Application of a Data Mining Model to Predict Customer Defection. Case of a Telecommunications Company in Peru

Mirko Bruno Vela López[1*], Maria Ysabel Arangurí García[2], Jessie Leila Bravo Jaico[3], Ángel Antonio Ruiz-Pico[4] and Ronald M. Hernández[5]

[1*]Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú. 73527613@usat.pe, Orcid: https://orcid.org/0009-0000-0973-0032

[2]Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú. maranguri@usat.edu.pe, Orcid: https://orcid.org/0000-0001-9220-5801

[3]Universidad Nacional Pedro Ruiz Gallo, Lambayeque, Perú. jbravo@unprg.edu.pe, Orcid: https://orcid.org/0000-0001-6841-2536

[4]Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú. aaruizpico@gmail.com, Orcid: https://orcid.org/0000-0003-2638-0593

[5]Universidad Continental, Lima, Perú. ronald.hernandez@outlook.com.pe, Orcid: https://orcid.org/0000-0003-1263-2454

Received: December 22, 2022; Accepted: January 26, 2023; Published: March 30, 2023

Abstract

In this research, a predictive model was developed using data mining techniques to analyze customer behavior, in order to identify and classify customers with a higher risk of defection in a Peruvian telecommunications company and thus, support the company in making accurate decisions and creating retention strategies. To achieve the main objective, the characteristics of the main data mining algorithms were analyzed based on the literature review, to determine the one that best suits the reality, obtaining the best performance in the proposed evaluation metrics with the XG Boost algorithm, which obtained 83% accuracy in determining potential customers at risk of defection. For the development of the prediction module based on the selected algorithm, the CRISP-DM methodology was used for the construction, evaluation and deployment. The deployment of the model was carried out by building a local web interface based on JavaScript and Python programming languages, using the Flask Framework to generate specific and global reports for the user. Finally, the degree of acceptable usability of the model was determined from two indicators; its effectiveness, demonstrated in the degree of precision obtained of 83%, the results in the evaluation metrics and the percentage of assertiveness of 80%; as well as the efficiency of the final interface.

Keywords: Data Mining, Algorithms, Supervised Learning, Classification, Telecommunications, Customer Defection.

1 Introduction

The telecommunications sector is being affected by high attrition rates (Zhang, T., 2022) due to intense competition, saturated markets, dynamic environments and the introduction of attractive offerings. All

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 14, number: 1 (March), pp. 144-158. DOI: *10.58346/JOWUA.2023.I1.012*

*Corresponding author: Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú.

these factors could result in a significant loss of revenue for the organization, but this can be controlled and maintained at acceptable levels (Mohamed, F.A., 2023).

In mobile telecommunications, the term "attrition" refers to the loss of subscribers who switch from one provider to another over a period of time and has a significant influence on the success of the company, as it is directly related to essential business objectives. (Gupta, K., 2022).

According to (Lu, N., 2012), the average mobile telecommunications churn rate is approximately 2.2% per month, which means, one out of every fifty subscribers of a given company discontinues its services each month. Furthermore, based on (Ahn, J.H., 2006) it has been shown that service providers as well as mobile operators lose 20% to 40% of their customers annually.

Customer churn is one of the main problems in many organizations and is more evident in the telecommunications sector, despite the efforts made and the strategies applied to keep customers. (Shabankareh, M.J., 2022).

It should be noted that telecommunications has become one of the key sectors of the global economy, but competition has increased as a result of technological advances and the increase in the number of service providers. In addition, operators are using a variety of strategies to survive in today's competitive marketplace. (Prakash, U., 2022).

This sector is known to handle large volumes of data. (Khalid, L.F., 2021) at high speed and the application of machine learning algorithms, therefore, allows a more accurate predictive analysis and supports better decision making, as stated by (Quasim, M.T., 2022). These machine learning algorithms currently focus, for example, on customer profiling, classification based on churn, and detection of features that affect churn (Hooda, P., 2022).

Some of the factors that affect customer defection are, according to (Ramadhanti, D., 2023), the type of contract, the number of monthly downloads, permanence, customer satisfaction value and add-ons. In addition, according to (Wanchai, P., 2017) today's customers can exercise their purchasing power by choosing among many operators or product providers to meet their communication needs.

Studies by researchers, such as the case of. (Ramadhanti, D., 2023) applies decision trees, according to this model. Customers of the study company tend to voluntarily unsubscribe, with some important factors affecting customer churn being contract type, number of monthly downloads, permanence, customer satisfaction value, and add-ons. In another study (Zhang, T., 2022) they use Fisher's discriminant equations and logistic regression analysis for attrition analysis in the telecommunications sector, where the latter obtained better results. While for (Shabankareh, M.J., 2022), they indicate that the integration of support vector machines (SVM) with the automatic chi-square interaction detection decision tree (Chaid) can generate the best result.

However, there are very few cases of studies carried out in Peru on this subject, such as the one conducted by (Melgarejo Galvan, A.R., 2017) who developed a Big Data architecture, using data from social networks to predict which customers may go to the competing company, based on their opinions.

Therefore, there is a clear need to know which customers will switch to a competing company in the near future, due to the high cost of acquiring new customers and the large existing offers and advertising campaigns, which increases the risk for companies in this sector.

Focusing on the reality of the telecommunications company in which this research was conducted, it was determined through a survey conducted with the data analyst of the business intelligence team, some of the possible causes of customer defection are: the low efficiency in analyzing the information collected from its customers. Being a top company in the country's telecommunications market, the

number of clients affiliated to its services is gigantic (approximately 20 million). At the same time, the amount of information provided or collected by each of these customers through its various systems is massive and cannot be handled by human analysts in its entirety. Moreover, not having systems capable of analyzing these large amounts of information to find patterns and create knowledge from it, the objective becomes unattainable, due to the time required to complete an analysis of customers using statistical techniques and other traditional analysis mechanisms. Another possible reason why a customer decides to drop out of the company is dissatisfaction with the service or timely attention to their complaints. Both reasons are expressed, for example, in direct complaints to the company, either for issues related to service subscription, service quality or irregular charges.

Given this, the importance of data mining is a useful tool for companies because it allows them to analyze their information from different perspectives, obtaining valuable information as soon as possible, supporting effective decision making and in turn, helping to build a customer profile based on their behavior.

For this reason, the purpose of this study is to develop a predictive model using data mining techniques to analyze customer behavior in order to identify and classify customers according to their probability of defection and thus support the company in making accurate decisions and creating retention strategies to improve their level of loyalty.

2 Literature Review

For the development of the present research, the results of related works were reviewed in order to identify the relevant methodology for this type of casuistry, being CRISP-DM, the one that not only raises in one of its phases the procedure for the prediction model, but also evaluates the business context, as highlighted by (Fareniuk, Y., 2022), as well as the importance of the deployment of an interface that allows, with a simple language, to present the scenarios, so that the user can define a correct strategy for customer retention, as stated in (Khalid, A., 2021).

However, the importance of the level of accuracy of the elaborated model is a concern expressed in different researches, which have applied the data mining (DM) approach, evaluating for example the decision tree (DT) methods. (Zhang, T., 2022), (Khalid, L.F., 2021), logistic regression (LR) reviewed in (Bogaert, M., 2023), (Chen, S.H., 2016) and others, such as artificial neural networks (ANN), k-nearest neighbors (k-NN) and support vector machine (SVM). (Yulianti, Y., 2020) in order to compare their performance for the analysis of defecting customers, based on their characteristics, in telecommunication companies.

These algorithms were applied to customer data, grouped by profile, i.e. behavioral patterns, based on data security using Blockchain technology. (Quasim, M.T., 2022) (Hooda, P., 2022), service quality (tariff, added value of customer service, etc.), brand or image (innovative and competitive aspects of the telecommunications company), consumer behavior (level of service consumption by different by-products, by time intervals, etc.), calling patterns, tariff information (Mitkees, I.M., 2017), costs, payment, customer service calls, demographic variables, to then establish on the basis of probabilities, defecting customers. (Kolli, N., 2020), recommending in other research, to complement customer interaction in their social networks, generating the multiplier effect of a positive or negative word-of-mouth.

Supervised machine learning algorithms were analyzed for the implementation process. (Mishra, K., 2017) using Python on a dataset in mobile telecommunications company and the use of XGBoost which provides comparatively more accurate prediction than other learning models. (Muthupriya, V., 2022).

The concept used in general is "customer churn", which is defined as the customer's decision to no longer remain in contact with the company, as highlighted by. (Huisheng, Z.H.U., 2020) (Gupta, K., 2022) y (Höppner, S., 2020).

3 Methodology

The research methodology applied in this study is based on the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. The phases that make up the methodology are: (1) business understanding, (2) data collection, (3) data preparation, (4) model development, (5) model evaluation, and (6) deployment. These are described in Figure 1 and are applied in a case of a telecommunications company in Peru.

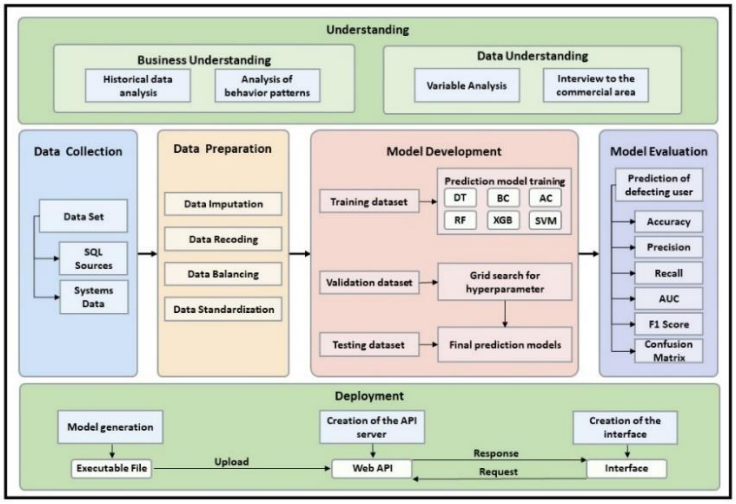


Figure 1: Research Methodology

Note: DT: Decision Trees, BC: Bagging Classifier, AC: Adaboost Classifier, RF: Random Forest, XGB: Xg boost, SVM: Support Vector Machine.

The data was extracted from the historical data of the telecommunications company, on a monthly basis and covers the period from January 2019 to July 2021, is in a *ccv format*, with 13 variables to work with and a total of a total of 65150 records (rows) that will be part of the análisis (see Table 1 and 2)

Table 1: Research Methods

Method	Description
Analytical	Study and analysis of the problem presented by the organization
Deductive	Strategy for the development of the solution to the problem
Implementation	The proposed solution will be implemented

Table 2: Data Collection Techniques and Instruments

Techniques	Instruments	Elements of the population	Purpose
Internal sources	Reports (Data Excel)	Customers	Knowing the customer profile
Literature review	Summary sheets	Scientific articles and research	To learn about the solution alternatives developed in previous studies.
Interview	Interview guide	Data Analyst/Business Intelligence Team Worker	To know the context and the variables of analysis

4 Results and Discussion

The research was developed in 6 phases: phase 1: business understanding, phase 2: data understanding, phase 3: data preparation, phase 4: modeling, phase 5: evaluation, and phase 6: deployment.

Phase 1: Understanding the Business

This phase aims to understand the business objectives, as well as the problem of customer "*attrition*" in the telecommunications field, and also seeks to shape the data mining objectives and the project plan to be carried out. To this end, the company pursues the financial consolidation of the company and accelerate the transformation to capture long-term value. Three objectives are established: (1) convergence and retention, through customer loyalty and capture, (2) improvement of the customer experience, through adaptation, digitization and delivering on the customer promise, and (3) operational efficiency, through automation and digitization of processes. Likewise, the term "*attrition*", in the telecommunications sector, refers to the loss of subscribers who switch from one provider to another during a given period of time.

From the analysis of the interviews conducted, it was observed that the main causes of client "*desertion*" in the sector are: (1) the lack of efficiency in analyzing the information collected from clients, since there are no automated mechanisms to process the data collected from the large number of clients available, (2) dissatisfaction with the service or timely attention to their complaints.

Consequently, the sector requires a system or model that allows it to automate the analysis of the company's information to find patterns of behavior based on customer information, and to identify potential customers at risk of dropping out in order to implement preventive measures (see Table 3)

Table 3: Data Mining Objectives

Objective	**Definition**
Objective 01	Use historical client information to identify relevant patterns, variables or trends that may explain the attrition rates presented.
Objective 02	Analyze the relationship between behavioral patterns, variables or trends manifested in the historical data of customer transactions.
Objective 03	Development of a predictive model using available customer behavior data to forecast the likelihood of churn for each customer.
Objective 04	Group and report customers based on their potential and probability of defection.

Phase 2: Understanding the Data

The main source of data is obtained from the "*data set*" offered by the telecommunications company, which provides historical data representing customer behavior in the company and their degree of satisfaction with the service received. The data set provided by the telecommunications company has 65150 records and 13 variables that represent the customer's historical behavior. Table 4. defines the variables analyzed in the study (see Table 4)

Table 4: Data Dictionary

Variable	Type	Scale of Measurement	Description
RUC	Qualitative Nominal	Reason	Unique customer identifier
CANT_LLAM_SAL_TOT	Quantitative Discrete	Reason	Indicates the total number of outgoing calls from the customer
MIN_SAL_TOT	Quantitative Continuous	Reason	Indicates the total outgoing minutes used by the customer.
CANT_LLAM_ENT_TOT	Quantitative Discrete	Reason	Indicates the total number of incoming calls from the customer
MIN_ENT_TOT	Quantitative Continuous	Reason	Indicates the total incoming minutes the customer received
QTY_SMS_TOT	Quantitative Discrete	Reason	Indicates the total number of messages the customer has sent.
MB_TOTAL	Quantitative Continuous	Reason	Indicates the total consumption of Gigabytes used by the client.
PERMANENCE	Quantitative Discrete	Reason	Indicates the customer's tenure with the company in months.
Q_MOBILE	Quantitative Discrete	Reason	Indicates the total number of the customer's mobile lines in the company.
SEGMENT	Qualitative Nominal	Nominal	Indicates the sector to which the customer belongs
Q_RECLAMOS	Quantitative Discrete	Reason	Indicates the number of claims and calls due to failures made by the customer during his permanence with the operator.
BILLING	Quantitative Continuous	Reason	Indicates the total amount to be paid in soles for customer consumption.
CLIENT	Quantitative Discrete	Reason	Indicates whether the customer has defected or not

The variables were grouped into 2 subgroups: (1) numerical variables or magnitudes and (2) categorical variables. Each subgroup of variables is intended to be processed independently due to the nature of the variables. The variable "RUC" was not grouped in this division does not represent any relevance. Table V shows the code established for processing the variables. (See figure 2)

```
columnsNumeric = ['CANT_LLAM_SAL_TOT', 'MIN_SAL_TOT', 'CANT_LLAM_ENT_TOT', 'MIN_ENT_TOT', 'CANT_SMS_TOT',
                  'MB_TOTAL', 'PERMANENCIA', 'Q_MOVILES', 'Q_RECLAMOS', 'CLIENTE', 'FACTURACION']
columnsString = ['SEGMENTO']
```

Figure 2: Excerpt from the Code in which the Variables are Grouped into Numerical and Categorical Data

Subsequently, we proceeded with the statistical-descriptive analysis of the variables using a function of the Python pandas library. It was established that the available data showed a large dispersion of values in certain variables, confirming the existence of outliers and missing values in the variable BILLING.

Phase 3: Data Preparation

Since there are missing records in the numerical variable INVOICING, it was decided to replace the empty values with the median. Thus, these parameters will not influence the final result. A library called SimpleImputer was used, which helped to generate a data imputer that had as parameters the missing value expression (NaN) in that variable and the imputation strategy, using the median

Likewise, the Label Encoder library was used to encode each of the records by providing a numeric value to each unique value within that variable, storing the results of the encoding in the data_recod variable (later df_imp).

The variables BILLING and SEGMENT of the initial data were recoded to provide the final consolidated data to be used in the construction of the model.

When analyzing the distribution of the data in the TARGET variable, it was observed that the proportion of clients identified as "non-defectors" was greater than the clients identified as "defectors", so as not to incur in unbalance problems that could affect the results of the model. Therefore, the ADASYN algorithm was used to create synthetic information from the data provided and reduce the imbalance present in the distribution of values of this variable.

The database was then divided into (1) training data and (2) test data. Data were used to train the predictive model and data were used to test whether the model generated from the training data performs well. A percentage of 80% and 20% was considered for the training data and test data respectively, as proposed by (Shabankareh, M.J., 2022). This process can be visualized in Figure 3 below.

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2)
X_train.shape, X_test.shape, Y_train.shape, Y_test.shape
```

```
((69258, 12), (17315, 12), (69258,), (17315,))
```

Figure 3: Determining the Training and Test Data

Phase 4: Modeling

It was established that the model to be followed should be based on supervised methods, since these have the characteristic of working on the basis of a target variable, predicting the output in this variable aided by the behavior of the other independent variables (Shobana, J., 2023).

Also, according to (Shen, Y., 2023), aligned classification methods are applicable to research data mining objectives, as they are used in situations where predicting an output on a target variable is required due to their high degree of accuracy.

In addition, the classification methods divide the data into 2 random sets to be worked one as a training set and the other as an evaluation set. Taking into consideration the previous analysis, experience and research by (Shen, Y., 2023) and (Qin, C., 2021), it was considered pertinent to use the models mentioned below, being these the most employed by the authors in the different bibliographies and considering the characteristics exposed by (Muthupriya, V., 2022) (Höppner, S., 2020) and (Shen, Y., 2023). Table 6 shows a summary of the characteristics of the algorithms used (see Table 5)

Table 5: Comparative Characteristics Of The Algorithms Used

Algorithm	Difficulty	Number of records	Level	Type of apprenticeship
Decision Tree	Easy	High	Medium	Supervised
Bagging	Media	Media	Medium	Supervised
Adabost	Media	Media	Medium	Supervised
Random Forest	Media	High	Medium	Supervised
Xgboost	Media	High	Medium	Supervised
SVM	Difficult	High	High	Supervised

Considering that the models to be implemented are part of the classification methods, the following evaluation metrics were considered relevant to evaluate the performance of the model: precision, accuracy, recall, auc, f1_score and the confusion matrix. These metrics were previously detailed in the theoretical bases.

Different models were built based on different algorithms in order to find the one that best adapted to the present reality, obtaining the best results in the proposed evaluation metrics.

The first of these algorithms was the Decision Trees algorithm and was trained using the training and test data previously obtained. When implementing certain parameters were used such as: the Gini criterion that was used to measure the quality of each division of the tree, min_sample_split with a minimum number of samples of 20 to divide a node, min_samples_left of 5 as minimum number of leaf node samples needed and a maximum max_depth. The results obtained were quite low and increased as the depth of the implemented tree increased, with the best result being 0.60 accuracy at its deepest point. These results can be seen in Figure 4.

```
# Mostramos los resultados obtenidos
do = pd.DataFrame({"Max Depth": depth_range, "Average Accuracy": accuracies})
do = do[["Max Depth", "Average Accuracy"]]
print(do.to_string(index=False))
```

```
 Max Depth  Average Accuracy
         1          0.352694
         2          0.364927
         3          0.505495
         4          0.515257
         5          0.566708
         6          0.546600
         7          0.590959
         8          0.569194
         9          0.589578
        10          0.596086
        11          0.602226
        12          0.603684
        13          0.604712
```

Figure 4: Decision Tree Algorithm

In order to improve the results of this algorithm, ensemble algorithms were used to help improve the performance of the models by randomly building a larger quantity of the same model to compare results.

The first ensemble algorithm used was the Random Forest algorithm. Likewise, the libraries to be used were imported and the method was executed with the parameters used in the previous model, providing as a premise the creation of 500 trees (n_estimators) to obtain different results and choose the best one. The results of this algorithm can be seen in Figure 5. The implementation of this algorithm offered acceptable results since it is based on building different decision trees to obtain the best results.

```
[ ]  rf.fit(X_train, Y_train)

     RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                            criterion='gini', max_depth=None, max_features='auto',
                            max_leaf_nodes=None, max_samples=None,
                            min_impurity_decrease=0.0, min_impurity_split=None,
                            min_samples_leaf=1, min_samples_split=2,
                            min_weight_fraction_leaf=0.0, n_estimators=500,
                            n_jobs=-1, oob_score=False, random_state=None, verbose=0,
                            warm_start=False)
     accuracy: 0.786601
     precision: 0.822437
     recall: 0.750111
     f1_score: 0.784611
     auc: 0.787977
```

Figure 5: Random Forest Algorithm

The second assembly algorithm used was Bagging. The libraries to be used were also imported and the method was executed with the default parameters in order to first visualize the results obtained. In turn, this algorithm works on the basis of a previous algorithm, so the previously implemented decision tree algorithm was used as the core. The results can be seen in Figure 6. This algorithm recalibrated the initial results and achieved an optimal final result in the results of the evaluation metrics. However, the final accuracy level of the model is not reflected in the consistency matrix, since in different scenarios it fails to meet the ideal expectations.

Figure 6: Bagging Algorithm

The third algorithm implemented to improve the initial model was AdaBoost. Following the same process, this algorithm was worked on the basis of the Decision Tree algorithm with the objective of improving its results by testing different scenarios (see Figure 7). As can be seen, this model decreased its effectiveness according to the evaluation metrics. Therefore, it was decided to implement an additional ensemble model in order to improve these results.

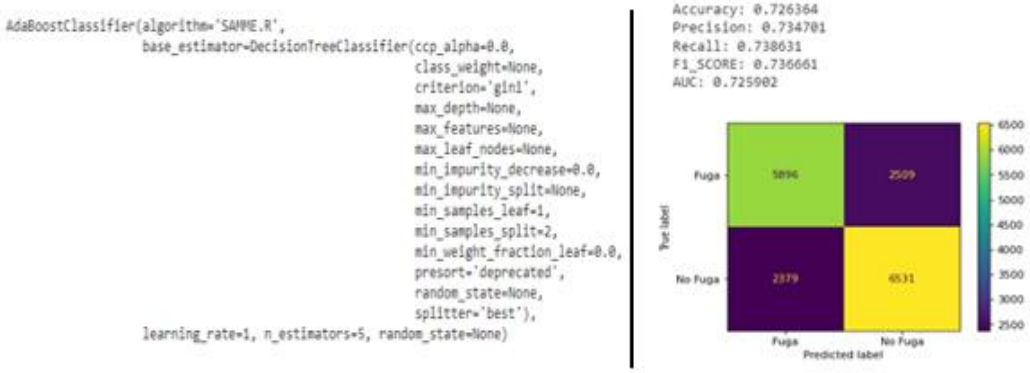


Figure 7: AdaBoost Algorithm

The fourth assembly algorithm implemented was "XG Boost". This algorithm, like the previous assembly algorithms, works on the basis of a previous model to improve the results. As this algorithm is much more accurate than the previous ones, it was decided to adjust the model parameters from the beginning, generating in first instance, a list with the parameters most used by this model and their possible values. This list was generated with a function "Grid Search CV" which is responsible for testing the model with the different parameters and scenarios to optimize the model. Figure 8 shows the results of the model. This model obtained very high results in the chosen evaluation metrics, reflected in the degree of prediction of the designed confusion matrix. This is a model with a great opportunity to be taken into account at the end of this stage.

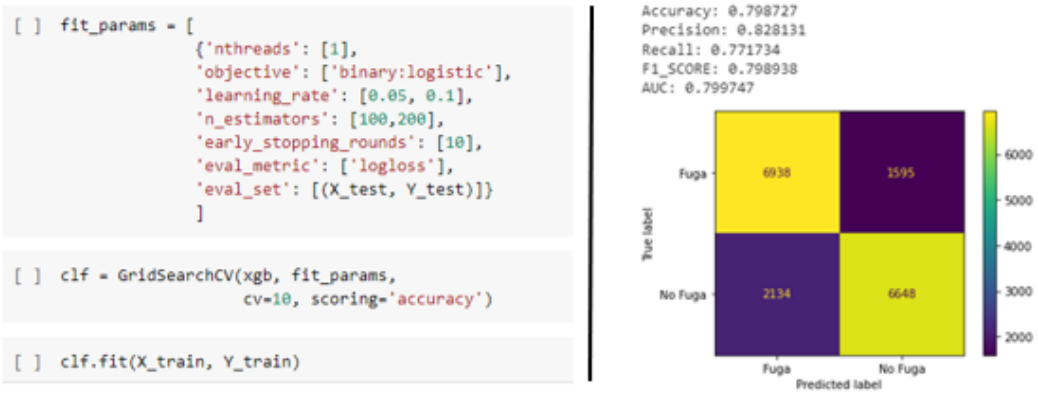


Figure 8: XG Boost Algorithm

The fifth ensemble model used was the Sector Vector Machine (SVM), which is an algorithm of greater scope compared to the previous ones. The first step was to import the libraries and use a scaling function to homogenize the information. The method selected to scale the data was the function par excellence called Standard Scaler (). Once the scaled data to be used was obtained, the Grid Search CV function used in the previous model was used to test the model with the different options and scenarios and to obtain the optimal parameters to be used. For this process, the most used parameters according to the bibliography of this algorithm were taken into consideration, which are the following: 'C' in charge of model regulation, 'Gamma' which represents the coefficient for the chosen kernel and 'Kernel', being 'rbf' the chosen one since it is the most popular kernel function for the implementation of this model. The results can be seen in Figure 9.

Figure 9: SVM Algorithm

Table 7 shows the comparative results of the 5 assembly models used to select the algorithm that best adapted to the reality of the organization and obtained the best results in the proposed evaluation metrics.

Table 6: Comparison of Model Results

Algorithm	Accuracy	AUC	F1_SCORE	Precision	Recall
XGBoost	0.80	0.80	0.80	0.83	0.77
Random Forest	0.79	0.79	0.78	0.82	0.75
Bagging	0.79	0.79	0.78	0.84	0.73
AdaBoost	0.73	0.73	0.74	0.73	0.74
SVM	0.63	0.63	0.65	0.64	0.67

Phase 5: Evaluation

Evaluate Results

From the information processed, it was possible to conclude that the model that best adapted to the information and the reality exposed was the model developed based on the XGBOOST algorithm compared to the other models developed, obtaining an accuracy of 80% and an AUC of 80%. Figure 10 shows the confusion matrix resulting from the evaluation of this model.

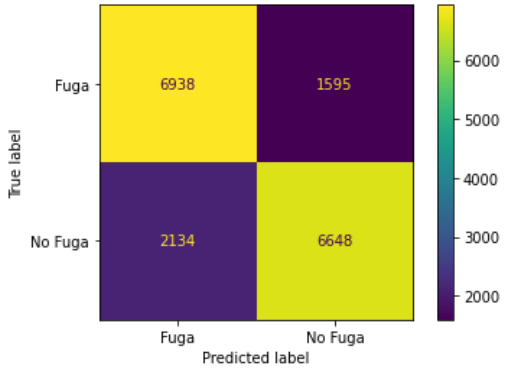


Figure 10: Confusion Matrix of the Selected Model

According to the confusion matrix, the model correctly classified 6938 remaining customers while 1595 were misclassified customers. It also correctly classified 6648 customers who absconded to competitors and 2134 customers who stayed with the company, but actually left with competitors.

To verify that the results of the model were favorable, a quality assurance (QA) test was performed, initially comparing the actual values of the CHURN variable with the values predicted by the model. Afterwards, random scenarios were assigned to the model to evaluate the prediction in each of them. This process resulted in the model behaving ideally and predicting the scenarios correctly by 83%, confirming the resulting values in the evaluation phase. Finally, the model was able to correctly predict the CHURN variable, receiving random scenarios from the base provided by the organization.

Once the model that adapted to the present reality was evaluated and selected, we proceeded to the development of the architecture necessary for its deployment and operation on the web. This step helped to integrate the results of the final model with the organization's business process and produce a report of results for the end user.

Phase 6: Deployment

Phase 1 of the deployment consisted of exporting the predictive model in the form of an executable file using the Python pickle library, which will be the basis of the web API to respond to customer requests and generate the requested predictions.

Phase 2 of the deployment consisted of the construction with the Flask development framework and the loading of the API on the web server that will serve as a link between the model and the final interface that will be displayed by the client. Once the tools to be used were selected, we proceeded with the construction of a test API. For this purpose, 2 methods were used: (1) the GET method, which will be in charge of responding to any direct request to the API, and (2) the POST method, which will be the method used for communication and sending data from the form. The objective is to attend to the client's requests and respond to each of them through the generated model.

Phase 3 of the deployment consisted of the development of the user interface and its direct connection to the API, so that the user can make requests directly from the interface and receive the required predictions, as well as the requested reports.

We monitored that the API was correctly uploaded to the server and that the GET and POST methods worked correctly. POST requests were made with certain clients chosen at random from the data used to train the model, in order to check that the correct prediction was made and ensure the correct functioning of the model and the API on the server.

For the development of the final interface, the Flask framework and the Python programming language were used, as well as pure HTML markup language, design language such as CSS, the Bootstrap 4 cross-platform library for designs and the JavaScript interpreted programming language for some functionalities within the interface.

The final interface has 2 sections: (1) the main section, called "Home", which offers the possibility of evaluating and predicting based on the probability of defection, a single client selected from the data provided and (2) the section called "Global Analysis", which offers the possibility of evaluating and predicting based on the probability of defection, an accumulation of clients provided based on the data. This section has the possibility of generating specific and global reports in Excel format (for one or several clients) as required by the end user.

5 Conclusions

By implementing the data mining model to predict customer defection in a telecommunications company, it was concluded:

When comparatively analyzing the algorithmic characteristics of the data mining techniques based on the main classification algorithms used in the literature, the XGBoost algorithm was identified as the one that best adapted to the present reality, obtaining 83% accuracy in predicting customers with a possible risk of dropping out of the company, and 80% sensitivity, which represents the percentage of true negatives that the model was able to predict correctly, i.e., the percentage of customers with no risk of defection predicted correctly, which means that this algorithm was able to adapt to the needs of the organization.

Based on the selected algorithm, the prediction module was built considering the variables that define the customer's behavior, using the CRISP-DM methodology as a development guide for the model construction, evaluation and deployment stage. At the same time, the connection between the prediction module and the final interface was established through an API developed with the Flask framework, so that the end user can use the model in an intuitive and friendly way, maintaining the 83% degree of accuracy in the positive cases detected correctly, achieved by the model.

We were able to report behavioral patterns based on the model variables that define customer behavior scenarios through the implementation of the predictive model in an intuitive local web interface

built from HTML markup language and JavaScript and Python programming languages, which has the ability to generate specific and global downloadable reports as required by the end user responsible for customer evaluation. In addition, the final interface provides the ability to assign an attrition probability percentage, which will allow the user to gain insight into which customers are most at risk and which should be targeted for retention strategies.

Finally, it was possible to determine the degree of acceptable usability of the model based on its effectiveness and efficiency. With respect to effectiveness, the model obtained 83% accuracy for the prediction of customers at risk of dropping out of the company and 0.80 of positive cases detected correctly, which represents a degree of assertiveness in the results of approximately 80%, if the model had been implemented previously, which is expected to help reduce the dropout rate in the applicable context. With respect to efficiency, we were able to build a simple and intuitive web interface, with which the end user can interact quickly and make an evaluation of the clients based on the data provided to obtain an answer in a matter of seconds (or minutes depending on the amount of information provided and the computational load) with respect to the client's current situation, that is, his potential risk of dropping out; In addition, white box and black box tests were performed to ensure the correct functioning of the execution flows and to verify the correct functionality of the final interface , as well as the evaluation and acceptance of the final product by the company's data analyst.

References

[1] Ahn, J.H., Han, S.P., & Lee, Y.S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy*, *30*(10-11), 552-568.

[2] Bogaert, M., & Delaere, L. (2023). Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art. *Mathematics*, *11*(5), 1-28.

[3] Chen, S.H. (2016). The gamma CUSUM chart method for online customer churn prediction. *Electronic Commerce Research and Applications*, *17*, 99-111.

[4] Fareniuk, Y., Zatonatska, T., Dluhopolskyi, O., & Kovalenko, O. (2022). Customer churn prediction model: a case of the telecommunication market. *Economics*, *10*(2), 109-130.

[5] Gupta, K., Hardikar, A., Gupta, D., & Loonkar, S. (2022). Forecasting Customer Churn in the Telecommunications Industry. *In IEEE Bombay Section Signature Conference (IBSSC)*, 1-5.

[6] Hooda, P., & Mittal, P. (2022). An Optimized Kernel MSVM Machine Learning-based Model for Churn Analysis. *International Journal of Advanced Computer Science and Applications*, *13*(5), 487-494.

[7] Höppner, S., Stripling, E., & Baesens, B. (2020). vanden Broucke S., Verdonck T. *Profit driven decision trees for churn prediction*, *European Journal of Operational Research*, *284*(3), 920-933.

[8] Huisheng, Z.H.U., & YUb, B. (2020). Customer Churn Prediction Based on HMM in Telecommunication Industry. *Fuzzy Systems and Data Mining VI: Proceedings of FSDM 2020*, *331*, 78-92.

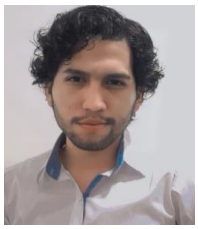
[9] Khalid, A., Khedr, A., Abdulrahman, H., & Zeki, A.M. (2021). Identifying State-Specific Customer Churn Patterns using DM Techniques. *In IEEE International Conference on Data Analytics for Business and Industry (ICDABI)*, 23-28.

[10] Khalid, L.F., Abdulazeez, A.M., Zeebaree, D.Q., Ahmed, F.Y., & Zebari, D.A. (2021). Customer churn prediction in telecommunications industry based on data mining. *In IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, 1-6.

[11] Kolli, N., & Balakrishnan, N. (2020). Hybrid features for churn prediction in mobile telecom networks with data constraints. *In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 734-741.

[12] Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, *10*(2), 1659-1665.
[13] Melgarejo Galvan, A.R., & Clavo Navarro, K.R. (2017). Big data architecture for predicting churn risk in mobile phone companies. *In Information Management and Big Data: Second Annual International Symposium, SIMBig 2015, Cusco, Peru, and Third Annual International Symposium, SIMBig 2016, Cusco, Peru, 2016, Revised Selected Papers 2*, 120-132. Springer International Publishing.
[14] Mishra, K., & Rani, R. (2017). Churn prediction in telecommunication using machine learning. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2252-2257.
[15] Mitkees, I.M., Badr, S.M., & ElSeddawy, A.I.B. (2017). Customer churn prediction model using data mining techniques. *In IEEE 13th International Computer Engineering Conference (ICENCO)*, 262-268.
[16] Mohamed, F.A., & Al-Khalifa, A.K. (2023). A Review of Machine Learning Methods For Predicting Churn in the Telecom Sector. *In IEEE International Conference On Cyber Management And Engineering (CyMaEn)*, 164-170.
[17] Muthupriya, V., Narayanan, R., Nakeeb, S., & Abhishek, A. (2022). Customer churn analysis using XGBoosted decision trees. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*.
[18] Prakash, U., Anila, A., Swetha, C., Vigneshwaran, K., & Kavinayaa, N. (2022). A Survey on Artificial Intelligence in Telecommunication for Churn Prediction. *In IEEE 6th International Conference on Electronics, Communication and Aerospace Technology*, 1261-1265.
[19] Qin, C., Wang, L., Ma, Q., Yin, Y., Wang, H., & Fu, Y. (2021). Contradictory structure learning for semi-supervised domain adaptation. *In Proceedings of the SIAM International Conference on Data Mining (SDM)*, 576-584. Society for Industrial and Applied Mathematics.
[20] Quasim, M.T., Sulaiman, A., Shaikh, A., & Younus, M. (2022). Blockchain in churn prediction based telecommunication system on climatic weather application. *Sustainable Computing: Informatics and Systems*, *35*.
[21] Ramadhanti, D., Larasati, A., Muid, A., & Mohamad, E. (2023). Building customer churn prediction models in Indonesian telecommunication company using decision tree algorithm. In *AIP Conference Proceedings*, 2654(1). AIP Publishing LLC.
[22] Shabankareh, M.J., Shabankareh, M.A., Nazarian, A., Ranjbaran, A., & Seyyedamiri, N. (2022). A stacking-based data mining solution to customer churn prediction. *Journal of Relationship Marketing*, *21*(2), 124-147.
[23] Shen, Y., Peng, M., Wu, Q., & Li, R. (2023). A machine learning method to variable classification in OpenMP. *Future Generation Computer Systems*, *140*, 67-78.
[24] Shobana, J., Gangadhar, C., Arora, R.K., Renjith, P.N., Bamini, J., & devidas Chincholkar, Y. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, *27*.
[25] Wanchai, P. (2017). Customer churn analysis: A case study on the telecommunication industry of Thailand. *In IEEE 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, 325-331.
[26] Yulianti, Y., & Saifudin, A. (2020). Sequential feature selection in customer churn prediction based on Naive Bayes. *In IOP Conference Series: Materials Science and Engineering*, 879(1), 1-7. IOP Publishing.
[27] Zhang, T., Moro, S., & Ramos, R.F. (2022). A data-driven approach to improve customer churn prediction based on telecom customer segmentation. *Future Internet*, *14*(3), 1-19.
[28] Zhao, Y., Hu, N., Zhang, C., & Cheng, X. (2020). DCG: A Client-side Protection Method for DNS Cache. *Journal of Internet Services and Information Security*, *10*(2), 103-121.

Authors Biography



Mirko Bruno Vela López
Professional with a degree in Systems and Computer Engineering from the Universidad Católica Santo Toribio de Mogrovejo. Specialized in IT auditing, IT risk management, SAP security, information security, ITGC and SOX controls.

Maria Ysabel Arangurí García
Mg. in Strategic Management of Information Technologies. Systems Engineer. Full-time professor at Universidad Católica Santo Toribio de Mogrovejo (Peru).

Jessie Leila Bravo Jaico
PhD in Computer Science and Systems. Master in Computer Science and Multimedia from Universidad de Los Lagos - Chile. Master in Business Administration with mention in Business Management from the Universidad Nacional Pedro Ruiz Gallo.

Ángel Antonio Ruiz-Pico
Dr. in Civil Engineering from the School of Civil Engineering of the University of A Coruña (Cum Laude) Degree in Geological Sciences. Faculty of Geological Sciences. Universities of Zaragoza and Granada. Director School of Environmental Civil Engineering, Universidad Católica Santo Toribio de Mogrovejo (Peru)

Ronald M. Hernández
Bachelor's Degree in Psychology. Master in Education. Qualified researcher in RENACYT (Concytec). Couator of the book Publishing in scientific journals. Has published scientific articles in international indexed journals in Scopus, Web of Science and SciELO