Project 1: Draft

Econ 1680: Machine Learning, Text Analysis, and Economics

Lena Kim

May 9, 2024

github link: https://github.com/wkim31/Econ1680-ML

1 Introduction and Motivation

Bank failure is often a direct signal of financial crises. Banks are not only a safe to park money for the average depositor, but also the financial drivers of social, economic, and technological change. This is why countless literature has sought to predict bank failure from its balance sheet metrics before they may occur- what range of specific metrics can signal a danger zone? (Meyer and Pifer, 1970, p. 860-868). While such research has indicated liquidity ratios, rates of interest, and operating revenue as the most salient metrics in determining whether a bank fails, this project seeks to go beyond and consider what happens after the actual failure.

There is no expectation of finding much difference in the metrics that play the most hand at bank failure. Rather, this projcet seeks to extend previous analysis to consider what metrics of the bank- such as funds left, total assets, amount due to creditors- contribute most to whether it is eventually acquired by a larger national bank, a regional bank, or not acquired at all.

For this end, both linear and nonlinear classification methods such as multinomial logit regression with year fixed effects, multilayer perceptrons, and a cluster analysis are used. The main evaluation metric to compare models will be accuracy scores on whether that model correctly classifies acquisition by regional, national, or none at all in order to determine the importance of the metrics at play.

Research Question: Given a failed bank's balance sheet metrics, how can we predict whether it is acquired by a national bank, a regional bank, or not acquired at all? Which metrics contribute the most weight to this decision?

2 Data Sources and Descriptions

The main data source is a publicly available dataset from FDIC archives on 126 failed banks and their selected balance sheet items over the years of 2008-2020. This set is joined with another dataset of the failed banks acquirers to output the complete dataset of interest. After some exploratory data analysis and preprocessing, the main features of interest are: 'Cash and Investments', 'Due from FDIC Corp and Receivables', 'Assets in Liquidation', 'Total Assets', 'Administrative Liabilities', 'Total Unpaid Other Claimants', 'Uninsured Deposit Claims', 'General Creditor', 'Total Liabilities', and 'Acquisition Type'. Summary statistics are shown at the end of this paper.

3 Method

The main methods implemented are: Multinomial Logit Regression with year fixed effects, Multilayer Perceptron classifier, and a cluster analysis.

Starting with the main model of interest for this project, multinomial logit was implemented, with year fixed effects to control for the large amount of banks that failed during or directly after the Great Financial Crisis of 2008. This model was selected because it allows for the multinomial nature of our target variable, as well as to explore the potential linearity of our data. An initial postulation was that the higher cash level and assets, the higher the classification (i.e. the higher the acquirer bank's level is, from none to national), because banks with more assets at the time of failure should be more attractive to acquire due to greater potential to grow. In addition, another initial postulation was that the higher the dues and liabilities, the lower the classification, as banks with more liabilities are probably less attractive to acquire.:

$$y_i = G(\beta_0 + \beta_1 X_{assets} + \beta_2 X_{cash} - \beta_3 X_{liabilities} - \beta_3 X_{duefromFDIC} + \varepsilon_i) \tag{1}$$

where, $y_i \in \{0, 2\}$ is the acquisition level of the acquirer (0 is none, 1 is regional, 2 is national), and G() represents the mean function of the prediction, as stipulated by the multinomial regression. An interpretation of the above would be: "For a 1 percent increase in X, the probability of bank acquisition by a national bank increases by β_i multiplied by the mean function (or the mean predicted probability of acquisition)". In the current example, we would require linearity of log-odds:

$$\log\left(\frac{P(Y = j|\mathbf{X})}{P(Y = K|\mathbf{X})}\right) = \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \ldots + \beta_{jp}X_p \tag{2}$$

In simpler terms, we assume that the relationship between the characteristics of a failed bank (represented by predictor variables like cash, liabilities) and the log-odds of being acquired by a specific type of bank (e.g., national or regional) is linear. That is, the way these characteristics influence the probability likelihood of acquisition by a certain type of bank follows a straight-line pattern, which seems plausible as stated above.

The second model is the Multilayer Perceptron, which leverages simple neural networks to explore the potential of a nonlinear model. An MLP classifier is used to allow for more complex relationships not necessarily explored in the multinomial logit to be captured through hidden layers and activation functions as well as to compare which features are most important in determining acquisition:

$$y_i = \sigma(\beta_0 + \beta_1 X_{\text{assets}} + \beta_2 X_{\text{cash}} - \beta_3 X_{\text{liabilities}} - \beta_4 X_{\text{due from FDIC}} + \epsilon_i) \tag{3}$$

where, $y_i \in \{0, 2\}$ is the acquisition level of the acquirer (0 is none, 1 is regional, 2 is national), and σ is the activation function log sigmoid. Unlike linear models, neural networks do not rely on underlying assumptions and instead learn patterns directly from the data. While the interpretation of these results are not as straightforward as the linear model, we can say that if β is more than 0, an increase in X_i is associated with an increase in the log-odds of being acquired by a national bank, and we can also analyze feature importance outputted by the neural network.

The final method was a cluster analysis using KMeans clustering. This method was not a results-oriented method, but rather a general method to get a sense of the data clusters in the initial onset of the project. After constructing a dendrogram for the optimal number of clusters at K = 2, distances between cluster centroids were minimized, where c_j is the centroid of cluster j, k is the number of clusters, and x_i is a data point:

$$\text{minimize} \sum_{i=1}^{n} \min_{j=1}^{k} \|x_i - c_j\|^2 \tag{4}$$

4 Results

After running a Multinomial Logit Regression, the results (attached in section 7) seem to indicate that most of the explanatory variables are not powerful predictors of bank acquisition, as indicated by statistically insignificant p-values. However, the most statistically significant predictor is Administrative Liabilities for acquisition type 2, with a p-value close to the traditional alpha level of 0.05.

These results indicate that a higher amount of administrative liabilities leads to a higher likelihood of being acquired by a national bank.One possible explanation that could be at play is that larger banks with greater financial resources may be more capable and willing to absorb and handle these administrative liabilities. A postulated hypothesis is that the failed bank's location, or other investing activities, or the possibility of greater synergies from the acquisition with that specific failed bank overshadowed the greater administrative liabilities.

A formal interpretation for Administrative Liabilities when acquisition type = 2 is "for a 1 percent increase in administrative liabilities, the probability of bank acquisition by a national bank increases by (1.0512 x mean G($y_i = G(\beta_0 + \beta_1 X_{assets} + \beta_2 X_{cash} - \beta_3 X_{liabilities} - \beta_3 X_{duefromFDIC} + \varepsilon_i)$.

A Multilayer Perceptron was trained after searching for the optimal learning rate. After training a model with 8 layers and 4 neurons and a learning rate of 0.01 with logistic sigmoid, the accuracy score found was around 42 percent. A naive visualization of feature importance using a shap kernel explainer indicated that conversely to the multinomial results, the most important features in indicating a higher level of acquisition was cash and investments.

The clustering analysis performed in the preliminary onset of the project indicated that clusters were not as conducive to the fit of the data as originally expected. Using a dendogram, a Kmeans algorithm with K initilized to 2 was used, but indicated 127 points in cluster 0 with only 2 points in cluster 1. This led to highly abberrant results for cluster 0, and were inconclusive.

5 Conclusion

Overall, the accuracy scores of the models are not the greatest. The strengths of the project thus far is the surprising finding from the multinomial logit model: administrative liabilities, contrary to expectations, increaes the likelihood of a failed bank getting acquired by a large national bank. The neural network feature importance reinforced an initial expectation- that cash and investments play the most positive weight in determining a higher level of acquisition. Overall, most balance sheet metrics are not highly predictive of future acquisition type. Future directions are to confirm the hypothesis outlined by the results of the multilogit by collecting income statements and cash flow statements of the failed bank to answer this: Does the failed bank's location, or other investing activities, or the possibility of greater synergies from the acquisition with that specific failed bank overshadowed the greater administrative liabilities?

6 References

References

Meyer, Paul and Howard Pifer (1970) "Prediction of Bank Failures," *Journal of Finance*, 25 (4), 860–68.

7 Graphs and Tables

Summary Statistics of Variables:

Table 1: Summary Statistics

	Cash and Investments	Due from FDIC Corp	Assets in Liquidation	Total Assets
Count	126	126	126	126
Mean	7209.11	1.05	281.96	7386.94
Stdv	18748.56	12.042496	1415.60	18979.5
Min	80.00	78.00	0.00	80.00
Max	153977.00	72.00	13001.00	153977.00

Note: Summary statistics for features of interest.

Table 2: Summary Statistics Continued

	Admin Liabilities	Total Unpaid Other Claimants	Uninsured Deposit	General Creditor
Count	126	126	126	126
Mean	1006.51	129828.40	924.31	29543.37
Stdv	2824.18	1319010.10	8448.840	26922.67
Min	0.00	0.00	0.00	14.0
Max	24452.00	14808670	94350.00	3025216

Table 3: Summary Statistics Final

	Total Liabilities	Acquisition Type
Count	126	126
Mean	432955.50	0.751938
Stdv	1721667	0.80
Min	6.00	0.00
Max	14809820.00	2.00

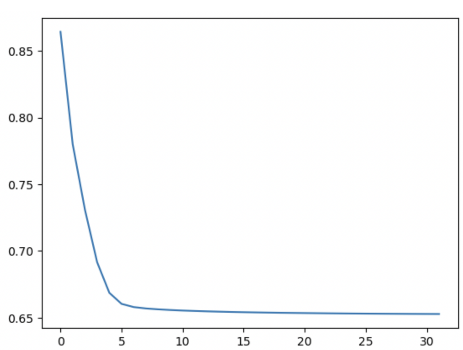
Figure 1: MN Logit Coefficient Plot

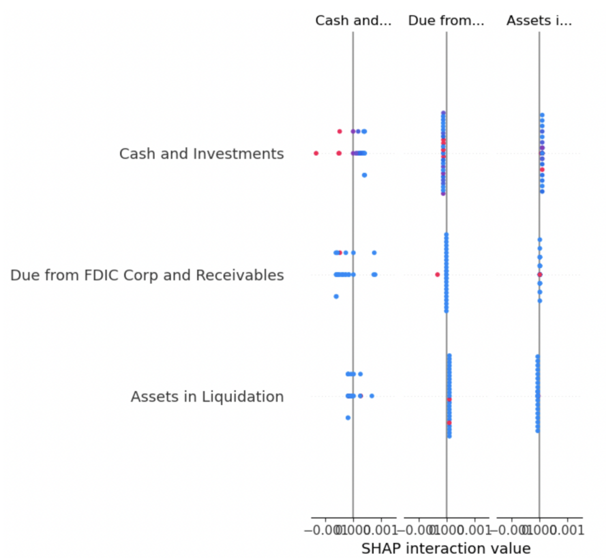
Table 1: MNLogit Regression Results

	Acquisition Type 1		2	
	coef	$P > \|z\|$	coef	$P > \|z\|$
const	-2.0615	0.228	-2.5275	0.112
Cash and Investments	0.1073	0.999	-2.5472	0.990
Due from FDIC Corp and Receivables	-0.2622	0.707	-0.3656	0.645
Assets in Liquidation	-0.2992	0.971	0.0368	0.997
Total Assets	0.0808	1.000	-2.3051	0.991
Administrative Liabilities	-0.5820	0.307	1.0512	0.051
Total Unpaid Other Claimants	-1.3792	0.961	0.1808	0.990
Uninsured Deposit Claims	-1.1893	0.626	0.8256	0.537
General Creditor	-2.2803	0.904	1.6074	0.897
Total Liabilities	1.4458	0.464	1.4525	0.406
Year_2009.0	0.9411	0.511	1.5052	0.366
Year_2010.0	1.2540	0.391	1.4309	0.402
Year_2011.0	0.9458	0.542	1.1449	0.527
Year_2012.0	1.5677	0.314	1.5614	0.397
Year_2013.0	1.4455	0.393	1.1936	0.565
Year_2014.0	-1.9565	0.759	-1.7699	0.775
Year_2015.0	-0.4925	0.982	4.4104	0.711
Year_2017.0	0.7339	0.703	-1.8369	0.809
Year_2019.0	-2.0161	0.766	-1.1705	0.876
Year_2020.0	4.8004	0.191	4.1356	0.292

Note: coefficients are not too significant.

Figure 2: Learning Rate and Feature Importances





Note: Best Learning rate at 0.01.

Most "postive" features (red dots) are cash investments