TELECOM CHURN ANALYSIS

INTRODUCTION

The Telecom Churn Prediction Project aims to identify customers who are likely to discontinue (churn) their telecom services. The main purpose of this project is to analyse customer data — including call usage, SMS activity, internet consumption, and account information — to understand the factors that influence customer retention and churn behaviour.

The classification objective is to predict whether a customer will stay (0) or churn (1) using machine learning algorithms. By training models on historical data, the system can classify new or existing customers based on their likelihood to leave the service.

In the real world, churn prediction is vital for telecom companies because retaining existing customers is more cost-effective than acquiring new ones. Early identification of potential churners enables businesses to take proactive measures such as offering discounts, improving service quality, or providing personalized plans — ultimately enhancing customer satisfaction and reducing revenue loss.

DATASET DESCRIPTION

The dataset used in this project is the Telecom Churn Dataset (telecom_churn.csv), which contains customer-related information collected from a telecom service provider. It includes demographic details, usage behaviour, and account information, which are used to predict whether a customer will leave (churn) or stay.

- Source:
  The dataset is collected from a telecom company's customer database (from open-source platform Kaggle).
- Total Records:
  The project uses 50,000 customer records for training and testing after data cleaning and preprocessing.
- Total Features (Attributes):
  There are 14 Total features.
- Input Variables:
  Input columns include a unique customer identifier, the telecom partner, gender, age, state, city, pincode, date of registration, number of dependents, aggregated data used, aggregated calls made.
- Target (Class) Variable:
  The churn column is the target variable, representing the classification label:

  - 1 → Churn (customer leaves the service)
  - 0 → Stay (customer remains subscribed)

DATA CLEANING

Data cleaning is an essential step in preparing the dataset for accurate and reliable model training. In this project, several preprocessing steps were performed to ensure data quality and consistency.

1. Handling Missing or Null Values

The dataset is generally clean with negligible missing values and no large-scale null problems.

2. Removal of Duplicate or Irrelevant Data

- Irrelevant identifiers were removed, and categorical variables were encoded into numerical format using appropriate encoders. Numerical features were standardized using scaling techniques to improve model performance. To address class imbalance, oversampling techniques like SMOTE were considered. The dataset was then split into training and testing subsets to ensure fair model evaluation..

EXPLORATORY DATA ANALYSIS (EDA)

After data cleaning, exploratory data analysis (EDA) was conducted to understand the structure, trends, and relationships in the telecom churn dataset.

Summary Statistics

- Descriptive statistics were generated for all numerical features to identify data distribution, outliers, and variation among customers.
- The churn target variable showed a clear **class imbalance**, with more customers staying than leaving.
- Analysis indicated that customers with shorter tenure and higher monthly charges were more likely to churn.

Visualization of Features and Target Distribution

- A count plot of churn visualized the imbalance between retained and churned customers.
- Histograms of key numerical variables such as **tenure**, **monthly charges**, and **total charges** highlighted distinct spending and usage patterns between the two groups.
- A correlation heatmap revealed that **tenure**, **contract type**, and **monthly charges** were among the most influential factors associated with customer churn.

DATA PREPROCESSING

- **Categorical Encoding:** All categorical features were converted to numerical values using LabelEncoder to make them compatible with machine learning models.'
- **Feature Scaling:** Numerical features were standardized using StandardScalar to ensure uniform contribution across features and improve model performance.
- **Train–Test Split:** The dataset was split into 80% training and 20% testing sets using to evaluate model performance on unseen data.

MODEL BUILDING AND EVALUATION

To predict customer churn effectively, several supervised machine learning classification algorithms were implemented and compared. The goal was to identify the best-performing model that can accurately classify customers as “Churn” (1) or “Stay” (0) based on behavioural and demographic features.

Several machine learning algorithms were implemented and compared, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost ,CatBoost and LightBoost Classifier. Each model was trained on the preprocessed dataset, and performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

MODEL EVALUATION

To assess how well each classification algorithm predicted customer churn, several evaluation metrics were used. These metrics provide a comprehensive understanding of the model’s performance in correctly identifying both churned and retained customers.

Accuracy

Accuracy measures the proportion of correctly predicted observations (both churned and non-churned) out of the total number of observations.

Precision
Measures how many customers predicted as churn are actually churners.

Recall
Measures how many actual churners were correctly identified by the model.

F1-Score
The harmonic mean of precision and recall, providing a balance between both.

Confusion Matrix
A 2×2 table summarizing the number of correct and incorrect predictions made by the classifier.

OPTIMIZATION TECHNIQUES

To improve model accuracy and robustness, several optimization methods were applied. **RandomizedSearchCV** was used for hyperparameter tuning, which efficiently optimized parameters such as learning rate, depth, and estimators to enhance model performance while avoiding overfitting. **Cross-validation** ensured that the results were consistent across different subsets of data, strengthening the model’s reliability.

Additionally, **feature selection** helped focus on the most significant attributes, including customer tenure, contract type, monthly charges, and payment method — all of which played key roles in predicting churn. These steps collectively improved accuracy, interpretability, and the overall stability of the final CatBoost model.

1. Hyperparameter Tuning (RandomizedSearchCV)

RandomizedSearchCV randomly samples a subset of parameter combinations, offering faster and more efficient optimization than GridSearchCV. Typical parameters tuned include learning rate, depth, and estimators for tree-based models, and C, penalty, or kernel parameters for Logistic Regression and SVM. This improved accuracy and F1-score while reducing overfitting.

2. Cross-Validation (Integrated with RandomizedSearchCV)

Although not explicitly coded with KFold, the cross-validation mechanism was integrated within RandomizedSearchCV. This ensured the model performance was validated across multiple data folds before finalizing hyperparameters. This Prevents overfitting on training data. It also Provides a more reliable estimate of real-world model performance.

3. Feature Engineering and Selection

Important features such as tenure, contract type, monthly charges, and payment method were prioritized, while less significant variables were removed. This improved computational efficiency, model interpretability, and reduced noise in predictions.

CONCLUSION

After evaluating multiple machine learning algorithms — including Logistic Regression, Random Forest, Support Vector Machine (SVM), Neural Network, XGBoost, LightGBM, and CatBoost — the final selected model for telecom churn prediction is:

CatBoost Classifier

- Achieved the highest overall accuracy (87%) and F1-score (0.82).
- Showed excellent precision-recall balance, effectively identifying churners without generating excessive false positives.
- Required minimal preprocessing and handled categorical variables automatically.
- Demonstrated robust generalization during cross-validation and hyperparameter tuning using RandomizedSearchCV.

Key Findings and Insights

The analysis of customer behaviour revealed that tenure, contract type, and payment method play major roles in churn. Customers with short-term or month-to-month contracts and those paying through electronic checks were more likely to leave, whereas long-term subscribers with bundled services showed higher loyalty. High monthly charges were also identified as a strong driver of churn. These findings can help telecom companies focus on customer retention by offering discounts, loyalty programs, and flexible plans.