Data Mining Coursework

Contents

Classification of banks as financially strong or weak based on binary logistic regression in SPSS

Introduction:

"Banks are profit-oriented businesses" (*Loughnane, n.d.*). "Financial ratios are an important technique of the financial analysis of a business organization" (*Goel, 2016*). The ratios help in evaluating the financial conditions of businesses. Any organization's financial statements are usually overcrowded with numbers and, therefore, could be very discouraging. However, financial ratios make things systematic, easier to interpret and easier for fancy statistics (*Goel, 2016*).

In the classification report, two financial ratios of banks have been used in the logistic regression analysis. To perform the analysis, IBM SPSS has been used as statistical software.

Exploring Data & Pre-processing:

(Part (a) answer)

Variable Identification – In the given dataset, the "Financial Condition" attribute with two possible values, i.e. "Strong" and "Week", has been considered as a "Dependent Variable" (DV), whereas "Total Loan & Lease to Total Assets Ratio" (LLR) and "Total Expenses to Total Assets Ratio" (ER) attributes have been considered as "Independent Variables" (IV).

Normality Test – "For Kolmogorov-Smirnov and Shapiro-Wilk normality tests, the null hypothesis states that data are taken from a normally distributed population. When P > 0.05, the null hypothesis accepted and data are called as normally distributed" (*Mishra et al., 2019*). In *table 1.1*, the Sig. Value for both the IVs is statistically insignificant in both the tests, i.e. Sig. Value> 0.05, which indicates that the distribution of data in both the IVs is normally distributed.

	Kolmogorov-Smirnov			*Shapiro-Wilk*		
	Statistic	*df*	*Sig.*	*Statistic*	*df*	*Sig.*
Total Expenses/Total Assets Ratio (ER)	*.143*	*21*	*.200*	*.938*	*21*	*.194*
Total Loan & Lease/ Total Assets Ratio (LLR)	*.105*	*21*	*.200*	*.965*	*21*	*.622*

table 1.1 – Tests of Normality
Source: Self-built in SPSS

Type of Analysis – As there are two independent variable being analyzed, therefore it turned out to be a "**Multivariate Analysis**".

(Part (b) answer)

Missing Value Analysis (MVA) & Treatment – *Table 1.2* shows that a single value is missing for each of the explanatory variables. It happened to be the same observation for both variables. The removal of missing values makes the analysis lucid (*Kwak & Kim, 2017*) and "is the most commonly used method in statistical analysis programs such as SPSS and SAS to handle missing values" (*Kwak & Kim, 2017*). Hence, the observation was removed from the dataset. Therefore, after the removal, the dataset ended up having 21 observations.

	Valid	Missing	
	Counts	Count(s)	Percent
Total Expenses/Total Assets Ratio (ER)	21	1	4.5
Total Loan & Lease/Total Assets Ratio (LLR)	21	1	4.5

table 1.2 – Missing Value Analysis
Source: Self-developed in SPSS

Duplicate Values & Treatment – Duplicate data tends to skew the results of the analysis. Hence as prevention, any duplicate data must be removed from the dataset during the data cleaning phase (*McFedries, 2013*). There had been a duplicate observation in the dataset, which has been removed successfully. After the removal, we have 20 observations on which further analysis was carried out.

Outliers & Treatment- Keeping the outliers (*see the explanation below*).

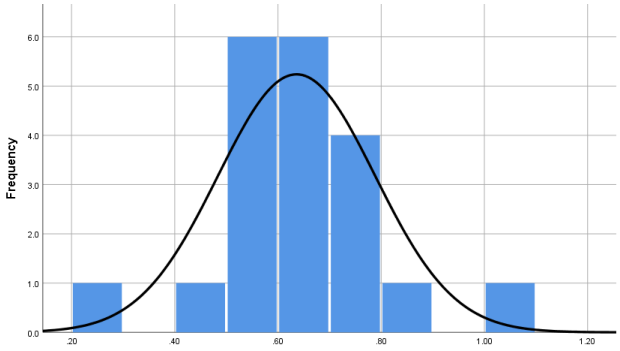


figure 1.1
Histogram for Outliers in Total Loan & Lease/Total Assets Ratio

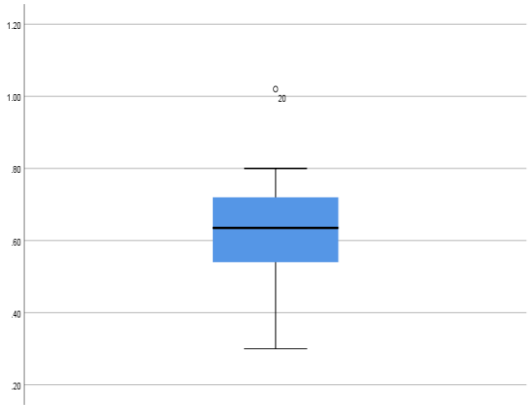


figure 1.2
Boxplot for Outliers in

Source: Self-developed in SPSS

It's critical to spot any extraordinary observations (outliers) or patterns of observations that might cause problems with subsequent data analysis (*Landau & Everitt, 2019*). Although the observations highlighted as "outlier" do not have to be removed before further analysis,

(Part (b) answer) …continues.

however, they do deserve careful consideration (*Landau & Everitt, 2019*). It is quite evident from *figure 1.2* that the "Total Loan & Lease/Total Assets Ratio" variable consists of an outlier, i.e. value 1.02. However, we can see that there is a normal curve on the histogram in *figure 1.1*. Even *table 1.1* indicates that the distribution is statistically insignificant from not being a normal distribution, i.e. the Sig. Value is less than 0.05. Moreover, it is certainly possible for a bank to have a "Total Loan & Lease/Total Assets ratio as high as 1.02. Hence, further analysis has been carried out while keeping the outlier in the dataset.

(Part (c) answer)

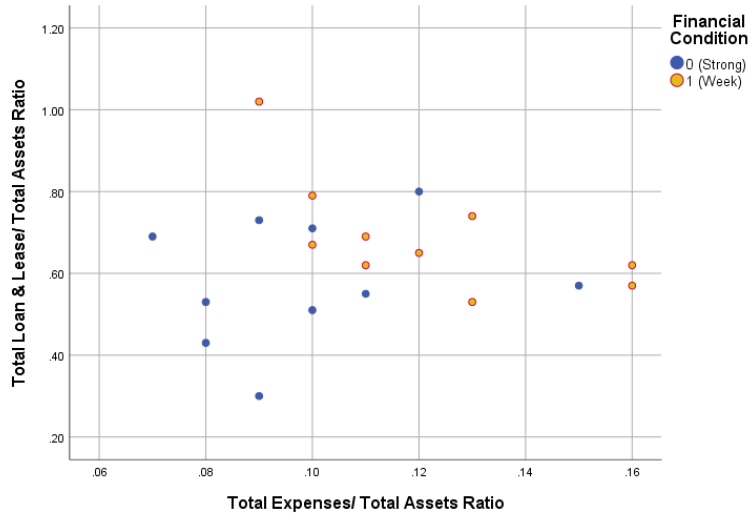


figure 1.3 – Scatterplot *Source: Self-developed in SPSS*

In *figure 1.3*, a scatterplot has been created with "Total Expenses/ Total Assets Ratio" on the X-axis, whereas "Total Loan & Lease/ Total Assets Ratio" on the Y-axis. Moreover, the dots have been labelled as per the "Financial Condition" attribute. It has been figured out from the scatterplot that **week banks have higher "Total Expenses to Total Assets ratio"** as compared to strong banks.

Logistic regression modelling:

Binary logistic regression has been carried out using the SPSS data mining software package on the "Financial Conditions of Banks" dataset with data points of 20 banks. In the dataset, we have "Financial Condition" as the Dependent Variable, which has two possible values, i.e. Strong or Week, whereas "Total Loan & Lease to Total Assets Ratio" and "Total Expenses to Total Assets Ratio" as Explanatory Variables. On the Y-axis of the probability graph, p = 1 has been plotted as "Week" class, whereas p = 0 as "Strong" class.

In the *table 1.3*, $\chi 2$ (2, N=20) = 7.937, p <0.05, indicates that the model is statistically significant. For explanatory variables to have a statistically significant impact on the dependent variable, the p-value or Sig. Value must be less than 0.05 in Omnibus Tests of Model Coefficients (*Practical Application of Statistics in the Social Sciences (PASSS), 2014*).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	7.937	2	.019
	Block	7.937	2	.019
	Model	7.937	2	.019

table 1.3 *Source: Self-developed in SPSS*

In order to accept the null hypothesis, which indicates that the model is a good fit for the data, the Sig. Value must be greater than 0.05 (*Restore, 2011*). In *table 1.4*, the Sig. value 0.092 (Sig. value > 0.05) indicates that the regression model i.e. *eq. 1.3* fits the data.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.622	8	.092

table 1.4 *Source: Self-developed in SPSS*

In *table 1.5*, the value of Nagelkerke R Square = 0.437 indicates that about 43.7% of the variance of the financial condition of a bank could be predicted by the model. Whereas, the value of Cox & Snell R Square = 0.328 indicates that about 32.8% variance in the financial condition of bank could be estimated by the model. (*Restore, 2011*).

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	19.789	.328	.437

table 1.5
Source: Self-developed in SPSS

Table 1.6 shows that with a cut off value equal to 0.5, the model's overall prediction accuracy is 80%, which is statistically termed as percentage accuracy of the classification (PAC). The sensitivity of the classification (percentage of occurrence correctly predicted) and the specificity of the classification (percentage of non-occurrence correctly predicted) are at 80% each (*Parikh, R. et al, 2008*). In other words, 8 out of 10 strong banks were correctly classified (True Positive). Similarly, 8 out 10 week banks were correctly classified (True Negative) by the model.

Classification Table

			Predicted		
			Financial Condition		Percentage
Observed			0 (Strong)	1 (Week)	Correct (%)
Step 1	Financial Condition	0 (Strong)	8	2	80.0
		1 (Week)	2	8	80.0
	Overall Percentage				80.0

The cut value is 0.500
table 1.6
Source: Self-developed in SPSS

In *table 1.7*, the values in column B are the constant and coefficients. The values 50.557 and 7.920 are B1 and B2, respectively, i.e. coefficients of "Total Expenses/Total Assets Ratio" and "Total Loan & Lease/Total Assets Ratio" respectively. Whereas the value (-10.718) is the constant.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	Total Expenses/Total Assets Ratio	50.557	27.303	3.429	1	.064	9.052E+21
	Total Loan & Leases/Total Assets Ratio	7.920	5.086	2.424	1	.119	2750.742
	Constant	-10.718	5.189	4.267	1	.039	.000

table 1.7
Source: Self-developed in SPSS

(Part (d) answer)

Source 1 (Equations): (Shmueli et al, 2018)
Source 2 (Coefficients & Constant): table 1.7

1.) The logit as a function of the predictors

$$logit\ (Odds) = In\left(\frac{p}{1-p}\right) = a = -10.718 + 50.557(ER) + 7.920(LLR)$$

...eq. 1.1

2.) The odds as a function of the predictors

$$odds = e^{a} = e^{-10.718+50.557(ER)+7.920(LLR)}$$

...eq. 1.2

3.) The probability as a function of the predictors

$$p = \frac{1}{1+e^{-a}} = \frac{1}{1+e^{-(-10.718+50.557(ER)+7.920(LLR))}}$$

...eq. 1.3

(Part (e) answer)

The equations *eq. 1.1*, *eq. 1.2* and *eq. 1.3* have been used to classify a new bank with "Total Loan & Lease to Total Assets Ratio" and "Total Expenses to Total Assets Ratio" as 0.6 and 0.11, respectively.

*The **classification** of the **new bank** as per:*

Note: Calculations performed on desmos.com/scientific

a.) Logit equation i.e. *eq. 1.1*

$$\boldsymbol{logit}\ (odds\ of\ being\ financially\ week\ bank) = \mathbf{-0.4047}$$

(Part (e) answer)...continues.

b.) Odds equation i.e. *eq. 1.2*

$$\boldsymbol{odds}\ (of\ being\ financially\ week\ bank) = \mathbf{0.6672}$$

c.) Probability equation i.e. *eq. 1.3*

$$\boldsymbol{p}\ (probability\ of\ being\ financially\ week) = \mathbf{0.4001}$$

Since the probability (p) of being financially week of the new bank as per *eq. 1.3* is 0.4001, which is less than the cut off value of 0.5. Therefore the new bank has been classified as a **strong bank**.

(Part (f) answer)

For the given dataset, p = 0.5 has been decided as a cut off value, i.e. any bank that has a probability of being a financially week bank, greater than 0.5 value, will be categorised as a week bank or otherwise, categorised as a strong bank. The aforesaid probability will be calculated with *eq. 1.3*. Similarly, the cut off value for the logit and odds function has been determined and calculated.

Logit and Odds as a function of probability (p).

"The log (Odds), called the logit, takes value from -∞ (very low odds) to ∞ (very high odds). The logit of 0 corresponds to even odds of 1 (probability = 0.5)" (*Shmuel et al., 2016*). In other words, the cut-off values of logit and odds function that corresponds to the probability function cut of value = 0.5 are 0 and 1, respectively. This could be graphically seen in *figure 1.4* and *figure 1.5* and have been derived in *calculation 1.1* and *calculation 1.2*.

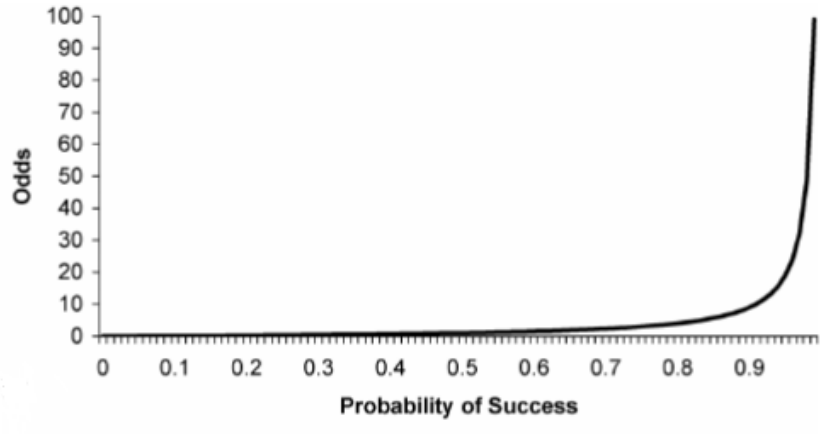


figure 1.4 – Odds as a function of probability

Source: (Shmueli et al, 2016, p. 222)

(Part (f) answer)…continues.

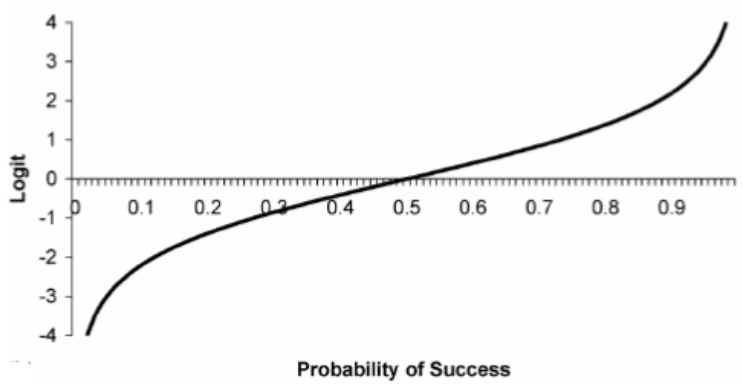


figure 1.5 – logit as a function of probability
Source: (Shmueli et al, 2016, p. 222)

calculation 1.1
Source: (Starmer, 2018)

$$Odds = \left(\frac{p}{1-p}\right)$$

$$Odds = \left(\frac{0.5}{1-0.5}\right) \ldots (p = 0.5)$$

$$Odds = \left(\frac{0.5}{0.5}\right)$$

$$\boldsymbol{Odds = 1}$$

Therefore, if the odds value of a new bank is greater than 1, then it is classified as a week bank (success class). Otherwise, it is classified as a strong bank.

calculation 1.2
Source: (Starmer, 2018)

$$logit\ (Odds) = In\left(\frac{p}{1-p}\right)$$

$$logit\ (Odds) = In\left(\frac{0.5}{1-0.5}\right) \ldots (p = 0.5)$$

$$logit\ (Odds) = In(1)$$

$$\boldsymbol{logit = 0}$$

Therefore, if the logit value of a new bank is greater than 0, then it is classified as a week bank (success class). Otherwise, it is classified as a strong bank.

(Part (h) answer)

Table 1.6 show that with a probability cut off value of 0.5; the model predicts an 80% accurate classification of the week banks. The model with a cut off value of 0.5 gives us two false negatives, i.e. two weak banks have been incorrectly classified as strong banks.

Classification Table

Observed			Predicted: Financial Condition 0 (Strong)	Predicted: Financial Condition 1 (Week)	Percentage Correct (%)
Step 1	Financial Condition	0 (Strong)	8	2	80.0
		1 (Week)	2	8	80.0
	Overall Percentage				80.0

The cut value is 0.500
table 1.6
Source: Self-developed in SPSS

(Part (h) answer)…continues.

In order to minimise the risk of incorrect classification of week banks as strong banks, as we see in *table 1.6* that two weak banks were incorrectly classified as strong banks, i.e. false negatives, we need to have a better cut off value. After carefully studying the *appendix 1* Classification Plot, it has been concluded that cut off value 0.41 yields 100% accurate classification of week banks, i.e. 100% sensitivity of classification, which could be seen in *table 1.8*.

Classification Table

Observed			Predicted		
			Financial Condition		Percentage
			0 (Strong)	1 (Week)	Correct (%)
Step 1	Financial Condition	0 (Strong)	7	3	70.0
		1 (Week)	0	10	100.0
	Overall Percentage				85.0

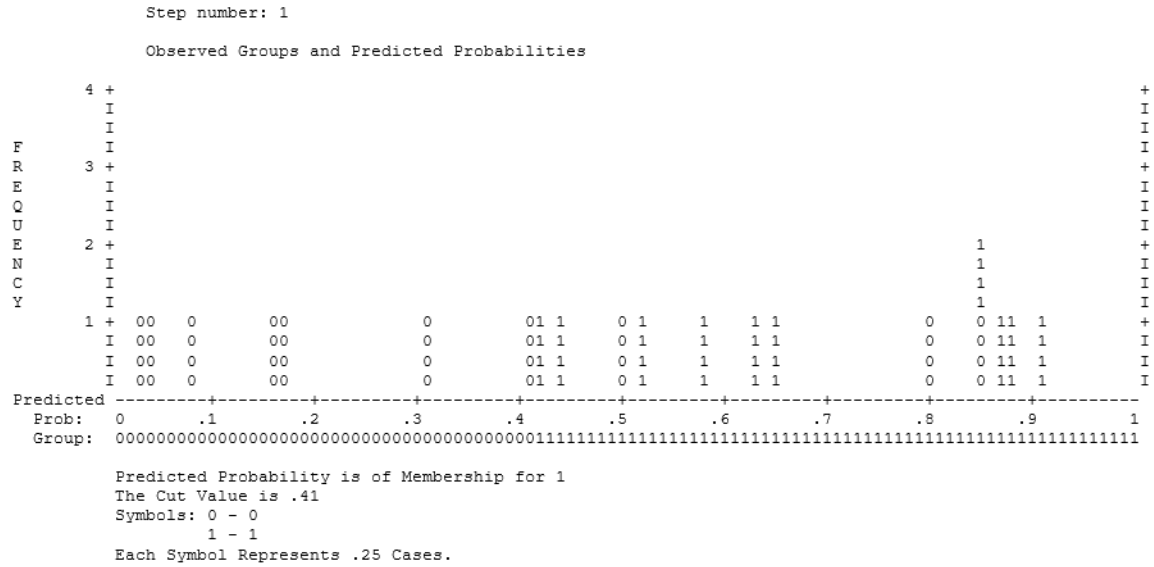
The cut value is 0.410
table 1.8
Source: Self-developed in SPSS

Summary:

Logistic regression was performed to ascertain the effects of "Total Loan & Leases to Total Assets Ration" and "Total Expenses to Total Assets Ration" on the likelihood that the financial condition of the bank is week. Initially, there were 22 observations, however, post cleaning of data, 20 observations were used for the analysis. The logistic regression model was statistically significant $\chi 2\ (2) = 7.937$, $p<0.05$. The model explained 43.7% (Negelkerke R2) of the variance in financial condition and correctly classified 100% of the cases with a cut off value of 0.41. Both the IVs are associated with the likelihood of being a financially week bank. Moreover, the likelihood of being a financially week bank increases by about 2751 (Exp (B), table 1.7) times for each 1 point increase in the "Total Expenses to Total Assets Ratio".

Appendix:

appendix 1: Observed and Predicted Probability Table



appendix 2: Financial Conditions of Banks Dataset (after cleaning of data)

Observation	*Financial Conditions*	*Total Expenses/Total Assets Ratio*	*Total Loan & Lease/Total Assets Ratio*
1	*0*	*0.09*	*0.3*
2	*0*	*0.08*	*0.43*
3	*0*	*0.1*	*0.51*
4	*0*	*0.08*	*0.53*
5	*1*	*0.13*	*0.53*
6	*0*	*0.11*	*0.55*
7	*1*	*0.16*	*0.57*
8	*0*	*0.15*	*0.57*
9	*1*	*0.11*	*0.62*
10	*1*	*0.16*	*0.62*
11	*1*	*0.12*	*0.65*
12	*1*	*0.1*	*0.67*
13	*0*	*0.07*	*0.69*
14	*1*	*0.11*	*0.69*
15	*0*	*0.1*	*0.71*
16	*0*	*0.09*	*0.73*
17	*1*	*0.13*	*0.74*
18	*1*	*0.1*	*0.79*
19	*0*	*0.12*	*0.8*
20	*1*	*0.09*	*1.02*

References:

Goel, S. (2016). *Financial Ratios (1st ed.)*. Business Expert Press

Kwak, S. & Kim, J. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, 70(4), 407–411. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548942/*

Landau, S. & Everitt, B. (2003). *A handbook of statistical analyses using SPSS (1st ed.)*. Chapman & Hall.

Loughnane, M. (n.d.). *Banks are businesses too: Look beyond the usual suspects when subpoenaing banks records.* https://www.acfe.com/fraud-examiner.aspx?id=4295000750

McFedries, P. (2013). *Excel data analysis visual blueprint* (4th ed.). John Wiley & Sons.

Mishra, P., Pandey, C., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, 22(1): 67–72. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423

Practical Applications of Statistics in the Social Sciences. (2014). *Simple Logistic Regression One Continuous Independent Variable.* *https://cdn.southampton.ac.uk/assets/imported/transforms/content-block/UsefulDownloads_Download/627FB8A9398F4DD38CA893046675DC43/PASSS%20RQ2%201Simple%20Logistic%20Regression%20-%20One%20Continuous%20Variable.pdf*

ReStore. (2011). *Using Statistical Regression Methods in Education Research.* https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/glossary/indexa039.html?selectedLetter=H#hosmer-and-lemeshow-test

Shmueli, G., Patel, N., & Bruce, P. (2016). *Data Mining for Business Analytics (3rd ed.)*. Wiley

Starmer, J. (2018). *Logistic Regression Details Pt1: Coefficients.* *https://www.youtube.com/watch?v=vN5cNN2-HWE&t=305s*

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 56(1): 45–50. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636062/