**Predicting mutual funds rating – Goodstone Rating**

1. Data has been provided in multiple files containing different information/attributes related to ratings of given mutual funds
2. The data has been studied and a consolidated input file is prepared having numeric values. Categorical values (Investment class/fund size) had been converted into numeric values based on OHE concept. Missing values had been replaced with median. Inception Dates had been divided in below categories :

* 0-5 years
* 5-10 years
* 10-15 years
* 15 years and older

Please find enclosed the input file:



1. There had been multiple iterations that took place to find the optimum set of attributes. The highest correlated attributes with ratings, were taken and arranged in descending order. With base of 20 fields, iterations started and went up till ~90 fields. Two things were observed in these iterations:
2. For certain set of records ratings are not changing and there is no ambiguity
3. For the remaining set, the ratings were changing based on the selection of attributes

The optimum set of attributes is as follows (55 fields):

'5yrs\_sharpe\_ratio\_fund','10yrs\_sharpe\_ratio\_fund','3yrs\_sharpe\_ratio\_fund','fund\_return\_3years','3\_years\_alpha\_fund','5\_years\_return\_fund','5\_years\_alpha\_fund','5\_years\_return\_mean\_annual\_fund','10\_years\_alpha\_fund','3\_years\_return\_mean\_annual\_fund','10\_years\_return\_fund','10\_years\_return\_mean\_annual\_fund','1\_year\_return\_fund','2015\_return\_fund','2014\_return\_fund','2010\_return\_fund','2012\_fund\_return','2018\_return\_fund','2011\_return\_fund','2017\_return\_fund','years\_up','10yrs\_sharpe\_ratio\_category','1\_month\_fund\_return','fund\_return\_3months','return\_ytd','ytd\_return\_fund','total\_assets','2016\_return\_fund','investment\_class\_Growth','2010\_return\_category','maturity\_bond','yield','ID-20 and older','portfolio\_tech\_allocation','mmc','ID-10 to 15','5yrs\_sharpe\_ratio\_category','ID-0 to 5 years','5years\_fund\_std','3years\_fund\_std','10years\_fund\_std','2012\_return\_category','10\_years\_return\_category','fund\_beta\_3years','ID-5 to 10','a\_rating','pb\_ratio','2013\_return\_fund','2014\_category\_return','ytd\_return\_category','3\_months\_return\_category','category\_return\_1year','category\_return\_1month','2017\_category\_return','2016\_return\_category'

Attaching the worksheet:



1. During iterations, it has been observed that the accuracy was going down using scaled data, hence used model on data without scaling.
2. RandomForestClassifier is been used to predict the ratings, as accuracy was low using other models.
3. With optimum set of attributes, model has been executed. It is observed that the output using nominal **encoding** was not fetching predictions for all the records. In first run, it gave values for only 4045 records out of 5000 records.
4. The output received had been included in training data and in 2nd iteration, predictions received for 4177 records followed by 4195 and kept on increasing following the same iterative process.

Please find enclosed couple of appended inputs files:



1. Did multiple iterations, please find enclosed the later stage input file



1. Some ~ 300 records had not been predicted by the model and saturated with iterations
2. The model is been trained on consolidated rating column (having values from 0 to 6, not on nominal encoded rating) with the last input sheet 🡪 This gave output for all the records (including the missing ~300 records)
3. Please find enclosed the ipynb file and the input training which gave **89.39** accuracy file:

