1. Selected Machine Learning Algorithm

For this project, we chose the **Random Forest Classifier**.

- **Rationale:** Random Forest is an ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It was chosen for its **strong performance with a minimal risk of overfitting**. While we also tested XGBoost, Random Forest provided a similar level of performance without the need for extensive hyperparameter tuning. It also handles multicollinearity and feature interactions well, making it a robust choice for the dataset.

---

2. Model Performance and Evaluation

Despite the use of advanced feature engineering and data balancing techniques (SMOTEENN), the final model's performance indicates that it is **not yet a reliable tool** for predicting customer churn. The key challenge remains the **lack of a strong predictive signal** within the available features.

Final Evaluation Metrics

Metric	Score	Interpretation
Accuracy	0.56	The model's overall correct predictions are just above a coin flip.
Precision (for Churn)	0.19	When the model predicts churn, it is only correct **19%** of the time.
Recall (for Churn)	0.34	The model correctly identifies only **34%** of customers who will actually churn.
ROC-AUC	0.482	The model's ability to distinguish between churners and non-churners is worse than random chance (0.5).

Confusion Matrix

The confusion matrix shows the model's performance on the test data.

Actual Churn	Predicted No Churn	Predicted Churn
No Churn	148 (True Negatives)	91 (False Positives)
Churn	40 (False Negatives)	21 (True Positives)

The high number of **False Positives (91)** is a significant issue, as it means the model is generating many false alarms. The model also misses a large portion of actual churners, as shown by the **False Negatives (40)**.

---

3. Business Utilisation and Areas for Improvement

Given its current state, the model should be used as an **exploratory tool** rather than a definitive decision-making system.

Utilisation for Business Decisions

- **Prioritised Engagement:** The model can rank customers by churn risk. The business can then focus limited resources on the top-tier customers with the highest risk, even with the model's low accuracy.
- **Targeted A/B Testing:** The model can be used to identify a specific group of at-risk customers for testing new retention strategies, such as special offers or personalized outreach.

Recommendations for Improvement

The most impactful way to improve this model is to **enrich the dataset with new, high-value features**. The current features lack the strong signal needed to predict churn.

1. **Acquire Behavioral Data**: Incorporate detailed data on customer usage and interactions. This is the single most important step for improving the model.
2. **Incorporate Qualitative Data**: Gather and integrate data from customer feedback surveys, support tickets, and sentiment analysis.
3. **Explore Customer Segmentation**: Develop separate models for different customer groups (e.g., high-value vs. low-value) as their churn drivers may be different.