Customer Churn Prediction Model Report

Acquiring new customers is more expensive than retaining existing ones. This report details a predictive model to identify high-risk churn customers, enabling proactive retention strategies.

Why Random Forest?

Non-linear Relationships

Handles complex interactions in customer behavior data.

Robustness

Resistant to noise and outliers common in real-world data.

Feature Importance

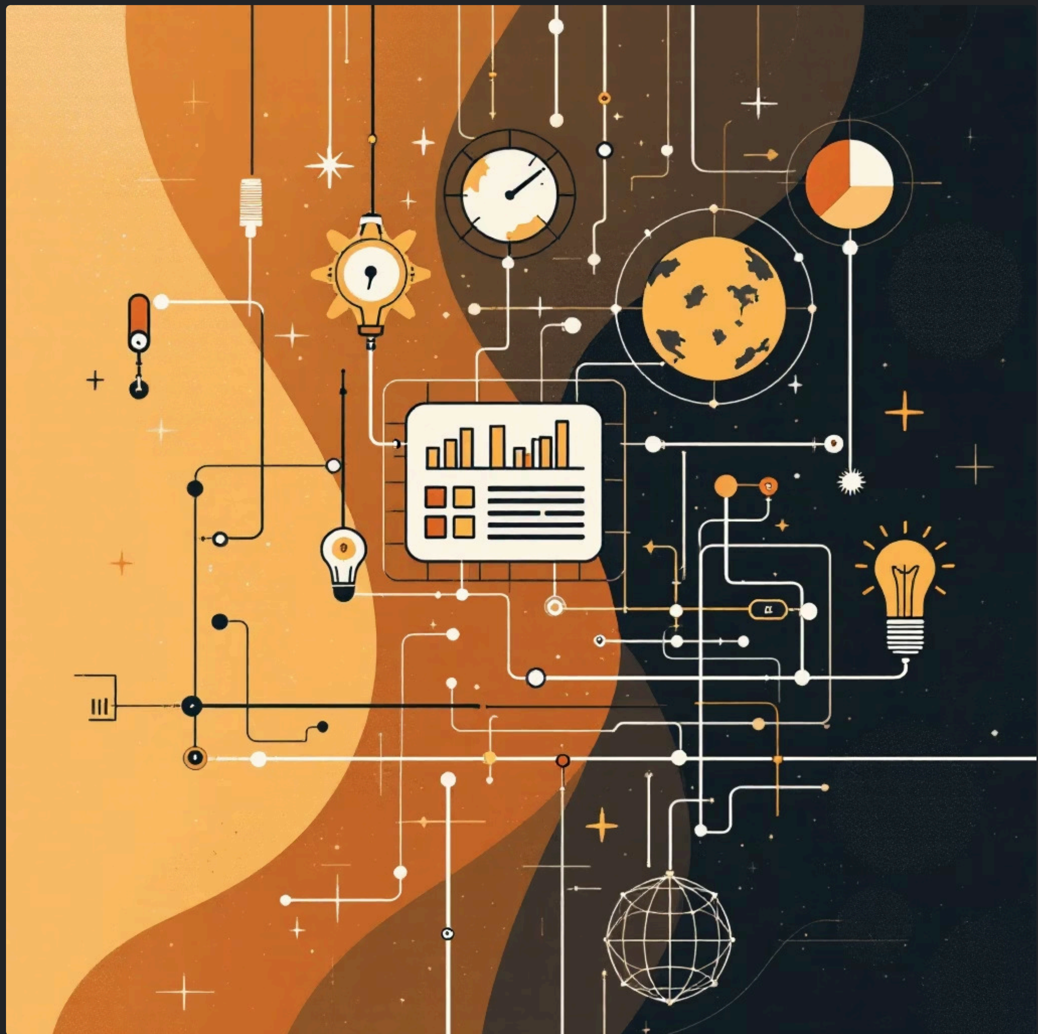
Provides insights into key churn drivers for better interpretability.

Tabular Data Performance

Highly effective for structured business datasets.

Random Forest balances accuracy, robustness, and explainability, making it ideal for this business context.

Model Training: Data Preparation



Target Variable

ChurnStatus (1 = Churned, 0 = Retained).

Data Cleaning

Removed non-predictive columns like CustomerID and date fields.

Encoding

Categorical variables (e.g., ServiceUsage) used one-hot encoding.

Data Split

Stratified train-test split to preserve class proportions.

Addressing Class Imbalance

Non-churn customers represent ~80% of observations, churn customers ~20%. This imbalance affects model behavior and makes accuracy unreliable.

Class Weighting

Applied `class_weight="balanced"` to the model.

Metric Prioritization

Prioritized recall and F1-score over accuracy.

Threshold Tuning

Introduced classification threshold tuning for optimal results.

Model Optimization

1

Feature Engineering

One-hot encoding for categorical variables, preventing data leakage.

2

Cross-Validation

5-fold cross-validation for model generalization.

3

Hyperparameter Tuning

Optimized parameters (trees, depth, samples) using GridSearchCV, focusing on recall.

Initial Model Challenges

The initial model showed high accuracy but failed to identify actual churners. It was "safe" but not "useful."

Problem:

Most customers didn't churn, so the model predicted "no churn" for almost everyone.

Business Impact: Missing a churner means lost revenue; flagging a loyal customer means extra attention.

Accuracy alone was insufficient; we needed to catch more churn customers.

Refining the Model: Focus on Recall

We shifted our focus from overall accuracy to recall, prioritizing the identification of actual churners.

1

Changed Success Metric

From "How often is the model correct?" to "How many churn customers does the model catch?"

2

Increased Sensitivity

Treated churn mistakes as more serious and adjusted confidence thresholds.

This made the model more willing to flag at-risk customers, accepting a trade-off for higher recall.

Model Performance: Key Metrics

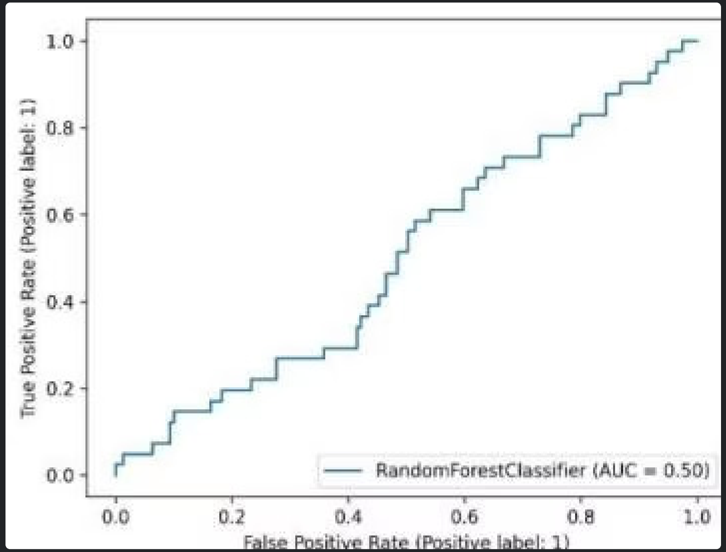
The final model prioritizes detecting churn, providing useful risk signals rather than perfect predictions.

Precision	0.80	0.21
Recall	0.50	0.51
F1-score	0.61	0.30

Overall Accuracy: 50% (Threshold = 0.3)

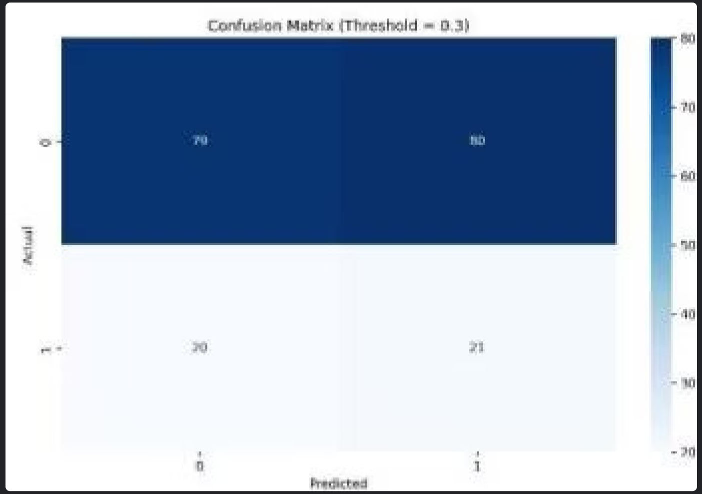
Interpretation of Results

ROC Curve (AUC = 0.50)



Indicates random classification; model struggles to separate churned from retained customers with current features.

Confusion Matrix (Threshold = 0.3)



High false positive rate; many non-churners flagged, while actual churners are still missed.

The Precision–Recall curve also shows consistently low precision, limiting practical usefulness.

Business Applications & Future Improvements

Applications:

- Proactive Retention Campaigns
- Customer Segmentation
- Operational Decision Support
- Strategic Planning

Areas for Improvement:

- Model Enhancements (e.g., XGBoost, SMOTE)
- Advanced Feature Engineering (e.g., CLV, sentiment)
- Threshold Optimization
- Continuous Model Monitoring and Retraining