Machine Learning for Bank Failure Early Warning Systems

Researchers: Katja Mathesius, Jaehyok Choi, Jacob Danner, Blythe Kelly, Gonzalo Valdenebro

Academic Mentors: Dr. Eric Manley, Dr. Sean Severe

Background

The global financial crisis from 2007 to 2008 was the worst economic crash since the great depression and was largely sparked by a sharp increase in bank failures. While there are measures for countering bank failures, in order to utilize them effectively one must be able to predict which banks are likely to fail. Due to the large amount of data collected in the field of banking, machine learning's strong predictive abilities has great potential to help address this problem.

Purpose

This research reveals if there is an accurate way to use machine learning as a means of predicting if a bank is going to fail based on previous bank health data. If the success of this technique can be proven, the model will be able to assist bank examiners in allocating resources to avert bank failures.

Dataset

The complete data set includes 5,461 attributes.
Ongoing work on the data includes:

- Storing and cleaning the data set.
- Examining forms and reaching out to organizations in charge of collecting the data to gain a more meaningful understanding of data included

Once the cleaning and storing has been completed this dataset will be made available for public access.

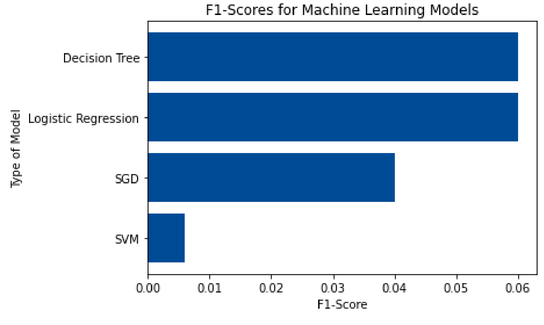
Variable	Definition
FDIC Fail	Official FDIC Failure
Wheelock and Wilson (2000)	(Equity-Goodwill)/Assets<0.02
CAPITAL	Equity/Assets
REALLOAN	Real Estate Loans/Assets
CILOAN	C&I Loans/Assets
CHARGEOFF	Net Chargeoffs/Assets
NONPERFORM	Nonperforming Assets/Assets
EXPENSE	Noninterest Expense/(Net Interest Income + Noninterest Income)
ROA	Net Income/Assets
LIQ	(Fed Funds Purchased - Fed Funds Sold)/Assets
SIZE	Ln(Assets)
BRDEP	Brokered Deposits/Assets
FEDCHART	Dummy Variable if Bank is Chartered by Federal Reserve
BHC	Dummy Variable if Bank is Part of a Bank Holding Company
BRANCH	Branches per 1,000 People (All Banks), Weighted by Deposits by County
BANKS	Total Branches in Operation
UNRATE	Unemployment Rate Weighted by Deposits by County
LABFORCE	Labor Force Participation Rate Weighted by Deposits by County
PCINCOME	Per Capita Income Weighted by Deposits by County (1000s)
POPULATION	Population Weighted by Deposits by County 1(000s)

Data currently cleaned for use with machine learning contains 27 unique attributes and 158,031 unique observations. Data is primary drawn from quarterly FDIC reports.

Each observation consists of a bank's unique attributes in a quarter. The Y variable for the data is a binary measure that is true if the bank failed within the next 6 months and false otherwise. The X variables contain information about the bank during that quarter such as number of assets, number of branches. It also includes information about the area the banks are located in such as the surrounding population and GDP, as these external forces are also predicted to have a strong impact on whether or not a bank fails.

General Analysis

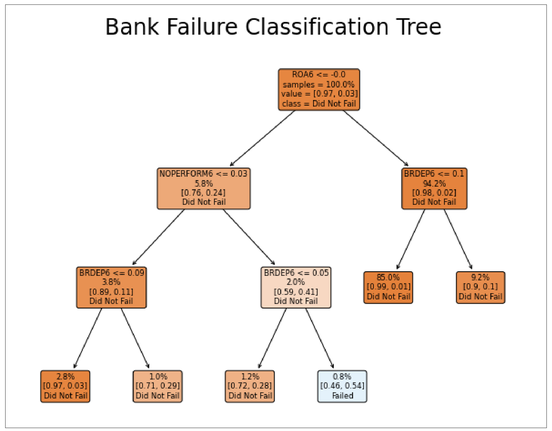
Current research has explored 4 different potential models for determining bank health: support vector machines (SVM), stochastic gradient descent (SGD), logistic regression, and decision trees. For this analysis we used the model's F1-scores to judge performance. The F1-score captures both the models ability to detect banks that did fail and how often it falsely flagged banks that didn't fail. Of the models, the decision tree model was favored for its ease of interpretation and high F1-score relative to other models.



Decision Trees

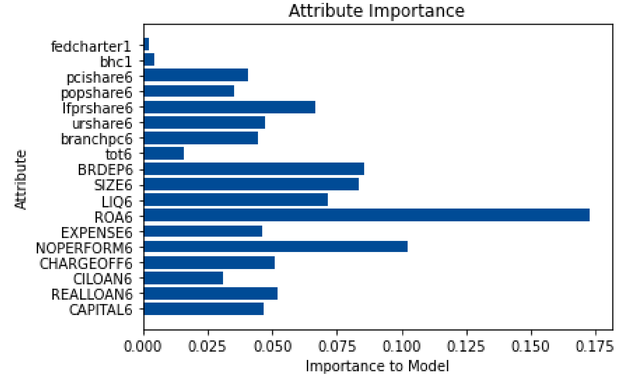
The decision tree model operates by repeatedly splitting the data based on provided attributes in order to maximize the purity of the data. The primary parameters involved in model tuning were class_weight and max_depth.

- Class_weight affects how significant the algorithm considers being in one classification over the other. The optimal tree had a class_weight of: Failed - .94, Didn't Fail - .06.
- Max_depth forces the model to terminate early and is utilized to prevent the model from over-fitting to the training data. The current optimal max_depth is 3.



Attribute Importance

The decision tree model is also capable of informing which attributes are most significant to predicting bank failure. Importance is calculated using a mixture of mean and standard deviation of the decrease in impurity that results from splitting the data on that attribute.



Important attributes to determining bank failure included a bank's return on assets, the amount of non-performing assets they have, and their brokered assets. Meanwhile less important attributes included if the bank was chartered by the federal reserve or part of a bank holding company. Noticeably the analysis reveals there are many variables in the dataset that prove to be relatively important to the model. While current optimal models aren't making use of all these attributes, this demonstrates there likely is value still to be extracted from these variables.

Conclusions

Our results demonstrated that it is possible to build machine learning models which correctly predicted most banks that failed during the 2007-2008 Global Financial Crisis, although there is still work to be done. Of the common machine learning techniques, decision trees proved to be the most effective at predicting failure. With further refinement this approach could be used to assist regulators in resource allocation to help banks avoid failure and react quickly to new causes of failure.

Future Research

- This research could be improved through exploring other models, further parameter tuning, and experimentation with the training and testing set split.
- We hypothesize that many false positives currently occurring in the model could be banks close to failure but that haven't officially failed yet. Future research aims to switch the binary failure variable to a numeric measure of bank health in order to a gain a more complete understanding of how well the model is performing.
- Currently testing has yet to be carried out using the full data set. Use of this expanded data set will allow for machine learning to guide which variables are the most important and work from this greater quantity of data to improve accuracy.