Charles F. Dolan School of Business
BA 545 – Data Mining, Competition#1, SP 2019

Advanced Preparation of Financial Data

Overview of the Competition

Understanding pricing strategies in the context of the *Initial Public Offering* (IPO) process has been receiving much attention. Most prior studies have however focused on information sources from post issuance periods, and understanding such strategies from the management's perspective during the IPO process is still an open research issue. Form 424 variants, as finalized IPO prospectus approved by *Security Exchange Committee* (SEC), contain rich and genuine information about the issuing firms. In this study, we analyze the inter-relationships between the management's confidence (through the proxy of sentiments expressed in textual contents in the *Management's Discussion & Analysis* (MD&A) sections in the prospectus) and the pre-/post-IPO valuations.

In this competition, you are provided with data regarding successful U.S. IPOs from more than 600 companies. Your client is seeking advanced and novel methods to prepare the collected data, for further predictive analysis of the "underpricing" phenomenon. You will mainly focus on the data understanding and data preparation phases in the CRISP-DM model. Advanced modeling techniques, and presentation skills will be needed for extra points.

General Information about the Competition

- Started: 12:00 am, Monday, February 4th, 2019 ETC
- Milestone Checkpoint: 11:59 pm, Tuesday, February 19th, 2019 ETC
- Submission Deadline: 12:00 am, Monday, March 4th, 2019 ETC
- Points: 200 points in total, 100 points in each part – see **Evaluation Rules** for details.

Competition Rules

Please find the generic competition rules below. Dr. Tao reserves all the rights to further explain the rules.

Participation Rules

- All participants have to be in groups; every group contains 3 students.
- Privately sharing data, codes, or Modeler streams outside of the groups is not permitted – once violated, the group will be disqualified from the competition.
- Group leaders have full authority over the group – the communications between groups, or between groups and Dr. Tao should only be conducted by group leaders.

Submission Rules

- Only one (1) submission can be made by each group.
- Unlimited submissions can be made before the checkpoint deadline – which will not be graded.
- No late submissions (after submission deadline) will be accepted.
- Each submission contains (fail to submit any part may lead to the disqualification of the competition):
    - The data file modified by the group;
    - The python workflow and/or any other code used by the group;
    - Detailed comments and explanations to all the code used.

Evaluation Rules

- Submissions are evaluated on two evaluation metrics of the modeling results; each part takes up to 100 points:

- *Prediction Accuracy*: we will use F-1 score as the measurement of how close the predicted values are to the factual outputs – please refer to this Wikipedia article for the computation of F-1 score (https://en.wikipedia.org/wiki/Precision_and_recall):

$$\text{precision} = \frac{true\ positive}{true\ positive + false\ positive}, recall = \frac{true\ positive}{true\ positive + false\ negative}, f1score = 2 \times \frac{precision \times recall}{precision + recall}$$

- *Predictive Power*: we will use ROC curve (Receiver Operating Characteristics), or similarly, AUC (Area Under Curve), is a graphical representation that demonstrate the predictive performances of models. Note that ROC/AUC can only be applied to **binary targets**. The rule-of-thumb for ROC is as follows (**the higher the better**):

AUC	Interpretation
1.0	Perfect test
0.9 to 0.99	Excellent test
0.8 to 0.89	Good test
0.7 to 0.79	Fair test
0.51 to 0.69	Poor test
0.5	Worthless test

- Submissions are ranked base on MAE and AUC, respectively. The rank and associated points can be found below (tied rank is allowed):

Rank	Points
1	96
2	94
3	92
4	90
5	88
6	86
7	84

- Every submission will be evaluated by Dr. Tao and at least one (1) independent judge – all disagreements need to be resolved before release the results to the participants. **Example evaluation code will be provided for your convenience.**

- Extra point opportunities: each group can make use of up to one (1) extra points opportunity – each opportunity worth up to 10 extra points:
  - More modeling techniques: the basic modeling technique being used for evaluation is (logistic) regression model(s) – if you can find more applicable modeling techniques (decision tree, neural network, etc.) and apply them as your evaluation technique(s), you will receive up to 10 extra points;
  - If you can use any data preprocessing step other than the suggested tasks on the next page – you will receive up to 10 extra points;
  - You need to notify Dr. Tao your group's decision regarding the extra point opportunities **no later than** 11:59 pm, Monday, February 11th, 2019 ETC.

Competition Guidelines

Research Question

The overarching research question is "What are the determinants of IPO underpricing phenomena?" In this competition, your main purpose is to prepare the data for predictive models answering the overarching research question.

Data Dictionary

See attached 'Data Dictionary' Document.

Suggested Tasks

Following tasks are normally conduct in data preparation phase. Not all steps are required/available in this particular dataset – and not in this particular order. Some additional step(s) might be required – which you might need to research on. Keep in mind that your decisions on different strategies below will determine the results.

1. *Descriptive statistics* – describing the data using minimum, maximum, 1st & 3rd quartile, mean, median, standard deviation, number of records, number of missing records, …
2. *Imputation* – dealing with missing data, you can choose from following strategies: i) drop the record with missing (highly discouraged); ii) replace the missing with mean/median/mode, determined on the data type; iii) replacing missing values in continuous field with linear regression predictions;
3. *Normalization* – You need to manipulate all continuous fields to follow normal distribution: which contains two steps: i) removing skewness (using logarithm, square root, etc.) ii) make sure the residual is randomly distributed;
4. *Correlation analysis* – you need to select predictor variables with low pair-wise correlation (i.e. *spearman's*) values – usually the threshold is 0.5 – one variable from the pair should be dropped;
5. *Standardization* – you need to convert the values at the same numeric level; one way of doing this is to use the *z-score* standardization, which is calculated as following:

$$z = \frac{(x-\bar{x})}{\sigma}$$, where $\bar{x}$ is the mean, and σ is the standard deviation

6. Recoding – for categorical data, you might want to recode them. For instance, since you need to use AUC as the evaluation metric, you should convert the target(s) to binary (two classes). Also, you should recode any categorical variable(s) with no more than 5 classes.

Additional Rules

- If you decide to include/exclude certain variable(s)/observation(s) from the dataset, you will need to get an approval from Dr. Tao;
- You will need to report your workload assignment (within the group) in the Milestone report and the final report – non-equal workload assignment will lead to the penalty to the whole group;
- If cross-group collaboration is identified, both groups will be disqualified from the competition (result in 0 points for this part of the class).