Customer Churn Classification Model

Table of Contents

1. Dataset Overview

This study was conducted to predict customer churn using data from a telecom operator. The dataset contains **66 features**, and the target variable is `churn` (1: churned customer, 0: active customer).

Key Features:

- **General User Information**: `user_account_id`, `year`, `month`, `user_lifetime`
- **Financial Information**: `user_account_balance_last`, `user_spendings`
- **Usage Information**: `user_no_outgoing_activity_in_days`, `user_has_outgoing_calls`, `user_has_outgoing_sms`
- **Last 100 Usage Records**: `last_100_calls_outgoing_duration`, `last_100_sms_outgoing_count`, `last_100_gprs_usage`

Data Balance:

- **79.1% active (non-churn) customers**
- **20.9% churned customers**
- Since the dataset is **imbalanced**, strategies to handle class imbalance were considered during modeling.

2. Assumptions

- **Recent usage patterns may be crucial for churn prediction.** Variables like `user_no_outgoing_activity_in_days`, `user_spendings`, and `user_has_outgoing_calls` may indicate churn tendencies.

- **Low balance or low spending could signal churn risk.** Customers with lower `user_account_balance_last` and `user_spendings` are more likely to churn.
- **Last 100 call, SMS, and internet usage data may play a critical role in churn prediction.** Identifying whether a user is still active in recent months can be key.

3. Exploratory Data Analysis (EDA) Findings

An analysis of missing values was conducted, and **no missing values** were found in the dataset. All features are complete, and no imputation was required during data preprocessing.

Key Findings:

- **Customers who churn generally have lower balance and lower spending profiles.**
- **Users with long periods of inactivity in calls and SMS have significantly higher churn rates.**
- **Users with low data usage are more likely to churn compared to those with high data usage.**
- **The last 100 transactions play a crucial role in churn prediction. Users with low call duration and SMS count in their last 100 transactions tend to have a higher churn risk.**
- **The initial subscription period (`user_intake`) is an important factor; users with lower spending in the early stages are more likely to churn.**

4. Feature Selection and Engineering

To maximize churn prediction performance, extensive **feature selection and engineering** was applied. Several new variables were created to better model customer behavior. Some key derived variables include:

- **Call and Spending Ratios:** `calls_outgoing_avg_spending` was created to determine the average spending per call duration.
- **SMS Usage Analysis:** `sms_onnet_ratio`, `sms_offnet_ratio`, and `sms_abroad_ratio` were calculated to understand the distribution of sent SMSs across different categories.
- **Data Usage vs. Call Duration Balance:** `data_to_call_ratio` was created to analyze the relationship between a user's data consumption and call duration.
- **Spending Habits:** `reload_to_spending_ratio` was derived to evaluate the relationship between top-up amounts and spending habits.
- **Inactivity Ratios:** Features such as `calls_inactive_ratio`, `sms_inactive_ratio`, and `gprs_inactive_ratio` were created to measure user inactivity over time.

- **Last 100 Usage Trends:** To assess churn risk based on recent activity, features like `last_100_calls_ratio`, `last_100_sms_ratio`, and `last_100_gprs_ratio` were computed.

These derived features significantly enhanced the model's churn prediction accuracy and provided better insights into customer behavior.

5. Hyperparameter Optimization and Modeling

To maximize model performance, **Grid Search CV** and **Bayesian Optimization** were used for hyperparameter tuning. A **Boruta-like algorithm** was employed to identify the most relevant features. Additionally, a **0.05 threshold** was used to minimize overfitting risk.

Best Hyperparameters:

- **Colsample by Tree**: 0.7
- **Enable Categorical**: False
- **Eval Metric**: 'auc'
- **Learning Rate**: 0.01
- **Max Depth**: 8
- **Min Child Weight**: 5
- **Missing**: NaN
- **N Estimators**: 500
- **N Jobs**: -1
- **Random State**: 42
- **Subsample**: 0.8
- **Gamma**: 0.2

These hyperparameters improved the model's performance while reducing overfitting.

6. Model Evaluation

Confusion Matrix:

[[8828 649]

 [811 1712]]

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.93	0.92	9477
1	0.73	0.68	0.70	2523
accuracy			0.88	12000

macro avg 0.82 0.81 0.81 12000

weighted avg 0.88 0.88 0.88 12000

- **Accuracy:** 87.83%
- **Gini Test Score:** 0.8275

7. Model Strength and Segmentation

Customer segmentation was performed to create **proactive intervention plans**. The following segmentation table illustrates different churn risk groups:

Segment	Customer Count	Churn Count	Churn Rate
1	1200	9	0.75%
2	1200	22	1.83%
3	1200	24	2.00%
4	1200	28	2.33%
5	1200	38	3.17%
6	1200	83	6.92%
7	1200	167	13.92%
8	1200	426	35.50%
9	1202	728	60.57%
10	1198	998	83.31%

This segmentation allows businesses to take early actions to reduce churn risk and improve customer retention. Customer segmentation was performed to create **proactive intervention plans**.

8. Future Improvements

- **Real-time predictions**
- **Advanced feature engineering**
- **Model variety**
- **Expanding customer segmentation**
- **Proactive action plans**

9. Summary

This study successfully applied advanced machine learning techniques for **churn prediction**. The model achieved **a Gini Test Score of 0.8275** and **an Accuracy of 87.83%**, demonstrating **strong predictive power**. Customer segmentation helped identify high-risk customers, allowing businesses to take preventive actions.