Modern Finance

Modern Finance Institute

Article

Forecasting the equity premium: Do deep neural network models work?

Xianzheng Zhou[1], Hui Zhou [2,3], and Huaigang Long [4]*

[1] Guosen Securities Co., Ltd. Shenzhen Internet Branch, China; zhouxianzheng@zju.edu.cn
[2] Tulane University;
[3] California State University, Sacramento; susan.zhou@csus.edu
[4] Zhejiang University of Finance and Economics; longhuaigang@zufe.edu.cn
* Correspondence: longhuaigang@zufe.edu.cn; School of Finance, Zhejiang University of Finance and Economics, 18 Xueyuan Street, Hangzhou City, Zhejiang Prov, China 310018.

Abstract: This paper constructs deep neural network (DNN) models for equity-premium forecasting. We compare the forecasting performance of DNN models with that of ordinary least squares (OLS) and historical average (HA) models. The DNN models robustly work best and significantly outperform both OLS and HA models in both in- and out-of-sample tests and asset allocation exercises. Specifically, DNN models generate monthly out-of-sample R^2 of 3.42% and an annual utility gain of 2.99% for a mean-variance investor from 2011:1 to 2016:12. Moreover, the forecasting performance of DNN models is enhanced by adding 14 further variables selected from finance literature.

Keywords: equity premium, return predictability, deep neural network, asset allocation, forecasting performance

JEL codes: C45, C53, G10, G12

1. Introduction

Equity premium forecasting is one of the core issues in financial research. It is closely related to many important financial issues, such as portfolio management, capital cost, and market effectiveness (Rapach & Zhou, 2013; Rapach et al., 2010). However, the out-of-sample predictability is still controversial. For example, Welch and Goyal (2008) find that 14 popular predictive variables do not outperform the simple historical average (HA) of returns. However, Campbell and Thompson (2008) point out that equity premium is predictable out-of-sample by adding parameter constraints based on financial theory. Neely et al. (2014) also show that combining information from both macroeconomic variables and technical indicators using principal components analysis (PCA) performs significantly better than the historical average forecast.

Among methods for stock return prediction, traditional linear regression methods have been widely adopted, e.g., OLS (Ordinary Least Squares), LASSO (Least Absolute Shrinkage and Selection Operator, see Tibshirani, 2011), Ridge regression (Tikhonov, 1998). However, literature applying nonlinear methods, especially deep learning, to extract information from the stock return time series is still limited (Bekiros et al., 2016; Gupta et al., 2018). The ability to extract and transform features from data, and to identify hidden nonlinear relations without relying on econometric assumptions and human expertise, makes deep learning much more attractive than other machine learning methods. On the other hand, the number of conditioning variables that are believed to have forecasting power for returns is large and has continued to increase over the last five decades. The traditional methods are reaching their limits on handling a large number of conditioning variables, so more advanced statistical tools, such as deep learning, can be a

Citation: Zhou, X., Zhou, H., & Long, H. (2023). Forecasting the equity premium: Do deep neural network models work? *Modern Finance*, 1(1), 1-11.

Accepting Editor: Adam Zaremba

Received: 30 June 2023
Accepted: 3 August 2023
Published: 5 August 2023

solution (Gu et al., 2018). As one of the most popular deep learning methods, Deep Neural Network (DNN) DNN does not require manual indicator selection and enables us to apply much more variables as inputs. In this paper, we apply DNN method to directly forecast the U.S. equity premium and compare the result with that of OLS regression method.

Specifically, following Neely et al. (2014), we compare the forecasting performance(measured by $MSFE_{\mathrm{OS}}$, R^2_{OS}, and MSFE-adjusted statistic) of the Ordinary Least Squares models using 28 input variables (OLS+28) with Deep Neural Network models using the same 28 input variables (DNN+28) and Deep Neural Network models using the same 28 factors and additional 14 variables (DNN+42). Next, following Kandel and Stambaugh (1996) and Welch and Goyal (2008), we use the out-of-sample forecasts to compute the Certainty Equivalent Return (CER) gain and Sharpe ratio for mean-variance investors who optimally allocate their wealth between equities and risk-free bills. Our results show that the OLS+28 model has a surprisingly poor performance over the out-of-sample period 2011:01-2016:12, which Neely et al. (2014) didn't test due to data availability. In contrast, the two DNN models both have good performances. The R^2_{OS} of DNN models are near 3%, and the DNN models generate large and robust economic gains for investors with an annualized CER gain at around 3%. The monthly Sharpe ratio of DNN models substantially outperforms HA and OLS+28 model.

Our study contributes to the existing literature in three ways. First, to the best of our knowledge, we are the first to apply deep learning——one of the hottest IT technologies——to forecast equity premiums in a finance academic paper. Unlike most studies focusing on traditional econometric models, we introduce a nonlinear machine learning model to forecast equity premiums. Our results show that DNN models can outperform HA models and OLS models. Especially, we find the poor predictive ability of OLS models during the period 2011:01-2016:12, which is beyond the period studied by Neely et al. (2014). However, the DNN models still work well in this period. Second, we test whether DNN models can incorporate more predictive information from additional 14 variables selected from existing finance literature. The results show that the forecasting performance of DNN can be improved by inputting more variables. These, in turn, verify the existing finance literature. Last but not least, our asset allocation results indicate that DNN models can be applied to practical investment management and produce a large number of economic values.

The rest of the paper is organized as follows. Section 2 presents the methodology and data. Section 3 discusses the empirical results. Section 4 concludes the paper.

2. Methodology and Data

2.1. HA model

Welch and Goyal(2008) argue that simple historical average(HA) forecasts equity premium better than regressions equity premium on predictors including 14 popular macroeconomic variables. So our first benchmark model is HA model, which can be expressed as follows:

$$R_{t+1} = \frac{1}{t}\sum_{s=1}^{t} R_s \tag{1}$$

where R_t is the equity premium at month t.

2.2. OLS model

Based on PCA and OLS predictive regression framework, Neely et al. (2014) find that, compared with HA model, combining information from both 14 macroeconomic variables and 14 technical variables significantly improves equity premium forecasts. We repeat their study and define OLS models as follows:

$$R_{t+1} = \alpha_i + \beta_i x_{i,t} + \varepsilon_{i,t+1} \tag{2}$$

where R_{t+1} is the equity premium at month t+1, $x_{i,t}$ is the predictor i at month t. Based on data through t, we can get $\hat{\alpha}_{i,t}, \hat{\beta}_{i,t}$ from the OLS estimate of $\alpha_{i,t}, \beta_{i,t}$. Then the out-of-sample forecast $\hat{R}_{t+1}$ is

$$\hat{R}_{t+1} = \hat{\alpha}_i + \hat{\beta}_i x_{i,t} \tag{3}$$

Especially, we denote the OLS regression on principal components extracted from these 28 variables studied by Neely et al. (2014) as ***"OLS+28"*** model.

2.3. DNN model

Our DNN models have the following general equations:

$$x_1^{(l)} = \text{ReLU}[\text{BN}(x^{(0)\prime})\theta_1^{(0)}] \tag{4}$$

$$x_n^{(l)} = \text{ReLU}(\text{BN}(x^{(l-1)\prime})\theta_n^{(l-1)}) \tag{5}$$

$$\hat{R}_{t+1} = x^{(N^{(l)}-1)\prime}\theta^{(N^{(l)}-1)} \tag{6}$$

where $N^{(l)}$ denotes the number of neurons in each layer $l \in \{1, \ldots, N^{(l)}\}$. We define the output of neuron n in layer l as $x_n^{(l)}$ and the vector of outputs for this layer (augmented to include a constant,$x_0^{(l)}$) as $x^{(l)} = (1, x_1^{(l)}, \ldots, x_{N^{(l)}}^{(l)})'$. The number of units in the input layer is equal to the dimension of the variables, and let $x^{(0)} = (1, x_1, \ldots, x_m)'$, where x_m is the m-th input variable. Let $\theta_n^{(l-1)}$ denotes weight and bias parameters in each layer $l \in \{1, \ldots, N^{(l)}\}$. $\hat{R}_{t+1}$ is the forecast of log equity premium at month t+1. Rectified linear unit (ReLU) is the most popular activation function (Nair and Hinton, 2010) and we use this at all nodes. Batch normalization (BN) is a simple regularization technique for controlling the variability of variables across different regions of the network and across different datasets (Nair and Hinton, 2010). Equation (4) states the relationship between the input variables in input layer and the output vectors in the first hidden layer. Equation (5) shows the recursively output formula for the neural network at each neuron in layer l. And equation (6) gives the final output of forecasting results. For comparing with HA and OLS+28 models, we first apply the same 28_variables as input to the OLS+28 model and DNN+28 model. Then, in order to examine whether DNN models can extract information from the 14 additional predictors to improve the forecast performance, we add 14 additional variables selected from existing finance literature and obtain the DNN+42 models.

At present, there is no uniform approach to determine the best parameters such as number of layers and neurons for DNN on a given problem. Since Gu et al. (2018) suggest that shallow learning outperforms the relatively deeper learning, we choose three or four hidden layers to start search in our study. To solve this nonlinearity and nonconvexity problem, we use adaptive moments method (Adam, Kingma, et al., 2014) to train our DNN models and grid search method to select the best one.[1] Finally, DNN+28 models take 200, 200, 200, and 128 neurons in four hidden layers and 0, 0.5, 20 as the values of the weight decay of Adam, dropout probability and epochs respectively. For DNN+42 models, these values are 600, 300, 300 in three hidden layers and 0, 0.5, 10, respectively. For robustness check, we will discuss the effect of those key parameters on forecasting performance.

DNN models tend to suffer from overfitting when tuning parameters to achieve satisfactory results. Four methods are applied to prevent overfitting: First, we shrink the weight parameters of DNN model via L2 penalized estimation method, because the method can control the weight of regularization term in loss function. Second, we apply

[1] Adam is a commonly used optimization method for deep learning, just like that of OLS for linear regression. Grid method research, which is a traditional way of performing hyperparameter optimization, is simply an exhausting searching through a manually specified subset of the hyperparameter space of a learning algorithm.

dropout technique to prevent overfitting and co-adaptations of neurons, and set the output of any neuron to zero with probability p. Models with dropout can be interpreted as an ensemble of models with different numbers of neurons in each layer, but also with weight sharing, and thus can enhance generalization ability (Srivastava et al. 2014). Third, early stopping method is adopted to determine the best training epoch. And we stop training once the model performance stops improving on test datasets. Finally, we use the batch normalization algorithm, which normalizes the input of each layer to ensure that the input data of each layer is stable, thus achieving the purpose of speeding up training and improving generalization ability.

2.4. Forecast Evaluation Measures

Following Neely et al. (2014) and Welch and Goyal (2008), we employ two kinds of forecast evaluation measures. *First,*OS_2 *and MSFE-adjusted Statistics.* R_{OS}^2 measures the forecasting accuracy versus benchmark HA model and a monthly R_{OS}^2 of 0.5% is economically significant (Campbell & Thompson, 2008). MSFE-adjusted statistic measures the statistical significance (Clark & West, 2007). *Second, Asset Allocation Performance measured by* following six measures: (1) certainty equivalent return gain [CER gain, $\triangle(ann\%)$], (2) CER gain in expansions [$\triangle(ann\%)$, EXP], (3) CER gain in recessions [$\triangle(ann\%)$, REC], (4) Sharpe ratio, (5) Relative average turnover, (6) CER gain with 50bps per transaction [$\triangle(ann\%)$, cost = 50bps].[2]

2.5. Data

The dataset used covers the monthly period from 1950:12 to 2016:12, based on data availability. The equity premium R_t is computed as the difference between the log return on the S&P 500 (including dividends) and the log return on a risk-free bill. As mentioned before, in order to compare the forecasting performance of our considered models, we select 48 predictors[3]. These are consisted of three groups: 14 macroeconomic variables from Welch and Goyal (2008), 14 technical variables from Neely et al. (2014), and 14 additional variables from existing finance literatures including investors sentiment changes (Wurgler and Baker, 2006), financial stress index (Cardarelli et al., 2011), ratio of 52-week high (George & Hwang, 2004),etc.[4]

Table 1 reports the summary statistics for the log equity premium (1950:12-2016:12), macroeconomic variables (1950:12-2016:12), technical variables (1950:12-2016:12), and additional variables (1965:08-2016:12). The average monthly equity premium (0.004) divided by its standard deviation (0.043) produces a monthly Sharpe ratio value of 0.088. Most of the macroeconomic variables and additional variables are strongly auto-correlated.

[2] For brevity, the details for the definition and computation of all these measures can be seen from Goyal & Welch (2008) and Neely et al. (2014) .

[3] Online Appendix Table A1 provides detailed information on the source and computation methods of the 48 variables.

[4] The macroeconomic variables are collected from Amit Goyal's webpage at http://www.hec.unil.ch/agoyal/. The technical variables are available from Guofu Zhou's webpage at http://apps.olin.wustl.edu/faculty/zhou/. The investor sentiment data are collected from http://people.stern.nyu.edu/jwurgler/.

Table 1. Summary Statistic

	Mean	Median	Std	Min	Max	Auto-cor	Skewness	Kurtosis		Mean	Median	Std	Min	Max	Auto-cor.	Skewness	Kurtosis
Panel A: Log equity premium, December 1950 to December 2016																	
	0.004	0.008	0.043	-0.248	0.149	0.049	-0.669	2.535									
Panel B: Macroeconomic variables, December 1950 to December 2016									*Panel C: Technical variables, December 1950 to December 2016*								
DP	-3.602	-3.531	0.412	-4.524	-2.753	0.994	-0.134	-0.872	MA(1,9)	0.677	1	0.468	0	1	0.703	-0.761	-1.425
DY	-3.597	-3.525	0.412	-4.531	-2.751	0.994	-0.139	-0.848	MA(1,12)	0.708	1	0.455	0	1	0.780	-0.919	-1.160
EP	-2.831	-2.860	0.449	-4.836	-1.899	0.989	-0.723	2.648	MA(2,9)	0.684	1	0.465	0	1	0.748	-0.793	-1.375
DE	-0.771	-0.815	0.320	-1.244	1.379	0.986	2.961	15.854	MA(2,12)	0.705	1	0.456	0	1	0.821	-0.901	-1.191
RVOL	0.145	0.135	0.051	0.055	0.316	0.963	0.799	0.549	MA(3,9)	0.686	1	0.465	0	1	0.785	-0.801	-1.362
BM	0.498	0.414	0.270	0.121	1.207	0.994	0.761	-0.465	MA(3,12)	0.703	1	0.457	0	1	0.817	-0.893	-1.207
NTIS	-0.010	-0.013	0.020	-0.051	0.058	0.979	0.650	0.265	MOM(9)	0.703	1	0.457	0	1	0.767	-0.893	-1.207
TBL	-4.866	-4.970	3.275	-16.300	-0.010	0.990	-0.527	0.596	MOM(12)	0.728	1	0.445	0	1	0.804	-1.026	-0.951
LTY	-6.772	-6.460	2.683	-14.820	-1.750	0.993	-0.589	0.132	VOL(1,9)	0.666	1	0.472	0	1	0.609	-0.706	-1.506
LTR	0.639	0.510	3.054	-11.240	15.230	0.037	0.380	2.275	VOL(1,12)	0.687	1	0.464	0	1	0.709	-0.809	-1.349
TMS	1.905	2.060	1.507	-3.650	4.550	0.955	-0.464	-0.172	VOL(2,9)	0.660	1	0.474	0	1	0.761	-0.675	-1.549
DFY	1.062	0.940	0.448	0.320	3.380	0.964	1.754	4.229	VOL(2,12)	0.690	1	0.463	0	1	0.825	-0.826	-1.322
DFR	0.011	0.060	1.498	-9.750	7.370	-0.064	-0.348	6.146	VOL(3,9)	0.676	1	0.468	0	1	0.770	-0.753	-1.437
INFL	-0.330	-0.305	0.359	-1.792	1.915	0.619	0.160	3.458	VOL(3,12)	0.682	1	0.466	0	1	0.835	-0.785	-1.388
Panel D: Additional variables, August 1965 to December 2016																	
PDND	-4.658	-6.194	13.58	-50.23	31.632	0.970	0.147	0.147	WH52_Ratio	0.936	0.965	0.083	0.51	1.04	0.891	-1.858	3.915
RIPO	16.808	12.700	19.44	-28.80	119.10	0.648	2.112	6.403	WH52_Abs	0.154	0.000	0.361	0.00	1.00	0.079	1.922	1.700
NIPO	25.916	19.000	23.23	-	122.00	0.862	1.203	1.079	DV	0.010	0.009	0.003	0.01	0.02	0.997	0.649	-0.287
CEFD	8.674	9.220	7.343	-10.91	25.28	0.962	-0.124	-0.327	WV	0.009	0.009	0.002	0.00	0.01	0.998	0.128	-0.778
S	0.172	0.151	0.086	0.045	0.430	0.994	0.946	0.348	AV	0.009	0.009	0.003	0.01	0.02	0.992	0.592	-0.595
ΔSENT	0.001	0.032	0.942	-3.616	5.416	0.086	0.289	2.882	VAR005	0.060	0.058	0.015	0.03	0.08	0.980	0.024	-1.063
FS	100.77	100.74	0.894	98.359	105.89	0.857	0.621	2.229	VAR001	0.078	0.080	0.020	0.04	0.11	0.981	-0.191	-1.054

These popular 14 macroeconomic variables including log of dividend-price ratio[log(DP)], log of Dividend yield[log(DY)], log of earnings-price ratio[log(EP)], log of dividend-payout ratio[log(DE)], equity risk premium volatility(RVOL), book-to-market ratio(BM), net equity expansion(NTIS), treasury bill rate(TBL), long-term yield(LTY), long-term return(LTR), term spread(TMS), default yield spread(DFY), default return spread(DFR), inflation(INFL). These 14 technical variables based on three popular trend

following strategies: moving average variables (MA), momentum variables (MOM), and volume variables (Vol). We choose different parameters and obtain 14 technical variables: MA(1,9), MA(1,12), MA(2,9), MA(2,12), MA(3,9), MA(3,12), MOM(9), MOM(12), VOL(1,9), VOL(1,12), VOL(2,9), VOL(2,12), VOL(3,9), VOL(3,12), and all these 14 technical variables are binary variable. These 14 additional variables including dividend premium (PDND), number of IPOs(RIPO), average first-day returns (NIPO), closed-end fund discount(CEFD), equity share in new issues(S), sentiment changes(ΔSENT), financial stress(FS), ratio of 52-week high(WH52_Ratio), absolute of 52-week high(WH52_Abs), daily volatility(DV), weekly volatility(WV), annual volatility(AV), 5% of the quantile in the past 60 months(VAR005), 1% of the quantile in the past 60 months(VAR001).

3. Empirical results

Similar to Neely et al. (2014), these models are estimated in-sample using recursively expanding windows with initial length of 15 years. We divide out-of-sample period into three panels: panel A (1966:01-2011:12), panel B (1980:09-2010:12), and panel C (2011:01-2016:12).[5] We report results in each panel for the whole period along with NBER-date business-cycle expansions and recessions period.[6]

3.1. In-sample test results

Table 2 reports the in-sample results of HA, OLS+28, DNN+28, DNN+42 models for the three panels. The results in Panel A, show that OLS+28 models can beat the HA models in terms of *MSFE* and R^2, which is consistent with Neely et al. (2014).[7] Overall, the in-sample results of DNN models outperform HA and OLS+28 models on all the three panels, which is not affected by business-cycle expansions or recessions

Table 2. In-sample Test Results

Model	$MSFE_{IS}$	R^2_{IS} (%)	R^2_{IS}. EXP (%)	R^2_{IS} REC (%)
		Panel A: January 1966 to December 2011		
HA	20.23			
OLS+28	15.15	0.05	0.42	0.52
DNN+28	15.47	3.03	1.08	5.48
		Panel B: September 1980 to December 2010		
HA	20.54			
OLS+28	16.24	0.04	0.29	0.41
DNN+28	15.47	3.03	1.08	5.48
DNN+42	18.56	3.72	0.50	6.96
		Panel C: January 2011 to December 2016		
HA	10.67			
OLS+28	17.49	1.81	1.81	-
DNN+28	17.37	2.62	2.62	-
DNN+42	18.59	3.81	3.81	-

This table reports the in-sample performance of various measures of forecast models estimated using recursively expanding windows with 180 initial months. MSFEIS is the in-sample mean squared forecast error. *R*IS 2 measures the mean of the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. *R*IS 2 statistics are also calculated separately for NBER-dated expansions (EXP) and recessions (REC).

[5] The reason why we divided according to panel A is to reproduce Christopher et al. (2014)'s result and facilitate the comparison with the DNN models. Panel B is because the forecasting start time of 14 additional variables is 1980:09, which is convenient compared with the DNN+48 model. Panel C is designed to test and compare the out-of-samples which nearly didin't be examined by Christopher et al. (2014) due to the time limited.

[6] More details for NBER-date business-cycle period can be found at https://www.nber.org/cycles/cyclesmain.html.

[7] We report estimated slope coefficients, MSFE and R2 of the bivariate predictive regression between the equity premium and one predictor variable as showed in Christopher et al. (2014) in Online Appendix Table A2.

3.2. Out-of-Sample forecasting results

Table 3 provides the out-of-sample forecasting results of models.[8] From Panel A of Table 3, in terms of R_{OS}^2 and $MSFE$OS, the OLS+28 model outperforms the HA model from 1966:01 to 2011:12, which have almost the same results as those of Neely et al. (2014). However, Panel B shows that the performance of OLS+28 model in each panel is worse than the HA model since 1980:09. This means that the OLS+28 model only performs better than the HA model in the former 15 years. Besides, the OLS+28 model obtains significantly large positive R_{OS}^2 (11.37%, 10.64% in panel A and panel B, respectively) during recessions, but disappointingly negative R_{OS}^2 during expansions (-2.63%, -4.14% in panel A and panel B, respectively). This suggests that the OLS+28 model's strong performance on the whole sample is largely due to high R_{OS}^2 values during recessions. From Panel C, it further shows that, surprisingly, the OLS+28 model displays no out-of-sample predictive ability in terms of R_{OS}^2 (- 5.02%) from 2011:01 to 2016:12, a period that has not been examined by Neely et al. (2014). Overall, the OLS+28 model does not have good predictive robustness.

Turning to our proposed DNN models, the results in Table 3 show that both DNN+28 and DNN+42 model strongly beat the simple HA benchmark and the OLS+28 model in terms of MSFE and R_{OS}^2. The out-of-sample MSFEs for DNN models are significantly less than that of HA and OLS+28 at the conventional confidence level. Impressively, it is worth pointing out that the R_{OS}^2, EXP statistics of DNN models overwhelmingly beat the OLS+28 model and are positive in each panel. These indicate that DNN models can outperform the HA model both in expansions and recessions, and have good robustness.

Table 3. Out-of-Sample Forecasting Results

Model	$\boldsymbol{MSFE}_{\text{IS}}$	$\boldsymbol{R}^2_{\text{IS}}$ (%)	$\boldsymbol{R}^2_{\text{IS}}$. EXP (%)	$\boldsymbol{R}^2_{\text{IS}}$ REC (%)
		Panel A: January 1966 to December 2011		
HA	20.23			
OLS+28	15.15	0.05	0.42	0.52
DNN+28	15.47	3.03	1.08	5.48
		Panel B: September 1980 to December 2010		
HA	20.54			
OLS+28	16.24	0.04	0.29	0.41
DNN+28	15.47	3.03	1.08	5.48
DNN+42	18.56	3.72	0.50	6.96
		Panel C: January 2011 to December 2016		
HA	10.67			
OLS+28	17.49	1.81	1.81	-
DNN+28	17.37	2.62	2.62	-
DNN+42	18.59	3.81	3.81	-

Note. This table reports the out-of-sample performance of various measures of forecast models estimated using recursively expanding windows with 180 initial months. *MSFEOS* is the out-of-sample mean squared forecast error. ROS 2 measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. *MSFE-adjusted* is the Clark and West (2007) statistic for testing the null hypothesis that the historical average forecast MSFE is less than or equal to the competing forecast MSFE against the one-sided (upper-tail) alternative hypothesis that the historical average forecast MSFE is greater than the competing forecast MSFE. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively. Ros^2 statistics is also reported separately for NBER-dated expansions (EXP) and recessions (REC). The out-of-sample evaluation periods vary in terms of different panels

[8] Similarly to Neely et al. (2014), we report out-of-sample forecasts for the bivariate OLS regression model between the equity risk premium and single variable, PC-ECON, PC-TECH in Online Appendix Table A3.

Moreover, it shows that, overall, the performance of DNN+42 model are relatively better than the DNN+28 models.[9] Especially, the DNN+42 model has an R_{OS}^2 of 3.37% in Panel B of Table 2, which significantly exceeds the R_{OS}^2 of 1.49% of DNN+28 model. The out-of-sample MSFEs of DNN+42 model are much less than that of HA models at the 1% confidence level. Thus, the results suggest that the forecasting performance of DNN modes is enhanced by incorporating 14 additional variables.

3.3. Asset allocation results

Table 4 reports the portfolio performance for asset allocation over 1966:01-2016:12. In accord with the R_{OS}^2 in Table 1, the OLS+28 model does not uniformly get robustness performance in terms of △(*ann*%), △(*ann*%), EXP, and △(*ann*%), REC in Table 4.

Turning to the performance of DNN models, Table 4 shows that CER gains in both recessions and expansions are positive. Besides, though the turnover is relatively high compared with HA and OLS+28 models, the CER gains with a proportional transactions cost of 50 basis points per transaction are still positive. From the perspective of asset allocation, the DNN+28 models also obtain good performance. Table 4 consistently confirms that the DNN+42 model outperforms the DNN+28 model in terms of CER gain and Sharpe ratio. DNN+42 models generate monthly out-of-sample $R2$ of 3.42% and annual utility gain of 2.99% for a mean-variance investor from 2011:1 to 2016:12. The asset allocation analysis demonstrates a substantial economic value of employing DNN models for equity premium forecasting.

Table 4. Portfolio Performance Measures (Risk aversion coefficient (γ) = 3)

Model	△(ann%)	△(ann%), EXP	△(ann%), REC	Sharpe ratio	Relative average turnover	△(ann%), cost =50 bps
Panel A: January 1966 to December 2011						
HA(CER)	4.87	9.33	-17.52	0.06	2.66%	4.70
OLS+28	5.07	0.05	30.33	0.16	6.43	4.20
DNN+28	4.40	1.46	18.99	0.14	13.64	2.37
Panel B: September 1980 to December 2010						
HA(CER)	7.12	11.54	-17.61	0.10	2.63%	6.95
OLS+28	2.77	-1.57	26.96	0.16	5.18	2.09
DNN+28	2.49	1.13	9.90	0.15	14.36	0.37
DNN+42	4.48	0.32	27.65	0.20	19.85	1.48
Panel C: January 2011 to December 2016						
HA(CER)	8.35	8.35	-	0.26	2.31%	8.21
OLS+28	-4.56	-4.56	-	0.16	12.50	-6.19
DNN+28	2.88	2.88	-	0.31	7.78	1.95
DNN+42	2.99	2.99	-	0.33	16.52	0.84

Note: This table reports the portfolio performance measures for a mean-variance investor who allocates capital monthly between equities and risk-free bills using the monthly out-of-sample forecast results of the U.S. equity premium based on different forecast models. The utility gain △(ann%) is the annualized certainty equivalent return gain for the investor with risk aversion coefficient of three. △(ann%) statistics are also reported separately for NBER-dated expansions (EXP) and recessions (REC). The monthly Sharpe ratio is the mean portfolio return in excess of the risk-free rate divided by its standard deviation. The out-of-sample evaluation periods are varies in terms of different panels. Relative average turnover is the average turnover for the portfolio based on the model forecast divided by the average turnover for the portfolio based on the historical average forecast. The △(ann%), cost=50bps is the CER gain assuming a proportional transactions

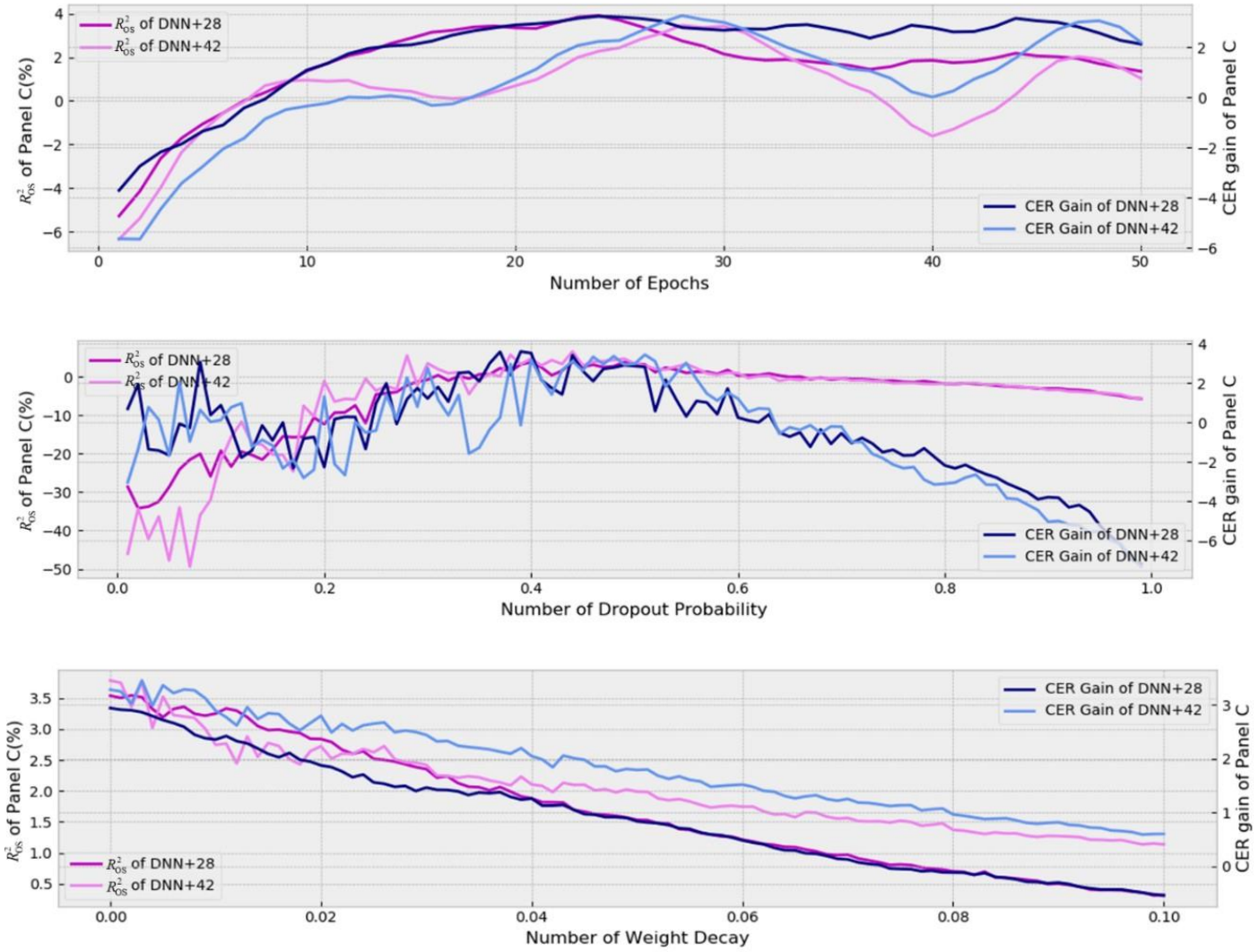
[9] To show the out-of-sample forecast result for these models visually, we depict out-of-sample forecast log equity premium for these models over three panels in Online Appendix Figure A1. It suggests that DNN models have a lower forecast variance and can more correctly reflect actual market fluctuations during expansions than OLS models.

cost of 50 basis points per transaction. For comparison, the historical average model [HA(CER)] is the annualized certainty equivalent return, using as a benchmark for other models to compute the utility gain.

3.4. Robustness checks

To further validate our results, we conduct the following robustness checks. First, the effects of the number of DNN models' epochs, dropout probability and weight decay on forecasting performance are displayed in **Figure 1**.[10] It shows that these key parameters have good performance near the optimal value. Second, we report the out-of-sample forecasting results year by year for our models (Table A11 in the Online Appendix). Finally, we check the results of asset allocation exercise with risk aversion coefficients equal to 4,5,6 (Table A4 – Table A10 in the Online Appendix). Overall, these robustness checks confirm that DNN models indeed work better than HA models and OLS models for forecasting equity premium.

Figure 1. Effects of the Number of DNN Models' Epochs, Dropout Probability and Weight Decay on R_{OS}^2 and CER Gains in Panel C



These figures depict how R_{OS}^2 and delta CER change with the number of' epochs, dropout probability and weight decay applied to DNN models in panel C (2011:01-2016:12). R_{OS}^2 measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative

[10] More details about the effects in panel B can be found at Figure A2 in the Online Appendix.

to the historical average benchmark forecast. The Delta CER [△(ann%)] measures the annualized certainty equivalent return gain for the investor with risk aversion coefficient of three for DNN models.

4. Conclusion

This study compares the predictive ability of deep neural network models with that of ordinary least squares models and historical average models. We find that DNN models robustly work the best and significantly outperform both OLS and HA models in both in- and out-of-sample tests and asset allocation exercises. Moreover, the forecasting performance of DNN is enhanced by adding 14 additional variables selected from finance literature, which indicates that the DNN comprehensively incorporates the predictive information contained in these variables. One possible explanation for their excellent performance is that the nonlinear DNN automatically extract high dimension features from data and discover different forecasting patterns in data. Our study is of great significance to portfolio construction and risk management for investors.

Supplementary Materials: Online Appendix is available from the authors.

Author Contributions: Conceptualization, Xianzheng Zhou, and Huaigang Long.; methodology, Xianzheng Zhou.; software, Xianzheng Zhou; validation, Xianzheng Zhou., Huaigang Long, and Hui Zhou.; formal analysis, Xianzheng Zhou; investigation, Xianzheng Zhou; resources, Hui Zhou; data curation, Hui Zhou ; writing—original draft preparation, Xianzheng Zhou; writing—review and editing, Huaigang Long; visualization, Hui Zhou; supervision, Huaigang Long; project administration, Huaigang Long; funding acquisition, Huaigang Long. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The processed data from this study are available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *Journal of Finance*, 61(4), 1645-1680.

Bekiros, S., Gupta, R., & Majumdar, A. (2016). Incorporating economic policy uncertainty in US equity premium models: A nonlinear predictability analysis. *Finance Research Letters*, 18, 291-296.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1-127.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average? *Review of Financial Studies*, 21(4), 1509-1531.

Cardarelli, R., Elekdag, S., & Lall, S. (2011). Financial stress and economic contractions. *Journal of Financial Stability*, 7(2), 78-97.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291-311.

George, T. J., & Hwang, C. Y. (2004). The 52-week high and momentum investing. *Journal of Finance*, 59(5), 2145-2176.

Gu, S., Kelly, B. T., & Xiu, D. (2018). Empirical Asset Pricing via Machine Learning. SSRN working paper. http://dx.doi.org/10.2139/ssrn.3159577.

Gupta, R., Mwamba, J. W. M., & Wohar, M. E. (2018). The role of partisan conflict in forecasting the U.S. equity premium: A nonparametric approach. *Finance Research Letters*, 25, 131-136.

Kandel, S., & Stambaugh, R. F. (1996). On the predictability of stock returns: an asset-allocation perspective. *Journal of Finance*, 51(2), 385-424.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Nair, V., & Hinton, G. E. (2010). *Rectified Linear Units Improve Restricted Boltzmann Machines*. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), 807-814.

Neely, C. J., Rapach, D. E., & Tu, J. et al. (2014). Forecasting the Equity Risk Premium: The Role of Technical Indicators. *Management Science*, 60(7), 1772-1791.

Rapach, D. E., Strauss, J. K., & Zhou, G. (2010). Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *Review of Financial Studies*, 23(2), 821-862.

Rapach, D., & Zhou, G. (2013). Forecasting stock returns. In *Handbook of Economic Forecasting* (pp. 328–383). Elsevier B.V.

Srivastava, N., Hinton, G., & Krizhevsky, A. et al. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958.

Tibshirani, R. (2011). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 73(3), 267-288.
Tikhonov, A. N., Leonov, A. S., & Yagola, A. G. (2018). *Nonlinear ill-posed problems*. London: Chapman & Hall. ISBN 0412786605.
Welch, I., & Goyal, A. (2008). A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *Review of Financial Studies*, 21(4), 1455-1508.