Project 1: Draft

Econ 1680: Machine Learning, Text Analysis, and Economics

Lena Kim

March 8, 2024

1 Introduction

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)). While such research has indicated liquidity ratios, rates of interest, and operating revenue as the most salient metrics in determining whether a bank fails, I seek to go beyond and consider what happens after the actual failure.

For this project, I do not expect to find much difference in the metrics that play the most hand at bank failure. Rather, I seek 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, I will use both linear and nonlinear classification methods such as multinomial logit, multilayer perceptrons, support vector classifiers, and random forest classifiers. My main evaluation metric will be accuracy scores, as it is important that the model correctly classifies acquistion by regional, national, or none at all in order to later 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

I use a publicly available Kaggle dataset on around 126 failed banks and their selected balance sheet items over the years of 2008-2020, which is pulled from the FDIC archives on failed banks by state. I then joined this set with another dataset of failed banks and their acquirers to output my complete dataset of interest. After some exploratory data analysis and preprocessing, my 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 below:

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

This is a random sample of FDIC failed bank data for just over 120 failed banks, which is computationally efficient enough for the ambitious number of methods I am planning to implement.

3 Method

This project implements several methods alike, and I plan to see which one produces the best model fit. Starting with the first model and the main model of interest for this writeup, multinomial logit was implement. An intial 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 when they fail should be more attractive to acquire due to synergies. In addition, 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 likelihood of acquisition by a certain type of bank follows a straight-line pattern, which seems plausible as stated above.

The second model I want to focus on is Multilayer Perceptron, which leverages simple neural networks to explore the potential of a nonlinear model:

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

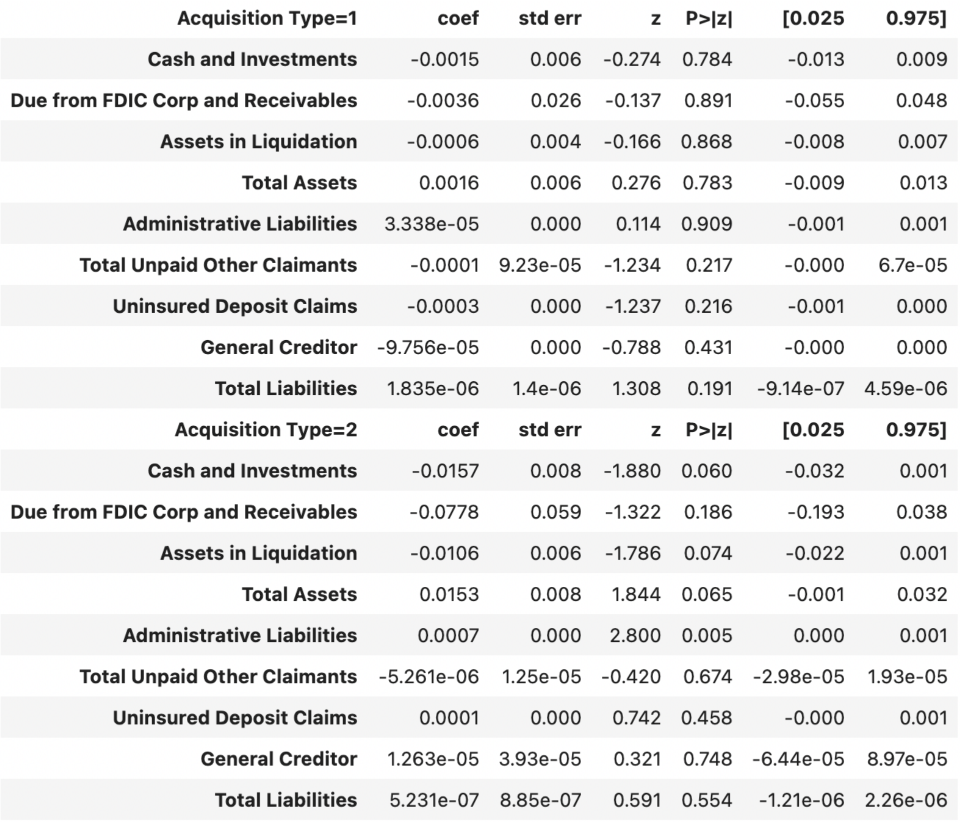
where, $y_i \in \{0, 2\}$ is the acquisition level of the acquirer (0 is none, 1 is regional, 2 is national), and σ is the activatation function log sigmoid. While the interpretation of these results are not as straightforward as the linear model, we can say that if β is more than 0, a one unit increase in X_i is associated with an increase in the log-odds of being acquired by a national bank.

4 Results or Expected Results

After running a Multinomial Logit Regression, the results seem to indicate rather a nonlinear relationship, as indicated by statisticaly insignificant p-values:

We can see that the trend is that none of the results are too statistically significant, except for a few in acquisition type = 2 (national bank). We can interpret such results as in Administrative Liabilities when acquisition type = 2 as "for a 1 percent increase in administrative liabilities, the probability of bank acquisition by a national bank increases by (0.0007 x mean G($y_i = G(\beta_0 + \beta_1 X_{assets} + \beta_2 X_{cash} - \beta_3 X_{liabilities} - \beta_3 X_{duefromFDIC} + \varepsilon_i)$.. I

Figure 1: MN Logit Coefficient Plot



Acquisition Type=1	coef	std err	z	P>|z|	[0.025	0.975]
Cash and Investments	-0.0015	0.006	-0.274	0.784	-0.013	0.009
Due from FDIC Corp and Receivables	-0.0036	0.026	-0.137	0.891	-0.055	0.048
Assets in Liquidation	-0.0006	0.004	-0.166	0.868	-0.008	0.007
Total Assets	0.0016	0.006	0.276	0.783	-0.009	0.013
Administrative Liabilities	3.338e-05	0.000	0.114	0.909	-0.001	0.001
Total Unpaid Other Claimants	-0.0001	9.23e-05	-1.234	0.217	-0.000	6.7e-05
Uninsured Deposit Claims	-0.0003	0.000	-1.237	0.216	-0.001	0.000
General Creditor	-9.756e-05	0.000	-0.788	0.431	-0.000	0.000
Total Liabilities	1.835e-06	1.4e-06	1.308	0.191	-9.14e-07	4.59e-06
Acquisition Type=2	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Cash and Investments	-0.0157	0.008	-1.880	0.060	-0.032	0.001
Due from FDIC Corp and Receivables	-0.0778	0.059	-1.322	0.186	-0.193	0.038
Assets in Liquidation	-0.0106	0.006	-1.786	0.074	-0.022	0.001
Total Assets	0.0153	0.008	1.844	0.065	-0.001	0.032
Administrative Liabilities	0.0007	0.000	2.800	0.005	0.000	0.001
Total Unpaid Other Claimants	-5.261e-06	1.25e-05	-0.420	0.674	-2.98e-05	1.93e-05
Uninsured Deposit Claims	0.0001	0.000	0.742	0.458	-0.000	0.001
General Creditor	1.263e-05	3.93e-05	0.321	0.748	-6.44e-05	8.97e-05
Total Liabilities	5.231e-07	8.85e-07	0.591	0.554	-1.21e-06	2.26e-06

Note: coefficients are not too significant.

have calculated this G() in stata as 0.8, so it increases the probability by 0.56 percent.

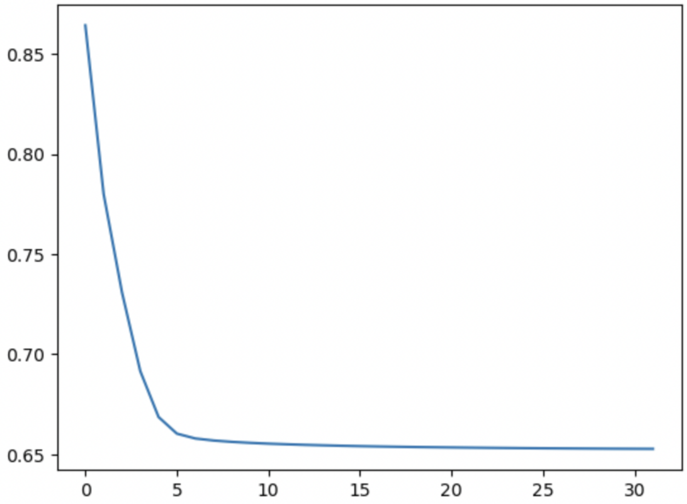
In this model, it is important to consider the effects of time, as indicated by the Great Financial Crisis of 2008, of which our dataset is part of. As next steps to improve the logit model, I will consider adding Fixed Effects for 2008 versus other years.

I have also trained a Multilayer Perceptron 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, I found an accuracy score of 49 percent. As a next step, I will consider tuning more hyperparameters to improve the evaluation metric.

In addition, my jupyter notebook contains the beginnings of my Random Forest Classifier as well as a Support Vector Classifier with a nonlinear kernel. The preliminary accuracy scores of training and testing, respectively, were:

- RFC, (0.97, 0.5)
- SVC, (0.3846)

Figure 2: Learning Rate for MLP Classifier



Note: best learnign rate at 0.01.

5 Conclusion

There is a general sense of which direction to head next (considering time series and more nonlinear models). We can see that the accuracy scores of my linear models are not the greatest, and as such we need to strengthen the accuracy scores by such improvements. The strengths of my project thus far is that my data preprocessing and cleaning steps- the bulk of my time spent thus far- and my kfolds have been implemented quite succesfully, freeing up the steps of analysis. The limitations could be the nonlinearity that seems to be present in my data, as evidenced by the initially low accuracy scores for these models. Future directions are:

- consider a time series analysis, or a Fixed Effects on years
- hyperparameter tuning: search more hyperparameters to tune for MLP other than the learning rate
- consider F test for multiple regressions to assess predictive power of features

References

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