Karlsruher Institut für Technologie

Department of Economics and Management

Institute for Finance (FBV)

Jun.-Prof. Dr. Julian Thimme

Bachelor's Thesis

Predicting Asset Prices with Text Data: A Machine Learning Approach

by

Simon Leiner

2215787

Wirtschaftsingenieurswesen (Bachelor)

Date of submission

May 19, 2022

Abstract

We investigate the predictability of monthly risk premia from text data from 1984 - 2017. We apply machine learning models to predict the monthly equity risk premium of the S&P 500 index from the full text of 800,000 Wall Street Journal articles, summarized via a topic model by Bybee et al. (2021). We show the non-predictability of the aggregate market equity risk premium on a monthly time horizon. Having negative R^2_{OOS} scores, both machine learning models get outperformed by a naive benchmark forecast of zero. Consistent with previous research and the theory of market efficiency, our results demonstrate that news data is fully reflected in prices within less than a month, thus eliminating their return-predictive information.

Contents

List of Abbreviations

OOS	Out-Of-Sample	HM	Historical Mean
ML	Machine Learning	CAPM	Capital Asset Pricing Model
OLS	Ordinary Least Squares		

List of Figures

List of Tables

1 Introduction

The fundamental goal of empirical asset pricing is to measure and understand the behavior of the risk premium (see, Gu et al., 2020, p.2). Measuring an asset's risk premium is fundamentally a problem of prediction (see, Gu et al., 2020, p.3). However, predicting the risk premium is a notoriously difficult task. Exploiting predictive information moves prices, thereby eliminating the predictability of the market. Hence, the predictive signal is weak and is constantly being pulled towards zero (see, Israel et al., 2020, p.9). The market efficiency hypothesis introduced by Fama (1970) suggests that the risk premium, being the conditional expected excess return, is dominated by unanticipated news, so-called "noise" (see, Ke et al., 2020, p.23). Therefore, we use news data instead of standard economic predictors to predict the aggregate market risk premium. Text data is inherently high dimensional, and predictors are strongly correlated. Conventional statistical methods are not suited for settings where the number of predictors is close to the number of observations. Thus, we need models that can handle high dimensionality and excel at prediction tasks. Machine learning methods, being specialized for prediction tasks, are hence ideally suited for the problem of equity risk premium measurement (see, Gu et al., 2020, p.3).

This paper investigates the prediction of the S&P 500's monthly risk premium from newspaper articles by applying different machine learning models. By employing an expanding window estimation framework, we mimic the situation of a real-time investor. We use the data set provided by Bybee et al. (2021), which condenses the full text of newspaper articles published in the Wall Street Journal (WSJ) from 1984 - 2017. The ultimate goal of this study is to research the predictability of the S&P 500's risk premium from text data.

Return prediction is crucial for the investing industry (see, Israel et al., 2020, p.1). It is accordingly vital to better predict and understand the risk premium. Thus, we investigate whether alternative data sources like text data can predict the risk premium.

Our main empirical finding is the non-predictability of the S&P 500's aggregate market equity risk premium by monthly news attention. Both machine learning models have negative R^2_{OOS} scores. The lasso and random forest model get outperformed by a naive forecast of zero with respectively R^2_{OOS} values of -0.5 % and -9.77 %. Regarding the theory of market efficiency, the immediate implication of the monthly risk premium's non-predictability is that news data is fully reflected in prices within less than a month.

Our study contributes to a growing literature in economics that utilizes text as data. While some papers, like (Bybee et al., 2021) and (Ellingsen et al., 2020), investigate news for macroeconomic time series forecasting and explainability, we focus on the applicability of text data in the field of empirical asset pricing. Our paper combines the two emerging areas of machine learning and textual data in the field of empirical asset pricing.

The rest of this paper is organized as follows. Section 2 introduces the general problem of equity risk premium prediction, describes the machine learning models, and explains the estimation framework and evaluation metric. Section 3 provides general information and descriptive statistics of the data. Section 4 reports the results of the empirical analysis and discusses the findings concerning previous research. Finally, section 5 concludes.

2 Methodology

This section introduces the general problem of empirical asset pricing and describes the machine learning models we employ in our study. First, we present its general idea and objective function for each machine learning model and briefly discuss its advantages. Next, we outline the estimation method, describing the sample splitting and estimation process. Then, we specify the process of tuning the models' hyperparameters. Finally, we explain the R^2_{OOS} metric we use to assess and compare the forecasting performance.

An asset's excess return can be expressed as the conditional expectation of future realized excess returns plus a measurement error (see, Gu et al., 2020, p.3). This implies the general structure to decompose excess returns in an additive manner as follows:

$$r_{t+1} = E_t[r_{t+1}|\mathscr{F}_t] + \varepsilon_{t+1} \tag{1}$$

where r_{t+1} is an asset's excess return and ε_{t+1} is a non-predictive error term. $E_t[r_{t+1}|\mathscr{F}_t]$ is the conditional expectation of r_{t+1} given an information set $\mathscr{F}_t$, which contains information up to time t.

The objective goal is to isolate a representation of $E_t[r_{t+1}|\mathscr{F}_t]$ as a function of variables that maximizes the OOS explanatory power for realized r_{t+1} (see, Gu et al., 2020, p.9). Hence, we try to best approximate the conditional expectation with an a priori unknown function, which can be expressed as follows:

$$E_t[r_{t+1}|\mathscr{F}_t] = \mu_t(\mathscr{F}_t) \tag{2}$$

where $\mu_t(\mathscr{F}_t)$ is an unknown function.

There is a variety of possible presentations for $\mu_t(\mathscr{F}_t)$. The unknown function can either be linear or non-linear as well as parametric or non-parametric and can incorporate few or many predictor variables. Compared to traditional factor models, like the CAPM, supervised machine learning models impose weaker assumptions on $\mu_t(\mathscr{F}_t)$ and instead let the data talk (see, Nagel, 2021, p.12). We mimic the situation of a real-time investor by using an expanding estimation procedure. Consequently, each time new information arrives, we re-estimate $\mu_t(\mathscr{F}_t)$. Hence, $\mu_t(\mathscr{F}_t)$ is time-dependent, allowing for structural changes within the function itself throughout time.

2.1 Models

This section establishes our model choices and briefly introduces the selected machine learning models by describing their general idea, stating the respective objective function, and the main advantages in our specific setting.

Gu et al. (2020) perform a comparative analysis of machine learning models suited for the problem of risk premium prediction. The proposed models can be roughly divided into two subgroups: linear and non-linear models. To cover a wide range of possible functions $\mu_t(\mathscr{F}_t)$ for approximating $E_t[r_{t+1}|\mathscr{F}_t]$, we pick one model of each subclass, such that the discrepancy between them is maximized. Predicting the equity risk premium with two different underlying functions $\mu_t(\mathscr{F}_t)$ allows us to compare their predictions and increases the robustness of our findings.

2.1.1 Lasso

Introduced by Tibshirani (1996), the lasso is a penalized least-squares linear regression model. Choosing the L_1 norm as a penalty function shrinks the coefficients and yields sparse coefficient estimates. It forces some of the coefficients to be exactly zero, hence performing variable selection and thus reducing the number of included predictors. The objective function to minimize is defined as follows:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \mu(x_i, \beta))^2 + \lambda \|\beta\|_1 \tag{3}$$

where $\{y_1, ..., y_N\}$ is a random sample, $\mu(x_i, \beta)$ a parametric function of $\{x_{11}, ..., x_{pN}\}$ with coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ and λ the hyperparameter.

The hyperparameter λ defines how heavily the penalty term is weighted and thus directly influences how many variables with non-zero coefficients are included in the model. λ is chosen by the hyperparameter selection process, described in section 2.3.

The most common and least complex linear model is the OLS regression model. Nevertheless, it's prone to overfitting, particularly because the "signal-to-noise" ratio of return data is low (see, Gu et al., 2020, p.11). Additionally, OLS has the risk of becoming inefficient or inconsistent when the number of predictors is relatively high, compared to the number of observations (see, Gu et al., 2020, p.11). The text data being inherently high dimensional results in the unsuitability of the OLS regression model in our empirical study. In contrast, penalized linear models are robust to overfitting and can deal with high-dimensional data sets. We prefer the lasso model for its superior interpretability. The parametric lasso model assumes linearity of $\mu_t(\mathscr{F}_t)$. However, this might be too restrictive for the potentially non-linear and complex data generating process of $E_t[r_{t+1}|\mathscr{F}_t]$(see, Gentzkow et al., 2019, p.17). As a result, linear models can be viewed as a first-order approximation to $E_t[r_{t+1}|\mathscr{F}_t]$. [1]

2.1.2 Random Forest

A regression tree is a non-linear, non-parametric machine learning model. It aims to approximate the a priori unknown function $\mu_t(\mathscr{F}_t)$ by sequentially dividing the predictor space into subspaces. The predictor variables are selected such that the reduction in impurity is maximized. This strategy may result in satisfactory predictions on the training set, but is likely to overfit, leading to inadequate OOS forecasts. Therefore, the tree must be regularized, which can be accomplished by ensemble methods. Random forest is such an ensemble method introduced by Breiman (2001). Random forest aggregates many independently estimated decision trees and combines their forecasts by using an advanced form of bootstrap aggregation. In contrast to the general "bagging" method, only a random subset of predictors is considered at each split in the tree.

[1] We employ the Lasso implementation from sklearn.

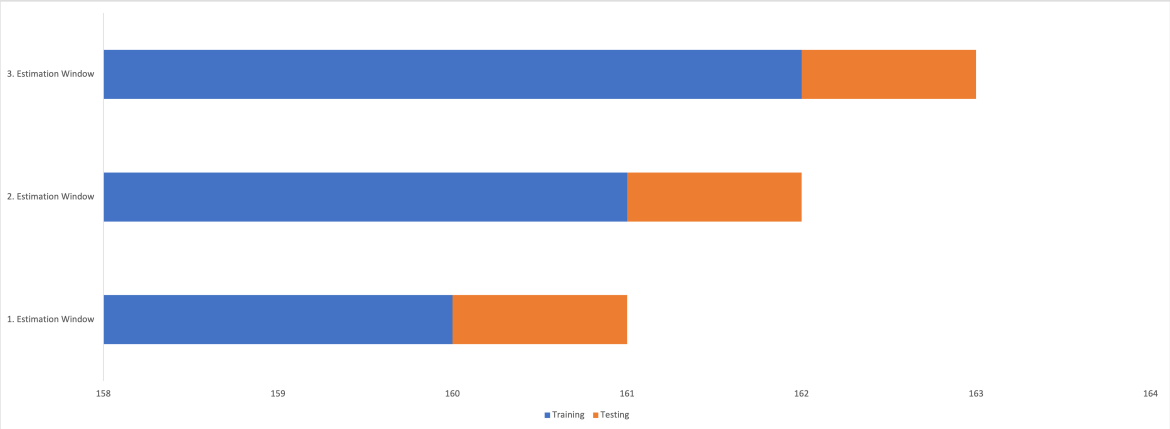
The most important hyperparameters are the number of trees in the forest B and the depth L of the individual trees. Both parameters control the model complexity of $\mu_t(\mathscr{F}_t)$. B and L are chosen by the hyperparameter selection process, described in section 2.3.

In contrast to decision trees, the random forest has several key advantages. First, "bagging" stabilizes and can substantially improve the trees' predictive performance. Second, randomly picking only a subset of predictor variables for splitting results in the reduction of correlation among the different trees and thus further variance reduction (see, Gu et al., 2020, p.18). In contrast to linear models, regression trees are non-parametric models. Not making any assumptions on the unknown function it tries to approximate, it is inherently different from traditional regression models (see, Gu et al., 2020, p.16). [2]

2.2 Estimation Method

Aiming to mimic the situation of a real-time investor, we incorporate all information in $\mathscr{F}_t$ that would have been available for the investor up to a given point in time. This can be accomplished by using an expanding window estimation framework.

Figure 1: Expanding Window Estimation Method



The figure displays the first three estimation windows. The initial training set consists of the first 160 data points, and each time we predict the next month's risk premium, we move one month ahead.

[2] We employ the RandomForestRegressor implementation from sklearn.

Figure 1 shows the first three estimation windows. The initial window comprises 160 data points, and each time we refit the model, we expand the training data by one month (one data point). The initial estimation window covers the period from 01-03-1984 to 01-06-1997 (13 years and four months). We recursively refit the model each month, adding the most recent observation to the training set. With each estimated model, we generate a forecast for the next month's excess return $\hat{r_{t+1}}$. Following this procedure, we obtain one month ahead OOS forecasts for the whole testing set from 01-07-1997 to 01-08-2017 (20 years and two months). Thus, we generate 242 OOS forecasts for the next month's excess return. This procedure enables a flexible representation of the function $\mu_t(\mathscr{F}_t)$, as we allow for potential 242 slightly different versions of $\mu_t(\mathscr{F}_t)$ over time. [3]

2.3 Hyperparameter Tuning

The machine learning algorithms we apply heavily depend on the choice of their hyperparameters. There is little theoretical guidance on selecting the hyperparameters, so we tune them in a data-driven way. Following the expanding window estimation procedure, described in 2.2, we fit the machine learning model 242 times. Each time the model gets estimated, we select the model's hyperparameters accordingly. Hence, we also allow the hyperparameters to change over time. Instead of tuning the parameters with a validation sample, we apply an expanding window estimation inside the "original" training set. [4] This comes with the advantages of not shorting the testing set and is more robust, as the evaluation concerning the objective function is performed multiple times. Inside the "original" training set, we divide the data into training and testing sets and fit the model multiple times with a specific combination of hyperparameter values.

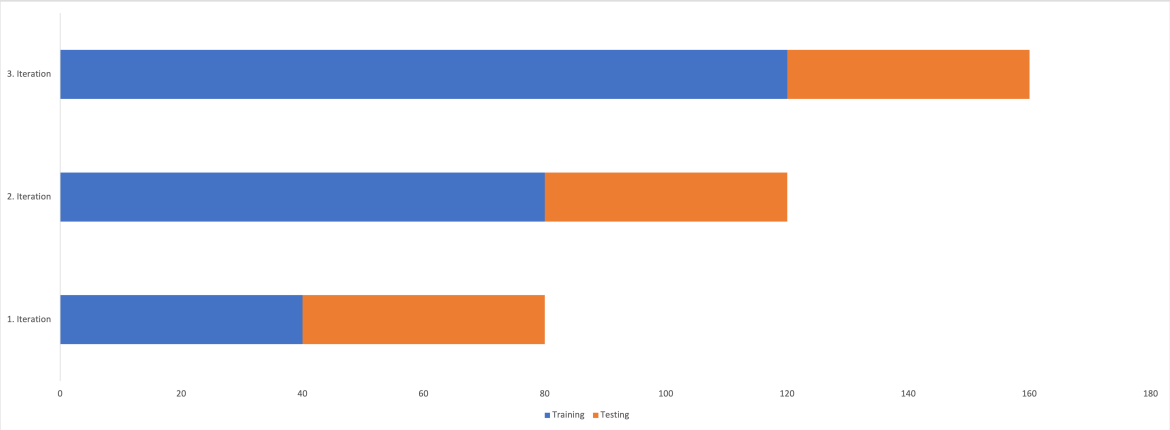
We obtain the predictions of the trained models on the testing sets and compute the R^2_{OOS} scores. We calculate their average score to estimate the expected prediction error. We proceed analogously for all the different parameter combinations given in a predetermined grid of possible values for each hyperparameter. We then select the best hyperparameters according to the highest R^2_{OOS} value. [5]

[3] We cannot use ordinary cross-validation, as it would fundamentally alter the temporal order of the time series.

[4] We employ the TimeSeriesSplit implementation from sklearn with three splits.

[5] We employ the GridSearchCV implementation from sklearn.

Figure 2: Expanding Window Method for Hyperparameter Tuning



The figure illustrates the estimation procedure inside the "original" initial training set, comprising the first 160 data points. Similar to cross-validation, the data is split into three folds, but successive training sets are supersets of previous ones.

This hyperparameter search requires us to refit the model three times the number of hyperparameter combinations. Figure 1 displays the splitting of the initial training set. Having to fit the model multiple times makes the estimation method computationally expensive. Therefore, we only tune the most crucial parameters of the machine learning models, the penalization parameter λ in the lasso model and the depth L and the number of trees B in the random forest model.

2.4 Performance Evaluation

The investor must know if it would have been beneficial for him to employ the machine learning model. Campbell and Thompson (2008) and Welch and Goyal (2007) state that the OOS performance is a useful model diagnostic for an investor. To assess the forecasting performance of the machine learning models, we select the OOS R^2 introduced by Campbell and Thompson (2008). To avoid illusive in-sample findings, we only evaluate OOS.

The R^2_{OOS} statistic is defined as follows:

$$R^2_{OOS} = 1 - \frac{\sum_{t=\in\Omega}(r_{t+1} - \hat{r_{ml,t+1}})^2}{\sum_{t=\in\Omega}(r_{t+1} - \hat{r_{be,t+1}})^2} \tag{4}$$

where r_{t+1} represents the realized excess returns, $\hat{r_{ml,t+1}}$ are the predictions of the ML model and $\hat{r_{be,t+1}}$ are the predictions of a benchmark model. Ω indicates that the fit is calculated on the testing sample, which has been excluded from training or hyperparameter tuning.

For the benchmark model, typically simple historical mean models are chosen (see Campbell and Thompson, 2008; Adämmer and Schüssler, 2020). However, predicting future excess returns with historical averages typically underperforms a naive forecast of zero, thus artificially increasing the R^2_{OOS} (see, Gu et al., 2020, p.22). Consequently, we select a naive forecast of zero as our benchmark model. [6] A positive R^2_{OOS} indicates the outperformance of the machine learning forecasts relative to the benchmark model in a mean squared error sense.

3 Data

We aim to predict the aggregate market equity risk premium and thus select the S&P 500 index as an approximation for a US stock market. We calculate monthly logarithmic returns [7] and subtract the monthly risk-free-rate, approximated by the 1-month Treasury-bill rate [8] to obtain excess returns. The data covers a period from 01-03-1984 to 01-08-2017.

For the predictor variables, we use the data set created by Bybee et al. (2021). They estimate a topic model from the full text of 800,000 Wall Street Journal articles from 01-01-1984 to 01-06-2017 [9]. The topic model summarizes and quantifies the proportion of news attention allocated to each topic over time. They use the full richness of text articles to quantify news data into 180 interpretable topics. By summing over all articles published within the same calendar month, Bybee et al. (2021) aggregate the media attention on a monthly frequency. The news attention is represented by weights $w \in (0,1)$, which when summed up

[6] In the appendix, we provide results for a different benchmark model.

[7] The raw price data is taken from `https://data.nasdaq.com/data/MULTPL/SP500_REAL_PRICE_MONTH-sp-500-real-price-by-month`.

[8] The data is available on Kenneth R. French's website: `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

[9] The text data is available on the website: `http://structureofnews.com`.

over all topics equals one [10]. The topic model already drastically reduces the dimensionality of the news data, but with 180 narratives, the data is much higher dimensional than the usual data used in economics and finance (see, Gentzkow et al., 2019, p.2). Additionally, the data set is high-dimensional relative to the number of monthly observations (see, Bybee et al., 2021, p.9f). Since the proportion of news attention over time is evolving, the text data undergoes structural shifts and exhibits time-series persistence (see, Bybee et al., 2021, p.2). In contrast to the text data, the returns are weak stationary and exhibit little autocorrelation.

The ultimate goal of our empirical study is to predict monthly excess returns. Consequently, we match the text articles published in month t with the excess return of the next month $t+1$. When combining the news and the S&P 500's excess returns, we end up with 402 monthly observations from 01-03-1984 to 01-08-2017. The following table displays some descriptive statistics of the S&P 500 index excess returns.

Table 1: Descriptive Statistics of the S&P 500 Excess Returns

	# observations	mean	std	min	25%	50%	75%	max
excess return	402	0.39 %	36 %	-23 %	-1.2 %	0.8 %	0.025 %	11 %

This table shows descriptive statistics of 402 realized excess returns for the period from 1-03-1984 to 1-08-2017. The values are represented in % and must be interpreted at a monthly level.

4 Empirical Study

In this section, we apply the proposed models within the estimation framework, described in section 2, to the problem of predicting the aggregate market equity risk premium of the S&P 500 index. First, we evaluate and report the monthly OOS predictability in terms of the R^2_{OOS} performance metric. Second, we explore the differently estimated models and their time-varying hyperparameters and analyze the topics the models thought most advantageous for predicting r_{t+1}. Finally, we discuss the results with a particular focus on previous research.

[10] For further information and the exact estimation procedure of the topic model, please refer to Bybee et al. (2021).

4.1 Risk Premium Prediction

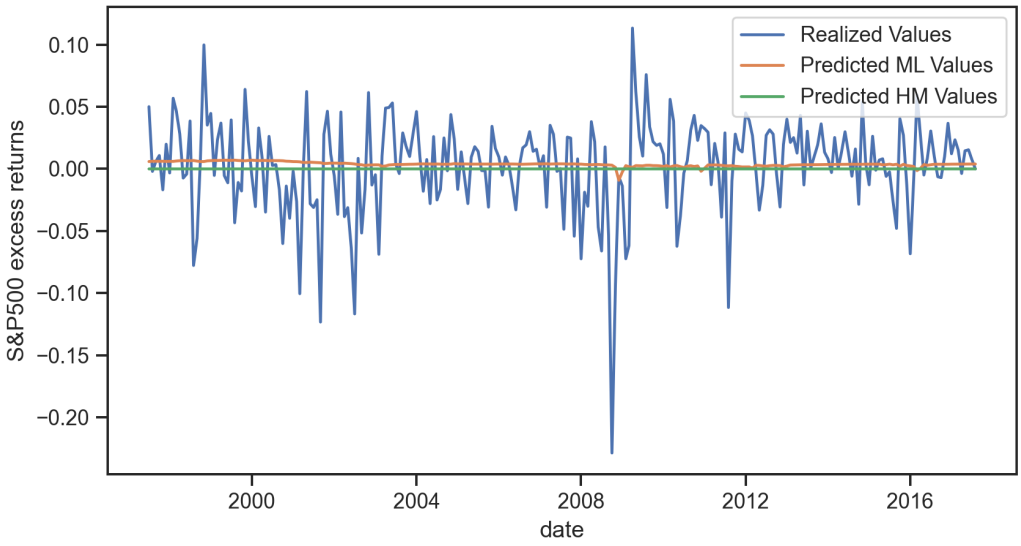
This section investigates the predictability of the monthly excess return of the S&P 500 index from text data. We present the OOS predictability for the lasso and random forest model statistically.

4.1.1 Lasso Predictions

Figure 3 presents the OOS forecasts from the recursively fitted lasso models from 01-07-1997 to 01-08-2017. Over the entire testing sample, the models' predictions are slightly positive and almost constant. All predictions are more or less equal throughout the whole OOS testing sample. The constant predictions of the lasso models result directly from the choice of the hyperparameter. λ controls the coefficients' magnitude and the number of predictors incorporated in the model, thus influencing how much information is included. Hence, λ decides how volatile the predictions are. The constancy of the models' predictions suggests the non-predictability of equity risk premium. To verify our suspicion, we provide further insight into the model and the chosen hyperparameters in section 4.2. The models rarely predict negative monthly excess returns. While the models completely fail to predict the negative excess returns following the crash of the Dot-com bubble in 2000, they adapt to the high volatile negative excess returns following the years of the Financial crisis in 2008. Nevertheless, the model's predictions are just a fraction of the true realized excess returns. Additionally, the negative risk premium predictions after 2008 occur with a delay of a few months, indicating non-predictability of the monthly risk premium. The mean of the realized excess returns in the OOS sample and forecasts is 0.26 % and 0.38 %, which is remarkably close to the mean of the entire period, being 0.39 %. Therefore, the mean prediction is captured, representing the market's overall tendency. With an R^2_{OOS} value of -0.5 % over the entire testing sample, the model performs disappointingly. A negative R^2_{OOS} means that the lasso model gets outperformed by a simple naive forecast of zero, indicating that news fails to predict the next month's excess return r_{t+1}.

Analyzing Figure 3 already provides insights into the OOS performance of the lasso models. Since the choice of the hyperparameters and thus the model complexity and our statistical evaluation heavily depend on the R^2_{OOS} metric, we want to provide deeper insights into the metric.

Figure 3: Monthly OOS Lasso Predictions



The figure shows the lasso models' OOS predicted monthly log excess returns. The predictions result from the estimation process described in 2.2 with an OOS horizon from 01-07-1997 to 01-08-2017. The blue line represents the realized excess returns, and the green line shows the predictions of the benchmark model, being constantly zero.

Therefore, we investigate if and when the forecasting gains and losses occurred in terms of the R^2_{OOS} metric. Accordingly, we proceed in a similar way like Goyal and Welch (2003) and calculate the difference in cumulative squared prediction errors of the machine learning and benchmark model. The metric is defined as follows:

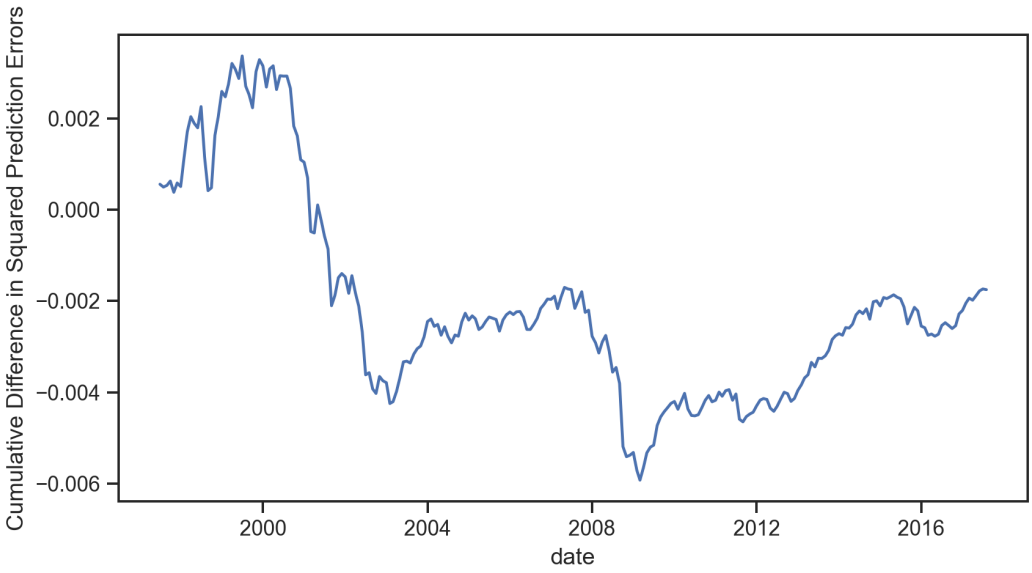
$$Diff(T) = \sum_{t=1}^{T} (r_{t+1} - \hat{r_{be,t+1}})^2 - (r_{t+1} - \hat{r_{ml,t+1}})^2 \tag{5}$$

where r_{t+1} represents the realized excess returns, $\hat{r_{ml,t+1}}$ is the prediction from the machine learning model and $\hat{r_{be,t+1}}$ is the prediction from a benchmark model.

When $Diff(T)$ is positive, the machine learning model outperforms the benchmark model and vice versa. Figure 4 presents the $Diff(T)$ of the predictions from the lasso models. Starting at zero, the lasso models slightly outperform the naive forecast of up to 2 % in the years before 2000. The models, always predicting a slightly more positive return than the benchmark model of zero, lose all of their forecasting gains when the realized excess returns

become negative in the subsequent years after 2000. Failing to predict negative returns, the models always underperform in times of financial struggle, like in the Dot-com bubble after 2000 or the Financial crisis after 2008. Because the overall mean forecast of the models is superior, the lasso models slightly but steadily outperform the naive forecast of zero over large timeframes following the years between 2004 and 2008 and the years after 2009. Due to the similarity between the lasso's and benchmark model's predictions, the $Diff(T)$ values are tiny. The differences in squared prediction errors indicate that the models cannot predict the next month's equity risk premium.

Figure 4: Monthly OOS Lasso Prediction Performance ($Diff(T)$)



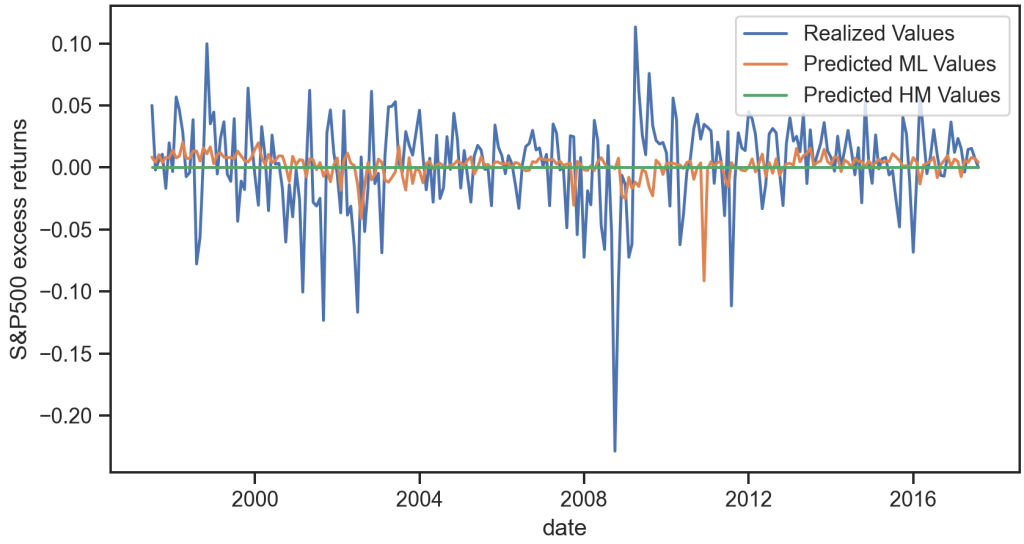
The graph shows the difference in cumulative squared prediction errors of the lasso models and a naive constant forecast of zero over the entire OOS period from 01-07-1997 to 01-08-2017.

4.1.2 Random Forest Predictions

Figure 5 presents the OOS forecasts from the recursively fitted random forest models. The predictions do not capture the trend of the realized monthly risk premium. Without a clear pattern, the predictions are like a random stochastic process evolving around their mean of 0.13 %, suggesting that news cannot predict the next month's risk premium. However, the predictions are not entirely random. Exceptions occur in times of financial drawdown, like

in the financial crisis beginning in 2008, where the predictions are more volatile. The models adjust accordingly and predict more negative excess returns. So, the predictions capture the next month's risk premium pattern, yet with a delay. We will further discuss this time delay in section 4.3 and offer a potential explanation. The R^2_{OOS} score over the entire OOS testing sample is -9.77 % and thus performs even worse than the lasso model, indicating that news fails to predict the next month's excess return r_{t+1}.

Figure 5: Monthly OOS Random Forest Predictions

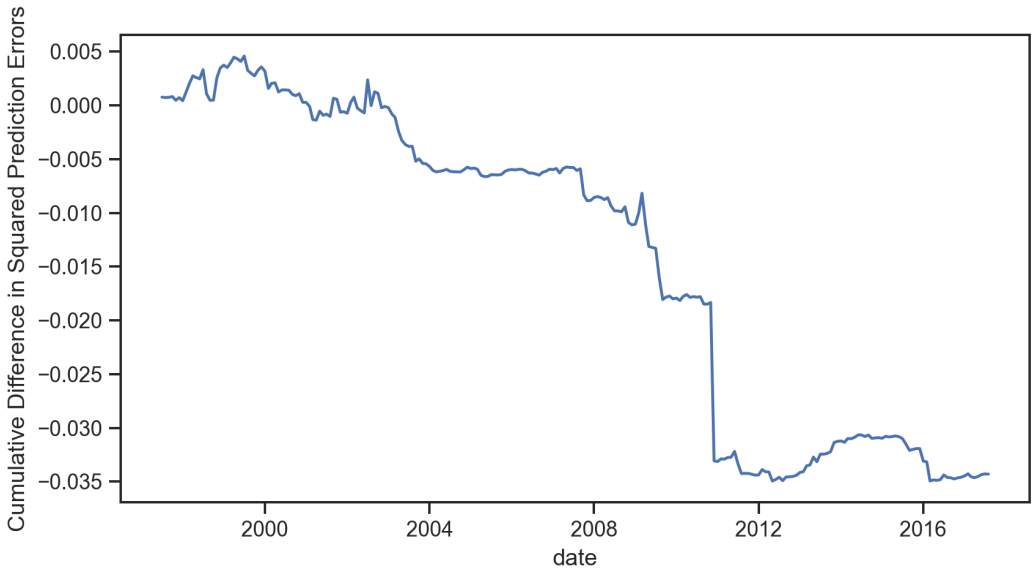


The figure shows the random forest models' OOS predicted monthly log excess returns. The OOS horizon is from 01-07-1997 to 01-08-2017. The blue line represents the realized excess returns, and the green line shows the predictions of the benchmark model.

Inspecting Figure 5 already expresses the unsatisfactory OOS performance of the random forest models. We proceed analogously and investigate the models' statistical performance regarding the cumulative differences in squared prediction errors compared to a naive constant forecast of zero. Figure 6 shows the $Diff(T)$ values of the predictions from the random forest models. Starting at zero, the random forest models get outperformed by the benchmark model, implying the random forest models are inferior to a constant naive forecast of zero over the first ten years of the OOS sample. In the following years of the financial crisis (2008 -2012), the models completely fail to correctly time and thus predict the monthly risk pre-

mium. The random forest models do not outperform the benchmark model significantly after 2012. The differences in squared prediction errors indicate that the models cannot predict the next month's equity risk premium.

Figure 6: Monthly OOS Random Forest Prediction Performance ($Diff(T)$)



The graph shows the difference in cumulative squared prediction errors of the random forest models and a naive constant forecast of zero over the entire OOS period from 01-07-1997 to 01-08-2017.

Due to the stochastic nature of the random forest model, different runs yield different results. The predictions and hence the R^2_{OOS} scores drastically vary between -16 % and 1 %. However, positive R^2_{OOS} values are the exception. The non-robustness of the random forest algorithm limits the validity of our findings. Nevertheless, it also indicates that the random forest model cannot robustly identify predictive information between news and excess returns, amplifying that the predictions can be viewed as a random stochastic process.

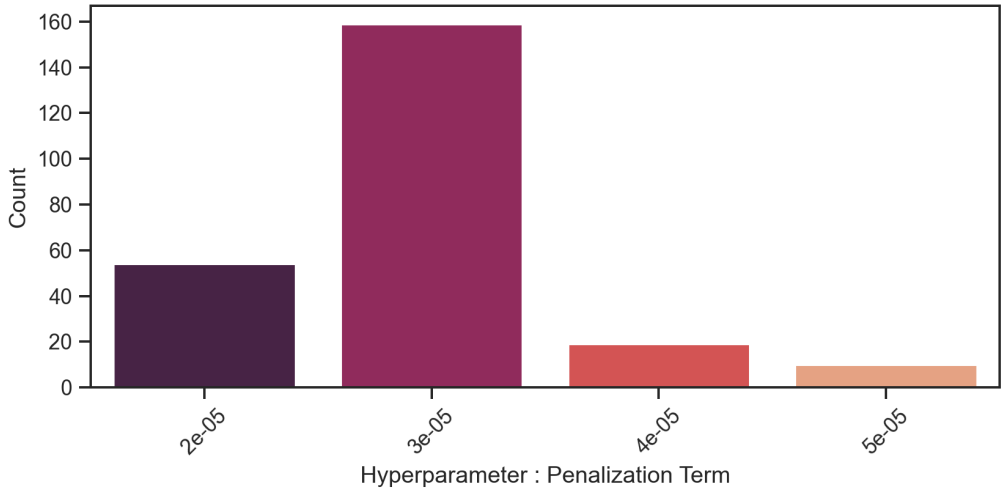
4.2 Model Insights

In this section, we investigate the employed machine learning models. For both the lasso and random forest models, we provide insights into the choice of the hyperparameters and their influence on the model structure and complexity. We identify and analyze the topics the models thought to be most important for predicting the next month's risk premium.

4.2.1 Lasso Inspection

The linear lasso model structure allows a clear interpretation of the estimated models. Therefore, we cannot only evaluate the models' predictions in a statistical sense but also gain valuable insights by analyzing the models' structure. As indicated in section 2.1.1, the hyperparameter λ controls the penalization of the objective function in equation 3. Generally speaking, the bigger λ, the heavier the penalization. The hyperparameters are selected as described in section 2.3 and are thus able to change over time. [11]

Figure 7: Hyperparameter λ Selection



The figure displays how often λ has been independently chosen. Since we refit the model a total of 242 times, the maximum number a single λ value can be picked is limited to 242. Possible values for λ not displayed in the figure are not selected at all.

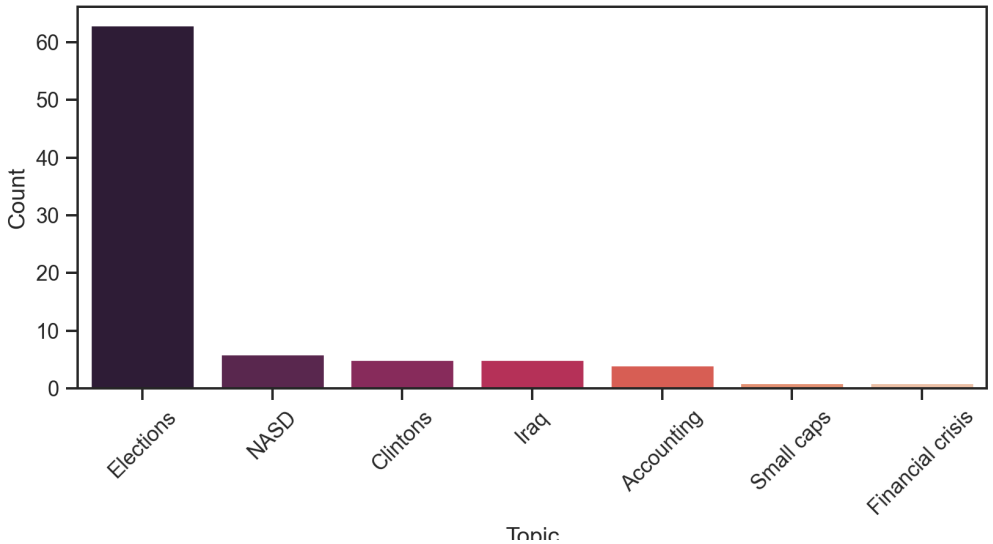
Figure 7 displays the different λ values and how often they have been selected. Instead of favoring the smallest λ values, the R^2_{OOS} scores select λ values that are relatively small, but still sufficient for heavily purging the parameter estimates. λ defines how many predictor coefficients get purged to zero, thus dropping out of the model, hence performing variable

[11] For the grid of values we select [0.000003,0.000005,0.000006,0.000007,0.00001,0.00002,0.00003,0.00004, 0.00005,0.0001,0.001,0.002,0.003, 0.005, 0.007,0.05,0.5,1]. Since the optimization for linear models is not computationally exhaustive, and the lasso model heavily depends on λ, we can try many different values for λ.

selection. To clarify whether the lasso model found a connection between news and future excess returns, we investigate the effect λ has on the number of parameters that get incorporated into the lasso models.

Figure 8 shows what and how often the topics are included in the estimated models. Figure 8 shows that "Elections" is by far the most included topic in the model. "Elections" is a seasonal topic, spiking every four years with secondary peaks every two years (see, Bybee et al., 2021, p.17). "Elections" has the most attention throughout the entire data set, with a mean topic proportion of 1.11 %. However, even the most chosen topic is only included 63 out of potentially 242 times, being about 26 %. Scarcely 7 out of 180 topics are incorporated into any of the 242 differently estimated lasso models. On average, 0.35 topics enter the model, implicating that the models do not include any topics most times, resulting in intercept-only regression models.

Figure 8: Number of included Topics



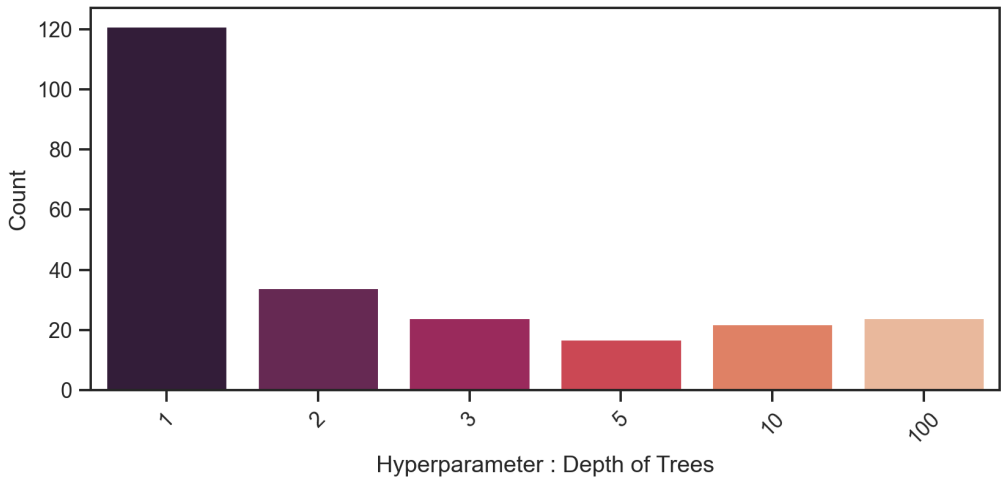
The graph depicts how often and which topics have been included by the lasso models. If the topic's coefficient gets purged to zero, the topic is not included in the model. There are a total of 180 topics (maximum x value), and a topic can be selected a maximum of 242 times (maximum y value), meaning it is part of every model. Topics not displayed in the figure were never incorporated into any model.

Hence, models that include fewer topics are preferred in terms of the R^2_{OOS} metric, suggesting the non-existence of predictive information in news data and amplifying the non-predictability of the risk premium. The analysis of figure 3 already suggested that the predictions' constancy indicates non-predictability. By investigating the models, we can now track this characteristic back to the fact that most lasso models are intercept-only regressions. Intercept-only regressions do not have any predictors and thus are constant models. These models do not incorporate any information from the news data into their predictions. Our findings conclude that it is impossible to predict the S&P 500 index next month's equity risk premium with monthly aggregated news data.

4.2.2 Random Forest Inspection

The random forest model is less interpretable than the lasso model. However, we can still investigate the selection of the hyperparameters, which define the complexity of the model and what topics the models found to be most important while training.

Figure 9: Hyperparameter L Selection

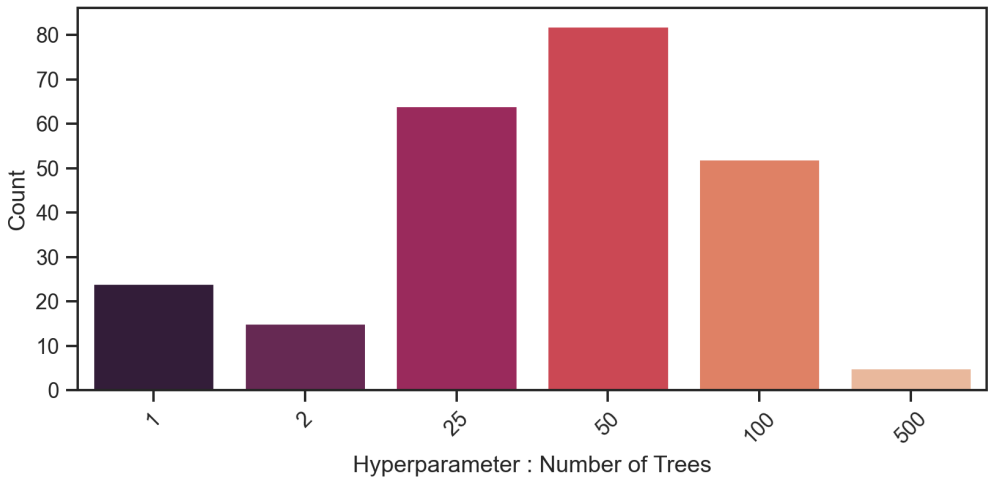


The figure displays how the hyperparameter L has been independently chosen by the 242 different random forest models. Possible values for L not displayed in the figure are not selected at all.

The first important hyperparameter, being the depth L, controls the number of splits in each tree and thus the depth of the individual decision trees. The deeper the tree, the more splits on topics and the more information about the news data is captured. [12] Figure 9 shows that swallow trees are preferred rather than deep trees in the sense of the R^2_{OOS} scores. Most trees are stumps, only splitting on one variable. Most decision trees splitting on only one topic indicates that most topics are not helpful for prediction. Considering that a tree of depth L can capture L - 1 interactions (see, Gu et al., 2020, p.17) , the model rarely improves when allowing interactions between topics.

The second crucial parameter B controls the number of decision trees in the forest. [13] Figure 10 shows that mainly 25 up to 100 trees are incorporated in the random forest models.

Figure 10: Hyperparameter B Selection



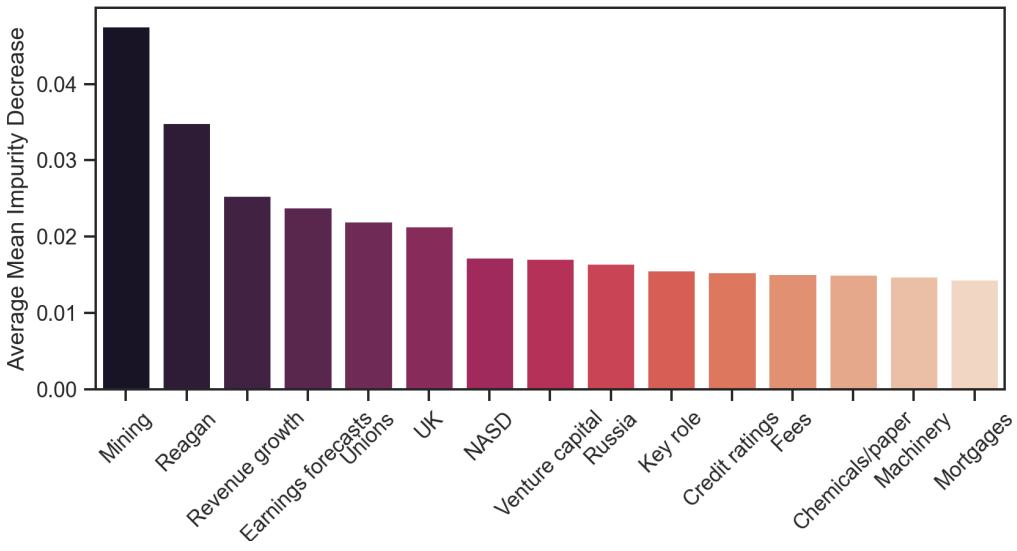
The figure displays how the hyperparameter B has been chosen by the different random forest models. The number of trees in the forest is selected independently each time the model gets refitted. The maximal number a hyperparameter value can be picked is limited to 242. Possible values for B not displayed in the figure are not selected at all.

[12] For the grid of possible values we select [1, 2, 3, 5, 10, 100].

[13] For the grid of possible values, we select [1, 2, 25, 50, 100, 500], allowing potential tiny and huge forests.

Feature importances provide insight into which topics are most relevant for prediction. Like the lasso model, we investigate what topics the models thought most important for predicting the risk premium. Figure 11 displays the averaged feature importances of all 242 trained random forest models. The figure shows the top 15 topics, sorted by their feature importance. "Mining" is the most significant feature while training. Surprisingly, news about minerals is somehow more crucial for predicting excess returns than topics like "Financial Crisis" or "Recession". With most individual trees being stumps, mostly 25 - 100 out of the 180 topics are included in the random forest models.

Figure 11: Aggregated Feature Importance Random Forest Models



The figure shows the averaged feature importance for the most informative 15 topics. The feature importance, being the mean impurity decrease in a decision tree, gets calculated for each run. We average the feature importance of all 242 refitted models.

The random forest is a non-parametric and non-linear model. With its higher complexity, the random forest model can include interactions between topics and hence should theoretically yield more accurate predictions than the lasso model. However, this only applies if the independent variables contain predictive information, which can be exploited to predict the dependent variable. In our empirical study, the mainly intercept-only lasso regressions possess more OOS forecasting power with their naive constant mean prediction than the random

forest models, which incorporate more topics and thus more information. The additional information included in the random forest models does not help to predict the risk premium. Although the additional information results in more volatile predictions, the timing does not fit, resulting in worse OOS forecast accuracy. These findings suggest that news data cannot predict the S&P 500 index next month's equity risk premium.

4.3 Discussion

The market efficiency hypothesis, introduced by Fama (1970), suggests that the risk premium is dominated by unanticipated news. In its strongest form, the news is immediately incorporated into prices, eliminating predictability. However, risk aversion and limits to arbitrage can result in the predictability of future asset prices. Consequently, the predictability of the risk premium heavily depends on how fast information is fully reflected in asset prices. Our results, the non-predictability of the S&P 500's monthly risk premium from news text, support the economic theory and indicate the validity of the market efficiency hypothesis at the monthly period. Furthermore, our findings imply that news data is fully reflected in prices within less than a month, thus eliminating the monthly equity risk premium predictability.

We are not the first ones investigating the relationship between text data and asset prices. With a novel text-mining approach, Ke et al. (2020) research the predictability of individual stocks on a daily and intra-day frequency. They find that compared to smaller firms, larger ones, measured by their market capitalization, respond less to news data and faster incorporate their predictive information into prices. Ke et al. (2020) not only investigate predictability but also provide results on the explainability and the prediction of the previous day's stock returns. They find that prices move ahead of the news, indicating that much of the daily news flow is already known to market participants. Hence, news has the most substantial connection with the previous month's stock return. The connection is weaker for returns in the same month and even more fragile for returns in the next month. However, due to limits in arbitrage and rationally limited attention, they find that news text is significantly predictive of future asset prices on the next day. In their analysis, Ke et al. (2020) differentiate between "stale" and "fresh" news and find that "stale" news gets fully incorporated within two days of arrival, while it takes four days for fresh news to be assimilated entirely. These findings are consistent with the finding that daily news predicts stock returns for only one to

two days by Heston and Sinha (2016). Our approach not only differentiates in the models we employ but also in how the data is aggregated. While we aggregate news monthly and aim to predict the S&P 500 index as an approximation for a portfolio, they perform a more in-depth analysis on single stocks at a shorter time horizon. Comparing results suggests that news data loses its predictive information on asset prices when aggregated into an overall index at the monthly level.

Similar to us, Bybee et al. (2021) employ the same text data as independent variables. While we focus on prediction rather than explainability, Bybee et al. (2021) investigate how news explains monthly aggregate market returns. They employ a 5-topic lasso model to regress AR(1) innovations in topic attention on a US stock market index. They report a monthly R^2_{OOS} of 17.8 %, showing that news can explain monthly excess returns. Comparing our and their R^2_{OOS} scores indicates that prices move ahead of the news, resulting in satisfactory explainability but poor predictability of index aggregated excess returns at the monthly level. This relationship between news data and the next month's risk premium thus explains the delay in the lasso and random forest predictions. Bybee et al. (2021) report the topics "Recession", "Problems", "Convertible/Preferred", "Record high" and "Options/VIX" as most meaningful at explaining aggregate market returns. Their topics differentiate from the topics we identified and are theoretically superior. This discrepancy amplifies the difference in risk premium explainability versus prediction.

(Bybee et al., 2022) also investigate the risk premium with news data, yet, they do so by estimating the risk exposures rather than just using the text data for excess return prediction. They utilize identical news data but aggregate the topic's attention daily instead of monthly. (Bybee et al., 2022) identify the topics "Recession", "Record High" and "Trading Activity" as the most influential risk factors. Their identified topics match with those of Bybee et al. (2021), supporting the explainability and non-predictability on a monthly basis.

The closest benchmark to our study is the paper by (Adämmer and Schüssler, 2020). Analogous, they investigate the predictability of the S&P 500 index monthly equity premium from news data. They utilize a 100-dimensional topic model to quantify the news data throughout time. With their model, consisting of flexible univariate regressions, they report extraordinary results with a monthly R^2_{OOS} of 6.52 %. They furthermore show the results for various machine learning algorithms. They find a monthly R^2_{OOS} of -2.97 % for tree-based

methods and R^2_{OOS} of 2.44 % for the lasso model. With an extraordinary R^2_{OOS} of 6.52 %, their results not only heavily contradict our empirical analysis but also previous research and economic theory. To mention just one example, (Gu et al., 2020) perform an analysis of various machine learning methods in the context of empirical asset pricing. They estimate the monthly risk premium of the S&P 500 index with a large collection of stock predictive characteristics for different machine learning models. They report a R^2_{OOS} of -0.86 % for penalized least squares models, 1.37 % for the random forest model and 1.80 % for their best performing neuronal network. Yet with 1.80 % their R^2_{OOS} are just a fraction of the results provided by (Adämmer and Schüssler, 2020). In order to validate our results, we investigate their conclusions and spot the following hints that emphasize the incorrectness of their findings. First, (Adämmer and Schüssler, 2020) choose a simple historical mean model as a benchmark, which can heavily inflate the R^2_{OOS} values. Second, the more complex tree-based methods are inferior to simple univariate linear regressions, which contradict the findings of (Gu et al., 2020). Third, most forecasting gains occur when their model selection strategy is active. This strategy selects only the best out of the 100 univariate regressions, thus just using one topic to predict the next month's risk premium. We highly doubt that a simple linear regression with only one independent variable can capture the complex data generating process of the risk premium and thus can predict the monthly risk premium. Since the model selection process is highly suspicious, we further investigate the process of how the best model gets chosen. (Adämmer and Schüssler, 2020) state that the best and thus selected model is the univariate regression with the lowest OOS mean squared prediction error. For that reason, they incorporate information into their model that is not available at that given point in time. Incorporating OOS information into the model selection process, (Adämmer and Schüssler, 2020) investigate explainability instead of predictability. The look-ahead bias hence invalidates their findings of strong OOS predictability.

In the appendix section A, we provide various robustness checks. We report the performance with a different benchmark model and estimate the model with a reversed data sequence.

5 Conclusion

We investigate the predictability of asset prices with text data on a monthly time horizon. Therefore, we apply two different machine learning models, the linear lasso and non-linear and non-parametric random forest model, to predict the monthly equity risk premium of the S&P 500 index. We show the non-predictability of the subsequent months' risk premium with news data, applying machine learning methods. With negative R^2_{OOS} scores, both machine learning models get outperformed by a naive forecast of zero.

Regarding other research and the theory of market efficiency, we conclude that there is no predictive information in monthly aggregated news data to predict the aggregate market's risk premium since news data is fully reflected in prices within less than a month.

We only investigate the risk premium's predictability on a monthly basis. As a result, our findings and conclusions are not only limited to a monthly time horizon, but also to the two machine learning models we employ. We provide insights and interpretation on the models, yet this is just what the models thought to be most significant for predicting the risk premium. Machine learning does not provide insights into causality, the economic mechanisms, or equilibria behind news and excess returns (see, Gu et al., 2020, p.7).

For future research, we propose a novel strategy that can exploit the predictability of news data on a daily level. We suggest predicting the risk premium with two different combined model forecasts. Well-established economic predictor variables are used to estimate the overall tendency of the market on a monthly level. For the second model, we propose to use AR(1) innovations on daily news data to estimate the risk premium more flexibly. Both forecasts are combined, resulting in daily risk premium forecasts.

A Appendix

The evaluation metric R^2_{OOS} heavily depends on the choice of the benchmark model. Therefore, we provide the results for another benchmark model, validating our results' robustness. Instead of a naive forecast of zero, we use a rolling mean model that always incorporates the last 12 months (data points). For the lasso model we get a R^2_{OOS} value of 4.6 %, which compared to the -0.5 %, is a sizable improvement of 5.1 percentage points. The R^2_{OOS} score of the random forest model, being -3.66 %, is still in negative territory. This underpins that predicting asset prices with historical asset prices typically underperforms a naive forecast of zero by a large margin (see, Gu et al., 2020, p.22), which results in an artificial improvement of the R^2_{OOS} scores.

The expanding window estimation framework is generally a walk forward backtest. We simulate the historical predictive performance of our models and show how they would have performed in the past. To check whether the models overfit the historical data, we reverse the sequence of historical data points (see, Lopez de Prado, 2018, p.162). By playing the information in reversed order, we change the sequence of data points and use different timeframes of the available data for training and testing. This automatically provides a robustness check against our specific choice of initial training and testing horizons. The lasso model has a R^2_{OOS} value of -0.32 %, similar to our original model, indicating the robustness of our findings. For the random forest model we get a R^2_{OOS} of -3.81 %.

References

ADÄMMER, P. AND R. A. SCHÜSSLER (2020): "Forecasting the Equity Premium: Mind the News!*," *Review of Finance*, 24, 1313–1355.

BREIMAN, L. (2001): "Random Forests," *Machine Learning*, 45, 5–32.

BYBEE, L., B. KELLY, A. MANELA, AND D. XIU (2021): "Business News and Business Cycles," Working Paper, National Bureau of Economic Research.

BYBEE, L., B. T. KELLY, AND Y. SU (2022): "Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text," *Johns Hopkins Carey Business School Research Paper*, 21-09.

CAMPBELL, J. Y. AND S. B. THOMPSON (2008): "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, 21, 1509–1531.

ELLINGSEN, J., V. LARSEN, AND L. THORSRUD (2020): "News media vs. FRED-MD for macroeconomic forecasting," .

FAMA, E. F. (1970): "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, 25, 383–417.

GENTZKOW, M., B. T. KELLY, AND M. TADDY (2019): "Text As Data," *Journal of Economic Literature*, 57, 535–574.

GOYAL, A. AND I. WELCH (2003): "Predicting the Equity Premium with Dividend Ratios," *Management Science*, 49, 639–654.

GU, S., B. KELLY, AND D. XIU (2020): "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 33, 2223–2273.

HESTON, S. L. AND N. R. SINHA (2016): "Predicting Stock Returns from News Stories," .

ISRAEL, R., B. KELLY, AND T. MOSKOWITZ (2020): "Can Machines 'Learn' Finance?" *Journal of Investment Management*.

KE, Z., B. T. KELLY, AND D. XIU (2020): "Predicting Returns with Text Data," Working Paper, University of Chicago.

LOPEZ DE PRADO, M. (2018): *Advances in Financial Machine Learning*, John Wiley and Sons.

NAGEL, S. V. (2021): *Machine learning in asset pricing*.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

WELCH, I. AND A. GOYAL (2007): "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction," *The Review of Financial Studies*, 21, 1455–1508.

Declaration of Authorship

I hereby confirm that I have authored this Bachelor's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Karlsruhe, July 30, 2020

Simon Leiner