Telecom Customer Churn Prediction

Using Machine Learning to Reduce Customer Attrition

A Comprehensive ML Pipeline for Predictive Analytics

Presented by: Masna Hansika
Date: 10 Nov 2025

Problem Statement

The Challenge

Customer churn costs the telecom industry **billions of dollars annually**. Acquiring new customers is 5-25x more expensive than retaining existing ones.

Our Solution

- **Predictive Analytics:** Identify at-risk customers before they leave
- **Machine Learning:** 6 advanced algorithms trained and compared
- **Actionable Insights:** Data-driven retention strategies
- **Real-time Prediction:** Production-ready deployment system

88%
Model Accuracy

$XXM
Potential Savings

1000+
Customers Analyzed

📁 Dataset Overview

1000
Customers

21
Features

~27%
Churn Rate

5
Categories

Feature Categories

👤

Demographics (5)
Gender, Senior Citizen, Partner, Dependents, Customer ID

📱

Services (9)
Phone, Internet, Security, Backup, Protection, Tech Support, Streaming

💼

Account Info (6)
Tenure, Contract, Billing, Payment, Monthly & Total Charges

🎯

Target Variable (1)
Churn - Whether customer left the service (Yes/No)

Data Source: Telco Customer Churn Dataset (Kaggle) / Generated Sample Dataset

Project Pipeline

End-to-End Machine Learning Workflow

Stage 1
Data Collection

↓

Stage 2
Preprocessing

↓

Stage 3
EDA

↓

Stage 4
Model Training

↓

💡

Stage 5
Insights

Key Pipeline Features

- ▶ Automated data quality checks and validation
- ▶ Advanced feature engineering and transformation
- ▶ Comprehensive exploratory data analysis
- ▶ Multi-model training and comparison
- ▶ Business-focused insights extraction

🔧 Data Preprocessing

1. Missing Value Treatment

- ▸ TotalCharges: Converted to numeric, imputed with median
- ▸ Numerical features: Median imputation
- ▸ Result: 0% missing values in final dataset

2. Feature Engineering

➕

Created Features

ChargePerTenure: Monthly charges divided by tenure
TenureGroup: Categorized tenure into 4 groups

🔢

Encoding

Binary: Yes/No → 1/0 (15 features)
One-Hot: Multi-category features

⚖️

Scaling

StandardScaler: Normalized all numerical features
Mean: 0, **Std Dev:** 1

Process	Before	Af ter
Shape	1000 × 21	1000 × 45+

Missing Values	20 rows	0 rows
Categorical Features	15 features	Fully encoded

📊 Exploratory Data Analysis

Key Findings

1. Churn Distribution

73%
No Churn

VS

27%
Churn

2. Top Correlations with Churn

- **Month-to-month contract:** +0.45 correlation
- **Fiber optic internet:** +0.31 correlation
- **Low tenure:** -0.35 correlation
- **Electronic check payment:** +0.28 correlation

Visual Analysis Generated

📊

Distribution Plots
Churn by contract type, tenure, charges

🔥

Correlation Heatmap
Feature relationships and dependencies

📦

Box Plots

Numerical feature distributions by churn

💡 Insert screenshots from your actual EDA visualizations here

🤖 Model Training & Comparison

Six Algorithms Trained

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	**0.8200**	**0.7100**	0.5800	0.6400	**0.8800**
XGBoost	0.8100	0.7000	0.5600	0.6200	0.8700
LightGBM	0.8050	0.6900	0.5500	0.6150	0.8650
Gradient Boosting	0.8000	0.6800	0.5400	0.6050	0.8600
Logistic Regression	0.7950	0.6700	0.5400	0.6000	0.8400
Decision Tree	0.7500	0.6000	0.5000	0.5450	0.7800

🏆 Best Model: Random Forest

Selected based on **ROC-AUC score of 0.88**, achieving the best balance between identifying churners (recall) and minimizing false positives (precision).

💡 Replace with your actual model results from the execution

🎯 Best Model Performance

Random Forest Classifier Results

82%
Accuracy

71%
Precision

58%
Recall

0.88
ROC-AUC

What This Means

✅ **Accuracy (82%)**
Model correctly predicts 82% of all customers

🎯 **Precision (71%)**
71% of predicted churners actually churn (low false alarms)

🔍 **Recall (58%)**
Model catches 58% of actual churners

📈 **ROC-AUC (0.88)**
Excellent discrimination between classes