Delivery Report — Churn Prediction Project in Telcos

1. Project Objective

Identify customers most likely to leave the mobile service by analysing data and developing machine learning models. The results help design strategies to reduce churn, improve user experience, and optimize internal processes.

2. Data Exploration and Preparation (EDA)

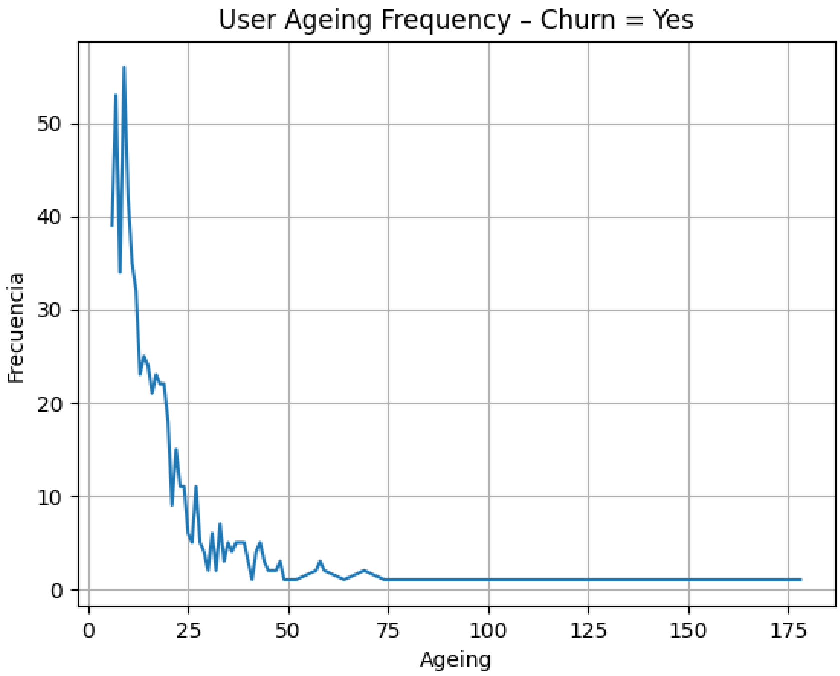
- The data was loaded and cleaned, removing duplicates and formatting variables where required.
- Variables **CX_QF, CX_RECURSO, CX_QS, CX_PETICION, CX_INTENCION_RET, CX_INTENCION_NO_RET** were excluded due to a high number of missing values and low variability (only taking the value zero).
- Categorical variables were encoded using **WoE (Weight of Evidence)** to improve model stability, since some customer experience variables had many missing values.
- The target variable showed strong imbalance:

Churn	Frecuencia
0	499,372
1	628

Therefore, **SMOTE balancing** was applied to train the supervised models.

3. Relationship Between Variables and Churn

- **Ageing:** The chart *"User Tenure Frequency – Churn = Yes"* shows that customers with shorter tenure are more likely to leave.



- **Contact Reasons:** The category with the highest churn rate is **Cancellation** (0.45%), compared to much lower rates in other reasons (e.g., Commercial 0.05%).

---

4. Net Promoter Score (NPS)

- **Global KPI:** 40.25 (moderate satisfaction level).

By region:

- Andina: 40.42
- Bogotá: 33.83
- Costa: 45.05
- Sur: 45.61

By category:

- Commercial: 64.51
- Requests: 52.94
- Cancellation: 44.93
- Billing: 36.61
- Technical Support: 31.62
- Abandoned: -50.00 (highly dissatisfied customers).

5. Contact Index

The calculated **Contact Index** was **0.19075**, meaning that ~19% of active customers made at least one contact during the period. Customers with higher interaction frequency are more likely to churn.

6. Variables Influencing Revenue

Spearman Correlation (Top):

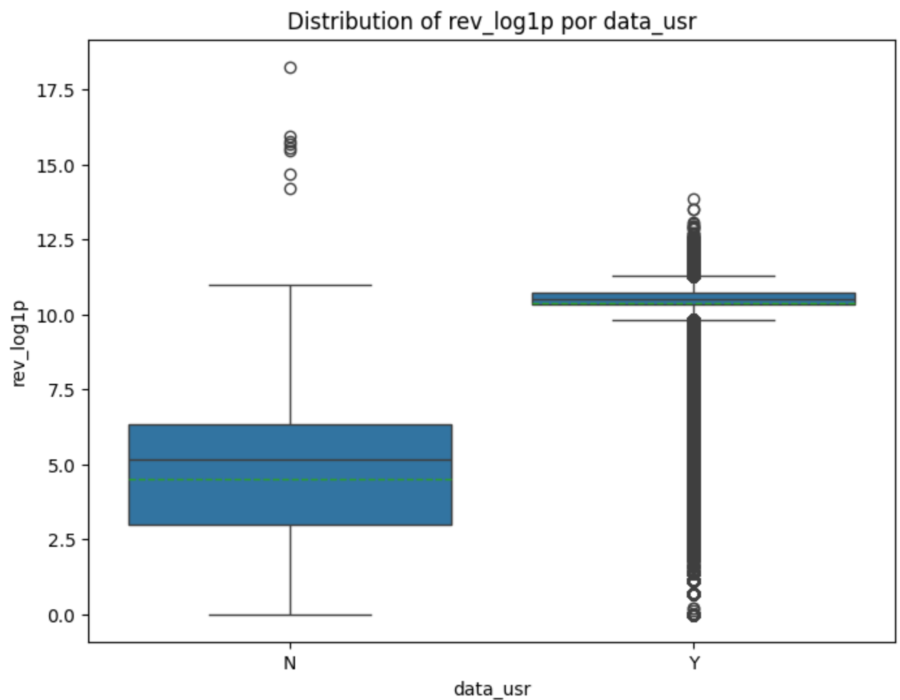
- ageing: 0.32
- data_mb: 0.16
- CX_SOP_TEC: 0.08
- CX_PETICIONES: 0.07

The strongest positive correlations with revenue were **tenure** and **data consumption**.

ANOVA Analysis: confirmed that categorical variables such as *region, data_usr, technology, device_technology, and device_subtype* significantly influence revenue (p-value = 0.0).

Descriptive Model (Random Forest) – Most Important Variables:

- data_usr_N (49%)
- data_mb (29%)
- data_usr_Y (17%)
- ageing (2.5%)



Conclusion: Revenue mainly depends on data usage, customer tenure, and user type.

7. Churn Predictive Modeling

Supervised models tested:

- Logistic Regression
- Random Forest
- Balanced Random Forest
- XGBoost
- Decision Tree

Unsupervised models:

- Isolation Forest
- One-Class SVM

After applying **SMOTE balancing** and validating across algorithms, the best results came from **Logistic Regression**.

Confusion Matrix:

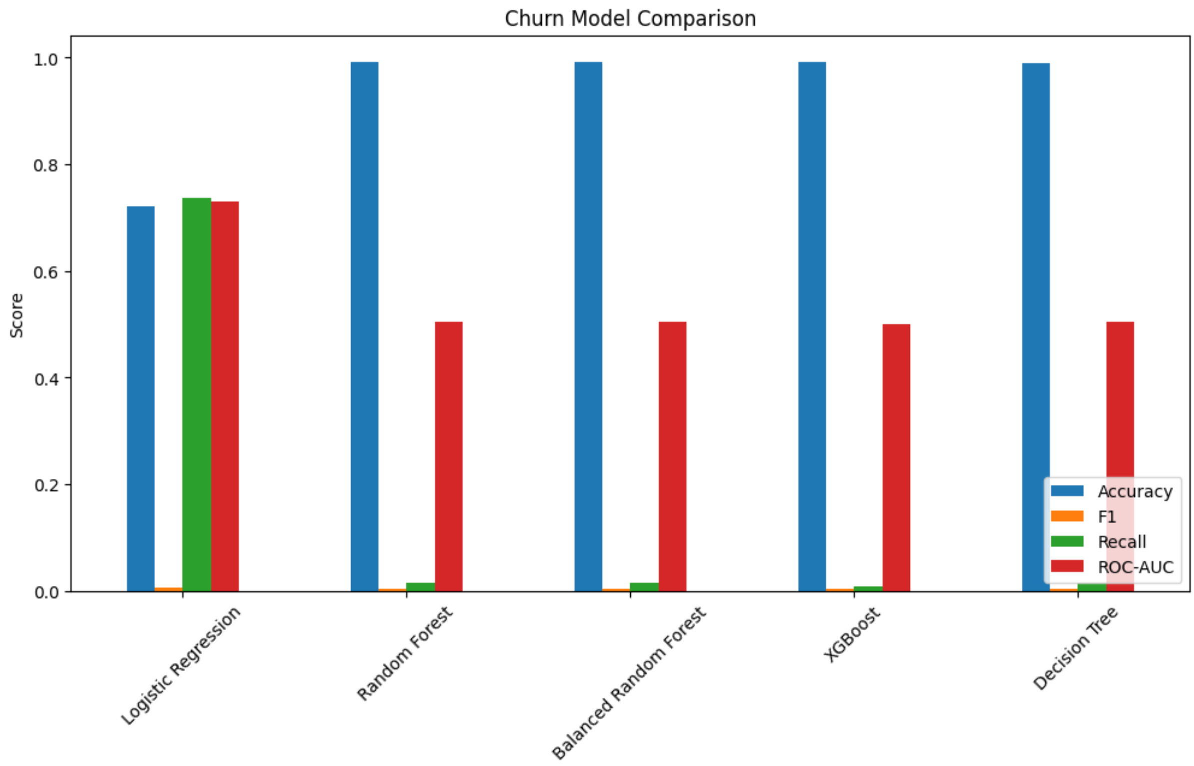
- True Negatives (TN): 72,133
- False Positives (FP): 27,741
- False Negatives (FN): 33
- True Positives (TP): 93

Main Metrics:

- Precision (class 0 – retained customers): 1.00
- Recall (class 0): 0.72
- Precision (class 1 – churn customers): 0.00
- Recall (class 1): 0.74
- Overall Accuracy: 72%
- F1-score (class 1): 0.01

Interpretation:

- The model identifies a significant portion of customers at risk of churn (recall = 74% in the minority class).
- However, precision for churn is very low due to the imbalance of the target variable.
- Despite this, the model meets the strategic goal of detecting at-risk customers, even at the cost of false positives, which can be managed with business strategies.



Influence of User Experience on Churn

Analysis of customer experience variables (with suffix **CX_**) showed:

- Many missing values and low variability (mostly zeros), limiting explanatory power.
- Coefficients must be interpreted cautiously.

Key findings:

- **Satisfaction (NPS):** Negative coefficients indicate higher satisfaction is associated with lower churn.
- **Interactions (CX_):** Negative coefficients don't necessarily imply churn reduction; instead, they reflect that customers without interactions differ from those with them, but due to missing data, patterns are inconsistent.

Conclusion:
Customer experience impacts churn mainly through **NPS**, which is a significant predictor. Operational CX_ variables do not consistently contribute due to missing values and low variability.

8. Retention Strategy

- **Proactive Segmentation:** Identify and prioritize high-risk customers.
- **Personalized Actions:**
    - Discounts or plan upgrades for customers with issues.
    - Priority for new customers and those contacting due to cancellation.
- **Multichannel Management:** Integrate digital and phone support for critical customers.

- **Continuous Monitoring:** Periodic recalibration of the model and data quality control.

9. Best Practices

- Applied **SMOTE** for class balancing.
- Used **stratified validation** for train/test splits.
- Compared models using **recall, F1-score, and precision**.
- Documented and reproducible pipeline.

10. General Conclusion

- The main churn drivers are **low tenure**, **technical support issues**, and **cancellation intent**.
- **NPS** and **Contact Index** show a clear correlation with churn probability, emphasizing the importance of improving customer satisfaction and reducing friction in touchpoints.
- The predictive model anticipates at-risk customers with recall levels suitable for retention strategies, making it a valuable tool for segmentation and prioritization of campaigns, even when false positives must be managed with business criteria.