

Predicting Company Bankruptcy

Parin Shaik, Yue Gao, Fares Sakaan,
Sherif Mohasseb






Goals & Purpose



Goal: Predict company bankruptcy based on financial indicators using machine learning models.

Importance: Build models to identify whether a company is at risk of bankruptcy

Predictors (18): current assets, cost of goods sold, depreciation and amortization, EBITDA, inventory, net income, total receivables, market value, net sales, total assets, total long-term debt, EBIT, gross profit, total current liabilities, retained earnings, total revenue, total liabilities, and total operating expenses



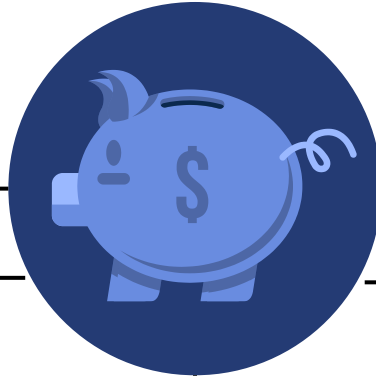
Preliminary Information

**78,682 rows,
21 columns.**

The columns are comprised of both quantitative and qualitative data.

**Average Yearly
Income**

The average company has a net income of \$129k a year.



**Initial Ratio of Alive
Companies vs. Failed
Companies**

93% Alive
7% Failed

**Outcome
Variable**

Binary Outcome:
'Alive' or 'Failed'

**Most Important
Predictors
(from Lasso)**

Cost of Goods Sold, Depreciation & Amortization, Inventory, Net Income, Total Receivables, Market Value, Total Assets, Gross Profits, Total Current Liabilities, Retained Earnings, Total Liabilities

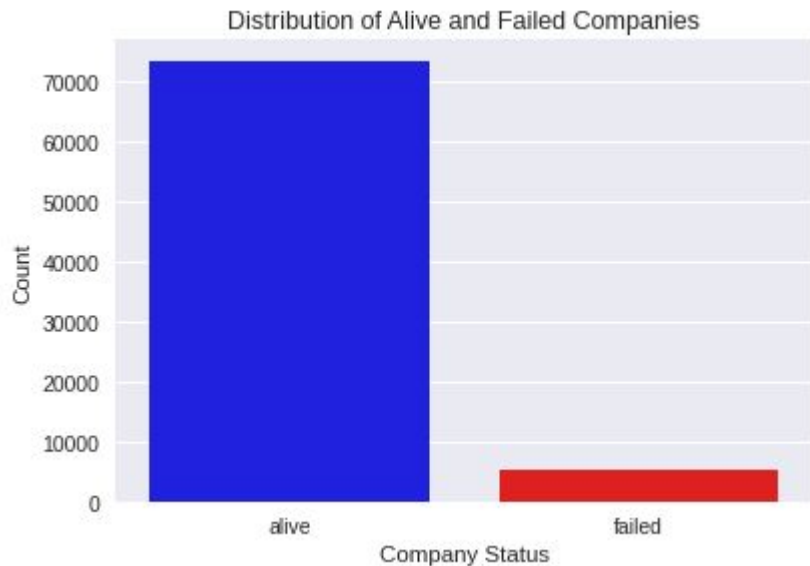
Methodology Overview

1. Data Cleaning

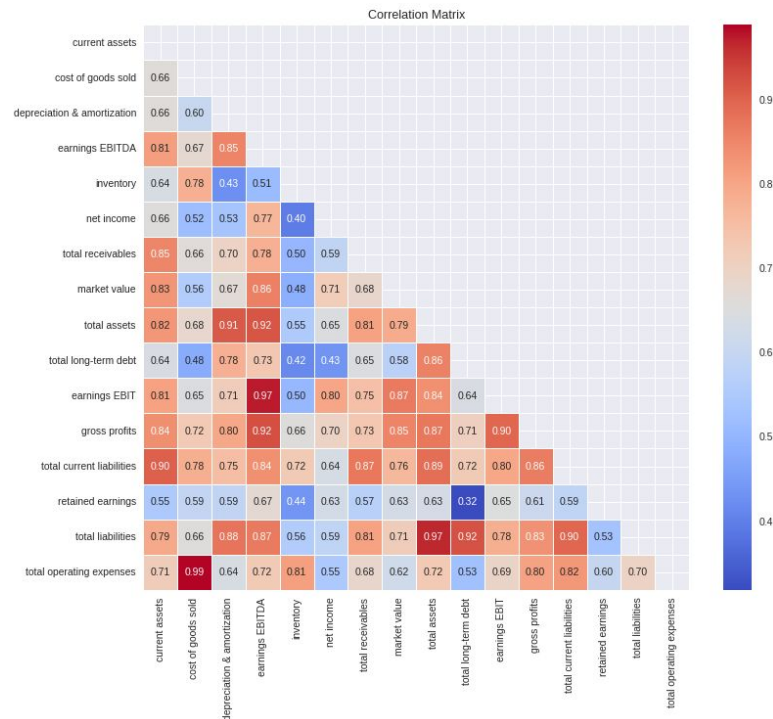
- Resampling
- Feature Deletion
- Compute multicollinearity to further eliminate features via Logistic Regression

2. Predictive Models used: Logistic Regression, Ridge, Lasso, Elastic Net, kNN, Random Forest (Bagging), Classification Tree, XGBoosting

Our Dataset

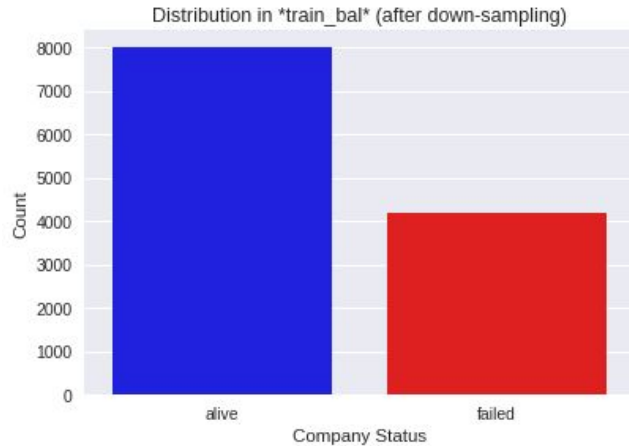


Bar Chart for Outcome Variable



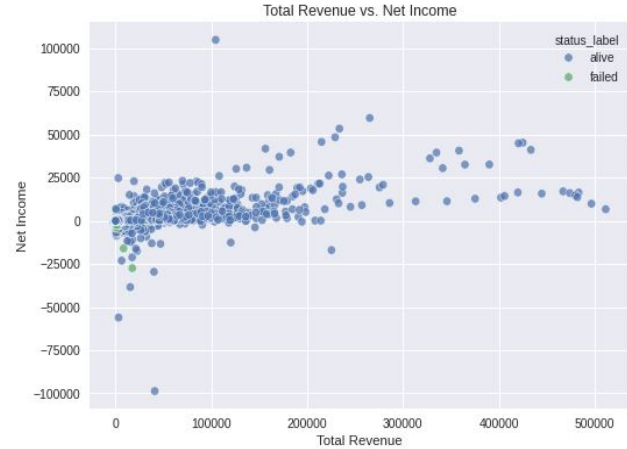
Feature Correlation Matrix

Data Cleaning Methodology



Downsampling

Because the original dataset is highly imbalanced (with much more 'alive' than 'failed'), we split the data into training and test (80/20) then downsampled 'Alive' in the training set to 8000 rows.

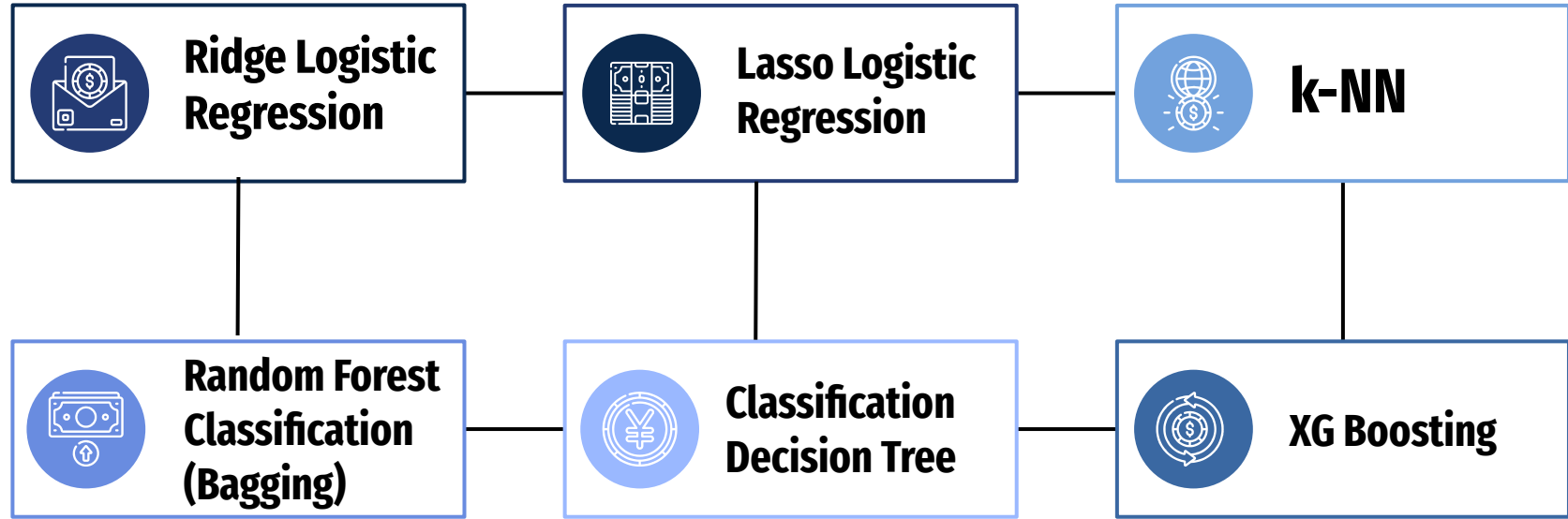


Dropped 2 Features

Earnings before Tax + Total Operating Expenses
= Net Sales = Total Revenue

Dropping these predictors allowed for a better focus on more relevant features, which led to improved model performance.

Classification Model Implementations



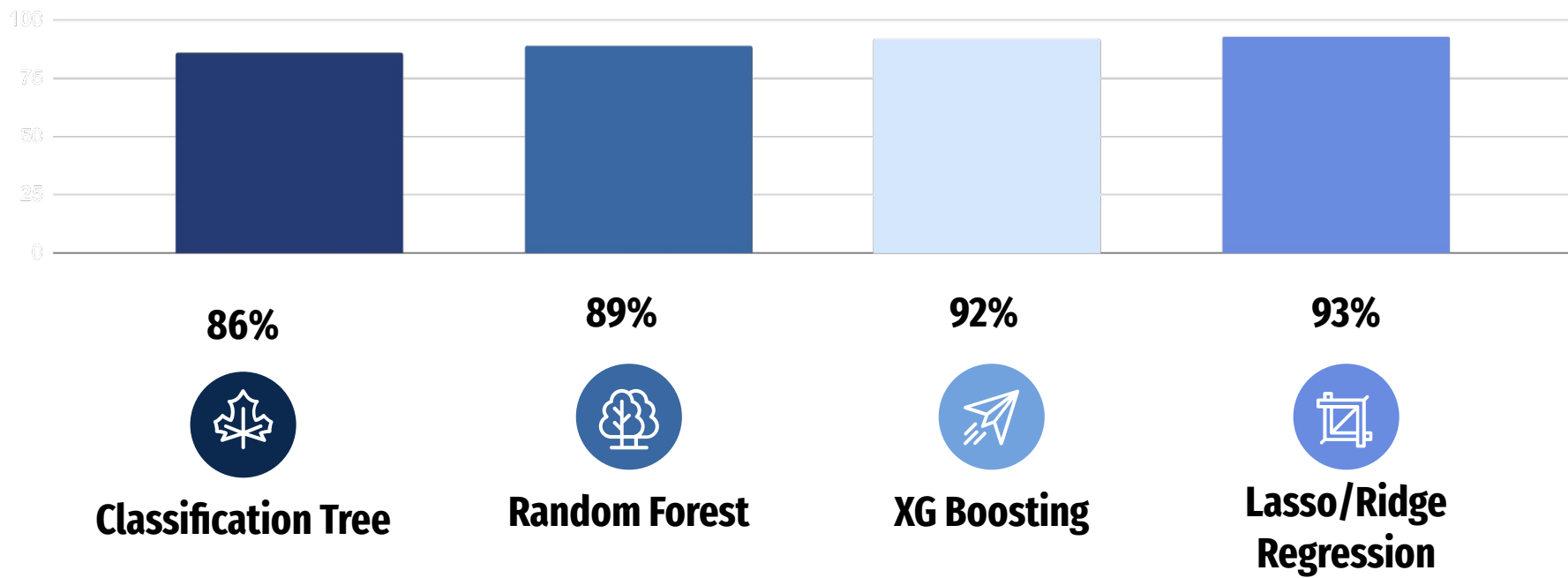
Comparison Criteria



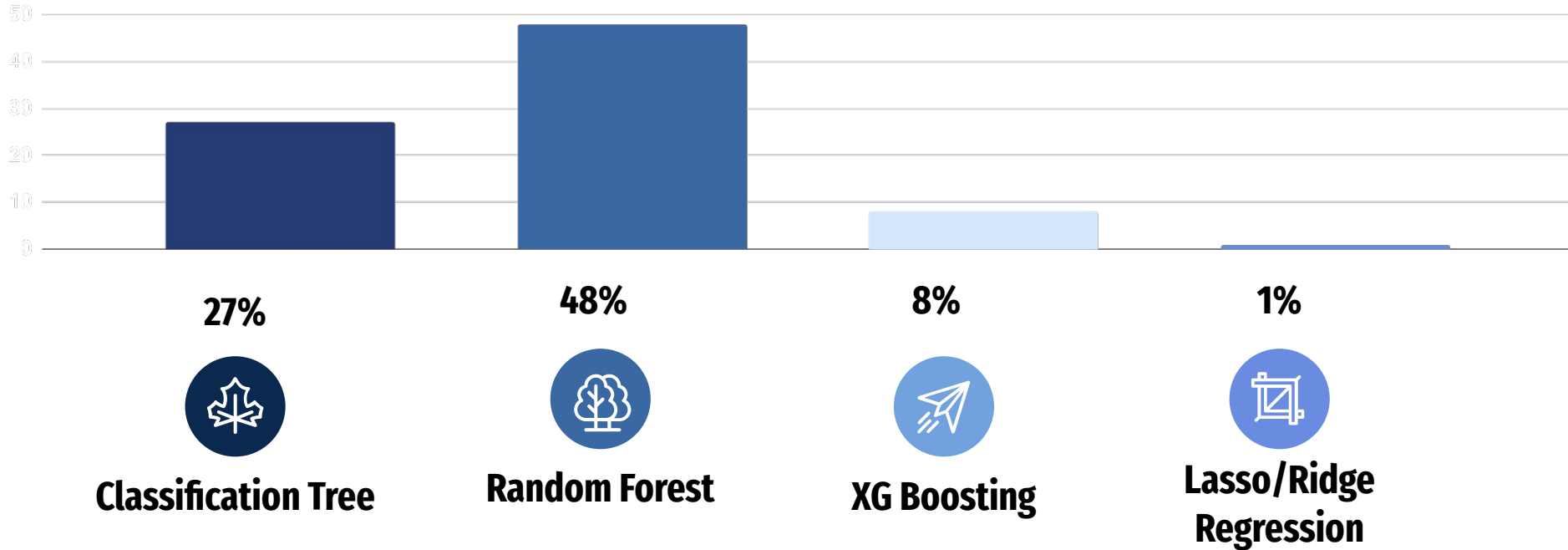
Goal: Choose the model that balances high overall accuracy with strong ability to catch “failed” firms while keeping false alarms tolerable.

- Train models on arbitrarily split, downscaled data
- Tune hyperparameters via Cross-Validation to optimize performance
- Make predictions with testing data, calculate accuracy and run classification report (precision, recall, F-1 score)
- Compare accuracy, as well as how it compares to recall and precision ratios
- Determine optimal model

Comparing Top Models by their Accuracies



Comparing Top Models by their Failure Predictions



Confusion Matrix

Models	Ridge Regression	Lasso Regression	kNN	Random Forest	Classification Tree	XG Boosting
Accuracy	0.93	0.93	0.79	0.89	0.86	0.92
Precision	0.57	0.56	0.59	0.70	0.55	0.59
Recall	0.51	0.51	0.76	0.63	0.59	0.53
F1-Score	0.49	0.49	0.60	0.655	0.57	0.54

Challenges

- Unequal reporting of failed companies to alive (prior to downsampling, ratio of alive to failed companies is 14:1)
- Finding the best balance between accuracy and the ability to find failures
- Downsampling to a specific number
- After downsampling, the ratio scaled down to 1.9.
- Accuracy Discrepancies due to imbalance
- Multicollinearity yielded high VIF for few features



Takeaways



- Highest Overall Accuracy: Random Forest achieved 89% test accuracy, among the best across all models.
- Strong Minority Class Detection: 48% recall and 30% precision for the critical "failed" class.
- Ensemble learning reduced overfitting and class weighting improved failure detection.
- In biased datasets, under- and oversampling are crucial to ensure fair and effective model training.
- Without proper sampling, high overall accuracy can be misleading, masking poor performance on critical groups.
- Good sampling strategies lead to models that generalize better and provide more reliable, actionable predictions.

Thank You!
Any Questions?

