Dynamic Asset Management
MFE-230K, Winter 2024

Instructor: Kevin Coldiron

Assignment 02 - Forecasting Beta with Machine Learning

Due: February 4, 2024

In this assignment, you will question a several decades-old methodology to compute beta and will leverage more modern ML models to more accurately forecast market betas. Your goal is not only to achieve superior forecasting performance, but also to show that you can scrutinize the ”black box” and practice explaining how it exactly works to your potential future skeptic stakeholders; be it your boss or your clients. Following a rigorous methodology and understanding its limitations will be key here as much as it will be in your profession. We are using the paper ”Estimating Stock Market Betas via Machine Learning” for inspiration.

The data consists of features known at a certain date for a certain company stock together with your target which is the market beta of the following year using daily returns (f_ols_1y_d). The features given and their description can be seen in Table 1.

Leave the last 5 years of data for testing purposes. Do not touch them until you have made all your modeling decisions (to avoid data leakage) . You will not be assessed on your performance on the test set but on your process and analysis. Leave a number of years before the test set for the validation set, i.e. not used for training the models but used to make modeling determinations upon examining forecasting performance (Used for Cross-Validation and Hyper-parameter tuning). You'll judge how many years is reasonable for the validation set. Before that will be your train set used for training your models.

You will compare the following models:

a. Baseline: CAPM beta computed using the last 5 years of monthly excess returns. The forecast is provided as a feature in ols_5y_m already.

b. OLS: Fit a linear model using OLS.

c. RF: Random Forest Regression.

d. GBT: Gradient Boosted Regression Trees, you can decide any boosting algorithm - either from sklearn or XGBoost.

e. LSTM: Long Short-Term Memory Network.

f. X: A model of your choosing. Name it as you will.

You do not have to use all features, nor use them exactly as provided, for OLS, RF, GBT, LSTM, or X. Be as consistent as reasonable in your feature inputs for OLS, RF, GBT, and LSTM. Since this will one of your last assignments before graduating, we are making the deliverable more flexible, similar to what might be a short project you pursue in your professional role.

Guidance is structured as in job openings ;)

Minimum deliverable:

a. Provides a table with forecasting errors. At a minimum includes average value-weighted mean squared error (MSE) and mean absolute percentage error (MAPE). Includes relative improvement of models against baseline model performance. You are welcome to include other performance measures you find useful and/or more directly interpretable.

b. Explains clearly how each model works, including definition, mechanics, and limitations.

c. Shows insights on the importance of features.

Desired deliverable:

a. Implements state-of-the-art model. Includes pipeline with proper feature selection, imputation, and engineering.

b. Provides a comprehensive analysis of model performance. For example, are there particular stock subsets where performance and/or relative improvements are markedly different?

c. Goes above and beyond in model interpretation. Finds ways to easily explain the role of each feature value in each prediction.

d. Can you use the same pipeline to predict the alpha?

Column	Description
log_size	Log of Size, being Size the total value of shares outstanding
log_bm	Log of Book to Market
log_pcf	Log of Price to Cashflow
mom	Momentum
strev	Current month return or Short term reversal
vol	Trailing 12-month Volatility
roa	ROA
roe	ROE
log_age_lb	Log of a lower bound for company age
price	Price
bid	Bid
ask	Ask
log_to	Log of dollar volume of month or turnover
rf	Current month risk-free rate
rm	Current month market rate of return
ols_3m_d	Beta of last 3 months using daily returns
ols_1y_d	Beta of last year using daily returns
ols_5y_m	Beta of last 5 years using monthly returns

Table 1: Features

References: `https://datascientest.com/en/data-leakage-definition-and-prevention`