Integration and Analysis of Telecommunications Customer Churn Data

Ricardo Andrés Chamorro Martínez, Oscar Stiven Muñoz Ramirez & Diego Armando Polanco Lozano
Algorithm and Programming III Laboratory 2,
ICESI University

Summary

This report presents the development of a customer churn analysis and prediction project for a telecommunications company, applying the CRISP-DM methodology. Multiple data sources were integrated and cleaned to construct a coherent and inconsistency-free dataset, on which exploratory analyses and supervised classification models were performed. The results show that factors such as contract type, customer tenure, monthly charges, and payment method significantly influence the probability of churn. Among the models evaluated,**Random Forest**It achieved the best performance (F1=0.906, AUC=0.981), demonstrating high accuracy and predictive capacity, which allows for the effective identification of at-risk customers and the strengthening of retention strategies.

I. INTRODUCTION

This lab prepares and analyzes customer churn data in a telecommunications company using CRISP-DM. First, integrate seven files into a unified master dataset, then clean, impute, and code variables. Finally, it develops supervised classification models (Logistic Regression, Decision Trees, Random Forest) to predict Churn, evaluating their performance with cross-validation and metrics such as accuracy, recall, F1-score and confusion matrix. The goal is to understand dropout factors in order to improve retention strategies.

II. THEORETICAL MARK

A. Customer Churn

It is the percentage of users who leave the service in a given period, directly affecting the company's profitability.

B. Dropout Factors

These include dissatisfaction, service quality, competition, and prices. Analyzing these aspects will help reduce churn.

C. Data Preparation and Normalization

Data preparation includes integrating, cleaning, and normalizing information from various sources. Techniques such as null imputation, scaling, and coding are used to ensure consistent data suitable for analysis.

D. Supervised Models

Supervised machine learning models, such as logistic regression, decision trees, random forest, or XGBoost, are crucial for predicting dropout rates. They handle large structured datasets and reduce noise, improving the accuracy in identifying dropout patterns.

E. Hyperparameters

Optimizing hyperparameters such as tree depth and learning rate is key to obtaining accurate, robust predictive models and avoiding overfitting/underfitting.

III. METHODOLOGY

Results of the integration process

The main Telco_customer_churn.csv file contained 7,043 records (unique customers) with no duplicates in the Customer ID key. After merging with the six subfiles, the final dataset retained exactly 7,043 rows, confirming that no records were lost or duplicated, as all joins were performed using left joins. The consolidated dataset reached 64 columns, combining demographic, geographic, service, customer status, and financial metrics, including variables such as Customer ID, Gender, Senior Citizen, Internet Service, Monthly Charges, Total Charges, Churn Label, Customer Status, Population, Satisfaction Score, CLTV, Contract, and Payment Method.

Handling duplicate keys

In all files, the join keys (Customer ID or Zip Code) were unique, with no duplicates. To prevent accidental row duplication, an automatic collapse rule was applied based on the first non-null occurrence in case of duplication. In practice, this aggregation was not necessary, as no secondary file contained repeated keys.

File	Union key	rows	columns	key duplicate	percentage of duplicates
CustomerChurn	Customer ID	7,043	21	0	0.0 %
T.Demographics	Customer ID	7,043	9	0	0.0 %
T.Location	Customer ID	7,043	10	0	0.0 %
T.Population	Zip Code	1,671	3	0	0.0 %
T.Services	Customer ID	7,043	31	0	0.0 %
T.Status	Customer	7,043	12	0	0.0 %

	ID				

Table 1. Percentage of rows without a match in each file

Key Observations on Data Integration:

1. **Customer Churn:**He enriched the dataset with crucial variables regarding customer retention, the services they have contracted, and the payment methods used.
2. **T.Demographics:**It added essential demographic information about customers, including age, marital status, and the number of dependents.
3. **T.Location:**It added a geographical dimension, incorporating data such as country, state, city, postal code, and coordinates.
4. **T.Population:**It was successfully integrated using the Zip Code as the key (instead of the Customer ID), allowing the population variable to be filled in accurately for all customers.
5. **T.Services:**It provided comprehensive details on the services contracted by customers, available offers, applied charges, and average consumption.
6. **T.Status:**It incorporated data on the current status of each customer, their level of satisfaction, and the specific churn category in which they are found.

Did merging the Telco_customer_churn.csv and CustomerChurn.csv files remove duplicate columns?Yes, using `resolve_collisions_left_priority()`, merging Telco_customer_churn.csv and CustomerChurn.csv eliminated duplicate columns such as "Senior Citizen" and "Total Charges," prioritizing those from the main file. Out of a total of more than 90 potential columns, 64 unique and non-redundant columns were obtained, confirming a successful merge without artificially inflating the dataset with repeated information. However, the cleaning section will search more thoroughly for duplicates.

Variable	Type	Nulls	Description
Gender	Categorical (object)	0	Client's sex (Male / Female).
Senior Citizen	Categorical (object)	0	Indicate if the customer is a senior citizen.
Internet Service	Categorical (object)	0	Type of Internet service contracted (DSL, Fiber optic, None).
Monthly Charges	Numeric (float)	0	Monthly value billed for the contracted services.
Churn Label	Categorical (object)	0	Indicate whether the customer deserted (Yes) or remains (No).

Table 2. Sample of the 5 most representative columns

IV. DATA CLEANING AND TRANSFORMATION

The data cleaning and transformation phase aimed to ensure the quality, consistency, and usability of the integrated dataset (Telecom_Customer_Churn_Complete.csv), guaranteeing that it was free of null values, semantic inconsistencies, and formatting errors that could affect subsequent analysis. The actions performed in each substage of the process are detailed below.

a. Identification and imputation of null values: First, null values were checked with df.isna().sum(). The "Churn Category" and "Churn Reason" columns (5,174 nulls each) were imputed with "Not Applicable" since they only apply to customers who defected. The remaining categorical variables (*Offer*, *Internet Type*) were filled with fashion. Numerical variables did not require imputation, except*Total Charges*. After applying these rules, the dataset had no null valuess.

b. Conversion of the Total Charges column: The "Total Charges" column was corrected for statistical analysis. Initially, it was text with non-numeric characters and empty values. The steps included: removing symbols, converting to numeric (NaN for invalid), and replacing NaN with 0.0 (customers without charges). The conversion was verified with a statistical summary, resulting in a numeric column (float) ready for analysis.

c. Semantic cleaning of service columns: The textual values of the service variables were standardized to eliminate ambiguities. "No internet service" was replaced with "No" in internet columns (Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Device Protection Plan) and "No phone service" was replaced with "No" in "Multiple Lines". Thus, all service variables were left with "Yes" and "No", simplifying the binary coding, avoiding confusion in the modeling and facilitating the interpretation in analysis and graphs.

d. Normalization of identifiers and column names:
To optimize data manipulation and ensure structural consistency, the following actions were taken:

1. The "Customer ID" column was renamed to "Customer_ID" and set as the primary index to ensure the uniqueness of each record.
2. The column names were cleaned up: spaces, opening parentheses and brackets were replaced with underscores; closing parentheses and commas were removed.

For example, "Monthly Charges (USD)" was transformed into "Monthly_Charges_USD", standardizing the names, eliminating conflicting characters, and facilitating integration with machine learning libraries.

e. Grouping of customers according to length of stay:"Tenure_Group" was created from "Tenure in Months" for comparative analysis and as a categorical variable. Customers were classified into five tenure groups: 0-12, 12-24, 24-48, 48-60 and more than 60 months, facilitating the identification of

dropout patterns based on tenure.

f. Population variable conversion: The variable "Population", originally presented as text with commas, was converted to a numeric format by removing the commas and applying the pd.to_numeric() function, in order to facilitate its use in subsequent analyses.

g. Elimination of unnecessary columns: The columns "Unnamed: 0", "ID", "Customer_ID", "Tenure_in_Months" (the latter being replaced by Tenure_Group), "Churn_Reason" and "Churn_Category" were removed. The last two were removed for the purpose of data leakage, that is, the inclusion of future or unavailable information at the time of the prediction. The resulting dataset consists of 61 columns, has no duplicates, and features a consistent naming convention.

V. EXPLORATORY DATA ANALYSIS (EDA).

Process: Once preprocessing was complete, in-depth analysis of the dataset, now comprising 48 variables, began. The analysis included exploring the class balance (26.5% churn vs. 73.5% non-churn), univariate analysis of demographic and service variables, bivariate analysis of each variable versus churn, and in-depth multivariate analysis evaluating interactions between contract type, internet services, monthly/total charges, customer tenure, additional services, payment methods, and combined demographic profiles. Multiple visualizations were used (histograms, boxplots,violinplots, heatmaps, scatter plots) and correlation analysis to identify patterns.

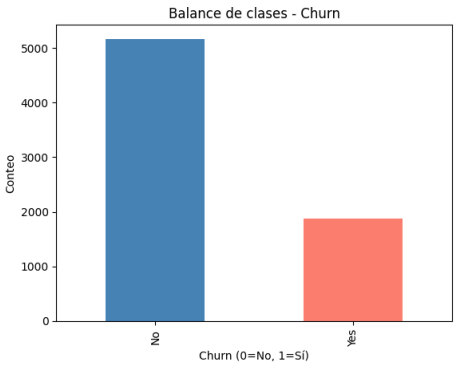

Chart 1. Target variable balance

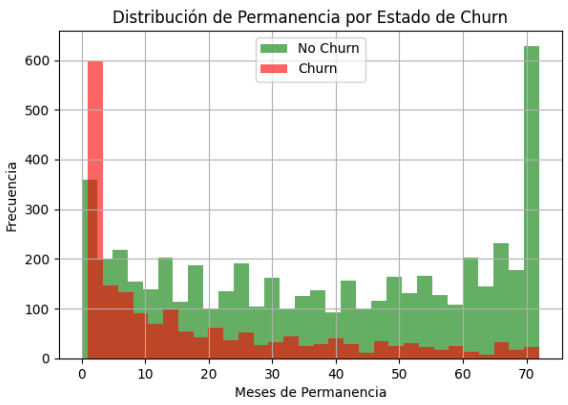

Figure 2. Distribution of churn over time

Key Findings: The analysis revealed that contract type is the most critical factor: month-to-month contracts have a 42% churn rate versus only 3% for two-year contracts. Tenure shows a strong inverse relationship: 47-50% churn in the first 12 months, dropping to less than 15% after 24 months. Monthly charges are significantly higher for customers who churn ($74 vs. $61), while total charges are lower (indicating early churn). Value-added services such as Online Security and Tech Support reduce churn from approximately 40% to approximately 15%. Payment method is also relevant: Electronic Check has a 45% churn rate versus 15-17% for automatic methods. Senior citizens without partners or dependents make up the most vulnerable segment (41-45% churn).

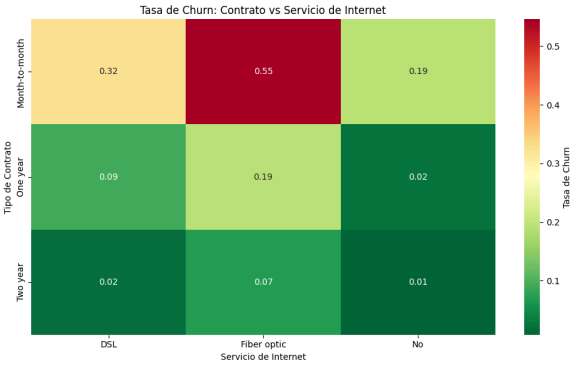

Chart 3. Churn percentage given payment frequency and contract service

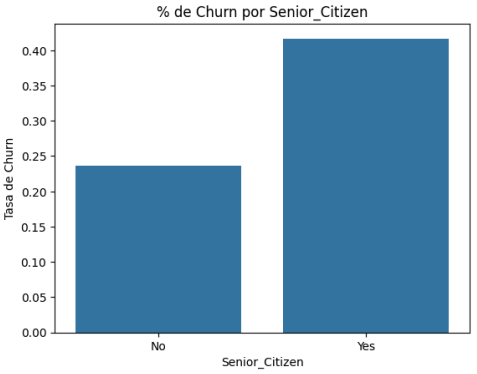

Chart 4. Churn rate given the type of citizen "senior"

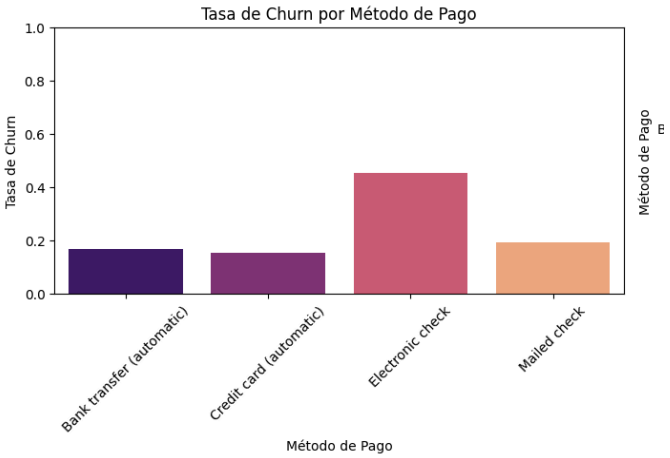

Chart 5. Churn rate given the type of payment in your contract

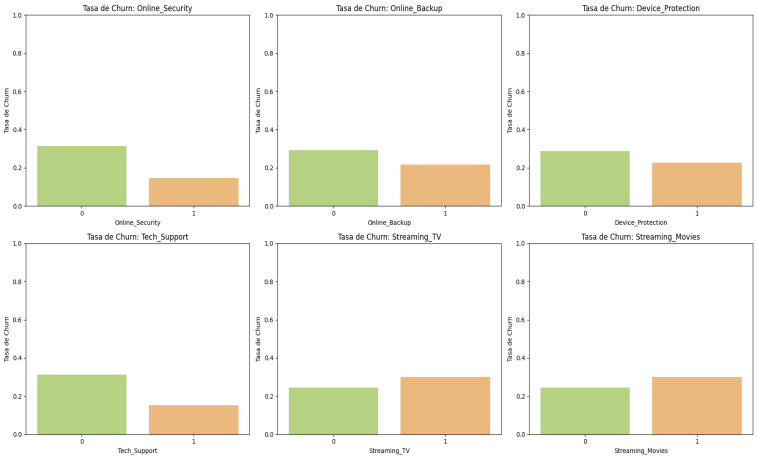

Chart 6. Churn rate given the presence or absence of a service in the contract

EDA Conclusion: The EDA clearly identified high-risk segments (flexible contracts + low tenure + no additional services + high charges + electronic checks + single demographic profiles) and established that retention strategies should focus heavily on the first 6–12 months of the customer lifecycle, promote long-term contracts, incentivize the adoption of security/technical support services, and transition customers to automatic payment methods. The strongest correlations with churn were Tenure (-0.35), Total Charges (-0.20), Monthly Charges (+0.19), and Paperless Billing (+0.19), validating these factors as key predictors for the subsequent model.

VI. EXPLORATION AND MONITORING OF CLASSIFICATION MODELS

The process began with the selection of three supervised classification algorithms designed to predict customer churn: Logistic Regression (linear and interpretable model), Decision Tree (nonlinear explanatory model), and Random Forest (robust ensemble of multiple trees). Before training, the dataset underwent rigorous cleaning, removing columns with information leakage (such as Churn_Reason and Churn_Category, which only exist after churn), unique identifiers without predictive value (Customer_ID, Location_ID), and duplicate or constant variables. Subsequently, the data was divided into a training set (80%) and a test set (20%) using stratification to maintain class proportions.

To robustly evaluate the models, stratified cross-validation was implemented.con k=5Folds, which divides the dataset into 5 subsets while maintaining the class proportion in each partition. Each model was trained 5 times using 4 folds for training and 1 for validation, rotating the folds and averaging the metrics (Accuracy, Precision, Recall, F1-score, and AUC-ROC). This approach ensures a reliable performance estimate without relying on a single data split, reducing the risk of overfitting and allowing for an objective comparison of the three models. Random Forest emerged as the best model with F1=0.914 and AUC=0.984 in cross-validation, outperforming Logistic Regression (F1=0.912, AUC=0.973) and Decision Tree (F1=0.890, AUC=0.925).

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.958	0.938	0.902	0.919	0.992
Decision Tree	0.942	0.888	0.896	0.892	0.928
Random Forest	0.957	0.971	0.864	0.914	0.984

VII. RESULTS OF THE BEST MODEL.

The Random Forest model demonstrated exceptional performance in the independent test set, achieving an accuracy of 95.3%, meaning it correctly classified over 95% of all customers. Key metrics reveal an excellent balance:PrecisionA value of 0.97 indicates that when the model predicts that a customer willdesertIt is correct 97% of the time (only 10 false positives), while the Recall score of 0.85 shows that it correctly detects 85% of customers who actually quit the service (56 false negatives). The F1 score of 0.906 confirms the balance between both metrics, and the AUC-ROC of 0.981 demonstrates an almost perfect ability to discriminate between quitters and non-quitters. The confusion matrix shows 1025 true negatives and 318 true positives, demonstrating robust classification.

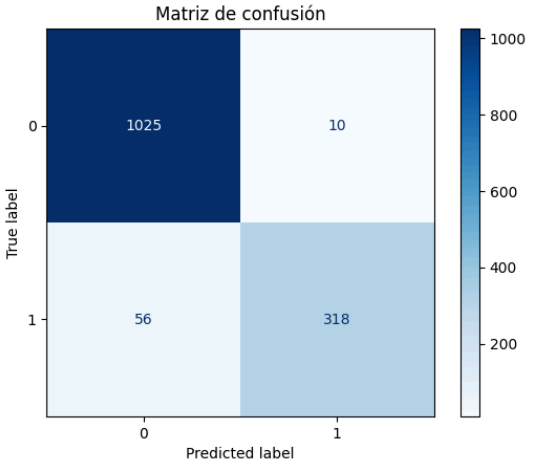

Figure 7. Confusion matrix of the random forest model

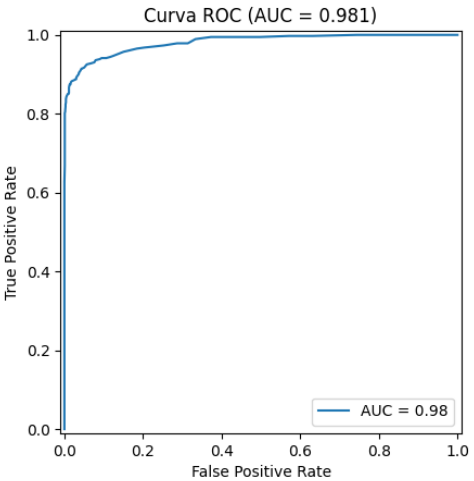

Figure 8. ROC curve of the random forest model.

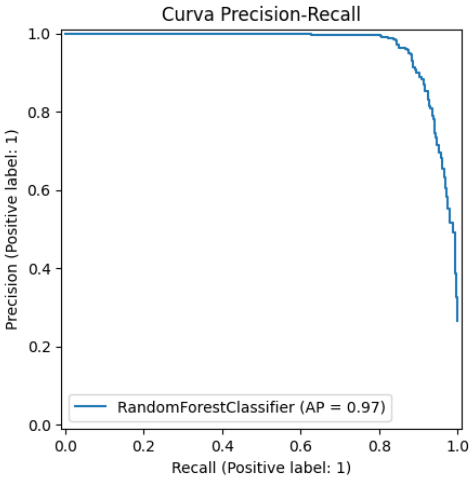

Figure 9. Precision-Recall Curve of the random forest model

These results position Random Forest as a highly reliable model for customer retention programs, allowing for the highly accurate identification of those at greatest risk of churn. The model is slightly conservative (prioritizing accuracy over recall), meaning it prefers to avoid contacting customers without real risk (false positives), even if it misses some potential churns. The ROC and Precision-Recall curves (both with areas greater than 0.97) confirm that the model maintains its excellent performance even when the decision threshold is adjusted, providing operational flexibility. This balance between accuracy and coverage allows the company to strategically focus its retention resources on customers truly vulnerable to churn, optimizing costs and maximizing the impact of interventions.

VIII. CONCLUSIONS.

The lab successfully integrated, cleaned, and analyzed a large customer dataset from a telecommunications company, applying the CRISP-DM methodology in a comprehensive and structured manner. Through the preparation and transformation process, a coherent dataset, free of duplicates and null values, was ensured, ready for predictive modeling. Exploratory analysis identified key factors in customer churn, such as contract type, customer tenure, monthly charges, additional services, and payment method, revealing clear behavioral patterns and high-risk segments.

In the modeling phase, the supervised algorithms—especially Random Forest—demonstrated outstanding performance, achieving an F1-score of 0.906 and an AUC of 0.981, which shows high predictive power and reliability. These results allow the company to anticipate customer churn with high accuracy, facilitating strategic decision-making focused on early retention, customer loyalty, and the optimization of sales resources.

In conclusion, the study not only built a robust predictive model, but also provided a deep understanding of the factors that influence churn, offering a solid basis for the design of effective policies that reduce churn and strengthen customer relationships.

IX. BIBLIOGRAPHY

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.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), 2902–2917

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3ra ed.). Morgan Kaufmann