Machine Learning for Bank Failure Early Warning Systems

Researchers: Katja Mathesius, Adam Lear, Sam Basala, Tommy Huelhorst, & Chi Dang
Academic Mentors: Dr. Eric Manley, Dr. Sean Severe

Drake UNIVERSITY

Background

The global financial crisis from 2007 to 2008 was the worst economic crash since the great depression and was largely sparked by mass bank failures. While there are measures for countering bank failures, in order to effectively utilize them one must be able to predict which banks are going to fail. Machine learning's strong predictive abilities has great potential to help address this problem.

Purpose

This research can assist bank examiners in allocating resources to avert bank failures and 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.

Dataset

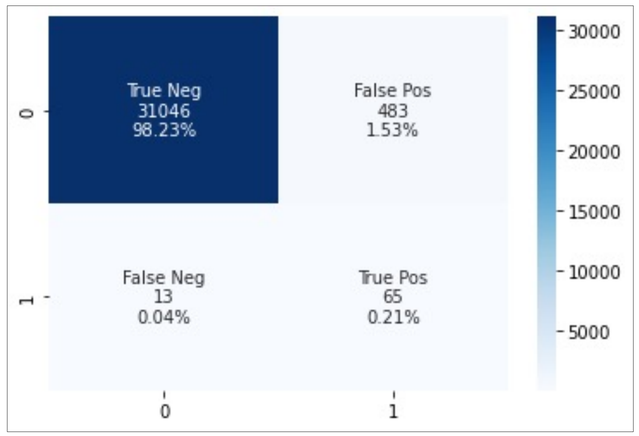
- Bank data from 2007-2012 that contains 158,031 unique observations and 27 unique attributes
- One observation depicts a bank's unique attributes in each quarter

Variable	Definition
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)

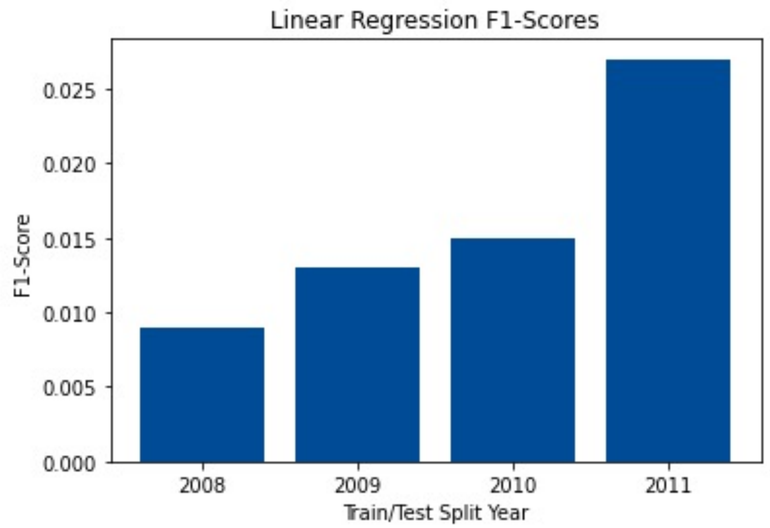
Methods and Results

The project used four different machine learning algorithms to make predictions through a basic procedure: data integration, data analysis, modeling, and model analysis.

- Support Vector Machine (SVM): An algorithm that separates classifications using a plan in the most even amount. Utilized parameters: C, kernel, degree, gamma, class_weights.

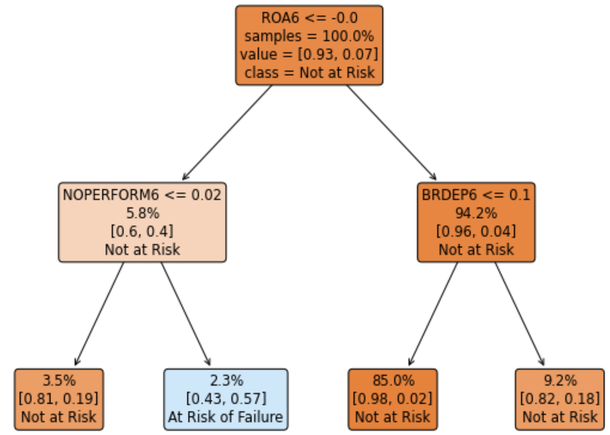


- Stochastic Gradient Descent (SGD): Uses linear models along with the SGD learning. Utilized parameters: loss, early_stopping, alpha, and max_iter.
- Linear Regression: A linear approach that models the relationship between a target and independent variables. Train data and test data were split by a given year. Parameters utilized: class_weight and max_iter.



- Decision Tree uses a series of decisions to attempt to maximize the purity of the data. The primary parameters used were class_weight and max_depth.

Bank Failure Classification Tree



Conclusions

The model that worked most efficiently in predicting bank failure was the SVM model. 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. This approach could be used to assist regulators in resource allocation to help banks avoid failure and react to new causes of failure.

Future Research

- This research could be improved through exploring other models, making use of different parameters, and through splitting the train and test data in a different manner.
- Investigate suspicion that the false positives contain many banks that may have been close to failure even if they did not actually fail, so the models may be performing better than our binary classification metrics are able to detect.