FACULTAD DE INGENIERÍA
REVISTA DE LA FACULTAD DE INGENIERIA
www.revistaingenieria.uda.cl

Identification and classification of *online* shopping products into categories for fraud prevention

Danilo Augusto Ganancin Faria[1], Erick Luciano Floriano Mendes[1], Fulvio Eduardo Ferreira[1], Gustavo Santos de Albuquerque[1], Gustavo Augusto de Sousa[2], Paola Fernanda S. Okuda[2], Pedro Henrique R. Abrahão[2], Rafael Belmiro Cristóvão[2], Teh Led Red[2], Daniel Camilo Fuentes Guzman[3], Milton Miranda Neto[3], Marcos Jardel Henriques[3], Oilson Alberto Gonzatto Junior[4], and Francisco Louzada Neto[4]

[1]Professional Master's in Applied Mathematics, Statistics, and Computing for Industry (MeCAI) at the Institute of Mathematical and Computer Sciences of the University of São Paulo (ICMC-USP) and the Research, Innovation, and Dissemination Center of the Center for Mathematical Sciences Applied to Industry (CEPID-CeMEAI).
[2]Undergraduate Course in Statistics at the Institute of Mathematical and Computer Sciences of the University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil.
[3]Interinstitutional Postgraduate Program in Statistics (PIPGEs) UFSCar-USP (Federal University of São Carlos (DES-UFSCar) and University of São Paulo (ICMC-USP)). São Carlos, São Paulo, Brazil. Email: jardel@usp.br
[4]Institute of Mathematical and Computer Sciences, University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil.

ABSTRACT

This paper aims to present the solution to a real problem (*e-commerce* data) involving the categorization of an unstructured database (text), as well as to demonstrate the process carried out to achieve satisfactory results that met the demands of client companies aiming to combat fraud in their online sales. To conduct this research, a database containing information about sales made by *e-commerce* companies over a one-week period was used. Descriptive analysis of the data was performed for subsequent use of statistical classification models. The techniques chosen for text and/or word categorization were: Multinomial Logistic Regression and Neural Networks. Using these techniques, a satisfactory result was achieved for the proposed problem, with an approach that allowed the identification of new categories of important variables. However, the work managed to improve data classification by approximately 96%, aligning with the company's objectives.

Keywords: Unstructured data. Logistic regression. Neural networks. E-commerce.

1 Introduction

The problem addressed originates from a company in the fraud detection industry. Among its range of clients are primarily e-commerce companies. Due to the growing popularity of commerce through electronic means, the volume of transactions has increased, and consequently, attempts at fraud have risen [3, 6, 20, 24, 26, 36, 37].

The company in question plays an important role in combating fraud related to online purchases. For the development of this research, a dataset was used focusing on online purchase transactions, based on the names (titles) of the products sold. Many frauds occur due to system failures. That is, a product with a swapped and/or similar name might be considered another product (cheaper) due to classification errors. Among the information contained in the dataset (which is, by the way, a text-based database and thus unstructured data) is the variable "Nome item", which is filled out independently by each company and contains a brief

description of the item involved in the transaction.

The origin of the problem lies in categorizing the product based on the information provided, and one factor that complicates classification is the nomenclature of the product name, which is done independently by partner companies. Each partner company has the freedom to describe the item in a way they deem most appropriate. However, for the anti-fraud company, it is important to identify which items with different descriptions might represent the same product and, consequently, belong to the same category. The variable "CategoriaCS", present in the database used in this research, indicates the product category to which each item belongs. This variable is often crucial when constructing predictive models for fraud risk. Therefore, the objective of this research is to appropriately infer the "CategoriaCS" from the "Nome item" variable.

In the literature, we find some authors who have solved text classification problems using *Support Vector Machines (SVM)* [1, 10, 12, 21, 21, 25, 30] and Naive Bayes classifier [7, 16, 17, 19, 23, 27, 35, 38, 39].

The aim of this work is to develop a solution that can satisfactorily classify the product and consequently assist in the subsequent stages of fraud detection. The paper continues with a description of the dataset and then the methodologies used during the project. Section **??** presents the results obtained. Section **??** concludes the work and then presents the bibliographic references.

2 The Methods

This section aims to describe all the statistical techniques used in this project. All analyses were performed using the programming languages **R** and *Python*.

2.1 Bag of Words

A "bag of words" is a text representation that describes the occurrence of words in a text document, involving a vocabulary of known words and a measure of the presence of those words. The method is called "bag of words" because any information about the order of words in the text document is discarded. The model aims only to understand if known words appear in the document, regardless of where the word is located in the text. In this approach, the histogram of words in the text is analyzed, considering each word as a feature [9].

A *Bag of Words* model can be quite simple or extremely complex, depending on the objective and the chosen approach. To implement this technique, one must first select the dataset, that is, the text document from which features are to be extracted. For the study in question, what was treated as a text document was the concatenation of the descriptions of all items, also considering the category in which the item was allocated. Once the dataset is defined, the desired vocabulary must be designed. For this step, it was decided to consider each word individually as an element belonging to the "bag of words," filtering out only *stopwords* (words like articles, prepositions, punctuation, and other connectors that, for the study in question, do not carry interpretation). The next step is to transform the sentences into vectors. Each element of the vector represents the number of times each of the words in the "bag of words" appears in the sentence. Each vector represents a sentence; in the study in question, a sentence is represented by the description of an item. At this stage, bigrams, trigrams, or other approaches could also be used, but the simplest to implement is considering each word as an element of the vocabulary.

With the vocabulary defined, the next step is to determine how to score the words, which now represent variables. A simple and easily interpretable approach is frequency or relative frequency, which is the number of times the words appear in the text. Since the goal of this work is to identify the words that most help discriminate the correct category for each item, the frequency with which a word appears in the description of a category seems to be a reasonable indicator of which category the item belongs to.

The "bag of words" approach is the first technique to be addressed in this study. Regardless of the algorithm or model used to discriminate product categories, it is of utmost importance to apply the "bag of words" technique.

2.2 Neural Networks

The development of artificial neural networks [4, 28], or simply Neural Networks, was initially motivated to reproduce the behavior of the human

brain, which processes information in a fundamentally different way from conventional computers. It can be said that the brain is a highly complex and nonlinear computer. Models based on neural network techniques aim to reproduce this form of human brain processing [29].

The main idea of neural networks is to construct a model composed of a large number of very simple processing units called neurons, with a large number of connections between them. The basic processing of information in the network occurs in the neurons. Information between neurons is transmitted through connections known as synapses [29].

Different topologies of neural networks are found in the literature [5, 15, 31–34]. In this work, multilayer neural networks were used. A multilayer neural network typically consists of aligned layers of neurons. In this type of network, inputs are presented in the first layer, known as the input layer. This layer distributes the input information to the hidden layer(s) of the network. The final layer is the output layer, where the problem's solution is obtained. The input layer and the output layer may be separated by one or more intermediate layers, called hidden layers. The neurons in a layer are connected only to the neurons in the immediately following layer, with no feedback or connections between neurons in the same layer. Additionally, the layers are typically fully connected.

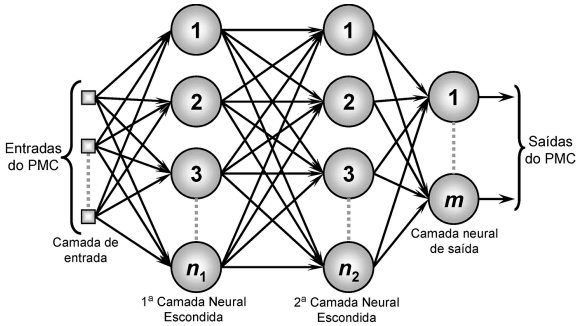


Figura 1: Example of a Multilayer Neural Network
Source: Moreira et al. [22]

A classic example of a multilayer neural network is illustrated in Figure 1. To implement an artificial neural network, the following parameters must be determined: the number of nodes in the input layer; the number of hidden layers and the number of neurons to be placed in these layers; and the number of neurons in the output layer. These parameters affect the performance of the artificial neural network model and should be carefully selected and tested.

The number of nodes in the input layer corresponds to the number of variables that will be used to feed the neural network. There is no general criterion for determining the number of neurons in the hidden layer(s). Generally, neural networks with fewer hidden neurons are preferred as they tend to have better generalization power, reducing the risk of overfitting.

In addition to the parameters mentioned above, it is necessary to define the activation function for the hidden layer neurons and the output layer, the training algorithm and its respective parameters, the data transformation or normalization method, the selection of the training and testing sets, the training stopping criterion, and the choice of a performance measure for the model.

The activation function is a mathematical function that, when applied to the linear combination of input variables and weights arriving at a specific neuron, returns its output value. There are various mathematical functions used as activation functions. The most commonly used activation functions are the logistic function and the hyperbolic tangent function.

When implementing a neural network, the dataset is usually divided into two sets: the training set and the test set. The training set (typically consisting of 70% of the data) represents the records from the database that were selected for generating the neural network model parameters. The test set (typically consisting of 30% of the data) is used to evaluate the model developed on the training set.

Training the neural network involves a nonlinear unconstrained minimization problem, where the network's synaptic weights are iteratively adjusted to minimize the mean squared error between the desired response from the input data and the output obtained at the output neuron. The most commonly used algorithm for this type of training is the backpropagation algorithm.

The application of the *backpropagation* algorithm requires the selection of a set of parameters (number of algorithm iterations, stopping criterion, initial weights, learning rate), whose influence can be decisive for the network's generalization capability.

The training stopping criterion must consider the network's generalization capability. Excessive training tends to lead to overfitting of the network.

The neural network technique applied to classification problems generally presents good results. Positive aspects of this technique include the lack of need for independent variables and its predictive power compared to other conventional techniques. However, it should also be noted that this technique often produces results that are difficult to interpret, as it is necessary to interpret a neuron in this network.

For the developed study, the artificial neural network technique was crucial for classifying items marked as "N.I."(Not Identified). Members of the group that conducted this study manually classified some items marked as "N.I.,"creating new categories. After this, an artificial neural network model was developed to classify these records marked as "N.I."into these new categories.

2.3 Logistic Regression

The logistic regression model [2, 8, 14] is a statistical model widely used in various studies, both in academia and the job market. In bivariate logistic regression, the response variable Y_i follows a Bernoulli probability distribution, assuming the value "1"with probability π_i and "0"with probability $1-\pi_i$. Thus, i varies according to the observations, resulting from the inverse logistic function applied to a vector x_i, which includes a constant and $k-1$ explanatory variables [18].

The Bernoulli distribution has the following probability function:

$$\mathbb{P}(Y_i|\pi_i) = \pi_i^{Y_i(1-\pi_i)^{(1-Y_i)}}.$$

The unknown parameters are present in the vector β, a $k \times 1$ vector with the first element representing a constant and the remaining elements representing the parameters corresponding to each explanatory variable of the model.

An alternative way to define the same model is to assume an unobserved continuous variable Y_i^* distributed according to the logistic density function with mean μ_i. In this case, μ_i varies according to the observation vector as a linear function of x_i. The model would be quite similar to a linear regression if Y_i^* were observable.

Thus, we would have $Y_i^* \sim \text{Logistic}(Y_i^*|i)$ with the following probability density function:

$$\mathbb{P}(Y_i^*) = \frac{\exp\{-(Y_i^* - \mu_i)\}}{(1+\exp\{-(Y_i^* - \mu_i)\})^2}$$

Which can be written as:

$$\begin{aligned}\mathbb{P}(Y_i^* = 1|\beta) &= \pi_i \\ &= \mathbb{P}(Y_i^* > 0|\beta) \\ &= \frac{1}{1-\exp\{x_i\beta\}}\end{aligned}$$

Logistic regression is a straightforward approach for solving classification problems, but it generally provides very satisfactory results, with a significant advantage in model interpretability. In the study in question, logistic regression was applied to predict product categories, and its results are presented in the following section.

3 Results

For the development of this research, data on transactions conducted via e-commerce between August 23 and 29, 2019, were used. The dataset contains 223,675 observations and 12 variables. Each observation in the dataset represents a transaction conducted in e-commerce across various companies.

The variables present in the dataset provide the following information:

- ID: transaction code (unique for each observation/transaction);
- Date: date and time of the transaction;
- Store: store code;
- Item Name: description of the item sold in the transaction (description provided by partner companies);
- Store Category: category of the item sold in the transaction (description provided by partner companies);
- Gender: gender of the consumer who made the transaction;
- ZIP Code: ZIP code of the consumer who made the transaction;

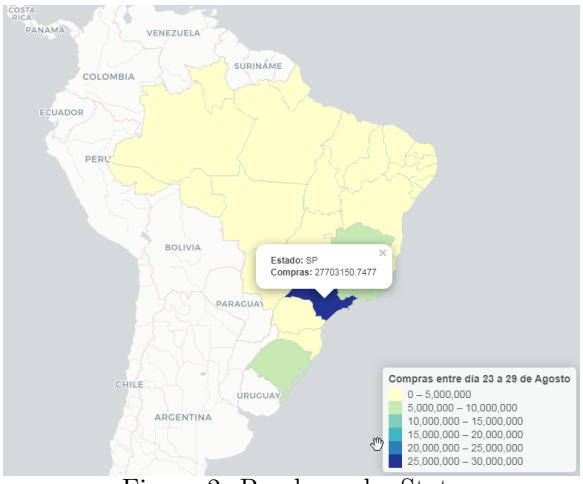


Figura 2: Purchases by State
Source: the authors

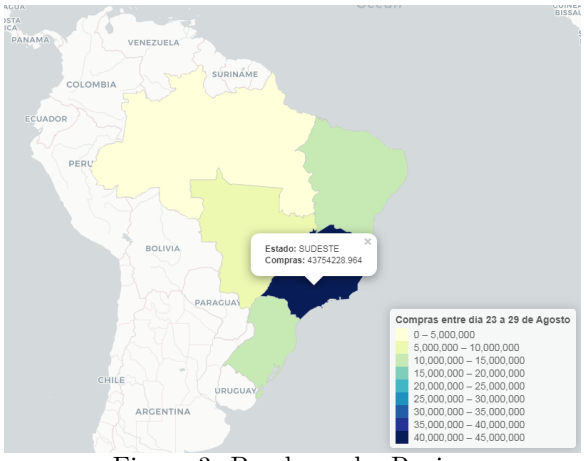


Figura 3: Purchases by Region
Source: the authors

- Neighborhood: neighborhood of the consumer who made the transaction;
- City: city of the consumer who made the transaction;
- State: state of the consumer who made the transaction;
- Value: value of the transaction;
- Category: category to which the item was allocated.

It can be observed that none of the variables in this database disclose any information about the company responsible for the transactions, let alone the consumers who made the purchases. Thus, there is no sensitive data, ensuring the privacy of both the end consumers and the companies responsible for the transactions.

3.1 Exploratory Analysis

Several analyses were conducted to better understand the distribution of the data and possible relationships between the variables in the dataset. The main results are presented in graphical forms and will be shown below.

The charts in Figures 2 and 3 show the distribution of purchase values by states and macroregions. Figure 2 highlights the importance of the state of São Paulo for e-commerce sales. São Paulo had sales values more than double those of Minas Gerais, the second state with the highest sales value during the period. São Paulo, Minas Gerais, and Rio de Janeiro are the three states with the highest sales values in the observed period. The results in Figure 3 are consistent: the Southeast Region shows sales values more than 4 times higher than any other Brazilian macroregion. The Northeast and South macroregions have similar values, while the North macroregion has the lowest sales value in e-commerce for the observed period.

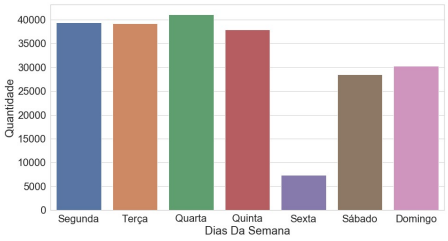


Figura 4: Purchases by Day of the Week
Source: the authors

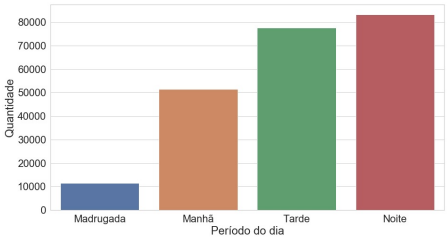


Figura 5: Purchases by Periodo
Source: the authors

The graphs in Figures 4 and 5 show, respectively, the number of purchases distributed across the days of the week and the periods during which

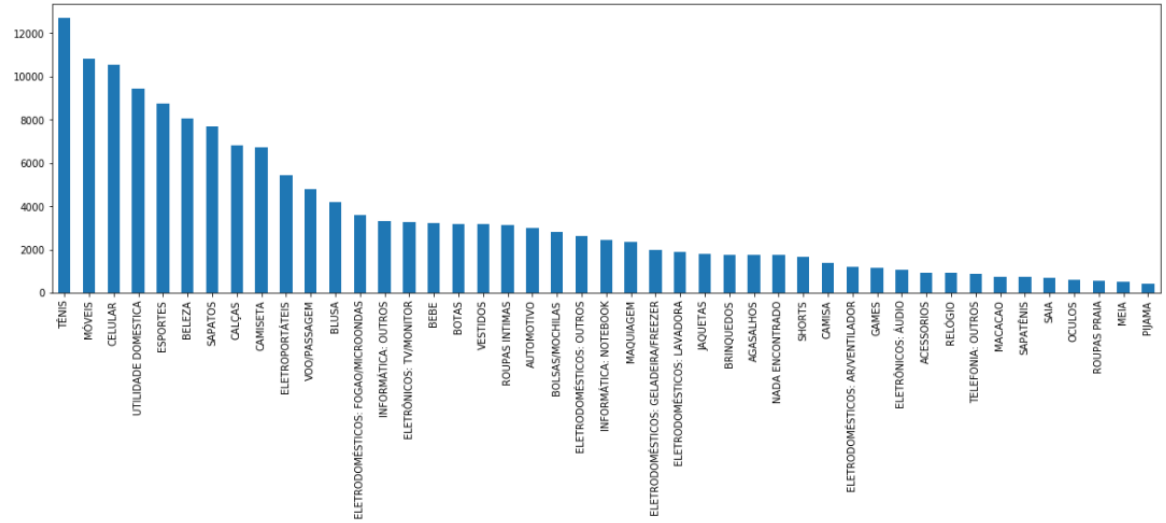


Figura 6: First word used in item descriptions

Source: the authors

the transactions took place. In Figure 4, it can be seen that on Friday, the volume of transactions is significantly lower compared to other days of the week. Additionally, Saturdays and Sundays appear to have slightly lower volumes compared to Monday, Tuesday, Wednesday, and Thursday. Figure 5 indicates the preferred shopping periods for consumers. It can be identified that consumers have a higher preference for making purchases on e-commerce platforms in the afternoon and evening periods.

The analyses presented in Figures 6 and 7 were conducted based on the text analysis of the item description variable. Figure 6 considers only the first word used in the description. It can be observed that the first word in some cases provides strong indications of the category to which the product or item belongs (for example: “Sneakers”, “Pants”, “Sandals”, among others). However, it should be noted that in some cases, understanding the correct category of the product may not be sufficient (for example, in cases like “Smart”, “Purchase”, “Mini”, among others).

The graph presented in Figure 7 represents the most frequent words in the item descriptions, regardless of their order. What is observed here is essentially the same pattern as in Figure 6, with some words strongly indicating the category to which the item belongs (e.g., “Apparel,” “Small Appliances,” “Household Appliances”), and others that may be used for items in different categories (e.g., “Casual,” “Men,” “Women”).

Another important evaluation was the distribution of the *Category* variable, which is intended to be predicted using some statistical or computational method. This variable contains 54 different types of categories.

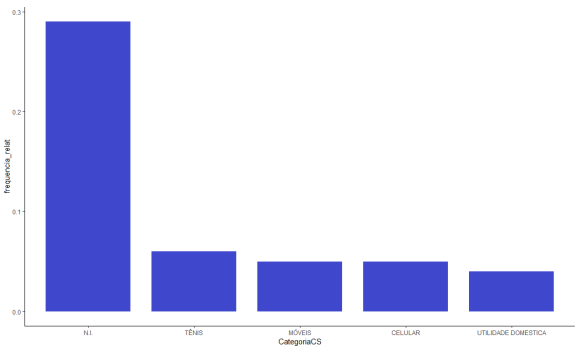


Figura 8: Distribution of the Top 5 Most Frequent Categories

Source: the authors

The chart in Figure 8 shows the distribution of the 5 most frequent categories, indicating a high concentration of records in a single category. The category with the highest frequency in the dataset is the "N.I."category. This indicates records that do not have an identified classification. This category alone represents 29% of the records in the dataset and tends to include records that do not fit into any of the other 53 pre-existing categories.

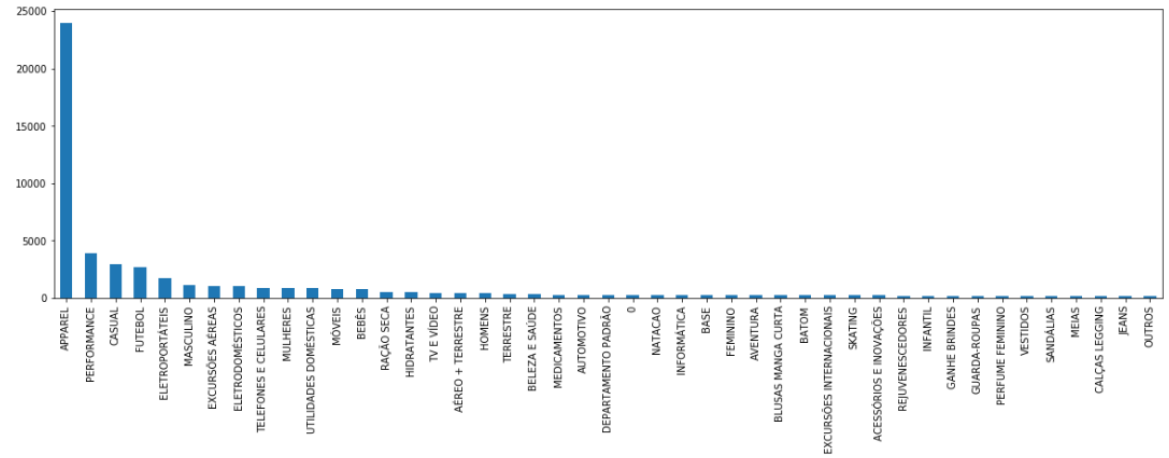


Figura 7: Most Frequent Words in Item Descriptions

Source: the authors

3.2 Overlook into the data

The first step involved applying a logistic classifier with the goal of predicting the *CategoriaCS* variable. For this purpose, the dataset was divided into training (75%) and testing (25%) sets. Even at this stage, it was possible to achieve an excellent result, as shown below: Distribution of the Top 5 Most Frequent Categories, as shown in Figure 9.

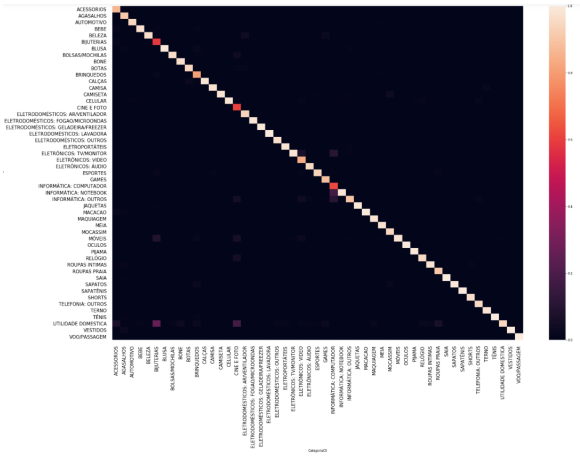


Figura 9: Confusion Matrix: Predicted Category vs Actual Category

Source: the authors

From the confusion matrix, it is possible to note that the classifier's accuracy within most of the known categories is good, achieving an accuracy of 0.968.

3.3 Predicting - Elements from the N.I. category

However, when applying the model results to the N.I. (Not Identified) category, it was observed that there were still items whose categories were not straightforward to predict because their descriptions did not exist within the initial *CategoryCS*. These items were predicted; however, with very low probability. Furthermore, the logistic classifier tended to classify them as *Household Utility*, the most diverse category among the existing ones, as shown in Figure 10.

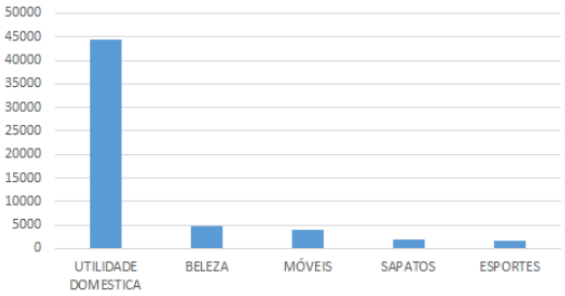


Figura 10: Frequency of the Most Common Categories for N.I. Records

Source: the authors

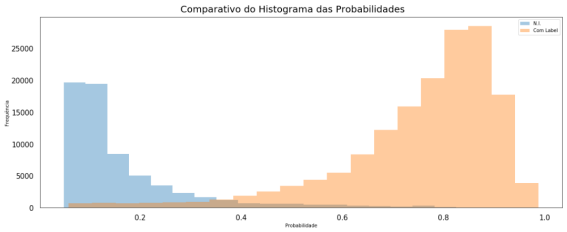


Figura 11: Comparison of Probability Histograms
Source: the authors

From these results (Figure 10), and furthermore, as can be seen in Figure 11, it was concluded that the creation of new product categories was necessary, which were developed in Step 3.

3.4 Clustering - New product categories

Through exploratory work and frequency analysis within items with a probability lower than 0.25, it was possible to identify potential new categories, as shown in Table 1.

Tabela 1: Possible new categories based on the first word of the item description

First Word	Quantity	Possible Category
Ração	3295	Pet
Compra	1193	Recarga
Kit	1149	-
Gift	938	GiftCard
Conjunto	816	-
Jogo	734	-
Livro	459	Livro
Antipulgas	302	Pet
Vinho	251	Bebidas
Cerveja	243	Bebidas
Remédios	-	Remédios

Source: the authors

With the definition of these new categories to be classified within the N.I.s, they were manually imputed into a portion of the dataset (500 cases) so that by applying a neural network model, it would be possible to classify all those without *labels*. The neural network model also performed well and achieved an accuracy of 0.877. In Figure 12, we can understand the illustration of the Confusion Matrix: Neural Network Predicted Category vs Manually Created Category.

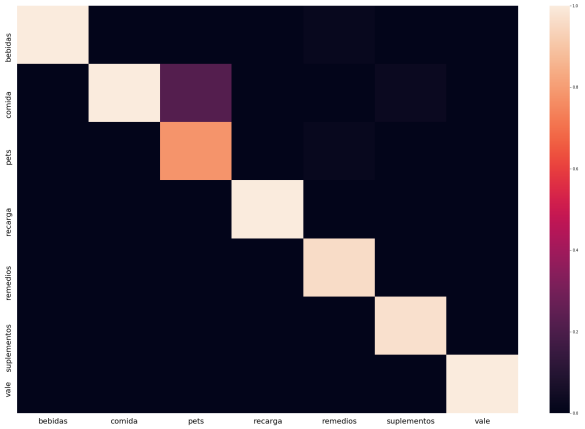


Figura 12: Confusion Matrix: Neural Network Predicted Category vs Manually Created Category
Source: the authors

From these new results, with the N.I. records marked with the new categories, it was possible to develop a Multinomial Logistic Regression model. The expectation was for improved classification, with a lower rate of items having probabilities below 0.4. The results are presented in Step 4.

3.5 Wrapping the results

Upon running the multinomial logistic model [11, 13] again (with the new categories), a noticeable improvement was observed when comparing the probability histograms, as shown in Figure 13. This improvement is even more evident when analyzing the overlaid probability density functions on the same graph, as illustrated in Figure 13(c), as well as in Figure 14.

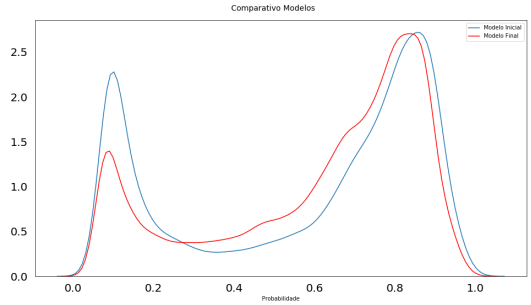
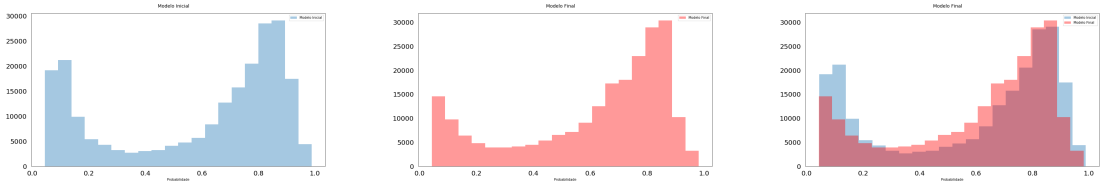


Figura 14: Probability Density Plot Before and After Adding New Categories (from N.I. Records)
Source: the authors

It is clear that there was a significant reduction in the lower probability ranges. Previously, 69% of



((a)) Initial model ((b)) Final model ((c)) Model Overlay

Figura 13: Comparison between the initial and final models

Source: the authors

the records had probabilities greater than 0.4, and later, with the addition of new categories, 77.5% of the records had probabilities greater than 0.4. This indicates that the process of defining new categories was important for making the responses more reliable.

Thus, considering that records with probabilities less than 0.4 would be labeled as N.I., after the entire modeling process, it would be possible to reduce the volume of N.I. records in the database from 29% to 22.5%, while maintaining an accuracy of 0.96 for the records that are already classified.

4 Conclusion

From the new classification model developed, it was possible to create some new product categories. Additionally, the results indicate that the model has a high capacity to correctly classify the categories of each item (accuracy of 0.96), which means that the objective of this research has been achieved. Considering a probability threshold of 40%, there is a 22% reduction in the volume of N.I. records.

Although the results are promising and the volume of N.I. records has decreased with this new approach, there is still the possibility of creating new categories manually and developing a new model that could present better results. However, it is worth noting that this requires significant manual effort.

Thus, a new method for classifying e-commerce order categories has been developed, which can bring benefits in various ways, such as using this variable more effectively in the construction of predictive credit and fraud models, as well as enabling tracking of new categories, which has an immediate impact on the business. Moreover, and perhaps more importantly, by correctly classifying 96% of the products, it is possible to avoid making incorrect sales due to a mere "misidentification"of a product.

Referências

[1] Chen, P.-H., Lin, C.-J., and Schölkopf, B. A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry 21*, 2 (2005), 111–136.

[2] David, K., and Mitchel, K. Logistic regression: A self learning text, 1994.

[3] DE LIMA, L. B. *DETECÇÃO DE ANOMALIAS EM TEMPO DE RESPOSTA DE SERVIDORES WEB: UMA ABORDAGEM AUTOMATIZADA PARA APRIMORAR A SEGURANÇA E A EFICIÊNCIA*. PhD thesis, Dissertação de Engenharia Elétrica e de Computação 2023. Tese de Doutorado ..., 2023.

[4] Farley, B., and Clark, W. Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory 4*, 4 (1954), 76–84.

[5] Fiesler, E., and Beale, R. Neural network topologies. *The Handbook of Neural Computation, E. Fiesler and R. Beale (Editors-in-Chief), Oxford University Press and IOP Publishing* (1996).

[6] Fiori, D. A. *Comércio eletrônico e e-business:: conceitos para entender a transformação digital*. Editora Intersaberes, 2023.

[7] Frank, E., Trigg, L., Holmes, G., and Witten, I. H. Naive bayes for regression. *Machine Learning 41* (2000), 5–25.

[8] Garson, G. Logistic regression: Statnotes, from north carolina state university. *Retrieved from http:/www2. chass. NCU. edu/garson* (2008).

[9] Goldberg, Y. *Neural network methods for natural language processing*. Synthesis Lectures on Human Language Technologies, 2017.

[10] Guenther, N., and Schonlau, M. Support vector machines. *The Stata Journal 16*, 4 (2016), 917–937.

[11] Hamilton, L. C., and Seyfrit, C. L. Interpreting multinomial logistic regression. *Stata Technical Bulletin* (1993).

[12] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their applications 13*, 4 (1998), 18–28.

[13] Hendrickx, J., et al. Special restrictions in multinomial logistic regression. *Stata Technical Bulletin 56* (2000), 18–26.

[14] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. *Applied logistic regression*. John Wiley & Sons, 2013.

[15] Ibnu, C. R. M., Santoso, J., and Surendro, K. Determining the neural network topology: A review. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications* (2019), pp. 357–362.

[16] Jiang, L., Wang, D., Cai, Z., and Yan, X. Survey of improving naive bayes for classification. In *Advanced Data Mining and Applications: Third International Conference, ADMA 2007 Harbin, China, August 6-8, 2007. Proceedings 3* (2007), Springer, pp. 134–145.

[17] Jiang, L., Zhang, H., and Cai, Z. A novel bayes model: Hidden naive bayes. *IEEE Transactions on knowledge and data engineering 21*, 10 (2008), 1361–1371.

[18] King, G., and Zeng, L. Logistic regression in rare events data. *Political analysis 9*, 2 (2001), 137–163.

[19] Lowd, D., and Domingos, P. Naive bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning* (2005), pp. 529–536.

[20] Massa, D., and Valverde, R. A fraud detection system based on anomaly intrusion detection systems for e-commerce applications. *Computer and Information Science 7*, 2 (2014), 117–140.

[21] Meyer, D., and Wien, F. Support vector machines. *R News 1*, 3 (2001), 23–26.

[22] Moreira, S. Rede neural perceptron multicamadas (diagrama do perceptron multicamadas), 2018. Published in Ensina.AI, Dec 24.

[23] Murphy, K. P., et al. Naive bayes classifiers. *University of British Columbia 18*, 60 (2006), 1–8.

[24] Niranjanamurthy, M., and Chahar, D. The study of e-commerce security issues and solutions. *International Journal of Advanced Research in Computer and Communication Engineering 2*, 7 (2013), 2885–2895.

[25] Noble, W. S. What is a support vector machine? *Nature biotechnology 24*, 12 (2006), 1565–1567.

[26] Ray, S. Fraud detection in e-commerce using machine learning. *BOHR International Journal of Advances in Management Research 1*, 1 (2022).

[27] Rish, I., et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (2001), vol. 3, Seattle, WA, USA;, pp. 41–46.

[28] Rochester, N., Holland, J., Haibt, L., and Duda, W. Tests on a cell assembly theory of the action of the brain, using a large digital computer. *IRE Transactions on information Theory 2*, 3 (1956), 80–93.

[29] S., H. *Neural Networks: A comparative Foundation*. New Jersey: Prentice Hall;, 1999.

[30] Shmilovici, A. Support vector machines. *Data mining and knowledge discovery handbook* (2010), 231–247.

[31] Stanley, K. O., and Miikkulainen, R. Efficient evolution of neural network topologies. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)* (2002), vol. 2, IEEE, pp. 1757–1762.

[32] Stanley, K. O., and Miikkulainen, R. Efficient reinforcement learning through evolving neural network topologies. In *Proceedings of the 4th Annual Conference on genetic and evolutionary computation* (2002), pp. 569–577.

[33] Stanley, K. O., and Miikkulainen, R. Evolving neural networks through augmenting topologies. *Evolutionary computation 10*, 2 (2002), 99–127.

[34] Stier, J., Gianini, G., Granitzer, M., and Ziegler, K. Analysing neural network topologies: a game theoretic approach. *Procedia Computer Science 126* (2018), 234–243.

[35] Webb, G. I., Keogh, E., and Miikkulainen, R. Naïve bayes. *Encyclopedia of machine learning 15*, 1 (2010), 713–714.

[36] Weng, H., Ji, S., Duan, F., Li, Z., Chen, J., He, Q., and Wang, T. Cats: cross-platform e-commerce fraud detection. In *2019 ieee 35th international conference on data engineering (icde)* (2019), IEEE, pp. 1874–1885.

[37] Weng, H., Li, Z., Ji, S., Chu, C., Lu, H., Du, T., and He, Q. Online e-commerce fraud: a large-scale detection and analysis. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)* (2018), IEEE, pp. 1435–1440.

[38] Yang, F.-J. An implementation of naive bayes classifier. In *2018 International conference on computational science and computational intelligence (CSCI)* (2018), IEEE, pp. 301–306.

[39] Zhang, H. The optimality of naive bayes. *Aa 1*, 2 (2004), 3.