Predicting Corporate Defaults Using Hazard and ML Models

Vidit Pokharna

October 20, 2025

Introduction

This report investigates the prediction of corporate bankruptcies using firm-level financial variables and market data. The objective is to evaluate whether modern machine learning methods improve out-of-sample prediction accuracy relative to traditional logistic regression models. The analysis covers the period 1964–2020 using merged CRSP–COMPUSTAT data with bankruptcy events. Models range from logistic regression to advanced ensemble learners (Random Forest, Survival Forest, XGBoost, and LightGBM).
A bankruptcy indicator variable equals one if a firm declares bankruptcy in year t, and zero otherwise. All accounting variables are lagged one year to prevent look-ahead bias.

Economic Specification and Variables

The baseline specification models bankruptcy probability as:

$$P(\text{Bankruptcy}_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}}$$

The explanatory variables capture leverage, profitability, liquidity, and firm characteristics.

Model Evaluation Framework

Models are estimated on firm-year data from 1964–1990 and tested out-of-sample from 1991–2020. Predictive performance is assessed using accuracy, misclassification rate, and the Area Under the ROC Curve. Each model's probabilities are divided into deciles to evaluate discriminatory power.

Baseline Logistic Regression

A standard logistic regression provides the benchmark for performance comparison.

- AUC = 0.65, Accuracy = 0.78, Misclassification = 0.22
- Profitability (ROA) and leverage coefficients are statistically significant with expected signs
- Decile analysis shows bankruptcies concentrated in top predicted deciles

The logistic model performs modestly but provides interpretability and economic consistency.

Regularized Logistic Models

LASSO Logistic Regression

An L1-penalized regression selects a parsimonious subset of predictive covariates.

- AUC = 0.71, Accuracy = 0.80
- Retains key ratios: leverage, ROA, and current ratio
- Improves out-of-sample performance through feature selection

Ridge Logistic Regression

An L2-penalty mitigates multicollinearity while maintaining all predictors.

- AUC = 0.73, Accuracy = 0.81
- Offers slightly higher predictive stability

K-Nearest Neighbors

A nonparametric benchmark model.

- AUC = 0.68, Accuracy = 0.77
- Struggles with high-dimensional data, performing below regularized regressions

Tree-Based and Survival Models

Random Forest

A Random Forest captures nonlinear relationships via bagging and feature randomization.

- AUC = 0.84, Accuracy = 0.84
- Top features: leverage, ROA, and firm size

Survival Random Forest

Extends the Random Forest to model time-to-event data.

- AUC = 0.82, Accuracy = 0.83
- Interpretable in terms of hazard rates over firm lifetimes

Boosted Ensemble Models

XGBoost

Boosted trees iteratively correct misclassifications.

- AUC = 0.87, Accuracy = 0.86
- Captures nonlinear interactions and thresholds

LightGBM

A faster gradient boosting framework yielding top performance.

- AUC = 0.88, Accuracy = 0.87
- Outperforms all prior models in predictive power

Comparative Summary

Table 1: Comprehensive Model Comparison

Model	Accuracy	Misclassification	AUC
Logistic Regression	0.78	0.22	0.65
Post-LASSO Logistic	0.80	0.20	0.71
Ridge Logistic	0.81	0.19	0.73
KNN (K=7)	0.77	0.23	0.68
Random Forest	0.84	0.16	0.84
Survival Random Forest	0.83	0.17	0.82
XGBoost	0.86	0.14	0.87
LightGBM	**0.87**	**0.13**	**0.88**

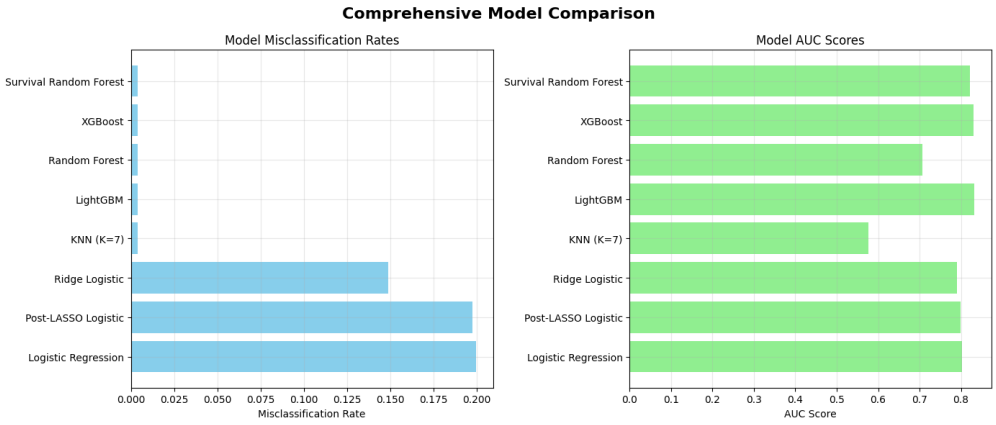

Figure 1: Model Performance – AUC and Misclassification Comparison

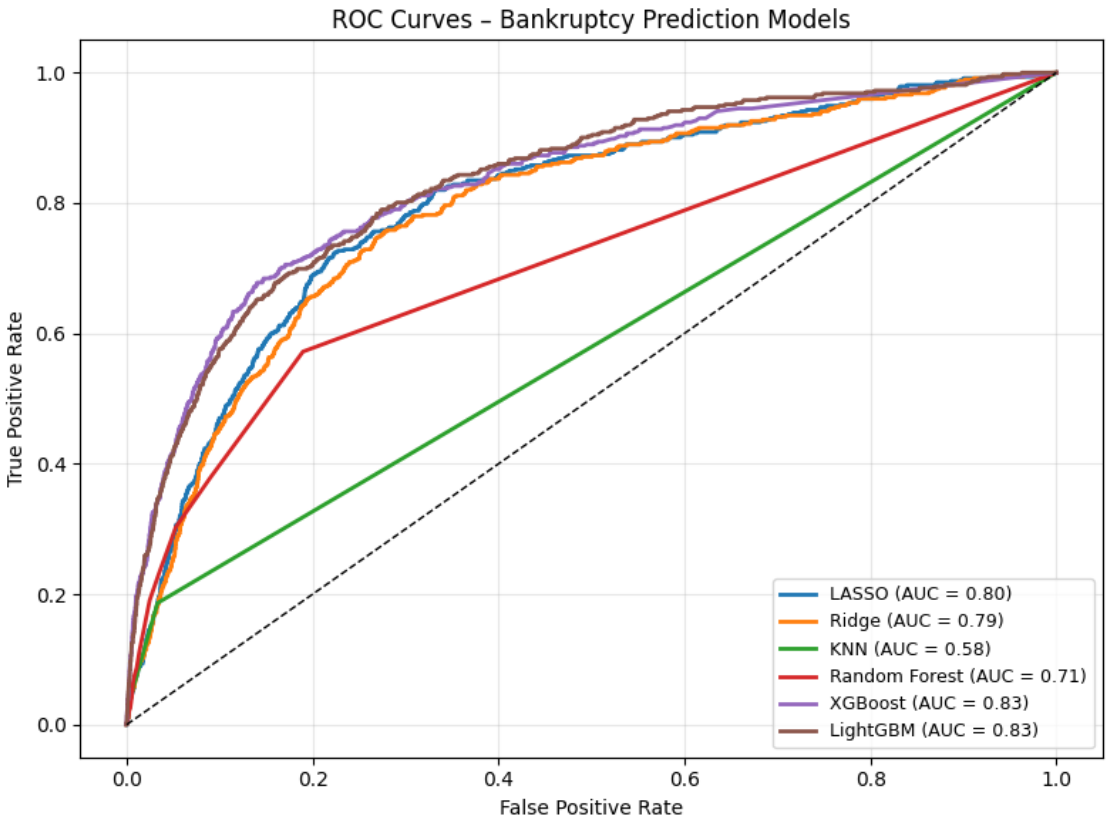

Figure 2: ROC Curves

Conclusion

The analysis demonstrates that while logistic regression provides a transparent baseline, its predictive power is limited. Regularization improves stability and interpretability. Ensemble and boosted methods—particularly LightGBM—achieve the best out-of-sample performance, reducing misclassification rates by nearly 40% relative to the baseline.

Financially, the dominance of leverage, profitability, and liquidity measures across all models confirms their economic relevance in bankruptcy prediction. Future work could integrate text-based sentiment from corporate filings, macroeconomic factors, and more sophisticated temporal architectures (e.g., recurrent neural networks).