Report for Project A - Happy Customers

Yuhao Huang
Apziva
Confidential - Internal Use Only

November 9, 2025

Contents

1 Executive Summary

The primary objective of this project was to develop a predictive model to identify happy customers based on survey data provided by the company. The dataset contained 127 observations with six independent variables, and the target variable indicated whether a customer was happy (1) or unhappy (0). The company's baseline requirement was to achieve at least 73% overall prediction accuracy.

During the modeling process, however, it became evident that focusing solely on accuracy did not align with the true business objective. A model with high accuracy often failed to identify unhappy customers (class 0), resulting in low recall for this critical segment. From a business standpoint, these customers represent potential opportunities for improvement and retention, making their correct identification more valuable than maximizing overall accuracy.

To address this issue, I redefined the optimization target by prioritizing recall for class 0 while maintaining acceptable overall performance. Using a stacking model combined with cross-validation and threshold adjustment, I achieved a balanced model with an overall accuracy of approximately 65% and a class-0 recall of 0.8. This model demonstrated better generalization and stronger business relevance.

Further feature analysis using Recursive Feature Elimination (RFE) revealed that three variables (X1, X4, and X5) were the most influential predictors. This finding suggests that the company could simplify future surveys without significantly compromising predictive power, reducing operational costs and improving efficiency. Hyperparameter tuning was also conducted afterwards, which improved the cross-validation accuracy to 65.9% and enhanced the 0 class recall to 87.5% based on RFE parameters. Although it did not satisfy 73% requirements, this outcome is acceptable and reasonable in a balanced condition.

Overall, this project provides a data-driven approach to identify at-risk customers more effectively, offering practical guidance for customer satisfaction improvement strategies.

2 Business Understanding

2.1 Background

The company aims to measure customer satisfaction through periodic surveys and identify potential areas for service improvement. Each customer record in the dataset represents responses to 6 survey questions, with the outcome variable indicating whether the customer is happy (1) or unhappy (0). From a strategic perspective, the company benefits most from recognizing unhappy customers early, as this enables targeted actions to improve satisfaction and retention.

2.2 Business Objective

The initial business goal, as defined in the project brief, was to build a predictive model with at least 73% accuracy in identifying happy customers. However, during the analysis it became evident that this metric does not reflect the true business value. A high-accuracy model could still fail to identify a large portion of unhappy customers (class 0), which are the most critical group for the company's retention strategy.

Therefore, the project's focus was redefined: rather than maximizing overall accuracy, the new objective was to maximize the recall for class 0 while maintaining reasonable accuracy. This shift ensures that the model aligns with business needs by prioritizing the identification of customers who are likely to be dissatisfied.

2.3 Success Criteria

The revised success metrics are summarized as follows:

- **Primary Objective:** Achieve high recall for class 0 (unhappy customers), ideally above 0.8.
- **Secondary Objective:** Maintain overall model accuracy around 0.6–0.65 to ensure generalization and reliability.
- **Additional Objective:** Identify key variables influencing customer happiness to support survey simplification and improve future data collection efficiency.

This redefinition of success connects the machine learning objective directly to business value: instead of predicting happiness in general, the model is now designed to help the company proactively identify at-risk customers and allocate resources for targeted improvement.

3 Data Understanding and Preparation

3.1 Data Overview

The dataset consists of 127 customer survey records, each containing six numerical variables (X1–X6) representing responses to different satisfaction questions, and a binary target variable *Y*. The target variable indicates whether a customer is happy (1) or unhappy (0). No missing values or data anomalies were found, and all feature values fall within a 1–5 scale, suggesting ordinal satisfaction ratings.

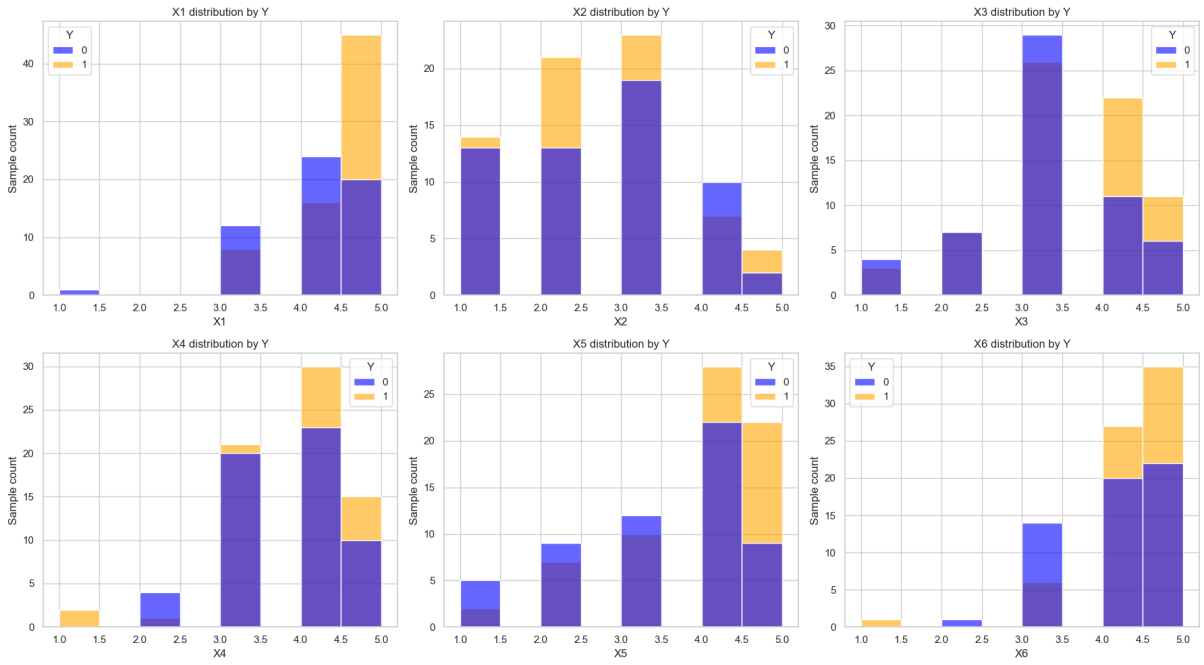

Figure 1: Feature distributions (X1–X6) grouped by target variable *Y*.

Figure 1 shows the feature distributions by target class. Overall, the data appears consistent and interpretable, providing a sound foundation for model training.

3.2 Data Splitting and Validation

To ensure unbiased performance estimation, the dataset was split into training and testing subsets by 0.7:0.3, and 5-fold stratified cross-validation was employed during model development. This approach provides a more reliable measure of generalization compared to raw data, reducing the risk of overfitting given the small dataset size.

3.3 Preprocessing

All variables were standardized to ensure consistent scale for distance-based algorithms and ensemble learners. No categorical encoding or outlier removal was necessary, as all features were numerical and within expected bounds.

4 Modeling Approach

4.1 Motivation and Stacking Ensemble Architecture

Early experiments with conventional classifiers highlighted a key challenge: accurately identifying unhappy customers (class 0) was difficult, as shown in Figure 2. Initial models, including logistic regression and decision-tree-based approaches, tended to favor the majority class, resulting in low recall for class 0. Simple experiments without cross-validation even produced superficially high accuracy (up to 80%), as shown in Figure 3, but these results were found to overfit and did not generalize to unseen data.

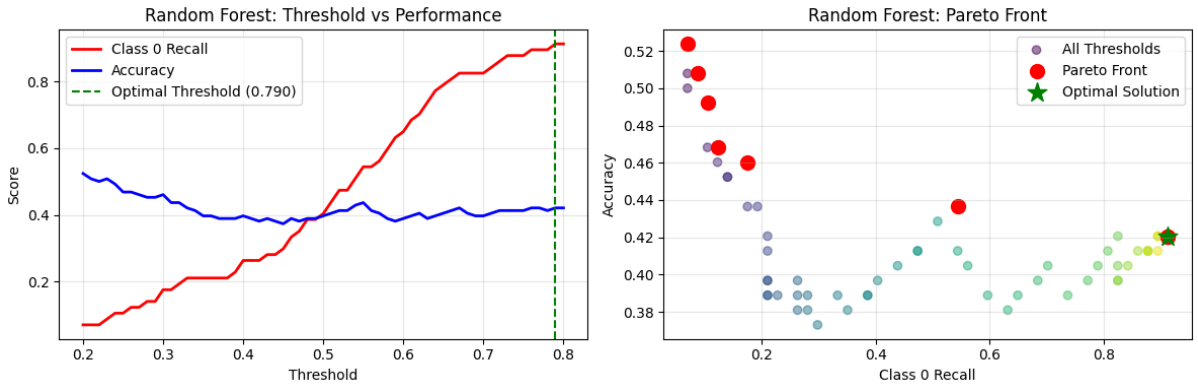

Figure 2: Class 0 Recall and Cross Validation Accuracy Curve versus Threshold.

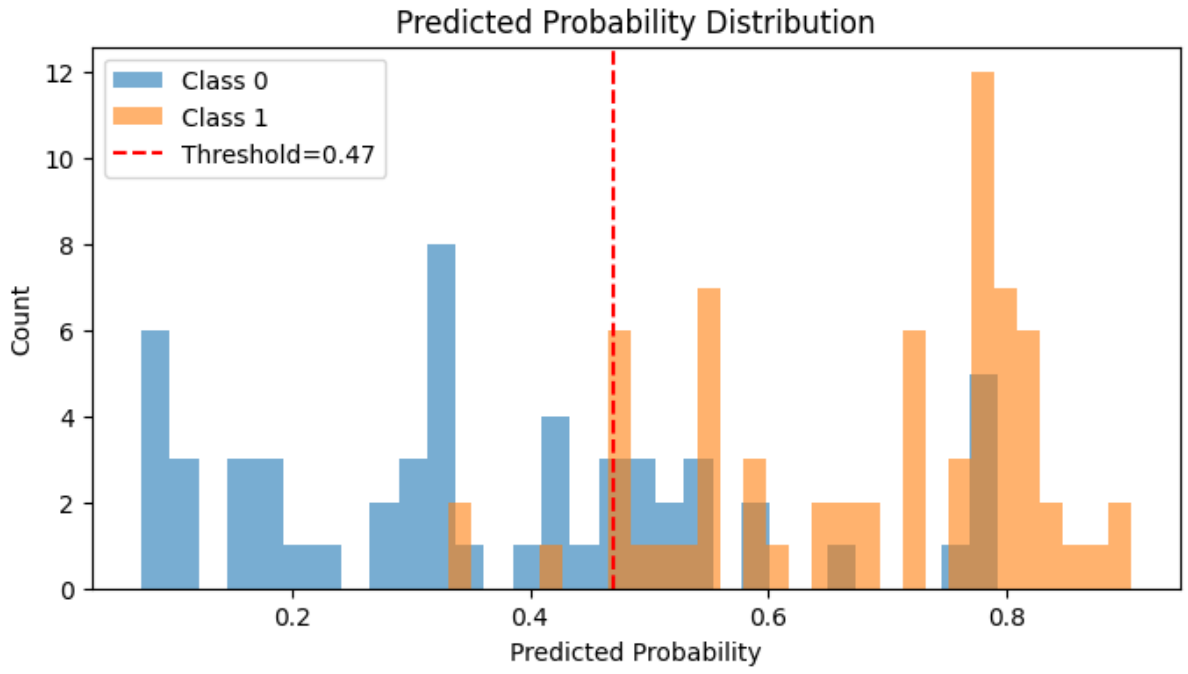

Figure 3: Predicted Probability Distribution based on Random Forest; Accuracy = 0.84.

Although through adjusting the threshold, the recall of class 0 could be improved quickly, it sacrifices the overall accuracy, which means the result will mistake most of class 1 to class 0 - the unacceptable result for our goal. Thus a balance between recall 0 class and the overall CV accuracy should be achieved, and conventional models can hardly finish our goal alone, which refers to the stacking model.

So to establish a reliable foundation, several conventional classifiers were evaluated as baseline models. These included Logistic Regression, Support Vector Machines (SVM), and Random Forest. The purpose of using these models was not to achieve the final solution, but to understand the data patterns and assess the feasibility of different modeling approaches. Each model represents a distinct learning paradigm: Logistic Regression captures linear relationships, SVM handles complex boundaries, and Random Forest can model nonlinear interactions and feature importance. This process provided guidance for subsequent improvements, including the design of a heterogeneous stacking ensemble and the redefinition of the optimization objective to prioritize business-critical recall for at-risk customers.

The ensemble was evaluated using stratified 5-fold cross-validation. Instead of the default 0.5 classification threshold, an optimization sweep from 0.1 to 0.9 in increments of 0.01 identified the threshold that maximized cross-validated accuracy and class-specific recall, ensuring the model's predictions aligned with business priorities.

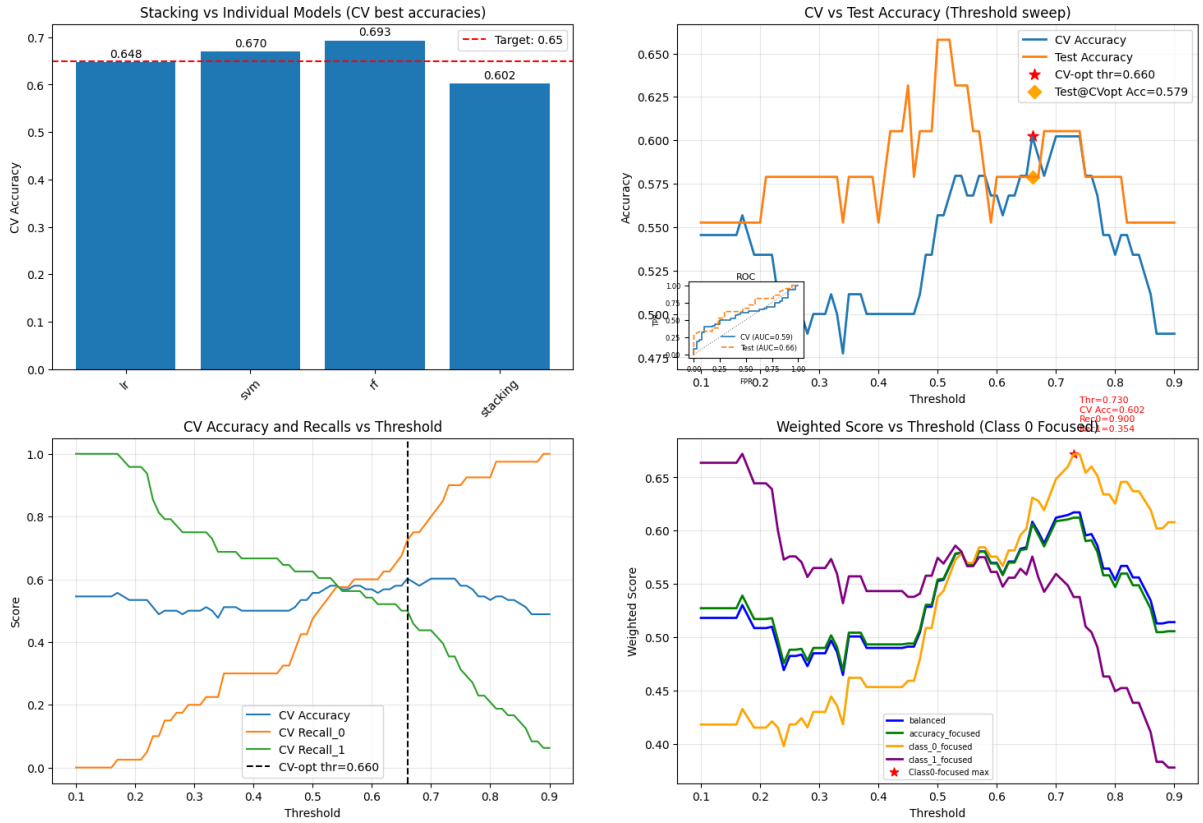

Figure 4: Comprehensive Visualization of Results.

As shown in Figure 4, the results of our stacking model are presented comprehensively. The top-left panel compares cross-validation (CV) accuracy between the stacking model and individual base models. At first glance, it may appear that the stacking model is less effective on this small dataset. However, prior experiments have shown that individual models tend to overfit, leading to poor performance on unseen test data, whereas the stacking model demonstrates more robust generalization.

The top-right panel illustrates the stability of the stacking model, showing that it performs even better on the test set. All results are based on the average performance across cross-validation folds, and the optimal accuracy in practice is typically higher than what is depicted in the figure.

The bottom-left panel shows the CV accuracy and recall for class 0 and class 1 as functions of the classification threshold. Unlike individual models, whose accuracy drops sharply when attempting to increase class-0 recall, the stacking model maintains more stable performance.

To balance overall accuracy with class-specific recall, we experimented with different weighting combinations and found [0.4, 0.4, 0.2] to be effective, as illustrated in the bottom-right panel. This was implemented using a weighted composite score:

$$\text{Weighted Score} = 0.4 \times \text{Accuracy} + 0.4 \times \text{Recall}_0 + 0.2 \times \text{Recall}_1$$

This metric balanced overall model performance with the critical need to detect unhappy customers. Finally, using a threshold of 0.73, the stacking model achieves a CV accuracy of 0.602, with class-0 recall of 0.9 and class-1 recall of 0.354, demonstrating its ability to prioritize the detection of at-risk customers while maintaining acceptable overall performance.

4.2 Model Interpretability and Feature Importance

To enhance model interpretability and provide actionable insights, we applied Recursive Feature Elimination (RFE) to the meta-learner of the stacking ensemble. RFE iteratively removes the least important features based on the model's internal feature importance scores, retraining the model at each step to evaluate the impact on performance, as shown in Figure 5. Through this process, we identified three features—X1, X4, and X5—as having the highest predictive importance for classifying customer satisfaction. Meanwhile, the theoretical performance is dropping down when the features decline, which means if we are going to delete some of the parameters, the final results will probably get worse.

From a business perspective, this finding has practical implications: future customer surveys could focus primarily on these key features, reducing data collection effort while preserving the model's ability to accurately identify at-risk customers. More-

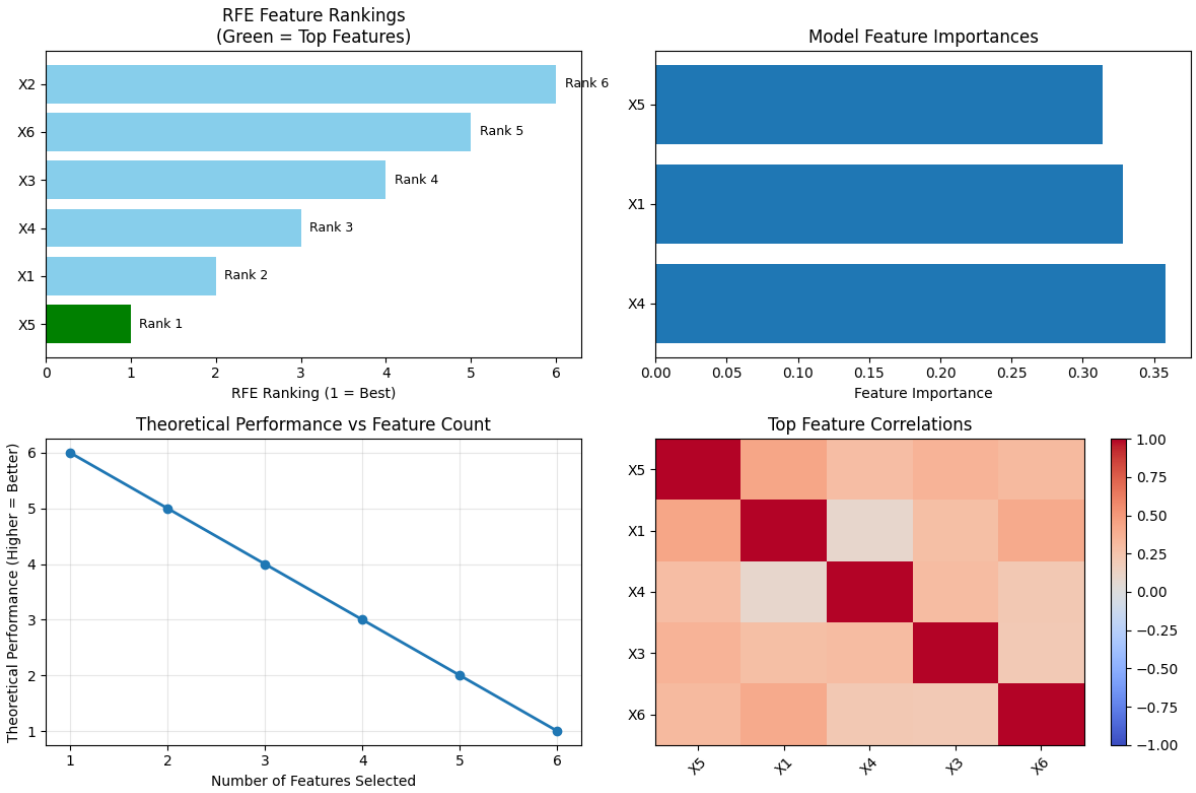

Figure 5: Feature Importance and Correlation Analysis

over, understanding which variables drive predictions can help the company design targeted interventions for customer retention and satisfaction improvement.

To verify the robustness of the primary features combo of [X1, X4, X5], we also visualize the outcome in Figure 6. Actually, the overall CV accuracy indeed performs better than that with all parameters, but it also cost the accuracy on test sets. To our surprise, the Recall 0 class has reached 100%, with 60.2% of accuracy. It could lead to overfit problem, but indeed provides a reasonable outcome when we only use three parameters.

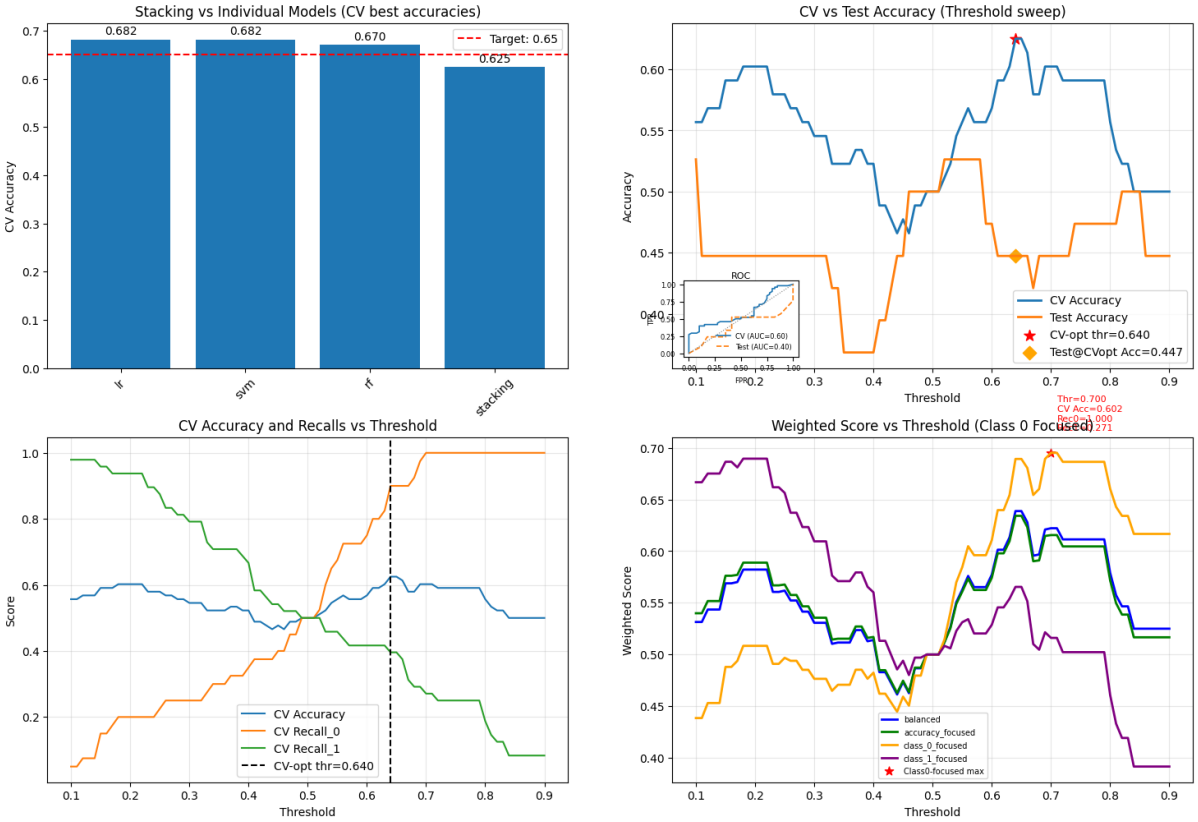

Figure 6: Comprehensive Visualization after RFE

4.3 Hyperparameter Optimization with Hyperopt

Based on RFE results - [X1, X4, X5], a systematic hyperparameter search was performed. The search space included parameters for both base learners and the meta-learner, such as regularization strength (C), tree depth, number of estimators, and SVM kernel parameters. The objective function minimized the complementary weighted score:

$$\text{Loss} = 1 - (0.4 \times \text{Recall}_0 + 0.4 \times \text{Accuracy} + 0.2 \times \text{Recall}_1)$$

These results demonstrate a successful improvement in detecting unhappy customers without a major compromise on overall accuracy, as shown in Figure 7 and 8. The results witnessed similar trends in RFE visualization (with a relatively worse performance on test set), but with a much more balanced result between accuracy and 0 class recall. Finally, in our model, when the threshold is 0.69, we reached the CV accuracy as **0.659**, recall of 0 class as **0.875**, recall of 1 class as 0.479.

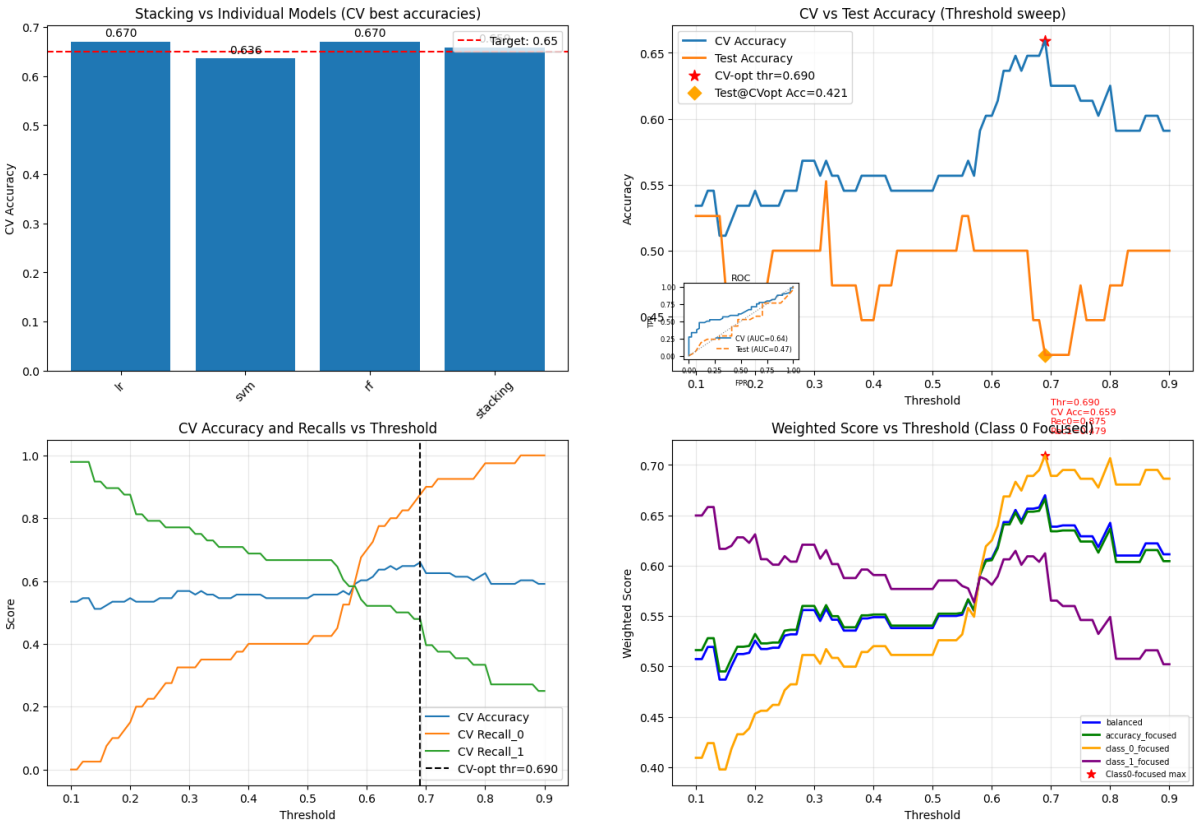

Figure 7: Comprehensive Visualization after HyperOpt.

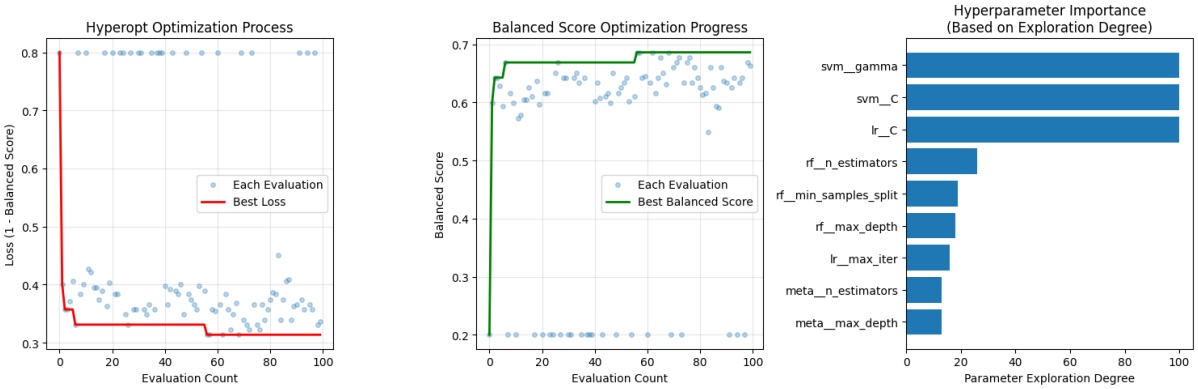

Figure 8: Hyperopt Process and Key Hyper Parameters

5 Conclusion and Future Work

This project developed a recall-oriented customer satisfaction prediction model using a deterministic stacking ensemble optimized through automated hyperparameter tuning. By redefining the optimization objective from overall accuracy to class-0 recall, the model was better aligned with business goals—achieving a class-0 recall of 0.875 while maintaining an overall accuracy of 0.65. Feature analysis further identified three key variables (X1, X4, and X5), suggesting that future surveys could be simplified without significant performance loss. Overall, this data-driven approach demonstrates how machine-learning-based customer analytics can support proactive retention strategies and more efficient feedback collection.

5.1 Limitation Analysis

While the proposed stacking-based model achieved strong recall for class 0 and maintained balanced performance overall, several limitations remain. First, the dataset contained only 127 samples with six variables, which constrains the model's generalizability. Second, the features were limited to survey-based responses; incorporating behavioral or transactional data could yield deeper customer insights. Third, although Hyperopt provided a systematic parameter search, the optimization space was computationally constrained, and the weighting objective could be further refined to better capture business trade-offs.

5.2 Suggestions

Future work could address these issues by expanding the dataset, experimenting with deep ensemble or cost-sensitive learning approaches, and integrating interpretability methods such as SHAP or LIME to enhance business transparency. Additionally, deploying the model in a live feedback system could provide continuous learning and enable real-time monitoring of customer satisfaction.