

Dataset: SyriaTel Churn Dataset. Characteristics: The dataset is well-structured with 3,333 records and 20 features (after dropping phone number). No missing values were present, simplifying the imputation process. Class Imbalance: The target variable (churn) is highly imbalanced, with ~85.5 Non-Churners and ~14.5 Churners. This confirms the necessity of prioritizing metrics like Recall and F1-score over simple Accuracy.

Key Data Insights (EDA)

International Plan is Critical: Customers with the International Plan had a churn rate of 42.4, compared to just 11.5 for those without. This feature is a powerful, simple predictor. Customer Service Calls: The average number of customer service calls was noticeably higher among churned customers, suggesting service friction is a major risk factor.

3. Modeling

Methodology: Pipelines and Tuning

To ensure a robust and leak-free process, all modeling utilized scikit-learn's Pipeline and ColumnTransformer:

- Data Split: Data was split into Training (75) and Test (25) sets, stratified by the churn target to preserve the class balance.
- Preprocessing (ColumnTransformer): Numeric: Standardized using StandardScaler. Categorical: One-Hot Encoded using OneHotEncoder.
- Hyperparameter Tuning: All models were tuned using GridSearchCV with 5-fold Cross-Validation and the primary scoring set to scoring="recall" to ensure optimal performance on the business objective. Models Evaluated

Model	Purpose	Tuning Parameter
Baseline Logistic Regression	Interpretability & Baseline. Provides initial understanding of linear feature effects.	N/A (Untuned)
Tuned Logistic Regression	Performance Improvement. Tuned the regularization strength, C.	model__C
Tuned Decision Tree	Non-Parametric & High Performance. Captures complex, non-linear relationships.	model__max_depth, model__min_samples_split, model__min_samples_leaf

4. Evaluation

Model	Test Recall (Churn=1)	Test Precision (Churn=1)	Test F1-Score	Test ROC-AUC
Tuned Decision Tree (Final)	0.645	0.876	0.743	0.822
Tuned Logistic Regression	0.273	0.55	0.365	0.792

Final Model Selection Rationale

The Tuned Decision Tree is the chosen production model. Superior Recall: It achieved a Recall of 0.645, a massive improvement over the 0.273 achieved by the best Logistic Regression model. This directly addresses the goal of minimizing costly missed churners (False Negatives). High Efficiency: It maintained a strong Precision of 0.876, meaning 87.6 of the customers flagged for intervention are genuine churn risks. This ensures that the retention budget is spent efficiently. Excellent Discrimination: The ROC-AUC of 0.822 confirms the model has strong ability to correctly rank customers by their probability of churn.

5. Conclusion

The Decision Tree provides SyriaTel with a powerful tool to generate accurate, high-recall risk scores. This tool shifts the retention strategy from being reactive to being proactive, significantly boosting the potential for customer retention and revenue protection.

Actionable Recommendations

- Immediate Deployment of the Decision Tree: Implement the Tuned Decision Tree to generate a weekly customer risk ranking. Use the churn=1 prediction as the trigger for a personalized retention offer.
- Target High-Risk Features: Prioritize the following customers for immediate support or special offers: Customers with the International Plan. Customers with 3 or more recent customer service calls.

3. Threshold Optimization (Next Step): Before full-scale deployment, quantify the exact cost of a False Negative (lost revenue) and a False Positive (cost of intervention). Use this financial information to set a profit-maximizing threshold for the model's output probability (instead of the default 0.5), ensuring the intervention program delivers the highest possible ROI.