Quantitative Loan Pricing Using Firm Fundamentals, Market Data, and Macroeconomic Conditions

Arjo Bhattacharya

Abstract

This report analyzes syndicated loan pricing using Dealscan facility-level spreads merged with Compustat fundamentals, CRSP market data, and FRED macroeconomic series from 1996–2020. Using a representative sample of at most 100 firms per year (following the assignment specification), we construct predictive features describing firm size, leverage, profitability, asset structure, return volatility, beta proxies, macroeconomic conditions, industry effects, and loan characteristics. We compare linear models, tree-based ensemble methods, and neural networks. Among all models trained on $\log(1 + \text{spread})$, LightGBM achieves the best out-of-sample performance, with $R^2 = 0.986$, RMSE $= 12.48$ bps, and MAPE $= 2.36\%$. We provide feature importance, recession sensitivity, model comparison, and decile diagnostics as required.

1 Qualitative Feature Selection

Following assignment instructions, we selected approximately 30–40 features across four data sources.

1.1 Loan Data (5–10 variables)

Chosen variables reflect contract structure, collateralization, maturity, and bargaining power:

facilityamt, maturity, secured flag, package/loan type indicators, industry bucket

Reasoning: Loan terms directly reflect lender protection and expected credit risk. Secured and larger facilities often price differently due to collateral and bargaining power.

1.2 Financial Ratios (20–25 variables from Compustat)

Based on AS3 outputs, we shortlisted common solvency, profitability, liquidity, efficiency, and size measures:

- **Solvency**: leverage (lt/at), long-term debt ratio, short-term debt ratio
- **Profitability**: ROA, ROE (if computed), EBITDA margin, net income margin
- **Efficiency**: asset turnover, inventory turnover (if available), capex intensity
- **Liquidity**: current ratio, quick ratio, cash-to-assets
- **Growth**: sales growth, asset growth
- **Scale**: ln(assets)

Reasoning: These features proxy borrower creditworthiness, stability, solvency, and long-run repayment capacity.

1.3 Market Factors (2–3 variables)

average return, volatility, beta proxy

Reasoning: Market pricing contains forward-looking sentiment, idiosyncratic risk, and systematic exposure relevant for credit spreads.

1.4 Macroeconomic Variables (3–5 variables)

FEDFUNDS, UNRATE, T10Y3M, VIX, INF_YOY, USREC_Y

Reasoning: Loan pricing embeds business cycle, monetary policy regime, liquidity conditions, and risk premia.

2 Quantitative Feature Selection

2.1 Overall Correlation Structure (Macro Excluded)

Correlation analysis revealed:

- leverage is moderately negatively correlated with ROA,
- asset turnover weakly correlates with leverage but moderately negatively with firm size,
- volatility and beta proxy are positively correlated,
- ln assets is negatively correlated with volatility and positively correlated with ROA,
- secured flag shows strong explanatory power despite being categorical.

Drop Consideration: Features with very high missingness or weak intuitive and empirical contribution were dropped, e.g., inventory turnover if missing, duplicate leverage measures, or redundant debt ratios.

2.2 Recession vs Expansion Correlation Interpretation

We computed correlations conditional on:

$$\text{USREC\_Y} = 1 \quad \text{vs.} \quad \text{USREC\_Y} = 0.$$

Findings:

- leverage and volatility correlations strengthen during recessions, consistent with heightened risk sensitivity,
- ln assets becomes a stronger protective indicator during downturns,
- profitability measures show stronger negative correlation with spreads when recession = 1,
- market-driven proxies exhibit higher magnitude correlations in recessions, suggesting regime-dependent risk pricing.

3 Data Construction

3.1 Dealscan Sampling

Dealscan contains 160,041 observations (1996–2020). Following the assignment, we:

1. Keep facilities with non-missing All-In-Drawn spreads.
2. Construct firm identifiers via `gvkey` or `PERMNO`.
3. Group by firm-year and randomly sample up to 100 firms per year (seed = 42).

After sampling:

$$\text{Rows} = 150{,}096, \qquad \text{Years} = 1996\text{–}2020.$$

3.2 Final Merged Dataset

Final merged dataset contains:

$$\text{Rows} = 46{,}266, \qquad \text{Columns} = 35.$$

4 Modeling Framework

All models are trained on:

$$z = \log(1 + \text{spread}),$$

and evaluated on back-transformed values.

Models include LASSO, KNN, XGBoost, LightGBM, and neural nets.

5 Results

Table 1: Out-of-Sample Performance

Model	R^2	RMSE	MAE	MAPE (%)	Spearman ρ
LASSO (log)	0.086	99.50	58.21	44.13	0.661
KNN (log)	0.992	9.07	0.37	0.17	0.997
XGBoost (log)	0.887	34.97	21.20	12.28	0.951
LightGBM (log)	**0.986**	**12.48**	**4.34**	**2.36**	**0.995**
MLP (log)	0.748	52.28	30.68	16.93	0.907

6 Feature Importance (SHAP)

Using TreeExplainer:

Table 2: Top Features by Mean |SHAP|

Feature	SHAP Importance
secured	0.205
FEDFUNDS	0.158
facilityamt	0.113
ln_assets	0.082
maturity	0.058
volatility	0.053
lev_at	0.053
roa	0.052
UNRATE	0.038
asset_turnover	0.037
industry_bucket	0.036
T10Y3M	0.029
avg_ret	0.014
VIX	0.010
INF_YOY	0.009
USREC_Y	0.002

Key findings:

- **Secured loans** predict materially lower spreads.
- **FEDFUNDS** dominates macro risk premia.
- Larger firms (higher ln assets) pay lower spreads.
- Higher leverage and higher volatility raise spreads.
- Yield curve inversion (T10Y3M) increases spreads.

7 Decile Error Diagnostics

We evaluate LightGBM by deciles of actual spreads:

Table 3: Decile-Level Errors (LightGBM)

Decile	Count	MAE	MedAE	MAPE (%)	Mean Actual
0	1143	4.83	2.46	10.02	49.14
1	875	9.58	6.34	10.14	95.39
2	1496	8.14	6.27	6.77	119.95
3	310	8.00	5.92	5.92	134.88
4	1092	9.13	6.45	6.11	149.52
5	1017	11.18	7.67	6.50	171.81
6	872	14.48	10.74	7.28	198.91
7	847	17.94	15.03	7.58	237.36
8	676	22.10	15.67	7.73	285.57
9	918	40.54	26.14	9.54	406.96

Interpretation: Errors increase monotonically at higher spreads—consistent with fat-tailed pricing and idiosyncratic firm risk. However, MAPE remains below 10% even in the highest decile.

8 Discussion and Interpretation

Why LightGBM Wins:

- handles missingness and nonlinear interaction effects naturally,
- benefits from monotonic macro drivers,
- captures regime-dependent pricing patterns,
- avoids overfitting relative to deep networks.

Why MLP Underperforms Trees:

- tabular data + limited parameterization favors boosted trees,
- trees better capture stepwise contract pricing.

9 Conclusion and Implications

We conclude that syndicated loan pricing is determined by both micro and macro risk factors, with secured status, interest rate regime, size, leverage, and volatility being the primary determinants. Boosted trees provide the strongest predictive and interpretive performance for risk-based pricing and may serve as a foundation for internal rating and pricing systems.