**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.

In bankrupt prediction case, any model training in the machine learning is highly variable sensitive, therefore, multicollinearity could decrease precision and estimates unreliable. Daoud (2017) and Chan et al. (2022) also believe that multicollinearity makes less explanatory of independent variables in the model. Apart from that, the model highly possible have poor generalisation ability and overfit the data issue if multicollinearity is presented in the model training, that is, it will perform poorly on testing set for bankrupt prediction.

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.

Overall, in this report, VIF test are applied in the determine of multicollinearity and feature selection. In the results of test, it shows that only 51 features can be remained in the subsequent model training and testing. It could increase the generalisation ability precision for prediction model.

*Step 1: Build logistic model without perfectly collinear variables.*

1. ### Build linear model

2.

3. train\_lm <- lm(Y ~ ., data = train)

4.

5. # Check for perfectly collinear independent variables

6. # Remove them as they cause an error in the VIF calculation

7. ld.vars <- attributes(alias(train\_lm)$Complete)$dimnames[[1]]

8.

9. # Create new formula

10. formula.new <- as.formula(paste("Y ~.", paste(ld.vars, sep="-"), sep= "-"))

11.

12. # Create new linear model without perfectly collinear variables

13. train\_lm <- lm(formula.new, data = train)

*Step 2: Compute VIF values and select variables with VIF > 10.*

1. ### Compute VIF values ###

2. library(car)

3. train\_vif\_value <- vif(train\_lm)

4.

5. # Plot VIF values for each independent variable

6. print(train\_vif\_value)

7.

8. # Select variables with VIF > 10

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

10.

11. ### Remove these variables

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

13.

14. # Reset original variables names

15. colnames(train) <- original\_names[temporary\_names %in% colnames(train)]

*Step 3: Build a new logistic model with available variables and generate VIF chart.*

1. train\_lm <- lm(formula.new, data = train)

2.

3. train\_vif\_value <- vif(train\_lm)

4.

5. vif\_table <- data.frame(variable = names(train\_vif\_value), vif = train\_vif\_value)

6.

7. library(ggplot2)

8. p <- ggplot(vif\_table, aes(x = variable, y = vif)) +

9. geom\_bar(stat="identity") +

10. geom\_hline(yintercept = 10, linetype = "dashed", color = "red") + # VIF

11. xlab("Variable") +

12. ylab("VIF") +

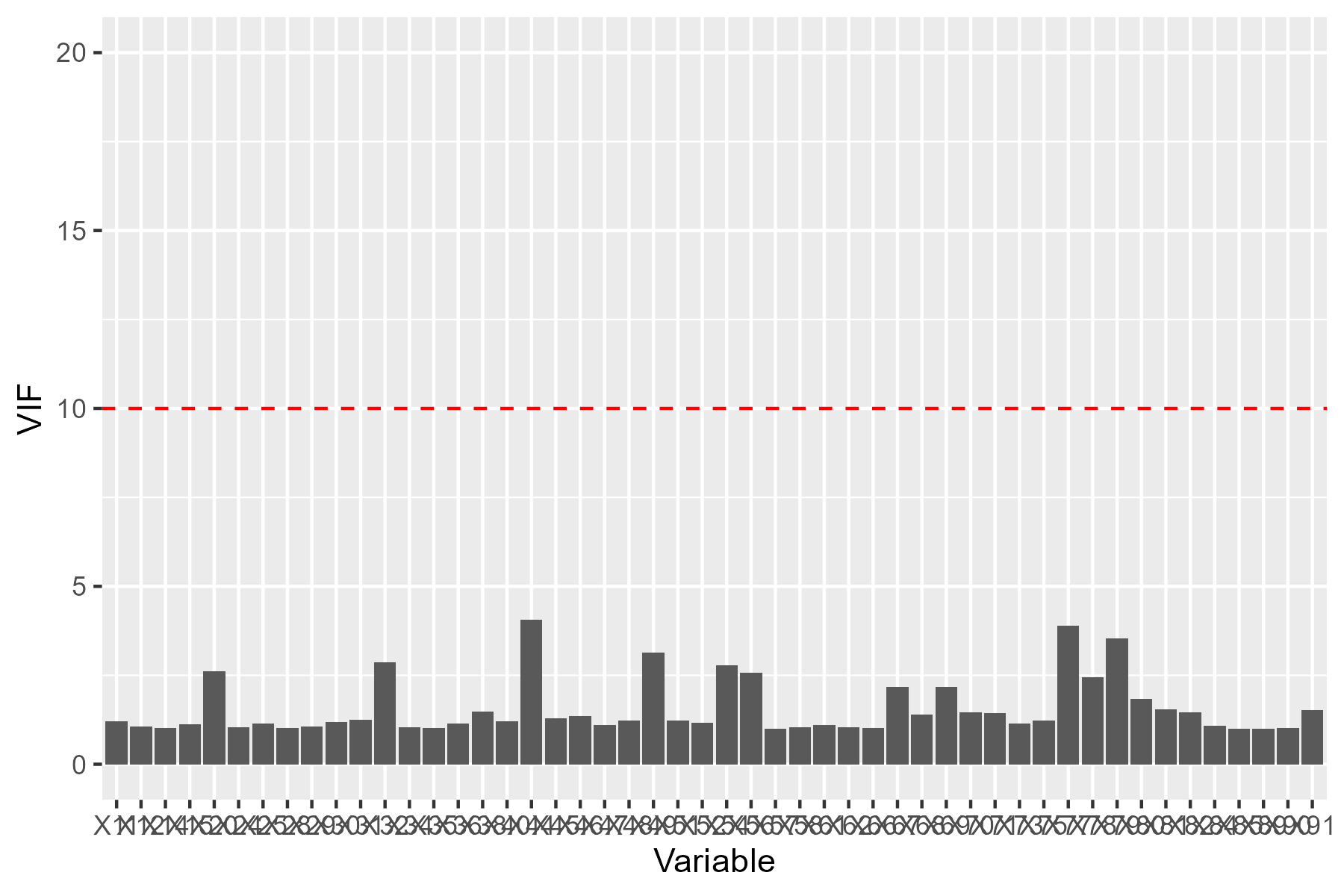
13. ylim(0, 20)

14. theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

15.

16. ggsave('D:/University of Birmingham/Big data management/Group assignment/plot\_vif.png', plot = p, width = 6, height = 4, units = "in")

Chart 1. The VIFs for each feature after drop.



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.

Daoud, J.I. (2017) ‘Multicollinearity and Regression Analysis’, *Journal of Physics: Conference Series*, 949. doi:10.1088/1742-6596/949/1/012009.

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.

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.

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.