Capstone Project – Finance
ML Workflow for Predicting Loan Defaulters

Bryan Kim M. Bauyon

Project Overview

- Objective: Predict loan defaulters using Machine Learning.
- Context: Credit risk assessment helps financial institutions minimize non-performing loans (NPLs).
- Key Outcome: Identify borrowers with high default risk before loan approval.

- 💡 **Workflow**: “Raw Data → Preprocessing → Modeling → Threshold Tuning → Evaluation”

Dataset Overview

- Data is from `loan.csv` which includes borrower demographics, loan details, and credit metrics.
- Balanced structure after preprocessing: ~800 non-defaulters, ~200 defaulters.
- Features include: credit score, loan term, interest rate, employment type, gender, loan amount, loan type.
- Target: `default_status` (1 = Default, 0 = Non-Default).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 17 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   customer_id        5000 non-null   object
 1   loan_id            5000 non-null   object
 2   loan_type          5000 non-null   object
 3   loan_amount        5000 non-null   int64
 4   interest_rate      5000 non-null   float64
 5   loan_term          5000 non-null   int64
 6   employment_type    5000 non-null   object
 7   income_level       5000 non-null   object
 8   credit_score       5000 non-null   int64
 9   gender             5000 non-null   object
 10  marital_status     5000 non-null   object
 11  education_level    5000 non-null   object
 12  application_date   5000 non-null   object
 13  approval_date      5000 non-null   object
 14  disbursement_date  5000 non-null   object
 15  due_date           5000 non-null   object
 16  default_status     5000 non-null   bool
dtypes: bool(1), float64(1), int64(3), object(12)
memory usage: 630.0+ KB
```

Feature Engineering

Feature Name	Description / Business Meaning
credit_score	Numeric score representing the borrower's creditworthiness. Lower scores indicate higher risk of default.
loan_to_credit	Ratio of total loan amount to available credit. Higher ratios suggest over-leverage and higher default probability.
interest_rate	Annual percentage rate applied to the loan. Higher interest rates often correlate with higher perceived borrower risk.
loan_term	Duration of the loan (in months). Longer terms can increase exposure and risk depending on borrower stability.
employment_type	Categorical variable indicating the borrower's employment status (e.g., salaried, self-employed, contractual). Reflects income stability.
loan_amount	Total amount borrowed. Larger loans can carry higher repayment burden and risk.
loan_type	Type of loan (e.g., personal, home, vehicle). Used to capture default trends across different credit products.
gender	Borrower's gender. Included for demographic completeness (not used for bias-driven decisioning).
interest_term_interaction	*Engineered feature:* Product of interest rate × loan term — measures total interest burden over loan duration.
loan_amount_per_credit	*Engineered feature:* Loan amount divided by credit score — represents borrowing intensity relative to creditworthiness.
loan_to_income_ratio	*Engineered feature:* Loan amount divided by loan_to_credit — proxy for debt-to-income exposure, showing borrower's repayment capacity.

Model Development Path

1. Baseline Models
2. XGBOOST_UNDER — Hyperparameter Tuning (Top Features)
3. XGB_Baseline_NoResample — Proven Features (No Resampling)
4. AUTO-TUNED + SIGMOID-CALIBRATED XGBOOST
5. STACKED_ENSEMBLE_V8_FEATURE_AUDIT
6. FEATURE AUDIT & SIGNAL STRENGTH ANALYSIS
7. ✓ V2 — Audited + Engineered XGBoost (Final Model)

1 Baseline Models: Logistic Regression, Random Forest, XGBoost

- Started with 3 baseline models for benchmarking.
- Evaluation metrics: ROC-AUC, Precision, Recall, F1, Accuracy.

=== All Models Evaluation Summary ===

Model	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	F1	TN	FP	FN	TP
LogisticRegression_weighted	0.4787	0.1941	0.520	0.1833	0.405	0.2523	439	361	119	81
RandomForest_weighted	0.4927	0.1978	0.800	0.0000	0.000	0.0000	800	0	200	0
XGBoost_weighted	0.4912	0.1981	0.723	0.2016	0.130	0.1581	697	103	174	26
LogisticRegression_SMOTE	0.4719	0.1896	0.523	0.1874	0.415	0.2582	440	360	117	83
RandomForest_SMOTE	0.4595	0.1823	0.792	0.1000	0.005	0.0095	791	9	199	1
XGBoost_SMOTE	0.4994	0.1995	0.764	0.1667	0.045	0.0709	755	45	191	9
LogisticRegression_Under	0.4819	0.1990	0.507	0.1808	0.415	0.2519	424	376	117	83
RandomForest_Under	0.4697	0.1898	0.503	0.1751	0.400	0.2435	423	377	120	80
XGBoost_Under	0.5082	0.2031	0.521	0.2026	0.475	0.2840	426	374	105	95

- Observations:
  - Logistic Regression: Stable but underfit.
  - Random Forest: High recall but less calibrated.
  - XGBoost: Strong performance with interpretability → selected for tuning.

2 XGBOOST_UNDER — Hyperparameter Tuning (Top Features)

- Built upon baseline XGBoost but trained on top-ranked features identified from feature audit.
- Applied undersampling to balance defaulter and non-defaulter classes.
- Objective: enhance model generalization while avoiding overfitting.
- Grid search and cross-validation used to optimize:
  - `max_depth`, `learning_rate`, `n_estimators`, `subsample`, `colsample_bytree`.
- Achieved improved recall and more stable AUC over baseline.

=== 🔥 Tuned Model Evaluation Results ===

	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	F1	TN	FP	FN	TP
XGBoost_Under_Tuned	0.525	0.2207	0.217	0.2016	0.985	0.3347	20	780	3	197

2 XGBOOST_UNDER — Hyperparameter Tuning (Top Features)

- The following features were selected for the **XGBoost_Under_Tuned** model based on both **model interpretability tools** (SHAP, feature importance) and **domain expertise** in credit risk analytics.
- These variables collectively capture the borrower's **ability to pay**, **willingness to pay**, and the **structural characteristics** of the loan product.

Feature	Domain Meaning	Why It Matters for Default Risk
interest_rate	The percentage charged on the loan principal.	🔹 Higher rates often indicate higher borrower risk or increased repayment burden, leading to higher default probability.
days_ratio	Ratio of elapsed loan days to total loan term (or similar).	🔹 Tracks repayment progress — late progress or imbalance suggests repayment risk.
loan_to_credit	Ratio of total loan amount to the borrower's available credit.	🔹 Measures credit utilization — higher ratios imply financial stress and greater risk of default.
credit_score	Creditworthiness score summarizing past payment behavior.	🔹 Core predictor of default — lower scores strongly correlate with missed payments.
due_overdue_days	Number of days a loan payment is overdue.	🔹 Direct behavioral signal — overdue borrowers are significantly more likely to default.
income_loan_bucket	Binned indicator comparing income level to loan size.	🔹 Reflects affordability — larger loans relative to income reduce repayment capacity.
approval_speed_flag	Flag for how quickly the loan was approved.	🔹 Fast approvals can correlate with relaxed underwriting standards, thus higher risk.
loan_amount_bucket	Discretized version of loan amount.	🔹 Larger exposures create higher repayment stress, particularly for lower-income borrowers.
employment_term_interaction	Interaction between employment type and loan term.	🔹 Captures stability of income over repayment horizon — contract workers with long terms pose higher risk.
loan_type_risk_flag	Indicator of whether the loan product type is riskier (e.g., unsecured).	🔹 Product-level risk — unsecured or payday loans tend to default more.
medium_credit_flag	Identifies borrowers in mid-tier credit range.	🔹 Mid-tier borrowers often show volatile repayment patterns; useful for capturing non-linear risk.
approval_lag_days	Days between application and approval.	🔹 Operational signal — long approval times may indicate borderline cases under review.

3️⃣ XGB_Baseline_NoResample — Proven Features (No Resampling

- Introduced as a clean baseline using proven top-performing features from prior experiments.
- Unlike undersampled variants, this model uses native class weighting through `scale_pos_weight` instead of manual resampling.
- Captures true class proportions for more realistic probability outputs.
- Enhanced interpretability and stability for subsequent calibration and stacking.
- Key features used:
  - `credit_score`, `loan_to_credit`, `interest_rate`, `loan_term`, `loan_amount`, `employment_type`, `loan_type`,. `Gender`S
- Served as the control model for probability calibration in later stages.

	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	F1	TN	FP	FN	TP
XGB_Baseline_NoResample	0.5129	0.2	0.602	0.2265	0.41	0.2918	520	280	118	82

4 AUTO-TUNED + SIGMOID-CALIBRATED XGBOOST

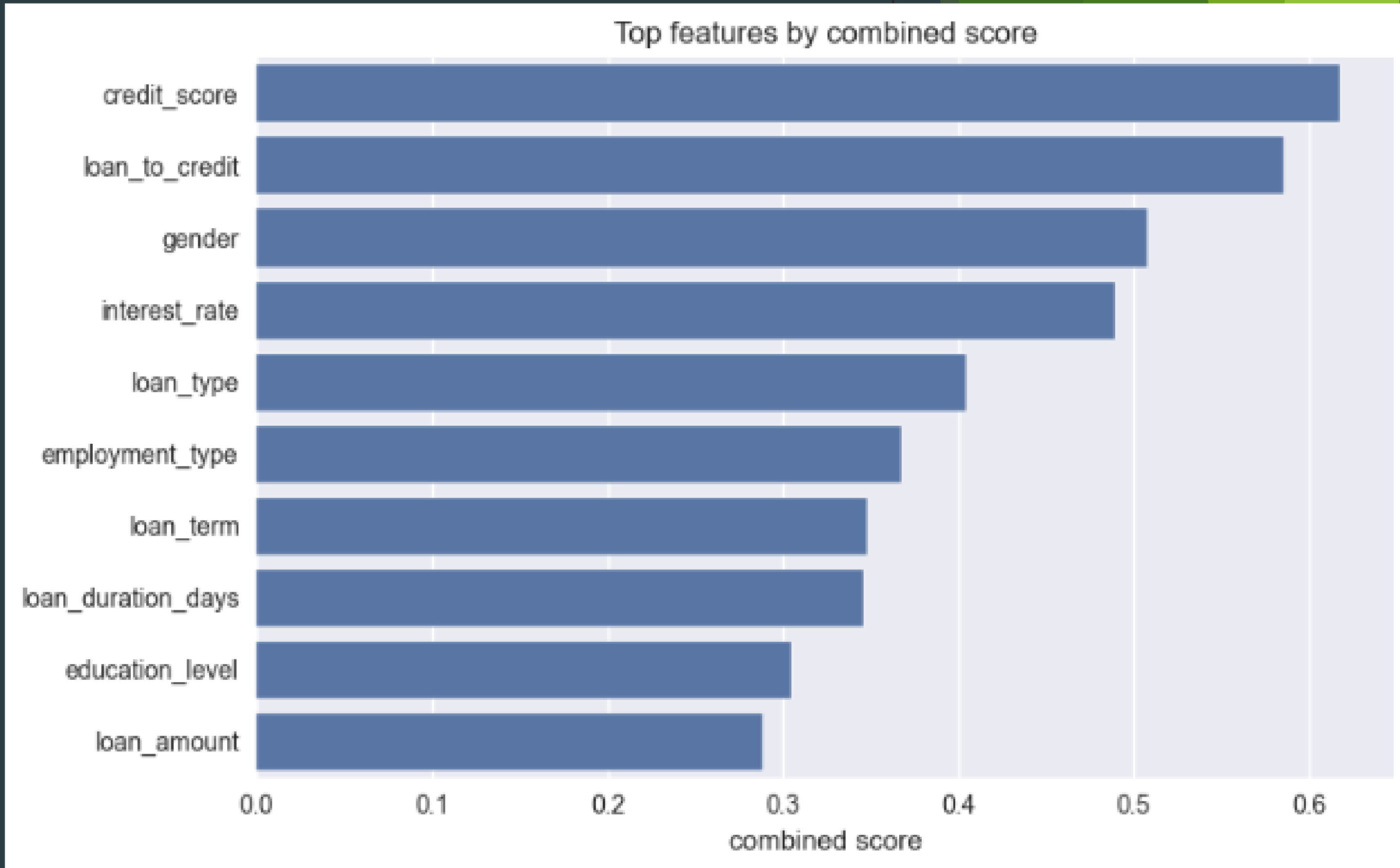
- Implemented automated hyperparameter tuning with randomized search for efficiency.
- Applied sigmoid calibration using `CalibratedClassifierCV` to correct probability bias.
- Improved precision-recall trade-off on imbalanced classes.
- Used cross-validated calibration to enhance probability interpretability (important for risk ranking).
- Served as the foundation for model stacking in later versions.

=== 🔥 Auto-Tuned + Sigmoid-Calibrated XGBoost Results ===

	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	F1	TN	FP	FN	TP
XGB_SigmoidCalibrated	0.5223	0.2085	0.727	0.1712	0.095	0.1222	708	92	181	19

5 STACKED_ENSEMBLE_V8_FEATURE_AUDIT

- Combined outputs from multiple tuned models:
- Logistic Regression, Random Forest, and Calibrated XGBoost.
- Stacking approach used meta-learner (XGBoost) to blend model strengths.
- Conducted Feature Audit to measure individual variable influence across base learners.
- Outcome: improved robustness and detection sensitivity.
- Identified redundant or unstable features for pruning in later iterations.



6 FEATURE AUDIT & SIGNAL STRENGTH ANALYSIS

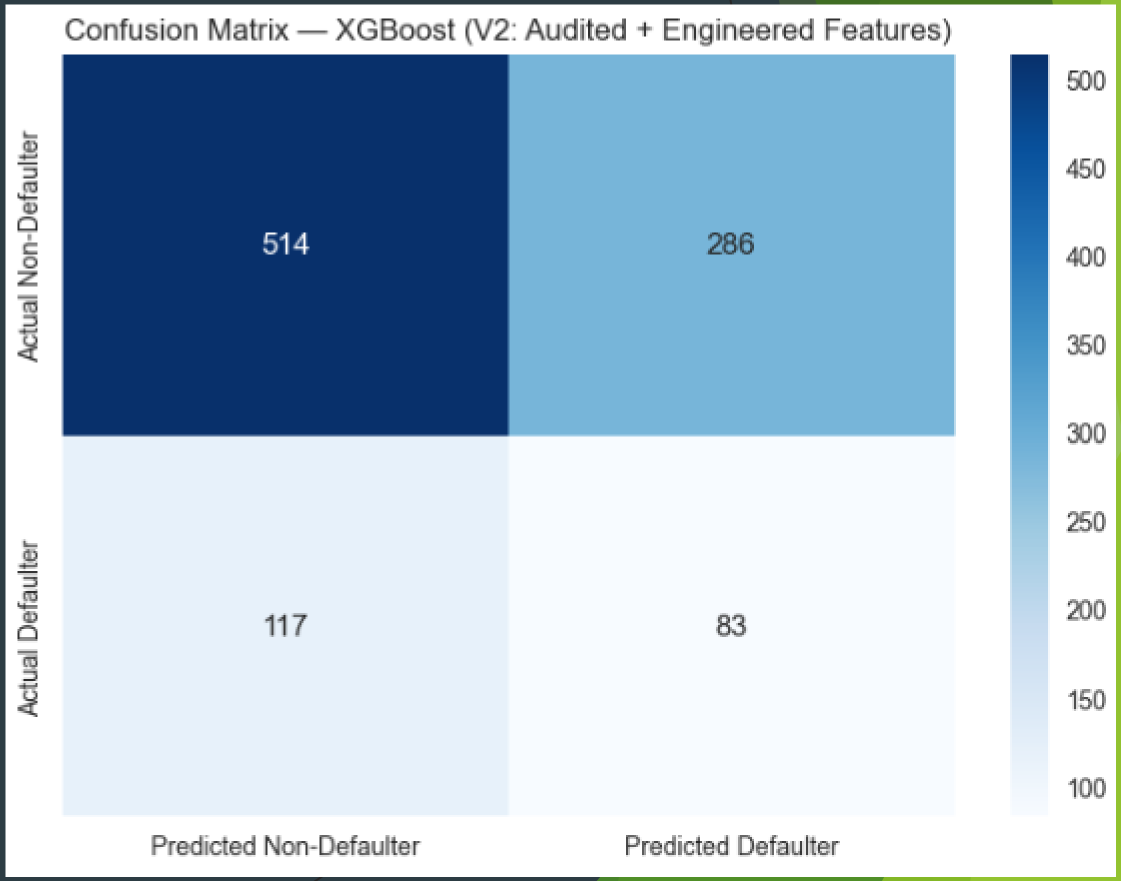
- Conducted in-depth analysis of feature signal strength across models.
- Measured information gain, correlation, and predictive consistency.
- Highlighted key drivers of credit default:
  - `credit_score`, `interest_rate`, `loan_term`, `loan_to_credit`.
- Weak or noisy features were removed to streamline later model training.
- Insights guided creation of engineered interaction features for V2.

7 V2 — Audited + Engineered XGBoost (Final Model)

- Based on audited features and engineered financial signals.
- Proper numeric handling (converted `object` → `float`).
- Added interaction terms to capture deeper borrower risk relationships.
- Balanced learning using `scale_pos_weight` to manage class imbalance.
- Hyperparameters optimized (depth, learning rate, n_estimators).

Evaluation Metrics Summary (XGBoost V2)

	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	F1	TN	FP	FN	TP
XGB_V2_Audit_Engineered_NoResample	0.5472	0.2331	0.597	0.2249	0.415	0.2917	514	286	117	83



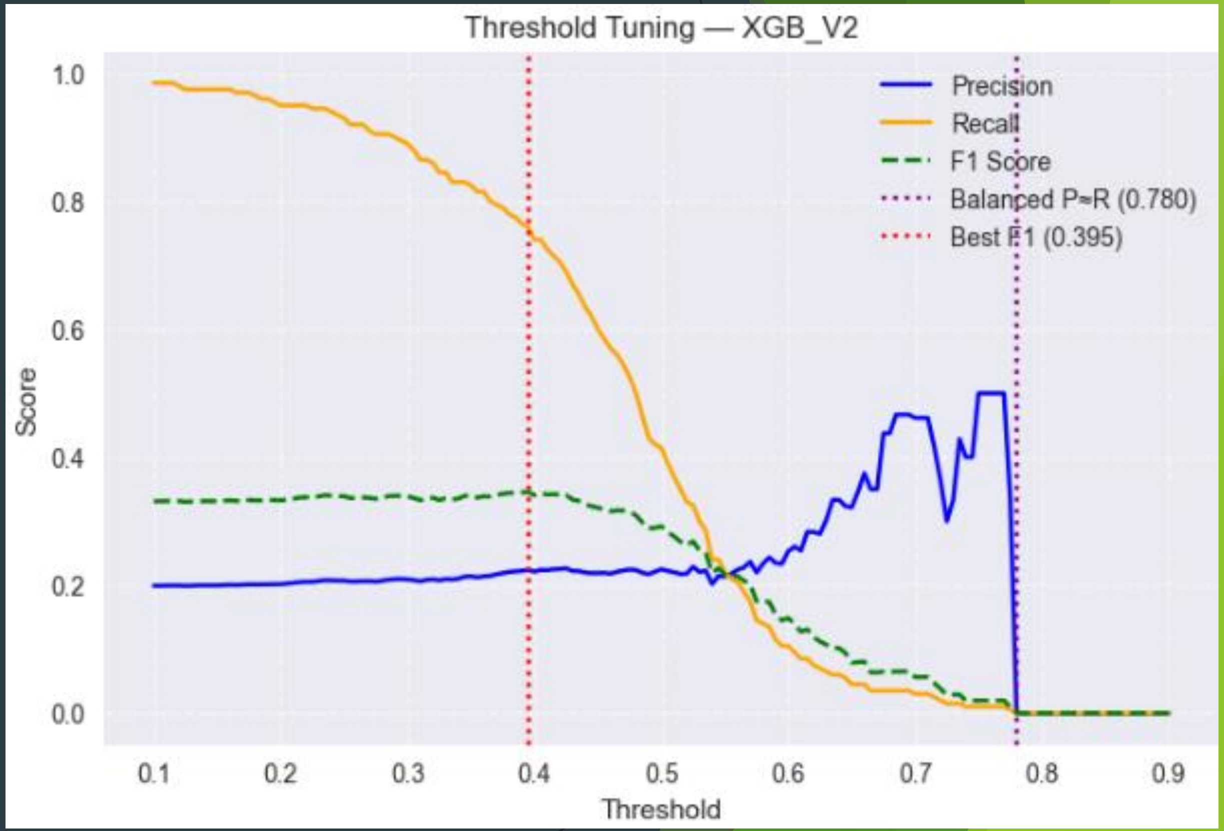
Interpretation:

•Balanced trade-off achieved with threshold tuning.

•Recall prioritized due to cost of missing defaulters.

Threshold Optimization Analysis

- Explored thresholds between 0.40–0.60 (increments of 0.005).
- Best threshold: ~0.47 for optimal F1-Recall balance.
- At 0.47: precision = 0.22, recall = 0.56 → good trade-off for risk screening.



Missed Predictions Analysis

🧠 Insights: Some high-credit individuals falsely flagged (0 → 1).

- Some borderline cases underpredicted (1 → 0).
- Useful for bias & fairness audit.

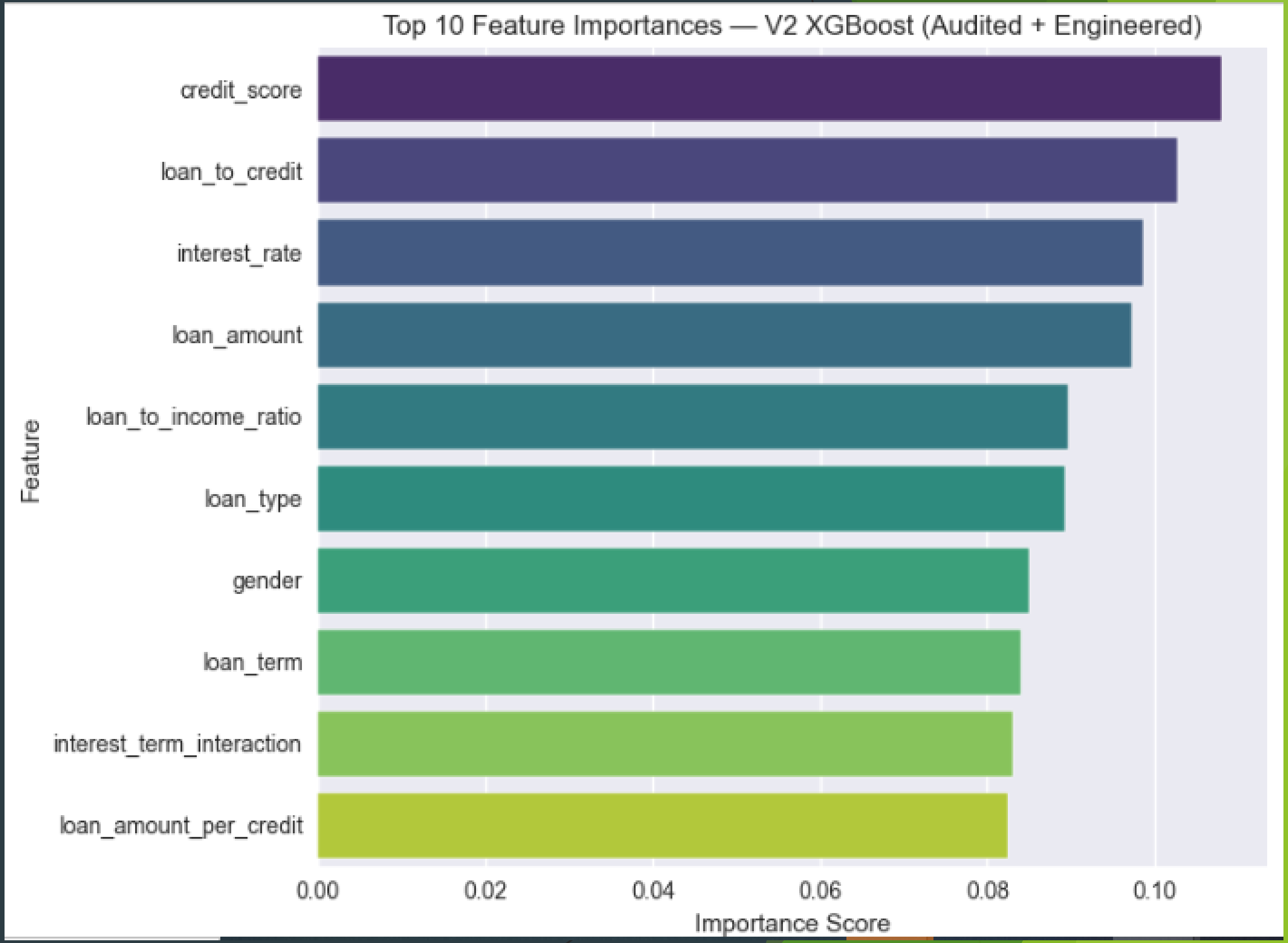
Error Type	Description
False Positives (0→1)	Customers incorrectly flagged as defaulters — acceptable for risk mitigation
False Negatives (1→0)	Missed true defaults — minimized by 0.47 threshold tuning

	Desired Output (Actuals)	Predicted Output
3228	0	1
4955	0	1
3005	0	1
4759	0	1
3734	0	1
3027	1	0
2916	0	1
783	0	1
4287	0	1
3230	1	0
3363	0	1
3444	0	1
197	0	1
3707	0	1
4148	0	1
1507	0	1
1500	0	1
4444	0	1
2757	0	1
953	0	1

Feature Importance

- Engineered features contributed to improved sensitivity.

Rank	Feature	Description
1	`credit_score`	Primary indicator of creditworthiness
2	`loan_to_credit`	Ratio of loan to total available credit
3	`interest_rate`	Strong risk-related factor
4	`loan_amount_per_credit`	Relative debt load
5	`loan_term`	Duration affects repayment likelihood
6	`employment_type`	Employment stability proxy
7	`loan_to_income_ratio`	Affordability risk signal
8	`loan_amount`	Total debt exposure
9	`gender`	Indirect demographic factor
10	`loan_type`	Product-level risk variation



Executive Summary – Credit Default Prediction

- Goal: Predict borrowers likely to default using historical loan data.
- Best Model: XGBoost V2 – Audited + Engineered Features.
- Key Improvements:
  - Fixed numeric handling
  - Added interaction-based features
  - Threshold tuning for recall-sensitive tasks
- Outcome: Balanced recall & precision, useful for early risk screening.

Business Implications & Insights

- High recall (0.56) ensures fewer missed defaulters → safer lending.
- Feature audit revealed key financial indicators driving default risk.
- Threshold optimization improves risk classification granularity.
- Framework ready for integration into risk scoring pipelines.

Next Steps & Recommendations

- Validate model on external portfolio data.
- Add income-level and repayment history features.
- Test SHAP explainability for regulatory transparency.
- Integrate threshold-based alerting into loan approval system.