Exam paper for: Concepts in Social Data Analytics (T) (LA E23 CBUSV2201U)

Student number: S159569

Date of submission: 18 December 2023

Normal pages: 15

Count of characters: 34.068

DATA-DRIVEN FORECASTING

Machine Learning's Impact on Asset Pricing

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Abstract

This paper evaluates the viability of using both an elastic net (EN) and a neural network (NN) for predicting excess stock returns in R. It applies principles from social data analytics and finance to process, analyse, and forecast excess stock returns. The data set utilised in this research consists of 7,630 individual stocks spanning a 17-year period from 2005 to 2021. The findings indicate that the shallow NN outperforms the deep NN. Nevertheless, neither the EN nor the shallow NN exhibits a dominant set of predictive signals. To reinforce these results, this study recalibrates the excess returns by incorporating relevant risk factors through the Capital Asset Pricing Model (CAPM) and the Fama-French Three-Factor Model (FF3) regression. Despite the inclusion of additional risk factors, the predictive performance of the models remains unimproved, thus supporting the hypothesis of efficient markets.

Keywords: Elastic Net, Neural Network, Excess Stock Returns, Social Data Analytics, Finance, Predictive Analytics.

Contents

1 Introduction

In today's dynamic financial landscape, the practice of forecasting stock prices has advanced significantly with the introduction of machine learning. The traditional analysis of excess stock returns, which requires considerable time and effort, has sparked an increasing interest in data-driven approaches. It is crucial to recognise, however, that stock prices reflect not only the financial health of an organisation but are also influenced by the collective emotions and actions of investors and the general public.

This paper addresses the following question: To what extent can a social data-driven approach utilising a neural network (NN) and an elastic net (EN) predict excess stock returns? This study employs concepts from social data analytics and finance to transform, analyse, and predict excess stock returns in R. It concentrates on five key stock characteristics: short-term reversal (MOM1M), medium-term reversal (MOM12M), momentum change (CHMOM), most recent maximum return (MAXRET), and log market equity (MVEL1). Furthermore, it incorporates four macroeconomic predictors, the market ratio (BM), Treasury bill rate (TBL), Net Equity Expansion (NTIS), and Dividend Price Ratio (DP). Additionally, the data set includes 70 industry dummies and 20 interactions terms, totalling 100 baseline signals. This research extends upon the framework introduced by Gu et al. (2020), integrating their most significant stock characteristics and macroeconomic predictors.

My contribution to the literature is threefold: i) Expanding upon the research conducted by Gu et al. (2020) by including only the most relevant stock characteristics and macro predictors, thereby preventing overfitting and enhancing computational efficiency. ii) Improving the accuracy of the estimation by incorporating a more relevant data set. Specifically, I have excluded years prior to 2005, as the financial landscape has evolved significantly over the decades, and included the years 2017-2021. iii) Incorporating an additional simple EN model to compare it with a more complex NN model. My principal hypothesis is that a shallow neural network model will outperform an elastic net in predicting excess stock returns. This paper aims to achieve similar results to those obtained by Gu et al. (2020), demonstrating that NN models are able to show a small set of dominant predictive signals.

This paper utilises public data from The Center for Research in Security Prices, (CRSP), to construct an EN and a NN model, with the EN being transformed into a Lasso regression following the tuning process. These models are trained on 7630 individual stocks spanning from 2005 to 2021. They are evaluated using an out-of-sample approach, where market capitalisation weights are calculated to create a zero-net portfolio, thereby assessing the models' effectiveness. In summary, this study rejects the principal hypothesis, since both the NN and lasso regression fail to demonstrate significant predictive signals. To bolster these findings, this paper recalibrates excess returns with relevant risk factors, employing the Capital Asset Pricing Model (CAPM) and the Fama-French Three-Factor Model (FF3). However, the regression results from both models, represented as NN_{CAPM}, NN_{FF3}, EN_{CAPM}, and EN_{FF3}, also reject the principal hypotheses, showing no predictive ability.

2 Literature Review

Economists and financial analysts have extensively explored various techniques to comprehend market behaviour and identify patterns in stock prices. A comprehensive study by Rapach et al. (2013) employs a lasso regression to predict global equity market returns, utilising lagged returns from all countries. Their findings demonstrate the significant predictive power of lagged U.S. returns. This remains true even after accounting for interest rates and dividend yields.

Several papers have applied neural networks to forecast derivative prices. For example, a study by Hutchinson et al. (1994) yields promising results, although it cautions against generalising their approach beyond predicting firms listed in the S&P-500 index. Meanwhile, Yao et al. (2000) demonstrate that, particularly in volatile markets, a neural network option pricing model outperforms the traditional Black-Scholes model. Another study by Song et al. (2018) evaluates five distinct machine learning models, including a least squares support vector machine and four variations of the neural network model. Their research highlights the promising forecasting accuracy of neural network models for individual stocks, with the back-propagation neural network model emerging as the top performer.

A recent empirical study by Gu et al. (2020) investigated various machine learning (ML) techniques to analyse the behaviour of expected stock returns from 1957 to 2016, using a data set comprising nearly $30,000$ individual stocks. The results of the study highlighted the superiority of neural network models compared to other methods. Interestingly, the findings diverged from the usual paradigms seen in other fields such as bioinformatics or computer vision, as shallow learning was found to be more effective than deep learning, which might be due to the low signal-to-noise ratio in asset pricing. This paper will also predict individual stock prices using a neural network and an elastic net model. However, Gu et al. (2020) suggested that the primary benefit of ML methods lies in forecasting returns for larger and more liquid stocks, as well as portfolios. The study also identified a convergence among the various ML techniques used, as they all showed a small set of dominant predictive signals.

Conversely, a small number of researchers have demonstrated that the application of neural networks does not consistently yield promising accuracy results. Pang et al. (2020) conducted a study focusing on the Long Short-Term Memory (LSTM) model, revealing an average prediction accuracy of 53.2% for three specific stocks and 57% for forecasting the A-share composite index, which is not considered particularly promising. However, it is important to note that this study primarily examines the Shanghai stock market, characterised by distinct features compared to Western markets, potentially leading to variations in outcomes across different market contexts. Additionally, Chudziak (2023) conducted a study underscoring that their neural network models failed to outperform the benchmark buy-and-hold strategy. This finding implies that the utilisation of neural networks does not yield abnormal excess returns.

A review of the literature reveals that most studies show positive results when employing various machine learning techniques. Neural networks, in particular, have been notably successful, as demonstrated by the

works of Gu et al. (2020), Hutchinson et al. (1994), Yao et al. (2000), and Song et al. (2018). However, there are studies, such as those by Pang et al. (2020) and Chudziak (2023), which report less successful outcomes for neural networks. This difference in findings highlights the importance of variable selection in predicting stock prices. For example, Gu et al. (2020) uses 94 stock characteristics, but only five variables seem to provide the most relevant information for stock prediction. Moreover, the effectiveness of stock return prediction varies across different financial markets, as evidenced by Pang et al. (2020) and Hutchinson et al. (1994), who observe conflicting results depending on the financial context being studied.

Furthermore, the timing of the data plays a crucial role in understanding these differences in results. It is evident that stock returns exhibit distinct patterns during crises, in contrast to more stable times. According to Hung et al. (2014), certain variables significantly impact stock prices in non-crisis periods, however, during crises, most of these variables lose their explanatory power. Additionally, Hong et al. (2021) demonstrates significant fluctuations in stock return predictability and price volatility during the COVID-19 crisis, underscoring the importance of temporal dynamics in predictive accuracy.

Based on the literature and the results stated above, I present my following hypotheses regarding the neural network model and the elastic net's ability to forecast excess stock returns. These hypotheses are motivated by (Gu et al., 2020) and the previously mentioned literature:

Hypothesis 1 (H1): *A lasso regression derived by the EN framework will be able to predict excess stock returns.*

Hypothesis 2 (H2): *A neural network will be able to predict excess stock returns.*

Hypothesis 3 (H3): *A neural network will outperform a lasso regression in predicting excess stock returns.*

3 Theoretical Framework & Methodology

In the following section, this paper presents both the theoretical framework and the methodology for the EN and NN models, along with the zero-net investment portfolio implementation in R. Section 3.1 outlines the definition behind hyperparameter tuning and overfitting in relation to the EN and NN models. Section 3.2 describes the weights used in the zero-net portfolio. Lastly, Sections 3.3 and 3.4 define the CAPM and FF3 used to include additional risk factors.

3.1 Hyperparameter Tuning & Overfitting

The primary objective of hyperparameter optimisation is to enhance the performance of machine learning models by meticulously selecting the optimal set of hyperparameters that minimise the Mean Squared Prediction Error (MSPE). The tuning process involves a comprehensive assessment of the model's performance across various hyperparameter combinations, playing a crucial role in determining the optimal

hyperparameters. These hyperparameters, inherently governing the model's complexity, significantly influence the performance of the machine learning models.

It is important to recognise that hyperparameter tuning must be executed with care. If this tuning procedure occurs using the same data set on which the model's performance is evaluated, there is a potential risk of overfitting. Overfitting happens when a model becomes excessively tailored to the training data, capturing the data's noise rather than its underlying patterns. Consequently, a model that has overfitted may show diminished predictive capability when applied to new, unseen data.

To evaluate the tuning process this paper uses MSPE to measure the fit.

$$\text{MSPE} = \frac{1}{T}\frac{1}{N}\sum_{t=0}^{T}\sum_{i=1}^{N}(\text{predicted excess return}_{i,t} - \text{excess return}_{i,t})^2 \tag{1}$$

where i is a unique combination of the stock, t denotes the time period in the model, and N signifies the number of observations. The tuning parameters yielding the lowest MSPE are considered the most effective for predicting future excess returns.

3.1.1 Elastic Net

The Elastic Net (EN) is a regularisation technique that combines elements of both ridge regression and lasso regularisation. Lasso regularisation is a method used for linear regression. Similar to lasso, the EN has the capability to produce parsimonious models by driving certain coefficients to zero, (MathWorks, 2023). The EN used in this paper is equivalent to equation (8) in Gu et al. (2020), with the following value function,

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

where the two hyperparameters are λ and ρ. This is evaluated in Figure 3. λ represents the penalty parameter, and ρ is the switching parameter, where $\rho = 0$ corresponds a ridge regression and $\rho = 1$ to a lasso regression.

3.1.2 Neural Network

A neural network (NN) is a nonlinear statistical model and can be used as a two-stage regression or classification model. It consists of an input layer of raw predictors, one or more hidden layers that interact and transform the predictors nonlinearly, and an output layer that aggregates the hidden layers' outputs into a final prediction. The number of neurons in the input layer is equal to the dimension of the predictors. Furthermore, the output layer in the NN models used in this paper consists of a single neuron, although it can contain multiple neurons for classification purposes (Hastie et al., 2009). Each neuron applies a nonlinear activation function f to its aggregated signal before sending its output to the

next layer, in which this paper utilise the rectified linear unit (ReLU) activation function:

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases} \tag{3}$$

The function is defined to return x for each positive neuron and 0 for negative neurons. According to Schmidt-Hieber (2020), applying the ReLU activation function in combination with a deep NN architecture makes it possible to achieve near minimax rates for arbitrary smoothness in the regression function. In regression tasks, achieving near minimax rates means that the model is approaching the theoretical limits of how well it can perform given the available data. This paper will explore two NN configurations for regression tasks. The first model contains one layer, representing a shallow NN model, while the second NN comprises three layers, representing a deep NN model.

To obtain a NN model that performs well on the data set, it is important to tune the model. The NN model contains unknown parameters, which are commonly referred to as weights. Tuning the model involves seeking values for the unknown parameters to make the model fit the training data (Hastie et al., 2009). θ denotes the complete set of weights, consisting of:

$$\{\alpha_{0m}, \alpha_m; m = 1, 2, ..., M\} \; M(p+1) \text{ weights}$$

$$\{\beta_{0k}, \beta_k; k = 1, 2, ..., K\} \; K(p+1) \text{ weights}$$

Where M represents the number of neurons in a hidden layer, where each neuron has its own set of weights, including the bias term α_{0m} and the weights α_m associated with each input feature. K represents the number of neurons in another specific hidden layer. p indicates the number of input features, (Hastie et al., 2009). The parameters to be tuned include the number of neurons, hidden layers, and the penalisation parameter λ. The penalty helps prevent overfitting the model, while the number of layers and neurons determine the depth of the NN.

3.2 Market Capitalisation Weights & Zero-net Portfolio

A zero-net portfolio strategy has been devised to leverage the predictive potential of the two ML forecasts. In accordance with the methodology outlined by Gu et al. (2020), out-of-sample predictions for stock excess returns over a 1-month horizon are computed. These predictions are then categorised into two groups: the top percentile $P90$, representing the highest-performing forecasts, and the bottom percentile $p10$, signifying the lowest-performing forecasts. The predictions are utilised to construct a zero-net investment portfolio. However, it is essential to first calculate the value weights to determine the optimal allocation for each stock within the portfolio.

Market capitalisation is employed as the basis for value-weighting the zero-net portfolio, a choice offering a greater degree of real-time relevance compared to minimum variance weights or naive portfolio weights. In the context of each monthly iteration, denoted as t, the weight assigned to individual stocks is determined

using the following formula:

$$\text{Value weight}_{i,t} = \frac{\text{Market cap}_{i,t} \cdot [\hat{r}_{i,t+1} > \text{p90}]}{\sum_{i=1}^{k} \text{Market cap}_{i,t} \cdot [\hat{r}_{i,t+1} > \text{p90}]} + \frac{\text{Market cap}_{i,t} \cdot [\hat{r}_{i,t+1} < \text{p10}]}{\sum_{i=1}^{k} \text{Market cap}_{i,t} \cdot [\hat{r}_{i,t+1} < \text{p10}]} \tag{4}$$

where $\hat{r}_{i,t+1}$ represents the predicted excess return and k denotes the total count of firms under consideration. The designations "p90" and "p10" correspond to the top ten percentile and the bottom ten percentile, respectively.

To evaluate the predictive efficacy of the lasso regression derived from the EN framework and the NN models, a zero-net investment portfolio strategy is executed. The underlying strategy entails purchasing stocks within the highest top ten percentile while simultaneously engaging in short selling of stocks within the bottom ten percentile, given as:

$$\text{Zero-Net Investment Return} = r_{i,t+1} \cdot \text{Value weight}_{i,t} \cdot [\mathbf{1}(\hat{r}_{i,t+1} > \text{p90}) - \mathbf{1}(\hat{r}_{i,t+1} < \text{p10})] \tag{5}$$

where $r_{i,t+1}$ is the excess return.

3.3 Capital Asset Pricing Model

The CAPM captures the relationship between the expected return of an asset and its systematic risk, quantified by the β coefficient. Originally introduced by Sharpe (1964), the CAPM demonstrates that investors can mitigate unsystematic risk through the construction of a well-diversified portfolio comprising numerous assets. This model operates on foundational assumptions that investors exhibit rationality and risk aversion. It also presumes the existence of a risk-free asset, denoted by a constant interest rate (R_f), and assumes no transaction costs or tax considerations.

The CAPM can be expressed in the following mathematical form, which demonstrates the linear relationship between the anticipated returns of individual assets and the expected return on the market portfolio.

$$E[R_{i,t} - R_f] = \beta_i \cdot (E[R_{m,t}] - R_f), \tag{6}$$

where $E[R_{i,t}]$ is the expected return for asset i in period t, R_f is the risk-free rate, and $E[R_{m,t}] - R_f$ is the market risk premium. The proportionality factor β_i is given by

$$\beta = \frac{Covariance(R_e, R_m)}{Variance(R_m)} \tag{7}$$

where R_e represents the return on an individual stock, while R_m represents the return on the overall market, (Verbeek, 2017).

Subsequently, equation (6) can be reformulated as a linear regression model, which serves to forecast the

excess returns with additional risk factors analysed in Section 4.3.2.

$$R_{i,t} - R_f = \alpha_i + \beta_i(R_{m,t} - R_F) + \epsilon_{i,t}. \tag{8}$$

Equation (8) represents a regression model with an intercept term, denoted by α_i. The error term, ϵ_{it}, which is a function of unexpected excess returns and has an average value of zero, (Verbeek, 2017). The full derivation and assumptions used are presented in the appendix "Estimating CAPM Regression".

3.4 Fama-French Three-Factor Model

The FF3 model is a quantitative framework for characterising stock returns, devised in 1992 by Eugene Fama and Kenneth French, (Fama & French, 1993). In contrast to the CAPM, which employs a single variable to contrast stock returns with market returns, the FF3 model introduces a pair of additional variables, as depicted in Equation (9). This framework is utilised in Section 4.3.2 to assess whether the additional risk factors from the FF3 model have greater predictive power compared to the CAPM.

$$R_{i,t} - R_f = \alpha_i + \beta_i(R_{m,t} - R_F) + b_s \cdot \text{SMB} + b_v \cdot \text{HML} + \epsilon_{i,t}. \tag{9}$$

SMB represents the differential between small and large market capitalisation, while *HML* signifies the distinction between high and low book-to-market ratios. These two variables assess the historical excess returns of small-cap versus large-cap stocks and value stocks as opposed to growth stocks. Furthermore, the coefficients b_s and b_v are ascertained through the linear regression procedures derived in Section 4.3.2, and can contain both negative and positive values.

4 Empirical Analysis

This section presents the empirical analysis. The first section, Section 4.1, provides a detailed description of the data utilised in this study and highlights the most relevant variables. Section 4.1.1 outlines the data transformation procedures employed in the analysis, while Section 4.1.2 explains the data split structure. Section 4.2 presents the tuning results derived from the NN and EN models, and the Section 4.3 presents the out-of-sample performance. Finally, Section 4.3.2 includes additional risk factors for the EN and NN models.

4.1 Data Description

The data set used in this paper has been sourced from the Center for Research in Security Prices, (CRSP, 2023). It encompasses 7,630 individual stocks, drawing upon their price data spanning from 2005 to 2021. Consequently, this data set comprises a total of 778,311 data points for the target variable, excess returns, in conjunction with the five selected stock characteristics outlined in Figure 1. Additionally, the data set includes 70 industry dummies derived from the industry classification variable, sic2, and 20 interaction terms between the five stock characteristics and four macro variables. This facilitates the model to capture

the combined effects of these variables.

Figure 1: Summary statistics of in-sample (left-panel) and out-of-sample (right-panel)

	mom1m	mvel1	maxret	mom12m	chmom	mom1m	mvel1	maxret	mom12m	chmom
SD	0.610	0.588	0.539	0.595	0.596	0.610	0.590	0.508	0.606	0.613
Mean	-0.004	0.026	0.120	-0.012	-0.001	0.008	0.087	0.126	0.016	0.000
q1, 25%	-0.557	-0.498	-0.317	-0.536	-0.528	-0.550	-0.433	-0.286	-0.526	-0.560
q3, 75%	0.549	0.543	0.584	0.516	0.527	0.568	0.603	0.555	0.575	0.560

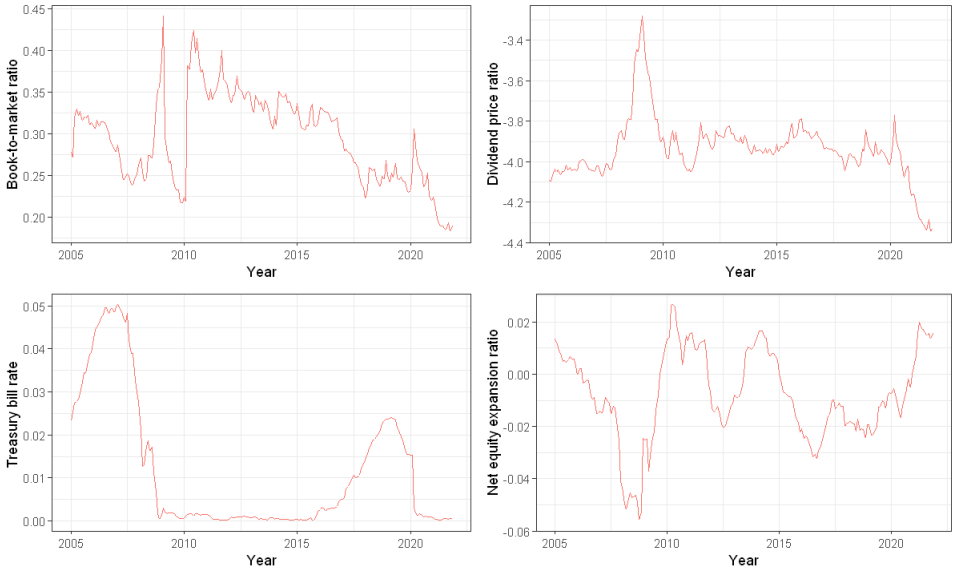
Source: CRSP (2023) and own calculations in R.
Note: The left panel shows in-sample characteristics while the right panel shows out-of-sample characteristics.

The selection of the variables in Figure 1 is based on the empirical research conducted by (Gu et al., 2020). According to their findings, one of the stock characteristics with the most significant predictive power is short-term reversal, denoted as MOM1M. It captures the immediate market reactions and can reflect rapid changes in public sentiment or investor behaviour by representing the prior month's return at time (t-1). MOM1M, displaying the highest in-sample standard deviation at 0.610, indicates higher volatility. Therefore, this variable might contain the most valuable information for predicting excess stock returns.

Following MOM1M, the medium-term reversal variable, MOM12M (representing returns from month t-12), is included. This aims to capture the overall financial health and performance of a company over an extended period, which is less influenced by short-term market sentiments. Additionally, the momentum change variable, CHMOM, is included to capture the directional trend of stock returns and exhibits the highest out-of-sample standard deviation at 0.613. This variable reflects how public sentiment and broader market trends can shift over time, influencing stock prices. The fourth variable, MAXRET, pertains to the stock's most recent maximum return and displays the lowest in-sample standard deviation at 0.539, alongside an out-of-sample standard deviation of 0.508. Lastly, the fifth variable, log market equity, denoted as MVEL1, represents market capitalisation. Notably, the mean value of MVEL1 increases from the in-sample to the out-of-sample period. This observation suggests that tuning the model in-sample might lead to a skewed fit for the out-of-sample period, a consideration that is also applicable to the MOM12M characteristic.

In addition, this paper includes macroeconomic predictors visualised in Figure 2, based on the study by (Gu et al., 2020). These predictors encompass the Book-to-Market ratio (BM) for the Dow Jones Industrial Average, reflecting the relationship between a company's book value and its market value. Additionally, the Treasury bill rate (TBL) and the Net Equity Expansion (NTIS) variable, which measure the ratio of 12-month moving sums of net issues by NYSE-listed stocks relative to the total market capitalisation of NYSE stocks (Welch & Goyal, 2008). Lastly, the Dividend Price Ratio (DP) serves as an indicator of stock attractiveness.

Figure 2: Visual presentation of the four macroeconomic predictors



Source: CRSP (2023) and own calculations in R.

In Figure 2, all variables experienced significant fluctuations during 2008-2010 due to the global financial crisis. For example, the BM ratio surged in 2008-2009, reflecting the impact of the crisis as stock prices dropped, which potentially caused the BM ratio to rise due to stable or increasing book values. Post-crisis, government interventions such as low interest rates and quantitative easing led to market value recovery, potentially lowering the BM ratio as market values caught up with or exceeded book values, thereby reshaping financial metrics.

In 2020, variables BM, DP, and TBL exhibited negative trends due to the COVID-19 pandemic. In contrast, the NTIS variable showed a positive trend. This positive trend in NTIS can be attributed to two main factors. First, governments worldwide implemented stimulus programs in response to the economic challenges posed by the pandemic, potentially contributing to the positive trend in NTIS. Second, some companies responded to the crisis by raising capital through methods such as issuing new shares or taking on debt. This capital infusion led to an expansion in net equity, further boosting the positive trend in NTIS.

Lastly, this paper performs an Augmented Dicky-Fuller (ADF) test developed by Dickey and Fuller (1979), to test whether the variable excess return is stationary. After normalising the dependent variable, the

Table 1: Augmented Dickey-Fuller Test

Data	Dickey-Fuller	Lag order	P-value
ret_excess	-884.7	0	0.01

Source: (Dickey & Fuller, 1979), and CRSP (2023) data used in R.

ADF test in Table 1 yields a Dickey-Fuller statistic of -884.7 with a lag order of 0 and a corresponding p-value of 0.01. The null hypothesis of the ADF test posits that the data are non-stationary, while the

alternative hypothesis contends that the data are stationary. In this case, the low p-value of 0.01 provides strong evidence against the null hypothesis, suggesting that the financial data are indeed stationary.

4.1.1 Data Transformation

To prepare the data set for subsequent analysis, dates before the year 2005 have been excluded. This action is motivated by the objective of obtaining a more contemporaneous data set, under the assumption that data points predating 2005 may not exhibit the same patterns and characteristics as those in the current period. The inclusion of such historical data could potentially introduce bias and distort the analytical outcomes.

Secondly, all NaN values are dropped since missing values in financial time series data may not be random. This can occur due to specific market conditions or events. Leaving these gaps unaddressed can introduce bias into your analysis, as the missing data points may carry different information than the available ones. Removing NaN values helps mitigate this potential bias.

Thirdly, a recipe is created using the excess return as the response variable. In the recipe, the variables month, stock ID (permno), and market capitalisation (mktcap_lag) are removed. Afterwards, the data are normalised by subtracting the mean and dividing by the standard deviation for each numeric predictor. The variable sic2 is not normalised since it is required for generating dummy variables. Normalising the data is important since the data set is a time series. Therefore, normalisation is essential to identify trends, patterns, and anomalies. By scaling data over time, the stability and convergence of the NN and EN models are improved, thereby enabling a more robust forecasting.

4.1.2 Data Split

The process of dividing the data into training, validation, and test sets is a fundamental step in machine learning predictive analysis. In this particular setup, the in-sample comprises 80% of the full data set, while the out-of-sample consists of the remaining 20 %:

- **The training set:** This set contains 80% of the in-sample data and is applied to train the models.
- **The validation set:** This set contains 20% of the in-sample data, which is used to tune the hyperparameters of the models.
- **The testing set:** This set contains the full out-of-sample data and is utilised to assess the performance of the models.

The use of distinct data sets for training and testing serves a crucial purpose in assessing the models' ability to generalise to unseen data. The test set acts as an independent evaluation to confirm that the model is not merely memorising the training data, which is known as overfitting. Additionally, the validation set plays a significant role by allowing for the iterative adjustment of the hyperparameters and the selection of the optimal model configuration, all without affecting the integrity of the test set. This separation ensures a reliable evaluation of model performance and facilitates effective model tuning.

4.2 Training and Validation Performance

4.2.1 Tuning Results of Neural Network

The results of the tuning process, presented in Table 2, are based on the training and validation data sets. Both the shallow and deep NN models are evaluated with 10 and 20 neurons. Following the methodology outlined in Gu et al. (2020), this study employs 100 epochs for comparison. However, a rule of thumb suggests using approximately three times the number of predictor variables as the number of epochs, which, in this case, would amount to 300 epochs. Consequently, an analysis was also conducted with 300 epochs, but this did not yield a significant alteration in the MSPE of the NN models. This indicates that the results remain robust to changes in the number of epochs. Notably, the NN model with a single hidden layer and a regularisation parameter (λ) of 0.001 exhibits the lowest MSPE, amounting to 0.028.

Table 2: MSPE for Single (Shallow) and Three Layer (Deep) Neural Network

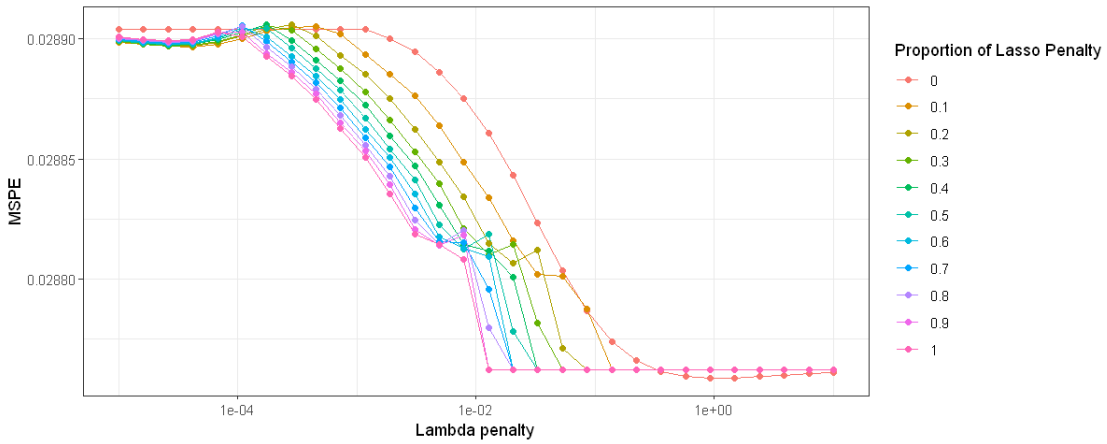
λ	Neurons	Single layer (MSPE)	Three layers (MSPE)
1e-05	10	0.030	0.035
1e-03	10	0.029	0.035
1e-05	20	0.0305	0.057
1e-03	20	0.030	0.031

Source: CRSP (2023) and own calculations in R.

4.2.2 Tuning Results of Elastic Net

For the EN, the following parameters are tuned $\lambda \in [1e^{-8}, 1]$ and $\rho \in [0, 1]$ on the validation data set only. Figure 3 illustrates the relationship between the MSPE and the lambda penalty in the EN model.

Figure 3: MSPE for Various Lambda and Lasso Penalties in the Elastic Net



Source: CRSP (2023) and own calculations in R.

As the lambda penalty increases, there is a noticeable decrease in MSPE. The lowest MSPE, at 0.02876, is achieved when the lambda penalty is approximately equal to one and the lasso parameter, denoted as ρ, is set to zero. This outcome implies that the lasso regression outperforms all other combinations within the EN framework as the penalty strength increases. Therefore, based on the findings presented

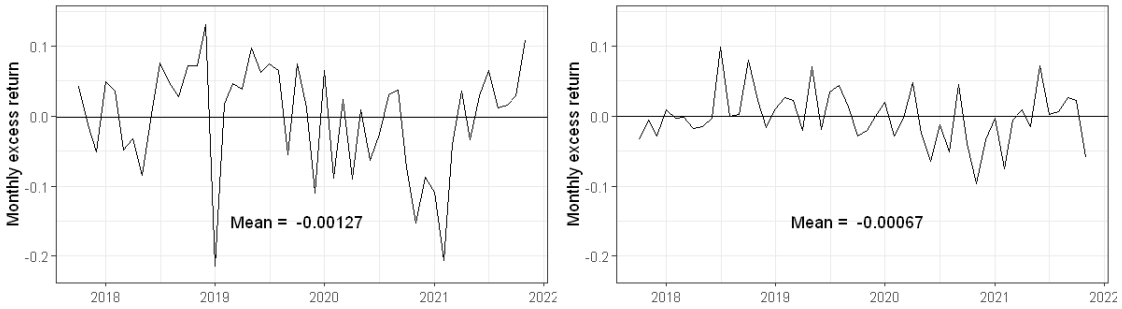
in Figure 3, this paper proceeds with the parameter settings of $\lambda = 1$ and $\rho = 0$ for the Lasso regression, as it appears to be the optimal choice.

4.3 Out-of-sample Performance and Portfolio Management

4.3.1 Lasso Regression versus Neural Network Performance

The predicted excess returns for each month is depicted in Figure 4. The computed mean excess return

Figure 4: Neural Network (left panel) and Lasso Regression (right panel) Predictions of Excess Returns



Source: CRSP (2023) and own calculations in R.

for the NN model is -0.00127, while the Lasso Regression model yields a mean excess return of -0.00067. It is noteworthy that neither of these mean excess returns exhibits statistically significant deviations from zero. Upon examining the observed outcomes, it becomes apparent that both models exhibit considerable volatility, with the NN demonstrating the highest volatility among them. This implies that neither model is able to accurately forecast excess returns.

Both models project an approximate zero excess return and they encounter notable challenges in predicting excess returns during periods characterised by macroeconomic instability. This challenge becomes particularly evident when assessing the models' performance during the period spanning from 2020 to mid-2021. It is worth considering that the limited predictive power of the models during this period may be attributed to the inherent limitations of the predictors. Factors such as stock characteristics and macroeconomic variables may struggle to anticipate significant exogenous macroeconomic shocks, such as the unprecedented COVID-19 pandemic.

4.3.2 Including Additional Risk Factors, CAPM & FF3

The findings appear to support the no-arbitrage theory, given that the excess returns exhibit fluctuations around a mean of zero. Nonetheless, before drawing any definitive conclusions, it is relevant to reevaluate these excess returns in light of the additional risk factors.

The computation of the average excess return and its corresponding standard errors for both the CAPM and the FF3 model is undertaken. This analytical approach facilitates an empirical examination of whether the zero-net portfolio strategy yields average excess returns that are statistically significant from

zero while accounting for relevant risk factors. Specifically, the CAPM incorporates the excess return of the market portfolio (MKT), while the FF3 model includes two additional risk factors, namely small minus big (SMB) and high minus low (HML). Subsequently, an FF3 regression analysis is performed for both the NN and EN model.

$$r_{\text{NN},t} = \alpha_{\text{NN}} + \beta_{\text{NN},1} \cdot MKT_t + \beta_{\text{NN},2} SMB_t + \beta_{\text{NN},3} HML_t + \epsilon_{\text{NN}} \tag{10}$$

and the same for the EN:

$$r_{\text{EN},t} = \alpha_{\text{EN}} + \beta_{\text{EN},1} \cdot MKT_t + \beta_{\text{EN},2} SMB_t + \beta_{\text{EN},3} HML_t + \epsilon_{\text{EN}} \tag{11}$$

The CAPM regression for both the NN and EN is also conducted. The results is depicted in Table 3.

Table 3: CAPM & FF3 Regression Results

	Dependent variable:			
	NN CAPM	NN FF3	EN CAPM	EN FF3
mkt_excess	-0.621^{**}	-0.164	-0.118	-0.078
	(0.356)	(0.205)	(0.167)	(0.170)
smb		-1.782^{***}		-0.548^{***}
		(0.135)		(0.123)
hml		-0.489		-0.086
		(0.313)		(0.168)
Constant	0.004	-0.003	0.003	-0.0005
	(0.009)	(0.015)	(0.013)	(0.004)

Source: CRSP (2023) and own calculations in R.
Note: *p<0.1; **p<0.05; ***p<0.01

Statistics from Newey and West (1987) were utilised to evaluate the shallow NN and EN models. In the NN CAPM model, the α_{NN} value registers at 0.004 and lacks statistical significance, as its p-value exceeds 0.10. This suggests there is no strong evidence to support abnormal excess returns after accounting for market-related risks. Notably, the $\beta_{NN,1}$ value is -0.621 and is statistically significant with a p-value less than 0.05. This indicates a negative relationship with the market portfolio and implies decreased volatility.

In contrast, the EN CAPM model shows an α_{EN} value of 0.003, which is not statistically significant given its p-value exceeds 0.10. This result aligns with the interpretation that it cannot be confidently stated that the average excess returns deviate from zero.

Regarding the NN FF3 outcomes, a statistically significant negative association is observed between the SMB factor and excess returns, evident at a 1% confidence level. The coefficient of -1.782 suggests that larger firms play a role in explaining variations in excess returns. However, in the FF3 approach, the excess return for the MKT factor no longer holds statistical significance in comparison to the CAPM.

In summary, the inclusion of additional risk factors does not lead to an improvement in portfolio performance for the machine learning models EN and NN. This observation is supported by the statistically insignificant values, as detailed in Table 3.

5 Discussion

5.1 Ethical Implication

Given this paper's focus on publicly available stock data, concerns typical of Big Social Data Analytics, such as legal issues related to personal and private data, are not pertinent in this context. However, ethical concerns arises when individuals are not fully aware of or do not understand the risks involved with stock trading, leading to potential financial harm. Both stock trading and gambling involve a degree of risk and uncertainty. In the stock market, even well-researched investments can result in losses due to unpredictable market dynamics. Similarly, gambling outcomes, such as lotto, are largely based on chance. Therefore, the ethical implications of stock trading and gambling raise questions about the responsibility of regulators to protect participants and maintain market integrity.

In conclusion, while stock trading is a fundamental component of the financial system and differs from gambling in its intention (investment vs. entertainment), the associated risks, potential for addiction, and impact on individual financial stability introduce ethical considerations.

5.2 Data Suitability

The data set from CRSP (2023) employed in this study comprises 7,630 individual stocks spanning a 17-year time period from 2005 to 2021. This extensive data set facilitates a robust and comprehensive analysis, thereby enhancing the credibility of the results obtained through the NN and EN. However, when compared to the research conducted by Gu et al. (2020), which covers the period from 1957 to 2016, it is essential to acknowledge that my data set may not capture long-term trends or rare events predating 2005, such as the late 1990s dot-com bubble.

Nonetheless, the decision to initiate the data set in 2005 aligns with significant changes in the financial data landscape over the decades. Additionally, the data set has been extended to include the years 2017 through 2021, encompassing the COVID-19 pandemic in 2020. However, this extension could introduce noise and outliers, as observed in the study by (Hong et al., 2021). To counter this, the data set has been normalised, scaling it to a common range. This process facilitates a clearer understanding of rel-

ative changes in the data points and mitigates the influence of outliers and extreme events such as the COVID-19 pandemic.

Lastly, this study follows standard protocol by excluding NaN values, a necessary step for accurate machine learning computations, despite the risk of information loss and potential bias.

5.3 Out-of-Sample Results and Link to Existing Literature

The out-of-sample results of the zero-net investment portfolio reject the principal hypothesis, indicating that the NN model does not outperform the Lasso regression, which is transformed by the EN. Both models exhibit no significant signs of superior predictive power, leading to the rejection of hypotheses H1, H2, and H3. These findings contradict several existing studies, such as those by Gu et al. (2020), Hutchinson et al. (1994), and Song et al. (2018), all of which suggest that neural networks demonstrate promising predictive capabilities. However, they align with the research conducted by Pang et al. (2020) and Chudziak (2023), which found that neural networks do not exhibit strong predictive power and cannot outperform a simple buy-and-hold strategy.

It is worth noting that these results may diverge from the existing literature because the excess returns are not adjusted for specific risk factors. To enhance the robustness of the findings, this paper applies the CAPM and the FF3 regression to the excess returns predicted by both the Lasso regression and the neural network. However, the regression results reveal that incorporating additional risk factors does not improve the performance of either the EN or the NN machine learning models. This reinforces the rejection of my hypotheses.

Finally, if it were possible to generate positive excess returns using EN and NN models with only five stock characteristics and four macro predictors, it would be reasonable to assume that other investors would have adopted this zero-net investment portfolio strategy. This would lead to a convergence of excess returns towards zero, resulting in a no-arbitrage situation. In essence, these findings align with the efficient market hypothesis proposed by Samuelson (1965) and Fama (1965), which posits that stocks always trade at their fair value.

6 Conclusion

This paper assesses the feasibility of employing an EN and a NN to predict excess stock returns in R. Similar to the findings in the study by Gu et al. (2020), this paper finds that the shallow NN outperforms the deep NN. However, even with the incorporation of additional risk factors, neither model surpasses the market's performance, thereby rejecting the hypotheses (H1) and (H2). Specifically, the NN model does not outperform the EN model when combined with the CAPM and FF3 Model, thereby rejecting the hypothesis (H3). These results challenge existing literature and the principal hypothesis, which states that the shallow NN model would outperform a simple EN approach. In essence, this paper suggests that attempts to enhance relevance by narrowing the data period and simplifying the models through

the reduction of reducing the number of predictor variables do not yield similar or improved prediction accuracy compared to the study by (Gu et al., 2020).

Appendix

Estimating CAPM Regression

To derive a linear regression model using equation (6), this paper will apply the premise of rational expectations, whereby the anticipations held by economic agents align with mathematical expectations. Consequently, it becomes feasible to establish a linkage from equation (6) to actual returns. Additionally, the Ordinary Least Squares (OLS) estimator is relied upon due to its consistency properties, thereby ensuring that the error term maintains an absence of correlation with the regressor.

To begin with, define the asset i's unexpected return as

$$u_{it} = R_{it} - E\{R_{it}\}$$

Next, define the market portfolio as,

$$u_{mt} = R_{mt} - E\{R_{mt}\}.$$

Now reformulating equation (6) as

$$R_{it} - R_f = \beta_i(R_{mt} - R_F) + \epsilon_{it} \tag{12}$$

where

$$\epsilon_{it} = u_{it} - \beta_i u_{mt}.$$

A regression model without an intercept is represented by equation (12). The error term ϵ_{it} has a mean of zero and is a function of unexpected returns.

$$E\{\epsilon_{it}\} = E\{u_{it}\} - \beta_i E\{u_{it}\} = 0$$

Above equation for the expected error term, explains that the OLS estimator is consistent, (Verbeek, 2017). Moreover, the definition of β_i, which may be expressed in the form of,

$$\beta_i = \frac{E\{u_{it}u_{mt}\}}{Variance\{u_{mt}\}}$$

which indicates that the error term is uncorrelated with the regressor $R_{it} - R_f$. This can be illustrated using by using the fact that R_f is not stochastic, which provides the following result:

$$E\{\epsilon_{it}(R_{mt} - R_f)\} = E\{(u_{it} - \beta_i u_{mt})u_{mt}\} = Eu_{it}u_{mt} - \beta_i Eu_{mt}^2 = 0. \tag{13}$$

References

Chudziak, A. (2023). Predictability of stock returns using neural networks: Elusive in the long term. *Expert systems with applications*, *213*, 119203.

CRSP. (2023). Mfe database. https://www.crsp.org

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427–431.

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, *38*(1), 34–105.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. https://doi.org/https://doi.org/10.1016/0304-405X(93)90023-5

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Hong, H., Bian, Z., & Lee, C.-C. (2021). Covid-19 and instability of stock market performance: Evidence from the us. *Financial Innovation*, *7*(1), 1–18.

Hung, C.-H. D., Azad, A. S., & Fang, V. (2014). Determinants of stock returns: Factors or systematic co-moments? crisis versus non-crisis periods. *Journal of International Financial Markets, Institutions and Money*, *31*, 14–29.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The journal of Finance*, *49*(3), 851–889.

MathWorks, T. (2023). Lasso and elastic net. https://www.mathworks.com/help/stats/lasso-and-elastic-net.html

Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777–787.

Pang, X., Zhou, Y., Wang, P., Lin, W., & Chang, V. (2020). An innovative neural network approach for stock market prediction. *The Journal of Supercomputing*, *76*, 2098–2118.

Rapach, D. E., Strauss, J. K., & Zhou, G. (2013). International stock return predictability: What is the role of the united states? *The Journal of Finance*, *68*(4), 1633–1662.

Samuelson, P. A. (1965). Rational theory of warrant pricing. In *Henry p. mckean jr. selecta* (pp. 195–232). Springer.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425–442.

Song, Y.-G., Zhou, Y.-L., & Han, R.-J. (2018). Neural networks for stock price prediction. *arXiv preprint arXiv:1805.11317*.

Verbeek, M. (2017). *A guide to modern econometrics*. John Wiley & Sons.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508. Retrieved September 13, 2023, from http://www.jstor.org/stable/40056859

Yao, J., Li, Y., & Tan, C. L. (2000). Option price forecasting using neural networks. *Omega*, *28*(4), 455–466. https://doi.org/https://doi.org/10.1016/S0305-0483(99)00066-3