Customer Churn Prediction

Amit Vasant Rangane- 38915109

1. Introduction

Objective:

This project uses machine learning models to predict customer churn in a telecom dataset, aiming to identify high-risk customers and provide actionable insights to reduce churn and enhance retention strategies.

Business Relevance:

Customer churn is a significant concern in the telecom industry. Predictive analytics enables businesses to proactively intervene by offering personalized incentives and improved services to retain at-risk customers.

---

2. Data Overview

Dataset Information:

- The dataset comprises **telecom customer data**, including demographics, usage behavior, and customer service interactions.
- **Target Variable:** Churn (1 = Customer Churned, 0 = Retained)
- **Key Features:**
    - **Behavioral Features:** Call Failures, Complaints, Seconds of Use, Distinct Called Numbers.
    - **Subscription Features:** Subscription Length, Tariff Plan, Customer Value.

Data Preprocessing Steps:

- Removed missing values & duplicates.
- Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance class distribution.

Data Quality Metrics:

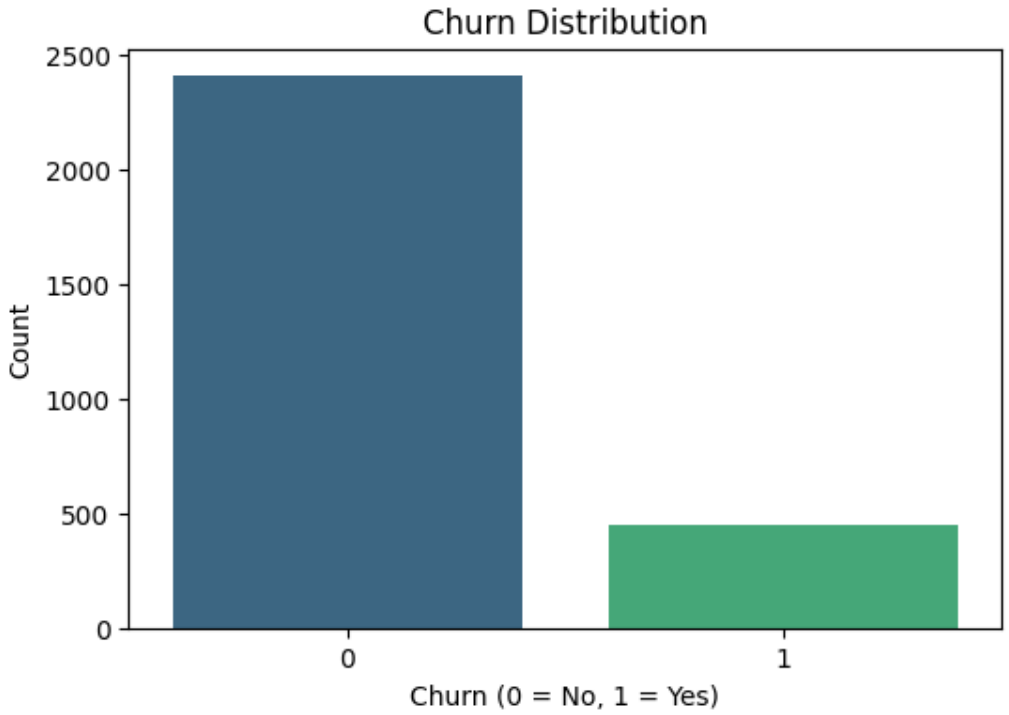
- Initial dataset size: 3,164 records
- Categorical features: 0
- Numerical features: 16
- Churned customers (%): 15.74%
- Retained customers (%): 84.26%
- Missing values (%): 0.13%
- Duplicate records: 297 records

---

3. Exploratory Data Analysis (EDA)

Churn Distribution Chart:

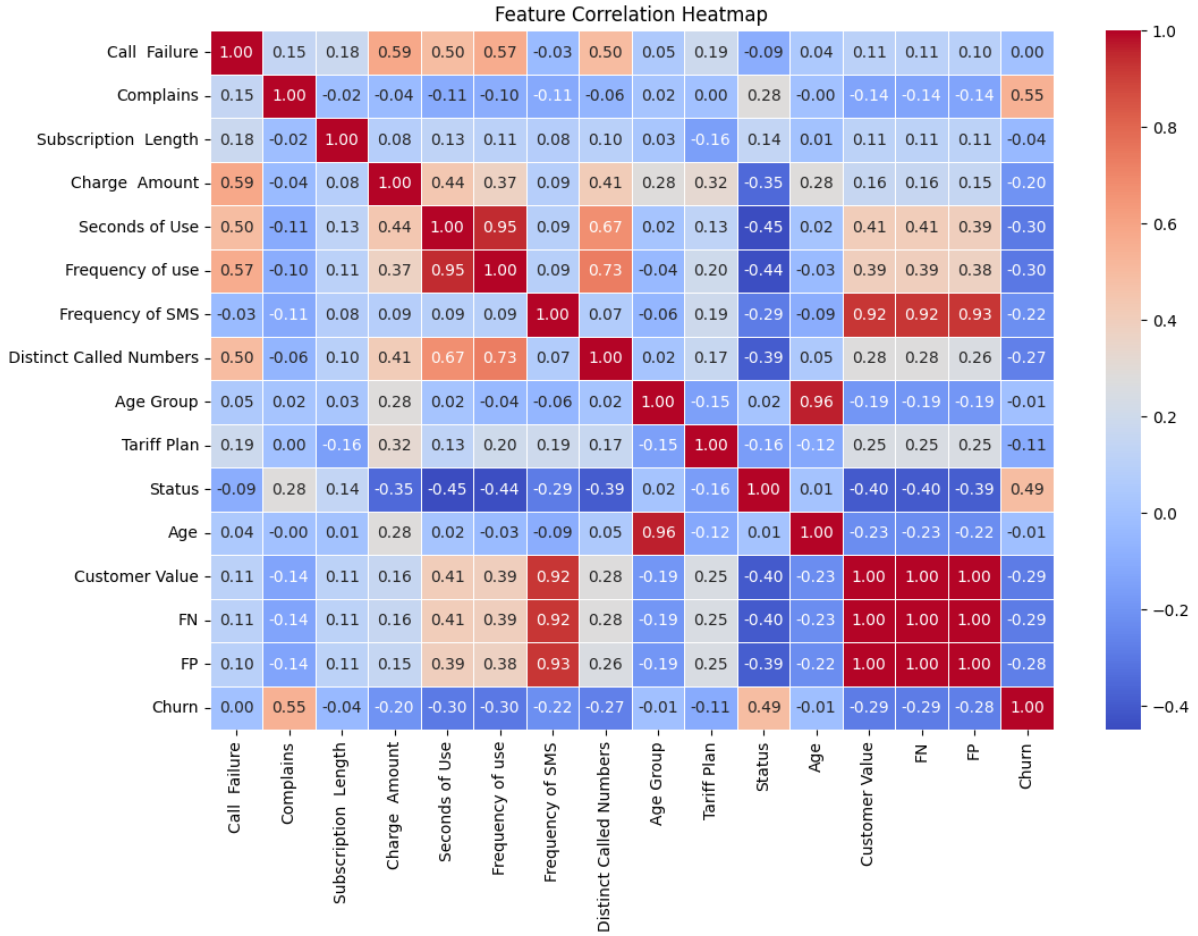
The churn distribution chart shows a significant imbalance between retained (0) and churned (1) customers, highlighting the need for class balancing techniques like SMOTE.



Feature Correlation:

- **Complaints, Call Failures, and Seconds of Use** are positively correlated with churn.
- **Subscription Length** is negatively correlated with churn.

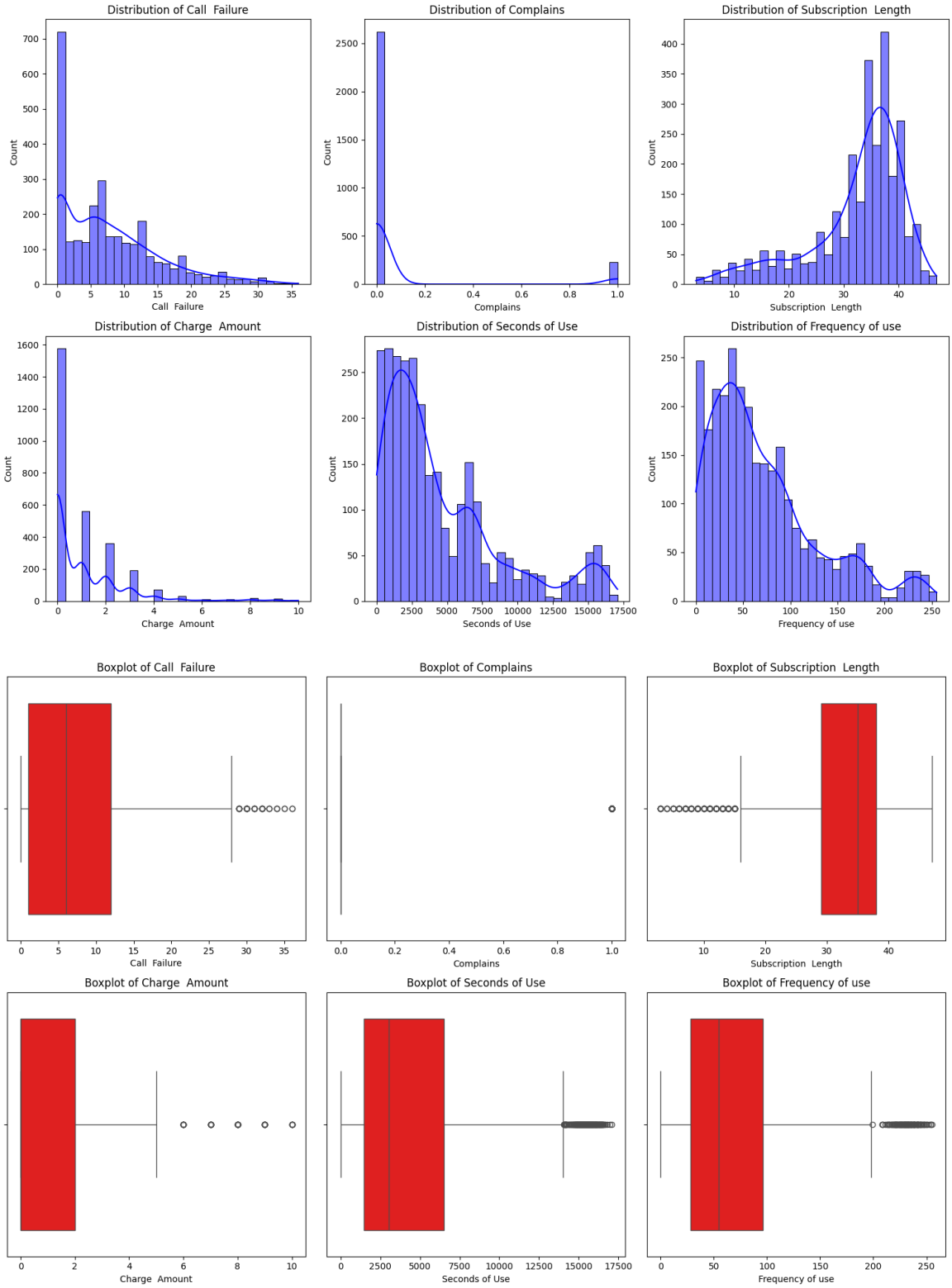
Feature Correlation Heatmap:



Feature Distributions:

- Customers with **high call failures and complaints** tend to churn.
- Subscription length and customer value are important factors

Histograms & Boxplots of Key Features:



4. Model Building & Evaluation

Machine Learning Models Used:

- **Random Forest**
- **XGBoost (Final Model)**
- **Gradient Boosting**

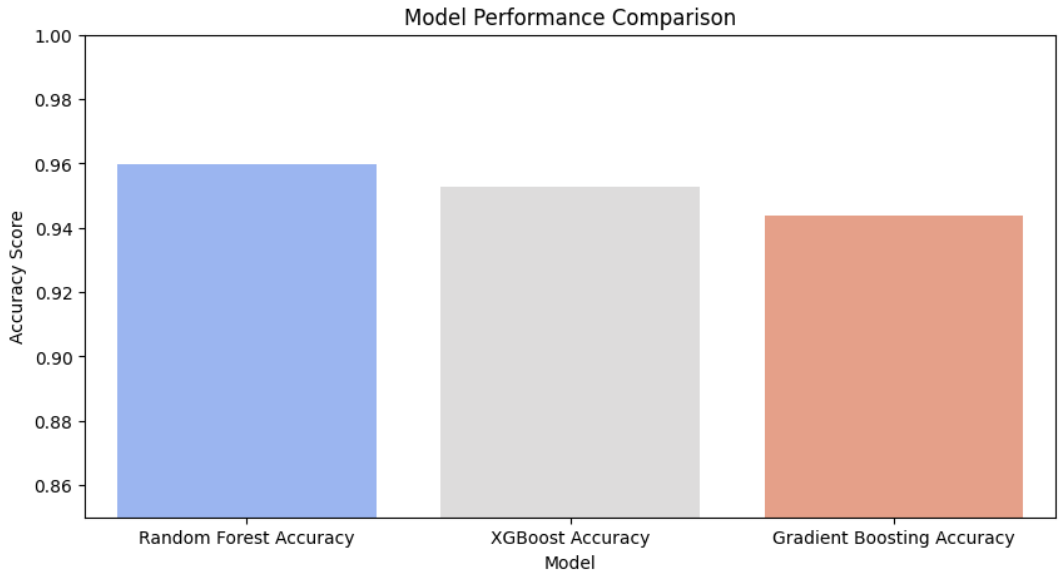
Random Forest, XGBoost, and Gradient Boosting were selected due to their robustness in handling imbalanced datasets and their ability to capture complex relationships in the data.

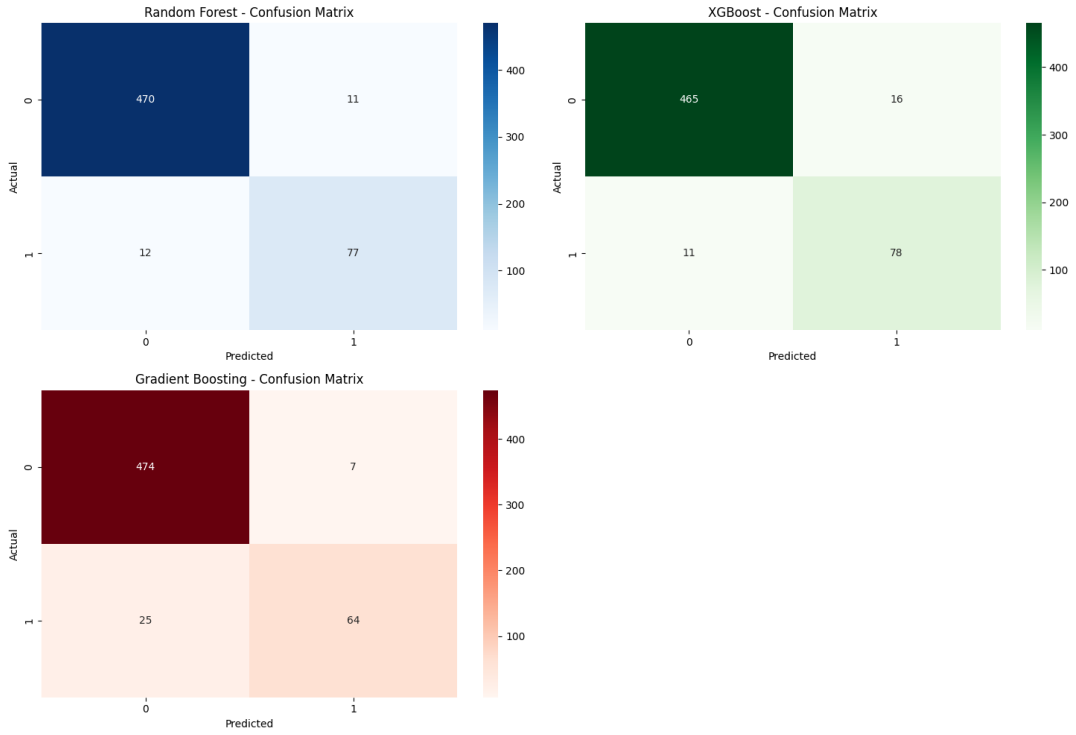
Model Performance Comparison:

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	95.96%	0.98	0.98	0.98
XGBoost	95.26%	0.98	0.97	0.97
Gradient Boosting	94.38%	0.95	0.99	0.97

The Random Forest and XGBoost models exhibit similar accuracy. Therefore, we will perform hyperparameter tuning on both models and select the one that delivers the best performance.

Model Performance Comparison Graph:



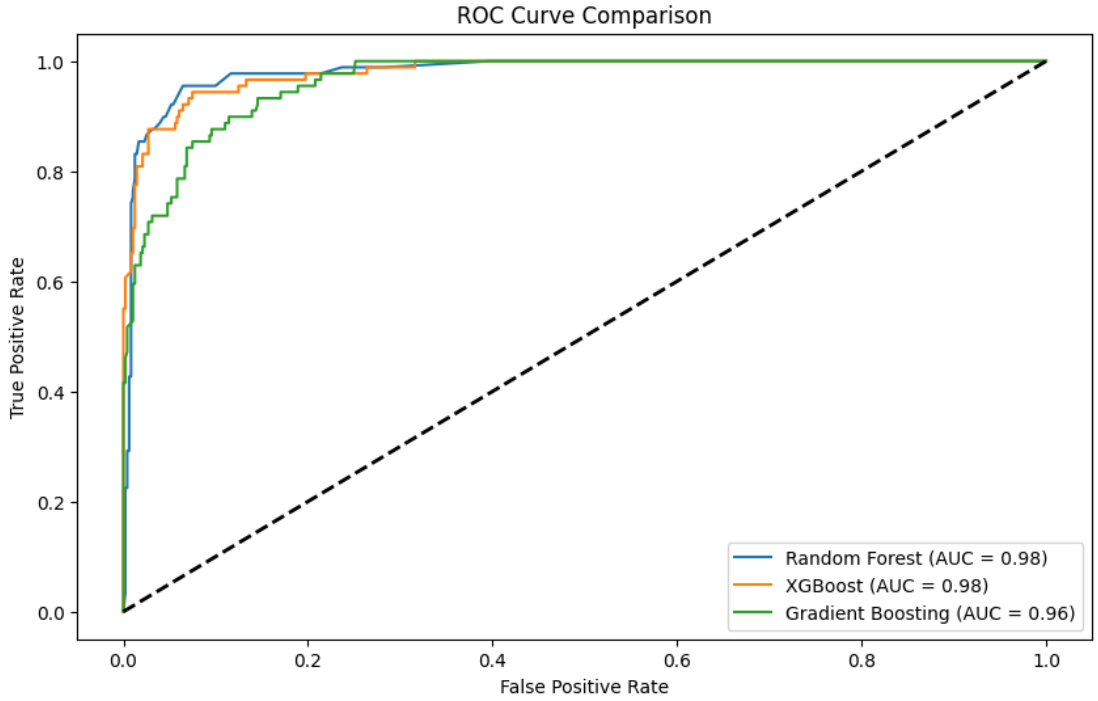


ROC Curve & AUC Score Comparison:

The ROC curve and AUC scores indicate that all models perform well, with Random Forest and XGBoost achieving the highest AUC scores of 0.9811 and 0.9799, respectively

Model	AUC Score
Random Forest	0.9811
XGBoost	0.9799
Gradient Boosting	0.9642

ROC Curve Graph:



---

5. Hyperparameter Tuning

Best Parameters After Tuning:

Random Forest:

- max_depth = 20, min_samples_split = 5, n_estimators = 200
- **Final Accuracy:** 95.61%

XGBoost (Final Model):

- learning_rate = 0.2, max_depth = 6, n_estimators = 50
- **Final Accuracy:** 95.79%

Best Model Selection:

Hyperparameter tuning was performed using grid search to optimize model performance. The best parameters for Random Forest and XGBoost were identified, with XGBoost achieving the highest accuracy.

---

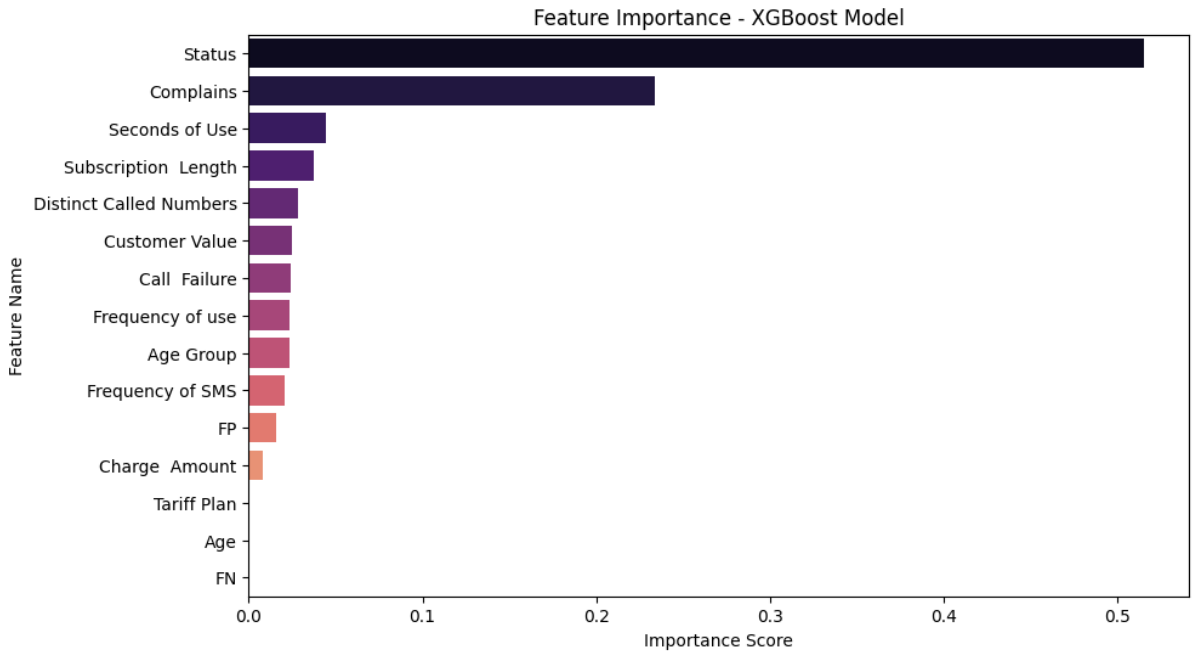
6. Feature Importance Analysis (XGBoost Model)

The top features influencing churn predictions are:

Rank	Feature	Importance Score
1	**Status**	0.5156
2	**Complaints**	0.2334
3	**Seconds of Use**	0.0443
4	**Subscription Length**	0.0373
5	**Distinct Called Numbers**	0.0280

The feature importance analysis reveals that 'Status' and 'Complaints' are the most influential factors in predicting churn, followed by 'Seconds of Use' and 'Subscription Length.'

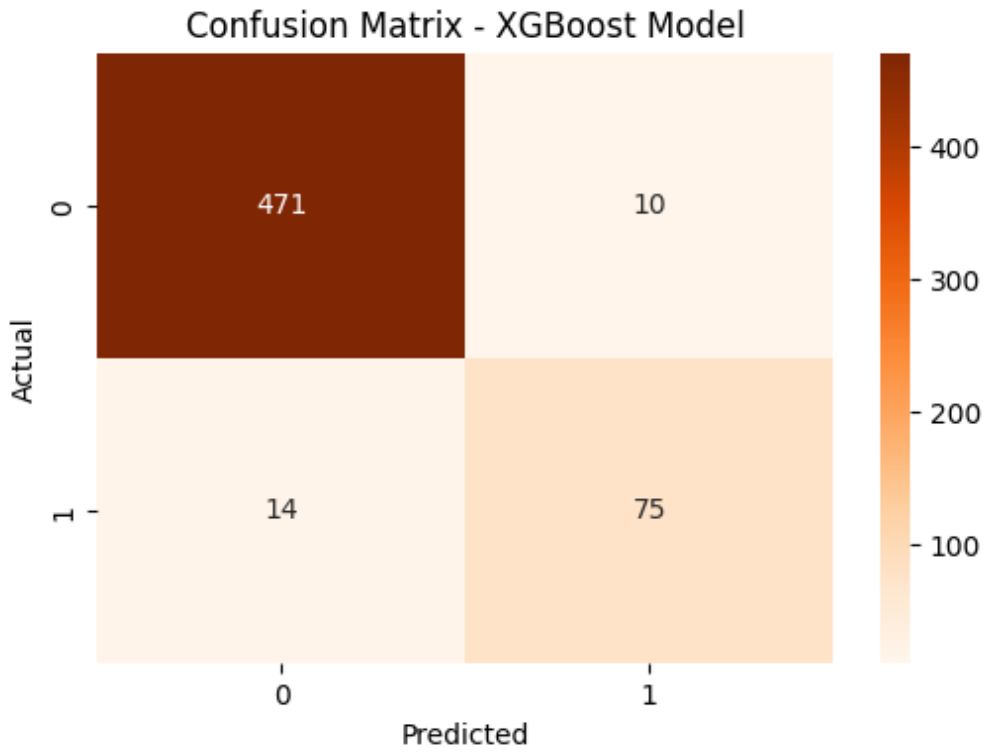
Feature Importance Chart:



---

7. Confusion Matrix after hyper-parameter tuning

Confusion Matrix Visualization:



Confusion Matrix for XGBoost:

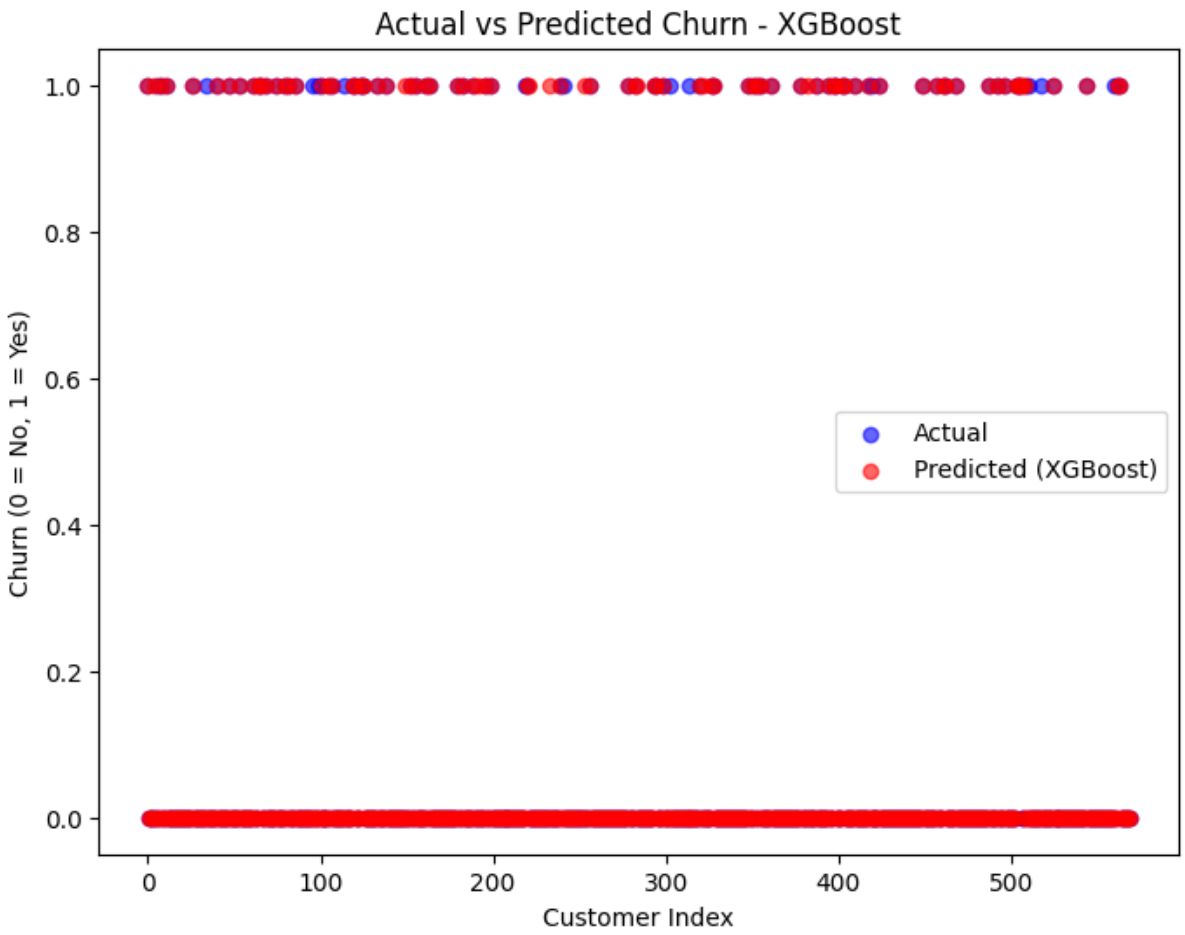
The confusion matrix shows that the model correctly predicted 471 non-churn cases and 75 churn cases, with 10 false positives and 14 false negatives.

Actual \ Predicted	0 (No Churn)	1 (Churn)
0 (No Churn)	471	10
1 (Churn)	14	75

---

8. Actual vs Predicted Churn Analysis

The scatter plot compares actual vs. predicted churn, showing a strong alignment between the model's predictions and the actual outcomes.



---

9. Business Recommendations

Key Takeaways from Churn Prediction:

- **High Complaint Customers Churn More**: Customer complaints strongly influence churn.
- **Subscription Length Matters**: Short-term subscribers are more likely to leave.
- **High-Usage Customers with Low Value Churn More**: Offer loyalty programs for high-usage, low-value customers.

Actionable Business Strategies:

- **Reduce Call Failures & Complaints** → Improve customer service response time.
- **Loyalty & Retention Programs** → Offer discounts for long-term subscription plans.
- **Targeted Marketing Campaigns** → Personalized engagement for customers with high churn probability

10. Conclusion

- **XGBoost was the best model** for predicting customer churn with **95.79% accuracy**.
- **Complaints, Call Failures, and Subscription Length** were the key indicators of churn.
- **Business Strategies** should focus on **improving customer experience & targeted retention offers**.

Future Work: Implement real-time churn prediction and customer segmentation to improve proactive intervention strategies.