Vinnie

Thanks Joe.

As mentioned from Josh and Joe, our group has performed five different methods to obtain the potential best model.

Now let’s make a comparison of all the methods. We will be comparing the model by looking at the confusion matrix, ROC and the Gini Score.

Firstly we have StepAIC. StepAIC has a high rate of predicting a default and high Gini Score of 0.408. However the prediction is biased to the non-defaulted side, meaning having low prediction on the defaulted subjects.

Next we have Random Forest Method

* Also having high rate of predicting a default, RF is slightly less bias compared to StepAIC with a high Gini Score of 0.397.

Next up we have Lasso

* It has similar rate of predicting a default as RF. But a slightly higher Gini Score: 0.401

Due to the bias of the data, upsampling and downsampling have been carried out. As one can see that the bias in the confusion matrix has been solved. However, the Gini Value has decreased by a small amount compared to the other methods.

Now let’s put the results together to see which one is the best.

The ROC graphs of the different methods has been shown with their respective colors. As mentioned, all the methods have similar AUC value, therefore we can’t really get any insights from the ROC graph. So we will need to look into their respective accuracy which calculated from the confusion matrix. From the table we can see that StepAIC gives the highest accuracy on predicting both defaulted and non-defaulted by having 84.8% of accuracy and at the same time, the highest auc score of 0.7.

Thus, the model from StepAIC has been chosen as the most suitable model.

From the results, the model indicates that client who have longer loan terms have a higher risk to default. Besides, clients who have lower grade tends to have a higher chance to default, as it supports that the clients might be graded according to the pass loaning behaviour or income.

Furthermore, client who loan the money for small business operation has a higher chance to default compared to the others. This might be due to the reason where small business is hard to make profit. Not surprisingly, the result has shown that people who has higher income will have a lower chance to default. Variables such as inquiry in the last 6months, number of public record, revolving utilization, employment length as well as the credit length also helps in indicating clients that have potential to default.

Conclusion

In conclusion, As from the graph we can see that the current model is doing a lot better than the previous model. Although the accuracy is marginally reached 85%, the high gini score confirms the adequacy as well as the superior performance of the model. Besides, the accuracy might be affected by other factors such as data bias.

Team has spent most of the project time on doing critical analysis as we believed that with a properly cleaned and structured data will help to ease out the job for modelling. Research papers were referred when strategically grouping variables and handling missing observations.

Recommendation:

Before ending the presentation, a few recommendations being proposed to improve the model. Firstly, we would like to work more on classifying data with more extensive method

If time is allowed, Reject Inference method can be incorporated in the modelling process to obtain a better model.Moreover, a better treatment for missing data will help improve the current model. Last but not least, we would recommend bank take further action to make a decision about what threshold is in their best interest from a business perspective. For instance, by having model that predicts non defaulted ensures earning for the bank whereas the model which predicts defaulted helps to avoid loss of money. However, as business risks and factors were not included this was outside of the scope of this project.