## Contents



# 1 Model Output

The three classification models logistic regression, decision tree and XG-Boost, as described in the previous chapter, were optimised using tuned parameters, tuned weights and empirical threshold values. The optimised models were then trained to develop relationships using the data in the training dataset. Thereafter, the testing dataset is utilised to examine the performance of the trained model in predicting defaults.

Using the cost related evaluation metrics, comparison of both CSL techniques is shown in table 11:

*Table 1: Model Performance Cost related metrics*

| Model | CSL Technique | Average Misclassification Cost (AMC) | MMCE | Total Costs |
|---|---|---|---|---|
| **Logistic Regression** | ***CSL – R*** | 433.88 | 0.111 | 44,74.51 |
| | ***CSL – A*** | 491.57 | 0.140 | 44,91.49 |
| **Decision tree** | ***CSL – R*** | 509.41 | 0.148 | 46,54.50 |
| | ***CSL – A*** | 568.28 | 0.125 | 51,56.15 |
| **XG-Boost** | ***CSL – R*** | 599.90 | 0.119 | 54,81.32 |
| | ***CSL – A*** | 469.27 | 0.136 | 44,64.61 |

*Note: mmce = Mean Misclassification error; total costs in thousands USD*

Based on the average misclassification cost, logistic regression demonstrates a good performance in both CSL methods. CSL – R logistic regression model shows the least cost against all models. The least AMC is 433.88. It shows logistic regression is able to learn patterns and correctly predict defaults.

Analysing CSL-A models, XG-Boost shows the least costs overall, with AMC = 469.27 and MMCE = 0.136.

Although literature suggests tree based models are well-equipped in cost sensitive tasks, they have a higher misclassification cost in both CSL techniques.

*Table 2: Model Performance with overall evaluation metrics*

| Model | CSL Technique | Recall | Precision | F1 | G-mean | Kappa | Accuracy | AUC |
|---|---|---|---|---|---|---|---|---|
| **Logistic Regression** | **CSL – R** | 0.741 | 0.516 | 0.660 | 0.625 | 0.527 | 0.878 | 0.811 |
| | **CSL – A** | 0.748 | 0.511 | 0.607 | 0.810 | 0.525 | 0.859 | 0.813 |
| **Decision tree** | **CSL – R** | 0.752 | 0.494 | 0.596 | 0.808 | 0.403 | 0.851 | 0.810 |
| | **CSL – A** | 0.505 | 0.590 | 0.544 | 0.689 | 0.473 | 0.876 | 0.723 |
| **XG-Boost** | **CSL – R** | 0.410 | 0.643 | 0.500 | 0.627 | 0.436 | 0.880 | 0.686 |
| | **CSL – A** | 0.748 | 0.514 | 0.609 | 0.811 | 0.528 | 0.860 | 0.840 |

Examining Table 12, the overall performance metrics of the models. Although, all models have a similar accuracy between 0.85 and 0.88, its recall, precision and AUC varied values.

Detailed discussion by each model is given below:

## 1.1 Logistic Regression

Comparing both the CSL techniques in logistic regression model, both show similar results. Logistic regression has a higher recall values (i.e. 0.741 and 0.748) than other models. It suggests logistic regression is able to predict defaults and have a better pattern recognition using CSL techniques. Although the performance metrics between both models (CSL – R and CSL – A) are very close to one another in Table 12, CSL – R has a lower AMC, MMCE and total cost in Table 11. Therefore, CSL – R technique has overall superior results with lower costs than CSL – A in the logistic regression algorithm.

## 1.2 Decision Tree

Comparing both the CSL techniques in decision tree models, both show different results. CSL – R show better performance than CSL – A in decision trees. From Table 12, CSL – R recall 0.752 is higher than CSL – A recall value of 0.505. It is interesting to note CSL – A has a higher precision value 0.590 than CSL – R value 0.494. This suggests CSL – R has a higher false negatives and CSL – A model has a higher false positives. In the case of default prediction, having higher false positives is not ideal in the real-life application. Hence, CSL – R decision tree model is better than CSL – A model. Additionally, CSL – R model has a higher AUC value 0.810, suggesting a better model performance in distinguishing positive and negative classes.

CSL – R decision tree model also has a lower AMC and MMCE from Table 11, making it a better model than CSL – A decision tree.

### 1.3 XG-Boost

Both CSL XG-Boost models show similar characteristics to CSL decision tree models. Just like decision tree models, XG-Boost CSL – R has a higher precision value and XG-Boost CSL – A has a higher recall values. Hence, CSL – A has lower false positives, which is ideal for default prediction dataset. XG-Boost CSL – A models also showcases a higher AUC 0.840, which is highest among all the models in Table 12. From Table 11, XG-Boost CSL – A model has lower misclassification cost (AMC = 469.27).

To achieve the research objective in finding the most effective CSL technique, in Table 11, CSL – R logistic regression and CSL – A XG-Boost models have lowest misclassification costs, and in Table 12 have a better overall model performance on other evaluation metrics. CSL – R technique has lower costs with logistic regression and decision trees. While CSL – A performs better with XG-Boost model. CSL – R have the lowest average misclassification cost (AMC) and misclassification errors, making it the best performing CSL model in this study.

In terms of the second research objective, to find the best model for default prediction, it is crucial to compare model's performance in predicting defaults correctly. CSL – R logistic regression and CSL – A XG-Boost models are the top performing models from Table 11 and Table 12. Both models have a same recall values of 0.748. It is interesting to understand the confusion matrix of each model to distinguish them. Table 13 & Table 14 shows confusion matrix of the respective models.

*Table 3: Confusion Matrix of Logistic Regression CSL – R Model*

| | **Actual Defaults** | **Actual non-defaults** |
|---|---|---|
| **Predicted defaults** | 989 | 925 |
| **Predicted non-defaults** | 345 | 6878 |

*Table 4: Confusion Matrix of XG-Boost CSL – A Model*

| | **Actual Defaults** | **Actual non-defaults** |
|---|---|---|
| **Predicted defaults** | 999 | 944 |
| **Predicted non-defaults** | 335 | 6859 |

From Table 14, it is evident CSL – A XG-Boost model has a lower false positives 335. Although the false negatives is higher in CSL – A XG-Boost model than CSL – R logistic regression model. In reality, having higher false positives will increase debt exposure for the organisation and hence it is ideal to keep it at the lowest possible. Therefore, CSL – A XG-Boost model is an ideal model for loan prediction using the CSL technique.

Furthermore, to understand model's performance in differentiating positive and negative classes (i.e. defaults and non-defaults), ROC plot for both the models is shown in figure 9 below:

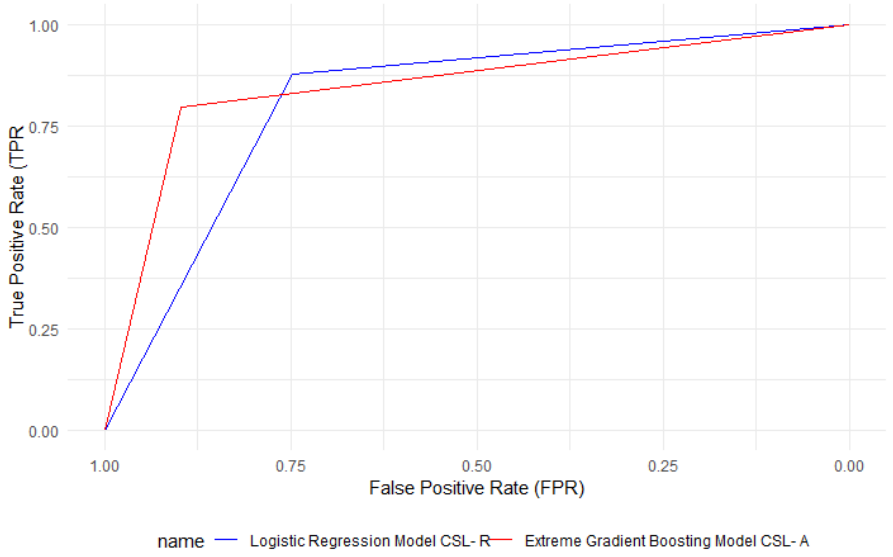


*Figure 1: ROC Plots*

CSL-A XG-Boost has a better ROC curve towards the top left area of the plot, making it the best performing model.