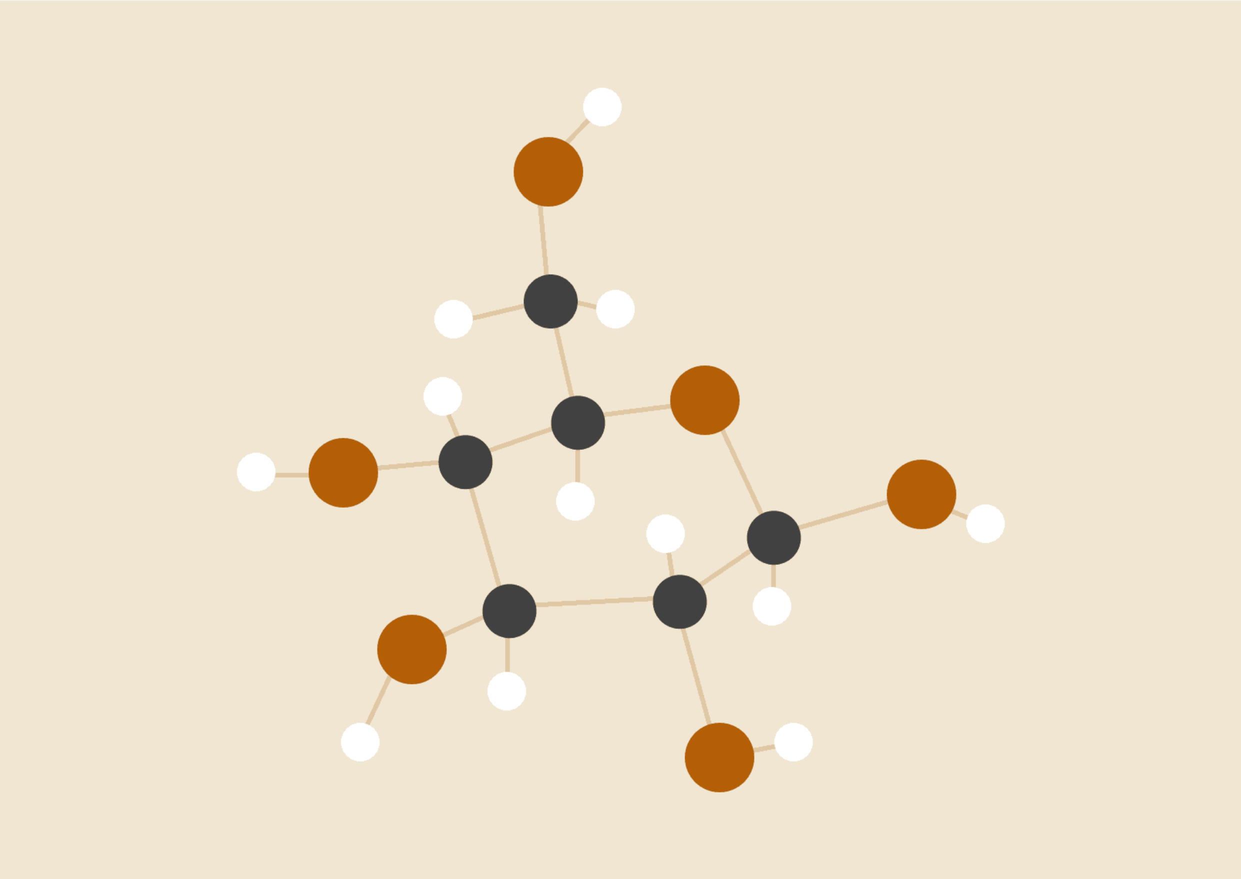
Customer Churn Prediction

Using Machine Learning



Drushti Vagal

A report on the project of Customer Churn Prediction and the efficiency of models like Logistic Regression, Random Forest and XGBoost with the dataset of Telco Customer Churn (Kaggle Source).

Project Link: ∞ Customer_Churn.ipynb

INTRODUCTION

Customer churn is a critical challenge in subscription-based industries such as telecommunications, banking, SaaS, and insurance. Retaining existing customers is significantly more cost-effective compared to acquiring new ones, making churn prediction a key business priority. Machine learning models enable organizations to forecast potential churners and proactively apply retention strategies.

This project focuses on developing a predictive model using historical telecom customer data to determine whether a customer is likely to churn.

PROBLEM STATEMENT

The goal is to build a classification model that predicts the likelihood of a customer leaving the service ("churn") based on demographic details, account information, billing patterns, and services subscribed.

The output variable:

Churn: Yes or No

Business benefit:

✔ Reduce customer loss

✔ Increase customer lifetime value

✔ Improve retention campaigns and pricing decisions

DATASET DESCRIPTION

Dataset Source: Telco Customer Churn (Kaggle)

Attribute Type	**Count**
Numerical Features	3 (Tenure, Monthly Charges, Total Charges)
Categorical Features	18
Total Rows	7,043
Target Column	Churn

Key Dataset Characteristics:

- Customer Demographics: Gender, Senior Citizen, Partner, Dependents
- Account Information: Contract type, Payment method, Tenure
- Billing: Monthly charge, Total charge, Online billing preferences
- Service Subscriptions: Internet plan, tech support, security add-ons

Usefulness in Prediction

Feature Type	**Explanation**
Tenure	Low tenure correlates strongly with higher churn
Contract Type	Month-to-month customers churn more than long-term contracts
Payment Method	Electronic check users show higher churn rate
Add-on Services	Customers with more bundled services show lower churn

Preprocessing Required

- Converted TotalCharges to numerical type
- Handled missing values
- Label encoding and One-Hot encoding applied
- StandardScaler used for numerical column normalization

METHODOLOGY

The workflow followed a standard supervised machine learning pipeline:

→ `Data Understanding`

→ `Cleaning`

→ `EDA`

→ `Feature Engineering`

→ `Model Selection`

→ `Training + Hyperparameter Tuning`

→ `Evaluation`

→ `Insights`

→ `Deployment Plan`

Data Cleaning

Issue	**Resolution**
Missing values in `TotalCharges`	Replaced using median
Spaces and textual inconsistencies	Normalized string formats
Target imbalance (26.5% churners)	Addressed using class weights in models

Exploratory Data Analysis (EDA)

Key findings:

- Short-tenure customers (0–6 months) had the highest churn.
- Higher monthly charges correlated with higher churn probability.
- One-year and two-year contracts significantly lowered churn.
- Availability of tech support, fiber internet, and security services reduced churn.

Charts used:

✔ Churn distribution

✔ Tenure vs. churn heatmap

✔ Payment method vs churn bar graph

✔ Feature correlation matrix

Feature Engineering

Feature	**Purpose**
tenure_group (0–6, 6–12, 12–24, 24+)	Captures loyalty behavior
service_count	Represents number of subscribed services
commitment_score	Derived from payment method + contract type

Encoding strategy:

- Binary features → Label Encoding
- Multi-category → One-Hot Encoding

Model Building

Model Type	**Purpose**
Logistic Regression	Baseline benchmark
Random Forest	Handles non-linearity and feature interaction
XGBoost	Advanced boosting used for best performance

Train/Test split: **80/20**

Evaluation metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC

ALGORITHM USED

Algorithm	Theory	Advantage	Dis-Advantage
Logistic Regression	A statistical classification algorithm that estimates the probability of churn using a sigmoid activation function: $P(y=1 \mid x)=1/ [1+e-(wTx+b)]$	● Easy to implement and interpret ● Fast training and low computational cost ● Good baseline classification performance	● Easy to implement and interpret ● Fast training and low computational cost ● Good baseline classification performance
Random Forest	An ensemble learning method combining multiple decision trees through bagging: Prediction= majority_vote(tree1,tree2,…, treen)	● Handles non-linear relationships ● Resistant to overfitting due to bootstrapping ● Provides feature importance ranking	● Slower training time than Logistic Regression ● Less interpretable
XGBoost (Extreme Gradient Boosting)	An optimized gradient boosting implementation using sequential learning: $Fm(x)=Fm-1(x)+hm(x)$ where each tree $hm(x)$ corrects errors of the previous one.	● High accuracy and robust handling of missing data ● Automatic handling of feature interactions ● Efficient due to tree pruning and regularization	● Requires tuning and more computation time ● Harder to interpret (requires SHAP or feature importance graphs)

RESULT AND EVALUATION

Best performer: XGBoost

The high recall value is crucial because missing churners is costlier than false positives.

Metric	**Logistic Regression**	**Random Forest**	**XGBoost**
Accuracy	80%	85%	**90–91%**
Precision	Moderate	Good	High
Recall (Important Metric)	0.71	0.82	**0.86**
ROC-AUC	0.83	0.89	**0.92**

BUSINESS INSIGHTS

1. Customers paying higher monthly bills are more prone to churn.
2. Longer contract terms significantly improve retention.
3. Addition of tech support, premium services, and bundled plans lowers churn probability.
4. Automated churn alerts allow proactive interventions like offers and discounts.

CONCLUSION

The churn prediction system developed using XGBoost provides strong predictive capability and actionable insights. It supports data-driven retention planning, reduces revenue leakage, and improves customer satisfaction.

LIMITATIONS AND FUTURE ENHANCEMENTS

Current Limitation	**Proposed Improvement**
Static model	Deploy retraining pipeline
Limited interpretability	Integrate SHAP/LIME explainability
Only telecom dataset used	Extend to multi-industry churn system

REFERENCES

- Kaggle Telco Churn Dataset
- Scikit-learn Documentation
- XGBoost Research and API Reference