Loan Pricing Prediction Using Machine Learning Models

Vidit Pokharna

November 16, 2025

1. Introduction

This project predicts loan rate spreads using firm, deal, and macroeconomic variables. Two neural models were required: a baseline supervised model and a multi–task model with an auxiliary output. The goal is to evaluate whether auxiliary learning improves predictive accuracy.

Benchmarks (LASSO, KNN, XGBoost, LightGBM) and two MLP architectures were trained and compared using mean squared error (MSE). Additional diagnostics—prediction plots, residuals, feature importance, and correlation analysis—support model interpretation.

2. Data

The dataset includes firm financial ratios, loan characteristics (facility amount, maturity, seniority, secured flag, revolver status), and macro indicators (GDP growth, credit spreads, rates). Numeric variables were standardized; categorical ones were one-hot encoded.

Missing values were imputed. A correlation heatmap (Figure 1) highlights relationships and multicollinearity patterns among predictors.

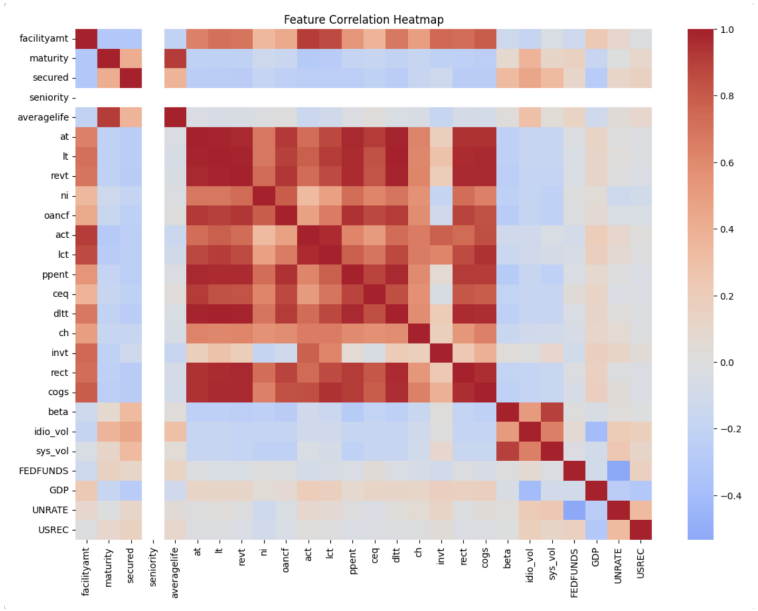

Figure 1: Correlation Heatmap of Input Features

3. Modeling Approach

3.1 Baseline Models

The following models were trained on an 80/20 train–test split with MSE as the evaluation metric:

- LASSO regression
- KNN
- XGBoost
- LightGBM
- Baseline MLP

3.2 Auxiliary Multi–Task Model

The auxiliary model uses:

- A shared MLP trunk,
- A main head predicting loan spread,
- A secondary head predicting an auxiliary target.

Losses are combined via a weighted sum. The intent is to test whether the auxiliary prediction stabilizes or improves the representation learned by the network.

4. Results

Model	Test MSE
LASSO	245867.79
XGBoost	4689.19
LightGBM	2377.37
KNN	2870.64
MLP Baseline	2204.34
MLP + Auxiliary Output	5642.95

Table 1: Model Performance on Test Set

The baseline MLP is the strongest performer. LightGBM also performs well. The auxiliary model performs noticeably worse, indicating that the auxiliary task did not help the primary objective.

4.1 Prediction vs Actual

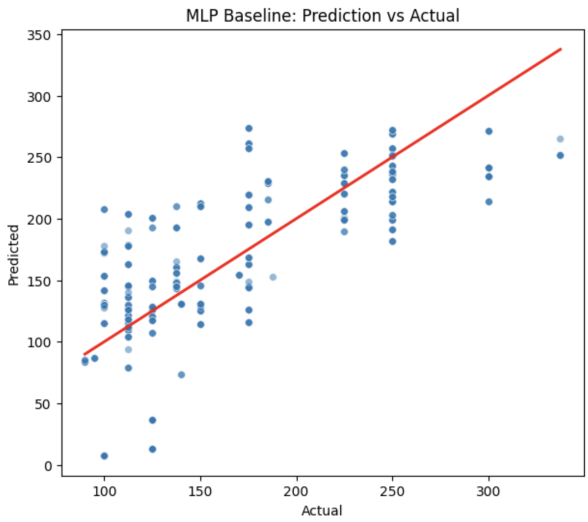

Figure 2: MLP Baseline

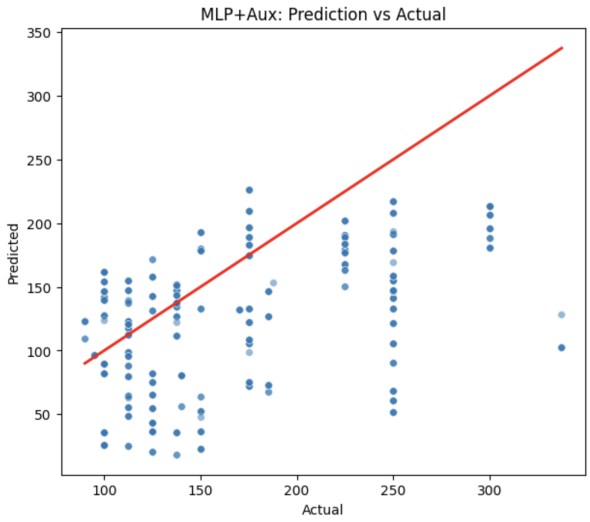

Figure 3: MLP + Aux

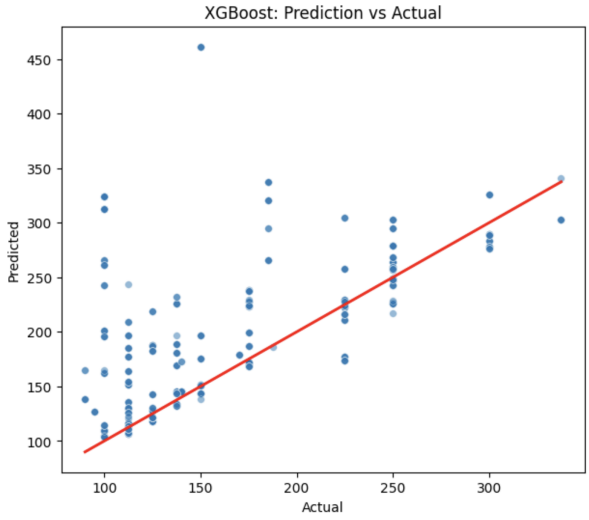

Figure 4: XGBoost

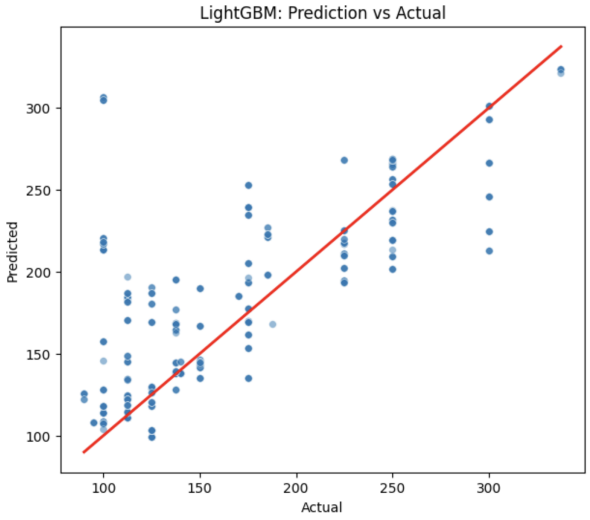

Figure 5: LightGBM

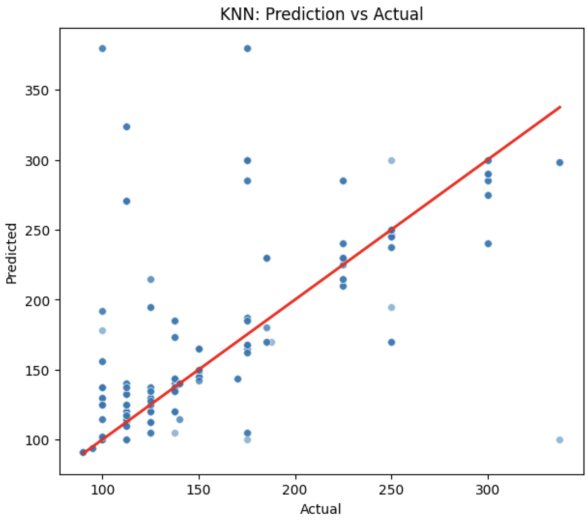

Figure 6: KNN

4.2 Residual Distributions

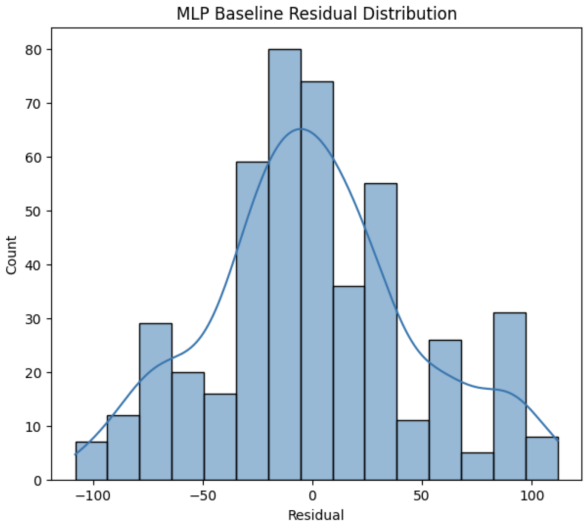

Figure 7: MLP Baseline Residuals

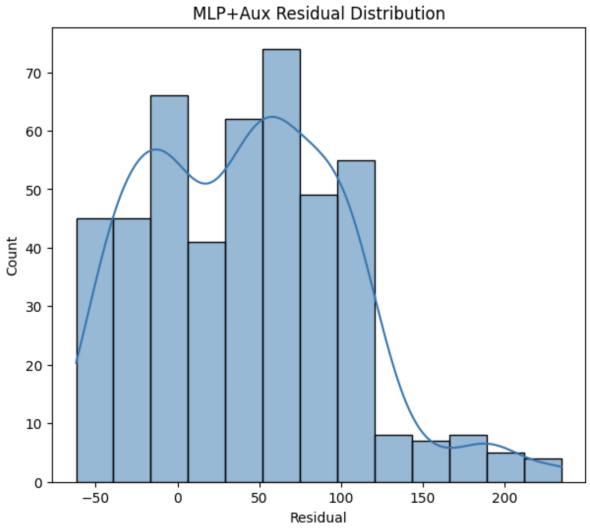

Figure 8: MLP + Aux Residuals

4.3 Feature Importance

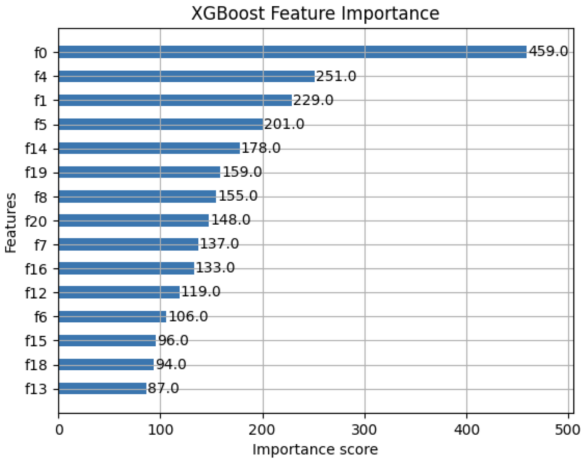

Figure 9: XGBoost Feature Importance

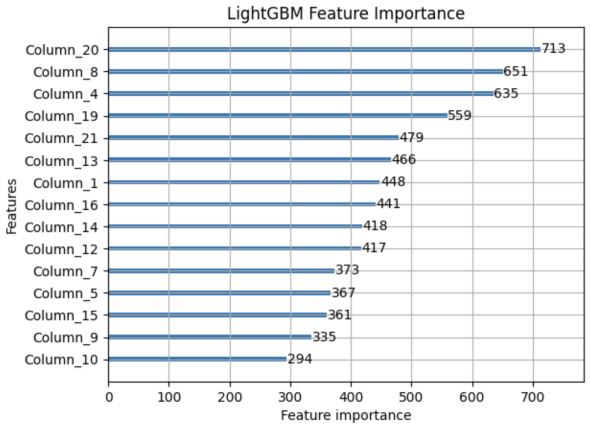

Figure 10: LightGBM Feature Importance

5. Discussion

The baseline MLP's strong performance suggests meaningful nonlinear structure in loan pricing data. LightGBM provides comparable accuracy with interpretable feature importance.

The auxiliary model underperforms the baseline. The auxiliary target ("allinundrawn") is only weakly correlated with the main target ("allindrawn"), so the shared representation is forced to learn two objectives that do not reinforce one another. In such cases, multi–task learning degrades rather than improves performance. This illustrates a key principle: auxiliary tasks only help when they share meaningful structure with the primary prediction problem.

Tree–based model importances confirm that facility size, maturity, seniority, and leverage drive loan spread variation, consistent with credit intuition.

6. Conclusion

Across all models, the baseline MLP achieves the lowest test error. The auxiliary head does not yield improvements due to the weak relationship between the auxiliary and main targets. LightGBM and XGBoost remain competitive alternatives with greater interpretability.

Future improvements include testing more informative auxiliary targets, applying stronger regularization, or extending the feature set with more macroeconomic or industry variables.