Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrie en Operationele Research

Predicting stock market returns using high-dimensional data

Dennis Huang (535119)

Supervisor:	Robin Lumsdaine
Second assessor:	Kole, HJWG
Date final version:	1st July 2024

Abstract

Making accurate predictions of stock market returns is crucial for both risk management and financial gains. Therefore, this paper evaluates the performance of the Partially Protected Bayesian LASSO compared to other techniques and models capable of handling high-dimensional data. One-quarter-ahead predictions are made using an expanding window to assess the performance of these models. Although the differences between the two Bayesian LASSO models were small, the results show that the regular Bayesian LASSO performed the best, demonstrating significant predictive ability over the Partially Protected Bayesian LASSO. The Partially Protected Bayesian LASSO still performed on par with Support Vector Regression, Random Forest, and PCA-OLS models. These promising results highlight the potential of the Partially Protected Bayesian LASSO for future research. Further investigation could focus on refining the selection of protected variables by making them less prone to shrinkage or improving the prior distributions of the model.

1 Introduction

Accurately predicting stock market returns is crucial in various fields, including portfolio building, option pricing, and risk management. Sound decision-making in these areas can lead to substantial profits for both individual and institutional investors. However, achieving accurate predictions remains a challenging endeavor. This is especially the case because the stock market is influenced by numerous macro economic variables such as central bank interest rates, investors expectation, political events and other economic condition indicators(Sigo, 2018). As a result, Big Data which is high-dimensional data from a variety of sources, can offer a solution by incorporating all this information and patterns when making predictions. Previously, Big Data has been successfully applied in the finance and banking industry leading to various improvements regarding customized solutions, fraud detection and risk management(Dulhare et al., 2020). However, these high-dimensional datasets come with various challenges like overfitting which negatively impacts the performance of the model with new data, as well as noise and multicollinearity(Fan et al., 2014). Therefore, models that can successfully extract the most information from these datasets and adequately address the aforementioned challenges, could significantly improve the forecasts of stock market returns.

The Partially Protected LASSO(Yaman et al., 2024) is a model that can address these challenges, by selecting the most important and relevant predictors, thereby reducing the sensitivity to noise in the data. Furthermore, it is also easier to interpret by shrinking the total number of predictors in high-dimensional datasets. This model was initially introduced because Yaman et al. (2024) wanted to bridge theory and forecast accuracy in the political science by 'protecting' theoretical significant variables in the political science when making predictions. The protection is achieved by making the theoretically significant variables less prone to shrinkage compared to others, leading to bigger coefficients for these variables. For forecasting stock market returns, this model serves as a interesting and novel technique, by 'protecting' variables that have demonstrated strong out-of-sample performance. Exploring new techniques for forecasting stock returns is particularly important, as stock returns include a significant portion of unpredictability where even the best forecasting models are only able to explain a small part of stock returns (D. Rapach & Zhou, 2013). Therefore, finding potentially good models and methods for

predicting stock market returns will be a great addition to the existing literature.

In the first part of this thesis, the method of Yaman et al. (2024) is replicated using the American National Election Study (ANES) dataset which is a public opinion survey about voting behavior in the US. In this survey participants were asked to rate the feelings toward political leaders from 0 to 100, referred to as the feeling thermometer. The thermometer value of Joe Biden will serve as the target variable. The aim of Yaman et al. (2024) was to asses the trade-off involved in protecting theoretically important variables from shrinkage. The replication results show that the Partially Protected Bayesian LASSO successfully makes the protected variables less prone to shrinkage compared to a Bayesian LASSO that shrinks variables indiscriminately while only slightly decreasing in predictive performance. These results are consistent with the findings of the original paper. This thesis then extends the study by the following research question:

Research Question

Does the Partially protected LASSO improve forecasting performance of existing stock market return models using high-dimensional data?

To answer this, a dataset consisting of 98 variables, including financial indicators, macroeconomic data and technical indicators is used to predict the S&P 500 stock market returns. The Bayesian LASSO is then compared to other models that have demonstrated strong performance in predicting stock market returns with high-dimensional data. The empirical results show that the regular Bayesian LASSO performed the best, showing significant predictive ability over the Partially Protected LASSO. However, the Partially Protected Bayesian LASSO has equal predictive ability with the two machine learning models (Random Forest and Support Vector Regression) and PCA-OLS. Additionally, all these models demonstrated a significant ability to predict the sign and generate excess returns of the S&P 500. However, the Bayesian LASSO models have a major advantage over other models and techniques capable of handling high-dimensional datasets, as it retains the interpretability of the model coefficients and comprehends the significance of variables in making accurate predictions.

The remainder of this paper is organized as follows. First, in Section 2 the literature is introduced. Then, in Sections 3 and 4, the data and the pre-processing of the replication and extension parts will be described. Section 5 will be dedicated to the methodology and performance measures. The results are reported in Section 6. Lastly, a conclusion will be given in Section 9.

2 Literature

Over the years, many researchers have tried to make accurate predictions of stock market returns, especially much emphasis has been given to identify the best predictors. However, most bivariate OLS models fail to beat the historical average forecast(Welch & Goyal, 2008). This problem remains even when all predictors are included in a multiple linear regression model for forecasting. This is likely due to overfitting and having too many parameters, which often leads to poor out-of-sample predictions. Fortunately, there has been an increasing number of liter-

ature on models and techniques that address these problems and can handle high dimensional data in predicting stock market returns.

The LASSO regularization method from Tibshirani (1996) is one of such methods. Chinco et al. (2019) uses the LASSO to make one-minute-ahead forecasts of stock returns with high-dimensional data. The LASSO enhances both the in-sample fit and Sharpe-ratios. The strength of the LASSO lies in its ability to create a parsimonious and selecting the most important variables from high-dimensional data, contributing to better out-of-sample predictions. This advantage is further emphasized by other papers that forecast stock market returns using LASSO(Feng et al., 2020)(Dai et al., 2022). The Bayesian LASSO however, has not been widely applied in predicting stock market returns despite the advantages over the regular LASSO model. This model provides better model interpretation and it incorporates prior beliefs, which can be advantageous in the presence of multicollinearity(Park & Casella, 2008).

In addition to overfitting and overparametrization, is the presence of non-linearities in stock market returns(Enke & Thawornwong, 2005). Models that can adequately incorporate these non-linearities could significantly improve forecasting performance. Vijh et al. (2020) have shown that machine learning models excel at capturing hidden patterns and non-linear relationships in high-dimensional data. Among these, the Support Vector Machine (SVM)(Vapnik, 2013) is commonly used. Huang et al. (2005) reviews several methods to predict the Nikkei 225 Index stock market index, finding that the Support Vector Machine (SVM) model beats the Linear Discriminant Analysis, Quadratic Discriminant Analysis and Neural Networks models. The good performance of the SVM model is due its ability to prevent overfitting and the SVM solution is always unique and optimal, in contrast with other machine learning models like neural networks which involves solving more complicated mathematical problems that can lead to suboptimal solutions.

Random Forest(RF)(Breiman, 2001) is another machine learning model frequently used. Given the noisy and fluctuation prone nature of stock market returns, the RF model, which uses ensemble methods to smooth out noise, serves as a suitable model for forecasting stock market returns. Akyildirim et al. (2022) compares various methods, including the Random Forest, on high-frequency data of 27 blue-chip stocks traded on the Istanbul Stock Exchange to forecast the sign and change of percentage, finding that the Random Forest performed the best. It outperformed artificial neural networks, k-nearest neighbors, logistic regression, Naïve Bayes and extreme gradient boosting classifier. This success is large due to the ability of the Random Forests to capture complex non-linear relationships between predictors, which might be missed by linear models such as logistic regression. Furthermore, most machine learning models are highly sensitive to hyperparameter changes but Probst et al. (2019) show that the effect of hyperparameter tuning is smaller for the RF model.

Moreover, D. Rapach & Zhou (2013) highlight the use of dimension reductions techniques, which summarizes the information contained in a large number of individual predictors. Principal Component Analysis (PCA) has proven to produce consistent estimators (Bai, 2003) (Stock & Watson, 2006). PCA not only extracts the most important information from the predictors but also prevents overfitting and multicollinearity, as the predictors produced by PCA are uncorrelated. These properties make PCA one of the best methods for dealing with high-dimensional

datasets. Ludvigson & Ng (2007) successfully applied PCA on a large dataset of 209 macroeconomic and 172 financial indicators which not only led to good in-sample performance but also significant out-of-sample forecasting of stock market returns. Similarly, Neely et al. (2014) applied PCA to a combination of 14 macroeconomic and 14 technical indicators, finding that the predictors constructed from this method performed the best when forecasting the US equity premium. These previous studies have mainly focused on applying PCA in combination with a linear regression which leaves room to study the effectiveness on non-linear models like machine learning models.

3 Data replication

The dataset from the American National Election Study (ANES) 2020 is used for the replication, just as in Yaman et al. (2024). In this survey participants were asked to rate the feelings toward political leader from 0 to 100 referred to as the feeling thermometer. The thermometer value of Joe Biden is the target variable for the replication part.

While traditional research often relies on theoretically agreed-upon variables like race and income to make predictions, recent literature focuses more on forecasting vote results without necessarily considering the theoretical relevance of these variables. The goal of Yaman et al. (2024) was to bridge these two approaches. In this paper, the variables identified by Argyle et al. (2023) are chosen to be protected. It contains variables related to ethnicity, gender, age, ideology, income and being evangelical or not.

The first step in preprocessing the ANES dataset is that only the pre-election variables will be used, thereby excluding the post-election variables. Then the missing observations and invalid responses of the survey takers are converted to NA values. Variables with more than 60% missing observations and those that have a variance of less than one are removed. Additionally, observations from the dataset are entirely removed from the dataset if they contain missing values for the target variable, Joe Biden's feeling thermometer. To address the remaining NA values, the Multivariate Imputation by Chained Equations (MICE) procedure in R is applied. This approach creates multiple imputed values for each missing entry, considering other variables and incorporating the uncertainty associated with the missing data. After preprocessing, the data consists of 449 columns and 8060 observations. Of this dataset, 70% of the data will be used for training the models and the remaining 30% to test.

4 Data extension

The dataset for the extension consists of 98 variables, which are macroeconomic variables related to the USA; technical indicator variables constructed based on leverage, volume, and volatility clustering (Liu & Pan, 2020); and several financial indicators. The target variable is the quarterly log returns of the S&P 500 price index, calculated using the formula $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$. Since, not all variables were available on a quarterly bases, the daily or monthly variables were aggregated by summation or averaging over the quarter. All of the variables span from 1990Q1 to 2020Q3. A full list of all the variables and their sources can be found in Table 5 in Appendix B.

As mentioned in the literature section, most variables fail to beat the historical average. Fortunately, there have been several variables that have shown good out-of-sample performance. Those are the purchasing managers index, yield curve rates and interest rate(McMillan, 2021), the T-bill rate(Qiu et al., 2016) and the S&P GSCI Commodity total return index(Black et al., 2014). The data is preprocessed by testing for stationarity, linear trend stationarity, and quadratic trend stationarity using an augmented Dickey-Fuller (ADF). Variables containing a trend, are detrended and when the variable has a unit root, the first difference was taken. No observations were imputed, as only variables spanning from 1990Q1 to 2020Q3 were selected.

5 Methodology

In this section, the proposed methods will be introduced. It starts with a discussion of the original LASSO Tibshirani (1996). Following this, the Bayesian LASSO and the use of priors in this paper will be discussed. After that, the benchmark models will be explained. Lastly, the performance metrics used in this paper are described.

5.1 Targeted predictors

Maintaining the best characteristics of the high-dimensional dataset while introducing a sparse structure by performing variable selection beforehand is suggested by Bai & Ng (2008). This is important because not all predictors are important for predicting stock market returns and could even introduce noise, thereby worsening the forecast. Therefore, the Elastic Net developed by Zou & Hastie (2005) is chosen as the variable selection method because of its good performance when predictors are highly correlated and will be used on the variables that are not chosen to be protected. Elastic Net incorporates both the LASSO(L_1) and Ridge(L_2) regression penalty terms, formulated in Equation 1. α is the weight between the L_1 and L_2 penalty terms. The Ridge penalty decreases the magnitude of the coefficient and the LASSO penalty can set them to zero, thereby providing variable selection. Following Bai & Ng (2008) $\alpha = 0.5$ and tune λ such that ten predictors are chosen. Thus, the final dataset will comprise of these ten predictors and the protected variables.

$$\hat{\beta}^{EN} = \underset{\beta}{\operatorname{argmin}} \left[\text{RSS} + \lambda \left((1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1 \right) \right], \tag{1}$$

5.2 LASSO

Equation 2 shows the formulation of the LASSO coefficients β. The first part is the residual sum of squares (RSS) which is the same as an ordinary least squares estimator. The second part is the L_1 norm, which shrinks the coefficients towards zero. The higher the value of λ the more sparsity in the model you get. Furthermore, a $\lambda = 0$ gives you the OLS estimator. There are methods for regularization models such as LASSO and Elastic Net to get an estimate of the coefficients standard errors, which are not shrunk to zero. However, for variables with coefficients set to zero, there has yet to be a method to get reliable estimates. This is exactly what the Bayesian LASSO can solve which in return gives a better insight into the importance of the variables.

$$\hat{\beta}_{LASSO} = \sum_{i=1}^{n} \left(y_i - \sum_j x_{ij} \hat{\beta}_j \right)^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \tag{2}$$

5.2.1 Bayesian LASSO

The Bayesian models make use of prior distributions on the coefficients, treating them as random variables, rather than fixed values. Furthermore, many Bayesian models rely on methods like a Gibbs sampler, a type of Monte Carlo Markov Chain (MCMC), to derive the posterior distribution: $P(\theta \mid x, y) = \frac{P(x,y|\theta) \cdot P(\theta)}{P(x,y)} \propto P(x, y \mid \theta) \cdot P(\theta)$. This method numerically approximates the posterior distribution by creating a sequence of samples using a Markov chain when direct sampling is not feasible. The Bayesian LASSO introduced by Park & Casella (2008) also applies these principles. In this model a prior distribution is placed on the coefficients β more specifically, the Laplacian distribution is assigned to the coefficients, which acts like the L_1 penalty term of the LASSO to introduce shrinkage. Yaman et al. (2024) showed an example of a Bayesian LASSO with a conditional Laplace prior specification, which can be seen in Equations 3.

$$\begin{aligned}
&\beta_j \mid \tau_j^2, \sigma^2 \sim N\left(0, \sigma^2 \tau_j^2\right) \\
&\pi\left(\sigma^2\right) = 1/\sigma^2 \\
&\tau_j^2 \,\middle|\, \lambda^2 \sim \text{Exponential}\left(\frac{\lambda^2}{2}\right), \text{ for } j = 1, \ldots, p, \\
&\lambda^2 \sim \Gamma(1, 0.1).
\end{aligned} \tag{3}$$

The conditional Laplace prior, combined with an uninformative prior on σ^2, ensures that the posterior is unimodal. If this were not the case, it would slow down the convergence of the Gibbs sampler. This property ensures that the posterior coefficients of the β are interpretable by means of the standard errors. Additionally, the inclusion of the intermediate parameter τ simplifies the sampling from the posterior distribution, thereby also improving the convergence of the Gibbs sampler. Moreover, integrating τ_j^2 out, yields the desired Laplacian prior:

$$\beta_j \mid \lambda, \sigma \sim \mathcal{L}\left(0, \frac{\sigma}{\lambda}\right), \text{ for } j = 1, \ldots, p$$

To get the Partially-Protected Bayesian LASSO, the specifications below are used. $\boldsymbol{X}$ denotes a matrix of predictors that are standardized. β is a vector with coefficients of the predictors and σ^2 are the residual variances. λ^2 performs as the L_1 penalty term with τ^2 as the intermediate parameter, n is the number of observations in the dataset and p is the number of predictors.

$$\begin{aligned}
&Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
&\epsilon_i \sim N\left(0, \sigma^2\right), i = 1, \ldots, n \\
&\beta_j \sim N\left(0, \tau_j^2 \sigma^2\right), j = 1, \ldots, p.
\end{aligned}$$

Equations 4, 5, 6 and 7 describe the distributions of the hyperparameters. σ^2 is assigned a uniform distribution and λ^2 a gamma distribution. The τ parameters are split into two categories to differentiate between protected variables $\tau^2_{\text{protected}}$ and non-protected variables $\tau^2_{\text{non-protected}}$. The distribution of $\tau^2_{\text{protected}}$ is chosen to have a distribution that is less concentrated around

zero, which should lead to bigger $\beta's$ and thus giving more importance to the protected variables. Note that when every variable is in the protected group, the Bayesian LASSO turns into a regular Bayesian linear regression. To assess the trade-off between theoretical relevance and prediction accuracy, this Bayesian linear regression and the Bayesian LASSO without protection will be assessed.

$$\tau^2_{\text{non-protected}} \sim \exp\left(\frac{\lambda^2}{2}\right) \tag{4}$$

$$\tau^2_{\text{protected}} \sim \Gamma(1,1) \tag{5}$$

$$\sigma^2 \sim \mathcal{U}(0.1,10) \tag{6}$$

$$\lambda^2 \sim \Gamma(1,0.1) \tag{7}$$

5.3 ARMA-(X)

The ARMA model(Heij et al., 2004) is a widely used time series model for forecasting. It consists of an auto-regression part (AR(p)) that captures the relationship between the current observation and the lagged observation. Secondly, it incorporates a moving average component MA(q) which depends on the dependencies of the current observation and past residual errors. Additionally, the model can be extended by including covariates, turning it into an ARMA-X model(Wang & Jain, 2003). To tune the parameters, the AIC is used with (p,q = 1,2,3,4) where the model with the smallest AIC and thus the best fit is chosen. In Equation 8 the full mathematical description of this model is formulated.

$$y_t = c + \sum_{p=1}^{P} \alpha_p y_{t-p} + \sum_{q=1}^{Q} \theta_q \epsilon_{t-q} + \sum_{m=1}^{M} \beta_m x_{m,t-1} + \epsilon_t \tag{8}$$

5.4 OLS

Typically, a standard linear regression for predictions is formulated, as in Equation 9 where r_t now denotes the log returns and $x_{i,t}$ a predictor at time t.

$$r_{t+1} = \alpha_i + \sum_{k=1}^{K} \beta_{i,k} x_{i,t,k} + \epsilon_{i,t+1} \tag{9}$$

To construct predictors using Principal Component Analysis(PCA), the original predictors $x_{i,t,k}$ are transformed into orthogonal components which are linear combinations of the original predictors. The principal components (PC) created from this, are ordered by the amount of variance they explain in the data. This offers a method for easily monitoring the significant correlations among multiple predictors of future returns that are uncorrelated with each other.

There are various methods for determining the number of principal components to retain, though no single optimal solution exists. In this thesis, the components are chosen based on the scree plot and Kaiser's Criterion, which suggests that PC's with eigenvalues larger than one should be chosen. Based on this criterion, it is determined to retain the first five components.

The regression with these predictors are then formulated as in Equation 10. Graph 2 in Appendix C shows a plot with the methods used to determine the number of factors.

$$r_{t+1} = \alpha_i + \sum_{j=1}^{q} \beta_j PC_{t,j} + \epsilon_{t+1} \tag{10}$$

For the OLS regression models, the Newey-West (NW) standard errors Newey & West (1986) are applied, which is a Heteroskedasticity and Autocorrelation Consistent (HAC) estimator. This can be particularly useful when predicting stock market returns because of momentum effects, where past returns can predict future returns, leading to autocorrelation. Furthermore, volatility clustering, which is a stylized fact of returns, can lead to heteroskedasticity.

5.5 Random Forest

The Random Forest(Breiman, 2001) makes use of multiple regression trees. Each regression tree is a simple model that makes predictions by recursively splitting the covariates into subsets, forming a tree-like structure. The predictions of multiple decision trees are then combined and averaged, which reduces the noise in the predictions of individual regression trees.

Randomization happens in two ways to reduce the variance and prevent overfitting. Initially, the regression tree selects a random subsample of the covariates at each node. The algorithm then determines the best split. It will continue until it meets a stopping criteria, such as the maximum depth of the tree or the minimum number of samples needed for splitting. Secondly, the RF model utilises a technique called bagging, which involves generating multiple bootstrap samples from the original dataset. Bootstrapping means training each regression tree on a different subsample of the data, where each subsample is created by randomly selecting data points. After fitting the regression trees on these bootstrap samples, bagging takes the average of their predictions to make the final prediction.

5.6 Support Vector Regression

The Support Vector Regression of Vapnik et al. (1997) works by identifying the best-fitting tube that approximates a continuous-valued function to balance between model simplicity and prediction accuracy. It starts off by introducing a symmetric convex ϵ-insensitive loss function that applies equal penalties to both overestimations and underestimations to minimize deviations within a set margin (ϵ). The goal is to identify the flattest tube that contains the majority of training data points. This is achieved by solving this convex optimization problem using numerical optimization techniques.

5.7 Hyperparameter tuning

Tuning the hyperparameters of machine learning models is crucial and can significantly impact the performance of the model(Probst et al. (2019), Van Rijn & Hutter (2018)). As a result, a grid search will be used to perform parameter tuning, which is one of the most commonly used methods (Bergstra & Bengio, 2012). Additionally, before each prediction, the machine learning models will be trained on k-fold cross validation. K-fold cross validation splits the data into

k-folds and chooses one fold for testing $k-1$ folds for training. This will be repeated until all folds have been used for testing. Traditionally, the folds are randomly shuffled but to preserve the temporal order of the time series, a blocked cross validation(Cerqueira et al., 2017) is used in this paper. The hyperparameters of the SVR and RF models that are optimised in this thesis are in Table 6 and 7 of Appendix C.

5.8 Performance Measures

To evaluate the performance of the regressions between models, several different metrics are used. The first is the Root Mean Squared Error (RMSE) which is the square root of the average of the squared errors. The RMSE punishes large errors harder, therefore can be sensitive to outliers. The Mean Absolute Error (MAE) is the average of the absolute value of the errors. It is thus less sensitive to outliers. Both are the same unit as the target variable which makes it easier to interpret.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{12}$$

Additionally, the out-of-sample R^2 (Campbell & Thompson, 2008) defined in Equation 13. Here $\hat{y}_t^h$ represents the h-step ahead forecast, y_t^h is the actual value and $\overline{y}_t^h$ is the mean average forecast. This widely used performance measure serves as a comparison of a model's performance against the historical average. The range of the $R^2_{OoS,h}$ is between $(-\infty, 1]$, where negative values indicate that the model performs worse than a historical average.

$$R^2_{OoS,h} = 1 - \frac{\sum_t (y_t^h - \hat{y}_t^h)^2}{\sum_t (y_t^h - \overline{y}_t^h)^2}, \tag{13}$$

5.8.1 AG and PT test

The above-mentioned performance measures focus on reducing a loss function rather than achieving a significant effect in terms of maximizing profits. When trading with options and for short sellers, it is crucial that the sign of the prediction is correct to make informed decisions. Therefore, a Pesaran-Timmermann (PT) test(Pesaran & Timmermann, 1992) is applied to test whether the model can significantly predict the direction of returns. Moreover, even if an investor can accurately forecast the market's movement, it does not guarantee that they will make more profits. This is because errors in predicting the market's direction can lead to a higher magnitude of returns than when no errors occur(Skouras, 2000). Consequently, the Anatolyev-Granger(AG) test was developed by Anatolyev & Gerko (2005) to test for excess profitability. Both the PT and AG tests have an asymptotic distribution of $\mathcal{N}(0,1)$ under the null and can be obtained by the dac_test() function in R.

5.8.2 Minzer-Zarnowitz

Another way of testing the performance of the models is by way of a Minzer-Zarnowitz regression(Mincer & Zarnowitz, 1969), given by Equation 14. In this regression y_t represents the vector of actual values and $\hat{y}_t$ a vector of the predictions. The joint hypothesis of $\alpha = 0$ and $\beta = 1$ is tested by means of a Wald test. The null of this test assumes an unbiased and efficient forecast against the alternative that it is not. This test is performed using R's minzar_test() function.

$$y_t = \alpha_i + \beta \hat{y}_t + \epsilon_t, \tag{14}$$

5.8.3 Diebold-Mariano

A Diebold-Mariano test (Diebold & Mariano, 2002) is used to determine whether the differences in predictive ability between models are significant. Typically, a Diebold-Mariano test statistic is formulated as shown in Equation 15, where $\bar{d}$ is the average loss differential. In this thesis, the mean squared errors will be used as the loss function. $\hat{V}$ is the variance of $\bar{d}$.

$$\text{DM} = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}}, \tag{15}$$

In the traditional DM test, all observations are treated as equal. However, for stock market returns, it can be sensible to put more weight on the negative returns. This approach aligns with behavioral finance theory, which suggests that investors suffer from loss aversion(Novemsky & Kahneman, 2005), meaning that they are more sensitive to losses than to gains, and it is also crucial from the perspective of risk management. To address this, Van Dijk & Franses (2003) made an adjustment to the DM-test by weighing the loss differentials by a function $w(\omega_t)$. Specifically, $w_{LT}(x_t) = 1 - \Phi(y_t)$, where $\Phi(\cdot)$ is the cumulative distribution function of y_t, is used to assign more weight to negative returns. As a result, the weighted Diebold-Mariano (W-DM) test becomes W-DM $= \frac{\bar{d}_w}{\sqrt{\hat{V}(\bar{d}_w)}}$. However, the DM test statistic can be oversized in small samples. Therefore, an adjustment is made to obtain the modified weighted Diebold-Mariano (MW-DM) test statistic, as shown in Equation 16. P denotes the number of out-of-sample observations and h represents the forecast horizon. Both the DM-test and the MW-DM test will have the Partially Protected Bayesian LASSO as the benchmark with a null of equal predictive ability against the alternative that the Partially Protected LASSO is better.

$$\text{MW-DM} = \sqrt{\frac{P + 1 - 2h + h(h-1)/P}{P}}\,\text{W-DM}. \tag{16}$$

6 Results

This section will present the results from both the replication and the extension of this thesis. For the replication part, the performance of the Partially Protected Bayesian LASSO is evaluated in terms of out-of-sample performance, model fit and most importantly, its the ability to protect certain variables from shrinkage. Then, in the extension part, the Partially Protected LASSO is compared to several other models and techniques capable of handling high-dimensional datasets

for predicting S&P 500 returns. Moreover, the ability to create profits and the performance during recession and expansion periods will be examined.

7 Results replication

Figure 1 presents the posterior mean coefficients (β) of the three models, along with their 95% confidence interval of the protected variables. Dummy variables were created for the variables race and gender to separate their categories. In line with, Yaman et al. (2024) the category 'multiple race' is excluded from the protected variables, resulting in its coefficient being shrunk to zero, just as in the original paper. The category female was removed for identifiability. Observing this figure reveals that the partially protected model effectively shifts the mean coefficients(β) further away from zero. This outcome aligns with the original paper, indicating that the partially protected model effectively reduces the shrinkage compared to the no-protection model. The fully protected model results in mean $\beta's$ being the furthest away from zero, which is not surprising since it introduces the least shrinkage. The protection of a variable ensures that there is a marginal posterior distribution for the coefficients but as it can be seen from the figure, it does not exclude zero from the interval. All these results are in line with Yaman et al. (2024). When comparing the mean β coefficients in this thesis to those in the original paper (Figure 3 in the appendix), some differences can be seen. Two main reasons could explain the observed differences. First, after preprocessing, the dataset of Yaman et al. (2024) contained 385 columns, whereas this paper has 449 columns. This can significantly impact the coefficients of the model, as more predictors lead to more parameters to estimate, resulting in different regularization and thus different β coefficients. Secondly, the MICE imputations involves a certain amount of randomness. Different seeds and iterations can produce varying imputed values, leading to different β coefficients.

Figure 1: The posterior mean coefficients of the protected variables with 95% credible intervals for the three different models when predicting the feelings thermometer values of Joe Biden.

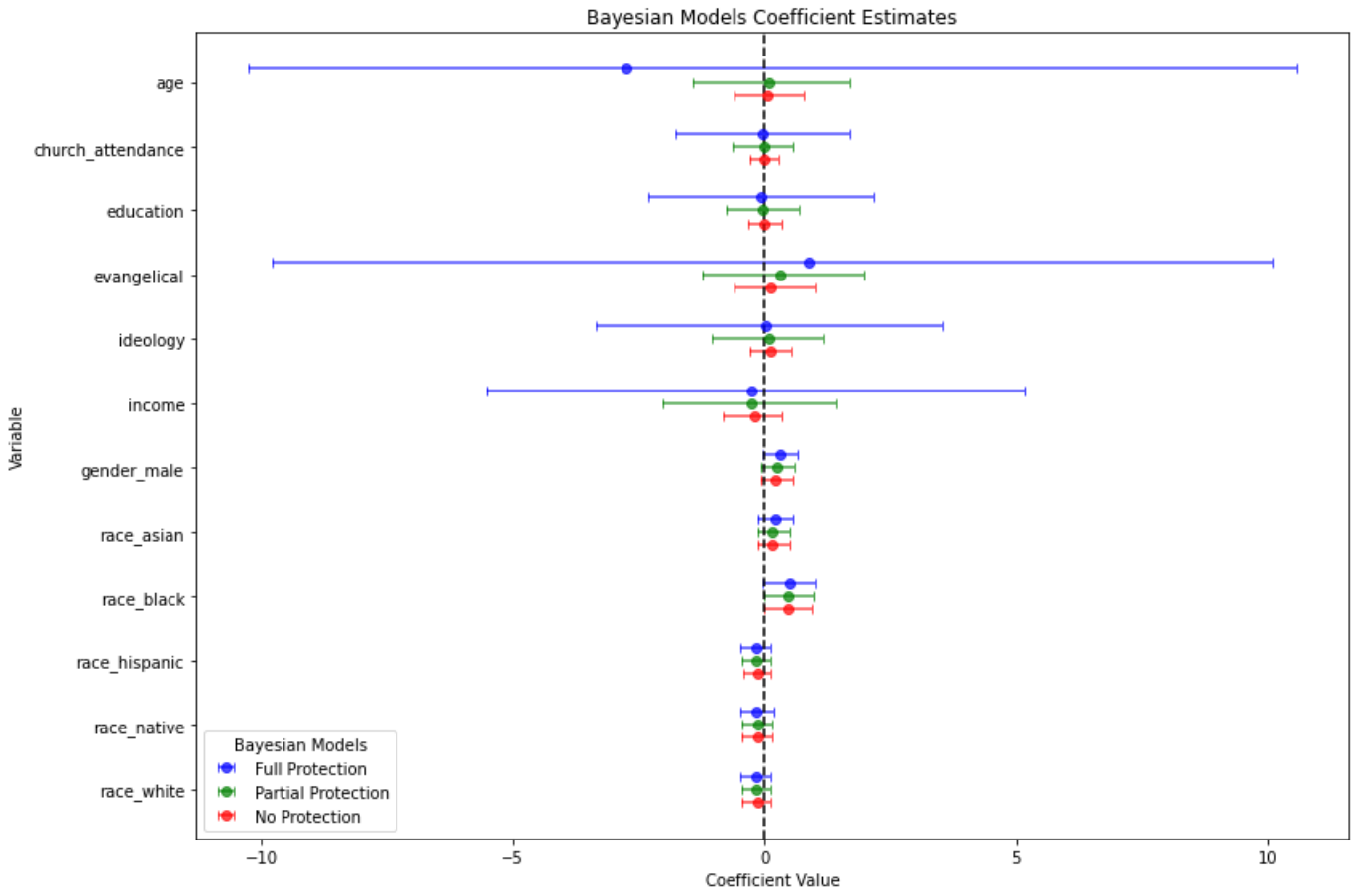


Table 1 shows the prediction measures for the model fit and out-of-sample predictions. The no protection model has the best fit and out-of-sample performance based on the BIC and MSE, while the partially protected model is not far behind based on these measures. In contrast, the full protection model, which offers the most protection to the variables, performs the worst.

Based on Table 1, the same conclusions are drawn as in Yaman et al. (2024). However, the performance metrics differ slightly from the original paper, as shown in Table 8 of the appendix. Across all three models, the MSE increases by approximately 7%, while the BIC decreases by roughly 17%. The worsening MSE results could be attributed to the MICE imputations, which involve a certain amount of randomness, potentially leading to less accurate imputations in this thesis. Furthermore, consistent with previous research (Bai & Ng, 2008), having more predictors does not necessarily make a (high-dimensional) dataset more informative and can lead to more noise, thereby worsening forecasting performance. The BIC values appear to improve compared to Yaman et al. (2024), likely because the dataset in this study led to a better model fit. However, a better fit does not guarantee improved forecasting performance, as can be seen from the MSE values.

The paper of Yaman et al. (2024) primarily focused on the ability of the Partially Protected LASSO to protect theoretically important variables from shrinkage. In terms of assessing the model's predictive capabilities, no definitive conclusions can be drawn, as a small or big difference in performance heavily depends on the target variable. Moreover, there were no tests of whether

the differences in predictive ability were significant. However, these results and the Partially Protected LASSO do serve as a great stepping stone to test the model in various areas, such as finance. This is exactly what will happen in the following sections, where the Partially Protected Bayesian LASSO is used to forecast stock market returns.

Table 1: Comparison of the models in terms of BIC and MSE

Metric	**Partial Protection**	**No Protection**	**Full Protection**
BIC	66272.3702	66291.20905	66209.26983
MSE Train	993.9151	992.84540	998.77920
MSE Test	985.8867	984.56840	993.18050

8 Results extension

This research primarily aims to examine the predictive ability of the Partially protected LASSO against other models described in Section 5. In continuation of the previous research such as those by Campbell & Thompson (2008) and D. E. Rapach et al. (2010), a recursive (expanding) estimation window is utilised for the out-of-sample prediction, where the first 40% of the dataset formed the initial training set.

Table 2 presents the results for several metrics: the RMSE, Minzer-Zarnowitz (MZ), the Diebold-Mariano (DM) and the Modified-Weighted Diebold-Mariano (MW-DM) tests. For the DM tests, the null hypothesis is that the predictive ability of the models is equal, with the alternative being that the protected model performs better in predicting stock market returns.

The RMSE penalizes larger errors harder, meaning models with lower RMSE indicate a reduction of risk on large, unexpected losses for the investor. Conversely, the MAE punishes errors equally giving a more general overview of the model's prediction accuracy of stock market returns. While these two measures focus on prediction errors, the out-of-sample R^2 evaluates the performance of the model on new data and the ability of the model to capture the underlying trends and patterns of stock market returns. The results of these three measures lead to the same rankings of the models, thus Table 2 only shows the RMSE but the MAE and R^2 are still displayed in Table 9 of the appendix. Additionally, the RMSE values are standardized using the protected model as the benchmark, with values smaller than 1.000 indicating improved performance relative to the benchmark.

Table 2: **Performance measures for predicting the S&P 500 returns**

Model	RMSE	MZ p-value	DM p-value	MWDM p-value
Historical	1.322	0.073	0.008	0.009
AR	1.360	0.001	0.006	0.007
Protected	1.000	0.976		
Non-protected	0.969	0.625	0.953	0.930
SVR	1.124	0.555	0.122	0.127
RF	1.084	0.712	0.139	0.146
ARIMA	1.381	0.007	0.004	0.004
ARIMA-X	1.311	0.000	0.088	0.085
OLS	1.881	0.000	0.000	0.000
PCA-OLS	1.003	0.017	0.481	0.583
PCA-RF	1.152	0.222	0.076	0.090

Note. This table presents the RMSE and p-values for the Minzer-Zarnowitz (MZ), Diebold-Mariano (DM), and Modified-Weighted Diebold-Mariano (MW-DM) tests for various predictive models. The null hypothesis for the DM tests is that the predictive ability of the models is equal, with the alternative being that the protected model performs better.

Based the RMSE, the no-protection model performs the best with a value of 0.969, indicating 3.1% improvement compared to the benchmark. When evaluating the p-values of the DM and MW-DM tests, the null of equal predictive ability is not rejected. However, the high p-values of 0.953 and 0.930 suggest it is worth testing whether the no-protection model significantly outperforms the protected model. This test results in a DM statistic of 1.694 with a p-values of 0.0473, indicating that the no-protection model has indeed significantly better predictive ability at a 5% significance level. An OLS model that incorporates all predictors to make forecasts performs the worst, with a RMSE that is 88.1% worse than the benchmark. However, when combined with the PCA method, it becomes the best model after Bayesian LASSO models, with a RMSE that is only 0.30% worse than the protected model. It is even the case that there is no significant difference in predictive ability compared to the protected model. This result aligns with previous literature, likely due to the prevention of overfitting by reducing the noise in the data and creating a more parsimonious model(D. Rapach & Zhou, 2013). The machine learning models, Random Forest and Support Vector Regression, also perform relatively well, despite having 8.4% and 12.4% worse RMSE than the protected model but the difference in performance is not significant based on the DM tests. An interesting result is that the Random Forest model utilising the factors extracted from PCA performs worse than the original RF model, so much so that the difference between the protected model becomes significant. One possible explanation is that while the PCA factors successfully extract the most important information from the variables, they also remove the model's ability to learn from hidden patterns in the features. The models that solely rely on the time series of the S&P 500 returns, historical average, AR and ARIMA, are the worst models, with RMSE values approximate 30% worse compared to the

benchmark. The performance of the ARIMA does seem to improve when covariates are added to the model, highlighting the additional value of the covariates used in this thesis but its predictive ability is still significantly worse than the protected model. Based on these results, no conclusion can be drawn about the performance of linear and non-linear models, as the Bayesian LASSO models are performing the best but do not significantly differ from the non-linear models.

When examining the p-values of the Minzer-Zarnowitz (MZ) test, which tests the null hypothesis of an unbiased and efficient forecast, the Bayesian LASSO models and the machine learning models do not reject the null hypothesis, with p-values well above 10%. However, for the PCA-OLS model, the MZ null hypothesis can be rejected at the 5% significance level, despite being one of the best models based on RMSE, MAE, and R^2. A possible explanation for this is that the model consistently overestimates or underestimates the returns of the S&P 500. If these biases are sufficiently small, it is still possible to achieve good performance metrics despite these systematic errors. Figure 4 supports this, as the residual plot shows most errors concentrated around zero. Additionally, the intercept of 0.0187 and slope of 1.15638 in the MZ regression indicate that the PCA-OLS model tends to predict lower returns when the actual returns are increasing and predict higher returns when bigger negative returns occur.

8.1 Profitability tests

In Table 3, the results of the PT test and AG test are displayed. The table also includes accuracy, which is the proportion of times the model correctly predicted the direction, a crucial metric for short sellers aiming to ensure profitability by capitalizing on price declines and for investors seeking to mitigate losses through timely portfolio adjustments.

The protected and no-protection models perform the best in terms of accuracy, with scores of 0.784 and 0.797, respectively. Additionally, the ranking of the best performers based on accuracy aligns closely with those based on RMSE, MAE, and out-of-sample R^2. One notable exception is the PCA-OLS model, which ranks among the worst in terms of accuracy. These differences in accuracy were further examined using a PT test, with a null hypothesis of no sign predictability. This test revealed that all models, except the ARIMA and AR models, exhibit significant sign predictability at the 1% significance level. Although the ARIMA and AR models have the lowest accuracy, the difference is less than 2% compared to the PCA-OLS model. Despite the small deviation, the PT test outcome differs because it considers more than just the accuracy of the predictions. The PT test also looks at the patterns of the predictions; if correct predictions are clustered in certain periods and incorrect predictions too, then this might affect the outcome of the test.

Similar conclusions can be drawn from the AG test, which has a null hypothesis of no excess returns. Given the similarities between the two tests, this is not surprising. One important note about the AG test is that even if the null hypothesis is rejected, it does not guarantee excess returns in practice. The test does not account for transaction costs and other trading frictions, which could lead to drastically different results.

Table 3: **Profitability tests results**

Model	PT stat	PT p-value	AG stat	AG p-value	Accuracy
AR	-0.638	0.738	-1.526	0.936	0.703
Protected	4.192	0.000	5.134	0.000	0.784
Non-protected	4.406	0.000	5.165	0.000	0.797
SVR	3.418	0.000	3.263	0.001	0.730
RF	3.838	0.000	4.108	0.000	0.770
ARIMA	0.601	0.274	0.636	0.262	0.703
ARIMA-X	3.601	0.000	3.860	0.000	0.743
OLS	3.560	0.000	4.229	0.000	0.784
PCA-OLS	3.626	0.000	4.019	0.000	0.716
PCA-RF	3.418	0.000	3.325	0.000	0.730

Note. This table presents the PT statistic and p-value, AG statistic and p-value, and accuracy for different models. The PT test has a null hypothesis of no sign predictability, and the AG test has a null hypothesis of no excess returns. Accuracy represents the proportion of times the model correctly predicted the direction of stock market returns.

8.2 Recession vs Expansion

In Table 4, the RMSE of the models in periods of recession and expansion are displayed along with their Diebold-Mariano(DM) test p-values. The periods of recession and expansion are defined by the National Bureau of Economic Research (NBER). The RMSE values are standardized using the protected model as the benchmark, with values smaller than 1.000 indicating improved performance. Furthermore, the DM test in this table has a null hypothesis of equal predictive ability compared to the protected model against the alternative that the protected model is better for predicting stock market returns.

The best model is once again the no-protection model, improving the performance by 2.8% and 4.2% compared to the RMSE of the benchmark in periods of expansion and recession. The performance of the models in periods of expansion seems to closely follow the results over the whole sample. One difference is that the Support Vector Regression performs better than the Random Forest and PCA-OLS during expansion phases, with RMSE just 2.3% worse than the benchmark. However, the null hypothesis of equal predictive ability with the benchmark is not rejected for all three models, with p-values ranging from 0.185 to 0.333. During recessions, the ARIMA-X model performs relatively well, surpassed only by the Bayesian LASSO models and PCA-OLS. Unlike its overall sample performance, the ARIMA-X model demonstrates equal predictive ability with the protected model during recessions, as indicated by a DM test p-value of 0.192. This suggests that ARIMA-X can a good model for investors during recessions. Furthermore, during recessions, PCA-OLS achieves the best RMSE for predicting S&P 500 returns, improving this metric by 16.5% compared to the benchmark. However, an additional DM test shows that PCA-OLS does not have significantly better forecasting ability than the protected model during recessions, with a DM statistic of 1.1384 and a p-value of 0.1462. It is

important to note that the recession periods analyzed consist of only eight quarters: 2007Q4 to 2009Q2 and 2020Q1. Consequently, one prediction can significantly skew the results. To address this, studentized residuals were used to test for significant outliers. Tables 10 and 11 in the appendix show that most models contain at least two or three significant outliers during these recession periods. In particular 2020Q1 seems difficult to predict which is not surprising given the sudden emergence of COVID-19.

Table 4: **Model Performance in Recession and Expansion Periods**

Model	RMSE (Recession)	RMSE (Expansion)	DM-Recession	DM-Expansion
Historical	1.759	1.165	0.037	0.042
AR	1.834	1.188	0.028	0.045
Protected	1.000	1.000		
Non-protected	0.958	0.972	0.747	0.948
SVR	1.416	1.023	0.154	0.333
RF	1.142	1.066	0.282	0.185
ARIMA	1.765	1.247	0.032	0.027
ARIMA-X	1.121	1.362	0.192	0.105
OLS	1.954	1.860	0.006	0.000
PCA-OLS	0.835	1.046	0.854	0.212
PCA-RF	1.152	1.153	0.303	0.089

Note. This table presents the RMSE of the models during periods of recession and expansion. The DM-Recession and DM-Expansion show the Diebold-Mariano test p-values for recession and expansion periods. The null hypothesis for the DM test is equal predictive ability compared to the protected model, with the alternative hypothesis being that the protected model performs better.

9 Conclusion

Predicting stock market returns is a very effective method for managing risk and portfolio building. Because many current variables and forecasting models explaining only a small part of stock market returns, it remains of great interest to both academics and investors to identify the best models and techniques. The aim of this study is to compare the model introduced by Yaman et al. (2024) to other methods using high-dimensional data and assess whether it leads to significant improvements. Based on performance metrics, the Partially Protected Bayesian LASSO is the second-best performer after the regular Bayesian LASSO. Although the differences between these two models are small, there is still a significant difference in forecasting ability. The machine learning models (Random Forest and Support Vector Regression) and the PCA-OLS models performed well, despite having worse values for the performance measures, they did not show a significant difference in forecasting ability compared to the Partially Protected LASSO. The PCA-OLS model, however, is biased and inefficient, potentially underestimating risks and failing to capture important market dynamics. Despite these issues, PCA-OLS still shows significant ability to predict the sign and create excess returns, but investors should

consider these limitations if they would use this model. Thus, the Partially Protected Bayesian LASSO fails to beat the regular Bayesian LASSO, PCA-OLS and the machine learning models but it does add a model in the literature with competitive forecasting ability for stock market returns in the presence of high dimensional data. Furthermore, while PCA creates predictors that capture the most variance in the data it removes the ability to interpret the contribution of individual variables in the original dataset and machine learning models lack the capacity to evaluate model parameters and comprehend the significance of variables in making correct predictions, known as the black box paradigm. The clear advantage of the Partially protected LASSO is that the coefficients of the variables are very easy to interpret by means of the posterior coefficients while keeping the original dataset intact.

There are several ways to enhance the performance of the Partially Protected LASSO for predicting stock market returns, particularly in the selection of variables included in the model. Given the volatile nature of the stock market, it may not be optimal to protect the same variables throughout the entire period. Evidence from Paye & Timmermann (2006) indicates structural breaks in the coefficients of predictors for stock market returns, suggesting that incorporating these breaks into the Partially Protected LASSO could significantly improve performance. Additionally, while previous studies have demonstrated strong out-of-sample performance for protected variables used in this thesis, the effectiveness is heavily influenced by the type of forecasting window. The performance of these variables will also vary depending on the country and out-of-sample period examined. For the S&P 500 index returns, selecting variables is particularly challenging due to the significant fluctuations in the index's industry weight over time, as illustrated in Table 12 in the appendix. Therefore, the performance of the Partially Protected LASSO cannot be generalized for all stock market returns.

Another important factor is the choice of prior distribution in the Bayesian model. Proper selection of these prior distributions ensures effective regularization. The Bayesian LASSO employs a prior distribution to address the issue of multicollinearity by shrinking the coefficients, leading to more stable estimates. However, the model may fail to converge if the shrinkage is insufficient. Overall, this paper introduces a novel and transparent model that can provide valuable insights for decision-making in this field.

References

Akyildirim, E., Bariviera, A. F., Nguyen, D. K. & Sensoy, A. (2022). Forecasting high-frequency stock returns: a comparison of alternative methods. *Annals of Operations Research*, *313*(2), 639–690.

Anatolyev, S. & Gerko, A. (2005). A trading approach to testing for predictability. *Journal of Business & Economic Statistics*, *23*(4), 455–461.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C. & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.

Bai, J. & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317.

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305.

Black, A. J., Klinkowska, O., McMillan, D. G. & McMillan, F. J. (2014). Forecasting stock returns: do commodity prices help? *Journal of Forecasting*, *33*(8), 627–639.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Campbell, J. Y. & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Cerqueira, V., Torgo, L., Smailović, J. & Mozetič, I. (2017). A comparative study of performance estimation methods for time series forecasting. In *2017 IEEE international conference on data science and advanced analytics (DSAA)* (pp. 529–538).

Chinco, A., Clark-Joseph, A. D. & Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance*, *74*(1), 449–492.

Dai, Z., Li, T. & Yang, M. (2022). Forecasting stock return volatility: the role of shrinkage approaches in a data-rich environment. *Journal of Forecasting*, *41*(5), 980–996.

Diebold, F. X. & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Dulhare, U. N., Ahmad, K. & Ahmad, K. A. B. (2020). *Machine learning and big data: concepts, algorithms, tools and applications*. Hoboken, USA: John Wiley & sons.

Enke, D. & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*(4), 927–940.

Fan, J., Han, F. & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314.

Feng, G., Giglio, S. & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

Heij, C., De Boer, P., Franses, P. H., Kloek, T. & Van Dijk, H. K. (2004). *Econometric Methods with Applications in Business and Economics*. Oxford, United Kingdom: Oxford University Press.

Huang, W., Nakamori, Y. & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, *32*(10), 2513–2522.

Liu, L. & Pan, Z. (2020). Forecasting stock market volatility: The role of technical variables. *Economic Modelling*, *84*, 55–65.

Ludvigson, S. C. & Ng, S. (2007). The empirical risk–return relation: A factor analysis approach. *Journal of Financial Economics*, *83*(1), 171–222.

McMillan, D. G. (2021). Forecasting us stock returns. *The European Journal of Finance*, *27*(1-2), 86–109.

Mincer, J. A. & Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance* (pp. 3–46). NBER.

Neely, C. J., Rapach, D. E., Tu, J. & Zhou, G. (2014). Forecasting the equity risk premium: the role of technical indicators. *Management Science*, *60*(7), 1772–1791.

Newey, W. K. & West, K. D. (1986, April). *A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix* (Working Paper No. 55). National Bureau of Economic Research.

Novemsky, N. & Kahneman, D. (2005). The boundaries of loss aversion. *Journal of Marketing Research*, *42*(2), 119–128.

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Paye, B. S. & Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, *13*(3), 274–315.

Pesaran, M. H. & Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, *10*(4), 461–465.

Probst, P., Wright, M. N. & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Qiu, M., Song, Y. & Akagi, F. (2016). Application of artificial neural network for the prediction of stock market returns: The case of the japanese stock market. *Chaos, Solitons & Fractals*, *85*, 1–7.

Rapach, D. & Zhou, G. (2013). Forecasting stock returns. In *Handbook of economic forecasting* (Vol. 2, pp. 328–383). Elsevier.

Rapach, D. E., Strauss, J. K. & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, *23*(2), 821–862.

Sigo, M. O. (2018). Big data analytics-application of artificial neural network in forecasting stock price trends in india. *Academy of Accounting and Financial Studies*, *22*(3).

Skouras, S. (2000). Risk neutral forecasting. *SSRN Electronic Journal*, 1–39.

Stock, J. H. & Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, *1*, 515–554.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Van Dijk, D. & Franses, P. H. (2003). Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. *Oxford Bulletin of Economics and Statistics*, *65*, 727–744.

Van Rijn, J. N. & Hutter, F. (2018). Hyperparameter importance across datasets. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2367–2376).

Vapnik, V. (2013). *The nature of statistical learning theory*. New York, USA: Springer Science & Business Media.

Vapnik, V., Golowich, S. E., Smola, A., First, G. & Shausee, R. (1997). Support vector method for function approximation, regression estimation, and signal processing. In *Advances in neural information processing systems 9: Proceedings of the 1996 conference* (Vol. 9, p. 281).

Vijh, M., Chandola, D., Tikkiwal, V. A. & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia Computer Science*, *167*, 599–606.

Wang, G. C. & Jain, C. L. (2003). *Regression analysis: modeling & forecasting*. New York, USA: Institute of Business Forecasting Planning.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). Hoboken, USA: 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.

Yaman, S., Atalan, Y. & Gill, J. (2024). Bridging prediction and theory: Introducing the bayesian partially-protected lasso. *Electoral Studies*, *87*, 102730.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.

A Programming code and Data

For the replication the final dataset used is named "updated_data" in the "code thesis" zip file. The protected variables have the following names in the datafile: "ideology", "church_attendance", "evangelical", "age", "education", "income", "race_white", "race_black", "race_hispanic", "race_asian", "race_native", "gender_male". The code for the models can be found in the map "replication code" also within the zip. This code for the models was provided by Yaman et al. (2024) and is publicly available on Github[1]. Lastly, the MICE imputations were done with the following settings: mice(dataframe , m=1, method='cart', maxit=15, seed=500)

For the extension the final dataset is named "extension_dataset" and the dataset with the predictors formed from PCA are called "pca_first_5_factors". The protected variable have the following names: "purchasing_managers_index", "interest_rate", "log_commodity", "us_treasury_bill_rate___3_month",

[1] https://github.com/selimyaman/protectR

"three_mo_yield_curve", "six_mo_yield_curve", "one_yr_yield_curve", "five_yr_yield_curve". The for the Bayesian LASSO models the R file "lasso for predictions" is used. Some code was added to this code to be able to make one-step-ahead predictions over a expanding window but the core of the code is still the same as Yaman et al. (2024).

B Data

Table 5: The data used for the extension with their variable names, type and sources

Variable	**Type**	**Source**
Log Returns	Target	kaggle.com[2]
US GDP CONA	Macroeconomical	Datastream
US PERSONAL CONSUMPTION EXPENDITURES CONA	Macroeconomical	Datastream
US GOVERNMENT CONSUMPTION and INVESTMENT CONA	Macroeconomical	Datastream
US PRIVATE DOMESTIC FIXED INVESTMENT CURA	Macroeconomical	Datastream
US CHANGE IN PRIVATE INVENTORIES CONA	Macroeconomical	Datastream
US EXPORTS OF GOODS and SERVICES CONA	Macroeconomical	Datastream
US IMPORTS OF GOODS and SERVICES CONA	Macroeconomical	Datastream
US GNP CONA	Macroeconomical	Datastream
US IPD OF GDP	Macroeconomical	Datastream
US CHAIN-TYPE PRICE INDEX OF GDP	Macroeconomical	Datastream
US THE CONFERENCE BOARD LEADING ECONOMIC INDICATORS INDEX	Macroeconomical	Datastream
US CURRENT ACCOUNT BALANCE CURA	Macroeconomical	Datastream

Continued on next page

[2] https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data/data

Table 5 continued from previous page

Variable	Type	Source
US GOODS and SERVICES BALANCE ON A BALANCE OF PAYMENTS BASIS CURA	Macroeconomical	Datastream
US CAPITAL AND FINANCIAL ACCOUNT BALANCE CURA	Macroeconomical	Datastream
US FOREIGN RESERVE ASSETS CURN	Macroeconomical	Datastream
US EXPORTS F.A.S. CURA	Macroeconomical	Datastream
US IMPORTS F.A.S. CURA	Macroeconomical	Datastream
US VISIBLE TRADE BALANCE CURA	Macroeconomical	Datastream
US MONETARY BASE CURN	Macroeconomical	Datastream
US MONEY SUPPLY M1 CURN	Macroeconomical	Datastream
US MONEY SUPPLY M2 CONA	Macroeconomical	Datastream
US FEDERAL FUNDS TARGET RATE	Macroeconomical	Datastream
US TREASURY BILL RATE - 3 MONTH	Macroeconomical	Datastream
US INTERBANK RATE - 3 MONTH	Macroeconomical	Datastream
US PRIME RATE CHARGED BY BANKS	Macroeconomical	Datastream
US TREASURY YIELD ADJUSTED TO CONSTANT MATURITY - 20 YEAR	Macroeconomical	Datastream
US DOW JONES INDUSTRIALS SHARE PRICE INDEX	Macroeconomical	Datastream
US FEDERAL GOVERNMENT BUDGET BALANCE CURN	Macroeconomical	Datastream

Continued on next page

Table 5 continued from previous page

Variable	Type	Source
US PUBLIC DEBT OUTSTANDING CURN	Macroeconomical	Datastream
US TOTAL TREASURY SECURITIES OUTSTANDING (PUBLIC DEBT) CURN	Macroeconomical	Datastream
US FOREIGN NET LONG TERM FLOWS IN SECURITIES CURN	Macroeconomical	Datastream
US CONSUMER CREDIT OUTSTANDING CURA	Macroeconomical	Datastream
US CONSUMER CONFIDENCE INDEX	Macroeconomical	Datastream
US NEW PASSENGER CARS - TOTAL REGISTRATIONS	Macroeconomical	Datastream
US SALES OF NEW ONE FAMILY HOUSES	Macroeconomical	Datastream
US EXISTING HOME SALES: SINGLE-FAMILY and CONDO	Macroeconomical	Datastream
US NATIONAL ASSOCIATION OF HOME BUILDERS HOUSING MARKET INDEX	Macroeconomical	Datastream
US PERSONAL INCOME CURA	Macroeconomical	Datastream
US PERSONAL SAVING AS percentage OF DISPOSABLE PERSONAL INCOME	Macroeconomical	Datastream
US DISPOSABLE PERSONAL INCOME CURA	Macroeconomical	Datastream
US PERSONAL CONSUMPTION EXPENDITURES CURA.1	Macroeconomical	Datastream
US POPULATION	Macroeconomical	Datastream

Continued on next page

Table 5 continued from previous page

Variable	Type	Source
NONFARM INDUSTRIES TOTAL	Macroeconomical	Datastream
US UNEMPLOYMENT RATE	Macroeconomical	Datastream
TOTAL PRIVATE CURA	Macroeconomical	Datastream
MANUFACTURING CURN	Macroeconomical	Datastream
TOTAL PRIVATE VOLA	Macroeconomical	Datastream
US OUTPUT PER HOUR OF ALL PERSONS	Macroeconomical	Datastream
US UNIT LABOR COSTS	Macroeconomical	Datastream
US OUTPUT PER HOUR OF ALL PERSONS NONFARM BUSINESS	Macroeconomical	Datastream
US UNIT LABOR COSTS - NONFARM BUSINESS SECTOR	Macroeconomical	Datastream
COMP FOR CIVIL WRKRS	Macroeconomical	Datastream
US CAPACITY UTILIZATION RATE	Macroeconomical	Datastream
LOANS and LEASES IN BANK CREDIT CURA	Macroeconomical	Datastream
COMMERCIAL and INDUSTRIAL LOANS CURA	Macroeconomical	Datastream
PURCHASING MANAGERS INDEX	Macroeconomical	Datastream
PURCHASING MANAGER BUSINESS BAROMETER	Macroeconomical	Datastream
GENL BUS ACTIV	Macroeconomical	Datastream
US INDUSTRIAL PRODUCTION	Macroeconomical	Datastream
US NEW PRIVATE HOUSING UNITS STARTED	Macroeconomical	Datastream
US NEW PRIVATE HOUSING UNITS	Macroeconomical	Datastream

Continued on next page

Table 5 continued from previous page

Variable	Type	Source
US HOUSING AUTHORIZED	Macroeconomical	Datastream
US CONSTRUCTION EXPENDITURES	Macroeconomical	Datastream
US CORPORATE PROFITS WITH IVA	Macroeconomical	Datastream
US BANKRUPTCY FILINGS	Macroeconomical	Datastream
US CPI ALL URBAN	Macroeconomical	Datastream
US CPI ALL ITEMS LESS FOOD and ENERGY	Macroeconomical	Datastream
PRICE INDEX FOR PERSONAL CON-SMPTN.EXPENDITURE	Macroeconomical	Datastream
PRICE INDEX FOR PCE LESS FOOD and ENERGY	Macroeconomical	Datastream
US export ALL COMMODITIES	Macroeconomical	Datastream
US import ALL COMMODITIES	Macroeconomical	Datastream
US TERMS OF TRADE	Macroeconomical	Datastream
three Mo yield curve	Macroeconomical	U.S. Department of the Treasury
six Mo yield curve	Macroeconomical	U.S. Department of the Treasury
one Yr yield curve	Macroeconomical	U.S. Department of the Treasury
two Yr yield curve	Macroeconomical	U.S. Department of the Treasury
three Yr yield curve	Macroeconomical	U.S. Department of the Treasury
five Yr yield curve	Macroeconomical	U.S. Department of the Treasury
seven Yr yield curve	Macroeconomical	U.S. Department of the Treasury
ten Yr yield curve	Macroeconomical	U.S. Department of the Treasury
PE ratio	Financial	multpl.com[3]
dividend yield	Financial	multpl.com[4]
earning yield	Financial	multpl.com[5]
VAR_t_1	Technical	Constructed as in Liu & Pan (2020)
VAR_t_2	Technical	Constructed as in Liu & Pan (2020)
VAR_t_3	Technical	Constructed as in Liu & Pan (2020)
VAR_t_4	Technical	Constructed as in Liu & Pan (2020)

Continued on next page

[3] https://www.multpl.com/s-p-500-pe-ratio/table/by-month
[4] https://www.multpl.com/s-p-500-dividend-yield/table/by-month
[5] https://www.multpl.com/s-p-500-earnings-yield/table/by-month

Table 5 continued from previous page

Variable	Type	Source
VOLM_t(1)	Technical	Constructed as in Liu & Pan (2020)
VOLM_t(2)	Technical	Constructed as in Liu & Pan (2020)
VOLM_t(3)	Technical	Constructed as in Liu & Pan (2020)
VOLM_t(4)	Technical	Constructed as in Liu & Pan (2020)
LV_t(1)	Technical	Constructed as in Liu & Pan (2020)
LV_t(2)	Technical	Constructed as in Liu & Pan (2020)
LV_t(3)	Technical	Constructed as in Liu & Pan (2020)
LV_t(4)	Technical	Constructed as in Liu & Pan (2020)
interest rate	Macroeconomical	Bank for International Settlements
log commodity price index	Financial	investing.com[6]

C Methodology

Figure 2: PCA results. The left panel displays the explained variance ratio for each principal component, including the cumulative explained variance (shown by a line) and the individual explained variance (represented by bars). The scree plot, shown on the right side, shows the eigenvalues of the principal components. The red dashed line indicates the Kaiser criteria, which signifies components with eigenvalues above 1.

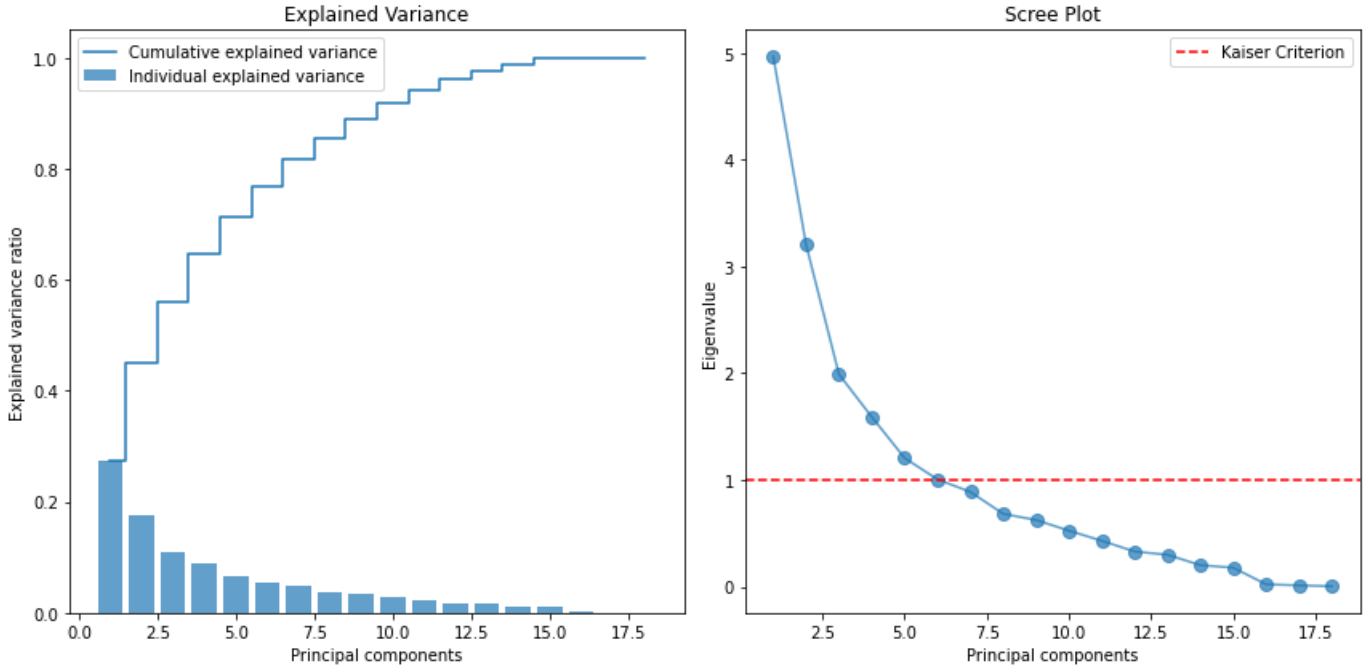


Table 6: Hyperparameter grid for Random Forest

Parameter	**Values**	**Description**
n_estimators	[100, 200, 300, 800, 1000, 1500]	Number of trees in the forest.
max_depth	[None, 10, 20, 40, 60, 80, 100]	Maximum depth of the tree.
min_samples_split	[2, 5, 10]	Minimum number of samples required to split an internal node.
min_samples_leaf	[1, 2, 4]	Minimum number of samples required to be at a leaf node.
bootstrap	[True, False]	Whether bootstrap samples are used when building trees.

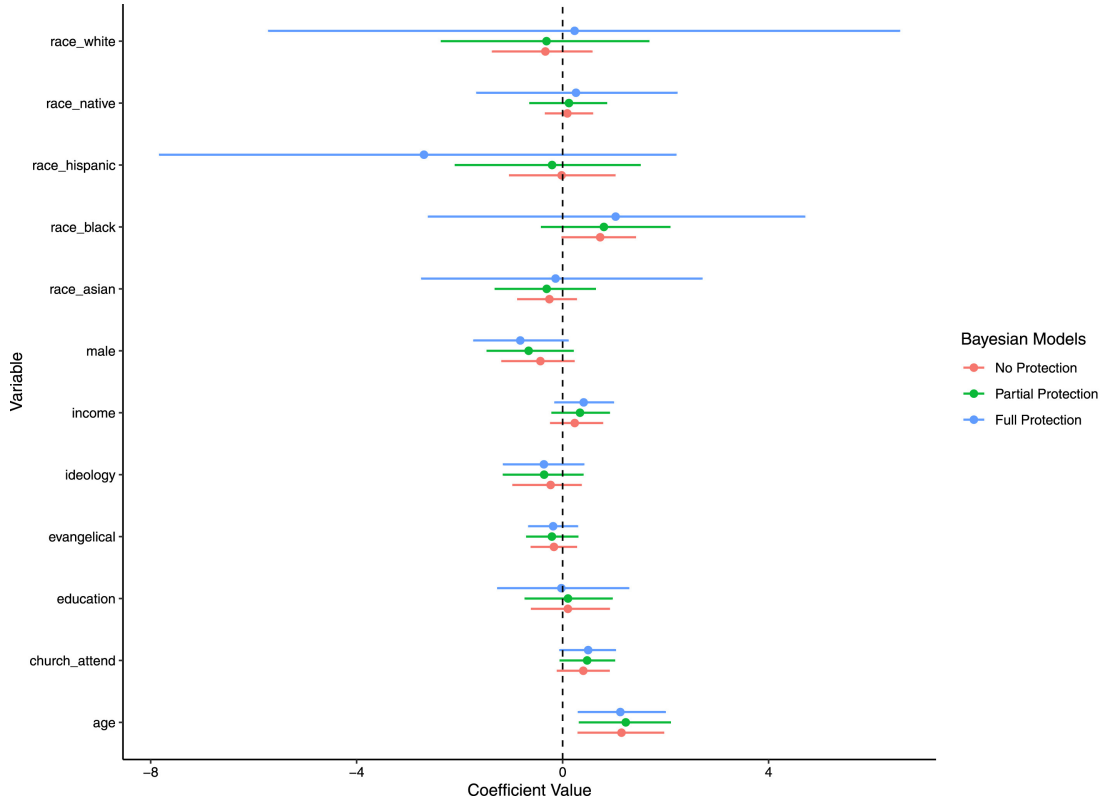
Table 7: Hyperparameter grid for SVR

Parameter	**Values**	**Description**
C	[0.1, 1, 10, 20, 50, 100]	Regularization parameter.
gamma	['scale', 'auto', 0.001, 0.01, 0.1, 1]	Kernel coefficient for 'rbf'.
epsilon	[0.001, 0.01, 0.1, 0.2, 0.5, 1]	Epsilon in the epsilon-SVR model.

[6] https://www.investing.com/indices/sp-gsci-commodity-total-return-historical-data

D Results

Figure 3: The posterior mean coefficients of the protected variables across the three different models as in Yaman et al. (2024).



Note: The error bars represent the 95% credible intervals for the posterior mean estimates.

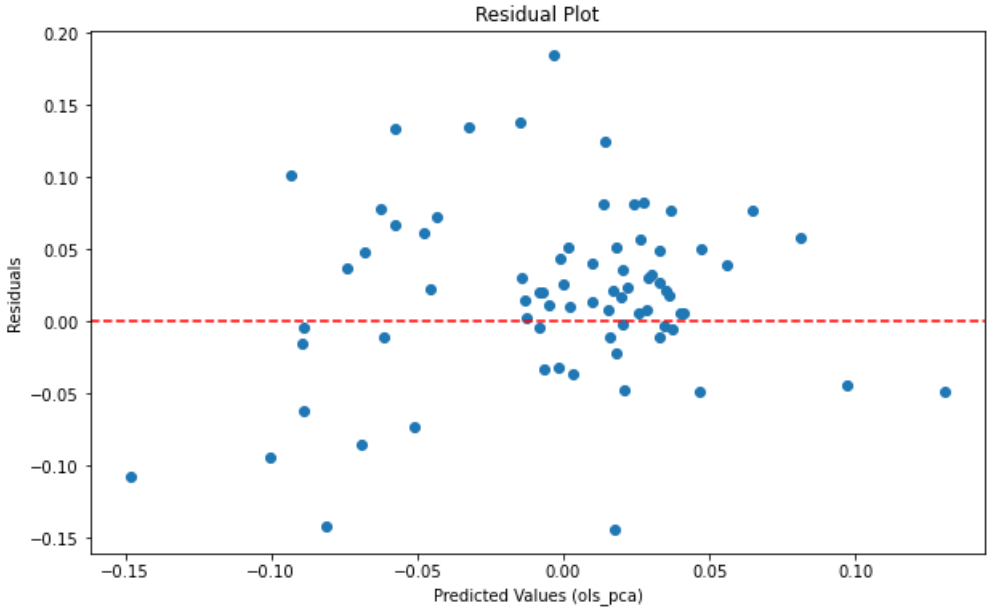
Table 8: MSE and BIC values from the paper of Yaman et al. (2024)

Model	MSE Training	MSE Testing	BIC
No Protection	906.33	915.31	79482.38
Partial Protection	910.07	918.98	80108.71
Full Protection	922.42	929.74	81657.27

Table 9: MAE and out-of-sample R^2 of the models when predicting the S&P 500 return

Model	R^2	MAE
Historical	-0.014	0.059
AR	-0.074	0.060
Protected	0.420	0.045
Non-protected	0.455	0.043
SVR	0.267	0.052
RF	0.318	0.049
ARIMA	-0.106	0.061
ARIMA-X	0.002	0.055
OLS	-1.054	0.094
PCA-OLS	0.417	0.048
PCA-RF	0.229	0.053

Figure 4: The residuals of the PCA-OLS model with the dots representing the residuals of the model



D.1 Studentized residuals

The studentized residuals(Weisberg, 2005) are calculated using the formula in Equation 17. e_i are the residuals of obervation i. $\hat{\sigma}_{(i)}$ the standard error of these residuals calculated as in Equation 18 with n the number of observations and p the numbers of predictors. Lastly, h_{ii} is the i-th diagonal element of the hat matrix $X(X^TX)^{-1}X^T$. The distribution under the null is a student t distribution with n-k-1 degrees of freedom. Rejecting this null indicates a significant outlier.

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \tag{17}$$

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2 \tag{18}$$

Table 10: Studentized Residuals with P-values (Part 1)

Date	ARIMA	ARIMAX	Historical	NoProtec	OLS	OLSPCA
2007-10-01	-0.65 (0.52)	-1.59 (0.12)	-0.81 (0.42)	-1.30 (0.20)	**-2.40** (0.02)	-0.53 (0.60)
2008-01-01	-1.77 (0.08)	-1.17 (0.25)	**-2.04** (0.05)	-1.52 (0.13)	-1.87 (0.07)	-0.28 (0.78)
2008-04-01	-0.69 (0.49)	**-2.49** (0.02)	-0.75 (0.46)	-1.72 (0.09)	**-2.52** (0.01)	-0.59 (0.56)
2008-07-01	-1.12 (0.27)	-0.67 (0.50)	-1.67 (0.10)	-0.49 (0.62)	-1.41 (0.17)	-0.07 (0.95)
2008-10-01	**-3.78** (0.00)	-1.15 (0.25)	**-4.49** (0.00)	**-2.47** (0.02)	-1.70 (0.09)	-1.90 (0.06)
2009-01-01	**-2.02** (0.05)	0.71 (0.48)	**-2.07** (0.04)	-1.04 (0.30)	-0.43 (0.67)	-1.25 (0.21)
2009-04-01	**2.80** (0.01)	-0.14 (0.89)	1.80 (0.08)	0.41 (0.68)	-0.79 (0.43)	1.37 (0.18)
2020-01-01	**-3.23** (0.00)	**-2.24** (0.03)	**-3.37** (0.00)	**-3.14** (0.00)	**-2.20** (0.03)	**-2.35** (0.02)

Note. Values in bold indicate p-values that are significant at the 5% level.

Table 11: Studentized Residuals with P-values (Part 2)

Date	Protec	RF	RFPCA	SVR	AR
2007-10-01	-1.58 (0.12)	-1.13 (0.26)	0.23 (0.82)	-1.30 (0.20)	-0.79 (0.44)
2008-01-01	-1.81 (0.08)	-0.10 (0.92)	-0.12 (0.90)	-1.73 (0.09)	**-1.98** (0.05)
2008-04-01	**-2.20** (0.03)	-1.07 (0.29)	-0.81 (0.42)	-0.73 (0.47)	-0.73 (0.47)
2008-07-01	-0.66 (0.51)	-0.26 (0.80)	-0.06 (0.95)	-0.63 (0.53)	-1.62 (0.11)
2008-10-01	**-2.21** (0.03)	**-3.92** (0.00)	**-2.09** (0.04)	**-5.08** (0.00)	**-4.36** (0.00)
2009-01-01	-0.50 (0.62)	-0.38 (0.70)	**-2.35** (0.02)	-0.84 (0.40)	**-2.01** (0.05)
2009-04-01	0.30 (0.77)	1.45 (0.15)	**2.34** (0.02)	1.68 (0.10)	**2.49** (0.02)
2020-01-01	**-2.92** (0.01)	**-3.00** (0.00)	-1.87 (0.07)	**-2.93** (0.01)	**-3.28** (0.00)

Note. Values in bold indicate p-values that are significant at the 5% level.

E Conclusion

Table 12: Sector Allocations in Percentages of the S&P 500 index[a]

Sector	**July 2023**	**2013**	**2003**	**Median**
Communication Services	9%	38%	5%	15%
Consumer Discretionary	11%	6%	6%	7%
Consumer Staples	7%	7%	12%	8%
Energy	4%	6%	7%	7%
Financials	13%	11%	20%	13%
Health Care	13%	9%	15%	12%
Industrials	8%	7%	12%	8%
Information Technology	27%	11%	17%	16%
Materials	2%	2%	2%	2%
Real Estate	2%	2%	1%	2%
Utilities	2%	2%	3%	3%
Grand Total	**100%**	**100%**	**100%**	**100%**

[a] https://einvestingforbeginners.com/historical-sp-500-industry-weights-20-years/