Loan Approval Prediction : Project Checkpoint 2

Team: Predictive Minds

Members: Suraj R. Rao | Om G. Patel | Darren A. Ferrao | Mohammad A. K. Dalwai

Objective: Expand and refine the loan-approval prediction pipeline through multiple models, feature engineering, class-balancing, and model tuning to achieve higher predictive accuracy and recall.

Checkpoint 1 Summary

Dataset

58 645 samples × 13 features.

Models

Logistic Regression | Decision Tree | Random Forest.

Best Model (CP1)

Random Forest → Accuracy 0.95 | F1 0.78.

Next Goal

Improve recall for *Default* class and add richer features.

Class Distribution & Balancing

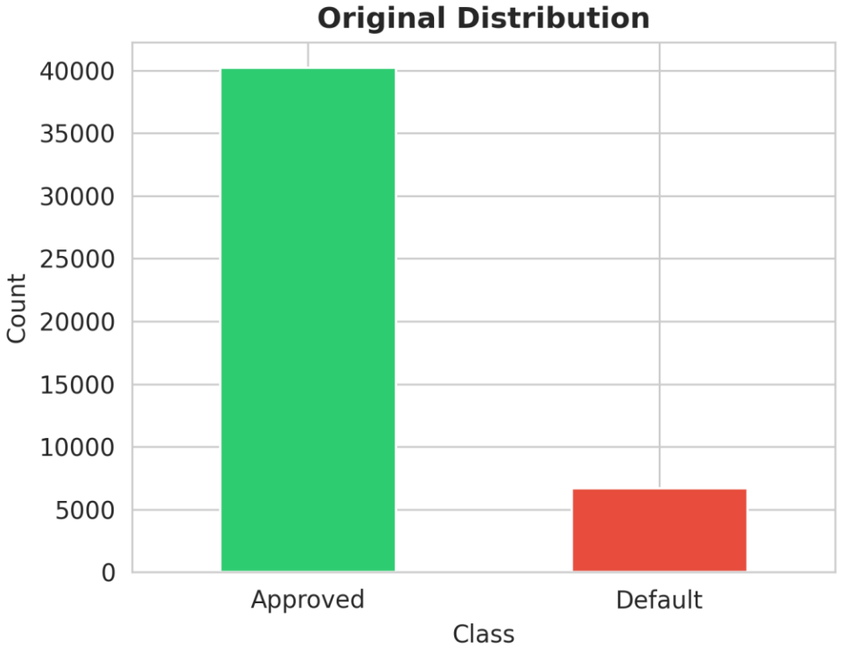
Original Distribution

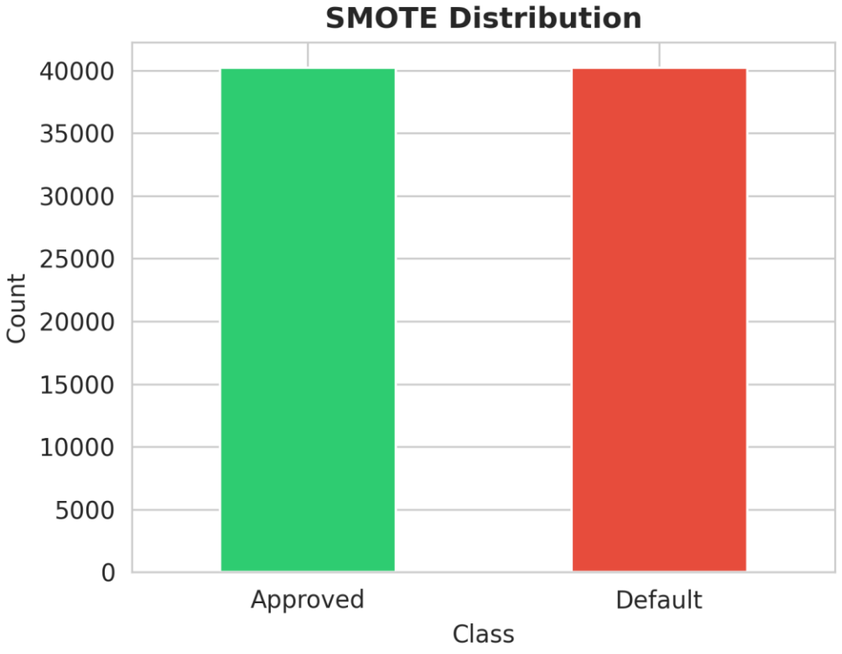
- Original ratio: Approved 50 295 | Default 8 350 (≈ 6 : 1).

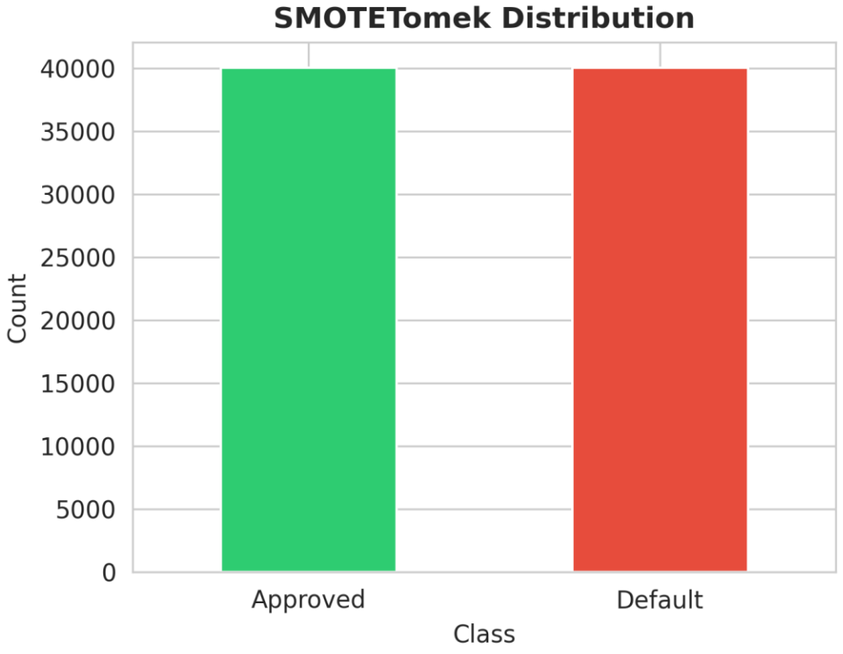
- After SMOTE: 80 472 samples (balanced).
- After SMOTETomek: 80 134 samples (balanced & clean).

Impact

Balancing substantially improved minority-class recall and F1.







Feature Engineering & Data Preparation

Engineered Features

Added 12 engineered features → 25 total variables.

Examples: income_to_loan_ratio, debt_to_income_ratio, age_employment_ratio, log & polynomial terms.

Data Processing

Label encoding + standard scaling for numerical features.

Train/Test Split

46 916 train | 11 729 test (80/20).

Models Evaluated (10 Algorithms)

Algorithms Tested

Logistic Regression | Decision Tree | Random Forest | XGBoost | LightGBM | Gradient Boost | AdaBoost | SVM | KNN | Naive Bayes.

Datasets Used

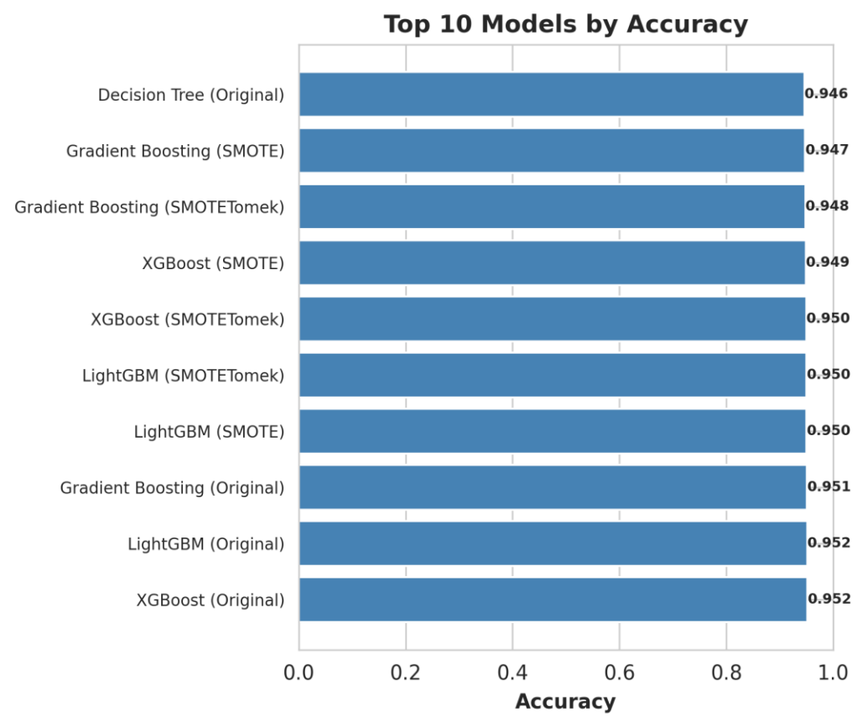
Original, SMOTE, SMOTETomek.

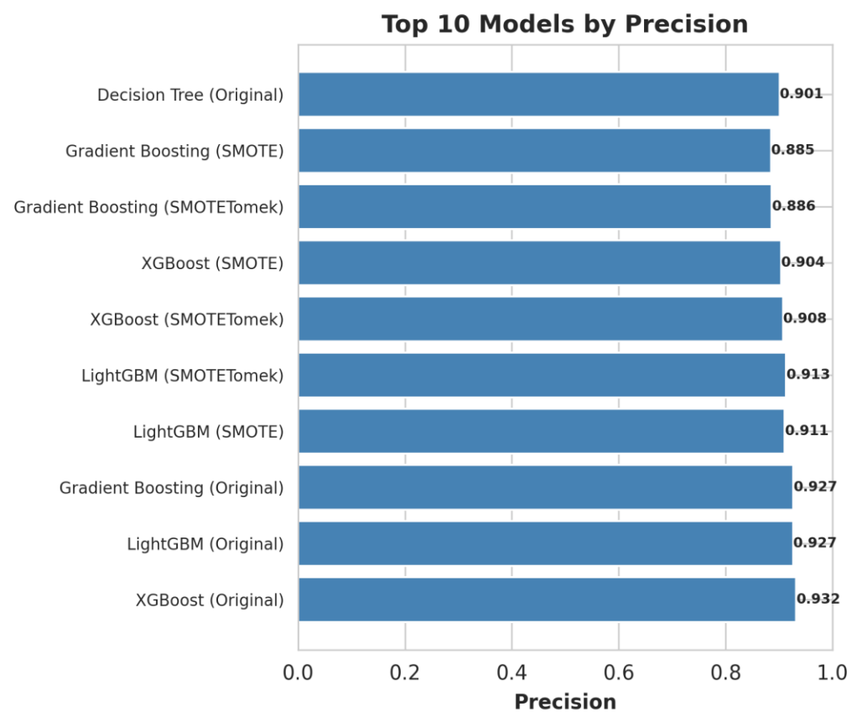
Evaluation Metrics

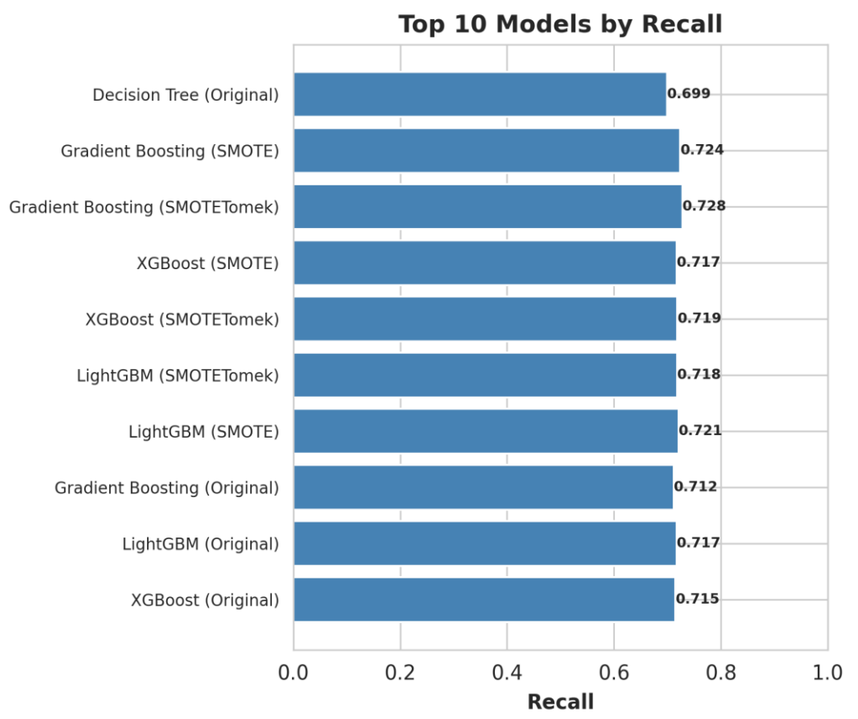
Accuracy | Precision | Recall | F1 | ROC-AUC.

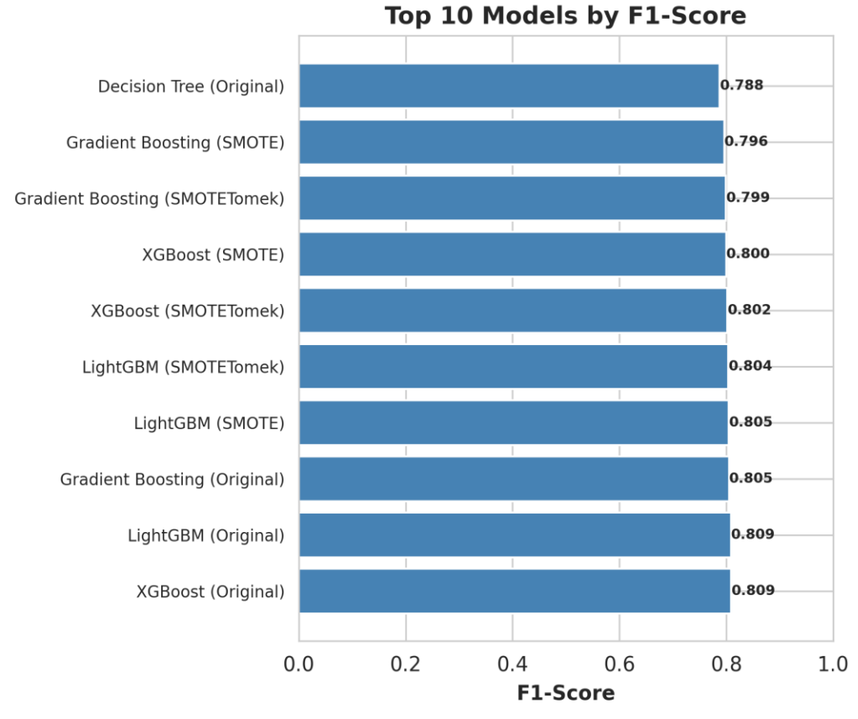
Best Performers

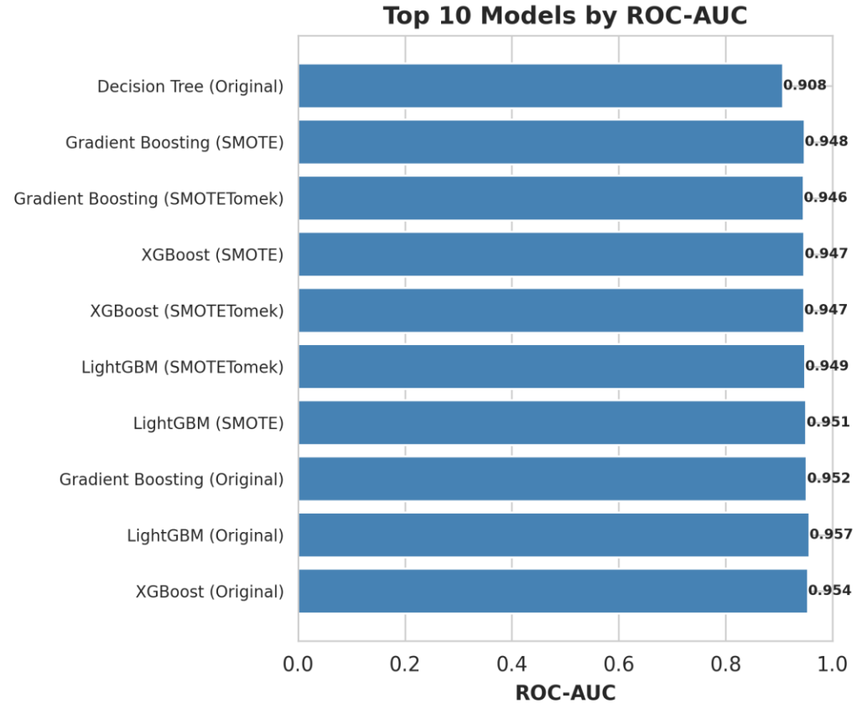
Ensemble models (XGBoost, LightGBM, Gradient Boost) performed best.


Top 10 Models by Accuracy
Decision Tree (Original)
Gradient Boosting (SMOTE)
Gradient Boosting (SMOTETomek)
XGBoost (SMOTE)
XGBoost (SMOTETomek)
LightGBM (SMOTETomek)
LightGBM (SMOTE)
Gradient Boosting (Original)
LightGBM (Original)
XGBoost (Original)
0.946
0.947
0.948
0.949
0.950
0.950
0.950
0.951
0.952
0.952
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy


Top 10 Models by Precision
Decision Tree (Original)
Gradient Boosting (SMOTE)
Gradient Boosting (SMOTETomek)
XGBoost (SMOTE)
XGBoost (SMOTETomek)
LightGBM (SMOTETomek)
LightGBM (SMOTE)
Gradient Boosting (Original)
LightGBM (Original)
XGBoost (Original)
0.901
0.885
0.886
0.904
0.908
0.913
0.911
0.927
0.927
0.932
0.0
0.2
0.4
0.6
0.8
1.0
Precision


Top 10 Models by Recall
Decision Tree (Original)
Gradient Boosting (SMOTE)
Gradient Boosting (SMOTETomek)
XGBoost (SMOTE)
XGBoost (SMOTETomek)
LightGBM (SMOTETomek)
LightGBM (SMOTE)
Gradient Boosting (Original)
LightGBM (Original)
XGBoost (Original)
0.699
0.724
0.728
0.717
0.719
0.718
0.721
0.712
0.717
0.715
0.0
0.2
0.4
0.6
0.8
1.0
Recall


Top 10 Models by F1-Score
Decision Tree (Original)
Gradient Boosting (SMOTE)
Gradient Boosting (SMOTETomek)
XGBoost (SMOTE)
XGBoost (SMOTETomek)
LightGBM (SMOTETomek)
LightGBM (SMOTE)
Gradient Boosting (Original)
LightGBM (Original)
XGBoost (Original)
0.788
0.796
0.799
0.800
0.802
0.804
0.805
0.805
0.809
0.809
0.0
0.2
0.4
0.6
0.8
1.0
F1-Score


Top 10 Models by ROC-AUC
Decision Tree (Original)
Gradient Boosting (SMOTE)
Gradient Boosting (SMOTETomek)
XGBoost (SMOTE)
XGBoost (SMOTETomek)
LightGBM (SMOTETomek)
LightGBM (SMOTE)
Gradient Boosting (Original)
LightGBM (Original)
XGBoost (Original)
0.908
0.948
0.946
0.947
0.947
0.949
0.951
0.952
0.957
0.954
0.0
0.2
0.4
0.6
0.8
1.0
ROC-AUC

Dataset-Level Performance Comparison

Compared Original, SMOTE, and SMOTETomek datasets for top models.

1

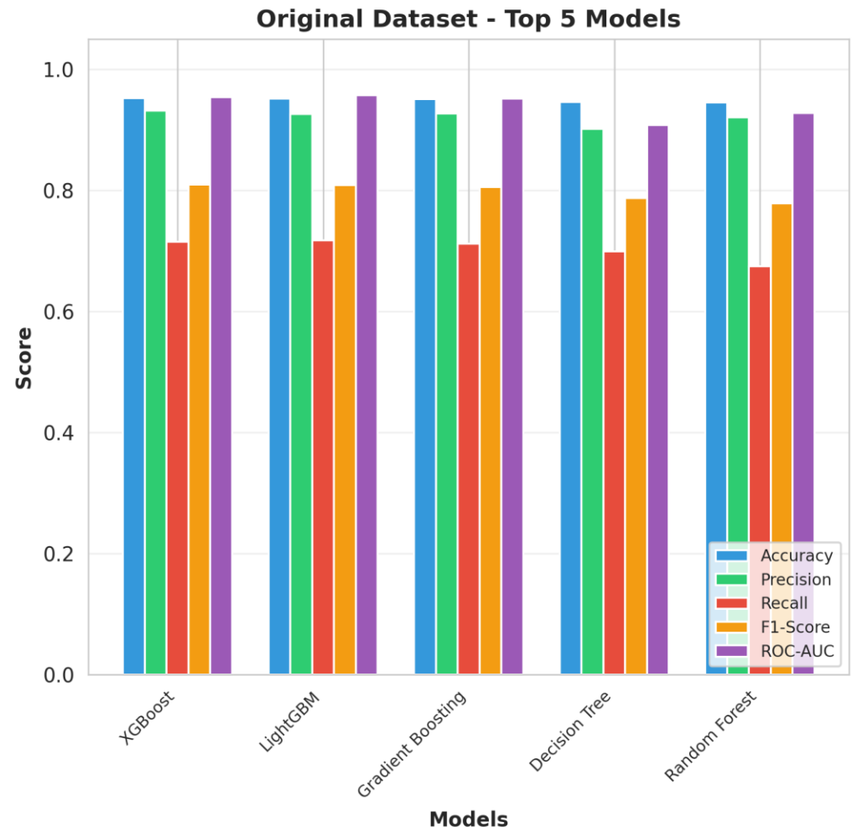
Balanced Datasets

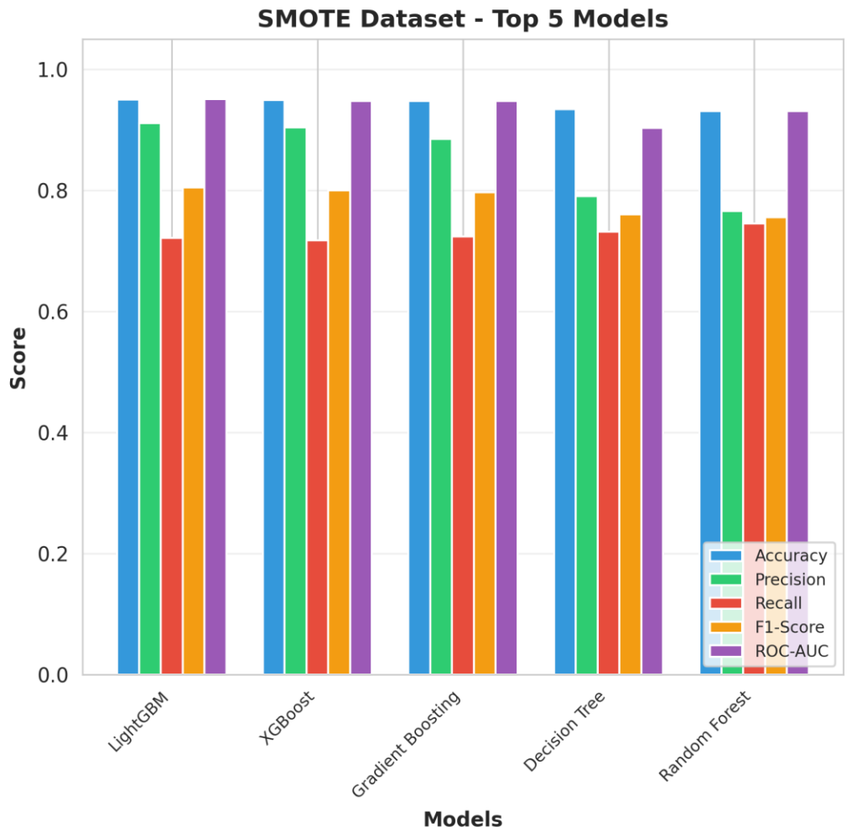
Enhanced recall without sacrificing accuracy.

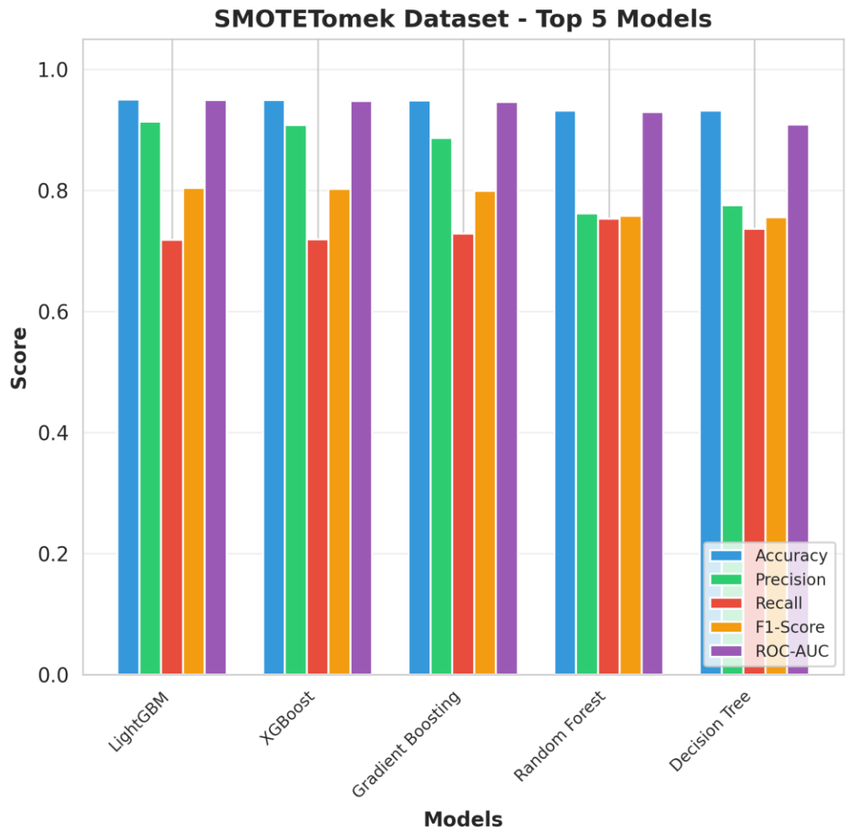
2

Original Dataset

XGBoost and LightGBM achieved highest overall metrics.







Top Model Results & Evaluation

Top 3 Models

XGBoost (Orig)

Achieved > 95 % accuracy and ≈ 0.81 F1 score.

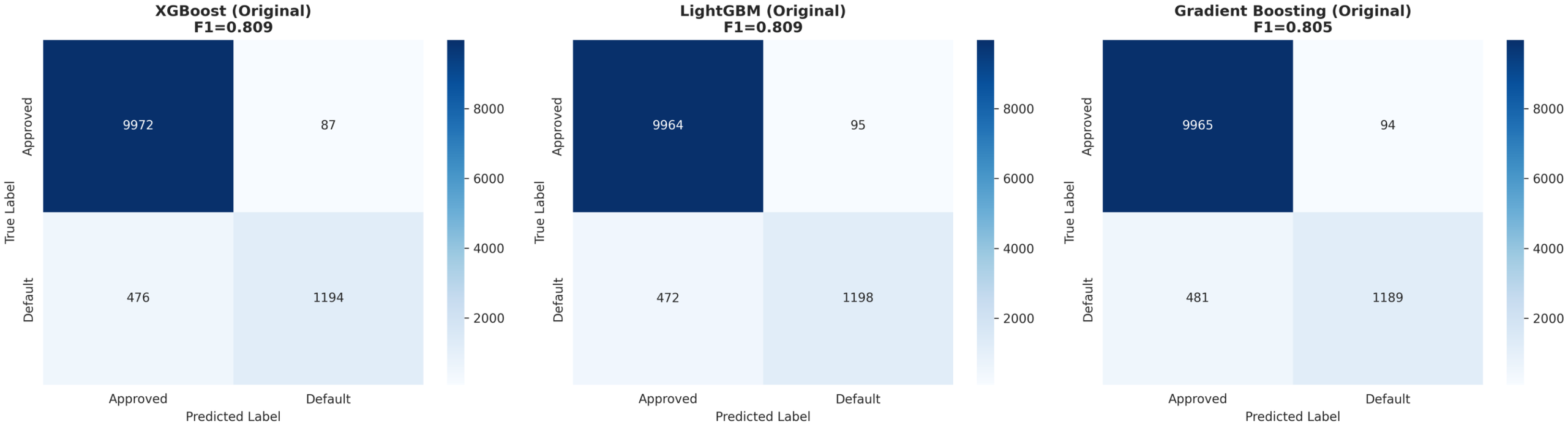
LightGBM (Orig)

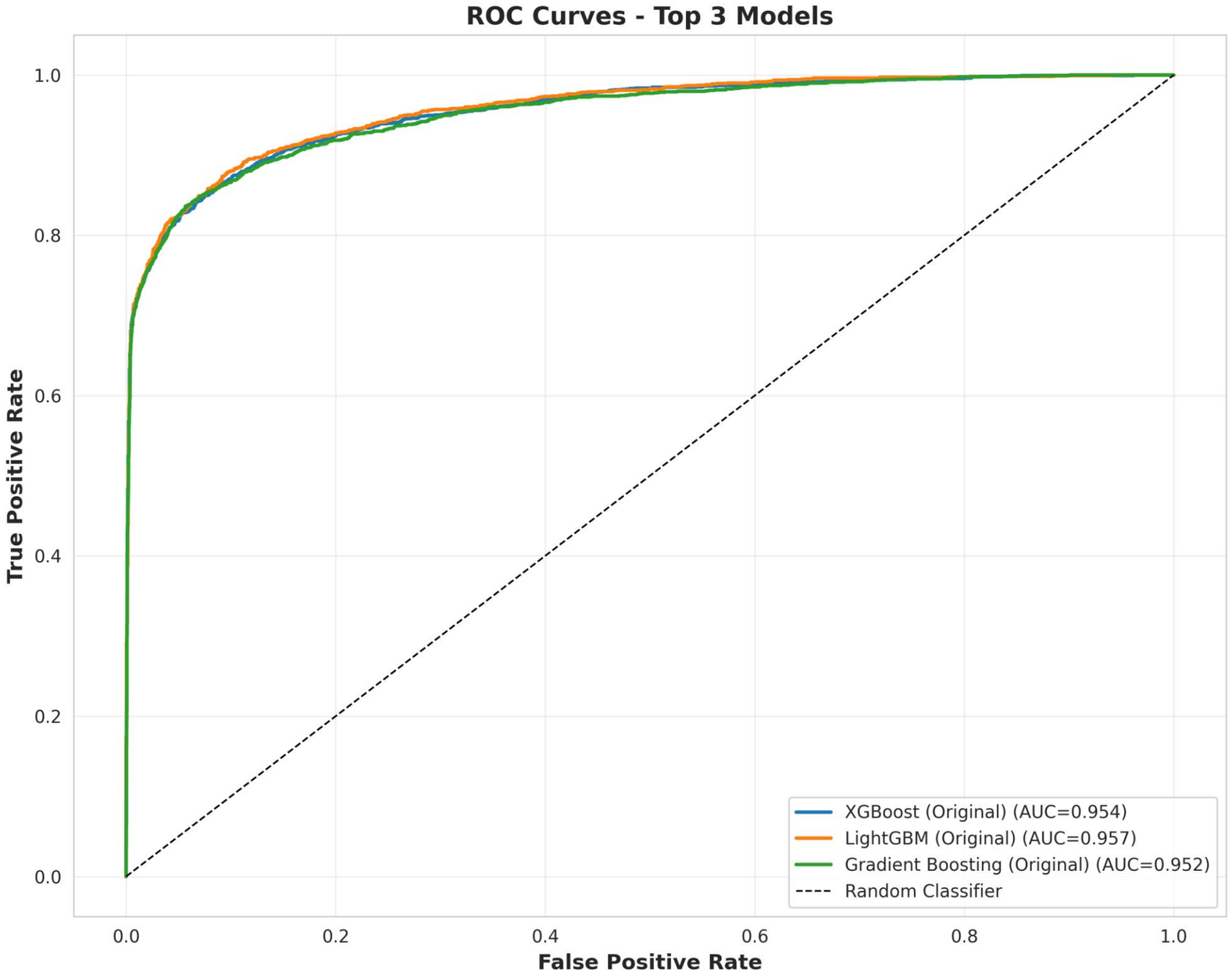
Achieved > 95 % accuracy and ≈ 0.81 F1 score.

Gradient Boost (Orig)

Achieved > 95 % accuracy and ≈ 0.81 F1 score.

High recall (~ 0.72–0.73) for the ***Default*** class.





Tuned Best Model (XGBoost) & Validation

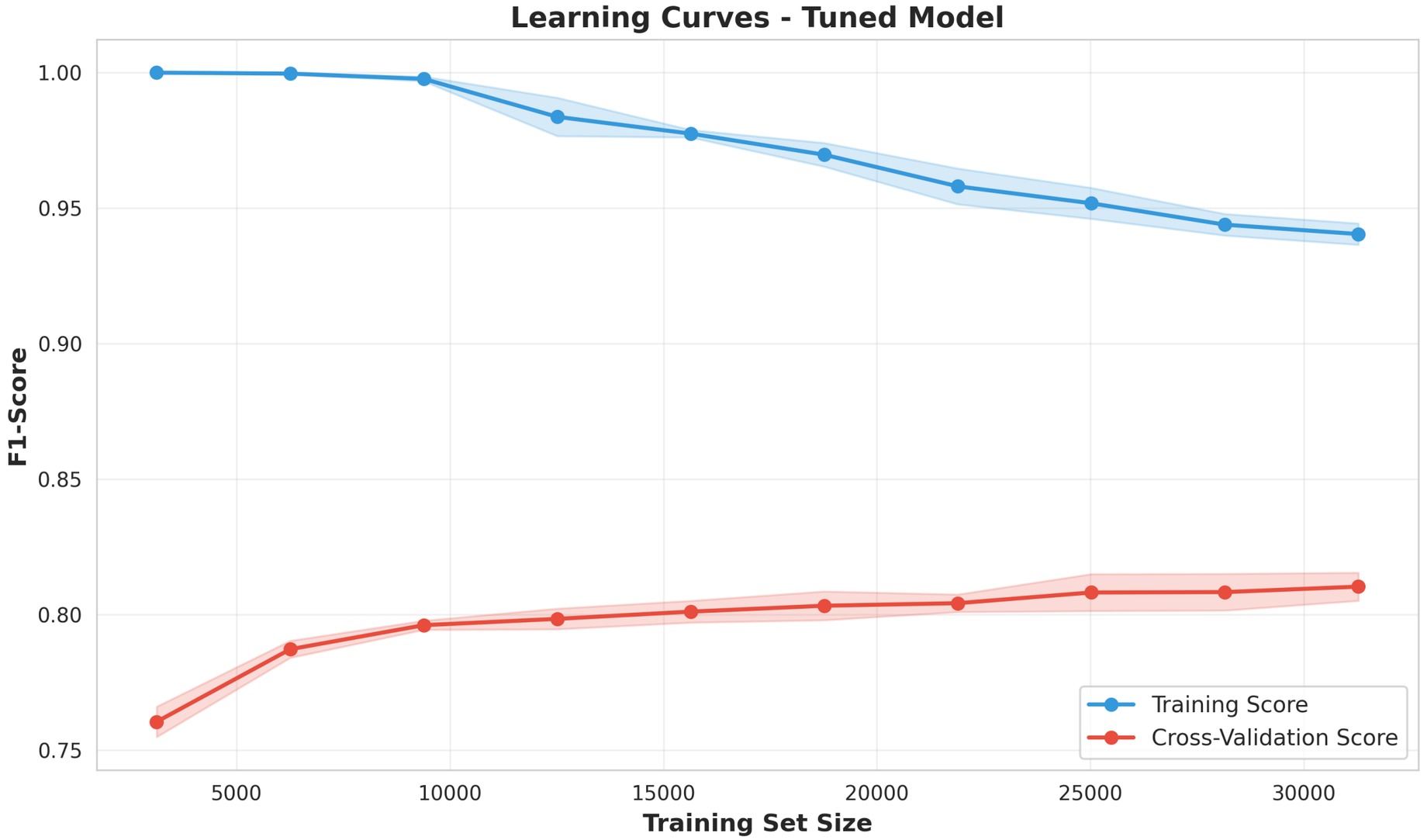
Tuning Process

Tuned via GridSearchCV → parameters: colsample_bytree 0.8, learning_rate 0.1, max_depth 7, n_estimators 300.

5-Fold CV F1 = 0.810 (± 0.025).

Final Test Metrics

- Accuracy 0.9517
- Precision 0.913
- Recall 0.731
- F1 0.812
- ROC-AUC 0.953



Feature Importance & Progress from CP1

Top Predictors

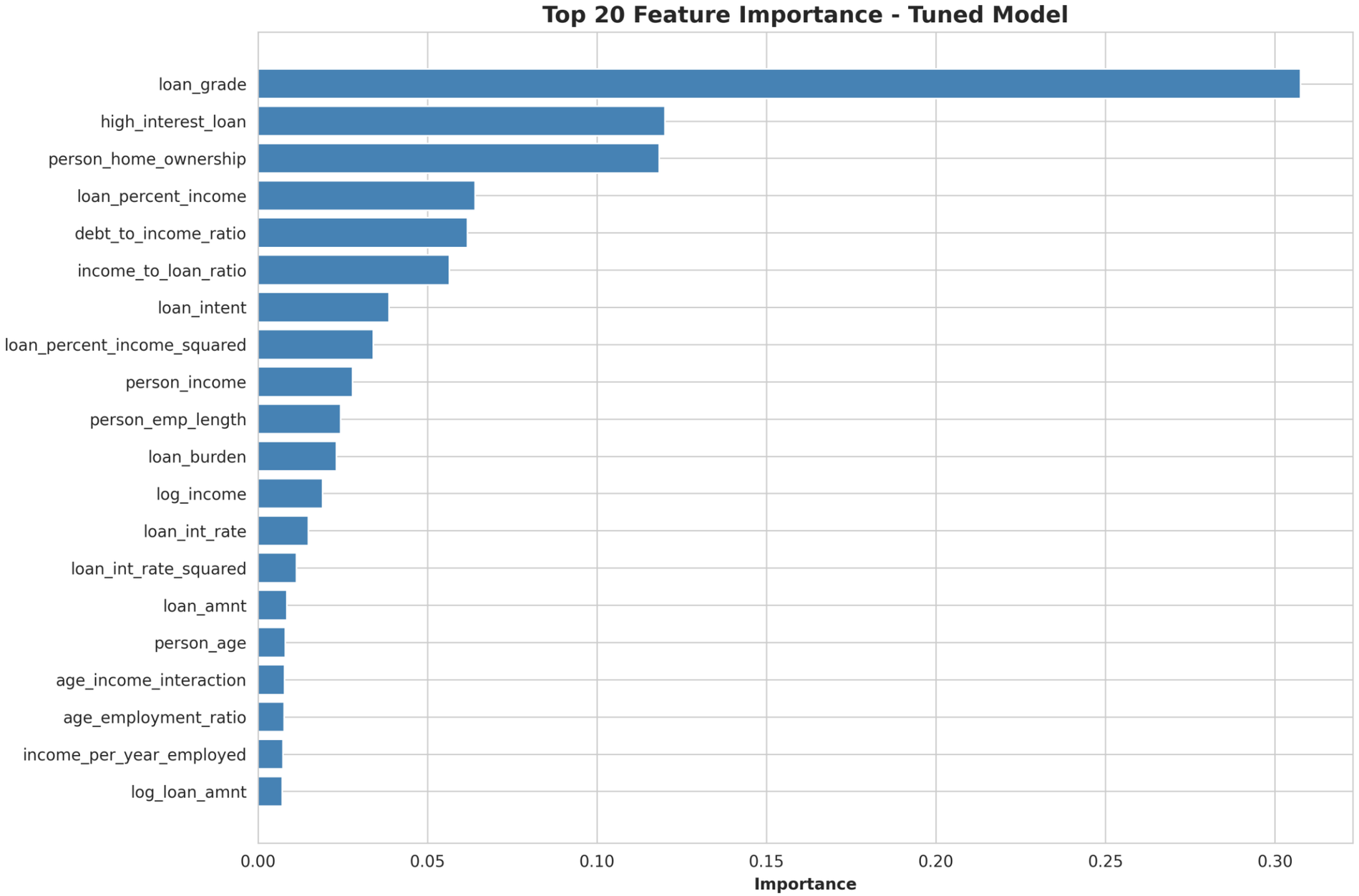
loan_grade, high_interest_loan, home_ownership, loan_percent_income, debt_to_income_ratio, income_to_loan_ratio.

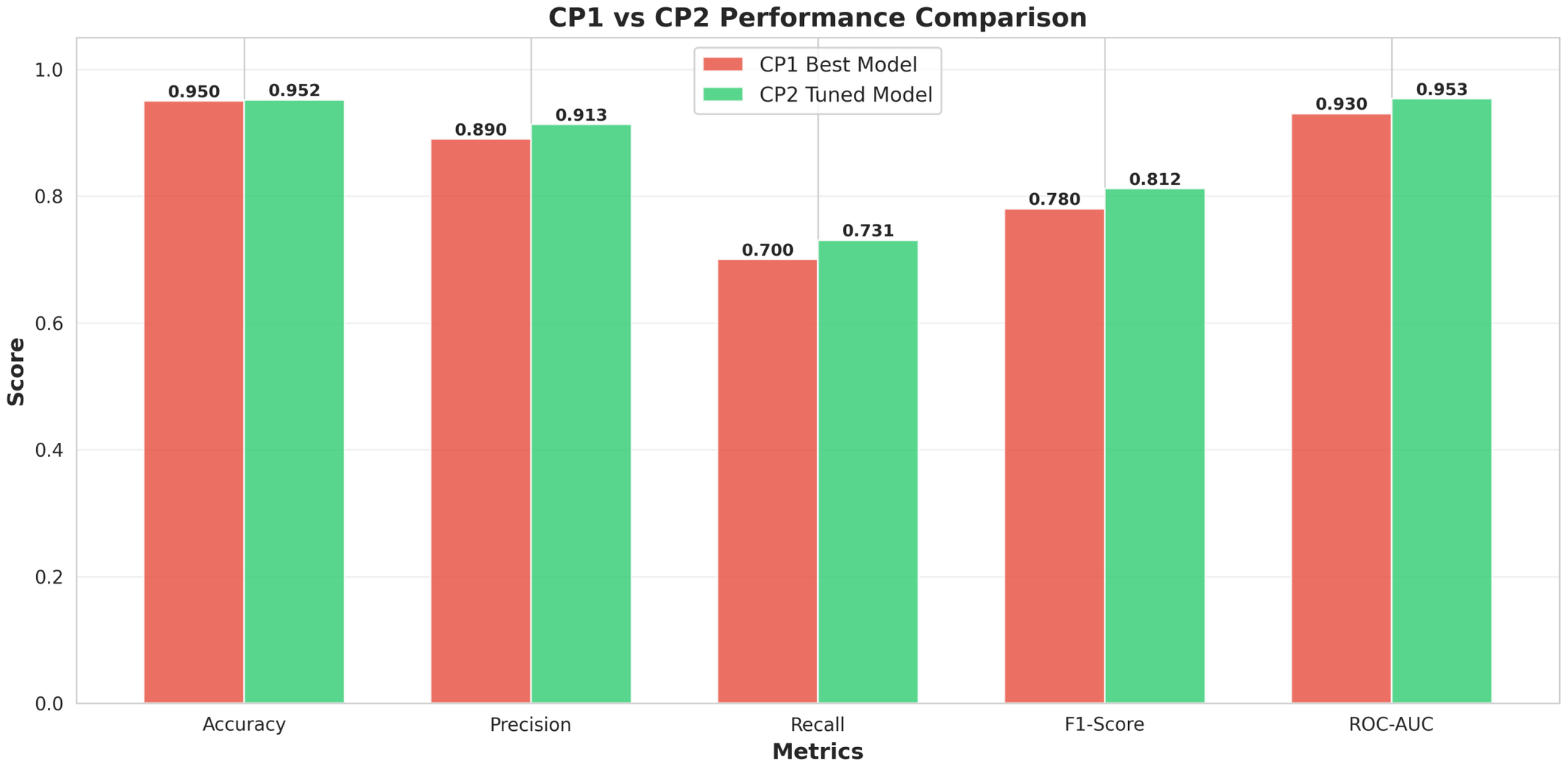
Enhancement Impact

Feature engineering and tuning enhanced recall and stability.

Improvement vs CP1

F1 + 4.1 % | Recall + 4.4 % | ROC-AUC + 2.5 %.





Next Steps (Short-Term Final Scope)

1 SHAP Visualizations

Finalize SHAP visualizations for model explainability.

2 Feature Selection

Integrate feature selection to streamline the final model.

3 Final Report

Prepare final report with business impact discussion and deployment outline.

Thank you

Video Link: https://drive.google.com/file/d/1n0VxR-CIltRnNrpPSafxPl_7Kl55jU8T/view?usp=sharing