Decision Tree Ensemble Classifiers in Predicting Corporate Bankruptcy

PRINCETON UNIVERSITY

Jay Lee

Department of Computer Science

Motivation

- Corporate bankruptcies can severely damage the economy, like in 2007-2008. Being able to forecast such events may help counteract and prevent them.
- Accurate forecasts of bankruptcy will also impact the credit scoring and lending industries. Better predictions can lead to fewer defaults, making the industry more stable and efficient.

Prior Approaches

- Problem first proposed by Edward Altman (1968). Altman used discriminant analysis on financial ratios to predict bankruptcy [1].
- In 1980, Ohlson proposes a logistic regression model to predict probabilities of firm insolvency, with applications in lending and credit scoring [2].
- Many different machine learning models show significantly improved accuracy compared to conventional models, such as SVM [3] and Artificial Neural Nets [4].
- More recently, using a small dataset in 2017, Barboza et. Al. showed that Ensemble methods such as bagging and random forest can reach accuracies of up to 87% [5].

Goal

- Explore the predictive capabilities of tree-based ensembles models in forecasting corporate bankruptcy

Data

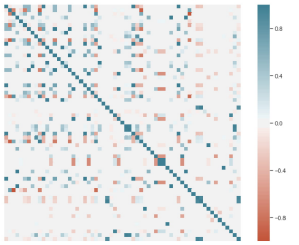
- We study 15702 different Polish companies and their bankruptcy status between 2000 and 2013, generously provided by the UCI Machine Learning Repository.
- The data has 15702 rows, 64 features that represent financial ratios, and whether or not the firm went bankrupt within two years since data collection.
- Missing data was imputed by mean, and dataset was split into 75/25 train/test to validate the model.
- In both train/test sets, bankrupt firms were very rare (about 5%). To prevent bias, the training set was oversampled using Synthetic Minority Over-Sampling Technique to reach 50/50.

Abstract

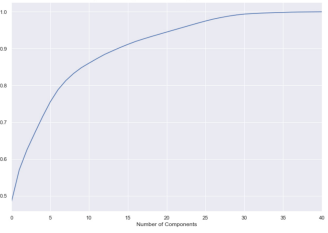
- Successfully predicting corporate bankruptcy can provide significant social and economic benefits to both consumers and lenders alike.
- The search for the "ideal" predictor started in 1968, and today, hundreds of papers exist on the topic. Recently, the most successful methods include SVM and artificial neural networks using back propagation.
- This paper explores tree-based ensemble methods Random Forest and Gradient Boosting. Both methods demonstrated significantly higher AUC and accuracy compared to traditional machine learning methods.
- Applying PCA to the raw dataset made performance noticeably worse.

Data Exploration

Correlation Heatmap of Input Variables



Percentage of feature variance explained by number of principal components



Models

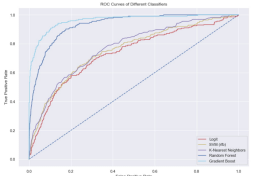
- Task: For each firm, given 64 input features, output either 0 (firm will not go bankrupt within 2 years) or 1 (firm will go bankrupt).
- Each classification was performed both with and without PCA (n=35) applied.
- The following models were used for the classification task:
  - Logistic Regression
  - SVM
    - RBF Kernel
  - K-Nearest Neighbors
    - Used $K = \sqrt{n} = 147$
  - Random Forest
  - Gradient Boosting
    - Used a decision tree as the learner function
    - Applied via the XGBoost library

Hyperparameter Tuning

- To maximize accuracy, we tuned the hyperparameters of the Random Forest and Gradient Boosting models using 3-fold Cross Validation.
- To save time, a randomized grid search was first used to narrow down the range, then an exhaustive grid search was performed to choose the exact parameters.
- This was done for both models, with and without PCA.

Results

- In our problem, correctly identifying the *bankrupt* firms is more important than correctly identifying the *healthy* firms.
- Thus, in general, we care more about recall: what proportion of the true 1's did we select?



Model	Precision (%)		Recall (%)		AUROC (%)		Accuracy (%)	
	No PCA	PCA	No PCA	PCA	No PCA	PCA	No PCA	PCA
Logistic Regression	66.5179	62.9464	13.0931	11.8887	73.1752	72.1609	72.8986	71.2685
SVM (rfb)	63.8393	63.3929	14.1304	12.4234	75.5010	73.5513	75.8023	72.4147
K-Nearest Neighbors	77.6786	78.5714	11.3874	11.3402	77.2050	76.8594	64.2384	63.7290
Random Forest	51.7857	37.9464	50.2165	35.2697	93.1984	85.6000	94.3199	92.4860
Gradient Boosting	67.4107	38.8393	75.1244	37.3391	96.2829	86.8251	96.8670	92.7916

- Gradient Boosting performed the best in all categories. It had over 75% recall and over 96% AUROC.
- Random Forest also performed well: 93% AUC, but 50% recall.
- PCA made almost every single measure worse. It decreased the AUROC of Gradient Boosting and Random Forest by almost 10%.

Discussion and Future Work

- Tree Ensemble Models significantly outperformed traditional models in every measure. However, recall still not ideal.
- PCA made every measure worse.
  - Although many of the features are highly correlated with each other, they might not be *linear*.
  - Studying each ratio in-depth to reason independence and correlation could help improve this problem.
- However, PCA significantly improved training time.
  - Random Forest improved 20% (115 min → 91.6 min)
  - Gradient Boosting improved 35% (58 min → 37.6 min)
- In future work, would like to look focus on feature selection/extraction to decrease dimension (extra trees?).
- Include macro-variables (GDP growth, inflation, etc.) as well.

References

1. Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 1968.
2. James A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109-131, 1980
3. Jae H. min and Young-Chan Lee. Bankruptcy prediction using support vector machines with optimal choice of kernel function parameters. *Expert Systems with Applications,* 28(4):603-614, 2005.
4. P. P. M. Pompe and A. J. Feelders. Using machine learning, neural networks, and statistics to predict corporate bankruptcy. *Computer-Aided Civil and Infrastructure Engineering,* 12(4):267-276.
5. Flavio Barboza, Herbert Kimura, and Edward Altman. Machine learning models and bankruptcy prediction. *Expert Systems with Applications,* 83:405-417, 2017.