Evaluation of Machine Learning in Finance

Ze Yu Zhong

Supervisor: David Frazier

Monash University

Main Motivation

To evaluate the application of machine learning to predicting financial asset returns, with specific regard to how they deal with the unique challenges present in financial data via simulation.

Motivation

Empirical finance is typically concerned with two main goals:

- Prediction accuracy - want to predict future returns!
- Statistical causal inference - want to understand *what* drives returns

Motivation

Recognise that these goals are very difficult, especially for traditional regression!

The underlying data generating process for financial data is unknown - returns are estimated by *risk factors*, defined by Harvey et al. (2016) as a collection of regressors that can proxy for underlying risk factors

Factors are often unsuitable for regression:

- Persistent
- Cross-sectionally correlated (multicollinearity)
- Non-stationary
- Pre-known, and therefore little time series variation

Motivation

Consequences when included in traditional regression well documented:

- Biased t-stats
- High variance for coefficient estimates
- Unstable coefficient estimates

These can adversely affect statistical inference.

The estimated coefficients, due to being imprecise, will also result in poor out of sample prediction, which are a function of coefficients. This is particularly so if the multicollinearity between regressors changes over time, which is likely in financial data

Background - Dividend Ratio Example

- Included due to good in sample performance in the 1990s (Goyal and Welch, 2003)
- *Persistent* (Goetzmann and Jorion (1993), Ang and Bekaert (2006))
  - Correlated with lagged dependent variables on the right hand side of the regression equation.
  - Violates assumptions of independent regressors of OLS: t stats are biased upwards due to autocorrelated errors
  - GMM and NW errors corrections are also biased, (Goetzmann and Jorion, 1993)
- Not robust and have poor out of sample performance since 2000s (Goyal and Welch (2003), Lettau and Ludvigson (2001), Schwert (2003))

Dividend Ratio Example

- Factors such as dividend ratios, earnings price ratio, interest and inflation etc. were "widely accepted" able to predict excess returns, (Lettau and Ludvigson, 2001)
- Welch and Goyal (2008) conclude that not a single variable had any statistical forecasting power, and the significance values of some factors change with the choice of sample periods.

Background

- More factors produced by literature: currently over 600 documented (Harvey and Liu, 2019)
  - False discovery problem, (Harvey et al., 2016)
  - Factors are cross sectionally correlated - inefficient covariances, factors may be subsumed within others, (Feng et al., 2019)
  - Number of factors may be more than sample size, making regression impossible

Why apply Machine Learning in Finance?

Machine learning methods have been used within the literature and appear to be well suited:

- High dimensional - more flexible than traditional regression models, which make strong functional form assumptions and are sensitive to outliers, (Freyberger et al., 2017)
- Explicit "regularization" methods for guarding against overfitting
- Methods to produce an optimal model from all possible at manageable computation cost

More able to manage the explosion in the number of factors suggested by the literature!

Applications in the Literature

Causal Analysis:

- Kozak et al. (2017), Rapach and Zhou (2013), Freyberger et al. (2017), and others apply shrinkage and selection methods to assist with factor selection

Prediction Performance:

- Gu et al. (2018), Feng et al. (2018), construct machine learning portfolios that historically outperform traditional portfolios in terms of prediction error and predictive R^2
- Attribute their success to machine learning's ability to find non-linear interactions

Motivations

However, little work has been done on how machine learning actually recognises and deals with the challenges in financial data.

Machine learning is mainly suited for specific contexts:

- Little interpretability in regressors
- Often assume i.i.d.
- Often require large amounts of data
- Struggle with state dependence

Methodology Issues

- Feng et al. (2018) cross validates their training set, destroying temporal aspect of data, and only explore a handful of factors
- Gu et al. (2018) only use data up until the 1970s to produce predictions in the last 30 years
- Gu et al. (2018)'s models do not have consistent importance metrics - only their tree based methods recognise dividend yield as important

Motivations

Can machine learning assist with returns prediction and causal inference?

- Persistent Regressors?
- Identify true factors from a high dimensional, cross sectionally correlated panel?
- Is regularization enough to handle non-robustness?
- Are their conclusions consistent?
- Do they perform better than traditional methods?

Motivations

We explored this via two studies, focusing on the prediction performance and factor selection performance of some common machine learning models:

Simulation study

- Explicitly explore how machine learning performs in controlled environments

Empirical study

- Validate our results
- Contrast them with literature

Model Specification

Returns are modelled as an additive error model

$$r_{i,t+1} = E(r_{i,t+1}|\mathcal{F}_t) + \epsilon_{i,t+1} \tag{1}$$

where

$$E(r_{i,t+1}|\mathcal{F}_t) = g^*(z_{i,t}) \tag{2}$$

Stocks are indexed as $i = 1, \ldots, N$ and months by $t = 1, \ldots, T$.

$g^*(z_{i,t})$ represents the model approximation using the P dimensional predictor set $z_{i,t}$.

Overview

Machine Learning Methodology consists of 3 overall components:

- Sample Splitting
- Loss Function(s)
- Models/Algorithms considered

Sample Splitting - Expanding Window Scheme

An expanding/growing window approach was used to

- Allow models to incorporate more data over time
- Preserve temporal ordering of data (compared to cross-validation)
- Allows the model to use the most recent data, in some sense

The training/validation split was chosen s.t. the size of the training set was 1.5 times the length of the validation set to begin with, consistent with Gu et al. (2018).

Sample Splitting - Expanding Window Scheme

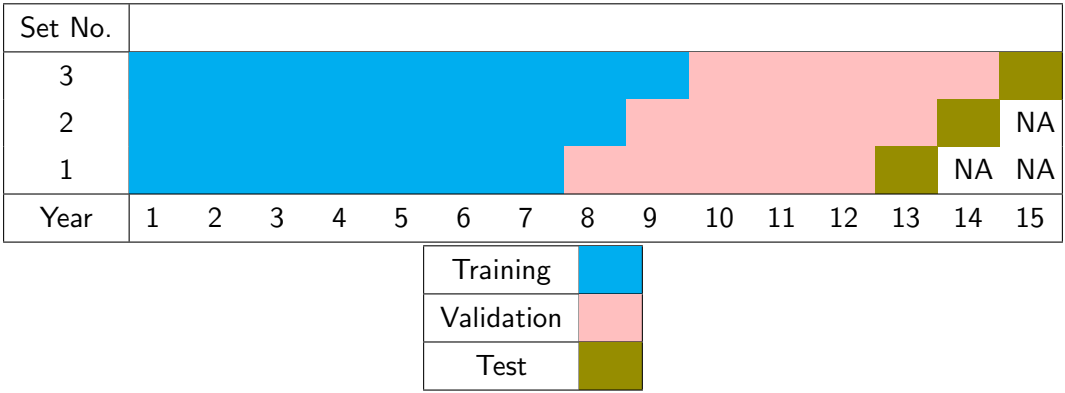

Figure 1: Sample Splitting Procedure

Loss Functions

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{j=i}^{n}|y_j - \hat{y}_j| \tag{3}$$

Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{j=i}^{n}(y_j - \hat{y}_j)^2 \tag{4}$$

Models

We focus on 4 common machine learning models:

- Linear Models
- Penalized Linear Models
- Random Forests
- Neural Networks

In general, we mimic the methodology of Gu et al. (2018).

Linear Models

Linear Models assume that the underlying conditional expectation $g^*(z_{i,t})$ can be modelled as a linear function of the predictors and the parameter vector θ:

$$g(z_{i,t}; \theta) = z_{i,t}'\theta \tag{5}$$

Optimizing θ w.r.t. MSE yields the Pooled OLS estimator

Penalized Linear Models

Elastic Net penalty aims to produce efficient and parsimonious via shrinkage and selection of factor coefficients

Linear Models + Penalty term (Elastic Net by Zou and Hastie (2005) shown):

$$\mathcal{L}(\theta; .) = \underbrace{\mathcal{L}(\theta)}_{\text{Loss Function}} + \underbrace{\phi(\theta; .)}_{\text{Penalty Term}} \tag{6}$$

$$\phi(\theta; \lambda, \rho) = \lambda(1-\rho)\sum_{j=1}^{P}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2 \tag{7}$$

Random Forests

Ensemble model of regression trees with low correlation, each with high depth

Able to capture multiway non-linearities

Tend to be quite robust

Depth of trees and number of predictors sampled for each tree tuned

Neural Networks - Overview

Most complex model considered

Able to capture non-linearities systematically

In theory able to approximate any function - Universal Approximation Theorem

Neural Network Specification

- Neural networks with up to 5 hidden layers were considered.
- The number of neurons is each layer determined by geometric pyramid rule (Masters, 1993)
- All units are fully connected

Neural Network Tuning

Neural Networks have the largest number of hyperparameters to tune - computationally infeasible to conduct comprehensive grid search.

A conservative gridsearch was conducted, but it was observed that in general, default values provided the best performance.

Observed that neural networks are high sensitive to hyperparameters - not easy to tune!

Hyperparameter Tuning

Hyperparameters gridsearched:

- Optimizer
- Learning Rate
- Batch Size
- l1 penalty
- Activation function
- Momentum

Neural Network - Activation Function

Gu et al. (2018) specified using the ReLU activation function for all layers

We found that this specification **does not work** due to the "dying ReLU" problem"

Alternatives such as Leaky ReLU also did not give very good performance.

A specification of the tanh activation function was observed to give the best results

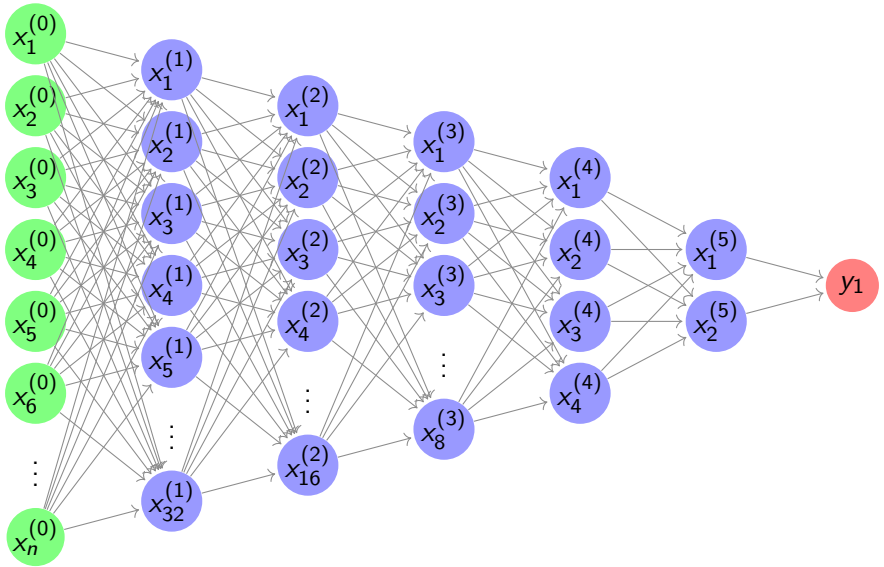

Figure 2: Neural Network 5 (most complex considered)

Loss Metrics

Overall predictive performance for individual excess stock returns were assessed using the following loss metrics:

- Mean Squared Error
- Mean Absolute Error
- Out of Sample Predictive R

Out of Sample R squared

An out of sample R squared metric was also reported, as is popular in the financial literature. No consensus as to how this metric is to be calculated, and we use the following formulation:

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - \widehat{r}_{i,t+1})}{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - \bar{r}_{i,t+1})^2} \tag{8}$$

where $\mathcal{T}_3$ indicates that the fits are only assessed on the test subsample

Interpretation of the numerical value is cautioned, as R squared was originally designed for use in assessing in sample fit for linear models.

Variable Importance

- Defined as the reduction in predictive R-Squared from setting all values of predictor j to 0, while holding the remaining model estimates fixed
- Note some minor transformations applied for numerical reasons

$$VI_{j,norm} = \frac{VI_j + \min(VI_j) + o}{\Sigma VI_j + \min(VI_j) + o} \quad ; \quad o = 10^{-100} \tag{9}$$

Simulation Study

Simulation study conducted see how well machine learning performs in a controlled environment which we understand, as opposed to empirical data which we do not understand
Characteristics of financial data to be captured

- Stochastic volatility in errors
- Low signal to noise ratio
- Persistence across time in regressors
- Cross sectionally correlated (collinear) regressors
- High number of regressors

Gu et al. (2018)'s Simulation Design

We replicate Gu et al. (2018)'s simulation design as a starting point General structure:

- Generate individual and macroeconomic factors
- These enter into a data generating process, ranging from simple to complex
- Final returns process will be this "true" latent process, plus a large error term

Several Problems:

- White noise, constant volatility specification
- No cross sectional correlation

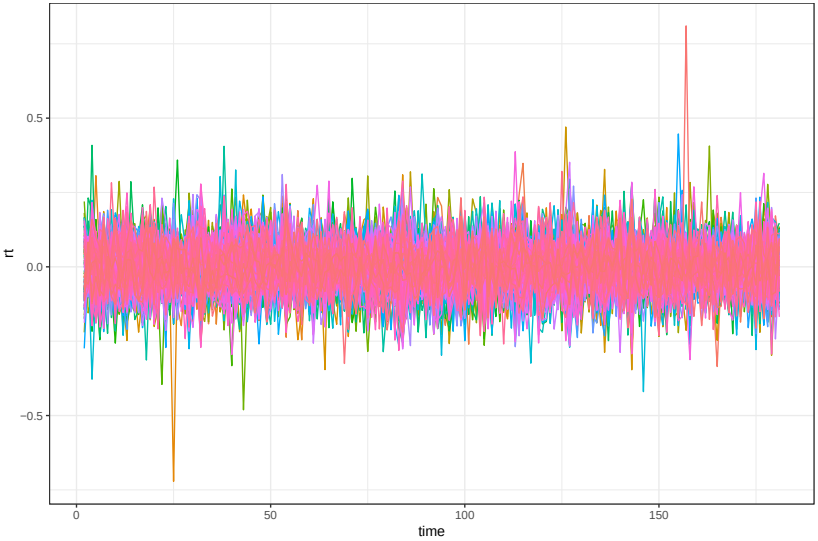

Figure 3: Gu et al. (2018)'s Simulation Design

Proposed Simulation Design

We propose a design which incorporates:

- Persistence in regressors
- Cross sectional correlation in regressors
- Stochastic volatility in errors

Proposed Simulation Design

Latent factor model with stochastic volatility for excess return, r_{t+1}, for $t = 1, \ldots, T$:

$$r_{i,t+1} = g\left(z_{i,t}\right) + \beta_{i,t+1} v_{t+1} + e_{i,t+1}; \tag{10}$$

$$z_{i,t} = \left(1, x_t\right)' \otimes c_{i,t}; \quad \beta_{i,t} = \left(c_{i1,t}, c_{i2,t}, c_{i3,t}\right); \tag{11}$$

$$e_{i,t+1} = \exp\left(\frac{\sigma^2_{i,t+1}}{2}\right) \varepsilon_{i,t+1}; \tag{12}$$

$$\sigma^2_{i,t+1} = \omega + \gamma_i \sigma^2_{t,i} + w_{i,t+1} \tag{13}$$

v_{t+1} is a 3×1 vector of errors, $w_{i,t+1}, \varepsilon_{i,t+1}$ are scalar error terms. Variances tuned such that the R squared for each individual return series was 50% and annualized volatility 30%.

Simulating Characteristics

Matrix C_t is an $N \times P_c$ vector of latent factors.

x_t is a 3×1 multivariate time series

ε_{t+1} is a $N \times 1$ vector of idiosyncratic errors.

A simulation mechanism for C_t that gives some correlation across the factors and across time was used. We build in correlation across time among factors by drawing normal random numbers for each $1 \leq i \leq N$ and $1 \leq j \leq P_c$, according to

$$\overline{c}_{ij,t} = \rho_j \overline{c}_{ij,t-1} + \epsilon_{ij,t}; \quad \rho_j \sim \mathcal{U}\left(\frac{1}{2}, 1\right) \tag{14}$$

Simulating Characteristics

Then, define the matrix B:

$$B := \Lambda\Lambda' + \frac{1}{10}\mathbb{I}_n, \quad \Lambda_i = (\lambda_{i1}, \ldots, \lambda_{i4}), \quad \lambda_{ik} \sim N(0, \lambda_{sd}),\ k = 1, \ldots, 4 \tag{15}$$

B is a p.s.d. matrix, serves as a vcov matrix with λ_{sd} controlling the density of the matrix. λ_{sd} values of 0.01, 0.1 and 1 were used to explore effects of **cross sectional correlation**. Simulate characteristics according to

$$\widehat{C}_t = L\overline{C}_t; \quad B = LL' \tag{16}$$

where L represents the lower triangle matrix of B using the Cholesky decomposition.

Simulating Characteristics

Finally, the "observed" characteristics for each $1 \leq i \leq N$ and for $j = 1, \ldots, P_c$ are constructed according to:

$$c_{ij,t} = \frac{2}{n+1} \operatorname{rank}\left(\hat{c}_{ij,t}\right) - 1. \tag{17}$$

with the rank transformation normalizing all predictors to be within $[-1, 1]$

Simulating Return Series

We will consider 3 different functions $g(\cdot)$:

$$(1)\ g_1(z_{i,t}) = \left(c_{i1,t}, c_{i2,t}, c_{i3,t} \times x'_{3,t}\right)\theta_0$$
$$(2)\ g_2(z_{i,t}) = \left(c^2_{i1,t}, c_{i1,t} \times c_{i2,t}, \text{sgn}\left(c_{i3,t} \times x'_{3,t}\right)\right)\theta_0$$
$$(3)\ g_3(z_{i,t}) = \left(1[c_{i3,t} > 0], c^3_{i2,t}, c_{i1,t} \times c_{i2,t} \times 1[c_{i3,t} > 0], \text{logit}\left(c_{i3,t}\right)\right)\theta_0$$

Tune θ^0 s.t. predictive R^2 is 5% to ensure **low signal to noise ratio**.

Note that g_2 is the most non-linear and **difficult to approximate**, given the observed characteristics.

Simulating Return Panel

Each realization results in a panel of:

- $N = 200$ stocks
- $T = 180$ periods
- $P_c = 100$ characteristics

A total of 10 realizations were simulated and had models fitted to them.

Simulating Returns Panel

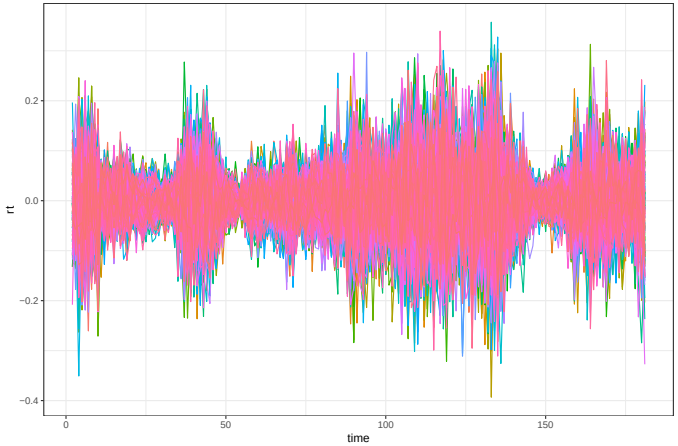

Figure 4: Proposed Simulation Design, 1 Realization

Simulation Results

Key Results:

- Minimizing quantile loss is better
- Elastic Net $>$ Random Forests $>$ Neural Networks $>$ Linear Models
- Cross sectional correlation **does not affect prediction performance** by much
- Elastic net is best for causal analysis, **even with multicollinearity**
- Random Forest is better in non-linear settings
- Neural Networks **did not outperform**

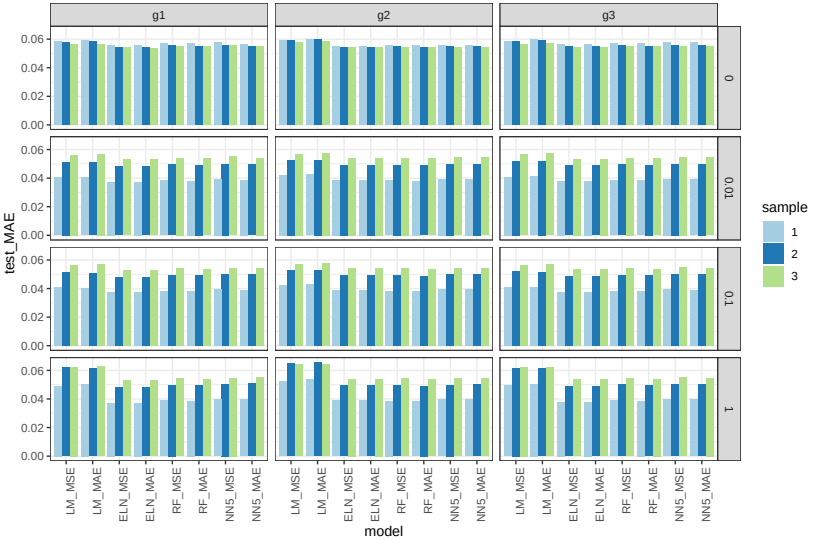

Figure 5: Test MAE across all simulation designs

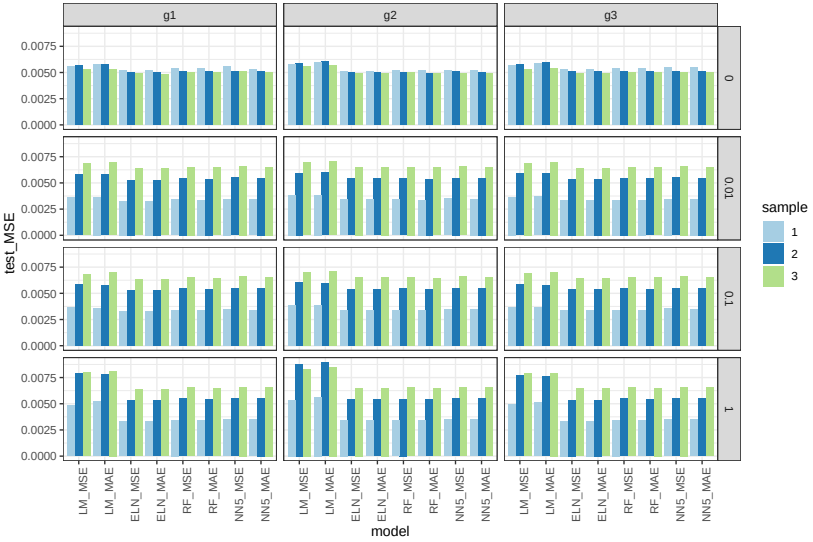

Figure 6: Test MSE across all simulation designs

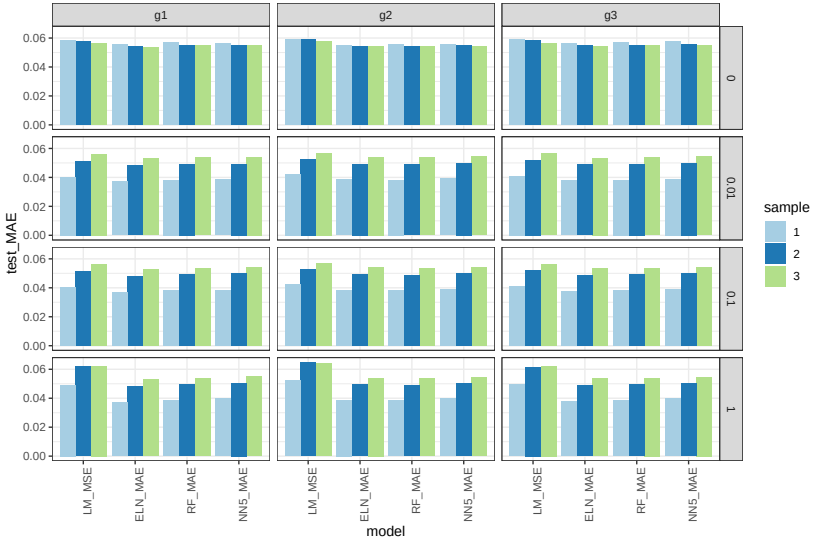

Figure 7: Test MAE across all simulation designs

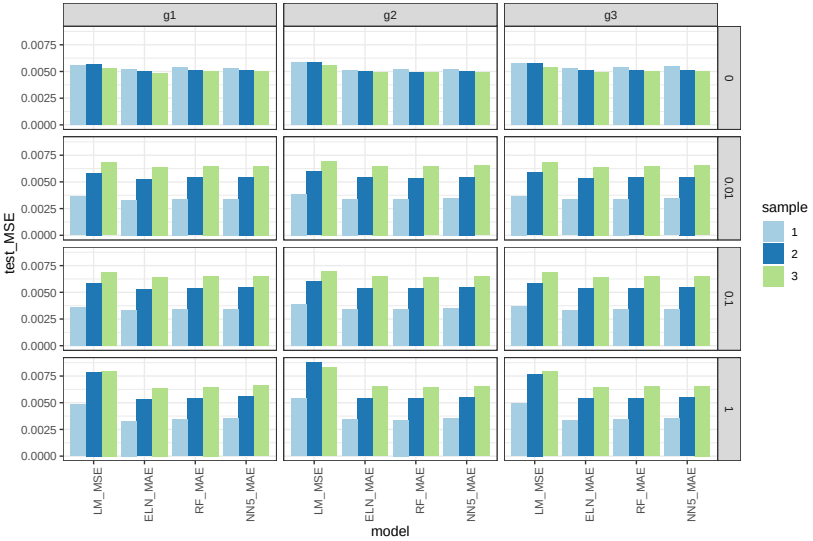

Figure 8: Test MSE across all simulation designs

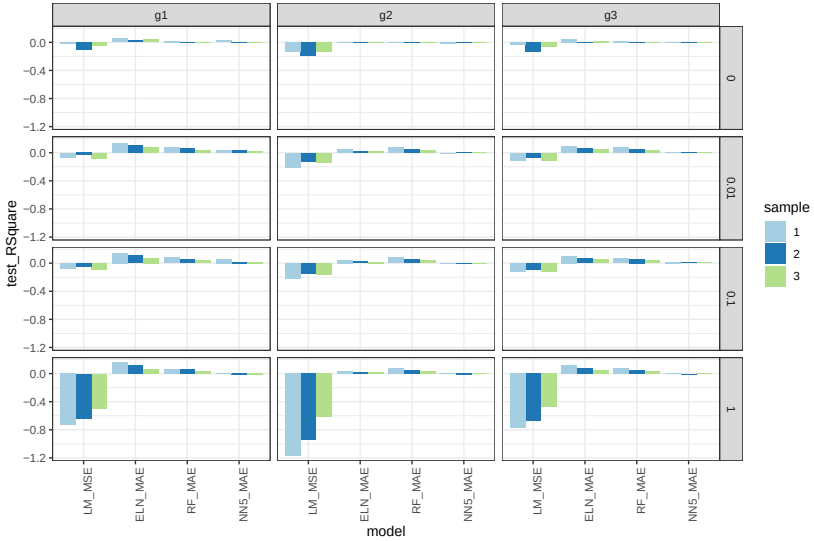

Figure 9: Test R Squared Across All Simulation Designs

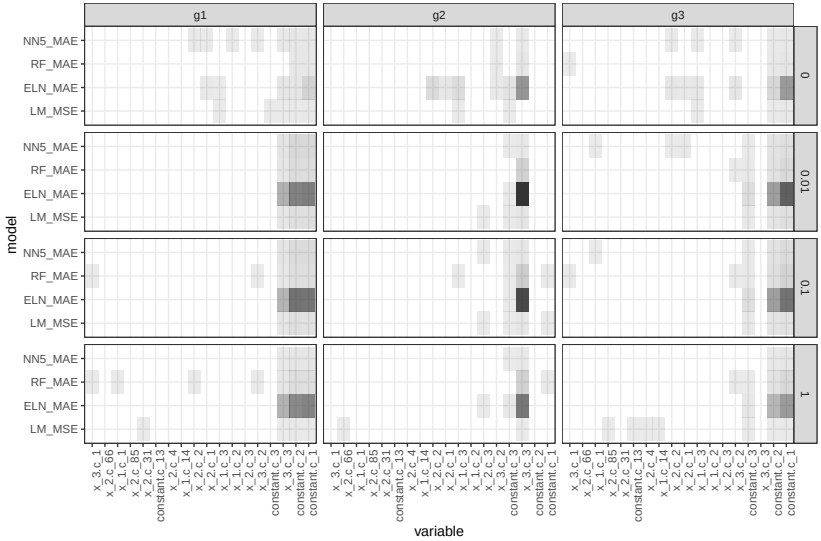

Figure 10: Variable Importance Across All simulation Designs

Empirical Data

We now focus on the Empirical Data

Will see the results from the simulation study are **largely the same**

Data Source

Gu et al. (2018)'s dataset of individual factors available from their website:

- March 1957 - December 2016
- 61 annual factors
- 13 quarterly factors
- 20 monthly factors
- Industry dummy with 74 levels

Data Cleaning Procedure

Reducing dataset size:

- Only include NASDAQ stocks
- Filter out penny stocks (microcaps)
- Keep only instruments with share code of 10, 11 (filtering out REITs, etc)
- Convert to a monthly format
- Industry dummy was dropped due to inaccuracy and high dimensionality

Data Cleaning Procedure

Missing Data

- Significant increase in data quality and availability after 1993 Q3
- Factors with more than 20% missing data were removed
- Remaining missing factors were imputed with cross sectional medians

Macroeconomic Factors

Table 1: Macroeconomic Factors, (Welch and Goyal (2008))

No.	Acronym	Macroeconomic Factor
1	macro_dp	Dividend Price Ratio
2	macro_ep	Earnings Price Ratio
3	macro_bm	Book to Market Ratio
4	macro_ntis	Net Equity Expansion
5	macro_tbl	Treasury Bill Rate
6	macro_tms	Term Spread
7	macro_dfy	Default Spread
8	macro_svar	Stock Variance

Cleaned Dataset

Baseline set of covariates defined as:

$$z_{i,t} - (1, x_t)' \otimes c_{i,t} \tag{18}$$

where $c_{i,t}$ is a P_c matrix of characteristics for each stock i, and $(1, x_t)'$ is a $P_x \times 1$ vector of macroeconomic predictors.

Number of covariates in this baseline set is $61 \times (8 + 1) = 549$.

Sample Splitting

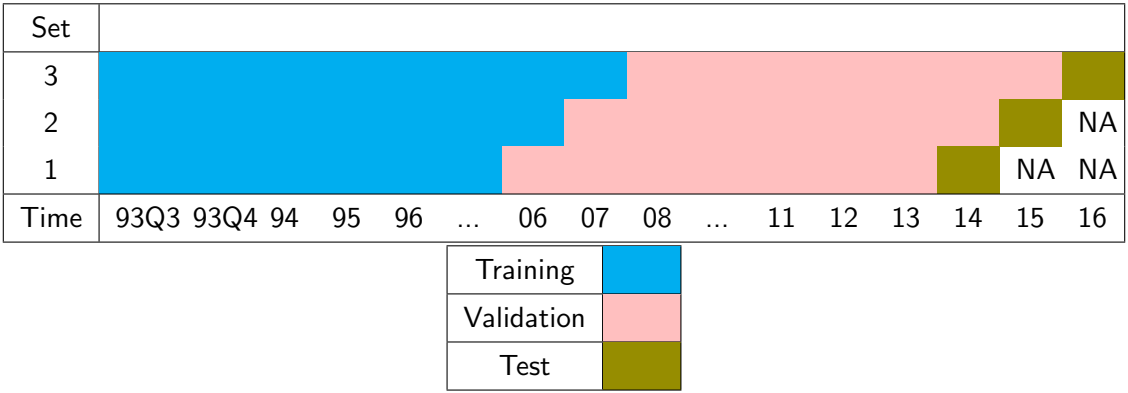

Figure 11: Empirical Data Sample Splitting Procedure

Overview

Results **largely the same** as the simulation study.

- Minimizing quantile loss is better
- Elastic Net $>$ Random Forests $>$ Neural Networks $>$ Linear Models
- Elastic Net and Random Forests tend to agree on same subset of predictors
- Neural Networks **did not outperform**
- Neural Networks are unstable

Suggests that

Contradicts Gu et al. (2018)'s findings.

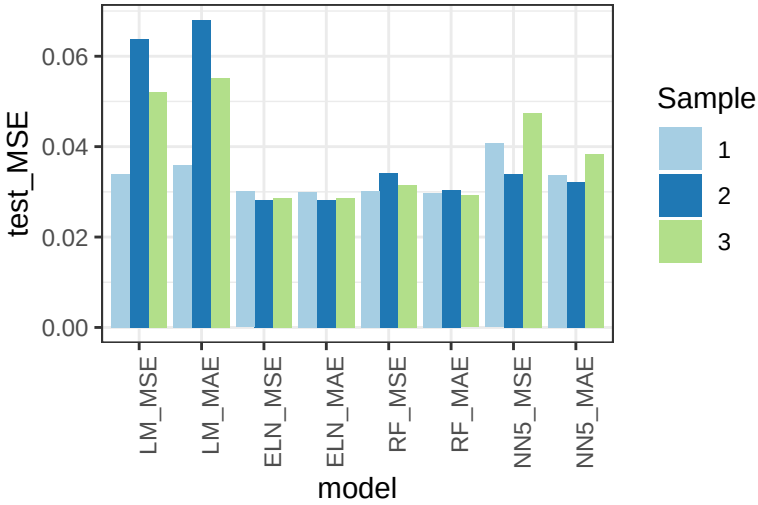

Figure 12: Empirical Data Test MSE

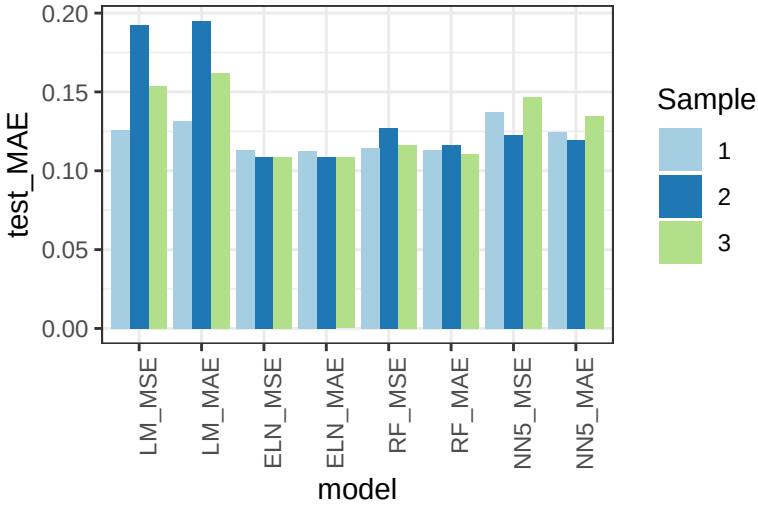

Figure 13: Empirical Data Test MAE

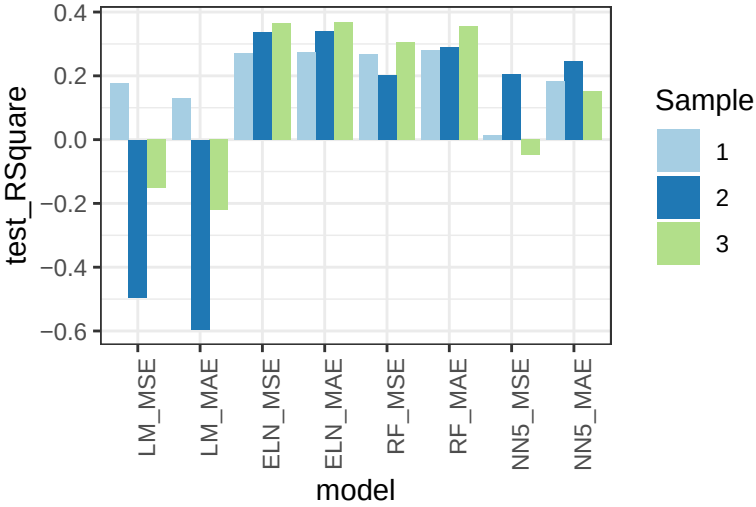

Figure 14: Empirical Data Test R Squared

Prediction Performance Results

Elastic Net $>$ Random Forests $>$ Neural Networks $>$ Linear Models

More data did not seem adversely affect prediction performance - data considered may have had less shocks than simulated datasets

Neural Networks are much more unstable on empirical data, especially when using MSE as a loss function

Weak evidence that deeper neural networks are better

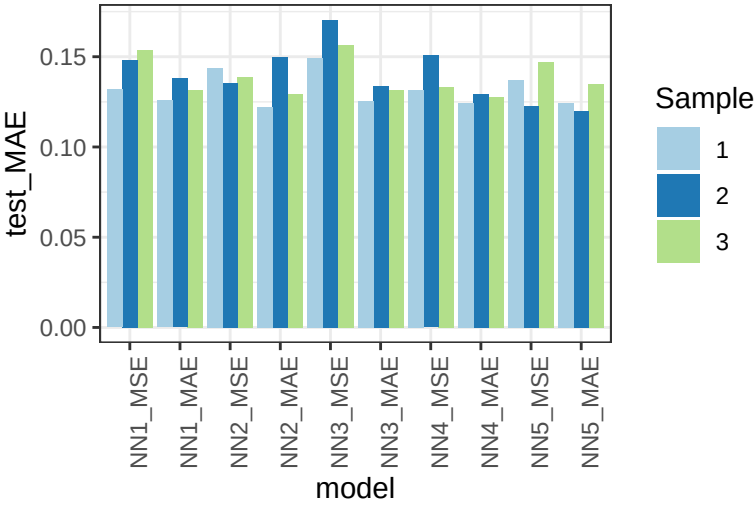

Figure 15: Empirical Data Test MAE for Neural Networks

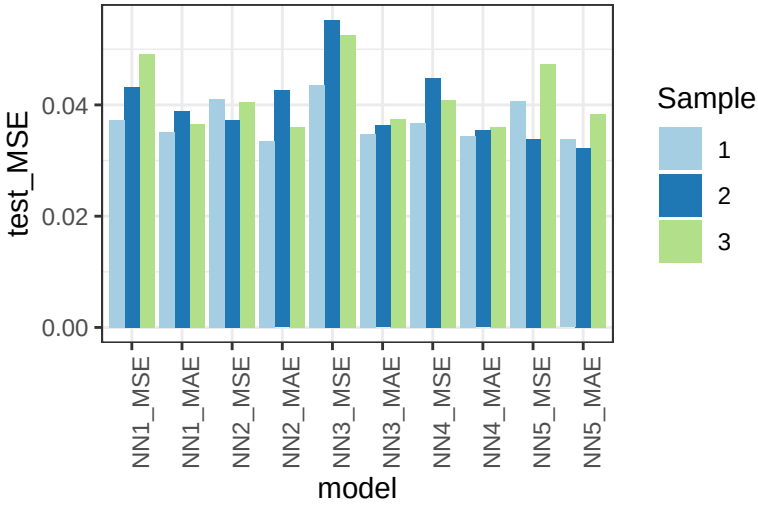

Figure 16: Empirical Data Test MSE for Neural Networks

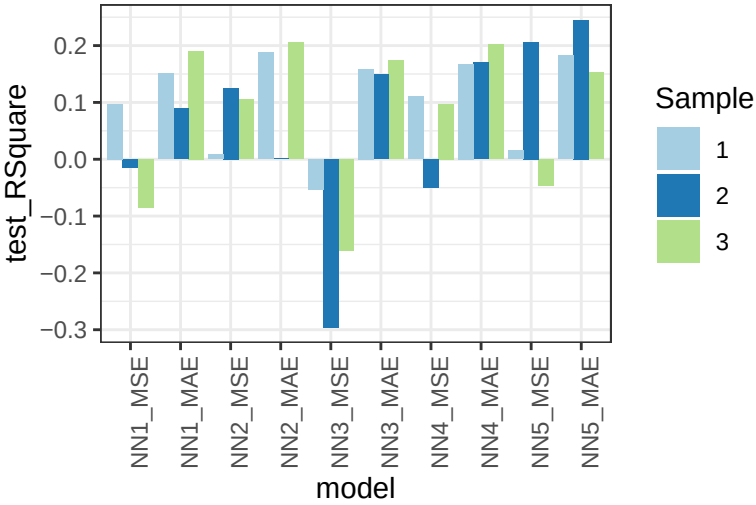

Figure 17: Empirical Data Test R Squared for Neural Networks

Variable Importance Results

Difficult to compare with simulation results, as underlying DGP is unknown

However, similar trends can be observed:

- Elastic Net and Random Forests agree on same subset
- Random Forest struggle to discern between factors
- Neural Networks results are completely different

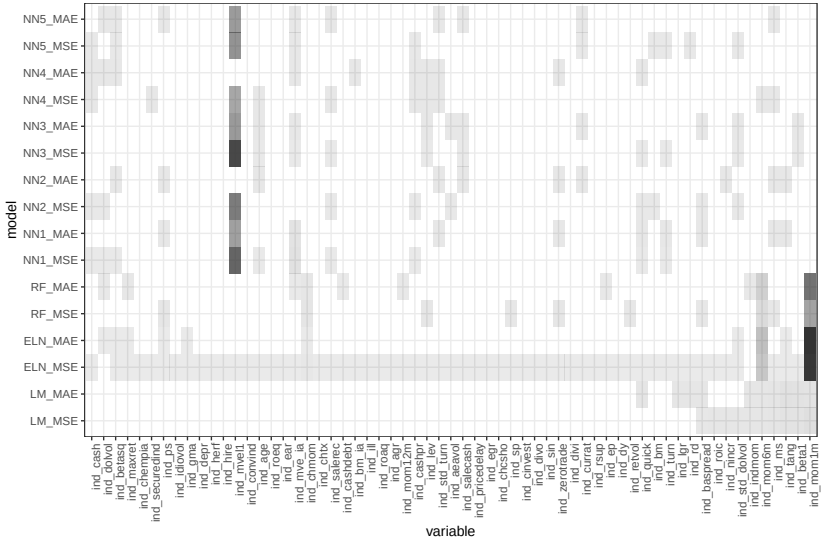

Figure 18: Empirical Individual Factor Importance, averaged over all samples

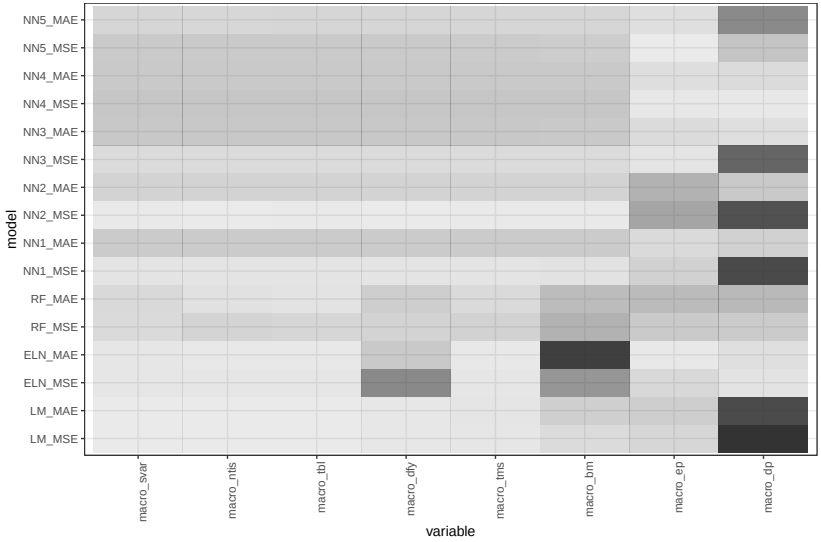

Figure 19: Empirical Macroeconomic Factor Importance, averaged over all samples

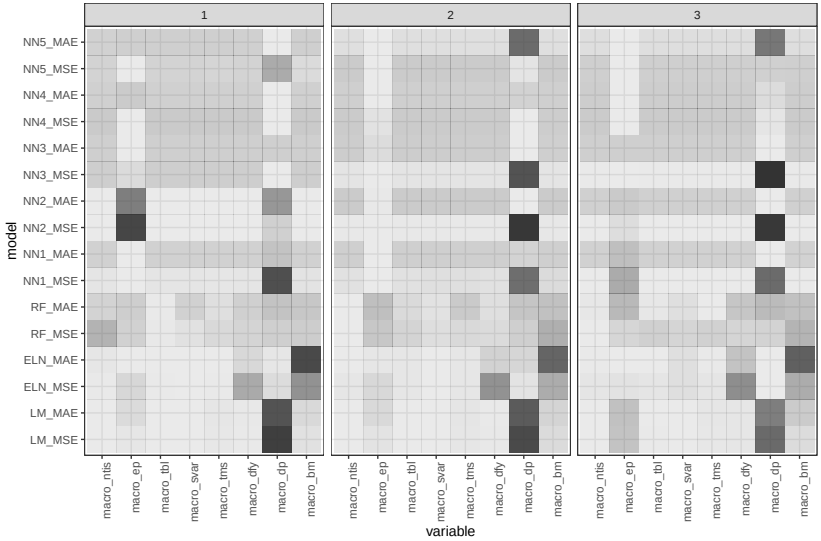

Figure 20: Empirical Macroeconomic Factor Importance, averaged over all samples

Variable Importance

Elastic Net and Random Forests:

- 1 and 6 month momentum factors
- Book-to-market Ratio and Default Spread

Factors in Neural Networks (**inconsistent**):

- Individual Market Value
- Dividend-price Ratio and Earnings-price Ratio

Conclusion

Machine learning offers tools to improve stock return prediction and identification of true underlying regressors.

Elastic Net and Random Forests are the best performing models.

Feed-forward neural networks considered did not outperform.

Elastic Net and Random Forests agree and correctly identify causal regressors. Neural networks only agree in simulated contexts.

Cross sectional correlation does not affect prediction performance of machine learning by much.

Minimizing quantile loss yields better prediction performance.

Contribution to Literature

Overall findings differ from sparse literature on similar topics.

Results are consistent across simulated data and empirical data, suggesting a robust set of results.

Overall, results are promising, especially for Elastic Net.

References

Ang, A., Bekaert, G., 2006. Stock return predictability: Is it there? The Review of Financial Studies 20, 651–707.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Feng, G., Giglio, S., Xiu, D., 2019. Taming the Factor Zoo: A Test of New Factors. Tech. Rep. w25481, National Bureau of Economic Research, Cambridge, MA.

Feng, G., He, J., Polson, N. G., 2018. Deep Learning for Predicting Asset Returns. arXiv:1804.09314 [cs, econ, stat] ArXiv: 1804.09314.

Freyberger, J., Neuhierl, A., Weber, M., 2017. Dissecting characteristics nonparametrically. Tech. rep., National Bureau of Economic Research.

Goetzmann, W. N., Jorion, P., 1993. Testing the predictive power of dividend yields. The Journal of Finance 48, 663–679.

Goyal, A., Welch, I., 2003. Predicting the equity premium with dividend ratios. Management

Questions and Answers