

Project Report: Telco Customer Churn Analysis

Link: <https://colab.research.google.com/drive/1OekNgDMhyN8zC8L3AYxLwyaB10MnOk1V?usp=sharing>

1. Executive Summary

The objective of this project was to analyze a telecommunications dataset to predict customer churn. Using various machine learning models and data preprocessing techniques, the study aimed to identify patterns and develop predictive models for customer retention strategies.

2. Problem Statement

The dataset contains customer demographics, services, and billing details, along with whether they churned. The goal is to predict churn and understand key drivers behind customer retention.

3. Data Description

The dataset consists of attributes such as:

- **Demographic details:** Gender, age, dependents, etc.
 - **Service details:** Internet services, online security, tech support, etc.
 - **Billing details:** Monthly charges, total charges, and payment methods.
 - **Target Variable:** Churn (Yes/No).
-

4. Data Cleaning and Preparation

- **Handling Missing Data:** Missing values in the TotalCharges column were replaced with the median.
 - **Encoding:** Categorical variables were encoded using techniques like Label Encoding.
 - **Feature Scaling:** StandardScaler was applied to normalize numeric features.
 - **Outlier Handling:** Boxplots identified and removed outliers.
 - **Balancing Classes:** Oversampling with SMOTE ensured class distribution balance for churn.
-

5. Exploratory Data Analysis (EDA)

- **Visualizations:**

- Histograms for data distribution.
 - Correlation heatmap to identify relationships.
 - Count Plots for churn distribution pre- and post-oversampling.
 - **Insights:**
 - Key features like tenure, contract type, and monthly charges have significant correlations with churn.
-

6. Models Evaluated

1. **Gradient Boosting Classifier:**
 - **Accuracy:** 0.83
 - Corrects errors iteratively for better predictions.
 2. **XGBoost Classifier:**
 - **Accuracy:** 0.83
 - Efficient and scalable for large datasets.
 3. **Random Forest Classifier:**
 - **Accuracy:** 0.83
 - Reduces overfitting by averaging decision trees.
 4. **Logistic Regression:**
 - **Accuracy:** 0.80
 - Provides baseline performance.
 5. **K-Nearest Neighbors (KNN):**
 - **Accuracy:** 0.80
 - Relies on feature proximity.
-

7. Key Findings

- The XGBoost model showed the best performance (83.3% accuracy), making it suitable for churn prediction.
 - Customer tenure and contract types strongly influence churn likelihood.
 - Data preprocessing, including outlier removal and class balancing, significantly improved model performance.
-

8. Recommendations

- Focus on retaining customers with month-to-month contracts.
- Improve engagement for customers with lower tenure to reduce churn.
- Optimize pricing and service bundling strategies based on customer preferences.
-

Conclusion

This project identified key churn predictors and developed a robust model for prediction. By implementing the recommendations, the company can retain more customers, reduce costs, and improve profitability.