Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Financial Economics

Constructing efficient portfolios with machine learning. An empirical analysis for the UK stock market.

Grace Tan Yen Rou (528973)

Supervisor:	Dr. Jens Soerlie Kvaerner
Date final version:	11th June 2023

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

I propose including aggregate indices of investor sentiment (SENT) and investor attention (ATTENT) as predictors, alongside firm and macroeconomic factors, in machine learning stock return prediction model. Therefore, I design a benchmark sample and an extended sample that includes the additional SENT and ATTENT predictors for comparison. Among the various forecasting models, the Deep Neural Network (DNN) exhibits the best predictive performance for stock-level returns, with R^2_{oos} of 1.70% in the benchmark sample and 1.90% in the extended sample. Graphically, the cumulative long-short portfolio return constructed using predictions from the DNN outperforms the UK cumulative excess market return in both samples. However, the majority of the portfolio's return comes from the short positions. After adjusting for the CAPM, Fama-French 3, 5, and 6 risk factor models, the risk-adjusted portfolio return for the benchmark sample remains highly significant. The portfolio exhibits a monthly adjusted return range of 1.97% to 1.83%, supported by t-statistics ranging from 3.89 to 6.18. In the extended sample, it yields a highly significant monthly adjusted return range of 1.75% to 2.13%, supported by t-statistics ranging from 3.14 to 3.70. Lastly, incorporating sentiment and attention (SENT/ATTENT) significantly enhances portfolio performance, with a 71.02% increase in the return of the portfolio for each one-unit increase in the portfolio's return with sentiment and attention, with a t-statistic of 8.05.

Keywords: asset pricing, machine learning, stock return prediction, investor sentiment, investor attention, portfolio analysis.

Preface

This master's thesis is written for the Erasmus School of Economics. It is part of the study program Master's Financial Economics. I am honored to present this thesis, which marks the culmination of my academic journey. It is with great pleasure that I share my research findings, methodologies, and reflections on this significant topic.

I would like to express my sincere gratitude to my supervisor, Dr. Jens Soerlie Kvaerner, for his unwavering support, guidance, and invaluable insights throughout the entire research process. His expertise and mentorship have been instrumental in shaping the direction and quality of this work.

Contents

1 Introduction

In the field of asset pricing, researchers have placed significant emphasis on explaining anomalies and utilizing them for predictive purposes. When it comes to machine learning-based stock return prediction, most studies have followed the mainstream literature established by Gu, Kelly and Xiu (2020). These studies incorporate both firm-specific and macroeconomic anomalies in their stock prediction models. However, the majority of these investigations have focused on the US market, which is known for its high tendency of home bias (Azevedo, Kaiser & Müller, 2022). As a result, there has been a recent surge of interest in studying markets outside the US.

While previous studies have successfully utilized firm-specific and macroeconomic anomalies in machine learning-based stock return prediction, the inclusion of investor behavior anomalies has been largely overlooked. This is surprising considering the empirical evidence from traditional regression analysis (OLS), which suggests that investor sentiment and attention have demonstrated predictive power on stock return (Hudson & Green, 2015; J. Chen, Tang, Yao & Zhou, 2022).These predictors capture the irrational behavior of investors, driven by heterogeneous beliefs and deviations from fundamental asset values. For instance, investor attention arises due to cognitive limitations, while investor sentiment reflects overoptimistic and overconfident attitudes, and vice versa.

In this study, I incorporate investor sentiment and investor attention predictors, in addition to formal firm and macroeconomic predictors, into machine learning models for stock prediction and prediction-sorted portfolio analysis in the UK market. Following the main literature Hudson and Green (2015) and J. Chen et al. (2022), I use Principal Component Analysis (PCA) to aggregate investor sentiment and investor attention proxies into composite indices called SENT and ATTENT. These indices are then used as predictors in the machine learning models for stock prediction.

Furthermore, I construct two groups of study subjects: the benchmark sample and the extended sample. The benchmark sample consists of 49 firm and macroeconomic predictors that are fit into the machine learning models to predict stock returns. In the extended sample, in addition to the traditional and macroeconomic predictors, the SENT and ATTENT indices are added, resulting in a total of 51 predictors used in the forecast models to predict stock returns at the individual stock level.

At the stock level analysis, my aim is to investigate whether flexible non-linear functional form machine learning models such as Generalized Linear Models (GLM), Gradient Boosting Method (GBM), and Deep Neural Networks (DNN) outperform the linear model Ordinary Least Squares regression (OLS). I also aim to determine the best forecast model under each respective benchmark sample and extended sample. The UK sample is divided into three different subperiods: from 2000 to 2022. The first 10 years are used to train the data, the next 5 years are used for hyperparameter tuning in the validation set, and the last 8 years are used for the out-of-sample testing set.

Among the machine learning models, the DNN with NN3 architecture outperforms the other forecast models, demonstrating the highest prediction performance. This result is consistent with various existing literature in both the US and non-US markets (Gu et al., 2020; Drobetz, Haller, Jasperneite & Otto, 2019; Hanauer & Kalsbach, 2023). However, the DNN achieves an out-of-sample R^2_{oos} of 1.90% under extended sample which is higher than the benchmark sample of R^2_{oos} of 1.70%. However, both groups consistently yield higher R^2_{oos} compared to OLS models in both benchmark and extended sample ranging from 0.60% to 0.67%. The Diebold-Mariano test further confirms the superior performance of the DNN model in terms of predictive accuracy, which is significant at a p-value 10% level for both benchmark and extended samples.

To explore the explainability of the forecast models, I analyze variable importance plot, and all models agree that STReversal is the most important factor shaping stock level predictions in both the benchmark and extended samples. Additionally, SENT and ATTENT are among the top-ranking variables across all forecast models. To increase the interpretability of the superior "black-box" forecast model DNN, I introduce Partial Dependence and Individual Conditional Expectation Plots to study the marginal effects of STReversal, SENT, and ATTENT on stock returns in the extended sample. SENT shows an increasing upward trend, whereas ATTENT exhibits an increasing downward trend, which is consistent with the literature on investor behavior and stock return prediction(Hudson & Green, 2015; J. Chen et al., 2022).

Having verified that the DNN is the best-performing forecast model for stock level prediction, I assess its return predictability using conventional portfolio sorts under both the benchmark and extended samples. The realized excess return is sorted based on the model's predicted excess return for the next month. Finally, I create an equal-weighted zero-net-investment portfolio, also known as a long-short portfolio, for both the benchmark sample and extended sample by buying stocks with the highest expected returns (decile 10) and selling stocks with the lowest expected returns (decile 1). The Sharpe ratio of the High minus Low decile (10-1) for the benchmark sample is 0.33, which is lower than the extended sample's (10-1) Sharpe ratio of 0.45.

The cumulative portfolio return for the DNN model in both the benchmark and extended samples outperforms the market return (FTSE All Share) over the 8-year period. However, the extended sample's final cumulative return as of 2022 is higher than the benchmark sample's cumulative return. Graphically, in the UK market, the long leg of the portfolio mostly incurs losses, however a significant spike is observed around 2020 to 2021. Most of the time, the portfolio relies on short selling to compensate for losses on the long side. The cumulative return under the extended sample has a smoother trendline compared to the benchmark sample. The incorporation of SENT and ATTENT factors allows the machine learning model to better adjust its predictions, especially during market turbulent periods, preventing a decrease in the overall cumulative return and consistently contributing to higher returns for that period.

Additionally, I assess the adjusted portfolio returns using the CAPM, Fama-French 3, 5, and 6 factor models. These factor models are unable to explain away the DNN long-short portfolio returns, as a highly significant alpha (risk-adjusted return) at the p-value of 1% level is present in both the benchmark and extended samples. The reported range of monthly adjusted return in the benchmark sample is 1.97%(CAPM)-1.83%(FF6), with t-statistics ranging from 3.89 to 6.18. As for the range of monthly adjusted return in the extended sample is 2.13%(CAPM) to 1.75% (FF6), with t-statistics ranging from 3.14 to 3.70. This indicates that machine learning models such as DNN has its uniqueness in generating unexplainable additional returns.

Lastly, I perform a regression analysis, regressing the extended sample's long-short portfolio return on the benchmark sample's long-short portfolio return. The results show that a one-unit increase in the long-short portfolio return with sentiment and attention (extended sample) is associated with an estimated increase of 71.02% with t-statistics of 8.05 in the long-short portfolio return without sentiment and attention (benchmark sample). This relationship is highly positive and significant at the p-value 1% level. The inclusion of sentiment and attention as stock level predictors indeed improves the portfolio's returns.

The rest of the paper is organized as follows. In section 2, I provide a literature review of machine learning stock return prediction and portfolio analysis, as well as the extensions associated to investor behavior predictors. Section 3 contains the methodology related to the research design, construction of stock level prediction models, and the procedures for both stock and portfolio level analysis. I present the results in section 4 and conclude in section 5.

2 Literature Review

2.1 Machine Learning Stock Return Prediction and Portfolio Analysis

The prediction of stock return has been a subject of extensive research in finance. Traditional method such as Ordinary Least Square regression (OLS), has been widely used to predict stock return based on firm characteristics and macroeconomics variables. For example, Fama and French (2008) and Lewellen (2014), they run cross sectional regression of future stock return on a several lagged firm-level stock characteristics' predictors. On the other hand, Welch and Goyal (2008), they conduct time-series regression for stock return on a small number of macroeconomics predictor variables.

The main limitation for using traditional method (OLS) to predict stock return, is that the model with simple linear functional form unable to accommodate the large number of predictors, unable to capture interaction variables as well as non-linear patterns among the predictors. Therefore, it results in weaker prediction performance.

One of the pioneering works in machine learning asset pricing literature by Gu et al. (2020) successfully leveraged the advantages of existing literature by constructing predictors that incorporate both firm characteristics and macroeconomic variables in a cross-sectional setting. Secondly, they utilized machine learning models with non-linear functional forms, allowing them to capture the non-linear and non-stationary relationship between stock returns and firm and macroeconomic characteristics. Additionally, these models effectively captured signals from a large number of predictors, thereby increasing the coverage and significantly improving the prediction performance for stock returns.

Subsequent to the study conducted by Gu et al. (2020), a multitude of researchers have replicated their framework by experimenting with various supervised and unsupervised models. These include techniques such as dimension reduction, penalized and generalized linear models, support vector machines, regression trees, and neural networks. These models have been applied using an extensive large number of firm and macroeconomic variables that demonstrate the potential to predict stock returns. Notable examples of such studies include Avramov, Cheng and Metzker (2023); Feng, He, Polson and Xu (2018); Feng and He (2022); Freyberger, Neuhierl and Weber (2020); Rapach and Zhou (2020); Chinco, Clark-Joseph and Ye (2019); L. Chen, Pelger and Zhu (2023). A consistent finding across all these studies is that both tree-based supervised models and neural network supervised models have been widely acknowledged as highly effective in predicting stock returns.

Rather than including an exponentially large number of stock predictors, which can introduce redundancy and noise in the machine learning models, recent research suggest an optimal range of predictors from 40 to 80. These predictors should focus on accessible stock characteristics without incorporating complex interactions or nonlinear variables. Beyond this range, the marginal predictive power tends to decline (Choi, Jiang & Zhang, 2022; Drobetz et al., 2019; Hanauer & Kalsbach, 2023). For instance,Crego, Soerlie Kvaerner and Stam (2023) conduct a study on long-short portfolios and observe a decrease in the average return as the number of predictors increase. They examine it from three different dimensions: actual and predicted yield, as well as realized return. Furthermore, their analysis indicates that around 50 predictors serve as the cutoff point for achieving zero positive return at the portfolio level.

However, the majority of studies tend to conduct their analyses solely based on one market, primarily the US market, which can introduce a home bias (Azevedo et al., 2022). Consequently, another recent strand of literature has emerged, expanding the scope of analysis beyond the US market. For example, Drobetz et al. (2019) focus on the European market, while Hanauer and Kalsbach (2023) concentrate on emerging markets. These studies consistently find similar evidence that machine learning (ML) models outperform the benchmark model of ordinary least squares (OLS) in predicting stock-level returns, irrespective of whether it is in developed or emerging markets. In addition to stock-level predictions, researchers, following the approach of Gu et al. (2020), have also extended their analyses to evaluate prediction-sorted portfolios. Furthermore, a few studies have explored risk-adjusted returns at the portfolio level, utilizing benchmark models such as the CAPM and Fama-French models (Drobetz et al., 2019; Hanauer & Kalsbach, 2023; Crego et al., 2023).

2.2 Extension of Machine Learning Prediction Models

2.2.1 Investor Sentiment as Predictors

The waves of irrational sentiment can produce both underreaction and overreaction to news (Barberis, Shleifer & Vishny, 1998). For example, investors with overly optimistic or pessimistic expectations can persist and divert asset prices from their rational, fundamental values for significant of time (Kahneman & Tversky, 1973; Odean, 1998). Most specifically, investor sentiment is defined as as the propensity to speculate (Baker & Wurgler, 2006). Under this definition, sentiment-based mispricing is based on an uninformed demand shock. Because correlated demand shocks persist among uninformed noise traders over time, this can lead to high levels of speculative activity, consequently causing persistent mispricing (Brown & Cliff, 2004).

Investor sentiment able to influence asset prices and encompasses explanatory power on some well-known asset pricing anomalies (Sun, Najand & Shen, 2016).For example,Lemmon and Portniaguina (2006) find that consumer confidence as the proxy of the sentiment indeed able to forecast small stock returns under time-series setting.Antoniou, Doukas and Subrahmanyam (2013) find that momentum profits only exist only under investor optimism. Baker and Wurgler (2006) demonstrate the cross sectional effect of investor sentiment on return prediction. They point out that when sentiment is high,the future return is relatively low for small, young, unprofitable, distressed, high growth, non-dividend-paying firms as well as firms with volatile stock return.

There are several empirical ways to measure investor sentiment. First, survey-based techniques related to acquiring the public's expectations and thoughts about the stock market, which aim to capture the mood of investors. For example, University of Michigan Consumer Sentiment Index, Consumer Confidence Index, the AAII investor sentiment survey, and the UBS/GALLUP Index (Brown, 1999; Fong, 2013; Schmeling, 2009). However, these indexes are available and suitable for US market analysis. Second method is media-based investor sentiment measure, which is based on textual analysis of media contents. For example, newspapers, blogs and google search results (Sun et al., 2016). This method highly relies on Natural Processing Language (NPL) for accurate measurements. Third is the composite sentiment index using principal component analysis (PCA) to extract a single sentiment measure from a range of pertinent financial market indicators.

Despite the discrepancy of the components of composite sentiment index have been suggested in the existing literature. The different versions of composite index still able to capture the variation within the component, with the renewal of information across time. It still shows consistent significance of sentiment on the predictability of the cross sectional stock return in various studies (Brown & Cliff, 2004; Baker & Wurgler, 2006; Hudson & Green, 2015).

2.2.2 Investor Attention as Predictors

Attention is a scarce cognitive resources (Kahneman & Tversky, 1973), information can be incorporated into asset prices only when investors pay sufficient attention (Huberman & Regev, 2001). However, in reality, investors, especially retail investors are attention constrained (Andrei & Hasler, 2015; Hirshleifer & Teoh, 2003; Kacperczyk, Van Nieuwerburgh & Veldkamp, 2016;

Peng & Xiong, 2006). They cannot fully understand all market level information due to bounded rationality with limited time and energy. Therefore, rational inattention investors incline to focus on major events, aggregate market level, sector-wide shocks rather than firm level (Huang, Huang & Lin, 2019; Peng & Xiong, 2006; J. Chen et al., 2022). It further reflects on investors' decision making, they tend to choose attention-grabbing stocks, creating temporary price pressure, deviate the stock price from fundamental values (J. Chen et al., 2022).

Investor attention has substantial asset pricing implication and the attention drawing effect depends on the position that a certain type of information is displayed (J. Chen et al., 2022). Fedyk (2018) find the front page news items on Bloomberg terminal induces greater trading volume, and the price changes rapidly after publication. Barber, Huang, Odean and Schwarz (2022) show that the compilation of a "top fluctuating stock list" within the Robinhood mobile phone trading app leads to concentrate trading. This concentrate trading pattern is characterized by intense buying behavior, which subsequently results in abnormal returns.

Empirically, the existing literature suggested using traditional investor attention proxies, such as abnormal trading volume, extreme trading volume (Barber & Odean, 2007);past returns (Aboody, Lehavy & Trueman, 2010),nearness to 52-week high and nearness to historical high (J. Li & Yu, 2012); analyst coverage (Hirshleifer & Teoh, 2003), which all based on the market level. However, with the presence of internet, researchers utilize the keyword-based social media search traffics. For example, Google Search Volume (Bijl, Kringhaug, Molnár & Sandvik, 2016; Han, He, Rapach & Zhou, 2018; X. Li, Ma, Wang & Zhang, 2015). Also, a new strand of literature suggesting aggregate upper limit hits as investor attention proxies. For example, Seasholes and Wu (2007) find that daily upper limit hits draw investor attention and temporarily drive stock price up from 2001 to 2003. Similarly, T. Chen, Gao, He, Jiang and Xiong (2019) argue that the retail buying after UP limit-hitting day leads to long-run price reversal for up to 120 days from 2012 to 2015. Finally, Cai, Jiang and Liu (2022) demonstrate a negative predictive relationship between cross-sectional stock returns and investor attention, as measured by aggregate upper limit hits. Moreover, implementing long-short trading strategies based on this attention measure generates substantial economic value. Instead of utilizing individual proxies,Chu, Goodell, Shen and Zhang (2022) and J. Chen et al. (2022) have suggested aggregating various investor attention proxies through principal component analysis (PCA) into a composite attention index.

Irrespective of whether individual proxies or aggregate indices are utilized, multiple studies consistently reveal the predictive power of investor attention when it comes to stock returns (Cai et al., 2022; Chu et al., 2022; J. Chen et al., 2022). Additionally, investor attention has demonstrated its independent forecasting power even when controlled for common return predictors, economic predictors, and investor sentiment predictors (J. Chen et al., 2022).

3 Methodology

3.1 Data

I obtain monthly equity return from DataStream for all the firms listed on FTSE All Share, which represents approximately 98%-99% UK stock capitalization. The sample period ranging from January 2000 to December 2022, compromising both active and dead firms as of December 2022, that resulted in 1,078,840 observations across 3929 firms. However, I remove trailing and leading return NAs (missing values), to include only those active period observations for the dead firms. Also, I further drop the observations with infinite returns, both positive and negative, as it is impossible for a firm to have an infinite return in terms of market valuation. After these adjustments, the sample is left with total of 406,969 observations across 3838 firms. Finally, to calculate the stock excess return, I subtract the risk free rate from the stock return. I prepare two sets of data for analysis. The first set consists of 38 firm predictors and 11 macroeconomic predictors. The second set is an extension of first set by including 2 additional predictors: SENT, an aggregated investor sentiment predictor, and ATTENT, an aggregated investor attention predictor.

3.1.1 Firm Predictors

I construct 38 firm predictors from raw accounting data retrieved from DataStream, following the methodology used by Hanauer and Kalsbach (2023). My predictors have a comprehensive coverage of different firm characteristic categories, including past returns, investments, profitability, intangibles, value, and trading frictions, and are based on either annual or monthly frequency. (Appendix A.1) outlines an overview of the 38 firm-level predictors. I did not exclude financial firms but set the following characteristics as missing for these firms, as they are not meaningfully defined for financials: ATO, C, D2A, DPI2A, F2CY, FreeCF, CF2P, GP2A, OA, PCM, PM, Prof, RNA, SG2A, and NOA.

3.1.2 Macroeconomiccs Predictors

For the macroeconomics predictors, I extract the macroeconomics data from various sources such as OECD, UK National Statistics, FRED, as well as DataStream. All of the data are based on a quarterly/monthly basis from January 2000 to December 2022. There is no missing data for the macroeconomics predictors. I have constructed 9 Macroeconomics predictors following the predictors mentioned in Welch and Goyal (2008), including the dividend-yield ratio (d/y), FTSE All Share earning price ratio (e/p), FTSE All Share book-to-market ratio (b/m), risk-free rate (rf), term spread (tms), stock variance (svar), investment to capital ratio (i/k), inflation (infl), as well as the long-term yield (lty). These predictors have been re-examined in Goyal, Welch and Zafirov (2021), and they have been proven to be robust in equity premium prediction. Additionally, I have constructed 2 variables based on Paye (2011): the option adjusted spread (oas), which serves as a substitute for the default spread, and the commercial paper-to-Treasury spread (cp). (Appendix A.2) provides detailed overviews of the definition of the 11 macroeconomics predictors.

3.1.3 Investor Sentiment and Attention Predictors

To construct investor attention and investor sentiment proxies at the market level, my focus lies on utilizing the FTSE 100 as a data source, as suggested by Hudson and Green (2015). This particular index has proven to attract significant attention and media coverage, thereby serving as a reliable reflection of UK investors' market sensitivity across diverse market conditions. Additionally, the FTSE 100 offers greater availability for futures and options, making it a more suitable option for constructing proxies compared to the FTSE All Share.

Concerns about the FTSE 100's representativeness in relation to the FTSE All Share can be set aside. Hudson and Green (2015) suggest that there is a contagious sentiment effect among different countries. Specifically, they find a relationship between US sentiment and UK sentiment. Therefore, it is reasonable to assume that both investor sentiment and attention within the same country can be highly reliable, and indices tend to influence one another from FTSE 100 to the FTSE All share. This contagious effect strengthens the reliability of using the FTSE 100 as the source for constructing proxies to gauge investor sentiment and attention.

Investor Sentiment

Following Hudson and Green (2015) I construct Advances-Decline Ratio (AVCD), Smart Money Flow Index (SMART. index), which is a substitute for Money Flow Index, (MFI), Put-call Volume ratio (PCV), Put-call Open Interest ratio (PCO), 30 Days Relative Strength Index (RSI.30D) and 30-days Implied Volatility Index (IVI.30D). Additionally, I include 2 sentiment indicators from European Central Bank (ECB), composite indicator of systematic stress (CISS) and country-level financial stress composite indicator (CLIFS) for the UK.

Prior to conducting Principal Component Analysis (PCA) on the 9 investor sentiment proxies and extracting the first component to construct a composite index SENT as predictor, I standardize the proxies to the [-1,1] range. By standardizing the variables, this ensures that no single proxy dominates the construction of the aggregate SENT predictor. PCA is a statistical technique used to reduce the dimensionality and identifying the most important patterns in the data. In our case, the first component of the PCA captures the overall variation in the 9 sentiment predictors, it accounts for 40.13% component variance. It has the capability to provide a summary measure of the underlying investor sentiment towards the UK stock market. By constructing this composite index SENT as a predictor to fit into machine learning models for stock return prediction, we can capture the overall sentiment of investors towards the market in a more concise and meaningful way. (Appendix A.3 and Appendix C.3) provide extensive overview of the definition and constructions of the proxies pre-PCA.

Investor Attention

I follow J. Chen et al. (2022) to construct abnormal trading volume (aavol), extreme returns (aeret), and Google search volume (GSV). Additionally, I refer to Cai et al. (2022) to construct upper aggregate limit-hits (ualhits). To ensure comparability across the different proxies, I standardize them within the [-1,1] range and conduct PCA on the 4 investor attention proxies. Finally, I extract the first principal component to construct an aggregate ATTENT predictor that captures overall signals from the underlying predictors. The first component (PC1) explains

34.76% of components variance, which means the ATTENT predictor able to capture significant part of common variation among the proxies. (Appendix A.4 and Appendix C.4) provided detailed definition and constructions of the proxies pre-PCA.

3.1.4 Data Pre-Processing

The majority of firm-level characteristics used in constructing predictors are made available to the public with a time lag, to avoid forward looking bias. I follow Gu et al. (2020) to lag annual variables by at least 6 months, and monthly variables by at least 1 month. Another issue in the dataset is that the presence of missing values (NAs). (Appendix B.1) showed that most of the missingness in each predictor is approximately around 20%-30%. Instead of imputing with cross-sectional median values, I implement Kalman filter imputation to replace the missing values. Kalman filter imputation (Moritz & Bartz-Beielstein, 2017) is a statistical method that can be used to estimate the values of missing data in a time-series dataset. It is particularly useful when dealing with missing data that have a temporal component, as it takes into account both the current and past values of the variable in question to make an estimate of the missing value. Secondly, it can provide more accurate estimates of missing values by incorporating additional information about the underlying structure of the data. Specifically, the Kalman filter takes into account the observed values of the variable, as well as any relevant trends or patterns in the data, to make a more informed estimate of the missing value. Furthermore, the benefit of Kalman filter imputation is that it can handle missing data that are not missing completely at random (MCAR). In other words, it can handle situations where the missingness may be related to other variables in the dataset. This is important because MCAR assumptions are often unrealistic in practice, and other imputation methods may not be appropriate in these situations. In short, I employ the Kalman filter imputation method to impute missing values based on other observed characteristics, past observations, and information from other firms cross-sectionally. The Kalman filter imputation shares a similar concept of missing value imputation as in Beckmeyer and Wiedemann (2023) model, which adapts a Natural Processing Language (NPL) model with a non-linearity assumption to impute financial missing data. However, the key difference between the two methods is that the Kalman filter method uses a linear dynamic system model to capture the time-varying nature of the data, while the NPL model in Beckmeyer and Wiedemann (2023) approach assumes non-linear relationships between the observed and missing variables. Furthermore, (Appendix B.1) reveals that some firm-level predictors exhibit highly skewed and leptokurtic distributions, and outliers have been detected. To address this issue, I follow the approach of Gu et al. (2020) and Freyberger et al. (2020) by cross-sectionally ranking all stock characteristics each month and standardizing features to the [-1,1] interval to prevent influence of outliers and ensure that each predictor has an equal impact on the analysis. (Appendix B.2) shows an overview of the summary statistic of firm and macroeconomics predictors after imputation and standardization, as well as (Appendix D.1) shows a correlation heatmap between across firm and macroeconomics predictors. Most of the predictors do not exhibit multicollinearity. The standardization procedure for SENT and ATTENT predictors differed slightly from that of the firm and macroeconomics predictors. Before PCA, the raw proxies for SENT and ATTENT have been standardized as mentioned above, and the resulting composite indices (SENT, ATTENT) are then standardize again within the [-1,1] interval. (Appendix C.1) presents the summary statistics for proxies, as well as the aggregate indices SENT and ATTENT. The pre-PCA correlation heatmap for the proxies, on the other hand, can be found in (Appendix D.1). Additionally, (Appendix C.2) showcases time series plots depicting investor sentiment and attention. Notably, significant fluctuations are observed in both SENT and ATTENT prior to 2010, following the aftermath of the 2008 financial crisis, as well as around 2020 during the Covid-19 crisis. These fluctuations vividly reflect the shifting beliefs of investors in response to major economic and market events at the aggregate level. I also checked the log transformed excess returns across years (2000-2022) and did not observe any extreme or abnormal patterns. The number of firms in the dataset is evenly distributed across the 23 years, ranging from 14,000 to 22,000. The median of the log excess return in each year is centered at 0. The most extreme excess return was observed in 2022 at the 100th percentile, with a value of 5.394. (Appendix B.3) shows the summary statistic for log excess return. After performing imputation, standardization, and exploratory data analysis, I am left with two sets of data. The first set consists of 49 predictors, which include both firm-level and macroeconomic predictors. The second set has 51 predictors, as I have extended it with the aggregated predictors SENT and ATTENT.

3.2 Research Design

I form 2 groups of study subject:
1. **Benchmark sample**: Firms and macroeconomics predictors, with 49 predictors are fit into the machine learning models to predict stock return.
2. **Extended sample**: composite indices SENT and ATTENT are added to the existing firms and macroeconomics predictors, resulting in total of 51 predictors, are then fit into machine learning models to predict stock return.
The benchmark sample and extended sample are then analyzed at the stock level and portfolio level, which I closely follow the methodology suggested by Gu et al. (2020).

Hypothesis Testing:
First hypothesis:
H0: One/more Machine Learning models do not outperform the ordinary least square regression (OLS) in terms of stock level return prediction, under both benchmark and extended sample.
H1: One/more Machine Learning models outperform the OLS in terms of stock level prediction, under both benchmark and extended sample.

Second hypothesis:
H0: Stock level prediction in extended sample does not outperform the benchmark sample.
H1: Stock level prediction in extended sample outperform the benchmark sample.

Third hypothesis:
H0: Extended sample's long-short portfolio return and/or cumulative return is lower or equal to benchmark sample long-short portfolio return.
H1: Extended sample's long-short portfolio return and/or cumulative return is higher than the benchmark sample long-short portfolio return.

Fourth hypothesis:
H0: Extended sample's adjusted portfolio return is lower than the benchmark sample's adjusted portfolio return.
H1: Extended sample's adjusted portfolio return is higher than the benchmark sample adjusted portfolio return.

Fifth hypothesis:
H0: Investor sentiment and attention do not improve machine learning prediction-based portfolio return.
H1: Investor sentiment and attention improve machine learning prediction-based portfolio return.

Stock Level Analysis:
The primary objective at the stock level analysis is to predict stock returns, and this is primarily achieved through the use of machine learning methods. To begin, I will discuss the objective function used to estimate model parameters, the process of data splits, and the importance of hyperparameter tuning through validation. Then, I will introduce statistical model which describes the general functional form of the method used for predicting risk premiums. Next, I will present various machine learning models and a benchmark regression model (OLS) that are employed in this context. Moreover, I will explore different techniques that measure the explanatory power and inference-making ability derived from machine learning models. Lastly, the study will delve into the evaluation metrics used to assess the out-of-sample performance of stock-level predictions made by machine learning models.

Portfolio Level Analysis:
In the portfolio level analysis, I begin by selecting the best prediction model for stock level returns from both the benchmark and extended samples. Then, the realized excess return is sorted based on the model's predicted excess return for the next month. Finally, I create an equal-weighted zero-net-investment portfolio, also known as a long-short portfolio, by buying stocks with the highest expected returns (decile 10) and selling stocks with the lowest expected returns (decile 1). All returns are based on an out-of-sample testing period of 8 years.

3.3 Data Splitting and Tuning via Validation

The data is time-series based, it contains 23 years from January 2000 to December 2022. I split the sample in three parts, 10 years for training set (January 2000- December 2009), 5 years for validation set (January 2010-December 2014), the remaining 8 years for out-of-sample testing set (January 2015-December 2022). It is crucial to retain the temporal ordering of the time series data, therefore, k-fold cross validation is not used in this study. This prevents the use of future information in my machine learning models. Since, the study data is large, the splitting techniques should be convincing enough to prevent overfitting issue in the machine learning models.

Another technique to prevent or reduce overfitting issue in the machine learning models is through regularization, which tune the optimal values for the hyperparameters. Hyperparameter tuning is more an art than knowledge, searching for an optimal values, might require exhaustive trial and errors, and there is limited theoretical advice for how to "optimize". Therefore, I utilize the random grid search for hyperparameter tuning, which is an automated process to search the optimal value within a specified hyperparameter grid. Hyperparameter tuning is critical, because they control the model complexity as well as reduce overfitting for the in-sample set, and enhance the out of performance predictability.

First, I use the training sample to estimate the model, considering a specific set of tuning parameters. Secondly, the validation sample is used to fine-tune the hyperparameters. The goal during validation is to minimize the root mean square error (RMSE), which serves as the objective. To accomplish this, an iterative Random Search is performed, exploring different combinations of hyperparameters. At each iteration, the model estimation is updated using the training data, based on the current set of hyperparameter values.

The final model is estimated using the optimal combination of hyperparameters that yield the lowest RMSE, as determined by the validation sample. In essence, the validation process acts as a pseudo-out-of-sample forecast and serves as an input to the estimation.

Finally, the out-of-sample testing sample is reserved solely to evaluate the predictive performance of the machine learning models. This sample is not used for estimation or hyperparameter tuning, ensuring an unbiased assessment of the models' performance.

3.4 Statistical model

I closely follow to the general prediction model outlined by Gu et al. (2020) and Hanauer and Kalsbach (2023) in my study. Specifically, the asset's excess return is determined by an additive prediction error model.

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1} \tag{1}$$

$$E_t(r_{i,t+1}) = g^*(z_{i,t}) \tag{2}$$

the $r_{i,t+1}$ represents the realized excess stock return, which is obtained by subtracting the risk-free rate from the stock return. $E_t(r_{i,t+1})$ is the conditional expected excess return.ϵ is the prediction error term. The $z_{i,t}$ is P dimensional vector of stock features known at time t for predicting individual stock return at $t + 1$. The function $g(\cdot)$ denotes a flexible model that incorporates these predictors for estimation purposes. In the case of machine learning models, $g(\cdot)$ is approximated by some function $g(z_{i,t}, \theta, \rho)$, where θ denotes a vector of coefficients derived from the underlying training data with respect to ρ and a specific loss function L. The hyperparameters ρ is based on user's configuration, it needs to be optimized with respect to L using the available data. The specific functional form of $g(\cdot)$ depends on the chosen model family, encompassing possibilities such as linear or non-linear, as well as parametric or non-parametric formulations.

3.5 Forecast models

3.5.1 Robust Ordinary Least Square Regression (OLS-Huber)

Ordinary Least Square (OLS) regression is the simplest predictive model in our study. This model assumes a linear relationship and does not incorporate non-linear effects or interactions between predictors (Gu et al., 2020). Consequently, OLS may perform poorly when a large number of predictors are included in the model. This is primarily due to the challenges posed by high dimensionality and a potentially low signal-to-noise ratio. In situations where the number of predictors is close to or exceeds the number of observations, Ordinary Least Square (OLS) regression has a tendency to overfit noise instead of capturing the underlying signal. This problem becomes even more pronounced in settings where the signal-to-noise ratio is low, as highlighted by Drobetz et al. (2019).

The simple linear model with the conditional expectations $g(\cdot)$ can be approximated by a linear function, with predictors and parameter vectors θ,

$$g(z_{i,t}; 0) = z'_{i,t}\theta \tag{3}$$

The primary objective of ordinary least squares (OLS) regression is to minimize the *l2* objective function, also known as the least square function.In financial returns and stock predictor variables, heavy-tailed distributions are common. However, the *l2* function tends to heavily penalize large errors compared to small errors, making linear regression sensitive to outliers.

This sensitivity undermines the stability of OLS regression forecasts. To address this issue, this study introduces the Huber Loss objective function, which mitigates the negative impact of heavy-tailed observations and ensures robust linear regression predictions. The Huber Loss objective function (Gu et al., 2020) is defined as follows:

$$L_H(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} H(r_{i,t+1} - g(z_{i,t}; \theta, \xi)) \tag{4}$$

where

$$H(x;\xi) = \begin{cases} x^2 & \text{if } |x| \leq \xi \\ 2\xi|x| - \xi^2 & \text{if } |x| > \xi \end{cases}$$

The Huber Loss function $H(\cdot)$ incorporates both squared errors for smaller errors and absolute errors for larger errors. In the R caret rlm package, the optimal Huber threshold value ξ is predetermined as 1.345 to tune the loss function. This threshold value helps balance the extended of different error magnitudes and enhances the robustness of the regression model.

Despite its simple model specification, ordinary least squares regression (OLS) with a robust Huber Loss function can be considered a suitable benchmark when comparing prediction performance against other machine learning models. By incorporating the Huber Loss function, OLS becomes more resilient to outliers and exhibits improved robustness in its predictions. Therefore, it can serve as a reliable point of reference when evaluating the performance of other more complex machine learning models.

3.5.2 Generalized Linear Model

The Generalized Linear Model (GLM) is a flexible modeling approach that overcomes the limitations of simple linear functional forms by incorporating a nonparametric model. GLM achieves this by introducing a K-term spline series expansion, which allows for non-linear transformations of the original predictors as additional terms in the model. This increases flexibility and further enhances the model's ability to capture complex relationships. However, to mitigate the risk of overfitting, regularization techniques are often employed as a complement to the GLM. The general concept of the generalized linear model is illustrated by Gu et al. (2020):

$$g(z; \theta, p(\cdot)) = \sum_{j=1}^{p} p(z_j)\theta_j \tag{5}$$

In order to construct the generalized linear model (GLM), I employ the "h2o: R Interface for H2O" (2022) open source machine learning library in R. Within H2O, regularization is implemented using Elastic Net Penalties, which combines the LASSO penalty (*l1*) and ridge regression penalty (*l2*). The *l1* penalty promotes sparsity, meaning that it encourages some coefficients to become exactly zero, effectively selecting a subset of variables. On the other hand, the *l2* penalty provides stability and encourages a grouping effect within the model. The grouping effect refers to correlated variables being either dropped or added together as a group.

To achieve the optimal outcome, the hyperparameter is introduced. It controls the distribution of the elastic net penalties between *l1* and *l2*. H2O automates the hyperparameter tuning process through Random Search and the hyperparameter's search range for α can be referred under Appendix E.1.

3.5.3 Boosted regression trees and random forests (GBM, DRF)

Regression trees inherently capture multi-way interactions and nonlinearity, eliminating the necessity for additional predictors to represent these effects (Drobetz et al., 2019). Since my study did not include interaction terms and non-linear predictors, it provides an opportunity to assess whether these advantageous specifications can increase the model's predictability. The process of regression trees involves adaptively splitting the dataset into groups of similar observations. Beginning with an initial node, the optimal split variable and value are determined to minimize the error (RMSE) within each partition. This iterative growth of the tree results in leaves with minimized impurity, and predictions are made based on the average of observed values within each leaf. While regression trees are capable of capturing interactions, being invariant to transformations of monotonic predictors and accommodating different datatypes, they are prone to overfitting and may require regularization. Ensemble methods such as bagging and boosting can be employed to aggregate predictions from multiple trees, addressing the issue of overfitting and improving forecasting performance (Gu et al., 2020; Drobetz et al., 2019; Azevedo et al., 2022).A general idea of the regression trees function is approximated as follows, for detailed specification explanation, see Gu et al. (2020):

$$g(z_{i,t};\theta,K,L)=\sum_{k=1}^{K}\theta_k\nVdash_{z_{i,t}\in\mathcal{C}_k(L)} \tag{6}$$

Distributed Random Forest (DRF) operates by constructing a multitude of decision trees using bootstrapped samples from the original dataset, which creates strong ensemble learner. It can be implemented either for classification or regression, in this study, it is specifically used for regression. Each tree as a weak learner is grown independently, utilizing a random subset of predictors at each split through dropout method. Dropout method is used to mitigate high correlations between bootstrap-replicated trees. The final prediction of a Random Forest model is obtained by aggregating the predictions of all the individual trees, typically through averaging. This approach helps to reduce overfitting, improve generalization, and provide robustness against outliers and noisy data. I have also implemented the DRF model using the H2O library. The main hyperparameters include the number of trees *(ntrees)*, maximum tree depth *(maxdepth)*, and the number of randomly selected predictors *(minrows)*. The optimal hyperparameters are then determined using H2O's Random Search Grid. See Appendix E.1 for the specified values of the hyperparameters.

On the other hand, Gradient Boosting Machine (GBM) builds an ensemble of trees in a sequential manner. Initially, a single tree is created, and its predictions are used to calculate the residuals or errors. The next tree is then built to predict the residuals of the previous tree, and this process continues iteratively. Each subsequent tree is designed to minimize the residuals from the previous trees. Finally, the predictions from all the trees are combined to form the final prediction. GBM is known for its ability to handle complex relationships and effectively model nonlinearity. GBM is prone to overfitting due to its sequential nature, where each tree is built to correct the mistakes of the previous trees. To mitigate overfitting, regularization is applied. Similar to random forest, I utilize the GBM model implemented in the H2O R library for our analysis. In order to enhance the accuracy of the GBM model, I tune several hyperparameters. These include the learning rate *(learntree)* to control the contribution of each tree through shrinkage, the maximum tree depth *(maxdepth)* the number of trees *(ntrees)* ,and the minimum number of samples required for splitting *(colsamplerate, samplerate).* See Appendix E.1, which provides grids of suggested hyperparameter values for an effective Random Grid Search approach. This approach aids in finding the best combination of hyperparameter values that maximize the performance of the GBM model.

3.5.4 Deep Neural Network (DNN)

Neural Networks are a class of artificial networks that draw inspiration from the structure and function of the human brain. They are highly parameterized and excel at solving complex problems, capture non-linearity and interaction effect by learning from large datasets, although they can be challenging to interpret. In this study, the H2O library's Deep Neural Network (DNN) is employed. Gu et al. (2020) demonstrated a general formula of feedforward neural networks model:

$$g(z;\theta) = \theta_0^{(1)} + \sum_{j=1}^{n} x_j^{(1)} \theta_j^{(1)} \tag{7}$$

According to Candel and LeDell (2022), DNN is a multi-layer feedforward artificial neural network trained with stochastic gradient descent using backpropagation. The basic principle behind multi-layer feedforward artificial neural networks is that DNNs consist of multiple layers of interconnected nodes called neurons. Each neuron performs a weighted computation on its inputs. Then, they map inputs to outputs in a unidirectional manner. For instance, by using predictors as a weighted average of input values (x), the DNN predicts the output (y) such as realized return. DNNs consist of multiple layers of interconnected nodes called neurons. However, Stochastic gradient descent (SGD) is a fundamental optimization algorithm used in training deep neural networks. During the training process, SGD iteratively updates the network's parameters by computing the gradients of the loss function with respect to these parameters. The ”stochastic” aspect refers to the fact that the gradients are estimated using small subsets of the training data, known as mini-batches, rather than the entire dataset. By employing SGD with backpropagation, DNNs can efficiently adjust their internal weights and biases to minimize the difference between the predicted output and the actual output. This iterative process of forward propagation (computing predictions) and backward propagation (updating parameters based on computed gradients) enables DNNs to learn intricate patterns and representations from the input data.

Complex models like Deep Neural Networks (DNNs) often suffer from overfitting issues. To mitigate this problem, it requires regularization through Random Grid Search. In this study, I explore various hyperparameters, including activation functions types *(activation)*, hidden layer size *(hidden)*, l1 and l2 regularization, and input dropout ratio, with the goal of enhancing generalization.

H2O's hidden layer size *(hidden)* parameter offers a unique advantage by allowing us to avoid constructing separate architectures for NN1-NN3. Instead, it allows me to predefine the number of neurons suggested by (Gu et al., 2020) within the NN1-NN3 architectures. The Random Search algorithm then searches for the best architecture among NN1-NN3 within the hyperparameter grid. The hidden layer size hyperparameter (hidden) can be specified as follows: (32, 16, 8), (32, 16), (32). The first option represents NN3, which consists of three hidden layers with 32, 16, and 8 neurons respectively. NN2 corresponds to two hidden layers with 32 and 16 neurons, while NN1 represents the shallowest architecture with a single hidden layer of 32 neurons. In this study, the network architecture of NN3 is selected during the regularization process. Therefore, the NN3 represents the DNN in the following result analysis. Regarding other parameters like epochs, learning rate, and batch size, I opt to utilize the default settings provided by H2O for the DNN model. See Appendix E.1 for the hyperparameter specification for DNN model.

3.6 Model Explainability

3.6.1 Variable Importance Plot

I aim to investigate the influential variables in the cross-section of expected return while accounting for other variables in a similar model. The variable importance graphs in this study provide an overview of all models, indicating which of the top 20 predictors in each model contribute to the its own prediction. Following the approach of Gu et al. (2020), variable importance in both ordinary least squares regression (OLS) and generalized linear models (GLM) is defined as the reduction in out-of-sample R-squared R^2_{oos} when setting all values of predictor j to zero while keeping the remaining model estimates fixed. For Random Forest and Gradient Boosting Machine, variable importance is defined as the mean decrease in impurity, as described by Breiman (2001) and Friedman (2001). Lastly, for Deep Neural Networks, variable importance is computed using the Gedeon Method, which assesses the contributions of input nodes to the output (Gedeon, 1997).

In many cases, Deep Neural Networks (DNN) are considered black-box sophisticated prediction models with low interpretability. However, recently various methods have been developed to offer interpretability from two perspectives: Global Model-Agnostic methods and Local Model Agnostic Methods. Global methods characterize the average behavior of the prediction derived from machine learning model. On the other hand, local methods explain individual predictions (Molnar, 2020). For example, the Partial Dependence Plot (PDP) provides a global viewpoint, while the Individual Conditional Explanation Plot (ICE) offers a local viewpoint.

3.6.2 Partial Dependence Plot

Partial dependence plot (PDP) gives a visual illustration of the average marginal effect of a predictor on the response, while holding other variable constant (*Model Categories x2014; H2O documentation — docs.h2o.ai*, n.d.). PDP strongly assumes independence between the features. However, PDP might suffer from hidden heterogenous effect with where data points for that predictor with half of both positive and negative impacts, ultimately canceling each other out and resulting in zero overall effect at average level. In such cases, Individual Conditional Expectation (ICE) plots can be used to address this issue. In short, PDP allows us to identify overall trends and understand the general impact of each input on the model's output.

3.6.3 Individual Conditional Expectation Plot

According to Molnar (2020), the Individual Conditional Expectation Plot (ICE) is the building block of the PDP plot. Both of them graphically demonstrate the marginal effect of a variable on the response. The PDP displays the average effect of the feature, whereas ICE demonstrates the effect for a specific instance. It captures the variation in fitted values across the range of a covariate, shedding light on potential heterogeneities and their extent (Goldstein, Kapelner, Bleich & Pitkin, 2015).

For example, in this study, the ICE plot function plots the effect for each decile. Each ICE percentile curve defines the conditional relationship between the interested variable x_s and the estimated response function $\hat{f}$ while holding other variables x_c constant. The shape, slope, and direction of the ICE curves can provide insights into the strength and direction of the marginal effect. This allows us to assess how changes in the predictor influence the predicted outcome. By examining ICE plots, it is possible to observe cases where increasing a certain covariate is associated with higher predicted values, indicating deviations from average behavior.

However, in situations where the curves in the plot have a wide range of intercepts and overlap, making it difficult to discern heterogeneity, a centered ICE (c-ICE) can be helpful. c-ICE helps to center the curves at a certain point $(x^*, \hat{f}(x^*, x_{ci}))$ in the feature and display only the difference in the prediction from this point. If x^* is the minimum value of the interest variable x_s , this ensures the ICE percentile curve originates from a common point, thus removing the differences in level due to different x_{ci}. However, if x^* is the maximum value, it reflects the cumulative effect relative to a base case (where x^* is the minimum value)(Goldstein et al., 2015). It is common to select the curves at the minimum and maximum points for ICE plot analysis. In my case, I anchor at both the minimum point and the maximum point for the particular covariate and see the variations of predictions across the 0th percentile, 50th percentile, and 100th percentile ICE curves.

Both PDP and ICE plots offer a comprehensive view of local and global interpretability of the model, specifically regarding the marginal effect of predictors on the response variable. In this study, I apply the PDP and ICE plots based on the best prediction model for both the benchmark and extended samples. I specifically focus on the effects of SENT, ATTENT, and the highest variable importance feature from the best prediction model. Finally, I visually compare how the results in benchmark sample and extended sample differ from each other.

3.7 Stock Level Performance Evaluation Metrics

3.7.1 Out-of-sample R-squared (R^2_{oos})

To measure the out-sample predictive performance for each forecasting model's stock excess return, I follow Gu et al. (2020) to use R^2_{oos} as evaluation metric:

$$R^2_{oos} = 1 - \frac{\sum_{(i,t)\in\mathcal{P}_3}(r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{P}_3} r^2_{i,t+1}} \tag{8}$$

Where $\mathcal{P}_3$ indicates the data sample never enter into models' training or validation.To evaluate each model, the R^2_{oos} metric combines prediction errors across firms and over time, offering a comprehensive assessment at the panel level. I benchmark the R^2_{oos} against a forecast value of zero instead of noisy historical averages. Because using historical mean stock return might artificially lower the threshold for determining good forecast performance (Gu et al., 2020).

3.7.2 Diebold-Mariano Test

I apply the Diebold and Mariano (2002) test to assess differences in out-of-sample predictive accuracy between two prediction models in stock level analysis. Based on the work of Gu et al. (2020), the Diebold-Mariano test has been modified to compare the average prediction errors across the cross-section of each model, as opposed to comparing errors among individual returns. The test statistics is defined as:

$$\mathrm{DM}_{12} = \frac{\bar{d}_{12}}{\sigma_{\bar{d}_{12}}} \tag{9}$$

The $\bar{d}_{12}$ denotes as the mean, $\hat{\sigma}(\bar{d}_{12})$ denotes as the Newey-West standard error of $d_{12,t}$ over the testing sample.

3.8 Portfolio return Performance Evaluation

To evaluate the performance of the portfolios, I focus primarily on the portfolio level returns for both the benchmark and extended samples. This assessment is started with an overview of metrics such as the average realized return, predicted return, Sharpe ratio and standard deviation of the high minus low portfolio deciles. Additionally, I analyze the cumulative portfolio returns graphically, considering the returns of the long leg, short leg, and overall portfolio. To provide a benchmark, I compare these cumulative returns to the cumulative market return of the FTSE All Share index. Furthermore, I assess the adjusted portfolio returns by using the CAPM, Fama-French 3, 5, and 6 factor models as benchmarks. These Fama-French models, derived from *Kenneth R. French - Data Library — mba.tuck.dartmouth.edu* (n.d.), are applicable to developed markets. Lastly, I investigate whether the inclusion of sentiment and attention as stock level predictors improves the portfolio's returns. To do this, I regress the extended sample's long-short portfolio return on the benchmark sample's long-short portfolio return.

4 Results and Discussions

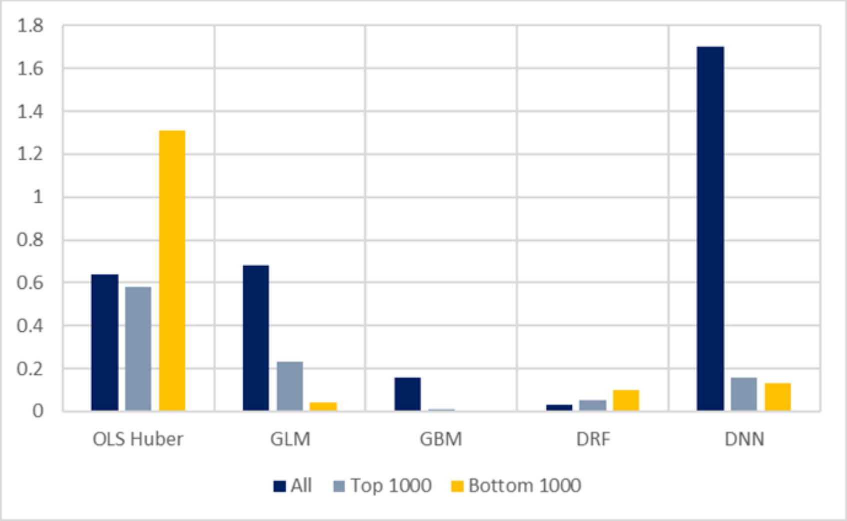
4.1 Stock Level Analysis

4.1.1 The cross-section of individual stocks

Table 1 provides the comparison of linear model and machine learning models in terms of their out-of-sample predictability R^2_{oos} for the benchmark sample. However, Table 2 provides the similar comparison for the extended sample with the inclusion of sentiment and attention predictors. I compare five models in total, starting with Ordinary Least Square with Huber Loss (OLS-Huber), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Distributed Random Forest (DRF) and Deep Neural Network (DNN).

Table 1:
Monthly level out- of -sample stock level prediction for benchmark sample (percentage R^2_{oos})

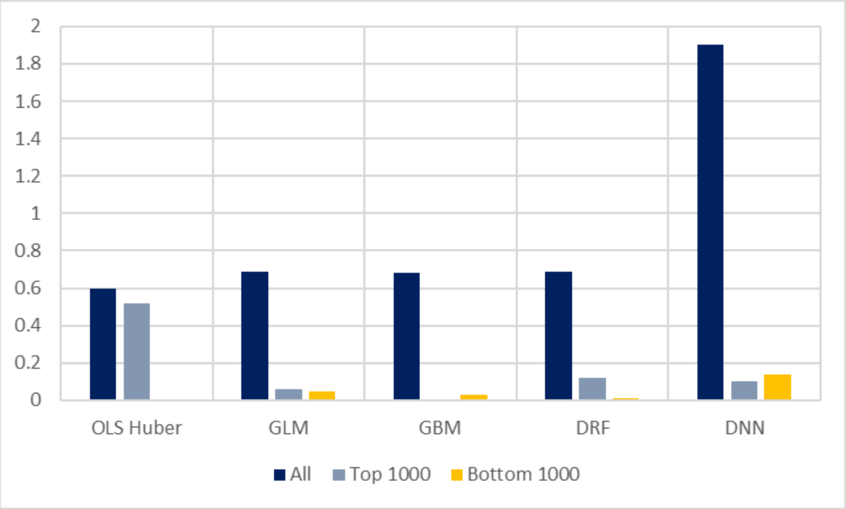
	OLS Huber	GLM	GBM	DRF	DNN
All	0.64	0.68	0.16	0.03	1.70
Top 1000	0.58	0.23	0.01	0.05	0.16
Bottom 1000	1.31	0.04	0.00	0.10	0.13



This table shows the monthly R^2_{oos} (in %) for OLS-Huber, GLM,GBM,DRF and DNN under the benchmark sample with 49 predictors. Huber indicates the use of Huber Loss instead of *l2* loss. I also report the top 1000 stocks and bottom 1000 stocks based on market value. The lower panel provides a graphical comparison of R^2_{oos}.

Table 2:
Monthly level out- of -sample stock level prediction for extended sample (percentage R^2_{oos})

	OLS Huber	GLM	GBM	DRF	DNN
All	0.60	0.69	0.68	0.69	1.90
Top 1000	0.52	0.06	0.00	0.12	0.10
Bottom 1000	0.00	0.05	0.03	0.01	0.14



This table shows the monthly R^2_{oos} (in %) for OLS-Huber, GLM,GBM,DRF and DNN under the extended sample with 51 predictors (with the inclusion of aggregated SENT ATTENT indices as predictors). Huber indicates the use of Huber Loss instead of *l2* loss. I also report the top 1000 stocks and bottom 1000 stocks based on market value. The lower panel provides a graphical comparison of R^2_{oos}.

The first row of Table 1 presents the monthly R^2_{oos} for all stocks from January 2015 to December 2022. Using OLS-Huber, the $R^2_{oos}$is determined to be 0.64%. For the extended sample in Table 2, it yields R^2_{oos} of 0.60%. This shows that the OLS-Huber in the benchmark sample with 49 predictors, has outperformed the R^2_{oos} benchmark of 0% and it achieves a satisfactory performance. However, when the total number of predictors included in the OLS-Huber model is increased under extended sample, the R^2_{oos} slightly decreases, but still remains close to the performance level of the benchmark sample. Both the benchmark and extended samples' OLS-Huber models exhibit a monthly R^2_{oos} in the range of 0.60-0.64%, which can be considered as a threshold for comparing with other machine learning models. In order to demonstrate the superiority of the machine learning models, these models need to first surpass the benchmark R^2_{oos} of 0% and then exceed the performance threshold set by the OLS-Huber model.

The GLM with flexible functional form further increases the out-of-sample predictive performance with R^2_{oos} of 0.68% for the benchmark sample, and 0.69% for the extended sample. The tree-based regression models GBM and DRF's out-of-sample prediction performance fails to outperform the OLS in the benchmark sample, with the reported R^2_{oos} of 0.16% for GBM, 0.03% for DRF. However, GBM yields R^2_{oos} of 0.68% and DRF yields R^2_{oos} of 0.69% in the extended sample. The tree-based regression models have inherently included interaction terms and non-linearity. However, under the benchmark sample, this specification does not enhance predictability and instead introduces redundancy to the models, resulting in underperformance compared to the OLS model. On the contrary, in the extended sample, this specification demonstrates its significance by capturing the effects of sentiment and attention. It suggests the presence of interaction and non-linearity with sentiment and attention, potentially improving performance. The extended sample's tree-based regression models outperform the OLS model and yield performance that is approximately similar to that of the GLM model.

The DNN model (with NN3 architecture) outperforms all the other prediction methods in both benchmark and extended sample. It generates R^2_{oos} of 1.70% for benchmark sample, whereas R^2_{oos} of 1.90% for the extended sample. The DNN model's ability to capture complex relationships, its deep architecture for hierarchical feature abstraction, the utilization of large-scale data collectively contribute to its best performance among the compared methods. The DNN model excels in integrating supplementary features like investor sentiment and investor attention, which serve as valuable signals of investor behavior. By combining these factors with other predictors, the DNN model effectively leverage their impact on stock returns, leading to improved predictive performance.

Additionally, I present the top 1000 and bottom 1000 monthly R^2_{oos} values under row 2 and 3 for both the benchmark sample in Table 1 and the extended sample in Table 2. It is noteworthy that both groups exhibit similar patterns, as they perform poorly in subsamples. The overall success of predictions is not driven by highest or lowest market capitalization stocks, except for the benchmark sample's OLS-Huber model, which shows high predictability for the bottom 1000 stocks. Lastly, the overall monthly R^2_{oos} for extended sample's models is higher compared to the benchmark sample, except the OLS has slight performance deterioration.

Table 3:
Comparison of monthly out-of-sample prediction via Diebold -Mariano Tests for benchmark sample

	GLM	GBM	DRF	DNN
OLS Huber	24.52***	3.67***	1.84*	1.87*
GLM		-8.06***	1.79*	1.82*
GBM			1.83*	1.85*
DRF				3.38***

Table 4:
Comparison of monthly out-of-sample prediction via Diebold -Mariano Tests for extended sample

	GLM	GBM	DRF	DNN
OLS Huber	67.21***	0.99	1.85*	1.88**
GLM		-40.80***	1.56	1.60
GBM			1.84*	1.88*
DRF				3.32***

Table 3,4 show the Diebold-Mariano test statistics comparing the out-of-sample stock level prediction performance among 5 models for benchmark sample and extended sample respectively. The positive numbers indicate the column model outperform the row model. $***p < 0.01$, $** \, p < 0.05$, and $*p < 0.1$ represent the significance level of p-value for difference forecast accuracy between models.

The Diebold-Mariano (DM) test is used to compare the forecast accuracy of two different methods by examining each column's model against each row's model in Table 3 and 4. It is evident that all the non-linear machine learning models outperform the ordinary least square model, as indicated by the positive DM values in the first row. The p-values obtained from the Diebold-Mariano test represent the probability of observing the realized forecast error difference between the two forecast methods, with the significance levels ranging from 1% to 10%. Moreover, the DNN model, which exhibits the highest monthly R^2_{oos},consistently outperforms all other models at a 10% significance level in both the benchmark and extended samples.

4.2 Model Interpretability

4.2.1 Variable Importance

The variable importance for each stock characteristic in each machine learning model is computed based on the techniques described in methodology section.The top 20 predictors with the highest importance for each machine learning model are then normalized to sum up to one. (Figure 1) illustrates the relative importance of these top 20 predictors among the total of 49 variables in each machine learning model for the benchmark sample. On the other hand, (Figure 3) displays the results for the top 20 predictors among the total of 51 variables in each machine learning model for the extended sample.

(Figure 2) however, reports the overall ranking characteristics for all models under benchmark sample, while (Figure 4) reports the results for extended sample. I rank the importance of each characteristic for each model, which sum to 100%. The color gradient shows the most influential characteristics in the darkest tone, the least influential characteristics in lightest tone.

The OLS-Huber, GLM, GBM, and DRF models exhibit a strong consensus regarding the categories of the top 20 stock-level predictors that exert the most significant influence in both the benchmark and extended samples (Figure 1, 3). These common categories include past returns (STReversal, LTReversal), trading friction (beta, SUV), macroeconomics (svar, rf, oas, cp), as well as value and profitability. However, the specific variables within the value and profitability categories vary among the models, and they are relatively less important in tree-based models. Tree-based models tend to prioritize trading friction and macroeconomics categories, displaying a noticeable skew towards them. On the other hand, while the DNN model also shows agreement in terms of the most important variable categories, it does not consider the trading friction category in its stock-level predictions for both the benchmark and extended samples. Additionally, the DNN model draws predictive signals from a broader range of characteristics, and the ranking weightage for each variable except the top ranked, the rest are evenly distributed. These could explain why the DNN model outperforms the other models in every instances. Despite these differences, all models agree that the most important variable is STReversal, with ranking weightages ranging from 6% to 21% across all models in both the benchmark and extended samples.

However, when additional predictors such as investor attention and sentiment (ATTENT, SENT) are included, all models, except OLS, show improved performance. One possible reason is that the inclusion of sentiment and attention as additional predictors alters the variables' composition and reshuffles the ranking within the value and profitability categories among all models. This effect is particularly prominent in the tree-based models. For example, in GBM, which previously placed high emphasis on macroeconomics category in the benchmark sample, the importance of macroeconomics variables now decreases in the extended sample, and the focus shifts to variables within the value and profitability categories. Similarly, in DRF, the composition of the value and profitability categories changes, and macroeconomics variables are replaced with more value category-related variables. These behaviors can be attributed to the selective nature of the models. OLS, GLM, GBM, and DRF tend to prioritize certain predictors' signals while sacrificing others, as evidenced by allocating 0% ranking weightage for some variables shown in (Figure 2, 4). This is why the inclusion of sentiment and attention improves their predictive ability, as it guides the models towards focusing more on firm-related characteristics rather than macroeconomics that improves prediction performance.

Furthermore, aside from changing the perspective on how the machine learning models analyze the original predictors, the new predictors themselves (sentiment and attention) emerge as among the top 20 most important variables. (Figure 3, 4)show that SENT is the second most important variable across all models after STReversal, except for GBM and DRF, which still tend to skew towards trading friction. SENT has a significant ranking weightage ranging from 3% to 14% across all models. ATTENT predictor, compared to SENT predictor, is less important in terms of prediction.

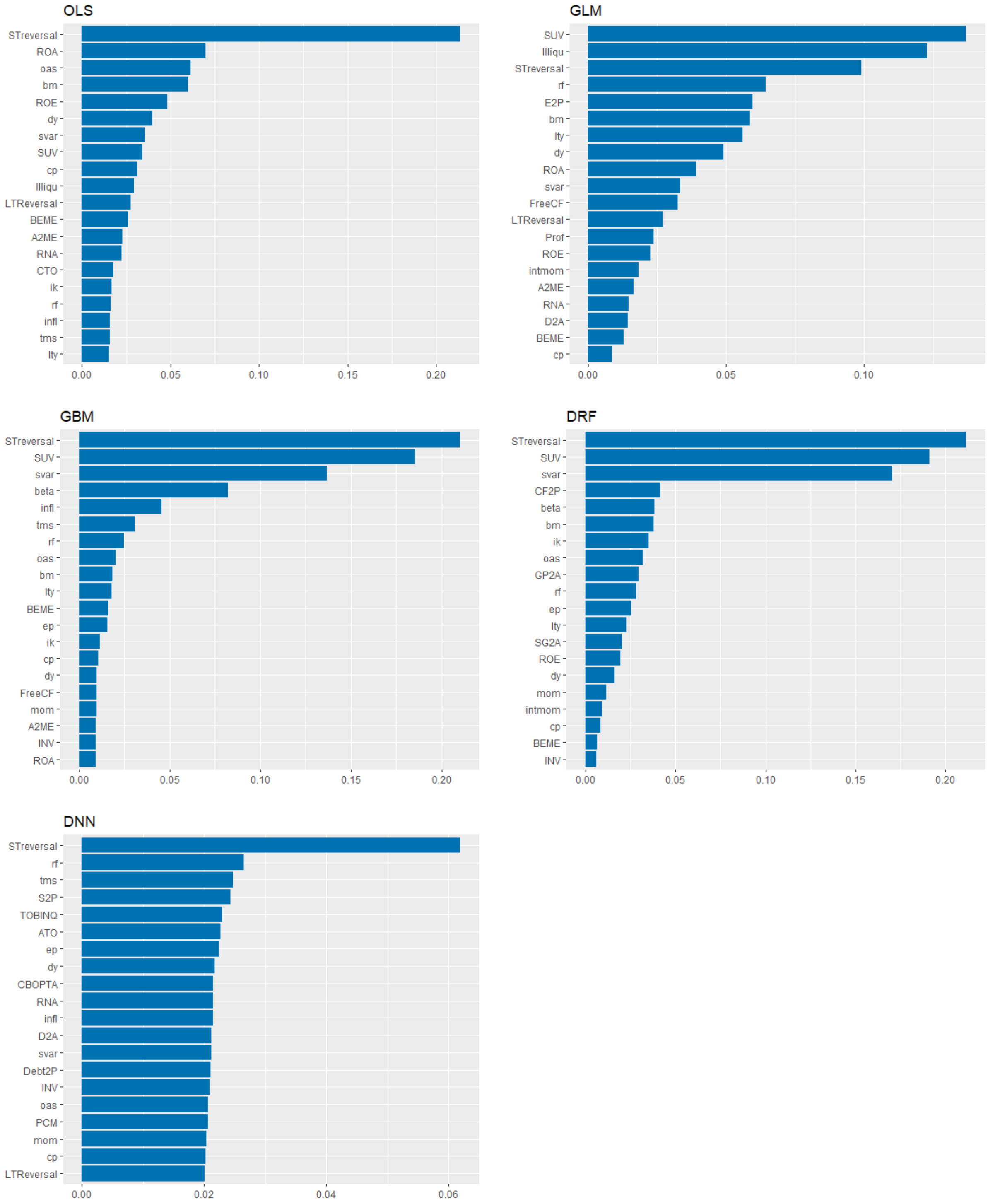


Figure 1: Variable Importance by models for benchmark sample

Variable Importance for the top 20 most important variables in each model for the benchmark sample. The variable importance is normalized to sum of one.

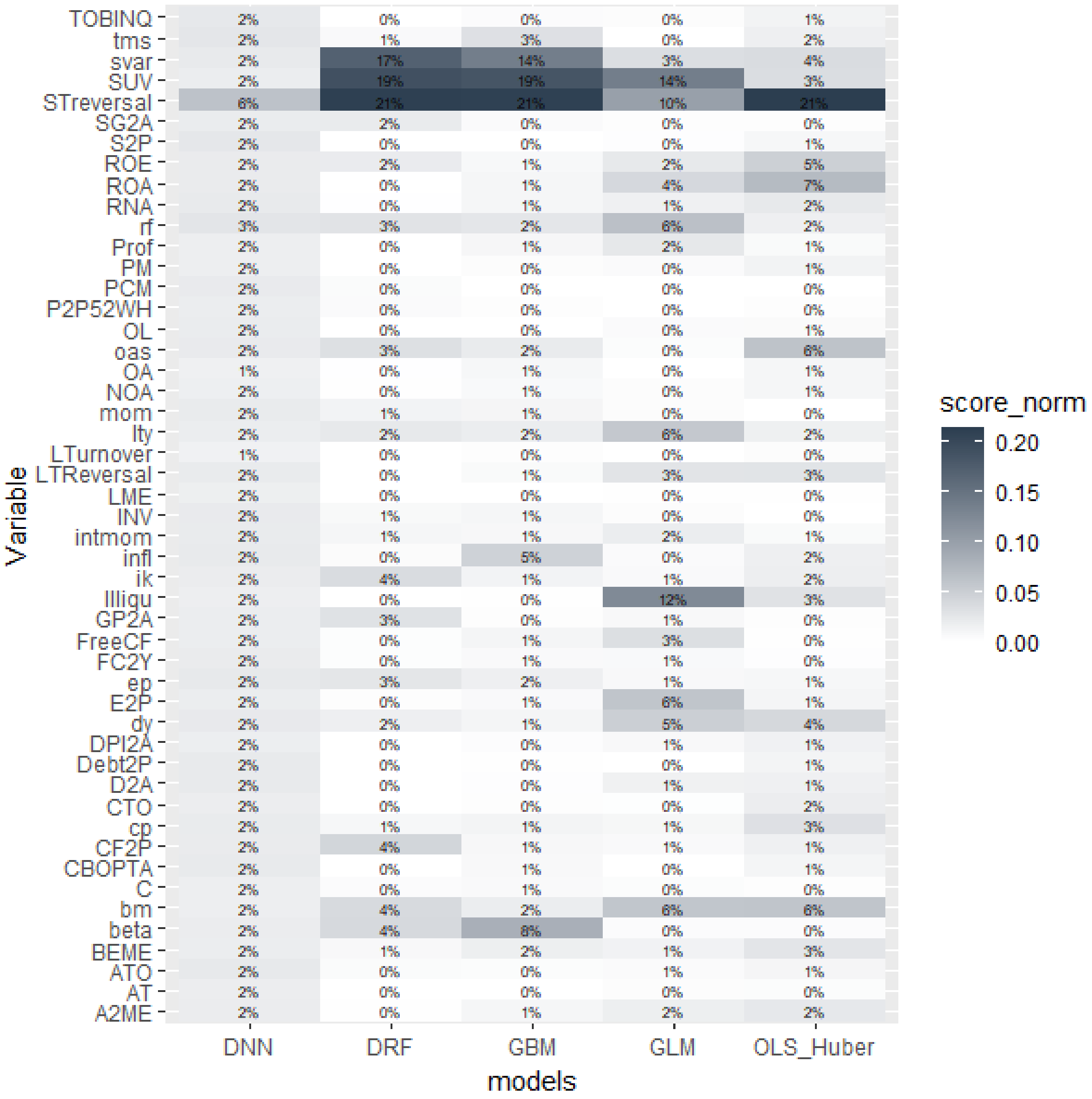


Figure 2: Characteristics importance for benchmark sample

It shows the all rankings of 49 firm and macroeconomics predictors in terms of overall contributions for each model for the benchmark sample. The columns represent all the prediction models. The most influential variable has the darkest color gradient, and the least has the lightest color gradient. The weightage for the variables is based on their rank on the sum of ranks of over each model, which sum to 100%.

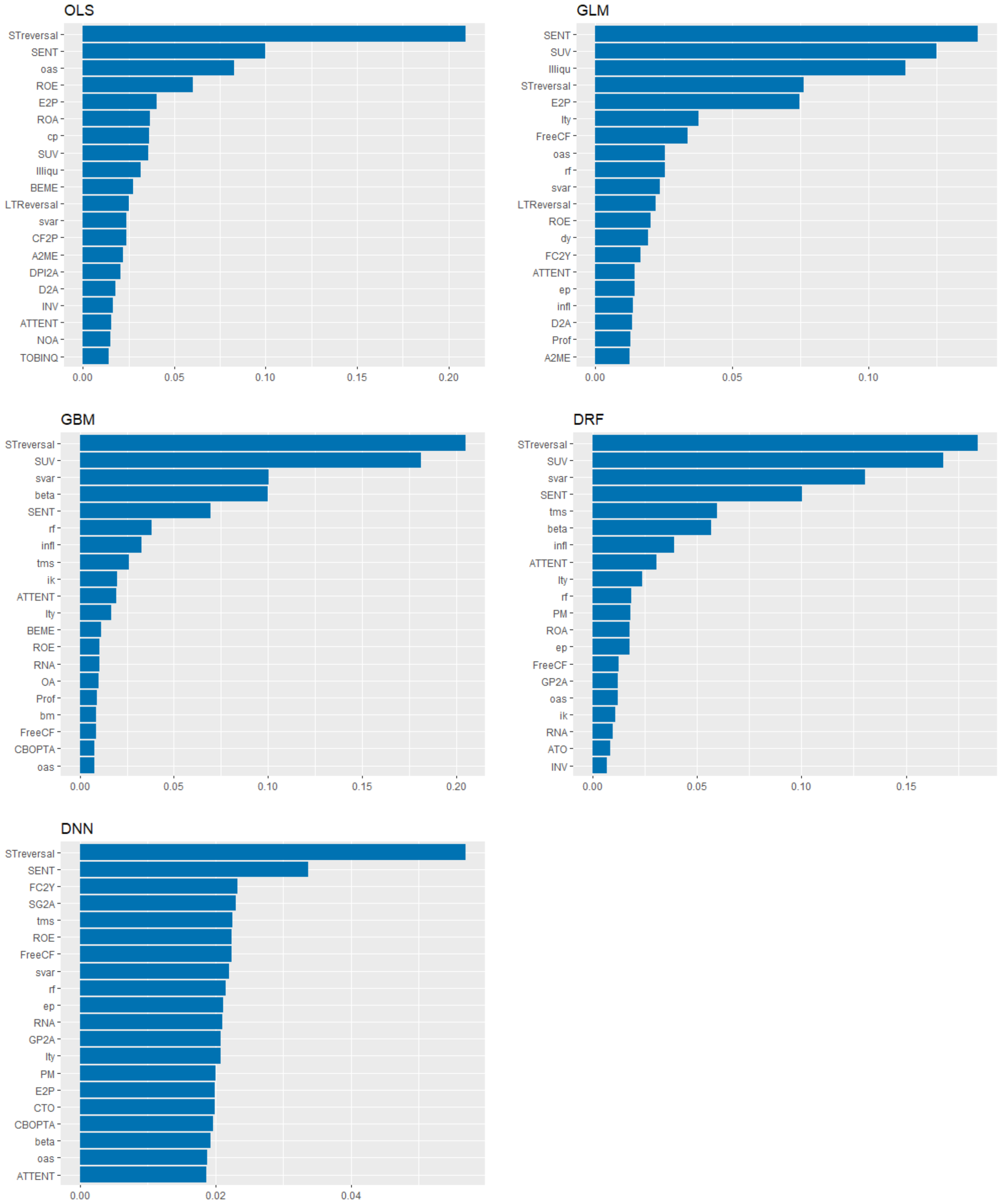


Figure 3: Variable Importance by models for extended sample

Variable Importance for the top 20 most important variables in each model for the extended sample. The variable importance is normalized to sum of one.

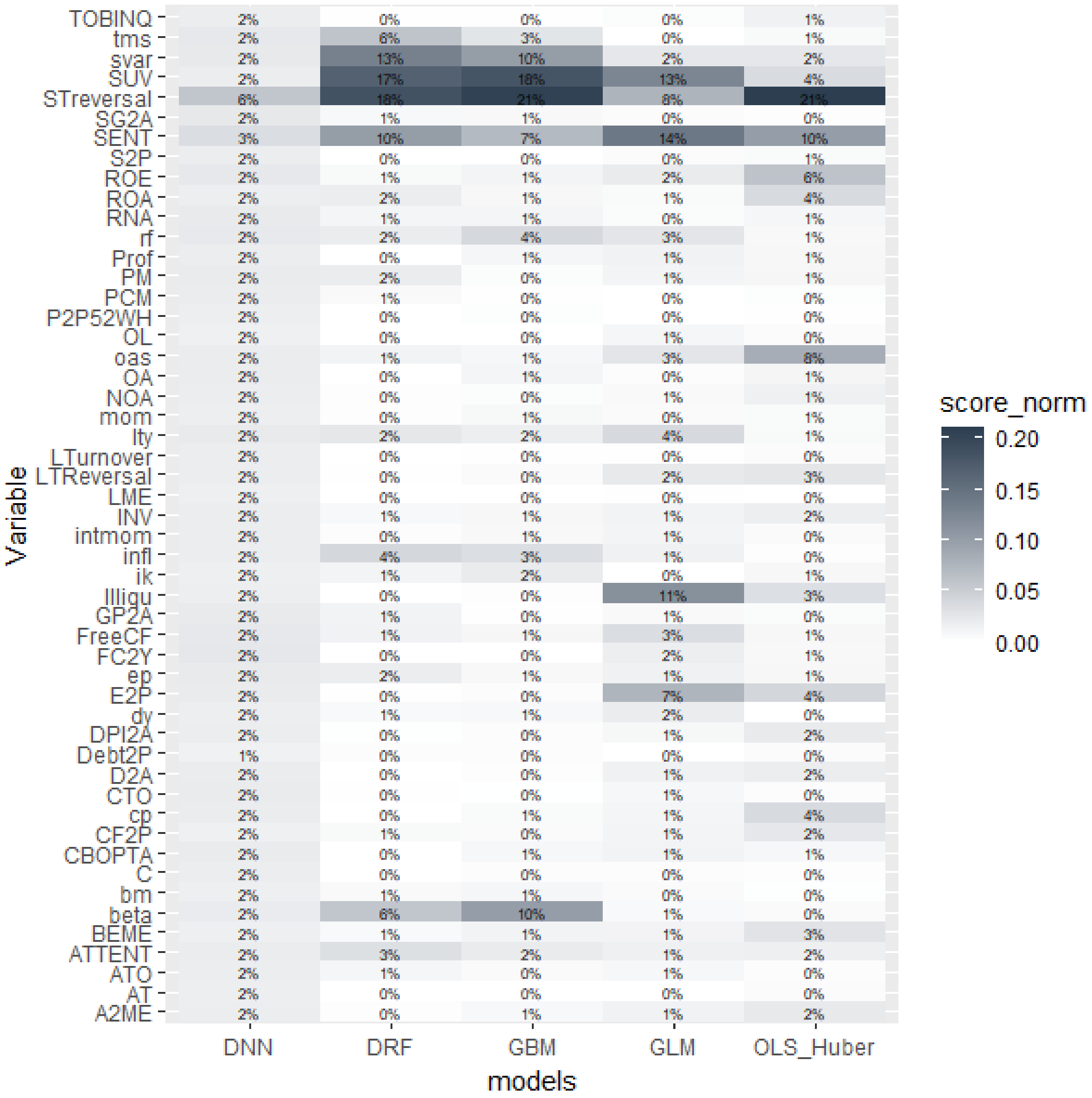


Figure 4: Characteristics importance for extended sample

It shows the all rankings of 51 firm, macroeconomics, aggregate indices SENT and ATTENT predictors in terms of overall contributions for each model for the extended sample. The columns represent all the prediction models. The most influential variable has the darkest color gradient, and the least has the lightest color gradient. The weightage for the variables is based on their rank on the sum of ranks of over each model, which sum up to 100%.

4.2.2 Partial Dependence Plot (PD) and Individual Conditional Plot (ICE)

Based on the out-of-sample prediction R^2_{oos} and Diebold-Mariano (DM) Test, it has been concluded that the DNN model demonstrates significant and superior performance compared to other forecast models. Therefore, the analysis of PD plot and ICE plot will focus exclusively on the best performing model, DNN with its NN3 architecture. The analysis will consider the most important variable, STReversal, under the benchmark sample, whereas, STReversal and additional variables, SENT and ATTENT, under the extended sample. (Figure 5) reveals several inferences. Firstly, the Partial Dependence Plot illustrates an upward increasing but gradual flattening trend for average marginal effect of STReversal under the benchmark sample. This indicates a nonlinear but positive relationship between the stock level return and STReversal. As the magnitude of STReversal marginally increases, the stock level return also increases. However, beyond a value of 0.5, the effect of STReversal on the stock level return prediction diminishes under the benchmark sample. When examining the Individual Conditional Plot, which breaks down the average marginal effect into percentiles ranging from 1 to 100th, all percentiles consistently align with the average marginal effect trend, suggesting no significant heterogeneity in the effects.

In (Figure 6), the average marginal effect of STReversal on the stock level return under the extended sample follows a similar trend to that observed in the benchmark sample (Figure 5). Nevertheless, The average marginal effect of SENT on the stock level return exhibits a positive incremental trend that is less steep compared to STReversal within the extended sample. However, the increasing trend still persists as the magnitude of SENT approaches 1.0.As for ATTENT, its average marginal effect on stock return decreases non-linearly as the magnitude of ATTENT increases. Nevertheless, the ICE plot does not exhibit any noticeable heterogeneous effects among STReversal, ATTENT, and SENT in relation to stock level return under the extended sample.

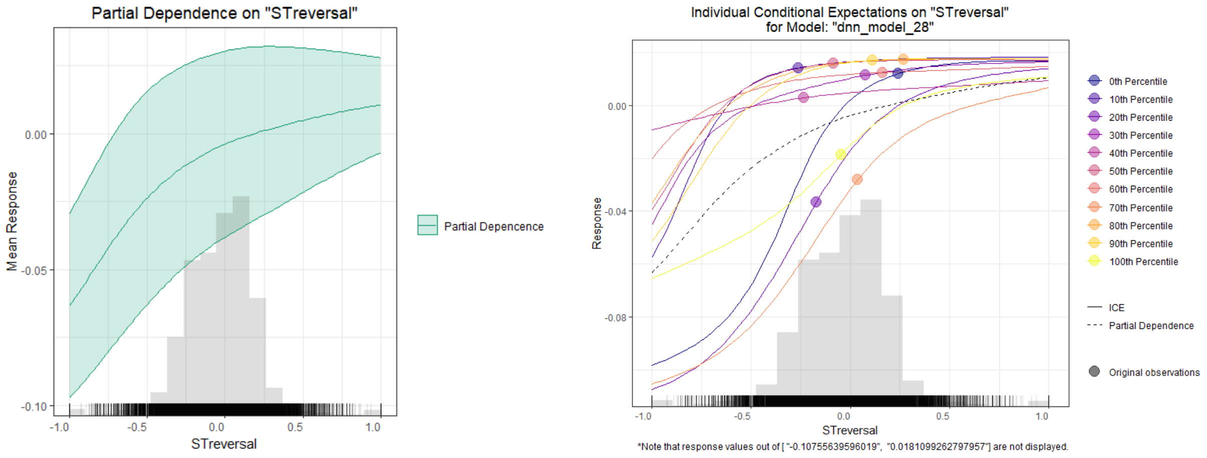
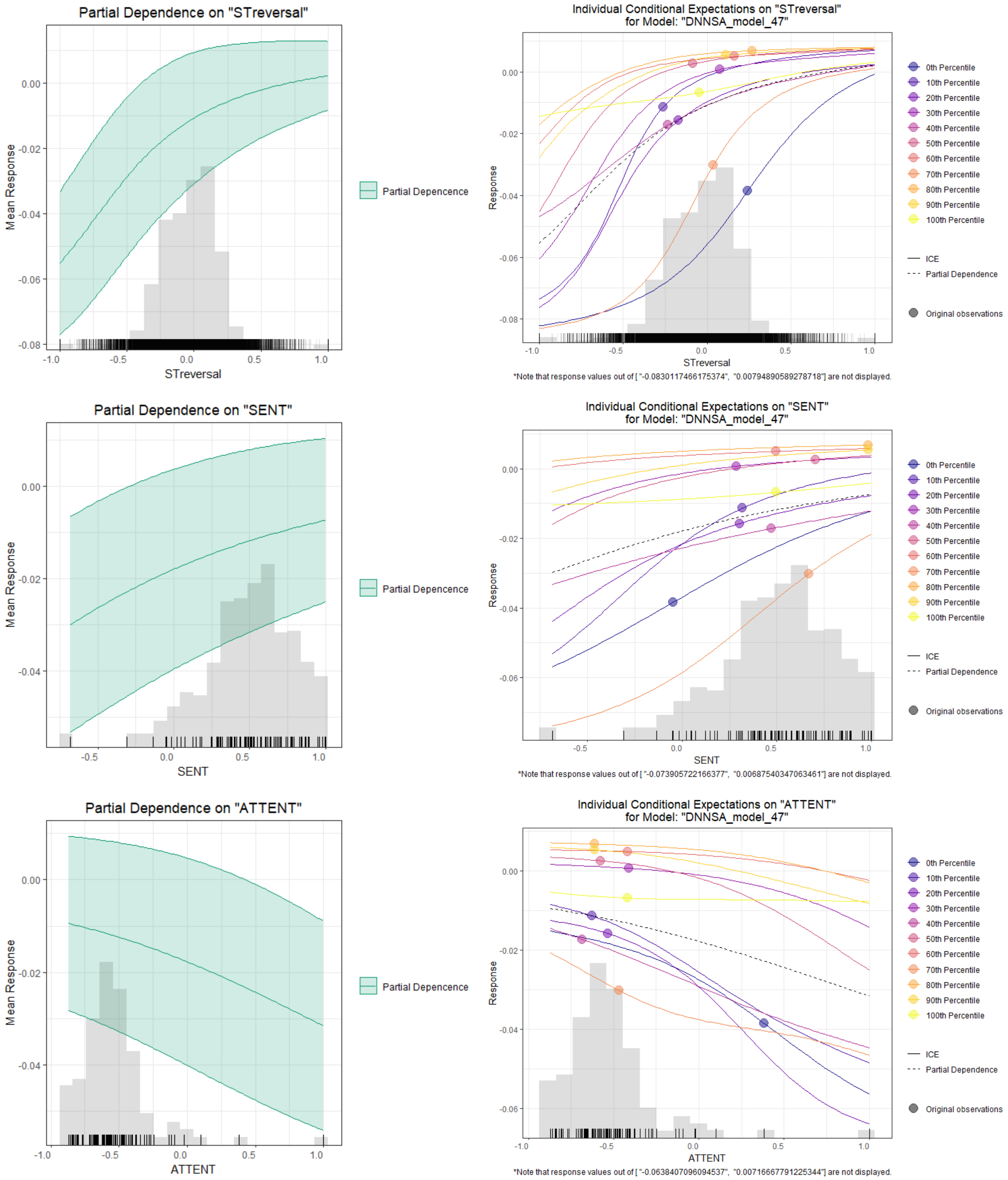


Figure 5: Partial Dependence plot (Left) and Individual Conditional Expectation plot (Right) for DNN's most important variable, STReversal under benchmark sample.

Left panel shows the average marginal effect of STReversal on the stock level return. Right panel shows the single instance marginal effect of STReversal on the stock level return, which based on 1-100th percentile.



Partial Dependence on "STreversal"

Individual Conditional Expectations on "STreversal"
for Model: "DNNSA_model_47"

*Note that response values out of ["-0.0830117466175374", "0.00794890589278718"] are not displayed.

Partial Dependence on "SENT"

Individual Conditional Expectations on "SENT"
for Model: "DNNSA_model_47"

*Note that response values out of ["-0.073905722166377", "0.00687540347063461"] are not displayed.

Partial Dependence on "ATTENT"

Individual Conditional Expectations on "ATTENT"
for Model: "DNNSA_model_47"

*Note that response values out of ["-0.0638407096094537", "0.00716667791225344"] are not displayed.

Figure 6: Partial dependence plot and Individual Conditional Expectation plot for DNN's most important variable STReversal, SENT and ATTENT under extended sample.

Left panels show the average marginal effect of STReversal, SENT and ATTENT on the stock level return. Right panels show the single instance marginal effect of STReversal,SENT and ATTENT on the stock level return, which based on 1-100th percentile.

Table 5:
c-ICE for benchmark sample's STreversal

Predictor's Magnitude (STreversal)	**P0**	**P50**	**P100**
-1	-0.10	-0.04	-0.07
1	0.02	0.02	0.01

This table presents the breakdown of the ICE plot percentile curves for the benchmark sample, focusing on the feature STReversal. It demonstrates the difference in predicted mean values for the 0th, 50th, and 100th percentiles of the response variable, logret, at two fixed point of the feature's magnitude (-1,1).

Table 6:
c-ICE for extended sample's STreversal

Predictor's Magnitude (STreversal)	**P0**	**P50**	**P100**
-1	-0.08	-0.05	-0.01
1	-0.00	0.01	0.00

This table presents the breakdown of the ICE plot percentile curves for the extended sample, focusing on the feature STReversal. It demonstrates the difference in predicted mean value for the 0th, 50th, and 100th percentiles of the response variable, logret, at two extreme fixed point of the feature's magnitude (-1, 1).

Table 7:
c-ICE for extended sample's SENT

Predictor's Magnitude (SENT)	**P0**	**P50**	**P100**
-0.68	-0.06	-0.02	-0.01
1	-0.01	0.00	-0.00

This table presents the breakdown of the ICE plot for the extended sample, focusing on the feature SENT. It demonstrates the difference in predicted mean values for the 0th, 50th, and 100th percentiles of the response variable, Return, at two extreme fixed point of the feature's magnitude (-0.68,1).

Table 8:
c-ICE for extended sample's ATTENT

Predictor's Magnitude (ATTENT)	**P0**	**P50**	**P100**
-0.87	-0.02	0.00	-0.01
1	-0.06	-0.03	-0.01

This table presents the breakdown of the ICE plot for the extended sample, focusing on the feature STReversal. It demonstrates the difference in predictions for the 0th, 50th, and 100th percentiles of the response variable, Return, at two extreme fixed point of the feature's magnitude (-0.87,1).

4.2.3 Centered ICE (c-ICE)

In Table 5, the benchmark's sample predicted values of the response variable ('logret') are illustrated at different percentiles (P0, P50, and P100) for varying magnitudes of the predictor STReversal at (-1 and 1).

At a magnitude of -1, the predicted values of 'logret' were -0.10, -0.04, and -0.07 for the P0, P50, and P100 percentiles, respectively. This suggests that lower magnitudes of STReversal are associated with negative predictions, with a slightly higher value at the P50 percentile. Conversely, at a magnitude of 1, the predicted values of 'logret' were 0.02, 0.02, and 0.01 for the P0, P50, and P100 percentiles, respectively. These results indicate a shift towards positive predictions as the magnitude of the predictor increases, although the differences between percentiles are relatively small.

As for Table 6, it reveals the extended sample's predicted values of the response variable 'logret' at different percentiles (P0, P50, and P100) for varying magnitudes of STReversal (-1 and 1).

At a magnitude of -1, the predicted values of the response variable are -0.08, -0.05, and -0.01 for the P0, P50, and P100 percentiles, respectively. This suggests a negative association between STReversal and the response variable, with slightly higher predicted values at the P50 percentile. On the other hand, at a magnitude of 1, the predicted values of the response variable are -0.00, 0.01, and 0.00 for the P0, P50, and P100 percentiles, respectively. These results indicate a weaker relationship between STReversal and the response variable, with predictions close to zero and minimal variation across percentiles.

In Table 7, it highlights the differences in predicted mean values for the 0th, 50th, and 100th percentiles of the response variable, logret, at two extreme fixed points of the SENT feature's magnitude (-0.68 and 1) in the extended sample.

At a magnitude of -0.68, the predicted mean values of the response variable are -0.06, -0.02, and -0.01 for the P0, P50, and P100 percentiles, respectively. This indicates a slightly negative association between SENT and the response variable, with a minimal change in mean values across the percentiles. Conversely, at a magnitude of 1, the predicted mean values of the response variable are -0.01, 0.00, and -0.00 for the P0, P50, and P100 percentiles, respectively. These results suggest a near-neutral relationship between SENT and the response variable, as the predicted mean values remain close to zero across the percentiles.

Table 8 illustrates the differences in predictions for the 0th, 50th, and 100th percentiles of the response variable, logret, at two extreme fixed points of the ATTENT feature's magnitude (-0.87 and 1) in the extended sample.

At a magnitude of -0.87, the predictions for the response variable are -0.02, 0.00, and -0.01 for the P0, P50, and P100 percentiles, respectively. This suggests a relatively stable prediction pattern, with slight variations in the predicted values across the percentiles. Conversely, at a magnitude of 1, the predictions for the response variable are -0.06, -0.03, and -0.01 for the P0, P50, and P100 percentiles, respectively. These results indicate a negative association between ATTENT and the response variable, with slightly lower predicted values as the magnitude increases.

It is worth noting that the differences between percentiles for (STReversal, SENT, ATTENT) are not substantial, indicating a relatively consistent prediction pattern for each predictor across the distribution of the response variable in both the benchmark and extended sample. Therefore, heterogeneity should not be a concern in this study.

4.3 DNN Portfolio Analysis

4.3.1 Prediction-sorted Portfolio Return

Table 9:
DNN Prediction-sorted Portfolio Performance under benchmark sample

	Pred	Avg	SD	SR
Low(L)	-0.03	-0.02	0.06	-1.62
2	-0.01	-0.01	0.07	-1.01
3	-0.01	-0.02	0.06	-1.48
4	-0.01	-0.02	0.06	-1.45
5	-0.01	-0.01	0.05	-0.91
6	-0.01	-0.00	0.05	-0.71
7	-0.00	0.00	0.05	-0.45
8	-0.00	0.00	0.05	-0.37
9	-0.00	-0.00	0.05	-0.66
High(H)	0.00	-0.00	0.05	-0.55
H-L	0.03	0.02	0.04	0.33

Table 10:
DNN Prediction-Sorted Portfolio Performance under extended sample

	Pred	Avg	SD	SR
Low(L)	-0.04	-0.02	0.06	-1.68
2	-0.02	-0.02	0.07	-1.32
3	-0.02	-0.02	0.06	-1.34
4	-0.02	-0.01	0.06	-1.02
5	-0.01	-0.01	0.05	-0.99
6	-0.01	-0.00	0.05	-0.59
7	-0.01	0.00	0.05	-0.50
8	-0.01	-0.00	0.04	-0.77
9	-0.01	-0.00	0.05	-0.71
High(H)	-0.01	0.00	0.05	-0.44
H-L	0.03	0.02	0.03	0.45

In (Table 9 and 10), I report the performance of prediction-sorted portfolio across the 8 years out-of-sample testing period for benchmark and extended sample. All realized stock returns are sorted into deciles based on their next month's predicted return. Pred,Avg,SD and SR, represent the monthly predicted return, average realized return, Standard deviation and Sharpe Ratio respectively.

Since the Deep Neural Network (DNN) model demonstrates superior predictive performance at the stock-level compared to other forecast models, both in the benchmark and extended samples (including the SENT and ATTENT predictors) in previous section. Consequently, I choose to exclusively rely on the DNN model for the portfolio return analysis. As outlined in the methodology section, the portfolios in both the benchmark and extended samples categorize their realized stock returns into deciles based on the corresponding 1-month ahead DNN stock-level predictions. An equal-weighted long-short portfolio is constructed for each benchmark and extended sample by buying stocks with the highest expected returns at the long leg (decile 10) and selling those with the lowest at the short leg (decile 1).

In both Table 9 and Table 10, the high minus low (H-L) strategy shows a similar monthly predicted return of 3% and average realized return of 2% for both the benchmark and extended samples. However, the benchmark sample exhibits a higher monthly volatility of 4% compared to the extended sample's monthly volatility of 3%. Consequently, the benchmark sample has a lower Sharpe ratio of 0.33, whereas the extended sample achieves a higher Sharpe ratio of 0.45.

4.3.2 Cumulative DNN Portfolio Return

In (Figure 7 and 8), the cumulative return of the long-short portfolio under the left panel shows that both the benchmark and extended samples outperform the market's cumulative excess return. Furthermore, there is a consistent upward increasing trend observed in both groups. However, it is important to highlight that the increasing trend is less steep in the extended sample compared to the benchmark sample until 2022.Starting in 2022, the trendline of the extended sample shows a steeper increase compared to the benchmark sample. By the end of 2022, the log cumulative return in the extended sample reaches approximately 0.80, which is higher than the log cumulative return of 0.70 in the benchmark sample.

Furthermore, a notable observation is that at the beginning of 2022, the DNN cumulative portfolio return displays a contrasting trendline direction compared to the market cumulative return. While the DNN cumulative portfolio return exhibits an increasing trend, the market cumulative return continuously declines throughout 2022, with only minor signs of recovery in the last few months. This shows that despite the overall market decline, the DNN model's ability to adapt and respond to changing conditions resulted in a positive and upward trajectory in the cumulative portfolio return.

Moreover, when comparing the cumulative portfolio return of the benchmark sample to the extended sample, several sharp dips can be observed in the benchmark sample. These dips occur during the start of 2019, throughout 2020, and at the beginning of 2021, coinciding with the occurrence of the Covid Crisis. In contrast, the extended sample exhibits a smoother trendline without significant dips in recent years. This suggests that these periods of market volatility and uncertainty had a more pronounced impact on the benchmark sample, highlighting the limitations of relying solely on traditional market indicators. In contrast, the extended sample, which incorporates sentiment and attention factors, demonstrated a more robust response during these challenging periods. By capturing and analyzing market sentiment and attention, the model was able to adapt and mitigate some of the adverse effects of the Covid Crisis, leading to a comparatively smoother performance and avoiding steep declines in the cumulative portfolio return.

4.3.3 DNN Long and Short Leg Comparisons

The right panels of (Figure 7 and 8), provide a breakdown of the cumulative return of the DNN long-short portfolio before buying and short-selling. These figures illustrate the comparison between the long leg, short leg, and the market excess return for both the benchmark and extended samples. It is evident that in both the benchmark and extended samples, the market excess return outperforms the long leg. However, the long leg in both groups closely follows the market trendline (FTSE All Share). The cumulative portfolio return of the long leg in both groups tends to hover around zero or slightly negative returns. Notably, there are significant spikes in the long legs of both the benchmark and extended samples during the period from 2020 to 2021, followed by a steep decline from the peak onwards.

The trendline for the short leg in the extended sample is relatively similar to that of the benchmark sample. The main difference is that, before the peak in 2021, the decreasing trendline of the short leg in the extended sample is less steep compared to the benchmark sample. After the peak in 2021, the short leg's trendline follows the overall market's decreasing trend as well. However, the downward trend in the extended sample is steeper than in the benchmark sample. This indicates that when the market trend is declining, the model incorporating sentiment and attention factors (SENT and ATTENT) is able to capture this trend, adjust its positioning, and suggest a more aggressive short-selling strategy (short sell the bottom decile stocks which have large persistent negative returns) after 2021. This adjustment helps compensate for the losses in the long leg and ultimately results in a positive overall cumulative return. Nevertheless, it is notable that the aggressive short-selling strategy induced by incorporation of sentiment and attention factors (SENT and ATTENT) provides more significant benefits to the model during market turbulent periods compared to normal periods. During normal periods, when the market is relatively stable, the impact of SENT and ATTENT factors on overall cumulative portfolio return may be less pronounced due to less aggressive short-selling strategy. This suggests that the model's ability to capture and leverage sentiment and attention signals is particularly valuable during market turbulent periods when traditional indicators may not fully capture the market dynamics.

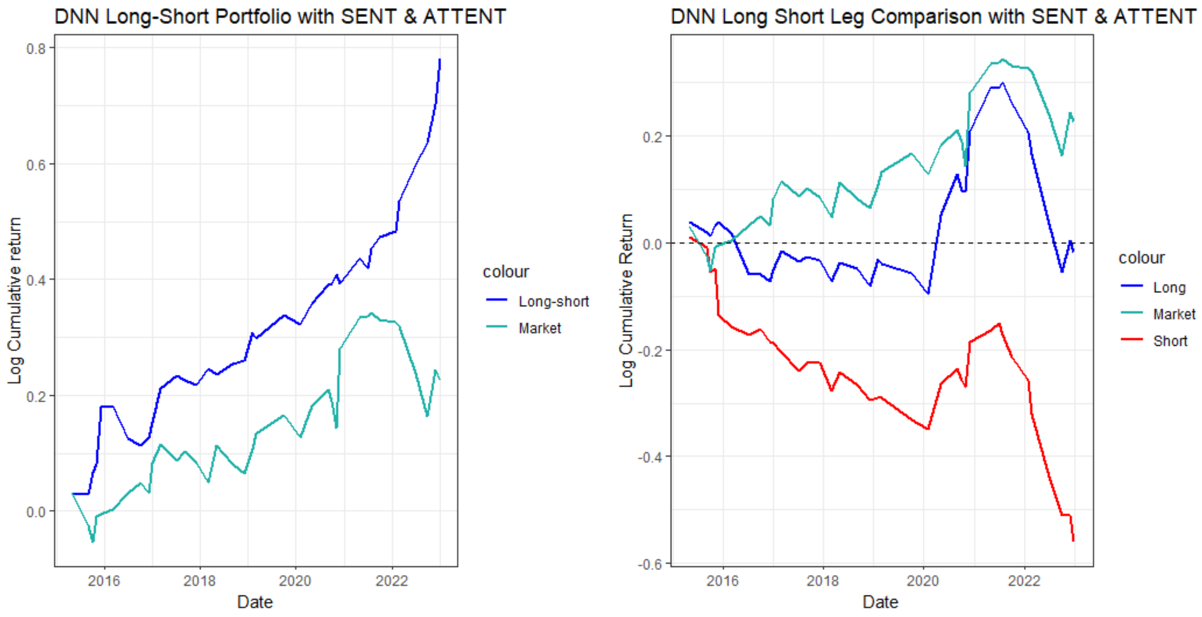


Figure 7: Cumulative return of DNN portfolio for the benchmark sample

Left panels show the average marginal effect of STReversal, SENT and ATTENT on the stock level return. Right panels show the single instance marginal effect of STReversal,SENT and ATTENT on the stock level return, which based on 1-100th percentile.

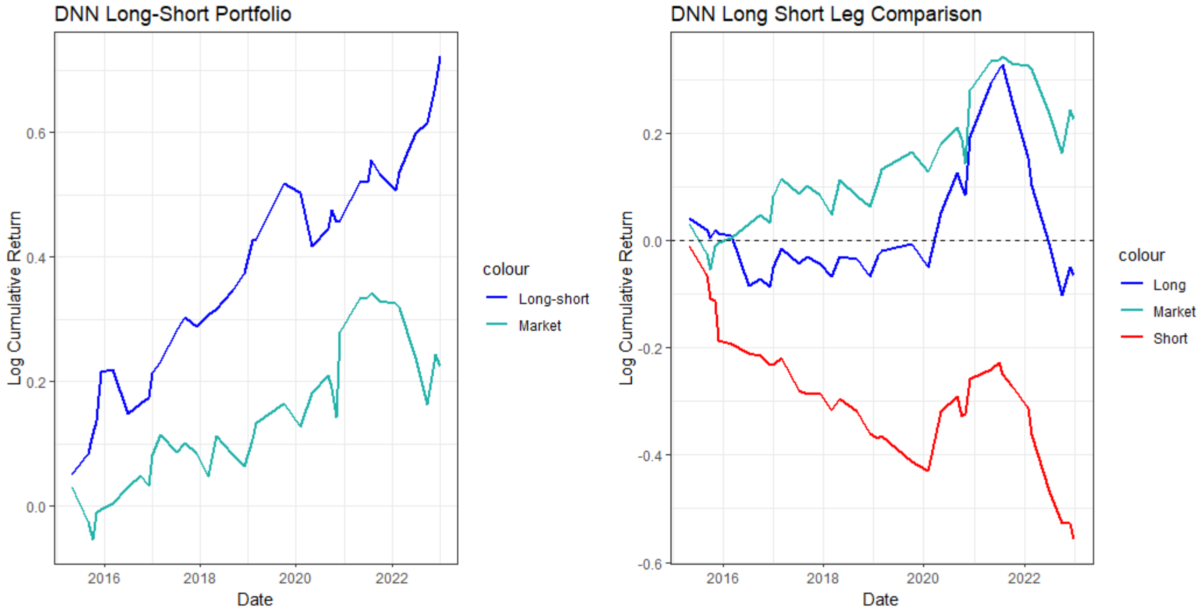


Figure 8: Cumulative return of DNN portfolio for the extended sample

The left panel shows the cumulative log return of long-short portfolio sorted on the out-of-sample predicted return over 8 years period for extended sample(SENT,ATTENT as predictors). In the right panel, the cumulative return of the portfolio is separated into the long leg (top decile) and short leg (bottom decile). The market excess return that based on FTSE All share is used as the benchmark.

4.3.4 Long-Short Portfolio Adjusted Return

Table 11:
DNN Portfolio Return based adjusted for CAPM, Fama-French 3/5/6 factors

α	MKT-RF	SMB	HML	RMW	CMA	MOM	r^2
0.0197***	-0.0003						0.0020
(3.89)	(-0.29)						
0.0193***	-0.0003	0.0021**	-0.0001				0.0139
(5.31)	(-0.43)	(0.98)	(-0.08)				
0.0177***	-0.0010**	0.0027	0.0071***	0.0066**	-0.0096**		0.1332
(6.18)	(-2.56)	(1.00)	(3.56)	(2.29)	(-2.08)		
0.0183***	-0.0019	0.0026	0.0047*	0.0061**	-0.0082*	-0.0024	0.1596
(4.93)	(-1.48)	(0.94)	(1.93)	(2.32)	(-1.73)	(-1.10)	

Table 12:
DNN Portfolio Return with SENT,ATTENT, adjusted for CAPM,Fama-French3/5/6 factors

α	MKT-RF	SMB	HML	RMW	CMA	MOM	R^2
0.0213***	-0.0003						0.0029
(3.70)	(-0.45)						
0.0206***	-0.0002	0.0033*	0.0001				0.0436
(3.19)	(-0.35)	(1.89)	(0.12)				
0.0177***	-0.0007	0.0045**	0.0061*	0.0080***	-0.0063		0.1833
(3.21)	(-1.29)	(2.10)	(1.74)	(4.60)	(-0.96)		
0.0175***	-0.0005	0.0045**	0.0068**	0.0081***	-0.0068	0.0007	0.1866
(3.14)	(-0.49)	(2.09)	(2.15)	(4.43)	(-1.16)	(0.36)	

Table 11 and 12 show the risk adjusted performance of DNN portfolios under benchmark sample and extended sample based on the factor pricing models, CAPM, Fama-French 3,5,6 factors. The t-test statistic is in bracket, $***p < 0.01$, $**p < 0.05$, and $*p < 0.1$ represent the significance level of p-value.

In (Table 11), the portfolio formed on the DNN forecast under the benchmark sample earns significant alphas at 1% level across CAPM, Fama-French 3 to 6 factors models. The alphas which represents risk-adjusted return gradually decline as the factor pricing models expanded with their factors from CAPM of 1.97% with t-statistics of 3.89 to Fama-French 5 factors models of 1.77% with t-statistic of 6.18. The 5 factors are MKT-RF for market, SMB for size, HML for value, RMW for operating profitability and CWA for investment. However, when adding the additional factor of momentum MOM to the 5 factors model, the monthly risk adjusted return rebounds to 1.83% with t-statistics of 4.93, meaning that the 5-factor model performs better than the 6-factor model in spanning away the risk-adjusted returns, it suggests that the additional factor (MOM) does not provide significant explanatory power or improvement. The spanning regression R^2 however, increases from 0.0020 to 0.1596 across the factors models of the benchmark sample.

In (Table 12), the portfolio formed based on the DNN forecast in the extended sample exhibits significant alphas at the p-value 1% level across the CAPM, Fama-French 3, and Fama-French 6 factors models. The monthly adjusted return for CAPM and Fama-French 3 factors models are approximately 2% with the t-statistics range from 3.70 to 3.19, clearly they are higher in the extended sample compared to the benchmark sample due to the presence of investor sentiment and attention. The monthly adjusted return also gradually decreases as the factor models expand from CAPM with 2.13% with t-statistic of 3.70 to Fama-French 6 factors model with 1.75% with t-statistic of 3.14. The alpha under the Fama-French 6 factors model is slightly lower than the alpha in the benchmark sample. This indicates that in the extended sample, the investor sentiment and attention components play a role in helping the Fama-French 6 factors model better explain the portfolio returns, resulting in a lower risk-adjusted return. Overall, neither the extended nor the benchmark sample's prediction sorted-portfolio returns can be fully explained by the factor pricing models. Although the investor behavior component embedded in the portfolio return may prompt the factor pricing models to capture returns better, further exploration is still required. Additionally, it is noteworthy that HML (High Minus Low) and RMW (Robust Minus Weak) factors play a significant role in explaining the portfolio returns under both the extended and benchmark samples.

4.3.5 Intervention Analysis on Long-Short Portfolio Return Improvement

Table 13:
Regression DNN Long-Short Portfolio Return with SENT/ATTENT on DNN Long-Short Portfolio Return without SENT/ATTENT

α	longshortSA	R^2
0.0045 (1.04)	0.7102*** (8.05)	0.359

The t-statistics is in brackets, $***p < 0.01$, $**p < 0.05$, and $*p < 0.1$ represent the significance level of p-value. This shows the regression of the extended sample's long short portfolio return on the benchmark sample's long short portfolio return.LongshortSA is the long-short portfolio return derived from the stock level prediction with the inclusion of sentiment and attention.

(Table 13) shows that the extended sample's DNN prediction sorted long-short portfolio is regressed on the benchmark sample's DNN prediction sorted long-short portfolio. The regression results suggest that the inclusion of sentiment and attention effect, as captured by the longshortSA variable, it significantly improves the performance of the long-short portfolio. A one-unit increase in the long-short portfolio return with sentiment and attention is associated with an estimated increase of 71.02% with t-statistic of 8.05 in the long-short portfolio return without sentiment and attention. This positive and statistically significant coefficient provides evidence that investor sentiment and attention information has a significant impact on the performance of the long-short portfolio.

5 Conclusion

In the UK market, Deep Neural Network (DNN) proves to be the most effective stock level prediction model, aligning with existing machine learning literature in both the US and other international markets (Gu et al., 2020; Drobetz et al., 2019; Hanauer & Kalsbach, 2023). Instead of following the past studies to rely on the OLS setting to predict stock returns based on investor sentiment and attention (Hudson & Green, 2015; J. Chen et al., 2022; Cai et al., 2022),I have incorporated these concepts into my machine learning models. The inclusion of investor sentiment and attention, along with firm and macroeconomic predictors, has significantly improved stock level prediction and the resulting long-short portfolio return. Both extended and benchmark sample's cumulative long-short portfolio return generated by the DNN model surpass the cumulative excess return of the UK market. Moreover, the machine learning (DNN) prediction-based portfolio has its uniqueness for generating alphas that cannot be explained away by the traditional factor models CAPM, Fama-French 3,5,6 models.

Many emerging markets face short-sale restrictions, which could pose a significant challenge in implementing the DNN model to achieve a positive increasing cumulative return. This difficulty arises when the portfolio's return pattern in these emerging markets, similar to the one observed in my UK study, predominantly relies on the positive cumulative return generated from the short leg of the portfolio. Nonetheless, when investor sentiment and attention are incorporated into the Deep Neural Network prediction, they smooth out the negative effects on cumulative portfolio return during market turbulence, such as the Covid crisis. The question remains whether the ability of sophisticated machine learning models like DNN to capture turbulent fluctuation signals informed by investor sentiment and attention is a result of sheer luck or superior model architecture. Is incorporating DNN with investor sentiment and attention during market turbulence a reliable option to achieve higher prediction returns? To what extent can machine learning models be useful during turbulent periods, and is the trend of outperformance consistent in every turbulent period? Lastly, my study also suggests that the investor sentiment and attention components in portfolio returns interact with the Fama-French factors, allowing these factors to capture a broader range of risk and return drivers in the market. However, I did not delve further into these aspects, leaving them as potential areas for future research and extrapolations.

References

Aboody, D., Lehavy, R. & Trueman, B. (2010). Limited attention and the earnings announcement returns of past stock market winners. *Review of Accounting Studies*, *15*, 317–344.

All Datasets - ECB Statistical Data Warehouse — sdw.ecb.europa.eu. (n.d.). `https://sdw.ecb.europa.eu/browse.do?node=9689727`. ([Accessed 02-Jun-2023])

Andrei, D. & Hasler, M. (2015). Investor attention and stock market volatility. *The review of financial studies*, *28*(1), 33–72.

Antoniou, C., Doukas, J. A. & Subrahmanyam, A. (2013). Cognitive dissonance, sentiment, and. *JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS*, *48*(1), 245–275.

Avramov, D., Cheng, S. & Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, *69*(5), 2587–2619.

Azevedo, V., Kaiser, S. & Müller, S. (2022). Stock market anomalies and machine learning across the globe. *Available at SSRN 4071852*.

Baker, M. & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance*, *61*(4), 1645–1680.

Barber, B. M., Huang, X., Odean, T. & Schwarz, C. (2022). Attention-induced trading and returns: Evidence from robinhood users. *The Journal of Finance*, *77*(6), 3141–3190.

Barber, B. M. & Odean, T. (2007, December). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, *21*(2), 785–818. Retrieved from `https://doi.org/10.1093/rfs/hhm079` doi: 10.1093/rfs/hhm079

Barberis, N., Shleifer, A. & Vishny, R. (1998). A model of investor sentiment. *Journal of financial economics*, *49*(3), 307–343.

Beckmeyer, H. & Wiedemann, T. (2023). Recovering missing firm characteristics with attention-based machine learning. *Available at SSRN 4003455*.

Bijl, L., Kringhaug, G., Molnár, P. & Sandvik, E. (2016, May). Google searches and stock returns. *International Review of Financial Analysis*, *45*, 150–156. Retrieved from `https://doi.org/10.1016/j.irfa.2016.03.015` doi: 10.1016/j.irfa.2016.03.015

Breiman, L. (2001).
Machine Learning, *45*(1), 5–32. Retrieved from `https://doi.org/10.1023/a:1010933404324` doi: 10.1023/a:1010933404324

Brown, G. W. (1999). Volatility, sentiment, and noise traders. *Financial Analysts Journal*, *55*(2), 82–90.

Brown, G. W. & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of empirical finance*, *11*(1), 1–27.

Cai, H., Jiang, Y. & Liu, X. (2022). Investor attention, aggregate limit-hits, and stock returns. *International Review of Financial Analysis*, *83*, 102265.

Candel, A. & LeDell, E. (2022, October). Deep learning with h2o [Computer software manual]. Retrieved from `https://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf`

Chen, J., Tang, G., Yao, J. & Zhou, G. (2022). Investor attention and stock returns. *Journal of Financial and Quantitative Analysis*, *57*(2), 455–484.

Chen, L., Pelger, M. & Zhu, J. (2023). Deep learning in asset pricing. *Management Science*.

Chen, T., Gao, Z., He, J., Jiang, W. & Xiong, W. (2019, January). Daily price limits and destructive market behavior. *Journal of Econometrics*, *208*(1), 249–264. Retrieved from `https://doi.org/10.1016/j.jeconom.2018.09.014` doi: 10.1016/j.jeconom.2018.09.014

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

Choi, D., Jiang, W. & Zhang, C. (2022). Alpha go everywhere: Machine learning and international stock returns. *Available at SSRN 3489679*.

Chu, G., Goodell, J. W., Shen, D. & Zhang, Y. (2022, August). Machine learning to establish proxies for investor attention: evidence of improved stock-return prediction. *Annals of Operations Research*, *318*(1), 103–128. Retrieved from `https://doi.org/10.1007/s10479-022-04892-0` doi: 10.1007/s10479-022-04892-0

Crego, J. A., Soerlie Kvaerner, J. & Stam, M. (2023). Machine learning and expected returns. *Available at SSRN 4345646*.

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

Drobetz, W., Haller, R., Jasperneite, C. & Otto, T. (2019). Predictability and the cross section of expected returns: evidence from the european stock market. *Journal of Asset Management*, *20*, 508–533.

Fama, E. F. & French, K. R. (2008). Dissecting anomalies. *The journal of finance*, *63*(4), 1653–1678.

Fedyk, A. (2018). *Front page news: The effect of news positioning on financial markets* (Tech. Rep.). Working paper.

Feng, G. & He, J. (2022). Factor investing: A bayesian hierarchical approach. *Journal of Econometrics*, *230*(1), 183–200.

Feng, G., He, J., Polson, N. G. & Xu, J. (2018). Deep learning in characteristics-sorted factor models. *arXiv preprint arXiv:1805.01104*.

Fong, W. M. (2013). Risk preferences, investor sentiment and lottery stocks: A stochastic dominance approach. *Journal of Behavioral Finance*, *14*(1), 42–52.

Freyberger, J., Neuhierl, A. & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *International journal of neural systems*, *8*(02), 209–218.

Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Goyal, A., Welch, I. & Zafirov, A. (2021). A comprehensive look at the empirical performance ofEquity premium prediction II. *SSRN Electronic Journal*. Retrieved from `https://doi.org/10.2139/ssrn.3929119` doi: 10.2139/ssrn.3929119

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

h2o: R interface for h2o [Computer software manual]. (2022, October). Retrieved from `https://www.h2o.ai` (R package version 3.38.0.2)

Han, Y., He, A., Rapach, D. & Zhou, G. (2018). What firm characteristics drive us stock returns. *Available at SSRN*, *3185335*.

Hanauer, M. X. & Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, *55*, 101022.

Hirshleifer, D. & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of accounting and economics*, *36*(1-3), 337–386.

Huang, S., Huang, Y. & Lin, T.-C. (2019, May). Attention allocation and return co-movement: Evidence from repeated natural experiments. *Journal of Financial Economics*, *132*(2), 369–383. Retrieved from `https://doi.org/10.1016/j.jfineco.2018.10.006` doi: 10.1016/j.jfineco.2018.10.006

Huberman, G. & Regev, T. (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *The Journal of Finance*, *56*(1), 387–396.

Hudson, Y. & Green, C. J. (2015). Is investor sentiment contagious? international sentiment and uk equity returns. *Journal of Behavioral and Experimental Finance*, *5*, 46–59.

Kacperczyk, M., Van Nieuwerburgh, S. & Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, *84*(2), 571–626.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Kenneth R. French - Data Library — mba.tuck.dartmouth.edu. (n.d.). `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. ([Accessed 04-Jun-2023])

Lemmon, M. & Portniaguina, E. (2006). Consumer confidence and asset prices: Some empirical evidence. *The Review of Financial Studies*, *19*(4), 1499–1529.

Lewellen, J. (2014). The cross section of expected stock returns. *Forthcoming in Critical Finance Review, Tuck School of Business Working Paper*(2511246).

Li, J. & Yu, J. (2012, May). Investor attention, psychological anchors, and stock return predictability. *Journal of Financial Economics*, *104*(2), 401–419. Retrieved from `https://doi.org/10.1016/j.jfineco.2011.04.003` doi: 10.1016/j.jfineco.2011.04.003

Li, X., Ma, J., Wang, S. & Zhang, X. (2015, September). How does google search affect trader positions and crude oil prices? *Economic Modelling*, *49*, 162–171. Retrieved from `https://doi.org/10.1016/j.econmod.2015.04.005` doi: 10.1016/j.econmod.2015.04.005

Model Categories x2014; H2O documentation — docs.h2o.ai. (n.d.). `https://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/model_categories.html`. ([Accessed 10-Jun-2023])

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Moritz, S. & Bartz-Beielstein, T. (2017). imputets: time series missing value imputation in r. *R J.*, *9*(1), 207.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *The* *journal of finance*, *53*(6), 1887–1934.

Paye, B. S. (2011). Deja vol: Predictive regressions for aggregate stock market volatility using macroeconomic variables. *SSRN Electronic Journal*. Retrieved from `https://doi.org/10.2139/ssrn.783986` doi: 10.2139/ssrn.783986

Peng, L. & Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, *80*(3), 563–602.

Rapach, D. E. & Zhou, G. (2020). Time-series and cross-sectional stock return forecasting: New machine learning methods. *Machine learning for asset management: New developments and financial applications*, 1–33.

Schmeling, M. (2009). Investor sentiment and stock returns: Some international evidence. *Journal of empirical finance*, *16*(3), 394–408.

Seasholes, M. S. & Wu, G. (2007, December). Predictable behavior, profits, and attention. *Journal of Empirical Finance*, *14*(5), 590–610. Retrieved from `https://doi.org/10.1016/j.jempfin.2007.03.002` doi: 10.1016/j.jempfin.2007.03.002

Sun, L., Najand, M. & Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance*, *73*, 147–164.

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

6 Appendices

6.1 Appendix A.1

#	Predictors	Category	Definition
1	A2ME	Value	Assets to market capitalization
2	AT	Trading Frictions	Total assets
3	ATO	Profitability	Sales to assets
4	BEME	Value	Book value to market value of equity
5	Beta	Trading Frictions	Market beta
6	C	Value	Cash and short term investments to assets
7	CbOPtA	Profitability	Cash based operating profits to assets
8	CF2P	Value	Cash flow from operating activities to market capitalization
9	CTO	Profitability	Capital turnover
10	D2A	Intangibles	Capital intensity
11	Debt2P	Value	Leverage
12	DPI2A	Investments	Ratio of change in property,plants equipment to total assets
13	E2P	Value	Earning to price
14	FC2Y	Profitability	Fixed costs to assets
15	FreeCF	Value	Cash flow to book value of equity
16	GP2A	Profitability	Gross profit to assets
17	INV	Investment	Investment
18	LME	Trading Frictions	Market capitalization
19	LTurnover	Trading Frictions	Stock's trading volume divided by shares outstanding
20	NOA	Investment	Net operating assets
21	OA	Intangibles	Operating Accruals
22	OL	Intangibles	Operating Leverages
23	P2P52WH	Trading Frictions	Price relative to its 52-week high
24	PCM	Profitability	Price to cost margin
25	PM	Profitability	Profit margin
26	Prof	Profitability	Gross Profitability
27	Q	Value	Tobin's Q
28	mom	Past returns	Momentum is the cumulative excess return from month t-12 to t-2
29	intmom	Past returns	Intermediate momentum is the cumulative excess return from month t-12 to t-7
30	STreversal	Past returns	Short term reversal is lagged one month excess return
31	LTreversal	Past returns	Long term reversal is the cumulative return from t-36 to t-13
32	RNA	Profitability	Return on net operating assets
33	ROA	Profitability	Return on assets
34	ROE	Profitability	Return on equity
35	S2P	Value	Sales to price
36	SGA2S	Intangibles	Sales and general administrative costs to sales
37	IIliqu	Trading Frictions	Illiquidity
38	SUV	Trading Frictions	Unexplained volume

Table A.1: Firm Predictors' Definitions
I replicate all the 38 firm predictors from Hanauer and Kalsbach (2023).See detailed explanations of definitions and constructions from their paper.

6.2 Appendix A.2

#	Predictors	Definition
1	b/m	Ratio of book value to market value of FTSE
2	d/y	Dividend-yield ratio of FTSE
3	e/p	Earning Price Ratio of FTSE
4	rf	UK risk free rate, equivalent to 3 month treasury bill rate.
5	tms	Term Spread is the differences between long-term yield on government bonds and the risk free rate
6	svar	Stock variance on FTSE All Share
7	i/k	Investment to capital ratio is the ratio of private investment to capital for the aggregate UK economy
8	infl	Inflation is the Consumer Price Index CPI for UK
9	lty	UK Long-term government bond yield
10	oas	Option Adjusted Spread is the spread of corporate bond rate and long term treasury bill rate which takes into account of the embedded option.
11	cp	Commercial paper to treasury spread is the spread between 3 month commercial paper rate on the 3 month UK treasury bills.

Table A.2: Macroeconomics Predictors' Definitions
I replicate the macroeconomics predictors from the Welch and Goyal (2008) and Paye (2011). See detailed explanations of the constructions of the variables from their papers

6.3 Appendix A.3

#	Proxies	Definition
1	AVCD	Advances-Decline Ratio is the number of stocks rising to the number of stocks declining in UK market
2	SMART.index	Smart Money Flow Index for FTSE100
3	PCV	Put-Call Trading Volume for FTSE100
4	PCO	Put-Call Open Interest ratio for FTSE100
5	RSI.30D	Relative Strength Index in 30 days for FTSE 100
6	IVI.30D	Implied Volatility in 30 days for FTSE100
7	CISS	Composite indicator of systematic stress in UK
8	CLIFS	Country-level financial stress composite indicator is a financial stress measure that accounts for co-movement of different market segments in UK

Table A.3: Investor Sentiment Predictors' Definitions
I follow Hudson and Green (2015) to construct the proxies of investor sentiment. See detailed explanations of the definition from their paper. I also incorporate the European Central Bank's financial stress indicators for the UK market as proxies. See detailed explanations from ECB (*All Datasets - ECB Statistical Data Warehouse — sdw.ecb.europa.eu*, n.d.).

6.4 Appendix A.4

#	Proxies	Definition
1	GSV	Google Search Volume related to FTSE
2	aavol	Abnormal Trading Volume for FTSE100
3	aeret	Extreme Returns for FTSE100
4	ualhits	Upper aggregate limit-hits represents the monthly count of sample stocks hitting their respective daily upper limit, wherein each stock's price experiences a 10% increases compared to the previous closing price

Table A.4: Investor Attention Predictors' Definitions
Cai et al. (2022) and J. Chen et al. (2022) provide references for construction of proxies for my study. See detailed explanations from their papers

6.5 Appendix B.1

Predictors	**n_missing**	**complete_rate**	**numeric.mean**	**numeric.sd**	**numeric.p0**	**numeric.p25**	**numeric.p50**	**numeric.p75**	**numeric.p100**
mom	30438	0.93	-0.06	0.58	-23.86	-0.41	-0.07	0.20	32.29
intmom	30289	0.93	-0.04	0.42	-15.63	-0.25	-0.03	0.15	15.73
STreversal	1777	1.00	-0.07	0.28	-1.00	-0.22	-0.05	0.10	1.00
LTReversal	107430	0.74	-0.05	0.83	-16.70	-0.61	-0.15	0.31	30.42
TOBINQ	17889	0.96	3.31	111.81	-1.94	0.56	0.93	1.69	19441.25
A2ME	42722	0.90	2.58	19.35	0.00	0.61	1.19	2.23	2577.14
AT	39838	0.90	6535722.83	70322840.35	0.00	12227.00	58197.00	369461.00	2447119464.00
NOA	60141	0.85	0.51	29.02	-438.00	0.35	0.57	0.73	13096.08
ATO	63341	0.84	1837993.08	79793312.76	-11679247555.56	6001.95	62563.74	463627.89	8288119674.29
BEME	10289	0.97	0.83	4.22	-100.00	0.26	0.55	1.04	100.00
C	50133	0.88	0.20	0.23	0.00	0.04	0.11	0.26	1.00
CBOPTA	40158	0.90	-0.05	2.43	-297.00	-0.03	0.07	0.16	383.75
CF2P	140204	0.66	47346.15	1514002.02	-54254900.00	-14.30	27.83	327.53	97103704.00
CTO	43672	0.89	0.89	3.35	-9.60	0.18	0.64	1.21	1636.58
D2A	53072	0.87	0.05	1.33	-0.42	0.01	0.03	0.06	535.92
Debt2P	139899	0.66	8345400.82	82193572.53	-558623469.00	10260.78	53784.83	417917.76	2447118172.11
DPI2A	99648	0.76	0.46	2.13	-0.95	0.10	0.30	0.71	996.71
E2P	139779	0.66	23628.45	928660.28	-17387888.00	-51.47	7.33	159.82	93258672.00
FC2Y	72540	0.82	150622.00	839150.86	-3872000.00	1080.00	6695.00	34000.00	20351588.00
FreeCF	59497	0.85	-625848.34	121906515.86	-38517142350.00	-7770.02	-50.60	599.99	1905346550.00
GP2A	40158	0.90	1398570.14	10280780.98	-89294000.00	4245.83	36116.00	253199.69	361936128.86
INV	11950	0.97	0.04	3.44	-1.00	0.00	0.00	0.00	1211.74
LME	109769	0.73	3775.42	122406.34	0.00	7.69	41.95	304.51	11748472.00
LTurnover	15618	0.96	0.26	15.28	0.00	0.01	0.02	0.06	6811.40
OA	213724	0.47	0.00	0.04	-5.22	0.00	0.00	0.00	6.34
OL	39931	0.90	781894.64	7837335.36	-135374.00	408.12	11453.00	103100.01	309129356.05
P2P52WH	46612	0.89	247.19	1458.95	0.00	36.06	71.91	124.47	135000.00
PCM	72540	0.82	1534312.61	10757387.64	-89294000.00	8643.80	49640.27	312061.52	361936129.15
Prof	42907	0.89	-111099.22	48192845.19	-1085230380.00	-15297.09	847.84	24548.75	8793330870.00
PM	80956	0.80	-4.58	139.81	-12424.00	-0.09	0.05	0.14	9877.00
RNA	70872	0.83	275343.89	11391564.04	-666472563.64	-2608.85	3123.18	43296.12	5599757580.95
ROA	43544	0.89	-0.30	6.59	-701.00	-0.11	0.01	0.06	151.26
ROE	46242	0.89	-29839643.59	7325585794.92	-2448591918750.00	-2486.04	1096.50	27410.81	518390384.10
S2P	139899	0.66	193043.39	3945666.23	-35716.18	62.86	683.43	5307.83	196470972.00
SG2A	72540	0.82	3.55	63.09	-79.67	0.07	0.25	0.54	5443.00
IIliqu	15615	0.96	36540.14	379624.26	-54.91	271.00	1904.96	11485.75	112009138.89
beta	7870	0.98	0.71	3.65	-213.28	-0.39	0.65	1.91	131.69
SUV	15686	0.96	0.00	1.00	-1.34	-0.09	-0.07	-0.06	295.11
lty	0	1.00	0.03	0.02	0.00	0.02	0.04	0.05	0.06
rf	0	1.00	0.03	0.02	0.00	0.01	0.01	0.05	0.07
svar	0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
e.p	0	1.00	0.06	0.02	0.03	0.05	0.06	0.07	0.12
d.y	0	1.00	3.36	0.61	2.06	2.99	3.31	3.63	5.53
b.m	0	1.00	0.67	0.26	0.32	0.42	0.64	0.88	1.32
i.k	0	1.00	0.04	0.08	-0.02	0.02	0.03	0.04	1.31
infl	0	1.00	0.02	0.01	0.00	0.01	0.02	0.03	0.10
cp	0	1.00	0.00	0.01	-0.04	0.00	0.00	0.00	0.05
tms	0	1.00	0.01	0.01	-0.02	0.00	0.01	0.01	0.03
oas	0	1.00	0.04	0.03	0.02	0.03	0.04	0.05	0.20

Table B.1: Summary Statistics for Firm and Macroeconomics predictors before imputation and standardization. Completeness, mean, standard deviation, and each quantile are shown.

6.6 Appendix B.2

Predictors	n_missing	complete_rate	numeric.mean	numeric.std	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100
mom	0	1	-0.20	0.34	-1.00	-0.45	-0.23	0.03	1.00
intmom	0	1	-0.13	0.31	-1.00	-0.34	-0.15	0.08	1.00
STreversal	0	1	-0.07	0.28	-1.00	-0.22	-0.05	0.10	1.00
LTReversal	0	1	-0.42	0.28	-1.00	-0.63	-0.48	-0.25	1.00
TOBINQ	0	1	-0.98	0.07	-1.00	-1.00	-0.99	-0.98	1.00
A2ME	0	1	-0.97	0.09	-1.00	-0.99	-0.99	-0.97	1.00
AT	0	1	-0.99	0.09	-1.00	-1.00	-1.00	-1.00	1.00
NOA	0	1	0.61	0.54	-1.00	0.51	0.86	0.96	1.00
ATO	0	1	-0.02	0.45	-1.00	-0.33	-0.02	0.28	1.00
BEME	0	1	0.05	0.31	-1.00	-0.02	0.01	0.23	1.00
C	0	1	-0.60	0.46	-1.00	-0.91	-0.77	-0.45	1.00
CBOPTA	0	1	0.74	0.29	-1.00	0.63	0.84	0.93	1.00
CF2P	0	1	-0.83	0.27	-1.00	-0.98	-0.96	-0.77	1.00
CTO	0	1	-0.84	0.19	-1.00	-0.96	-0.90	-0.80	1.00
D2A	0	1	-0.95	0.09	-1.00	-1.00	-0.99	-0.93	1.00
Debt2P	0	1	-0.93	0.10	-1.00	-0.98	-0.94	-0.90	1.00
DPI2A	0	1	-0.92	0.11	-1.00	-0.99	-0.96	-0.90	1.00
E2P	0	1	-0.58	0.55	-1.00	-0.98	-0.91	-0.45	1.00
FC2Y	0	1	-0.96	0.16	-1.00	-1.00	-1.00	-0.99	1.00
FreeCF	0	1	0.04	0.75	-1.00	-0.74	-0.10	0.90	1.00
GP2A	0	1	-0.96	0.13	-1.00	-1.00	-1.00	-0.98	1.00
INV	0	1	-0.64	0.35	-1.00	-0.92	-0.74	-0.47	1.00
LME	0	1	-0.99	0.07	-1.00	-1.00	-1.00	-1.00	1.00
LTurnover	0	1	-0.99	0.06	-1.00	-1.00	-1.00	-1.00	1.00
OA	0	1	0.02	0.43	-1.00	-0.29	0.00	0.34	1.00
OL	0	1	-0.99	0.07	-1.00	-1.00	-1.00	-1.00	1.00
P2P52WH	0	1	-0.98	0.09	-1.00	-1.00	-1.00	-0.98	1.00
PCM	0	1	-0.96	0.13	-1.00	-1.00	-0.99	-0.98	1.00
Prof	0	1	0.01	0.52	-1.00	-0.49	0.07	0.47	1.00
PM	0	1	0.83	0.29	-1.00	0.80	0.94	0.98	1.00
RNA	0	1	-0.43	0.45	-1.00	-0.89	-0.51	-0.12	1.00
ROA	0	1	0.62	0.41	-1.00	0.46	0.75	0.93	1.00
ROE	0	1	0.02	0.66	-1.00	-0.62	0.02	0.50	1.00
S2P	0	1	-0.99	0.07	-1.00	-1.00	-1.00	-1.00	1.00
SG2A	0	1	-0.97	0.09	-1.00	-1.00	-0.99	-0.96	1.00
IIliqu	0	1	-0.99	0.08	-1.00	-1.00	-1.00	-1.00	1.00
beta	0	1	0.04	0.26	-1.00	-0.13	0.03	0.20	1.00
SUV	0	1	-0.93	0.08	-1.00	-0.97	-0.95	-0.92	1.00
lty	0	1	0.10	0.57	-1.00	-0.44	0.25	0.60	1.00
rf	0	1	-0.21	0.68	-1.00	-0.83	-0.63	0.45	1.00
svar	0	1	-0.86	0.25	-1.00	-0.98	-0.94	-0.84	1.00
e.p	0	1	-0.26	0.38	-1.00	-0.52	-0.26	-0.05	1.00
d.y	0	1	-0.25	0.35	-1.00	-0.46	-0.28	-0.10	1.00
b.m	0	1	-0.29	0.53	-1.00	-0.80	-0.37	0.12	1.00
i.k	0	1	-0.91	0.12	-1.00	-0.93	-0.91	-0.90	1.00
infl	0	1	-0.57	0.30	-1.00	-0.74	-0.62	-0.49	1.00
cp	0	1	-0.13	0.19	-1.00	-0.14	-0.12	-0.11	1.00
tms	0	1	-0.10	0.45	-1.00	-0.45	-0.10	0.12	1.00
oas	0	1	-0.73	0.30	-1.00	-0.89	-0.79	-0.65	1.00

Table B.2: Summary Statistics for Firm and Macroeconomics predictors after standardization between [-1,1] and imputation. Completeness, mean, standard deviation, and each quantile are shown.

6.7 Appendix B.3

year_	variable	n	p0	p25	p50	p75	p100
2000	logret	19109	-2.313	-0.068	0	0.057	1.9
2001	logret	19657	-3.081	-0.068	0	0.057	1.951
2002	logret	19722	-2.386	-0.072	0	0.055	3.128
2003	logret	18753	-4.318	-0.073	0	0.055	3.218
2004	logret	19006	-2.722	-0.078	0	0.054	2.311
2005	logret	21306	-3.384	-0.076	0	0.057	4.633
2006	logret	22755	-4.091	-0.076	0	0.055	2.565
2007	logret	22616	-3.628	-0.076	0	0.055	4.614
2008	logret	21490	-4.609	-0.072	0	0.054	2.211
2009	logret	19061	-2.464	-0.073	0	0.054	2.664
2010	logret	17527	-2.069	-0.07	0	0.058	3.08
2011	logret	16766	-2.53	-0.071	0	0.056	2.662
2012	logret	16212	-3.377	-0.071	0	0.059	2.666
2013	logret	15864	-4.126	-0.073	0	0.057	3.08
2014	logret	16229	-3.24	-0.073	0	0.055	1.755
2015	logret	16280	-4.556	-0.072	0	0.056	2.755
2016	logret	15804	-3.296	-0.074	0	0.056	2.4
2017	logret	15520	-2.61	-0.071	0	0.057	1.609
2018	logret	15487	-5.235	-0.071	0	0.059	3.842
2019	logret	15013	-2.574	-0.071	0	0.058	2.014
2020	logret	14205	-4.818	-0.071	0	0.058	2.337
2021	logret	14263	-2.111	-0.072	0	0.057	1.824
2022	logret	14324	-2.306	-0.069	0	0.056	5.394

Table B.3: Summary Statistics of excess return expressed as logret

6.8 Appendix C.1

Proxies	complete_rate	mean	std	p0	p25	p50	p75	p100
AVDC	1	-0.04	0.36	-1.00	-0.29	-0.06	0.19	1.00
PCO	1	-0.46	0.22	-1.00	-0.60	-0.51	-0.32	1.00
SMART.index	1	-0.02	0.36	-1.00	-0.26	0.00	0.18	1.00
PCV	1	-0.72	0.21	-1.00	-0.84	-0.77	-0.65	1.00
RSI.30D	1	0.10	0.41	-1.00	-0.15	0.11	0.41	1.00
IVI.30D	1	-0.55	0.32	-1.00	-0.78	-0.64	-0.40	1.00
CLIFS	1	-0.59	0.38	-1.00	-0.85	-0.71	-0.45	1.00
CISS	1	-0.68	0.42	-1.00	-0.96	-0.87	-0.58	1.00
GSV	1	-0.77	0.32	-1.00	-0.95	-0.93	-0.70	1.00
aeret	1	0.10	0.12	-1.00	0.09	0.10	0.11	1.00
aavol	1	-0.24	0.26	-1.00	-0.40	-0.26	-0.10	1.00
ualhits	1	-0.76	0.22	-1.00	-1.00	-0.74	-0.66	1.00

Table C.1: Descriptive Statistics for SENT and ATTENT proxies before PCA

Predictors	complete_rate	mean	std	p0	p25	p50	p75	p100
SENT	1	0.00	1.79	-6.64	-0.79	0.37	1.34	2.86
ATTENT	1	0.00	1.18	-1.94	-0.73	-0.17	0.45	8.75

Table C.1.2: Descriptive Statistics for SENT and ATTENT indices after PCA

Predictors	complete_rate	mean	std	p0	p25	p50	p75
SENT	1	0.39	0.39	-1.00	0.20	0.48	0.68
ATTENT	1	-0.65	0.21	-1.00	-0.77	-0.68	-0.56

Table C.1.3: Descriptive Statistics for standardized SENT and ATTENT indices within[-1,1]

6.9 Appendix C.2

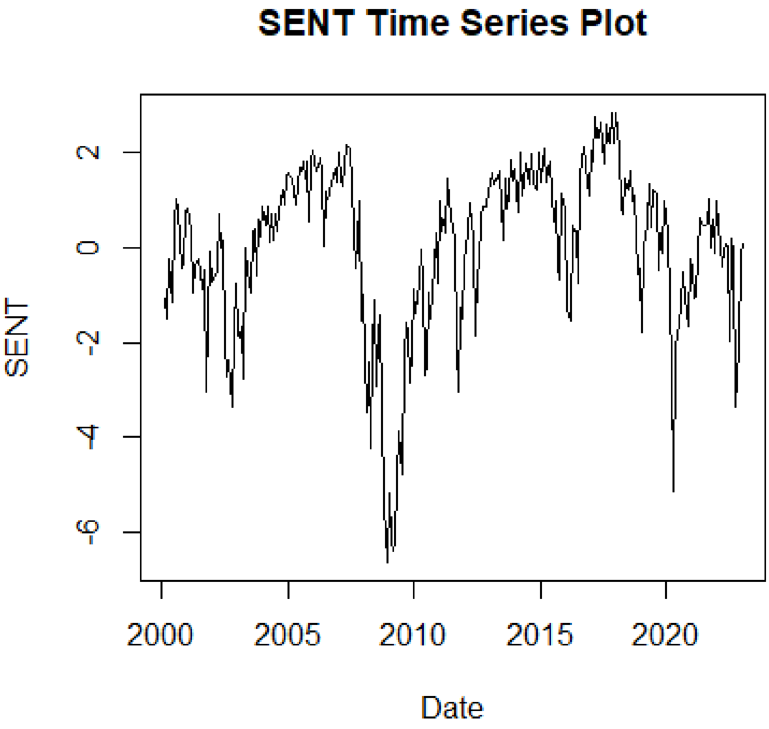


Figure C.2.1: Investor sentiment time series plot

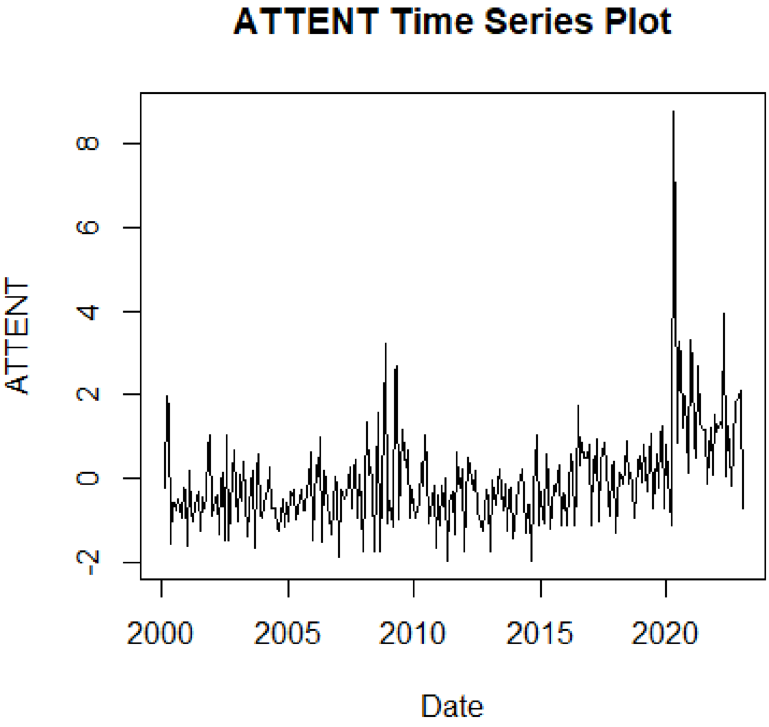


Figure C.2.2: Investor attention time series plot

6.10 Appendix C.3

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.79	1.16	0.94	0.93	0.81	0.78	0.50	0.41
Proportion of Variance	0.40	0.17	0.11	0.11	0.082	0.08	0.03	0.02
Cumulative Proportion	0.40	0.57	0.68	0.79	0.87	0.95	0.98	1.00

Table C.3.1: Importance of components PCA for SENT

AVDC	PCO	PCV	CLIFS	CISS	RSI.30D	IVI.30D	SMART.INDEX
0.24	0.17	0.09	-0.46	-0.47	0.40	-0.50	0.27

Table C.3.2: The element loadings for the first principal component (PC1) to construct SENT composite index

6.11 Appendix C.4

	PC1	PC2	PC3	PC4
Standard deviation	1.18	1.02	0.98	0.77
Proportion of Variance	0.35	0.26	0.24	0.15
Cumulative Proportion	0.35	0.61	0.85	1.00

Table C.4.1: Importance of components PCA for ATTENT

GSV	aeret	aavol	ualhits
0.48	-0.03	0.52	0.71

Table C.4.2: The element loadings for the first principal component (PC1) to construct ATTENT composite index.

6.12 Appendix D.1

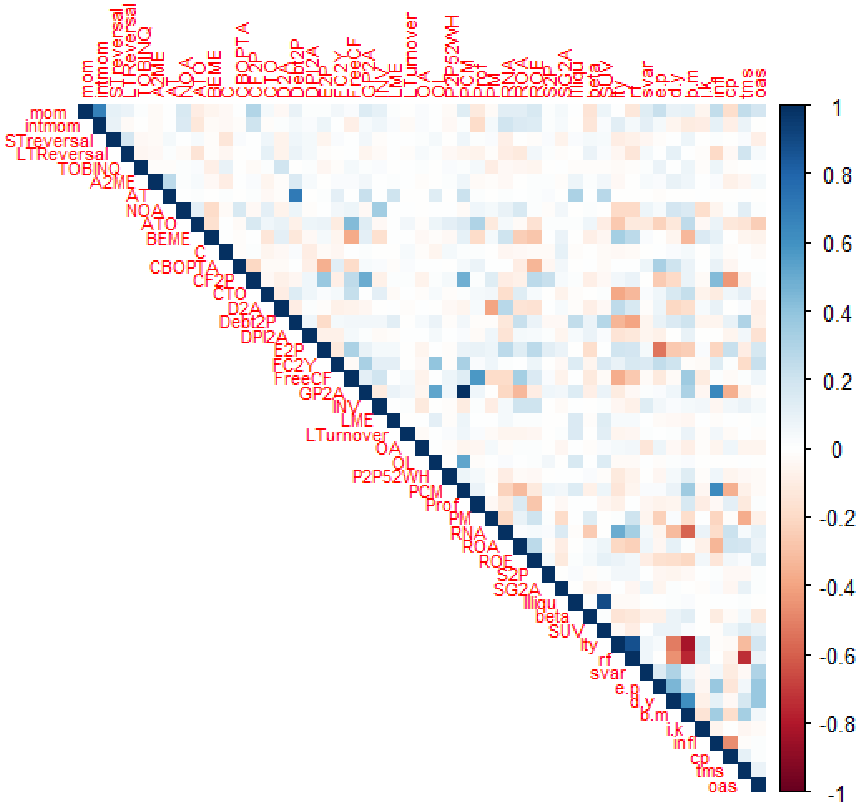


Figure D.1: Firm and macroeconomics Correlation Heatmap

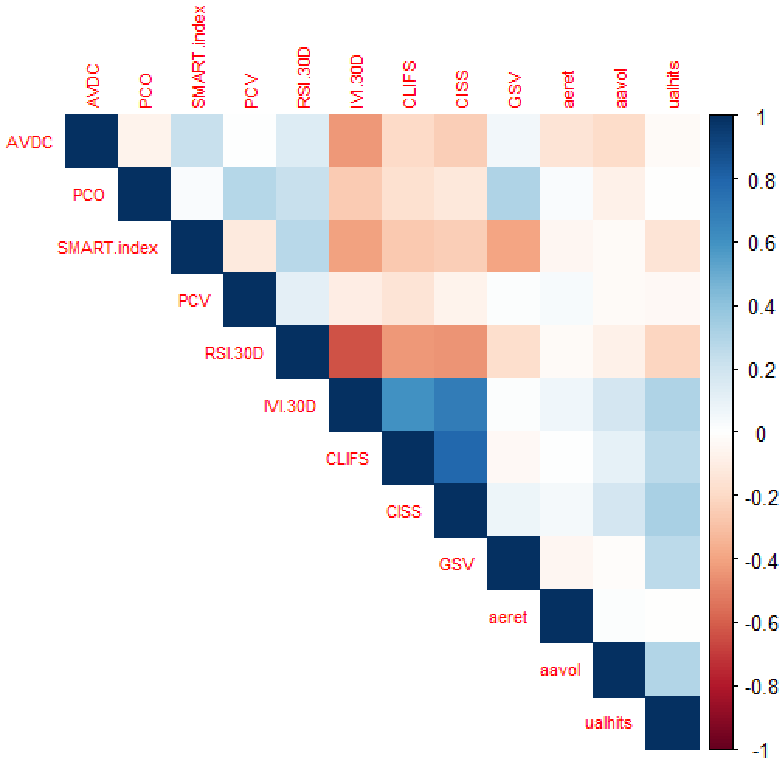


Figure D.1: Investor Sentiment and Investor Attention Correlation Heatmap

6.13 Appendix E.1

	Hyperparameter	**Specification**	**Definition**
OLS-Huber Loss	-	-	-
GLM	α	{0.0, 0.2, 0.4, 0.6, 0.8, 1.0}	alpha controls the distribution between the 1 (LASSO) and 2 (ridge regression) penalties.
DRF	n_trees	{50, 100, 250, 500, 750, 1000}	Number of trees
	max_depth	{5, 10, 15, 20, 25, 30}	Maximum tree depth
	min_rows	{1, 5, 10, 20, 50}	Minimum number of observations for a leaf
GBM	learn_rate	{0.01, 0.1, 0.01}	Learning rate
	max_depth	{2, 10, 1}	Maximum tree depth
	sample_rate	{0.1,0.5 , 1.0}	Row sampling rate
	col_sample rate	{0.1, 1.0, 0.1}	Column sampling rate
	ntrees	50 (default)	Number of trees to build
DNN	activation	{Rectifier","Tanh","Maxout", "RectifierWithDropout", "TanhWithDropout", "MaxoutWithDropout}	Activation function
	hidden	{(32,16,8),(32,16),(32)}	Hidden layer sizes with number of neurons within each layer
	input_dropout_ratio	{0, 0.05, 0.1}	Input layer dropout ratio
	L1	{0, 1e-5, 1e-4}	L1 regularization
	L2	{0, 1e-5, 1e-4}	L2 regularization

Table E.1: Prediction Models' Hyperparameters