

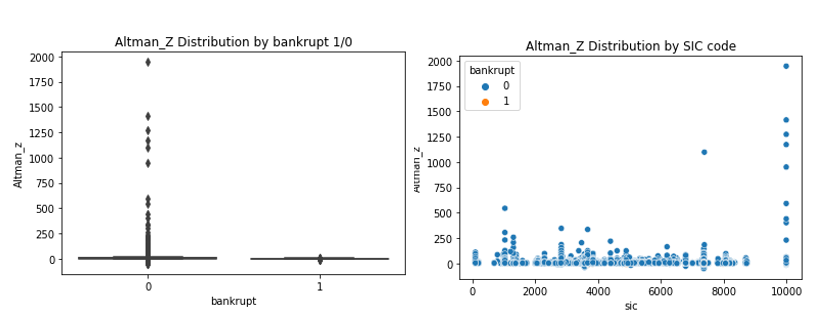
BT4016 Assignment 2

Kaustubh Jagtap (A0168820B)



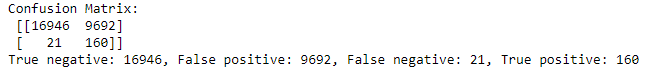
national university of singapore

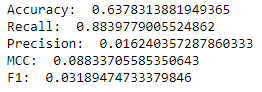
# EDA



From the charts, it is clear that there are some SIC's which exhibit higher bankruptcy rates. The statistical distribution of Altman's Z-score is also very different for companies which went bankrupt, versus those that didn't.  
Digging deeper, the mean Altman's z-score for bankrupt=1 is 0.565, and the median is 0.773, whereas for bankrupt=0, the mean and median are 6.554 and 3.678 respectively - clearly, Altman's is a decent guage of predicting bankruptcy.  
However, I foresee that there will be many false positives just by using Altman's Z-score, since there are many good companies (bankrupt=0) with a low score - as can be seen by the 25th percentile at 2.07, which is below our threshold cutoff of 2.675.

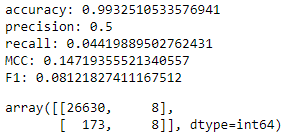
# Question 1



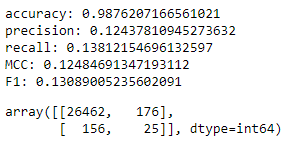


# Question 2

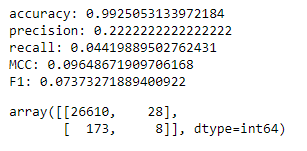
**LOGISTIC REGRESSION**



**CART**

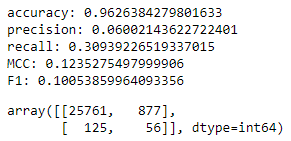


**XGBOOST**

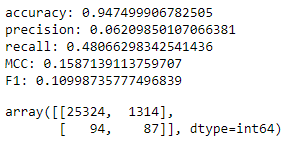


# Question 3

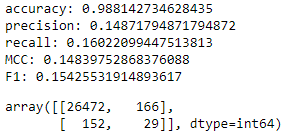
**CART, SMOTE ONLY**



**CART, SMOTE + ENN**



**CART, OVERWEIGHT CLASS WEIGHT**



Smote is not that good at improving the F1 score, since the plain CART model had a score of 0.131, as opposed to 0.101 and 0.110 for SMOTE and SMOTE+ENN respectively. Overweighting the class weight within the CART as a hyperparameter, however, did improve the F1 score to 0.154.

# Question 4

For this question, I built 2 models.

The first model was with feature engineering – I added in the following columns:

1. Boolean for dividend paid or not
2. Ratio of dividend to earnings ratio for each security
3. Ratio of operating expenses to revenue for each security

The rationale for including these was that they are not addressed in the Altman’s Score. I also included the one-hot-encoded first SIC (first digit only, since the EDA tells us only first digits make a difference).

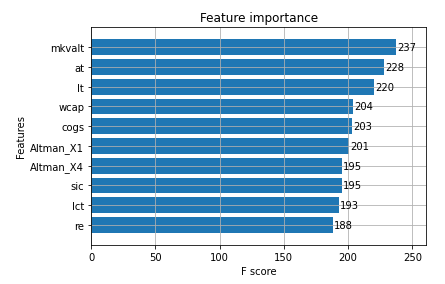
The second model was a plain vanilla one with all the features, and one-hot-encoded first digit of SIC score. (since based on EDA, only the first digit makes a significant difference). I used pos\_scale\_weight to balance the classes. This model was primarily to comply with the grading criteria of not being allowed to do feature engineering.

To tune the hyperparameters, I first visualized the ROC\_AUC across different folds, and from this I arrived at a narrow range of values for which to tune. For tuning, I used Bayesian Hyperparameter Optimization, with 5 folds. The ROC\_AUC score achieved during cross\_val on training set was 0.939.

The test predictions are attached in a csv file.

It is also worth mentioning that in a real credit investment scenario, the consequence of a false negative (i.e. predict bankrupt = 0 but it actually goes bankrupt and we lose all out money), is much more severe than the consequence of a false positive, where we predict that a company is going to go bankrupt, and avoid investing in it. Hence, we aim for a high recall, but can be more lenient in our precision metric.

# Question 5



None of the top 5 are Altman’s ratios. Rather, the top 5 comprise of the market value, total assets, total liabilities, working capital and cost of goods sold. These figures generally increase as the size of the firm increases 🡺the model presumably is picking up the trend that larger companies tend to default less frequently than smaller ones.

The 6th and 7th most important features, however, are the X1 and X4 ratios.