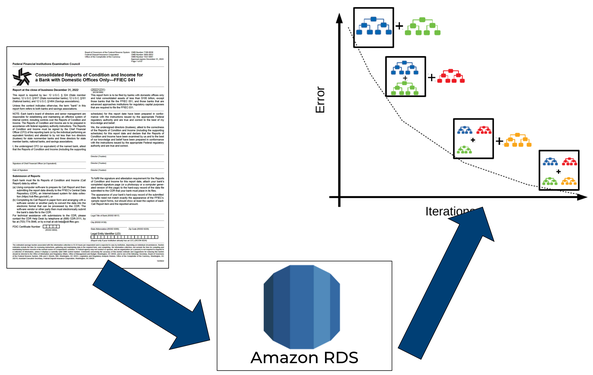
Building a Machine Learning Model to Predict Bank Failure

Gonzalo Valdenebro, Riley Rongere, Jacob Danner, Katja Mathesius, Jaehyeok Choi & Eric Manley, Ph.D.
Department Of Computer Science & Mathematics, College of Arts & Sciences
Drake University

CCSC CONSORTIUM FOR COMPUTING SCIENCES IN COLLEGES
Midwest Region

Abstract

This study aims to develop a machine learning model that predicts the risk of bank failure using historical bank call reports data. More specifically, we attempt to predict the total equity capital less goodwill as a ratio of the bank's total assets for the next calendar year. We use this as a proxy for bank health following the work of Wheelock and Wilson in 2000 [1] who defined bank failure as occurring when this value fell below 2%. To this end, we built a database that stores quarterly data from every US bank call report. The data was processed and analyzed using statistical modeling techniques to identify the most critical variables that correlate with future bank health.



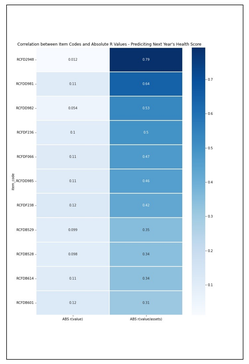
We are building machine learning algorithms to help isolate these metrics and predict bank failure risk accurately. The primary objective of this research is to help banks and investors understand the variables that lead to bank failure and to monitor them carefully. We believe that our study will provide valuable insights into the financial industry, which can help prevent bank failures and promote economic stability.

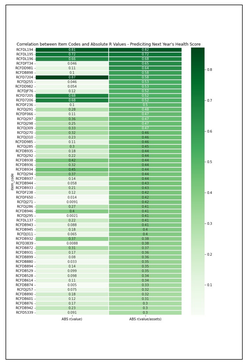
Methodology

We used 553 variables out of the 5,000 in the bank call report to measure statistical significance by running a simple linear regression against the next year health column to check what variables might have more significance predicting our target variable.

Variable Signal Search:

- We used the linregress from the Scipy [2] package in Python
- Selected the 53 variables that had the highest R^2 squared.





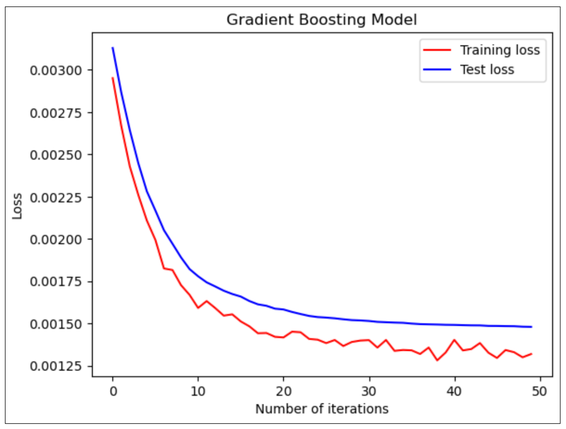
Modelling:

- After selecting the 53 variables, we queried the database and made a data frame with all those variables for each bank.
- We then made all the variables a ratio of their Total Assets [1].
- For this analysis we have tested our models using the Scikit – Learn [3] framework:
  - Random Forest Regressor: R^2 = 55%
  - Gradient Boosting Regressor: R^2 = 66 %

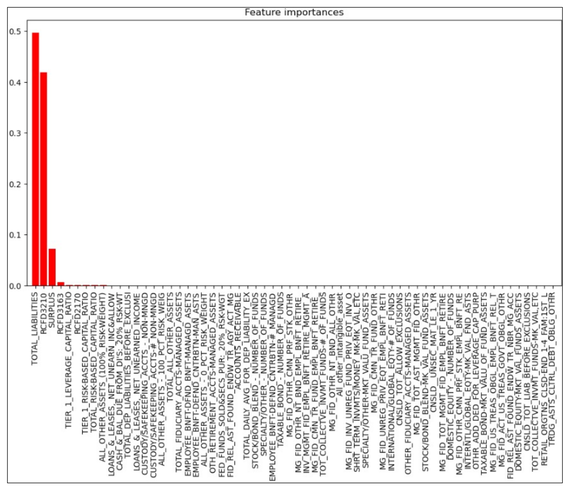
Results

We have compared results and determined that the best model was our Gradient Boosting Regressor resulting in an R^2 of 66%

Training/ Test Loss Graph:



Model Feature Importance Graph:



Conclusion

- We must consider adding more data to out models in order to gain more signal, our testing data is not doing a great job in the predictions
- Models that have higher "n_estimators" tend to score better, although there is a risk of overfitting

Recommendations

- Testing for significance on the reminder variables in the bank call report
- Implementation of a Neural Network given the complexity of the data

Acknowledgements

- [1] "Why Do Banks Disappear? The Determinants of U.S. Bank Failures and Acquisitions" Author(s): David C. Wheelock and Paul W. Wilson Source: The Review of Economics and Statistics, Feb., 2000, Vol. 82, No. 1 (Feb., 2000)
- [2] Scipy Package
- [3] Scikit-Learn
- Sean Severe, Ph.D.