***Part I: Data Preparation***

Graphical user interface

Description automatically generated with low confidence

Since the fraud dataset is a csv file, to analyse the data in SAS, PROC IMPORT must be used to import the data into SAS.

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Since our objective is to evaluate on Dechow et al. (2011) model’s predictive power and effectiveness in predicting fraud, we constructed the variables required to perform the Dechow et al. (2011)’s predicting model.

Table

Description automatically generated

Before constructing variables, the fraud data has 125095 observations. However, to construct the variables, company’s previous two years data are also needed and thus many observations are deleted in the process of using SQL self-joining code to construct the variables. As a result, number of observations in fraud dataset significantly decreased to 96245.

Graphical user interface, text

Description automatically generated

Table

Description automatically generated

Since observations with missing constructed variables are not useful in predicting accounting fraud according to Dechow et al (2011)’s model, do loop is applied to delete those observations to prevent those observations from interfering our prediction. After deleting observations with missing variables, number of observations in dataset decreased to 71036. However, some variables have huge difference between mean and median which indicates that existence of extreme values in the dataset. Therefore, the dataset must be winsorized to minimize the extreme values’ impact on our data analysis’s results.

Graphical user interface, text, application, email

Description automatically generated

This SAS macro is created to winsorize variables’ extreme values at top and bottom 1% and then it is applied to the FRAUDRAW4 dataset. This step is necessary because as mentioned above, the huge difference between mean and median for some variables signals existence of extreme values which may potentially adversely impact our analysis of the dataset. Therefore, the dataset must be winsorized to minimise the impact of variables’ extreme values on data analysis’s results.

Table

Description automatically generated

After winsorizing the dataset, the difference between median and mean for the variables narrowed and therefore signalling winsorization’s effectiveness in limiting extreme values’ impact on our data analysis.

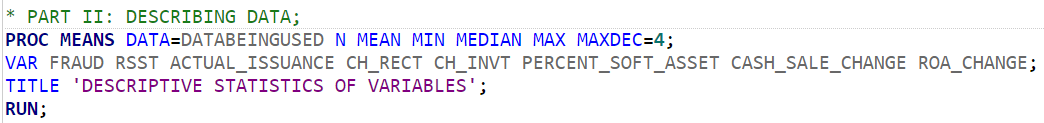
Graphical user interface, text, application, email

Description automatically generated

The winsorized fraud dataset is separated into two datasets according to existence of frauds in those observations. For the dataset that includes no fraud observations, we use RANUNI function to randomly select 20000 observations and merge that dataset with the dataset with only fraud observations. The resulting dataset will be the dataset we use to evaluate Dechow et al. (2011) model’s effectiveness in predicting accounting fraud.

***Part II: Describing Data***

*Descriptive Statistics of the Variables*



Table

Description automatically generated

PROC MEANS function is used to generate the descriptive statistics of important variables in the following steps. Fraud has a mean of 0.0168 means that only 1.68% of observations have fraud and the median of 0 means the median observation has no fraud.

RSST’s mean of -0.0109 indicates on average, firm’s RSST accruals decrease by 1.09% of average total assets whilst RSST’s median of 0.015 indicates the median observation’s firm’s RSST accruals increase by 1.5% of average total assets.

ACTUAL\_ISSUANCE’s mean of 0.8587 means that 85.87% of observations’ firms have either issued new shares or borrow money to raise funds and the median of 1 means the median observation’s firm has either borrowed money or issued new shares to raise funds.

CH\_RECT’s mean of 0.0070 means on average, firm’s receivables increase by 0.7% of average total assets whilst CH\_RECT’s median of 0.0032 means the median observation’s firm’s receivables increase by 0.32% of average total assets.

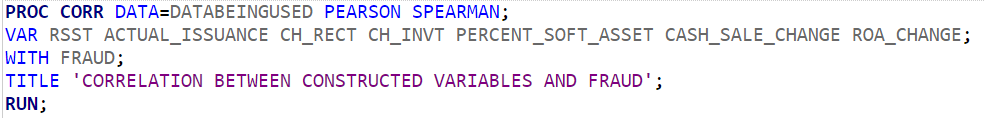
CH\_INVT’s mean of 0.0037 means on average, firm’s inventory level increase by 0.37% of average total assets whilst CH\_INVT’s median of 0 means the median observation’s firm’s inventory level does not change.

PERCENT\_SOFT\_ASSET’s mean of 0.5261 means on average, 52.61% of firm’s assets are soft assets whilst PERCENT\_SOFT\_ASSET’s median of 0.5480 means 54.8% of median observation’s firm’s assets are soft assets.

CASH\_SALE\_CHANGE’s mean of 0.1553 means on average, firms’ cash sales increase by 15.53% compared to previous year whilst the median of 0.0505 means the median observation’s firm’s cash sale only increased by 5.05% compared to previous year.

ROA\_CHANGE’s mean of -0.0028 means on average, firms’ ROA decrease by 0.0028 compared with previous year whilst ROA\_CHANGE’s median of 0.0004 means the median observation’s firm’s ROA increases by 0.0004 compared with previous year.

*Examine Correlation Between the Variables*



Since this project’s main objective is predicting accounting fraud using Dechow et al. (2011) model and evaluate on its effectiveness in predicting accounting fraud, Pearson and Spearman correlation between fraud and various constructed variables are examined.

Graphical user interface, text, application, email

Description automatically generated

For both Pearson and Spearman correlation results, since p-values of receivables change, change in inventory level, changes in cash sale and change in ROA are greater than 0.1, the null hypothesis that there is no correlation between these variables and fraud cannot be rejected at 10% significance level. However, for RSST accruals, actual issuance, and percentage of soft assets, since their p-values are less than 0.05, we are 95% confident to reject the null hypothesis that there is no correlation between these variables and fraud and conclude that their correlation coefficient is not zero, but the correlation coefficient values on the tables.

Although RSST accruals, actual issuance and percentage of soft assets are shown to be statistically correlated with fraud, the low Pearson correlation values and low Spearman correlation values indicate a weak correlation between these variables and occurrence of fraud.

According to Dechow et al. (2011), the model used to predict accounting frauds using the constructed variables is a logistic regression model. Therefore, the relationship between fraud’s log odd ratio and constructed variables is not linear, but as constructed variables increase, likelihood of fraud occurring increases. Therefore, it is expected that spearman correlation coefficient to demonstrate a stronger relationship. In fact, the result shows that both Pearson and Spearman correlation indicates a weak correlation between constructed variables and occurrence of fraud. Therefore, this is unexpected. However, considering that fraud is a binary variable which may affect the effectiveness of using spearman correlation to explain the correlation between occurrence of fraud and constructed variables, this correlation coefficient may be reasonable.

*Examine Constructed Variables Correlation with Fraud Using T-TEST*

Graphical user interface, text, application

Description automatically generated

As demonstrated by correlation results, only RSST accruals, actual issuance, and percentage of soft assets are statistically correlated with fraud, therefore T-Test is only applied to these variables to examine the effect of these variables on occurrence of fraud.

Table

Description automatically generated

The TTEST procedure shows that we are 99% confident that on average, misstate firms’ RSST accrual is higher than no-misstate firm’s RSST accrual by 5.17% of its average total assets assuming all else equal.

Table

Description automatically generated

The TTEST procedure demonstrates that we are 99% confident that on average, misstate firms are 5.75% more likely to borrow money or issue new shares to raise funds when compared to no-misstate firms holding all else equal.

Table

Description automatically generated

The TTEST procedure demonstrates that we are 99% confident that on average, misstate firms’ proportion of soft assets are higher than no-misstate firms by 15.05% of total average assets.

***Part III: Logistic Regression***

*Estimating Logistic Regression and Examine Regression Results*

A picture containing text

Description automatically generated

By using PROC LOGISTIC to estimate Dechow et al. (2011)’s logistic regression model, the following result is produced.

Table

Description automatically generatedTable

Description automatically generated

Since convergence criterion is satisfied and the p-value for global null hypothesis is less than 1%, the null hypothesis that the model has no predictive power can be rejected at 1% significance level and therefore concluding that at least one of the constructed variables’ regression coefficients in the logistic model is not equal to zero.

Considering that the p-values of change in receivables, change in inventory, cash sale change and ROA change are greater than 10%, the null hypothesis that these variables’ regression coefficients are zero cannot be rejected at 10% significance level. Therefore, these variables are not statistically significant in estimating the logistic regression to predict accounting fraud. However, since we are applying Dechow et al. (2011) model to our dataset, these variables will still be used to estimate the logistic regression.

On the other hand, since the intercept, RSST accruals, actual issuance, and percentage of soft assets have a p-value lower than 5%, we are 95% confident to reject the null hypothesis that their regression coefficients are zero and conclude their regression coefficient to be the one calculated by the above logistic regression model at 5% significance level.

Therefore, according to our estimate, the formula for odds ratio is:

.

Therefore, .

The odds ratio estimates results show that percentage of soft asset has the strongest positive association with occurrence of fraud with actual issuance coming second and RSST accruals coming third. On the other hand, change in receivables seems to be negatively associated with occurrence of fraud despite it being statistically insignificant in estimating the logistic model. As for the other variables, since their odd ratio estimates are close to 1, they have weak association with occurrence of fraud.

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

The above code first calculates the F-score of the observations and separate the observations into 5 groups according to magnitude of their F-scores. Results are then constructed and presented in Dechow et al. (2011)’s Table 7 Panel B’s format. Results are presented below.

Table

Description automatically generated

The above results provide us some insights into the effectiveness of Dechow et al. (2011)’s model. If the model is effective in predicting accounting frauds, then the fraudulent observations should be gathering at the highest quintile (RANK\_F\_SCORE=4) where top 20% of observations with highest F-scores are gathered. From the results, only 35.67% of observations with fraud are observed in the highest quintile. Although it is much higher than the expected level of 20%, this model is a lot less effective in this data than Dechow et al. (2011)’s dataset which has 51.01% of misstatement firms in the highest quintile.

In addition, the results show that for the highest rank, the minimum F-score for misstate firms is 1.553 whilst the minimum F-score for no-misstate firms is 1.5536, the gap between these two values are minimal. In contrast, Dechow et al. (2011)’s result demonstrates that for quintile 5, the cut-off F-scores for misstate firms and no-misstate firms are 1.397 and 0.933 respectively, the gap between these two values is significant. This indicates that Dechow et al. (2011) model works better in Dechow et al. (2011)’s dataset in terms of differentiating misstate firms from no-misstate firms using F-score.

However, the higher proportion of misstate firms at higher rank is still consistent with the expectation that the higher the predicted probability of a firm committing a fraud, the higher the chance that a company is committing a fraud. This is also a solid proof of some predictive power of Dechow et al. (2011) model in our dataset.

*Reasons For Variation from Dechow et al. (2011)’s Results*

The variation of Dechow et al. (2011) model’s performance in Dechow et al. (2011)’s dataset and this dataset may be caused by the changes in the way accounting fraud is implemented. Since Dechow et al. (2011)’s dataset only includes observations from 1982 to 2005 whilst our dataset includes observations from 2000 to 2018. Considering that when accounting fraud was exposed, people become more aware of characteristics of misstate firms. Therefore, firms have to implement accounting frauds with new methods and thus leading to misstate firms having different characteristics. As a result, Dechow et al. (2011) model which is effective in predicting accounting frauds during 1982 to 2005 will be less effective in predicting accounting frauds happened in 2000 to 2018.

*Selecting Cut-off Points and Discuss Predictive Power of the Model*

A picture containing text

Description automatically generated

When estimating the logistic regression, a classification table and a ROC curve have also been created. The classification table and ROC curve help determines an appropriate cut-off point and analyze predictive power of the Dechow et al. (2011) model.

*Classification Table Results*

Table

Description automatically generated

To determine the cut-off point, nature of event must be considered. Since accounting fraud is extremely rare and failure to predict an accounting fraud when there is one is going to lead to investors suffering tremendous losses. Therefore, type II error must be minimized to protect shareholders and investors interest. However, accuracy is also a very important factor to consider since 0 type II error can be easily done by setting probability level to 0 and predicting all observations have an accounting fraud. This result is obviously meaningless. Therefore, the cut-off point must be set at a point which balances accuracy and sensitivity.

Considering that if we fail to predict an accounting fraud (type II error), investors will be suffering tremendous losses whilst the consequence of predicting an accounting fraud when there is none (type I error) is merely investors losing an opportunity cost to invest in a good stock when there are so many good stocks on the market. Therefore, the loss caused by type I error is trivial compared to loss caused by type II error and thus sensitivity must be high while maintaining accuracy and specificity at a reasonable level.

Therefore, from the classification table, the cut-off point could be set at 0.01 which has a sensitivity of 92.1%, accuracy of 35.3% and specificity of 34.1%. By using a cut-off point of 1%, 92.1% of accounting frauds will be correctly predicted, type II error rate will be maintained at 7.9%, type I error rate will be 65.9% and the accuracy of the model’s prediction will be 35.3%. Although the accuracy of this model’s prediction is relatively low, the risk of failing to predict an accounting fraud is also low, therefore, picking 0.01 as cut-off point minimizes investors risk of failing to predict an accounting fraud while ascertaining investors can benefit by applying this model to predict accounting fraud.

*Predictive Power of the Model*

Table

Description automatically generated

Chart, line chart

Description automatically generated

Area under the ROC curve is often used to evaluate the predictive power of a model. As a rule of thumb, for a model to be considered to possess acceptable predictive power, the area under the ROC curve must be at least 0.7. However, in this case, the area under the ROC curve is only 0.6764. Therefore, the predictive power of Dechow et al. (2011) model on this dataset is not considered acceptable.

***Part IV: Model Selection and Back Testing***

*Using Stepwise Logistic Regression to Select Model to Predict Accounting Fraud*

Text

Description automatically generated

Table

Description automatically generated

By using the above code, all independent variables that are not statistically significant at 5% significance level will be excluded from estimating the logistic regression model. The result shows that there are 4 steps in reaching the summary of stepwise selection, so there could be 4 models and for the purpose of model selection, the one with strongest predictive power will be selected.

Table

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated

Since model produced in step 0 has a comparatively greater model fit statistics and thus signalling that it has the weakest predictive power. However, the models produced in step 1, step 2 and step 3 have similar model fit statistics which indicate similar predictive power, therefore, model cannot be selected based on model fit statistics.

Chart, line chart

Description automatically generated

Area under the ROC (AUC) curve is also an excellent tool used to evaluate model’s predictive power. Since the model produced in step 3 has the highest AUC, this model has the strongest predictive power among the four models.

However, the more variables to be included in a model, the more complex the model is and therefore will be harder to be implemented, especially when analyzing a large dataset which is usually the case for predicting accounting frauds. Since model with highest predictive power only has 3 variables and thus is not complex. Therefore, this model is suitable for predicting accounting fraud.

Therefore, the logistic regression model includes variables of RSST accruals, actual issuance, and percentage of soft assets. However, it is worth noting that since the AUC is below 0.7, the model’s predictive power is not satisfactory, this model is merely the best model amongst the four.

*Back Testing of Selected Model*

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

This code first uses RANUNI function to randomly select 10 years of data into estimation sample and the rest of the data into the test sample. The selected logistic regression model is then estimated using the estimation sample and its logistic regression coefficients are then output to EST1 and is merged with the test sample. F-score of the observations in the test dataset are then calculated and the cut-off point is selected to be F-score greater than 1 and thus all observations with a F-score greater than 1 is predicted to be misstate firms. The accuracy and sensitivity of the selected model in the test dataset are then calculated and are output to the dataset PERFORMANCESTAT. The above process is repeated 100 times to gives us a larger dataset and thus allowing us to better assess the performance of the selected model.

*Back Testing Results and Discussions*

Table

Description automatically generated

The results show that on average, this model has an accuracy of 57.82% which means that on average, this model has a 57.82% chance of predicting whether a firm has fraud right. Since the model’s median accuracy is greater than the model’s mean accuracy, the distribution of the model’s accuracy is left skewed and thus there are more observations with an accuracy greater than 57.82% and thus if we take a larger sample, the mean accuracy may be even higher.

In addition, the result also shows that the model’s mean sensitivity is 67.84% which means that on average, this model correctly predicts 67.84% of misstate firms. However, this also means that the type II error of this model is 32.16%. Considering the consequences of failing to correctly predict an accounting fraud will lead to investors suffering tremendous losses, the type II error of 32.16% is considerably too high. Also, considering that the market has many good companies, thus the opportunity cost of not investing in a company because of inaccurate accounting fraud prediction is low.

Therefore, for the best interest of investors, it is reasonable to sacrifice accuracy for lower type II error. Thus, the cut-off points of F-score greater than 1 is too high and must be lowered to lower type II error and increase sensitivity while maintaining a reasonable accuracy.

(Word Count: 2964 words)

***Reference List:***

Dechow, P.M., Ge, W., Larson, C.R., & Sloan, R.G. (2011). Predicting Material Accounting Misstatements. *Contemporary Accounting Research, 28*(1), 17-82.