**BT 4016: Assignment 2: Predicting Bankruptcy**

Due: 11:59pm of 12 November 2020 Thursday Midnight

Total Marks (10 marks)

1. **Overview of the problem and dataset**

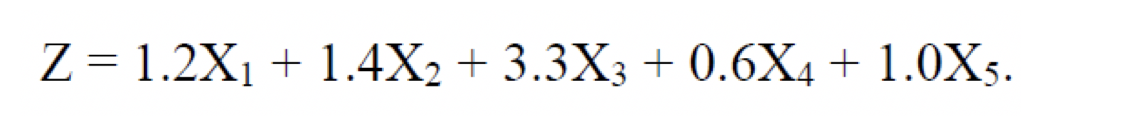
The purpose of this assignment is to practice predicting large USA firm’s bankruptcy cases. The primary data covers all large public bankruptcy cases in the USA. The sample period of this assignment is from 2002 to 2014. A company is considered “large” if its Annual Report with the USA Securities and Exchange Commission reported assets worth $100 million or more, measured in 1980 dollars.

In the following analysis, we use data from 2002 to 2011 as the training dataset (including validation dataset) and the file name is train.csv. We use data from 2012-01-01 onwards as the test dataset and the file name is test.csv. There are 181 bankruptcy cases in the train set. You do not have the label in the test dataset.

The variable list and their descriptions can be found in “Variable\_List.csv”. This is only a small subset of all accounting variables that we prepare for this assignment.

1. **Questions**

Q1. (1 mark) Please apply Altman’s Z-score with a cut-off value at 2.675 for classification.



The variables have already been computed for you (i.e Altman\_X1, Altman\_X2…Altman\_X5).

* Altman\_X1 = working capital / total assets = "wcap"/ "at"
* Altman\_X2 = retained earnings / total assets = "re" / "at"
* Altman\_X3 = earnings before interest and tax / total assets = "ebit" / "at"
* Altman\_X4 = market value of equity / total liabilities = "mkvalt" / "lt"
* Altman\_X5 = sales / total assets = "sale" / "at"

Please report one confusion matrix for the training data and the F1-score. This serves as the baseline case for comparison.

We did not have time to go over the ratios in class. Let me use this space to explain the meaning and intuition behind 5 ratios. These 5 variables are common financial ratios that are widely used in financial accounting. Let me quickly explain some more about the meaning of these ratios.

1. “working capital” is a typical proxy variable about the liquid asset owned by the firm. Working capital is defined as the difference between current assets and current liabilities and therefore this term can be quite negative. We also need to normalize it by “firm size” and total asset is the most common variable that represents firm size. This ratio is one common ratio that shows the **short-term company's solvency**. Working capital is not “cash”. But for layman’s English, you can think of it as “cash” owned by a company. In finance, some companies may have positive firm value and be very profitable but because of no cash at hand, they are in financial trouble because they cannot pay back large loans that just mature.
2. “retained earnings” is a common term in accounting. You can think of this as the accumulated profits that is still owned by the shareholders (i.e., if the firms pays out all profits as dividend, then this term is still 0). Recall that Total Asset = Total Debt + Total Equity. Equity=original shareholder investments + RE. This is a financial ratio that measure cumulative profitability, or say the proportion of total asset funded by cumulative profits. When the paper first published, this was a relatively new ratio. Also, the author showed that this ratio is very useful in predicting bankruptcy. So inventing this ratio is one contribution of this study. We will try and see whether it is still an important factor in XGBoost.
3. This term is a measure of ROA (return on asset). The only difference is what kind of earning we use as the numerator. The most common ROA uses net income. Here, the author used another widely-used variable (EBIT) as the numerator. ROA and ROE are two most popular general-purpose firm-performance measurement.
4. This term is conceptually similar to Total Equity/Total Debt. This is one type of “debt ratio”. This is a financial ratio that directly relates to bankruptcy in the long run. This is because if TA(by market value)<TL(by market value), then the firm should declare bankruptcy.
5. This is a famous ratio called **asset turnover rate**. The intuition of this one is different from the first 4 ratios. This ratio is a quick measurement for the selling efficiency of a company.

In sum, I would say these 5 variables represent the following concepts: (1) short-term ability to pay back debt, (2) Long-term ability to pay back debt by profits, (3) ROA is the profitability of a firm. (4) Long-term ability to pay back debt, measured by stock market valuation of equity. (5) Efficiency in generating sales.

**For Q2 and Q3, to lessen the workload, no hyper-parameter tuning is expected. The primary key that uniquely identifies each row is “GVKey and datadate” (which is in each row but should not be used for training features). The variable ‘bankrupt’ is the label we are trying to predict.**

**For Q4: The other columns are candidates of the input features for prediction. Make sure you specify the data type of each column correctly. Particularly, industry classification code (SIC code) is a categorical variable but its value is numerical. Therefore you need to handle it.**

Q2 (2 marks) Please train and build a prediction model by Logistic regression, CART decision tree, and XGBoost (with XGBoost random\_state=1). The input features are the 5 financial ratios in Q1.

**Report the 5-fold stratified cross validation performance results of all training data in one confusion matrix and the associated F1-score. Set shuffle=false.**

Again, you don’t need to tune in this question. More comments below:

1. You never need to tune logistic regression.
2. For CART decision tree, you just use DecisionTreeClassifier in sk-learn
3. For XGBoost, you can use the XGBoostClassifier sk-learn wrapper so that you can use default parameters. There is no default parameter for the low-level booster object for XGBoost, but you may wish to use the booster object for Q4 instead for more flexibility in tuning.

Q3 (2 Marks) Practice applying techniques to handle the unbalanced classification problems. Train only CART Decision Tree with the settings you used in Q2 and try the following three methods. (1) SMOTE for oversampling to target a 1:2 minority: majority ratio by setting sampling\_strategy = 0.5 (2) SMOTE (1:2 ratio) + Edited NN (5-NN method) (3) over-weighting rows with label=1 by 40 times (you should use class\_weight variable in DecisionTreeClassifier to do this). The learning objective is to help you practice handling unbalanced classification in this problem.

Again, report the 5-fold stratified cross validation confusion matrix and F1 score. Did these three methods help you get better results with CART Decision Tree?

CAN ADD IN 2s10s??

Q4 (4 Marks) In this question, you will practice as if you were participating a (simplified) data competition for this prediction problem.

* Method: you only need to try XGBoost.
* Handling unbalanced classification: you should try all three methods and only report the results by using one of the three methods above. You are not allowed to other methods to handle unbalanced dataset but feel free to configure the parameters for the algorithms as you deem fit.
  + You can change “scale\_pos\_weight” parameter for xgboost for the weighting method.
* Input features: you use 5 financial ratios, all accounting variables, and industry classification code. **You do not use firm ID (GVKey), or date variables for training.**
* No features engineering is allowed. The purpose is to practice parameter tuning and forecasting in a more realistic scenario. To limit your time spent in this exercise, you are not allowed to use additional variables for prediction. Although in a real data competition, this step is the most critical and time-consuming step if you want to win the competition. Regarding categorical variable, please note that:

You are welcome to try one of the methods discussed in the class - namely - one hot encoding, label encoding into a different integer space, no encode, or dropping it altogether. Please email me for approval if you have other intentions to treat the categorical variable of SIC differently from what is discussed above. To limit the time spent on the assignment the only transformation you should be doing is restricted to your treatment of the SIC categorical variable - no other variable should be touched.

* No stacking is needed. You are allowed to use only one classifier to make prediction.
* You may use hyperparameter tuning packages such as Bayesian Optimization or GridSearch to help you with the process of tuning.
* Output: submit a file with your predicted probability for the TA to evaluate your AUC. Your ranking will be decided by AUC. Please use the template file “submission\_A01XXXXX.csv” with file name submission\_yourID.csv.
* Grading: 1.5 mark is reserved for TA to deduct marks about your mistakes in coding. 3 marks are based on your AUC performance. For the performance part, if you performance is above median, then 3 out of 3. If you are in 25%-50% quantile (worse part), you may get 2 out of 3. For few cases that are very close to median, they may get 2.5. For the prediction performance in the worst quantile, we will give you a grade based on how poor the performance is. It could range from 0.5 to 1.5 out of 3.
* To reduce the potential of cheating, please also submit the final code with seeds so that we can re-train to derive the same model that delivers the prediction results. Please also explain briefly how your reach the conclusion to choose that classifier (and hyper-parameters) out of many options. If you carefully tune, you can see that you lost only 1 out of 10 by this grading rule. If you get caught cheating in this question, you will get 0 out of 10 in A2. You should not do the following two
  + Type 1 cheating: you just copy your friends’ parameters. Both get 0 out of 10.
  + Type 2 cheating: you find out which company went bankrupt and train on the test set.

Q5 (1 Mark) For your best model in Q4, report the importance score of top 10 features. You can use XGBoost’s inbuilt importance score feature to do so. Professor Altman claimed that X2 is the most important factor. Is it still the most important factor? Are those 5 factors still the most useful factors in XGBoost model?

Optional and not graded.

* You can try using the importance score as a simple method to do features selection. Given the importance score of your best XGBoost classifier, you can subjectively identify those features that are useless. Next, you drop those features and re-train your model with thorough parameter tuning. You final classifier usually gives you very similar prediction performance on the training set by 5-fold CV. Also, because your model is simpler, it is likely that you are less overfitting and your prediction performance on test dataset can be slightly better.
* If you have time and are interested in this prediction problem,
  + you can practice more when you have time.
  + You can try the 7 factor Z-Score model (but you need longer time horizon) or O-model.
  + You can train a model without using 5 factors, but use only raw accounting variables to see how useful in improving the performance after we add 5 ratios. This will also show you (1) XGBoost cannot learn ratio directly. (2) the usefulness of using ratio to do features engineering in this kind of problem.
  + You can also try more complicated features engineering.