A Tool for Spatially Based Prediction of Consumer Lawsuits against Electric Power Companies

Domingos A. Dias Junior, Johnatan C. Souza, João O. B. Diniz, Geraldo Braz Junior, João D. S. Almeida, Anselmo Cardoso de Paiva
Applied Computing Group (NCA)
Federal University of Maranhão (UFMA)
São Luís Brazil
Email: {domingos.adj; johnatancarvalho; joao.bandeira; geraldo; jdallyson; paiva}@nca.ufma.br

Erika W. B. A. L. Alves
Equatorial Energy, Brazil
São Luís, Brazil
Email: erika.assis@equatorialenergia.com.br

Abstract—**The main purpose of an energy company is the provision of services to the final consumer. This does not mean that completely avoiding failures of the system is an easy task. These failures could lead to problems in the relationship between the energy company and the clients, resulting in judicial dispute. Thus, it is interesting for a power company to preemptively identify the consumers who are dissatisfied. Therefore, it is important for company executives to identify regions or groups of consumers that are related to the same cause that motivated a lawsuit. Thus, the present work aims to propose a tool for spatial analysis and spatially based prediction of consumer lawsuits against an electric power company using data mining and machine learning techniques. The results obtained from a database of an electric power company in Brazil showed results with 96.52% sensitivity in identifying consumers with lawsuits related to Unregistered Power Consumption (UPC). Also, a tool for visualization and spatial analysis of this group of clients is presented.**

Keywords - lawsuit; electricity sector; consumer profile; geoanalysis; geovisualization.

I. INTRODUCTION

Nowadays, many researchers have been studying problems in large volumes of data to analyze the consumer profile, using the concepts of both data mining and machine learning. We can define data mining as the process of exploring large amounts of data looking for consistent patterns. One way to find consistent patterns and associations within databases is to use machine learning techniques [2].

The energy market in several countries is changing, basically due to the deregulation of the market and the emergence of judicial and administrative mechanisms for consumer protection [3]. The main purpose of an energy company is to provide services to the final consumer, even though it is very difficult to completely eliminate failures of the system [4]. System failures could result in a judicial dispute between the customer and the energy company. For that reason, it is useful for a power company to preemptively identify the consumers who are dissatisfied and also to identify the reasons of this in order to generate action plans to avoid lawsuits and increase the quality of service. Thus, there is a need for a prediction service for lawsuits.

By analyzing large data in companies and providing prediction services based on machine learning techniques, companies have valuable data from their customers, so they can act effectively and accurately on certain issues. However, the prediction of isolated client's lawsuits often does not reflect the robustness of a lawsuit prediction service. When it comes to electric power companies, the service provided encompasses a great number of customers at the same time (neighborhoods, cities, states). For example, the Equatorial Energy Group is responsible for the distribution of electricity in the states of Para, Maranhão, Piauı, and Alagoas, with over 5 million customers in Brazil.

Often, a customer's dissatisfaction with the service delivered can be motivating for other neighboring customers to file a lawsuit against the company. A practical example could be when a customer who has had the power supply interrupted for 24 hours is appointed as a potential customer to initiate a lawsuit through a prediction system. Usually, this interruption is not individual, so neighboring customers may also have this same profile and thus could potentially seek justice.

A tool for spatial analysis and spatially based prediction of consumer lawsuits against an electric power company that enables the company and its executives to look at the prediction of a group of customers in different or even risky locations provides a much more robust mechanism for addressing the problem. Because executives have such information, they can address more accurately the concerns of customers (or neighboring customer groups) at greater risk of filing lawsuits, thereby reducing costs to the company and, above all, increasing customer satisfaction and positive relationships with the company.

For the reasons mentioned above, the present work propose a tool for spatial analysis and spatially based prediction of consumer lawsuits against an electric power company using data mining and machine learning techniques. Because it is a tool for prediction of lawsuits, the paper presents a series of benefits, which we can highlighted as follows:

- A way of customer group prediction visualization and analysis for electric power companies.

 ISBN: 978-1-61208-762-7

- The system is applied industrially and with real data.
- Assists the company in establishing the risk of receiving legal proceedings.
- Helps managers and executives understand the motivation of lawsuits to generate planning and prevention actions.
- Helps improve customer service and response based on a robust and intelligent method.
- Helps improve client's satisfaction by avoiding lawsuits.

The rest of the paper is organized as follows. Section II presents the main related works. Section III presents the proposed tool of prediction of consumer lawsuits. The results are given in Section IV. Finally, a conclusion on the results obtained is given in Section V.

II. RELATED WORK

Despite being techniques with huge social appeal and great concern among power companies, computational methodologies based on data mining and machine learning are still very scarce, especially for the purpose of providing a completely automatic method of predicting lawsuits. However, there is some work in the literature dealing with customer satisfaction, consumer profile analysis, and consumer churn in companies. These works will still be listed as related works because they were fundamental for the basis of the work proposed here, even if it is not possible to compare them directly with our proposed method.

Customers file a lawsuit because they want to be treated fairly by the company when a service failure occurs. Identifying customers who are likely to file a lawsuit against the company can also be considered as an identification of customer dissatisfaction. The company suffers a considerable monetary loss when some clients leave it, in addition to the procedural costs generated. This is classified as Customer Churn Prediction (CCP), where techniques based on machine learning, regression analysis, and predictive modeling are used to estimate the likelihood that customers will leave the company. CCP is a common problem in sectors such as telecommunications, [5], banking [6], e-commerce [7], and gaming [8].

A growing demand for new customers intensifies competition among commercial banks. To increase profits for ongoing operations and increase core competitiveness, commercial banks must avoid losing customers while at the same time acquiring new customers. In [6], the churn of commercial bank customers is predicted based on the Support Vector Machine (SVM) model and uses the random sampling method to improve the SVM model. When the ratio is 1:10 (churners:non-churners), the model has better results. Gordini and Veglio [7] developed a tailor-made churn prediction model for the SVM-based Business-to-Consumer (B2C) industry.

Amin et al. [9] propose a decision-making technique based on the Approximate Set Theory (AST) to extract important decision rules related to customer turnover. The AST classification based on genetic algorithms outperforms other rule generation algorithms used. Subsequently, the authors proposed a new CCP approach based on the concept of estimation of classifier certainty using distance as a factor [5]. It was found that the distance factor is strongly related to the certainty of the classifier.

Milosevic et al. [8] also presented a methodology for preventing customer churn, however, focusing on the game industry freemium. They evaluated learning models, such as decision trees, logistic regression, random forest, gradient boosting, and naive Bayes for churn prediction. The gradient boosting model showed better performance than the others. They also proposed a personalized churn prevention approach, identifying game features that are potentially interesting to the user and using them to customize notifications.

The relationship between previous events, which could be handled individually, with the quality and fidelity of the relationship is the purpose of Francisco et al. [10] proposed work involving a mobile phone company. The study confirmed (directly and indirectly) the assumptions of a positive and significant effect between satisfaction, trust and commitment and their antecedents.

According to Siu et al. [11], customers complain because they want to be treated fairly by the company when a service failure occurs. This study investigates the role of justice in retaining customers who had failed restaurant service experiences. As a result, the authors confirm the relationship of fairness between prior and subsequent satisfaction.

In a previous study, we proposed a method of predicting unregistered power consumption lawsuits and related variables [12]. The method proved to be robust in the task of classifying these types of customers, however, in our first study, we attempted to build an effective computational method and not a tool for corporate use. Also, the method in our previous work was not an extensible method, using only eXtreme Gradient Boosting. Thus, in this work, we propose the development of a tool for spatially based prediction of consumer lawsuits against electric power companies in a way that is intuitive and helps companies' managers increase customer satisfaction. This paper differs from our first work in several points, mainly in the fact that our present work is an extensible method which allows the insertion of new classifiers and offers a usable tool. Also, our previous work did not use the entire database, only a small proportion for the elaboration of the method. In the current work, we use the entire database of an electricity company.

Based on what is observed in the literature, and the importance of preventing customer lawsuits, the method proposed in this paper is based on the construction of a predictive model. For this, we use features extracted from the temporal relationship data of consumers of an electric company. The goal is to detect, in advance, cases where the customer may be dissatisfied and may go to court against the power company, and thus provide preventive information for managers and technicians to best address the problem.

We can see that many works propose ways to predict possible customer complaints, intention to leave the company or prosecute it. The works above mentioned do not offer a tool for individual or joint customer analysis. It is observed that the joint relationship of customer groups is something to be explored since features of customer groups in certain regions can be crucial in deciding lawsuits, especially in the case of customers of electric power services. The dissatisfaction of a customer can motivate his neighbors to file lawsuits against the company.

This paper presents a work done to provide electrical power companies and their executives with a tool for spatial analysis and spatially based prediction of consumer lawsuits. We propose a prediction method based on the customer behavior within the company; these predictions are plotted and visualized spatially. Thus, the company executives have a precise view of the problem and may take action to reduce customer dissatisfaction.

III. MATERIALS AND PROPOSED METHOD

The methodology is organized into the following five steps: data acquisition, feature extraction, training, prediction, analysis and spatial visualization. These steps are described in the next subsections.

A. Data acquisition

This work uses a private database from Equatorial Maranhão Energy Distributor. The database has customer information from various company sectors. Such information involves consumption history, supply discontinuity, financial information, customer complaints, equipment used for each one, etc.

The database used is accompanied by the ground truth for each customer, that is used in the prediction step to calculate validation metrics. Some customers already had a lawsuit against the company. Thus, it is possible to build a method to learn these patterns and later apply the pattern model to new customers.

After the data acquisition step, the data is organized into semantic groups to identify which features are relevant for the prediction of lawsuits. Also, we consider the generation of new features from these. This is done as good predictions require representative features for the classifier.

B. Fature extraction

In this step, we analyze what features are considered relevant to build the electric company customer profile. Also, we describe the techniques used to refine and create new representative features. The features that we used are:

- General Information: individual features information of each client like spatial location, neighborhood, type of client (residential or commercial).
- Power consumption: features about consumption profile of each client.
- Power Loss: features related to occurrences of loss of energy or failure of distribution.
- Invoices: features that show if there were invoices caused by having UPC with that customer.
- Financial: historical payment behavior of a particular customer.
- Law: previous legal actions taken by the customer.

All features are preprocessed before the recognition step. In addition, it is necessary to analyze customers with a temporal perspective, that is, from customer information up to a certain period and predict their possible behavior for the later period. Thus, in addition to making data processing necessary, it is relevant to simulate time intervals in the database. To do this, we modeled the techniques for handling features with Feature Tools (an open-source Python tool for automated feature engineering), as described below:

- Numeric features are normalized, i.e. their values are converted to a single scale, that is, a single range of values;
- Categorical features are converted to numeric, using one hot encoding [13], which creates binary variables for each category;
- Mutual features are created using descriptive statistics: mean, median, standard deviation. Moreover, such mechanisms are used because of a temporal perspective of the information, for example, the standard deviation of consumption of each customer is calculated taking into account their history in the last 18 months.

Thus, the data were processed to aggregate temporal information using statistics. New features were created from the initial information, enabling the analysis of customer behavior at various time intervals. Below, we present the procedure of training of classifiers of the methodology.

C. Training

The training step is responsible for choosing the best algorithm to recognize, based of previously computed features, if the client has intentions to file a lawsuit. We evaluate two algorithms: Extreme Gradient Boosting and Balanced Bagging Classifier. To evaluate, it is important to define metrics. In this paper, the chosen metric was sensitivity, since, in the case of lawsuits, the important thing is to be able to predict the largest number of possible lawsuits. The following are the two classifiers evaluated so far by the method:

- Extreme Gradient Boosting [14]: Tree-based machine learning algorithm that implements an optimization of the Gradient Boosting method, which is a technique that produces a predictive model from the union of less robust classifiers, showing better performance than if they were used in isolation.
- Balanced Bagging Classifier [15]: It consists in an unbalanced data handling technique, which selects by default 10 subsets of the initial data and performs random sampling.

The large number of individuals present in the database makes it impossible to train multiple models due to time and memory limitations. Thus, we divide the database randomly into 60% training and 40% for testing. From training, 10% of the data is taken to validate the model. Sampling is made keeping all individuals who filed a lawsuit and three times more individuals who did not file a lawsuit (1:3). The test dataset is used in the next step and remains as it is. All models are submitted to the same database and ranked according to the specified metric.

In the training step, the best classifier is chosen. As already mentioned, two classifiers were evaluated in this step. These are Extreme Gradient Boosting and Balanced Bagging Classifier. During training, each model generated by these classifiers is validated, and the classifier parameters are estimated using Random Search [16]. After a series of iterations, the best result achieved in validation is maintained as the best estimated classifier.

Once this is done, the classifier is submitted to the next step of the method, that is the prediction on a test dataset, as described below.

D. Prediction

After the training step, with the best classifier selected, the best model is applied to the rest of the database. Thus, for each customer of the company, it is possible to measure the percentage probability that the customer will initiate a lawsuit or not. A client who is going to initiate a UPC lawsuit is considered to be one that has more than 50% probability, according to the previously estimated model.

To validate the prediction results of the model as a whole, global metrics are calculated in the test dataset to verify the robustness of the generated model. For this, metrics of sensitivity, specificity, and accuracy are extracted.

Accuracy (Acc) represents the proportion of true results (true positives and true negatives) in the population.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensitivity (Sen) expresses the number of true positives (clients with UPC correctly classified) divided by the total number of positive cases.

$$Sen = \frac{TP}{TP + FN} \tag{2}$$

Specificity (Spec) represents the number of true negatives (customers without UPC correctly classified) divided by the total number of negative cases.

$$Spec = \frac{TN}{TN + FP} \tag{3}$$

After all predictions of all consumers are generated and stored, we propose a tool for analysis and visualization of these data. Based on the latitude and the longitude information of each client, it is possible to show them on the map, and, with the prediction values of each one, it is possible to make a spatial analysis of the lawsuits.

E. Spatial analysis and visualization

As a central point in the prediction of lawsuits, we have the identification of the clients with high probability to file a lawsuit and the determination of the more important causes that lead the predictor to this classification.

The Tobler's first law of geography states that "everything is related to everything else, but near things are more related than distant things". Based on this, we have that spatially grouped clients in general receive the same service quality and experience the same troubles in the consumer/provider relationship. In this work we used a geographic information system to manage and visualize several aspects of the lawsuits prediction, thus improving the company capacity to understand the problems and launch action plans to avoid the consumer dissatisfaction.

We theorize that the geovisualization of lawsuit predictions and their variables that strongly influence this prediction may offer an understanding of the dissatisfaction cause and the need actions the must be performed to increase the service quality. In this sense, we propose and develop a visualization tool that offers a visualization of prediction evaluation in different ways. This tool is shown in Figure 1.

Equatorial Maranhão's consumers are defined by contract accounts, and several users are linked to a traffic, so in the analysis, users of the tool can search for isolated consumers and/or traffic (Figure 1(A)). As an end-user visualization and analysis tool, it has a number of mechanisms that make it easy to search processes through filters (Figure 1(B)). Because it comprises an entire state of Brazil, there is a filtering option by Region (North, South, East, Northwest, Center); when selecting the region, the sectional and city areas are defined, which the user can also select through the filters; type of fare used; and the number of consumers that will be displayed (5000, 10000, or 50000). Because the prediction presents the probability by consumers, the user is free to define the percentage of probability they want to see on the map (Figure 1(C)). Furthermore, it is known that several features are extracted in the training step, and the prediction is evaluated by the features that influence the most the classifier decision, i.e., the variables that strongly influence this prediction. Therefore, the tool allows the visualization of the 10 most important features for each client and/or group of clients (Figure 1(D)). By clicking on the clusters (green, yellow and blue circles with number of consumers), this tab is redefined with the features of that customer group.

Another analysis that can be done is to select the client individually (blue marker). When this is done, a new field displays the contract accounts, the total probability, and the top ten features associated with the likelihood of triggering a lawsuit (Figure 2). Finally, the visualization and analysis is dispersed on the map defined by clusters or individual consumers, from the selection of filters. Through this series of mechanisms, managers are able to act accurately on the problem, trying to remedy problems with lawsuits and consequently reduce customer dissatisfaction.

In the next section, we will discuss the results of the proposed method steps and how the use of a geoanalysis and visualization tool can be crucial in basic service delivery companies such as Equatorial Maranhão.

IV. RESULTS AND DISCUSSION

This section presents and discusses the results obtained with the proposed method for the prediction of power consumption lawsuits. For evaluation, the acquired database was divided into two sets: training and test.

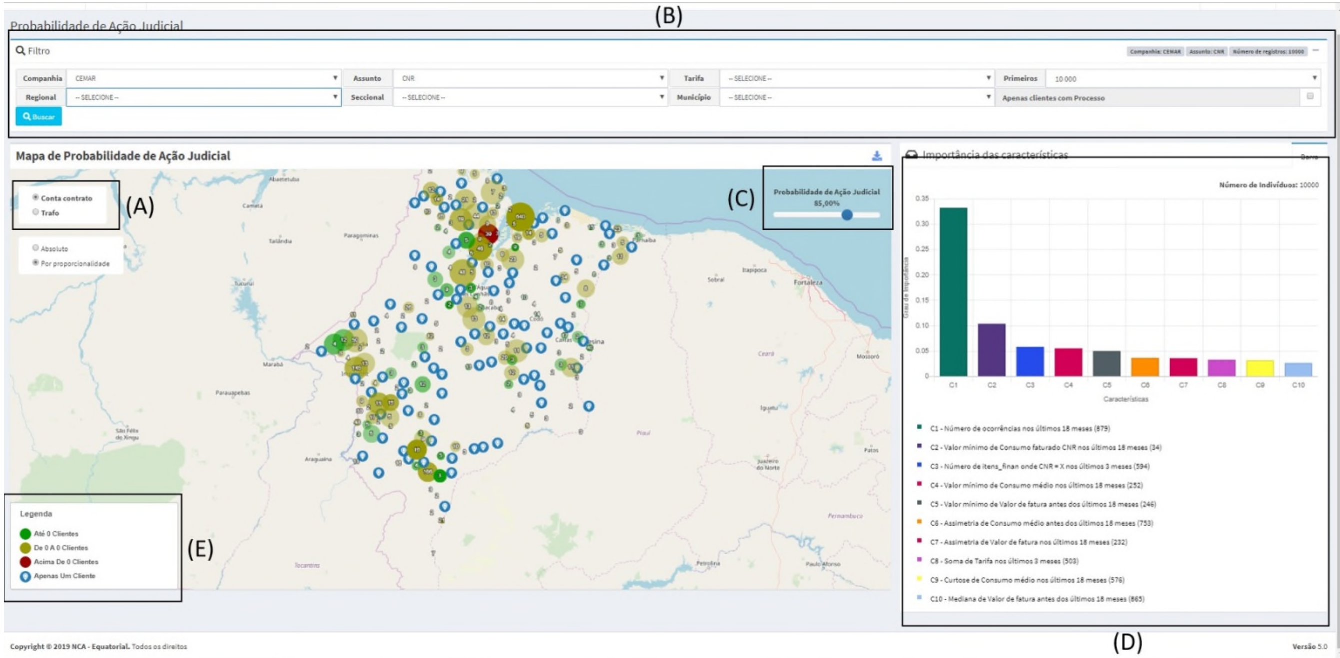


Figure 1. Spatial analysis and visualization tool for customer lawsuits prediction.

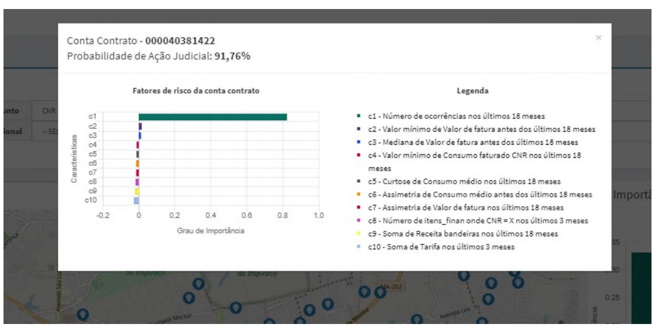


Figure 2. Individual analysis by contract account.

The client can go to court for a number of issues, which are defined during the court proceedings. In this paper the subject approached for prediction of lawsuit was Unregistered Power Consumption (UPC). This matter was chosen because it is the subject with the most lawsuits and that generates the most expenses for Equatorial Maranhão. The distribution of the proportion of clients with and without UPC is shown in Table I.

TABLE I. THE PROPORTION OF CLIENTS WITH UNREGISTERED CONSUMPTION SUBJECT IN TRAINING AND TEST DATASET.

Dataset	Consumer with UPC	Consumer without UPC	Total
Train	8.560	1.476.042	1.484.602
Test	5.714	998.442	1.004.156
Total	14.274	2.474.484	2.488.758

Approximately 2.5 million consumers were analyzed with the proposed method, however, a large imbalance in the datasets is noticeable. In the feature extraction step (Subsection III-B), 245 variables were retrieved directly from the database, while 925 were created from these by their temporal analysis, the transformation of categorical variables with one hot encoding and the use of descriptive statistics, totaling 1170 features.

Then, in the training step (Section III-C), the best classifier was decided when evaluating the 10% reserved for validation. Extreme Gradient Boosting yielded 90.3% sensitivity results in this step, while Balanced Bagging reached 93.6%. For this reason, Balanced Bagging was chosen to be the classifier that will make the prediction in the next step.

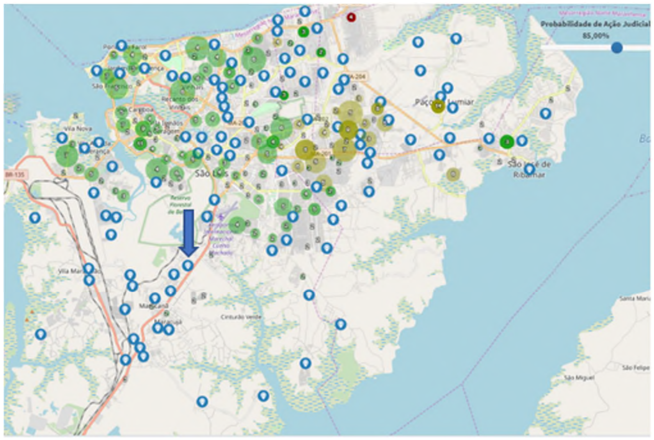
At the prediction step, the remainder of the database was evaluated by the model created in the training step, and thus generated the likelihood of a client to go to court against the company. To assess the robustness of the method, global validation metrics were generated, which are shown in Table II.

TABLE II. RESULT OF PREDICTION STEP USING BALANCED BAGGING.

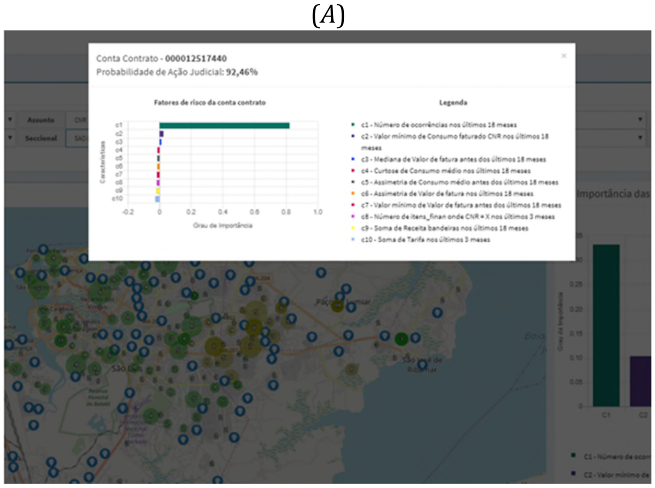
Subject	Acc (%)	Sen (%)	Spec (%)
UPC	91.86	96.52	91.83

As can be seen from Table II, the Balanced Bagging classifier proved robust results in the classification of new customers. Company managers can use this information, which has 96.52% sensitivity, in predicting UPC to avoid legal costs and improve quality.

However, despite this important information, there is a need to present the information more accurately, not just numerically. So, we use spatial analysis to visualize more effectively large or micro relationships between clients. As presented in Section III-E, a number of filtering mechanisms are presented for better map viewing. These mechanisms are crucial for geoanalysis, as executives can go directly to high-risk regions that are most likely to start legal action.



(A)



(B)

Figure 3. Case study: (A) Selecting an individual consumer, blue marker and (B) Display of variables that influence that consumer to file a lawsuit.

Based on an individual analysis, the manager can see which variables will influence that customer to start legal action against the company and act incisively. An example of this analysis in a case study can be seen in Figure 3 where, after filtering the interface, the user selects the individual consumer (blue marker, Figure 3A) and a new field with information of the most influential variables is presented (Figure 3B). By analyzing the variables, it is possible to obtain what influences the client the most, for example, the lack of energy in recent days. We can note that the customer has certain features that can influence him to file a lawsuit against the company, and that there is a 92.46% probability of going to court (Figure 3B). Also, the most probable cause is, for example, power outages. Then, the tool user can select the customer group (green, yellow and red circles), which that individual customer is contained in and the variables will be updated. So, managers will realize that not only a single customer, but the neighborhood is also going through the same problem and that the company's performance needs to reach that group of consumers to increase customer satisfaction with the company. Thus, the tool not only shows the percentage of dissatisfaction generated by the classifier, but also presents the possible variables that influenced consumers and groups of consumers to trigger actions, all combined with a friendly and intuitive geovisualization and analysis tool.

It is important to note that this is just a case study in which the tool can facilitate the use of company executives. Because it is a basic service delivery company and encompasses an entire state of the federation, the use of computational tools to help improve satisfaction is increasingly needed. A spatial analysis tool that aggregates lawsuit prediction information, as well as facilitating the company's operations, also proves to be an ally in the constant search for improvements in customer service.

V. CONCLUSION

This paper presents a spatial geoanalysis tool which predicts an electric power company customer's behavior. The prediction system is designed to allow its results to be used to avoid public disputes, e.g., in court trials. The work presents a prediction system based on computational intelligence techniques for prediction in a company from Maranhão state, Brazil. The method was applied to a database of more than two million customers and proved to be robust in hitting customers coming with UPC lawsuits against the company, resulting in a sensitivity of 96.52%.

Also, the geoanalysis tool presents the predictions of each client along with the variables that most influenced each of them. This tool has a user-friendly interface and several features for customer and customer group analysis, allowing company managers to act incisively on issues in a variety of areas that the company understands by reducing legal costs and increasing customer satisfaction. The tool also allows a complete visualization and analysis of the most diverse areas and the most diverse consumers of the state.

However, improvements are suggested as future work, such as acting on new judicial subjects (not only in UPC), using more classifiers in the training and prediction step, and generating other forms of data visualization such as heatmaps. Other spatial features can be added to search for improvements, analyze the various features of the database and verify the importance of spatial information for the robustness of the method, implement a tool in other basic service companies and, finally, create new features that can improve the classifier.

ACKNOWLEDGMENTS

The authors would like to thank Equatorial Energy for the financial support provided through the National Electric Energy Agency (ANEEL) Research and Development Program (R&D), PD-00037-0031/2019.

REFERENCES

[1] P.-N. Tan, Introduction to data mining. Pearson Education India, 2018.

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. MIT press, 2018.

[3] V. A. Ibanez, P. Hartmann, and P. Z. Calvo, “Antecedents of customer˜ loyalty in residential energy markets: Service quality, satisfaction, trust and switching costs,” The Service Industries Journal, vol. 26, no. 6, 2006, pp. 633–650.

[4] R. Johnston and A. Fern, “Service recovery strategies for single and double deviation scenarios,” Service Industries Journal, vol. 19, no. 2, 1999, pp. 69–82.

[5] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, “Customer churn prediction in telecommunication industry using data certainty,” Journal of Business Research, vol. 94, 2019, pp. 290 – 301.

[6] B. He, Y. Shi, Q. Wan, and X. Zhao, “Prediction of customer attrition of commercial banks based on svm model,” Procedia Computer Science, vol. 31, 2014, pp. 423 – 430, 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014.

[7] N. Gordini and V. Veglio, “Customers churn prediction and marketing retention strategies. an application of support vector machines based on the auc parameter-selection technique in b2b e-commerce industry,” Industrial Marketing Management, vol. 62, 2017, pp. 100–107.

[8] M. Milosevic, N. Zivic, and I. Andjelkovic, “Early churn prediction´ with personalized targeting in mobile social games,” Expert Systems with Applications, vol. 83, 2017, pp. 326 – 332.

[9] A. Amin et al., “Customer churn prediction in the telecommunication sector using a rough set approach,” Neurocomputing, vol. 237, 2017, pp. 242–254.

[10] E. C. Francisco-Maffezzolli, P. H. M. Prado, W. V. da Silva, and R. Z. Marchetti, “Evaluation of the customer relationship quality and propensity to change mobile telephone operators,” Brazilian Business Review, vol. 8, no. 4, 2011, pp. 1–22.

[11] N. Y.-M. Siu, T. J.-F. Zhang, and C.-Y. J. Yau, “The roles of justice and customer satisfaction in customer retention: A lesson from service recovery,” Journal of business ethics, vol. 114, no. 4, 2013, pp. 675–686.

[12] F. Y. Oliveira et al., “Prediction of unregistered power consumption lawsuits and its correlated factors based on customer data using extreme gradient boosting model,” in 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019, pp. 2059–2064.

[13] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, “Thermometer encoding: One hot way to resist adversarial examples,” in International Conference on Learning Representations, 2018.

[14] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining. ACM, 2016, pp. 785–794.

[15] G. Lemaıtre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” The Journal of Machine Learning Research, vol. 18, no. 1, 2017, pp. 559–563.

[16] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” Journal of machine learning research, vol. 13, no. Feb. 2012, pp. 281–305.