**Customer Churn Predictive Analysis: Phase 2 Report**

**Author: Costas Pinto**

Data Science Team

**Date:** August 3, 2025

**Status:** Phase 2 Complete

**1.0 Executive Summary**

* **Objective:** This report details the second phase of the customer churn prediction project. The primary goal was to build, tune, and evaluate a suite of machine learning models using the prepared dataset from Phase 1 to identify the most effective algorithm for predicting customer churn.
* **Methodology:** Four distinct machine learning algorithms were evaluated: Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). Each model was trained on the engineered dataset, incorporating SMOTE to handle class imbalance, with hyperparameters tuned using a 5-fold stratified cross-validation process to maximize the ROC-AUC score.
* **Key Findings:**
  + The comprehensive modeling pipeline ran successfully, but the resulting predictive performance across all tested algorithms was low. The best-performing model, a **Random Forest**, achieved a **Test AUC score of 0.5504**, which is only marginally better than random guessing.
  + This indicates that while the modeling framework is robust, the features currently available in the dataset, even after advanced engineering, do not contain strong enough signals to accurately predict churn.
  + The feature importance analysis from the best model suggests that behavioral recency features (DaysSinceLastLogin, DaysSinceLastTransaction) are the most influential, confirming that customer engagement is a critical area for further investigation.
* **Recommendations & Next Steps:**
  + **Do Not Deploy:** The current model is **not recommended for production deployment** due to its low predictive power.
  + **Focus on Data Enrichment:** The primary recommendation is to revisit the data gathering phase to acquire more predictive features. Priority should be given to data that captures customer sentiment, detailed service usage patterns, and competitor interactions.
  + **Refine Feature Engineering:** Further exploration of feature engineering, such as analyzing the sequence of customer interactions over time, is warranted.

**2.0 Model Development and Evaluation**

* **2.1 Dataset:** The analysis utilized the prepared\_churn\_dataset.csv file generated in Phase 1. The data was split into an 80% training set and a 20% testing set. A synthetic validation set was also generated to monitor for overfitting.
* **2.2 Algorithm Selection:** A diverse set of four industry-standard classification algorithms was chosen to ensure a comprehensive evaluation.
* **2.3 Handling Class Imbalance:** The SMOTE (Synthetic Minority Over-sampling Technique) was integrated into the training pipeline to create a balanced set of training data for each cross-validation fold.
* **2.4 Hyperparameter Tuning:** GridSearchCV was employed to systematically test different combinations of parameters for each algorithm, optimizing for the highest ROC-AUC score during cross-validation.

**3.0 Model Performance Results**

The table below summarizes the performance of each tuned model on the unseen test data.

|  |  |  |
| --- | --- | --- |
| **Model** | **Test Accuracy** | **Test AUC** |
| **Random Forest** | **0.795** | **0.5504** |
| Gradient Boosting | 0.680 | 0.5735 |
| SVM | 0.685 | 0.5074 |
| Logistic Regression | 0.465 | 0.5331 |

* **3.1 Analysis of Best Model (Random Forest):** While the Random Forest was the top performer, its low AUC score indicates a limited ability to reliably distinguish between churning and non-churning customers. The confusion matrix below shows that while it correctly identified many non-churners, it struggled significantly to identify actual churners.

[INSERT PLOT: confusion\_matrix\_phase2.png]

* **3.2 ROC Curve:** The ROC curve for the Random Forest model visually confirms its proximity to the baseline (random guess) line, reinforcing the low AUC score.

[INSERT PLOT: roc\_curve\_phase2.png]

* **3.3 Feature Importance Insights:** The feature importance analysis from our best model, Random Forest, consistently highlighted the following features as the most influential predictors:

[INSERT PLOT: feature\_importance\_phase2.png]

* 1. **DaysSinceLastLogin:** How recently a customer has engaged with online services.
  2. **CustomerTenure:** How long the customer has been with the bank.
  3. **DaysSinceLastTransaction:** How recently the customer has made a transaction.

This provides a clear, actionable insight: **customer engagement and tenure are the most critical areas influencing churn.**

**4.0 Conclusion and Recommendations**

The Phase 2 modeling process has been successfully executed. While a best-performing model (Random Forest) was identified, its predictive power is currently insufficient for reliable production use.

The key takeaway from this phase is that the project's success is contingent on enriching the dataset with more powerful predictive features.

**Recommendations:**

1. **Prioritize Data Acquisition:** Engage with business units to identify and integrate new data sources. Potential sources include:
   * Customer survey data (e.g., Net Promoter Score - NPS).
   * Detailed product/service usage logs.
   * Marketing campaign interaction data.
2. **Iterate on Feature Engineering:** Dedicate further resources to creating more sophisticated features, such as time-series analysis of transaction frequency or identifying shifts in customer spending behavior over time.
3. **Re-evaluate Models:** Once the dataset has been enhanced, the modeling pipeline should be re-run to evaluate the performance uplift.