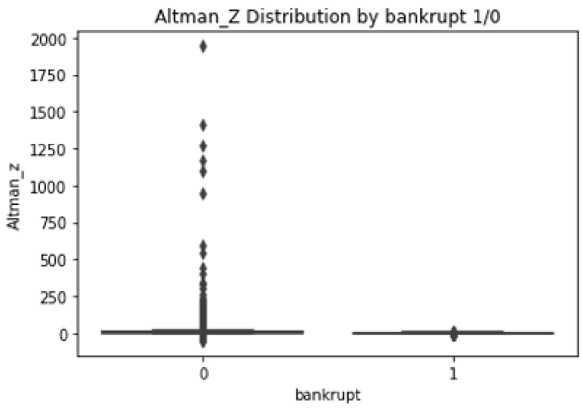
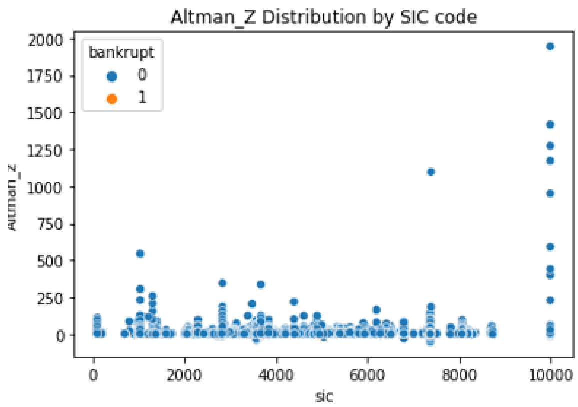
BT4016 ASSIGNMENT 2

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

```
Confusion Matrix:
 [[16946  9692]
 [   21   160]]
True negative: 16946, False positive: 9692, False negative: 21, True positive: 160

Accuracy:  0.6378313881949365
Recall:  0.8839779005524862
Precision:  0.016240357287860333
MCC:  0.08833705585350643
F1:  0.03189474733379846
```

Question 2

<u>LOGISTIC REGRESSION</u>

```
accuracy: 0.9932510533576941
precision: 0.5
recall: 0.04419889502762431
MCC: 0.14719355521340557
F1: 0.08121827411167512

array([[26630,     8],
       [  173,     8]], dtype=int64)
```

CART

```
accuracy: 0.9876207166561021
precision: 0.12437810945273632
recall: 0.13812154696132597
MCC: 0.12484691347193112
F1: 0.13089005235602091
```

```
array([[26462,   176],
       [  156,    25]], dtype=int64)
```

XGBOOST

```
accuracy: 0.9925053133972184
precision: 0.2222222222222222
recall: 0.04419889502762431
MCC: 0.09648671909706168
F1: 0.07373271889400922
```

```
array([[26610,    28],
       [  173,     8]], dtype=int64)
```

Question 3

CART, SMOTE ONLY

```
accuracy: 0.9626384279801633
precision: 0.06002143622722401
recall: 0.30939226519337015
MCC: 0.1235275497999906
F1: 0.10053859964093356
```

```
array([[25761,   877],
       [  125,    56]], dtype=int64)
```

CART, SMOTE + ENN

```
accuracy: 0.947499906782505
precision: 0.06209850107066381
recall: 0.48066298342541436
MCC: 0.1587139113759707
F1: 0.10998735777496839
```

```
array([[25324,  1314],
       [   94,    87]], dtype=int64)
```

CART, OVERWEIGHT CLASS WEIGHT

```
accuracy: 0.988142734628435
precision: 0.14871794871794872
recall: 0.16022099447513813
MCC: 0.14839752868376088
F1: 0.15425531914893617

array([[26472,   166],
       [  152,    29]], dtype=int64)
```

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).

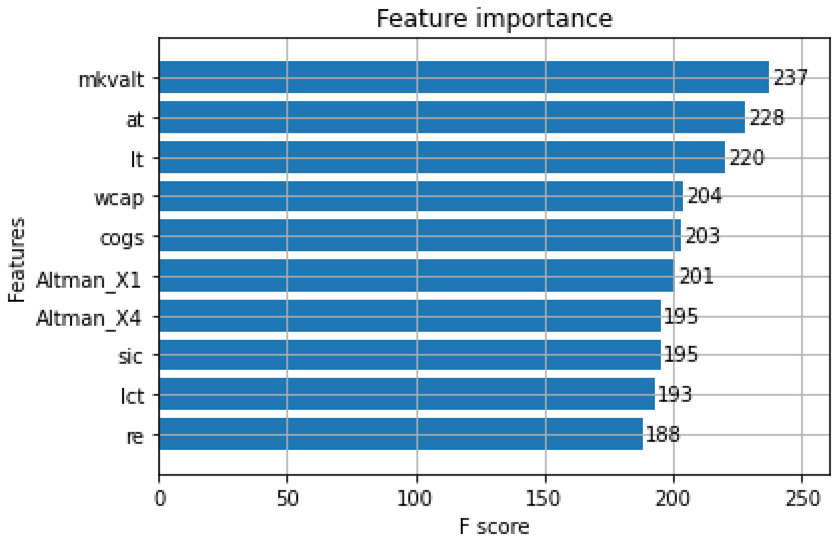
The second model was a plain vanilla one with all the features, and one-hot-encoded first digit of SIC score. (since based on EDA, only the first digit makes a significant difference). I used pos_scale_weight to balance the classes. This model was primarily to comply with the grading criteria of not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