Norwegian School of Economics
Bergen, Spring 2024

Recurrent Neural Networks in Diverse Market Conditions

Exploring predictive performance for government bond excess return

Aslak Troberg Djuve & Eilef Johansen Meyer

Supervisor: Walter Pohl

Master thesis, Economics and Business Administration

Major: Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

Writing this thesis has been both a challenging and rewarding experience, and the knowledge and insights gained during this process are invaluable. We are deeply grateful to our supervisor, Walter Pohl, for his guidance and invaluable feedback. Additionally, we extend our heartfelt thanks to our friends and family for their unwavering patience, encouragement, and support, which have been a constant source of motivation throughout the writing process.

Norwegian School of Economics

Bergen, December 2024

Aslak Troberg Djuve

Eilef Johansen Meyer

Abstract

This thesis investigates the predictive performance of Recurrent Neural Networks (RNNs) in forecasting excess return in zero-coupon bonds. We evaluate their performance using data from the U.S. and German bond markets. The study assesses predictive accuracy and the economic value in different market conditions.

Technically, we implement various models, including linear regressions, Random Forest Regressors, Principal Component Regression (PCR), Partial Least Squares (PLS), and Recurrent Neural Networks (RNNs). Forward rates and macroeconomic variables are integrated to enhance predictive accuracy, and their impact is analyzed across different market conditions, including the COVID-19 pandemic.

Our analysis shows that RNNs achieved statistically significant improvements (R^2_{OOS}) over the benchmark. For longer maturities, we found improvements of up to 35%, much of it as a result of out-performance in 2020 and 2021. These forecasting accuracy gains translated into significant economic value for the U.S. market. Although promising prediction results were also observed for German bonds(bunds), they did not yield the same economic utility.

This study highlights the promise of RNNs for financial forecasting, but also emphasizes the challenge of making models that generalize to the characteristics of multiple markets.

Contents

Appendices

List of Figures

List of Tables

1 Introduction

The ability to predict bond excess returns has long been a central challenge in financial economics, with significant implications for both academic research and investment practices. For academics, understanding the dynamics of bond excess returns provides insight into risk premia and market efficiency. For practitioners, accurate predictions enhance investment strategies, aiding in asset allocation and portfolio optimization. Given that bonds constitute a fundamental component of global financial markets, improving the ability to forecast excess returns is of significant economic and practical importance.

Historically, two major ideas have shaped this field: the Expectations Hypothesis and the Spanning Hypothesis. The Expectations Hypothesis suggests that bond returns cannot be predicted beyond the level of current long-term rates, and that the historical average serves as the best guide for future returns (Fisher, 1896). However, this idea has been widely challenged. For example, Cochrane and Piazzesi (2005) showed that forward rates and yield spreads have significant predictive power for bond returns. Similarly, the Spanning Hypothesis, which claims that all useful information for predicting bond returns is captured by the yield curve's level, slope, and curvature, has been questioned. Research, such as Ludvigson and Ng (2009), found that adding macroeconomic variables, like inflation and employment data, can improve predictions, suggesting that the yield curve alone is not enough.

In recent years, advances in machine learning has provided new methods to forecast bond excess return. Traditional econometric methods are useful but often struggle to account for the complex relationships found in financial data (Thornton & Valente, 2012). Machine learning models have shown that they can handle these complexities better and provide more accurate predictions. For instance, Bianchi et al. (2021) demonstrated that neural networks incorporating macroeconomic and yield curve data outperform traditional models in predicting bond excess returns in the U.S. However, most studies focus on the U.S, which has abundant data and a relatively unified economic structure. European markets, such as Germany, are less studied.

Although machine learning has improved bond prediction models, gaps remain. Dense feed-forward neural networks, which are commonly used, do not account for the order

of events over time. Recurrent neural networks (RNNs) are built for sequential data, and this can make them more effective for financial time-series predictions. Studies like Gudelek et al. (2020) and Brezak et al. (2012) have highlighted the potential of RNNs for forecasting. However, RNNs application to bond excess returns remains limited. Additionally, their performance across regions, such as the U.S. and Germany, has not been thoroughly explored.

Building on this, the primary objective of this thesis is to evaluate the predictive capabilities of RNNs for forecasting bond excess returns. By leveraging forward rates and macroeconomic variables, this study compares the performance of these models in two distinct markets. The U.S. market, which benefits from a long historical sample spanning August 1972 to December 2022, and the German bond market, where the available macroeconomic data covers a shorter period from January 2000 to December 2022. Despite this disparity, both markets share a out-of-sample period from January 2014 to December 2022, ensuring a comparable evaluation of model performance. To achieve these objectives, we construct a forecasting framework that predicts one-year holding period bond excess returns. Furthermore, we assess the performance of the models during the COVID-19 pandemic, a period marked by volatility and economic uncertainty. In addition to predictive accuracy, we measure the economic value of the forecasts using utility-based frameworks, including mean-variance utility and power utility, to quantify their practical benefits in portfolio decision-making. Through this analysis, the thesis aims to provide a deeper understanding of how RNNs can enhance bond excess return predictions while addressing the challenges posed by regional economic differences, data limitations, and the market disruptions caused by Covid-19.

The results of our research suggest that RNNs deliver notable predictive improvements compared to traditional econometric models in the U.S. market. These gains significantly increase after the onset of COVID-19. This indicates that RNNs may possess a degree of adaptability that allows them to handle rapidly changing market environments more effectively than traditional methods.

In contrast, the predictive gains in the German market are more modest, reflecting the challenges posed by a shorter sample period and data sparsity. Despite these limitations, the improvements remain economically meaningful in some settings, suggesting that

RNNs could still offer value in markets with less data. Nevertheless, the relatively smaller enhancements compared to the U.S. market raise important questions about the generalizability of RNNs across regions. Factors such as structural differences in bond markets, economic variability, or limitations in macroeconomic data availability seem to constrain the predictive capabilities of machine learning models in non-U.S. settings.

Overall, the results highlight the potential of RNNs to improve bond return forecasting, particularly in the U.S. market. However, they also emphasize the need for cautious interpretation. The observed improvements may not generalize across all market conditions or regions. Further research is needed to understand the factors driving these outcomes. This includes exploring alternative model architectures and refining data preprocessing techniques. It also involves examining performance under a wider range of economic and financial environments.

The remainder of this thesis is structured as follows. Section 2 reviews related literature and develops the hypotheses. Section 3 describes the data sources and preprocessing steps. Section 4 outlines the methodology, detailing the economic and technical techniques employed. Section 5 presents the analysis of the results. Section 6 is a discussion of the findings, their implications, limitations and directions for future research. Finally, section 7 is the conclusion.

2 Background

2.1 Literature Review

The predictability of bond excess returns is a key topic in financial economics. It challenges two fundamental theories: the Expectations Hypothesis and the Spanning Hypothesis. These theories have significantly shaped our understanding of the term structure of interest rates. However, their assumptions are increasingly questioned by empirical research.

The Expectations Hypothesis argues that bond excess returns cannot be predicted. It assumes that the term premium, which is the extra return investors require for holding long-term bonds, stays constant over time. According to this view, forward rates are unbiased predictors of future interest rates. The best forecast for bond excess returns is therefore the historical average. However, empirical studies have repeatedly contradicted this assumption. For example, Fama and Bliss (1987) showed that forward rates contain valuable information about future bond excess returns. This directly challenges the Expectations Hypothesis. The finding marked a major shift in the literature by proving that bond excess return change over time.

Building on this, Campbell and Shiller (1991) showed that the yield spread is a strong predictor of bond returns. The yield spread is defined as the difference between long-term and short-term bond yields. It reflects expectations about future interest rate changes and economic activity. This provides an intuitive explanation for time-varying bond excess return. Later, Cochrane and Piazzesi (2005) introduced the CP-factor. The CP-factor is a linear combination of forward rates and the short-term yield. It achieved an out-of-sample R^2 of 44 percent. This provided strong evidence against the Expectations Hypothesis and reinforced the idea that bond excess returns can be predicted.

The Spanning Hypothesis states that the yield curve captures all relevant information for predicting bond excess returns. It focuses on three primary components: the level, the slope, and the curvature. The level captures the overall interest rate environment. The slope reflects the steepness of the curve. The curvature describes the shape of the curve. These components form the foundation of models like the Nelson-Siegel framework and are also a common interpretation of the first three principal components derived

from principal component analysis (Nelson & Siegel, 1987). However, this view has been challenged. Cochrane and Piazzesi (2005) also found that higher-order components of the yield curve provide additional predictive power for bond excess returns. This suggests that conventional measures of the yield curve may fail to capture all relevant risk factors.

Further work by Ludvigson and Ng (2009) expanded on these findings. They introduced macroeconomic variables as predictors of bond excess returns. Their analysis showed that factors like inflation, labor market conditions, and consumption growth improve forecasting accuracy. This is especially true when these factors are combined with yield curve data. These unspanned macroeconomic factors reveal a key limitation of the Spanning Hypothesis. The yield curve alone does not fully capture the broader economic dynamics that influence bond excess return.

Despite these advances, challenges remain in predicting bond excess returns. Small-sample biases can distort results, as noted by Bauer and Hamilton (2018). This raises concerns about the reliability of conclusions drawn from limited data. Additionally, predictors like forward rates and macroeconomic variables often perform well within the sample. However, their out-of-sample predictability tends to weaken, as shown by Goyal and Welch (2008). Another challenge is the instability of models over time. Economic relationships that appear stable in one period may break down in another due to changing market conditions. These issues make achieving consistent and reliable predictions difficult. They also prompt researchers to explore more robust and flexible modeling techniques.

The introduction of machine learning (ML), has created new opportunities for bond return forecasting. Traditional regression models are interpretable and theoretically grounded. However, they often fail to capture the non-linear relationships inherent in financial data (Thornton & Valente, 2012). ML models, such as tree-based methods and neural networks, overcome these limitations. They excel by leveraging their ability to model complex interactions (Gu et al., 2020).

Bianchi et al. (2021) demonstrate that machine learning methods, particularly neural networks, outperform traditional models in predicting bond returns. Their findings show that dense feed-forward neural networks achieve higher out-of-sample R^2 values than other traditional methods. Furthermore, Feng et al. (2024) highlight the effectiveness of machine learning in capturing shifts in risk premia. This is especially true during market

downturns.

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, are promising tools for time-series forecasting. Unlike feedforward networks, RNNs aim to capture the temporal structure in the datasets. While Bianchi et al. (2021) focused on feedforward architectures, other studies offer a different perspective. Fischer and Krauss (2018), Chen et al. (2019), Gudelek et al. (2020) and Brezak et al. (2012) have shown that RNNs often outperform feedforward models in forecasting settings. This is arguably due to their memory-like capabilities.

Regional differences between the U.S. and German bond markets create unique challenges for predictive modeling. The U.S. market is unified, with a single monetary authority (the Federal Reserve) and dollar-denominated bonds. This structure ensures consistent data collection and results in extensive, high-quality datasets. It is also highly liquid. In contrast, Germany's bond market is part of the eurozone. Where monetary policy is overlooked by the Eurosystem (Deutsche Bundesbank, 2024), and economic conditions vary across member states. This adds complexity, such as exposure to sovereign risks and less historical data due to the transition from the Deutsche Mark to the euro in 2002. These differences can make Germany's market more complex and harder to predict.

When it comes to determinants in the U.S. and Germany, research by Beckmann et al. (2023) shows that the slope of the yield curve is a key predictor in both regions for bond excess returns. However, additional macroeconomic variables, such as labor markets and prices, play a more prominent role in Germany than in the U.S. Notably, housing market variables are significant predictors in the U.S. but have minimal predictive power in Germany. This reflects regional differences in economic dynamics.

Evaluating the economic value of predictive models goes beyond statistical metrics like out-of-sample R^2. Thornton and Valente (2012) highlighted that improved forecast accuracy does not always lead to higher investor utility. This is particularly true under quadratic utility frameworks. Rooted in Markowitz's Modern Portfolio Theory (Markowitz, 1952), the mean-variance utility framework is a cornerstone of portfolio optimization. It assesses the trade-off between risk and return by considering the mean and variance of returns. This approach is widely adopted as it provides a tractable method for maximizing expected returns at a given level of risk. However, it has limitations. The framework

assumes constant risk aversion and treats gains and losses symmetrically, penalizing upside deviations as much as downside ones.

To address these limitations, alternative frameworks like power utility offer more flexibility in reflecting investor behavior. Power utility, based on expected utility theory formalized by von Neumann and Morgenstern (1944), allows risk aversion to vary with wealth levels. Unlike the static assumptions of mean-variance utility, power utility accounts for dynamic decision-making across different wealth states. It captures the reality that investors often show heightened risk aversion during wealth declines and greater risk-taking when wealth increases. This framework is especially relevant for long-term investment horizons and intertemporal decision-making. In such cases, wealth fluctuations significantly influence risk preferences. However, the flexibility of power utility comes with a trade-off. It is less analytically tractable and often requires computational methods to solve optimization problems.

The practical application of predictive models in trading strategies provides a tangible measure of their effectiveness. It links statistical performance to economic outcomes. By integrating predictive models into portfolio management, studies like Huang and Shi (2023) have shown improved portfolio performance. These improvements include optimized risk-return trade-offs and enhanced risk management. Such applications highlight the importance of evaluating predictive models beyond statistical accuracy. They emphasize the need to assess their ability to deliver meaningful economic value in real-world investment contexts.

2.2 Hypothesis Development

This thesis aims to advance the understanding of bond excess return predictability by exploring the potential of RNNs. We investigate two main hypotheses, focusing on the performance of RNNs in capturing temporal dependencies, their performance during the Covid-19 period, and their generalizability across financial markets.

H1A. ***Recurrent Neural Networks outperform traditional models.***

Financial and macroeconomic data often exhibit temporal dependencies, where past trends influence future outcomes. LSTMs, with their ability to maintain an internal state across time steps, are particularly suited to modeling both short- and long-term dependencies, making them a valuable tool for financial forecasting (Hochreiter & Schmidhuber, 1997). As discussed in the literature review, studies such as Fischer and Krauss (2018), Chen et al. (2019), Gudelek et al. (2020), and Brezak et al. (2012) demonstrate the better performance of RNNs in various forecasting contexts. Building on this, our study seeks to extend these findings to bond markets to evaluate whether similar findings can be achieved.

H1B. ***Recurrent Neural Networks adapt better to challenging market conditions, such as the onset of COVID-19.***

The rationale for H1B is based on the unique bond market dynamics observed during recessions and results in Bianchi et al. (2021) and (Feng et al., 2024). They found that their results was substantially higher during recessions. Economic downturns often cause abrupt shifts in excess return. These shifts are driven by investor uncertainty, changes in monetary policy, and economic distress. A key phenomenon during recessions is the flight to quality. In this scenario, investors move capital from risky assets to safer assets, like government bonds. This reallocation leads to significant deviations in yields and excess returns. The RNN architecture allows the model to update the internal state based on new information, enabling them to capture the sudden shifts in market dynamics more effectively than traditional models.

We suspect that recessionary dynamics may introduce new temporal dependencies. Traditional models often struggle to capture these complexities. We suggest that RNNs, particularly LSTM models, may be better suited for this task. They can retain and use

long-term sequential information effectively.

H2. ***If Recurrent Neural Networks indeed demonstrate superior performance, this advantage should extend to other markets, specifically the German bond market.***

Our second hypothesis, H2, is grounded in the idea of model generalizability.

Although there are obvious regional differences, the fundamental dynamics of bond markets are similar. Yield curve dynamics, macroeconomic indicators, and investor behavior are key drivers of bond excess returns in both the U.S. and German markets. RNNs can capture these universal relationships, potentially making them effective across different markets.

RNNs are designed to handle sequential data and capture temporal dependencies, which are universal characteristics of financial time series. This capability is not market-specific but rather a function of the model's architecture. Therefore, we believe it is reasonable to expect that RNNs can perform equally well in different markets, provided that the data quality and availability are sufficient.

3 Data

3.1 Yield Data

3.1.1 US Yield Data

The Liu and Wu (2021) yield curve dataset provides a comprehensive monthly series of zero-coupon yields for U.S. Treasury securities. It is reconstructed to reflect historical market conditions. Spanning from June 1961, the dataset offers a consistent and reliable foundation for studying term structures, bond market dynamics, and excess returns. Using this dataset, we calculate forward rates and excess bond returns for maturities up to 10 years. This enables a detailed analysis of U.S. Treasury market behavior. Our sample period starts in January 1972, aligning with the issuance of 10-year U.S. Treasury bonds.

We exclude bonds with maturities longer than 10 years from this analysis due to data limitations. The U.S. Treasury began issuing 20-year bonds in July 1981. This creates a significant gap in the dataset for these maturities prior to that time. Similarly, 30-year Treasury bonds were introduced in November 1985, further limiting data availability for longer maturities in earlier years. By focusing on maturities up to 10 years, we maintain dataset continuity and yield structure consistency. This allows for robust forward rate and excess return calculations over the entire historical range. Additionally, this approach centers the analysis on the most liquid and frequently traded segment of the U.S. Treasury market. This provides a reliable foundation for exploring bond excess returns over time.

3.1.2 German Yield Data

For the German market, we use the Bundesbank zero-coupon yield curve dataset. This dataset is derived from listed federal securities using the Svensson method. The Svensson method is a parametric approach widely used to create smooth yield curves from observable bond prices (Svensson, 1994). The data is sourced from the Bundesbank's time series database (Bundesbank, 2024). It spans from September 1972 to 2024 and includes maturities of up to 30 years. Although the dataset offers a longer historical period, our sample for Germany begins in January 2000. This ensures alignment with the availability of German macroeconomic data.

To ensure comparability across regions, we focus on bonds with maturities up to 10 years. Although longer-maturity bonds exist in Germany, their issuance frequency and continuity vary. This makes direct comparison across markets difficult. By concentrating on core maturities, we maintain consistency between the U.S. and German datasets. This approach supports a robust cross-market analysis of bond excess returns.

3.1.3 Forward Rates

We use forward rates as explanatory variables to capture the yield curve information in our analysis, corresponding with Cochrane and Piazzesi (2005).

To compute the forward rates for both the Liu and Wu (2021) and the German yield dataset, we apply a standard methodology based on consecutive spot rates across different maturities. Following established financial literature (see, for instance, Campbell and Shiller, 1991; Fama and Bliss, 1987), we used the following formula to calculate the one-year forward rate at time t for an n-month maturity:

$$f_t^{(n)} = \frac{n}{12} y_t^{(n)} - \left(\frac{n}{12} - 1\right) y_t^{(n-12)}, \tag{3.1}$$

where $y_t^{(n)}$ represents the n-month spot rate at time t, and $f_t^{(n)}$ denotes the implied forward rate from month $n-12$ to month n. This approach isolates the one-year forward rate between consecutive maturities, allowing for a direct interpretation of yield expectations over specific time intervals.

For both datasets, we computed a series of forward rates up to a ten-year horizon. These covered maturities from one to ten years. This ensures a consistent comparison of the term structure across the U.S. and Germany. By generating these forward rates, we examined the dynamics of the yield curves. This allow us to explore the predictive power of forward rates for bond excess returns in different regions. See figure 3.1 for plot of forward rates of all maturities for both markets.

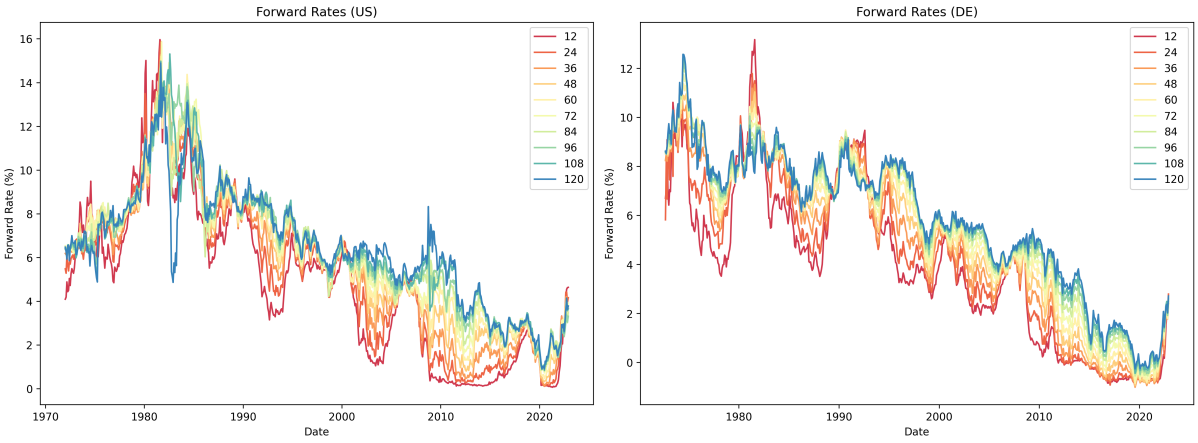


Figure 3.1: Plot of Forward Rates
US (left) & Germany (right) for the period 1972:01 - 2022:12.

3.1.4 Excess Return

The dependent variable for our analysis is excess bond returns. The excess return of holding a bond with n month maturity for 1 year is the return of holding this bond for 1 year minus the 1 year yield. Excess bond returns are widely used as a dependent variable in financial studies for three key reasons. First, they are measured in real terms, so adjustments for inflation or interest rates are unnecessary (Cochrane & Piazzesi, 2005). Second, they provide a clear benchmark through the Expectations Hypothesis, which is difficult to outperform. Third, they represent the risk premium earned by holding a bond for less than its full maturity.

To compute the one-year holding period excess return for a bond with n-month maturity ($n \geq 12$), we adopt the following approach, consistent with Bianchi et al. (2020), refined by Hoogteijling et al. (2021). Let $p_t^{(n)}$ denote the log price of a zero-coupon bond at time t (in months) with a payoff of 1 and maturity at $t+n$. The continuously compounded yield of this bond is given by $y_t^{(n)} = -\frac{12}{n} p_t^{(n)}$. The one-year log excess return for an n-month bond can then be expressed as:

$$\begin{aligned} \mathrm{xr}_{t:t+12}^{(n)} &= p_{t+12}^{(n-12)} - p_t^{(n)} - y_t^{(12)} \\ &= -\frac{n-12}{12} y_{t+12}^{(n-12)} + \frac{n}{12} y_t^{(n)} - y_t^{(12)} \\ &= -\left(\frac{n}{12} - 1\right)\left(y_{t+12}^{(n-12)} - y_t^{(n)}\right) + \left(y_t^{(n)} - y_t^{(12)}\right). \end{aligned} \tag{3.2}$$

where $\text{xr}_{t,t+12}^{(12n)}$ denotes the excess return on an n-year bond from time t to $t+12$ months. In this formula, $y_t^{(12n)}$ is the spot yield for an n-year maturity bond at time t, $y_{t+12}^{(12(n-1))}$ represents the yield on the $(n-1)$-year maturity bond observed one year later, and $y_t^{(12)}$ is the one-year yield, serving as the risk-free alternative.

This calculation was applied to both the Liu and Wu (2021) U.S. yield curve dataset and the Bundesbank yield dataset for Germany. We compute excess returns for the 24, 36, 48, 60, 84, and 120-month maturities to capture the predictive power of excess returns across a range of bond maturities. Including multiple maturities allows us to analyze how risk premia and return dynamics evolve over different time horizons. It also helps identify whether predictive relationships vary with bond maturity, providing a more comprehensive understanding of the term structure and excess return behavior. See figure 3.2 for plots of excess returns for all maturities for both markets.

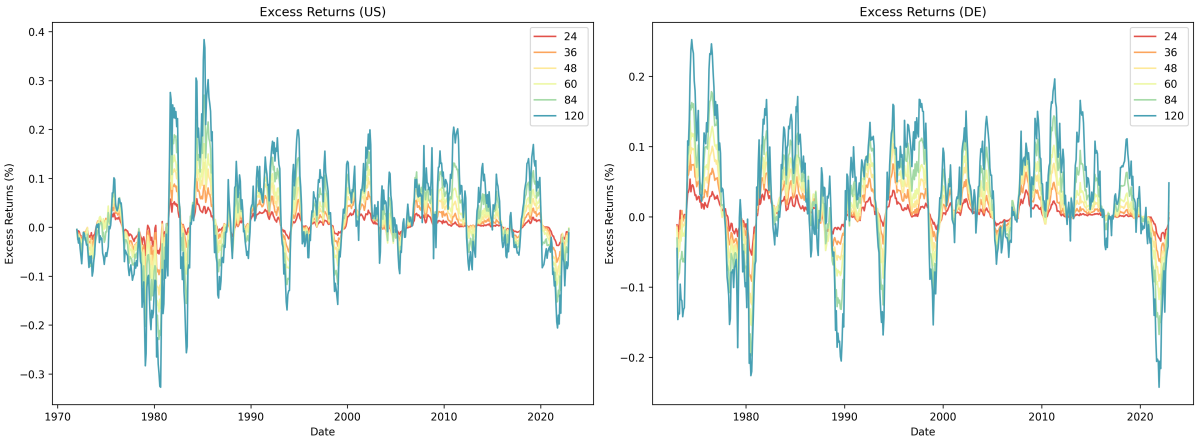


Figure 3.2: Plot of Excess Returns
US (left) & Germany (right) for the period 1972:01 - 2022:12.

3.2 Macroeconomic Data

3.2.1 US. Macroeconomic Data

For the U.S. market, we utilize the FRED-MD database, a comprehensive monthly dataset of 128 macroeconomic variables published by the Federal Reserve Bank of St. Louis (McCracken & Ng, 2016). The FRED-MD dataset includes a wide array of economic indicators, such as measures of economic activity, inflation, interest rates, and employment, among others. Each variable is provided at a monthly frequency and has been pre-processed

according to transformation codes recommended by McCracken and Ng (2016) to ensure stationarity, making it suitable for time-series analysis.

The dataset spans from January 1959 to the most recent available month; however, our analysis uses data starting in January 1972, aligning with the issuance of the 10-year bond. For specific details on the macroeconomic variables, including transformation methods and variable definitions, refer to Table A.2.

3.2.2 German Macroeconomic Data

For the German market, we use the EA-MD-QD dataset, a macroeconomic dataset for the euro area and its member countries designed specifically for research purposes (Barigozzi & Lissona, 2024). The EA-MD-QD dataset includes a total of 107 monthly and quarterly economic variables. This dataset is specifically designed to resemble the FRED-MD dataset for the U.S. market, allowing for a comprehensive analysis across different regions.

The dataset spans from January 2000 to the most recent available date. Each variable is provided at either a monthly or quarterly frequency. The data undergoes transformations to ensure stationarity, following the methodology outlined in McCracken and Ng, 2016, as detailed in the Table A.1.

For variables reported quarterly, we use linear interpolation to convert the data to a monthly frequency. This ensures consistent temporal granularity across the dataset. However, linear interpolation introduces a look-ahead bias. Future values influence the estimation of intermediate months, which can distort results. This bias is particularly problematic in out-of-sample testing, as interpolated values may include information unavailable in real-time. Additionally, interpolation may fail to capture shifts occurring between quarters, leading to missed signals in the data. This approach ensures uniformity for model training and comparison and the trade-offs and implications is carefully considered.

3.3 Splitting and out-of-sample data

For our data splitting, we use January 2014 to December 2022 as our out-of-sample period. As we calculate excess return for a 1-year holding period, we create a 12-month window between the training and testing, which means that 2013 is neither used for training nor

testing. Although this negatively impacts the performance of some of our models, we implement it this way to ensure that the model does not have training information on the true value of xr_{t+12} in month T (the last month in training) to gain information on xr_{t+11} in month $T+1$(the first month of testing). This is crucial because it is reasonable to believe that xr_{t+12} in month T would have predictive power over xr_{t+11} in month $T+1$, which again may explain excess returns for other maturities, including xr_{t+12} in month $T+1$, which is what we are trying to predict. As we use a 1-year holding period to calculate our excess returns, we create this gap of 1 year to ensure that the model doesn't have information that was not yet available at the time of the prediction.

For our neural networks, a separate validation set is created, to readjust some hyper-parameters like the learning-rate, or stop the training if the model starts over-fitting. The validation period is 2010 to 2012. In the pre processing, the validation set is treated like the test set, transformed with scalers fitted to the training set only, and with a gap to exclude overlapping maturities with the training data.

3.4 Stationarity

In our analysis of bond excess returns, we carefully considered the stationarity of our datasets. For macroeconomic variables, we applied appropriate transformations (detailed in the appendix) to ensure stationarity. However, for yield data, we made a deliberate choice to use forward rates $f_t(n)$ in their original form, despite their known non-stationary nature, characterized by a downward trend since the 1980s. This decision to retain non-stationarity in yield data has important implications:

1. Linear models and traditional regressions may underperform due to the presence of trends in the data.

2. The Recurrent Neural Network (RNN) model demonstrated a significant performance boost when leveraging these trends.

We conducted comparative tests using both stationary yield data ($\Delta f_t(n) = f_t(n) - f_{t-1}(n)$) and non-stationary data across all models. The results showed that the RNN's performance improvement when using non-stationary data was substantial and statistically significant. This finding justified our decision to maintain the original, non-stationary form of the

forward rates in our analysis. While this approach introduces challenges for some models, it allows us to capture and utilize long-term trends in yield data, potentially leading to more accurate predictions of bond excess returns, particularly with advanced machine learning techniques like RNNs.

4 Methodology

4.1 Economic Methodology

Studies by Thornton and Valente (2012) and Sarno et al. (2016) emphasize that statistical predictability of bond returns does not inherently translate into economic value for investors. To address this, we employ an economic methodology to evaluate the practical utility of our predictive models from an investor's perspective. This involves utilizing performance metrics such as Mean-Variance Utility and Power Utility framework to assess whether our models can enhance portfolio performance and provide meaningful economic value over the EH benchmark.

4.1.1 Mean-Variance Utility

The first economic methodology evaluates the economic value of bond return predictability by assessing the asset allocation decisions of a mean-variance investor. The investor dynamically allocates wealth between a k-period bond and a risk-free 1-year bond to maximize expected utility. This framework incorporates forecasts of bond excess returns from all the predictive models and evaluates their performance relative to the Expectations Hypothesis (EH) benchmark.

The investor's utility function is grounded in mean-variance preferences, balancing expected returns and risk. At each time t, the utility is expressed as:

$$U_t = \mathbb{E}_t \left[r_{p,t+1}^{(k)} \right] - 0.5 \cdot \gamma \cdot \mathrm{Var}_t \left[r_{p,t+1}^{(k)} \right], \tag{4.1}$$

where U_t is the utility at time t, $r_{p,t+1}^{(k)}$ represents the portfolio return for the k-period bond at time $t+1$, γ denotes the investor's coefficient of risk aversion, $\mathbb{E}_t$ is the conditional expectation based on information available at t, and Var_t is the conditional variance of the portfolio return.

The optimal portfolio weight $\omega_t^{(k)}$ allocated to the k-period bond is derived from Markowitz optimization, given by:

$$\omega_t^{(k)} = \frac{1}{\gamma} \cdot \frac{\hat{xr}_{t+1}^{(k)}}{\hat{\sigma}_{t+1}^2}, \tag{4.2}$$

where $\hat{xr}_{t+1}^{(k)}$ represents the forecasted excess return for the k-period bond, and $\hat{\sigma}_{t+1}^2$ is the forecasted variance of the bond's excess returns, estimated using a rolling window of past excess returns. The parameter γ reflects the investor's risk aversion, where a higher value implies more conservative investment behavior. To ensure realistic portfolio strategies, weights are constrained within $[-1, 2]$, limiting excessive shorting or leveraging and we set $\gamma = 5$. This approach follows practices in literature (Gargano et al., 2019; Thornton & Valente, 2012)

The portfolio return is calculated as:

$$r_{p,t+1}^{(k)} = y_t^{(1)} + \omega_t^{(k)} xr_{t+1}^{(k)}, \tag{4.3}$$

where $y_t^{(1)}$ denotes the yield on the one-year risk-free bond at time t, and $xr_{t+1}^{(k)}$ is the realized excess return for the k-period bond at time $t+1$. This formulation represents the return from a dynamically managed portfolio that optimally combines the risk-free bond and the k-period bond.

The economic value of the predictive models is assessed through the certainty equivalent return (CER), which represents the fixed return an investor would accept instead of the portfolio's actual return with its associated risk. CER is computed as:

$$\text{CER} = \mu_p - 0.5 \cdot \gamma \cdot \sigma_p^2, \tag{4.4}$$

where μ_p and σ_p^2 are the mean and variance of portfolio returns over the evaluation period, respectively. The CER gain is defined as the difference between the CER of the predictive models and the CER of the EH benchmark, capturing the economic value added by incorporating predictive information into portfolio decisions.

To statistically evaluate the CER gains, we employ the Diebold-Mariano (DM) test, comparing the time series of realized utilities from the predictive models against the EH benchmark. Realized utilities for each time t are derived from the portfolio returns and

used to calculate utility differences. The DM test assesses whether these differences are statistically significant, accounting for autocorrelation and heteroscedasticity using the Newey-West estimator (Newey & West, 1987).

4.1.2 Power Utility

Unlike the mean-variance utility framework, which assumes constant absolute risk aversion, the Power Utility framework models investor preferences under Constant Relative Risk Aversion (CRRA). The utility function for this investor is given by

$$U(w_t, xr_{t+1}) = \frac{\left[(1 - w_t)\exp(y_t^{(1)}) + w_t \exp\left(y_t^{(1)} + xr_{t+1}\right)\right]^{1-\gamma}}{1-\gamma}, \quad \gamma > 0, \tag{4.5}$$

where w_t is the weight on the risky bond, $y_t^{(1)}$ represents the one-year risk-free rate, xr_{t+1} is the excess return on the risky bond, and γ denotes the investor's relative risk aversion. This formulation captures the investor's trade-off between risk and return, reflecting wealth-dependent risk preferences under the log-normal assumption of excess returns.

Following Campbell and Viceira (1999), excess returns are assumed to be log-normally distributed, and the return on a portfolio with weight w_t allocated to the risky bond can be approximated as

$$R_{p,t+1} = 1 + y_t^{(1)} + w_t \cdot xr_{t+1} + \frac{1}{2} w_t \cdot \sigma^2_{t+1|t}, \tag{4.6}$$

where $\sigma^2_{t+1|t}$ represents the conditional variance of the excess return. This equation adjusts for the expected return and incorporates a variance term arising from the log-normal distribution of returns. Substituting this return into the CRRA utility function highlights the role of risk and return in shaping optimal portfolio choices.

Under the Power Utility framework, the investor's optimal weight on the single risky bond can be derived as

$$w_t^{(n)} = \frac{1}{\gamma\left(\sigma^{(n)}_{t+1|t}\right)^2}\left[\hat{xr}^{(n)}_{t+1} + \frac{\left(\sigma^{(n)}_{t+1|t}\right)^2}{2}\right], \tag{4.7}$$

where $\hat{xr}_{t+1}^{(n)}$ is the forecasted excess return on the risky bond, $\sigma_{t+1|t}^{(n)}$ is the conditional standard deviation of the excess return, and γ represents the investor's relative risk aversion. The term $\frac{\left(\sigma_{t+1|t}^{(n)}\right)^2}{2}$ captures the variance adjustment that arises due to the log-normal approximation of excess returns. This expression highlights how the optimal weight balances expected returns with the risk captured by the conditional variance, scaled by the degree of risk aversion.

The economic value of the predictive models under the Power Utility framework is assessed through the certainty equivalent return (CER), similar to the approach used in the Mean-Variance Utility framework. CER gains are calculated as the difference between the CER of the predictive models and the CER of the EH benchmark.

Statistical significance of the CER gains is evaluated using the Diebold-Mariano (DM) test, following the same methodology as in the Mean-Variance Utility section. This involves comparing the realized utilities under the predictive models to those of the EH benchmark while accounting for autocorrelation and heteroscedasticity using the Newey-West estimator.

The Power Utility framework provides a more flexible and realistic approach to portfolio optimization by incorporating wealth-dependent risk preferences and adjusting for the log-normal distribution of excess returns. The derived optimal weights reflect both the forecasted returns and the variability of the risky bond, scaled by the investor's risk aversion.

4.2 Technical Methodology

4.2.1 Regression models

Linear Regression

The multivariate linear regression is a linear model that predicts the value of a continuous response variable based on one or more predictor variables. The model can be expressed as:

$$Y = X\beta + \epsilon \tag{4.8}$$

where Y is the dependent variable, X is the matrix of independent variables, β is the vector of coefficients, and ϵ is the error term. The coefficients β are estimated by minimizing the sum of squared errors (SSE):

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \tag{4.9}$$

This optimization problem can be solved using ordinary least squares (OLS) method, which yields the closed-form solution:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4.10}$$

Ridge Regression

Ridge regression adds a penalty term to the OLS loss function to prevent overfitting. The penalty term is proportional to the magnitude of the coefficients. The Ritdge regression model is defined as:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4.11}$$

where λ is the regularization parameter that controls the strength of the penalty. The solution to this optimization problem is given by:

$$\hat{\beta}_{\text{Ridge}} = (X^T X + \lambda I)^{-1} X^T Y \tag{4.12}$$

where I is the identity matrix.

Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression uses an L1 penalty instead of the L2 penalty used in Ridge regression. The Lasso regression model is defined as:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.13}$$

The L1 penalty can shrink some coefficients to exactly zero, performing feature selection.

Elastic Net Regression

Elastic Net regression combines the L1 and L2 penalties of Lasso and Ridge regression, respectively. The Elastic Net regression model is defined as:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{4.14}$$

where λ_1 and λ_2 are the regularization parameters for the L1 and L2 penalties, respectively.

Elastic Net balances the benefits of both Lasso and Ridge regression, allowing for both feature selection and shrinkage of coefficients.

Principal Component Analysis (PCA)

PCA is a method for reducing the dimensionality of a dataset while retaining as much variability as possible. It involves transforming the original variables into new, orthogonal variables called principal components. Mathematically, this can be represented as:

$$X = TP^T + E \tag{4.15}$$

where X is the original data matrix, T is the score matrix, P is the loading matrix, and E is the residual matrix. The principal components are derived from the eigenvectors of the covariance matrix of X, sorted in descending order of their corresponding. We use 3 yield components (traditionally interpreted as level, slope and curvature) and 8 macroeconomic components.

Principal Component Regression (PCR)

PCR combines PCA with multivariate linear regression (MLR) to predict a quantitative

target variable. The process involves selecting a subset of the PCA components based on their explanatory power and conducting regression using the selected principal components as predictors.

The model can be expressed as:

$$Y = TC + E \tag{4.16}$$

where Y is the response variable, T is the score matrix from PCA, C is the coefficient matrix, and E is the error term. PCR reduces multicollinearity and noise by using principal components, but it weights predictor variables based on variance rather than their correlation with the target variable. When fitting the PCA model, the response variable is not considered at all.

Principal Least Squares (PLS)

PLS is a method that reduces the predictor variables to a smaller set of latent variables that maximize the covariance between the predictor and response variables. Unlike PCR, PLS models both the predictor matrix X and the response matrix Y simultaneously.

The PLS model can be represented as:

$$X = TP^T + E \tag{4.17}$$

$$Y = UQ^T + F \tag{4.18}$$

where T and U are the score matrices for X and Y, respectively, P and Q are the loading matrices, and E and F are the residual matrices. The PLS algorithm iteratively finds the directions w and c that maximize the covariance between X and Y.

The PLS regression model equation is given by:

$$Y = XW(P^TW)^{-1}C + F \tag{4.19}$$

PLS is particularly useful when dealing with high-dimensional data and correlated predictor variables, as it handles these issues more effectively than PCR by directly incorporating the covariance structure between X and Y.

While both PCR and PLS reduce dimensionality, PLS is more efficient in capturing the

relationship between the predictor and response variables due to its focus on maximizing covariance. PLS often requires fewer latent variables than PCR and can handle nonlinear relationships better. However, PCR is simpler to interpret and can be more robust in certain scenarios where the primary concern is reducing noise and multicollinearity.

4.2.2 Random Forest Regressor

At the core of a Random Forest Regressor are decision trees, which are simple yet effective models for both classification and regression tasks. A decision tree works by recursively partitioning the data into smaller subsets based on the values of the input features. Each internal node in the tree represents a feature or attribute, and each leaf node represents a class label or a predicted value.

If we have B decision trees $T_1, T_2, \ldots, T_B$, the final prediction $\hat{y}$ for a new input x is:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

4.2.3 Neural Networks

Neural networks (NNs) are computational models inspired by the architecture and function of biological neural systems (Hinton, 1990). They are structured as layers of interconnected nodes or "neurons" that process data by passing inputs through successive transformations, each of which is guided by a set of weighted connections. These weights are learned during training, allowing the network to adjust based on the data it encounters. NNs are particularly valued for their ability to capture complex, non-linear relationships, which makes them well-suited to tasks like financial prediction, where the relationships among variables are often intricate and non-linear (Goodfellow et al., 2016).

In the context of bond excess return forecasting, the architecture of the neural networks has often been variants of feedforward dense neural networks. We propose that Recurrent neural nets (RNN) could be able to explain more of the excess returns due to their ability to capture more of the temporal structures in sequential data. We assume that value at any given time point can be influenced by its historical values. Dense networks, while powerful in capturing complex relationships, treat inputs independently and do not

inherently account for the temporal order of data. This limitation can lead to suboptimal performance when modeling sequences where past information is crucial. RNNs maintain a hidden state that is updated at each time step, allowing them to retain information about previous inputs.

An RNN processes sequences by iterating through the time steps of the input sequence while maintaining a hidden state vector h_t.

The **hidden state** at time t is computed as:

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b) \tag{4.20}$$

where W_h and W_x are weight matrices, b is a bias vector, and σ is an activation function such as tanh or ReLU. Despite their strengths, standard RNNs struggle with long-term dependencies due to issues like vanishing gradients. LSTMs (Long-Short Term Memory) address this by introducing memory cells and gating mechanisms that regulate the flow of information.

The **forget gate** decides what information to discard from the cell state:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4.21}$$

The **input gate** determines what new information to store in the cell state:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4.22}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4.23}$$

The cell state is updated by combining the old cell state and the new candidate values:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{4.24}$$

Where the operator $\odot$ is the element-wise product (Hadamard product).

The **output gate** determines the output based on the cell state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4.25}$$

$$h_t = o_t \odot \tanh(C_t) \tag{4.26}$$

These mechanisms allow LSTMs to effectively capture long-range dependencies by controlling what information is preserved or discarded over time.

The model initiates by setting the initial hidden state `h0` and cell state `c0` to zero, which are essential for the LSTM layer to start processing the input sequence. The input sequence `x` is then fed into the two LSTM layers. During forward propagation, the input sequence is processed by the LSTM layers, and only the output from the last time step is used for further processing through the fully connected layer. The fully connected layer is incorporated to transform the high-dimensional output into the predictions for time T+1, and mitigate risk of overfitting by adding another step of dropout. The final output layer is designed to produce a specific number of outputs corresponding to the predicted excess returns for different maturities.

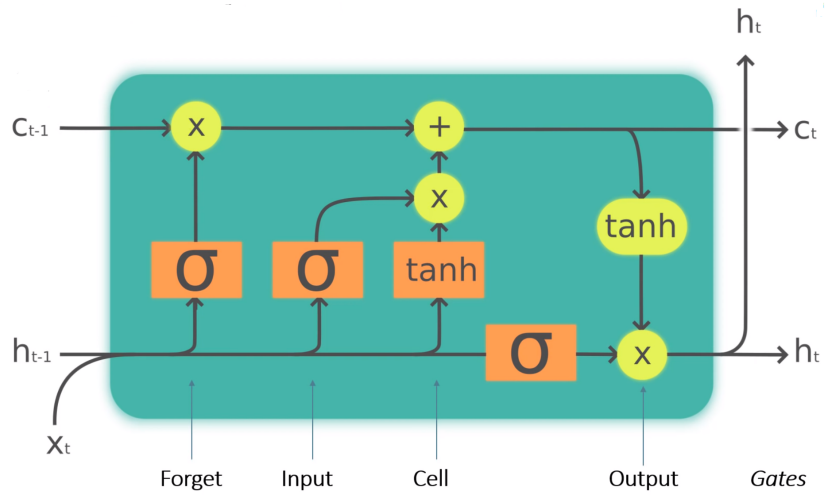


Figure 4.1: LSTM cell
Structure of an LSTM cell retrieved from Jain (2019)

The RNN model is trained using PyTorch's (Paszke et al., 2019) implementation of the Adam (Adaptive Moment Estimation) optimizer with a learning rate starting at $1e-3$,

and a scheduler that adjusts the learning rate once the learning stagnates . As our problem is one of forecasting, we use Mean Squared Error (MSE) as the loss function, and add some L2 penalty. The training loop involves multiple epochs, where in each epoch, the model processes all the training data in batches. During the Adam optimizer updates the model parameters based on the gradients of the loss with respect to the parameters.

$$\text{Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.27}$$

The **Adam update rule** is:

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{4.28}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of samples, θ are the model parameters, α is the learning rate, m_t and v_t are the first and second moments of the gradients, and ϵ is a small constant to prevent division by zero.

4.2.4 Performance measures and statistical testing

The predictive performance of the models is evaluated using the out-of-sample R^2 measure (R^2_{oos}), as proposed by Campbell and Thompson (2008). This metric compares the prediction accuracy of a model against a EH benchmark, calculated as the historical mean of the dependent variable. The formula for R^2_{oos} is:

$$R^2_{\text{oos}} = 1 - \frac{\sum_{t=0}^{T-1}\left(xr_{t+1}^{(n)} - \hat{xr}_{t+1}^{(n)}\right)^2}{\sum_{t=0}^{T-1}\left(xr_{t+1}^{(n)} - \bar{xr}_{t+1}^{(n)}\right)^2}, \tag{4.29}$$

where $xr_{t+1}^{(n)}$ is the realized bond excess return with maturity n over the period $t+1$, $\hat{xr}_{t+1}^{(n)}$ is the model's predicted return, and $\bar{xr}_{t+1}^{(n)}$ is the EH benchmark forecast based on the historical mean of returns calculated up to $t-1$. The variable T represents the number of periods in the test sample. A positive R^2_{oos} indicates better model performance compared to the benchmark, while a negative value suggests otherwise.

To test the significance of R^2_{oos}, the null hypothesis $R^2_{\text{oos}} \leq 0$ is tested against the alternative

$R^2_{\text{oos}} > 0$, indicating whether the predictive model significantly outperforms the benchmark. The MSPE-adjusted statistic by Clark and West (2007) is used, addressing bias in nested model comparisons. The test compares the squared prediction errors of the model and the benchmark, with an adjustment factor to account for their overlap. Significance is determined using a standard normal distribution, ensuring robust evaluation of predictive performance.

4.3 Software and Computation

The computational framework for this thesis involved a combination of programming languages and libraries, including Python, R, PyTorch, SciKit-Learn, NumPy, Pandas, and visualization tools such as Matplotlib and Seaborn. The models were trained on Google Colab, which provided access to GPU resources to accelerate the training and hyper-parameter search (Grid-search).

To effectively train our models on time series data, we utilized a time series split strategy. This approach is essential in financial applications where data points are temporally ordered. Specifically, we employed the 'TimeSeriesSplit' function from SciKit-Learn(Pedregosa et al., 2011), which allows us to create training and testing datasets while respecting the temporal order of observations. This method ensures that our model is trained on past data to predict future outcomes without any data leakage.

In our implementation, we refit the model iteratively to allow it to utilize the largest dataset available at each time. After training the model to predict xr_{t+12}, we appended the corresponding X and y values to the dataset as for subsequent predictions. This technique was particularly beneficial for the German dataset, where we had less historical data. By frequently refitting the model with updated predictions, we enhanced its performance and adaptability to new information.

We established a gap of 12 months between the training and testing datasets. This gap was crucial in ensuring that the model did not have access to future information when making predictions, thereby simulating a more realistic forecasting scenario. The training process involved several iterations where the model was retrained with new data points as they became available.

The overall training procedure can be summarized as follows:

1. Data Preparation: Calculations of excess returns, forward rates and transformation and standarization of macroeconomic features. The datasets were merged and cleaned to remove any missing values. Macroeconomic data is shifted one month, so the data reflects last month's values.

2. Time Series Split: Using 'TimeSeriesSplit', we created multiple train-test splits while maintaining the temporal order of data.

3. Model fitting and prediction: The model is fitted on the training set which includes true values until xr_{t-12}, and the prediction is then made based on last month's macroeconomic numbers, and current forward rates. The model is then refitted to include xr_{t-13} which was previously dropped from the training set. This method implies that the last 11 months of macroeconomic data and yield data is also dropped in the training set. To omit this we could have utilized a rolling window. We initially implemented a rolling window as suggested in (James et al., 2013) but the performance did not significantly increase although the computation became far heavier, especially for models like the RFR.

For the neural network the process is slightly different. As the model updates the hidden and cell state continuously, the previous timestep might include crucial information, and we therefore implement a sliding window. Data is managed through a PyTorch DataLoader. In the training set, variables $X[t : t + 12]$ corresponds to xr_{t+12}, which effectively drops the first 12 values of xr. The same is done for the validation set and testing set, solving the problem of look-ahead bias without dropping the months leading up to the prediction.

5 Analysis

5.1 Factor loadings

We begin by examining the different macroeconomic factor-loadings found in our PCA analysis. We are interested in parameters that either has a strong positive loading, or strong negative loading when constructing the first component. The factor loadings tells us how the variables are weighted when we reduce the dimensionality of our dataset. For the U.S., we find that the labor market contains many of the largest positive factors, along with some output/income variables. The same is true for the negative factors, where the unemployment rate has large negative values, along with the VIX rate, as well as Moody's Baa-Fed funds spread.

We find that stock market variables, such as S&P stocks' dividend yield or P/E does not score among the important variables.

Table 5.1: US factor loadings - Positive Loadings

Factor	Loading score
PAYEMS	0.187607
USGOOD	0.184982
MANEMP	0.176989
IPMANSICS	0.176909
DMANEMP	0.173689
INDPRO	0.170629
USTPU	0.169967
IPFPNSS	0.168639
CUMFNS	0.161284
IPDMAT	0.161270

Table 5.2: US factor loadings - Negative Loadings

Factor	Loading score
UNRATE	-0.127885
UEMP15OV	-0.113081
UEMP27OV	-0.092765
VIXCLSx	-0.090755
ISRATIOx	-0.082571
CLAIMSx	-0.082309
UEMP15T26	-0.072426
UEMP5TO14	-0.064106
BAAFFM	-0.061386
UEMPMEAN	-0.050601

Although employment, or unemployment, seems to be an important variable for Germany as well, we note that 'Economic sentiment indicator' (ESENTIX) tops the table for german factor loadings. On the other end of the scale, 'Self employment' has the highest negative factor loading, followed by other employment scales and long term debt, both as an liability(TLB.LDB) and an asset (TASS.LDB).

Table 5.3: German factor loadings - Positive Loadings

Factor	Loading score
ESENTIX	0.1856
BCI	0.1752
ICONFIX	0.1748
GFCF	0.1733
GDP	0.1728
THOURS	0.1713
SCONFIX	0.1695
TEMP	0.1694
EMP	0.1660
TRNING	0.1630

Table 5.4: German factor loadings - Negative Loadings

Factor	Loading score
SEMP	-0.0948
EMPFC	-0.0882
TLB.LDB	-0.0543
ULCCON	-0.0408
TASS.LDB	-0.0351
UNEO25	-0.0348
NFCL.LLN	-0.0341
UNETOT	-0.0306
EMPPA	-0.0298
CONSD	-0.0294

As PCA does not include any information about the response variable, the loading scores does not tell us anything about how the variables are correlated with Y. These are the coefficients that represent the correlation between the original variables and the principal components. The magnitude of the loading factors indicates the strength of the relationship between the variable and the first principal component. The sign (positive or negative) indicates the direction of the relationship.

As for forward rates, we find the same factor loadings in both markets, the short rates have the largest factor loading in $PC1$ and $PC2$, while the long rates explains $PC3$.

5.2 Out-of-sample performance

The R^2_{OOS} values provide valuable insights into the predictive performance of our models for excess bond returns in both the U.S. and German markets, compared to an EH benchmark. The Recurrent Neural Network (RNN) model consistently outperforms other models across all maturities for both markets. In the U.S. market, the RNN achieves the highest R^2_{OOS} values, ranging from 16.17% for 2-year bonds to 35.37% for 10-year bonds. Similarly, in the German market, the RNN shows superior performance, especially for medium to long-term maturities.

However, we see that the superior performance drastically vanishes when excluding 2020-2022 (see table B.2 in Appendix). The EH benchmark is particularly good before 2020 as the interest rates worldwide was consistently abnormally low, and it is likewise particularly

bad in Covid, when sudden changes made excess return far more volatile. Table 5.5 shows the R^2_{OOS} predictive performance over EH in the complete testing window 2014-2022.

A notable pattern across both markets is the tendency for predictability to improve for longer maturities. This trend is particularly pronounced for the RNN model, aligning with previous research suggesting that non-linear machine learning methods can capture more complex relationships in longer-term bond returns.

US bond market

For the classic coefficient of determination R^2 metric, as implemented in Scikit-learn's 'r2_score', PCA linear regression yields negative R^2 values for all maturities, but still beats the benchmark for longer maturities. After showcasing the factors, we might suspect that even though the PCA factors explain some of the variance in the X dataset, many of the factors are quite similar, and does not necessarily translate well to the variance in our Y variable. We report classic R^2 scores in Appendix B.1.

In contrast, PLS regression generates some of the best results for all maturities. PLS is similar to PCA, but aims to explain the variance (and correlation) with the Y variable as well. PLS is in fact the top performing models of linear combinations. Findings also align with those of Bianchi et al. (2021) that machine learning methods generally perform better for longer maturities. PLS is the second-best performer in our Out-of-sample test.

The RNN outperforms the other forecasting methods by a significant margin. We also note that most of our findings outperform the findings presented in earlier studies on the U.S. bond market. However, as we use different out-of-sample periods, we also tested the model in an out-of-sample period ending in 2019 to exclude the effects of COVID-19. We find that while our classic in-sample R^2 improves, our R^2_{OOS} (as introduced in 4.2.4) does the opposite. The neural network is beaten by the benchmark for the short maturities, but does still outperform for medium and long maturities.

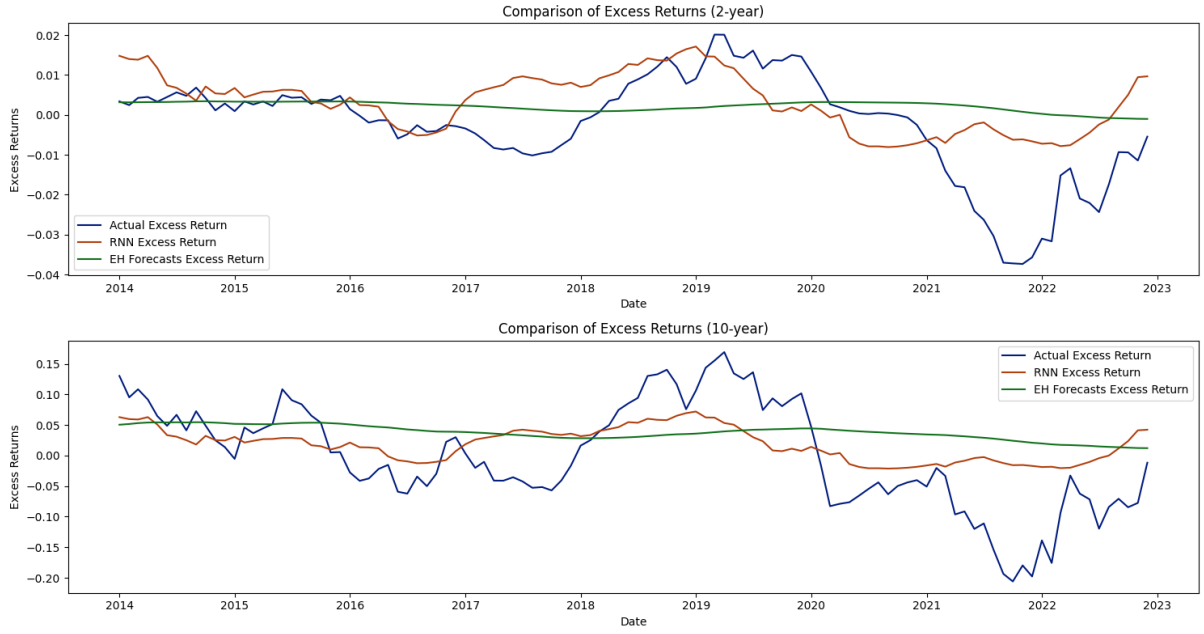


Figure 5.1: Plot of Actual , EH forecasts and RNN Predictions Excess Returns 2-year (over) & 10-year (under) for the U.S in the out-of-sample period 2014:01 - 2022:12.

German bond market

In the German market, the RNN model again demonstrates superior performance, especially for medium to long-term maturities, with R^2_{OOS} values ranging from 12.63% for 3-year bonds to 22.22% for 7-year bonds. Interestingly, the Elastic Net model performs well in the German market, outperforming PLS and PCR, particularly for longer maturities. The RFR model shows consistent positive performance across all maturities in the German market, unlike its performance in the U.S. market.

Interestingly, the Elastic Net model performs well in the German market, outperforming PLS and PCR, particularly for longer maturities. The Elastic Net achieves R^2_{OOS} values of 18.86% for 7-year bonds and 16.01% for 10-year bonds, both significant at the 1% level. This suggests that the Elastic Net's ability to extract a limited feature selection is particularly valuable in the German context.

We find the same trend as in the U.S. market for prediction accuracy when ending the out-of-sample window early. The R^2_{OOS} deteriorates, and we can not beat a moving average for all maturities, despite the coefficient of determination getting stronger.

Table 5.5: R^2_{OOS} values for different models predicting xr_{t+1}

Models	$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(10)}_{t+1}$
Panel A: U.S.						
RNN	16.17%***	19.51%***	28.73%***	30.88%***	33.88%***	35.37%***
	(0.003)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
Elastic Net	4.39%**	5.99%**	10.01%***	9.44%***	9.19%***	3.77%**
	(0.0178)	(0.0117)	(0.009)	(0.009)	(0.008)	(0.013)
PCR 8+3	−10.71%**	0.07%**	7.83%**	8.16%**	8.51%**	6.11%**
	(0.033)	(0.022)	(0.019)	(0.021)	(0.017)	(0.011)
PLS	−1.76%*	10.03%**	16.80%**	19.73%**	22.04%***	23.93%***
	(0.094)	(0.034)	(0.020)	(0.016)	(0.01)	(0.005)
RFR	−16.74%	−7.97%	−2.85%	−0.26%	3.49%*	8.34%**
	(0.613)	(0.376)	(0.198)	(0.131)	(0.077)	(0.037)
Panel B: Germany						
RNN	3.56%	12.63%***	18.15%***	22.18%***	22.22%***	19.01%***
	(0.800)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)
Elastic Net	−12.17%	1.26%***	10.41%***	15.63%***	18.86%***	16.01%***
	(0.107)	(0.007)	(0.001)	(0.000)	(0.000)	(0.000)
PCR 8+3	−29.62%	−16.43%***	−8.16%***	−3.53%***	1.72%*	3.27%
	(0.987)	(0.001)	(0.001)	(0.009)	(0.092)	(0.876)
PLS	−95.55%	−78.10%	−64.04%	−53.77%	−40.63%	−32.15%
	(0.808)	(0.675)	(0.578)	(0.533)	(0.533)	(0.614)
RFR	0.57%*	6.57%***	10.61%***	12.90%***	14.40%***	12.62%***
	(0.066)	(0.003)	(0.000)	(0.000)	(0.000)	(0.001)

This table reports out-of-sample R^2 values with the models for the U.S. market shown in Panel A and the models for German market shown in Panel B. The out-of-sample period start in January 2014 and ends in December 2022. We measure statistical significance relative to the expectation hypothesis model using the Clark and West (2007) test statistic. * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

5.3 Mean-Variance Utility

This section examines the economic value of predictive models through a mean-variance utility framework. When an investor uses the mean-variance utility framework, they construct a portfolio that maximizes expected return for a given level of risk or minimizes risk for a desired return. The certainty equivalent return (CER) measures the risk-adjusted performance, reflecting the guaranteed return an investor would accept instead of uncertain outcomes. Positive CER changes, compared to the benchmark, Expectations Hypothesis (EH), indicate improved risk-return trade-offs, with statistical significance confirming these improvements are not due to chance.

Table 5.6: Mean-Variance Utility

Models	$xr_{t+1}^{(2)}$	$xr_{t+1}^{(3)}$	$xr_{t+1}^{(4)}$	$xr_{t+1}^{(5)}$	$xr_{t+1}^{(7)}$	$xr_{t+1}^{(10)}$
Panel A: U.S.						
RNN	0.12%***	0.27%***	0.30%***	0.32%***	0.39%***	0.43%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Elastic Net	0.10%***	0.19%***	0.25%***	0.31%***	0.44%***	0.42%**
	(0.005)	(0.004)	(0.008)	(0.005)	(0.007)	(0.040)
PCR 8+3	0.06%*	0.11%*	0.12%	0.14%	0.21%	0.08%
	(0.050)	(0.073)	(0.224)	(0.255)	(0.348)	(0.939)
PLS	0.07%***	0.13%**	0.15%*	0.17%	0.27%	0.24%
	(0.008)	(0.011)	(0.085)	(0.162)	(0.198)	(0.468)
RFR	0.03%	0.04%	0.05%	0.04%	0.05%	−0.00%
	(0.237)	(0.191)	(0.270)	(0.482)	(0.582)	(0.794)
Panel B: Germany						
RNN	−0.08%***	−0.09%***	−0.18%***	−0.15%***	−0.45%***	−0.53%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Elastic Net	−0.11%***	−0.22%***	−0.25%***	−0.27%***	−0.29%***	−0.33%**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PCR 8+3	−0.11%***	−0.26%***	−0.34%***	−0.37%***	−0.41%***	−0.43%***
	(0.001)	(0.000)	(0.001)	(0.001)	(0.003)	(0.010)
PLS	−0.17%***	−0.32%***	−0.41%***	−0.47%***	−0.58%***	−0.76%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
RFR	−0.03%***	−0.05%**	−0.06%*	−0.07%*	−0.06%	−0.03%
	(0.001)	(0.017)	(0.050)	(0.086)	(0.195)	(0.359)

This table report annualized percentage change in certainty equivalent return for mean-variance utility investors for the U.S. market (Panel A) and the German market (Panel B) with a risk aversion coefficient of 5 and constraints for the weigths between -1 and 2. Statistical significance is based on a one-sided Diebold-Mariano test applied to the out-of-sample period 2014-2022. * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

Panel A: U.S.

For the U.S. models the RNN and Elastic Net models stand out as strong performers, with consistently higher CER values across most maturities compared to the EH benchmark. Notably, while PLS performed well in previous evaluations, shows weaker results in this setting, with lower and not significant CER values across the maturities.

For RNN and Elastic Net the CER gains tend to increase with bond maturity. This aligns with those of Bianchi et al. (2021), suggesting that machine learning methods perform better for longer maturities. This may also indicate that their ability to capture non-linear patterns is more valuable at longer maturities.

The performance of PCR, PLS, and RFR reveals notable differences in predictive ability. PLS outperforms the other two models, delivering higher CER values, particularly for

shorter horizons with statistically significant results. However, its performance declines for longer horizons as it is not able to maintain it's statistical significance. PCR, while showing some significance for shorter periods, achieves lower utility values overall and struggles as the prediction horizon increases. RFR performs the weakest, with consistently low and statistically insignificant results across all horizons, indicating limited predictive capability in this context.

Overall, the results emphasize that RNNs seem to offer economically significant improvements over the EH benchmark and performs generally better than the other models.

Panel B: Germany

In clear contrast to the U.S. results, all predictive models examined here fail to deliver positive economic value. Specifically, the RNN, Elastic Net, PLS and PCR models produce negative and statistically significant CER values across all examined maturities. This pattern indicates that, despite their ability to achieve positive results in the U.S., these methods do not uncover economically meaningful predictive information in the German bond market. The RFR model, although generating less negative CER values than the other methods, does not offer statistically significant improvements and thus cannot be considered a viable enhancement.

Together, these results imply that none of the predictive models tested provide any meaningful improvement over the simple EH benchmark in the German setting.

5.4 Power Utility

This section examines the economic value of predictive models through a power utility framework. When an investor uses the power utility framework, they account for risk preferences in a non-linear manner, allowing for varying degrees of risk aversion across the entire distribution of returns. The certainty equivalent return (CER) measures the risk-adjusted performance, reflecting the guaranteed return an investor would accept instead of uncertain outcomes. Positive CER changes, compared to the benchmark, Expectations Hypothesis (EH), indicate improved risk-return trade-offs, demonstrating that the predictive model enhances outcomes for different levels of risk aversion.

Table 5.7: Power Utility

Models	$xr_{t+1}^{(2)}$	$xr_{t+1}^{(3)}$	$xr_{t+1}^{(4)}$	$xr_{t+1}^{(5)}$	$xr_{t+1}^{(7)}$	$xr_{t+1}^{(10)}$
Panel A: U.S.						
RNN	0.82%***	2.20%***	3.71%***	4.86%***	6.86%***	8.48%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Elastic Net	0.57%**	1.81%***	2.90%***	3.69%***	5.09%***	5.44%***
	(0.031)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)
PCR 8+3	0.36%	1.58%***	2.67%***	3.52%***	4.61%***	4.92%**
	(0.205)	(0.006)	(0.003)	(0.002)	(0.004)	(0.019)
PLS	0.36%	1.34%**	2.53%***	3.44%***	4.93%***	5.85%***
	(0.179)	(0.018)	(0.005)	(0.004)	(0.003)	(0.005)
RFR	−0.05%	0.72%**	1.49%**	2.20%***	3.17%***	3.83%***
	(0.801)	(0.074)	(0.018)	(0.009)	(0.005)	(0.006)
Panel B: Germany						
RNN	−0.35%*	−0.36%***	−0.05%	−0.02%	0.68%***	6.28%***
	(0.084)	(0.001)	(0.124)	(0.422)	(0.175)	(0.002)
Elastic Net	−0.05%	0.76%*	1.57%***	1.95%***	2.75%***	3.84%**
	(0.866)	(0.063)	(0.004)	(0.002)	(0.004)	(0.011)
PCR 8+3	0.23%	0.68%	1.13%*	1.28%*	1.82%*	2.02%
	(0.309)	(0.108)	(0.072)	(0.082)	(0.060)	(0.191)
PLS	−0.63%***	−0.63%***	−0.28%	−0.02%	0.10%	−3.49%**
	(0.000)	(0.000)	(0.128)	(0.904)	(0.736)	(0.034)
RFR	0.03%	0.62%**	1.29%***	1.71%***	2.89%***	5.27%***
	(0.832)	(0.020)	(0.003)	(0.002)	(0.002)	(0.001)

This table report annualized percentage change in certainty equivalent return for power utility investors for the U.S. market (Panel A) and the German market (Panel B) with a risk aversion coefficient of 5 and constraints for the weigths between -1 and 2. Statistical significance is based on a one-sided Diebold-Mariano test applied to the out-of-sample period 2014-2022. * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

Panel A: U.S.

The RNN model again consistently delivers the best performance across all maturities, achieving the highest CER values at every maturitiy with statistically significant results at the 1% level. At shorter maturities such as the 2-year bond, the RNN achieves a CER improvement of 0.82%, which increases steadily with longer maturities, reaching 8.48% at the 10-year horizon. The Elastic Net model also shows strong results, with statistically significant CER improvements across all maturities. However, its performance consistently lags behind the RNN model, with CER values of 0.57% at the 2-year maturity and 5.44% at the 10-year maturity.

The simpler models, such as PCR and PLS, provide modest CER improvements, particularly at intermediate maturities, but their performance is less competitive. For

example, PCR achieves a CER of 4.92% at the 10-year maturity, which is lower than both RNN and Elastic Net. The RFR model delivers the lowest CER values across all maturities, with improvements reaching only 3.83% at the 10-year horizon.

In conclusion, the results for the U.S. market suggest that RNN may offer advantages over the EH benchmark and the other models. The model show promising improvements, particularly at longer maturities. Among all of the models, the RNN appears to be the most effective, with potential to provide meaningful economic value to investors, under the power utility framework, by improving bond return forecasts.

Panel B: Germany

Unlike in the U.S., where certain models consistently yielded positive CERs across all maturities, this pattern does not hold for the German market.

At shorter horizons, all of the models fail to add value, delivering mostly negative and mostly not statistically significant CER gains. This underperformance likely reflects the complexity and noise in short-term yield movements, where transient factors and short-lived market shocks dominate. Predictive models may struggle to isolate reliable signals among these rapid fluctuations, reducing their ability to outperform a simple benchmark at the short end of the curve.

By contrast, as maturities lengthen, the RNN delivers increasingly positive and significant results. It moves from negative outcomes at 2-, 3-,4- and 5-year horizons to becoming a top performer at the 10-year maturity. Elastic Net and the simpler RFR model show strong, statistically significant gains at intermediate and longer horizons. This suggests that persistent macroeconomic and term structure signals are better captured at longer maturities. In comparison, PCR rarely achieves significance. PLS consistently fails to generate improvements. This indicates that not all approaches adapt equally well to the German yield environment.

In contrast to the U.S. results, Germany's predictive landscape appears more challenging, and no single model dominates the entire term structure. Instead, the utility enhancements materialize primarily at longer horizons, where stable, long-term signals overshadow short-term noise.

In summary, these findings suggest that while certain predictive models can extract

valuable information, their benefits largely emerge over extended time horizons. Short-term maturities seem to be difficult to forecast.

5.5 Additional Results Economic Value

In addition to the benchmark case, which constrains the weights to the range $-1 \leq w_t^{(n)} \leq 2$ to avoid extreme investments, we also examine an alternative scenario. This scenario restricts the optimal weights to non-negative values, specifically $w_t^{(n)} \in (0, 0.99)$. These constraints ensure that the expected utility remains finite, even in cases of unbounded returns (see Geweke (2001) for further details).

The findings from this additional analysis (see Appendix C) further reinforce the results drawn from the main results, providing stronger support for the overall insights.

6 Discussion

In this section, we discuss our findings in relation to the hypotheses outlined in this study. First, we hypothesized that Recurrent Neural Networks would outperform traditional methods in predicting bond excess returns. Second, we hypothesized that Recurrent Neural Networks adapt better to challenging market conditions, such as the onset of COVID-19. Third, we hypothesized that this predictive ability, demonstrated in the U.S. market, would extend to other developed markets, specifically the German bond market.

We evaluate our results within the context of these hypotheses, examining the relative performance of RNNs across different maturities and market conditions. Additionally, we address potential limitations and the broader implications of our findings for bond return prediction and cross-market generalization.

6.1 Hypothesis 1 - RNN Performance

Our first hypothesis was that the RNN model would be able to outperform traditional models by capturing sequential nonlinear combinations between macroecononmic data, forward rates, and excess return. Our out-of-sample tests confirm that recurrent neural networks, convincingly outperform the EH benchmark. The results for R^2_{OOS} are statistically significant for almost all maturities in both markets.

The economic utility analysis for the U.S. market supports the statistical findings. The RNN achieves the highest certainty equivalent return (CER) improvements in both mean-variance and power utility frameworks. Under the mean-variance framework, CER improvements for the RNN range from 0.12% for 2-year bonds to 0.43% for 10-year bonds, all statistically significant at the 1% level. Under the power utility framework, the RNN further solidifies its dominance, with CER values increasing steadily from 0.82% at 2 years to 8.48% at 10 years. These results demonstrate the RNN's ability to provide risk-adjusted returns that significantly outperform the EH benchmark, particularly at longer maturities. Elastic Net also delivers meaningful economic value, especially at intermediate maturities, though its performance consistently lags behind the RNN. Overall, the U.S. market highlights the strength of machine learning models, with the RNN yielding promising results both statistically and economically.

Structural breaks, such as the onset of the COVID-19 pandemic, represent abrupt and significant changes in economic conditions that challenge the assumptions underpinning many financial models. These breaks disrupt established relationships between macroeconomic variables and financial returns, creating a testing ground for models like the RNN and the EH benchmark.

The pandemic induced sharp shifts in market dynamics, such as sudden changes in bond yields and rapid government interventions. In the U.S., these structural breaks highlighted the strength of the RNN's adaptability. By dynamically adjusting to changing conditions, the RNN was able to recognize emerging patterns, such as increased risk in specific asset classes, and reallocate capital effectively. This capability allowed the RNN to mitigate losses during the pandemic, as illustrated in figure 6.1.

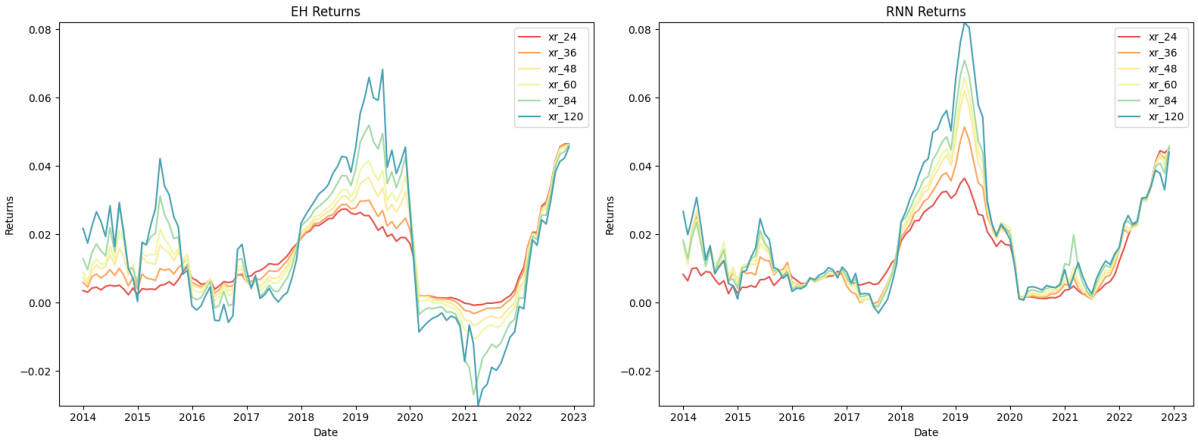


Figure 6.1: Plot of Returns from Mean-Variance Utility Portfolio - U.S. EH (left) & RNN (right) for the out-of-sample period 2014:01 - 2022:12.

When we end the OOS period before Covid, the RNN underperforms compared with earlier studies. This might suggest that dense networks with expanding windows are able to capture as much of the temporal change as our Recurrent model does. However, other models has often used more complex ensamble methods, including training separate networks for groups of macroeconomic variables and yields. Earlier studies also usually include one or more recession periods in the out-of-sample period. We chose to feed the network all the variables in the same input layer, with the hope of capturing complex temporal structures.

To summarize, the model shows superior performance in the U.S. in our out-of-sample window. It also better adapts to the rapid changes in the wake of COVID-19. In a period

without any structural brakes, it shows strong performance for long maturities, but fail to outperform at shorter ones.

6.2 Hypothesis 2 - Cross-Market Generalizability of RNN

Our second hypothesis proposed that if the RNN model outperforms the Expectations Hypothesis (EH) benchmark in the U.S. market, it should exhibit similar results in the German market.

The combined analysis of R^2_{OOS} and economic utility results highlights the Recurrent Neural Network (RNN) as the best performer in both markets, although with varying degrees of success. In the U.S., the RNN consistently delivers strong statistical and economic results across all maturities, making it the most robust model for forecasting bond returns for our analysis. These findings align with the results of Bianchi et al. (2021), who also identified neural networks as superior performers for capturing the complex dynamics of bond return predictability in the U.S. market.

In Germany, the RNN shows potential at longer maturities but struggles at shorter horizons. Some of the explanation might lie with the explanation power domestic macro variables has in the US compared to an European country. In fact, when we performed a principal component analysis on the whole Euro-zone, the results were more similar to the U.S. with regards to factor loadings in the first component. We also have to consider the shortened training window used for the german bond market, as well as the inaccuracies that might arise when we interpolate the macroeconomic data to convert many of the macroeconomic variables from quarterly to monthly. We could have chosen to limit the US data sources to match the ones we managed to obtain from the german market, but to make the results as comparable to previous findings as possible, we chose to proceed with all the data.

The economic utility results for Germany further illustrate the challenges of forecasting in this market. Under the mean-variance utility framework, all models, including the RNN, fail to deliver positive CER values. The RNN's CER improvements are negative across all maturities, ranging from -0.08% for 2-year bonds to -0.53% for 10-year bonds, indicating

that its statistical performance does not translate into economic value in this context. In the power utility framework, the RNN exhibits more promise at longer maturities, achieving positive CER values, such as 6.28% at 10 years. However, these gains are not consistent, and short-term predictions remain negative. Elastic Net and RFR display similar patterns, with occasional positive CER values at longer horizons, but their overall contributions to economic utility are limited. These results suggest that the structural differences in the German market, such as its distinct yield curve dynamics and challenging interest rate environment, hinder the effectiveness of predictive models.

In summary, the RNN model shows some of the same trends for both markets, significantly improving for longer maturities. However, there were clear differences in performance for the two markets, both in statistical and economical terms.

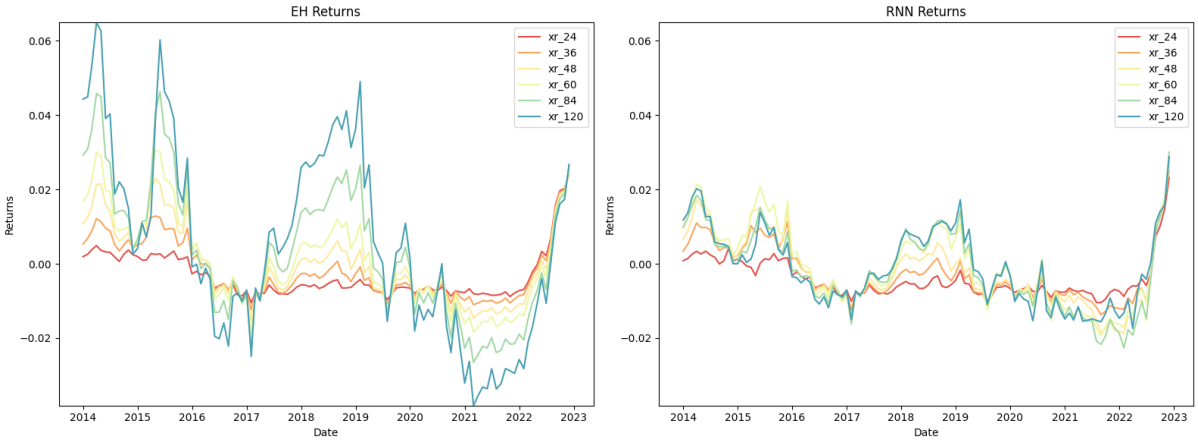


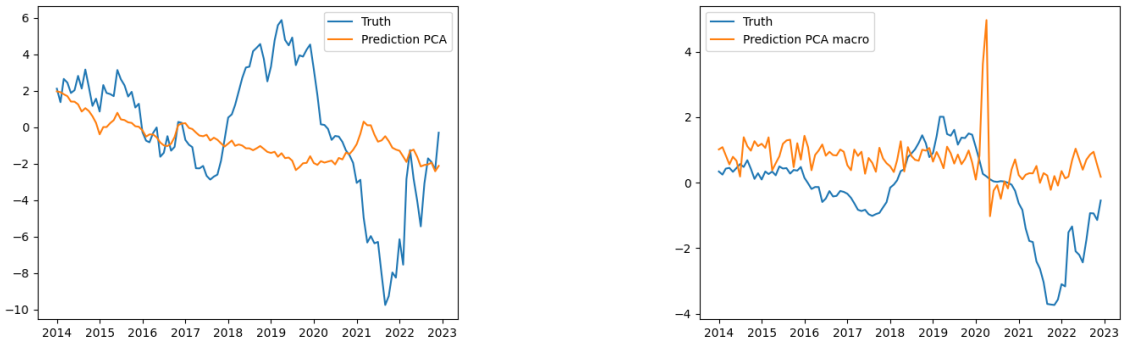
Figure 6.2: Plot of Returns from Mean-Variance Utility Portfolio - Germany EH (left) & RNN (right) for the out-of-sample period 2014:01 - 2022:12.

6.2.1 Pros and cons of the inclusion of macroeconomic data

To further investigate the differences in results between Germany and the U.S., we conducted a closer examination of the principal components. When replacing the variables with the principal components, the U.S. neural network model continues to demonstrate strong performance, with R^2_{OOS} values ranging from 10% to 17%. In contrast, the German model struggles, yielding negative R^2_{OOS} values.

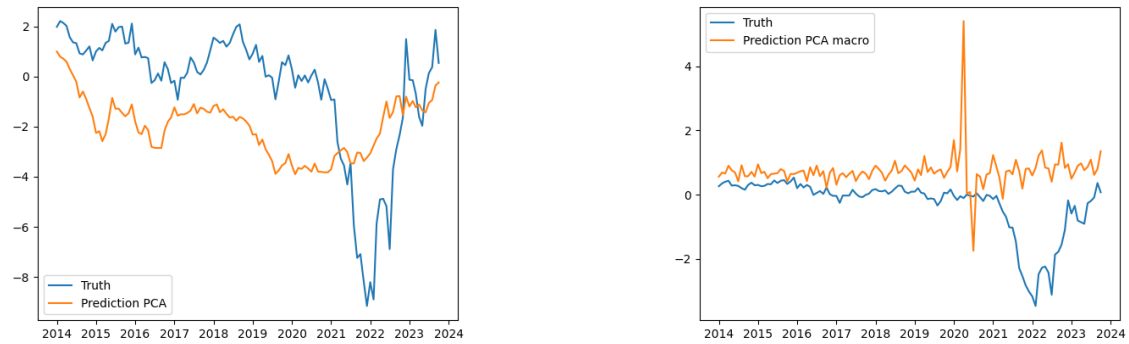
Our investigation of the Principal Component Analysis (PCA) reveals that for both the U.S. and German markets, the inclusion of macroeconomic variables generally improves model performance. However, this improvement comes with a caveat. In a linear modeling

context, both U.S. and German models exhibited significant prediction errors at the onset of the COVID-19 pandemic. Further analysis indicates that these errors are primarily attributable to the macroeconomic components in the models. Interestingly, we see almost the exact same spike in the predictions for linear models for both U.S. and Germany.



(a) Forward rates only PCA **(b)** Macro only PCA

Figure 6.3: Comparison of predictions while only using yields and macroeconomic variables in the U.S.



(a) Forward rates only PCA **(b)** Macro only PCA

Figure 6.4: Comparison of predictions while only using yields and macroeconomic variables in Germany.

These findings lead to an important conclusion regarding the role of macroeconomic variables in predicting bond excess returns. While macroeconomic factors can provide additional predictive power, their inclusion also introduces potential vulnerabilities to the models. This is particularly evident during periods of significant economic disruption, such as the COVID-19 pandemic.

The improved performance from including macroeconomic variables suggests that they contain information relevant to bond pricing that is not fully captured by the yield curve

alone. This challenges the Spanning Hypothesis, which posits that all relevant information for bond return prediction is contained within the yield curve's level, slope, and curvature.

However, it is crucial to consider that, in theory, all publicly available information, including macroeconomic data, should already be priced into bond yields. The fact that macroeconomic variables can enhance predictive power indicates potential market inefficiencies or limitations in how quickly this information is incorporated into prices.

The observation that linear models including macroeconomic variables performed poorly at the onset of the COVID-19 pandemic highlights a critical point: while these variables can enhance predictive power under normal circumstances, they may lead to less robust models during periods of economic stress or structural breaks. In theory, "redundant" information that was correctly not factored in to the bond prices, can disturb the model significantly.

6.3 Limitations and further research

The fact that linear models and RFR produce less accurate predictions does not render them useless. If we take the 10-year US bond as an example, it is not obvious that the 18% R^2 produced by the PLS regression explains "the same variance" as the RNN with 30% R^2. In the next iteration of our model, we would like to include more ensemble models, to see if we can explain different parts of the variance with the combination of econometric and machine learning techniques.

Despite the findings of this study, several limitations warrant careful consideration, as they may shape the interpretation and reliability of our results.

First and foremost, data constraints present significant challenges, especially in the German setting. Our German dataset spans a relatively short historical window starting in the year 2000, capturing fewer economic cycles and recessions compared to the U.S. dataset. This shorter sample limits the model's exposure to diverse market conditions, potentially making it less adept at forecasting bond returns during structural breaks, such as the COVID-19 pandemic. The weaker performance observed for Germany, particularly during the pandemic period, may partly reflect this limited training history. In contrast, the U.S. dataset includes multiple downturns and recessions, allowing the RNN to better

learn patterns associated with crisis periods and thereby adapt more effectively during Covid-19.

Additionally, to keep the same timesteps for both markets, we interpolated quarterly macroeconomic data to a monthly frequency. This procedure risks distorting crucial timing signals and smoothing out important economic fluctuations. As a result, some meaningful predictive relationships between macroeconomic factors and bond returns may have been lost. It also creates a look-ahead bias, as discussed earlier, but that does not seem to be very effective in our analysis.

These data and modeling constraints have practical implications for our findings. Our results, while suggestive of the potential for recurrent neural networks to improve forecasts, may understate the performance that more finely tuned, ensemble-based methods or richer datasets could achieve. The limited interpretability and computational intensity of RNNs, coupled with the constrained dataset, means that our conclusions should be viewed as a starting point rather than a definitive end goal. Different results might emerge if more comprehensive data were available, if more specialized modeling techniques were employed, or if the models were tested across a broader range of economic regimes.

In sum, these limitations imply that our findings, though promising, must be interpreted with caution.

7 Conclusion

This study set out to evaluate the predictive capabilities of machine learning models, particularly Recurrent Neural Networks with LSTM architecture. We use our model to forecast excess returns from zero coupon bonds across two distinct markets: the U.S. and Germany. Our findings provide some insights into the potential of advanced machine learning methods in financial forecasting.

Our results support the first hypothesis, as the RNN model significantly outperform traditional methods, particularly during periods of high market volatility and structural breaks, such as the COVID-19 pandemic. In the U.S. market, RNNs achieved statistically significant out-of-sample improvements across all maturities, with notable gains in economic utility, as reflected in the certainty equivalent return analysis. This suggests that we do indeed find some improvement in performance, due to the RNNs ability to dynamically capture complex temporal patterns and nonlinear relationships between macroeconomic variables, forward rates, and excess returns.

However, the second hypothesis—that our model would achieve similar success in the German market—yielded mixed results. While the model showed promise at longer maturities in Germany, the performance at shorter maturities and in terms of economic utility was less convincing. This disparity makes it difficult to make the claim that the model truly "understand" the dynamics of the market of government bonds.

Our findings suggest that while RNNs represent a promising tool for bond excess return forecasting, their effectiveness is influenced by the quality and granularity of the available data, as well as the economic characteristics of the market in question. The German dataset's relatively short historical window and composed macroeconomic environment may hinder the ability of RNNs to fully capture predictive relationships, limiting their practical application in this context.

Despite these challenges, this study contributes to the growing body of literature on machine learning in finance, providing evidence of the potential of RNNs to enhance bond return predictability, particularly in data-rich environments like the U.S.

In conclusion, while this study highlights the promise of RNNs for financial forecasting, it

also emphasizes the challenge of making models that generalize to the characteristics of multiple markets. The findings serve as a foundation for further exploration of advanced machine learning methods in financial research, offering opportunities for future research.

8 Declaration on the use of AI tools

Name (and version) of the AI tool: ChatGPT, 4.0o

Purpose of using the tool: Utilized primarily to provide structural guidance for both chapters and paragraphs, a task at which we experienced it to excel at. However, when prompted to rewrite text for improved language or academic writing, AI often introduced illogical statements or assertions that could not be supported by the provided information. As a result, we exercised caution and limited our use of AI to mainly structural assistance. It is also used for research support, concept explanation and brainstorming ideas.

Name (and version) of the AI tool: Perplexity Labs (Free version)

Purpose of using the tool: Utilized primarily to provide structural guidance for both chapters and paragraphs, a task at which we experienced it to excel at. However, when prompted to rewrite text for improved language or academic writing, AI often introduced illogical statements or assertions that could not be supported by the provided information. As a result, we exercised caution and limited our use of AI to structural assistance.

Name (and version) of the AI tool: Gemini - Colab Assistant

Purpose of using the tool: Built in to Google Colabs IDE, with autocomplete and generative support. Autocomplete was mainly used for repetitive tasks, such as loading and data cleaning for multiple datasets.

Name (and version) of the AI tool: Grammarly

Purpose of using the tool: Correcting language

We are aware that we are responsible for all content of this master's thesis, including the parts where AI tools are used. We are responsible for ensuring that the thesis complies with ethical rules for privacy and publication.

References

Barigozzi, M., & Lissona, C. (2024). EA-MD-QD: A large euro area and euro member countries datasets for macroeconomic research.

Bauer, M. D., & Hamilton, J. D. (2018). Robust bond risk premia. *The Review of Financial Studies*, *31*(2), 399–448.

Beckmann, L., Debener, J., & Kriebel, J. (2023). Understanding the determinants of bond excess returns using explainable ai. *Journal of Business Economics*, *93*, 1553–1590.

Bianchi, D., Büchner, M., Hoogteijling, T., & Tamoni, A. (2021). Corrigendum: Bond risk premiums with machine learning. *The Review of Financial Studies*, *34*(2), 1090–1103.

Bianchi, D., Büchner, M., & Tamoni, A. (2020). Machine learning for bond return predictability. *Journal of Financial Economics*, *137*(3), 641–665.

Brezak, D., Bacek, T., Majetic, D., Kasac, J., & Novakovic, B. (2012). A comparison of feed-forward and recurrent neural networks in time series forecasting. *2012 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 1–6.

Bundesbank, D. (2024). *Yield curve for listed federal securities (svensson method)* [Accessed: 2024-12-19]. https://www.bundesbank.de/en/statistics/time-series-databases

Campbell, J. Y., & Shiller, R. J. (1991). Yield spreads and interest rate movements: A bird's eye view. *The Review of Economic Studies*, *58*(3), 495–514.

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.

Campbell, J. Y., & Viceira, L. M. (1999). Consumption and portfolio decisions when expected returns are time varying. *The Quarterly Journal of Economics*, *114*(2), 433–495.

Chen, Y., Zhang, Y., & Chen, X. (2019). Stock price prediction using lstm and technical indicators. *Journal of Intelligent Information Systems*, *55*(2), 257–273.

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

Cochrane, J. H., & Piazzesi, M. (2005). Bond risk premia. *American Economic Review*, *95*(1), 138–160.

Deutsche Bundesbank. (2024). Monetary Policy - Deutsche Bundesbank [Accessed: 2024-12-19]. https://www.bundesbank.de/en/tasks/monetary-policy/monetary-policy-625914

Fama, E. F., & Bliss, R. R. (1987). The information in long-maturity forward rates. *The American Economic Review*, *77*(4), 680–692.

Feng, G., He, X., Wang, Y., & Wu, C. (2024). Predicting individual corporate bond returns. *SSRN Electronic Journal*.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market prediction. *European Journal of Operational Research*, *270*(2), 654–663.

Fisher, I. (1896). *Appreciation and interest: A study of the influence of monetary appreciation and depreciation on the rate of interest with applications to the bimetallic controversy and the theory of interest* [Publications of the American

Economic Association, Vol. 11, no. 4, pp 331-442]. American Economic Association by the Macmillan Company.

Gargano, A., Pettenuzzo, D., & Timmermann, A. (2019). Bond return predictability: Economic value and links to the macroeconomy. *Management Science*, *65*(2), 508–540.

Geweke, J. (2001). A note on some limitations of crra utility. *Economics Letters*, *71*(3), 341–345.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*(4), 1455–1508.

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

Gudelek, M. U., Ozbayoglu, A. M., & Sezer, O. B. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, *90*, 106181.

Hinton, G. E. (1990). Connectionist learning procedures. *Artificial Intelligence*, *40*(1-3), 185–234.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hoogteijling, T., Martens, M., & van der Wel, M. (2021). Forecasting bond risk premia using stationary yield factors. *SSRN Electronic Journal*.

Huang, J.-Z., & Shi, Y. (2023). Predicting bond risk premia using machine learning: The role of macroeconomic information. *Journal of Financial Economics*, *147*(2), 456–475.

Jain, K. (2019, December). Understanding lstms [Downloaded 19.12.2024; Includes an illustration of an LSTM cell with labeled gates and computations.]. https://kushalj001.github.io/black-box-ml/lstm/pytorch/gates/vanishing%20gradient/2019/12/28/Understanding-LSTMs.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Liu, Y., & Wu, J. C. (2021). Reconstructing the yield curve. *Journal of Financial Economics*, *142*(3), 1395–1425.

Ludvigson, S. C., & Ng, S. (2009). Macro factors in bond risk premia. *Review of Financial Studies*, *22*(12), 5027–5067.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *The Journal of Business*, *60*(4), 473–489.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Sarno, L., Schneider, P., & Wagner, C. (2016). The economic value of predicting bond risk premia. *Journal of Empirical Finance*, *37*, 247–267.

Svensson, L. E. (1994). Estimating and interpreting forward interest rates: Sweden 1992-1994 [Centre for Economic Policy Research, Discussion Paper No. 1051].

Thornton, D. L., & Valente, G. (2012). Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *Review of Financial Studies*, *25*(10), 3141–3168.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Appendices

A Variable Description

A.1 German Macroeconomic Data

The table includes several columns detailing the variables in the dataset. **N** provides the serial number, while **ID** is the unique identifier for each variable. **Series** describes the economic variable, and **Unit** specifies the measurement (e.g., CLV indexed to 2015). **SA** indicates the type of seasonal adjustment applied (e.g., NSA, SA, SCA, MSA), and **F** denotes the data frequency (Q for quarterly, M for monthly). **LT** and **HT** represent the light and heavy transformations applied, respectively, with codes ranging from 0 to 5. In this dataset, only **Light Transformations (LT)** are used, with codes as follows: **0** (no transformation, x_t), **1** ($100 \times \log(x_t)$), **2** ($100 \times \Delta \log(x_t)$), **3** ($100 \times \Delta^2 \log(x_t)$), **4** ($\Delta x_t$), and **5** ($\Delta^2 x_t$).

Table A.1: Data Description: Germany

N	ID	Series	Unit	SA	F	LT	HT
(1) National Accounts							
1	GDP	Real Gross Domestic Product	CLV(2015)	SCA	Q	2	2
2	EXPGS	Real Export Goods and services	CLV(2015)	SCA	Q	2	2
3	IMPGS	Real Import Goods and services	CLV(2015)	SCA	Q	2	2
4	GFCE	Real Government Final consumption expenditure	CLV(2015)	SCA	Q	2	2
5	HFCE	Real Households consumption expenditure	CLV(2015)	SCA	Q	2	2
6	CONSND	Real Households consumption expenditure: Durable Goods	CLV(2015)	SCA	Q	2	2
7	CONSSD	Real Households consumption expenditure: Semi-Durable Goods	CLV(2015)	SCA	Q	2	2
8	CONSND	Real Households consumption expenditure: Non-Durable Goods	CLV(2015)	SCA	Q	2	2
9	CONSSV	Real Households consumption expenditure: Services	CLV(2015)	SCA	Q	2	2
10	GCF	Real Gross capital formation	CLV(2015)	SCA	Q	2	2
11	GCFC	Real Gross fixed capital formation	CLV(2015)	SCA	Q	2	2
12	GFACON	Real Gross Fixed Capital Formation: Construction	CLV(2015)	SCA	Q	2	2
13	GFAMG	Real Gross Fixed Capital Formation: Machinery and Equipment	CLV(2015)	SCA	Q	2	2
14	GNFCPS	Gross Profit Share of Non-Financial Corporations	Percent	SCA	Q	0	0
15	GNFCIR	Gross Investment Share of Non-Financial Corporations	Percent	SCA	Q	0	0
16	GHIR	Gross Investment Rate of Households	Percent	SCA	Q	0	0

Continued on next page

Table A.1 – continued from previous page

N	ID	Series	Unit	SA	F	LT	HT
17	GHSR	Gross Households Savings Rate	Percent	SCA	Q	0	0
		(2) Labor Market Indicators					
18	TEMP	Total Employment (domestic concept)	1000-ppl	SCA	Q	2	2
19	EMP	Employees (domestic concept)	1000-ppl	SCA	Q	2	2
20	SEMP	Self Employment (domestic concept)	1000-ppl	SCA	Q	2	2
21	THOURS	Hours Worked: Total	2015=100	SCA	Q	2	0
22	EMPAG	Quarterly Employment: Agriculture, Forestry, Fishing	1000-ppl	SCA	Q	2	2
23	EMPIN	Quarterly Employment: Industry	1000-ppl	SCA	Q	2	2
24	EMPMN	Quarterly Employment: Manufacturing	1000-ppl	SCA	Q	2	2
25	EMPCON	Quarterly Employment: Construction	1000-ppl	SCA	Q	2	2
26	EMPRT	Quarterly Employment: Wholesale/Retail trade, transport, food	1000-ppl	SCA	Q	2	2
27	EMPIT	Quarterly Employment: Information and Communication	1000-ppl	SCA	Q	2	2
28	EMPFC	Quarterly Employment: Financial and Insurance activities	1000-ppl	SCA	Q	2	2
29	EMPRE	Quarterly Employment: Real Estate	1000-ppl	SCA	Q	2	2
30	EMPPR	Quarterly Employment: Professional, Scientific, Technical activities	1000-ppl	SCA	Q	2	2
31	EMPPA	Quarterly Employment: PA, education, health and social services	1000-ppl	SCA	Q	2	2
32	EMPENT	Quarterly Employment: Arts and recreational activities	1000-ppl	SCA	Q	2	2
33	UNETOT	Unemployment: Total	%active	SA	M	0	4.5
34	UNEO25	Unemployment: Over 25 years	%active	SA	M	0	4.5
35	UNEU25	Unemployment: Under 25 years	%active	SA	M	0	4.5
36	RPRP	Real Labour Productivity (person)	2015=100	SCA	Q	2	2
37	WS	Wages and salaries	CP	SA	Q	2	2
38	ESC	Employers' Social Contributions	CP	SA	Q	2	2
		(3) Credit Aggregates					
39	TASS.SDB	Total Economy - Assets: Short-Term Debt Securities	MLN€	MSA	Q	2	2
40	TASS.LDB	Total Economy - Assets: Long-Term Debt Securities	MLN€	MSA	Q	2	2
41	TASS.SLN	Total Economy - Assets: Short-Term Loans	MLN€	MSA	Q	2	2
42	TASS.LLN	Total Economy - Assets: Long-Term Loans	MLN€	MSA	Q	2	2
43	TLB.SDB	Total Economy - Liabilities: Short-Term Debt Securities	MLN€	MSA	Q	2	2
44	TLB.LDB	Total Economy - Liabilities: Long-Term Debt Securities	MLN€	MSA	Q	2	2
45	TLB.SLN	Total Economy - Liabilities: Short-Term Loans	MLN€	MSA	Q	2	2
46	TLB.LLN	Total Economy - Liabilities: Long-Term Loans	MLN€	MSA	Q	2	2
47	NFCASS	Non-Financial Corporations: Total Financial Assets	MLN€	MSA	Q	2	2
48	NFCASS.SLN	Non-Financial Corporations - Assets: Short-Term Loans	MLN€	MSA	Q	2	2
49	NFCASS.LLN	Non-Financial Corporations - Assets: Long-Term Loans	MLN€	MSA	Q	2	2
50	NFCLB	Non-Financial Corporations: Total Financial Liabilities	MLN€	MSA	Q	2	2

Continued on next page

Table A.1 – continued from previous page

N	ID	Series	Unit	SA	F	LT	HT
51	NFCLB.SLN	Non-Financial Corporations - Liabilities: Short-Term Loans	MLN€	MSA	Q	2	2
52	NFCLB.LLN	Non-Financial Corporations - Liabilities: Long-Term Loans	MLN€	MSA	Q	2	2
53	GGASS	General Government: Total Financial Assets	MLN€	MSA	Q	2	2
54	GGASS.LLN	General Government - Assets: Short-Term Loans	MLN€	MSA	Q	2	2
55	GGLB	General Government: Total Financial Liabilities	MLN€	MSA	Q	2	2
56	GGLB.SLN	General Government - Liabilities: Short-Term Loans	MLN€	MSA	Q	2	1
57	GGLB.LLN	General Government - Liabilities: Long-Term Loans	MLN€	MSA	Q	2	2
58	HHASS	Households: Total Financial Assets	MLN€	MSA	Q	2	2
59	HHLB	Households: Total Financial Liabilities	MLN€	MSA	Q	2	3
60	HHLB.SLN	Households - Liabilities: Short-Term Loans	MLN€	MSA	Q	2	2
61	HHLB.LLN	Households - Liabilities: Long-Term Loans	MLN€	MSA	Q	2	3
		(4) Labor Costs					
62	ULCIN	Nominal Unit Labor Costs: Industry	2016=100	SCA	Q	2	2
63	ULCMN	Nominal Unit Labor Costs: Manufacturing	2016=100	SCA	Q	2	2
64	ULCCON	Nominal Unit Labor Costs: Construction	2016=100	SCA	Q	2	2
65	ULCRT	Nominal Unit Labor Costs: Wholesale/Retail Trade, Transport, Food, IT	2016=100	SCA	Q	2	2
66	ULCFC	Nominal Unit Labor Costs: Financial Activities	2016=100	SCA	Q	2	2
67	ULCRE	Nominal Unit Labor Costs: Real Estate	2016=100	SCA	Q	2	2
68	ULCPR	Nominal Unit Labor Costs: Professional, Scientific, Technical activities	2016=100	SCA	Q	2	2
		(5) Exchange Rates					
69	REER42	Real Exchange Rate (42 main industrial countries)	2010=100	NSA	M	2	2
		(6) Interest Rates					
70	LTIRT	Long-Term Interest Rates (EMU Criterion)	Percent	NSA	M	4	4
		(7) Industrial Production and Turnover					
71	IPMN	Industrial Production Index: Manufacturing	2021=100	SCA	M	2	2
72	IPCAG	Industrial Production Index: Capital Goods	2021=100	SCA	M	2	2
73	IPCOG	Industrial Production Index: Consumer Goods	2021=100	SCA	M	2	2
74	IPDCOG	Industrial Production Index: Durable Consumer Goods	2021=100	SCA	M	2	2
75	IPNDCOG	Industrial Production Index: Non Durable Consumer Goods	2021=100	SCA	M	2	2
76	IPING	Industrial Production Index: Intermediate Goods	2021=100	SCA	M	2	2
77	IPNRG	Industrial Production Index: Energy	2021=100	SCA	M	2	2
78	TRNMN	Turnover Index: Manufacturing	2021=100	SCA	M	2	2
79	TRNCAG	Turnover Index: Capital Goods	2021=100	SCA	M	2	2
80	TRNCOG	Turnover Index: Consumer Goods	2021=100	SCA	M	2	2
81	TRNDCOG	Turnover Index: Durable Consumer Goods	2021=100	SCA	M	2	2
82	TRNNDCOG	Turnover Index: Non Durable Consumer Goods	2021=100	SCA	M	2	2
83	TRNING	Turnover Index: Intermediate Goods	2021=100	SCA	M	2	2
84	TRNNRG	Turnover Index: Energy	2021=100	SCA	M	2	2
		(8) Prices					
85	PPICAG	Producer Price Index: Capital Goods	2021=100	MSA	M	2	3
86	PPICOG	Producer Price Index: Consumer Goods	2021=100	MSA	M	2	3

Continued on next page

Table A.1 – continued from previous page

N	ID	Series	Unit	SA	F	LT	HT
87	PPIDCOG	Producer Price Index: Durable Consumer Goods	2021=100	MSA	M	2	3
88	PPINDCOG	Producer Price Index: Non Durable Consumer Goods	2021=100	MSA	M	2	3
89	PPIING	Producer Price Index: Intermediate Goods	2021=100	MSA	M	2	3
90	PPINRG	Producer Price Index: Energy	2021=100	MSA	M	2	3
91	HICPOV	Harmonized Index of Consumer Prices: Overall Index	2010=100	SCA	M	2	3
92	HICPNEF	Harmonized Index of Consumer Prices: All Items: no Energy & Food	2010=100	SCA	M	2	3
93	HICPG	Harmonized Index of Consumer Prices: Goods	2010=100	SCA	M	2	3
94	HICPIN	Harmonized Index of Consumer Prices: Industrial Goods	2010=100	SCA	M	2	3
95	HICPSV	Harmonized Index of Consumer Prices: Services	2010=100	SCA	M	2	3
96	HICPNG	Harmonized Index of Consumer Prices: Energy	2010=100	SCA	M	2	3
97	DFGDP	Real Gross Domestic Product Deflator	2015=100	SCA	Q	2	3
98	HPRC	Residential Property Prices (BIS)	MLN€	SCA	Q	2	3
(9) Confidence Indicators							
99	ICONFIX	Industrial Confidence Indicator	Index	SA	M	0	0
100	CCONFIX	Consumer Confidence Indicator	Index	SA	M	0	0
101	ESENTIX	Economic Sentiment Indicator	Index	SA	M	0	4
102	KCONFIX	Construction Confidence Indicator	Index	SA	M	0	0
103	RTCONFIX	Retail Confidence Indicator	Index	SA	M	0	0
104	SCONFIX	Services Confidence Indicator	Index	SA	M	0	0
105	BCI	Composite Business Confidence Index	2010=100	SA	M	2	0
106	CCI	Composite Consumer Confidence Index	2010=100	SA	M	2	0
(11) Others							
107	SHIX	Share Prices	2010=100	SA	M	2	2

A.2 U.S. Macroeconomic Data

The column *tcode* denotes the following data transformations for a series x: no transformation for $tcode = 1$; first difference, Δx_t, for $tcode = 2$; second difference, $\Delta^2 x_t$, for $tcode = 3$; logarithm, $\log(x_t)$, for $tcode = 4$; first difference of the logarithm, $\Delta \log(x_t)$, for $tcode = 5$; second difference of the logarithm, $\Delta^2 \log(x_t)$, for $tcode = 6$; and $\Delta \left(\frac{x_t}{x_{t-1}} - 1.0 \right)$ for $tcode = 7$.

Table A.2: Data Description: US

id	tcode	fred	description
Group 1: Output and income			
1	5	RPI	Real Personal Income
2	5	W875RX1	Real personal income ex transfer receipts
6	5	INDPRO	IP Index
7	5	IPFPNSS	IP: Final Products and Nonindustrial Supplies

Continued on next page

Table A.2 – continued from previous page

id	tcode	fred	description
8	5	IPFINAL	IP: Final Products (Market Group)
9	5	IPCONGD	IP: Consumer Goods
10	5	IPDCONGD	IP: Durable Consumer Goods
11	5	IPNCONGD	IP: Nondurable Consumer Goods
12	5	IPBUSEQ	IP: Business Equipment
13	5	IPMAT	IP: Materials
14	5	IPDMAT	IP: Durable Materials
15	5	IPNMAT	IP: Nondurable Materials
16	5	IPMANSICS	IP: Manufacturing (SIC)
17	5	IPB51222s	IP: Residential Utilities
18	5	IPFUELS	IP: Fuels
20	2	CUMFNS	Capacity Utilization: Manufacturing
Group 2: Labor market			
21*	2	HWI	Help-Wanted Index for United States
22*	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed
23	5	CLF16OV	Civilian Labor Force
24	5	CE16OV	Civilian Employment
25	2	UNRATE	Civilian Unemployment Rate
26	2	UEMPMEAN	Average Duration of Unemployment (Weeks)
27	5	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks
28	5	UEMP5TO14	Civilians Unemployed for 5-14 Weeks
29	5	UEMP15OV	Civilians Unemployed - 15 Weeks & Over
30	5	UEMP15T26	Civilians Unemployed for 15-26 Weeks
31	5	UEMP27OV	Civilians Unemployed for 27 Weeks and Over
32*	5	CLAIMSx	Initial Claims
33	5	PAYEMS	All Employees: Total nonfarm
34	5	USGOOD	All Employees: Goods-Producing Industries
35	5	CES1021000001	All Employees: Mining and Logging: Mining
36	5	USCONS	All Employees: Construction
37	5	MANEMP	All Employees: Manufacturing
38	5	DMANEMP	All Employees: Durable goods
39	5	NDMANEMP	All Employees: Nondurable goods
40	5	SRVPRD	All Employees: Service-Providing Industries
41	5	USTPU	All Employees: Trade, Transportation & Utilities
42	5	USWTRADE	All Employees: Wholesale Trade
43	5	USTRADE	All Employees: Retail Trade
44	5	USFIRE	All Employees: Financial Activities
45	5	USGOVT	All Employees: Government
46	1	CES0600000007	Avg Weekly Hours : Goods-Producing
47	2	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing
48	1	AWHMAN	Avg Weekly Hours : Manufacturing
127	6	CES0600000008	Avg Hourly Earnings : Goods-Producing
128	6	CES2000000008	Avg Hourly Earnings : Construction
129	6	CES3000000008	Avg Hourly Earnings : Manufacturing
Group 3: Housing			
50	4	HOUST	Housing Starts: Total New Privately Owned

Continued on next page

Table A.2 – continued from previous page

id	tcode	fred	description
51	4	HOUSTNE	Housing Starts, Northeast
52	4	HOUSTMW	Housing Starts, Midwest
53	4	HOUSTS	Housing Starts, South
54	4	HOUSTW	Housing Starts, West
55	4	PERMIT	New Private Housing Permits (SAAR)
56	4	PERMITNE	New Private Housing Permits, Northeast (SAAR)
57	4	PERMITMW	New Private Housing Permits, Midwest (SAAR)
58	4	PERMITS	New Private Housing Permits, South (SAAR)
59	4	PERMITW	New Private Housing Permits, West (SAAR)
Group 4: Consumption, orders, and inventories			
3	5	DPCERA3M086SBEA	Real personal consumption expenditures
4*	5	CMRMTSPLx	Real Manu. and Trade Industries Sales
5*	5	RETAILx	Retail and Food Services Sales
64	5	ACOGNO	New Orders for Consumer Goods
65*	5	AMDMNOx	New Orders for Durable Goods
66*	5	ANDENOx	New Orders for Nondefense Capital Goods
67*	5	AMDMUOx	Unfilled Orders for Durable Goods
68*	5	BUSINVx	Total Business Inventories
69*	2	ISRATIOx	Total Business: Inventories to Sales Ratio
130*	2	UMCSENTx	Consumer Sentiment Index
Group 5: Money and Credit			
70	6	M1SL	M1 Money Stock
71	6	M2SL	M2 Money Stock
72	5	M2REAL	Real M2 Money Stock
73	6	BOGMBASE	Monetary Base
74	6	TOTRESNS	Total Reserves of Depository Institutions
75	7	NONBORRES	Reserves Of Depository Institutions
76	6	BUSLOANS	Commercial and Industrial Loans
77	6	REALLN	Real Estate Loans at All Commercial Banks
78	6	NONREVSL	Total Nonrevolving Credit
79*	2	CONSPI	Nonrevolving consumer credit to Personal Income
132	6	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding
133	6	DTCTHFNM	Total Consumer Loans and Leases Outstanding
134	6	INVEST	Securities in Bank Credit at All Commercial Banks
Group 6: Interest and Exchange Rates			
84	2	FEDFUNDS	Effective Federal Funds Rate
85*	2	CP3Mx	3-Month AA Financial Commercial Paper Rate
86	2	TB3MS	3-Month Treasury Bill
87	2	TB6MS	6-Month Treasury Bill
88	2	GS1	1-Year Treasury Rate
89	2	GS5	5-Year Treasury Rate
90	2	GS10	10-Year Treasury Rate
91	2	AAA	Moody's Seasoned Aaa Corporate Bond Yield
92	2	BAA	Moody's Seasoned Baa Corporate Bond Yield
93*	1	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS
94	1	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS

Continued on next page

Table A.2 – continued from previous page

id	tcode	fred	description
95	1	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS
96	1	T1YFFM	1-Year Treasury C Minus FEDFUNDS
97	1	T5YFFM	5-Year Treasury C Minus FEDFUNDS
98	1	T10YFFM	10-Year Treasury C Minus FEDFUNDS
99	1	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS
100	1	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS
101	5	TWEXAFEGSMTHx	Trade Weighted U.S. Dollar Index
102*	5	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate
103*	5	EXJPUSx	Japan / U.S. Foreign Exchange Rate
104*	5	EXUSUKx	U.S. / U.K. Foreign Exchange Rate
105*	5	EXCAUSx	Canada / U.S. Foreign Exchange Rate
Group 7: Prices			
106	6	WPSFD49207	PPI: Finished Goods
107	6	WPSFD49502	PPI: Finished Consumer Goods
108	6	WPSID61	PPI: Intermediate Materials
109	6	WPSID62	PPI: Crude Materials
110*	6	OILPRICEx	Crude Oil, spliced WTI and Cushing
111	6	PPICMM	PPI: Metals and metal products
113	6	CPIAUCSL	CPI : All Items
114	6	CPIAPPNSL	CPI : Apparel
115	6	CPITRNSL	CPI : Transportation
116	6	CPIMEDSL	CPI : Medical Care
117	6	CUSR0000SAC	CPI : Commodities
118	6	CUSR0000SAD	CPI : Durables
119	6	CUSR0000SAS	CPI : Services
120	6	CPIULFSL	CPI : All Items Less Food
121	6	CUSR0000SA0L2	CPI : All items less shelter
122	6	CUSR0000SA0L5	CPI : All items less medical care
123	6	PCEPI	Personal Cons. Expend.: Chain Index
124	6	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods
125	6	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods
126	6	DSERRG3M086SBEA	Personal Cons. Exp: Services
Group 8: Stock Market			
80*	5	S&P 500	S&P's Common Stock Price Index: Composite
82*	2	S&P div yield	S&P's Composite Common Stock: Dividend Yield
83*	5	S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio
135*	1	VIXCLSx	VIX

B R^2

B.1 SciKitLearn R^2

Table B.1: R^2_{OOS} values from SciKitLearn

Models	$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(10)}_{t+1}$
Panel A: U.S.						
RNN	12.919%	16.378%	26.146%	27.926%	30.479%	30.958%
Elastic Net	0.678%	2.329%	6.750%	5.574%	4.522%	-2.806%
PCR 8+3	-15.009%	-3.822%	4.497%	4.241%	3.801%	-0.303%
PLS	-5.715%	6.523%	13.785%	16.309%	18.032%	18.739%
RFR	-16.022%	-8.140%	-3.105%	-1.578%	0.719%	3.569%

The R^2 scores are calculated using SciKit-Learn's `r2_score` metric for all models across all maturities.

B.2 R^2_{OOS} without pandemic

Table B.2: R^2_{OOS} values without the Pandemic

Models	$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(10)}_{t+1}$
Panel A: U.S.						
RNN	−25.65%*	−6.42%*	11.24%**	18.56%**	19.94%**	22.50%***
	(0.088)	(0.075)	(0.043)	(0.031)	(0.017)	(0.008)
Elastic Net	−28.80%	−23.37%	−24.85%	−26.51%	−27.00%	−23.95%
	(0.633)	(0.598)	(0.634)	(0.672)	(0.667)	(0.628)
PCR 8+3	−14.74%	−13.60%	−20.12%	−22.69%	−33.54%	−45.66%
	(0.144)	(0.261)	(0.498)	(0.598)	(0.640)	(0.551)
PLS	−55.23%	−34.40%	−31.90%	−32.52%	−36.85%	−37.86%
	(0.787)	(0.780)	(0.765)	(0.758)	(0.711)	(0.593)
RFR	−34.81%	−24.21%	−21.44%	−20.68%	−18.16%	−14.95%
	(0.355)	(0.445)	(0.532)	(0.605)	(0.650)	(0.623)

This table reports out-of-sample R^2 values with the models for the U.S market shown in Panel A. The out-of-sample period start in January 2014 and ends in December 2019 (excluding the effects of the pandemic). We measure statistical significance relative to the expectation hypothesis model using the Clark and West (2007) test statistic. * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

C Further Results

C.1 Utility - No negative weights

Table C.1: Mean-Variance Utility - No negative weights

Models	$xr_{t+1}^{(2)}$	$xr_{t+1}^{(3)}$	$xr_{t+1}^{(4)}$	$xr_{t+1}^{(5)}$	$xr_{t+1}^{(7)}$	$xr_{t+1}^{(10)}$
Panel A: U.S.						
RNN	0.05%**	0.16%***	0.20%***	0.21%***	0.23%***	0.27%**
	(0.049)	(0.000)	(0.000)	(0.000)	(0.007)	(0.014)
Elastic Net	0.01%	0.05%	0.08%	0.12%*	0.19%*	0.17%
	(0.379)	(0.119)	(0.135)	(0.070)	(0.079)	(0.287)
PCR 8+3	−0.00%	0.01%	0.01%	0.02%	0.02%	−0.07%
	(0.915)	(0.742)	(0.991)	(0.971)	(0.846)	(0.304)
PLS	0.01%	0.06%*	0.06%	0.05%	0.06%	−0.02%
	(0.609)	(0.071)	(0.278)	(0.644)	(0.891)	(0.470)
RFR	0.01%	0.03%	0.04%	0.02%	0.03%	−0.03%
	(0.537)	(0.276)	(0.337)	(0.612)	(0.686)	(0.532)
Panel B: Germany						
RNN	−0.07%***	−0.07%***	−0.15%***	−0.14%***	−0.45%***	−0.53%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Elastic Net	−0.08%***	−0.16%***	−0.23%***	−0.26%***	−0.29%***	−0.33%**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PCR 8+3	−0.07%***	−0.15%***	−0.23%***	−0.29%***	−0.41%***	−0.49%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
PLS	−0.11%***	−0.25%***	−0.37%***	−0.45%***	−0.58%***	−0.74%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
RFR	−0.03%***	−0.04%**	−0.06%**	−0.07%**	−0.08%	−0.05%
	(0.000)	(0.019)	(0.031)	(0.043)	(0.114)	(0.291)

This table report annualized percentage change in certainty equivalent return for mean-variance utility investors for the U.S market (Panel A) and the German market (Panel B) with a risk aversion coefficient of 5, with weight constraints [0, 0.99]. Statistical significance is based on a one-sided Diebold-Mariano test applied to the out-of-sample period 2014-2022. * significance at 10% level; ** significance at 5% level; *** significance at 1% level.

Table C.2: Power Utility - No negative weights

Models	$xr_{t+1}^{(2)}$	$xr_{t+1}^{(3)}$	$xr_{t+1}^{(4)}$	$xr_{t+1}^{(5)}$	$xr_{t+1}^{(7)}$	$xr_{t+1}^{(10)}$
Panel A: U.S.						
RNN	0.29%***	0.78%***	1.24%***	1.62%***	2.35%***	3.49%***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Elastic Net	0.21%**	0.64%***	0.97%***	1.22%***	1.77%***	2.35%***
	(0.029)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
PCR 8+3	0.14%	0.54%***	0.86%***	1.11%***	1.61%***	2.25%***
	(0.167)	(0.010)	(0.002)	(0.002)	(0.001)	(0.001)
PLS	0.14%	0.45%**	0.83%***	1.12%***	1.67%***	2.69%***
	(0.147)	(0.027)	(0.003)	(0.002)	(0.001)	(0.000)
RFR	0.01%	0.24%	0.44%**	0.62%***	0.97%***	1.62%***
	(0.854)	(0.119)	(0.023)	(0.012)	(0.007)	(0.003)
Panel B: Germany						
RNN	−0.10%*	−0.03%	0.00%	0.00%	0.00%	1.17%**
	(0.130)	(0.129)	(0.450)	(0.315)	(0.354)	(0.012)
Elastic Net	0.01%	0.33%**	0.49%***	0.40%**	0.29%**	0.35%*
	(0.930)	(0.022)	(0.006)	(0.013)	(0.013)	(0.096)
PCR 8+3	0.23%	0.68%	1.13%*	1.28%*	1.82%*	2.02%
	(0.309)	(0.108)	(0.072)	(0.082)	(0.060)	(0.191)
PLS	0.10%	0.28%**	0.42%*	0.26%	0.21%	0.39%
	(0.245)	(0.046)	(0.051)	(0.239)	(0.376)	(0.292)
RFR	0.03%	0.30%**	0.37%***	0.37%***	0.44%**	0.97%***
	(0.672)	(0.012)	(0.004)	(0.008)	(0.012)	(0.005)

This table report annualized percentage change in certainty equivalent return for power utility investors for the U.S market (Panel A) and the German market (Panel B) with a risk aversion coefficient of 5, with weight constraints [0, 0.99]. Statistical significance is based on a one-sided Diebold-Mariano test applied to the out-of-sample period 2014-2022. * significance at 10% level; ** significance at 5% level; *** significance at 1% level.