USE OF MACHINE LEARNING FOR ACTIVE PUBLIC DEBT COLLECTION WITH RECOMMENDATION FOR THE METHOD OF COLLECTION VIA PROTEST

Álvaro Farias Pinheiro[1], Denis Silva da Silveira[2]
and Fernando Buarque de Lima Neto[1]

[1]Polytechnic School, University of Pernambuco, Recife, Brazil
[2]Department of Administrative Sciences,
Federal University of Pernambuco, Recife, Brazil

ABSTRACT

This work consists of applying supervised Machine Learning techniques to identify which types of active debts are appropriate for the collection method called protest, one of the means of collection used by the Attorney General of the State of Pernambuco. For research, the following techniques were applied, Neural Network (NN), Logistic Regression (LR), and Support Vector Machine (SVM). The NN model obtained more satisfactory results among the other classification techniques, achieving better values in the following metrics: Accuracy (AC), F-Measure (F1), Precision (PR), and Recall (RC) with indexes above 97% in the evaluation with these metrics. The results showed that the construction of an Artificial Intelligence/Machine Learning model to choose which debts can succeed in the collection process via protest could bring benefits to the government of Pernambuco increasing its efficiency and effectiveness.

KEYWORDS

Data Mining, Artificial Intelligence, Machine Learning & Public Debt Collection.

1. INTRODUCTION

According to Witten, it is estimated that data from organizations double every 20 months, and this large amount of data is increasingly difficult to use for decision making [1]. According to Wirtz, this problem is even more pronounced in the public sector [2]. In this context, the application of Artificial Intelligence (AI) and Machine Learning (ML) techniques has been shown to be increasingly appropriate to solve this problem [1]. And, as pointed out by Gousios, there is more and more data to be used in software engineering associated with data science techniques [3] to find descriptive and predictive information, in several areas, from credit analysis [4] to the billing process [5].

According to Hunt, the debt collection process is increasingly using Artificial Intelligence for better results. And the application of these techniques in the public sector that deal with debt collection becomes even more relevant for two reasons: (1) the need to differentiate debtors from tax evaders; and (2) the legal obligation to collect, regardless of the amount. The first is related to the need for greater assertiveness in the process, and the second is related to efficiency. Both are expected contributions using AI [6].

David C. Wyld et al. (Eds): AIAPP, NLPML, DMA, CRIS, SEC, CoSIT, SIGL - 2022
pp. 99-108, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120909

Arising from the above argument, the following challenges are commonly encountered in government agencies to adequately carry out public debt collections: (1) dealing with the large volume of data [1], and (2) using reliable and agile mechanisms to carry out debt [6].

This article is organized into six sections. This section presented the objective for the development of this work. The second section dealt with motivation. The third section described the theoretical foundation. The fourth section presented the research method used. The fifth section presented the results. And the sixth section presented the conclusions, with applicability for future work. Finalizing the document with the references.

2. MOTIVATION

The Attorney General of the State of Pernambuco (AGS/PE) has been using a tool for Business Intelligence (BI), called Qlik Sense since 2019, and with it, descriptive analyses were performed to obtain knowledge of the data, which are available to the organization stored in the Oracle database of this institution, and data inserted and maintained by transactional application called the Justice Automation System (JAS) implemented since 2006.

With these data, added to several others stored in the SQL Server database of this body, from applications integrated in the platform called Portal-AGS/PE, it was possible to better understand the evolution of active debt over the years, and in this understanding, it was observed the need to better understand how data behaves.

AGS/PE being the state government agency, responsible for collecting active debts, with the knowledge obtained with the BI tool, observed the annual increase in the amount of debt and the number of debtors, which motivated the use of Business Analytics (BA) to identify the causes and consequences and in the search for its solutions, with a focus on improving the collection processes.

With the purpose of assisting in this process, in this article, we use AI and ML techniques to identify which debts are most appropriate for the protest collection modality. Thus, the AGS/PE Active Debt Center will have a less costly tool for public coffers than electronic court. This was possible by training smart techniques based on data from the Active Debt Registration (ADR).

3. THEORETICAL BASIS

This section provides an overview of the techniques used to perform Supervised Machine Learning. In addition, it also presents the metrics used for the analysis of classification techniques.

3.1. Supervised Learning Techniques

Neural Network (NN) allows you to perform complex computations through a training function on a dataset. Thus, NNs can be seen as approximations of nonlinear functions, e.g., classification or regression. There are several parameterization and models, usually have an input layer and one, output, and one or more intermediate layers [7].

Logistic Regression (LR) allows estimating the probability associated with the occurrence of a given event in front of a set of explanatory variables, being a statistical technique that aims to model, from a set of observations, the logistic relationship between responses and a series of numerical or categorical explanatory variables [8].

Support Vector Machine (SVM) allows you to generate a representation of examples as points in space, mapped so that the examples in each category are divided clearly and accurately. Thus, new input cases are then mapped appropriately as belonging to one of the categories of the output space. Therefore, what an SVM does is find a separation line, a hyperplane, between data from multiple classes. That is, the hyperplane seeks to maximize the distance between the closest points relating to each of the existing classes [9].

3.2. Classification Analysis Metrics

Accuracy (AC) is the ratio between true positives (TP) and true negatives (TN) for the sum of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [10].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (1)$$

Precision (PR) is the ratio between true positives (TP) for the sum of the number of true positives (TP) and false positives (FP) [10].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2)$$

Recall (RC) is the ratio between true positives (TP) for the sum of the number of true positives (TP) and false negatives (FN) [10].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

F-Measure (F1) is twice the ratio between multiplying precision by the recall to precision and recall [10].

$$\text{F1} = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

Receiver Operating Characteristic (ROC) is a metric for comparing the performance of the models, represented by the area under the ROC curve. The ROC curve is plotted as a diagram of true positive values (TP) as a function of the false positive ratio (FP). The closer the value is to 1, the better the classifier performance [10].

Finishing, True Positives (TP) are tests that are passing because the application is behaving as expected. True Negatives (TN) are tests failing due to real failures. False Positives (FP) are tests that pass, but they shouldn't pass. And, False Negatives (FN) are tests failing, however, due to inconsistency in the test itself [10].

4. METHOD

It is through the scientific method that the researcher's way of proceeding to the conclusion of research to achieve a goal is defined [11]. Thus, the goal of this section is to describe in the broadest sense, the application of the above techniques based on the data obtained belonging to AGS/PE and referring to the registration of active debts.

The first step of the method was the identification of the problem. Thus, at this stage, we sought to identify problems in the tax foreclosure that prevented the financial recovery of the state.

Therefore, a classification of the debts entered was carried out, to be able to learn, based on the occurrences, which debts were appropriate to the protest collection modality. Being more specific, several pieces of training of techniques already named was carried out to identify in the Active Debt Registration (ADR) which debts are most appropriate.

The idea here was: the characteristics that pointed to a greater assertiveness in the automated sending of lots of ADR to protest, would be decreased in the number of ADR returned because they did not fit the rules of this modality.

The second step was data mining because to apply Artificial Intelligence / Machine Learning it is necessary to first perform data mining. The Cross-Industry Standard Process for Data Mining (CRISP/DM) technique was used, performing the activities, as described by Chapman, which are: (1) understanding of the business; (2) understanding of the data; (3) data preparation; (4) data modeling; (5) evaluation of the data; and (6) deployment. Repeating the process as many times as necessary, until the mined data are satisfactory [12].

The goal of this step was to determine which data set is the most appropriate to solve the problem, and how this data should be standardized, and balanced, to avoid bias in the learning process. And for this problem, which aims to predict which active debt (ADR) should be protested electronically or not, the following data were selected:

- CAD: number of the certificate of the active debt.
- COMPANYNAME: name (company name).
- OCCURRENCE, type of occurrence of the debt, being canceled, returned, paid, protested, withdrawn, and held.
- SITUATION: debt situation, being it: awaiting regularization, with active installment, settled, exhausted installment, under-defense and under-defense, and guaranteed judicial.
- PROCESS: the type of the process: active, canceled, and liquidated.
- BALANCE: the amount of the outstanding balance.
- NOTARYPUBLICOFFICE: the identification of the protest office.
- PROTOCOL: the protest protocol number.
- DTOCCURRENCE: the date of occurrence.
- DTSUBSCRIPTION: the date of registration of the active debt.
- IRREGULARITY: the description of the irregularity, in most occurrences, were: invalid transferor, incompatible ZIP Code, insufficient address, uninformed, other jurisdiction, and insufficient time.
- PROVIDENCE: the type of providence, which can be a cancellation with charge, cancellation without charge, and not informed.

After data selection, the next phase was data elaboration, and the activities of cleaning, creation, integration, and formatting of the data were carried out, to identify which are the independent variables, and the ones identified were: OCCURRING, SITUATION, PROCESS, BALANCE, NOTARYPUBLICOFFICE, DTOCCURRENCE, DTSUBSCRIPTION, and IRREGULARITY.

With the purpose of supporting the definition of the new variables created, with the goal of being used in the calculation of dependent variables. They are, OCCURRENCEDAY and SUBSCRIPTIONDAY, were created.

In this process, we chose to remove the PROTOCOL variable because it was not adding any value to the problem. And, the CAD variable because it is a unique identifier that does not have duplicity and does not contribute to the process of extracting characteristics, was anonymized, and left as a goal only for the identification.

This is to keep the debtor's identification in absolute secrecy, due to the confidential nature of the taxpayer's data and the legal requirements of the General Data Protection Act (GDPA). Therefore, as the data from this sample are real, we chose not to show the debtor, and the COMPANYNAME variable was also extracted, thus ensuring its anonymity.

The variables SITUATION, IRREGULARITY, and PROVIDENCE were divided into several binary variables with the application of normalization. Finally, the variable OCCURS was marked as the target variable with the values representing success or failure in the protest process.

After this elaboration, all independent variables were normalized to balance the values that were at different or very heterogeneous scales. For this process, the method of obtaining the Y of X was used, based on the rule Y = (X-MIN)/(MAX-MIN), to be used as data entry in the tested models, avoiding bias.

In the next step, the selection process was applied, using the vertical selection technique, selecting only the records whose PROCESS criterion was equal to 'active', ignoring the 'canceled' and 'settled' because there is no point in calculating the dependent variable of processes that are not in progress. Following the same logic, the tuples that had the variables with null values were also extracted, resulting in 6966 tuples available for learning.

With the cleaning and transformation of the data performed in all variables, to better adapt the classification techniques for supervised training, the Google Colaboratory tool [14] was used for the test stage of the models, aiming to verify which of the chosen techniques is the most appropriate.

The tests of the models were performed with varying hyperparameters, to identify which configuration was the most applicable to the data treated. Emphasizing that the classification techniques applied to the data were NN, LR, and SVM, and all training was performed using the Orange Canvas tool [13].

During the training of the selected models, the evaluation stage was performed. In this step, it was possible to observe how the results obtained with the techniques met the specified problem. In this study, the evaluation was performed through techniques that seek to influence better decision-making, through training for classification through metrics, such as Accuracy, Precision, Recall, F-measure, and area under the Receiver Characteristic Operating (ROC) curve. The results obtained in this evaluation phase are presented in the results section.

Finally, the last phase performed was the implementation, in this phase all the learning obtained, through data mining, training, testing, and validations with the choice of the best model based on the best values in the analyzed metrics were saved, generating a pickle file to be able to use it when necessary, allowing the serialization of objects to be used in python applications.

5. RESULTS

In this section, we present the results of the application of the techniques with the evaluation by the metrics Accuracy (AC), F1, Precision (PR), and Recall (RC). Thus, after the application of the techniques, it was observed that the Neural Network that obtained the highest RATE of AC was 98%, with F1 with 97%, PR with 97%, and RC with 98%.

Table 1 shows the settings of the hyperparameters that were applied to the neural network until they reach the best value. The column 'L' represents the number of layers used in the experiment, the column 'N' the number of neurons per layer, the column ‘F’ is the function, the column 'S' the method applied, and the column 'R' the learning rate used that was 0.0001 for all, with the best results in bold.

Table 1. Values used in NN hyperparameters and the results evaluated by AC, F1, PR and RC metrics

L	N	F	S	R	AC	F1	PR	RC
1	32	ReLu	Gradient	0.0001	0.978	0.971	0.971	0.978
1	64	ReLu	Gradient	0.0001	0.978	0.971	0.970	0.978
1	128	ReLu	BFGS	0.0001	0.978	0.970	0.969	0.978
1	256	ReLu	Adam	0.0001	0.978	0.970	0.969	0.978
1	512	ReLu	Adam	0.0001	0.978	0.976	0.974	0.978
2	64	Adam	Gradient	0.0001	0.977	0.968	0.967	0.977
2	64	ReLu	Gradient	0.0001	0.979	0.972	0.972	0.979
2	64	Hyperbolic	Gradient	0.0001	0.978	0.971	0.971	0.978
2	64	ReLu	BFGS	0.0001	0.979	0.977	0.975	0.979
3	64	ReLu	Gradient	0.0001	0.978	0.971	0.970	0.978
3	128	ReLu	Gradient	0.0001	0.978	0.971	0.971	0.978
3	128	ReLu	Gradient	0.0001	0.979	0.972	0.972	0.978
4	256	ReLu	Gradient	0.0001	0.979	0.972	0.972	0.978
4	**256**	**ReLu**	**Adam**	**0.0001**	**0.980**	**0.978**	**0.977**	**0.980**

Table 2 shows the regularization hyperparameters represented by column 'R' and Strength by column 'S' used in the LR model with the application of the metrics Accuracy (AC), F1, Precision (PR), and Recall (RC), with the best results in bold.

Table 2. Values used in LR hyperparameters and the results evaluated by AC, F1, PR and RC metrics

R	S	AC	F1	PR	RC
L1	80	0.978	0.977	0.971	0.975
L1	70	0.978	0.977	0.971	0.975
L1	60	0.978	0.977	0.971	0.975
L1	50	0.978	0.977	0.971	0.975
L1	40	0.978	0.977	0.975	0.975
L2	50	0.977	0.967	0.965	0.975
L2	40	0.979	0.977	0.975	0.975
L2	30	0.978	0.977	0.975	0.975
L2	20	0.979	0.977	0.975	0.975
L2	10	0.978	0.978	0.976	0.978
L1	30	0.978	0.978	0.976	0.978
L1	20	0.979	0.978	0.976	0.978
L1	10	0.979	0.978	0.976	0.978
L1	**1**	**0.980**	**0.978**	**0.977**	**0.980**

Table 3 shows the Cost hyperparameters represented by column 'C' and Regression by column 'R' used in the SVM model with the application of Accuracy (AC), F1, Precision (PR), and Recall (RC) metrics, with the best bold results.

Table 3. Values used in VMS hyperparameters, and the results evaluated by AC, F1, PR and RC metrics

C	R	AC	F1	PR	RC
0.10	0.40	0.974	0.970	0.970	0.977
0.10	0.30	0.974	0.970	0.970	0.977
0.10	0.20	0.974	0.970	0.970	0.977
0.10	0.10	0.974	0.970	0.970	0.977

0.50	0.50	0.975	0.970	0.970	0.977
0.50	0.40	0.975	0.960	0.960	0.977
0.50	0.30	0.975	0.970	0.970	0.977
0.50	0.20	0.975	0.970	0.970	0.977
0.50	0.10	0.975	0.970	0.970	0.977
1.00	0.50	0.975	0.971	0.971	0.978
1.00	0.40	0.976	0.971	0.971	0.978
1.00	0.30	0.976	0.971	0.971	0.978
1.00	0.20	0.976	0.971	0.971	0.978
1.00	**0.10**	**0.978**	**0.971**	**0.971**	**0.978**

Table 4 presents the results of the metrics applied to compare performance between the 3 classification techniques used: the neural network, logistic regression, and the support vector machine. With the configuration of 4 layers, with 256 neurons in each layer, using the ReLu activation function with the Adam technique and learning rate of 0.0001, the neural network, proved to be the best model among the 3 tested, as can be observed in the following table, with the best results in bold.

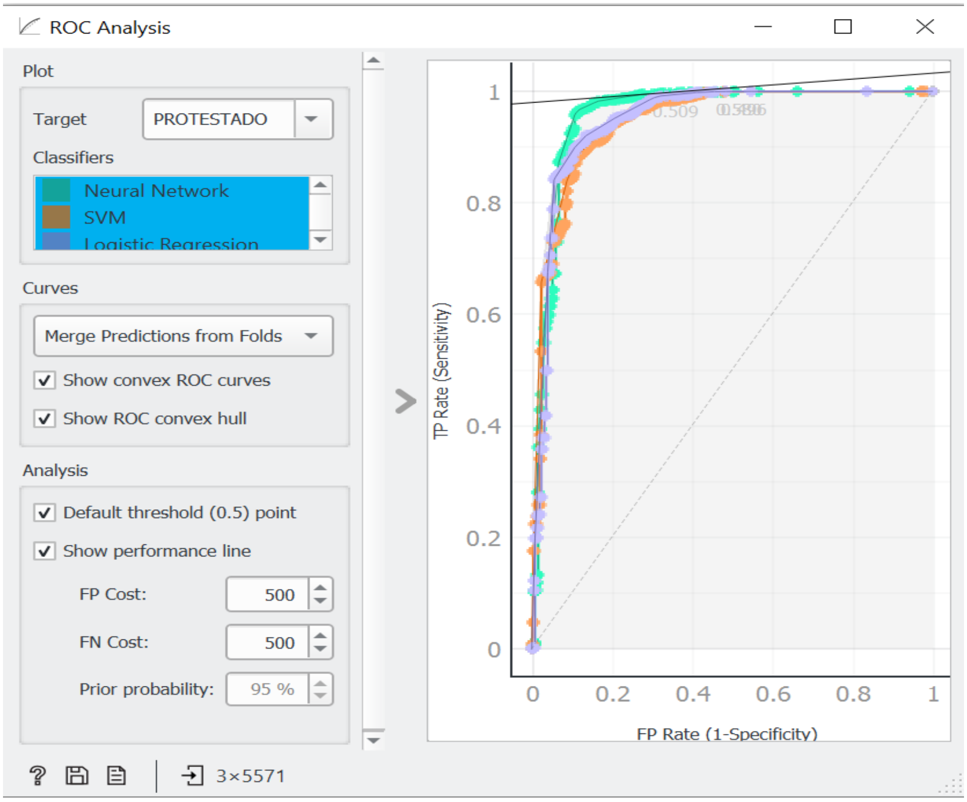
Table 4. Comparison of the results obtained with the experiments between the techniques of NN, LR and SVM

Technique	**AC**	**F1**	**PR**	**RC**
LR	0.978	0.971	0.970	0.978
SVM	0.978	0.970	0.971	0.978
NN	**0.980**	**0.978**	**0.977**	**0.980**

The data set used in this experiment had 6966 records, with the variable OCCURRING being the target, with 5 categorical variables, 7 numerical, and 3 textual variables. Being used for this experiment 6 variables for extraction of characteristics, the independent variables: SITUATION, BALANCE, EVENTS, SUBSCRIPTIONDAY, IRREGULARITY, and PROVIDENCE.

Targeting the VARIABLE OCCURS and the target variables, only for identification, CAD_ANONYMOUS, and NOTARYPUBLICOFFICE. For the set, only the paid occurrence was discarded because there were not enough instances to balance the training, since the data set included only the data from the years 2018, 2019, and 2020. Being used in training and evaluating a subset corresponding to 80% of the data, resulting in 5571 records, with the remaining 1392 for validation.

To perform the tests, the Cross-Validation technique was used with the number of folds equal to 3 in a stratified way. Thus, observing the ROC curve of these 3 techniques compared, it is possible to notice that the neural network model had the most representative area, as can be seen in Figures 1 and 2. In the validation it was also possible to observe that the neural networks model was able to predict with a certain rate of success, the data that have not yet been presented to the model, obtaining 98% accuracy.



Plot

Target: PROTESTADO

Classifiers

- Neural Network
- SVM
- Logistic Regression

Curves

Merge Predictions from Folds

☑ Show convex ROC curves

☑ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

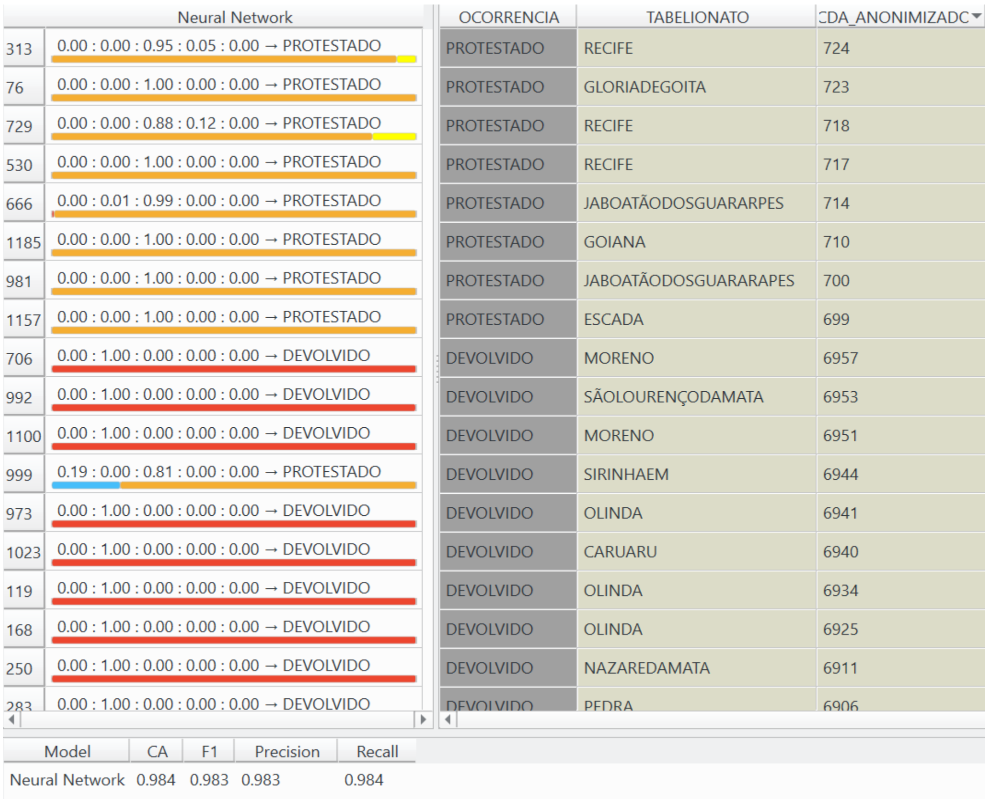
FP Cost: 500

FN Cost: 500

Prior probability: 95 %

3×5571

Figure 1. ROC curve resulting from the evaluation performed between NN, LR and SVM



	Neural Network	OCORRENCIA	TABELIONATO	CDA_ANONIMIZADC
313	0.00 : 0.00 : 0.95 : 0.05 : 0.00 → PROTESTADO	PROTESTADO	RECIFE	724
76	0.00 : 0.00 : 1.00 : 0.00 : 0.00 → PROTESTADO	PROTESTADO	GLORIADEGOITA	723
729	0.00 : 0.00 : 0.88 : 0.12 : 0.00 → PROTESTADO	PROTESTADO	RECIFE	718
530	0.00 : 0.00 : 1.00 : 0.00 : 0.00 → PROTESTADO	PROTESTADO	RECIFE	717
666	0.00 : 0.01 : 0.99 : 0.00 : 0.00 → PROTESTADO	PROTESTADO	JABOATÃODOSGUARARPES	714
1185	0.00 : 0.00 : 1.00 : 0.00 : 0.00 → PROTESTADO	PROTESTADO	GOIANA	710
981	0.00 : 0.00 : 1.00 : 0.00 : 0.00 → PROTESTADO	PROTESTADO	JABOATÃODOSGUARARAPES	700
1157	0.00 : 0.00 : 1.00 : 0.00 : 0.00 → PROTESTADO	PROTESTADO	ESCADA	699
706	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	MORENO	6957
992	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	SÃOLOURENÇODAMATA	6953
1100	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	MORENO	6951
999	0.19 : 0.00 : 0.81 : 0.00 : 0.00 → PROTESTADO	DEVOLVIDO	SIRINHAEM	6944
973	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	OLINDA	6941
1023	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	CARUARU	6940
119	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	OLINDA	6934
168	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	OLINDA	6925
250	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	NAZAREDAMATA	6911
283	0.00 : 1.00 : 0.00 : 0.00 : 0.00 → DEVOLVIDO	DEVOLVIDO	PEDRA	6906

Model	CA	F1	Precision	Recall
Neural Network	0.984	0.983	0.983	0.984

Figure 2. Result of the application of the NN model with the AC, F1, PR and RC metrics

6. CONCLUSION

This article sought to analyze the best Model of AI/ML that can be used to recommend CADs that should be protested or directed to another type of collection, optimizing the debt collection process, due to the gain of assertiveness, using a database of debts from the state of Pernambuco. With the results, it was possible to verify that the classification methods used achieved good accuracy results, being above 97%. Thus, it is possible to infer that the proposed model can be considered dependable since all search metrics achieved a satisfactory result.

The method used to select characteristics allowed better attributes to be used in the learning process and this, consequently, can produce more assertive results. Thus, the AGS may reduce or even eliminate the number of CADs that are returned by the notary offices, allowing them, which do not suit this form of collection, can be directed in advance to the most appropriate method of collection to their characteristics.

6.1. Future Works

As future work, training will be carried out with the completeness of the data, which are since 2006, the year of implementation of the justice system, since for this experiment was used a sample corresponding to the years 2018, 2019, and 2020.

With the use of the complete database, it will be possible to apply the model to the CADs that are indicated as not suitable for protest can be directed to other debt collection modalities. Being possible the from the return of failure, to indicate what would be the other modality of judicial collection, considering the rule for sending to electronic filing, which can be represented by logic: $Y = ((X1 \cap (X3 > 4000)) \cup (X2 \cap (X3 > 2000))$, where (Y) represents the range of life, (X1) represents ICMS, (X2) Other taxes, (X3) outstanding balance, allowing for electronic judgement only if the outstanding balance exceeds $4,000.00 in the case of TAX ON MOVEMENT or $2,000.00 in the case of other taxes, being able to send to BANKNOTE PROTEST SERVICE otherwise.

ACKNOWLEDGEMENTS

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001, and with the collaboration of the Attorney General of the State of Pernambuco, in particular the Coordination of Active Debt of the Attorney of Finance.

REFERENCES

[1] Witten, I. H., Frank, E., Hall, M. A., & Pal C. J. (2011) Mining: Practical tools and techniques of registering. Morgan kaufmann.

[2] Wirtz, B. W., & Müller, W.M. (2019) An integrated artificial intelligence structure for public management. Public Management Review.

[3] Gousios, G., & Spinellis, D. (2017) Github Mining Software Engineering Data at the 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C). IEEE.

[4] Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015) A hybrid data mining model of resource selection algorithms and set learning classifiers for credit score, Journal of Retailing and Consumer Services.

[5] van de Geer, R., Wang, Q., & Bhulai, S. (2018) Data-driven consumer debt collection via machine learning and approximate dynamic programming.

[6] Hunt, R. M. (2007) Collecting consumer debt in America.

[7] Rumelhart, D. E. & Geoffrey E. H., & Ronald J. W. (1986) Learning representations by back-propagating errors.

[8] Tolles, J., & William, J. (2016) Logistic Regression Relating Patient Characteristics to Outcomes. JAMA.

[9] Cortes, C., & Vapnik, V. N. Support-vector networks. Machine Learning. 20 (3): 273–297. CiteSeerX, 1995.

[10] Powers, D. M. (2011) Evaluation: from precision, recall and f-measure to roc, informed, striking.

[11] Richardson, R. J. (2017) Social research - Methods and Techniques. 4. Ed. São Paulo: Atlas.

[12] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000) Crisp-dm 1.0 step by step data mining guide.

[13] Demšar, J., Jet, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., & Štajdohar, M. (2013) The Journal of Machine Learning Research.

[14] Schumann, M. A (2015) Book about Colab: and related activities. New York, N.Y.: Printed Matter.

AUTHOR

Álvaro Farias Pinheiro is an Analyst in Information and Communication Technology Management at the State Agency for Information and Communication Technology of the State, Coordinator of Systems, Digital Automation and Innovation of the Attorney General of the State of Pernambuco, and Institutional Relations Director of the Pernambuco State Civil Servants Association. He is a member of the National Artificial Intelligence Technical Group. He is currently a Ph.D. Student in Computer Engineering at the Polytechnic School of Engineering of Pernambuco, University of Pernambuco. He holds an MBA in Artificial Intelligence applied to Marketing at Unyleya Faculty. He holds a Master's in Software Engineering from the Recife Center for Advanced Studies and Systems. He holds a Specialization in Software Engineering Development Methodologies from the Brazilian Union of Technology. He holds a Bachelor's degree in Information Systems with Emphasis in Software Engineering from Recife Integrated Faculty.

Denis Silva da Silveira is an associate professor in the Management Department at Federal University of Pernambuco. He received his Ph.D. degree in Production Engineering (2009) at Federal University of Rio de Janeiro, with postdoctoral research at New University of Lisbon (2016). He has experience in Information Systems and Business Process Management. His research interests include Business Processes Management, Information Systems Architecture, Conceptual Modelling, Semantic Models and User Models.

Fernando Buarque de Lima Neto is Ph.D. in Artificial Intelligence from the University of London (2002), with a degree from the Imperial College London - DIC, in Artificial Neural Networks (2002), a master's degree in Computer Science from the Federal University of Pernambuco (1998) and graduation in Computer Science from the Catholic University of Pernambuco (1990). He is an Associate Professor at UPE at the Polytechnic School of Pernambuco, and a permanent member of the Doctoral Program in Computer Engineering. The lines of research are (1) Artificial/Computational Intelligence, (2) Modeling/Simulation of Real Complex Problems, and (3) Decision Systems explainable via Computational Semiotics.