Telecom Customer Churn Prediction and Customer Segmentation Using ML

1. Problem Statement

Telecommunication companies operate in a highly competitive market where customer retention is as important as acquiring new customers. Losing existing customers (churn) directly impacts revenue, marketing costs, and long-term growth.

The objective of this project is to:

- Predict the probability of customer churn using machine learning.
- Identify high-risk customers early for proactive retention strategies.
- Segment customers into meaningful behavioral groups for targeted business actions.
- Provide a business-friendly analytical view through Power BI dashboards.
- Deploy the churn prediction model as a scalable web service.

This solution enables business teams to move from reactive churn handling to **proactive customer retention and personalized engagement**.

2. Dataset Overview

- Source: Telecom customer dataset (~7,043 customers)
- Features include:
  - Demographics (gender, senior citizen, family status)
  - Services subscribed (internet, streaming, security, support, etc.)
  - Contract and payment information
  - Billing details (monthly charges, total charges)
  - Target variable: Churn (Yes/No)

3. Project Workflow

Step 1: Data Wrangling & Cleaning

- Converted TotalCharges to numeric and handled missing values.
- Replaced missing TotalCharges for tenure = 0 customers with 0.
- Removed inconsistent and duplicate records.
- Standardized categorical values.

Step 2: Feature Engineering

- Converted binary categorical variables to numeric (0/1).
- Created support_services_count feature from multiple service columns.
- Engineered family stability features.
- Encoded categorical variables using One-Hot Encoding where appropriate.
- Generated churn probability instead of binary churn output.

Step 3: Exploratory Data Analysis (EDA)

Key EDA Insights:

- Customers with **month-to-month contracts** churn significantly more than long-term contract customers.
- Customers with **fiber optic internet** have higher churn compared to DSL users.
- Customers **without online security, backup, or tech support** are more likely to churn.
- High monthly charges increase churn risk, but value-added services reduce it.
- New customers (low tenure) have higher churn probability.
- Electronic check payment method has higher churn compared to automatic payments.

Step 4: Churn Prediction Modeling

- Built baseline Logistic Regression model.
- Compared with Random Forest and XGBoost.
- Logistic Regression chosen for:
  - Stability
  - Interpretability
  - Strong recall for churn customers

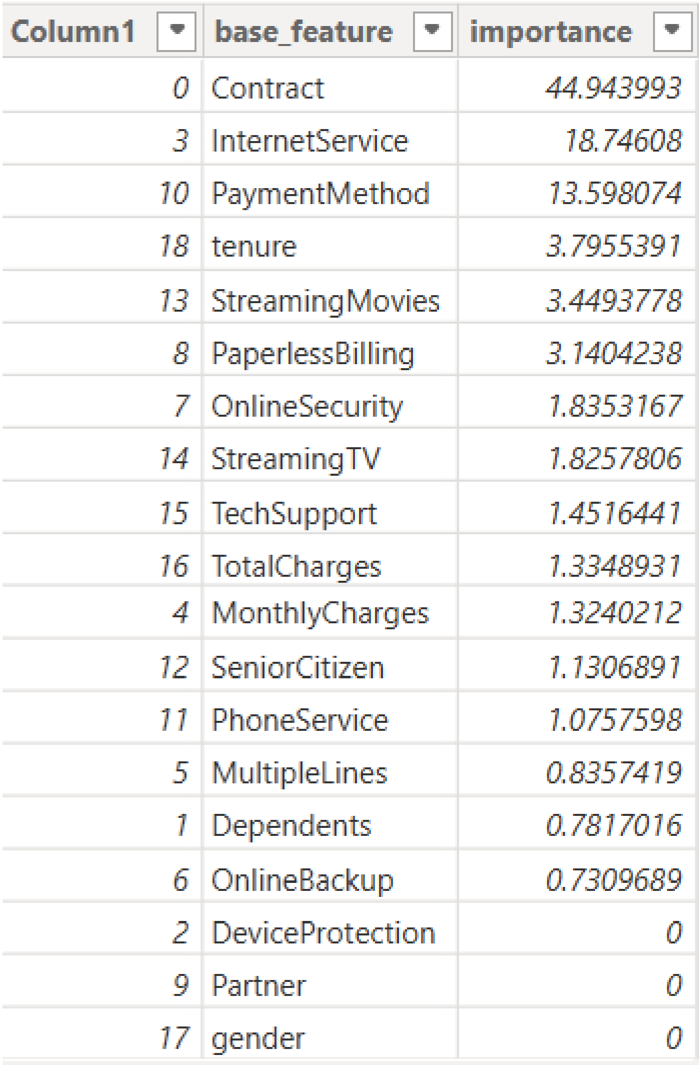
Model Optimization:

- Target imbalance handled using class weights.
- Probability threshold tuned to **0.35** to maximize recall.
- Recall prioritized over accuracy to minimize missed churn cases.

Step 5: Feature Importance (XGBoost)

Top churn drivers:

- Contract type
- Tenure
- Monthly charges
- Support service usage
- Payment method
- Internet service type



Column1	base_feature	importance
0	Contract	44.943993
3	InternetService	18.74608
10	PaymentMethod	13.598074
18	tenure	3.7955391
13	StreamingMovies	3.4493778
8	PaperlessBilling	3.1404238
7	OnlineSecurity	1.8353167
14	StreamingTV	1.8257806
15	TechSupport	1.4516441
16	TotalCharges	1.3348931
4	MonthlyCharges	1.3240212
12	SeniorCitizen	1.1306891
11	PhoneService	1.0757598
5	MultipleLines	0.8357419
1	Dependents	0.7817016
6	OnlineBackup	0.7309689
2	DeviceProtection	0
9	Partner	0
17	gender	0

Step 6: Customer Segmentation (K-Means)

- Applied K-Means on behavioral and financial features:
  - tenure
  - MonthlyCharges
  - TotalCharges
  - support_services_count
- Optimal clusters = 4
- Cluster profiles:

1. New & Low Engagement Customers
2. High-Value Loyal Customers

3. Price Sensitive Customers

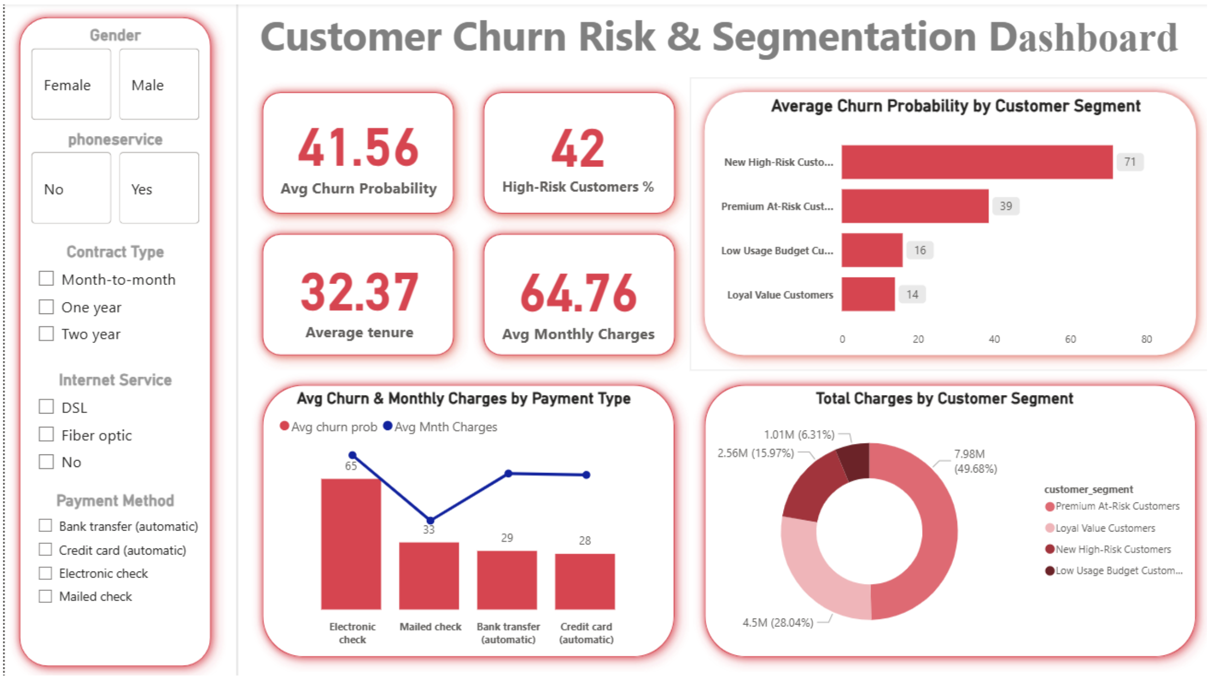
4. Long-Term Budget Customers

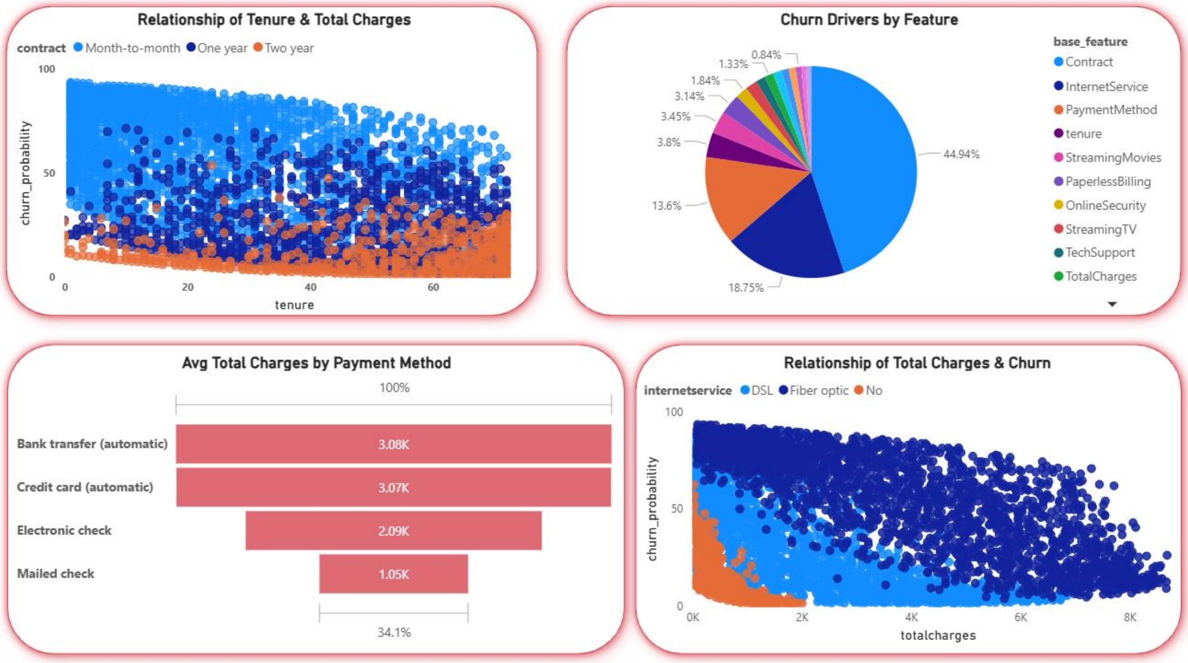
Step 7: Business Intelligence (Power BI)

Power BI dashboard shows:

- Overall churn trends
- Churn by service, contract, payment type
- Cluster-based churn analysis
- High-risk customer segments
- Revenue vs churn patterns

Used churn probability instead of churn label for risk-based analysis.





Step 8: Deployment

- Trained model and preprocessing pipeline saved using joblib.
- API built using FastAPI.
- Containerized using Docker.
- Deployed on AWS EC2 for scalable inference.

4. Power BI Dashboard Patterns

- Month-to-month contracts show highest churn risk.
- High churn among electronic check users.
- Fiber optic customers churn more than DSL.
- Clusters with low engagement show higher churn probability.
- High-value loyal customers show lowest churn.

5. Business Recommendations

1. **Target high-risk customers early** using churn probability scoring.
2. Promote long-term contracts to reduce churn.

3. Bundle support services to increase perceived value.
4. Offer incentives for electronic check users to shift to auto payments.
5. Focus onboarding programs for new customers in their first 3 months.
6. Prioritize retention campaigns for high-value but high-risk segments.

---

6. Conclusion

This project successfully integrates machine learning, customer segmentation, and business intelligence to create a comprehensive churn management system. By predicting churn risk, segmenting customers, and visualizing insights, the solution enables data-driven retention strategies that improve customer lifetime value and reduce revenue loss.

The deployment ensures scalability, and the dashboard ensures accessibility for business stakeholders, making this an end-to-end, production-ready churn analytics platform.

---

7. Technologies Used

- Python, Pandas, NumPy
- Scikit-learn, XGBoost
- K-Means Clustering
- FastAPI
- Docker
- AWS EC2
- Power BI
- SQLAlchemy, PostgreSQL