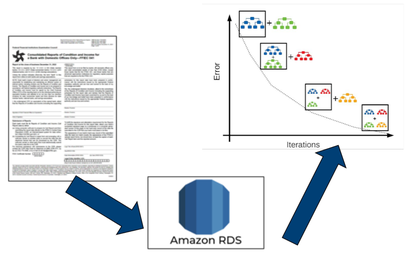
Bank Failure Early Warning Systems Using Machine Learning

Jackson Becker, Vu Bui, Haris Mehuljic, Jose Usina Mogro, & Gonzalo Valdenebro
Academic Mentors: Eric Manley, Ph.D. & Sean Severe, Ph.D.
Department Of Computer Science & Mathematics, College of Arts & Sciences, Drake University

Abstract

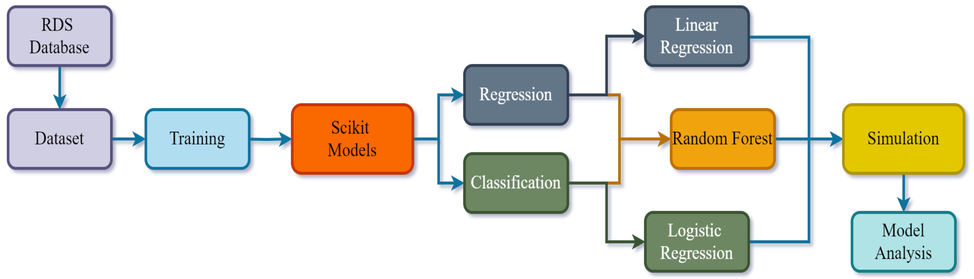
The global financial crisis of 2007 to 2008 was the worst economic crash since the Great Depression and was largely caused by an unforeseen spike in bank failures. Machine learning's strong predictive ability has great potential to address this issue. Therefore, we seek to avert future disasters by creating predictive machine learning models using data gathered from historical bank call reports.

We utilize the bank health metric as defined in Wheelock and Wilson[1] as our prediction target. Our research considers both classification models and regression models to predict bank health.



The purpose of this research is to develop a predictive model for assessing the health of financial institutions in real-time, establishing thresholds to detect at-risk banks early. This model will offer valuable insights into variables contributing to bank failure, aiding in promoting economic stability, and preventing potential future crises.

Methodology



- The dataset we use for machine learning is requested from our AWS RDS instance and saved as a CSV.
- Data preparation
  - Define features.
  - Define time series variables (referred to as lags).
  - Use the bank health metric[1] as prediction target.
- Simulated the bank failure prediction using data from 2006-2016, focusing on the two preceding years' worth of data.
- Analyzed the performance of the models

Conclusion

Classification:

- Logistic Regression performs well in predicting failures, improving over time with more training data and achieving high accuracy with low false positive rates. Random Forest does not predict well until 2010 and has a high variability in accuracy, with an interesting cyclical pattern.

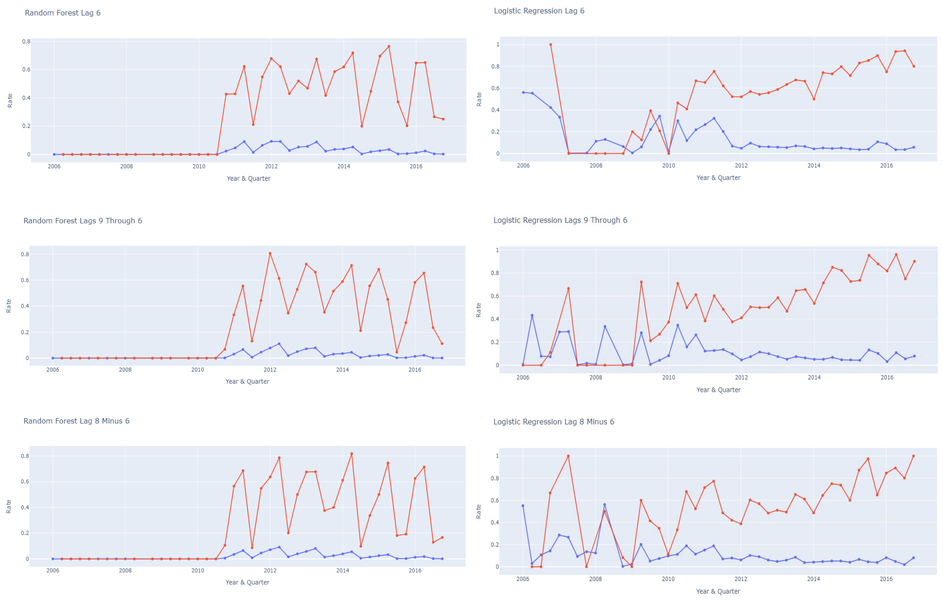
Regression:

- Both classification and regression models show strong performance, achieving an $R^2 > 0.8$, even early on, indicating an improvement on their effectiveness in capturing relevant patterns for prediction over time.

Both:

- Using one quarter's worth of data for predictions is as effective as using a time series with two years' worth of data, highlighting the efficiency of the models in capturing predictive insights.

Results



Classification *(left)*:

- Used logistic regression and random forest models.
- Random forest displays a cyclical pattern which occurs every year, with the amount of bank failures predicted at the end of the year peaks, before going back down at the beginning of the next year.
- Logistic regression appears to be more stable.

Regression *(right)*:

- Used linear regression and random forest models.
- Ran the same three tests as we did when testing classification.
  - Included specific time series variables.
  - The graph is a combined graph of the results for each algorithm.

Future Research

- Investigate the cyclical nature of the random forest classification results
  - Perhaps there is an underlying phenomenon which is the cause.
  - Do bank failures occur more often at certain times of year?
- Investigate ordering of regression predictions to flag a fixed proportion of banks for regulators to look more closely at.

References

- [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)
- Python Libraries:
  - [2] Pandas
  - [3] Scikit-Learn