2017 16th IEEE International Conference on Machine Learning and Applications

Identifying anomalies in parliamentary expenditures of Brazilian Chamber of Deputies with Deep Autoencoders

Thiago Alencar Gomes, Rommel N. Carvalho, Ricardo S. Carvalho
Department of Computer Science (CIC)
University of Brasília (UnB)
Brasília, Brazil
Email: {alencargomesthiago, rommel.carvalho, ricardosc}@gmail.com

***Abstract*—Each Brazilian Deputy receives a quota of money quota to cover the politician activity expenses, besides their salary. The amount of money reserved for that quota can sum up to almost 1 billion of Brazilian currency (approximately 300 million US Dollars) in a 4 year legislature. Civic society is using that data to perform independent auditing to verify expenses that are against the rules. This article presents the application of deep Autoencoders to identify anomalies in that data. The anomalies found indicate new suspicious expenses and several data quality problems in the data opened to the society.**

***Keywords*-Deep Learning; Autoencoder; Anomaly Detection; Open Data**

I. Introduction

This paper reports on an experimental study of the application of Autoenconders neural networks [1] to detect suspicious expenditures of Brazilian deputies. In recent years the Brazilian Chamber of Deputies (CD) made significant progress towards opening its data. Through an Application Interface Protocol (API) the CD provide structured information about deputies, voting sessions, legislative bills and expenditures of the quota for the exercise of parliamentary activity (CEAP - Cota para Exerccio da Atividade Parlamentar in portuguese).

Cruvinel [2] evidences that this data is relevant because the CD is an institution responsible to formulate laws that affect all the Brazilian citizen. Hence, it is used in several society segments like the Academic, Private Sector, Media and Government itself. In this work we focus on the CEAP data. CEAP refers to the extra value in each deputy monthly budget, besides the salary and gratifications. The rules for approving and the values limits are regulated in the Act nº 43 from 2009 [3]. These values are paid through reimbursement, so the parliamentary spends the money and submit an invoice to the CD to get the money back. The values limits are calculated based on the representation state of the deputy. This variation occurs because airline tickets prices vary from the capital of that state and Brasília, the federal capital of Brazil where the legislative work is done. The mean of the maximum values is R$40,256.17, times 513 deputies, times 12 months equals R$247,816,982.52. In a 4 years mandate it sums up to R$991,267,930.08, almost 1 billion of Brazilian currency (approximately 300 million US Dollars).

Besides the materiality, the normative has loose rules and the process of approval is not public or transparent. So the scenario is a lot of data available (9 years of daily expenditures in a total of 2,851,283 observations that are increasing daily) and a limited amount of resources to audit suspicious expenditures. Also, detecting fraud in this dataset is a problem of multivariate outliers: observations that are outlying in multiple dimensions. For instance, a fraudulent meal reimbursement may be outlying in its price dimension and in its place dimension, in other words, a expenditure with bananas of R$300.00 in a gas station. Thus, unsupervised learning is a proponent choice for this scenario. Therefore, the present work aims to explore and apply new algorithms to generate a model that identifies anomalies and outliers in the reimbursement of the quota for the exercise of parliamentary activity. We choose Autoencoders based on two successful applications in fraud detection [4], [5] in the Brazilian government. Hence, we use H2O and R to perform our study on the observations that refer to 9 years of reimbursement, from 2009 to 2017. To guide the research and generate more consistent results the Cross Industry Standard Process for Data Mining (CRISP-DM) [6] will be used. We expect that the new model will serve as a prioritizing tool to help the team, other citizens and other public agencies in auditing and monitoring the deputies expenditures. This article is structured as follow: Section II presents the background knowledge about anomaly detection, Autoencoders and its application in Brazilian government data. Section III presents the detailed execution of the CRISP-DM process, starting from the Business Understanding to the Model Evaluation activities. Finally, in Section IV we present our conclusion and the possibles future works.

II. Background

Baesens, Vlasselaer and Verbeke [7] explain that there are two types of outliers: valid and invalid. An example of the first one is a person with salary of US$1,000,000.00, its a true fact but it stands out from the rest of

0-7695-6321-X/17/31.00 ©2017 IEEE
DOI 10.1109/ICMLA.2017.00-33

An invalid observation is a person with 300 years old. Those are univariate outliers. Nonetheless, the analysis can be more complex as the outliers cannot be detected in an unidimensional view of the data. Hence, analysis are carried out by more than one dimension, which characterize multivariate outliers. Autoencoders aim to reduce the original feature space in order to extract the essential aspects of the data. Fig. 1 shows a schematic architecture of an autoenconder network with a single hidden layer. The input data has 5 variables or features that serve as input for the the input layer. Then the network tries to represent these 5 features in 3 layers: the first layer has the same number of neurons as the number of features. The second layer, called hidden layer, represents those 5 features with 3 neurons (focusing on identifying general patterns on the underlying data) and, finally, the neural net tries to reconstruct the data in the output layer with the same number of neurons as the number of features.

In the Brazilian government, several works have been made with the use of data mining to detect outliers and focus auditing efforts on the most suspicious transactions. Domingos et al. [4] used Autoencoders to isolate suspicious Information Technology (IT) purchases from Brazilian Federal Government. The presented experiment followed CRIPS-DM process and compared the performance of the algorithm on a single-thread versus a multi-thread environment using H2O platform. As results, it generated a model that can prioritize suspicious IT purchases by the analysis of the highest Mean Square Error (MSE) and it was possible to gain 35,1% in time cost with the application of the algorithm running in four threads. With the same approach, but with highly dimensional data derived from 10 different datasets, Paula et al. applied Autoencoder in databases of foreign trade of the Secretariat of Federal Revenue of Brazil to identify exporting corporations whose explanatory variables show signs of divergence (anomalies) compared to regular patterns found. First, the authors used Gradient Boosting Machines(GBM) to filter the most important variables that explain the variability of exported volumes: from the initial 80 variables, the GBM selected 18 as most important. Then, the variables were transformed in relative indexes that represent the participation of each attribute in the amount exported. Finally, the Autoenconder with 6-3-6 hidden layers were used, MSE was used to select the more anomalous transactions and finally performance was compared with 1, 2, 3 and 4 processors cores. The last setup showed a decrease of 11.8% in time cost. There was no academic publication that uses data mining techniques in the CEAP data. Our work aims to build shallow and deep autoencoders models that can detect suspicious deputies expenditures for further analysis.

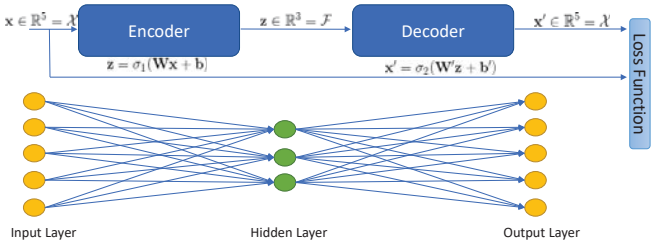

Figure 1. Scheme of a shallow Autoencoder with 3 neurons in the hidden layer

III. Methodolgy and Results

This study used the well-known data mining reference model Cross Industry Standard Process for Data Mining (CRISP-DM) [6]. This section presents a summarized description of each phase and a report of activities and results obtained. The business understand phase is not described because the results were obtained by the interaction with the SME team and the CD documentation and are presented in the Section I.

In the Data understanding and Data preparation phases the data was collect from three sources: Chamber of Deputies CEAP files: Comma Separated Files (CSV) files, made available in the website, divided by year. Actually, there are 9 files that contain data from 2009 to 2017. Chamber of Deputies open data API: from that, we read XML files that contain detailed information about each deputy through the endpoint. Companies data from Secretariat of Federal Revenue of Brazil (SFRB): this data refers to data about each company that received a payment from any deputy. This three datasets were cleaned and joined resulting in a dataset with 1,473,358 observations and 21 variables. The Table I explains each variable.

In the Modeling and Evaluation phases modeling techniques are selected, applied and their parameters adjusted to approximate the best values. As stated before, the model selected was Autoenconder neural networks. The choice was made because this is an efficient and simple technique to detect outliers and the results can be applied to enhance other models in two ways: first, the model reduce and shows the principal variables or components that represent the dataset. Second, the observations with lower reconstruct mean-square error (RMSE) can be used to enhance a classification model. The environment setup used for running the model was a laptop with Intel Core i7-5500U processor with 2 cores, 4 threads and 16GB of RAM memory. The H2O cluster used all 4 threads and a maximum of 12GB of the RAM memory. Furthermore, we use undercomplete Autoencoders, in other words, with a lower dimensionality than our data. This characteristic takes care of the regularization for the model. Hence, we did not use dropouts as a technique for avoiding overfitting. We used a random split

Variable	Type	Description
sgUF	Categorical	Brazilian State that the deputy represented at the time of the expense
sgPartigo	Categorical	Abbreviation of the party name represented by the deputy at the time of the expense
codLegislatura	Categorical	4 years period that represents the deputy mandate
numSubcota	Categorical	Type of expense. There are 21 types, e.g. Food and Air Tickets
numEspecSubCota	Categorical	A more detailed description of the expense type
txtCNPJCPF	Categorical	Identification number that represents each company. This identifier was maintained because an outlier can be detected for a specific company
indTipoDocumento	Categorical	Type of the receipt presented by the deputy. '0' for bill of sale, '1' for simple receipt and '3' for a combination of the above.
datEmissao	Date	Date of the expense. Represents the day of the year
numMes	Categorical	Month of financial competence of the expense
numAno	Categorical	Year of financial competence of the expense
vlrUpdated	Numeric	Updated value of the expense by the IPCA with 2017 as reference
sexo	Categorical	Gender of the deputy that made the expense
situacao	Categorical	Status of the deputy at the time of expense. There are 4 levels: 'In exercise', 'In license', 'In substitution', 'In vacancy'
ageD	Numeric	Deputy age at the time of expense
ageC	Numeric	Company age at the time of expense
capital_social	Numeric	Value of company's monetary value
city	Categorical	City where the company is situated
legal_entity	Categorical	Type of the company in the SFRB
main_activity_code	Categorical	Principal activity of the company
type	Categorical	Indicates if the company is the parent company or subsidiary
state	Categorical	Brazilian State where the company is situated

Table I
VARIABLES FROM THE FINAL DATASET WITH TYPES AND DESCRIPTIONS

of 80% for training and 20% for validation for all model generation. The validation in this case is used to verify if there was overfitting.

We used the H2O framework random grid search to build 29 models. We ordered the models by MSE and selected the two with the lower MSE. The random grid search was selected based on Bergstra and Bengio [8] work that explains: "Compared with neural networks configured by a pure grid search, we find that random search over the same domain is able to find models that are as good or better within a small fraction of the computation time. Granting random search the same computational budget, random search finds better models by effectively searching a larger, less promising configuration space". Table II summarizes the information about the two models with the best MSE. For hyperparameters tuning we used:

1) Number of neurons in each hidden layer: {15},{18},{5,3,5},{15,9,15},{9,9,9},{7,7,7},{6,3,6}
2) Number of Epochs: 1 to 10 epochs
3) Activation: Tahn and Rectifire
4) Maximum Number of Models: 100
5) Stopping Metric: MSE
6) Stopping Tolerance: 0.00001
7) Stopping Rounds: 5

Table II
MODELS BUILT WITH THE WHOLE DATA AND THE SUMMARY FOR EACH ONE

Model	Training MSE	Validation MSE	Training RMSE	Validation RMSE
Shallow	0.0002009042	0.0002009143	0.01417407	0.01417443
Deep	0.0001636939	0.0001634949	0.01279429	0.01278651

The two best models were a shallow and a deep model. The shallow was built with 1 epoch and 18 neurons in the hidden layer, and the deep one with 7 epochs and 6-3-6 neurons in each of the 3 hidden layers, respectively.At a first look, it seems that two good models were built. Both with very low MSE and with no overfitting. With further analysis in the models scoring history, the shallow model showed a low variance in the mean rates and weights. To verify that, we plotted a histogram of the distribution of the MSE for each model in training and validation data with a dotted line representing the 95% percentile. Fig. 2 shows that the deep model is better to distinct each observation and we get more reliable outliers: the distribution of the errors indicates that the deep model learned the variance within the data and points to specific observations. Where the shallow model points to more than 400,000 observations. Fig. 3 shows the distribution error using a different color for each expense type and shows that the model generalizes and identifies anomalies in heterogeneous types in the validation dataset. The horizontal line delimits a tail in the 95% percentile, based on Wiley's [9] methodology all expenses plotted above are considered anomalies because are harder to reconstruct compared to the rest of it and are good candidates for further analysis.

IV. CONCLUSION

The application of Data Mining in government open data is a good opportunity to enhance civic participation and public accountability. Using deep Autoencoders it was possible

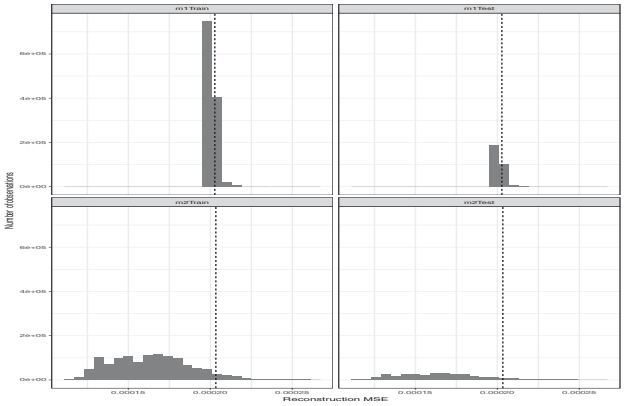

Figure 2. Histogram of the error from shallow and deep model. "m1" label represents the shallow model and "m2" the deep one

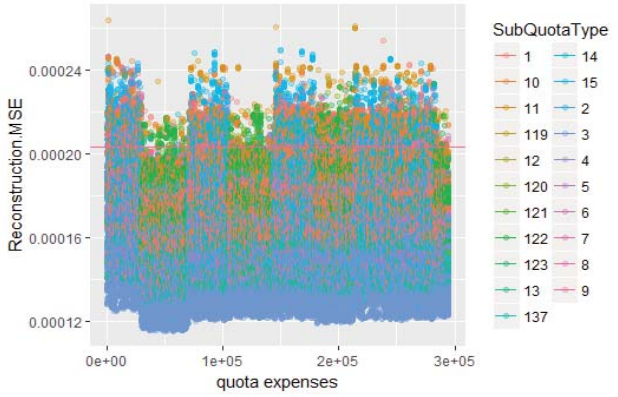

Figure 3. Scatter plot of the reconstruction MSE for the validation dataset

to generate models with good generalization performance that will be useful to civil society and the government itself for better analysis of congresspersons expenses with public money. To choose the best model we verified the distribution of errors for each model and selected the one that presented more reliable anomalies. We built 29 models and selected the two with lowest MSE and made the same comparison to select the best one. We considered anomalies the records that was above the 95% percentile of the errors. Through the analysis of those records we discover suspicious expenses that are against CEAP rules. Furthermore, we highlight several records that seem to contain error on the input of the data by the Chamber of Deputies. The construction of these models and analyses was facilitated by the use of R and H2O. Our dataset has 16 categorical variables that results on a 80.425 different levels, which makes the process slow for our available computational resources. The H2o parallelization and random grid made possible to build and analyze the models in feasible time. For the deep model, built with the whole data, our model detected 14.667 new suspicious observations. For instance, we identified a meal with a value of R$823.45, from a restaurant with a median of R$289 for an individual meal, that indicates this was not a meal for only one person as the defined by the rules. Another suspicious meal detected was a meal paid for an energy company. In another expense type, the model detected telephony bills with high values within election months, it indicates that the money could be used for election expenses and not for parliamentary activity. So, we can see that the model can detect anomalies in different dimensions. For future work we intend to analyse all detected anomalies to build a dataset of suspicious expenditures validated by the Chamber of Deputies and use to make and evaluate classification models.

REFERENCES

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[2] G. W. F. Cruvinel, "Dados governamentais abertos: um diagnóstico da demanda na Câmara dos Deputados," Master's thesis, Centro de Formao, Treinamento e Aperfeioamento da Cmara dos Deputados/Cefor, 2016. [Online]. Available: http://bd.camara.gov.br/bd/handle/bdcamara/28579

[3] BRASIL, Câmara dos Deputados, "Ato da mesa nº 43, de 21/05/2009 - Institui a Cota para o Exercício da Atividade Parlamentar," May 2009.

[4] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, and G. N. Ramos, "Identifying it purchases anomalies in the brazilian government procurement system using deep learning," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016.

[5] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagão, "Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering," in *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE, 2016, pp. 954–960.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," 2000.

[7] B. Baesens, V. V. Vlasselaer, and W. Verbeke, "Predictive analytics for fraud detection," *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*, pp. 121–206.

[8] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[9] J. Wiley, *R Deep Learning Essentials*, ser. Community experience distilled. Packt Publishing, 2016. [Online]. Available: https://books.google.com.br/books?id=U5njCwAAQBAJ