1. **Multicollinearity**

Before generates the model to make predictions of company bankruptcy, data pre-processing is necessary to mitigate the influence of bias to model, otherwise the model is not reliable (Chan et al., 2022). In this case, there are 95 financial features for each company, and there is an extremely high degree of similarity in some of the variables.

For solving potential multicollinearity issue, Drobnič et al. (2020) suggested that variance inflation factor (VIF) test is one of the most frequently used methods for the solution. Sundus et al. (2022) agreed with Drobnič et al. (2020), and applied VIF test to determine the correlation between independent variables in the regression model. Drobnič et al. (2020) and Srisa-An (2021) both recommend set VIF thresholds value as 10, it means that there are collinear variables if VIF value over 10. In this report, the threshold is settings are consistent with these scholars. Whereas O’brien (2007) found that VIF need to be determined in the context of features and factors in the regression model. It means that the threshold is set by researchers also influence the conclusion of multicollinearity measurement. Any researchers need to focus on the potential bias introduced by VIF.

Another solution of multicollinearity problem is correlation matrix. Srisa-An (2021) created a correlation matrix to measure the correlation between independent variables before data analysis for avoiding autocorrelation. P. Obite et al. (2020) also used correlation matrix to measure potential multicollinearity problem in the artificial neural network. In this test, high correlation coefficients suggest that high probability of multicollinearity in the independent variables.

Al-Mamun et al. (2020) employed Farrar-Glauber test for each model for determining potential multicollinearity issue among the independent variables, which is another great method to examining whether exist collinear variables. Muhammad Hussain et al. (2022) applied both VIF and Farrar-Glauber test in their research since there is no sufficient condition to determine multicollinearity. Similarly, Imdadullah et al. (2016) point out both Farrar-Glauber test and VIF have no unique or standard critical values in their measures, hence, the critical values are set by researchers can influence the results regard the multicollinearity.

Overall, in this report, correlation matrix, VIF and Farrar-Glauber test are applied in the determine of multicollinearity and feature selection. In the results of test, it shows that only 18 features can be remained in the subsequent model training and testing. It could increase the accuracy of bankrupt prediction model in this case.

*Step 1: Create a correlation matrix and generate a heatmap.*

1. ### Correlation matrix between independent variables (x1-x95) ###

2. train\_correlation\_matrix <- cor(df[, -1])

3.

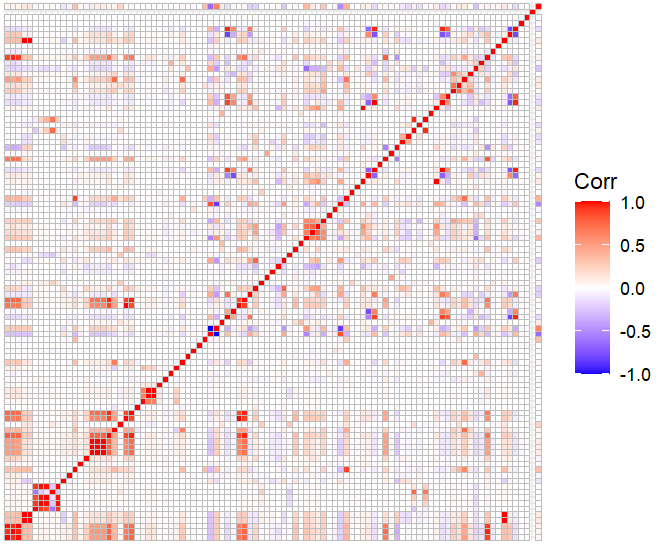
4. print(train\_correlation\_matrix)

5.

6. library(ggcorrplot)

7. ggcorrplot(correlation\_matrix)

**Chart 1. The heatmap between the independent variables in training set**



*Step 2: Find highly correlated variables then drop all similar features.*

1. ### Find highly correlated variables ###

2. threshold <- 0.9

3. train\_similar\_variables <- which(abs(train\_correlation\_matrix) > threshold & train\_correlation\_matrix != 1, arr.ind = TRUE)

4.

5. print(train\_similar\_variables)

6.

7. train\_highly\_correlated\_pairs <- data.frame(row = similar\_variables[, 'row'],

8. col = similar\_variables[, 'col'])

9.

10. print(train\_highly\_correlated\_pairs)

11.

12. for (i in 1:nrow(train\_highly\_correlated\_pairs)) {

13. row\_index <- train\_highly\_correlated\_pairs[i, 'row']

14. col\_index <- train\_highly\_correlated\_pairs[i, 'col']

15.

16. train <- train[, -col\_index]

17. }

18.

19. print(train)

*Step 3: Drop features with aliased issue and run VIF test.*

1. processed\_train <- processed\_train[, !names(processed\_train) %in% c('x77', 'x78')]

2.

3. train\_lmdf <- lm(y1~ ., data = processed\_train)

4. print(train\_lmdf)

5.

6. library(car)

7. train\_vif\_value <- vif(train\_lmdf)

8.

9. print(train\_vif\_value)

*Step 4: Drop features with VIFs over 10.*

1. ### Sort the variables with higher than 10 vif value ###

2. train\_variable\_to\_remove <- names(train\_vif\_value)[train\_vif\_value > 10]

3.

4. ### Remove the these variables

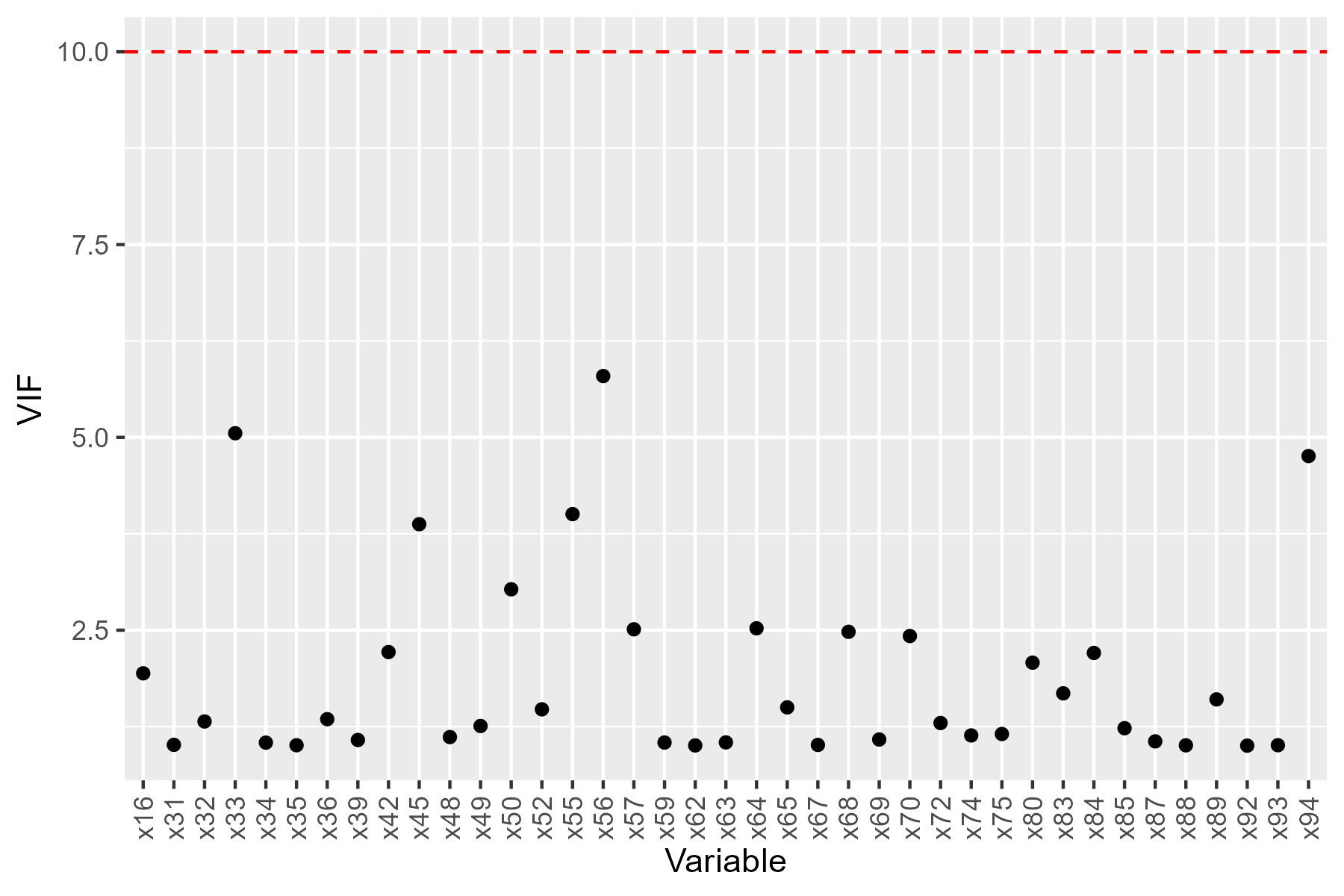
5. processed\_train <- processed\_train[, !names(processed\_train) %in% train\_variable\_to\_remove]

6.

7. train\_processed\_lmdf <- lm(y1~ ., data = processed\_train)

8. print(train\_processed\_lmdf)

Chart 2. The VIFs for each features after drop



*Step 5: Run Farrar-Glauber test as secondary testing to ensure model accuracy.*

1. ### Farrar-Glauber Test ###

2. library(mctest)

3. train\_omcdiag\_result <- omcdiag(train\_processed\_lmdf)

4.

5. train\_imcdiag\_result <- imcdiag(train\_processed\_lmdf)

6.

7. print(train\_omcdiag\_result)

8. print(train\_imcdiag\_result)

Al-Mamun, M.A., Brothers, T. and Newsome, A.S. (2020) ‘Development of machine learning models to validate a medication regimen complexity scoring tool for critically ill patients’, *Annals of Pharmacotherapy*, 55(4), pp. 421–429. doi:10.1177/1060028020959042.

Chan, J.Y.-L. *et al.* (2022) ‘Mitigating the multicollinearity problem and its Machine Learning Approach: A Review’, *Mathematics*, 10(8). doi:10.3390/math10081283.

Drobnič, F., Kos, A. and Pustišek, M. (2020) ‘On the interpretability of Machine Learning Models and experimental feature selection in case of Multicollinear Data’, *Electronics*, 9(5), p. 761. doi:10.3390/electronics9050761.

Imdadullah, M., Aslam, M. and Altaf, S. (2016) ‘Mctest: An R package for detection of collinearity among regressors’, *The R Journal*, 8(2), pp. 495–505. doi:10.32614/rj-2016-062.

Muhammad Hussain, N. *et al.* (2022) ‘Accessing artificial intelligence for fetus health status using hybrid deep learning algorithm (AlexNet-SVM) on Cardiotocographic Data’, *Sensors*, 22(14), p. 5103. doi:10.3390/s22145103.

O’brien, R.M. (2007) ‘A caution regarding rules of thumb for variance inflation factors’, *Quality & Quantity*, 41, pp. 673–690. doi:10.1007/s11135-006-9018-6.

P. Obite, C. *et al.* (2020) ‘Multicollinearity effect in regression analysis: A feed forward artificial neural network approach’, *Asian Journal of Probability and Statistics*, 6(1), pp. 22–33. doi:10.9734/ajpas/2020/v6i130151.

Srisa-An, C. (2021) ‘Guideline of collinearity - avoidable regression models on time-series analysis’, *2021 2nd International Conference on Big Data Analytics and Practices (IBDAP)* [Preprint]. doi:10.1109/ibdap52511.2021.9552165.

Sundus, K.I. *et al.* (2022) ‘Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset’, *Informatics in Medicine Unlocked*, 33, p. 101088. doi:10.1016/j.imu.2022.101088.