Classification (Cross Validation)

Hayden Hawley

Overview

The objective of this project was to build a classification model for predicting loan approval using cross-validation and grid search to identify the best-performing model.

Data Exploration

During my initial manual review of the dataset, some noteworthy patterns seemed to emerge. These are the features I decided to use:

Home Ownership Status (person_home_ownership)

Renters showed a higher likelihood of loan approval, whereas those who own their homes are generally less likely to be approved. Applicants with a mortgage appeared more neutral.

Loan Amount (loan_amnt)

Loans with amounts less than or equal to $6,000 were predominantly not approved; lower loan amounts may not meet the lender's criteria.

Loan Interest Rate (loan_int_rate)

An interest rate of 15% or higher is strongly correlated with loan approval; higher rates might compensate for perceived risk or may be a hard policy criterion for approval.

Loan-to-Income Ratio (loan_percent_income)

A loan amount of 30% or more of the applicant's income is correlated approval; the relative size of the loan compared to income is a significant factor in the decision process.

Credit Default Indicator (cb_person_default_on_file)

A recorded default ("Yes" in this field) is a strong predictor of a rejected loan.

Baseline Results

To establish a benchmark, I trained an initial model using the default SGD with the selected features. The results were:

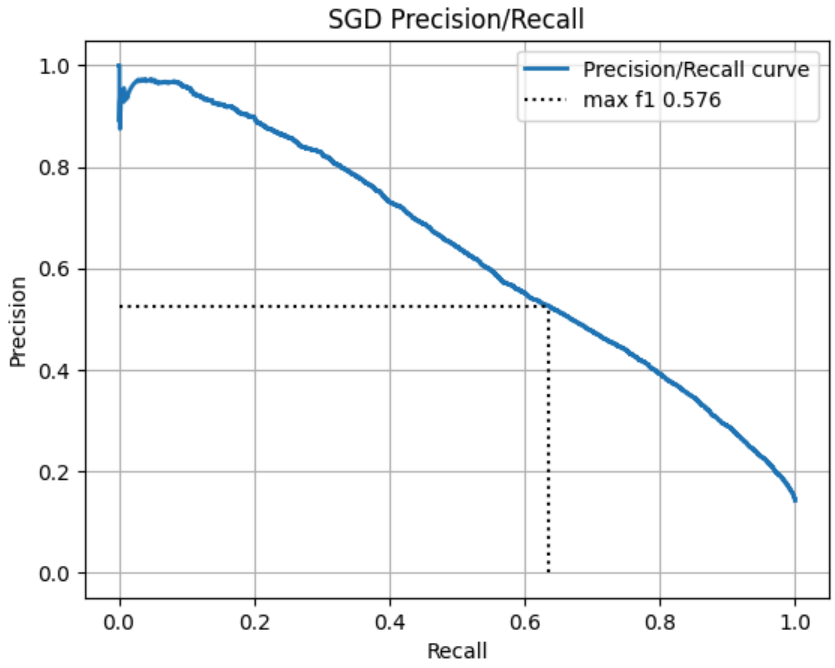
Cross Validation Score: 0.894 : ['0.896', '0.892', '0.895']

Confusion Matrix:

Precision: 0.772

Recall: 0.364

F1: 0.495



Model Selection

To improve upon the baseline model, I conducted a grid search across four different model types. The results are summarized below:

Model	# Candidates	Train Score	Cross Val. Score	Precision	Recall	F1
SGD	24	89.34%	89.38%	0.762	0.376	0.503
Linear	32	88.37%	88.37%	0.865	0.217	0.347
Boost	27	93.24%	92.83%	0.828	0.611	0.703
Forest	72	93.30%	92.54%	0.749	0.61	0.673

Gradient Boosting (Boost) and Random Forest (Forest) showed the best performance, so I refined these models further by expanding the hyperparameter search. The updated results are:

Model	# Candidates	Train Score	Cross Val. Score	Precision	Recall	F1	Kaggle Score
Boost v2	50	93.22%	92.84%	0.828	0.611	0.703	**0.814**
Forest v2	72	94.09%	92.62%	0.752	0.613	0.675	**0.805**

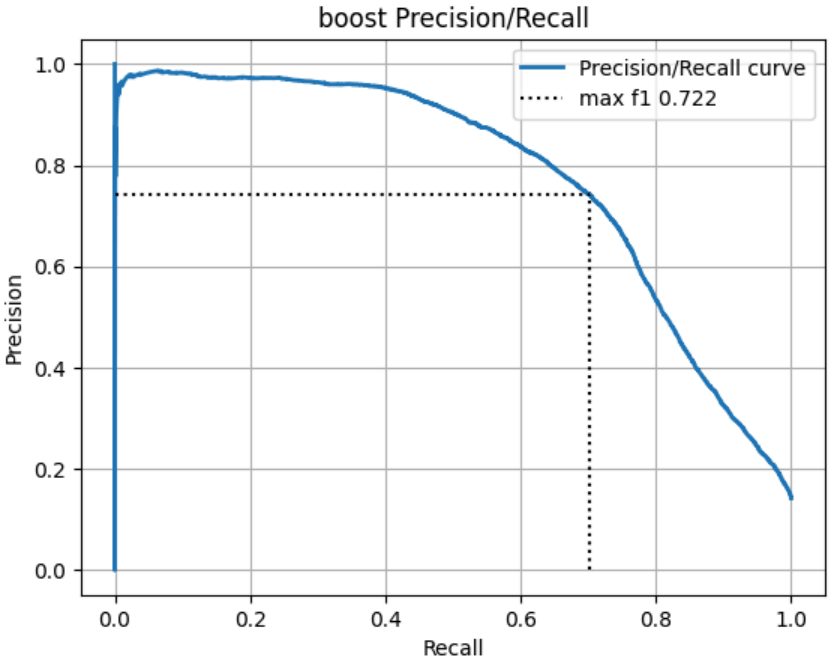
After tuning, the best hyperparameters for each model were:

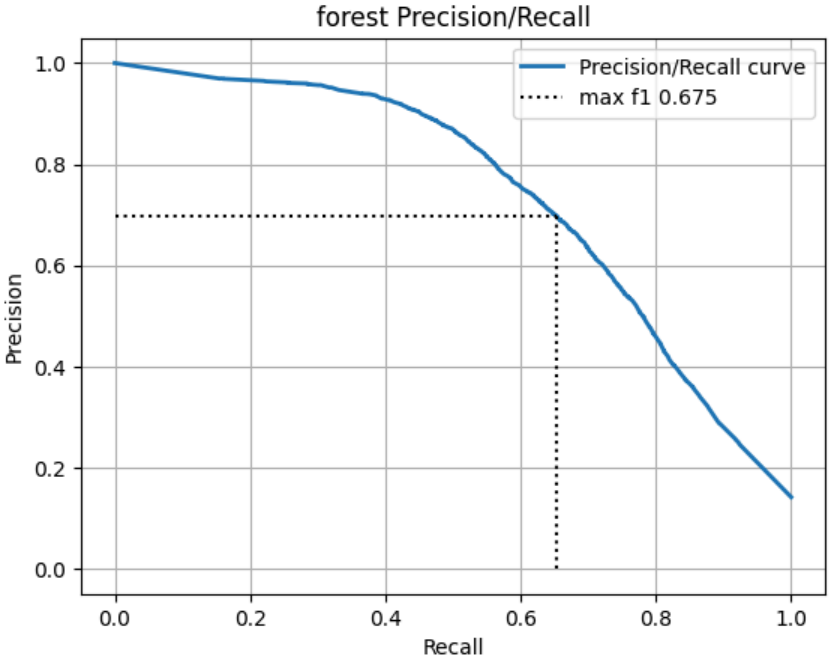
Boost - Best Params:

{ 'features__categorical__categorical-features-only__do_numerical': False,
'features__categorical__categorical-features-only__do_predictors': True,
'features__categorical__encode-category-bits__categories': 'auto',
'features__categorical__encode-category-bits__handle_unknown': 'ignore',
'features__numerical__numerical-features-only__do_numerical': True,
'features__numerical__numerical-features-only__do_predictors': True,
'model__learning_rate': 0.175,
'model__max_depth': 3,
'model__n_estimators': 200}

Forest - Best Params:

{ 'features__categorical__categorical-features-only__do_numerical': False,
'features__categorical__categorical-features-only__do_predictors': True,
'features__categorical__encode-category-bits__categories': 'auto',
'features__categorical__encode-category-bits__handle_unknown': 'ignore',
'features__numerical__numerical-features-only__do_numerical': True,
'features__numerical__numerical-features-only__do_predictors': True,
'model__class_weight': None,
'model__max_depth': 12,
'model__max_features': 'sqrt',
'model__min_samples_split': 3,
'model__n_estimators': 190}


boost Precision/Recall
Precision/Recall curve
max f1 0.722
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0


forest Precision/Recall
Precision/Recall curve
max f1 0.675
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Evaluation

Expanding the grid search to include additional hyperparameter combinations yielded negligible improvements. While the cross-validation scores were strong, the final Kaggle scores were disappointing.

If I were to tackle this project again, I would reconsider some of the features that were initially excluded. It's possible that they have a greater impact on loan approval than initially assumed.