Proposal for NBER Workshop on Big Data

"Predicting Corporate Bankruptcy: A Machine Learning Approach"[1]

James F. Albertus[2]

April 5, 2019

This proposal describes my paper studying the prediction of corporate bankruptcy using big data and machine learning techniques. The paper makes two core contributions. First, from a set of nearly 20,000 bankruptcy predictors derived from the literature, it identifies the subset of these that offer new predictive information.[3] Second, it generates a novel, more accurate, firm-year level panel of corporate bankruptcy scores, similar to those widely used in the finance literature.

In addition to academics, predicting corporate bankruptcy is an important topic of research for regulatory institutions and industry participants. For example, as evidenced in their post-financial crisis stress test exercises, the Federal Reserve System in keenly focused on anticipating corporate financial distress. Moreover, my research draws on (in the form of both proposed bankruptcy predictors and otherwise) work by the largest credit rating agencies, such as Moody's. Corporate bankruptcy prediction is relevant to a broad audience and has important social implications.

Beaver (1966) is the first paper to propose a set of variables to predict corporate bankruptcy. Since these first predictors were formulated, numerous others have appeared in the literature. This has resulted in a dilemma similar to that in the factor-based asset pricing literature, as discussed by Cochrane (2011). Specifically, it is often unclear whether the improvements offered by newly proposed predictor variables can be replicated with combinations of predictor variables already known in the existing literature.

Drawing on the literature, I assemble a firm-year level panel of bankruptcy predictors using CRSP-Compustat data over the period 1980-2017. I construct an indicator variable for corporate bankruptcy from the UCLA-Lopucki bankruptcy database. Consequently, the machine learning algorithms I consider perform a supervised binary classification task.

In my paper, I implement a set of five standard machine learning algorithms to identify the features with the greatest predictive power. While I identify 65 features from the existing literature, I analyze a much larger set. First, as the literature considers the dynamics of features' predictive power (Altman, 1968), I also include in the analysis each feature's first

[1]This research relies on the Pittsburgh Supercomputer's XSEDE. The paper is available upon request.

[2]Carnegie Mellon University, Tepper School of Business; Email: albertus@cmu.edu.

[3]The predictors considered include lags and interactions, as discussed subsequently.

and second lag. Second, to allow for interactions and nonlinearities potentially overlooked in existing work, I include interactions among the features and their lags up to order two. This expands the set of features dramatically. In order to implement the machine learning algorithms (particularly the neural network), I rely on the XSEDE available on the Pittsburgh Supercomputer.

The analysis yields three main findings. First, unregularized logistic regression – a leading conventional methodology for bankruptcy prediction – is outperformed on all metrics considered, including precision, recall, F_1-score, and AUC ROC. Machine learning techniques indeed improve predictive accuracy in the context of corporate bankruptcy.

Second, among the five algorithms I consider – ridge regression, LASSO regression, decision trees, random forests, and neural networks – the neural network and LASSO regression perform best on a held-out test set, as reflected in their F_1-scores. I obtain these results after tuning each algorithm's hyperparameters on a validation set. Ensemble methods that combine the algorithms' predictions offer even better performance.

Third, I find that many of the variables thought to predict bankruptcy do not provide marginal predictive information when orthogonalized against other predictors. Features measured in the year preceding bankruptcy tend to provide more information than features measured with greater lags. However, the dynamics of some features contain valuable information, consistent with the existing literature.

An additional result from my analysis is an improved firm-year level panel of corporate bankruptcy probabilities. These probabilities may find application in many settings, including, for example, academic studies of the association between bankruptcy risk and asset returns.

Together, my paper's results demonstrate the potential of high-performance computing to avoid the mistaken discovery of new corporate bankruptcy predictors, while at the same time improving predictive accuracy. Taken broadly, my study exemplifies the potential of big data and machine learning techniques to advance knowledge in financial economics.

References

Altman, E. I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23, 589–609.

Beaver, W. H., 1966. Financial ratios as predictors of failure. Journal of Accounting Research pp. 71–111.

Cochrane, J. H., 2011. Presidential address: Discount rates. The Journal of Finance 66, 1047–1108.