Final Project: Bankruptcy Prediction & Investment Optimization

DS| Team 34: Jiayi Yu; Nina Zhuang; Xin'ai Hao; Doan Le; Yanze Liu; Yang Xiao

Business Understanding

As a team of business analysts in private equity, our project addresses a fundamental challenge in investment management — the risk of unexpected corporate bankruptcies. Bankruptcy not only leads to economic losses and social disruption but also represents a major source of financial risk for institutional investors such as mutual funds and pension funds, where a single default can translate into millions in losses. Yet, being overly conservative and avoiding all risky firms may also mean missing valuable investment opportunities.

Our core business question is thus: How do the techniques of data mining predict risk of bankruptcy and turn such forecasts into practical investment thresholds that fit different risk tolerances? Investors vary in risk toleration — conservative mutual funds look for safety, whereas speculative hedge funds take higher exposures in hopes of higher returns. A generic solution is thus inapplicable.

To solve this, we develop machine learning models that predict the probability of firm bankruptcy based on financial statement data and embed them into a cost–benefit framework that mirrors real-world payoffs. This framework connects prediction outcomes (true positives, false negatives, etc.) with financial consequences, converting statistical results into meaningful investment insights. By adjusting decision thresholds, investors can customize portfolio strategies ranging from conservative to speculative.

Ultimately, this model provides a systematic, data-driven method for screening firms, reducing exposure to high-risk companies while preserving growth potential. Beyond investor decision-making, such a predictive system can also serve as an early-warning

mechanism for regulators and financial institutions, demonstrating how data science can bridge the gap between statistical prediction and actionable financial strategy.

Data Understanding

We use the American Companies Bankruptcy Prediction Dataset from Kaggle, which covers the years 1999–2018. It contains firm-level financial data suitable for bankruptcy prediction, with each observation representing one firm in one year. The outcome label shows whether the firm survived or failed. The dataset has about 78,682 observations from roughly 9,000 firms, with only 6.6 percent labeled as bankrupt. This imbalance reflects reality—defaults are rare—but also makes it more difficult to build an interpretable predictive model.

The dataset includes a range of financial indicators linked to company health. Liquidity measures such as current assets and liabilities show whether a firm can pay its short-term bills. Leverage indicators like debt-to-equity capture exposure to long-term debt. Profitability metrics such as EBITDA and net income measure the ability to generate earnings, while size and valuation measures like total assets and market value provide context on scale. Together, these categories of variables cover liquidity, leverage, profitability, and valuation, giving a comprehensive picture of corporate stability and risk.

Still, there are limitations. Firms that failed before 1999 are missing, which introduces survivorship bias. The 2008 financial crisis represents a structural shock, and patterns from that period may not generalize to more stable years. Because the dataset combines all industries, sector-specific bankruptcy signals may be muted. And over nearly two decades, evolving accounting standards may affect comparability.

To improve reliability, we apply preprocessing steps. Missing values are imputed, variables are normalized to allow comparisons across firms of different sizes, and SMOTE resampling is used to address the imbalance between bankrupt and non-bankrupt cases. Resampling helps ensure the model does not ignore rare bankruptcy events, while normalization keeps financial ratios consistent across firms of different scales. These adjustments make the dataset more suitable for modeling. Despite its imperfections, it remains a strong foundation for both descriptive analysis and predictive modeling, supporting our goal of building bankruptcy forecasts that can guide investors in practice.

Data Preparation

Before building different models and measuring their individual performances, we cleaned and transformed some variables in the dataset to reduce bias and inaccuracies in our models.

Integration and Cleaning

We first generated the correlation table for all the independent variables, and removed redundant or collinear features - Net sales (X9), EBIT (X12), Gross Profit (X13), and Total Operating Expenses (X18) - to prevent distorted coefficient estimated and improve model stability.

Since variables like net income and total assets are highly related to the size of the firm, which would incur difficulty when building a predictive for all firms, we engineered some new ratio-based variables to capture normalized firm-level financial health:

- Net Profit Margin = Net Income / Total Revenue
- Return on Assets (ROA) = Net Income / Total Assets

- Current Ratio = Current Assets / Current Liabilities
- Quick Ratio = (Current Assets – Inventory) / Current Liabilities
- Debt-to-Asset Ratio = Total Liabilities / Total Assets

Transformation and Encoding

We applied transformation to the target variable, turning status_label from strings to binary (0 = alive, 1 = failed). We also standardized all numeric features using StandardScaler, as well as applying SMOTE (Synthetic Minority Oversampling Technique) to correct class imbalance before training, as 93% of the firms in the dataset are labeled as alive, compared to 7% as failed.

Dataset Partitioning

After all the cleaning and transformation, we split the dataset into the train set and the test set by 80/20. The split is stratified by class proportion to preserve data distribution integrity.

Modeling

Models Evaluated

We compared four models in this study: Random Forest, XGBoost, Logistic Regression with L2 regularization, Support Vector Machines and Lasso Logistic Regression with L1 regularization. The Random Forest model was used as a basic non-linear option that’s quite good at handling correlated variables, while XGBoost was included because it’s a strong performer for structured, tabular data.

Algorithm Selection Rationale

Each model used in this study has distinct advantages and limitations. The Random Forest model effectively captures nonlinear relationships and demonstrates high robustness and stability, making it well-suited for complex financial data; however, it offers limited interpretability and can be computationally slower when applied to large datasets. XGBoost provides strong predictive performance and flexible parameter tuning, allowing it to adapt well to diverse data structures, yet its complexity increases the risk of overfitting and makes parameter optimization more challenging. Logistic Regression serves as a simple and interpretable baseline model, providing clear insights and comparability across experiments, but it lacks the expressive capacity to model nonlinear interactions present in financial variables. Support Vector Machines (SVM) perform well in high-dimensional feature spaces and excel when clear decision boundaries exist, though they are computationally intensive, less interpretable, and highly sensitive to parameter selection. Overall, these models complement each other, balancing predictive power, interpretability, and computational efficiency.

Evaluation

Model Performance Metrics

The evaluation of our data mining results focuses primarily on economic utility rather than conventional statistical accuracy.

Since the purpose of our model is to improve investment decisions, we assess performance by examining how the predicted bankruptcy probabilities translate into expected financial returns across different investor risk profiles.

Instead of relying on metrics such as accuracy or F1-score, which do not reflect the economic consequences of decision errors, we evaluate the model's performance through a cost–benefit matrix that assigns explicit monetary gains or losses to each prediction outcome—true positives, false positives, true negatives, and false negatives.

Conservative Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest ($\hat{y} = 1$)	0 (TP)	−0.05 (FP)
Invest ($\hat{y} = 0$)	−1.50 (FN)	+0.10 (TN)

Neutral Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest ($\hat{y} = 1$)	0 (TP)	−0.10 (FP)
Invest ($\hat{y} = 0$)	−1.00 (FN)	+0.10 (TN)

Aggressive Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest ($\hat{y} = 1$)	0 (TP)	−0.20 (FP)
Invest ($\hat{y} = 0$)	−0.80 (FN)	+0.10 (TN)

Speculative Investor

Investment Decision	Bankrupt (1)	Not Bankrupt (0)
Do Not Invest (ŷ = 1)	0 (TP)	−0.30 (FP)
Invest (ŷ = 0)	−0.50 (FN)	+0.10 (TN)

This approach allows us to directly measure how the model's predictions affect portfolio profitability, thereby connecting machine learning outputs to tangible financial outcomes.

During analysis, the model was subjected to four different styles of investment—Conservative, Neutral, Aggressive, and Speculative—each signifying some trade-off between risk-bearing capacity and expected return.Each style's optimum decision limit was then found by maximizing the expected return of the portfolio based on its given cost–benefit schedule.

Among various styles of investment, the models indicate obvious differences in their performance. The Random Forest model provides the most stable and realistic results with optimal values of 0.37, 0.45, 0.52, and 0.66 in the case of Conservative, Neutral, Aggressive, and Speculative investment styles and the corresponding expected returns are 0.041, 0.050, 0.052, and 0.064. It is the most stable, precise, andinterpretable model. XGBoost indicates slightly higher returns of 0.039 to 0.064 but its large optimal values render it less realistic for practical implementation. Logistic Regression (L2) and SVM are interpretable baseline models, though both do poorly in the case of nonlinear relationships and in the calibration of probabilities.

The Random Forest model shows the strongest economic improvement among conservative investors by successfully lowering bankruptcy exposure with minimal loss of profit potential. With more aggressive or speculative styles, the returns improve with excess risk exposure, which points out the implication that the economic performance of the model hinges on investors' risk preferences. This analysis framework provides a practical, business-minded view by converting statistical forecasts into risk-adjusted financial returns.

To create the business case for the model, we contrasted Random Forest's expected return curves with naive baselines like "Always Invest" and "Always Avoid." The model beats the latter consistently, particularly for risk-averse investors, with 3–4% improvement in expected portfolio return per company investment. This enhancement is used as the proxy for the ROI of data-driven decisions. Calculating precise ROI is difficult with simplified cost–benefit assumptions and the unavailability of real transaction data. The framework principally captures theoretical return under idealized market conditions without adjustments for transaction costs, liquidity, or diversification. Thus, relative improvement metrics like changes in expected return or downside risk reduction become more realistic indicators of value. Nevertheless, the simplified cost–benefit construction makes the model unreal; it is best highlighted for conservative investors with returns for more aggressive styles potentially overestimated because the latter do not consider drawdown and volatility risk.

Deployment

The bankruptcy prediction model developed in this project can help investment firms make decisions. It estimates each company's bankruptcy probability, ranging from 0 to 1, based on financial statements. It uses an optimal threshold that reflects investor risk preferences through a cost-benefit framework. Companies that exceed this threshold are identified as high-risk, which means they may be excluded or shorted. Others can still be considered for investment. This approach relies on data and economic principles to guide capital allocation. It reduces personal bias in investment choices. The model needs to be retrained regularly as new data comes in to stay stable and respond to changing market conditions.

Before using the model, firms should tackle potential issues such as model drift, data quality, and interpretability. The connections between financial indicators and default risk shift over time, making retraining necessary. Additionally, financial reports differ in their timing and consistency, so proper data validation and normalization are required. While Random Forest gives more transparency than deep learning, its opaque nature still requires clear tools like feature importance or SHAP analysis for regulatory accountability.

From both ethical and practical angles, the model should support—not replace—human judgment. Too much reliance on automation can blur lines of responsibility. Data bias may unfairly impact smaller or less transparent firms. If many firms adopt this model, it could increase market volatility due to herding behavior, stressing the importance of strong governance. Though the model has strong interpretability and clear benefits for conservative investors, it is limited by its focus on structured data and simplified cost-benefit assumptions. Its effectiveness mainly shines in conservative strategies. Results

for aggressive or speculative approaches may exaggerate potential gains while ignoring maximum drawdown and risk-adjusted returns.

To make the model more useful in practice, future work should involve regular governance and recalibration. It should integrate additional data sources like credit ratings or market sentiment. Keeping human oversight in the process and being transparent about methods, performance, and limitations will be essential for building trust, ensuring compliance, and promoting sustainable use in investment practices.

Appendix

Team Member	Contribution
Jiayi Yu	Led data preprocessing, feature engineering, and model development; contributed to KNN analysis.
Yanze Liu	Focused on model evaluation and threshold optimization; analyzed investment-style performance under the cost–benefit framework.
Yang Xiao	Conducted exploratory data analysis and statistical validation; contributed to Lasso analysis.
Nina Zhuang	Responsible for slides design and visualization; summarized analytical findings into clear business insights.
Doan Le	Drafted and refined report content; organized project documentation and ensured consistency in writing style.
Xin’ai Hao	Handled the data modeling section in the report; translated technical results into business language.