Identifying IT purchases anomalies in the Brazilian Government Procurement System using Deep Learning

Silvio L. Domingos*, Rommel N. Carvalho*†, Ricardo S. Carvalho† and Guilherme N. Ramos*
*Department of Computer Science (CIC)
University of Brasilia (UnB), Brasilia, DF, Brazil
Email: silviold@gmail.com, gnramos@unb.br and rommelnc@unb.br
†Department of Research and Strategic Information (DIE)
Ministry of Transparency, Monitoring and Control (MTFC), Brasilia, DF, Brazil
Email: rommel.carvalho@cgu.gov.br and ricardo.carvalho@cgu.gov.br

Abstract—**The Department of Research and Strategic Information (DIE), from the Brazilian Office of the Comptroller General (CGU), is responsible for investigating potential problems related to federal expenditures. To pursue this goal, DIE regularly has to analyze large volumes of data to search for anomalies that can reveal suspicious activities. With the growing demand from the citizens for transparency and corruption prevention, DIE is constantly looking for new methods to automate these processes. In this work, we investigate IT purchases anomalies in the Federal Government Procurement System by using a deep learning algorithm to generate a predictive model. This model will be used to prioritize actions carried out by the office in its pursuit of problems related to this kind of purchases. The data mining process followed the CRISP-DM methodology and the modeling phase tested the parallel resources of the H2O tool. We evaluated the performance of twelve deep learning with auto-encoder models, each one generated under a different set of parameters, in order to find the best input data reconstruction model. The best model achieved a mean squared error (MSE) of 0.0012775 and was used to predict the anomalies over the test file samples.**

1. Introduction

The Brazilian Office of the Comptroller General (CGU) was instituted in 2003, as an anti-corruption agency, with the main responsibility of helping the President in its attributions related to the national public assets protection and transparency [1]. The Department of Research and Strategic Information (DIE) is in charge of the activities of strategic information production to help the investigation of possible irregularities in federal expenditures.

In 2013, some changes were made [2] to prepare the department to cope with the growing public demand on transparency and anti-corruption measures. Improvements in its organizational structure were approved in order to better equip the department to cope with the investigative tasks under its responsibility.

Although being an government agency, like all enterprises, the CGU works with a limited amount of resources. So the investigation activities need to be prioritized according to the level of suspicion of any fact originated in the public federal administration. In this scenario, the DIE needs to prospect, evaluate, and select the IT tools capable of monitoring public expenses and discover relevant information that can point out a fraud or suspicious act. At this moment, a special interest resides on searching anomalies or misbehavior on federal government IT purchases.

In Brazil, the government procurement is regulated by the law 8,666 [3]. This law establishes all the processes, power of rights, types of purchases, and contract rules to be followed by all public administrators in the country. Unfortunately, despite of it, some cases of misbehavior and fraud have occurred during the the past years [4].

To prevent and avoid future problems of this nature, the DIE is constantly working on new methods of data mining. One of these goals is to automate processes within the CGU that help to prioritize the investigative actions in the Agency.

Among the various algorithms available, those related to machine learning are known to offer good results on finding relevant insights over structured and unstructured data. According to [5], these algorithms can be classified in tree main types of learning:

- Supervised: observes some samples of information (pairs input/output) previously classified (labeled data), and identifies the function that maps them;
- Unsupervised: learns the input data patterns without the need of previously knowing the output data labels (unlabeled data);
- By reinforcement: learns through a series of reinforcement information (feedback), consisted basically of rewards or penalties.

Depending on the data characteristics, it can be necessary to use different kinds of algorithms and this is one of the challenges faced by the CGU nowadays. The huge amount of data and the impossibility to previously label the suspicious or fraudulent samples requires other approaches.

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.106

The unsupervised algorithms pose as a promising choice in this context.

The present work concentrates on investigating and defining a new method to generate a predictive model to detect anomalies in the Federal Government Procurement System. To help organize and coordinate this effort, the CRISP-DM will be used as the data mining process. The modeling phase will analyze all IT purchases made between 2014 and 2015 by the Brazilian federal government, using the SIASG database, that CGU has access to. SIASG (http://www.comprasgovernamentais.gov.br/acesso-aos-sistemas/comprasnet-siasg) is the integrated system that manages all the bidding and contracting information about purchases made by the Brazilian federal government.

The predictive model generated is expected to be applied as a prioritization tool to help the CGU on selecting investigation initiatives that have more probability of success, contributing to the effectiveness of the Agency and helping to reduce the budget needed to carry out these tasks.

This paper is structured as follows. First, an introduction 1 about the context is provided to clarify the problem. Second, a brief explanation of the related work found 2 followed by the methodology 3 and data preparation 4 sections. Then, the modeling phase 5 will be described and, finally, we'll present the evaluation 6 and the conclusion 7 that we have reached through this work.

2. Related work

In this section we are going to present the related work on machine learning and procurement systems fraud identification found during our research.

Despite the use of data mining techniques in helping fraud detection and investigation, the fraudsters will eventually find a way to bypass the security measures implemented [6]. New approaches are being constantly studied and, in this context, various statistics and machine learning initiatives have presented good results [7] [8] [9]. Specifically on procurement fraud detection, some efforts have been made in the last years in the brazilian government with good results [10] [11] [12].

The most common initiatives, as these cited above, involve supervised learning or classification techniques, once labeled data is available. Nevertheless, some researchers are applying unsupervised learning with success, as well [13] [14], showing that it is possible to find patterns and uncover useful hidden information even in the absence of labeled data.

In the field of machine learning, there are recent studies showing good results obtained with the use of deep learning algorithms in various situations. According to LeCun [15] these methods have dramatically improved the state-of-the-art in speech recognition, visual objects recognition, object detection, drug discovery, and genomics. He concludes that unsupervised learning is expected to become far more important, in the longer term, since human learning is largely unsupervised, dependable of observation, not knowing the name of everything.

The application of deep learning algorithms covers a wide domain of problems:

- handwriting recognition [16];
- face detection [16];
- speech recognition and detection [16], [17]; [18] [19];
- object recognition [16], [20] and computer vision [21];
- natural language processing [16];
- sentiment analysis [22];
- audio [20] [23] and text [20];
- robotics [16];
- fraud detection [24] [25].

In 2002, Hand [26] stated that pattern detection was a new science, concerned with defining and detecting local anomalies within large data sets. An early application of unsupervised technique for anomaly detection can be found in the work developed in 2001 [13] over credit card fraud.

For Chandola [27], anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. In his study, he presented the advantages and disadvantages of several techniques and pointed out the power of clustering algorithms to find out patterns in unlabeled data.

There are some recent implementation of clustering algorithms to find suspicious activities in the health care domain. One example is the search for outliers in medical claims [28] that demonstrated the efficiency of the k-means clustering algorithm in assessing fraudulent behavior of the claimers. On another study over health care insurance fraud [29], k-means was successful applied as an unsupervised outlier technique to point out suspicious claims.

The focus of this experiment was to search for anomalies in purchases made by federal government departments and offices around the country. Once the available data was unlabeled, and considering previous studies over fraud detection, cited above, we decided to use an unsupervised algorithm to perform this task.

In addition, our interest was to evaluate a state-of-art algorithm that could be executed in an open stable platform with parallel processing resources. In our quest, we found out a recent implementation of the deep learning algorithm in a fraud detection initiative [24] using the H2O platform. Furthermore, the availability of this platform as a library in the R environment, one of the main tools used in CGU, helped to make this experiment possible.

3. Methodology

Following, we will detail the methodology, presenting the phases and the steps performed in this work.

The experiment described in this paper was developed considering the CRISP-DM framework [30], [31], [32]. CRISP-DM (Cross Industry Standard Process for Data Mining) is a complete and documented process and a common framework used in a lot of projects related to data mining.

All of his stages are well organized, structured and defined, helping to conduct and assess researches on this area.

One of its strong points is the focus on business problems. Nevertheless, the framework describes techniques of analysis, modeling, data mining, implementation, and delivering, all of them important to any data mining project [30]. It is important to note that the stages of implementation and delivery were not covered by this experiment. They are planned to be done afterwards.

After some discussion with the CGU specialists, the analysis, modeling, and data mining phases were conducted as described in the following sections.

4. Data Preparation

In this section we are going to show the preparation of the data (cleaning and transformation) to make it ready to be used in the modeling phase.

The data used in the experiment was provided by CGU and came from real procurement transactions occurred during the years of 2014 and 2015, in all federal government offices around the country. The initial dataset was composed of 137,035 records with 31 columns.

We performed an evaluation of the attributes contained in each record, checking if any column could be excluded from the input file. Initially, we verified that the purchase id, a unique number that identifies each purchasing event, should be omitted because of it this characteristic. In addition, two columns that where filled with identical values were not considered in the output file.

Then, since this data was related to real transactions, to avoid any secrecy problems, three columns that named departments or offices were excluded from the dataset. As the file had the identification codes for all of them, this deletion did not mean any information loss. The same procedure was done with six other categorical data that had an associated numerical identification code.

After that, the remaining columns were checked to see if there was any correlation among them. Some of the columns showed relevant correlations and were excluded from the dataset, once they would not give any further power of prediction to the model. One example was the correlation found between the superior department code and the budget unity code that presented a strong coefficient of 0,925266. Another two correlated attributes were identified during this process and excluded from the file.

Continuing the dataset preparation, a brief discussion on the need of any transformation in specific columns revealed that the purchase data should be treated to make this information useful to the modeling phase. So this column was derived in three, as described below:

- Month-of-year;
- Day-of-month; and
- Day-of-week.

Another important treatment realized over this file was the cleaning of the lines that had missing values. The business specialists considered that this procedure, that excluded 25,158 rows, would make the dataset more appropriated to produce a better anomaly detection model.

All of these data treatment tasks resulted in a dataset with 111,877 rows and 18 columns.

5. Modeling

In the following lines, we will present the details and discuss the steps that were performed during the modeling phase.

All the subsequent steps were done using the following software on a standard workstation running Windows 10 Operating System (Intel Core i5 64 bits processor and 8GB of RAM memory):

- R version 3.3.0 (a GNU project that provides a language and an environment for statistical and machine learning tasks) [33] [34];
- RStudio version 0.99.893 (an open source IDE-Integrated Development Environment that helps to run R commands and visualize the results) [35] [36];
- H2O build project version 3.8.2.6 (open source machine learning platform that can be called from R environment) [37] [38].

The file derived from the data treatment procedures were split in two parts by bootstrapping, in a proportion of 80% for training and 20% for testing. The number of rows in each set resulted to be 89,618 and 22,259, respectively.

These files were processed in the H2O platform, that offers a complete implementation of the deep learning algorithm with an auto-encoder option and an easy integration with RStudio. In addition, the H2O R library makes it possible to run an anomaly function to detect outliers on a given dataset using a chosen deep learning model available in the environment.

Before continuing the modeling phase explanation, it is important to briefly clarify how an auto-encoder works. An auto-encoder is an artificial neural network used to learn a representation of a set of data. It is an unsupervised learning algorithm that tries to approximate an identity function of the input information. So, the target output is the input itself, with a calculated margin of error, and the resulting model is capable of reconstructing the learned data with the least possible distortion [39], [40].

As this work was intended to run multiple learning tasks varying some algorithm parameters, in order to maximize the performance of the final model, it was necessary to search for an automatized process to support this strategy. An modeling with this characteristic was also possible because of the parallelism resources of the H2O platform allied to the R environment. The function chosen to do this task was $h2o.grid$, which performs a grid search over the provided hyper-parameters, automatically generating several different models and their respective results.

After an analysis of the parameters available in the H2O implementation of the deep learning auto-encoder algorithm, it was decided to run a combination of the following ones.

!ht

TABLE 1. HYPER-PARAMETER SEARCH SUMMARY

epochs	hidden	activation	MSE
10	15	Tanh	0.0013221
5	15	Tanh	0.0014603
5	10	Tanh	0.0016430
10	10	Tanh	0.0016440
10	5	Tanh	0.0021219
5	5	Tanh	0.0021662
10	15	TanhWithDropout	0.0021792
5	15	TanhWithDropout	0.0022387
5	10	TanhWithDropout	0.0024609
10	10	TanhWithDropout	0.0024774
10	5	TanhWithDropout	0.0028214
5	5	TanhWithDropout	0.0028697

The number of possible combinations of these parameters was twelve (2 x 3 x 2):

- Activation (neuron activation function): $Tanh$ and $TanhWithDropout$;
- Hidden (number of hidden layers): 5, 10 and 15; and
- Epochs (how many times the dataset should be iterated): 5 and 10.

The execution of the grid was then submitted to R and H2O over the training dataset. First, the model was executed with the H2O platform configured to use just one computing thread. With this configuration, the time needed to run all the combinations and generate the models was 48 minutes and 9 seconds. Afterwards, the number of threads was raised to four, causing a reduction in the execution time to 16 minutes and 54 seconds. These results confirmed the performance gain when using parallelism and the suitability of the platform to this kind of task.

The characteristics and performance of the models generated are detailed in Table 1, sorted by the their mean squared error (mse) in ascending order.

The MSE is a measure that evaluates the deviations between the input and the output data estimated by the model. In this experiment, we looked for the model with the lowest MSE because this value demonstrates its capability to reconstruct the input data with the minimum loss.

Among the 12 models generated, the one that used the neuron activation function $Tanh$, with 15 hidden layers and 5 epochs (iterations) presented the lowest MSE (0.0013221) and was chosen as the best model.

After that, the anomaly function over the testing file was executed to reconstruct each input record using the best model above. The overall MSE obtained from this execution was 0.0013238. To allow an individual analysis of the deviation between input and output test records, the function also provided the MSE (reconstruction error) for each row.

At last, we also executed the anomaly function using the training dataset to obtain its MSE deviation curve and compare it to the one generated with the testing data. The overall MSE reached at this time was 0.0013245.

6. Evaluation and Outcomes

In this section we are going to analyze the results and outcomes obtained in the modeling phase.

An evaluation of the parameters used to generate the models made it possible to see that the activation function $Tanh$ produced the best ones. Besides, the models with more hidden layers performed better than the ones with less.

Comparing the best model mean squared error (MSE) with the values generated by the anomaly test, we found out a gap of only 0.0000017 for the testing file and and 0.0000024 for the training file. These values confirmed that the model has the ability to generalize to a set of unknown data, what can be seen in the charts presented in the figures 1 and 2.

Figure 1. Training Data Deviation from MSE

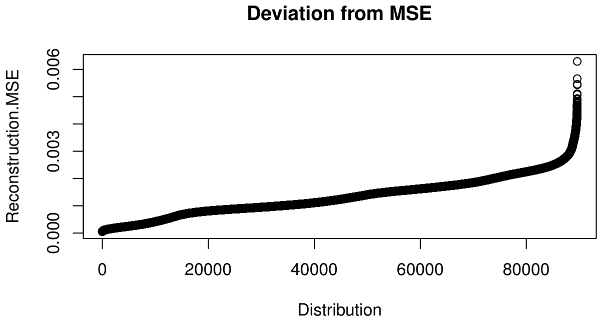
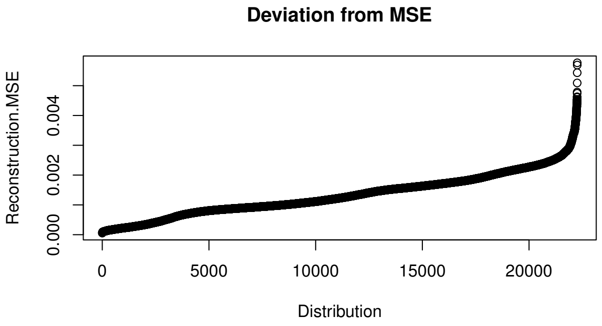


Figure 2. Testing Data Deviation from MSE



Another important information that can be visually found in these charts are the clear presence of outliers (input observations with an abnormal distance from the model's MSE) at the right upper corner, that corresponds to anomalies on the reconstructed purchase transactions. Hence, through this model it is possible to identify suspicious purchases in the input dataset and then evaluate its attributes to check if the event deserves to be prioritized for further investigation.

In order to check these findings, we selected five of the most suspicious transactions (higher MSE) and compared its attributes to a set of rows located around the test data's MSE. To compose this set, we extracted 19 records with MSE between 0.0013200 and 0.0013499 from the test file. A brief investigation showed that all the suspicious transactions presented differences in the program, action and credit codes. The availability of a model that discovers this type of hidden pattern is highly valuable to help investigation efforts.

Since the outcome of this work is to build an anomaly detection model that could help CGU to prioritize investigative actions, an adequate selection of the anomalous purchases must be implemented. A typical choice would be starting the analysis from the higher MSE, but there are other strategies that can be implemented, as establishing a cutoff or defining specific types of purchase to begin the investigation. These strategies will be discussed with the area specialists to evaluate which one is the best.

In addition, evaluating the parallel performance of the H2O environment, we found out that, in fact, there is a lot of gain when running the model with more than one thread. The computing time running with four threads corresponded to 35,1% when compared to the execution using with just one thread.

All the information collected in this experiment was very important and will be useful for future work on anomaly detection on this kind of data.

7. Conclusion and Future Work

In the next paragraphs, we will present our conclusion about the steps performed in this experiment and the results obtained, finalizing with some proposals for future work.

To help CGU in its pursuit to better allocate investigation resources on the most suspicious cases it is necessary to have good data mining prediction models. Using the deep learning with auto-encoder algorithm it was possible to generate an anomaly model with a good generalization performance. This will certainly be an useful tool that CGU can implement to discover relevant patterns and uncover suspicious and possibly fraudulent purchases.

The best model generated by the deep learning algorithm presented a low mean squared error (MSE) and was applied as an anomaly detection tool to the training and to the testing datasets, that showed very small MSE differences when compared to the base model. To better assess the results, we selected five rows with the highest MSE (most distant) and compared them with some samples located near the overall MSE, all of them from the testing dataset. We found out that the rows with the highest MSE had some attributes with no coincidence with the other samples. Those five rows were considered possible suspicious events that deserved further analysis, and a demonstration of the model's capability in pointing out anomalies on unlabeled data.

As was showed in this paper, the use of R and H2O together make it possible to implement a Deep Learning algorithm without much effort. Through the available functions a grid of auto-encoder models were generated and the best one was used to uncover anomalies in a real dataset containing purchases from a federal government procurement system. As showed in the processing results, the time spent to build all the models was much lower when using parallel processing and the H2O platform was capable of managing the computing resources to achieve this goal. Once the purchases in the Brazilian federal government represent a huge amount of data, it will only be possible to mine and find valuable information on it through the use of parallel processing.

As future work, it will be necessary to further analyze and investigate the anomalies identified to confirm its real behavior (fraud or not) and assess the model's performance. This task should be conducted by CGU specialists and we are sure that the results will be much more valuable than the one obtained here. Moreover, the resulting analysis could be used to better understand the model in order to improve it.

In addition, we consider it would be a desirable approach to broaden the grid variation with other parameters and to aggregate more attributes in the input data. This certainly will help the search for anomalous patterns on this kind of data.

Finally, considering the parallel processing outcomes discussed in this paper, it would be an important initiative to make more experiments on this matter. There are some configuration not experimented here, as the use of more cores/threads and clustering techniques, that deserves to be implemented. Besides, it would be interesting to work with bigger dataset sizes to analyze and compare differences in computing times and better evaluate the performance gain.

References

[1] Presidencia da Republica Federativa do Brasil, "Lei 10683," May 2003.

[2] ——, "Decreto 8109," Sep. 2013.

[3] ——, "Lei 8666," Jun. 1993.

[4] John Lyons and David Luhnow, "Brazils Giant Problem," Apr. 2016. [Online]. Available: http://www.wsj.com/articles/brazils-giant-problem-1461359723

[5] S. J. Russell, P. Norvig, and E. Davis, *Artificial intelligence: a modern approach*, 3rd ed., ser. Prentice Hall series in artificial intelligence. Upper Saddle River: Prentice Hall, 2010.

[6] Richard J Bolton and David J Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, Jan. 2002.

[7] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, 2010. [Online]. Available: http://arxiv.org/abs/1009.6119

[8] K. Chaudhary, J. Yadav, and B. Mallick, "A review of fraud detection techniques: Credit card," *International Journal of Computer Applications (09758887) Volume*, 2012. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.677.970&rep=rep1&type=pdf

[9] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *Networking, sensing and control, 2004 IEEE international conference on*, vol. 2. IEEE, 2004, pp. 749–754. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1297040

[10] R. Carvalho, E. de Paiva, H. da Rocha, G. Mendes, and B.-D. FederalBrazil, “Methodology for Creating the Brazilian Government Reference Price Database,” Dec. 2012. [Online]. Available: https://www.researchgate.net/profile/Rommel_Carvalho/publication/270820493_Methodology_for_Creating_the_Brazilian_Government_Reference_Price_Database/links/560739ac08aea25fce399974.pdf

[11] R. N. Carvalho, L. Sales, H. A. Da Rocha, and G. L. Mendes, “Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil.” in *BMA@ UAI*, 2014, pp. 70–78. [Online]. Available: http://ceur-ws.org/Vol-1218/bmaw2014_paper_7.pdf

[12] Rommel N Carvalho, 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*, ser. Lecture Notes in Computer Science, 2013, vol. 7123, pp. 19–40.

[13] R. J. Bolton, D. J. Hand, and others, “Unsupervised profiling methods for fraud detection,” *Credit Scoring and Credit Control VII*, pp. 235–255, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.5743&rep=rep1&type=pdf

[14] G. Cabanes, Y. Bennani, and N. Grozavu, “Unsupervised Learning for Analyzing the Dynamic Behavior of Online Banking Fraud.” IEEE, Dec. 2013, pp. 513–520. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6753964

[15] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[16] I. Arel, D. C. Rose, and T. P. Karnowski, “Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier],” *IEEE Computational Intelligence Magazine*, vol. 5, no. 4, pp. 13–18, Nov. 2010.

[17] G. E. Dahl, Dong Yu, Li Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[19] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6645–6649.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[21] D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

[22] X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 513–520.

[23] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[24] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” *Journal of Big Data*, vol. 2, no. 1, Dec. 2015. [Online]. Available: http://www.journalofbigdata.com/content/2/1/1

[25] Venkatatesh Ramanathan, “Fraud Detection with Deep Learning at Paypal,” h2o.university. [Online]. Available: http://university.h2o.ai/cds-lp/cds02.html?mkt_tok=3RkMMJWWfF9wsRonvanAZKXonjHpfsX56%2BkqUaG0lMI%2F0ER3fOvrPUfGjI4ATsBlI%2BSLDwEYGJlv6SgFTLTBMbBrwrgKXBk%3D

[26] D. J. Hand, *Pattern detection and discovery*. Springer, 2002.

[27] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Jul. 2009. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1541880.1541882

[28] T. P. Hillerman, R. N. Carvalho, and A. C. B. Reis, “Analyzing Suspicious Medical Visit Claims from Individual Healthcare Service Providers Using K-Means Clustering,” in *Electronic Government and the Information Systems Perspective*, A. K and E. Francesconi, Eds. Cham: Springer International Publishing, 2015, vol. 9265, pp. 191–205.

[29] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton, and J. van Hillegersberg, “Outlier detection in healthcare fraud: A case study in the Medicaid dental domain,” *International Journal of Accounting Information Systems*, vol. 21, pp. 18–31, Jun. 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1467089515300324

[30] Colin Shearer, “The CRISP-DM model - The new blueprint for data mining,” p. 8, 2000.

[31] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Citeseer, 2000, pp. 29–39.

[32] P. Chapman, J. Clinton, R. Kerber, Khabasa, Thomas, Thomas Reynartz, C. Shearer, and R. Wirth, “CRISP-DM 1.0 - Step-by-step data mining guide,” 2000. [Online]. Available: http://www.citeulike.org/group/1598/article/1025172

[33] “R.” [Online]. Available: https://www.r-project.org/

[34] Jared P Lander, *R for Everyone*, ser. Addison Wesley Data & Analytics Series, 2014.

[35] “RStudio,” Boston, MA. [Online]. Available: https://www.rstudio.com

[36] Julian Hillebrand and Maximilian H Nierhoff, *Mastering RStudio - Develop, Communicate, and Collaborate with R*, Birmingham, UK, Nov. 2015.

[37] “H2o,” Mountain View, CA. [Online]. Available: http://www.h2o.ai/#/

[38] Arno Candel and Viraj Parmar, “Deep Learning with H2o.pdf,” Feb. 2015. [Online]. Available: https://t.co/kWzyFMGJ2S

[39] Y. Bengio, “Learning Deep Architectures for AI,” *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009. [Online]. Available: http://www.nowpublishers.com/article/Details/MAL-006

[40] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures.” *ICML unsupervised and transfer learning*, vol. 27, no. 37-50, p. 1, 2012. [Online]. Available: http://www.jmlr.org/proceedings/papers/v27/baldi12a/baldi12a.pdf