Master's Thesis of Daniel Martin Bensley

Cross-Industry Return Forecasting and Economic Linkages: A Machine Learning Approach

산업간수익예측및경제연계: 머신러닝접근방식

August 2022

Graduate School of International Studies

Seoul National University

International Commerce Major

Bensley, Daniel Martin

Cross-Industry Return Forecasting and Economic Linkages: A Machine Learning Approach

Ahn, JaeBin

Submitting a master's thesis of International Studies

August 2022

Graduate School of International Studies

Seoul National University

International Commerce Major

Bensley, Daniel Martin

Confirming the master's thesis was written by

Bensley, Daniel Martin

August 2022

Chair Lee, Soohyung (Seal)

Vice Chair Rhee, Yeongseop (Seal)

Examiner Ahn, JaeBin. (Seal)

Abstract (English)

In this thesis, I adopt a machine learning approach to investigate the predictive power of industry returns using information from lagged sector returns covering the entire U.S. production network. The predictive regression framework identifies key industries able to forecast another individual industry's return lagged by one-month, revealing many economically intuitive customer-supplier relationships between sectors. Network analysis is carried out to examine the relationship between a sector's predictive power and this industry's importance as a customer and supplier in a web of industries. Constructing five out-of-sample industry portfolios, the resulting unprecedented high annualized risk-adjusted returns compared to previous studies, highlight the relevance of the machine learning technique used in this thesis. In accordance with the theory of gradual diffusion of information between interconnected industries, the results are supportive of the existence of information frictions in equity markets.

Keywords: Financial machine learning; Industry interdependencies; Economic network analysis; Post-Elastic-Net-XGBoost regression; Multifactor regression; Industry portfolio

Student Number: 2022-23348

Abstract (Korean)

본 연구는 머신 러닝 기법을 토대로 미국 전체 생산네트워크 상에서 지연 효과를 동반한 산업 분야별 수익률 정보를 이용하여 산업별 수익률의 예측효과를 확인하고자 한다. 같은 정보를 예측 분석 회귀 모형에 대입하여 한달 후의 관련 산업의 수익률을 예측할 수 있는 기간 산업의 종류를 확정하였고, 이를 통해 통상적으로 산업분야 간에 존재하는 소비자-공급자 관계성을 검증할 수 있었다. 또한, 해당 기간 산업이 타 산업의 소비자 또는 공급자 역할을 수행한다는 것을 네트워크 분석을 통해 확인했다. 추가적으로 표본 외 오차에 해당하는 5개 산업의 포트폴리오를 분석한 결과 전례 없이 높은 연율로 환산된 위험조정수익 (Risk Adjusted Return, RAR) 이 존재한다는 사실을 알아냈고, 머신 러닝 모형이 연구 결과를 입증하는데 효과적이었음도 밝히고자 한다. 상호연관성이 높은 산업분야 간에 존재한다고 알려진 정보의 점진적 확산 이론에 기초하여 본 연구 결과를 분석했을 때, 본 연구는 자본시장에 정보 마찰이 존재한다는 통설의 설명과도 부합한다.

Table of Contents

1 Introduction 1

2 Theoretical Background 5

2.1 Neoclassical Asset Pricing Models 5

2.1.1 Expected Utility Theory and Bayesian Updating 5

2.1.2 Market Efficiency Hypothesis 7

2.1.3 Markowitz Portfolio Selection Theory 8

2.1.4 Capital Asset Pricing Model (CAPM) 9

2.2 Behavioural Asset Pricing Theory 10

2.2.1 Gradual Diffusion of Information 11

3 Data & Methodology 11

3.1 Data 12

3.2 Standard Regression Framework 14

3.3 Elastic-Net Regression 14

3.4 XGBoost – Extreme Gradient Boosting 15

3.5 Multifactor Regression 16

3.6 Economic Network Analysis 17

4 Results 20

4.1 In-Sample Regression Results 20

4.2 Economic Network Analysis Results (In-Sample) 23

4.3 Out-Of-Sample Results 26

4.3.1 Industry Portfolio 26

5 Discussion 27

6 Conclusion 29

Reference List 31

1 Introduction

Several economists have investigated stock market predictability in the past and different theories stemming from the neoclassical view and the behavioural finance view have been adopted to determine the driving forces behind asset pricing, with both groups respectively conducting empirical studies to support their stance. While these studies often focus on forecasting the aggregate market returns using economic predictor variables like yield spreads (Ferson and Korajczyk 1995), this thesis analyses industry return predictability using information of lagged industry returns of 30 sectors across the U.S. economy, allowing to investigate complex inter-industry relationships. The idea behind using lagged industry returns to extract information about the returns of interconnected industries one month later, stems from the behavioural finance theory of Hong, Torous and Valkanow (2007). In their model information frictions across several interconnected industries within the economy exist, relaxing important assumptions of neoclassical asset pricing models such as the existence of fully efficient markets and the presence of rational economic agents. Furthermore, cash flow shocks that originate in one industry for any reasons such as rising commodity prices, a recessionary shock, or the emergence of a new technology in one sector, can influence the expected cash flow of linked industries.

In this setting the investors can be viewed as respective industry experts in one sector (think of analysts or asset management firms covering only stocks of firms limited to a few industries) and are unable to process all the information of cash-flow shocks across the whole economy (especially for sectors they do not trade or cover) which in turn hinders them from figuring out all the price-relevant information so that prices across all industries immediately adjust following a cash flow shock. Consequently, information about equity prices gradually diffuses across these related sectors with a certain delay which results in the possibility of predicting returns of industries using lagged industry returns of interconnected sectors.

Inspired by the work of Rapach, Strauss, Tu and Zhou (2015) the predictive power of lagged industry returns is modelled in a way that all the industries can theoretically predict the following month's return of industry n with $N = 30$. This ensures that the regression framework used contains every possible direct and indirect link between the sectors in the U.S. production network. Allowing potentially 30 industries to serve as explanatory variables and predict the t+1 return of a respective dependent variable among all these industries poses serious statistical challenges when applying the ordinary least squares (OLS) method. In this environment the large number of predictors would not only result in overfitted estimators, but another problem would also arise in identifying which industries are the most important ones. With the goal of yielding good out-of-sample predictions, both problems can be tackled using a machine learning approach.

This thesis aims to replicate and extend the main findings of the target paper published by Rapach et. al (2015) by modifying the methodology and broadening the scope of analysis. Using the estimation method called Elastic-Net regression helps to reduce the dimensionality and prevents overfitting. This takes place through the implementation of a regularization parameter, coefficients that are not important are penalized and excluded from the model, thus improving the overall predictive accuracy. Once this estimations method has resulted in a sparser model of important candidate predictors of industry returns, another predictive machine learning model, namely XGBoost is utilized based only on the previously estimated industry sub-selection. XGBoost stands for extreme gradient boosting and is frequently used for time-series forecasting because of its computational efficacy and its superior performance proven during various machine learning competitions. The method developed by Chen and Guestrin (2016) relies on an ensemble of decision trees, where new trees are added to the model to improve the accuracy, while also making use of a regularization parameter to prevent overfitting. Since coefficients are firstly estimated using the Elastic-Net method and afterwards using the state-of-the-art machine learning technique XGBoost, the predictive regression model is labelled Post-Elastic-Net-XGBoost regression.

Regarding the in-sample analysis, I estimate a regression framework via Post-Elastic-Net-XGBoost, using industry return data for 30 sectors from the Kenneth French Data Library during a period starting from 1959 until 2019. The Post-Elastic-Net-XGBoost model selects a total of 190 lagged industry returns as forecasting variables and at least one such predictor is chosen for each of the 30 industries. Furthermore 70 predictors prove to be statistically significant according to a 90 % bootstrapped confidence interval. Next to these findings the results also seem to be economically intuitive while supporting the theory of Hong et al. (2007) that information frictions and boundedly rational investors with limited processing capabilities allow lagged sector returns to affect returns of related sectors gradually over time. This phenomenon can be observed for several industries including the finance sector. Given the importance of this sector as a credit intermediatory to several firms across the U.S. production network, a positive cash-flow shock within this sector is expected to increase the ability to make credit available and improve borrowing conditions for related firms, thus boosting their equity prices. Exactly these linkages are found in the context of the in-sample analysis as lagged financial sector returns are selected 15 times as predictor variables for a wide range of industries.

Another part of the in-sample analysis is the examination of the relationship between the predictive power of lagged-industry returns and importance of those predictor industries as measured by their eigenvector centrality score across the U.S. economy. Similar to Carvalho (2014), I conduct two network analyses using U.S. input-output data from the OECD website for 36 industries for a benchmark year 2008 and 2018. Being able to match 20 of the 36 industries to those industry definitions in Kenneth French’s industry return database, I find a statistically significant relationship between an individual industry’s ability to predict another industry’s return using the Post-Elastic-Net-XGBoost coefficient estimates and that industry’s centrality score. Furthermore, by comparing the two network analyses for different years, I find economically meaningful interpretations of structural changes over time for certain industries like Coal and Finance with the respect of how often these industries were selected as predictor variables and changes in their centrality score for the benchmark year 2008 and 10 years later.

For the out-of-sample analysis I compute forecasts of monthly industry returns for a period of almost 50 years to simulate the situation of an investor and quantify the economic usefulness. Forecasts estimated via Post-Elastic-Net-XGBoost are produced by dividing the return dataset into two different parts, the training and testing set. The training set is made up of data from the beginning of the sample in December 1959 until the last month of December 1969 with the resulting estimates then being used to predict and generate a set of returns for all 30 industries for January 1970. Using a sliding-window or walk-forward method the model is trained again to always include the previously predicted month and forecast the next month's return. Using this approach 5 equally weighted quintile portfolios are constructed by sorting the 30 forecasted industry returns and going short (long) on the bottom/lowest (highest/top) forecasted returns. In the context of a multifactor analysis, the portfolios prove to have, on average, negative or insignificant exposure to the broad equity market as measured by common risk factors with annualized alpha values of up to 18 %, discarding a risk-based explanation for the behaviour of the industry portfolio. For the 1970:01 to 2018:12 out-of-sample period, I compare industry portfolio performances based on Post-Lasso-OLS estimates and Post-Elastic-Net-XGBoost estimates. With the latter generating a 28 percent higher average annualized monthly returns compared to the Post-Lasso-OLS estimates used by Rapach et al. (2015). This demonstrates the usefulness of utilizing the Elastic-Net technique together with XGBoost as it appears to help extracting information of lagged-industry returns to a higher degree.

The structure of this paper is as follows. Section 2 provides a theoretical background of the related literature and core concepts. Section 3 explains the methodology and data used and

Section 4 reports both the in-sample and out-of-sample results. The thesis ends with section 5 and section 6 containing the discussion part and concluding comments.

2 Theoretical Background

Two of the key assumptions of classical asset pricing theory are fully rational investors and the existence of efficient markets. Capital market models built on these assumptions constitute modern finance theory not only in academia but also in practice. Behavioural finance rejects and challenges these key assumptions and has emerged as an alternative theoretical framework aiming to explain the formation of asset prices through empirical research. This chapter will briefly introduce both asset pricing theories, specifically the gradual diffusion of information theory (Hong & Stein, 1999).

2.1 Neoclassical Asset Pricing Models

Proponents of the classical asset pricing theory consider investors to be always rational in a setting where if stocks deviate from their fundamental value, rational investors have countless arbitrage opportunities to immediately identify and correct the mispriced assets back to their fundamental value (Nanayakkara, Nimal, Weerakoon, 2019). Investors are faced with the decision to invest and select assets based on the expected risk-return profile of different assets or portfolios given the existence of an individual utility function (Von Neumann, Morgenstern, 1947) and the belief, that new information will instantly and correctly update their beliefs in a way of maximizing the expected utility according to the Bayes Theorem.

2.1.1 Expected Utility Theory and Bayesian Updating

Von Neumann and Morgenstern (1947) model the behavior of rational investors when making a risky decision considering the investors individual risk preference. In the neoclassical world market participants are generally considered to be risk-averse and must be compensated with additional returns in order for them to be willing to bear more risk. In this setting a utility theory can be postulated, so that rational decision makers choose between different action alternatives A. Each alternative A yields a maximum expected utility a_i according to the individual utility function u and the probability p_i of each n consequences that come with choosing a certain action over another. The expected utility can therefore be expressed following McClave, Benson, Sincich (1998, p.966):

$$\mathrm{EU(A)} = \textstyle\sum_i^n p_i u(a_i) \tag{1}$$

Important to note here is that utility functions must exist which model the individual preferences of market participants and determine the scope of utility achieved. Furthermore, individual utility is determined from using lotteries where the respondent is faced with the decision between a secure payment and a lottery, i.e., a combination of uncertain payment flows. From the results, an approximation procedure is then used to derive the utility function (Von Neumann, Morgenstern, 1947). This theory further assumes that rational investors incorporate all available information in determining probabilities of each consequence. The well-known Bayes Theorem or Bayesian updating refers to an approach by which rationally acting market participants can promptly and correctly incorporate new information. Initially assumed (a priori) probabilities are updated according to new information so that the resulting (a posteriori) probabilities again include all available information (McClave et. al, 1998).

From the neoclassical point of view, it remains to be said that the image of humankind is that of the *homo oeconomicus*. This fictional person represents an aggregation of all individual market participants and computes every decision problem without errors, processes all available information without distortion, thus making it possible to always choose the action that promises the greatest benefit (Brav, Heaton, 2002). While observed isolated, unsystematic deviation from this principal behaviour is possible. But because several effective correction mechanisms implemented by rational, price-determining market participants and arbitrageurs are in place, these deviations can be considered negligible (Ross, 2002).

2.1.2 Market Efficiency Hypothesis

Another building block of classical asset pricing theory is the concept of market efficiency. Fama (1965) states that an efficient capital market exists, where prices of assets automatically entail all historic price data and immediately incorporate new information. Changes in prices are thus a consequence of the release of new information like earnings or macroeconomic indicators, which randomly facilitates itself in the market. The notion of market efficiency as a continuum can be made more concrete by specifying the definition of market efficiency as relative efficiency vis-à-vis a well-defined set of information θ_t. The market is efficient vis-à-vis the information set θ_t, provided that this market incorporates all the information it contains into valuations so quickly that no economically profitable information advantage over the market can be obtained by adding to the information in θ_t (Fama 2014). It can be shown that the pricing process of a market that is efficient in this way corresponds to a martingale with respect to θ_t. That is:

$$\mathrm{E}[p_t + 1|\theta_t] = p_t \tag{2}$$

so that the expected future market price at time t ∈ [1; ... ; T - 1] is p_{t+1} given the information set θ_t is equal to the current market price p_t (Fama, 2014). The current market price is thus graphically the "best prediction" of the future market price, so that adding the information set θ_t does not change the expected value.

One can differentiate between three different forms of market efficiency (Fama, 1970):

- If θ_t entails only the historic past prices of asset, the market is set to be weakly efficient.

- If θ_t is now the quantity of all public price-relevant information, the market is said to be semi-strongly efficient. This implies the weak-form efficiency plus publicly available information such as government statistics, macroeconomic indicators, and accounting information of companies as part of the information set (Ross, 2002)

- If θ_t is the quantity of all, i.e., also non-public, price-relevant information, the market is highly or strongly efficient. Even private information not known to most of the market participants is reflected in asset prices immediately (Ross, 2002).

2.1.3 Markowitz Portfolio Selection Theory

The central idea of the modern portfolio theory as presented by Markowitz (1959) is that all investors, acting according to the neoclassical assumptions and being rational and risk-averse, prefer a selection of efficient investment alternatives independent of their personal preference functions. Two main metrics are introduced to evaluate and rank investment alternatives, thus abstracting from individual investor preference using the principle of μ and σ. Among investment alternatives with comparable expected return (μ), those with the lowest risk (σ) (variance of the returns) are preferred, or with comparable risk, those with maximum expected return are preferred. Efficient investment alternatives are those that have an optimal risk-return ratio according to the $\mu\sigma$-principle (Markowitz 1959).

Another core principle of the theory is the realization that the possibility of investment risk reduction can be achieved through diversification due to the covariance of individual value returns. This diversification effect depends on the correlation of the returns of the individual assets that the portfolio constitutes of, while the expected return of the portfolio is simply the weighted average of the expected individual asset returns (Markowitz 1959). The correlation coefficient can take values between +1 and -1. If two stocks e.g., the Kakao Corp. and Naver Corp. assets are perfectly correlated (+1) they move identically, if one asset gains 7 percent the other gains 7 percent and vice versa. In this case no reduction of the investment risk is achieved through portfolio formation. Perfectly negatively correlated assets (-1) would move inversely to each other, if one gains 5 percent the other one drops 5 percent, thus reducing the overall risk of the portfolio substantially.

2.1.4 Capital Asset Pricing Model (CAPM)

Building on the modern portfolio theory, Sharpe (1964) developed the Capital Asset Pricing Model, which extends the statement of the portfolio selection theory in a way that all investors in the market equilibrium hold the same investment portfolio regardless of their individual preference function. If all individual securities available in the market are considered, homogeneous investor expectations are assumed and the possibility of investing or borrowing at the risk-free interest rate is taken into account, then the selection of efficient investment alternatives is reduced to combinations of the so-called market portfolio and investment or borrowing at the risk-free interest rate (Fama, French 2004). The Sharp-Lintner CAPM captures this risk-return relationship through the following equation (Ross, Westerfield, Jaffe, 1999, p.260):

$$\mathrm{E}(R_i) = R_f + (\sigma_{im}/{\sigma_m}^2) * [\mathrm{E}(R_m - R_f)], \tag{3}$$

where the expected return $\mathrm{E}(R_i)$ of an individual asset i belonging to the market portfolio is shown as a function of the expected market return $\mathrm{E}(R_m)$ minus risk-free interest R_f, the covariance of the returns of the individual asset and the total market σ_{im} as well as the total market risk σ_m. The ratio of $(\sigma_{im}/{\sigma_m}^2)$ is also called the market beta. The assumption of market efficiency imply that all securities are quoted at market clearing prices and that the market is in equilibrium. In market equilibrium, a linear relationship between return and risk (beta) applies, whereby only the systematic risk that cannot be eliminated through diversification is being assessed.

2.2 Behavioural Asset Pricing Theory

Behavioural finance, which has been developing since the late 1970, breaks with the core beliefs of the neoclassical asset pricing theory, especially the assumption of rational investors and efficient markets. By drawing on insights from psychology and decision research, the theory attempts to explain observable financial market inefficiencies through systematic irrational behaviour patterns of market participants. Simon (1955) was the first to establish a behavioural finance theory questioning the existence of rational agents. One starting point of the research direction is the result of numerous laboratory experiments in the field of cognitive psychology, which confirm situational irrational investor behaviour. Another starting point are the numerous empirical capital market studies, which prove that securities prices in practice deviate from the specifications of the established capital market models. These inefficiencies of financial markets lead to asset mispricing and are often referred as market anomalies. According to the theory, these market anomalies are caused by market participants acting irrationally and these irrational behavioural patterns being of systematic nature. In sufficiently liquid capital markets, isolated irrational investor behaviour does not affect the prevailing market equilibrium prices. Only synchronous irrational behaviour by many market participants can cause at least temporary price deviations from the equilibrium level. Then, for example, the over- or underreaction of market participants to certain new information translates into an over- or underreaction of asset prices (Hirshleifer, 2001).

2.2.1 Gradual Diffusion of Information

In the paper “A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets” Hong and Stein (1999) develop a theory that grounds on the behavioural abnormality of limited cognitive capacity information processing of market participants. The authors develop a model based on the distinction between two investor groups, each of which processes only part of the available information: Newswatchers, who form expectations based on private, forward-looking fundamental information, and momentum traders, who simply extrapolate past price developments into the future. It is assumed that new information is only slowly absorbed by newswatchers (this is especially true for small company stocks and stocks with low analyst coverage), which leads to a delayed price reaction (underreaction). However, if short-term momentum traders recognise a price trend in the delayed correction of the underreaction and enter the market, the market successively overreacts. In this setting market participants are boundedly rational and cannot extract all information from prices, thus under-reacting in the light of new information which results in stock return predictability. In their later study, Hong and Stein (2007) investigate the ability of industries in forecasting the stock market. The underlying hypothesis is that based on the idea that valuable information such as macroeconomic fundamentals originates from selected industries and diffuses only gradually (with a lag of one to two months) into the aggregate stock market, leading to return predictability across the economy. Key assumption for this theorem includes limited information processing capabilities and the notion of agents to ignore or be inattentive to asset pricing information arising in industries in which they do not specialize.

3 Data & Methodology

The following chapter will clearly lay out the methodology of this master thesis. Different regression models will be explained in the context of the research question. Given the complexity of some of the machine learning models, the goal here is not only to provide an intuition of how they work mathematically but also explain why they were selected and how they differ compared to the original target paper methodology. The same accounts for data used in my work, specifying which data sets from which sources and over what time span was utilized.

3.1 Data

To retrieve return data for different industries across the U.S. economy, this thesis makes use of the Kenneth French database[1], where monthly excess returns (in excess of the 1-month U.S. treasury bill) of 30 value-weighted industry portfolios can be downloaded from. The extensive dataset spans from July 1926 to December 2021 and industry portfolios are constructed by assigning each NYSE, AMEX and NASDAQ stock to an industry portfolio and updating the portfolios each year. In contrast to Rapach et al. (2015) an updated data set of monthly returns ranging from January 1960 until December 2018 is being used. **Table 1** reports standard metrics and summary statistics. Compustat or CRSP SIC codes are used to match industries and abbreviations are as follows: **Food** = Food Products; **Beer** = Beer and Liquor; **Smoke** = Tobacco Products; **Games** = Recreation; **Books** = Printing and Publishing; **Hshld** = Consumer Goods; **Clths** = Apparel; **Hlth** = Healthcare, Medical Equipment and Pharmaceutical Products; **Chems** = Chemical Products; **Txtls** = Textiles; **Cnstr** = Construction and Construction Materials; **Steel** = Steel Works Etc.; **Fabpr** = Fabricated Products and Machinery; **Elceq** = Electrical Equipment; **Mines** = Precious Metals, Non-Metallic, and Industrial Metal Mining; **Coal** = Coal; **Oil** = Petroleum and Natural Gas; **Util** = Utilities; **Telcm** = Communication; **Servs** = Personal and Business Services; **BusEq** = Business Equipment; **Autos** = Automobiles and

[1] Available here: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library

Trucks; **Paper** = Business Supplies and Shipping Containers; **Trans** = Transportation and Warehousing; **Wholesale** = Wholesale Trade; **Rtail** = Retail Trade; **Meals** = Restaurants, Hotels, Motels; **Fin** = Banking, Insurance, Real Estate, Trading; **Carry** = Aircraft, Ships, and Railroad Equipment; **Other** = Everything Else. The results are very much in line with the target paper results. However, while **Smoke** (Tobacco products) also displays the highest average annualized returns with 9.53%, **Food** (Food products) has the highest Sharpe ratio of 0.48. The lowest Sharpe ratio (0.04) and average annualized return (1.05%) during this period is observable for the Steel industry (Steel works).

The OECD input-output table provides information about the transaction value of intermediate and final expenditures between industries in an economy measured in US dollars. To derive a measure of importance of individual sectors within the U.S. production network, I calculate the eigenvector centrality score for the 20 industries from the Kenneth French database that can be reasonably matched according to the OECD industry classification. Following Rapach et al. (2015) the matching was conducted as follows (with OECD definitions in brackets): **Food** (Agriculture, Hunting, Forestry, and Fishing), **Books** (Pulp, Paper, Paper Products, Printing, and Publishing), **Hlth** (Health and Social Work), **Chems** (Chemicals and Chemical Products), **Txtls** (Textiles, Textile Products, Leather, and Footwear), **Cnstr** (Construction), **Steel** (Basic Metals), **Fabpr** (Fabricated Metal Products Except Machinery and Equipment), **Elceq** (Electrical Machinery and Apparatus N.E.C.), **Autos** (Motor Vehicles, Trailers, and Semi-Trailers), **Carry** (Other Transport Equipment), **Coal** (Mining and Quarrying), **Oil** (Coke, Refined Petroleum Products, and Nuclear Fuel), **Util** (Electricity, Gas, and Water Supply), **Telcm** (Post and Telecommunications), **Servs** (Other Community, Social, and Personal Services), **Buseq** (Computer, electronic and optical equipment), **Trans** (Transport and Storage), **Meals** (Hotels and Restaurants), **Fin** (Financial Intermediation).

3.2 Standard Regression Framework

The main regression framework is identical to the regression formula found in Rapach et. al (2015, p.6):

$$r_{i,t+1} = a_i + \sum_{j=i}^{N} b_{i,j} r_{j,t} + \varepsilon_{i,t+1} \text{ for } t = 1, \ldots., T-1 \text{ and } N = 30, \quad (4)$$

where $r_{i,t}$ represents the return of a respective industry portfolio i beyond the return of the one-month Treasury bill return at time t. The analysis entails the returns of 30 industries and thus N is set to 30. To account for randomness a zero-mean error term representing the margin of error within the model $\varepsilon_{i,t}$ is introduced. Since this model in theory allows all 30 industries one-month lagged excess returns to influence and predict all other industry's returns in the following time period, the equation can be considered a vector autoregression of the first order (Rapach et. al 2015).

3.3 Elastic-Net Regression

The Elastic-Net model combines the features of the two regularization methods Ridge- and Lasso Regression. One challenge to overcome using a basic regression model with many predictors (independent variables) are the poor out-of-sample results. Instead of using only the adaptive Lasso regularization method introduced in the target paper (Rapach et. al, 2015), the idea is to combine both the variable selection and coefficient shrinkage effects and identify a penalty term that eliminates and shrinks lagged industry coefficients depending on their quality of improving the model’s out-of-sample performance. Comparable with the work of Varian (2014), when determining appropriate coefficients for the predictor variables a function of the sum of squared residuals plus a penalty term must be minimized:

$$\arg\min[(r_{i,t+1} - \textstyle\sum_{j=i}^{N} b_{i,j}\tilde{r}_{j,t})^2 + \lambda_i \sum_{j=1}^{N}[(1 - \alpha)\left|b_{i,j}\right| + \alpha\left|b_{i,j}\right|^2], \tag{5}$$

where $\tilde{r}_{j,t}$ is the excess return for each industry, λ_i the parameter controlling the degree of regularization and a determining the relative weight and thereby the ratio of the type of penalty (l_1 or l_2 penalty). The goal here is to identify a sub-sample of industries deemed the most important for predicting lagged-industry returns based on the Elastic-Net coefficient estimation.

3.4 XGBoost – Extreme Gradient Boosting

The open-source machine learning model XGBoost was first published and described by Tianqi Chen and Carlos Guestrin in the paper “XGBoost: A Scalable Tree Boosting system” in 2016. The highly-efficient system has been widely recognised and several data mining and machine learning competitions have been won utilizing XGBoost (Chen and Guestrin, 2016). XGBoost is a special implementation of the gradient boosting algorithm and refers to a class of ensemble machine learning models which can also be used for regression predictions. This technique yields predictions in the form of several weak prediction models which are typically decision trees. Using an optimization algorithm and an appropriate cost function, weak learners are iteratively combined into a single strong and accurate learner (Friedman 2002). In addition, XGBoost also uses a regularization parameter to control the model’s complexity and tackle overfitting issues, which make it a suitable candidate in the context of this thesis research question given the plethora of possible lagged industry returns to choose from. According to Chen and Guestrin (2016, p.2) the following regularized learning objective is being minimized:

$$\left[\sum_{i=1}^{N} L(\tilde{r}_j, \qquad r_i)\right] + \gamma T + \frac{1}{2}\lambda\|\omega\|^2, \tag{6}$$

with L being a differentiable loss function measuring the difference between the predicted industry return $\tilde{r}_j$ and the target industry return r_i and λ being a nonnegative regularization parameter. It is important to note that in the process of determining the final regression coefficients and predicting the next month's industry returns, this machine learning method described is used on a sub-set of important predictor industries of lagged industry returns for each of the 30 industries which were determined during the preceding Elastic-Net regression. Given the pre-selection of important industries, the XGBoost linear algorithm is employed to estimate coefficients displayed in **Table 2**.

3.5 Multifactor Regression

There are numerous empirical capital market studies about market anomalies, which prove that in practice, the prices of securities deviate from the specifications of the established capital market models and by no means directly include all available information, as postulated by the EMH (De Bondt, Thaler 1985). The performance of the out-of-sample industry portfolios based on the Elastic-Net-XGBoost regression forecasts constitute such an example of presumably unexplained market anomalies. To test for the exposure to common risk factors, multifactor models can be utilized to express the industry portfolio returns as a function of different explanatory variables. Similar as in Rapach et. al (2015) the following regression model is estimated:

$$r_{p,t} = \alpha + \beta_{MKT} MKT_t + \beta_{HML} HML_t + \beta_{QMJ} QMJ_t + \beta_{SMB} SMB_t + \beta_{UMD} UMD_t + e_{p,t}, \quad (7)$$

where $r_{p,t}$ is the industry portfolio return for the five out-of-sample portfolios, MKT_t represents the market factor, SMB_t and HML_t the Fama and French (1993) "small-minus-big" and "high-minus-low" size and value factors, UMD_t ("up-minus-down") a momentum portfolio and QMJ_t representing the "quality-minus-junk" factor.[2] By the time calculations were performed, data for the Pastor and Stambaugh (2003) liquidity factor was not available. Results of the multifactor regression are reported in **Table 4.1/Table 4.2** and discussed in chapter 4.2.1

3.6 Economic Network Analysis

One of the major further developments of this thesis in contrast to the target paper is to perform an economic network analysis using results of the redefined and methodological differing Elastic-Net and XGBoost machine learning models, as well as using different time periods and comparing the results for 2008 and 2018. Performing an economic network analysis should provide insight as to which industries are considered central hubs formed by their dependencies on one another for input and output factors (Aobdia, Caskey and Ozel 2012). Given the underlying theory of gradual diffusion of information (Hong et al. 2007), special emphasis is placed on measuring an industry's centrality and deriving a measure to quantify the extent to which an

[2] All factor data is extracted from Kenneth French's Data Library (MKT, HML, SMB, UMD) and the AQR website (QMJ) available at https://www.aqr.com/library/data-sets

individual industry has strong or weak ties to the entirety of industries constituting this network of industries. Characterizing the U.S. production network as a matrix allows for the use of a degree of network centrality called eigenvector centrality (Katz, 1953). As opposed to other measures of centrality like degree centrality, the eigenvector measure captures complex direct and indirect relationships between industries. Next to measuring the direct network effects of a node's importance, this measure also reflects indirect links where an industry can be important and central not only because of its own role and respective customer and supplier relationships but because this industry in an input supplier to a few nodes that themselves have strong links to other industries. In the presence of boundedly rational economic agents and information frictions the propagation of cash-flow shocks from one industry to another and therefore the ability of an individual's industry's return predictability is likely to be a function of the respective industry's centrality (importance) in the industry production network (Rapach et. al, 2015).

Network analysis is based on graph theory and in principle, every graph G consists of two sets, the set V of nodes and the set E of edges, which can be expressed as $G = (V, E)$. In this setting, each industry in a network is a node. If an industry is a supplier or customer to another industry, this is represented by the fact that there is an edge between these two nodes. For each industry the flow of inputs can be modelled using matrix notation where matrix W consists of $w_{a,b}$ elements representing the fraction of industry b in the total inputs used by industry a. Each nonzero $w_{a,b}$ element can be considered a directed edge (intersection of two industry nodes) with certain edge weights assigned (Rapach et. al, 2015). Following Cavarlho (2014) the centrality score can be calculated according to this formula:

$$\boldsymbol{c} = (0.5/\mathrm{N})(I - 0.5W')^{-1}\mathbf{1}_N, \quad (8)$$

where $\boldsymbol{c}$ is composed of all 30 industries eigenvector centrality scores and high centrality scores highlight important nodes in the web of industries (Rapach et. al, 2015). While the results in **Table 2** reveal which industries predict lagged cross-industry return how often and to what extent for the year 2018, the OECD input-output table for the United States is used to compute each industries eigenvector centrality score. Currently three editions of the input-output tables are available (2015, 2018 and 2021 edition) and since the OECD industry classification does not align perfectly to the industry definition for Kenneth French's return data, some of the 36 industries available from the OECD cannot be reasonably matched. Comparing the predictive regression results up until the end of the benchmark year 2008 with OECD data for the same period, the Kenneth French Data Library industry definitions can be matched to 20 OECD industry definitions following the methodology used by Rapach et. al (2015, p.13) and described in Section 3.1. The most recent data available from OECD is the 2021 edition of input-output data for the year 2018. Deviations of industry definitions of different OECD input-output table editions led to a different mapping methodology used in some cases. These include the following ones, while the rest remains unchanged compared to the 2008 version (Brackets include corresponding OECD definition). **Books** (Paper products and printing*)*, **Elceq** (Electrical Equipment), **Servs** (Other service activities*)*, **Games** (Arts, entertainment and recreation) **Meals** (Accommodation and food service activities*)*, **Fin** (Financial and insurance activities*)*.

Estimation results for the Post-Elastic-Net-XGBoost regression framework trained until the end of 2008 and 2018 are related to the centrality score of industries using the input-output table for the respective years 2008 and 2018. Their relationship can be visualized by drawing a scatterplot. Repeating this process for updated data in 2018 against results for the benchmark year 2008 allows to test the validity of possible evidence for the existence of gradual diffusion of information in the context of predictability of lagged industry returns and allows to reveal interesting industry patterns which will discussed in **Section 4** in detail.

4 Results

4.1 In-Sample Regression Results

Following the approach of Rapach et. al (2015) to estimate predictions for the Post-Elastic-Net-XGBoost predictive regression, monthly excess returns of 30 value-weighted industry portfolios were used. **Table 2** presents the Post-Elastic-Net-XGBoost coefficient estimates of each industry covering an estimation period from January 1960 until December 2018 (709 observations). In this setting, employing the Elastic-Net algorithm to select a sub-sample of important predictor industries, has certain advantages compared to running a simple Lasso Regression as it combines both the l_1-and l_2-penalty terms. In the same fashion as Lasso Regression, the algorithm yields coefficient estimates that result in sparser estimates through an l_1-penalty term which shrinks the slope coefficient estimates even to zero for some industries, thus excluding them from the model. This results in a reduction of complexity and prevents overfitting. A disadvantage of the l_1-penalty is that while it is generally good at selecting only the most relevant predictor variables (Zhang and Huang 2008), models can suffer from downward biases and "overshrinking" (Rapach et al. 2018). Instead of risking that important coefficients and industries are being removed from the model, potentially resulting in these statistical drawbacks, this paper relies on utilizing both penalty terms. The combination of a moderate l_1-penalty term only reducing the dimension of the model to a certain extent and a l_2-penalty term to penalize but not leave out less important slope coefficients used in this thesis, generally performs well in choosing the most relevant predictor industries.

Different statistical issues revolving around post-selection inference and multiple testing of several isolated null hypotheses arise in the context of the in-sample tests (Rapach et. al 2018). Since these topics are beyond the scope of this work but resemble a similar methodology used in the target paper, reference to the original paper section is recommended, where the authors tackle this issue in detail using the Benjamini and Hochberg (2000) procedure. **Table 2** reveals that a total of 190 Post-Elastic-Net-XGBoost regression estimates were selected and according to a 90 % bootstrapped confidence interval that was constructed to check for statistical significance, 70 (marked bold in the table) coefficient estimates are significant at the 10 % level. The results together with the fact that the machine learning model selects at least one return predictor for all the industries and in some cases up to 11 predictors for a single industry, underscores the relevance and predictive power that one-month lagged industry returns entail.

In accordance with the results of Rapach et al. (2018), autocorrelation does not seem to play a big role when applying the Post-EN-XGBoost model since only six among the 190 lagged industry return predictors were picked as an industry's own lagged return. Taking the main underlying theory of gradual diffusion of information across inter-connected industries (Hong et al. 2007) into consideration, most of the coefficient estimates appear to be economically intuitive and confirm relationships inferred by the authors in the target paper. While this paper extends the methodology and compares estimation results for the periods ranging up to 2008 and 2018 the financial industry (**Fin**) remains among the most selected lagged industry predictors in both cases, being picked 17 and 15 times respectively. Lagged coal returns also seem to entail considerable predictive power as they are selected 23 (2008) and 24 times (2018). The authors point to certain industries like **Fin**, **Coal** and **Oil** to explain how estimations results are plausible within the US network of industries from an economic perspective (Rapach et. al 2015).

The results of this paper line up closely and reveal similar predictive relationships between related industries in accordance with the theory of gradual diffusion. For the updated period ranging until 2018, lagged returns coefficients for this industry are positive and 7 of them significant, resulting in estimates up to 0.18 in scope for the **Steel**, **Txtls** and **Serv** industry. Since financial intermediaries provide credit to a wide range of firms, they usually play a central role in the economy. Any positive increase in returns for financial firms increase the credit availability and borrowing capacity of interconnected industries, subsequently boosting returns in these sectors (Rapach et. al 2015). Not only positive, also negatively related relationships between industries can be observed in **Table 2**. Among these a very interesting pattern emerged, lagged returns for **Hshld** (Consumer Goods) are highly negatively related to **Autos** (Motor Vehicles etc.) returns. A potential explanation of this relationship could be derived from delayed consumption behavior. Economic agents will usually save some funds prior to making a big purchase, e.g. buying a car, and thus forgoing or delaying the consumption of other consumer goods or vice versa prior to that purchase. The **Hshld** industry also includes expensive goods like jewelry, watches, and furniture. Increased returns in the **Hhsld** sector could mean that consumers that spent a considerable amount of their funds for these consumer goods in month t, do not have enough disposable income in month t+1 to afford purchasing a motor vehicle which in turn negatively affects the sales of cars in month t+1. Another interesting pattern that can be observed is that lagged **Food** (Food products) returns are positively related to **Beer** (Beer & Liquor) returns. **Food** products consists of meat and dairy products, but also agricultural services and production (crops). Economically intuitively one can imagine different factors like a particularly good (bad) harvest or favorable (unfavorable) trading agreements increasing (decreasing) the supply of wheat and malt, or other materials used in the production of beer. The translating effect of commodity price shocks and the dependencies of beer producers on agricultural products can be captured by the positive relationship between lagged **Food** product returns and increased equity prices in the **Beer** industry one period later due to the interdependencies between those industries.

4.2 Economic Network Analysis Results (In-Sample)

While subchapter 3.5 covers the data and methodology part of the economic network analysis, this section investigates the relationship between the number of times lagged returns of industries based on estimates from the Post-Elastic-Net-XGBRegression were chosen as variables to predict the following month's return and the centrality score of these respective sectors in terms of eigenvector centrality. **Figure 1** displays in-sample estimation results for 20 industries up to 2008 plotted against the same industry's centrality scores and calculated based on the 2008 OECD input-output table of the U.S. industry network. Repeating this exercise for the year 2018 yields similar but notably different results, which are displayed in **Figure 2**. Having the possibility to compare these two figures allows to examine changes in the composition of important industries over time and explore which industries were selected most frequently in 2008 versus 10 years later.

Regressing these two variables reveals a linear relationship that is significant at the 5% level for both 2008 and 2018 with p-values of 0.02 and 0.05 respectively. Both figures point to a positive connection between the two variables and showcase a high degree of explanatory power (R^2 metric of 50% in 2008 and 43.9% in 2018), confirming the findings of Rapach et al. (2015). I selected the year 2008 as the benchmark year because of the emergence of the Global Financial Crisis of 2007 to 2008 and the corresponding economic disruption across the world. In terms of eigenvector centrality score and the number of times lagged returns of an industry were selected via the Post-Elastic-Net XGBoost regression method, a few differences between the two scatterplots emerge. The top three industries **Coal**, **Steel** and **Fin** remain the same for both years, but as shown in **Figure 1**, while the **Coal** industry was the most frequently selected sector with the highest centrality score in 2008, in 2018 **Fin** was ranked first and **Coal** third in terms of the centrality score. A likely explanation for this change could have been that many financial and insurance companies were struggling in the aftermath of the Global Financial Crisis and numerous big players of the industry suffered big losses or bankruptcy which in turn could have led to a reduction in input received from and output supplied to other industries during this period. In 2018 however there was no substantial economic recession to the scope of the Global Financial Crisis and the size of the financial markets in the United States amounted to a total of \$1.5 trillion or 7.4 percent measured in U.S gross domestic product[3]. The size of this sector likely resulted in a higher centrality score of the **Fin** sector compared to the score value influenced by crisis conditions 10 years prior. According to data provided by the U.S. Energy Information Administration[4] coal power consumption peaked in 2008 with over one billion tons used per year and decreased by almost 40 percent in 2018 through the promotion of renewable sources of energy and natural gas. Considering the diminishing role of the **Coal** industry as an energy provider for related industries over time, this could explain the decreasing centrality score showcased in **Figure 2**.

In **Figure 1** the R^2 statistic is higher than in **Figure 2** and the same accounts for differences in the estimated slope steepness (0.41 and 0.32). Although it is hard to pinpoint exactly which underlying mechanism are causing these differences and the US economy likely experienced a considerable structural shift over the 10-year period, the comparatively higher explanatory power for the 2008 data between the high level of information of lagged cross-industry returns and the centrality score measure could indicate a stronger predictive accuracy of lagged industry

[3] Source: selectusa.gov

[4] EIA, Monthly Energy Review – March 2021, table 6.2

returns in times of economic distress and coincides with the target paper's discoveries. Instead of suffering financial losses, Rapach et. al (2015) find that their out-of-sample industry-rotation portfolio gains in times of economic recessions including during the Global Financial Crisis. Considering information frictions and industry interdependencies among firms operating in different sectors, this could indicate that in times where shocks propagate throughout the whole U.S. production network the pervasiveness of lagged industry returns, originating from industry interdependencies and gradual diffusion of information, as well as boundedly rational investors unable to extract pricing information, is higher in crisis times. (Rapach et. al 2015)

The result of the attempt to draw an exact comparison with the economic network analysis methodology used in the target paper by using input-output data from the mid 2000s and out-of-sample portfolio returns until the end of 2014 can be seen in **Figure 3**. While the plot looks similar, some differences between the importance of a few industries as measured by their eigenvector centrality score as well as the number of times the model selects predictor sectors as forecasting variables arise. While the latter can boil down to different algorithms used in the process of determining critical predictor industries, some differences in the degree of eigenvector centrality could be the result of slight modifications in calculating the eigenvector values used by the authors. Interestingly, the COAL and FIN sectors remain the most important industries in both the original and the replicated version of the scatter plot. However, the attempt of using a benchmark year and comparing the results of the scatter plots over time remains to offer an advantage as it allows to investigate the changes of both variables and allows for the interpretation of transformational shifts in the economy.

4.3 Out-Of-Sample Results

4.3.1 Industry Portfolio

To compute the out-of-sample forecasts based on the predictive Elastic-Net-XGBoost regression estimates for monthly industry returns, five zero-investment portfolios are constructed and evaluated regarding certain metrics such as annualized returns, Sharpe ratio and in the context of a multifactor regression to check for the exposure to common risk factors. To assess the predictive ability of lagged industry returns, data from December 1959 until December 1969 is being used as the testing set, producing out-of-sample forecast for each of the 30 industries for the month of January 1970. Quintile portfolios are constructed by ordering the 30 forecasted returns into five equally weighted portfolios (6 industries per portfolio) and going short on the bottom part and long on the highest forecasted returns. This process is repeated in a walk-forward fashion, meaning that the model is again trained up until January 1970 to predict the following month of February 1970 over the whole sample period until December 2018 (580 months). This method ensures that no information in the testing set that was not available at the time the variables were selected by the machine learning is used, effectively simulating investors behaviour in real time and avoiding biases (Rapach et. al 2018).

Table 3 reports out-of-sample summary statistics for the industry portfolios and compares the performance of the constructed portfolios for the Post-Lasso-OLS forecasts, the machine learning technique used by Rapach et al. (2015), and the Post-Elastic-Net-XGBoost forecasts. In both cases the fifth portfolio (going long on the 6 highest forecasted returns) performs the best. However striking differences between the two estimates exist. Forecasts based on the Post-Elastic-Net-XGBoost estimates deliver a mean monthly annualized return of 14,50 % over the sample period versus only 10,33 % for Post-Lasso-OLS forecasts. Furthermore, the annualized volatility observed is around 17 % in both cases which results in substantially higher risk-adjusted returns as measured by the Sharpe ratio (0,826 versus 0,598 respectively). This increase in performance highlights the value-added implementing the Post-Elastic-Net-XGBoost regression framework as it seems that these forecasts appear to be even more informative than Post-Lasso-OLS forecasts.

After the construction of the zero-investment quintile portfolios based on the out-of-sample estimates, the exposure to common risk factors is tested to analyse the behaviour of the industry portfolio and whether it is possible that a risk-based explanation for the observed returns can be discovered. **Table 4.1** reports the estimation results for Equation 7 for the industry portfolio based on Post-Lasso-OLS forecasts, and **Table 4.2** reports the estimation results for Equation 5 but based on Post-Elastic-Net-XGBoost forecasts. Both forecasts exhibit insignificant negative exposure to the market factor, significant negative exposure to the quality minus junk and momentum factor but also positive exposure to both the value and size factor. This paints a slightly different picture compared to the findings of Rapach et al. (2018) who find that the betas for all factors tested are statistically and economically insignificant. However, the industry portfolios still generate substantial significant annual alpha values of 14.4 % for the Post-Lasso-OLS (Table 4.1) and even 18 % for the Post-Elastic-Net-XGBoost forecasts (**Table 4.2**) producing extremely informative risk-adjusted average returns. These results together with the findings of Section 4.1 highlight the predictive power lagged industry returns entail for forecasting individual sector returns as a result of complex interdependencies of industries and gradually diffusing information of equity prices following cash-flow shocks.

5 Discussion

The findings of this work contribute to the existing literature not only by verifying most of the target papers results but also by yielding new ones. Incorporating a different methodology and combining the two machine learning models XGBoost and Elastic-Net further improves the out-of-sample performance as mentioned in the section above. Another refinement of the methodology used in this thesis is achieved by performing an economic network analysis for a benchmark year 2008 and comparing the results for data from 2018 to discuss changes in the predictive ability of lagged industry returns and the underlying economic structure of the U.S production network.

Overall, the results reported in **Table 2, 3, 4** and in both **Figure 1** and **Figure 2** cast further doubt regarding the existence of fully rational investors and a frictionless equilibrium, in which equity prices immediately incorporate all relevant price forming information. In contrast, more evidence is produced to support the theory of Hong and Stein (1999). Lagged cross-industry returns do seem to inherit a certain predictive power aptitude, supporting the claim that limited information-processing capabilities of boundedly rational investors exist. Investors who are specialized in trading assets of one industry tend to overlook cash-flow shocks that affect other sectors through complex industry interdependency mechanisms and since information about equity prices thereafter gradually diffuses across the base of investors, these frictions give rise to industry return predictability (Hong, Torous, Valkanov, 2007).

Although several economically reasonable links between lag and lead industries can be established this is not the case for every individual industry and their ability to predict monthly-lagged returns of another sector. In **Table 2** we can e.g., observe that lagged **Clths** returns predict **Cnstr** returns one month later. Although there might exist a logical explanation for these interdependencies, it is less intuitive than the other relationships presented in the results section. The use of machine learning tools in econometrics is still comparatively new, also because results of complex models are often difficult to interpret and often increased predictive performance of a machine learning model comes at the cost of decreased interpretability (Linardatos, Papastefanopoulos, Kotsiantis 2021). In this regard more research has to be conducted to shed light on the patterns discovered by applying machine learning techniques to economic data sets.

Publications of replication studies in economics are usually only being accepted by journals if they falsify an often-cited original study. The verification of findings stemming from complicated empirical studies goes unnoticed by many economists even if it can be considered a valuable service to both the literature and society itself. By replicating and extending existing studies, the goal is to gain a deeper understanding of the topic and to assess whether the data and methodology can be fine-tuned, and the results can be checked for robustness. The fact that findings from the target paper can be reproduced and new ones yielded using a fine-tuned methodology and updated data sets, increases the value of the original empirical work and should encourage economist to further explore variations, extensions, and limitations of topics in financial machine learning.

6 Conclusion

Combining state-of-the-art machine learning models, I analyze industry return predictability using information from 30 lagged industry returns covering a wide array of sectors across the U.S. economy. Tackling statistical challenges of overfitting when using an abundance of independent variables, the regularization method Elastic-Net combined with the gradient boosting technique XGBoostRegression are employed yielding sparse, robust, and accurate coefficient estimates. In-sample results reveal that over a period from 1959 until 2019, 190 lagged industry returns are selected to forecast another individual sector's return in the following period. Using network analysis, I find a strong positive relationship between the estimated predictor variables and the importance of those predictors in the U.S. production network as measured by their respective centrality score. Comparing the results of the economic network analysis for a benchmark year 2008 and a control year 2018 also reveals interesting patterns in terms of transformational change within the U.S. production network. Hereby, the data is pointing to structural shifts of key industries like the **Coal** and **Fin** sector in terms of their overall sector importance and their ability to predict the next month's return of several other related industries.

As part of the out-of-sample analysis, I construct five industry portfolios and monthly industry returns over a period of 50 years are computed in a walk-forward fashion testing for the exposure to common risk factors. The portfolio exhibits significant annualized alpha values of up to 18% and tends to reject a risk-based explanation of the portfolio performance. Furthermore, summary statistics for the industry portfolios based on the Post-Lasso-OLS technique used by Rapach et al. (2015) perform significantly worse in terms of mean monthly annualized return compared to the machine learning methodology used in this paper. In contrast, using the Elastic-Net-XGBRegression framework improves the risk-adjusted performance by nearly 30%.

Overall, the results further support the idea of information frictions in equity markets and the predictive power lagged-industry returns entail. Following the propagation of cash-flow shocks among interdependent industries, lagged-industry returns seem to entail asset pricing information which enables cross-industry equity price forecasting. These findings further support the idea of gradual diffusion of information and the existence of boundedly rational agents.

Reference List

A. Brav, J. Heaton (2002) Competing Theories of Financial Anomalies, Review of Financial Studies, Vol. 15 [2], 575-606

Aobdia, Daniel and Caskey, Judson and Ozel, N. Bugra, Inter-Industry Network Structure and the Cross- Predictability of Earnings and Stock Returns (January 9, 2014). Review of Accounting Studies, Vol. 19, No. 3, pp. 1191-1224, 2014. SSRN: https://ssrn.com/abstract=2196196 or http://dx.doi.org/10.2139/ssrn.2196196

Benjamini, Y., & Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. Journal of Educational and Behavioral Statistics, 25(1), 60–83. https://doi.org/10.3102/10769986025001060

Carvalho, Vasco M. 2014. "From Micro to Macro via Production Networks." Journal of Economic Perspectives, 28 (4): 23-48.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

D. Hirshleifer (2001) Investor Psychology and Asset Pricing, Journal of Finance, Vol. 56 [4], 1533-1597

E. Fama, K. French (1993) Common Risk Factors in the Returns on Stocks and Bonds, Journal of Financial Economics, Vol. 33, 3-56

E. Fama, K. French (2004) The Capital Asset Pricing Model: Theory and Evidence, Working Paper, University of Chicago (USA)

Fama, E. F. (1965). The Behavior of Stock-Market Prices. The Journal of Business, 38(1), 34–105. http://www.jstor.org/stable/2350752

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.2307/2325486

Fama, Eugene F. 2014. "Two Pillars of Asset Pricing." American Economic Review, 104 (6): 1467-85.

Ferson, Wayne E. and Korajczyk, Robert A., Do Arbitrage Pricing Models Explain the Predictability of Stock Returns?. Journal of Business, Vol. 68, No. 3, July 1995.

Friedman, J.H. (2002) Stochastic Gradient Boosting. Computational Statistics and Data Analysis, 28, 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

H. Simon (1955) A Behavioral Model of Rational Choice, Quarterly Journal of Economics, Vol. 69, 99-118

Hong, H., & Stein, J. C. (1999). A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. The Journal of Finance, 54(6), 2143–2184. http://www.jstor.org/stable/797990

Hong, H., W. Torous, and R. Valkanov (2007). Do Industries Lead Stock Markets? Journal of Financial Economics 82:2, 367-396

J. McClave, G. Benson, T. Sincich (1998) Statistics for Business and Economics, 7th Edition, Prentice Hall, London (UK)

J. von Neumann, O. Morgenstern (1947) Theory of Games and Economic Behavior, 3rd Edition (1953), Princeton University Press, Princeton (USA)

Katz, L. 1953. A new status index derived from sociometric analysis. Psychometrika 18, 39–43.

Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 2021, 23, 18. https://dx.doi.org/10.3390/e23010018

Markowitz, H. (1959). Portfolio selection: Efficient diversification of investments. New York: Wiley.

N. S. Nanayakkara & P. D. Nimal & Y. K. Weerakoon, 2019. "Behavioural Asset Pricing: A Review," International Journal of Economics and Financial Issues, Econjournals, vol. 9(4), pages 101-108.

Pástor, Ľ., & Stambaugh, R. F. (2003). Liquidity Risk and Expected Stock Returns. Journal of Political Economy, 111(3), 642–685. https://doi.org/10.1086/374184

Rapach, David E.; Strauss, Jack; Tu, Jun; and Zhou, Guofu. Industry Interdependencies and Cross-Industry Return Predictability. (2015). Research Collection Lee Kong Chian School Of Business.

S. Ross (2002) A Neoclassical Look at Behavioral Finance, The Princeton Lectures in Finance III, Massachusetts Institute of Technology, Cambridge (USA)

S. Ross, R. Westerfield, J. Jaffe (1999) Corporate Finance, 5th Edition, McGraw-Hill International, Boston (USA)

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*, 28 (2): 3-28.

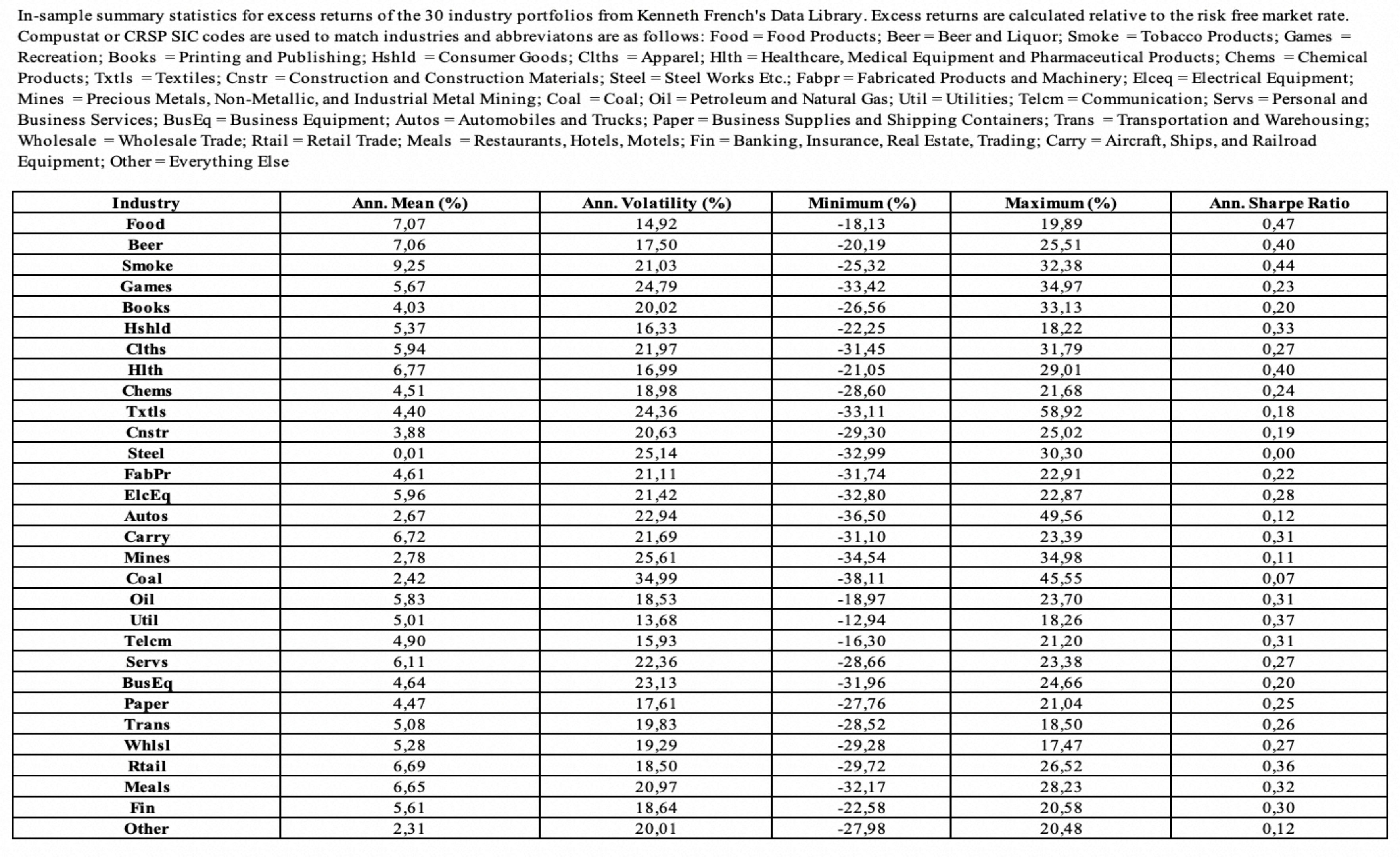
W. De Bondt, R. Thaler (1985) Does The Stock Market Overreact?, Journal of Finance, Vol. 40 [3], 793-805

W. Sharpe (1964) Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk , Journal of Finance, Vol. 19 [3], 425-442

Zhang, C.-H., & Huang, J. (2008). The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. The Annals of Statistics, 36(4), 1567–1594. http://www.jstor.org/stable/25464684

Tables and Figures

Table 1: Summary statistics, monthly industry excess returns, 1959/12 - 2018/12



In-sample summary statistics for excess returns of the 30 industry portfolios from Kenneth French's Data Library. Excess returns are calculated relative to the risk free market rate. Compustat or CRSP SIC codes are used to match industries and abbreviatons are as follows: Food = Food Products; Beer = Beer and Liquor; Smoke = Tobacco Products; Games = Recreation; Books = Printing and Publishing; Hshld = Consumer Goods; Clths = Apparel; Hlth = Healthcare, Medical Equipment and Pharmaceutical Products; Chems = Chemical Products; Txtls = Textiles; Cnstr = Construction and Construction Materials; Steel = Steel Works Etc.; Fabpr = Fabricated Products and Machinery; Elceq = Electrical Equipment; Mines = Precious Metals, Non-Metallic, and Industrial Metal Mining; Coal = Coal; Oil = Petroleum and Natural Gas; Util = Utilities; Telcm = Communication; Servs = Personal and Business Services; BusEq = Business Equipment; Autos = Automobiles and Trucks; Paper = Business Supplies and Shipping Containers; Trans = Transportation and Warehousing; Wholesale = Wholesale Trade; Rtail = Retail Trade; Meals = Restaurants, Hotels, Motels; Fin = Banking, Insurance, Real Estate, Trading; Carry = Aircraft, Ships, and Railroad Equipment; Other = Everything Else

Industry	Ann. Mean (%)	Ann. Volatility (%)	Minimum (%)	Maximum (%)	Ann. Sharpe Ratio
Food	7,07	14,92	-18,13	19,89	0,47
Beer	7,06	17,50	-20,19	25,51	0,40
Smoke	9,25	21,03	-25,32	32,38	0,44
Games	5,67	24,79	-33,42	34,97	0,23
Books	4,03	20,02	-26,56	33,13	0,20
Hshld	5,37	16,33	-22,25	18,22	0,33
Clths	5,94	21,97	-31,45	31,79	0,27
Hlth	6,77	16,99	-21,05	29,01	0,40
Chems	4,51	18,98	-28,60	21,68	0,24
Txtls	4,40	24,36	-33,11	58,92	0,18
Cnstr	3,88	20,63	-29,30	25,02	0,19
Steel	0,01	25,14	-32,99	30,30	0,00
FabPr	4,61	21,11	-31,74	22,91	0,22
ElcEq	5,96	21,42	-32,80	22,87	0,28
Autos	2,67	22,94	-36,50	49,56	0,12
Carry	6,72	21,69	-31,10	23,39	0,31
Mines	2,78	25,61	-34,54	34,98	0,11
Coal	2,42	34,99	-38,11	45,55	0,07
Oil	5,83	18,53	-18,97	23,70	0,31
Util	5,01	13,68	-12,94	18,26	0,37
Telcm	4,90	15,93	-16,30	21,20	0,31
Servs	6,11	22,36	-28,66	23,38	0,27
BusEq	4,64	23,13	-31,96	24,66	0,20
Paper	4,47	17,61	-27,76	21,04	0,25
Trans	5,08	19,83	-28,52	18,50	0,26
Whlsl	5,28	19,29	-29,28	17,47	0,27
Rtail	6,69	18,50	-29,72	26,52	0,36
Meals	6,65	20,97	-32,17	28,23	0,32
Fin	5,61	18,64	-22,58	20,58	0,30
Other	2,31	20,01	-27,98	20,48	0,12

Table 2.1: Results of the Post-Elastic-Net-XGBoost regression, monthly industry portfolio excess returns, 1960/01 - 2018/12

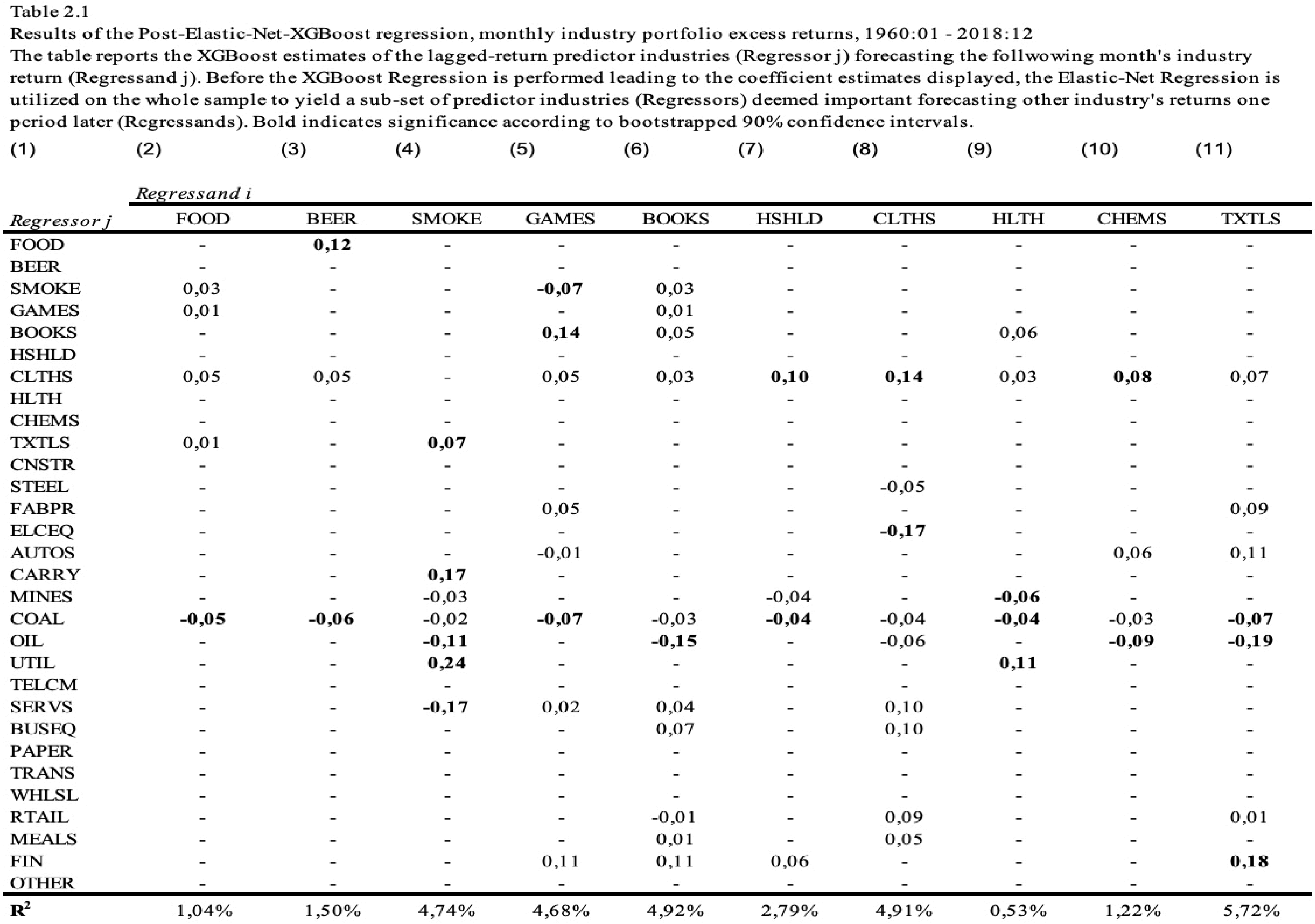


Table 2.1

Results of the Post-Elastic-Net-XGBoost regression, monthly industry portfolio excess returns, 1960:01 - 2018:12

The table reports the XGBoost estimates of the lagged-return predictor industries (Regressor j) forecasting the follwowing month's industry return (Regressand j). Before the XGBoost Regression is performed leading to the coefficient estimates displayed, the Elastic-Net Regression is utilized on the whole sample to yield a sub-set of predictor industries (Regressors) deemed important forecasting other industry's returns one period later (Regressands). Bold indicates significance according to bootstrapped 90% confidence intervals.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Regressand i									
Regressor j	FOOD	BEER	SMOKE	GAMES	BOOKS	HSHLD	CLTHS	HLTH	CHEMS	TXTLS
FOOD	-	**0,12**	-	-	-	-	-	-	-	-
BEER	-	-	-	-	-	-	-	-	-	-
SMOKE	0,03	-	-	**-0,07**	0,03	-	-	-	-	-
GAMES	0,01	-	-	-	0,01	-	-	-	-	-
BOOKS	-	-	-	**0,14**	0,05	-	-	0,06	-	-
HSHLD	-	-	-	-	-	-	-	-	-	-
CLTHS	0,05	0,05	-	0,05	0,03	**0,10**	**0,14**	0,03	**0,08**	0,07
HLTH	-	-	-	-	-	-	-	-	-	-
CHEMS	-	-	-	-	-	-	-	-	-	-
TXTLS	0,01	-	**0,07**	-	-	-	-	-	-	-
CNSTR	-	-	-	-	-	-	-	-	-	-
STEEL	-	-	-	-	-	-	-0,05	-	-	-
FABPR	-	-	-	0,05	-	-	-	-	-	0,09
ELCEQ	-	-	-	-	-	-	**-0,17**	-	-	-
AUTOS	-	-	-	-0,01	-	-	-	-	0,06	0,11
CARRY	-	-	**0,17**	-	-	-	-	-	-	-
MINES	-	-	-0,03	-	-	-0,04	-	**-0,06**	-	-
COAL	**-0,05**	**-0,06**	-0,02	**-0,07**	-0,03	**-0,04**	-0,04	**-0,04**	-0,03	**-0,07**
OIL	-	-	**-0,11**	-	**-0,15**	-	-0,06	-	**-0,09**	**-0,19**
UTIL	-	-	**0,24**	-	-	-	-	**0,11**	-	-
TELCM	-	-	-	-	-	-	-	-	-	-
SERVS	-	-	**-0,17**	0,02	0,04	-	0,10	-	-	-
BUSEQ	-	-	-	-	0,07	-	0,10	-	-	-
PAPER	-	-	-	-	-	-	-	-	-	-
TRANS	-	-	-	-	-	-	-	-	-	-
WHLSL	-	-	-	-	-	-	-	-	-	-
RTAIL	-	-	-	-	-0,01	-	0,09	-	-	0,01
MEALS	-	-	-	-	0,01	-	0,05	-	-	-
FIN	-	-	-	0,11	0,11	0,06	-	-	-	**0,18**
OTHER	-	-	-	-	-	-	-	-	-	-
R^2	1,04%	1,50%	4,74%	4,68%	4,92%	2,79%	4,91%	0,53%	1,22%	5,72%

Table 2.2: Results of the Post-Elastic-Net-XGBoost regression, monthly industry portfolio excess returns, 1960/01 - 2018/12

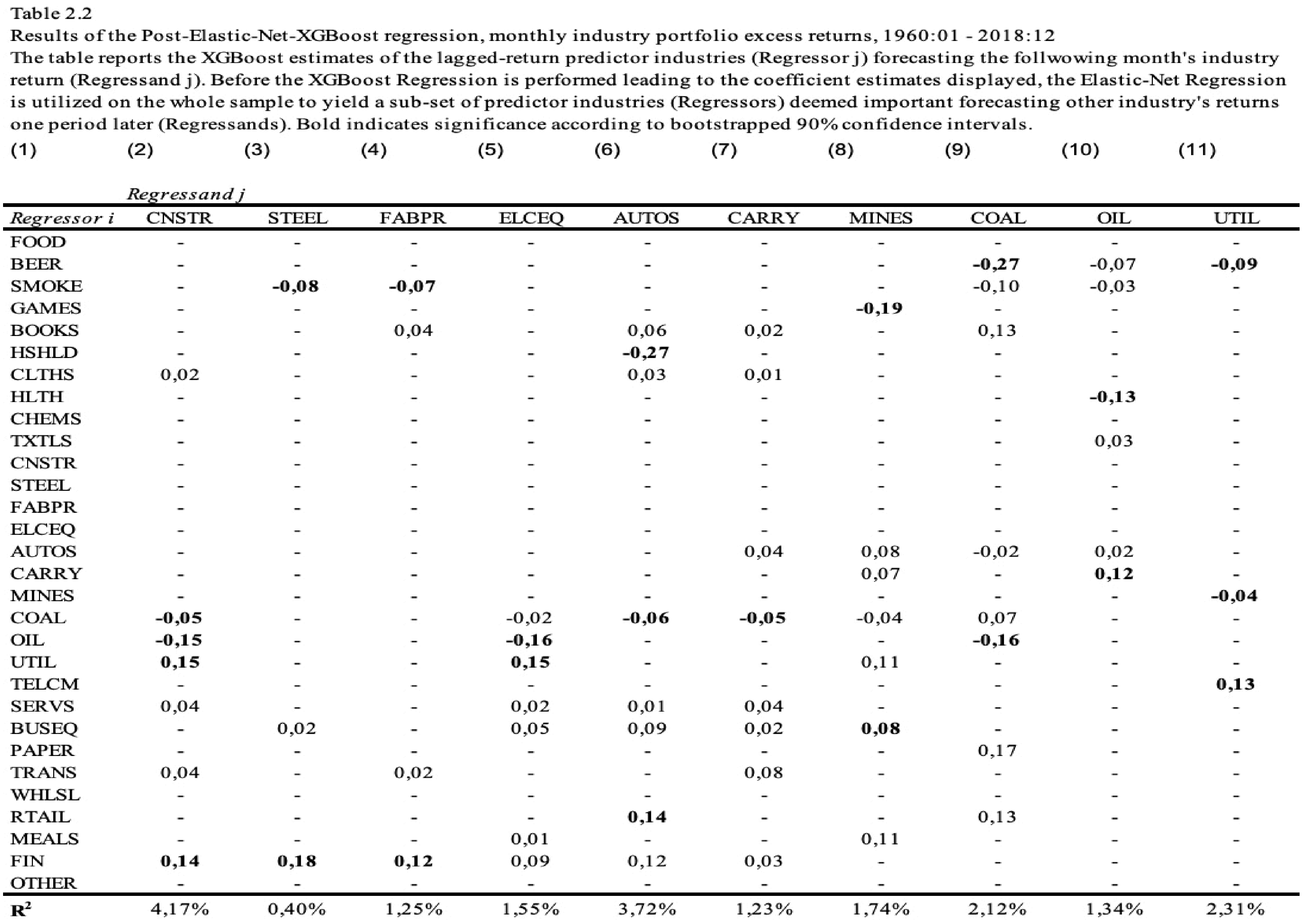


Table 2.2

Results of the Post-Elastic-Net-XGBoost regression, monthly industry portfolio excess returns, 1960:01 - 2018:12

The table reports the XGBoost estimates of the lagged-return predictor industries (Regressor j) forecasting the follwowing month's industry return (Regressand j). Before the XGBoost Regression is performed leading to the coefficient estimates displayed, the Elastic-Net Regression is utilized on the whole sample to yield a sub-set of predictor industries (Regressors) deemed important forecasting other industry's returns one period later (Regressands). Bold indicates significance according to bootstrapped 90% confidence intervals.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Regressand j									
Regressor i	CNSTR	STEEL	FABPR	ELCEQ	AUTOS	CARRY	MINES	COAL	OIL	UTIL
FOOD	-	-	-	-	-	-	-	-	-	-
BEER	-	-	-	-	-	-	-	**-0,27**	-0,07	**-0,09**
SMOKE	-	**-0,08**	**-0,07**	-	-	-	-	-0,10	-0,03	-
GAMES	-	-	-	-	-	-	**-0,19**	-	-	-
BOOKS	-	-	0,04	-	0,06	0,02	-	0,13	-	-
HSHLD	-	-	-	-	**-0,27**	-	-	-	-	-
CLTHS	0,02	-	-	-	0,03	0,01	-	-	-	-
HLTH	-	-	-	-	-	-	-	-	**-0,13**	-
CHEMS	-	-	-	-	-	-	-	-	-	-
TXTLS	-	-	-	-	-	-	-	-	0,03	-
CNSTR	-	-	-	-	-	-	-	-	-	-
STEEL	-	-	-	-	-	-	-	-	-	-
FABPR	-	-	-	-	-	-	-	-	-	-
ELCEQ	-	-	-	-	-	-	-	-	-	-
AUTOS	-	-	-	-	-	0,04	0,08	-0,02	0,02	-
CARRY	-	-	-	-	-	-	0,07	-	**0,12**	-
MINES	-	-	-	-	-	-	-	-	-	**-0,04**
COAL	**-0,05**	-	-	-0,02	**-0,06**	**-0,05**	-0,04	0,07	-	-
OIL	**-0,15**	-	-	**-0,16**	-	-	-	**-0,16**	-	-
UTIL	**0,15**	-	-	**0,15**	-	-	0,11	-	-	-
TELCM	-	-	-	-	-	-	-	-	-	**0,13**
SERVS	0,04	-	-	0,02	0,01	0,04	-	-	-	-
BUSEQ	-	0,02	-	0,05	0,09	0,02	**0,08**	-	-	-
PAPER	-	-	-	-	-	-	-	0,17	-	-
TRANS	0,04	-	0,02	-	-	0,08	-	-	-	-
WHLSL	-	-	-	-	-	-	-	-	-	-
RTAIL	-	-	-	-	**0,14**	-	-	0,13	-	-
MEALS	-	-	-	0,01	-	-	0,11	-	-	-
FIN	**0,14**	**0,18**	**0,12**	0,09	0,12	0,03	-	-	-	-
OTHER	-	-	-	-	-	-	-	-	-	-
R^2	4,17%	0,40%	1,25%	1,55%	3,72%	1,23%	1,74%	2,12%	1,34%	2,31%

Table 2.3: Results of the Post-Elastic-Net-XGBoost regression, monthly industry portfolio excess returns, 1960/01 - 2018/12

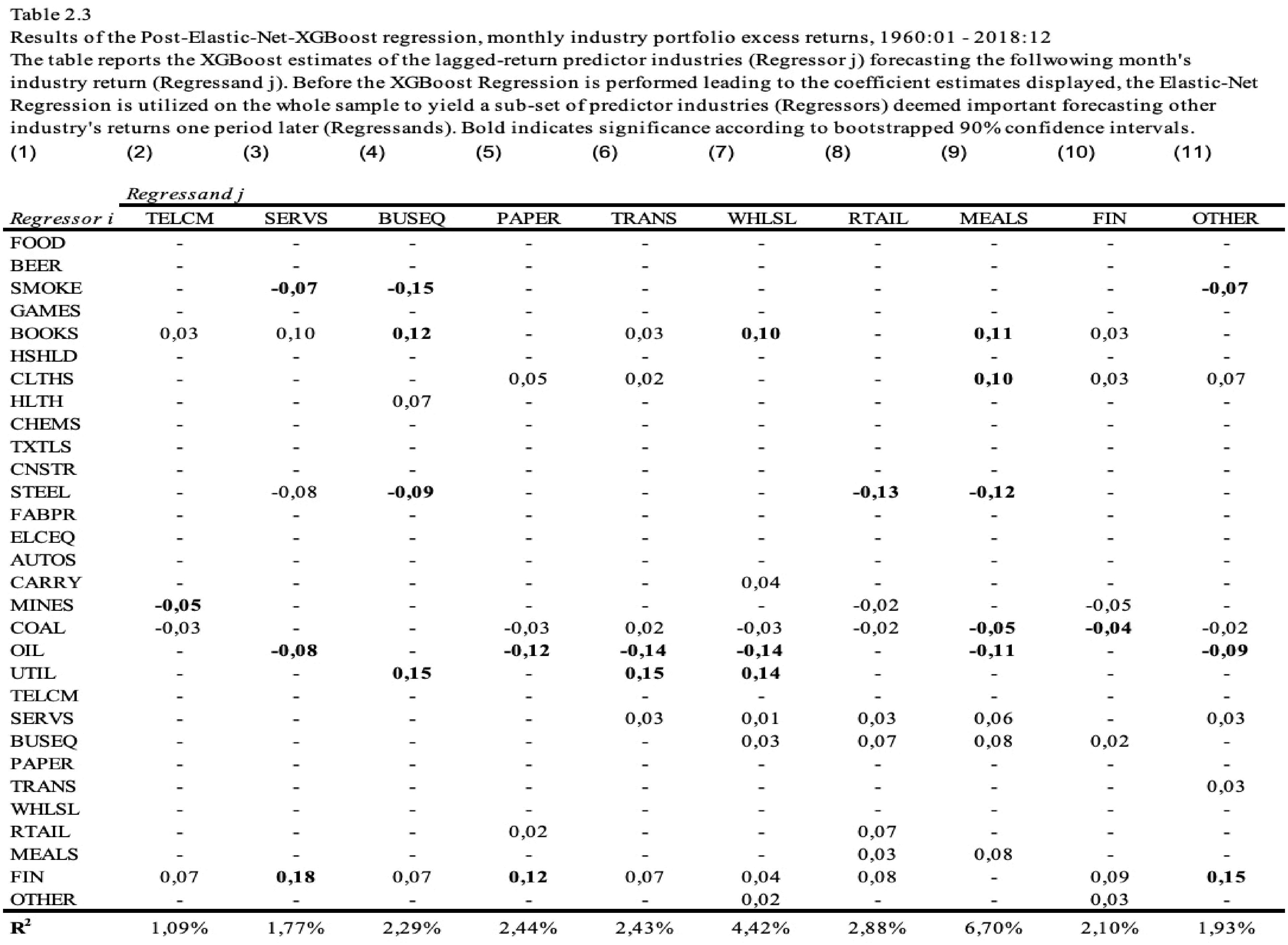


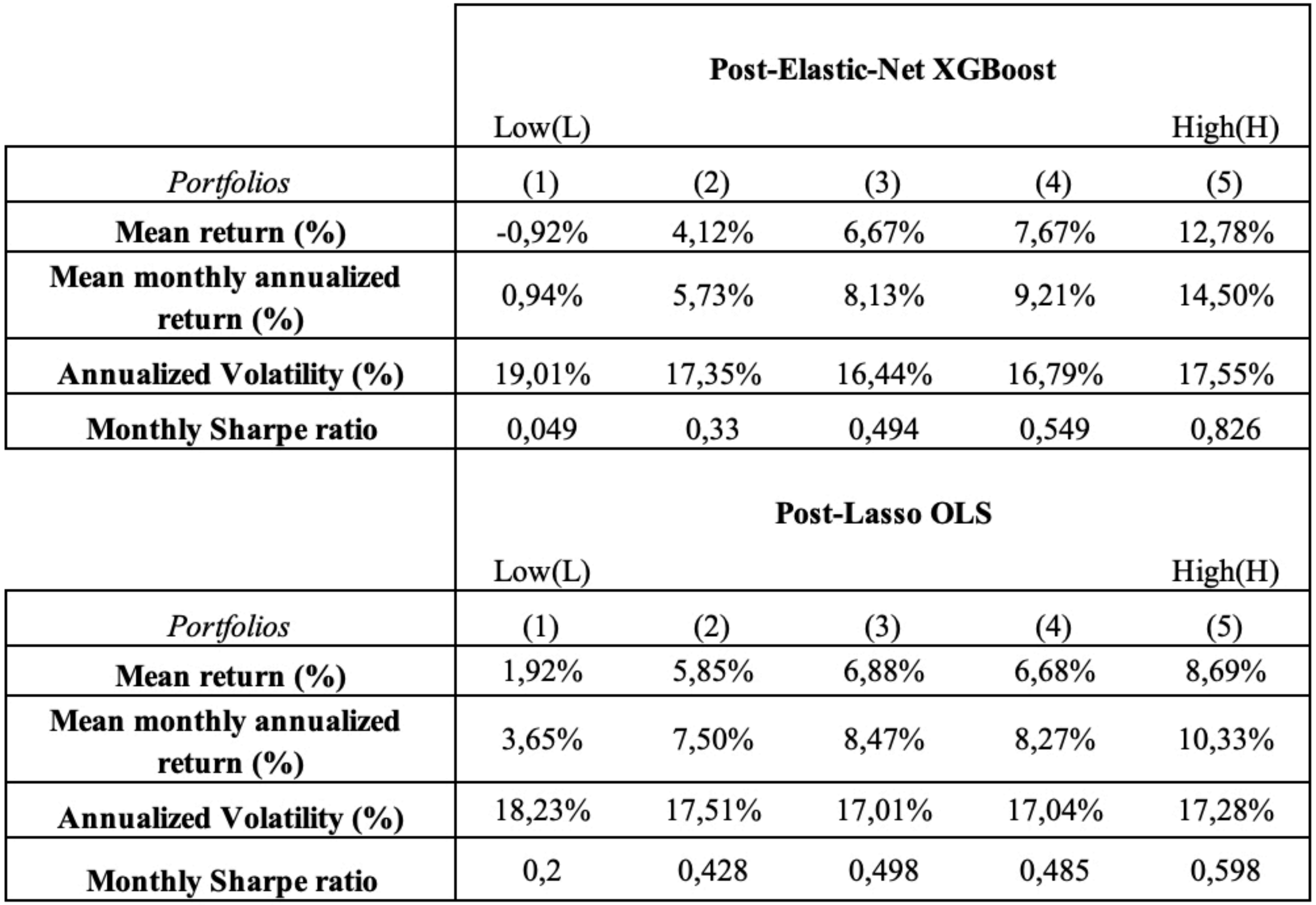
Table 2.3

Results of the Post-Elastic-Net-XGBoost regression, monthly industry portfolio excess returns, 1960:01 - 2018:12

The table reports the XGBoost estimates of the lagged-return predictor industries (Regressor j) forecasting the follwowing month's industry return (Regressand j). Before the XGBoost Regression is performed leading to the coefficient estimates displayed, the Elastic-Net Regression is utilized on the whole sample to yield a sub-set of predictor industries (Regressors) deemed important forecasting other industry's returns one period later (Regressands). Bold indicates significance according to bootstrapped 90% confidence intervals.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Regressand j									
Regressor i	TELCM	SERVS	BUSEQ	PAPER	TRANS	WHLSL	RTAIL	MEALS	FIN	OTHER
FOOD	-	-	-	-	-	-	-	-	-	-
BEER	-	-	-	-	-	-	-	-	-	-
SMOKE	-	**-0,07**	**-0,15**	-	-	-	-	-	-	**-0,07**
GAMES	-	-	-	-	-	-	-	-	-	-
BOOKS	0,03	0,10	**0,12**	-	0,03	**0,10**	-	**0,11**	0,03	-
HSHLD	-	-	-	-	-	-	-	-	-	-
CLTHS	-	-	-	0,05	0,02	-	-	**0,10**	0,03	0,07
HLTH	-	-	0,07	-	-	-	-	-	-	-
CHEMS	-	-	-	-	-	-	-	-	-	-
TXTLS	-	-	-	-	-	-	-	-	-	-
CNSTR	-	-	-	-	-	-	-	-	-	-
STEEL	-	-0,08	**-0,09**	-	-	-	**-0,13**	**-0,12**	-	-
FABPR	-	-	-	-	-	-	-	-	-	-
ELCEQ	-	-	-	-	-	-	-	-	-	-
AUTOS	-	-	-	-	-	-	-	-	-	-
CARRY	-	-	-	-	-	0,04	-	-	-	-
MINES	**-0,05**	-	-	-	-	-	-0,02	-	-0,05	-
COAL	-0,03	-	-	-0,03	0,02	-0,03	-0,02	**-0,05**	**-0,04**	-0,02
OIL	-	**-0,08**	-	**-0,12**	**-0,14**	**-0,14**	-	**-0,11**	-	**-0,09**
UTIL	-	-	**0,15**	-	**0,15**	**0,14**	-	-	-	-
TELCM	-	-	-	-	-	-	-	-	-	-
SERVS	-	-	-	-	0,03	0,01	0,03	0,06	-	0,03
BUSEQ	-	-	-	-	-	0,03	0,07	0,08	0,02	-
PAPER	-	-	-	-	-	-	-	-	-	-
TRANS	-	-	-	-	-	-	-	-	-	0,03
WHLSL	-	-	-	-	-	-	-	-	-	-
RTAIL	-	-	-	0,02	-	-	0,07	-	-	-
MEALS	-	-	-	-	-	-	0,03	0,08	-	-
FIN	0,07	**0,18**	0,07	**0,12**	0,07	0,04	0,08	-	0,09	**0,15**
OTHER	-	-	-	-	-	0,02	-	-	0,03	-
R^2	1,09%	1,77%	2,29%	2,44%	2,43%	4,42%	2,88%	6,70%	2,10%	1,93%

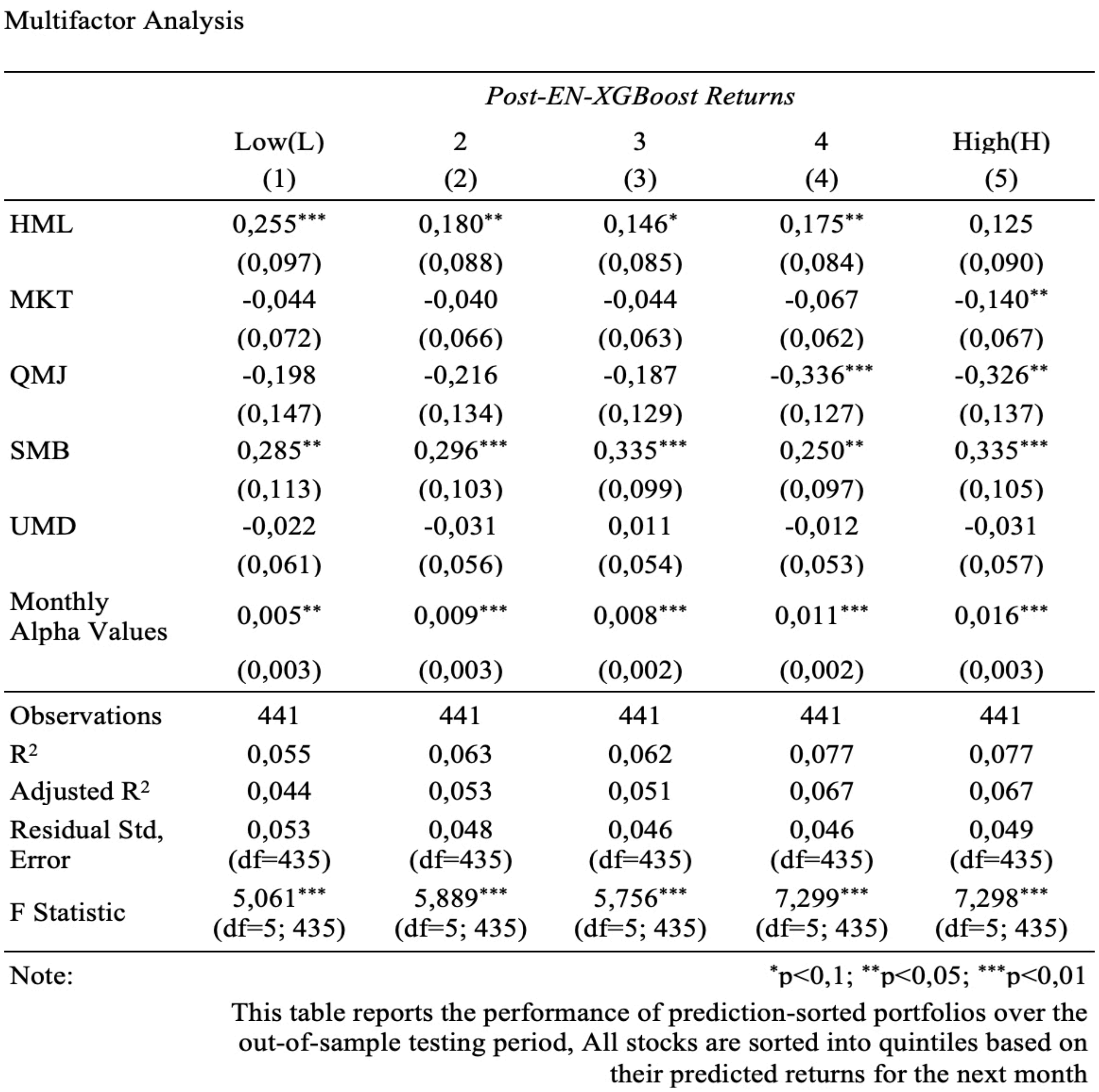
Table 3: Out-of-Sample Industry Portfolio Summary Statistics, 1970/01 - 2018/12



	Post-Elastic-Net XGBoost				
	Low(L)				High(H)
Portfolios	(1)	(2)	(3)	(4)	(5)
Mean return (%)	-0,92%	4,12%	6,67%	7,67%	12,78%
Mean monthly annualized return (%)	0,94%	5,73%	8,13%	9,21%	14,50%
Annualized Volatility (%)	19,01%	17,35%	16,44%	16,79%	17,55%
Monthly Sharpe ratio	0,049	0,33	0,494	0,549	0,826

	Post-Lasso OLS				
	Low(L)				High(H)
Portfolios	(1)	(2)	(3)	(4)	(5)
Mean return (%)	1,92%	5,85%	6,88%	6,68%	8,69%
Mean monthly annualized return (%)	3,65%	7,50%	8,47%	8,27%	10,33%
Annualized Volatility (%)	18,23%	17,51%	17,01%	17,04%	17,28%
Monthly Sharpe ratio	0,2	0,428	0,498	0,485	0,598

Table 4.1: Multifactor Analysis



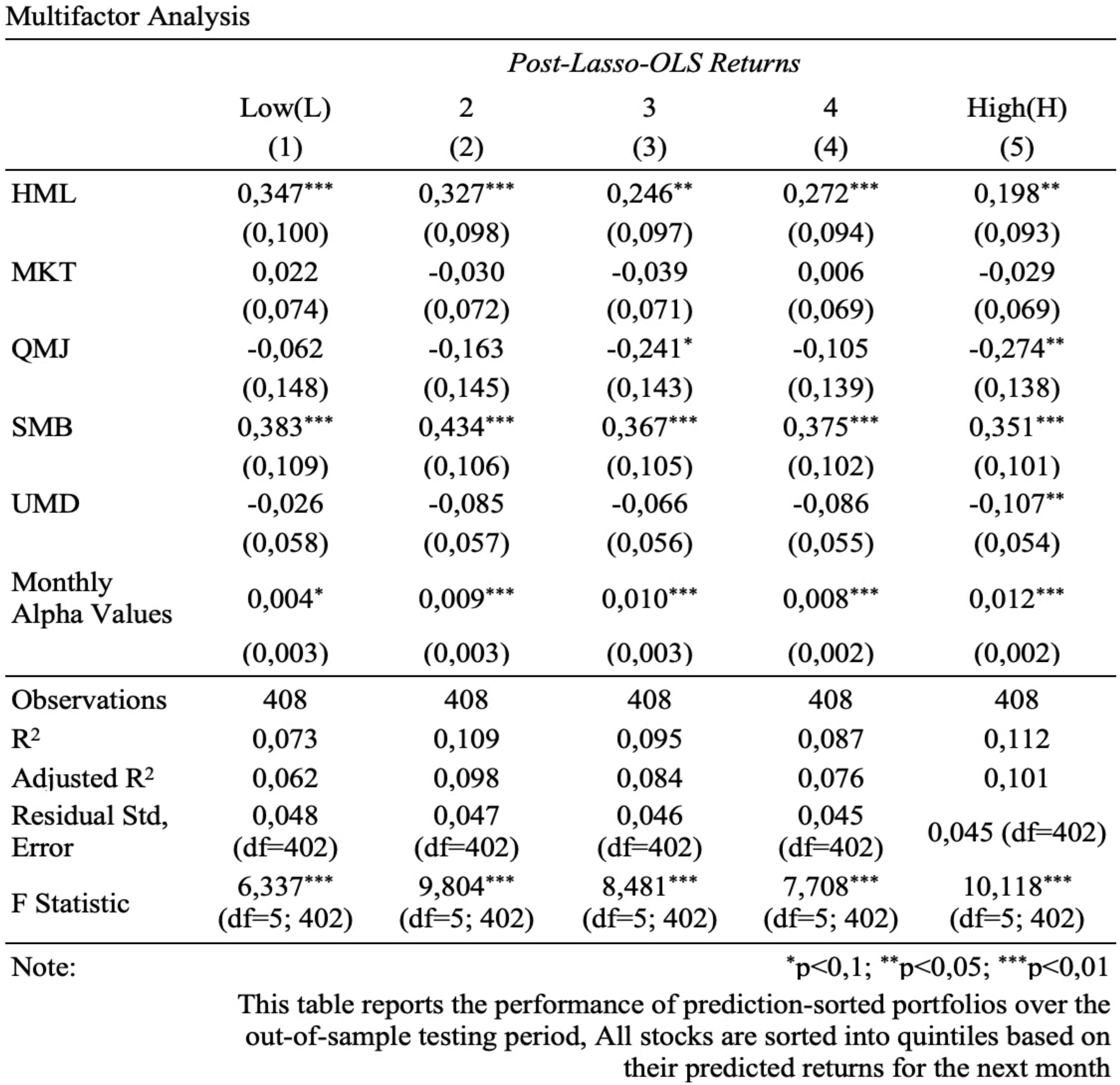
Multifactor Analysis

	Post-EN-XGBoost Returns				
	Low(L)	2	3	4	High(H)
	(1)	(2)	(3)	(4)	(5)
HML	0,255***	0,180**	0,146*	0,175**	0,125
	(0,097)	(0,088)	(0,085)	(0,084)	(0,090)
MKT	-0,044	-0,040	-0,044	-0,067	-0,140**
	(0,072)	(0,066)	(0,063)	(0,062)	(0,067)
QMJ	-0,198	-0,216	-0,187	-0,336***	-0,326**
	(0,147)	(0,134)	(0,129)	(0,127)	(0,137)
SMB	0,285**	0,296***	0,335***	0,250**	0,335***
	(0,113)	(0,103)	(0,099)	(0,097)	(0,105)
UMD	-0,022	-0,031	0,011	-0,012	-0,031
	(0,061)	(0,056)	(0,054)	(0,053)	(0,057)
Monthly Alpha Values	0,005**	0,009***	0,008***	0,011***	0,016***
	(0,003)	(0,003)	(0,002)	(0,002)	(0,003)
Observations	441	441	441	441	441
R^2	0,055	0,063	0,062	0,077	0,077
Adjusted R^2	0,044	0,053	0,051	0,067	0,067
Residual Std, Error	0,053 (df=435)	0,048 (df=435)	0,046 (df=435)	0,046 (df=435)	0,049 (df=435)
F Statistic	5,061*** (df=5; 435)	5,889*** (df=5; 435)	5,756*** (df=5; 435)	7,299*** (df=5; 435)	7,298*** (df=5; 435)

Note: *p<0,1; **p<0,05; ***p<0,01

This table reports the performance of prediction-sorted portfolios over the out-of-sample testing period, All stocks are sorted into quintiles based on their predicted returns for the next month

Table 4.2: Multifactor Analysis



Multifactor Analysis

	Post-Lasso-OLS Returns				
	Low(L)	2	3	4	High(H)
	(1)	(2)	(3)	(4)	(5)
HML	0,347***	0,327***	0,246**	0,272***	0,198**
	(0,100)	(0,098)	(0,097)	(0,094)	(0,093)
MKT	0,022	-0,030	-0,039	0,006	-0,029
	(0,074)	(0,072)	(0,071)	(0,069)	(0,069)
QMJ	-0,062	-0,163	-0,241*	-0,105	-0,274**
	(0,148)	(0,145)	(0,143)	(0,139)	(0,138)
SMB	0,383***	0,434***	0,367***	0,375***	0,351***
	(0,109)	(0,106)	(0,105)	(0,102)	(0,101)
UMD	-0,026	-0,085	-0,066	-0,086	-0,107**
	(0,058)	(0,057)	(0,056)	(0,055)	(0,054)
Monthly Alpha Values	0,004*	0,009***	0,010***	0,008***	0,012***
	(0,003)	(0,003)	(0,003)	(0,002)	(0,002)
Observations	408	408	408	408	408
R^2	0,073	0,109	0,095	0,087	0,112
Adjusted R^2	0,062	0,098	0,084	0,076	0,101
Residual Std, Error	0,048 (df=402)	0,047 (df=402)	0,046 (df=402)	0,045 (df=402)	0,045 (df=402)
F Statistic	6,337*** (df=5; 402)	9,804*** (df=5; 402)	8,481*** (df=5; 402)	7,708*** (df=5; 402)	10,118*** (df=5; 402)

Note: *p<0,1; **p<0,05; ***p<0,01

This table reports the performance of prediction-sorted portfolios over the out-of-sample testing period, All stocks are sorted into quintiles based on their predicted returns for the next month

Figure 1: Centrality Scatterplot (2008)

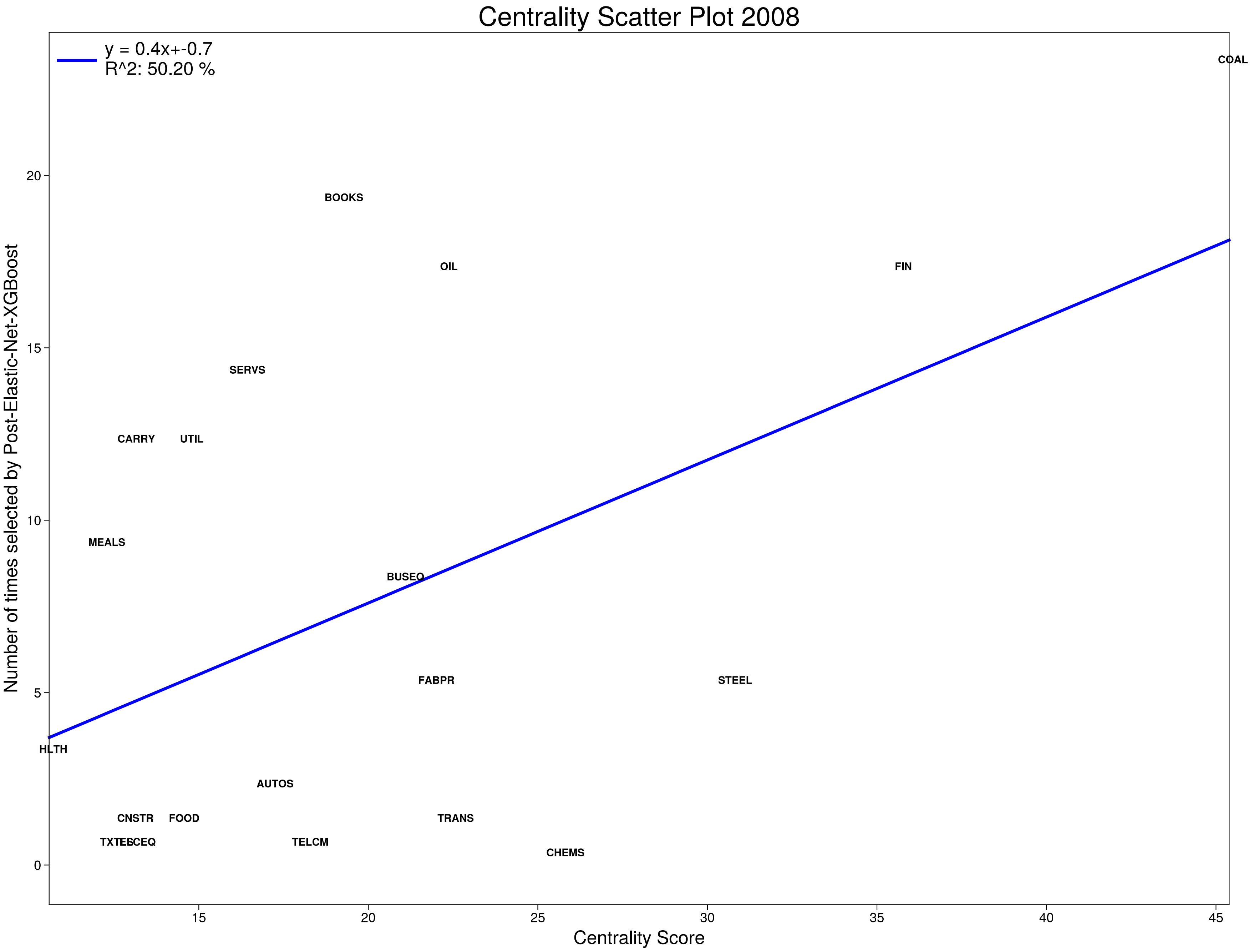


Figure 2: Centrality Scatterplot (2018)

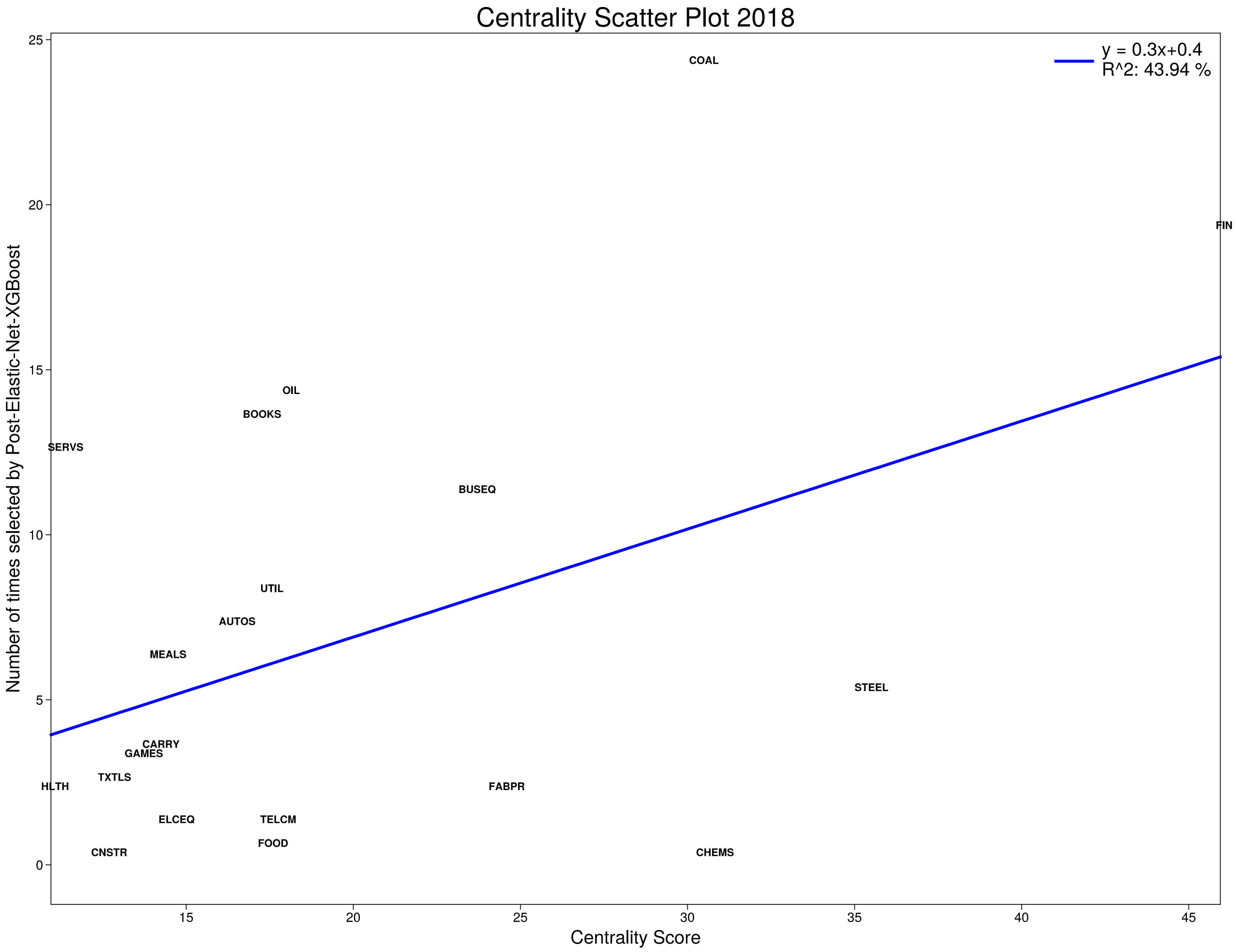
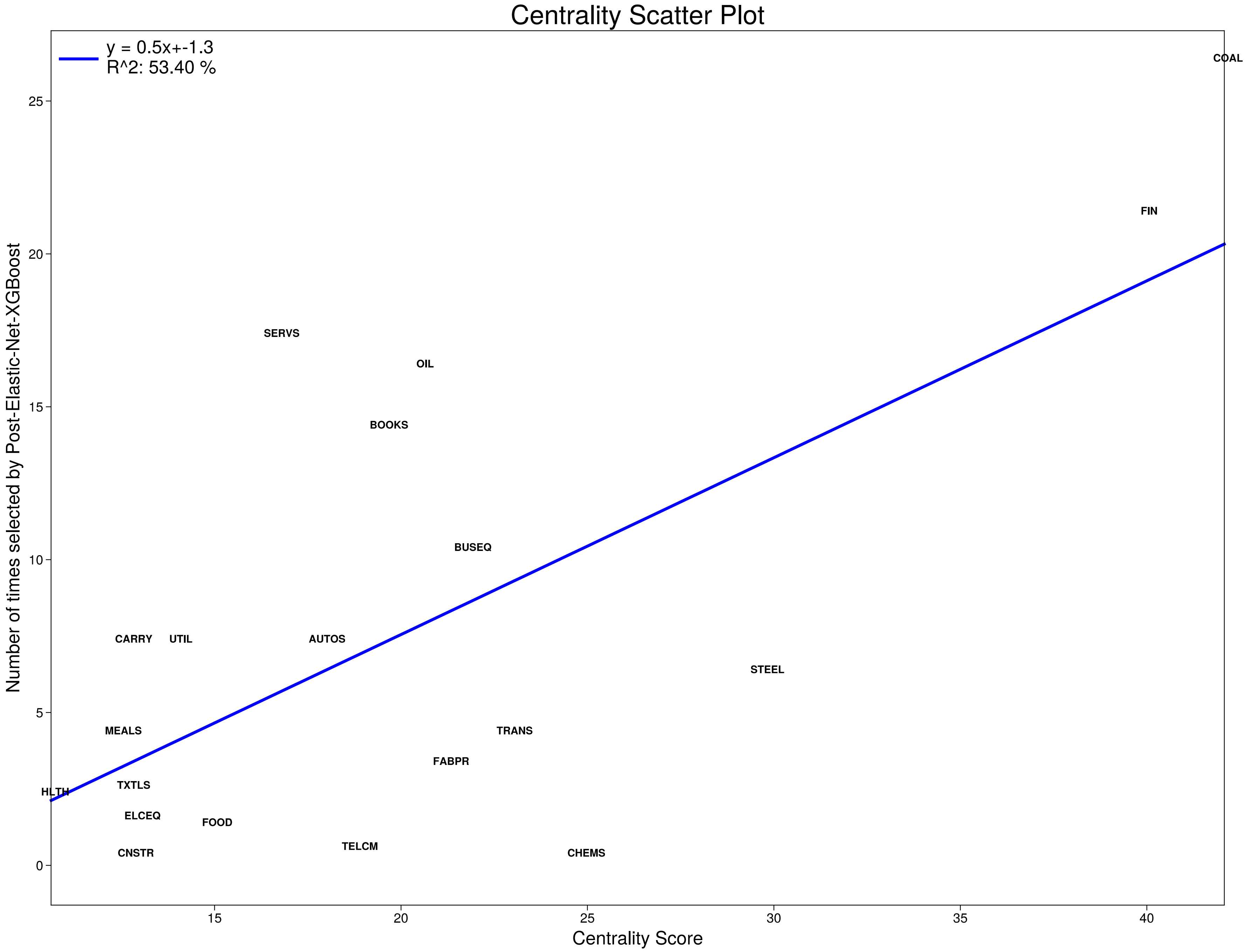


Figure 3: Centrality Scatterplot Replication


Centrality Scatter Plot
y = 0.5x+-1.3
R^2: 53.40 %
COAL
FIN
SERVS
OIL
BOOKS
BUSEQ
CARRY
UTIL
AUTOS
STEEL
MEALS
TRANS
FABPR
TXTLS
HLTH
ELCEQ
FOOD
TELCM
CNSTR
CHEMS
Number of times selected by Post-Elastic-Net-XGBoost
Centrality Score