RETURN PREDICTABILITY WITH AGNOSTIC FUNDAMENTAL ANALYSIS (AND MACHINE LEARNING)

1. Project Instructions

In this project you will test the performance of an "agnostic fundamental analysis" trading strategy similar to Bartram and Grinblatt (2018). You will use three different models (OLS, OLS post-Lasso, and Random Forest) to predict firm "fundamental values" with 27 firm accounting variables.

This is a team assignment. Please submit one set of code and one set of results per team. Teams can include up to 4 people. Projects completed in Python will receive a max score of 100%. Projects completed in Matlab or R will receive a max score of 90%.

1.1. Initial Reading:

(1) Read Bartram and Grinblatt's paper "Agnostic fundamental analysis works." **(Link)**.

1.2. Download Data:

(1) Download the cleaned data using the Dropbox link ("student_data.csv")

(2) Download the "student data instruction.csv" for detailed data information

1.3. Notes about Data:

The data provided has already been cleaned. Detailed data information and cleaning procedures are shown in the Appendix.

1.4. Analysis: (with Python)

(1) Calculate the mispricing signals with three different models. For each month t, estimate the model using only the observations for month t. Use the obtained signals to form portfolios to hold in month $t+1$. In other words, you will have to re-estimate the model each month using only the current month's cross section of firms.

(a) Model 1: OLS

(i) Package & Function: sklearn-LinearRegression

(ii) Regression formula: $MarketValue_{j,t} = \beta_{0,t} + \sum_{i=1}^{27} \beta_{i,t} * X_{i,j,t} + \varepsilon_{j,t}$ where $MarketValue_{j,t}$ is the total market value of firm j on date t. $X_{i,j,t} (i = 1, ...27)$ stands for each of the 27 explanatory variables from the financial statements of firm j on date t.

(iii) Firm's fair value prediction formula: $FairValuePrediction_j, t = \hat{\beta}_{0,t} + \sum_{i=1}^{27} \hat{\beta}_{i,t} * X_{i,j,t}$

(iv) Mispricing signal: $M_j, t = \frac{FairValuePrediction_{j,t} - MarketValue_{j,t}}{MarketValue_{j,t}}$

(b) Model 2: OLS Post-Lasso

(i) Package & Functions: sklearn-Lasso, LassoCV

(ii) Parameters:

(A) Lasso: max iter = 10000, normalize = True

(B) LassoCV: alphas = None, cv=10, max iter = 100000, normalize = True

(iii) Explanatory Variables Selection:

(A) Get the optimal alpha by LassoCV

(B) Fit the Lasso model with the optimal alpha, get the coefficients of the fitted results, then pick the explanatory variables which have nonzero coefficients

(C) Repeat OLS procedures with the picked explanatory variables

(D) Calculate the corresponding mispricing signal

(c) Model 3: Random Forest

(i) Package & Function: sklearn-RandomForestRegressor

(ii) Parameters:

(A) RandomForestRegressor: n estimators = 1000, random state = 42, min samples leaf=20

(iii) Fit the monthly data with RandomForestRegressor model

(iv) Predict the fair value with the same explanatory data as in the previous step

(v) Calculate the corresponding mispricing signal

(2) For each of the three models, form portfolios based on the mispricing signals:

(a) Rank stocks by mispricing signals within each month t

(b) Q5 denotes the most underpriced quintile of stock and Q1 the most overpriced quintile.

(c) Build portfolios for each month $t + 1$:

(i) Buy stocks in the most underpriced quintile (Q5) and short the ones in the most overpriced quintile (Q1).

(ii) Two portfolios:

- **Equal-weighted portfolio**: the mean of the selected-stock returns in the next month (drop missing values). Remember short positions will be the negative of the stock's realized return.

- **Signal-weighted portfolio**: the average of the selected-stock returns in the next month weighted by mispricing signals. Remember short positions will be the negative of the stock's realized return.

Intuition of signal-weighted portfolio: we will buy more shares of firm A than those of firm B, if firm A is more undervalued than firm B. Similar logic for overvalued case.

(3) Calculate the cumulative returns with each of the three models:

(a) Assuming the initial asset price is \$1

(b) Cumulative asset price: $AssetPrice_{i+1} = AssetPrice_i * PortfolioReturn_{i+1}$

(4) Compare the performance of the OLS, OLS post-Lasso, and Random Forest models in this trading strategy.

2. Appendix

Data information & cleaning procedures:

(1) Data source and sample period:

- Financial statements: Compustat quarterly data

- Stock price and return: CRSP stock monthly data

- Sample period: 198703 - 201212

(2) The data have been cleaned through the following steps:

(a) CRSP data:

- U.S. corporations only: delete the stock records with share class not equal to 10 or 11.

- Stocks should be listed on the New York Stock Exchange, American Stock Exchange,or Nasdaq Stock Market–National Market System: only keep stocks with exchange code 1-3.

- Delete stock records with prices less than $5.

- Only keep stock records with a positive number of common share outstanding.

- Delete stocks in financial service sector (SIC code 60-69).

(b) Compustat data:

- Dropped 'TEQQ' variable (stockholders equity, total) in regression data: TEQQ is missing in 721,306 records among 840,276 records in sample data during 198703 - 201212

- Dividend is calculated as the most recent dividend disclosure. Other items on cash and income statements are valued based on the four most recently released quarterly reports.

- Delete records with missing values for any of the 27 explanatory variables.