Machine learning for empirical asset pricing and risk premia forecasting

Dingtian Zhu, Xin Gu
NYU Tandon, Department of Finance and Risk Engineering
Jul 1, 2020

Abstract:

Relative to traditional empirical methods in asset pricing, machine learning accommodates a far more expansive list of potential predictor variables, as well as richer specifications of functional form. Machine learning methods can be successfully applied to the two canonical problems of empirical asset pricing: predicting returns in the cross section and time series.

The literature has accumulated a long list of predictors that various researchers have argued possess forecasting power for returns. The number of stock-level predictive characteristics reported in the literature numbers in the hundreds and macroeconomic predictors of the aggregate market number in the dozens. Additionally, predictors are often close cousins and highly correlated. Traditional prediction methods break down when the predictor count approaches the observation count or predictors are highly correlated. While machine learning methods may have the advantage of dealing with big data and non-linearity, it may help to enhance new predictors.

Here we want to use machine learning methods, neural networks(NN) and decision trees to realize variable selection and dimension reduction techniques, which will lead to an automated prediction model for stock return.

Key Words:

Asset pricing, portfolios, cross-section of expected returns, neural networks, stock characteristics, variable selection, machine learning.

1. Introduction

Compared with traditional empirical methods in asset pricing, machine learning can accommodate a more expansive list of potential predictor variables and a richer specification of functional form. The ML methods are able to be applied to predict returns in the cross section as well as in time series.
Stochastic discount factor was introduced in the 1990s to describe the difference in asset price. Consider SDF model here:

$$m_t := \gamma_0^{-1} - \gamma_0^{-1}\lambda_v^\intercal v_t := \gamma_0^{-1}(1 - \lambda_g^\intercal g_t - \lambda_h^\intercal h_t),$$

In particular, we show how to estimate and test the marginal importance of any factor gt in pricing the cross section of expected returns beyond what is explained by a high- dimensional set of potential factors h_t, where g_t and h_t could be tradable or non-tradable factors.
We assume the true asset pricing model is approximately low-dimensional; however, in addition to relevant asset pricing factors, g_t and h_t include redundant ones that add no explanatory power to the model, as well as useless ones that have no explanatory power at all.
Our methodology can be thought of as a conservative test for new factors, which benchmarks them against a large-dimensional set of existing ones.

2. Literature review

For dimension reduction, most recent literature uses NN to abstract information from rich available indicators.
Feng et al. (2018) employ a two-pass regression approach without assuming prior knowledge of which factors to include as controls from among the possible hundreds of factors found in the literature.
Kozak et al. (2019) provide a shrinkage approach to model fitness, and Feng et al. (2019) test new factors through model selections. In dimension reduction through principal components (PCA).
Kelly et al. (2019) employ characteristics as instruments in NN to get deep-layer factors.
Chen et al. (2019) states macroeconomic information dynamics are summarized by macroeconomic state variables which are obtained by a Recurrent Neural Network (RNN) with Long-Short-Term-Memory units.

Our paper follows their research directions and provides a deep learning framework of the SDF model with dimension reduction.

3. Methods

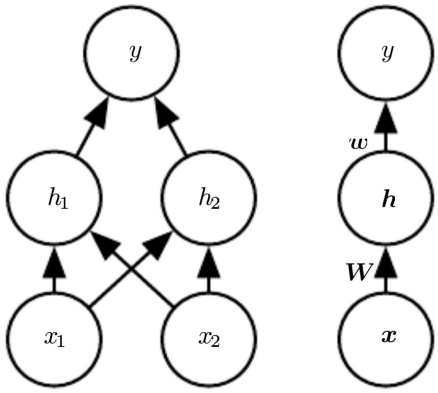
Brief introduction of the ML methods we want to use in the study and why/how.

3.1 Shrinkage

Because of the results of some previous research, here we will focus on the Lasso method.

3.2 Deep learning methods

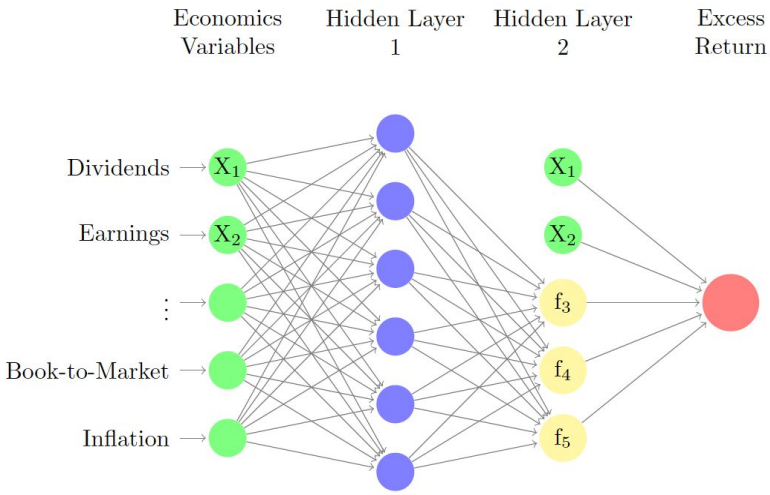
a. Deep Feedforward neural networks (DFN)
The basic DFN follows an hierarchical structure and consists of three layers: Input layer, hidden layer and output layer.



(Messer, 2017, page 7)

The activation function can have different forms. Common choices are sigmoid activation and rectified linear units.

b. Multi-layer deep learners: LSTM



(Feng et al., 2018a, page 5)
Deep learning is a data reduction scheme that uses L layers of "hidden" factors, which can be highly nonlinear. From a finance viewpoint, we have a hierarchical model of the form:

$$R_{t+1} = \alpha + \beta X_t + \beta_f F_t + \epsilon_{t+1}$$

$$F_t = F^{W,b}(X_t)$$

$$F^{W,b} := f_1^{W_1,b_1} \circ \cdots \circ f_L^{W_L,b_L}$$

$$f^{W_l,b_l}(Z) := f_l(W_l Z + b_l), \quad \forall 1 \le l \le L$$

where *F* is a multivariate data reduction map represented as a deep learner. The network parameters *(W, b)* are weights and offsets to be trained. t are the usual idiosyncratic pricing errors. The major difference between DL and traditional factor models is the usage of compositions of factors rather than shallow additive models.
Unlike traditional estimation of Ft and coefficients by regression with a two-step procedure, deep learning will estimate coefficients and factors jointly.
We need to use a loss function to minimize the MSE of the in-sample fit.

4. Metrics for assessing forecast performance

This part will introduce the optimizing metrics we will use in our study.
For reference:

4.1 Performance measures for point forecasts for M4 competition:

For M4 competition it was decided to use the average, referred to as the overall weighted average (OWA), of two of the most popular accuracy measures:

- symmetric mean absolute percentage error sMAPE

$$sMAPE \triangleq \frac{2}{h} \sum_{t=n+1}^{n+h} \frac{\left|Y_t - \widehat{Y}_t\right|}{|Y_t| + \left|\widehat{Y}_t\right|} * 100\ ()$$

- mean absolute scaled error M

$$sMAPE \triangleq \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} \left|Y_t - \widehat{Y}_t\right|}{\frac{1}{n-m} \sum_{t=m+1}^{n} |Y_t - Y_{t-m}|}$$

where Yt is the value of the time series at point t , Yt hat the estimated forecast, h the forecasting horizon, n the number of the data points available in-sample, and m the time interval between successive observations considered for each data frequency, i.e., 12 for monthly, four for quarterly, 24 for hourly and one for yearly, weekly and daily data.

The first measure uses percentage errors that are scale independent, intuitive to understand and part of an everyday vocabulary. The second measure aims to correct some potential problems of the first and to provide an alternative with better mathematical properties.

4.2 Performance measures for prediction intervals for M4 competition:

The M4 Competition adopted a 95% prediction interval (PI) for estimating the uncertainty around the point forecasts. The performances of the generated PIs were evaluated using the Mean Scaled Interval Score (MSIS) described in [Makridakis-et-al-2019a]

5. Empirical Study

In this section, we will introduce the empirical study of our research, including the dataset/ software we used, the preliminary analysis we referred to and and their assumptions, the results of our study and a comparison of performance.

5.1 Data used

Our empirical analysis is based on a data set of all available U.S. stocks from CRSP with monthly returns from 1970 to 2019 combined with 40+ time-varying firm-specific characteristics and 100+ macroeconomic time series. It includes the most relevant pricing anomalies and forecasting variables for the equity risk premium.

For factors claimed in academia, we also take into account 50+ factors from abundant resources. We test whether these factors have predictive value, and use existing factor models as the performance benchmark for our ML model.

5.2 Preliminary analysis: analysis of existing methods/ results;

5.3 Results and Comparison

6. Conclusion

6.1 The main conclusion of our study

Which method(s) has a better performance over others, how much better; the advantage and disadvantages of the methods.

6.2 discussion of further improvement

References

Bianchi, D., Buchner, M., and Tamoni, A. (2019). "Bond Risk Premia with Machine Learning". In: SSRN Electronic Journal.

Brogaard, J. and Zareei, A. (2018). "Machine learning and the stock market". In: SSRN

Bryzgalova, S., Pelger, M., and Zhu, J. (2019b). "Forest Through the Trees: Building Cross-Sections of Stock Returns". In: SSRN Electronic Journal.

Chen, L., Pelger, M., and Zhu, J. (2019). "Deep learning in asset pricing". In: SSRN Electronic Journal.

Choi, D., Jiang, W., and Zhang, C. (2019). "Alpha go everywhere: machine learning and international stock returns". In: SSRN Electronic Journal.

Feng, G., Giglio, S., and Xiu, D. (2019a). "Taming the Factor Zoo: A Test of New Factors". In: SSRN Electronic Journal.

Feng, G., He, J., and Polson, N. G. (2018a). "Deep Learning for Predicting Asset Returns". In: arXiv Electronic Journal.

Feng, G., Polson, N., and Xu, J. (2018b). "Deep learning factor alpha". In: SSRN Electronic Journal.

Granziera, E. and Sekhposyan, T. (2019). "Predicting relative forecasting performance: An empirical investigation". In: International Journal of Forecasting.

Gu, S., Kelly, B. T., and Xiu, D. (2019a). "Autoencoder asset pricing models". In: SSRN Electronic Journal.

Gu, S., Kelly, B. T., and Xiu, D. (2019b). "Empirical asset pricing via machine learning". In: SSRN Electronic Journal.

Jin, S., Corradi, V., and Swanson, N. R. (2015). "Robust Forecast Comparison". In: SSRN Electronic Journal. Kim, H. and Durmaz, N. (2012). "Bias correction and out-of-sample forecast accuracy". In: International Journal of Forecasting 28(3), pp. 575–586.

Kozak, S., Nagel, S., and Santosh, S. (2019). "Shrinking the cross-section". In: Journal of financial economics.

Li, Y., Turkington, D., and Yazdani, A. (2020). "Beyond the Black Box: An Intuitive Approach to Investment Prediction with Machine Learning". In: The Journal of Financial Data Science.

Liew, J. K. and Mayster, B. (2017). "Forecasting ETFs with Machine Learning Algorithms". In: The Journal of Alternative Investments.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). "The M4 Competition: Results, findings, conclusion and way forward". In: International Journal of Forecasting 34(4), pp. 802–808.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2019). "The M4 Competition: 100,000 time series and 61 forecasting methods". In: International Journal of Forecasting.

Petropoulos, F., Wang, X., and Disney, S. M. (2018). "The inventory performance of forecasting methods: Evidence from the M3 competition data". In: International Journal of Forecasting.

Popescu, A. V. (2019). "The Macroeconomy and the Cross-Section of International Equity Index Returns: A Machine Learning Approach". In: SSRN Electronic Journal.

Rapach, D. E., Strauss, J. K., Tu, J., and Zhou, G. (2019). "Industry return predictability: A machine learning approach". In: The Journal of Financial Data Science.

Ryll, L. and Seidens, S. (2019). "Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey". In: arXiv Electronic Journal.

Samuels, J. D. and Sekkel, R. M. (2017). “Model Confidence Sets and forecast combination”. In: International Journal of Forecasting 33(1), pp. 48–60.
Siliverstovs, B. (2017). “Dissecting models’ forecasting performance”. In: Economic Modelling.
Simonian, J., Wu, C., Itano, D., and Narayanam, V. (2019). “A Machine Learning Approach to Risk Factors: A Case Study Using the Fama-French-Carhart Model”. In: The Journal of Financial Data Science.
Weigand, A. (2019). “Machine learning in empirical asset pricing”. In: Financial Markets and Portfolio Management.

Appendix

A. Computational Remarks
   (More figures/ Tables, etc)
B. Complete Results of the Asset Return Prediction
C. (LSTM?)