
# Telecommunication Churn

Classification Project

24-10-2025

**Report By Sourabh Biradar**

## Group members :


| | | |
|---|---|---|
| Sivuni Chandra Mahesh | sivunichandramahesh@gmail.com | 9010386994 |
| Devasetty Kiran | kiru17032003@gmail.com | 9052329561 |
| P R Gururaj | prguru2002@gmail.com | 6302293614 |
| Jay Shirpurwar | jaiishirpurwar@gmail.com | 7387731960 |
| Shubham Dnyaneshwarrao Reche | shubham121.reche@gmail.com | 7066157636 |
| Sourabh Biradar | sourabhbiradar@outlook.com | 9148945067 |

## Overview

High **customer churn** is endemic to the telecommunications sector, with typical annual rates surpassing **10%**. This critical business challenge necessitates **robust client retention strategies**. Our work addresses this through a **binary classification projec**t, focused on predicting the key outcome of whether a customer will **churn or remain loyal**.

## Goals

1. The primary goal is to build, evaluate, and optimize a **binary classification** model that accurately predicts future customer behavior—specifically, distinguishing between **loyal customers** and those who will **churn.**
2. To construct a highly accurate **predictive classification model** (Churn/Loyal) that enables the company to proactively intervene with targeted strategies, thereby reducing the annual **churn rate** (currently >10%) and improving overall business stability.

## Data Source

- Dataset : 'P585 Churn.xlsx'
- Target variable : **'churn'** (no/yes)

## Methodology

### Data Exploration and Preprocessing (EDA)

- Initial Data State : The dataset contained **5000 records and 21 features**.
- Target Imbalance : **'no'** *[4293 (85.86%)]* & **'yes'** *[707 (14.14%)]*
- Cleaning Steps :
    - The dataset had **no missing or duplicate** records.
    - There were **2 miss categorise**d numerical features *'day.charge' & 'eve.mins'* , handled using Pandas in-built method **'to_numeric'.**
    - Every feature had outliers , used **'Manual Clipping**' with limits (lower&upper).
    - *'voice.messages'*,*'intl.calls'* features were heavily **Right skewed** so applied 'log' transformation on them.
    - Further '*voice.messages*' was scaled using RobustScaler as it was still right skewed

    - Ran **chi-square test of independence** on categorical features to determine their association with target variable

- Feature Engineering :

    - Created a **new feature** *['total.mins']* merging all 'mins' features.
    - Likewise created few **new ratio features** *&* dropped original base features.
    - *['eve.charge_per_min']* had some **'inf'** values imputed with '0'.
    - Dropped **highly correlated** features.

- Data Visualization :

    - **Box-plots** : for outliers analysis (pre & post Clipping).
    - **Histograms** : for distribution & skew analysis.
    - **HeatMaps** : to check *Multicollinearity*.
    - **Pair Plots** : for *Univariate (The Diagonal Plots) & Bivariate Analysis (Non-Diagonal plots)*.
    - **Count plots** : for distribution of categorical columns.

- Train-Test split :

    - **test_size=0.2** : Test set *20%* and Train set *80%*.
    - **'straitify=y'** : to ensure that the *target class distribution* is *identical* in both training and testing datasets.
    - Used **SMOTE** technique to handle imbalance in target class *(Over-sampled training set )* inside **pipeline.**

- Preprocessing :

    - **Label Encoded** target variable *'churn'*.
    - **Frequency Encoded** *'state'* feature as it has 51 classes , too much to OneHotEncode as it would lead to overfitting & False weightage if Ordinal Encoded.
    - **Ordinal encoded** remaining categorical features.
    - *Standardization* of numerical features using **Standard Scaler** (Z-score normalization).
    - **Robust Scaled** *'voice.messages'*

## Model training and Hyperparameters tuning

**Used GridSearchCV & RandomizedSearchCV  with cv = 5 (cross validation k-folds)**

**Pipeline >> SMOTE + models**

- Best params for **Logistic Regression** model
  - *C = 0.1 ,*
  - *penalty = 'l1' ,*
  - *solver = 'saga'*
- Best params for **Decision Tree Classifier** model
  - *criterion = 'entropy' ,*
  - *max_depth = 5 ,*
  - *max_features = None,*
  - *min_samples_leaf=15,*
  - *min_samples_split=10*
- Best params for **K-Nearest Neighbors Classifier** model
  - *n_neighbors = 31 ,*
  - *weights = 'distance ,*
  - *algorithm ='auto,*
  - *p =1 ,*
  - *metric='minkowski' (minkowski_distance (l_p) : manhattan_distance (l1) and euclidean_distance (l2))*
- Best params for **Support Vector Classifier**  model
  - *C = 1 ,*
  - *gamma = 'auto' ,*
  - *class_weight= 'balanced'*
  - *kernel = 'rbf'*
- Best params for **Random Forest Classifier** model
  - *n_estimators=300 ,*
  - *max_depth=20 ,*
  - *min_samples_leaf=2 ,*
  - *min_samples_split=5,*
  - *max_features='sqrt'*

- Best params for **XGBoost Classifier** model
    - *n_estimators=200,*
    - *max_depth=15,*
    - *min_child_weight=1,*
    - *gamma=0.0,*
    - *colsample_bytree=1.0,*
    - *subsample=0.9*
- Best params for **LightGBM Classifier** model
    - *n_estimators=200,*
    - *max_depth=7,*
    - *learning_rate=0.1,*
    - *num_leaves = 31,*
    - *subsample=1.0*

## Model evaluation on training set

| **Models** | **ROU-AUC score** |
|---|---|
| Logistic Regression | 0.80 |
| Decision Tree Classifier | 0.89 |
| K-Neighbors Classifier (KNN) | 0.82 |
| Support Vector Classifier (SVC) | 0.87 |
| Random Forest Classifier (Bagging) | 0.919 |
| Extreme Gradient Boosting Classifier (XGB) | 0.924 |
| Light Gradient Boosting Classifier (LightGB) | 0.921 |

NOTE : Used ***'roc_auc'*** as the scoring metric in GridSearchCV & RandomizedSearchCV to find the best combination of hyperparameters

## Model evaluation on test set

### Model Performance Comparison

| Models | ROU-AUC score | Precision | Recall | F1 score |
|---|---|---|---|---|
| Logistic Regression | 0.795 | 0.32 | 0.72 | 0.44 |
| Decision Tree Classifier | 0.918 | 0.73 | 0.69 | 0.71 |
| K-Neighbors Classifier (KNN) | 0.804 | 0.34 | 0.72 | 0.47 |
| Support Vector Classifier (SVC) | 0.89 | 0.51 | 0.81 | 0.63 |
| Random Forest Classifier (Bagging) | 0.912 | 0.87 | 0.77 | 0.82 |
| **Extreme Gradient Boosting Classifier (XGB)** | **0.915** | **0.89** | **0.80** | **0.84** |
| Light Gradient Boosting Classifier (LightGB) | 0.928 | 0.88 | 0.74 | 0.80 |

BEST MODEL : *Extreme Gradient Boosting Classifier (XGB)*

**NOTE** : *Precision* , *Recall & F1 score* are for 'churn' (Class : 1)

### Evaluation Metrics

- **Area Under the ROC Curve (AUC):** A measure of the model's ability to distinguish between the two classes across various thresholds.
- **Recall (Sensitivity):** The ability of the model to correctly identify *all* customers who will churn (minimizing False Negatives, the most costly error).
- **Precision:** The accuracy of the model's positive predictions (how many predicted churners *actually* churned).
- **F1-Score**: The harmonic mean of Precision and Recall.

## Conclusion on Performance :

The *Extreme Gradient Boosting Classifier (XGB)* classifier demonstrated superior performance, achieving the best balance between identifying actual churners (Recall) and maintaining reliable positive predictions (Precision). Its Recall of ***0.80*** means that the model successfully captured ***80%*** of all customers who ultimately left the company.

## Key Feature Importance

The model provided significant insight into the main drivers of customer churn. The top 4 features influencing the prediction were:

1. **total.mins** : High Usage = High Churn Risk
2. **customer.calls** : High calls = High Churn Risk
3. **intl.plan** : active international plan =  High Churn Risk
4. **voice.plan** : active voice plan =  Lesser Churn Risk

## Conclusion and Recommendations

### Project Conclusion

The classification project successfully addressed the challenge of telecommunications customer churn. The optimized ***XGBClassifier*** model provides the company with a powerful tool to predict customer attrition with an effective Recall of ***80%***. This level of predictive capability enables the transition from reactive to proactive client management.

## Business Recommendations

Based on the model results and feature analysis, the following actions are recommended for the telecom company:

1. **Targeted Retention Campaigns** :
   Deploy personalized retention offers (e.g., loyalty discounts, custom plans) for customers identified with *high total usage minutes and frequent customer service calls*, as these segments show a higher likelihood of churn. Tailored engagement and proactive support may reduce dissatisfaction among heavy users.
2. **Contract Structure Review :**
   Focus retention efforts on customers with Month-to-Month contracts by offering incentives to switch to long-term contracts (e.g., annual or biannual plans). Longer commitments help stabilize the customer base and reduce churn volatility.
3. **Plan Optimization Strategy** :

   Customers with an *active international plan exhibit higher churn risk* — likely due to high costs or dissatisfaction with international call rates. Reassess pricing or provide competitive international bundles to retain this segment.

   Conversely, customers with an *active voice plan are less likely to churn*. Promote or cross-sell voice plans to at-risk customers as part of a retention bundle.

4. **Service Enhancement :**
   Enhance the quality and reliability of customer support and technical services, especially for users making *frequent customer calls*. Quick resolution and personalized care can directly reduce churn among this group.
5. **Early Intervention Program :**
   Implement a “first 90 days” loyalty program for new or low-tenure customers. Proactive communication, onboarding support, and early engagement can help prevent churn before dissatisfaction develops.