**Final Project Guidelines**

This semester, we have been learning about how to use machine learning to predict returns and create portfolios based on these predictions. For the final project, I want you to put what you have learned into practice. Your team will come up with a recommendation for a portfolio of stocks in your industry to be held for one year. The portfolio will consist of the 20 stocks that your machine learning algorithm predicts will perform the best in the next year.

In coming up with your recommendation, you should employ each of the machine learning techniques we have learned during the course:

1. Linear regression
2. Lasso penalized regression
3. Random Forests
4. Gradient Boosting Regression
5. Neural Networks

The paper for the course should be structured as follows:

1. Executive summary. The executive summary should summarize in no more than a page the portfolio that you are recommending and how you arrived at that decision. You don’t need to go into details here of exactly what you did, but discuss the approach you used, how you generally structured your investigation, and include a table of the tickers of the stocks that you recommend with their predicted returns.
2. Methodology.
   1. First, discuss the machine learning approaches that you used. I’m looking here for understanding of the various approaches we covered. You can go into detail, but also keep in mind that it should be explained well enough that your grandparents can understand what you are doing.
   2. Independently or in the first part, explain the objective that you are trying to target. In class I discussed how maximizing out-of-sample R2 doesn’t necessarily lead to the best portfolio returns. That doesn’t mean that you should ignore out-of-sample R2 or not use it as a metric. You might come up with some alternative metric entirely – remember that I will reward creativity. But justify why you are using this criterion. And remember that the criterion has to be out-of-sample. Choosing the approach that generates the highest returns *ex post* is not in the spirit of this analysis.
   3. Next, discuss how you selected or tuned your hyperparameters. This discussion should include how you split the sample into training, validation, and testing sets, whether you took a rolling or expanding approach, etc. Don’t simply use the approach that we’ve been taking with 5-year training, 3-year validation, and 1-year testing. You can use that approach, but I’m looking for an explanation *why*. In the case of neural nets, you will be selecting a set of rather than tuning hyperparameters, so explain this selection process.
3. Results. The next section should explain the results and provide tables and/or figures documenting the performance of the machine learning strategy. Performance metric should be similar to what we used in class and include the following:
   1. Total return on $1 invested
   2. Comparable return on a value-weighted benchmark of stocks in your industry
   3. Test-sample R-squared[[1]](#footnote-1)
   4. Geometric average return
   5. Alpha
   6. **Beta**
   7. Maximum one-, three-, and 12-month drawdowns
   8. **Sharpe ratio**

If there are other metrics that you think would be appropriate to discuss, please include these as well.

In addition to one-month holding period results, I would like you to replicate the above analysis for a holding period of 12 months. That is, rather than re-forming portfolios every month in the test sample, consider the performance of a value-weighted portfolio formed at the beginning of the test sample and held through a year of the test sample.

The results should also include a discussion of which signals are important for the methodologies for which signal importance can be quantified.

1. Predictions. The variables that we have used thus far are in the U.S. Master Screen used in the Tozzi Center. Based on the in-sample results from the data, recommend a portfolio of 20 firms in your group to hold over the next twelve months using Factset to obtain rankings on the signals. I will ask Kai to provide assistance in using the Master Screen.
2. Code appendix. You should include your Python code in an appendix to the paper.
3. Presentation. Prepare a 10-minute presentation, which should represent a high-level overview of the contents of your paper – discuss what you did, what you found, and what you recommend going forward.

**Additional Guidelines**

* Page length. I hate putting a number on the minimum number of pages. My guess is that it should take around 10 pages to explain the above plus tables, figures, and code. If, however, you don’t write 10 pages of text, please don’t feel the need to stretch the narrative to 10 pages!
* Creativity. We’ve all learned the same basic techniques in class. Therefore, I am particularly interested in creative solutions to approaching the task. This could include thinking about alternative signals to those we’ve already considered, or some kind of different machine learning or adaptation of the machine learning approaches we’ve studied.
* I am here to provide assistance. Starting March 23, we will not have class sessions in order to provide you with additional time to work on the project. During that class time, I will be available to consult. Further, I’m more than happy to schedule other times to discuss the project or issues you are having.
* The presentation is before the paper due date. This means that I expect you to present what you have discovered thus far, not necessarily your final recommendation. Again, I’m more than happy to discuss beforehand if what you are thinking about for the presentation will be enough.

1. The R-squared should just be a single number, even though you have multiple testing sets. Remember that the R-squared is just one, minus the ratio of the variance of prediction errors to the total variance in the sample. So you can calculate the R-squared by taking the variance of prediction errors in the *entire* set of testing data, divided by total variance in the testing data. [↑](#footnote-ref-1)