Bayesian Models to Assess Risk of Corruption of Federal Management Units

Ricardo S. Carvalho[1] **and Rommel N. Carvalho**[1,2]
[1]Department of Research and Strategic Information at the Brazilian Office of the Comptroller General [*]
[2]Department of Computer Science at the University of Brasília [†]
{ricardo.carvalho,rommel.carvalho}@cgu.gov.br

Abstract

This paper presents a data mining project that generated Bayesian models to assess risk of corruption of federal management units. With thousands of extracted features related to corruptibility, the data were processed using techniques like correlation analysis and variance per class. We also compared two different discretization methods: Minimum Description Length Principle (MDLP) and Class-Attribute Contingency Coefficient (CACC). The feature selection process used Adaptive Lasso. To choose our final model we evaluated three different algorithms: Naïve Bayes, Tree Augmented Naïve Bayes, and Attribute Weighted Naïve Bayes. Finally, we analyzed the rules generated by the model in order to support knowledge discovery.

1 INTRODUCTION

Currently, it is known that corruption is a recurrent and primary subject on the Brazilian government agenda, fundamentally requiring its ostensive and efficient combat. Public corruption can be defined – supported by Brazilian Law no. 8,429, of June 1992[1] – as the act of misconduct or improper use of public office that leads to illicit enrichment, causing injury to the public treasury or infringing upon the principles of the public administration.

The Brazilian Office of the Comptroller General (CGU), an agency incorporated in the Presidency structure, has as one of its competences the role of assisting directly and immediately the President on matters and measures related to preventing and fighting corruption. Through activities of strategic information production, the Department of Research and Strategic Information (DIE) is the area responsible for investigating possible irregularities involving federal civil servants working in management units.

Nowadays, there are more than thirty thousand active federal management units[2], all subject to investigation. Due to this large number of units, most of the time DIE is limited to performing only investigations of those involved on large federal operations or recurrent complaints, often restricting its activities to cases triggered externally. Thus, it is important to have prioritization of activities based on risks of involvement in corruption so that DIE can act more effectively and proactively.

This work has two main objectives and contributions. The first is to build a Bayesian model to assess risk of corruption of federal management units. To this end, we seek to apply data mining techniques based on the state-of-the-art, along with a practical study of the information related to corruption. Therefore, we wish to contribute to CGU's activities in fighting corruption by building an useful model for their work priorization. Also, the step-by-step of this data mining project might be interesting for other practitioners, since it involves the combination of several different methods. We show how we applied correlation analysis and two discretization methods to process features, Adaptive Lasso for feature selection, and end up comparing three different algorithms to choose our final Bayesian model. Hence, this work gives contribution to practitioners while describing the application of data mining techniques with a practical objective and singular combination of techniques.

[*] SAS, Quadra 01, Bloco A, Edificio Darcy Ribeiro Brasilia, DF, Brazil

[†] Campus Darcy Ribeiro Brasília, DF, Brazil

[1] Brazilian Law no. 8.429, June 1992: http://www.planalto.gov.br/ccivil_03/leis/l8429.htm

[2] Management Units dataset: http://www.tesourotransparente.gov.br/ckan/dataset/siafi-relatorio-unidades-gestoras

The second objective is to achieve knowledge discovery in relation to information about corruptibility of federal management units, seeking to extract new rules in this domain. To this end, the information of management units available – as well as its direct and indirect relationships with the federal civil servants working there – are analyzed with the support of DIE experts in fighting corruption. After building our final model, we analyzed its derived rules. With this in mind, we wish to contribute to the enrichment of the experts' knowledge in fighting corruption.

In Section 2, we depict works most closely related to fighting corruption and how data mining has been used, while in Section 3 we give an overview of the information selected by DIE experts that will be used to build our models. Section 4 describes steps taken to pre-process data, such as correlation analysis, discretization, and also feature selection. In Section 5 we show how we used machine learning to build several models and Section 6 depicts our evaluation strategies. Section 7 discusses our deployment efforts related to the products of this work and we end this paper with a conclusion in Section 8.

2 RELATED WORK

In the last decade, observing current research areas, a topic closely related to risk of corruption is fraud detection. The main objective of fraud detection is to reveal trends of suspicious acts. For example, an emerging theme is to use data mining to detect financial fraud. A review of the academic literature of such application (Ngai et al., 2011) shows its successful use in detecting credit card fraud, money laundering, bankruptcy prediction, among others. This review also identifies common data mining techniques used in fraud detection, including Artificial Neural Networks, Decision Trees, Logistic Regression, and Naïve Bayes.

In this context, a recent survey on the subject of data mining-based fraud detection (Phua et al., 2010) displays a summary of published technical articles and a review on the topic. This survey, as well as other works (Kou et al., 2004), includes comments on similar applications. Also, an individual-oriented corruption analysis (Carvalho et al., 2014) was done building a corruption risk model for affiliated civil servants with algorithms like Random Forest and Bayesian Networks.

Regarding aspects of corruption, research related to public bidding and contracting processes has also been carried out, though not as widely as in fraud detection. The use of clustering and association rules to the problem of cartels in public bidding processes (Silva and Ralha, 2010) found results that corroborate the application of data mining in the prevention of corruption. Another paper (Balaniuk et al., 2012) shows the use of Naïve Bayes to evaluate the risk of corruption in public procurement. The authors applied natural logarithm to discretize attributes and based their assessment on the results of the conditional probabilities defined by experts.

In addition, a recent paper (Carvalho et al., 2013) presents the use of probabilistic ontologies to design and test a model that performs the fusion of information to detect possible fraud in bidding processes involving federal money in Brazil.

With respect to discretization algorithms, it has currently received a lot of focus as a pre-processing technique, mostly since many machine learning algorithms are known to produce better models by discretizing continuous attributes (Garcia et al., 2013). Two algorithms have received generally great performance, namely: CACC (Class-Attribute Contingency Coefficient) (Tsai et al., 2008) and MDLP (Minimum Description Length Principle) (Irani, 1993). In this work we compare the results of these algorithms after feature selection by creating models to allow us to choose the best results.

For feature selection, a recent review (Tang et al., 2014) shows several different widely used techniques, such as Adaptive Lasso (Zou, 2006). The Adaptive Lasso has basically two steps. First, an initial estimator is obtained, usually using Ridge Regression (Zou, 2006). Then a optimization problem with a weighted L1 penalty is carried out. The initial estimator generally puts more weight on the zero coefficients and less on nonzero ones to improve upon its predecessor: the Lasso (Zou, 2006). Compared to the Lasso, the adaptive Lasso has the advantage of the oracle property (Zou, 2006), resulting in a performance as well as if the true underlying model were given in advance. Compared to the SCAD and bridge methods (Tang et al., 2014), which also have the oracle property, the advantage of the adaptive Lasso is its computational efficiency.

3 DATA UNDERSTANDING

Seeking to analyze corruptibility of federal management units, various databases that DIE has access have been identified as useful for this work. For a better understanding of the data, the available information were divided into four dimensions, namely: Corruption; Employment; Sanctions; and Political.

Some of the information treated in this work are related to the federal civil servants that work in the management units. These information can give an idea of how much power a certain unit concentrates or how much influence the civil servants bring to the unit environment.

Due to the limited size of this paper, we present each dimension giving only an overview of the existing databases and relevant information identified by DIE experts regarding possible relationships with corruptibility.

3.1 CORRUPTION DIMENSION

CGU maintains the Federal Administration Registry of Expelled (CEAF)[3], which is a database with information that gathers expulsion penalties (expel, retirement abrogation, and dismissal) of federal civil servants since the year of 2003.

This database will be used to define management units that are corrupt, namely the positive class in our machine learning algorithms. The first paragraph of the Section 4 describes how this is done.

3.2 EMPLOYMENT DIMENSION

The employment dimension covers the information of management units regarding the federal civil servants that work there. It may be related to basic information such as office time and income, or even data that exposes the importance of the unit the servant is working – such as number of coordination roles or critic public offices like those that deal directly with public resources or financial assets.

Most of the information comes from the Human Resources Integrated System (SIAPE) of the Brazilian Federal Government[4].

For the employment dimension, the experts in fighting corruption of DIE selected 16 different information, that later can be transformed in 16 or more different features in the data preparation phase. Examples of these information are: mean, maximum, and minimum monthly income; number of coordination roles that deal with public contracts; number of roles for specific activities such as head of regional agency.

3.3 SANCTIONS DIMENSION

The sanctions dimension covers the information of management units that got sanctioned, due to practices of bad management of public money. We used sanctions in the Accounts Judged Irregular (CADIRREG) from the Federal Court of Accounts (TCU)[5], that judges the accounts of each management unit, deciding about its regularity according to Brazilian laws. Similarly, we used CGU's certificates of management irregularity[6].

Therefore, the experts in fighting corruption of DIE selected four different information, that later can be transformed in four or more different features in the data preparation phase. Examples of these information are: number of accounts judged irregular from TCU; and number of regularity certificates from CGU.

3.4 POLITICAL DIMENSION

The political dimension covers data of federal civil servants related to political activities, namely analyzing information of affiliation to political parties. By getting the affiliated servants of each management unit, we can measure how much each political party influences the units and if this will relate to corruption. The main database comes from Superior Electoral Court (TSE)[7].

Taking into account the knowledge of DIE experts, from the data provided by TSE we selected nine different information. Examples are: number of affiliations for a given political party and total number of affiliated servants in each management unit.

4 DATA PREPARATION

The data to be prepared are extracted for two classes, called "Corrupt" and "Non Corrupt". On one hand, "Corrupt" management units are those that throughout its history have had at least one civil servant who was expelled due specifically to corruption. In other words, units that had corrupt civil servants, which are those registered in CEAF whose legal basis for expulsion is consistent with our definition for corruption, as stated in Section 1.

On the other hand, to build the "Non Corrupt" group, we sampled a large group of management units and removed those considered "Corrupt" by definition, keeping the random sample proportion.

Thus, the dataset for non corrupt was created with a random sample of approximately 4,800 federal management units – amount approximately 8 times greater than the number of corrupt units.

[3] CEAF – Link: http://www.portaldatransparencia.gov.br/expulsoes/entrada

[4] Website for the Human Resources Integrated System (SIAPE) of the Brazilian Federal Government: http://www.siapenet.gov.br

[5] CADIRREG: http://contas.tcu.gov.br/cadirreg/CadirregConsultaNome

[6] CGU's audits reports: http://sistemas.cgu.gov.br/relats/relatorios.php

[7] TSE repositories: http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais

The data preparation phase includes feature selection and goes through the following steps, which will be described in the next sections:

- Data Cleaning and Feature Engineering: Adjusts the dataset;
- Preliminary Analysis: Treats variance zero per class and correlation;
- Data Separation: Segregates data for training and testing;
- Intermediary Analysis: Variance and correlation filtering;
- Feature Selection: Uses Adaptive Lasso;
- Discretization: Applies MDLP and CACC;

4.1 DATA CLEANING AND FEATURE ENGINEERING

Besides usual data cleaning activities – such as adjustment of inconsistencies, data conversion, and standardizing data types – the treatment of missing values was also conducted. For categorical variables we created a category "NA" representing the absence of values for a given variable. As for counting numerical variables, missing values represent the actual value of zero, so they were replaced by such value. In addition, other fields with missing values were treated individually. For example, date of cancellation of party affiliation, when affiliation still active, were replaced by a current date in order to create features for time of affiliation.

On feature engineering, first we created binarized features for all the categorical variables. Then, since some information can be registered more than once for a given management unit – for example, one can have several regularity certificates – we had to summarize the features for each unit. With only numerical features, a few of them were summarized by creating features with maximum, minimum, average, and total. For example, annual income was transformed into maximum annual income, minimum annual income, and mean annual income.

After this step, we had created 2,238 different features.

4.2 PRELIMINARY ANALYSIS

At first we removed features that had variance, within one of the classes, equal to zero, since with zero class-variance algorithms might bring estimates of coefficients that do not generalize (Hosmer et al., 2013). After calculating class-variance for each of the 2,238 features, 747 of them were removed – most of these being related to binarized categorical variables.

We also preliminarily addressed perfect pairwise correlation, which accounts for redundant information and may give biased estimates. Perfectly correlated features may have been added accidentally, or may have arisen after feature engineering.

Among the 1,495 variables analyzed, 96 – 48 pairs – returned perfect correlation. DIE experts chose which to eliminate in each pair.

4.3 DATA SEPARATION

At this point, our complete dataset had 688 corrupt units and 4,792 non corrupt units, with 1,447 features.

In this step we created two different datasets: Training Data (DT) and Testing Data (DTE). The first will be used through all data preparation and modeling, while the second will only be used as a final test after choosing the best final model.

To keep the original balance, DTE was created using a random sample of 20% of corrupts plus 20% of the non corrupts, and DT stayed with the remaining data, corresponding to 80% of the complete dataset.

4.4 INTERMEDIARY ANALYSIS

Similarly to the Preliminary Analysis, we again analyzed the class-variance. This resulted in removing 62 features with zero variance in one of the classes.

Nevertheless, in the intermediary analysis we did a different correlation analysis, following the well known hypothesis (Hall, 1999): "A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other".

Initially we calculated the correlation matrix of the remaining 1,376 features, also adding their correlation with the class column indicating corruptibility – 0 to non corrupt units and 1 to corrupt units. Then we filtered pairs of features with correlation equal or greater than 0.70 (absolute value) – number generally considered high correlation (Taylor, 1990). After that, the resulting matrix was sorted in descending order regarding the correlation of the features in relation to the class.

Thereafter, the rows of the matrix were traversed from the features with the largest correlation to the class. In each row, we kept the feature with the highest correlation with the class and removed the remaining features – from the dataset and the matrix – that had inter-correlation higher than 0.70 (absolute value).

With this algorithm we eliminated 468 features that had absolute correlation equal or greater than 0.70, thus remaining 910 features.

Such an approach was used to try to avoid the collinearity problem, mainly due to the fact that it is impossible to analyze all the possible combinations of feature groups, involved in this work. Thus, the correlation heuristic of each feature with its class – although not fully reflected in a model due to interactions between the features – serves as a technique to try to keep the theoretically most significant features – considering the correlation with class[8].

4.5 FEATURE SELECTION

To perform feature selection, each dataset passes through a regularized regression, specifically using Adaptive Lasso. For this purpose, we start by performing Ridge Regression with 10-fold cross-validation on the DT dataset. The estimates of the coefficients are used to construct an adaptive weights vector. With this vector introduced as the penalty factor, we implement Adaptive Lasso with 10-fold cross-validation. It is worth noticing that the Adaptive Lasso can force some of the coefficients to have estimates exactly equal to zero, thereby reducing the number of features.

After feature selection with Adaptive Lasso, we selected 144 features. The 10-fold cross-validation resulted in a AUC (Area Under the ROC Curve) (Bradley, 1997) of 0.85, considered satisfactory.

4.6 DISCRETIZATION

In recent years, discretization has received increasing research attention (Garcia et al., 2013). In the case of non-monotonic variables, the use of discretization techniques proves to be essential since it makes it possible to separate an original non-monotonic variable in various monotonous derived covariates (Tufféry, 2011). Also, when thinking about Bayesian models, some algorithms need all the features to be categorical, and discretization is a method of doing so.

In recent research (Garcia et al., 2013), two algorithms have received generally great performance, namely: MDLP (Minimum Description Length Principle) (Irani, 1993) and CACC(Class-Attribute Contingency Class) (Garcia et al., 2013). We compare these algorithms by later creating models for groups of features discretized with each method.

Accordingly, we have generated two different datasets from DT, one dataset for each discretization method used. The dataset discretized with MDLP algorithm returned 23 binary features, while CACC returned 66 – the reason these datasets have less features than the original is due to the fact that constant features were automatically removed.

5 MODELING

In the modeling phase we started by creating models for each of the datasets discretized with MDLP and CACC. For this, we created Bayesian models using three different algorithms: Naïve Bayes (Lowd and Domingos, 2005), Tree Augmented Naïve Bayes (Zheng and Webb, 2011), and Attribute Weighted Naïve Bayes (Taheri et al., 2014).

This task was done using the R Package named caret[9]. We used 10-fold cross-validation to evaluate AUC and tried several different combinations of parameters for each of the three algorithms – from 20 to 60 combinations. For example, for Tree Augmented Naïve Bayes we used three score functions (loglik, bic, aic) each along side 20 different values for smoothing (from 0 to 19). After these models were built, caret selects the one with the combination of parameters that resulted in the best AUC value for each algorithm.

5.1 DISCRETIZATION SELECTION

The first step is to choose the most suitable discretization. With this in mind, for each discretized dataset we take the average results of AUC for the three algorithms used, again using 10-fold cross validation to try to estimate the out-of-sample results. The mean AUC outcomes are depicted in Table 1, along side the number of features each dataset has.

Table 1: Mean Results of Bayesian Models for each Discretized Dataset

Discretization	No. of features	**AUC**
MDLP	23	0.82
CACC	66	0.83

Although the results for the dataset with CACC discretization were slightly better, it is desirable to minimize the number of features considered in a model. Mainly models with less features tend to be more numerically stable and be adopted more easily. Also, a model with less features can avoid overfitting and increase its interpretability.

[8] It may be useful to use different methods to analyze correlation in future work.

[9] R Package caret: `https://cran.r-project.org/web/packages/caret/index.html`

Therefore, we chose to select the features discretized with MDLP, since the respective model achieved results close to CACC but kept almost three times less features.

5.2 MODEL SELECTION

With the discretized dataset chosen, we now evaluate the Bayesian models built with the three algorithms: TAN (Tree Augmented Naïve Bayes) AW-NB (Attribute Weighted Naïve Bayes) and NB (Naïve Bayes). The AUC outcomes are showed in Table 2.

Table 2: Results of Bayesian Models for MDLP Dataset

Algorithm	**AUC**
TAN	0.8272
AW-NB	0.8207
NB	0.8244

Observing the results we chose the Bayesian model created with NB (Naïve Bayes) to be our final model, since it is more interpretable and simpler, while keeping practically the same results as the other two models.

6 EVALUATION

In the evaluation phase, we start by analyzing the results of the final model on the testing data separated on the beginning of this work. Finally, we analyzed the conditional probabilities of the features to extract useful knowledge regarding fighting corruption.

6.1 TESTING DATA

To ultimately validate our final model, we used the dataset separated in the data preparation phase for this purpose: the testing dataset (DTE). The first step here is to adjust DTE to have the same 23 final features selected from MDLP discretization.

Then, applying the final model on DTE we got AUC of approximately 0.76. Hence, we consider the results satisfactory. The reason being that the results are just a little below those obtained in the training dataset and are higher than 0.70, considered to be a threshold of good models.

6.2 KNOWLEDGE DISCOVERY

Observing the conditional probabilities of the final model, we extracted the rules it follows to define corruptibility for federal management units. This knowledge discovery aims to give a contribution to the activities of fighting corruption. Some of the main rules extracted that indicate an increase of risk of corruption are showed below.

- Accounts judged irregular by TCU;
- Responsibilities related to financial activities;
- Substitution public functions for controlling expenses;
- Number of requested civil servants allocated;
- Heading roles on regional agencies;
- Political party affiliations;
- Activities spread by multiple municipalities; and
- Number of public offices occupied by designation (without a selective process).

After discussing the main rules with DIE experts, they made a few comments in order to rationalize upon the knowledge discovered by the model.

- Accounts judged irregular by TCU are themselves by definition scenarios that involve inadequacies or irregularities;
- Responsibilities related to expenses and financial activities are critical, since they involve public resources and possible embezzlements;
- A management unit with several civil servants allocated by request might show a scenario of poor strength of the internal career;
- The heading roles related to regional management units usually have civil servants holding a relatively high amount of decision-making power with greater discretion, displaying a scenario of high propensity to corruption;
- Political party affiliations are related to greater political influence in decisions of public interest on the federal management units;
- Units with activities on many municipalities have to deal with decentralization problems; and
- Public offices employed by designation are occupied in the government due to nomination from discretionary authorities, not necessarily related to merit.

Therefore, by analyzing the rules together with the experts' comments, we see that the results have reasonable suitability in scenarios involving federal management units.

7 DEPLOYMENT

In the deployment phase, we created a Web application to allow managers at CGU to query management units and analyze their risk of corruption. With paths of grouped queries, managers can now view management units organized by their agencies. They are also able to perform ad-hoc queries, using as input unique identifiers of management units to obtain risk of corruption analysis for an individual unit or groups of them.

To deploy the predictive model to assess risk of corruption we simply implemented the calculation of Naïve Bayes with the conditional probabilities for the features selected on our final model. Using the output probabilities given by the model, we then discretized the results manually to only show risk categories, specifically: less than 0.20 as Very Low; equal or greater than 0.20 but less than 0.40 as Low; equal or greater than 0.40 but less than 0.60 as Medium; equal or greater than 0.60 but less than 0.80 as High; and equal or greater than 0.80 as Very High.

The Web application also generates pdf reports containing, for a given management unit: risk of corruption, average and maximum risk of corruption of the management units on the same agency. The application not only shows risk results, but also several other government data related to each management unit, allowing a general view of each unit.

With the application running, we started to present this work to all areas of CGU. Currently, several activities involving management units are being prioritized using our risk of corruption predictive model together with other information.

8 CONCLUSION

This paper described a data mining project that generated Bayesian models to assess risk of corruption of federal management units. We analyzed data from several government databases and, with the help of DIE experts, we developed thousands of important features. These variables were prepared and pre-processed removing those with zero class-variance and high inter-correlation.

Feature selection was done using Adaptive Lasso, which selected the 144 most relevant features. We compared two different discretization methods: CACC and MDLP. Bayesian models were built for datasets discretized with the two methods using the following algorithms: Naïve Bayes, Tree Augmented Naïve Bayes, and Attribute Weighted Naïve Bayes. To first choose the best discretization method we evaluated our results obtaining the average of the 10-fold cross-validation metrics performed per dataset. MDLP was chosen due to great results aligned with a considerable reduction of the number of features selected – from 144 to 23.

After choosing the dataset discretized with MDLP we evaluated the AUC for the three algorithms used on modeling. The results were very close, approximately 0.82. Therefore, we chose the model created with Naïve Bayes to be our final model, since it is more interpretable and simpler.

The dataset labeled Testing (DTE) separated on data preparation was then used to confirm the validity of the final model. DTE showed AUC of approximately 0.76.

Finally, the rules of the final model were extracted. With help from DIE experts, we derived knowledge for corruption fight activities. Rules generated and experts' comments were outlined to give an overview of the results.

The predictive model from this project was also deployed in a Web application, allowing managers from CGU to query and analyze federal management units regarding their risk of corruption. With the results of our model, CGU is already prioritizing corruption related activities to help maximize audits efficacy.

Therefore, this work contributed with an end-to-end data mining project overview, with application of several state-of-the-art techniques. We reinforced CGU's activities in fighting corruption by building an useful model to assess risk of corruption of federal management units. The knowledge discovered is also increasing the expertise of DIE analysts. With the Web application developed from this project, we help potentially save millions in public resources. Additionally, with risk assessment we encourage proactive audits, helping managers plan their work. To that end, we generate impact nationwide in fighting corruption.

Acknowledgements

The authors would like to thank the corruption fighting expert Victor Steytler for providing useful insights for the development of this work. Finally, the authors would like to thank CGU for providing the resources necessary to work in this research, as well as for allowing its publication.

References

R. Balaniuk, P. Bessiere, E. Mazer, and P. Cobbe. Risk based Government Audit Planning using Naïve Bayes Classifiers. *Advances in Knowledge-Based and Intelligent Information and Engineering Systems*, 2012.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Ricardo Carvalho, Rommel Carvalho, Marcelo Ladeira, Fernando Monteiro, and Gilson Mendes. Using political party affiliation data to measure civil servants' risk of corruption. In *2014 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 166–171. IEEE, 2014.

Rommel Carvalho, Shou Matsumoto, Kathryn B. Laskey, Paulo C. G. Costa, Marcelo Ladeira, and Lacio L. Santos. Probabilistic ontology and knowledge fusion for procurement fraud detection in brazil. In *Uncertainty Reasoning for the Semantic Web II*, pages 19–40. Springer, 2013.

S. Garcia, J. Luengo, J. A. Sáez, V. Lopez, and F. Herrera. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *Knowledge and Data Engineering, IEEE Transactions on*, 25 (4):734–750, 2013.

Mark A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999. URL `https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf`.

David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Keki B Irani. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.

Yufeng Kou, Chang-Tien Lu, Sirirat Sirwongwattana, and Yo-Ping Huang. Survey of fraud detection techniques. In *Networking, sensing and control, 2004 IEEE international conference on*, volume 2, pages 749–754. IEEE, 2004.

Daniel Lowd and Pedro Domingos. Naive bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 529–536. ACM, 2005.

EWT Ngai, Yong Hu, YH Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010. URL `http://arxiv.org/abs/1009.6119`.

Carlos Vinícius Silva and Célia Ralha. Utilização de Técnicas de Mineração de Dados como Auxílio na Detecção de Cartéis em Licitações. In *WCGE - II Workshop de Computação Aplicada em Governo Eletrônico*, 2010.

Sona Taheri, John Yearwood, Musa Mammadov, and Sattar Seifollahi. Attribute weighted naive bayes classifier using a local optimization. *Neural Computing and Applications*, 24(5):995–1002, 2014.

Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014.

Richard Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990.

Cheng-Jung Tsai, Chien-I Lee, and Wei-Pang Yang. A discretization algorithm based on class-attribute contingency coefficient. *Information Sciences*, 178(3): 714–731, 2008.

Stéphane Tufféry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011.

Fei Zheng and Geoffrey I Webb. Tree augmented naive bayes. In *Encyclopedia of Machine Learning*, pages 990–991. Springer, 2011.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006.