Final Project Report: Customer Churn Prediction

1. Business Understanding

Customer churn prediction is essential for companies aiming to retain users and reduce service cancellations. Retaining existing customers is often more cost-effective than acquiring new ones. Therefore, predicting potential churners allows for proactive engagement and personalized retention strategies.

In this project, we aim to build a model that identifies whether a customer will continue using the service or stop, enabling companies to take effective measures to prevent churn.

---

2. Data Understanding

*The dataset used is the **Customer Churn Dataset** from Kaggle, comprising 505,256 rows and 12 features:*

- ***CustomerID**: Unique customer identifier*
- ***Age**: Age of the customer*
- ***Gender**: Male/Female*
- ***Tenure**: Duration of service usage*
- ***Usage Frequency**: Service usage frequency per month*
- ***Support Calls**: Frequency of support requests*
- ***Payment Delay**: Delay in bill payments*
- ***Subscription Type**: Basic/Standard/Premium*
- ***Contract Length**: Monthly/Quarterly/Annual*
- ***Total Spend**: Total amount spent by the customer*

- ***Last Interaction**: Months since last activity*
- ***Churn**: Target label (1 for churned, 0 for active customers)*

Data is split into:

- ***Training set**: 440,882 rows*
- ***Testing set**: 64,374 rows*

📌 *[Insert pie chart showing churn distribution]*

---

3. Data Pre-Processing

- *Checked for **missing values** and handled them appropriately.*
- *Encoded **categorical variables** using Label Encoding and One-Hot Encoding.*
- ***Scaled numerical features** using StandardScaler to ensure uniform feature scaling.*
- *Performed **train-validation-test split** using Stratified Sampling to preserve class proportions.*
- *Removed any duplicate entries found in the dataset.*

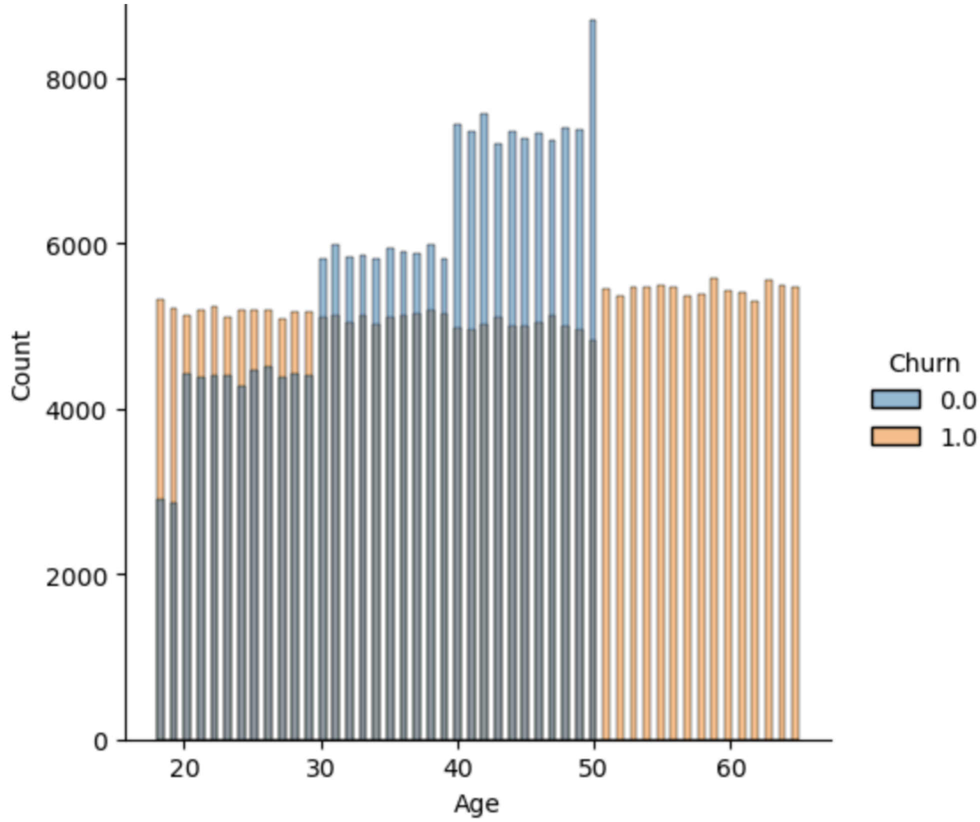
📌 *[Insert bar chart of missing values before cleaning]*

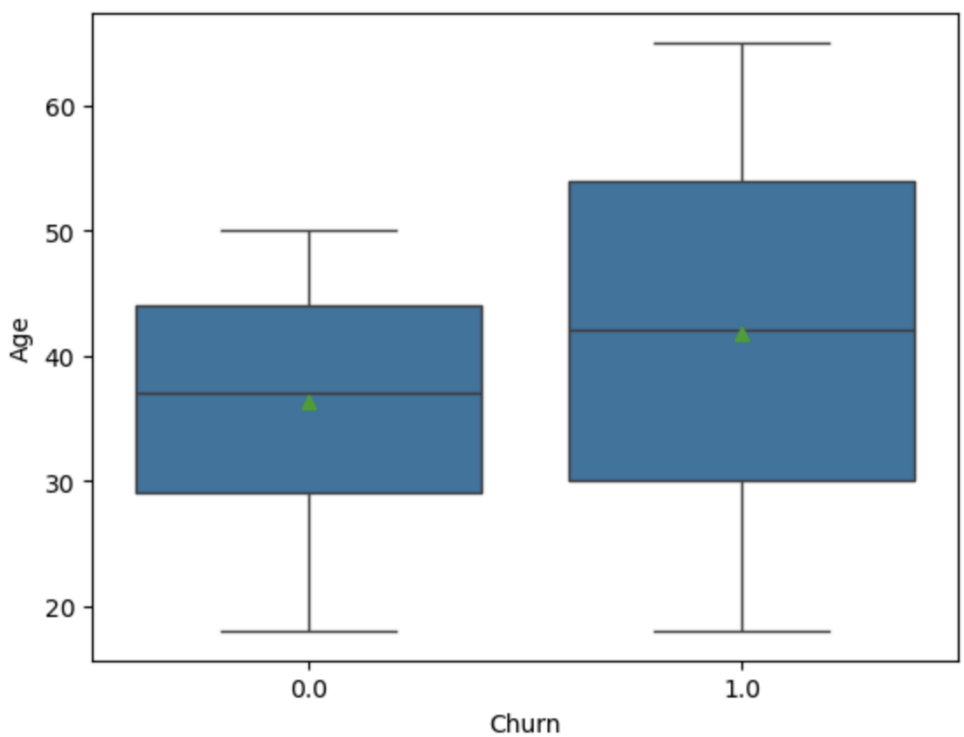
---

4. Exploratory Data Analysis (EDA)

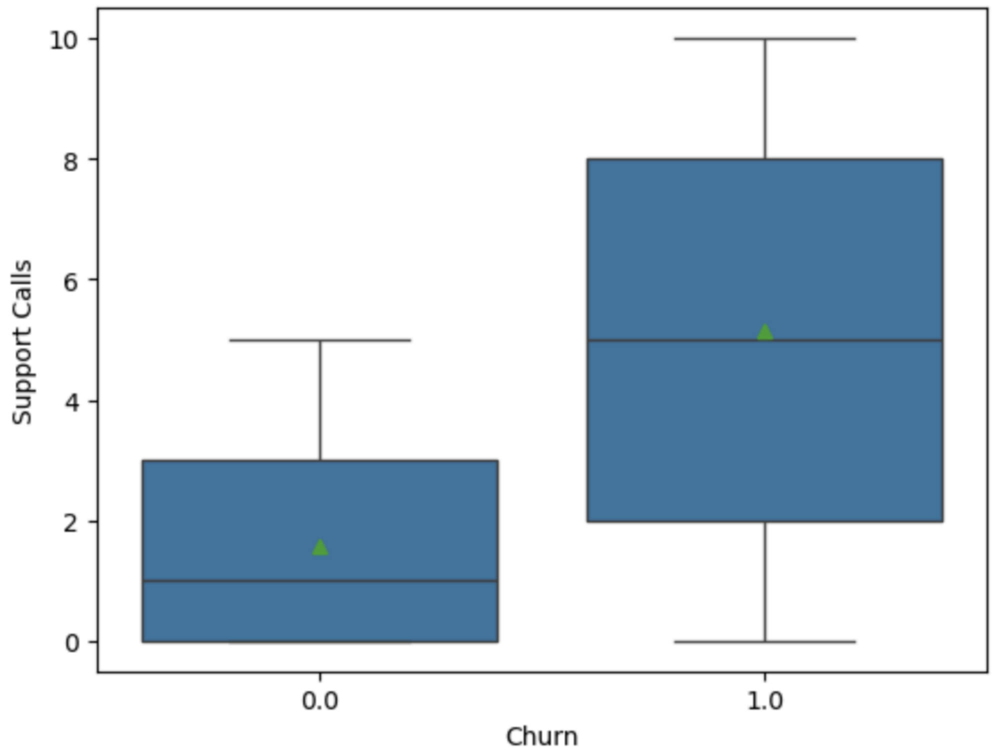
Key Observations:

- ***Age and Churn**: Customers aged **40–50** are less likely to churn. Those aged under 30 or over 60 have higher churn rates.*

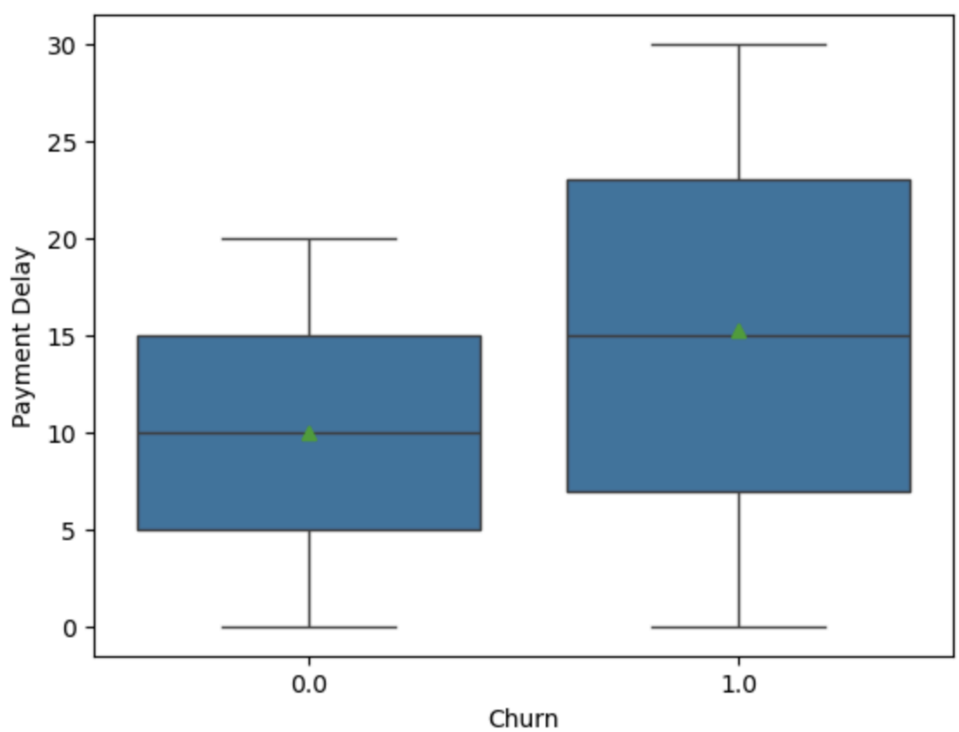
- 

- 

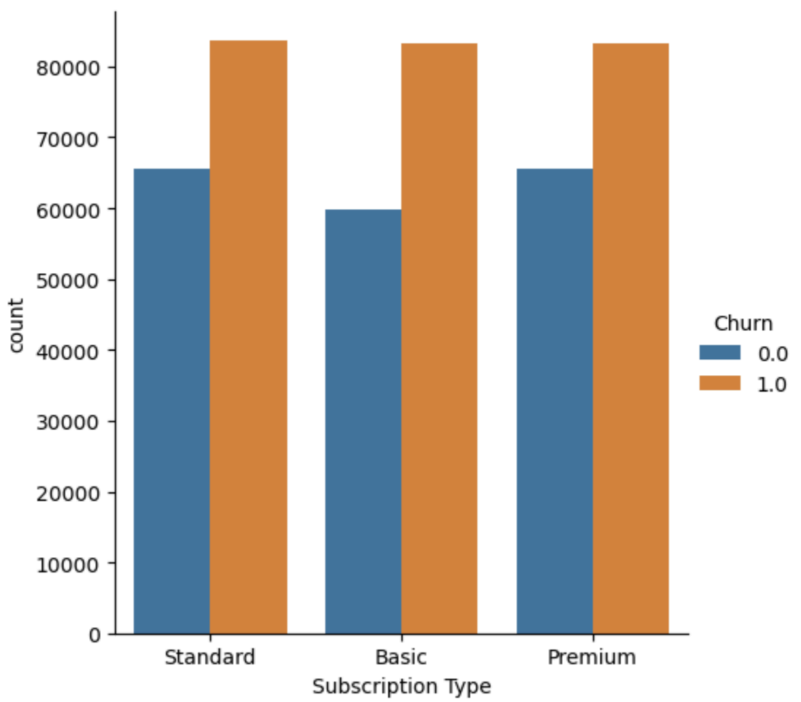
- ***Support Calls****: A major predictor. Users with more than 5 support calls show significantly higher churn probability.*

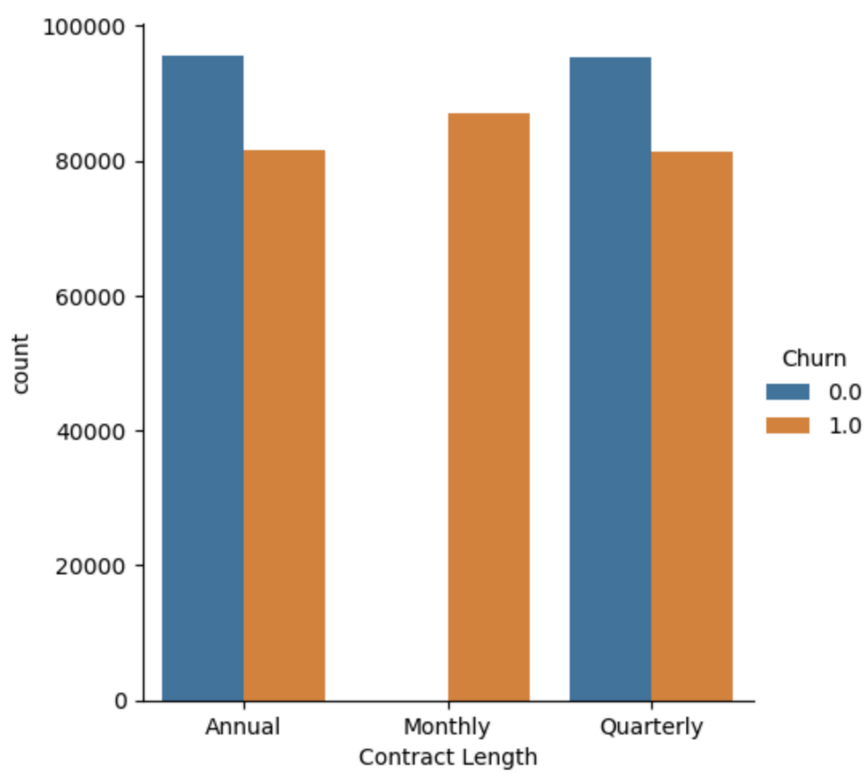


- ***Payment Delay****: Longer delays are associated with churned customers.*

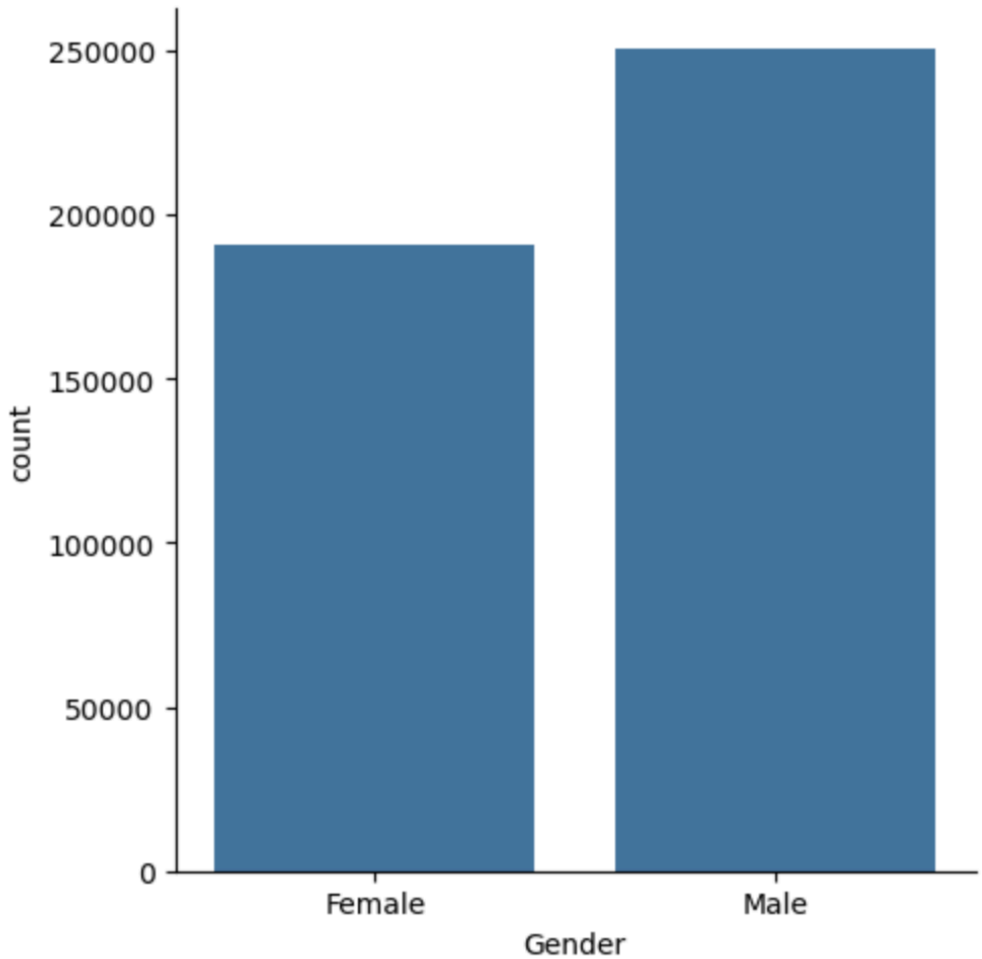


- ***Subscription Type & Contract Length**: Customers with **Premium subscriptions** and **Annual contracts** are more loyal.*





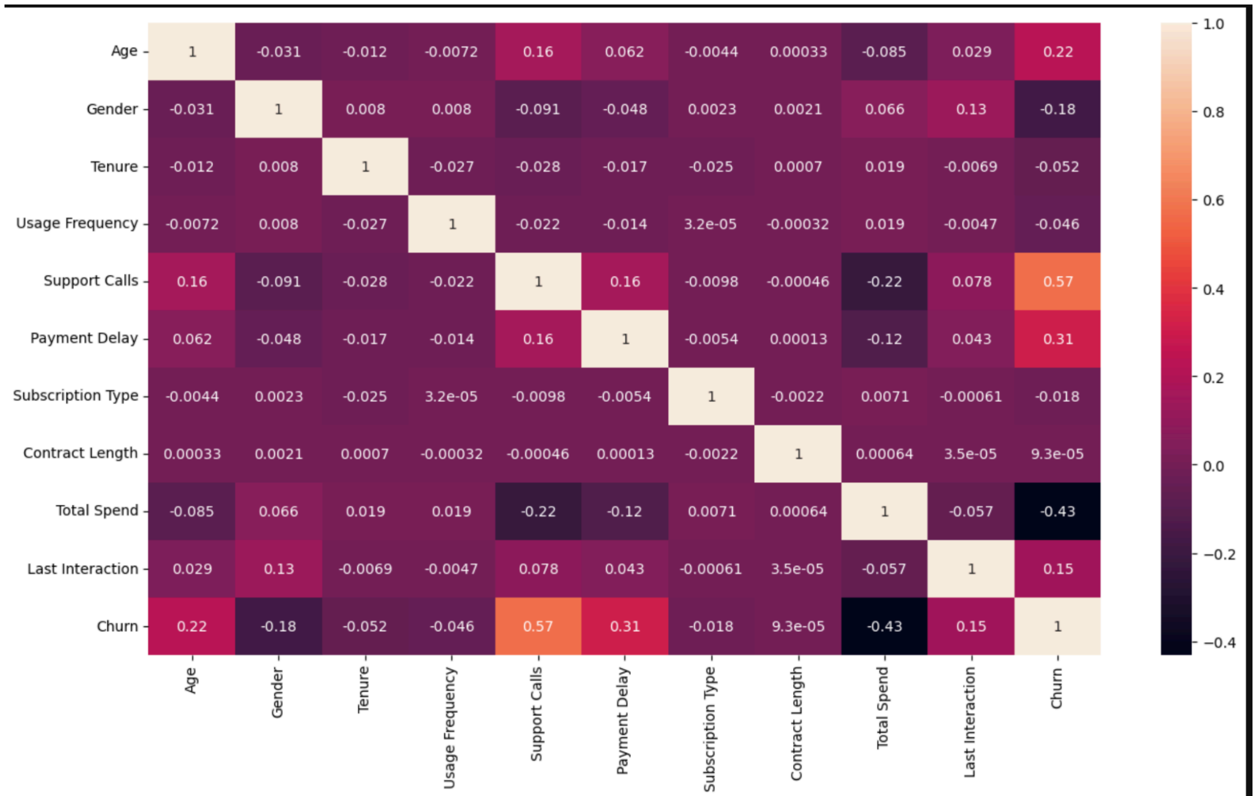
- ***Gender and Churn**: No significant difference observed.*



---

5. Feature Analysis

- *Generated a **correlation matrix** to assess relationships between features.*
- *Found high correlation between `Churn` and `Support Calls`, `Payment Delay`, and `Last Interaction`.*



6. Modeling (Baseline Models)

We built three baseline models:

a) Logistic Regression

A simple, interpretable model used as a benchmark. It assumes linear relationship between input features and the log odds of the outcome.

- *Strength: Fast, interpretable.*
- *Limitation: May underperform on non-linear problems.*

b) Random Forest Classifier

An ensemble of decision trees trained via bagging.

- *Strength: Handles non-linear relationships and overfitting well.*

- *Limitation: Less interpretable.*

c) XGBoost Classifier

An advanced gradient-boosting technique optimized for performance.

- *Strength: High predictive power and handles imbalanced data well.*
- *Limitation: Longer training time, tuning required.*

Evaluation Metrics:

- *Accuracy, Precision, Recall, ROC-AUC*
- *Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE)*

Model Performance Comparison (Original Data)

Model	***ROC-AUC***	***MAE***	***RMSE***	***MSE***	***Recall***
Logistic Regression	*0.82*	*0.21*	*0.45*	*0.20*	*98%*
Random Forest	*0.91*	*0.16*	*0.36*	*0.13*	*100%*
XGBoost	*0.93*	*0.14*	*0.32*	*0.10*	*100%*

---

7. Modeling with Oversampling (SMOTE)

To address class imbalance:

- *Applied **SMOTE (Synthetic Minority Oversampling Technique)** to oversample minority class.*
- *Re-trained models on balanced data.*

Results with SMOTE:

- ***Logistic Regression**: Recall improved to 97%*
- ***Random Forest**: Maintained 100% recall*
- ***XGBoost**: Maintained 100% recall*

---

8. Key Insights & Recommendations

- ***Support Calls**: A critical churn indicator. Higher calls = higher dissatisfaction.*
    - *Recommendation: Improve service quality and resolve issues on first contact.*
- ***Age Group 40–50**: Most loyal segment.*
    - *Recommendation: Implement reward programs for this demographic.*
- ***Payment Delay**: Positively correlated with churn.*
    - *Recommendation: Simplify payment processes and send automated reminders.*
- ***Top Performers**: XGBoost and Random Forest achieved high accuracy and recall.*
    - *Recommendation: Deploy models in production for real-time churn monitoring.*

---

9. Conclusion & Next Steps

We successfully built a machine learning pipeline for customer churn prediction:

- *Trained baseline and advanced models (Logistic Regression, Random Forest, XGBoost)*

- *Evaluated models on both original and balanced data*
- *Identified top predictors: Support Calls, Payment Delay, Last Interaction*

Best Model: XGBoost (100% recall with SMOTE)

Business Impact:

- *Enables **targeted retention campaigns***
- *Reduces **customer acquisition costs***

Next Steps:

- *Deploy XGBoost via a RESTful API or business dashboard*
- *Incorporate additional features (demographics, engagement metrics)*
- *Conduct A/B testing on retention offers*
- *Explore explainability with SHAP or LIME for actionable insights*