NII	Non-interest income to earning assets
3	NIX	Non-interest expense to earning assets
4	NOI	Net operating income to assets
5	ROA	Return on assets.
6	ROE	Return on equity
7	ER	Efficiency ratio
8	NCA	Non-current assets plus other real estate owned to assets
9	CUS	Cash plus US treasury and government
10	EC	Equity capital to assets
11	CR	Core capital leverage ratio

Table 2. Summary of Parameters for running EDR

Parameter	Value
Population size	500
Initialization method	Growth
Generations	50
Crossover Rate	0.8
Mutation Rate	0.05
Selection	Tournament (size 2)
Control bloat growing	Tarpeian method
Evaluate rules varying precision	.02

r_i var_k is a variable where $k = 1, 2, \dots, 11$

$$x_{ik} = \begin{cases} 1 & \text{if } var_k \in r_i \\ 0 & \text{if } var_k \notin r_i \end{cases}$$

Table 3 shows the usefulness index for each variable and the percentage of repositories whose rules did not use that variable at all.

As can be observed from table 3 the index average was 0.192, then the variables whose performance was lower than the average could be candidates to be eliminated from the data set. The variables NII, ROA, ER, CUS y CR are in that case. In the same vein those variables were not used at all in the 11%, 15%, 33%, 5% and 17% of the cases respectively. However, it is difficult to establish a minimum threshold for that index, for that reason it was needed to complement this measure with a qualitative study.

The qualitative test consists in analyzing the performance of each variables by measuring the AUC when the variable is presented and when this is absent. Let us describe the experiment in more detail, for each data set D_i where $i = 1, 2, \dots, 10$ a repository R_i was taken randomly. Notice that the repositories were created in the first part of this experiment. Every repository R_i was evaluated by calculating the AUC, this measure was taken as a goldstone to compare the performance of the

Table 3. Usefulness index

Num Var	short name	Description	% R_i that did not use the variable
1	NIM	0.2764	0%
2	NII	0.0435	11%
3	NIX	0.1904	0%
4	NOI	0.5183	0%
5	ROA	0.0259	15%
6	ROE	0.2912	0%
7	ER	0.0141	33%
8	NCA	0.3784	0%
9	CUS	0.1165	5%
10	EC	0.2194	0%
11	CR	0.0378	17%
Average		0.1925	7%

same repository but eliminating those rules which contain a specific variable var_k . Let us define $R_i^{-var_k}$ as the repository of rules R_i removing those rules which contain the variable k and $R_i^{var_k}$ as the set of rules from R_i whose conditions contain variable k . Table 4 shows the AUC of each repository R_i and the percentage of the difference between $AUC(R_i)$ and $AUC(R_i^{-var_k})$, the formula of this variation is:

$$Variation(R_i, R_i^{-var_k}) = \frac{AUC(R_i^{-var_k}) - AUC(R_i)}{AUC(R_i)}$$

$$\forall i = 1, 2, \dots, 10, \forall k = 1, 2, \dots, 11$$

If the variation is negative it means that removing variable k caused a lost of useful patterns to detect the bankruptcy.

In order to illustrate our analysis, let us explain the Figure 1, which shows the ROC curve of R_7 and R_7^{-ROE} . Additionally, the individual performance of rules $\in R_7^{ROE}$ has been plotted in the ROC space. As can be noticed the AUC of R_7 is 0.88, which outperforms the AUC of R_7^{-ROE} (0.83). Besides, it is clear that the majority of rules in R_7^{ROE} surpass the curve R_7^{-ROE} . This means that the set of rules R_7^{ROE} is certainly contributing to the classification function. From this analysis we can assume that ROE is an important feature to create our bankruptcy prediction model. The bonus of rules in R_7^{ROE} can be easily observed in table 4, which indicates that those rules represents about the 5% of the AUC.

On the other hand, as can be seen from Figure 2 the AUC of R_2 is 0.89 and it increased the AUC to .91 by removing the rules in R_2^{NIX} . This can be noticed from table 2, that variable NIX is no contributing to the classification work, with the exception of R_4 . Furthermore, this variable was not used in the 11% of the repositories, thus we can conclude that NIX is not an important feature to describe the model.

Based on the two previous analysis, the variables NIX, NII, ER and CR were removed from the data set D_0 , which will be used to train and test our approach. The fact that some variables were eliminated from the data set because these were unhelpful or disadvantageous does not mean that those variables could not be useful to create a bankruptcy models using other structure o representation. The reason is very easy, as we mentioned previously, EDR system is able to create a set of rules, which have a specific format, these compare variables against variables or thresholds. If the predictive power of a specific variable can not be represented using this structure the variable will be not relevant.

4.3 Experiment 2

The objective of the second experiment is to evaluate the performance of EDR to create a set of decision rules for predicting bankruptcy, as well as show a sample of the rules produced by this approach. To validate the results of this experiment we used bootstrapping to re-sample the data set D_0 . Notice that D_0 has been formed by just using the approved variables in the previous experiment. Thus, D_0 was re-sampled 100 times, given that EDR is a stochastic process it was run ten times per each data set $D_i \forall i = 1, 2..100$, it means that the EDR was executed 1,000 times producing the same number of solutions or repositories.

Table 4. EDR performance eliminating variables

Num	AUC	-NIM	-NII	-NIX	-NOI
R_1	0.79	1.66%		0.53%	-44.06%
R_2	0.89		0.05%	2.21%	-41.11%
R_3	0.88				-42.78%
R_4	0.87	-0.59%		-0.67%	-9.13%
R_5	0.84			1.28%	-39.42%
R_6	0.83		-0.02%		-40.21%
R_7	0.88		-0.11%	0.61%	-42.66%
R_8	0.87		0.02%	1.87%	-43.23%
R_9	0.85			2.02%	-44.06%
R_{10}	0.89			3.78%	-36.43%
Num	AUC	-ROA	-ROE	-ER	-NCA
R_1	0.79		-1.94%		2.6%
R_2	0.89	-0.02%	-3.69%		-1.69%
R_3	0.88		-3.87%		-0.2%
R_4	0.87		0.14%		-0.39%
R_5	0.84		-3.61%		-0.45%
R_6	0.83	-0.24%	-0.14%		-1.52%
R_7	0.88		-5.07%		0.24%
R_8	0.87	-0.07%	-4.37%		0.49%
R_9	0.85		-8.75%		-2.09%
R_{10}	0.89	0.18%	0.15%	0.11%	-2.73%
Num	AUC	-CUS	-EC	-CR	
R_1	0.79	-0.28%			
R_2	0.89	-1.65%			
R_3	0.88	-0.47%	-0.22%		
R_4	0.87		0.52%		
R_5	0.84	-1.36%	1.14%	-0.67%	
R_6	0.83	-3.92%	-0.14%		
R_7	0.88	0.29%	0.24%		
R_8	0.87	-2.19%		0.09%	
R_9	0.85	-0.65%	0.02%	0.21%	
R_{10}	0.89	0.37%	0.31%		

The statistics of the results are displayed in table 5. The first column shows the statistics of the results grouped by data set and the second column displays the statistics of the complete set of repositories $R_j \forall j = 1, 2..1000$. As can be seen from Table 5, the average AUC is 0.8758 when results have been arranged by data set and 0.8797 when this is calculated using the complete set of results and each repository is independent of each other. As it is well known, the standard deviation is a widely used measure of the variability or dispersion, in our results the deviation was 0.027 and 0.034. Which is low taking into account that the application was validated using bootstrapping and EDR is a stochastic process.

In order to illustrate our results, the figures 3 and 4 show the performance of the best, the worst and an average ROC curve generated by our approach. The first graph shows the results grouped by data set and the second the results without grouping.

As can be seen, the individual performance of the rules in the best and the worst solution have been plotted in Figure 5. Let us call them R_{best} and R_{worst} respectively, notice that

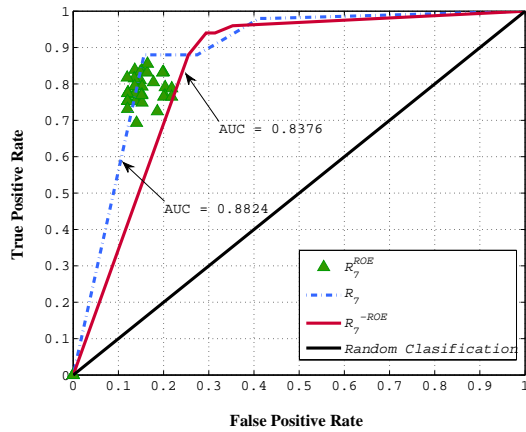


Figure 1. ROC curve of R_7 and R_7^{-ROE}

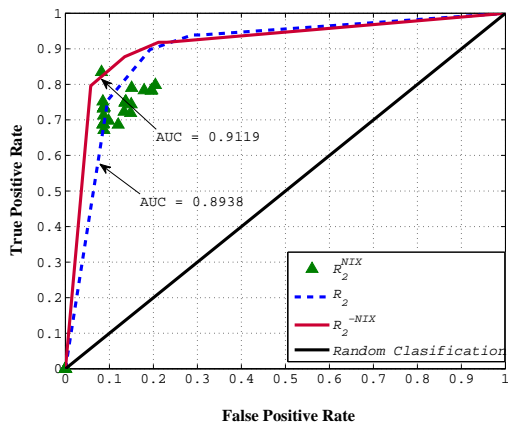


Figure 2. ROC curve of R_2 and R_2^{-NOI}

each solution is composed by 50 rules.

It is clear that the performance of rules in the best repository surpasses the performance of those in the worst repository. As can be observed, the rules in R_{best} are able to classify more true positives cases discriminating the false alarms. For example, $r_1 \in R_{best}$ is able to classify 84% of the positives cases producing just a 2% of false alarms (see table 7) and r_3 is capable to classify correctly 91% of the true positives and just 12% of false positives.

Finally, notice that the average performance of EDR has improved in comparison to the results reported before the variable selection (Table 5). As can be observed the average AUC reported by grouping and individually have slightly increased and the deviation in both cases has decreased. This exhibits the consistency of the results and evidences that the rejected variables were not needed to create the model.

5 Conclusions

In conclusion the experimental results have demonstrated that the EDR is able to generate more understandable solutions that other machine learning approaches, as Neural Net-

Table 5. EDR performance using the AUC

Before Variable Selection		
	Grouped by data set	All results
Average	0.8653	0.8702
Deviation	0.0369	0.0385
Minimum	0.7928	0.7702
Maximum	0.9165	0.9284
After Variable Selection		
Average	0.8758	0.8797
Deviation	0.0277	0.0342
Minimum	0.7915	0.7434
Maximum	0.9560	0.9683

Table 6. Sample of decision Rules from R_{Best}

Num	Decision Rule
r_1	ROE < NCA and NCA > CUS and NOI < 0.494251
r_2	NCA > CUS and NOI < 0.155161 and ROE < 0.553151
r_3	ROE < NCA and NOI < 0.643337 and NCA > -0.41476
r_4	ROE < NCA and NOI < 0.361927
r_5	ROE < NCA and NOI < 0.386157 and ROA > -1.30336
r_6	ROE < NCA and NIM > -1.18823 and NOI < 0.386157
r_7	ROE < EC and NOI < 0.338693
r_8	ROE < NCA and NOI < 0.306450 and CUS < 4.861662
r_9	ROE < NCA and NOI < 0.386157 and ROA < 4.247958
r_{10}	ROE < NCA and NOI < 0.344490 and ROE < 0.256092

works, are not able to provide. The main advantage of the solution is the understandability, which facilitates the analysis of the resultant model. Furthermore, as it was observed in the experiment I, a series of preliminary runs aids to identify the relevant features of the data set, which helps us to speed up the evolutionary process as well as to eliminate futile information. Moreover, the experiments were conducted using very simple publicly available information which can be extended easily to include either more financial ratios or a different set of also public information. This one very important aspect of our approach: the transparency of the rule construction by means of the evolutionary process is not tied to any particular school of thought regarding the information which can be useful on the predictability of bankruptcy for banks. Finally, our approach was able to produce a understandable set of rules, whose usage can be tuned using the precision of the rule and satisfying different levels of uncertainty.

References

- [1] B. C. on Banking Supervision., "International convergence of capital measurement and capital standards. a revised framework."
- [2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. United States of America: Wadsworth International Group, 1984.

Table 7. False Positive Rate and True Positive Rate of rules in Table 6

Rule	False Positive Rate X	True Positive Rate Y
r_1	0.02381	0.84483
r_2	0.02381	0.81034
r_3	0.11905	0.91379
r_4	0.11905	0.87931
r_5	0.11905	0.87931
r_6	0.11905	0.87931
r_7	0.11905	0.86207
r_8	0.11905	0.86207
r_9	0.11905	0.86207
r_{10}	0.11905	0.86207

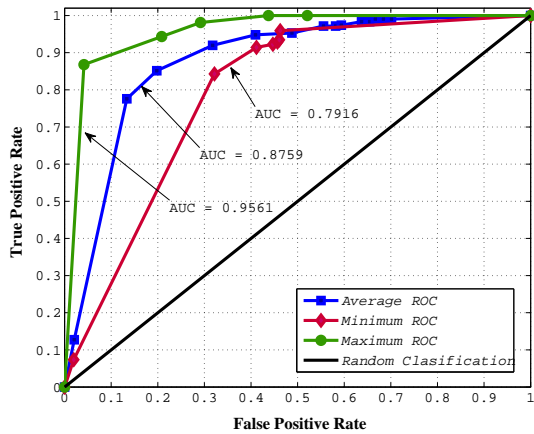


Figure 3. ROC curves of 100 re-samplings

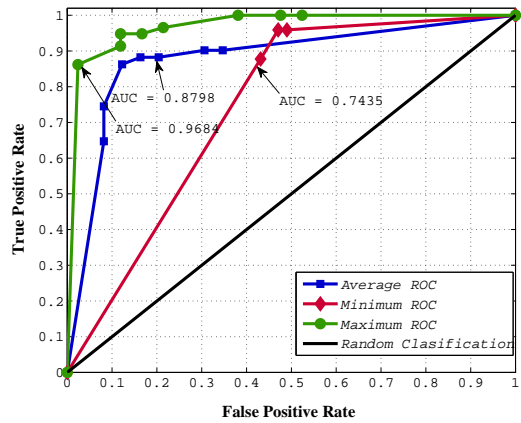


Figure 4. Best and worst ROC curve of 1,000 solutions

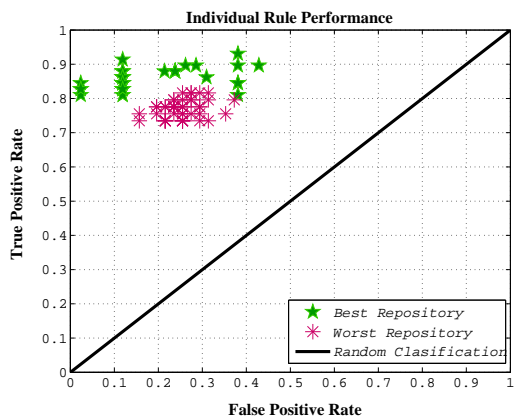


Figure 5. Individual rule performance of the best and worst solution (repository)

[3] P. Flach and N. Lavrac, *Intelligent Data Analysis*, M. Berthold and D. J. Hand, Eds. Berlin, Germany: Springer-Verlag, 2003.

[4] R. Callan, *Essence of Neural Networks*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

[5] A. L. Garcia-Almanza and E. P. Tsang, "Evolving decision rules to predict investment opportunities," *International Journal of Automation and Computing*, pp. 22–31, 2008.

[6] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, Massachusetts: The MIT Press, 1992.

[7] M. Greiner, D. Pfeiffer, and R. Smith, "Principles and practical application of receiver-operating characteristic analysis for diagnostic tests," *Prevent Veterinary Med*, vol. 45, pp. 23–41, 2000.

[8] T. Pena Centeno, S. Martinez Jaramillo, and B. Abudu, "Bankruptcy prediction: a comparison of some statistical and machine learning techniques," Banco de México, Working Paper 2009-18, Dec. 2009.

[9] W. H. Beaver, "Financial ratios as predictors of failures," *Journal of Accounting Research*, vol. 4, pp. 71–111, 1966.

[10] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *Journal of Finance*, vol. 23, no. 4, pp. 589–609, September 1968.

[11] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, p. 179, 1936.

[12] A. L. Garcia-Almanza, E. P. Tsang, and E. Galvan-Lopez, "Evolving decision rules to discover patterns in financial data sets," *Computational Methods in Financial Engineering*.

[13] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979. [Online]. Available: <http://www.jstor.org/stable/2958830>