Articles

Machine Learning Models Based on Geospatial Data for Customer Churn Prediction in ISP

Modelos de Aprendizaje Automático Basados en Datos Geoespaciales para la Predicción de Fuga de Clientes en Proveedores de Servicios de Internet (ISP)

Gabriela Solano Aguilar*
Néstor Estrada Brito *

REVISTA TECNOLÓGICA
ciencia y educación
Edwards Deming

ISSN: 2600-5867

Edited by: Tecnológico Superior Corporativo Edwards Deming

July – December Vol. 9 - 2 - 2025

https://revista-edwardsdeming.com/index.php/es

e-ISSN: 2576-0971

Received: March 29, 2025

Approved: May 1, 2025

Page 1-32

ABSTRACT

The purpose of this study is to address the problem of customer churn in the telecommunications sector, specifically in an Internet Service Provider (ISP) operating in Ecuador, through a machine learning model and the application of geospatial analysis. To carry it out, it was necessary to apply class balancing techniques, since the dataset presented a smaller number of customers in abandonment, in contrast to those who remain in active service. In addition, customer segmentation was carried out using K-Means. In terms of churn prediction, several classification models were evaluated and the one with the best performance was selected. Based on their predictions, geospatial analysis was applied to examine the territorial distribution of customers, identify patterns and enable the development of more effective retention strategies. To evaluate each model, ROC-AUC and recall metrics were applied, where Random Forest in combination with Random Downsampling presented the best performance. While, through segmentation, four groups of customers with similar characteristics were identified. In conclusion, this study demonstrates that

*Magíster en Ingeniería de Sistemas, Docente de la Escuela Superior Politécnica de Chimborazo - ESPOCH, Riobamba, Ecuador
gabriela.solano@espoch.edu.ec, https://orcid.org/0009-0007-7565-4702
*Magíster en Ingeniería Electrónica en Telecomunicaciones y Redes, Docente en la Escuela Superior Politécnica de Chimborazo- ESPOCH, Riobamba, Ecuador.
nestor.estrada@espoch.edu.ec, https://orcid.org/0000-0002-4100-7351

Articles
Revista tecnológica ciencia y educación Edwards Deming
ISSN: 2600-5867

integrating machine learning and geospatial analysis is an effective combination for predicting customer churn in the telecommunications sector. The combination of Random Forest with data balancing techniques, together with customer segmentation using K-Means, resulted in a robust and accurate model.

Keywords: Machine learning, geospatial analysis, customer churn, Random Forest, K-Means

RESUMEN

El propósito de este estudio es abordar el problema de la pérdida de clientes en el sector de las telecomunicaciones, específicamente en un Proveedor de Servicios de Internet (ISP) que opera en Ecuador, mediante un modelo de aprendizaje automático y la aplicación de análisis geoespacial. Para llevarlo a cabo, fue necesario aplicar técnicas de balanceo de clases, ya que el conjunto de datos presentaba una menor cantidad de clientes que abandonaban el servicio, en contraste con aquellos que permanecían activos. Además, se realizó una segmentación de clientes utilizando K-Means. En cuanto a la predicción de abandono, se evaluaron varios modelos de clasificación y se seleccionó el de mejor rendimiento. Con base en sus predicciones, se aplicó análisis geoespacial para examinar la distribución territorial de los clientes, identificar patrones y permitir el desarrollo de estrategias de retención más efectivas. Para evaluar cada modelo se utilizaron las métricas ROC-AUC y recall, siendo Random Forest en combinación con Random Downsampling el que presentó el mejor desempeño. A través de la segmentación, se identificaron cuatro grupos de clientes con características similares. En conclusión, este estudio demuestra que integrar el aprendizaje automático con el análisis geoespacial es una combinación eficaz para predecir la pérdida de clientes en el sector de las telecomunicaciones. La combinación de Random Forest con técnicas de balanceo de datos, junto con la

segmentación de clientes utilizando K-Means, dio como resultado un modelo robusto y preciso.

Palabras clave: Aprendizaje automático, análisis geoespacial, pérdida de clientes, Random Forest, K-Means

INTRODUCTION

According to *Statista*, the number of internet users has grown from 1.023 billion in 2005 to more than 5.5 billion projected for 2024. While this increase has fostered global connectivity, access to information, and digital transformation, it has also generated significant challenges at a global level. These include increasing demands for quality of service, unequal access, and pressure on Internet Service Providers (ISPs) to remain competitive in a saturated market.

In Ecuador, the increase in Internet use has also been notable. According to the Ministry of Telecommunications report published in 2024, the percentage of rural parishes and cantonal capitals with access to fixed internet via fiber optics grew from 75.82 % in 2022 to 80.80 % in the first quarter of 2024 (Aguilar Cazar Miguel, 2024). However, the Telecommunications Regulation and Control Agency (Arcotel) points out that the most frequent complaints related to internet service include communication quality problems and intermittency, poor coverage, charges different from those agreed, service interruptions, and the obligation to stay with the provider (Arcotel). This problem deteriorates the customer experience and weakens trust in ISPs, which could cause the customer to abandon the company and switch service providers (Pejić Bach et al., 2021).

Customer churn or abandonment understood as the unilateral severance of the contractual relationship with a provider (Cenggoro et al., 2021), is a crucial challenge for ISPs, whose regular subscription base constitutes their business model (Adhikary & Gupta, 2021; Cenggoro et al., 2021; Pejić Bach et al., 2021). In a competitive market, in which obtaining new customers costs more than retaining existing ones (Khoh et al., 2023), ISPs need to detect when a customer is at risk of churn, in addition to analyzing what impacted them to decide to leave. This is to implement strategies to make the customer stay with the company.

Traditionally, retention strategies do not normally embrace user behavior analysis, nor geographical factors that could motivate the decision to desert the service. Faced with these difficulties, this study presents a novel solution founded on two inherent vectors: machine learning and geospatial analysis, whose complementarity allows the creation of an exhaustive solution to prevent customer defection in ISPs. Both machine learning and geospatial analysis are accompanied by probability theory and

statistics, which provide the mathematical basis for model development that enables strong and accurate prediction.

Machine learning focuses on identifying patterns that can be difficult and complex for humans to find through large amounts of data. Then, through unsupervised learning, such as the clustering technique, customers can be segmented into common characteristics, which provides useful information about their behavior (Pejić Bach et al., 2021), on the other hand, supervised learning allows for addressing binary classification problems, such as the one present in this study, i.e. whether or not a customer will abandon the ISP (Usman-Hamza, Balogun, Amosa, et al., 2024).

Geospatial analysis, on the other hand, provides a strategic facet to the study, as it explores how geographic factors affect the client's decision, which in turn allows identifying patterns related to abandonment and locating key areas that need special attention (Peng et al., 2024).

To address this challenge, the study proposes as research objectives, first, the preprocessing of the data provided by the ISP, to improve the quality of these data. Secondly, the application of clustering techniques allows for segmenting customers into groups with similar characteristics. Thirdly, the evaluation of different classification models to select the most appropriate one according to their performance, taking into account different metrics and also using cross-validation. Finally, the interpretation of the results of the selected classification model, through geospatial analysis, to identify spatial patterns that are related to client attrition.

Several studies have addressed methods to predict customer churn in the telecommunications sector, for example; (Pejić Bach et al., 2021) developed an approach based on cluster analysis and decision trees, which allowed them to perform market segmentation and also predict customer churn. (Geiler et al., 2022) took eight supervised learning methods, which they combined with sampling techniques in the presence of data imbalance associated with churn. Other research highlights the use of geospatial analysis to understand how customers behave. (Peng et al., 2024) demonstrated how the use of big geo-data, such as mobile signaling data and Points of Interest (POI), can extract representative spatial patterns in customer flow.

The present study focuses on an ISP located in Ecuador, which has activities in the provinces of Chimborazo and Tungurahua. The ISP has a lack of systems for identifying and predicting clients in abandonment. As for the data provided by this company, they correspond to residential customers with fiber optic plans, from them, this study addresses the problem of churn with a technological and institutional approach, in addition to applying machine learning and geospatial analysis to understand the factors related to churn and predict when customers are in danger of falling into it.

This study is oriented to answer the research question: How can a predictive model that integrates machine learning and geospatial analysis predict customer churn in an

ISP operating in Ecuador?

MATERIALS AND METHODS

Customer Churn detection

There are two ways in which the problem of customer defection can be addressed: reactive or proactive. The first is when the company acts after the customer has expressed a desire to cancel the service. On the other hand, in the proactive approach, the company seeks to anticipate abandonment through predictive techniques (Lalwani et al., 2022). This approach has been gaining relevance, as it allows the company to identify customers at risk of churn in time and to apply retention techniques to them.

Within the proactive paradigm, customer defection in the telecommunications sector has been extensively addressed with supervised classification models. Decision trees are among the models presented because they allow for the interpretation of interaction between various characteristics. (Bachan & Gaber, 2021) experimented with: Decision Tree, Logistic Regression, and Support Vector Machine (SVM). They used data from a Trinidad and Tobago ISP and compared that Decision Tree performed the best.

(Adhikary & Gupta, 2021) developed tests, applying more than 100 classifiers, and concluded that the Regularized Random Forest presented superior accuracy, on the other hand, the Bagging Random Forest obtained the best AUC-ROC in unbalanced data sets.

(Colot et al., 2021) proposed a variant of Random Forets, called Essence Random Forest, which was designed to reduce redundancy in the data, which improves the stability of the model. This study was applied in a European telecommunications company and showed that using geospatial and mobility data, in combination with the model proposed, improves the prediction of abandonment.

(Krishna et al., 2024) studied several machine learning techniques for churn prediction, and based on their results, they found that Random Forest presented higher accuracy. Class imbalance was also addressed in this study, where they applied the Synthetic Minority Oversampling Technique (SMOTE), which improved the performance of the model and consolidated Random Forest as the best option for predicting customer churn.

(Wagh et al., 2024) also conducted a similar study in which they concluded that Random Forest performed better compared to other models employed. Similarly, for handling class imbalance, they utilized techniques such as SMOTE and Edited Nearest Neighbors (ENN), which helped improve the identification of clients who were prone

to dropping out. Their research is also another example of using data-balancing techniques in ensemble models to improve the accuracy of prediction.

One such prominent study was performed by (Ortakci & Seker, 2024) who formulated a model which not only predicts defection but also recommends customized service rates to optimize customer retention. In their paper, they compared different classification models like K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and SVM. Comparing them on parameters like accuracy and AUC, they found that Random Forest outperformed the other models.

(Poudel et al., 2024) compared a few machine learning algorithms, wherein the Gradient Boosting Machine (GBM) was the most effective. Model interpretation was also emphasized by them, with Shapley Additive Explanations (SHAP), which analyzes how the different variables contribute to customer churn prediction, and in their paper, they highlighted call time, plan type, and contract length.

Another of the most widely used models in the classification of customer churn is XGBoost, since it can adequately deal with unbalanced data sets and also optimizes computational performance.

XGBoost was applied in the study presented by (Shrestha & Shakya, 2022), whose data set corresponds to a telecommunications company in Nepal. This model outperformed previous studies performed on public data in both accuracy and F1-score. Similarly, (Lalwani et al., 2022) compared several machine learning algorithms, XGBoost and AdaBoost stood out by obtaining the highest AUC value.

The study by (Chong et al., 2023) was conducted on a telecommunication firm in California, where XGBoost outperformed along with accuracy, recall, and F1-score, compared to other models such as: KNN, Random Forest, AdaBoost, Logistic Regression, and SVM.. Similarly, (Alteer & Alariyibi, 2024) applied XGBoost using the historical data of an ISP in Libya. They determined that XGBoost works optimally in real-world scenarios that involve imbalanced datasets where methodologies such as SMOTE and Smoteenn (Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors) could be applied in order to make class balance improvements. Their performance in AUC was superior when compared to alternative algorithms such as Random Forest and Gradient Boosting.

In turn, (Ouf et al., 2024) proposed a hybrid approach based on Smoteenn and XGBoost, which improved the accuracy of class imbalance. They also experimented with several data sets, concluding that the combination of these techniques improves the model in terms of accuracy, precision, recall, f1-score, and AUC-ROC.

Other studies have integrated models with esemble or hybrid schemes, thereby combining the strengths of different algorithms. For example, (Xu et al., 2021) proposed a system involving stacking and soft voting, in which XGBoost, Logistic

Regression, Decision Tree, and Naïve Bayes were combined as a two-level model. Their study showed that clustering of different features as well as ensemble techniques improve the accuracy of predictions.

In (Wahul et al., 2023) they found that ensemble models can outperform individual classifiers in metrics such as AUC, precision, and recall, for which the following were tested: Stochastic Gradient Boost, Random Forest, Gradient Boosting, and AdaBoost.

Another study that proposes an ensemble approach is that of (Khoh et al., 2023), which was combined with an optimized weighting technique. They also used SMOTE to address class imbalance and Powell's Optimization Algorithm to determine the weights that influence the base classifiers that are part of the ensemble. The authors conclude that this method was able to achieve better results, outperforming even deep learning models.

(Soleiman-garmabaki & Rezvani, 2024) studied the effect of combining individual classifiers using ensemble procedures such as AdaBoost and XGBoost focused on improving dropout prediction. In this regard, they experimented with six base models, including neural networks, Logistic Regression, Decision Tree, and Random Forest, before and after data balancing. In their results, the authors indicate that the best model was obtained by combining logistic regression, neural networks, and AdaBoost, with better performance in AUC and accuracy.

In the study by (Usman-Hamza, Balogun, Nasiru, et al., 2024) they took datasets from Kaggle and Machine Learning Repository (UCI), where they experimented with thirteen classifiers. The authors found that the tree-based models outperformed other traditional algorithms in accuracy and stability. On the other hand, the study highlighted the negative impact of class imbalance on model performance, so they applied SMOTE, which together with homogeneous ensemble methods improved the effectiveness of the models. Hybrid classifiers such as SysFor, CS-Forest, and ADTree showed resistance to data quality.

(Cenggoro et al., 2021) investigated the use of Deep Learning with vector embedding models to predict attrition in telecommunications. This model generated representations that discriminate customers who have a high probability of canceling from those who might remain in service.

(Saha et al., 2023) went beyond the application of traditional machine learning techniques, as they also used deep learning, including models such as Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Random Forest, XGBoost, and AdaBoost, on two customer churn data sets. According to their results, CNN and ANN are presented as the best performers. On the other hand, the study also highlights that data imbalance negatively affects the performance of the models and that, although techniques such as SMOTE can help with the problem, if the proportion of defective customers is very low, the impact can still be significant.

Customer segmentation and churn detection

There is also research that has explored the use of clustering techniques to segment customers based on their characteristics and behavioral patterns. Segmentation allows a better understanding of user-profiles and their interaction with the service, which can be useful both for retention strategies and for improving predictive models of attrition.

K-Means is one of the clustering algorithms used in recent studies related to customer defection prediction. An example is the work of (Pejić Bach et al., 2021), who propose a three-stage hybrid approach. In the first, they prepare the data. Then, they apply K-Means to segment the market and, finally, they use CHAID decision trees to predict attrition in each group. With the help of this analysis, they identify the group with the highest dropout rate. As for segmentation, it not only helps to identify the customers most likely to drop out of the service but also to analyze their behavior and thereby improve the accuracy of the predictions.

(Zhao et al., 2023) and (Xu et al., 2023) also used K-Means for customer segmentation. (Xu et al., 2023) integrated the RFM (Recency, Frequency, Monetary) model, as it allows enriching the segmentation analysis before the application of XGBoost for churn classification. This approach combined segmentation and key feature extraction, which allowed optimizing data quality before training the predictive model.

K-Medoids are also present as an alternative to K-Means since it is presented as a more robust option to outliers. In fact, (Bilal et al., 2022) found that combining K-Medoids with Gradient Boosted Trees, Decision Tree, and Deep Learning, presents better accuracy compared to traditional methods. (Liu et al., 2022) carried out a study in which they applied the same combination of models and confirmed that this integration obtains a better performance. However, they also point out that it has a higher computational cost.

On the other hand, (Fatima et al., 2024) applied SMOTE-NC and SMOTE-ENC before the training of classification models to address the class imbalance problem. In this study, Random Forest stands out together with SMOTE-NC with higher accuracy metrics in terms of attrition prediction. After classification, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was implemented for those clients that were detected as dropouts. This allows segmenting customers and understanding their behavior, which provides valuable information for the development of personalized retention strategies.

Alternatively, existen enfoques más avanzados en los cuales se aplica aprendizaje profundo para realizar la segmentación de clientes. El estudio realizado por (Geiler et al., 2022) es un ejemplo de ello, en donde se aplicaron distintas alterativas para el balanceo de clases, como: SMOTE, Adasyn, Random Undersampling, Neighborhood

Cleaning Rule (NCR) y Tomek Links, para mejorar el rendimiento de los modelos de clasificación en la predicción de la deserción. Los autores en su estudio emplearon ocho modelos supervisados, entre ellos: Random Forest, XGBoost, SVM y Decision Trees, además de enfoques ensamble, los cuales resultaron ser más efectivos en términos de la métrica AUC. Para la segmentación de clientes usaron Deep Autoencoder-based Clustering, el cual permitió mejorar la precisión del perfil de los clientes, así como una mejor interpretación de los patrones de deserción.

Geospatial analysis

Geospatial analysis is a tool that allows an analysis to be performed in such a way that location-related characteristics relate to and determine certain phenomena. It adds an extra dimension when dealing with certain behaviors or patterns in different contexts. In this framework, the prediction of customer defection has a high potential, since aspects such as the geographical environment, local infrastructure, the presence of competitors, or the socioeconomic characteristics of an area, can influence the abandonment of services.

Even though, up to the date of this study, there have not been numerous investigations that incorporate dropout prediction models and geospatial analysis, the work of (Colot et al., 2021) highlights how geospatial data, in combination with information about the use of mobile services and web browsing, seems to increase the predictive capacity of dropout in telecommunications. This study highlights how geographic features can complement other sources of information.

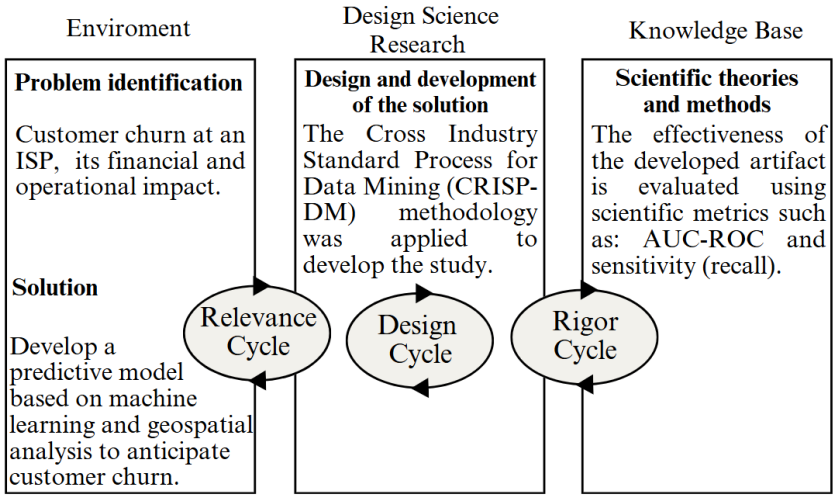
In other sectors, such as e-commerce, the authors (Matuszelański & Kopczewska, 2022) also show the value of geospatial data, as they analyzed socio-geo-demographic characteristics to determine factors associated with attrition and revealed how geographic variables can be predictive in different contexts. While this case does not pertain to the telecommunications sector, it confirms the contribution of geospatial analysis as a promising tool for modeling customer behavior.

In sectors outside the telecommunications sector, geospatial analysis has also been used to identify spatial patterns. An example of this is the research (Peng et al., 2024), which studied road density, as well as transport access, and the location of points of interest (POIs) that influence customer flow.

Through this literature review, several papers have addressed customer defection prediction in telecommunications using supervised classification models, some of which have chosen to integrate segmentation techniques. However, this study proposes a global approach, where in addition to predicting churn and segmenting customers, geospatial analysis is considered to identify spatial patterns associated with churn. This approach has not been addressed in previous studies, and is of great help, as it provides valuable information for the development of retention strategies and decision-making.

This study formulates a design science methodology based on the principles of design science, which consists of three interconnected phases: the relevance cycle, the design cycle, and the rigor cycle, as illustrated in **Error! Reference source not found.**. These are core to solving the problem under study and ensuring findings are scientifically valid and practically meaningful.

Figure 1: Design Science Methodology



Relevance cycle

A central problem in ISPs is client attrition and its financial and operational impact, which is the starting point for this study. To address the problem, we worked through a systematic literature review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Prisma) methodology, which provides a rigorous and transparent framework for the selection of studies, to have a scientifically backed knowledge base.

The implementation of the Prisma process in this study results in an exhaustive search process, developed in recognized academic databases such as Scopus and SpringerLink, from which publications are collected for the period 2021-2024, prioritizing those studies that are related to the telecommunications sector, where machine learning is applied in customer segmentation and attrition prediction, as well as geospatial analysis approaches.

Articles that were retracted or unrelated to the study were also excluded. Relevant algorithms, methodologies, and applications, as well as metrics, forms of validation, and analytical tools, were identified.

Research Design

Articles
Revista tecnológica ciencia y educación Edwards Deming
ISSN: 2600-5867

The present research follows a quantitative, correlational-predictive approach and non-experimental design, where the aim is to predict customer defection using historical and geo-referenced data.

Population and sample

Population: ISP residential customers operating in the provinces of Chimborazo and Tungurahua, Ecuador, with fiber optic plans.

Sample: Sub-set of active and defective clients selected through non-probability convenience sampling based on historical records provided by the ISP.

Inclusion criteria: Client records provided by the ISP, including both loyal and defective clients, were selected through non-probability sampling.

Exclusion criteria: Client records that are incomplete or contain erroneous information received by the ISP.

Collection instruments

All data used in this study were provided by the ISP through its internal database

Design cycle

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology has been used for the development of this study. This standard is widely used in the structuring of projects related to data analysis, given that it presents an iterative and systematic process, which allows the objectives of this study to be addressed in an organized manner. The main phases of CRISP-DM adapted to this environment are the following:

- **Understanding the problem:** This is where the objectives of the study are established, considering the needs of the ISP. The primary concern identified is customer attrition and therefore the requirement for a model that will leverage geospatial and customer data. The solution tries to find significant patterns and characteristics that differentiate customers who stick around from ones that leave.
- **Understanding the data:** Exploratory data analysis of data provided by the ISP, identification of key variables, and verification of data quality is conducted.
- **Data preparation:** The process of cleaning and transforming features. In addition, this phase addresses class imbalance, a recurring situation in customer defection prediction, exclusively in the classification task. To address the imbalance, three main strategies are applied: class weighting, by adjusting the weights in the classification algorithms through class balancing in Scikit-learn (Chong et al., 2023); downsampling, by reducing the amount of majority class

Articles
Revista tecnológica ciencia y educación Edwards Deming
ISSN: 2600-5867

samples through methods such as Random Downsampling, NCR and Tomek Links (Geiler et al., 2022); oversampling, increasing the representation of the minority class with techniques such as SMOTE (Agasti & Satpathy, 2024; Geiler et al., 2022; Khoh et al., 2023; Krishna et al., 2024; Soleiman-garmabaki & Rezvani, 2024; Usman-Hamza, Balogun, Amosa, et al., 2024; Usman-Hamza, Balogun, Nasiru, et al., 2024; Wagh et al., 2024) and Adasyn (Geiler et al., 2022); and combinations between oversampling and downsampling.

- **Modelling:** This is where predictive models are selected and trained. Techniques include:

  Segmentation: K-Means is applied because it has been the most mentioned in the literature review (Pejić Bach et al., 2021; Xu et al., 2023; Zhao et al., 2023) and also because it is a popular technique for grouping customers with common characteristics. This algorithm allows for the discovery of behavioral patterns and is widely used in marketing research, specifically for market segmentation in telecommunications (Pejić Bach et al., 2021).

  Classification: Models based on decision trees and ensemble techniques, which are highly effective in the literature reviewed, are employed. Decision Trees (Bachan & Gaber, 2021; Pejić Bach et al., 2021) are included for their interpretability and ability to handle heterogeneous data. In a decision tree, the impurity of a node can be measured through the entropy present in Equation (1), where $H(S)$ represents the entropy of a set S, c is the total number of classes in the dataset and p_i is the portion of observations belonging to class i.

  $$H(S) = -\sum_{i=1}^{c} p_i \, log_2 \, (p_i) \qquad (1)$$

The Random Forest model (Adhikary & Gupta, 2021; Colot et al., 2021; Krishna et al., 2024; Ortakci & Seker, 2024; Wagh et al., 2024) performs generalization via ensemble of many decision trees. The Random Forest prediction is obtained by majority voting as in Equation (2), where $\hat{y}$ is the resulting predicted class, while $\hat{y}_j$ is the prediction of *j*-th tree in the collection of T trees.

$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_T) \qquad (2)$$

Gradient Boosted Trees models (Poudel et al., 2024) s are responsible for optimizing learning sequentially, which reduces the error at each iteration. To update the model at each iteration called m, Equation (3) is used. In it, $F_m(x)$ represents a new prediction, $F_{m-1}(x)$ likewise a prediction, but from the previous iteration, $h_m(x)$ is the decision tree adjusted in the current iteration, and finally γ is the learning coefficient, which regulates the influence of the new trees.

$$F_m(x) = F_{m-1}(x) + \gamma h_m(x) \qquad (3)$$

On the other hand, XGBoost (Alteer & Alariyibi, 2024; Chong et al., 2023; Lalwani et al., 2022; Ouf et al., 2024; Shrestha & Shakya, 2022) is an advanced boosting method widely used due to its high performance in classification tasks. The gain function presented in Equation (4), is used to evaluate the quality of a split within the tree. In this equation G_L and G_R represent the sum of the gradient values at the left and right child nodes, respectively; H_L y H_R correspond to the sum of the second derivatives of the loss function at the same nodes; λ is a regularization parameter that controls the complexity of the model; and γ is the penalty term for the addition of new nodes in the tree.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \qquad (4)$$

In order to ensure that the best performance metrics were obtained, stratified cross-validation (StratifiedKFold) was applied, which ensures a balanced number of classes in each fold (Geiler et al., 2022). In addition to this, hyperparameter optimization (model-dependent) was obtained via GridSearchCV, which allowed us to search for the best combination of configurations for each model and maximize its predictive capability.

Model evaluation: The elbow method is employed to identify the optimal number of groups for K-Means by analyzing the variation in inertia (sum of distances to the centroid) as the number of groups or clusters increases (Xu et al., 2023). On the other hand, the Silhouette index measures the internal coherence in each cluster and the separation from other clusters.

In the case of classification models, having a dataset with unbalanced classes, the AUC-ROC metric has been chosen as it provides an aggregate measure of model performance at all possible decision thresholds; this metric, derived from the ROC curve, assesses the model's ability to distinguish between defecting and non-defaulting customers by measuring the ratio between the rate of true positives and false positives. Its usefulness is particularly prominent in unbalanced datasets, where the number of positive cases (churned customers) is significantly lower than the number of negative cases (Geiler et al., 2022; Krishna et al., 2024; Lalwani et al., 2022).

However, since the AUC-ROC metric does not allow us to observe in detail how well the model can identify at-risk clients, we incorporate an additional metric, Recall (sensitivity), which measures the proportion of clients labeled as dropouts who were correctly identified, thus prioritizing the reduction of false negatives.

Deployment: An interactive map is proposed to visualize the results of the final model. This involves the geospatial variables of the customers and shows both the customer segments and the geospatial patterns associated with defection.

Validation and Reproducibility

Cross-validation: In the case of classification algorithms, a 5-partition stratified cross-validation scheme (StratifiedKFold) is applied, ensuring that the class distribution is maintained in each partition equally, ensuring a strong model evaluation, particularly in an imbalanced class dataset.

Comparison of results: K-Means was used for segmentation, and two tests were performed one without geospatial data such as latitude and longitude, and one with them. This was to analyze which groups presented better metrics and thus better segmentation.

As for the classification models, the performance is compared with the results obtained by the different algorithms tested in this study to select those that offer the best performance.

Reproducibility: Since the project is getting data and information from an ISP, where there is sensitive information, the code of the models developed will not be open-sourced. However, the development process is described in this report, and reproduction can be done in similar datasets or other business environments.

Design Limitations

Limited data: Lack of demographic characteristics can restrict analysis of other potential factors that could be relevant.

Geographic bias: The results are determined by characteristics present in Ecuador, specifically in provinces Chimborazo and Tungurahua, and therefore if the research is to be generalized across varying contexts, then some modification will have to be done accordingly.

Technological dependence: When deploying the models, a scalable technological infrastructure may be required at the ISP.

Rigour cycle

The scientific rigor of the study is based on a Prisma review, which defines a robust theoretical framework, validation, and reproducibility methods documenting all phases of the process that will allow replicability along the same lines, an assessment of limitations and potential for generalization while facilitating continuous improvement.

RESULTS

A cleaning and validation process was applied to the data provided by the ISP to ensure the quality of the data. On the one hand, duplicate records were eliminated,

while null records were corrected by requesting information from the ISP. Likewise, an outlier analysis was carried out, which allows for the purging of inconsistent records present in the data capture.

Description of the dataset

As a result of the first data processing step, a cleaned and formatted dataset was obtained, comprising customer information regarding residential fiber optic plans, faulty and active.

Error! Reference source not found. has the dataset structure as 9653 instances with 11 attributes, 8 of them numeric and 3 nominal. Class imbalance can be noticed as dropouts are only 7.32% of customers. This distribution of imbalance should be considered, since it can affect the performance of the churn prediction model. Another solution to this issue is at the data preparation stage, where class balancing techniques have been used, whose impact will be discussed later.

Table 1: Summary of the dataset

# Instances	# Characteristics	# C. Num.	# C. Cat.	$\frac{Churn}{Not\ churn}$
9653	11	8	3	0.073

On the other hand, **Error! Reference source not found.** presents each of the variables included in the dataset. These characteristics contain information about the contracted service, as well as the history of customer interactions with the ISP.

Table 2: Description of the Variables contained in the Dataset

Variable	Description
Product_name	Type of internet plan contracted by the customer.
Product_speed	Data transmission capacity measured in megabits

	per second.
Product_price	Monthly cost in dollars of the service contracted by the customer.
Geo_latitud	Geographic position of the customer with respect to the horizontal axis (latitude).
Geo_longitud	Geographic position of the customer with respect to the vertical axis (longitude).
Nodo_name	Primary network equipment to which the customer is connected.
Total_invoices_loyalty	Number of months that the customer keeps their service active.
Compliance_index	Relación entre pagos realizados a tiempo y la cantidad total de pagos.
Tickets_resolved	Ratio of payments made on time to the total amount of payments.
Average_hours_tickets_resolved	Average time in hours to resolve critical tickets.
Churn	Indicates whether the customer has abandoned (1) or not (0).

Exploratory data analysis

In any study it is vital to carry out an exploratory analysis, given that in this way it is possible to get to know the data set better in terms of its structure, to find certain

Articles
Revista tecnológica ciencia y educación Edwards Deming
ISSN: 2600-5867

patterns, or to identify anomalies. It is at this point that an interesting factor was discovered when applying correlation analysis between numerical variables.

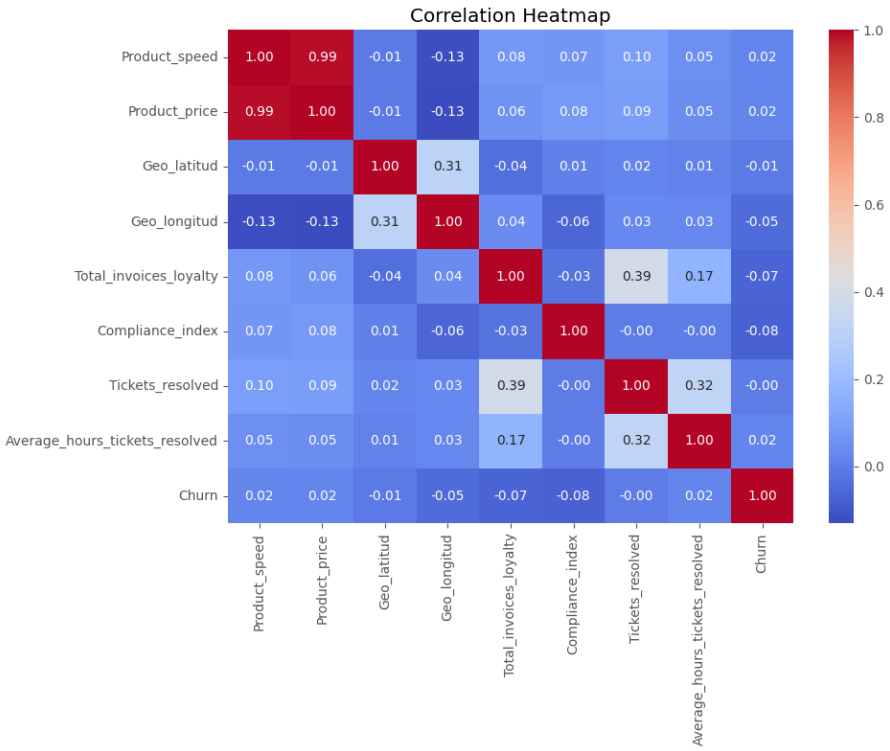
As shown in **Error! Reference source not found.**, the correlation analysis yielded a very high value (0.99) among the variables: Product_speed and Product_price. This shows high dependence among them, and hence it was agreed to exclude Product speed from analysis, so that collinearity problems could be avoided. On the other hand, no high correlations between the predictor variables and the target variable (Churn) existed, indicating that the utilization of machine learning models is inevitable to detect non-linear relationships.

Customer segmentation

In business management, customer segmentation is key to identifying behavioral patterns and thereby optimizing retention strategies. Understanding how customers interact with the service and deciphering the reasons that influence their retention, or churn improves business decision-making in the telecommunications sector. Designing marketing campaigns that are better targeted through customer segmentation allows them to be effective, as well as personalizing offers and anticipating potential churn, which contributes to customer satisfaction and thus to business sustainability.

In this study, the variables below were selected for segmentation: customer dwell time, payment fulfillment rate, number of critical support tickets and average time in hours for the resolution of those tickets, and geographic location.

Figure 2: Correlation heat map between quantitative variables



These variables were chosen for their relevance in characterizing customer behavior, in contrast to other variables that only describe the characteristics of the contracted service.

Two tests were carried out to obtain information on whether geographic coordinates influence customer segmentation. In the first one, these variables were incorporated and in the second one they were omitted, and K-Means was used, where k was given values from 2 to 10. The results showed that the test carried out without the geographical characteristics obtained a higher Silhouette Score, which means that there is greater separation between groups and that the elements within each of them are more similar.

It is also important to note that the elbow method was applied, without including the geospatial variables, to obtain the optimal number of clusters. **Error! Reference source not found.** shows the result of this analysis, where it can be observed that the inertia decreases progressively as the number of clusters increases, but a reduction can also be noted up to K=4, indicating that this number represents a good balance between explained variability and model complexity. The Silhouette Score confirmed that K=4 obtained the highest value (0.3406), for this ratio 4 clusters were selected.

To better understand the behavior of customers in each cluster, an analysis was carried out, according to the average of each characteristic used to perform the

segmentation. **Error! Reference source not found.** presents these values, of which clusters 0 and 2 are made up of customers with the longest time with the ISP (23 months), with a high rate of on-time payment.

On the other hand, clusters 1 and 3 group together customers with shorter tenure and lower payment compliance. In particular, cluster 3 stands out for having the lowest tenure and, in turn, a low number of support tickets (0.194 on average).

Figure 3: Elbow method for determining the optimal number of clusters

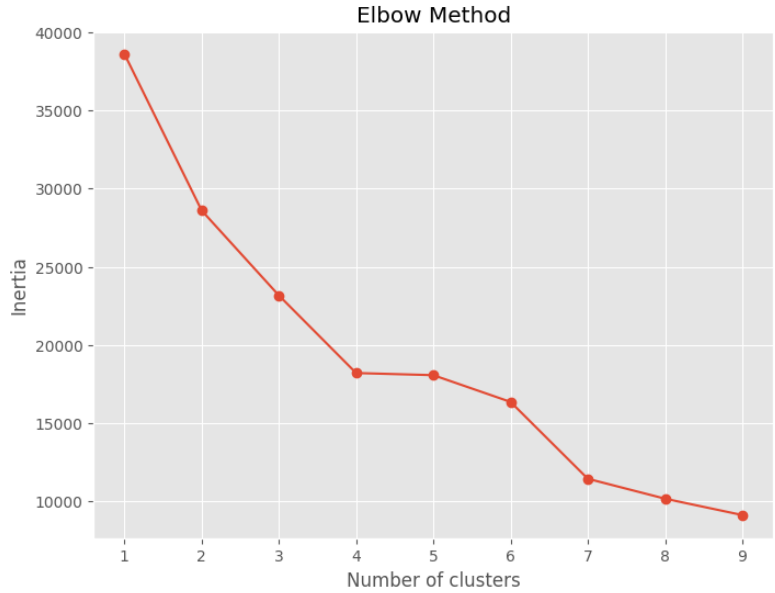


Table 3: Average number of relevant characteristics in each customer cluster

Cluster	Total_invoices_loyalty	Total_invoices_loyalty	Tickets – resolved	Average_hours_tickets_resolved
0	23.203	0.881	0.681	7.590
1	15.034	0.449	0.464	5.912
2	23.090	0.845	2.943	66.646
3	7.589	0.928	0.194	2.860

In addition, an analysis of attrition distribution for each cluster was conducted, which indicates that Cluster 1 has higher attrition. This suggests that clients with lower

payment compliance and shorter dwell time have a higher propensity to drop out of the service.

Customer churn prediction

For the development of the predictive models, the previously explained dataset was used, without the variable Product_speed, due to its high correlation with Product_price. Then, different classification models were evaluated: Decision Tree, Random Forest, Gradient Boosted Trees, and XGBoost.

Performance measures of all the models are presented in **Error! Reference source not found.**. Initially, the models were evaluated without using any balancing or baseline sampling strategy on the data, which revealed a majority class imbalance. Random Forest in this instance gave the best AUC-ROC (0.794), but it also yielded a very low recall (0.051), reflecting low capacity to identify churned clients. Similarly, GBT and XGBoost had AUC-ROC of 0.778 and 0.776 respectively, with even lesser recall.

Table 4: Evaluation of models.

Model	Sampling technique	AUC-ROC	Recall / Sensibility
Decision Tree	Baseline	0.714	0.074
Random Forest		0.794	0.051
GBT		0.778	0.029
XGBoost		0.776	0.044
Decision Tree	Sklearn's class_weight ‘balanced’	0.720	0.875
Random Forest		0.784	0.287
GBT		0.779	0.757
XGBoost		0.778	0.743
Decision Tree	SMOTE	0.671	0.360
Random Forest		0.768	0.360
GBT		0.761	0.162
XGBoost		0.754	0.213

Decision Tree	ADASYN	0.674	0.426
Random Forest		0.759	0.324
GBT		0.741	0.118
XGBoost		0.740	0.162
Decision Tree	Random downsampling	0.720	0.610
Random Forest		0.777	0.809
GBT		0.764	0.757
XGBoost		0.763	0.772
Decision Tree	NCR	0.755	0.117
Random Forest		0.781	0.081
GBT		0.779	0.066
XGBoost		0.776	0.088
Decision Tree	Tomek Links	0.745	0.147
Random Forest		0.783	0.015
GBT		0.781	0.029
XGBoost		0.781	0.044
Decision Tree	SMOTE + Random downsampling	0.663	0.397
Random Forest		0.773	0.294
GBT		0.757	0.176
XGBoost		0.748	0.206
Decision Tree	SMOTE + NCR	0.666	0.397

Random Forest		0.763	0.478
GBT		0.744	0.265
XGBoost		0.735	0.324
Decision Tree	SMOTE + Tomek Links	0.647	0.375
Random Forest		0.769	0.331
GBT		0.756	0.154
XGBoost		0.754	0.221
Decision Tree	ADASYN + Random downsampling	0.632	0.250
Random Forest		0.760	0.250
GBT		0.747	0.162
XGBoost		0.747	0.206
Decision Tree	ADASYN + NCR	0.675	0.449
Random Forest		0.757	0.485
GBT		0.744	0.228
XGBoost		0.746	0.331
Decision Tree	ADASYN + Tomek Links	0.680	0.404
Random Forest		0.760	0.324
GBT		0.751	0.162
XGBoost		0.746	0.154

The first class-balancing technique was sklearn's class_weight = 'balanced', where Decision Tree achieved the highest recall (0.875), with AUC-ROC at 0.720, showing high sensitivity, although a higher risk of false positives. GBT and XGBoost also improved their recall to 0.757 and 0.743 respectively, and maintained an AUC-ROC close to 0.778.

The impact of oversampling and undersampling techniques on the performance of the classification models can also be seen, as significant differences in terms of AUC-ROC and recall were obtained.

For example, moderate improvements in the recall were obtained when applying techniques such as SMOTE and Adasyn, however, the AUC-ROC was negatively affected, this could be seen in the GBT and XGBoost models. If we consider the performance of Random Forest using SMOTE, we have AUC-ROC as 0.768 with recall 0.360, whereas this model with Adasyn, in AUC-ROC reduced to 0.759, and in recall 0.324.

Among the tested algorithms, Random Downsampling and Random Forest performed best with the highest recall (0.809) and AUC-ROC of 0.777, which was the most suitable model for customer dropout risk prediction. XGBoost and GBT also enhanced recall (0.772 and 0.757, respectively), though losing a bit of AUC-ROC.

The more advanced undersampling strategies, such as NCR and Tomek Links, showed high AUC-ROC in Random Forest (0.781 and 0.783, respectively), but extremely low recall, indicating a lower ability to detect at-risk customers.

Interestingly, the combination of oversampling and undersampling techniques failed to improve the evaluation metrics. Recall obtained lower values, which shows that these combinations were not effective in addressing the problem with the data set under study. The best result in this section was achieved by Random Forest with SMOTE + NCR, with an AUC-ROC of 0.763 and a recall of 0.478, but it did not exceed the results previously observed.

Taking into account all the results of the models applied together with the oversampling and undersampling techniques, it was found that the model with the best combination of AUC-ROC and recall evaluation was Random Forest with Random Downsampling (AUC-ROC = 0.777, Recall = 0.809). The measured values indicate a higher capacity to identify customers prone to drop out, making it the best option in this study for dropout prediction.

The best-fit model obtained may be further validated by the Receiver Operating Characteristic (ROC) curve, depicted in **Error! Reference source not found.Error! Reference source not found.**. The continuous line plots the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different decision thresholds.

The area under the curve is 0.777, indicating high moderate discriminatory ability. In other words, this model can differentiate between defecting and loyal customers with higher performance than chance. And indeed change is represented by the dashed diagonal line, where a model would have no predictive ability, with an AUC of 0.5

Meanwhile, **Error! Reference source not found.** shows the five most relevant characteristics in the prediction of customer defection, according to the Random Forest model with Random Downsampling. These variables stand out for having the highest values of importance within the model.

Figure 5: ROC curve for Random Forest model with Random Downsampling

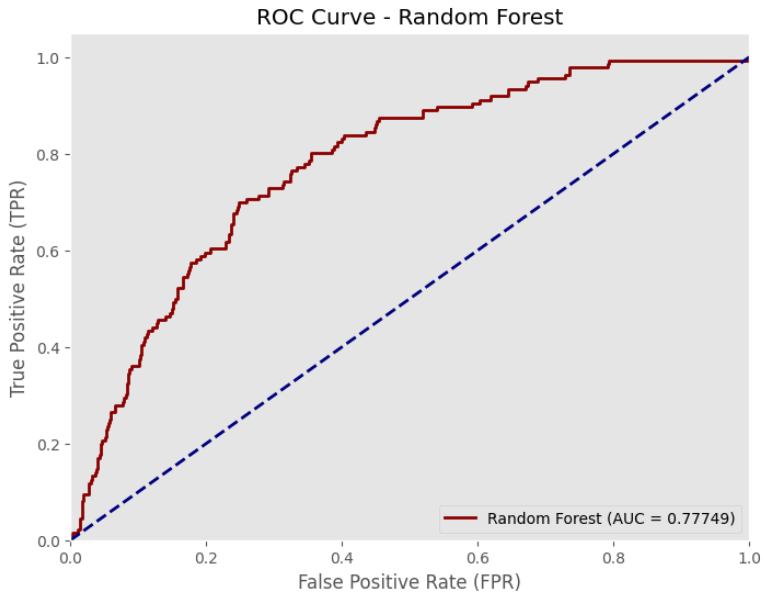
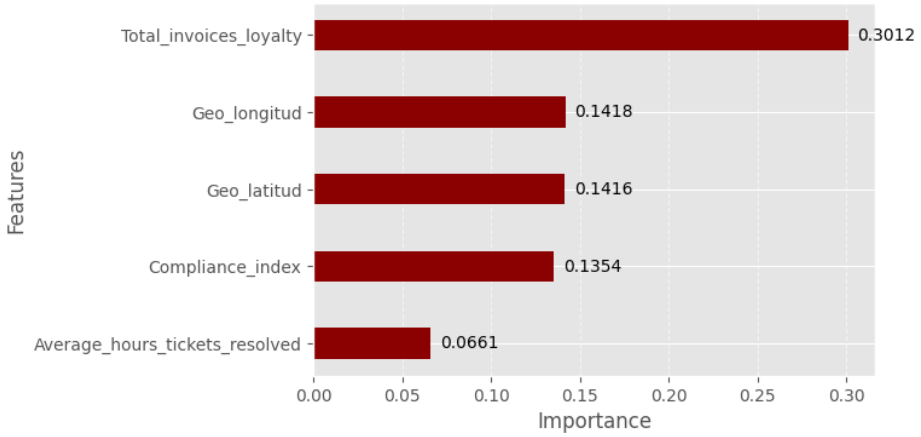


Figure 5: Most influential characteristics in the best-performing attrition model



Active service months are the highest among all the features, which suggests that billing history and customer loyalty are most likely to induce attrition. Likewise, longitude and latitude suggest that geographic location influences customer tenure. Payments made on time over total payments and average hours to close critical tickets also play an important role.

Geospatial analysis of client attrition

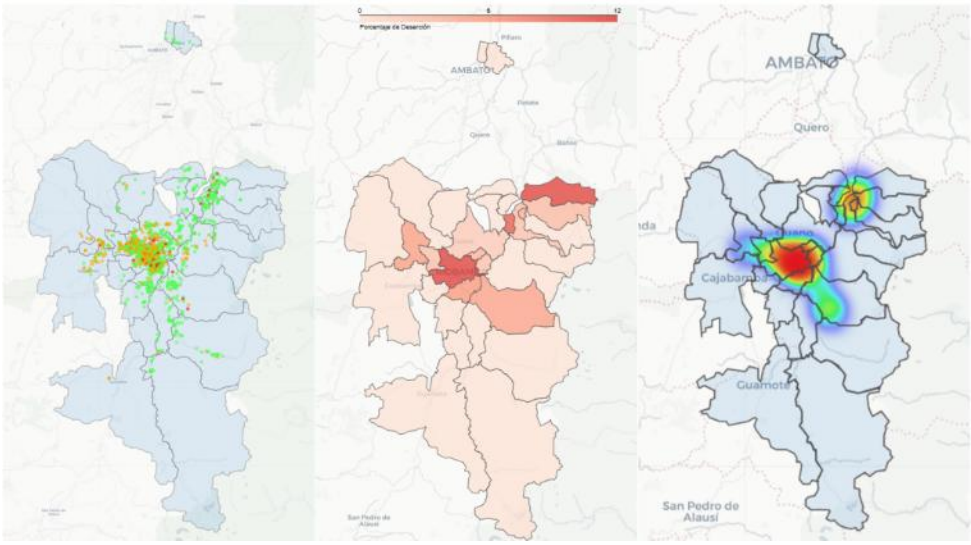
To understand the spatial distribution of customer churn, geospatial analysis tools were used in relationships generated by the Random Forest with a Random Downsampling model. Through this analysis, it was possible to identify spatial patterns, as well as areas of higher risk of churn, which allows the results to be interpreted in a geographical context and also provides strategic information to the ISP to optimize the service.

In fact, the geospatial analysis of customer churn predictions is illustrated in **Error! Reference source not found.**, which shows three interactive maps, in the areas where the ISP operates. These allow the identification of spatial patterns related to defection.

The first map (left) shows the geographical spread of predictions for each of the customers using their positioning data. Three colors are shown in this map, associated with the attrition probability: low in green, medium in orange, and high in red. From this visualization, it can be seen that there is a higher concentration of clients likely to drop out in urban areas, and a lower concentration in rural areas.

The second map (center) provides an attrition analysis by parishes the ISP is present in. For this purpose, individual forecasts have been integrated, and each parish's probability of customers who are likely to churn has been estimated. In this map, darker shades of color represent higher chances of churn. Moreover, the use of this data on territorial patterns leads to the generation of strategies and decisions.

Figure 6: Visualizing customer churn prediction on maps from the model



The third is a heat map (right), and it provides information on the distribution of clients with high drop-out propensity based on predictions from the best model.

Regions of red and yellow hues tend to drop out. The visualization helps to locate hotspots that require extra effort to earn customer loyalty.

DISCUSSION

The research question posed in this study aims to determine how a predictive model that integrates machine learning and geospatial analytics can predict customer churn at an ISP operating in Ecuador. This discussion analyses the results achieved in the execution of this study, from data processing, customer segmentation, and predictive modeling to geospatial analysis.

Through the results obtained, it can be demonstrated that churn prediction is a complex problem, which can be addressed and benefit from multiple data sources and methodologies. Undoubtedly, one of the most important phases in this study was the pre-processing of the data, as through this it was possible to ensure the quality and also the consistency of the dataset. To structure the robust dataset for further analysis and training, the elimination of duplicate records, the handling of missing values, and the identification of outliers were applied. In addition, one of the most prominent challenges related to the dataset was class imbalance, as only 7.32% of the clients had dropped out. This aspect is noteworthy as it significantly influenced the performance of the classification models, necessitating the application of class balancing techniques to improve detection rates, in line with the findings of (Chong et al., 2023; Geiler et al., 2022; Krishna et al., 2024; Wagh et al., 2024), who demonstrated the need for SMOTE and other balancing techniques to address class imbalance in telecommunications.

The segmentation allowed the identification of four clusters, which have similar characteristics and were analyzed to understand customer behavior. These clusters emerged from customer data once they were part of the company, i.e. dwell time, payment compliance, and interaction with the service. After analysis, cluster 1, which is composed of customers with irregular payments and less time with the company, showed the highest dropout. In contrast, clusters 0 and 2, which are composed of customers with more time in the service and timely payments, showed lower attrition. These results relate to studies by (Pejić Bach et al., 2021), (Zhao et al., 2023), (Xu et al., 2023) and (Fatima et al., 2024), who highlighted the relevance of segmentation for analyzing customer behavior and predicting attrition.

The evaluations of each of the classification models, presented in **Error! Reference source not found.**, highlight Random Forest with Random Downsampling as the one that achieved the best balance between AUC-ROC (0.777) and recall (0.809) metrics, which maintains a reasonable false positive rate. Also, the effectiveness of Random Downsampling in addressing class imbalance was highlighted, as it allowed the model to learn from the minority class, without affecting the overall model performance.

Studies by (Adhikary & Gupta, 2021), (Colot et al., 2021), (Krishna et al., 2024), (Wagh et al., 2024) y (Ortakci & Seker, 2024), show similar results, where Random Forest was identified as the best model in predicting attrition in the telecommunications sector, and in scenarios with unbalanced classes.

The ability of the Random Forest model to measure the feature importance in churn prediction was of paramount importance, as it stated customer subscription length to be the most important factor in predicting churn, a factor that corroborates the hypothesis that customers with longer subscription times on the service are less likely to churn. Likewise, geospatial characteristics, such as longitude and latitude, were highly relevant, pointing towards there being geographical determinants of whether to exit the service or not.. This aligns with (Colot et al., 2021), whose study pointed towards the possibility of utilizing geospatial data in the form of improved classification models to forecast attrition when paired with additional information.

Additional information regarding the pattern of dropout predictions became available with the use of geospatial analysis. Interactive maps allowed for easy visualization of how clients behave in accordance with model predictions. For example, heat map visualization revealed urban areas with higher attrition rates, which would imply that there may be service quality issues or aggressive commercial pressures. Conversely, parish-level data identified areas with a greater likelihood of clients to fall away, pointing towards the necessity of localized retention efforts. These results are in line with (Matuszelański & Kopczewska, 2022), who proved that geographic and socio-economic determinants influence customer defection in the Retail E-Commerce sector.

To ISPs, geospatial information is relevant since the combination of geospatial information with segmentation modeling and churn forecasting allows them to develop focused retention programs, which could be more effective than normal ones. For example, using focused loyalty programs where attrition could occur, improving infrastructure and quality of services, or investigating the potential of adjusting the price of internet plans. In addition, with the recognition of geographic trends, there is the potential to predict churn patterns and a more effective resource allocation can be attained.

The outcomes of this study to predict churn in ISP highlight the importance of combining different methodologies such as machine learning and geospatial analysis. Customer segmentation is critical to understand customer behavior, while class balancing as well as applying classification models improves discrimination between loyal customers and defectors. And even though geospatial characteristics may not be the best predictor, they are used in decision-making. This result corresponds with the study of (Ortakci & Seker, 2024), who demonstrated that by integrating machine learning and business expertise, it is achievable to develop more efficient retention plans.

Articles
Revista tecnológica ciencia y educación Edwards Deming
ISSN: 2600-5867

For future research, it would be important to address ensemble and hybrid techniques, such as those used by (Xu et al., 2021), (Wahul et al., 2023), (Khoh et al., 2023) and (Geiler et al., 2022). Furthermore, the application of deep learning in this scenario could lead to the identification of more complex patterns in customer behavior. Another approach that would be interesting to address is to integrate external characteristics or variables into the dataset, such as socio-economic indicators or the presence of competitors, which could help in the refinement of the predictive model. Finally, for class imbalance, other resampling strategies could be used, such as those applied in the study by (Ouf et al., 2024).

This study demonstrates the effectiveness of combining data pre-processing, segmentation, classification modeling, and geospatial analysis in customer churn prediction in ISP. The findings of this study add to the cumulative literature of predictive churn analysis in the telecommunication sector and also give valuable insights to the company to devise strategies to target customers who are at higher risk.

This research examined the prediction of customer churn in an internet service provider in central Ecuador, using a model based on machine learning techniques and geospatial analysis. It was possible to determine significant patterns of customer churn and select the optimal model for its prediction based on pre-processing of data, customer segmentation, and evaluation of different classification models.

Based on the results, short-time customers and low compliance customers in paying on time were more likely to leave the service. On the other hand, the application of data balancing techniques improved the performance of the model by allowing it to discriminate between customers likely to drop out of the service and those that are not. The best model in this study was a Random Forest with Random Downsampling, which performed higher ROC-AUC and recall scores.

On the other hand, geospatial analysis allowed for the identification of spatial patterns related to attrition. Such information is key for ISPs to design effective retention strategies. Overall, the combination of all these approaches proved to be useful in anticipating customer attrition.

Lastly, the nexus between machine learning and geospatial analysis represents a crucial tool for elevating customer management in the telecom sector, and there is a suggestion that further research on advanced modeling paradigms and data enrichment is conducted to improve the accuracy of the projections and hence the effectiveness of retention strategies.

REFERENCES

Adhikary, D. Das, & Gupta, D. (2021). Applying over 100 classifiers for churn prediction in telecom companies. Multimedia Tools and Applications, 80(28–29), 35123–35144. https://doi.org/10.1007/S11042-020-09658-Z/METRICS

Agasti, B. R., & Satpathy, S. (2024). Predicting customer churn in telecommunication sector using Naïve Bayes algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 35(3), 1610–1617. https://doi.org/10.11591/ijeecs.v35.i3.pp1610-1617

Aguilar Cazar Miguel. (2024). Informe de Avance del Indicador 8.1.2.: Porcentaje de parroquias rurales y cabeceras cantonales con presencia del servicio de internet fijo a través de enlaces de fibra óptica. https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2024/07/Informe-de-Avance-del-Indicador_Primer-Trimestre_Indicador-8.1.2.pdf

Alteer, S. A., & Alariyibi, A. (2024). Customer Churn Prediction Using Machine Learning: A Case Study of Libyan Internet Service Provider Company. 2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, MI-STA 2024 - Proceeding, 605–612. https://doi.org/10.1109/MI-STA61267.2024.10599671

Bachan, L., & Gaber, T. (2021). Predicting Customer Churn in the Internet Service Provider Industry of Developing Nations: A Single, Explanatory Case Study of Trinidad and Tobago. Advances in Intelligent Systems and Computing, 1339, 835–844. https://doi.org/10.1007/978-3-030-69717-4_77

Bilal, S. F., Almazroi, A. A., Bashir, S., Khan, F. H., & Almazroi, A. A. (2022). An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. PeerJ Computer Science, 8, e854. https://doi.org/10.7717/PEERJ-CS.854/SUPP-2

Cenggoro, T. W., Wirastari, R. A., Rudianto, E., Mohadi, M. I., Ratj, D., & Pardamean, B. (2021). Deep Learning as a Vector Embedding Model for Customer Churn. Procedia Computer Science, 179, 624–631. https://doi.org/10.1016/J.PROCS.2021.01.048

Chong, A. Y. W., Khaw, K. W., Yeong, W. C., & Chuah, W. X. (2023). Customer Churn Prediction of Telecom Company Using Machine Learning Algorithms. Journal of Soft Computing and Data Mining, 4(2), 1–22. https://doi.org/10.30880/jscdm.2023.04.02.001

Colot, C., Baecke, P., & Linden, I. (2021). Leveraging fine-grained mobile data for churn detection through Essence Random Forest. Journal of Big Data, 8(1), 1–26. https://doi.org/10.1186/S40537-021-00451-9/FIGURES/7

Fatima, G., Khan, S., Aadil, F., Kim, D. H., Atteia, G., & Alabdulhafith, M. (2024). An autonomous mixed data oversampling method for AIOT-based churn recognition and personalized recommendations using behavioral segmentation. PeerJ Computer Science, 10, 1–32. https://doi.org/10.7717/PEERJ-CS.1756/SUPP-1

Geiler, L., Affeldt, S., & Nadif, M. (2022). An effective strategy for churn prediction and customer profiling. Data & Knowledge Engineering, 142, 102100. https://doi.org/10.1016/J.DATAK.2022.102100

Khoh, W. H., Pang, Y. H., Ooi, S. Y., Wang, L. Y. K., & Poh, Q. W. (2023). Predictive Churn Modeling for Sustainable Business in the Telecommunication Industry: Optimized Weighted Ensemble Machine Learning. Sustainability 2023, Vol. 15, Page 8631, 15(11), 8631. https://doi.org/10.3390/SU15118631

Krishna, R., Jayanthi, D., Shylu Sam, D. S., Kavitha, K., Maurya, N. K., & Benil, T. (2024). Application of machine learning techniques for churn prediction in the telecom business. Results in Engineering, 24, 103165. https://doi.org/10.1016/J.RINENG.2024.103165

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 104(2), 271–294. https://doi.org/10.1007/S00607-021-00908-Y/METRICS

Liu, R., Ali, S., Bilal, S. F., Sakhawat, Z., Imran, A., Almuhaimeed, A., Alzahrani, A., & Sun, G. (2022). An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms. Applied Sciences 2022, Vol. 12, Page 9355, 12(18), 9355. https://doi.org/10.3390/APP12189355

Matuszelański, K., & Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. Journal of Theoretical and Applied Electronic Commerce Research 2022, Vol. 17, Pages 165-198, 17(1), 165–198. https://doi.org/10.3390/JTAER17010009

Number of internet users worldwide 2024 | Statista. (n.d.). Retrieved January 22, 2025, from https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Ortakci, Y., & Seker, H. (2024). Optimising customer retention: An AI-driven personalised pricing approach. Computers and Industrial Engineering, 188, 109920. https://doi.org/10.1016/j.cie.2024.109920

Ouf, S., Mahmoud, K. T., & Abdel-Fattah, M. A. (2024). A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry. Journal of Big Data, 11(1), 70. https://doi.org/10.1186/s40537-024-00922-9

Pejić Bach, M., Pivar, J., & Jaković, B. (2021). Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees. Journal of Risk and Financial Management 2021, Vol. 14, Page 544, 14(11), 544. https://doi.org/10.3390/JRFM14110544

Peng, X., Niu, Y. yan, Meng, B., Tao, Y., & Huang, Z. (2024). Big geo-data unveils influencing factors on customer flow dynamics within urban commercial districts. International Journal of Applied Earth Observation and Geoinformation, 134, 104231. https://doi.org/10.1016/J.JAG.2024.104231

Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. Machine Learning with Applications, 17, 100567. https://doi.org/10.1016/J.MLWA.2024.100567

Saha, L., Tripathy, H. K., Gaber, T., El-Gohary, H., & El-kenawy, E. S. M. (2023). Deep Churn Prediction Method for Telecommunication Industry. Sustainability 2023, Vol. 15, Page 4543, 15(5), 4543. https://doi.org/10.3390/SU15054543

Shrestha, S. M., & Shakya, A. (2022). A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal. Procedia Computer Science, 215, 652–661. https://doi.org/10.1016/J.PROCS.2022.12.067

Soleiman-garmabaki, O., & Rezvani, M. H. (2024). Ensemble classification using balanced data to predict customer churn: a case study on the telecom industry. Multimedia Tools and Applications, 83(15), 44799–44831. https://doi.org/10.1007/s11042-023-17267-9

Usman-Hamza, F. E., Balogun, A. O., Amosa, R. T., Capretz, L. F., Mojeed, H. A., Salihu, S. A., Akintola, A. G., & Mabayoje, M. A. (2024). Sampling-based novel heterogeneous multi-layer stacking ensemble method for telecom

customer churn prediction. Scientific African, 24, e02223. https://doi.org/10.1016/J.SCIAF.2024.E02223

Usman-Hamza, F. E., Balogun, A. O., Nasiru, S. K., Capretz, L. F., Mojeed, H. A., Salihu, S. A., Akintola, A. G., Mabayoje, M. A., & Awotunde, J. B. (2024). Empirical analysis of tree-based classification models for customer churn prediction. Scientific African, 23, e02054. https://doi.org/10.1016/J.SCIAF.2023.E02054

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14, 100342. https://doi.org/10.1016/J.RICO.2023.100342

Wahul, R. M., Kale, A. P., & Kota, P. N. (2023). An Ensemble Learning Approach to Enhance Customer Churn Prediction in Telecom Industry. International Journal of Intelligent Systems and Applications in Engineering, 11(9s), 258–266.

Xu, T., Ma, Y., Ao, C., Qu, M., & Meng, X. H. (2023). A NOVEL TELECOM CUSTOMER CHURN ANALYSIS SYSTEM BASED ON RFM MODEL AND FEATURE IMPORTANCE RANKING. Interdisciplinary Journal of Information, Knowledge, and Management, 18, 719–737. https://doi.org/10.28945/5192

Xu, T., Ma, Y., & Kim, K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. Applied Sciences 2021, Vol. 11, Page 4742, 11(11), 4742. https://doi.org/10.3390/APP11114742

Zhao, Y., Shao, Z., Zhao, W., Han, J., Zheng, Q., & Jing, R. (2023). Combining unsupervised and supervised classification for customer value discovery in the telecom industry: a deep learning approach. Computing, 105(7), 1395–1417. https://doi.org/10.1007/s00607-023-01150-4