fit\_predict output files

Each run of the fit\_predict.py program produces two output files:

* importances.csv
* predictions.csv

The invocation of fit\_predict takes these arguments: {ticker} {cusip} {hpset} {effective\_date}. The predictions produces are for the specified CUSIP on the effective date. Often there is more than one print on a given date.

The {hpset} specifies the set of hyperparameters that are tested. The hyperparameters are set by the program and are not learned from the data. A specific set of hyperparameters is called a “model spec”.

**predictions.csv**

All the predictions for a specific CUSIP on an effective\_date are in the predictions.csv file that is filed in the folder named fit\_predict\_{ticker}\_{cusip}\_{hpset}\_{effective\_date}.

Each predictions.csv file has these columns:

* Indices for pandas DataFrames are in columns A, B, and C. The info is repeated in other columns.
* id\_modelspec\_str: a string specifying the hyperparameters. It looks like {model\_name}-{number\_of\_trades\_back}-<other fields>, where
  + model\_name is the name of the model: “en” means elastic net, “n” means naive, “rf” means random forests
  + number\_of\_trades\_back is how much history is used in training the models. You’ll see values like 1 (only the most query trade is used to train the model), 10, 30, and so forth.
  + other\_fields are hyperparameters specific to a model\_name. We will want to refine these once we understand model performance.
* id\_target\_feature\_name: which oas spread we are predicting: B, D, or S. We are also predicting these spread of the next print for the same CUSIP.
* actual: the actual oasspread for the next print of the same CUSIP
* predicted: the prediction of the model specified by the id\_modelspec\_str for the id\_target\_feature\_name.
* There are some other fields that are meant to help find mistakes.

If the predictions.csv file is empty (has zero length), then no predictions were made for the CUSIP on that date. That could be because the date was a holiday.

**importances.csv**

This file is layout out like similarly to the predictions.csv file. It contains the name of the feature used for the prediction and the importance assigned to that feature by the training procedure. These importances are the “weights” that the machine learning has learned. Many times, the training procedure has assigned a zero value to an importance. That’s probably because our look back period is usually small.

Some additional columns to examine are:

* id\_feature\_name: the name of the feature used in predictions. Ignore this for naive models. For the other models, you will see a string beginning with either “p\_” or “otr1\_”. The “p\_” feature values are from the print itself. The “otr1\_” values are from the most recent previous on-the-run CUSIP matched to the query CUSIP.
* Importance: a number in the range [0,1]. These are supposed to sum to one.

We are particularly interested in importances for high performing models.