# **Telecom Customer Churn - Predicting behavior to Retain Customer**

**ABSTRACT**

Customer churn is a pressing issue for telecom companies, impacting their revenue and growth. This project aims to predict customer churn by analyzing a telecom customer dataset, employing data preprocessing techniques, and applying machine learning models. The dataset includes various customer attributes, such as demographics, account details, and service usage patterns. Key preprocessing steps involve handling missing values, encoding categorical variables, and feature scaling. The project utilizes logistic regression and random forest classifiers to develop predictive models. The performance of these models is evaluated using metrics like accuracy, precision, recall, F1 score, and confusion matrix. The project also addresses class imbalance with oversampling techniques. Results indicate that the models can accurately predict churn, with random forest providing robust performance and feature importance insights. Key factors influencing churn include contract type, monthly charges, and tenure. By identifying at-risk customers, telecom companies can implement targeted retention strategies, ultimately enhancing customer loyalty and reducing churn rates. This project not only demonstrates the application of machine learning in churn prediction but also provides actionable insights for improving customer retention in the telecom industry.

**INTRODUCTION**

In the competitive landscape of the telecom industry, customer churn is a critical issue as it directly affects the profitability and sustainability of a company. Customer churn refers to the phenomenon where customers stop using a company’s services. The ability to predict which customers are likely to churn and understand the factors leading to churn is invaluable for telecom companies. This knowledge allows for targeted interventions to retain at-risk customers, thereby enhancing customer loyalty and reducing revenue loss. This project focuses on using data analytics and machine learning to predict customer churn based on various customer attributes.

**A) PROBLEM DEFINITION**

The primary problem addressed in this project is to predict whether a customer will churn or not based on their historical data and current status with the telecom company. This involves analyzing customer behavior patterns and identifying key indicators that lead to churn. By building predictive models, the project aims to provide actionable insights for customer retention strategies.

**Aim of this project**

The main aim of this project is to develop a robust predictive model that can accurately identify customers who are likely to churn. Additionally, the project seeks to understand the key factors influencing churn and provide recommendations for reducing churn rates. The ultimate goal is to enable telecom companies to take proactive measures in retaining customers and improving overall customer satisfaction.

**B) BACKGROUND STUDY**

Understanding customer churn and its implications is essential for developing effective retention strategies. The background study involves exploring the concept of churn, its impact on the telecom industry, and reviewing existing methodologies for churn prediction. Customer churn is influenced by various factors, including service quality, pricing, customer service, and competition. Previous studies have employed statistical methods and machine learning techniques to predict churn, providing a foundation for the current project.

*Related works*

Several studies have explored the use of different machine learning algorithms for churn prediction. Techniques such as logistic regression, decision trees, random forests, and neural networks have been used to build predictive models. These studies highlight the importance of feature selection, data preprocessing, and handling class imbalance in developing accurate churn prediction models. By reviewing related works, this project builds upon existing knowledge and identifies the best practices for churn prediction.

**C) BACKGROUND STUDY**

In this section, the focus is on a deeper understanding of customer churn specifically in the telecom industry. Telecom companies often face high churn rates due to intense competition, evolving customer preferences, and technological advancements. Factors such as customer service quality, billing issues, service outages, and pricing plans play a significant role in customer retention. By studying these factors, the project aims to identify key indicators of churn and develop strategies to mitigate them.

*Related works*

Further exploration of existing literature reveals various approaches to churn prediction. Studies have shown that combining multiple data sources, such as usage patterns, demographic information, and customer feedback, can improve prediction accuracy. Additionally, techniques like oversampling, under sampling, and ensemble methods have been used to address class imbalance in churn prediction. These insights inform the selection of models and methodologies used in this project.

**D) OBJECTIVES AND CONTRIBUTIONS**

*Objectives:*

1. Develop a predictive model for customer churn using telecom customer data.

2. Identify key factors influencing customer churn.

3. Provide actionable insights and recommendations for reducing churn rates.

4. Compare the performance of different machine learning models.

*Contributions:*

1. A comprehensive analysis of telecom customer data.

2. Development and evaluation of predictive models.

3. Identification of significant predictors of customer churn.

4. Practical recommendations for churn mitigation strategies.

**DATA PREPROCESSING**

Data preprocessing is a crucial step in building a predictive model. The dataset used in this project contains various customer attributes, including demographic information, account details, and service usage patterns. The preprocessing steps include:

**1. Data Import and Exploration:** Importing the dataset and performing an initial exploration to understand the structure and content.

**2. Handling Missing Values:** Using techniques like SimpleImputer to handle missing values in the dataset.

**3. Encoding Categorical Variables:** Converting categorical variables into numerical format using Label Encoder.

**4. Converting Data Types:** Ensuring all relevant columns are in the appropriate numerical format.

**5. Feature Scaling:** Standardizing the features to ensure consistent scales for model training.

**6. Splitting the Dataset:** Dividing the data into training and testing sets for model evaluation.

**DESCRIPTIVE STATISTICS**

Descriptive statistics provide a summary of the dataset, helping to understand the distribution and relationships between variables. Key steps include:

**1. Summary Statistics:** Calculating mean, median, mode, and standard deviation for numerical variables.

**2. Distribution Plots:** Visualizing the distribution of variables using histograms and KDE plots.

**3. Correlation Analysis:** Using heatmaps to visualize correlations between variables and identify significant predictors of churn.

**MODEL TRAINING**

The project involves training various machine learning models to predict customer churn. The models used include:

**1. Logistic Regression:** A statistical model used for binary classification problems. It estimates the probability of a binary outcome based on input features.

**2. Random Forest Classifier:** An ensemble method that builds multiple decision trees and combines their predictions for improved accuracy and robustness.

**MODEL EVALUATION**

Model evaluation is critical to assess the performance and reliability of the predictive models. The evaluation metrics used include:

**1. Accuracy:** The proportion of correctly predicted instances out of the total instances.

**2. Precision and Recall:** Metrics that evaluate the model’s performance in predicting positive instances.

**3. F1 Score:** The harmonic mean of precision and recall, providing a single metric for model performance.

**4. Confusion Matrix:** A table that summarizes the performance of the model by showing true positives, false positives, true negatives, and false negatives.

**5. Classification Report:** A detailed report that includes precision, recall, F1 score, and support for each class.

**E) METHODOLOGY**

The methodology involves a systematic approach to data preprocessing, model training, and evaluation. Key steps include:

**1. Data Preprocessing:** Handling missing values, encoding categorical variables, feature scaling, and splitting the dataset.

**2. Feature Engineering:** Creating new features or modifying existing ones to improve model performance.

**3. Model Selection:** Choosing appropriate machine learning models based on the problem and dataset characteristics.

**4. Handling Class Imbalance:** Using techniques like RandomOverSampler to balance the dataset and improve model performance.

**5. Model Training:** Training the selected models on the preprocessed data.

**6. Model Evaluation:** Evaluating the models using appropriate metrics and visualization techniques.

**F) MODEL DESCRIPTION**

**Logistic Regression:**

*-Theoretical Background:* Logistic regression is used for binary classification problems. It models the probability of a binary outcome based on input features using the logistic function.

*- Implementation:* The model is trained using the training dataset and evaluated using the testing dataset. Feature scaling is applied to ensure consistent scales for the input features.

- *Advantages:* Logistic regression is easy to implement, interpretable, and works well for binary classification problems.

- *Limitations:* It assumes a linear relationship between the input features and the log-odds of the outcome.

**Random Forest Classifier:**

*- Theoretical Background*: Random forest is an ensemble method that builds multiple decision trees and combines their predictions. It uses bootstrap sampling and random feature selection to create diverse trees.

- *Implementation:* The model is trained using the training dataset with feature scaling applied. It is evaluated using the testing dataset.

- *Advantages:* Random forest is robust to overfitting, handles high-dimensional data well, and provides feature importance metrics.

- *Limitations:* It can be computationally intensive and less interpretable compared to simpler models.

**CONCLUSION**

Customer churn prediction is a critical task for telecom companies aiming to retain customers and reduce revenue loss. By leveraging data preprocessing techniques and machine learning models, this project successfully predicts customer churn and identifies key factors influencing churn. The insights gained from this analysis provide valuable recommendations for churn mitigation strategies, helping telecom companies improve customer retention and satisfaction. The project demonstrates the effectiveness of logistic regression and random forest classifiers in predicting churn, highlighting the importance of data preprocessing and feature scaling in developing accurate predictive models.