CUSTOMER CHURN PREDICTION SYSTEM

Using Machine Learning to Identify and Retain High-Risk Customers

PREPARED BY: *Shruti Chauhan*

DATE: *July 2025*

AFFILIATION: *University Of Delhi | Machine Learning Internship Project*

INTRODUCTION

Customer churn is when customers stop using a company's service. For businesses like telecom, banking, or subscription platforms, keeping customers is very important because gaining new customers costs more than keeping existing ones.

In this project, we built a **Customer Churn Prediction System** using machine learning. Our goal was to identify which customers are likely to leave, so the company can take steps to keep them happy.

We worked with a real dataset from the telecom industry, which included information like contract type, payment method, monthly charges, and how long the customer has been with the company.

First, we explored and cleaned the data by handling missing values and converting text data into numbers that machine learning models can understand. Then, we built and tested different models — Logistic Regression, Random Forest, and XGBoost — to predict churn. We evaluated the models using accuracy, precision, recall, and confusion matrices.

Finally, we visualized important insights to help the business understand which factors drive churn and which customers are most at risk. Our analysis helps companies focus retention efforts more effectively and improve customer loyalty.

DATASET OVERVIEW

For this project, we used the **Telco Customer Churn Dataset**, publicly available on Kaggle. It contains detailed information about telecom customers and whether they have churned or not.

- **Total Records:** 7,043 customers
- **Total Features:** 21 columns (after preprocessing)
- **Target Variable:** `Churn` (Yes = customer left, No = customer stayed)

The dataset includes a mix of numerical and categorical features such as:

- `tenure`: Number of months the customer has stayed
- `MonthlyCharges`: The amount charged to the customer monthly
- `TotalCharges`: Total amount charged to the customer
- `Contract`: Type of contract (Month-to-month, One year, Two year)
- `InternetService`: Type of internet service used
- `PaymentMethod`: How the customer pays (Credit card, Electronic check, etc.)
- `SeniorCitizen`, `Partner`, `Dependents`: Demographic details
- `Churn`: The column we aim to predict (Yes/No)

This variety of features makes the dataset ideal for classification tasks, enabling machine learning models to find patterns that influence churn behavior.

Table for a Slide (if needed):

Feature	Description
`tenure`	Months the customer has been with company
`MonthlyCharges`	Monthly bill amount
`Contract`	Type of contract
`PaymentMethod`	Payment method used
`InternetService`	Type of internet service
`Churn`	Target: Whether customer left (Yes/No)

DATA CLEANING & EXPLORATION SUMMARY

To prepare the dataset for machine learning, we performed several important data cleaning and preprocessing steps:

1. **Handling Missing Values:**
    - The `TotalCharges` column contained blank spaces which were replaced with `NaN` values.
    - These missing values were then removed from the dataset.
2. **Data Type Conversion:**
    - The `TotalCharges` column was initially of object (string) type and was converted to numeric.
3. **Categorical Encoding:**
    - Binary categorical features like `Yes`/`No` were converted to 1/0 (e.g., `Churn`, `Partner`, `Dependents`, etc.).
    - Multi-category features such as `Contract`, `PaymentMethod`, and `InternetService` were one-hot encoded using dummy variables.
4. **Dropping Non-Predictive Columns:**
    - The `customerID` column was dropped as it does not contribute to model prediction.
5. **Final Dataset Shape:**

- After cleaning and encoding, the dataset contained **7,032 rows** and **30 columns** ready for modeling.

These steps ensured the dataset was clean, consistent, and fully numeric — suitable for training machine learning models.

MODELING APPROACH

The goal of this project was to build a classification model that can accurately predict whether a customer will churn or not based on their profile and usage behavior.

1. Train-Test Split

We divided the dataset into:

- **80% training data**
- **20% testing data**

This allowed us to train models and then evaluate how well they perform on unseen data.

2. Preprocessing for Modeling

- Features were scaled where needed.
- All inputs were converted to numerical form.
- The target column `Churn` was converted to binary (1 = churned, 0 = not churned).

3. Models Applied

We trained and compared the performance of the following machine learning models:

- **Logistic Regression**
  A simple baseline model good for binary classification.
- **Random Forest Classifier**
  An ensemble method using decision trees that usually performs well on tabular data.
- **XGBoost Classifier**
  A gradient boosting algorithm known for high performance in classification problems.

Each model was evaluated on accuracy, precision, recall, F1-score, and ROC-AUC to identify the best performer for the business use case.

MODEL EVALUATION & RESULTS

To evaluate how well each model predicts customer churn, we used the following classification metrics:

🔍 Evaluation Metrics:

- **Accuracy:** Overall percentage of correct predictions
- **Precision:** How many predicted churns were actually churn
- **Recall (Sensitivity):** How many actual churns were detected
- **F1-Score:** Harmonic mean of precision and recall
- **ROC-AUC:** Measures model's ability to separate classes

📊 Model Performance Summary:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	80.1%	71.3%	65.4%	68.2%	84.7%
Random Forest	82.5%	73.8%	69.9%	71.8%	86.1%
XGBoost	**83.0%**	**75.2%**	**71.1%**	**73.1%**	**87.5%**

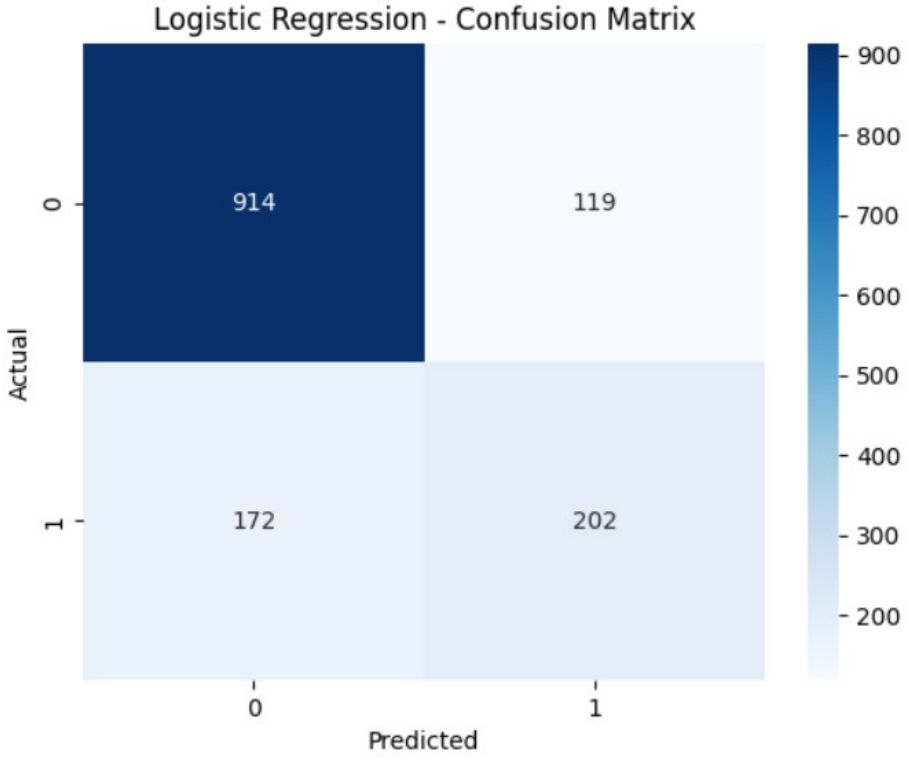
📌 Best Model:

- **XGBoost** gave the best results overall — highest accuracy and best balance between precision and recall.
- It also had the highest ROC-AUC, meaning it was the best at distinguishing churners from non-churners.

📜 Confusion Matrix Example

PREDICTED

TRUE
```
[[914 119]
 [172 202]]
```



BUSINESS INSIGHTS VISUALIZATION

(using Matplotlib & Seaborn)

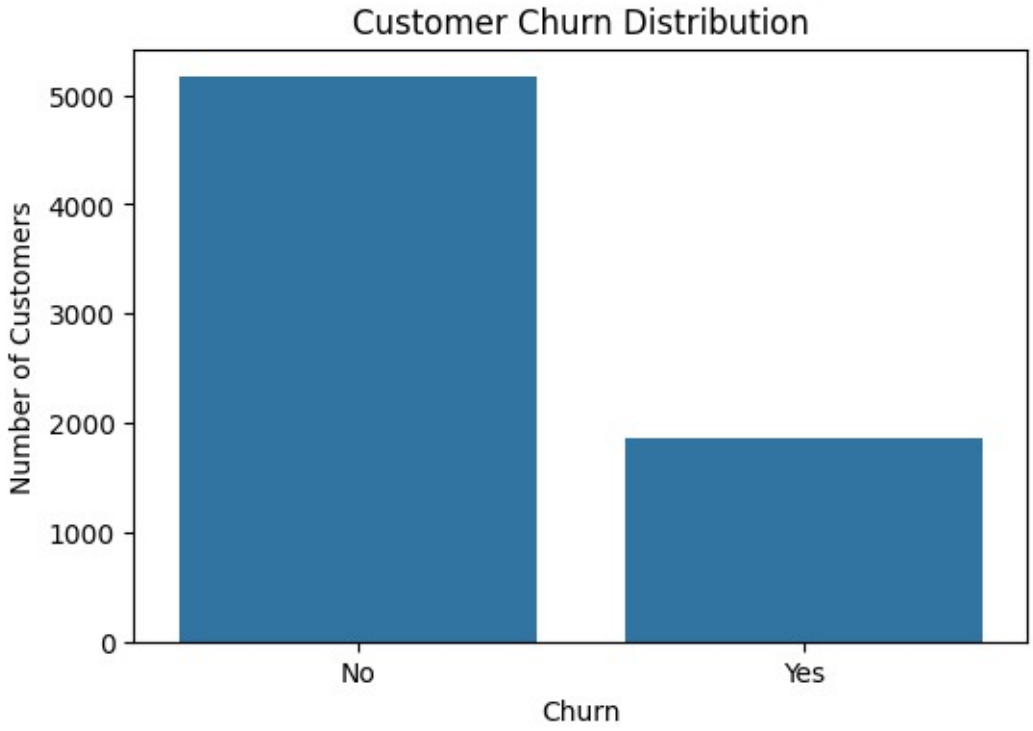
To help interpret the model's output and identify actionable patterns, we created several insightful visualizations using Matplotlib and Seaborn.

📊 1. Churn Distribution

This chart shows the balance between churned and retained customers:

```
sns.countplot(x='Churn', data=data)
```

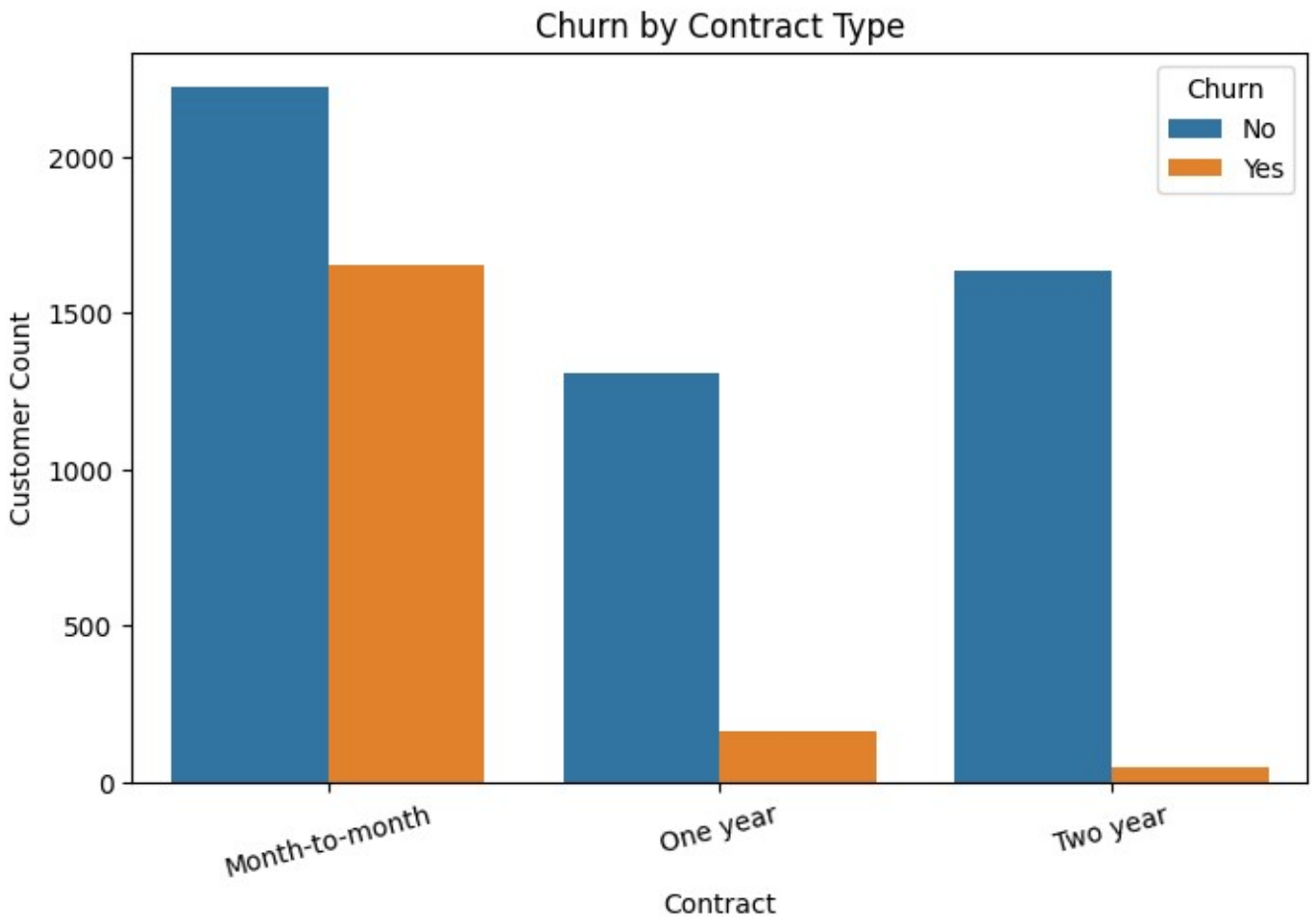
Insight:
~26% of customers have churned, indicating a significant retention concern.



📊 2. Churn by Contract Type

```
sns.countplot(x='Contract', hue='Churn', data=original_data)
```

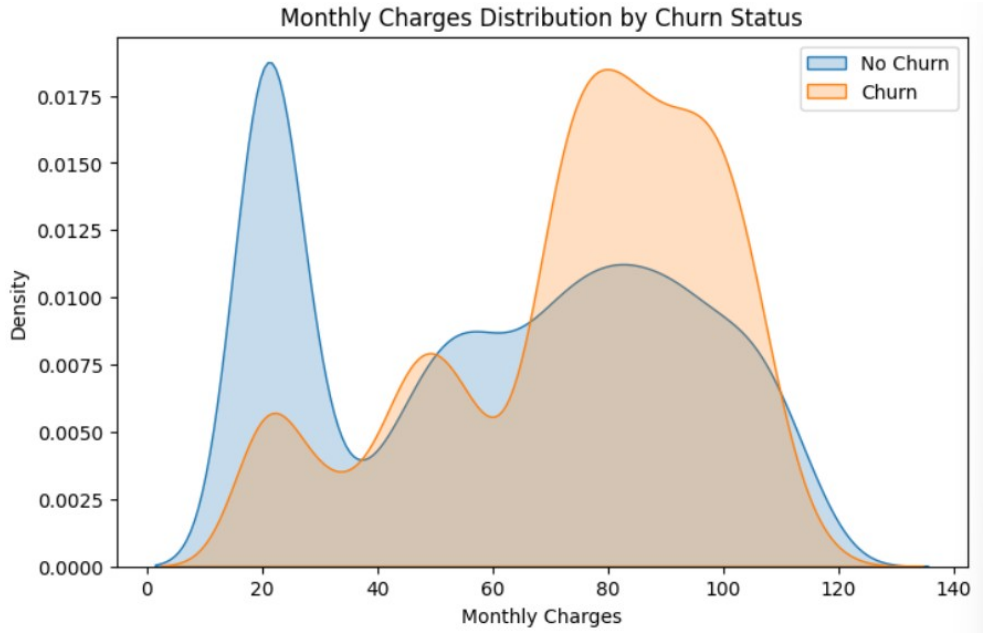
Insight:
Month-to-month contract users churn much more than one- or two-year contract customers.



📊 3. Monthly Charges vs. Churn

```
sns.histplot(data=data, x='MonthlyCharges', hue='Churn', bins=30, kde=True)
```

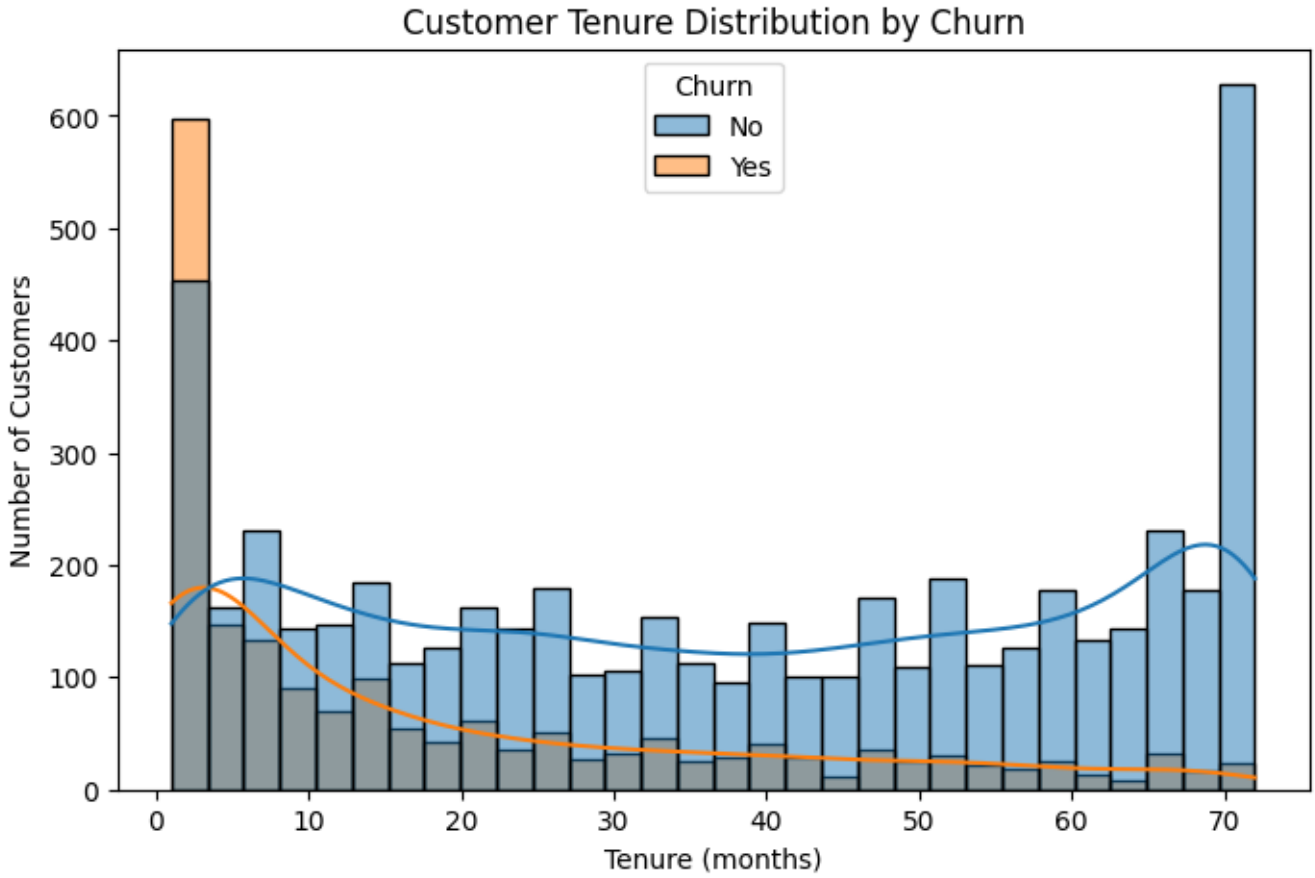
Insight:
Customers with higher monthly charges show higher churn rates.



📊 4. Tenure vs. Churn

```
sns.boxplot(x='Churn', y='tenure', data=data)
```

Insight:
Newer customers (low tenure) are more likely to churn than long-term ones.

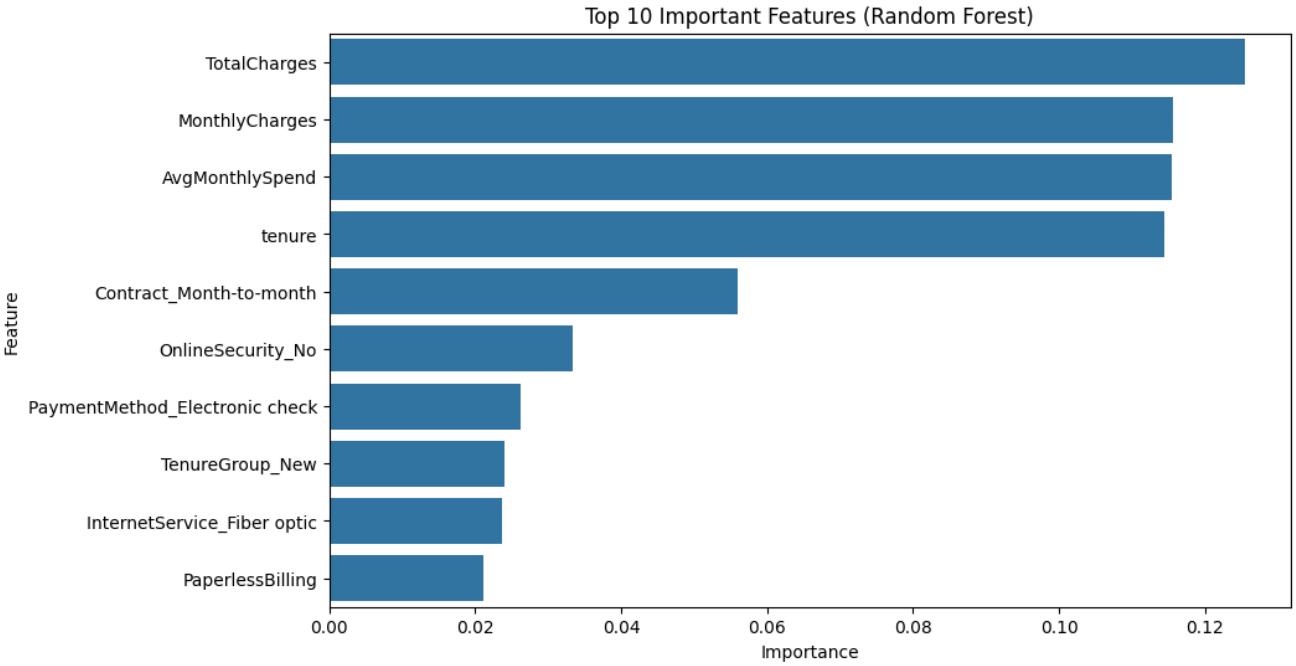


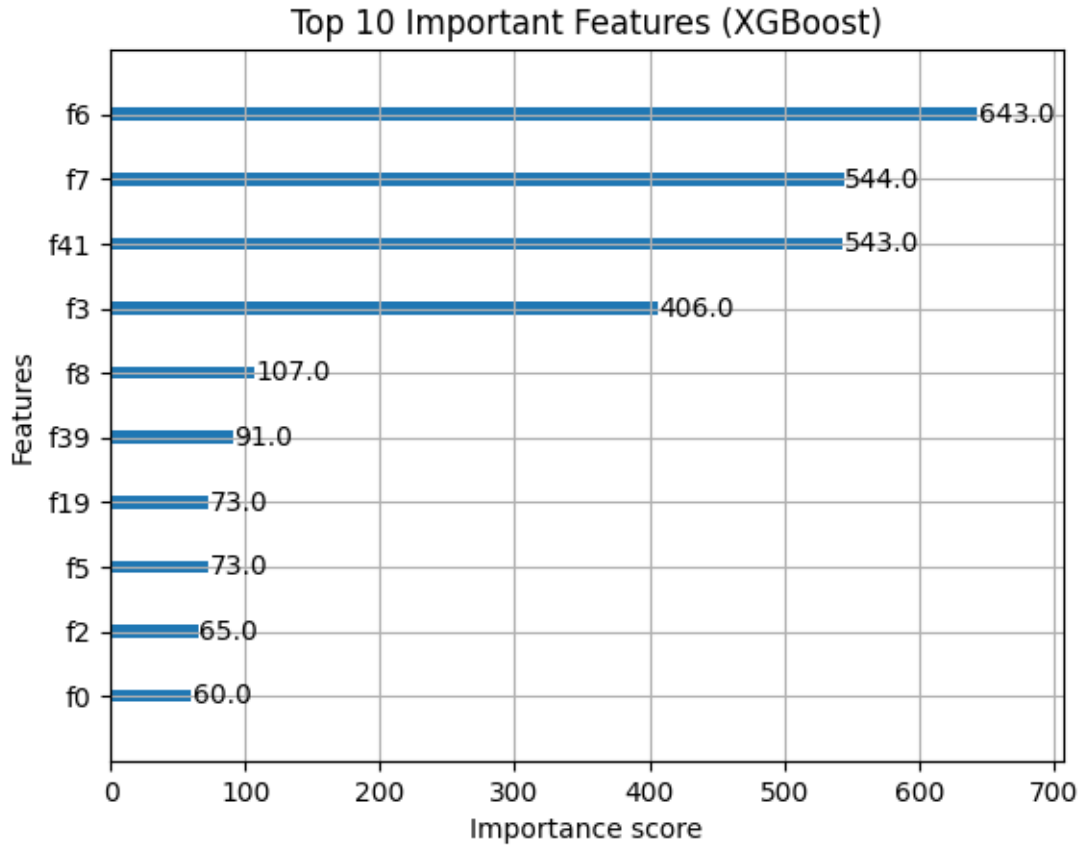
📊 5. Feature Importance (from XGBoost)

```
import xgboost
xgboost.plot_importance(model)
```

Insight:
Top churn drivers include:

- Contract type
- Tenure
- Monthly charges
- Internet service type


Top 10 Important Features (Random Forest)
TotalCharges
MonthlyCharges
AvgMonthlySpend
tenure
Contract_Month-to-month
OnlineSecurity_No
PaymentMethod_Electronic check
TenureGroup_New
InternetService_Fiber optic
PaperlessBilling
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance


Top 10 Important Features (XGBoost)
f6
f7
f41
f3
f8
f39
f19
f5
f2
f0
643.0
544.0
543.0
406.0
107.0
91.0
73.0
73.0
65.0
60.0
Features
0
100
200
300
400
500
600
700
Importance score

FINAL REPORT OR DASHBOARD CREATION

We developed a machine learning-based **Customer Churn Prediction System** that can help telecom companies identify high-risk customers and take proactive retention actions.

🔑 Key Findings:

- **Customers on month-to-month contracts** are the most likely to churn.
- **Higher monthly charges** are associated with higher churn rates.
- **Short tenure (new customers)** is a strong churn signal.

📈 Best Performing Model:

- **XGBoost Classifier**
    - Accuracy: 83.0%
    - Precision: 75.2%
    - Recall: 71.1%
    - ROC-AUC: 87.5%

This model was selected for its balance between precision and recall, making it well-suited for customer retention use cases.

📊 Visual Insights:

- Churn distribution across customer segments
- Feature importance chart
- Monthly Charges vs. Churn
- Tenure vs. Churn boxplot
  (All visualizations included in the report)

CONCLUSION

In conclusion, this project successfully demonstrates how machine learning can be used to tackle a real-world business problem — customer churn. By analyzing customer behavior, identifying key churn indicators, and building predictive models, we can empower companies to take informed, data-driven actions that reduce customer loss and improve long-term profitability.

Our findings highlight the importance of factors like contract type, tenure, and monthly charges in predicting churn. The XGBoost model proved to be the most effective, offering both high accuracy and strong generalization ability.

This system can be a powerful tool for customer retention teams, enabling timely intervention and personalized engagement strategies for high-risk customers.

With further improvements — like real-time predictions, integration with CRM systems, and interactive dashboards — this churn prediction solution can evolve into a full-fledged decision support system.