
Measuring the Risk of Public Contracts Using Bayesian Classifiers

Leonardo J. Sales^{1,3} and Rommel N. Carvalho^{1,2}

¹Department of Research and Strategic Information at the Brazilian Office of the Comptroller General *

²Department of Computer Science at the University of Brasília †

³Department of Economics at the University of Brasília ‡
{leonardo.sales,rommel.carvalho}@cgu.gov.br

Abstract

Bayesian Classifiers are widely used in machine learning supervised models where there is a reasonable reliability in the dependent variable. This work aims to create a risk measurement model of companies that negotiate with the government using indicators grouped into four risk dimensions: operational capacity, history of penalties and findings, bidding profile, and political ties. It is expected that this model contributes to the selection of contracts to be audited under the central unit of internal control of the Brazilian government, responsible for auditing more than 30,000 public contracts per year.

1 INTRODUCTION

Public contracts can be understood as adjustments made between public administration and private sector for the attainment of public interest objectives (Di Pietro, 1999). The contract terms are set by the governmental unit, this being understood as any body or public authority of federal, municipal, or state level.

Government spending coming from public contracts and direct purchases of goods and services account for approximately 19% of the Brazilian GDP in recent years. Data from the Brazilian Institute of Geography and Statistics (IBGE), published in National Accounts Report in the 2015 fourth quarter, quantifies in R\$ 1.07 trillion the amount of government consumption expenditure in that year (IBGE, 2016). The bidding and procurement

are the institutional means by which consumption materializes, having important role in the search for efficiency and effectiveness of public spending.

Given the huge number of contracts and purchasing processes to audit, this context raises the challenge of acting effectively in the pursuit of management problems, fraud, and corruption. This is the responsibility of the governmental control units, which specially in Brazil has limited resources.

Take the example of the Office of the Comptroller General (CGU), the central unit of internal control of the Brazilian federal government, which is responsible for auditing any transaction that represents federal spending. The CGU should audit both spending conducted directly (by the central units of the ministries) as the ones conducted indirectly (by almost 20,000 decentralized units), including all payments made by any state or municipality that receives federal funds through voluntary transfers (Brazil, 2003). Nevertheless, the CGU has only 1,200 auditors working directly in the oversight of these expenditures.

In this context, a big issue arises involving the need to rationalize the use of auditing capabilities. There is a clear need to optimize the choice of what will be effectively audited, since the complete census is impossible and uneconomical. Acting in a preventive way to avoid future problems is also important since most of the errors found generate irrecoverable damage, such as paralysis of an engineering project or the need to redo it.

Both the rationalization of choices (in a subsequent operation) and the understanding and treatment of vulnerability (in preventive action) can be analyzed within the more general concept of risk assessment. After all, what is sought in both cases is to identify factors or characteristics of purchases or contracts which increase the chance of future problems such as mismanagement or even fraud.

*SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro Brasília, DF, Brazil

†Campus Darcy Ribeiro Brasília, DF, Brazil

‡Campus Darcy Ribeiro Brasília, DF, Brazil

Supervised learning models have been used in similar problems in private sector. Financial institutions assess the risk of potential borrowers, among many suitors with different characteristics and history using such models, in this case called credit scoring (Lessmann et al., 2015). Insurance companies also use such statistical models to assign the value of insurance for a certain good. The techniques learn from the transaction history and quantify the weight of certain characteristics in determining the risk of a client or specific process. Thus, the auto insurance company knows that unmarried young men offer more risk than married women with children.

In practice, these models are applications of statistical and computational techniques of regression and classification using databases that have information of transaction history and labeled cases of “success” and “failure” (Friedman et al., 2001). A good condition in the construction of this type of risk analysis model is the existence of information on transaction history, with variables representing different characteristics of each transaction. Thus, one can distinguish and identify correlations between groups.

This paper proposes to create a predictive model of risk in contracts based on Bayesian classifiers. It will result in the quantification of the propensity that a supplier has problems in government contracts, according to the company’s characteristics. Learning models using Bayesian networks are especially useful when you need to organize or discover the knowledge of a particular area through the construction of cause and effect relationships captured from a set of data (Spiegelhalter et al., 1993). Besides this, Bayesian Classifiers have been incorporated into risk measurement studies, especially when it is important to capture and explain the relationships of cause and effect between the different prediction parameters, avoiding the “black box” issue, common in other techniques.

The model will be used to select high-risk contracts to be audited by the CGU and will be based on the estimation of the relations of cause and effect between various indicators that are related to the propensity of contractual risk. The dependent variable is the occurrence of more severe punishment that can be given to a supplier in Brazil: the impediment to bidding. The indicators that will be used as predictors represent characteristics grouped into four risk dimensions: operational capacity, history of penalties and findings, bidding profile, and political ties.

This work is divided into 5 sections. Besides this introduction, Section 2 presents the theoretical framework that supports the central idea of the work and the methodological approach adopted. Section 3 contains the de-

tails of the methodology used in the study, including the understanding of data modeling, the creation of the networks, and the validation of the models. Section 4 presents and discusses the results. Finally, Section 5 provides conclusions and considerations on gaps and opportunities for future work.

2 THEORETICAL REFERENCES

In this section we describe the public bidding process in Brazil, the Bayesian classifiers used for learning the predictive models, and some related works.

2.1 PUBLIC BIDDING IN BRAZIL

The whole process of buying products or hiring services in the Brazilian federal government takes place according to the rules of Law 8666/1993 (Brazil, 1993), called Procurement Law. Other regulatory acts complement this law, such as Law 10520/2005 (Brazil, 2002), establishing the types of Auction and Complementary Law 123/2006 (Brazil, 2006) establishing privileges for micro and small businesses in bidding. Law 8666/1993 (Brazil, 1993) details the stages of the bidding process itself, the bidding types allowed, types of contracts, aspects of qualification of companies, and also defines administrative and criminal penalties to be applied to suppliers in case of noncompliance.

The Procurement Law, together with other mentioned legislation, defines the following administrative penalties to suppliers, due to total or partial non-performance of contracts:

- warning;
- pecuniary penalty;
- temporary suspension of bid;
- declaration of non-trustworthiness; and
- impediment to bid and hire.

The whole process of procurement and contracting in the federal government is done using the government’s General Services Administration System (SIASG). Each purchase or contract is recorded in this system, since the opening of the process to the issue of commitment.

Existing since 1994, the SIASG started to be used by the government gradually and it already has more than 5 million purchases. All federal administration is required to use this system. Annually it records over 700,000 bids. Some of these bids representing continued provision of services or delivery of goods turns into contracts, generating nearly 30,000 new contracts per year.

2.2 BAYESIAN CLASSIFIERS MODELS

Since Bayesian networks (BNs) have been successfully used in classification problems – *e.g.*, see (Sahami et al., 1998; Friedman et al., 1997; Goldszmidt et al., 2010; Friedman and Goldszmidt, 1996; Cheng and Greiner, 1999; Ceccon et al., 2014; Ye et al., 2014; Shi et al., 2013) –, we decided to experiment with different BN learning algorithms in order to classify the companies that sell service and goods to the government with high likelihood of noncompliance.

Score-based learning is a popular method for inducing BNs. The main idea is to assign a score to a model based on how well it represents the data set used for learning. Thus, the purpose of the algorithm is to maximize the goodness-of-fit score.

In this work we use standard and well-known Bayesian network classifiers, which are aimed at classification. More specifically, we use two algorithms available in the *bnlearn* R package¹ (Scutari, 2009):

- Naïve Bayes (*naive.bayes*): a simple algorithm that assumes that all explanatory variables are independent of each other. In other words, the target variable is the only parent of all other variables.
- Tree-Augmented Naïve Bayes (*tree.bayes*): an algorithm that relaxes the simple Naïve Bayes assumption of independence, by allowing the explanatory variables to have one other variable as parent besides the target one.

Besides that, we also tried two different score-based learning algorithms, which are also available in the *bnlearn* R package used in this work (Scutari, 2009):

- Hill-Climbing (*hc*): a hill climbing greedy search on the space of the directed graphs.
- Tabu Search (*tabu*): a modified hill-climbing able to escape local optima.

The *bnlearn* package implements random restart with configurable perturbing operations for both algorithms.

A number of different scores were used to fine tune the models learned from the score-based algorithms and to improve their performance, which are also available in the *bnlearn* package (Scutari, 2009):

- the Akaike Information Criterion score (*aic*);
- the Bayesian Information Criterion score (*bic*);

¹The package is available at <http://www.bnlearn.com/>.

- the logarithm of the Bayesian Dirichlet equivalent score (*bde*); and
- the logarithm of the modified Bayesian Dirichlet equivalent score (*mbde*).

2.3 RELATED WORKS

Many studies use supervised learning models in order to predict risk in business transactions. The area where it is more common this type of approach is the bank credit (Lessmann et al., 2015; Hand and Henley, 1997).

These learning models attempt to quantify how the characteristics of potential borrowers influence the probability of default. Classically, the techniques most used for this purpose are Logistic Regression and Discriminant Analysis (Ghodselahe, 2011). Other studies have been testing and comparing some modern techniques (Baesens et al., 2002). In other areas, such as insurance, such models are also widely used.

Bayesian Classifiers have been incorporated into these studies, especially when you want to capture and explain the relationships of cause and effect between the different prediction parameters, avoiding the “black box” issue, common in other techniques (Jiang and Wu, 2009; Zonneveldt et al., 2010; Baesens et al., 2002). Bayesian algorithms provide more clear insights when modeling causal relationships.

A new approach to credit scoring by synthesizing Simple Naïve Bayesian Classifier (SNBC) and the Rough Set Theory is presented by (Jiang and Wu, 2009). A comparison between Naïve Bayes (NB) models, different augmented NB models, and a handcrafted causal network is made by (Zonneveldt et al., 2010).

In the context of public procurement, some initiatives already exist in order to implement similar models in predicting irregularities or contractual problems. For example, Naïve Bayes algorithms are used by (Balaniuk et al., 2012) in an unsupervised approach to quantify the combined risk of private companies and government units in the execution of contracts.

(Sales, 2014) built a model with the same objective of this work (to measure the risk of public contracts) and with similar data. In that case the accuracy using Logistic Regression and Decision Tree were compared, resulting in the best accuracy of 64%.

3 METHODS AND PROCEDURES

The first step in building the Bayesian classification model was the definition of the criteria for characterization of the companies with the highest risk (the “Bad”).

In this sense, we chose to characterize the “Bad” group all companies that suffered the following punishments in the years 2015 and 2016: temporary suspension of bid, declaration of non-trustworthiness, and impediment to bid and hire. The group of low-risk companies (hereinafter “Good”) are companies with existing contracts in the same period but without such punishment.

The database used contained 1,448 companies, of which 724 were previously classified as “Bad” and other 724 previously classified as “Good”².

From this initial setting, the second step was the creation of risk indicators, which cover the past of relations between companies and government, considering the period since 2011, as well as other information that are independent of the period, such as those from the registry of companies. The idea is to answer the following question: What happened in the recent past of the companies that contributed to its contractual default in 2015 and 2016?

These indicators were obtained from the four dimensions of risk: operational capacity, history of penalties and findings, bidding profile, and political ties. The meaning of each of the risk dimensions and some indicators used are described below:

- Operational capacity: irregularities related to the existence or insufficient physical and operational structure of the contracted company.
 - Quantity of indicators: 11.
 - Examples of indicators: number of employees, number of partners, the total amount received from the government, amount received from the government per employee, value received from the government for partner, average salary of employees, average salary of the partners, company size, number of activities carried out by the company, age from the company.
- History of penalties and findings: pre-existence of punishment or audit findings related to the company.
 - Quantity of indicators: 04.
 - Examples of indicators: quantity of received punishments, number of alerts generated in CGU monitoring.

²The 724 companies in the “Bad” group are all companies that meet the criteria described for this class. The 724 companies in “Good” group was obtained by sampling in the set of 41,000 companies that meet the requirements described. Sampling the second group was made in order to solve the dominant class issue, in a process called undersampling (see (Japkowicz et al., 2000) for more details of this process).

- Bidding profile: company profile when participating in bids, as the average quantity of offers, and the degree of success of business (percentage of wins).
 - Quantity of indicators: 12.
 - Examples of indicators: quantity of purchases, purchase quantity of items, average amount of offers, number of units of the federation, number of wins, percentage of victory, value of contracts, the difference in days between the opening of the company and the first participation in a public procurement.
- Political ties: company relationship with politicians, via donations in campaigns.
 - Quantity of indicators: 01.
 - Examples of indicators: amount donated in political campaigns.

The next step was the transformation of all variables in factors (categories), using a simple process of discretization, where values of each variable were divided into three intervals of equal size. Once complete, the database has been divided in training set (70%) and test (30%). The discretization was carried out due to the limitation of some algorithms used. In future experiments, we will learn models using algorithms that allows continuous variables.

At first, we used standard Bayesian classifiers available in the *bnlearn* R package, Naïve Bayes (NB) and Tree-Augmented Naïve Bayes (TAN).

As the database does not have a very large number of observations, we used a process of estimation with cross-validation in the training subset for both algorithms. The Cross-Validation procedure applied was the random division of training based on 10 sample partitions of equal size, for use in cycles of modeling where 9 partitions are used for training and one for testing. Error measures are then combined to have a single measurement error.

The estimation with cross-validation was performed using a Score-based learning algorithm, which ranks the network structures created with emphasis on model fit. In these algorithms, various parameters can be adjusted in search of the best results forecast.

The loss function used to measure the model results was the misclassification, where the dependent variable value is the result of local distributions (from its parents) and the error function is measured by coincidence or not with the actual values (hit rate).

Since an important aspect of machine learning is the parameter tuning and both NB and TAN in *bnlearn* do

not have any parameters to be tuned, we decided to also try another set of algorithms. In *bnlearn*, a set of algorithms that allow many different configurations is the score-based learning algorithms, namely: Hill-Climbing (HC) and Tabu Search (Tabu), both using incremental search. Tabu introduces changes in HC in order to avoid local optima.

In score-based algorithms, it is critical to set the network score calculation method, which measures the quality of the network created using the quantification of posterior probability. Two variables were used in the score parameterization: type of score and penalty parameter. The tested scores types were AIC (Akaike Information Criterion Score), BIC (Bayesian Information Criterion Score), BDE (Bayesian Dirichlet Equivalent Score), and MBDE (Modified Bayesian Dirichlet Equivalent Score), suitable for categorical variables. Besides that, we also tried many different penalty parameters.

The central idea was to try different values of each parameter in order to find the setting that present the best predictive ability. For better understanding, Table 1 shows some of these tested settings and its accuracy measure, aiming to compare the Naïve Bayes (NB) algorithm setting with different configurations³ of Score-Based algorithms.

Table 1: The table shows that despite our efforts in setting up the Score-Based Algorithms, there was no significant difference than the Naïve Bayes and TAN algorithms. The Accuracy here is the proportion of true results, either true positive or true negative.

| Algorithm | Setting | Accuracy - 95% CI |
|-------------|--------------|-------------------|
| NB | - | (0.70, 0.76) |
| TAN | - | (0.72, 0.78) |
| Tabu | AIC, K=0.1 | (0.67, 0.73) |
| Tabu | BDE, ISS=25 | (0.65, 0.70) |
| HC | MBDE, ISS=10 | (0.64, 0.70) |

4 RESULTS

Since the best models did not present a statistically significant difference in performance and usually the simpler the model the better the generalization, we chose the Naïve Bayes algorithm to run the final model with all the data from the training set in order to check the

³The parameters used to set the algorithm were the score-based algorithm, Hill-Climbing (HC) or Tabu Search (Tabu), the score types (AIC, BIC, BDE or MBDE) and the penalty parameter (ISS or K).

performance with the test set. The 95% confidence interval of the accuracy was (0.69, 0.77), which shows that the model generalizes well. The sensitivity of the model (prediction ability of “BAD” companies) was 76%. Table 2 shows the results of prediction on the test set.

Table 2: The table shows the model results, that presented a total accuracy of 73%, with higher quality in the identification of “Bad” cases.

| Prediction | Real values | |
|-------------|-------------------|-------------------|
| | Good | Bad |
| Good | 170 | 47 |
| Bad | 69 | 148 |
| % | 71% (specificity) | 76% (sensitivity) |

We consider this a good result in the context of government contracts, especially when compared with other similar works. Taking as reference the results obtained by (Sales, 2014), you can see a reasonable gain in predictive ability. The sensitivity of the model is particularly important since what really matters is the identification of high-risk cases, even assuming the cost of auditing some low risk contracts, which were misclassified.

5 CONCLUSION AND FUTURE WORK

This work is consistent with a great effort that has been developed by government control institutions to rationalize the use of their human and material resources in order to provide more effective results at lower operating and financial costs.

Considering the current Brazilian context, where a severe economic crisis has been treated through large cuts in public budgets (reducing the sending of resources to control bodies), the efficient use of resources should be a permanent goal.

The attempt to use statistical models based on Bayesian networks is in addition to other initiatives presented in Section 2. The main purpose of these studies is to extract knowledge from various databases that government control institutions have access in order to facilitate the selection of audit objects more likely to present problems.

The classification results are slightly better than other supervised models applied in government databases with the same goal (see (Sales, 2014), described in section 2.3). However, we believe that there is room for improvement in two possible ways: the inclusion of new indicators that capture aspects ignored by this model and the use of optimization algorithms in the parameterization of score-based networks.

Each step in direction of improving these models is a permanent gain for the public auditing activity, and consequently to society.

References

- Bart Baesens, Michael Egmont-Petersen, Robert Castelo, and Jan Vanthienen. Learning bayesian network classifiers for credit scoring using markov chain monte carlo search. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 49–52. IEEE, 2002.
- Remis Balaniuk, Pierre Bessiere, Emmanuel Mazer, and Paulo Cobbe. Risk based Government Audit Planning using Nave Bayes Classifiers. In *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, 2012. URL <https://hal.archives-ouvertes.fr/hal-00746198/>.
- Brazil. Lei n 8666, de 1993, 1993.
- Brazil. Lei n 10520, de 2002, 2002.
- Brazil. Lei n 10683, de 2003, 2003.
- Brazil. Lei Complementar n 123, de 2006, 2006.
- S. Ceccon, D.F. Garway-Heath, D.P. Crabb, and A. Tucker. Exploring early glaucoma and the visual field test: Classification and clustering using bayesian networks. *IEEE Journal of Biomedical and Health Informatics*, 18(3):1008–1014, May 2014. ISSN 2168-2194. doi: 10.1109/JBHI.2013.2289367.
- Jie Cheng and Russell Greiner. Comparing bayesian network classifiers. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI’99, page 101108, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-614-9. URL <http://dl.acm.org/citation.cfm?id=2073796.2073808>.
- Maria Sylvia Zanella Di Pietro. *Direito administrativo*, volume 22. Atlas São Paulo, 1999.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. URL <http://statweb.stanford.edu/~tibs/book/preface.ps>.
- Nir Friedman and Moises Goldszmidt. Building classifiers using bayesian networks. In *Proceedings of the national conference on artificial intelligence*, page 12771284, 1996.
- Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, November 1997. ISSN 0885-6125, 1573-0565. doi: 10.1023/A:1007465528199. URL <http://link.springer.com/article/10.1023/A%3A1007465528199>.
- Ahmad Ghodselahi. A hybrid support vector machine ensemble model for credit scoring. *International Journal of Computer Applications*, 17(5):1–5, 2011.
- Moises Goldszmidt, James J. Cochran, Louis A. Cox, Pinar Keskinocak, Jeffrey P. Kharoufeh, and J. Cole Smith. Bayesian network classifiers. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc., 2010. ISBN 9780470400531. URL <http://onlinelibrary.wiley.com/doi/10.1002/9780470400531.eorms0099/abstract>.
- David J Hand and William E Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.
- IBGE. Indicadores do Instituto Brasileiro de Geografia Estatística, Contas Nacionais Trimestrais, 2016.
- Nathalie Japkowicz et al. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. Menlo Park, CA, 2000.
- Yi Jiang and Li Hua Wu. Credit scoring model based on simple naïve bayesian classifier and a rough set. In *2009 International Conference on Computational Intelligence and Software Engineering*, 2009.
- Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, November 2015. ISSN 03772217. doi: 10.1016/j.ejor.2015.05.030. URL <http://linkinghub.elsevier.com/retrieve/pii/S0377221715004208>.
- Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, page 98105, 1998.
- Leonardo Jorge Sales. Risk prevention brazilian government contracts using credit scoring. In *Interdisciplinary Insights on Fraud*, chapter 11, pages 264–286. Cambridge Scholars Publishing, 2014.
- Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.
- Wei Shi, Yao Wu Pei, Liang Sun, Jian Guo Wang, and Shao Qing Ren. The defect identification of LED chips based on bayesian classifier. *Applied Mechanics*

and Materials, 333-335:1564–1568, July 2013. ISSN 1662-7482. doi: 10.4028/www.scientific.net/AMM.333-335.1564. URL <http://www.scientific.net/AMM.333-335.1564>.

David J. Spiegelhalter, A. Philip Dawid, Steffen L. Lauritzen, and Robert G. Cowell. Bayesian Analysis in Expert Systems. *Statistical Science*, 8(3):219–247, 1993. URL <http://www.jstor.org/stable/2245959>.

Ye Ye, Fuchiang (Rich) Tsui, Michael Wagner, Jeremy U. Espino, and Qi Li. Influenza detection from emergency department reports using natural language processing and bayesian network classifiers. *Journal of the American Medical Informatics Association*, pages amiajnl-2013-001934, January 2014. ISSN , 1527-974X. doi: 10.1136/amiajnl-2013-001934. URL <http://jamia.bmj.com/content/early/2014/01/09/amiajnl-2013-001934>.

S Zonneveldt, K Korb, and A Nicholson. Bayesian network classifiers for the german credit data. Technical report, Technical report, 2010/1, Bayesian Intelligence. <http://www.Bayesian-intelligence.com/publications.php>, 2010.