1. Executive Summary

This report details the third milestone of the churn prediction project: machine learning model development and optimization. A total of seven different classification models were trained, optimized, and evaluated to find the most accurate predictor of customer churn.

To ensure model fairness and accuracy, the training data was first balanced using the ADASYN oversampling technique. Following a comprehensive GridSearchCV tuning process, the XGBoost classifier was identified as the best-performing model, achieving the highest test accuracy of 97.03%. This model demonstrates excellent generalization and is the recommended model for deployment.

2. Methodology

The evaluation process followed a structured machine learning pipeline:

- **Data Preparation:** The pre-scaled and processed training and test datasets were loaded (X_train_scaled.csv, X_test_scaled.csv, y_train.csv, y_test.csv).
- **Class Imbalance Handling:** The ADASYN (Adaptive Synthetic Sampling) algorithm was applied to the training data to correct for its imbalanced nature. All models were trained on this resampled dataset.
- **Model Selection:** Seven distinct classifiers were selected for comparison:
    - XGBoost
    - Random Forest
    - Decision Tree
    - Support Vector Machine (SVC)
    - K-Nearest Neighbors (KNN)
    - Logistic Regression
    - Gaussian Naive Bayes (GaussianNB)
- **Hyperparameter Tuning:** GridSearchCV was used to find the optimal hyperparameters for each model, using 5-fold cross-validation and accuracy as the primary scoring metric.

3. Model Performance Comparison

All optimized models were evaluated on the unseen test set. The summary table below, generated by the notebook, ranks the models based on their final test accuracy.

3.1. Overall Performance Ranking (by Test Accuracy)

- XGBoost: 0.97
- Random Forest: 0.96
- Decision Tree: 0.92
- SVC: 0.88
- KNN: 0.85
- Logistic Regression: 0.84
- GaussianNB: 0.84

3.2. ROC Curve Analysis

The ROC (Receiver Operating Characteristic) curve plots the models' ability to distinguish between classes (churn vs. no churn). The models with curves closest to the top-left corner (highest Area Under the Curve, or AUC) are the best. The plot confirms that Random Forest (AUC = 0.92) and XGBoost (AUC = 0.90) are the most effective classifiers.

4. Detailed Model Analysis

The following is a model-by-model analysis of train vs. test performance, highlighting overfitting and potential tuning strategies.

XGBoost

- Train Accuracy: 1.00
- Test Accuracy: 0.97
- Analysis: This model achieved the highest test accuracy in the evaluation. The small gap between train and test scores indicates strong generalization. Performance can be further balanced with speed by tuning the learning_rate, max_depth, subsample, and n_estimators.

Random Forest

- Train Accuracy: 1.00
- Test Accuracy: 0.96
- Analysis: This model shows excellent generalization with only a minor signal of overfitting. It is a very strong alternative to XGBoost. Variance and training time can be controlled by adjusting n_estimators, max_depth, and max_features.

Decision Tree

- Train Accuracy: 1.00
- Test Accuracy: 0.92
- Analysis: The large gap between train and test scores indicates strong overfitting, which is common for decision trees as they tend to split aggressively to fit the training data perfectly. Limiting the max_depth or increasing min_samples_split would help stabilize performance.

SVC (Support Vector Machine)

- Train Accuracy: 1.00
- Test Accuracy: 0.88
- Analysis: A wide gap exists between the train and test accuracy, indicating overfitting. This variance can be reduced by tuning the C (regularization), gamma, and kernel parameters.

KNN (K-Nearest Neighbors)

- Train Accuracy: 1.00
- Test Accuracy: 0.85
- Analysis: This model shows heavy overfitting, which likely means a small k (like k=1) was chosen, causing the model to "memorize" the training set. Stability can be improved by increasing k and experimenting with different distance metrics.

Logistic Regression

- Train Accuracy: 0.85
- Test Accuracy: 0.84
- Analysis: This model is very stable, with almost no gap between train and test performance. While the data was already scaled, further performance gains (if needed) would come from tuning the C parameter and penalty type.

GaussianNB (Gaussian Naive Bayes)

- Train Accuracy: 0.84
- Test Accuracy: 0.84
- Analysis: This model is the most balanced, showing no overfitting. Its performance is likely limited by its core assumption of feature independence, which may not be true for this dataset.

5. Visual Error Analysis (Confusion Matrices)

The confusion matrices for each model provide a visual breakdown of correct and incorrect predictions on the test data. The top-row models (XGBoost and Random Forest) clearly show the smallest number of errors, particularly in False Negatives (customers who churned but were predicted not to), which is often the most costly error in a churn system.

6. Conclusion & Discussion

The XGBoost classifier is the recommended model, as it provides the best balance of high accuracy (97.0%) and strong generalization. This model was saved as best.joblib for deployment.