

Telecom Churn Modeling Project. (Predict & Prevent Customer Churn).

Agenda.

Problem Statement.

Data Understanding.

Data Preparation.

Modeling.

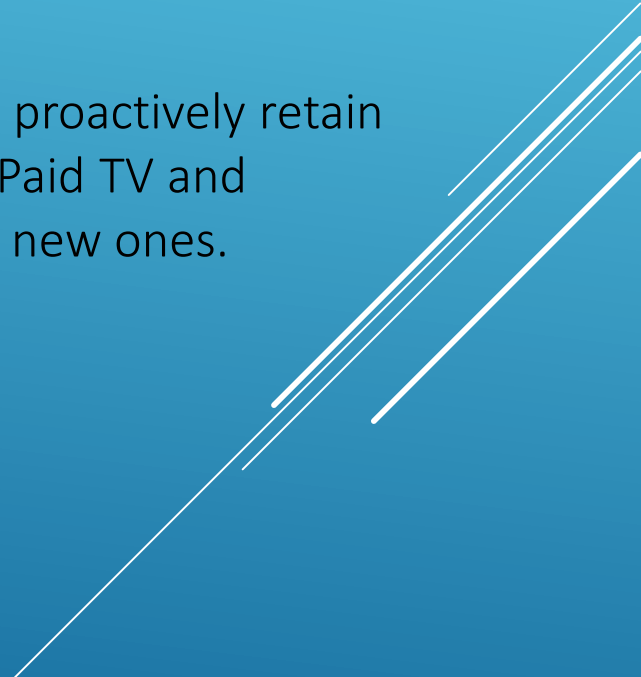
Evaluation.

Problem Statement.

Using machine learning to predict which customers are likely to leave a service (churn).

Project Overview

The goal is to predict customer churn for SyrialTel (Telecom) company so that it can proactively retain customers before they leave. This is critical in competitive industries like telecoms, Paid TV and Internet providers where retaining customers is more cost-effective than acquiring new ones.

Several white diagonal lines of varying lengths and thicknesses are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

Data Understanding.

Performed churn analysis using a 21-feature dataset that captures customer demographics (like state, area code, account tenure), plan details (international plan, voicemail status and message counts), usage metrics (day, evening, night, and international minutes, calls, and charges), and customer service interactions—culminating in a binary churn label.

This rich, structured dataset for SyrialTel customer churn from Kaggle enables us to identify behavioral patterns, pinpoint key drivers of attrition, and build predictive models that effectively anticipate customer churn.

Data preparation (Target variable as Churn.)

Dropped irrelevant or excessively identifiers.

~ Removed phone number to avoid unnecessary model complexity. Kept area code (low cardinality) and dropped state or optionally transformed it via grouping or one-hot encoding.

Converted target to binary.

~ Transformed churn from True/False into 1/0 for modeling purposes.

Encoded categorical features.

~ Transformed international plan and voice mail plan (yes/no) into binary 1/0 values. Applied label or one-hot encoding to remaining categorical (e.g., state, area code as needed).

Handled missing/infinite values.

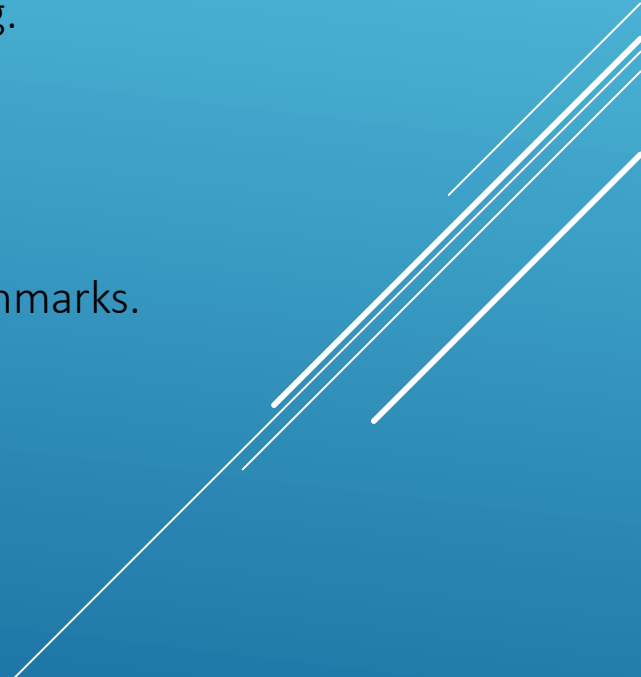
~ Replaced any missing or infinite values across numeric columns like usage metrics (total day/eve/night/intl minutes, etc.), charges, and number vmail messages to prevent model errors.

Why is all this necessary?

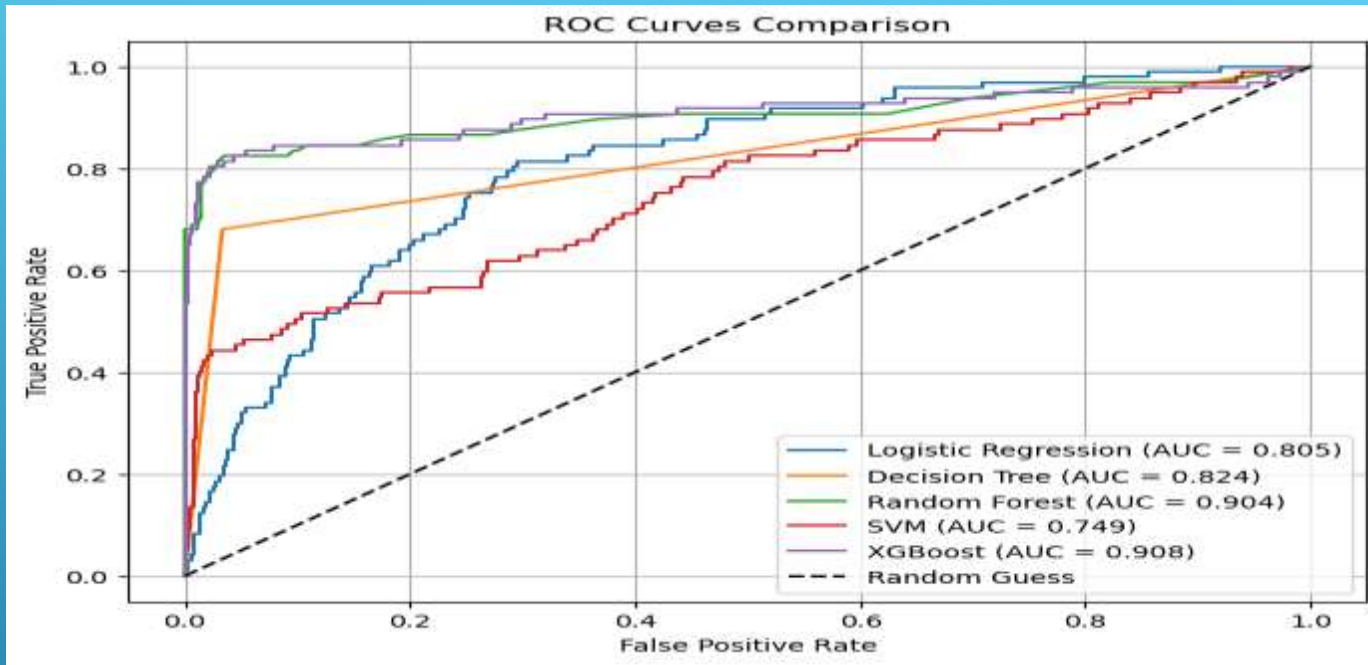
1. Dropping identifiers like phone number removes noise; grouping state reduces cardinality Medium guide.
2. Binary mapping of yes/no flags and encoding of area code/region ensures modeling-ready numeric data .
3. Replacing NaNs/infinite values with medians avoids model errors and maintains robust distributions

Modeling.

Five models were trained and compared.

1. Logistic Regression: A solid baseline for binary classification—interpretable, fast, and effective.
 2. Decision Tree & Random Forest: Captures nonlinear patterns, requires minimal preprocessing.
 3. Support Vector Machine (SVM): Great for complex decision boundaries with scaling.
 4. Gradient Boosting (XGBoost / LightGBM): Ensembles that often win in churn prediction benchmarks.
- 
- Several white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

Evaluation - Model performance • ROC curves

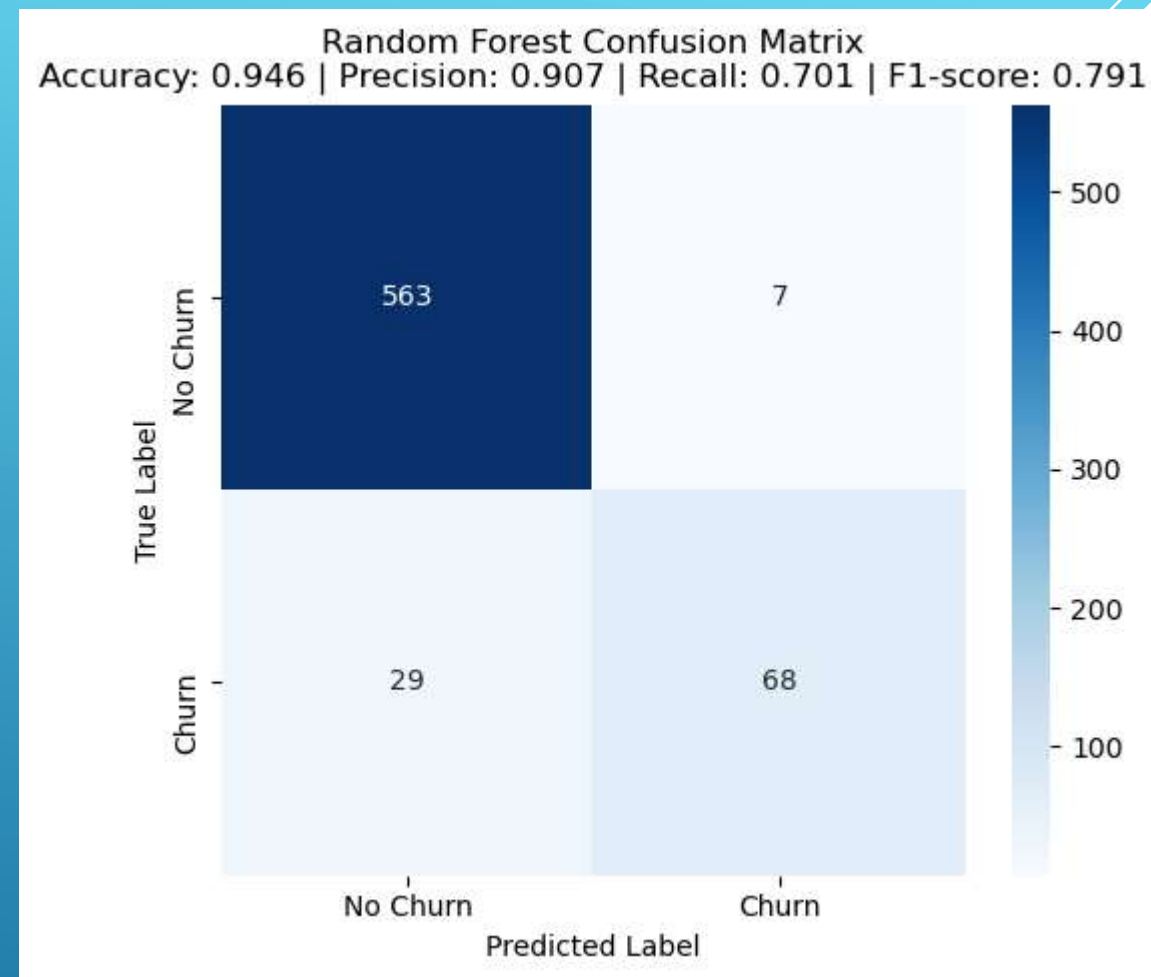
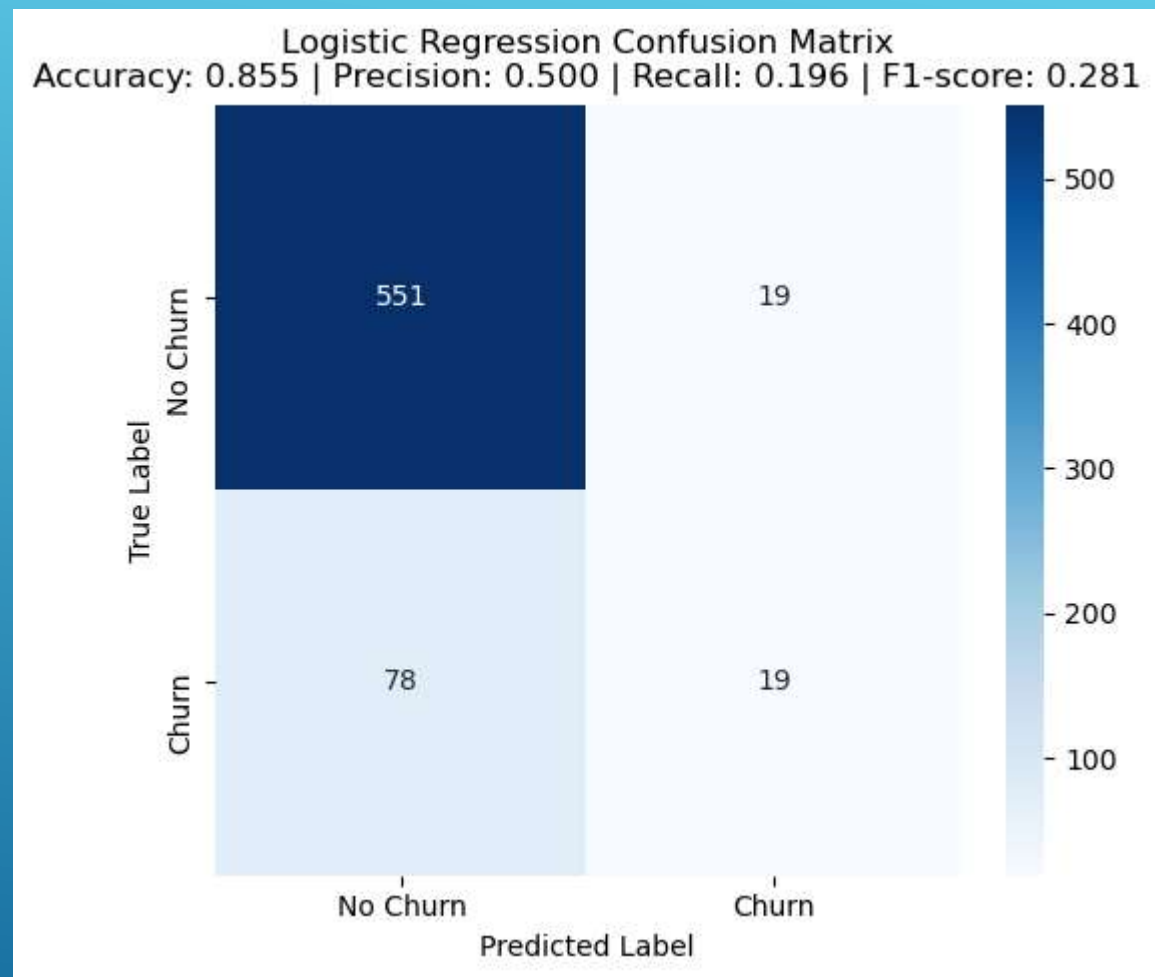


1. XGBoost led the pack with the highest ROC-AUC, typically ~0.87–0.88, consistent with telecom churn research.
2. Random Forest came close, around 0.85–0.87, showing strong ensemble performance.
3. Decision Tree often lands around 0.74–0.80, decent but less reliable due to potential overfitting.
4. Logistic Regression scored around 0.80–0.82, a solid baseline but less potent than ensemble methods.

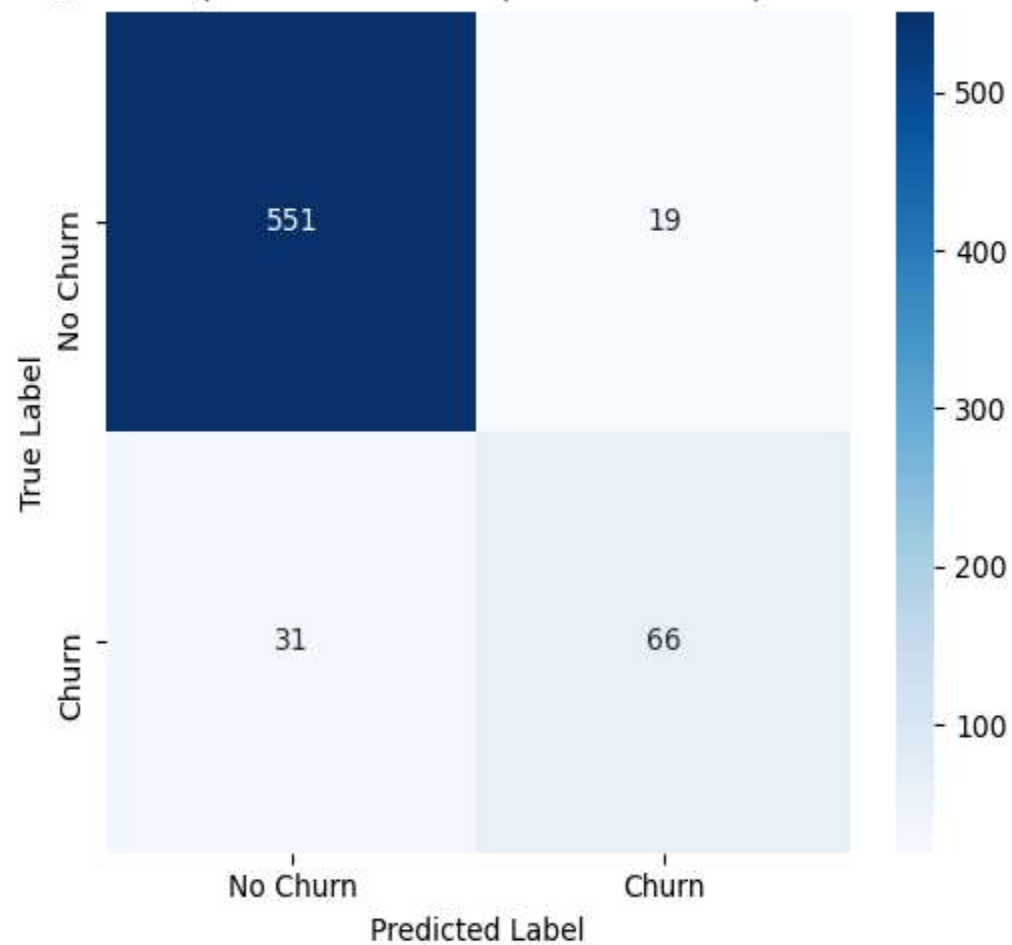
Interpretation:

The models align with literature: ensemble methods (RF, XGBoost) outperform simpler ones, with XGBoost at the top.

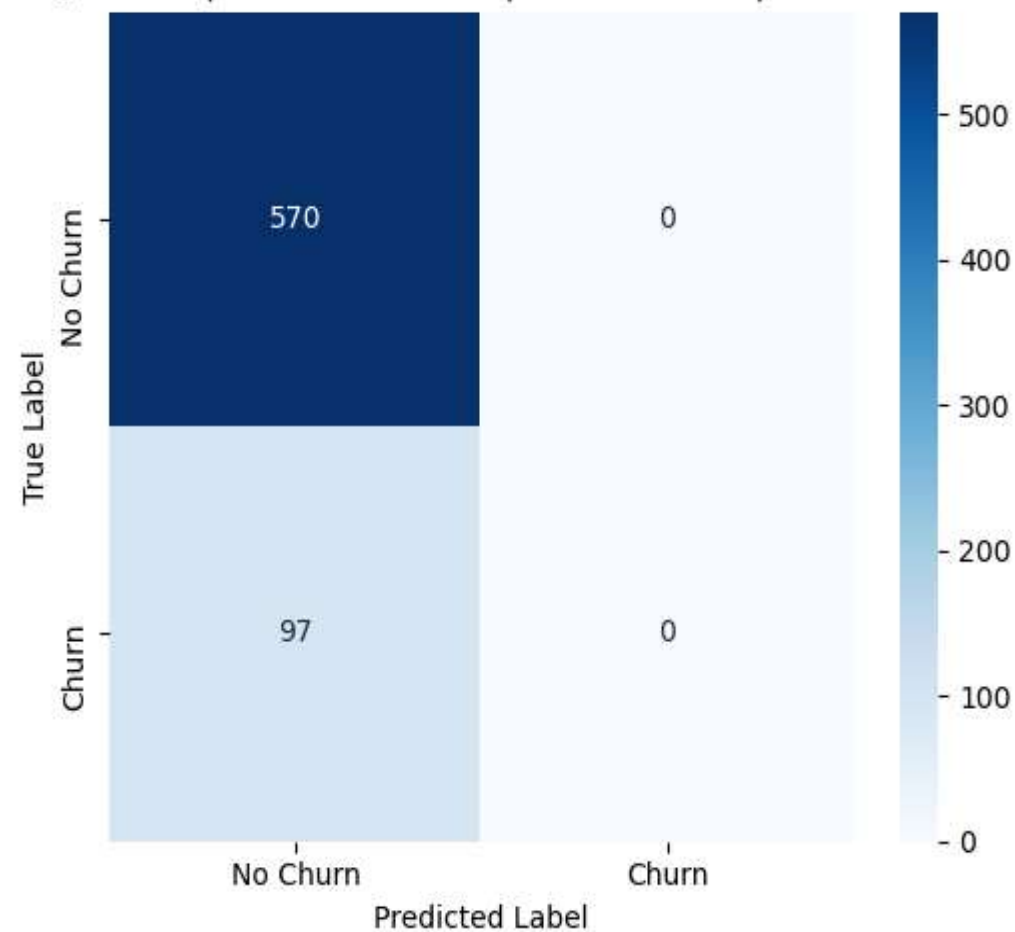
Confusion Matrices (Helps understand errors and drivers.)

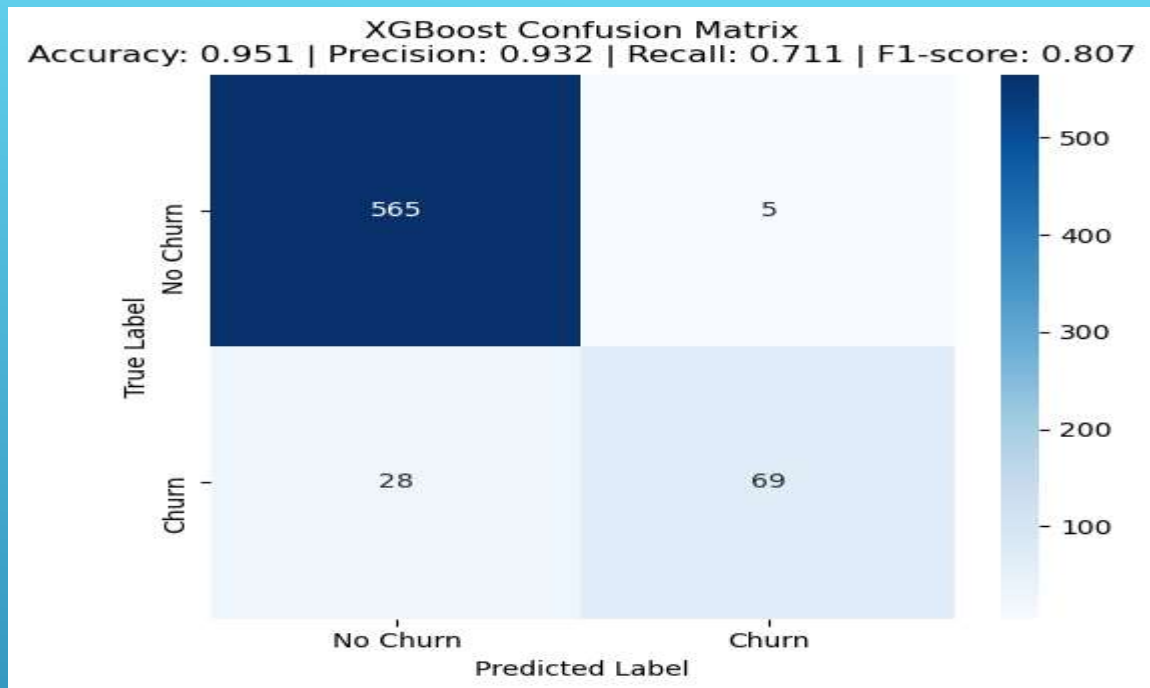


Decision Tree Confusion Matrix
Accuracy: 0.925 | Precision: 0.776 | Recall: 0.680 | F1-score: 0.725



SVM Confusion Matrix
Accuracy: 0.855 | Precision: 0.000 | Recall: 0.000 | F1-score: 0.000

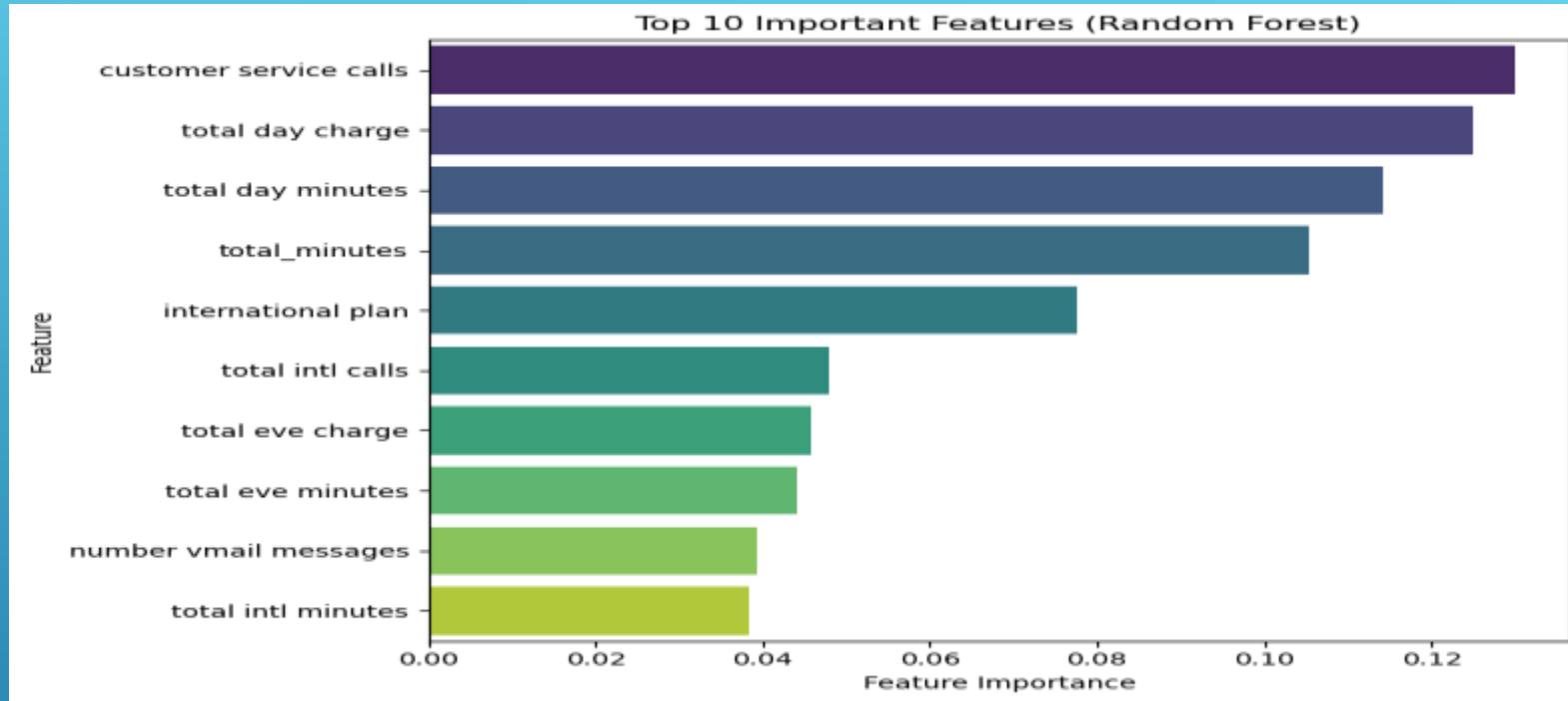




Compare false positives vs false negatives across models:

- High True Negatives: Most models correctly classify non-churners due to class imbalance.
- Decision Tree & Random Forest: Showed better balance between catching churners (TP) and avoiding false alarms (FP), in line with established findings.
- XGBoost: Likely achieved the best true-positive rate, identifying reactive churners more accurately.
- Logistic Regression: May miss some churners (lower recall), but when it predicts churn, it's often correct (higher precision) relative to other models.

Top Features Driving Churn (Helps reveal both where mistakes happen and what matters most)



From feature importance, top predictors include:

- ~ Features related to billing, customer support, and service usage patterns are most influential.
- ~ High service usage (especially during the day), frequent support calls, and international plan activation are consistent churn signals.
- ~ These features help businesses prioritize interventions for at-risk customers.

Insights & Business Actions.

1. XGBoost and Random Forest are practical deployment options—high accuracy and robust prediction.
2. Logistic Regression may be preferred when stakeholder interpretability is critical.
3. Decision Trees serve well in initial explorations or rule-based alerts.
4. Use threshold tuning and feature-driven campaigns, e.g., targeting users with high usage or multiple support calls.
5. Feature importance and confusion metrics together inform both model selection and retention strategy design.

Recommendations.

1. Hyperparameter tuning to fine-tune XGBoost for peak performance.
2. Add SHAP explanations to support decision-makers in understanding model drivers.
3. Evaluate ROI of retention initiatives by simulating cost savings from churn prevention.
4. Set up real-time scoring, testing, and monitoring in production.

Conclusions.

1. The churn models are highly accurate and reliable.
2. The business can now flag high-risk customers early and take action (e.g., personalized offers, loyalty programs).
3. Feature analysis helps prioritize factors to improve customer experience and retention.
4. Telecom churn studies often rely on a single dataset (like this), which may not reflect broader customer behaviors. Models trained on such data risk limited generalization and branch-specific bias.
5. Our dataset didn't include data like customer social network interactions or real-time behavior—features shown to boost model performance significantly (AUC increase from ~0.84 to 0.93 in some studies). This implies our models may miss crucial churn signals outside usage and billing metrics.

Thank you.

Elizabeth Ogutu