Predicting Corporate Defaults Using Hazard and Machine Learning Models

MGT 8803 Assignment 6

December 1, 2025

1 Introduction and Objective

This report implements and compares eight statistical and machine learning models for predicting corporate bankruptcy using a comprehensive dataset spanning 1964–2020. The analysis leverages CRSP market data, COMPUSTAT fundamental data, and bankruptcy filings to construct a robust early warning system for corporate default risk. We estimate models on 1964–1990 data (88,381 firm-years, 435 defaults) and evaluate out-of-sample performance on 1991–2020 (202,070 firm-years, 973 defaults). The extremely low base rate (0.48%) creates significant class imbalance, making this a challenging prediction task requiring careful model selection and evaluation.

2 Data Description and Preprocessing

2.1 Data Sources and Merging

We merge three primary datasets: (1) CRSP monthly stock files (3.16M observations, 1964–2023) containing market data (prices, returns, volume), (2) COMPUSTAT annual fundamentals (539K observations) with accounting variables, and (3) UCLA-LoPucki Bankruptcy Research Database (3,878 bankruptcy events, 1964–2020). The merge proceeds in three steps: (i) convert CRSP to annual frequency using December observations, (ii) lag COMPUSTAT by one year (merge_year = fyear + 1) to eliminate look-ahead bias, and (iii) match on PERMNO after creating CUSIP-PERMNO crosswalk. The final dataset contains 290,451 unique firm-year observations with 1,408 bankruptcy events (0.48% base rate).

2.2 Variable Construction and Economic Rationale

We construct 13 explanatory variables spanning market-based measures, profitability, leverage, liquidity, and valuation. All variables were selected based on financial theory and prior empirical research *before* examining model coefficients. Table 1 presents each variable with economic rationale and expected coefficient sign.

3 Empirical Methodology

We implement eight models spanning linear, regularized, non-parametric, and tree-ensemble approaches:

Model 1: Logistic Regression. Standard maximum likelihood estimation on all 13 features. We estimate both (a) in-sample on full training period and (b) out-of-sample with 1964–1990 train, 1991–2020 test split.

Model 2: LASSO Logistic. L1-penalized logistic regression with penalty parameter λ (inverse of C) selected via 3-fold cross-validation on 30,000-observation subset, testing $C \in \{0.0001, 0.001, 0.01, 0.1\}$. Optimal $C = 0.1$ selected by AUC. We also implement Post-LASSO by refitting unpenalized logistic

Table 1: Variable Definitions and Economic Rationale

Variable	Economic Rationale	Expected Sign
Market Variables		
EXCESS_RETURN	Negative returns signal deteriorating fundamentals and investor pessimism	−
LOG_MKTCAP	Larger firms have better capital access, diversification, economies of scale	−
VOLUME	Higher volume indicates liquidity and sustained investor interest	−
Profitability		
ROA	Net Income/Assets measures overall profitability and cash generation	−
EBIT/Assets	Operating profitability before financing, indicates core business health	−
Leverage		
LEVERAGE	Total Liabilities/Assets; higher leverage increases default boundary per Merton (1974)	+
DEBT/EQUITY	Financial leverage from creditor perspective; high ratios signal distress risk	+
Liquidity		
CURRENT_RATIO	Current Assets/Current Liabilities; measures short-term solvency	−
NWC/Assets	Net Working Capital/Assets; positive NWC funds operations	−
Altman Components		
RE/Assets	Retained Earnings/Assets; accumulated profitability cushion (Altman 1968)	−
SALES/Assets	Asset turnover efficiency and competitive position	−
Coverage & Valuation		
INT_COVERAGE	EBIT/Interest Expense; ability to service debt obligations	−
MTB	Market-to-Book ratio; market perception of firm value	−

regression on LASSO-selected features.

Model 3: Ridge Logistic. L2-penalized logistic regression with $C \in \{0.01, 0.1, 1.0, 10.0\}$ tested via CV. Optimal $C = 0.01$ selected.

Model 4: K-Nearest Neighbors. Non-parametric classifier using Euclidean distance on standardized features. We test $K \in \{3, 5, 7, 9, 11, 15, 21\}$ on 20,000-observation subset, selecting $K = 21$ by AUC.

Model 5: Random Forest. Bootstrap-aggregated decision trees with hyperparameter grid search over $(n\_estimators, max\_depth) \in \{50, 100, 200\} \times \{5, 10, 15\}$ on 30,000-observation subset. Optimal parameters: $n\_estimators = 200$, $max\_depth = 10$. We use `class_weight='balanced_subsample'` to address class imbalance.

Model 6: Survival Random Forest. Time-to-event model using `scikit-survival` package treating bankruptcy as event and constructing time-to-event from year of observation to year of default (or censoring at 2020). We use $n\_estimators = 50$ based on preliminary testing showing 50–100 yield similar performance.

Model 7: XGBoost. Gradient-boosted trees with $n\_estimators \in \{50, 100, 150, 200\}$ tested, optimal $n\_estimators = 50$. We set `scale_pos_weight` $= 202.17$ to handle class imbalance.

Model 8: LightGBM. Microsoft's gradient boosting framework with $max\_depth \in \{3, 6, 9, 12\}$ tested, optimal $max\_depth = 6$.

All models evaluated on identical test set using misclassification rate, AUC, Kolmogorov-Smirnov statistic,

precision, recall, F1-score, and decile analysis.

4 Empirical Results

4.1 In-Sample Logistic Regression

Table 2 presents in-sample logistic regression results. All coefficients align with *a priori* expectations: LEVERAGE has the largest positive coefficient (1.794), confirming Merton's structural model prediction, while ROA (−1.116) and RE/Assets (−1.016) have large negative coefficients, indicating profitability and accumulated earnings strongly reduce default risk. The model achieves in-sample AUC of 0.899 with 77.2% recall, successfully identifying most defaults at the cost of elevated false positives (misclassification rate 12.5%).

Table 2: Logistic Regression In-Sample Results (1964–1990)

Variable	**Coefficient**	**—Coefficient—**
Intercept	1.048	—
LEVERAGE	**1.794**	1.794
ROA	−1.116	1.116
RE/Assets	−1.016	1.016
EBIT/Assets	−0.736	0.736
EXCESS_RETURN	−0.679	0.679
NWC/Assets	−0.413	0.413
LOG_MKTCAP	−0.238	0.238
CURRENT_RATIO	−0.202	0.202
SALES/Assets	0.089	0.089
DEBT/EQUITY	0.011	0.011
Performance	**Value**	
Observations	88,381	
Defaults	435	
AUC	0.899	
KS Statistic	0.673	
Recall	0.772	

4.2 Out-of-Sample Model Comparison

Table 3 presents comprehensive out-of-sample results for all eight models. Four key findings emerge:

(1) Tree ensemble methods dominate linear models. Survival Random Forest achieves highest AUC (0.919) and KS statistic (0.737), followed closely by Random Forest (AUC = 0.917, KS = 0.719) and XGBoost (AUC = 0.890, KS = 0.677). In contrast, the best linear model (LASSO) achieves AUC = 0.853, KS = 0.643 — a performance gap of 6.6 AUC points.

(2) Regularization improves upon standard logistic regression. LASSO (AUC = 0.853) and Ridge (AUC = 0.852) outperform unregularized logistic regression (AUC = 0.823) by 3 AUC points, demonstrating benefit of shrinkage for out-of-sample prediction. However, LASSO selected all 13 features at optimal $C = 0.1$, suggesting weak regularization. Post-LASSO performs identically to LASSO since no features were eliminated.

(3) KNN fails catastrophically due to class imbalance. Despite achieving lowest raw misclassification rate (0.48%), KNN predicts default for only 1 observation out of 202,070 (recall = 0.1%), rendering it useless for early warning. This illustrates the danger of relying solely on accuracy metrics with severe class imbalance.

(4) Model complexity-performance tradeoff. Survival RF achieves best performance but requires

time-to-event data structure and 15,000-sample stratified training subset for computational feasibility. Random Forest provides nearly identical performance (2 AUC points lower) with simpler implementation, suggesting RF as optimal production model.

Table 3: Out-of-Sample Model Performance Comparison (1991–2020)

Model	Misclass	AUC	KS Stat	Precision	Recall	F1
Linear Models						
Logistic Regression	0.277	0.823	0.589	0.015	**0.862**	0.029
LASSO	0.229	0.853	0.643	0.018	0.861	0.035
Post-LASSO	0.228	0.849	0.635	0.018	0.853	0.035
Ridge	0.226	0.852	0.639	0.018	0.855	0.035
Non-Parametric						
KNN ($K = 21$)	**0.005**	0.694	0.373	1.000	0.001	0.002
Tree Ensemble Methods						
Random Forest	0.028	0.917	0.719	0.098	0.585	**0.168**
Survival RF	0.049	**0.919**	**0.737**	0.062	0.640	0.113
XGBoost	0.045	0.890	0.677	0.069	0.658	0.125
LightGBM	0.227	0.768	0.603	0.017	0.825	0.034

Test set: 202,070 observations, 973 defaults (0.48% base rate)

4.3 Decile Analysis and Rank Ordering

We rank test set observations by predicted default probability and partition into deciles. Table 4 presents results for top-3 performing models. All three successfully concentrate defaults in top decile: Random Forest identifies 79.8% of defaults in top decile (default rate 3.84%), Survival RF captures 82.3% (default rate 3.96%), and Logistic Regression captures 38.2% (default rate 1.84%). This rank-ordering ability is critical for portfolio credit risk management, enabling concentration of monitoring resources on highest-risk firms.

Table 4: Decile Analysis for Top Models (Out-of-Sample)

	Random Forest		Survival RF		Logistic Reg	
Decile	Defaults	Rate (%)	Defaults	Rate (%)	Defaults	Rate (%)
1 (Lowest Risk)	8	0.04	10	0.05	12	0.06
2	8	0.04	8	0.04	17	0.08
3	9	0.04	8	0.03	11	0.05
4	8	0.04	7	0.04	11	0.05
5	5	0.02	2	0.01	17	0.08
6	18	0.09	12	0.06	18	0.09
7	31	0.15	22	0.11	30	0.15
8	40	0.20	23	0.11	151	0.75
9	70	0.35	80	0.40	334	1.65
10 (Highest Risk)	**776**	**3.84**	**801**	**3.96**	**372**	**1.84**
Top Decile Capture	79.8%		82.3%		38.2%	

4.4 Feature Importance Analysis

Figure 5 presents variable importance rankings from the three tree-based models. Remarkable consistency emerges: ROA (return on assets) dominates across all three methods, accounting for 19.1% (RF), 42.5% (XGBoost), and highest importance (385) in LightGBM. This confirms profitability as the single most powerful default predictor. Interest coverage and EBIT/Assets rank 2nd and 3rd in Random Forest, highlighting debt servicing ability and operating profitability. Market-to-book ratio emerges as 4th most important in RF and 2nd in XGBoost, suggesting market valuation contains incremental information beyond accounting fundamentals. Leverage ranks 5th despite having largest logistic regression coefficient, indicating nonlinear interactions captured by tree methods reduce its standalone importance.

Table 5: Feature Importance Rankings (Top 5 Variables)

Random Forest			XGBoost			LightGBM		
Rank	**Variable**	**Imp.**	**Rank**	**Variable**	**Imp.**	**Rank**	**Variable**	**Imp.**
1	ROA	19.1%	1	ROA	42.5%	1	ROA	385
2	INT_COV	13.8%	2	MTB	8.7%	2	EBIT/Assets	271
3	EBIT/Assets	10.4%	3	LEVERAGE	6.8%	3	MTB	267
4	MTB	9.2%	4	INT_COV	6.1%	4	RE/Assets	232
5	LEVERAGE	9.1%	5	EXCESS_RET	5.8%	5	LOG_MKTCAP	220

5 Conclusions and Recommendations

This analysis demonstrates that modern machine learning methods significantly outperform traditional logistic regression for corporate default prediction. Our findings support recent academic research (Agarwal & Zhang 2022; Alanis, Chava & Shah 2022) showing tree-ensemble methods capture nonlinear relationships and variable interactions missed by linear models. The 9.6 AUC point improvement (Survival RF: 0.919 vs. Logistic: 0.823) translates to substantial economic value in credit portfolio management.

Key Recommendations:

(1) Deploy Random Forest for production. Random Forest achieves 91.7% AUC with interpretable feature importances, simpler implementation than Survival RF, and robust performance across evaluation metrics. The model correctly identifies 58.5% of defaults (recall) while maintaining 9.8% precision — an acceptable tradeoff for early warning systems prioritizing sensitivity over specificity.

(2) Focus monitoring on ROA, interest coverage, and operating profitability. These three variables collectively explain 43.3% of Random Forest's predictive power. Firms with declining ROA, deteriorating interest coverage, or negative EBIT/Assets warrant immediate attention regardless of other fundamentals.

(3) Avoid KNN for imbalanced classification. KNN's catastrophic failure (0.1% recall) despite 99.5% accuracy demonstrates that distance-based methods cannot handle severe class imbalance without extensive resampling — a costly preprocessing step tree methods avoid.

(4) Implement quarterly retraining. The 26–30 year out-of-sample period (1991–2020) tests model stability but production systems should retrain quarterly on expanding windows to capture evolving default determinants, especially during regime changes (e.g., 2008 financial crisis, COVID-19 pandemic).

(5) Combine with market-based measures. While accounting variables dominate importance rankings, market-to-book ratio and excess returns contribute 14.3% of RF importance. A two-stage system using high-frequency market signals to trigger deep fundamental analysis could optimize early detection.

Future Extensions: This analysis could be extended by (i) implementing rolling-window forecasts to simulate real-time deployment, (ii) incorporating macroeconomic variables (GDP growth, credit spreads, VIX) to capture systematic risk, (iii) testing ensemble stacking combining multiple model predictions, and (iv) conducting cost-benefit analysis weighting Type I vs. Type II errors by economic losses.