Executive Summary: Risk Prediction and Actionable Collection Strategy

Machine Learning Pipeline Analysis by Sayandip Bhattacharyya

November 4, 2025

Abstract

This report summarizes the design, methodology, and performance of a machine learning pipeline developed for the financial sector to optimize debt collection efforts. The project addresses the challenge of **class imbalance** by employing under-sampling and identifies customers at high risk of non-repayment. The final stage integrates a **rule-based recommendation engine** with the predictive model to translate risk scores into concrete, cost-efficient contact strategies. The model performance, as demonstrated by the most recent metrics, indicates the pipeline is currently **underperforming** and requires urgent remediation steps detailed in the recommendations section.

1 Business Problem Interpretation

The primary business challenge addressed is the efficient allocation of limited collections resources to maximize recovery while managing customer relationships. The problem is twofold:

P1 **Prediction of Default Risk (Outcome=1):** The core task is to accurately predict the small subset of customers ($\sim$**20%** in the simulated data) who will fail to repay their outstanding loan balance. Since the cost of a **False Negative** (failing to identify a defaulter) is high, the model must be sensitive to the **minority class**.

P2 **Actionable Strategy Generation:** Simply predicting risk is insufficient. The second problem is translating that risk score into an **optimal contact channel** and **strategy**. The goal is to move from a general strategy to a **personalized, risk-adjusted approach** (e.g., using a gentle email for low-risk early-stage delinquency versus an urgent phone call for high-risk, non-responsive accounts).

The project is designed to minimize resource waste on customers who are likely to repay (minimizing **False Positives**) and ensure aggressive action on those who pose the highest financial risk (maximizing **Recall**).

2 Model Methodology and Design

The project followed a robust ML pipeline, starting with a synthetic dataset designed to mimic real-world distributions (e.g., **Income** was normally distributed, **Days_Past_Due** was skewed).

A. Data Preparation and Feature Engineering

Key preparatory steps ensured data quality and complexity:

- **Scaling & Encoding:** All numerical features were scaled using **StandardScaler**. Categorical features (e.g., **Occupation**, **Bank_Code**) were processed using **One-Hot Encoding**.
- **High-Value Features:** Two critical features were engineered:
  1. **Financial Ratios:** Calculated metrics like the **Debt_to_Income_Ratio** for deeper financial health insights.
  2. **High-Risk Flag:** A composite binary feature established by business rules:

$$(\text{Days Past Due} > 30)\ \text{AND}\ (\text{Credit Score} < 500)\ \text{OR}\ (\text{Complaint Flag} = \text{True})$$

B. Imbalance Handling and Classifier Selection

To address the $80\%/20\%$ imbalance in the training data, the following approach was taken:

- **Undersampling Strategy:** The training set was balanced using **RandomUnderSampler** (RUS). The RUS method was chosen to provide the best-case separation by creating a clean, **balanced decision boundary**, which is highly beneficial for distance-based classifiers.
- **Model Comparison:** Multiple models (Logistic Regression, GBC, XGBoost, and KNN) were benchmarked. The **KNeighborsClassifier (KNN)** was retrained on the **undersampled data**, demonstrating improved performance for the minority class.

C. Recommendation Engine Architecture

The prediction model is integrated with a final, rule-based layer to produce actionable strategies:

1. **Input:** Takes the model's **Predicted Outcome** (0 or 1) and raw customer data (**Complaint_Flag**, **Days_Past_Due**, **Last_Contact_Channel**).
2. **Prioritization Logic:** Uses a cascading set of **IF-THEN rules**, prioritizing customer sensitivity first (e.g., **Complaint Flag** $\rightarrow$ Mail/Email) before escalating based on risk and non-responsiveness.
3. **Output:** Generates a specific **Recommended Channel** (e.g., Phone, SMS) and an associated **Strategy** (e.g., 'Urgent, direct conversation').

3 Results and Key Insights

Model performance was evaluated on the unseen test set, with a focus on metrics relevant to collections optimization.

Table 1: Top Model Performance Comparison (Actual Results)

Model	ROC-AUC	F1-Score	Precision	Recall
Logistic Regression	0.4329	0.2584	0.1870	0.4182
GradientBoostingClassifier	0.5262	0.1370	0.2778	0.0909
XGBoostClassifier	0.5677	0.2195	0.3333	0.1636
KNeighborsClassifier	**0.5990**	**0.2716**	**0.4231**	**0.2000**

A. Performance Overview (Actual Results)

The benchmarked models achieved the following performance metrics on the test set.

- **F1-Score Focus:** The **KNeighborsClassifier** model, despite a low overall performance, was selected as the **best-performing operational model** based on its highest F1-Score (**0.2716**). This performance suggests that the current pipeline design is significantly **underperforming** and requires urgent revision.
- **Strategy Implication:** The low **Recall (0.2000)** is a major concern. It indicates the model is only identifying **20%** of the true defaulters (False Negatives are high), failing the core business requirement of protecting against **high-risk losses**. The relatively higher **Precision (0.4231)** suggests that while the model is accurate when it does predict default, it is too conservative and misses most defaulters, which is currently an unacceptable operational risk.

B. SHAP Explainability Insights

The SHAP analysis revealed the **most impactful features** driving the "Not Repay" prediction (Outcome=1):

1. **Days_Past_Due:** As expected, this was the **most influential feature** globally, with higher scaled values strongly correlating with a higher risk of non-repayment.
2. **Credit_Score:** Low credit scores were a major driver for the positive outcome prediction.
3. **Engineered Features:** The engineered **Debt_to_Income_Ratio** and the **High_Risk_Flag** also ranked highly, confirming their business utility in combining multiple signals into a cohesive **risk factor**.

4 Future Scope and Recommendations

To enhance the pipeline's robustness and efficiency, the following steps are recommended:

F1 **Advanced Sampling Methods:** Replace **Random Under-Sampling** with more sophisticated techniques like **SMOTE** (Synthetic Minority Over-sampling Technique) or **Tomek Links**. This would help retain more information from the majority class while still addressing the imbalance, potentially improving Precision without sacrificing Recall.

F2 **Hyperparameter Optimization:** Conduct comprehensive **Hyperparameter Tuning** for the selected KNN model (e.g., optimizing the number of neighbors, distance metric, and weights) to maximize both **ROC-AUC** and **F1-Score** simultaneously.

- **F3** **Feature Engineering Audit:** Given the weak performance, a full audit of the existing features, particularly the **High-Risk Flag**, must be performed. New features capturing **recent payment history variance** or **contact response rates** should be introduced to provide stronger signals to the model.
- **F4** **Recommendation Engine Evolution:** The current engine is rule-based. The long-term scope should involve integrating a **Deep Reinforcement Learning (DRL)** approach to dynamically learn the **optimal contact sequence and timing** that maximizes recovery based on observed customer responses, moving beyond **static IF-THEN rules**.
- **F5** **External Data Validation:** Validate model distributions and performance using real, anonymized historical transaction and collection data to ensure the assumptions made during the synthetic data generation phase hold true in a **live environment**.