Research, Society and Development, v. 9, n. 11, e86691110491, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i11.10491

Optimization of operational costs of Call centers employing classification techniques

A otimização dos custos operacionais do Call center empregando técnicas de classificação

La optimización de los costes operativos del Call Center empleando técnicas de clasificación

Received: 11/22/2020 | Reviewed: 11/23/2020 | Accept: 12/02/2020 | Published: 12/07/2020

Amanda Ferreira de Moura
ORCID: https://orcid.org/0000-0002-7875-2259
Universidade Nove de Julho, Brazil
E-mail: amandasff@yahoo.com.br
Cíntia Maria de Araújo Pinho
ORCID: https://orcid.org/0000-0003-0525-5072
Universidade Nove de Julho, Brazil
E-mail: cintia.pinho01@gmail.com
Domingos Márcio Rodrigues Napolitano
ORCID: https://orcid.org/0000-0001-5840-6757
Universidade Nove de Julho, Brazil
E-mail: d.napolitano@uni9.pro.br
Fellipe Silva Martins
ORCID: https://orcid.org/0000-0003-3918-7231
Universidade Nove de Julho, Brazil
E-mail: fellipemartins@uni9.pro.br
João Carlos Franco de Barros Fornari Junior
ORCID: https://orcid.org/0000-0003-4564-9818
Universidade Nove de Julho, Brazil
E-mail: joaocarlosfornarijr@gmail.com

Abstract

The provision of credit to customers of banking chains through call center services has always been one of the resources that generate significant income for financial institutions, however, the service offers a cost, which is often above desirable to guarantee profitable contracting to

Bank. Based on this, this work aims to evaluate the optimization of operational costs of call center, using classification techniques, through experimentation of supervised machine learning techniques to perform the classification task, in order to generate a predictive model, which offers a better performance in the operation of offering bank credit, to carry out an effective and productive action, conceiving greater savings for the company in identifying the public with greater adherence. For this, a database comprising 11,162 call records made from a bank offering its customers a letter of credit was employed. The results showed value correlations between variables, such as duration of the call, marital status, education level and even recurrence in adhering to subscribers' credit agreements. Through the application of the PCA to reduce dimensionality and classification models, such as AdaBoost, Gradient Boosting, SVM RBF, Naive Bayes, Random Forest, it was possible to perceive the consumer profile with good acquiescence for the investment proposal and a group of people with a high probability of not adhering to the letter of credit, so it was possible to outline an action directed to the public predisposed to the offer, minimizing expenses reaching greater profitability.

Keywords: Call center; Machine learning; Supervised models; Classification and predictive model.

Resumo

A oferta de crédito aos clientes de redes bancárias, através de serviços de *call center* sempre foi um dos recursos que geram uma renda significativa as instituições financeiras, porém, o serviço oferece um custo, que muitas vezes está acima do desejável para garantir contratações rentáveis ao banco. Baseado nisso este trabalho tem por **objetivo avaliar a otimização de *custos operacionais de call center, empregando técnicas de classificação***, através de experimentação de técnicas de aprendizado de máquina supervisionado para realizar a tarefa de classificação, a fim de gerar um modelo preditivo, que ofereça um melhor desempenho na operação de oferta de crédito bancário, para a realização de uma ação eficaz e produtiva, concebendo maior economia a empresa na identificação do público com maior aderência. Para isso foi utilizada uma base de dados com 11.162 registros de ligações feitas de um banco, oferecendo aos seus clientes uma carta de crédito. Os resultados apresentaram correlações de valor entre as variáveis, como tempo de duração da ligação, estado civil, nível escolaridade e até recorrência na adesão à contratos de crédito dos assinantes. Com a aplicação da PCA para redução da dimensionalidade e dos modelos de classificação, como: *AdaBoost, Gradient Boosting, SVM RBF, Naive Bayes, Random Forest*, foi possível perceber o perfil do consumidor com boa aquiescência para proposta de investimento e um grupo de pessoas com probabilidade alta de não adesão à carta de crédito, assim pôde-se delinear uma ação direcionada ao público predisposto à oferta, minimizando gastos atingindo maior lucratividade.

Palavras-chave: *Call center*; Aprendizado de máquina; Modelos supervisionados; Classificação e modelo preditivo.

Resumen

La provisión de crédito a los clientes de las redes bancarias a través de los servicios de call center siempre ha sido uno de los recursos que generan importantes ingresos para las instituciones financieras, sin embargo, el servicio ofrece un costo, muchas veces por encima de lo deseable para garantizar una contratación rentable con Banco. En base a esto, este trabajo tiene como objetivo evaluar la optimización de los costos operacionales del centro de llamadas, utilizando técnicas de clasificación, mediante la experimentación de técnicas de aprendizaje automático supervisado para realizar la tarea de clasificación, con el fin de generar un modelo predictivo, que ofrece una mejor desempeño en la operación de oferta de crédito bancario, para llevar a cabo una acción eficaz y productiva, concibiendo mayores ahorros para la empresa en la identificación de los públicos con mayor adhesión. Para ello, se seleccionó una base de datos con 11.162 registros de llamadas realizadas desde un banco, ofreciendo a sus clientes una carta de crédito. Los resultados arrojaron correlaciones de valor entre variables, como duración de la llamada, estado civil, nivel educativo e incluso recurrencia en la adhesión a contratos de crédito de suscriptores, con la aplicación del PCA para reducir modelos de dimensionalidad y clasificación, como : AdaBoost, Gradient Boosting, SVM RBF, Naive Bayes, Random Forest, se pudo percibir el perfil del consumidor con buena aquiescencia para la propuesta de inversión y un grupo de personas con alta probabilidad de no adherencia a la carta de crédito, por lo que se pudo perfilar una acción dirigida al público predispuesto a la oferta, minimizando los gastos alcanzando una mayor rentabilidad.

Palabras clave: Call center; Machine learning; Modelos supervisados; Clasificación y modelo predictivo.

1. Introduction

Call centers (or telephone exchanges) are an inseparable element of most large-scale businesses today (Ibrahim *et al*., 2016) The use of call centers has become unquestionable in recent decades (Mwendwa, 2017) and its adoption in several industries is due to reasons such as cost cutting, increased marketing capillarization, greater access to consumer markets and potentially improving the efficiency of companies. On the other hand, the indiscriminate use of call centers may raise problems that did not exist previously - especially in the management of portfolios and the effectiveness of their sales techniques (Song, Du and Zhang, 2018). Thus, the advantages of implementing and maintaining call centers in operations can be neutralized by their poor management (Bateh and Farah, 2017). This balance between efficiency and quality is difficult to obtain (Brown *et al*., 2005; Ibrahim *et al*., 2016) and consequently, this study aims at investigating the operational costs of call centers, using supervised machine learning techniques to perform classifications. Thus, consumer profiles are considered and split into groups of people predisposed to accept credit offers and customers who have a high probability of not accepting such deals, thus enabling decision-makers to target intelligently the financial assets for each marketing action.

Similar problems to risk minimization exist (queuing modeling, agent modeling, time modeling, systems modeling, etc.) and hard modeling dissociated from the purpose of minimizing losses in the business are commonly found. In Moro, Laureano and Cortez’s "Using data mining for bank direct marketing: An application of the crisp-dm methodology" (2011), a database was used to carry out a study on the success in offering bank letters of credit via telephone, based on the CRISP-DM (methodology Cross Industry Standard Process for Data Mining) that employs data collected in the bank, in order to improve your techniques for more assertive decision making. There is also the question of information uncertainty and asymmetry (Clark *et al*., 2019). Uncertainty because control over the outcome of the process is minimal and asymmetry because the set of data on each customer / sale is very small. Traditional statistical techniques may not be successful because of these intrinsic problems of such a database.

Loss modeling, from a business point of view, however, still need more research. In this sense, tests will be carried out by subjecting the database to a treatment of its variables, identifying the most relevant ones, analyzing their correlations, and applying models from the Confusion Matrix, the application of PCA to reduce dimensionality, to classification models, such as: AdaBoost, Gradient Boosting, SVM RBF, Naive Bayes and Random Forest in order to obtain the best result for potential consumers and possible non-contractors, aiming at reducing costs in the applied actions. The results demonstrate that with the application of the classification techniques and models and seeing the results with better performance, it can be understood that with the help of a predictive models in the definition of profiles more susceptible to supply, it will be possible to optimize the redistribution in directing costs aimed at approaching the customer, which indicates a new way of working, with marketing directed to this group of investors, less costly action and more assertive approaches.

2. Theoretical Reference

In order to understand the context to which the research is inserted, a bibliographic research was carried out on the use of machine learning in the marketing area, as well as predictive models based on classification techniques used in the experiments of this work. The use of call centers in marketing activities proved to be an efficient way of approaching the public in several areas, however, with the increase in marketing campaigns in recent years, it reduced the number of people adhering to proposals of this nature. Studies were carried out to understand the dynamics and the profile of the public addressed by these actions. Moro, Laureano and Cortez (2011) carried out a study to understand and explain the success of a telephone contact, for this, they used a data mining project based on the CRISP-DM methodology (Cross Industry Standard Process for Data Mining), which implies the crossing of information through data mining for a better understanding of the business, in order to improve its techniques for more assertive decision making.

The application of techniques based on machine learning to improve actions in different sectors has been adding value and minimizing expenses in several areas of business performance. Moro, Cortez and Rita (2014) dedicated a study comparing four data mining models: logistic regression, decision tree, neural network and support vector machine, in order to predict the success of telemarketing calls for deposit sales.

2.1 Machine Learning and Methods Employed

Machine learning itself is an area that has been increasingly explored, this is a part of Artificial Intelligence (AI) that aims at the development of computational techniques, and decision making through well experiments - successes of previous training (Monard and Baranauskas, 2003). One of the ways of working with machine learning is supervised models that seek to predict behavior through a known database. With the help of supervised models, methods can be found to solve a multitude of problems, such as, for example, the research by Dumortier *et al.* (2016), in which different models for smokers were investigated in situations associated with wanting or not wanting to smoke during an attempt to quit smoking, with the aim of predicting high-impulse states.

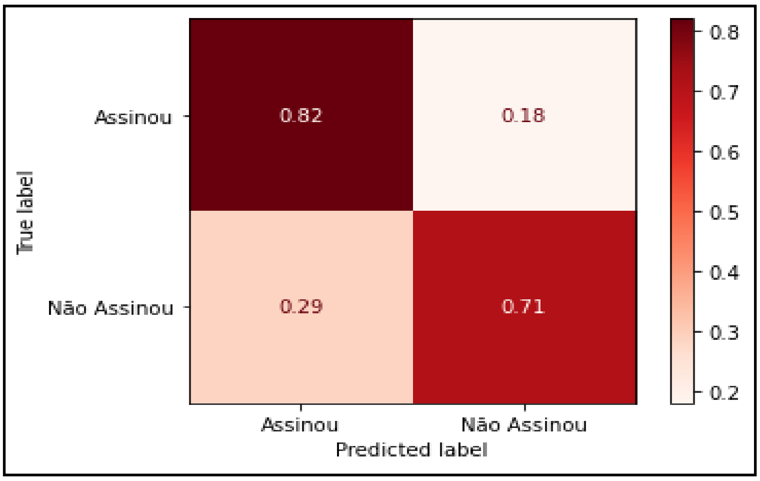
Decision Tree is a predictive model that makes partitions in subspaces, where each subspace serves as the basis for a different function. This can be used in different tasks such as: classification, regression and other analyses, as they improve forecasting models and can also make combinations between trees (Rokach, 2016). There are several studies on how to make a decision tree such as Luštrek *et al* (2016), Levatić *et al* (2017), Strnad and Nerat (2016), among others. A situation can be modeled in order to direct more efficient decision-making, where its predictive performance is slightly better than the standard algorithms (González, Herrera and Garcia, 2015).

The name "AdaBoost" is derived from *Adaptive Boosting* (meaning, adaptive impulse or stimulus), created by Yoav Freund and Robert Schapire, and is a meta-heuristic algorithm, which aims at increasing the performance of other learning algorithms. It is sensitive to noise in isolated data and cases, however, it does not suffer loss in generalization capacity after learning differently from most algorithms (Zhu *et al*, 2009). used as a comparative in several studies such as Xiao, Dong, Y., and Dong, Y. (2018), Sun *et al.* (2016), Lu, Hu and Bai (2015) among others.

The concept of *Gradient Boosting* emerged in Leo Breiman's observation that *boosting* can be interpreted as an optimization algorithm in a more suitable cost function. The *Gradient Boosting* is a machine learning technique for classification problems and regression, that generates a prediction model, usually decision trees, where it builds the model in steps, like other boosting methods, and generalizes, allowing optimization an arbitrary differentiable loss function (Hastie *et al*, 2009). One of the techniques used to evaluate the performance of a classifier is the confusion matrix – it is a table that presents correct and incorrect classifications, according to a previously adopted standard (Napolitano, 2019). From it, performance indicators can be generated. One of these indicators is the sensitivity or rate of true positives and the specificity or rate of false positives (Carvalho, *et al.*, 2011; Napolitano, 2019).

As can be seen in Figure 1, there are 4 different groups that we seek to evaluate to obtain a good development of our predictive model. Where each group is treated according to the priorities and the assertiveness margin to be reached by the company.

Figure 1. Confusion matrix.



Source: Authors, (2020).

According to the results seen in Figure 1, potential consumers are part of the True Positive (TP) with 82%, the False Positive (FP) group with 18%, deals with consumers who do not have adherence the credit offer, along with the True Negative (TN) with 29%. Finally, the False Negative (FN) group with 71% that in case the company does not contact wins the value of the connection, but loses the value of the offer.

2.2 Foster and Provost's expected value model

The relationship between the business problem and its solution through data mining can be broken down into treatable subproblems through the expected value analysis framework. Dividing the business problem into components corresponding to probability estimation and computation or estimation of values, together with a framework for recombining the components, is largely useful. For our example of call centers, the value added by a consumer must be taken into account, in addition to the accuracy of the model. It is difficult to realistically evaluate any customer segmentation solution without formulating the problem as an expected value (Provost; Fawcett, 2013a).

According to Foster, Melville and Tsechansky, (2007) a model is usually implemented considering the precision demonstrated by the algorithm in a classification task using a metric as a confusion matrix. However, this approach can mask important results when observing the results for the business (Foster, Melville and Tsechansky, 2007). In a consumer approach made without the application of a predictive model, the success of this approach tends to have results that are equivalent to the probability of a customer accepting or not, which can be defined a priori, as 50%. However, a classifier may considerably improve this approach, as long as data on available customers is available. However, it is necessary to observe the proportions of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (Provost; Fawcett, 2013b).

The idea of the expected value can be understood as the weighted average of the cost benefit between the portions of a Confusion Matrix by the value of the cost or benefit of each situation, for example a TP has a benefit that the acquisition of a product, a TN, does not it has a cost if it was decided not to approach consumers classified as negative, but a false positive has the cost of missed opportunity to acquire a new customer and an FP, the cost of approaching a customer who will not purchase the product.

3. Methodology

To achieve the objective proposed in this exploratory and experimental research, computational experiments were performed using Python version 3.7 as the main tool. The computer used to process the information was an Acer with AMD A12 Quad-Core processor, 8GB RAM and 1TB HD. The experiments carried out used the following Python libraries *matplotlib* for data visualization, *seaborn* that acts on top of *matplotlib* and helps to improve graphs in visual form, pandas for data management, manipulation and analysis, *scikit learn* for attribute extraction and machine learning algorithms such as decision trees, confusion matrix, among other analyzes and *numpy* for applying linear algebra and matrix operations.

To achieve the objectives of this research, the following steps were defined: 1) Review of the literature on and the operation of call centers and the use of learning and machine techniques; 2) Identification of supervised machine learning models for classifying bank credit offers through call centers; 3) Conducting experiments with the application of techniques to databases, as well as evaluation metrics; 4) Analysis of the results of the experiments involving identifying the performance of each technique and the respective impacts on the operation.

The literature review on the study area sought theories that strengthen the need for this study, as addressed by Gil (2002), this preliminary survey is an exploratory study with the purpose of providing familiarity with the study area and understanding the contributions of different authors given subject. This research selected the documents dealing with call centers, performance, machine learning in Brazil, supervised models, decision tree, classification models, algorithms, evaluation metrics, confusion matrix and predictive model.

The database used for the experiment was taken from the database whose data were related to a direct marketing campaign (phone calls) from a Portuguese banking institution. It has 11,162 records, with 17 variables (10 categorical and 7 numerical) listed below in Table 1. The purpose of this classification is to predict whether the customer will sign a letter of credit with the bank. To be able to apply the classification models later, it was necessary to treat categorical variables with the help of the function *pandas.get_dummies*, through *Python* to transform them into binary variables, reaching the total of 27 variables after the procedure.

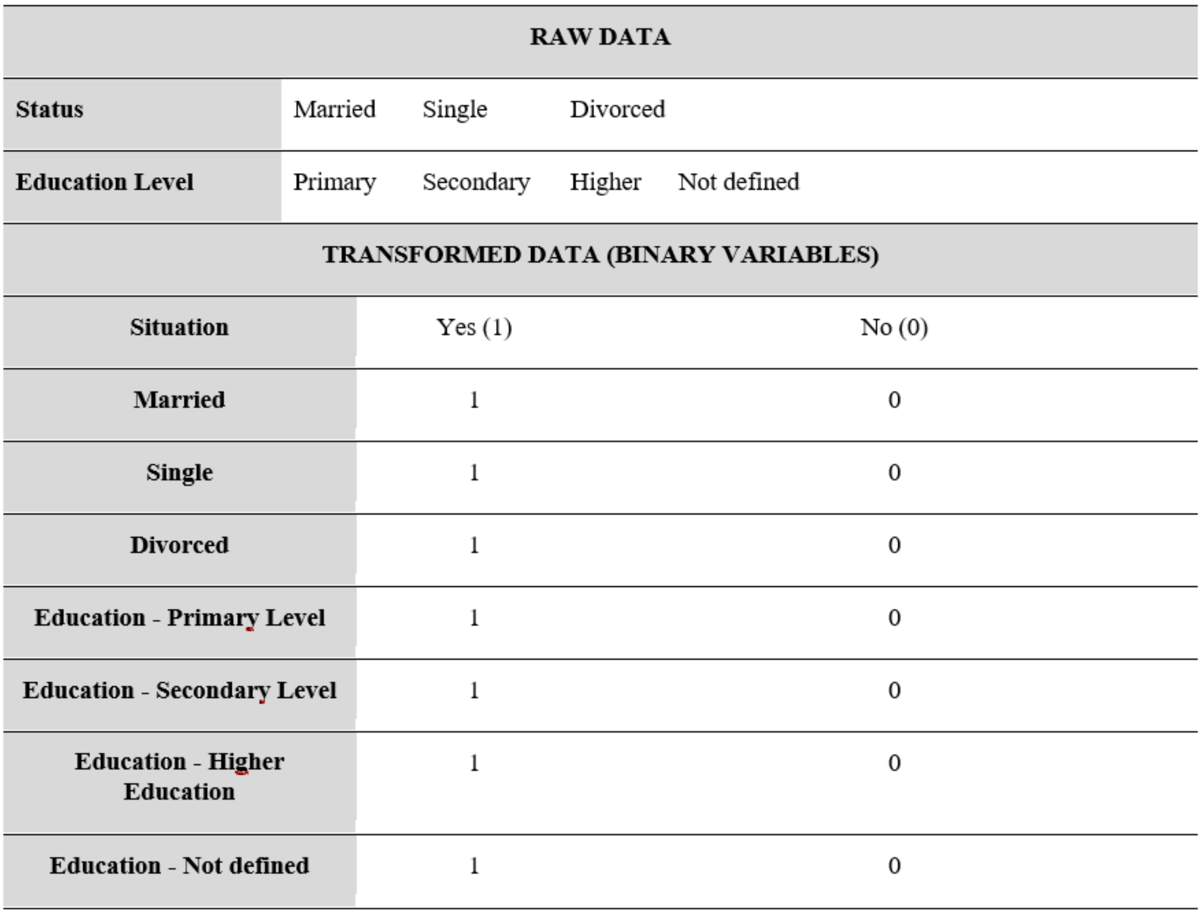
Table 1. Categorical and numerical variables.

Categorical Variable		Numerical variables	
Variable	**description**	**Variable**	**description**
job	admin, technician, services, management, retired, blue-collar, unemployed, entrepreneur, housemaid, unknown, self-employed, student	**age**	Age
marital	married, single, divorced	**balance**	Balance Banking
education	secondary, tertiary, primary, unknown	**day**	Last contact day
default	yes, no	**duration**	Contact
housing	yes, no	**campaign**	How many contacts during the campaign
loan	yes, no	**pdays**	How many days have passed since a previous campaign contact
deposit	yes, no (PREDICTOR VARIABLE)	**previous**	Contacts before the campaign
contact	unknown, cellular, telephone		
month	Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec		
poutcome	unknown, other, failure, success		

Fonte: Authors, (2020).

The realization of this approach allows to generate the algorithms with the predictions through decision trees and confusion matrix. Table 2 shows how this transformation was made in the variables Marital Status and Education Level. In it, it is possible to see how two variables that corresponded to marital status and educational level changed to 7 numerical variables, thus making it possible to apply the algorithms to make the predictions that will be demonstrated in the results. In this same way, all other categorical variables were transformed into numerical ones. All necessary precautions for the use of secondary data were taken (Martins *et al*., 2018).

Table 2. Examples of transforming Variables into Binary.



RAW DATA				
Status	Married	Single	Divorced	
Education Level	Primary	Secondary	Higher	Not defined

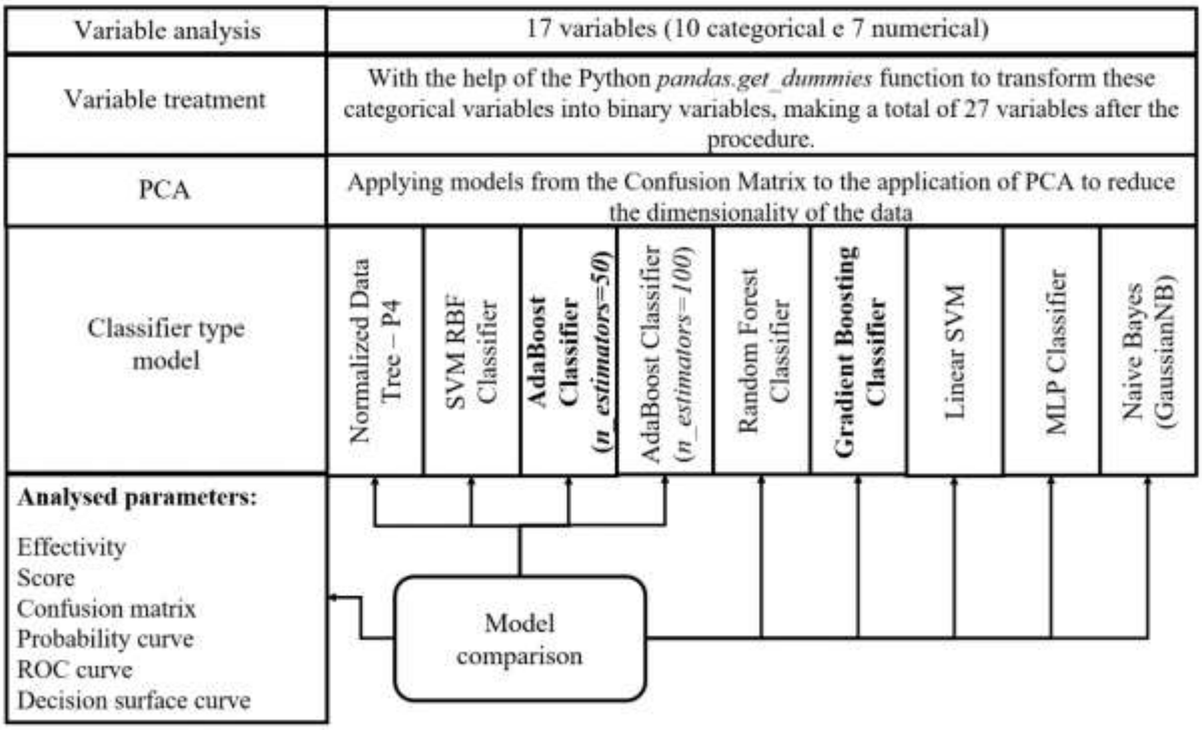
TRANSFORMED DATA (BINARY VARIABLES)		
Situation	Yes (1)	No (0)
Married	1	0
Single	1	0
Divorced	1	0
Education - Primary Level	1	0
Education - Secondary Level	1	0
Education - Higher Education	1	0
Education - Not defined	1	0

Source: Authors, (2020).

3.1 Model used to conduct the experiments

In Table 3 it is shown the order in which the experiments were conducted, initially the analysis and treatment of the variables, then the application of the models applying the PCA to reduce the dimensionality of the data and finally a comparison and analysis to understand which predictive model was more adequate to fulfill the objective proposed in this project.

Table 3. Techniques employed.



Variable analysis	17 variables (10 categorical e 7 numerical)								
Variable treatment	With the help of the Python *pandas.get_dummies* function to transform these categorical variables into binary variables, making a total of 27 variables after the procedure.								
PCA	Applying models from the Confusion Matrix to the application of PCA to reduce the dimensionality of the data								
Classifier type model	Normalized Data Tree – P4	SVM RBF Classifier	**AdaBoost Classifier *(n_estimators=50)***	AdaBoost Classifier *(n_estimators=100)*	Random Forest Classifier	**Gradient Boosting Classifier**	Linear SVM	MLP Classifier	Naive Bayes (GaussianNB)

Analysed parameters:

Effectivity
Score
Confusion matrix
Probability curve
ROC curve
Decision surface curve

Source: Authors, (2020).

In this Table 3, the steps followed to carry out the experiment were presented, from the analysis of the variables, the treatment with the function *get_dummies* for transformation into binary variables, application of the PCA to reduce the dimensionality of the data used, application of classifiers and parameters used to evaluate model performance.

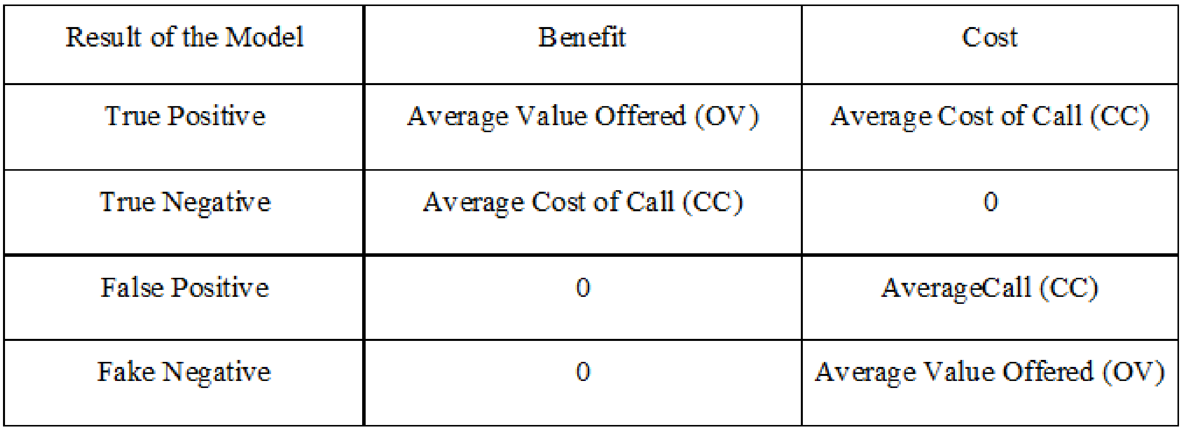
3.2 Analysis of the expected value of the models

In the next step after comparing the models, financial calculations were performed to understand which classifier offers the most savings for the banking institution, the results found in the models can be seen in the results presented. To do so, it compares the result without applying the model, considering the database. In this case the expected value (EV) consists of the difference between the average value of an offered contract (OV) and the average cost of the call (CC), multiplied by the percentage of successful calls (TS). This value must be subtracted and the average call cost (CC) multiplied by the rate of unsuccessful calls (NS). Therefore, we have:

$$EV = (OV - CC) \times TS - CMC \times NS \quad (1)$$

In an analogous way, you can calculate the expected value of the model using the results of the confusion matrix, but first you must verify the costs and benefits of each situation when the model is applied, which is presented in Table 4 below:

Table 4. Benefits and costs of the models.



Result of the Model	Benefit	Cost
True Positive	Average Value Offered (OV)	Average Cost of Call (CC)
True Negative	Average Cost of Call (CC)	0
False Positive	0	AverageCall (CC)
Fake Negative	0	Average Value Offered (OV)

Source: Authors, (2020).

In the case of a true positive (TP) the costs and benefits remain the same, but the success rate must with the use of the model must be higher without using any model. The true negative (TN) has no cost, but a benefit equivalent to the CC, while the false negative (FN) has only the cost of the call and the false positive has the cost of the lost sales opportunity. Therefore, as each model can have different TP, TF, FP, FN percentages, it is also necessary to define the rates of each data by the number of instances classified in each of these situations divided by the total number of instances in the test dataset, which are defined as TPR (TP rate), TFR (TF rate), FPR (FP rate) and FNR (FN rate). And the expected value of each model will be:

$$EV = (OV - CC) \times TPR + (-CC \times TPR) + (-CC \times FPR) + (-OV \times FNR) \quad (2)$$

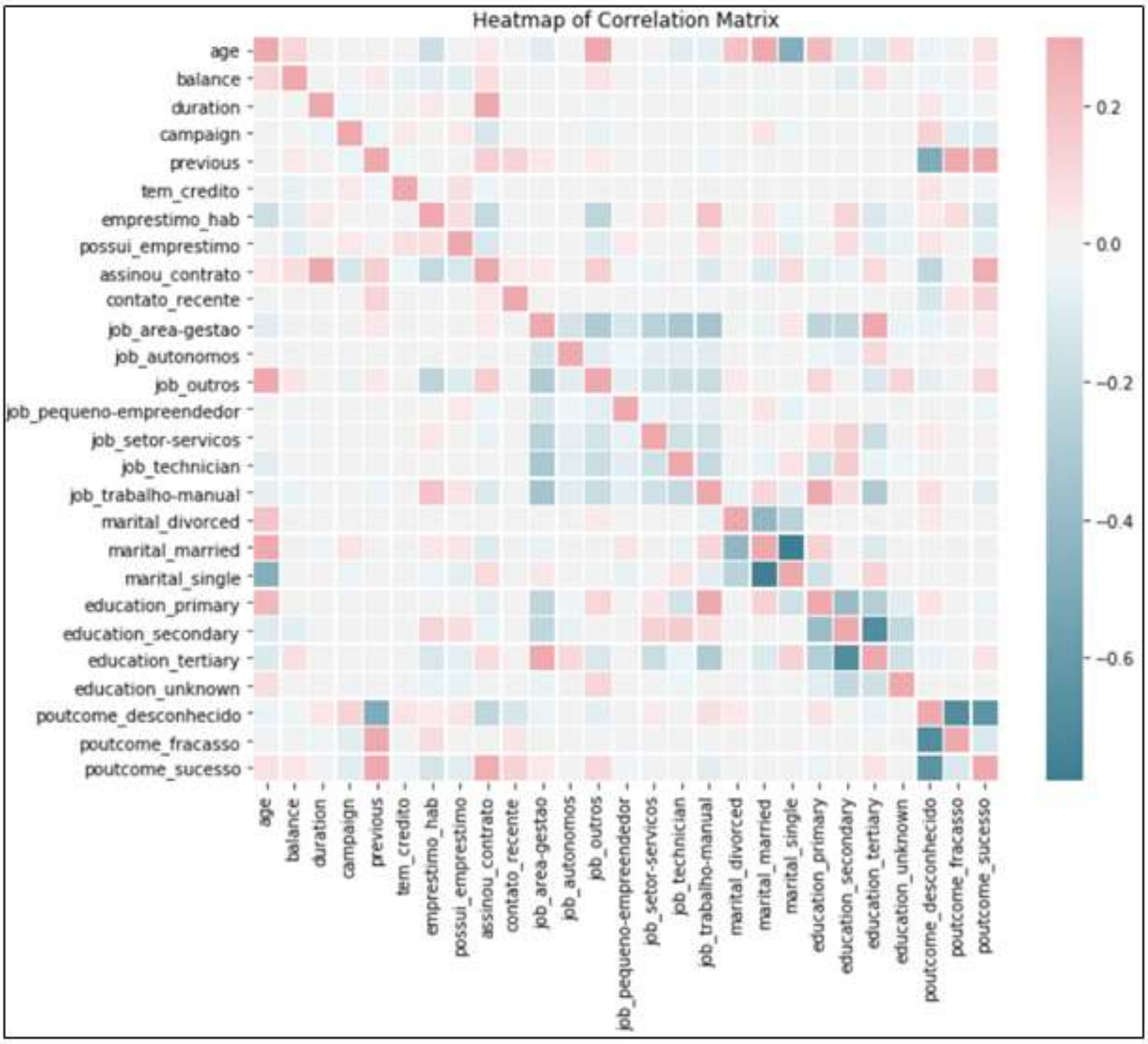
4. Results

After applying the mentioned techniques previously, it sought to ascertain and understand the profile of investors and to generate a predictive model offering greater savings to institutions. A sample with 11,162 records was used to conduct the experiments. This database is related to a marketing campaign through call centers in which the customer is offered to make an investment deposit. The results will demonstrate the profile of the customers that invest the most and then the classification models applied.

4.1. Exploratory data analysis

Of the 11,162 records analyzed, 5,289 (47%) signed an investment contract by telephone. Based on this first information, we tried to measure the degree of relationship between the variables found in the records. Figure 2 below will show the results of this Correlation Matrix and the variables that obtained the highest relationship index, thus understanding the importance of certain variables for analyzing and creating a predictive model.

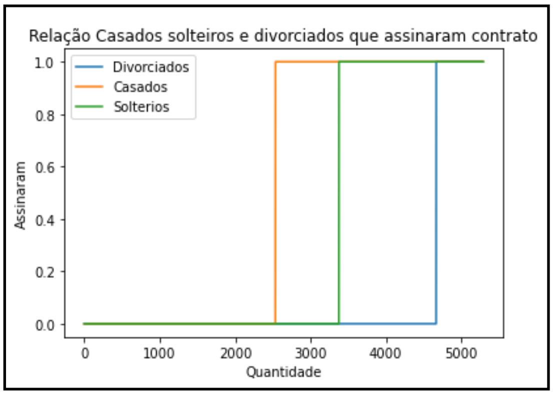
Figure 2. Correlation Matrix.



Source: Authors, (2020).

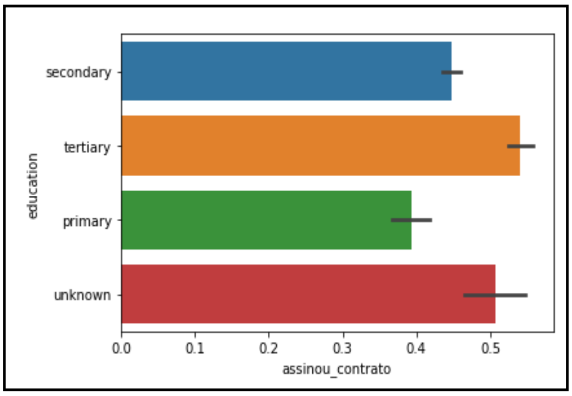
As illustrated in Figure 2, the results of this Correlation Matrix, thus, the variables that can bring greater knowledge to the client were: The length of time that the client is on the phone listening to the proposal; Single and divorced people make greater investment; Graduated people sign more contracts and there was strong evidence that it is always the same customers who make the investment. Next in Figures 3 and 4, it is possible to analyze the client profile.

Figure 3. Relationship between acceptance of the letter of credit and marital status.



Source: Authors, (2020).

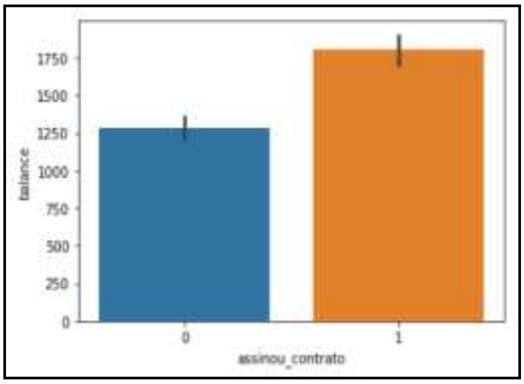
Figure 4. Highly educated customers tend to have better adherence.



Source: Authors, (2020).

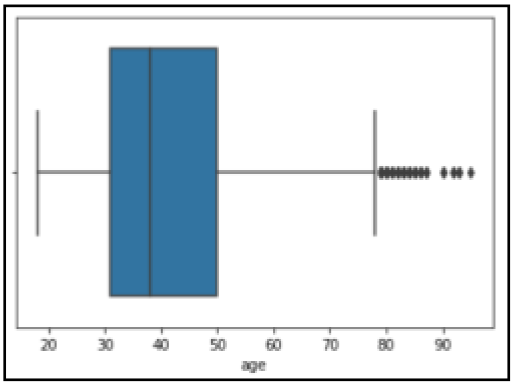
The next graphs were necessary to survey the customer's profile. In Figure 3, it can be seen that people with a single marital status (green line) receive more bank letter of credit proposals than married (orange) or divorced (blue) people. In Figure 4 we can see that people with a higher level of education (variable *tertiary*, in orange) have a better acceptance of bank investment proposals. Next in Figures 5 and 6, it is possible to analyze the investor profile.

Figure 5 - Average Age of Customers who sign a contract.



Source: Authors, (2020).

Figure 6 - Bank balance related to Contract Signing.



Source: Authors, (2020).

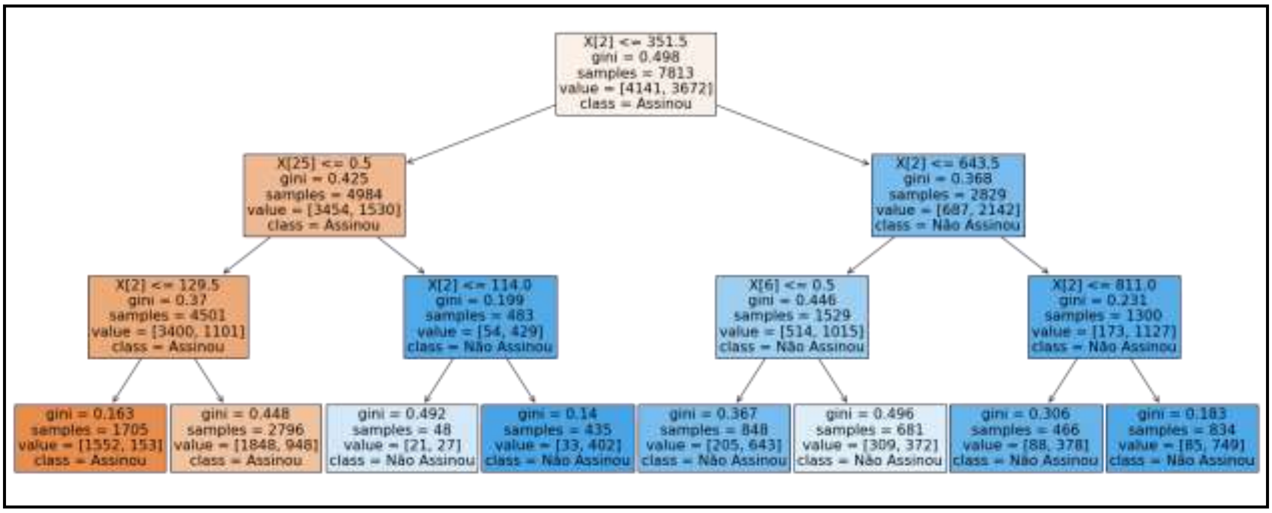
In Figures 5 and 6, still with the purpose of analyzing the investor profile, it was noticed that the people who signed the contract are between 30 and 50 years old and have a higher bank balance.

Other analyzes continued to be surveyed by means of graphics and as a final result of the characteristics of customers even before applying the predictive models, the following was noticed investment profile: The biggest consumers of this campaign were the people who had the highest bank balance, between 30 and 40 years of age, the higher the level of education the more you invest, the customers who signed a contract in previous campaigns also hired in the current one, clients who do not have a personal or housing loan sign more contracts and the customers who normally sign are those in which the institution took more than 60 days to contact them again offering a new service.

4.2 Predictive Models

The next step after surveying the consumer profile was to develop the predictive models using machine learning techniques, the data was separated so that 70% of the records were for training and 30% for testing. The first Decision Tree generated can be seen in Figure 7.

Figure 7. Decision Tree.

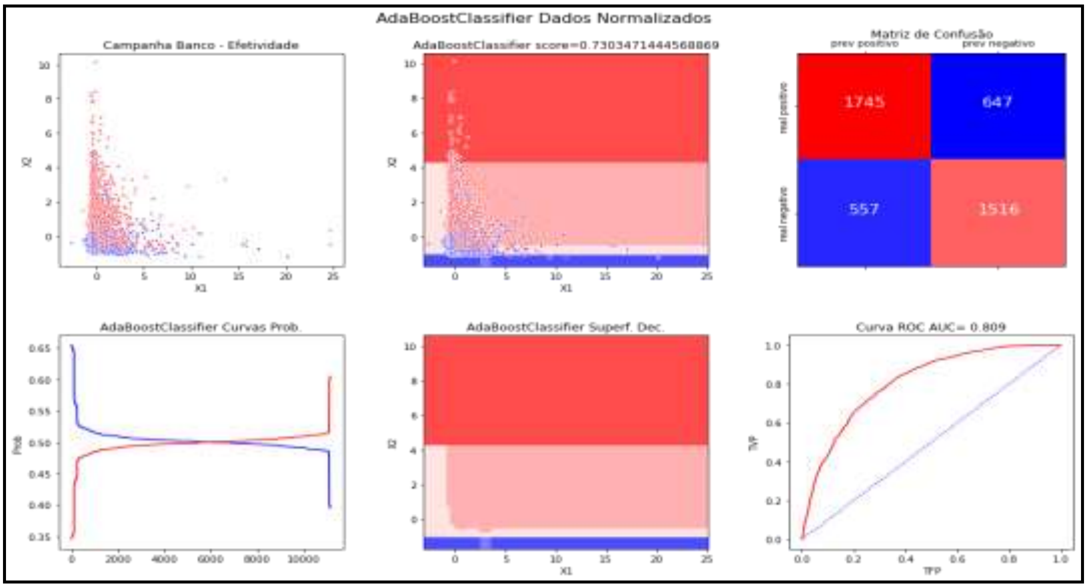


Source: Authors, (2020).

In Figure 7, the first rectangle the variable connection time (X [2]) which is more relevant to the number of people who joined proposal, followed by variable (X [25]) that deals with customers who signed a contract previously and have a predisposition to accept new bank credit offers.

This was followed by the application of the PCA to reduce dimensionality and tests with the classifiers. Figure 8 shows the result of the experiment with the AdaBoost classifier, its Confusion Matrix detected 1745 true positives and 1516 false negatives, which predicts a group with a high adherence profile to the credit proposal and a group with a high probability profile of not adhering. to the proposal, which leads us to a favorable scenario for action with greater efficiency, directed at groups with a greater propensity for acceptance. Its accuracy of 73.03% and ROC curve of 80.9%, was one of the best performances among the tested classifiers.

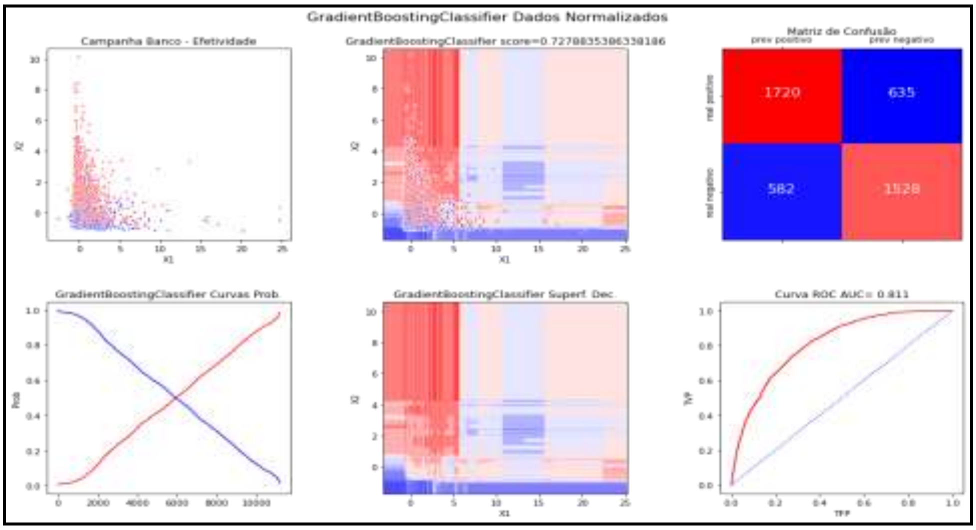
Figure 8. AdaBoostClassifier - n_estimators = 50 (Score 73.03%).



Source: Authors, (2020).

In the confusion matrix exemplified in the models of images 8 and 9, respectively *AdaBoost Classifier* and *GradientBoosting Classifier*, the 4 groups must be observed: VP (True Positive), FP (False Positive), VN (True Negative), FN (False Negative). For the analysis in question, the most relevant items are the False Negatives, so the company would not miss the opportunity to contract and the True Negatives, reducing the costs ofcalls by *call center* predicting the people who should not call.

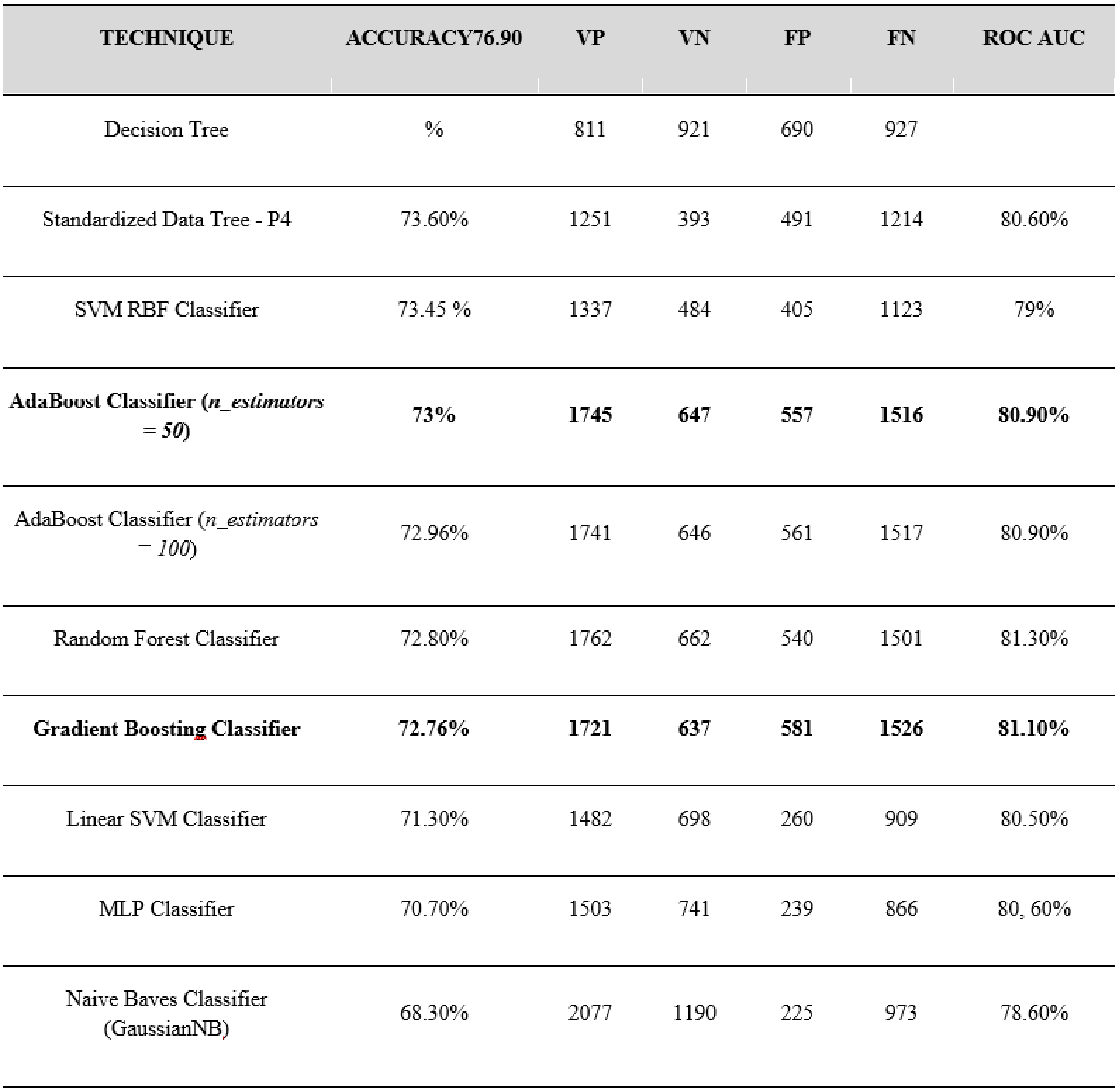
Figure 9. Gradient Boosting Classifier (Score: 72.76%).



Source: Authors, (2020).

Figure 9 shows the result of the experiment with GradientBoost classifier, as well as the AdaBoost classifier obtained good results in its Confusion Matrix which detected 1720 true positives and 1528 true negatives, which we are predicted by a group with a high adherence profile to the credit proposal and a group with a high probability profile of not adhering to the proposal, which also leads us to a favorable scenario for more effective action, directed at groups with a greater acceptance propensity. . Its score of 72.76% and ROC curve of 81.1%, was one of the best performances among the tested classifiers. Next, in Table 5, the results of accuracy, confusion matrix values and ROC curve of each classifier tested in the experiment.

Table 5. Comparative Table between Models.



TECHNIQUE	ACCURACY76.90	VP	VN	FP	FN	ROC AUC
Decision Tree	%	811	921	690	927	
Standardized Data Tree - P4	73.60%	1251	393	491	1214	80.60%
SVM RBF Classifier	73.45 %	1337	484	405	1123	79%
AdaBoost Classifier (*n_estimators = 50*)	**73%**	**1745**	**647**	**557**	**1516**	**80.90%**
AdaBoost Classifier (*n_estimators − 100*)	72.96%	1741	646	561	1517	80.90%
Random Forest Classifier	72.80%	1762	662	540	1501	81.30%
Gradient Boosting Classifier	**72.76%**	**1721**	**637**	**581**	**1526**	**81.10%**
Linear SVM Classifier	71.30%	1482	698	260	909	80.50%
MLP Classifier	70.70%	1503	741	239	866	80, 60%
Naive Baves Classifier (GaussianNB)	68.30%	2077	1190	225	973	78.60%

Fonte: Authors, (2020).

Table 5 presents results of accuracy, confusion matrix values and ROC curve of each classifier tested in the experiment, two classifiers were considered as the best performance among the ten tested models. The ***AdaBoost Classifier (n_estimators = 50)*** with 73% accuracy and 80.90% ROC curve and ***Gradient Boosting Classifier*** with 72.76% accuracy and 81.10% ROC curve. In Table 6, we can analyze the applicability of the models used within the study.

Table 6. Analysis Model Applicability.

Technical	**VP CMC**	**VN CMC**	**FP CMC**	**FN CMC**	**VP benefit**	**VNbenefit**	**benefit FP**	**FN benefit**	**expected value**
AdaBoost classifier (n_estimators = 50)	**US $ 27,274**	**US $ 10,113**	**US $ 8,706**	**US $ 23,695**	**US $ 8697. 726**	**R $ 0**	**-R $ 8,706**	**R $ 7,556,305**	**R $ 16,245,325**
Gradient Boosting Classifier	**R $ 26,899**	**R $ 9,956**	**R $ 9,081**	**R $ 23,851**	**R $ 8,578,101**	**R $ 0**	**-R $ 9,081**	**R $ 7,606,149**	**R $ 16,175,168**
Random Forest classifier	R $ 27,540	US $ 10,347	US $ 8,440	US $ 23,461	US $ 8,782,460	US $0	8,440 $R	R$ 7,481,539	R$ 16,255,559
AdaBoost classifier (n_estimators = 100)	R $ 27,212	R $ 10,097	R $ 8,768	R $ 23,711	R $ 8,677,788	R $ 0	-R $	8,768	R $ 7,561,289 R $ 16,230,309
Classifier Naive Bayes (GaussianNB)	R $ 32,464	R $ 18,600	R $ 3,517	R $ 15,208	R $ 10,352,536	R $ 0	-R $ 3,517	R $ 4,849,792	R $ 15,198,812
Tree Normalized Data - P4	R $ 19,553	R $ 6,143	R $ 7,674	R $ 18,975	R $ 6,235,447	R $ 0	-R $ 7,674	R $ 6,051,025	R $ 12,278,798
SVM RBF Classifier	R $ 20,897	R $ 7,565	R $ 6,330	R $ 17,552	R $ 6,664,103	R $ 0	-R $ 6,330	R $ 5,597,448	R $ 12,255,220
Linear SVM Classifier	R $ 23,164	R $ 10,910	R $ 4,064	R $ 14,208	R $ 7,386,836	R $ 0	-R $ 4,064	R $ 4,530,792	R $ 11,913,565
Classifier MLP Classifier	R $ 23,492	R $ 11,582	R $ 3,736	R $ 13,536	R $ 7,491,508	R $ 0	-R $ 3,736	R $ 4,316,464	R $ 11,804,237
Decision tree	R $ 12,676	R $ 14,395	R $ 10,785	R $ 14,489	R $ 4,042,324	R $ 0	-R $ 10,785	R $ 4,620,511	R $ 8,652,050

Fonte: Authors, (2020).

In Table 6, the applicability of the models was analyzed. Considering the provision of a full-time call center service (4 workers plus replacements), it costs at least R $ 18,000.00 each (PRESTUS Secretárias Compartilhadas, 2020), so each minute amounts to R$ 3.13. The average duration of a call in this study was 5 minutes. So, one may infer that a five-minute call costs the contracting company R$ 15.63. Multiplying this value by the TP, TN, FP and FN groups, we reach the cost of the call (CC). We used the amount of R$ 5,000 as an example of a credit offer. Then the value of the offer multiplied by each group of TP, TN, FP and FN, less the cost of the call, resulted in the values indicated as a benefit. Finally, the sum of the benefits presented resulted in the expected value.

5. Conclusions

In this article, a database analysis used by a call center company to offer credit to customers was proposed after refining and reorganizing the database leaving the information. With data ready to be analyzed, it was possible to start the work of data science. The objective was to understand the data and profiles of investors, and what the results would be to obtain through this analysis a better profitability of services and offers to customers.

In general, the results presented showed varying levels of significant correlation between variables, such as duration of the call, marital status, educational level, and even recurrence in adherence to credits, with the application of PCA (Principal Component Analysis) to reduce of dimensionality and the use of classification models such as: AdaBoost, Gradient Boosting, SVM RBF, Naive Bayes, Random Forest, it was possible to identify the correlation of the data that can best meet and mainly obtain a lower investment result to capture new investors .

Considering the results and gains obtained with the models, there is a very great opportunity to use these classification models to operate in the Telemarketing and call center markets, where the use of technology favors assertiveness in the best results in obtaining gains on product and service offerings. This article favors continuity for other studies related to the use of machine learning, prediction algorithms and classification models that can be addressed in different areas, markets and industries.

This work contributes to the development of the theory, by demonstrating that, even in situations of information ambiguity and asymmetry, it is possible - through data science techniques, to extract knowledge in order to make better informed decisions. In addition, this work contributes to the practice, by presenting refined results for immediate use in analogous situations - which can be expanded to credit areas and *call centers* of other types of businesses.

On the other hand, this work has some limitations. First, the data obtained came from secondary sources - that is, adaptation is needed (as could be explained previously and shown in Table 2) so that they are useful for the research in progress. Another limitation, arising from the same issue, is that it is not possible to complement the secondary data with extra information extracted from the same company, which may explain the forecast levels estimated in the models performed. A third aspect is that the database deals with a banking reality different from the Brazilian one - which means that the adoption of strategic planning based on the results obtained here must be carried out considering such differences. Therefore, further studies on the subject are recommended to improve the results obtained from the models, better adaptation to the local reality and possible future replication.

In terms of the continuity of research, one can list its evolution involving the use of hybrid models using Unsupervised and Supervised Learning, as well as evaluating the use of more complex algorithms that allow to increase the accuracy of the model and enable more optimized solutions.

References

Bateh, J., & Farah, J. (2017). *Reducing call center wait times through Six Sigma*. The Journal of Business Inquiry, 17(2), 131-148.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). *Statistical analysis of a telephone call center: A queueing-science perspective*. Journal of the American statistical association, 100(469), 36-50.

Clark, C. M., Tan, M. L., Murfett, U. M., Rogers, P. S., & Ang, S. (2019). *The call center agent's performance paradox: A mixed-methods study of discourse strategies and paradox resolution*. Academy of Management Discoveries, 5(2), 152-170.

Dumortier, A., Beckjord, E., Shiffman, S., & Sejdić, E. (2016). *Classifying smoking urges via machine learning*. Computer methods and programs in biomedicine, 137, 203-213.

Freund, Y., & Schapire, R. E. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of computer and system sciences, 55(1), 119-139.

Gião, P. R., Borini, F. M., & Júnior, M. D. M. O. (2010). *A influência da tecnologia no desempenho dos call centers brasileiros*. JISTEM-Journal of Information Systems and Technology Management (Online), 7(2), 335-352.

Gil, A. C. (2002). *Como elaborar projetos de pesquisa*, 4, 175. São Paulo: Atlas.

González, S., Herrera, F., & García, S. (2015). *Monotonic random forest with an ensemble pruning mechanism based on the degree of monotonicity*. New Generation Computing, 33(4), 367-388.

Gualandi, S., & Toscani, G. (2018). *Call center service times are lognormal: A Fokker–Planck description*. Mathematical Models and Methods in Applied Sciences, 28(08), 1513-1527.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Boosting and additive trees.* In The elements of statistical learning (pp. 337-387). Springer, New York, NY.

Hawkins, L., Meier, T., Nainis, S., & James, H. (2001). *The Evolution of the Call Center to Customer Contact Center*. Information Technology Support Center, White Paper.

Ibrahim, R., Ye, H., L'Ecuyer, P., & Shen, H. (2016). *Modeling and forecasting call center arrivals: A literature survey and a case study*. International Journal of Forecasting, 32(3), 865-874.

Levatić, J., Ceci, M., Kocev, D., & Džeroski, S. (2017). *Semi-supervised classification trees*. Journal of Intelligent Information Systems, 49(3), 461-486.

Lu, J., Hu, H., & Bai, Y. (2015). *Generalized radial basis function neural network based on an improved dynamic particle swarm optimization and AdaBoost algorithm*. Neurocomputing, 152, 305-315.

Luštrek, M., Gams, M., & Martinčić-Ipšić, S. (2016). *What makes classification trees comprehensible?*. Expert Systems with Applications, 62, 333-346.

Martins, F. S., da Cunha, J. A. C., & Serra, F. A. R. (2018). *Secondary data in research–uses and opportunities*. PODIUM Sport, Leisure and Tourism Review, 7(3).

Monard, M. C., & Baranauskas, J. A. (2003). *Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações*, 1(1), 32.

Moro, S., Cortez, P., & Rita, P. (2014). *A data-driven approach to predict the success of bank telemarketing*. Decision Support Systems, 62, 22-31.

Moro, S., Laureano, R., & Cortez, P. (2011). *Using data mining for bank direct marketing: An application of the crisp-dm methodology*.

Mwendwa, L. (2017). *Factors influencing call center agent attrition: A case of Kenya Power call center* (Doctoral dissertation, University of Nairobi).

PreRESTUS Secretárias Compartilhadas [Site institucional], Recuperado de <https://www.prestus.com.br/call-center/>.

Provost, F., & Fawcett, T. (2013). *Data science and its relationship to big data and data-driven decision making*. Big data, 1(1), 51-59.

Provost, F., & Fawcett, T. (2013). *Data Science for Business*: What you need to know about data mining and data-analytic thinking. "O'Reilly Media, Inc.".

Provost, F., Melville, P., & Saar-Tsechansky, M. (2007, August). *Data acquisition and cost-effective predictive modeling: targeting offers for electronic commerce*. In Proceedings of the ninth international conference on Electronic commerce (pp. 389-398).

Kaggle. Base de Dados Bank. Recuperado de <https://www.kaggle.com/henriqueyamahata/bank-marketing>.

Rokach, L. (2016). Decision forest: *Twenty years of research*. Information Fusion, 27, 111-125.

Strnad, D., & Nerat, A. (2016). *Parallel construction of classification trees on a GPU*. Concurrency and Computation: Practice and Experience, 28(5), 1417-1436.
Song, W., Du, C., & Zhang, C. (2018). *Research and Practice on Performance Test of Call Center Platform System*. JPhCS, 1069(1), 012088.

Sun, B., Chen, S., Wang, J., & Chen, H. (2016). *A robust multi-class AdaBoost algorithm for mislabeled noisy data*. Knowledge-Based Systems, 102, 87-102.

Xiao, L., Dong, Y., & Dong, Y. (2018). *An improved combination approach based on Adaboost algorithm for wind speed time series forecasting*. Energy Conversion and Management, 160, 273-288.

Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). *Class weights random forest algorithm for processing class imbalanced medical data*. IEEE Access, 6, 4641-4652.

Percentage of contribution of each author in the manuscript

Amanda Ferreira de Moura – 20%

Cíntia Maria de Araújo Pinho - 20%

Domingos Márcio Rodrigues Napolitano - 20%

Fellipe Silva Martins - 20%

João Carlos Franco de Barros Fornari Junior - 20%