Customer Churn Prediction Report

(with Python and Machine Learning)

Name: Shivika Gupta
Institution: Gokhale Institute of Politics and Economics
GitHub: github.com/shivikagupta21
LinkedIn: linkedin.com/in/shivika-gupta-6a5ab3217
Dataset Size: 7,043 customer records
Date: 18th August, 2025

I. Introduction

Customer churn, also known as customer attrition, refers to the phenomenon where customers discontinue their relationship with a company by canceling subscriptions, switching to competitors or ceasing purchases altogether. In the telecom industry, churn is a critical challenge because customers have **low switching costs** and **multiple alternative providers** to choose from. High churn rates directly impact a company's **revenue stability, profitability and growth prospects**. Studies show that the **cost of acquiring a new customer is 5–10 times higher** than retaining an existing one. Thus, retaining customers is far more cost-effective than constantly acquiring new ones. Churn

prediction is therefore a vital business problem. By leveraging **data-driven predictive analytics**, companies can identify customers who are most likely to leave, understand the reasons behind their dissatisfaction and take proactive measures to retain them. This project applies **machine learning models** to the Telco Customer Churn dataset to predict churn probability. The process involves:

1. Data cleaning and preprocessing
2. Exploratory data analysis (EDA) to identify churn patterns
3. Training predictive models (Decision Tree, Random Forest and XGBoost)
4. Evaluating their performance
5. Extracting actionable business insights.

II. Research Objective

The primary goals of this project are:

1. To build a churn prediction model using machine learning algorithms.
2. To evaluate models on accuracy, precision, recall, F1-score and ROC-AUC.
3. To identify key factors driving churn.
4. To provide business insights and actionable strategies.

III. Dataset and Columns introduction

Dataset Overview:

- Rows: **7,043**
- Columns: **21**

Column Name	Data type	Description
gender	Text	Male/Female
SeniorCitizen	Integer	Indicates if the customer is a senior citizen
Partner	Text	Whether the customer has a partner
Dependents	Text	Whether the customer has dependents
tenure	Integer	Number of months with the company

PhoneService	Text	Subscription to phone service
InternetService	Text	Internet type (DSL, Fiber optic, None)
OnlineService	Text	Availability of online security service
TechSupport	Text	Availability of technical support
Contract	Text	Contract type
PaymentMethod	Text	Payment mode
MonthlyCharges	Float	Monthly billing charges
TotalCharges	Float	Lifetime charges paid
Churn	Text	Target variable (Yes/No)

IV. Theory

Churn prediction is framed as a **binary classification problem** where the target variable (Churn) has two classes: **Yes (customer leaves)** and **No (customer stays)**. Since the dataset is **imbalanced (73% No, 27% Yes)**, the **SMOTE oversampling technique** was applied to ensure balanced learning.

Algorithms Applied

1. **Decision Tree Classifier** – simple and interpretable, but prone to overfitting.
2. **Random Forest Classifier** – ensemble of decision trees, robust and accurate, reduces overfitting.
3. **XGBoost Classifier** – gradient boosting method, highly accurate but more complex.

Evaluation Metrics

Model performance was assessed using:

1. **Accuracy** (overall correctness)
2. **Precision** (validity of churn predictions)
3. **Recall** (ability to detect actual churners)
4. **F1-Score** (balance of precision and recall)

5. **ROC-AUC** (discriminative power between churners and non-churners)

V. Data Cleaning and Transformation

Before building models, several preprocessing steps were applied to ensure data quality and suitability for machine learning:

1. **Column Removal:** Dropped the customerID column as it served only as an identifier and provided no predictive value.
2. **Handling Missing Values:**
   a. Detected blank values in the TotalCharges column.
   b. Replaced blanks with 0.0 and converted the column from string to **numeric (float)** for analysis.
3. **Data Type Standardization:** Ensured numerical columns (tenure, MonthlyCharges, TotalCharges) were correctly typed as integers/floats.
4. **Categorical Encoding:** Transformed categorical features (gender, Partner, Dependents, InternetService, Contract, PaymentMethod) into numerical format using **Label Encoding**.
5. **Target Class Balancing:**
   a. Observed class imbalance in the target variable Churn (73% No, 27% Yes).
   b. Applied **SMOTE (Synthetic Minority Oversampling Technique)** to synthetically generate minority class samples and balance the dataset.
6. **Feature Inspection**
   a. Verified all columns for consistency, removed redundancies and checked distributions.
   b. Identified key numerical variables (tenure, MonthlyCharges, TotalCharges) and categorical service-related variables as potential churn predictors.
7. **Train-Test Split:** Partitioned the dataset into **80% training** and **20% testing** to evaluate model generalization.

VI. Key Performance Indicators (KPIs)

KPI	Value
Overall Churn Rate	26.5%

Average Tenure	32.4 months
Average Monthly Charges	$64.8
Average Total Charges	$2,283

VII. Model Development & Comparison

Cross-validation and Test Accuracy

Model	CV Accuracy	Test Accuracy	Precision (Churn)	Recall (Churn)	F1 (Churn)	ROC-AUC
Decision Tree	0.78	0.77	0.58	0.58	0.58	~0.75
Random Forest	0.84	0.78	0.85	0.85	0.85	0.82
XGBoost	0.83	0.77	0.80	0.73	0.76	~0.81

Random Forest performed best overall.

VIII. Results: Random Forest

Confusion Matrix

	Predicted No	Predicted Yes
Actual No	880	156
Actual Yes	158	215

Classification Report

Metric	No Churn	Churn	Overall

Precision	0.85	0.58	0.78
Recall	0.85	0.58	0.78
F1-Score	0.85	0.58	0.78
Accuracy			77.7%

- **Macro Avg:** Precision = 0.71, Recall = 0.71, F1 = 0.71
- **Weighted Avg:** Precision = 0.78, Recall = 0.78, F1 = 0.78
- **AUC** = 0.8217, strong separability

IX. Interpretation

The Random Forest classifier delivered the best performance among the models tested with a cross-validation accuracy of 84%, test accuracy of 77.7% and a ROC-AUC of 0.82. This indicates that the model is reliable in distinguishing between customers who churn and those who stay.

Confusion Matrix Insights

- Out of all non-churners (Actual No), 880 were correctly predicted while 156 were incorrectly classified as churners.
- Out of all churners (Actual Yes), 215 were correctly identified while 158 were missed.
- This shows that the model is fairly good at detecting churn but still struggles with false negatives (customers who churn but are predicted as staying).

Classification Metrics

- **Precision (Churn = 0.58):** Of the customers predicted as churners, 58% actually churned. This suggests some over-prediction of churn.
- **Recall (Churn = 0.58):** The model identified 58% of actual churners. While not perfect, it's better than the Decision Tree and comparable to XGBoost.
- **F1-Score (0.58 for churn):** Balancing precision and recall, the model performs moderately well in identifying churners.
- **Overall Accuracy (77.7%):** Reflects good general classification performance across both classes.

- **ROC-AUC (0.82):** Confirms strong model capability in separating churn vs. non-churn customers.

Feature importance insights

The Random Forest highlighted the following as the most significant drivers of churn:

1. **Contract Type** – Customers on month-to-month contracts are much more likely to churn compared to those on one- or two-year contracts.
2. **Monthly Charges** – Higher monthly bills are strongly associated with churn, indicating possible dissatisfaction with pricing.
3. **Total Charges** – Lower cumulative spending (i.e., newer/shorter-tenure customers) correlates with higher churn.
4. **Tenure** – Customers in their first 12 months are at greater risk of leaving.
5. **Online Security & Tech Support** – Lack of these add-on services increases churn likelihood.

The Random Forest model shows that contract type and service features play a central role in predicting churn. Customers on short-term, expensive plans without additional services are the most at-risk group. Although the model does not capture every churner (recall = 0.58), it provides a reliable framework for targeted retention strategies.

X. Strategic Insights & Recommendations

1. **Encourage Long-Term Contracts**
   a. Customers on **month-to-month contracts** show the highest churn rate (~43%). This indicates that contractual flexibility, while attractive increases the likelihood of customers switching providers.
   b. Recommendation: Introduce **discounted one-year and two-year plans**, bundled offers or loyalty benefits (e.g., free months, device upgrades) to encourage customers to commit long-term. Longer contracts reduce churn risk by locking in customers and improving predictability of revenue.
2. **Bundle Value-Added Services**
   a. Lack of **Online Security** and **Tech Support** emerged as strong churn drivers. These services not only add value but also create switching barriers.
   b. Recommendation: Promote **bundled packages** that include internet, phone, and add-on services at discounted rates. For example, offering a “Premium Pack” with Tech Support and Security can enhance customer stickiness and perceived value.
3. **Strengthen Early Engagement (First-Year Customers)**
   a. Tenure analysis revealed that **first-year customers are most vulnerable to churn**. Many customers leave before establishing loyalty.

b. Recommendation: Implement **onboarding programs**, such as personalized welcome offers, periodic engagement calls and service usage tips. Proactive customer support during the first 12 months can significantly reduce attrition.

4. **Target High-Charge Customers**
   a. Customers with **monthly charges exceeding $80** face a disproportionately higher churn rate, likely due to cost sensitivity or lack of perceived value.
   b. Recommendation: Introduce **tiered loyalty programs** that reward long-term, high-paying customers with benefits like priority support, special discounts or additional data/services. This ensures they feel valued and reduces the risk of losing high-revenue customers.
5. **Deploy the Machine Learning Model**
   a. The Random Forest model achieved strong predictive performance (AUC = 0.82) making it reliable for **operational deployment**.
   b. Recommendation: Integrate the churn prediction model into CRM systems to **flag high-risk customers in real time**. Customer service teams can then prioritize interventions (e.g., retention calls, personalized offers) for those predicted to churn.
6. **Data-Driven Retention Campaigns**
   a. Instead of generic promotions, use the model outputs to **segment customers** into low, medium and high-risk groups.
   b. Recommendation: Tailor retention campaigns accordingly, e.g., high-risk customers get stronger incentives while low-risk customers may simply receive engagement messages to maintain satisfaction.
7. **Continuous Monitoring & Model Updating**
   a. Customer behavior and market dynamics evolve over time.
   b. Recommendation: Retrain and validate the churn prediction model **periodically** with fresh data to maintain accuracy.

XI. Conclusion

This churn prediction project underscores the importance of applying machine learning techniques to address one of the most pressing challenges in the telecom industry. Among the models tested, Random Forest proved to be the most reliable in predicting customer churn, offering a balance of accuracy, robustness and interpretability. The analysis highlighted several key drivers of churn, including contract type, monthly and total charges, customer tenure and the availability of value-added services such as online security and technical support. These insights suggest that churn is influenced not only by pricing and contractual flexibility but also by the overall quality and perceived value of services. By leveraging these findings, telecom companies can implement proactive retention

strategies, such as promoting long-term contracts, offering tailored loyalty benefits and bundling essential services to address customer needs more effectively. Doing so will enable firms to strengthen retention, reduce revenue losses and enhance long-term profitability while improving customer lifetime value.