Predicting Loan Payback

Rajorshi Sarkar · Surabhi Keesara · Nidhi Vinodbhai Patel

Northeastern University

Introduction

Predicting whether a borrower will repay a loan is critical to financial risk management. It reduces defaults and helps evaluate borrower credibility. Many scoring systems miss deeper relationships between borrower behavior and loan characteristics, making machine learning useful to improve accuracy.

Using ML, we analyze borrower demographics, financial indicators, and loan attributes and estimate the likelihood of successful loan repayment. By exploring linear and non-linear relationships, we maximize model accuracy.

Goal of Project

Our goal is to better understand the loan prediction process. We raise the following questions:

(1) What features are most important for predicting loan repayment probability?
(2) What model characteristics best take these features into account?
(3) More generally, what are the best methods of training and tuning regression ML models?

Methods

Data Pre-processing

- Cleaned dataset and confirmed no critical missing values.
- Applied one-hot encoding to all categorical columns.
- Checked numeric distributions for skew and outliers.
- Created engineered features:
- monthly_debt_payment =loan_amount*interest_rate
- grade_num and subgrade_num
- Performed feature selection using:
- Pearson correlation (removed highly correlated features)
- RFE (removed annual_income, loan_amount)
- Random Forest importance (removed grade_num,subgrade_num)

Dataset

- Kaggle Playground S5E11 Loan Repayment Dataset
- 593,994 training rows, 254,569 test rows
- Target: loan_paid_back (binary)

Modeling Approach

Our approach consisted of 4 different models. We tried variations with early stopping, variable learning rate, various tuning methods, and different feature encodings. The best performing parameters are detailed below with the best variation highlighted

Logistic Regression

Variations: L1, L2, Manual Regularization Tuning

penalty		"L2"
C		2.0
solver		"lbfgs"
max_iter		2000
n_jobs		1

Light GBM

Variations: Variable learning rate, Randomized Search, Optuna, Early Stop

n_estimators	1786		
learning_rate	0.01956	colsample_bytree	0.5325
num_leaves	21	reg_alpha	3.467127
min_child_samples	129	reg_lambda	4.90555
subsample	1	max_bin	231

CATBOOST

Variations: Standard, Randomized Search

iterations	3000	loss_function	Logloss
learning_rate	0.1	random_seed	53
depth	4	bagging_temperature	0.5
l2_leaf_reg	1		

MLP – Neural Network

Variations: Standard, Manual Regularization

batch_size	1024	activation	Relu
learning_rate	Adaptive	solver	Adam
early_stopping	True	alpha	1e-4

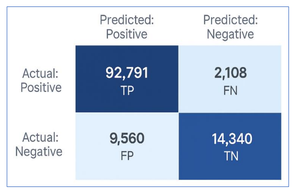
We also experimented with combining the best iterations of the above models into a ensemble stacked ML model.

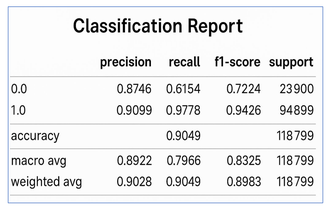
Experiments & Results

We used ROC Curve, Brier Calculations, and Gain & Lift Charts to analyze our models. Below is a leaderboard ranking of our top models by ROC Curve (based on Kaggle Competition)

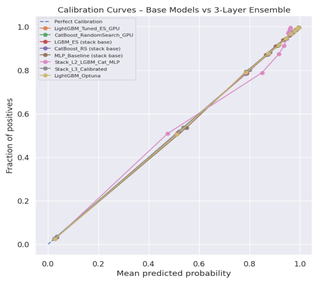
MODEL	ROC_AUC
STACK_L3	0.922294
LGBM_OPTUMA	0.921750
LGBM_Tuned_ES	0.921658
CATBOOST_RS	0.919403
LogReg_LASSO	0.910602
MLP_Standard	0.909912

Confusion Matrix and Classification Report for 3 Layer Stack Ensemble (Our Best Model)

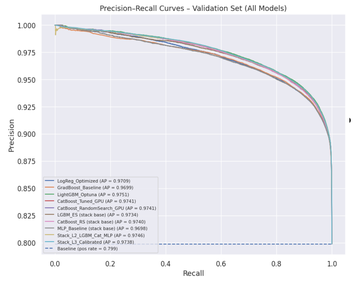




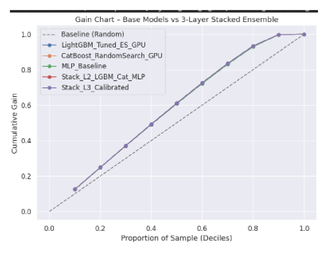
BRIAR CALIBRATION



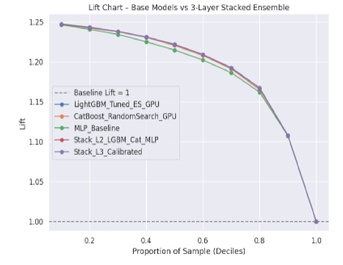
ROC CURVE



GAIN



LIFT



Conclusion

- All strong ML models achieved ROC–AUC between **0.91–0.92**, showing the loan repayment task is well-captured by boosted-tree models.
- The **best single model**, based on **LightGBM** reached **AUC = 0.921750** and **AP ≈ 0.975**, with excellent recall for "paid back."
- A **3-layer stacked ensemble** (LGBM + CatBoost + MLP → Logistic Regression → Isotonic calibration) achieved **AUC = 0.922294**, giving a *consistent* improvement in ranking and calibration.
- Class imbalance (~80% paid back) leads to **high recall but moderate specificity (**due to class imbalance) across models.
- Calibration curves and Brier scores show that our stacked ensemble model produces **highly reliable probability estimates**, essential for credit-risk decisions.
- Gain & Lift analysis confirms that the top 20–40% of model-ranked customers contain the majority of "good borrowers," demonstrating strong business utility.
- Our final submission scored **0.92280 on Kaggle**, exceeding validation performance and ranking within the **top ~200 teams**. demonstrating the importance of novel approaches in model selection, data interpretation, feature engineering and tuning ,
- **Through this Kaggle competition, we learned how dataset engineering, model choice, tuning, stacking, calibration, and rigorous validation drive real-world leaderboard performance, and how even small gains in each area can have a big impact in the real world.**

References

1] Kaggle Dataset: Loan Repayment (Playground S5E11)
https://www.kaggle.com/competitions/playground-series-s5e11