SYRIATEL CUSTOMER CHURN PREDICTION

SUMAIYA ABDULLAHI OSMAN

22/02/2025

BUSINESS UNDERSTANDING

Problem statement

Syriatel, a leading telecommunications company, faces a major challenge: **customer churn**. Customers leaving the network results in revenue loss and increased costs to acquire new users. To maintain profitability and customer loyalty, syriatel needs a reliable way to **predict which customers are likely to churn** and take proactive retention measures.

Objective

The goal of this project is to **build a classification model** to **predict** whether customers are at risk of churning based on their **call usage, account features, and service complaints**. Syriatel can then implement **targeted retention strategies**, such as offering discounts or improving service.

DATA UNDERSTANDING

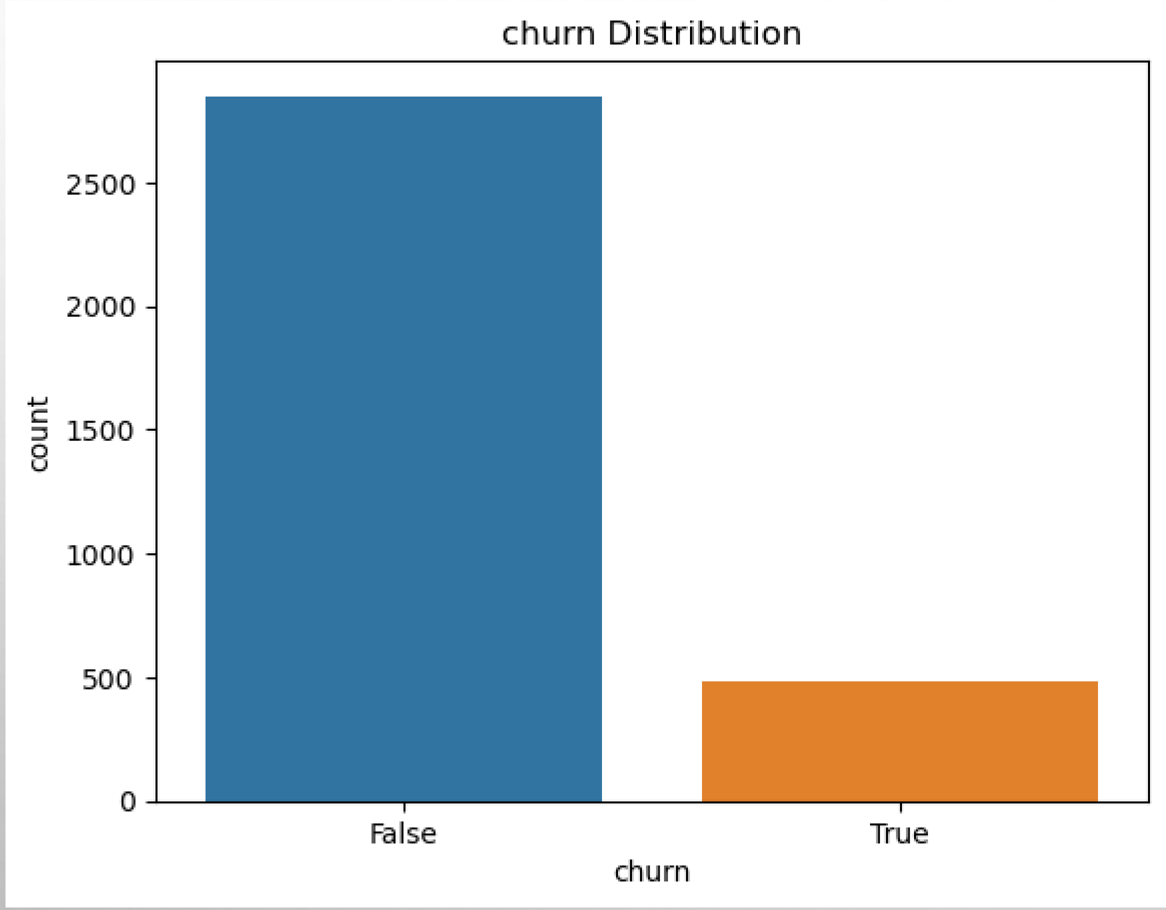
Dataset Source: Kaggle

- Features include: state, account length, area code, phone number, international plan, voice mail plan, and more.
- Target variable: `churn` (1 = churn, 0 = not churn)

Data Preprocessing & Feature Engineering

- Handled missing values and outliers.
- Scaled numerical features using minmaxscaler.
- Applied smote to handle class imbalance.
- Scaled train and test set using standardscaler.
- Selected important features using feature importance from xgboost and created new features.

Customer Distribution: Bar chart of churn vs. non-churn customers



- **The dataset is imbalanced** – only **14.5%** of customers churn.
- **Why it matters:** Models trained on imbalanced data may **struggle to predict churn correctly** because they are biased towards the majority class.
- **How we handle this:** Techniques like **SMOTE (oversampling), class weighting, or balanced metrics** are used to improve model performance.

Model Performance Comparison (F1-Score)

Random Forest performs best (~0.62)

•Strikes the best balance between Precision & Recall

•Effectively identifies churners while minimizing false positives

Decision Tree is a close second (~0.61)

•Slightly less balanced than Random Forest

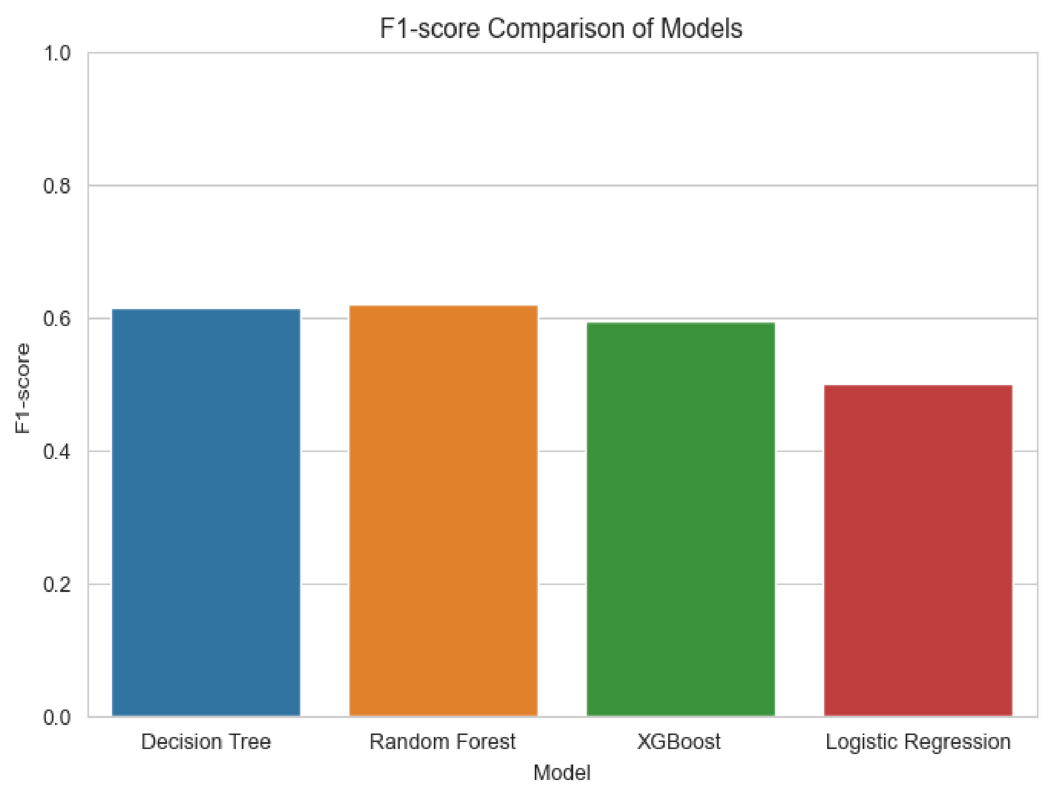
XGBoost performs slightly worse (~0.60)

•Despite being a strong model, hyperparameter tuning or data characteristics may have affected performance

Logistic Regression performs the worst (~0.50)

•Struggles to balance Precision & Recall

•High false negatives make it less suitable for this problem



Model Performance Comparison (Precision)

Decision Tree has the highest Precision (~0.55–0.56)

•More accurate when predicting churn

•Fewer false positives (misclassified non-churners)

Random Forest follows closely (~0.50–0.52)

•Still strong but slightly lower than Decision Tree

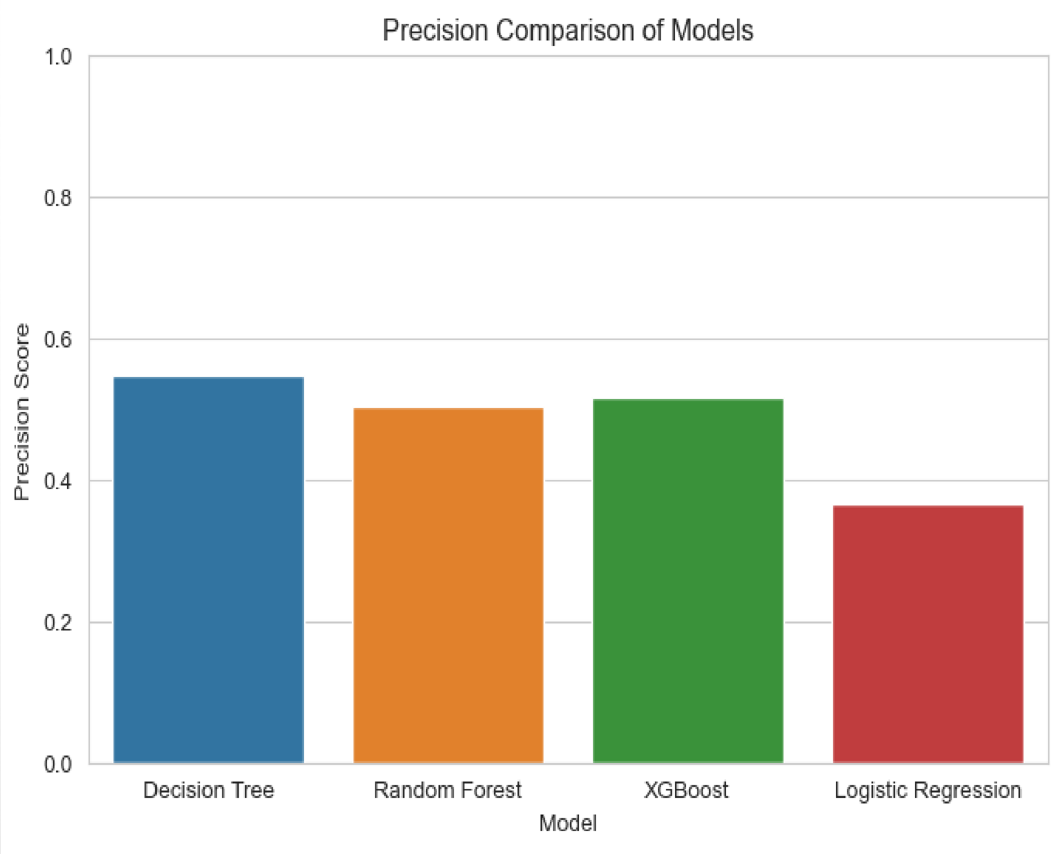
XGBoost has a slightly lower Precision (~0.48–0.50)

•Good Recall but misclassifies some non-churners as churners

Logistic Regression has the lowest Precision (~0.35–0.37)

•High false positives → wrongly predicts churn for many non-churners

•Could lead to **unnecessary interventions** for customers



Model Performance Comparison (Recall)

Random Forest & Logistic Regression have the highest Recall (~0.80)

•Best at identifying actual churners

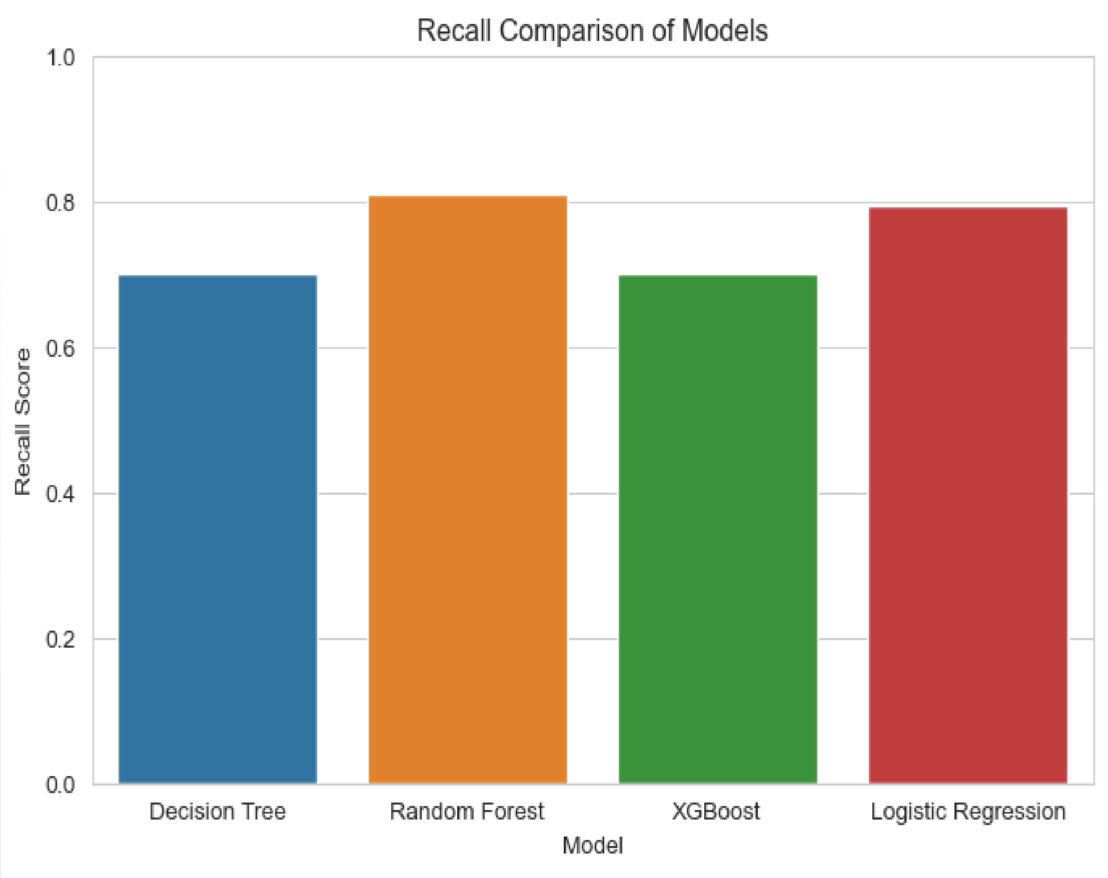
•**Fewer false negatives**, meaning fewer missed churners.

•Helps businesses detect and retain at-risk customers.

Decision Tree & XGBoost have slightly lower Recall (~0.70)

•Still effective but miss **more actual churners**

•Some at-risk customers may not be flagged.



Recommended Model – XGBoost

Model	Precision	Recall	F1-score
Decision Tree	0.548780	0.703125	0.616438
Random Forest	0.504854	0.812500	0.622754
XGBoost	**0.517241**	**0.703125**	**0.596026**
Logistic Regression	0.366906	0.796875	0.502463

Why Choose XGBoost?

✓ **Best Balance** between detecting churners (Recall) and avoiding false alarms (Precision)

✓ **Consistent Performance** across different validation sets

✓ **More Reliable** than Decision Tree & Random Forest for deployment

Feature Importance (from XGBoost)

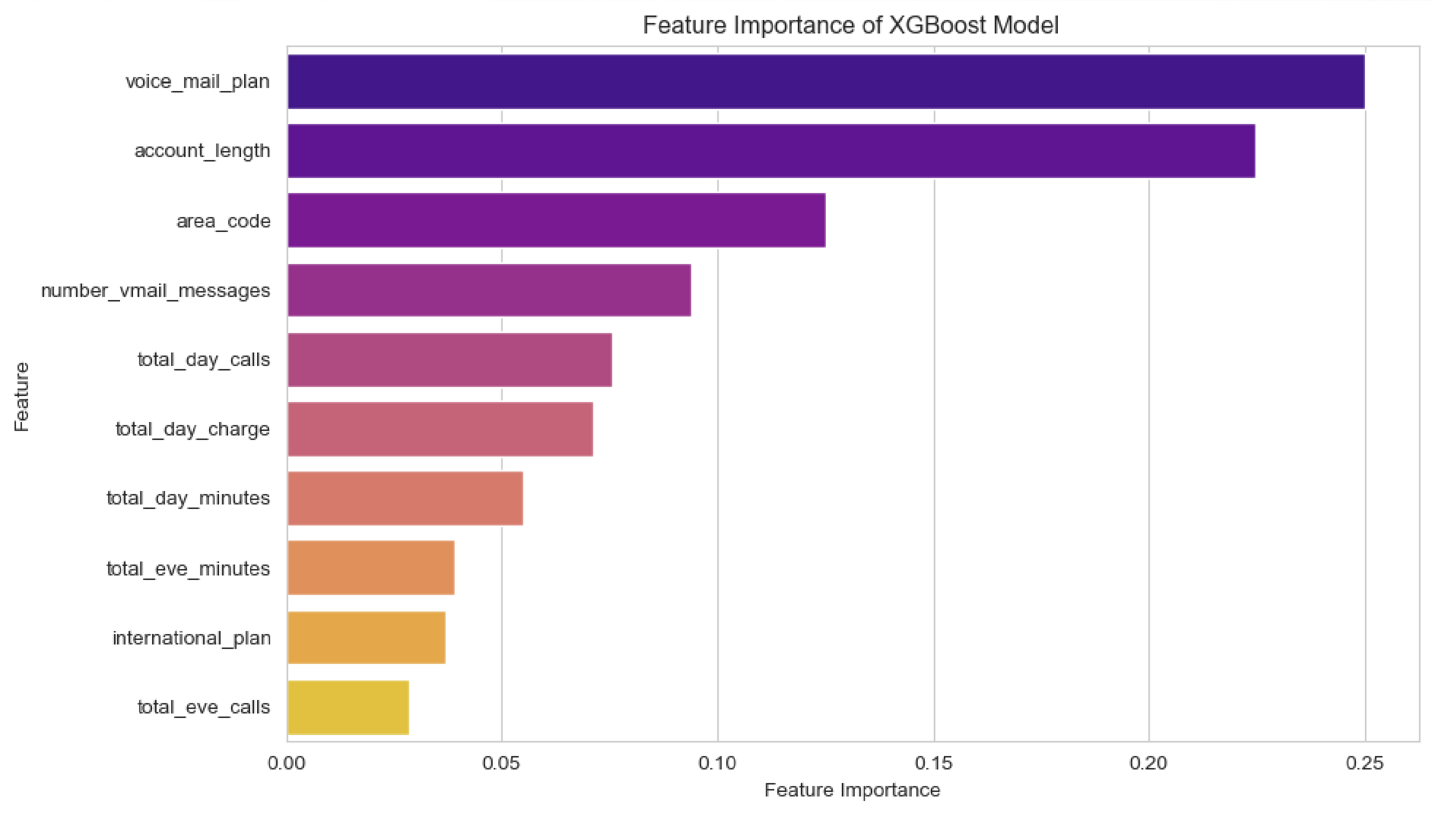
Findings

1.Voice Mail Plan is the most important feature, highly correlated with churn.

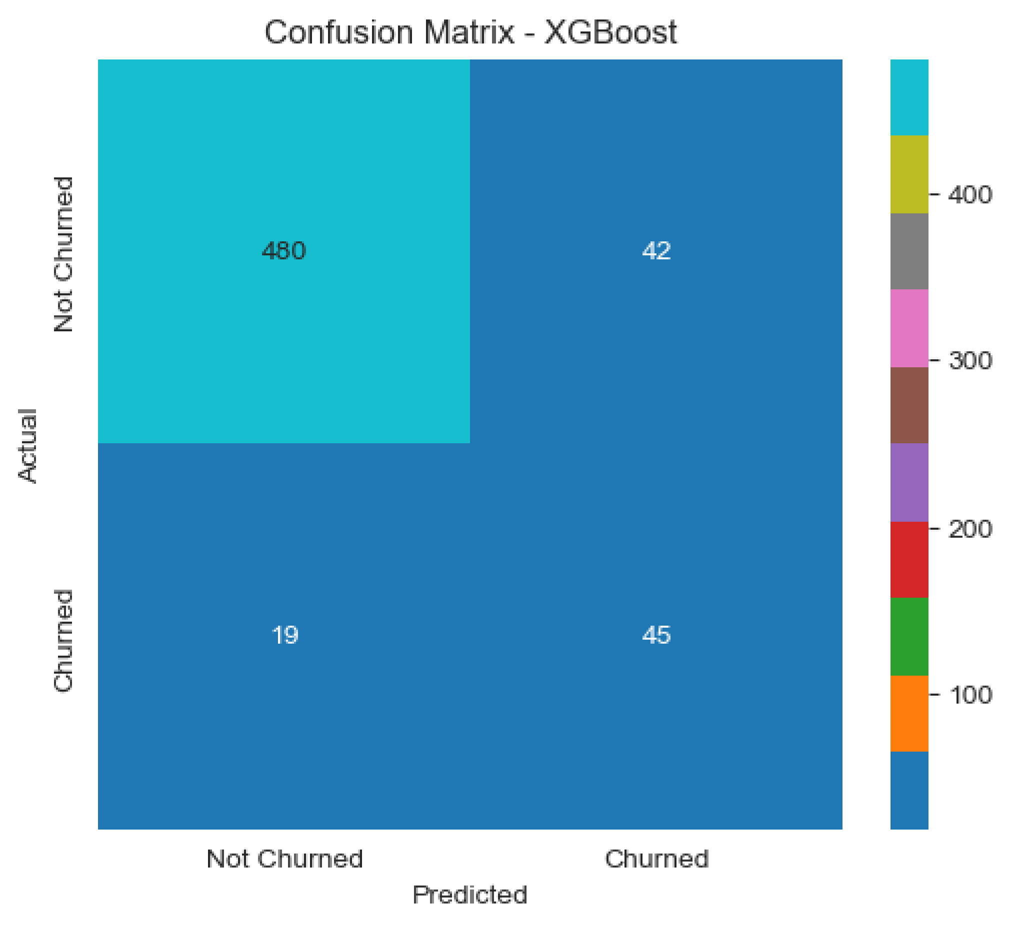
2. Account Length is the second most influential—longer tenure impacts churn.

3. International Plan & Voicemail Messages also play a role in churn prediction.

4. Other features (area code, total charges, calls, minutes) contribute but have less impact.



Confusion Matrix Heatmap



This heatmap visualizes the performance of the XGBoost model in predicting customer churn.

•**True Negatives (480)**: Correctly predicted non-churners.

•**False Positives (42)**: Wrongly predicted churn when the customer didn't churn.

•**False Negatives (19)**: Missed actual churners, predicting them as non-churners.

•**True Positives (45)**: Correctly identified churners.

Key Insight: The model performs well, but some churners are misclassified as non-churners, which could impact retention efforts.

Conclusion & Recommendations

Conclusion:

- **XGBoost** outperforms other models in churn prediction.
- High churn customers tend to have month-to-month contracts and high support calls.

Business Recommendations:

- Offer incentives for long-term contracts.
- Improve customer support experience to reduce churn risk.
- Implement targeted retention strategies for at-risk customers.

Next Steps

- **Deploy** the model into a production environment.
- **Integrate** predictive insights with CRM for proactive customer engagement.
- **Monitor** and **retrain** the model periodically to maintain accuracy.

Q&A

- **Thank You!**
- sumaiyaabdullahiosman@gmail.com