A Proactive Retention Engine for Telecom Customer Churn

Customer churn presents a significant and ongoing cost to the telecommunications industry. This project moves beyond reactive analysis to build a proactive, 4-phase machine learning system designed to **Predict**, **Explain**, **Prescribe**, and **Quantify** customer churn.

By training an XGBoost classifier, the system identifies high-risk customers with an **84% ROC-AUC score**. More importantly, it uses SHAP (SHapley Additive exPlanations) to understand the *specific reasons* for each customer's risk. This insight powers a prescription engine that suggests targeted retention offers.

A simulation on the test dataset, which included 1,407 customers, identified 188 high-risk individuals for intervention. The proposed strategy is projected to **retain 56 of these customers**, generating **$65,783 in saved revenue** at a cost of only $4,700, resulting in a **net profit of $61,083**.

1. Introduction

1.1. The Business Problem

The cost of acquiring a new customer is 5 to 10 times higher than the cost of retaining an existing one. For a subscription-based business like a telecom company, customer churn is a direct drain on revenue and profitability. The challenge is twofold:

1. We must accurately identify customers who are likely to leave *before* they do.
2. We must intervene in a cost-effective way, as offering discounts to all customers is financially unfeasible.

1.2. Project Objectives

The primary goal of this project was to develop an end-to-end data-driven system that:

- **Predicts:** Accurately classifies customers as 'Will Churn' or 'No Churn'.
- **Explains:** Opens the "black box" of the model to understand the key drivers of churn for the entire customer base and for each individual.
- **Prescribes:** Automatically suggests the most appropriate, cost-effective retention strategy for each high-risk customer.
- **Quantifies:** Provides a clear ROI and net profit analysis to justify the business case for this strategy.

2. Data & Key Insights (Exploratory Data Analysis)

The analysis was performed on the IBM Telco Customer Churn dataset, which contains 7,032 customer records and 20 features related to demographics, services, and billing.

Key insights from the EDA established the primary profile of an at-risk customer:

- **Finding 1: Contract is King.** Customers on **Month-to-Month contracts** are overwhelmingly the largest source of churn. Customers on "One year" or "Two year" contracts are significantly more loyal.
- **Finding 2: The New Customer Risk.** There is a strong negative correlation between **tenure** and churn. The newest customers (1-10 months) are far more likely to leave, while loyalty increases significantly over time.
- **Finding 3: The Service Gap.** Customers lacking key "sticky" add-on services like **Online Security** and **Tech Support** churn at a much higher rate. This suggests a perceived gap in service value.
- **Finding 4: The Price Point.** Customers with **Fiber Optic** internet service, which generally corresponds to **higher Monthly Charges**, also have a higher churn rate.

3. Methodology

A robust sklearn pipeline was developed to ensure all data processing and modeling steps are reproducible and scalable.

Phase 1: Prediction (XGBoost)

An XGBoost (Extreme Gradient Boosting) classifier was chosen for its high performance and scalability.

- **Imbalance Handling:** The dataset is imbalanced (73% 'No' vs. 27% 'Yes'). The scale_pos_weight parameter was used to force the model to treat misclassifying a "Churn" customer as more severe.
- **Hyperparameter Tuning:** A baseline model achieved a strong 0.81 ROC-AUC score. To improve this, RandomizedSearchCV was used to test 50 different parameter combinations. This "tuned" model became the final engine, achieving an improved **ROC-AUC score of 0.84** and a "Will Churn" **F1-Score of 0.75**.

Phase 2: Explanation (SHAP)

To understand the model's logic, SHAP (SHapley Additive exPlanations) was implemented. The global feature importance plot confirms the EDA findings, identifying Contract_Month-to-month and tenure as the most powerful decision-drivers.

Phase 3: Prescription Engine

This is the system's "brain." By applying SHAP at an *individual customer* level, we can see their unique risk profile.

The force plot for a high-risk customer (Index 1) clearly showed that their Contract and tenure were the primary reasons for their high churn score. This insight powers a simple, rule-based prescription engine:

If Churn Reason is...	**Prescribed Action**
Month-to-month Contract	"Offer 15% discount to switch to a 1-year contract."
No Tech Support	"Offer 6-month free trial of Tech Support bundle."
Low Tenure	"Send 'Welcome' email with a free movie rental."

Phase 4: ROI Calculation

A financial model was built to connect the model's output to a profit-and-loss statement. The following conservative assumptions were made:

- **Customer Lifetime Value (CLV):** $1,166.37 (Calculated from Avg. Monthly Revenue * Avg. Retained Lifetime)
- **Average Intervention Cost:** $25.00
- **Intervention Success Rate:** 30% (Assumed 30% of customers who receive an offer will accept it and stay)

4. Financial Impact Analysis

When the tuned model was applied to the 1,407 customers in the test set, the financial impact was as follows:

Metric	**Value**
High-Risk Customers Identified (Prob > 80%)	188
Total Cost of Intervention (188 * $25)	**$4,700.00**
Customers Retained (188 * 30% rate)	56

Total Value Saved (56 * $1,166.37 CLV)	**$65,783.14**
Net Profit (Value Saved - Cost)	**$61,083.14**

This demonstrates a significant and positive return on investment for the project.

5. Conclusion & Recommendations

This project successfully demonstrates that a data-driven retention strategy is not only feasible but highly profitable. The 4-phase (Predict, Explain, Prescribe, Quantify) system provides an actionable and scalable framework for reducing customer churn.

Recommendations:

1. **Deploy:** The best_model pipeline should be deployed into a production environment to score customers in real-time.
2. **Act:** High-risk (Prob > 80%) customers should be automatically flagged and their "Prescription" routed to the marketing or customer service teams for intervention.