Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

Adrian Maximilian Gjerstad Hatletvedt
Petter Solberg Svingen

Predicting financial markets with Google search categories

Master's thesis in Industrial Economics and Technology Management
Supervisor: Peter Molnár

June 2021

NTNU
Norwegian University of
Science and Technology

Adrian Maximilian Gjerstad Hatletvedt
Petter Solberg Svingen

Predicting financial markets with Google search categories

Master's thesis in Industrial Economics and Technology Management
Supervisor: Peter Molnár
June 2021

Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

NTNU
Norwegian University of
Science and Technology

Preface

This Master's thesis examines the usefulness of internet search data for predicting financial markets. The thesis concludes our Master of Science degree in Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU) in the spring of 2021. We would like to thank our supervisor, Peter Molnár, for exceptional guidance and feedback. His input has been invaluable, and we are grateful for the flexibility and presence he has shown.

Trondheim, June 2021

Abstract

In this paper we analyse the usefulness of Google search data for predicting the returns of industry stock indices in five countries, the US, Japan, Great Britain, France and India. Since we use a large set of Google Trends categories as predictors, the forecasting models need to be able to deal with high-dimensional data. We therefore consider Principal Component Regression, Principal Component Random Forest Regression, Ridge Regression, Lasso, and Elastic Net as our prediction models. The best performing model in our calibration sample (the first part of the US data) is the Elastic Net. A simple long-short strategy based on the Elastic Net model significantly outperforms the stock market in all five countries, after including transaction costs. Furthermore, we find that the model achieves most of its excess returns during weeks where the overall stock market drops. We also find that over time, the relationship between a search category and industry returns can change both in magnitude and direction. Our model automatically accounts for this, since it is re-fitted every week. Lastly, we find that the abnormal returns of our model are only weakly correlated across countries, suggesting that our trading approach can be most beneficial when applied internationally.

Sammendrag

I denne oppgaven bruker vi Google søkevolum til å predikere den ukentlige avkastningen til industri-indekser i fem land, USA, Japan, Storbritannia, Frankrike og India. Fordi vi bruker et stort antall Google søkekategorier som forklaringsvariabler er alle prediksjonsmodellene valgt for å kunne håndtere høydimensjonalitetsdata. Vi benytter Principal Component Regression, Principal Component Random Forest Regression, Ridge Regression, Lasso, og Elastic Net som våre prediksjonsmodeller. Den beste modellen på kalibreringsdataen (første del av dataen for USA) ble Elastic Net modellen. En enkel long/short strategi basert på denne Elastic Net modellen oppnår bedre resultater enn den utvalgte referanseindeksen i alle de fem landene, etter å ha trukket fra transaksjonskostnader. Vi finner at modellen oppnår best resultater sammenlignet med referanseindeksen i nedgangstider. I tillegg finner vi at relasjonen mellom søkekategori og industriavkastning endres over tid, både i retning og i størrelse. Modellen vi har laget tar automatisk høyde for dette, ettersom den rekalibreres ukentlig. Avslutningsvis finner vi at den abnormale avkasntningen til modellen kun er svakt korrelert på kryss av landegrenser, noe som tyder på at vår metode er ekstra godt egnet i en multinasjonal porteføljestrategi.

Contents

1 Introduction

It is widely accepted today that stock markets are influenced by investor sentiment. This idea debuted in mainstream economic theory as early as Keynes (1936). Since then, several researchers have explored how investor sentiment affects stock markets (De Long et al. 1990, Brown & Cliff 2004, Joseph et al. 2010, Dimpfl & Jank 2015, Tantaopas et al. 2016). Baker & Wurgler (2006) developed a composite sentiment index based on six sentiment proxies, which they found to successfully predict the cross-section of stock returns. Specifically, they found that stocks whose valuation are highly subjective, for example small stocks, extreme growth stocks and distressed stocks, are strongly influenced by changes in investor sentiment. These findings were later supported by Baele et al. (2010), Stambaugh et al. (2012), Sibley et al. (2016) and Chen et al. (2019). Vozlyublennaia (2014) looked at the interaction between stock returns and investor attention and found that not only did investor attention affect future prices, but that past performance can also drive investor attention in a feedback loop. Hirshleifer & Shumway (2003) even suggest that stock returns are higher on sunny days, a clear indication that the mood of investors influences markets.

Nonetheless, detecting changes in investor sentiment is not an easy task. Several attempts have been made to indirectly measure sentiment changes, including headlines and news (Barber & Odean 2008, Yuan 2012), advertising expense (Grullon et al. 2003, Lou 2014, Chemmanur & Yan 2019), extreme returns (Barber & Odean 2008), price limits (Seasholes & Wu 2007) and trading volume (Gervais et al. 2001, Barber & Odean 2008, Hou et al. 2009). This millennium, the internet era has opened up a new world for researchers in the search for accurate proxies of investor sentiment. This includes website traffic (Rajgopal et al. 2000, Lazer et al. 2001), social media activity (Broadstock & Zhang 2019, Sul et al. 2017, Duz Tan & Tas 2020) and traffic on online community forums (Dondio 2013, Hu et al. 2021). Another proxy candidate that has attracted a lot of attention is internet search activity. Google is the giant within internet search, accounting for as much as 90% of all search activity worldwide (Nadler & Cicilline 2020). In 2006, this data was made available to the public through the Google Trends (GT) service. Since then, it has been successfully used in several fields of research, for instance to predict unemployment (Choi & Varian 2009, Barreira et al. 2013), gasoline prices (Molnar & Basta 2017), private consumption (Vosen & Schmidt 2011), tourism, automotive and home sales (Choi & Varian 2012) and detection of seasonal flu outbreaks (Ginsberg et al. 2009).

Many have also tried to make use of internet searches in financial applications. Several studies have used GT to successfully forecast stock volatility (Vlastakis

& Markellos 2012, Goddard et al. 2015, Bijl et al. 2016). Additionally, Preis et al. (2010) find evidence that weekly search volumes for S&P 500 company names predict trading volume of the same company's stocks. These results have been replicated in the Norwegian (Kim et al. 2019), German (Bank et al. 2011, Fink & Johann 2012) and French (Aouadi et al. 2013) market. Perhaps unsurprisingly, using GT to forecast stock returns has proven to be a harder task. Kim et al. (2019) find no significant contemporaneous or next week relationship between company stock returns and google searches for the corresponding companies. Bijl et al. (2016) developed a trading strategy using Google searches but found it to be profitable only before transaction costs. Challet & Bel Hadj Ayed (2013) challenge a study that constructed a profitable trading strategy with single GT search terms by replicating the results with a set of equally many completely arbitrary search terms.

However, there are studies that have successfully used GT data to forecast stock markets. Gwilym et al. (2014) found that higher search activity for the search term 'Penny stock' predicts lower returns of certain US stock indices, and implement a simple trading strategy that generates significant excess returns over a buy-and-hold approach. Dzielinski (2012) found that abnormal search volumes for the search term 'Economy' predict a decrease in S&P 500 returns within one week and a subsequent reversal the week after. Preis et al. (2013) looks at a number of financially-related search terms, and find that many of them work well as buy/sell signals in a simple trading strategy of the DJIA. More recently, Hu et al. (2018) use the search terms 'S&P 500' and 'DJIA' to successfully improve the accuracy of a sophisticated neural net model that predicts the opening direction of the S&P 500 and DJIA indices.

The aforementioned studies all consider single search terms only. Other researchers have taken this a step further by including multiple search terms in their models. Jiang (2016) uses the set of search terms from Preis et al. (2013) in a Lasso model to successfully predict weekly returns of three US stock indices. Similar to the findings of Bijl et al. (2016), they find that for most of the significant search terms, increased search activity is followed by lower future returns. Curme et al. (2014) create multiple single keyword strategies and group the GT predictors into different semantic categories. They find that the strategies using finance- or politics-related search terms significantly outperformed the buy-and-hold portfolio. Lyocsa & Molnar (2020) use the average abnormal search volume for a set of COVID-19 related search terms as a measure of attention to the coronavirus. They use this measure as a transition variable in a regime switching model to successfully predict US stock market returns during turbulent market conditions. Another notable multiple-keyword approach was developed by Da et al. (2014), who construct an index from financially-related search terms that negatively correlate with future returns. The index is updated every six months with the most significant search terms to stay up to date with changing markets. They found that the index predicts short-term return reversals for a set of US stock indices.

This paper investigates the relationship between GT *categories* and the returns of industry stock indices. This is in contrast to the vast majority of preceding GT work, where regular *search terms* are used. *Categories* can be used in two ways. The first way is to use a category as an attachment to a search term, for instance to distinguish between 'apple' the fruit and 'apple' the technology company. Alternatively, categories can be used not as an attachment to a specific search term, but simply as an aggregated measure that includes all search terms for a particular topic. The *fruit* category would here include all searches for apples, oranges, pears etc. This first type of categories has already been used in several studies that investigate the relationship between GT and stock returns, see for instance Curme et al. (2014), Dzielinski (2012) and Vaughan & Chen (2015). The second type of categories have also been used in several fields, for instance to predict unemployment, tourism, automobile sales and private consumption (Choi & Varian 2009, Vosen & Schmidt 2011). However, to the best of our knowledge, we are the first to use it for stock prediction. This second type of categories have two main advantages over search terms. First, categories are aggregated from multiple search terms within the same topic. This makes it a more reliable measure for general interest than single search terms, which are prone to sudden fluctuations, often unrelated to the researcher's intended meaning of the search term. Second, GT categories are language neutral, which means our approach can easily be extended to non-English speaking markets. This enables us to test our approach for several countries.

In this analysis, we develop multiple prediction models that use a large set of GT categories to predict next week returns of several industry indices. The predictions are used to form a trading strategy where we go long (short) a set of the best (worst) performing indices. The predictive models are developed for the US market, and the best model, an Elastic Net model, is subsequently tested on the US out-of-sample period and four additional countries, Japan, Great Britain, France and India. We show that the Elastic Net model significantly outperforms the total return index for all five countries, suggesting that GT categories have predictive power on next-week industry returns. We find that our approach is most profitable during weeks where the overall stock market drops, and that over time, the relationship between a search category and industry returns can change both in magnitude and direction.

The rest of this paper is structured as follows: In section 2 we describe the data used, and define key variables. Section 3 introduces the predictive models, and defines the backtesting framework used. In section 4, we present our main findings. Section 5 concludes.

2 Data

The financial data is obtained from Refinitiv Eikon, and Kenneth R. French's online data library (*Kenneth R. French* 2021). The search data is obtained from Google Trends. All our analyses are conducted using weekly data. The industry stock indices we look at were first introduced in late 2006, so our full sample period is from January 1, 2007 to December 31, 2020.

2.1 Financial data

In this study we are interested in returns on an industry-level. For the US market, which we use to develop our models, we use the S&P Composite 1500 Select Industry Indices. The industries are defined by the Global Industry Classification Standard (GICS) scheme, created by MSCI and S&P. For the remaining countries, we use the Refinitiv industry indices, based on the The Refinitv Business Classification (TRBC) scheme. The indices were launched at varying dates, however the majority had been launched by the end of 2006. Because of this, we include all industry indices that were launched before January 2007. Table 1 shows an overview of the different countries and specifies the number of valid industry indices together with the evaluation benchmark used.

Country	**Market Capitalization (bnUS$)**	**Number of industry indices**	**Benchmark**
United States	30,440	64	S&P 1500 Composite Total Return Index
Japan	5,300	55	Refinitiv Japan Total Return Index
Great Britain	4,700*	29	Refinitiv Great Britain Total Return Index
France	2,370	19	Refinitiv France Total Return Index
India	2,080	38	Refinitiv India Total Return Index

Table 1: Country overview, ranked by total stock market capitalization in the year 2018.
*) No country total available, estimated based on size of London Stock Exchange

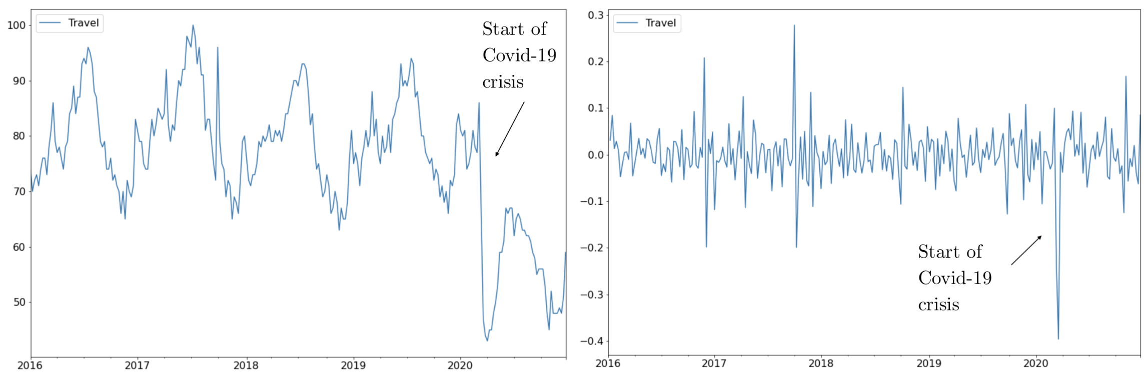
The weekly price data is retrieved from Refinitiv Eikon. All indices are total return indices, so prices are adjusted for stock splits, dividends and similar events. All indices are denoted in dollars, so for all countries we use the 1 Month Treasury Bill rate as a proxy for the risk free rate. Weekly returns are calculated as:

$$r_{i,t} = \frac{P_{i,t+1}}{P_{i,t}} \tag{1}$$

where $r_{i,t}$ is the return of week t for industry index i, and $P_{i,t}$ is the open price of the first trading day of week t for industry index i. We choose to use open prices because the weekly Google Trends data is released on Sunday afternoon, and we want to trade on the new information at the first opportunity.

2.2 Google Trends data

Google Trends is a service allowing users to retrieve data on the relative popularity of google searches over time. The service allows for retrieval of both historical and real-time data, starting from 2004 up to the time of the query. The frequency of the retrieved data can be monthly, weekly, daily or hourly - depending on the time frame length of the specific query. The searches are stored in a hierarchy of categories. This means a search like 'How to obtain a credit card?' is categorized into 'Credit Cards' which is at the bottom of the hierarchy 'Finance → Credit & Lending → Credit Cards.' It is the search interest for these categories we use in our analysis, as opposed to specific search terms. The US market is used in order to develop our prediction models, and is therefore the basis for category selection. We go through each GT category and select the ones we believe could have predictive power on the return of at least one of the 64 GICS industries. In total, 249 out of 1400 possible categories were selected. For each country, we retrieve the weekly search volumes for this set of categories. GT lets you delimit queries geographically, which means that for a specific country, we only consider the search volume for that particular country.



(a) SVI (b) ASVI

Figure 1: SVI and ASVI for the category *Travel* from 2016 to 2021.

Given a certain category, country, time frame and a number of other parameters, GT outputs a time series showing the search volume index (SVI) over time. The SVI does not represent absolute search volumes. Instead, the maximum search intensity of a query is assigned a value of 100, and the other values represent the relative popularity compared to this maximum. This means that SVI is not comparable across queries. Based on the raw SVI, we therefore construct an abnormal search volume index (ASVI) for each category. We take the first log difference of the SVI-series. In figure 1 we see the SVI of the category 'travel' change cyclically from a high in the summer vacation to a low in the winter. Such type of yearly seasonality is present for the majority of categories. We therefore subtract the mean of the log-differenced value from the same week one and two years ago. We find two years to be a good compromise between reducing noise and conserving data. This means we calculate ASVI as:

$$ASVI_{j,t} = \ln \frac{SVI_{j,t}}{SVI_{j,t-1}} - \frac{1}{2} \sum_{n=1}^{2} \ln \frac{SVI_{j,t-52*i}}{SVI_{j,t-1-52*i}} \tag{2}$$

where $SVI_{j,t}$ represents the search volume of category j for week t. Note that SVI data for week t is available on Sunday of the same week.

2.2.1 Benefit of using Google Trends categories

There are two major benefits of using GT *categories* instead of *search terms*. First, single search terms are susceptible to sudden interest spikes, often unrelated to the researcher's intended meaning of the search term. This issue is also discussed by Preis et al. (2013) and Bortoli & Combes (2015), who both argue that single search terms are prone to noise. For example, the movie 'The Social Network' was released around September 2010, which created a very large spike in searches for the single search term 'social networks', illustrated by figure 2. Even though the movie most likely increased the general interest for social networks around this time, the SVI of the single search term undoubtedly overstates this interest surge. By instead using categories, google filters out irrelevant searches, such as 'cinema tickets for the Social Network'. In this way, single events have a less drastic impact, making categories a more reliable measure for the general interest of a topic. The SVI for the category 'Social networks' is shown by figure 3, where we see that the increase in SVI is much less drastic.

Second, GT categories group relevant search terms regardless of language. Consider the case where we are interested in the search volume of the *Business* category in two separate languages, English and German. Without categories, we would have to find business-related search terms in both languages. However, with GT categories this is done for us. If we limit our searches to Germany, German business-related words, such as unternehmen (company), are automatically included in the *Business* category. This language neutrality has two advantages. It allows us to analyse countries for which we do not speak the language. Additionally, it allows us to use

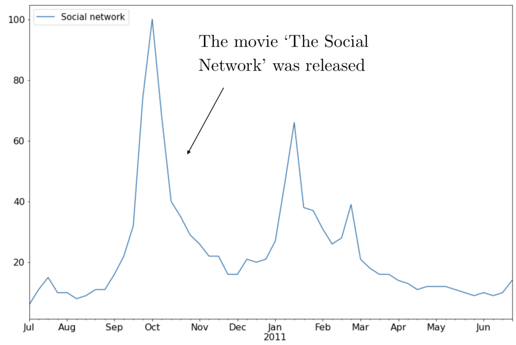


Figure 2: SVI for the search term *Social network* from July 2010 to July 2011.

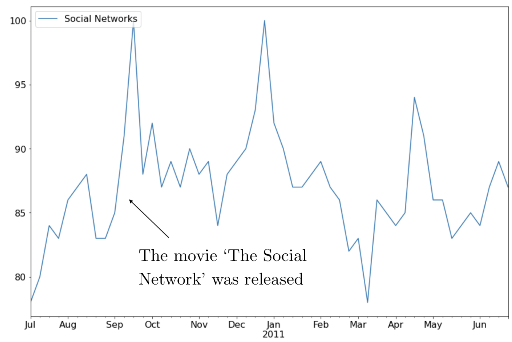


Figure 3: SVI for the category *Social networks* from July 2010 to July 2011.

the same set of categories for all countries, making our approach easily transferable to other countries.

3 Methodology

In this section we first present the prediction models used in our analysis. Because we work with many predictors per sample, the models we consider are all selected to deal with high dimensionality data. We try both filter methods and embedded methods. We then present the backtesting framework we use to develop and test our models, including the trading strategy and evaluation metrics used.

3.1 Prediction models

In this subsection, we present the prediction models we use. All methods use a forward-rolling approach with a fixed-size sliding estimation window of one year, which means we only use information available at estimation time. The target variable is the next-week return of the specific industry index, denoted $r_{i,t+1}$. The set of predictors, X_t, consists of the current week return of the specific industry index, and the abnormal search volumes of the GT categories $ASVI_{1,t}, ..., ASVI_{249,t}$.This means that the same set of GT categories are used for all industries. All models include an intercept term. With a window size of one year, there are 250 predictors and only 52 samples for each model fit, hence we consider models that deal with high-dimensional data. All models are implemented using the Python library Scikit-learn (Pedregosa et al. 2011), v. 0.24.2. All non-default parameters are specified in appendix A. Since three of the models (Random Forest, Lasso, and Elastic Net) have a degree of randomness, the results presented for these models are the average from 10 runs. The variance in performance between runs seems to be negligible.

3.1.1 Filter methods

Filter methods separate the process of feature selection and model prediction. The resulting feature subset is therefore unrelated to any specific prediction model. The goal of the feature selection process is to reduce dimensionality by removing the least relevant and redundant information from the data set. We use Principle Component Analysis (PCA) as our feature selection method. PCA aims to reduce the dimensionality of the data set while retaining as much information as possible. It does so by constructing a smaller set of principal components that are linear combinations of the original variables. Before the principal components are computed, the data is standardized by subtracting the mean and dividing by the variance. The first component is the axis that when the data is projected onto it, has the greatest variance among all candidates axes. Similarly, the second component is the axis that maximizes the variance, while being orthogonal to the first components. This process is repeated until one has the desired number of components. We try out five different numbers of principal components (10, 20, 30, 40, 50), and based on in-sample performance, we end up using 20 components. The resulting set of principal components are used as predictors in two different models: a linear Ordinary Least Squares Regression, and a nonlinear Random Forest Regression model.

3.1.1.1 Ordinary Least Squares

Our first model is an Ordinary Least Squares regression model on the form:

$$y_i = X_i\beta + \epsilon \tag{3}$$

where y_t is the dependent variable, X_t is the input set, and β is the vector of regression coefficients. The model assumes a linear dependency between the variables, and has the advantage of being both simple and easily interpretable.

3.1.1.2 Random Forest

We also consider a more advanced model to capture potential nonlinear relationships in our data. Borup & Schütte (2020) successfully used a feature reduction technique together with a Random Forest algorithm to predict unemployment growth using GT data. Inspired by their approach, we chose to include a Random Forest Regression model in our analysis. Random Forest is an ensemble technique that makes predictions based on the average prediction from a large number of simple regression trees. A single regression tree is a sequence of if-else rules that splits the training data into several regions. Figure 4 shows an example of a simple regression tree. The prediction for a new example, x_i, is the average target value, $\bar{y}$, for the region that x_i falls into.

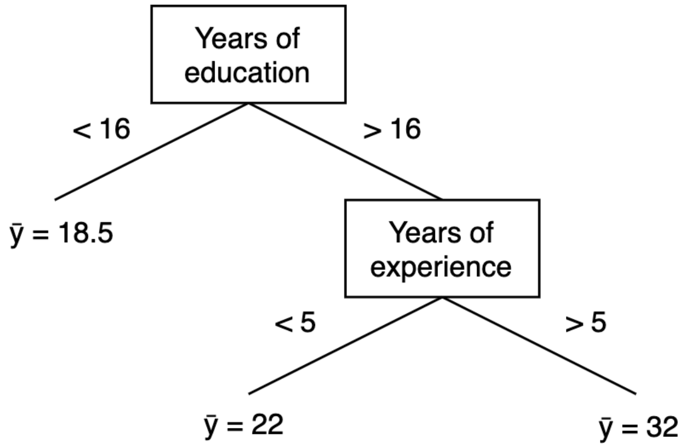


Figure 4: Example of a simple regression tree, where the target value is hourly wage ($).

These if-else rules are formed through a sequence of binary splits. The aim of each iteration is to find the split that minimises the resulting residual sum of squares (RSS), given by:

$$RSS = \sum_{y_t \in R_1} (y_i - \bar{y}_{R_1})^2 + \sum_{y_t \in R_2} (y_i - \bar{y}_{R_2})^2 \tag{4}$$

where R_1 and R_2 are the regions resulting from the binary split, and $\bar{y}_{R_i}$ is the average target value of the samples in region i. In the random forest algorithm, only a random subset of the predictors are considered for each split. This ensures variability in the resulting 'forest' of regression trees. The final prediction of the Random Forest algorithm is the mean of the predictions of all the individual regression trees.

3.1.2 Embedded methods

Embedded methods are another way of dealing with high-dimensional data. Like filter methods, embedded methods search for the optimal set of features, but the selection process is now closely tied to the specific learning algorithm, and tries to take advantage of the algorithm's characteristics. Regularization techniques are a common approach within this methodology, and two of the most common methods are Lasso and Ridge regression. These methods have been successfully used in financial applications, see for example Tian et al. (2015), Buncic & Melecky (2014), Nazemi & Fabozzi (2018). Zou & Hastie (2005) developed a model called Elastic

Net, which combines the advantages of Lasso and Ridge regression. We include all three methods in our analysis.

3.1.2.1 Ridge regression

Many prediction models, including standard OLS, are unsuited for situations where the number of predictors, M, is larger than the number of observations, N. Additionally, we suspect the GT data to be correlated, which will increase the variance of the OLS estimator and make it unstable. Ridge regression is a modified version of OLS, set to deal with these issues. Ridge regression differs from OLS in the loss function, which consists of the normal OLS loss term plus an L2-regularization term:

$$L_{Ridge}(\hat{\beta}) = L_{OLS}(\hat{\beta}) + L2loss = \sum_{i=1}^{N}(y_i + X_i\hat{\beta})^2 + \lambda\sum_{j=1}^{M}\hat{\beta}_j^2 \tag{5}$$

where y_i is the target value, X_i is the input set, and λ is the regularization penalty. The Ridge model is penalized for the sum of the squared coefficients, discouraging large coefficients. If $\lambda \to 0$, this becomes equivalent to standard OLS, and if $\lambda \to \infty$, all coefficients are forced to zero. Generally, this will lead to a more parsimonious model, and the variance will decrease but at the cost of some bias. As we increase the λ, the solution will be more stable but the bias will increase. Setting a value that is too large will cause underfitting, and setting a value that is too small will result in the same problems as standard OLS. Being able to control this bias-variance tradeoff will help us deal with both high-dimensionality and multicollinearity. We try out five different λ-values (0.0001, 0.001, 0.01, 0.1, 1). Based on in-sample performance, we ended up with $\lambda = 0.01$.

3.1.2.2 Lasso

Least Absolute Shrinkage Selector Operator (Lasso) is similar to Ridge regression in that it modifies the OLS loss function with an additional regularization term, this time L1-loss, specified as:

$$L_{Lasso}(\hat{\beta}) = L_{OLS}(\hat{\beta}) + L1loss = \sum_{i=1}^{N}(y_i + X_i\hat{\beta})^2 + \lambda\sum_{j=1}^{M}|\hat{\beta}_j| \tag{6}$$

In general, Lasso has many of the same advantages as Ridge. The main difference lies in how predictor coefficients are penalized. Lasso penalizes the sum of absolute values of the weights, which means some of the coefficients can be set exactly to zero. In this way, Lasso will actually perform feature selection. In Ridge, the squared term penalizes extreme values. This will lead to more evenly distributed coefficients, however they are never zeroed. This difference also affects how the methods deal with multicollinearity. While Lasso might randomly eliminate relevant predictors that are correlated with other predictors, in Ridge these predictors will all be retained, but are instead evenly diminished. Again, we try out five different λ-values (0.0001, 0.001, 0.01, 0.1, 1). Based on in-sample performance, we ended up with $\lambda = 0.0001$.

3.1.2.3 Elastic Net

Elastic Net was proposed by Zou & Hastie (2005) as a hybrid between Lasso and Ridge that tries to combine the advantages of the L1 and L2 loss. As with Ridge and Lasso, the model is a modified OLS, where the loss is now given by:

$$L_{ENET}(\hat{\beta}) = \sum_{i=1}^{N}(y_i + X_i\hat{\beta})^2 + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{M}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{M}|\hat{\beta}_j|\right) \tag{7}$$

where λ is the regularization-parameter, and α is the mixing parameter, which specifies the relative weight we give to the L1-loss. Zou & Hastie (2005) pointed out some potential problems with Ridge and Lasso that Elastic Net tries to overcome. In a setting where the number of predictors is greater than the number of observations, Lasso will select at maximum n predictors, potentially leaving out important information. The same restriction does not apply to Ridge. On the other hand, in the case where there are only a few significant predictors, Lasso will have a clear advantage over Ridge, since it can eliminate insignificant features altogether. By combining L1- and L2-loss, the Elastic Net tries to include the advantages of both Lasso and Ridge. We try out five different λ-values (0.0001, 0.001, 0.01, 0.1, 1) and nine different α-values (0.1, 0.2, ..., 0.9). Based on in-sample performance, we ended up with $\lambda = 0.01$ and $\alpha = 0.2$.

3.2 Backtesting procedure

Before conducting any backtesting, we calculate all variables to be used in the models and settle on a trading strategy. We look at the following countries in our analysis: USA, Japan, Great Britain, France, India. The US market is used to develop and evaluate our models. We split the US sample into an in-sample (67%) and an out-of-sample period (33%). The in-sample set is used to assess a larger set of models and parameters. Eventually, the best model is selected and tested on the out-of-sample data for all five countries.

3.2.1 Trading strategy

For each country, the industries are modelled separately. For instance, we have 64 separate models for the US market. The next week's return is predicted in a forward rolling approach based on a sliding window of historical data. Each week we get a predicted next-week return for each industry. On the open of the first trading day of the week, we go long the industries within the top 10% highest predicted next week returns, and short the bottom 10%. To maintain the exposure to the market equal to approximately one, the long position is weighted 130% and the short position 30%, a very common approach for long-short strategies (Clarke et al. 2008). For the US, we therefore go long and short 6 indices each week. However, France, for instance, only has 19 indices, and 10% would in this case be too few. We therefore increase the relative portfolio size slightly for some countries. Table 2 shows the size of the long and the short portfolio for each country.

Country	**Number of industry indices**	**Long/short portfolio size**
United States	64	6 industries
Japan	55	5 industries
Great Britain	29	4 industries
France	19	4 industries
India	38	5 industries

Table 2: Number of industry indices and portfolio size for each country.

3.2.2 Transaction costs

Since we cannot trade the industry indices directly, we need to trade its components (individual stocks), and therefore estimate the transaction cost of trading those stocks. Transaction costs (commissions + bid/ask spread) are set to 3 basis points for each one-way trade. Garveya & Wu (2009) find that commissions have declined gradually to as low as 0.15-0.3 cents per share in 2010 for professional investors. As a conservative measure, we therefore set the commission cost to 0.5 cents per share, and if we use notional stock prices from 2020, this results in a weighted average commission of around 0.75 basis points. Hagströmer (2021) finds that the weighted average effective spread of S&P 500 companies was around 2.84 basis points in 2015.

The market cap of the S&P 500 has been 90% of the S&P composite 1500 in the last couple of years, so it is fair to assume that the value-weighted average spread of the S&P composite 1500 index is not significantly higher than for S&P 500, as the remaining 1000 companies only contribute around 10% of the total market capitalization. We would rather be on the conservative side, so we set the effective spread to 4 basis points, i.e. 2 basis points for each one way trade. This results in a total cost of 2.75 basis points per one-way trade, which we round up to 3 basis points. Our strategy involves short positions, hence we set the yearly shorting costs equal to the risk free rate plus 22.5 basis points, based on findings by Kim & Lee (2019). Note that this level of costs are only realistic for institutional investors. Also note that we neglect transaction costs for the benchmarks. Although they are not directly tradeable, a replicating portfolio does not require very frequent rebalancing.

3.2.3 Evaluation metrics

We compare our models against two benchmarks for each country. The first is the equally weighted portfolio containing all industry indices for the specific country. The second is a benchmark index that reflects the return of the total market, where all companies are market cap-weighted. The specific benchmark used for each country is presented in table 1. The reason why these benchmarks might perform differently is that in the equally weighted portfolio, smaller industries will be over represented compared to the total return index. We only present before-cost performance for the benchmark, since transaction costs are neglected for these portfolios.

We evaluate our trading results using the following metrics: *Return* is the annualized average weekly return. *Volatility* is the annualized weekly standard deviation. *Sharpe* is the annualized weekly Sharpe ratio. *Max drawdown* is the weekly maximum drawdown, defined as the maximum loss from a peak to a low of a portfolio, before a new peak is achieved. *Jensen's alpha* is the annualized weekly Jensen's alpha using the benchmark index as the market proxy. *FF alpha* is the Fama-French 5 factor alpha, calculated using the Fama French 5-factor model, which includes the following pricing factors: the market premium factor (*Mkt-Rf*), the size factor (*SMB*), the value factor (*HML*), the profitability factor (*RMW*) and the investment factor (*CMA*). The Fama-French model is specified as:

$$\begin{aligned} r_{i,t} - r_{f,t} = \alpha_i + \beta_{i,Mkt-RF} * (Mkt - RF) + \beta_{i,SMB} * SMB + \\ \beta_{i,HML} * HML + \beta_{i,RMW} * RMW + \beta_{i,CMA} * CMA + \epsilon_{i,t} \end{aligned} \tag{8}$$

where $r_{f,t}$ is the risk free rate at time t and the βs are the pricing factor loadings.

Note that Jensen's alpha is calculated using a similar approach but only includes the first factor, *Mkt-Rf.* For all countries, the market proxy is the total return index, as specified in table 1. The remaining factor data is retrieved from Kenneth R. French's online data library. Due to insufficient factor data for the remaining countries, the Fama-French 5-factor alpha is only presented for the US market.

4 Results

In this section we present our main findings. First, the in-sample results are presented. We then choose the best performing model, the Elastic Net model, and test it on the out-of-sample data for all five countries. Furthermore, the results are validated using two robustness checks. Lastly, we take a closer look at the performance of the selected Elastic Net model.

4.1 In-sample results

We now present the results of the backtesting procedure on the in-sample data set for the US market. Table 3 shows a selection of trading metrics from the in-sample performance. All metrics are explained in section 3.2.3.

We see that all models achieve relatively good results, indicating that the GT categories contain valuable predictive information. Further, we see that the models perform well in terms of Sharpe ratio, but even better in terms of alpha. Hence, we suggest that such strategies would be best placed in a portfolio of other strategies. We base our final assessment of the models on the Fama French 5-factor (FF) alpha, which means the Elastic Net model (bolded) is the best performer, achieving an annualised FF alpha of around 15% after costs. In our further analysis, we will look more closely at this Elastic Net model and see how it performs on so-far unseen data.

4.2 Out-of-sample results

We now present results from testing the Elastic Net model on unseen data, which gives a better indication of its actual predictive power. For the US, the out-of-sample period is the last third of the period. For the remaining countries, the out-of-sample period is the full sample period. Table 4-8 shows the trading metrics for each country. Figure 5 shows the logarithmic cumulative return of the model and the benchmarks.

Model	Type	Return	Volatility	Sharpe	Max draw-down	Jensen's alpha	FF alpha
S&P 1500 Composite Total Return Index	Before costs	10.3%	20.7%	0.46	-3.5%	0.00%	0.00%
Equally-weighted buy-and-hold	Before costs	12.4%	21.3%	0.54	-3.3%	1.8%	1.3%
PCA OLS	Before costs	28.5%	27.2%	0.89	-4.5%	17.3%	17.7%
	After costs	24.8%	27.2%	0.79	-4.5%	14.2%	14.6%
PCA RF	Before costs	22.4%	27.3%	0.73	-5.1%	10.5%	10.8%
	After costs	19.1%	27.2%	0.63	-5.2%	7.5%	7.8%
Ridge	Before costs	23.4%	29.0%	0.73	-2.8%	11.4%	14.0%
	After costs	20.5%	29.0%	0.64	-2.8%	8.8%	11.3%
Lasso	Before costs	19.2%	28.2%	0.62	-5.4%	7.7%	9.1%
	After costs	16.2%	28.2%	0.52	-5.6%	5.0%	6.3%
Elastic Net	**Before costs**	**28.6%**	**28.4%**	**0.90**	**-3.4%**	**16.4%**	**18.1%**
	After costs	**25.4%**	**28.3%**	**0.80**	**-3.4%**	**13.5%**	**15.2%**

Table 3: Trading metrics for the US, in-sample (01-01-2008 to 31-12-2016).

The Elastic Net model performs very well in all countries, with consistently higher Sharpe ratios and annualised alphas ranging from 4.4% to 11.0% after transaction costs. We also note that for most countries the maximum drawdown is either similar or lower compared to the benchmark, suggesting that our model achieves superior returns without increasing the downside risk. There are a few other studies that successfully use GT data in a trading strategy. Preis et al. (2013) and Curme et al. (2014) both implement multiple single keyword strategies that buy and sell a US stock index weekly in the period from 2004-2011, achieving an average excess cumulative return of 20% and 40% over a buy-and-hold portfolio. Gwilym et al. (2014) present significantly stronger results in the same time period, achieving an after-cost annualised excess return of around 20%, using the single search term “Penny Stock”. As seen in figure 4, our Elastic Net model achieved an after-fee annualised excess return of around 9% for the US out-of-sample period. Our strategy does have a very different risk profile, so comparing simple returns can be somewhat misleading. However, our results are also strong in terms of risk-adjusted alpha. Unfortunately, the other studies only report their results in terms of simple returns. Compared to the other studies, our analysis covers a much longer time span and larger geographical scope. Furthermore, the existing literature is lacking in terms of robustness checks. We address this by having a more distinct separation between in- and out-of-sample, and by including two additional robustness checks.

Model	Type	Return	Volatility	Sharpe	Max drawdown	Jensen's alpha	FF alpha
S&P 1500 Composite Total Return Index	Before costs	17.3%	19.7%	0.75	-4.5%	0.00%	0.00%
Equally-weighted buy-and-hold	Before costs	15.5%	20.6%	0.64	-3.5%	-1.8%	0.8%
Elastic Net	Before costs	29.9%	20.3%	1.23	-2.5%	14.4%	12.3%
	After costs	26.0%	20.3%	1.07	-2.5%	11.0%	8.8%

Table 4: Trading metrics for the **US**, out-of-sample (01-01-2017 to 31-12-2020).

Model	Type	Return	Volatility	Sharpe	Max drawdown	Jensen's alpha
Refinitiv Japan Total Return Index	Before costs	6.4%	22.4%	0.25	-3.1%	0.00%
Equally-weighted buy-and-hold	Before costs	7.8%	21.4%	0.32	-3.6%	2.0%
Elastic Net	Before costs	18.3%	28.3%	0.57	-2.9%	11.1%
	After costs	15.0%	28.2%	0.48	-2.9%	8.0%

Table 5: Trading metrics for **Japan**, full sample (01-01-2008 to 31-12-2020).

Model	Type	Return	Volatility	Sharpe	Max drawdown	Jensen's alpha
Refinitiv GB Total Return Index	Before costs	6.8%	19.0%	0.32	-3.0%	0.00%
Equally-weighted buy-and-hold	Before costs	8.9%	19.0%	0.42	-2.4%	2.7%
Elastic Net	Before costs	16.0%	25.6%	0.56	-3.6%	8.5%
	After costs	13.0%	25.6%	0.46	-3.6%	5.8%

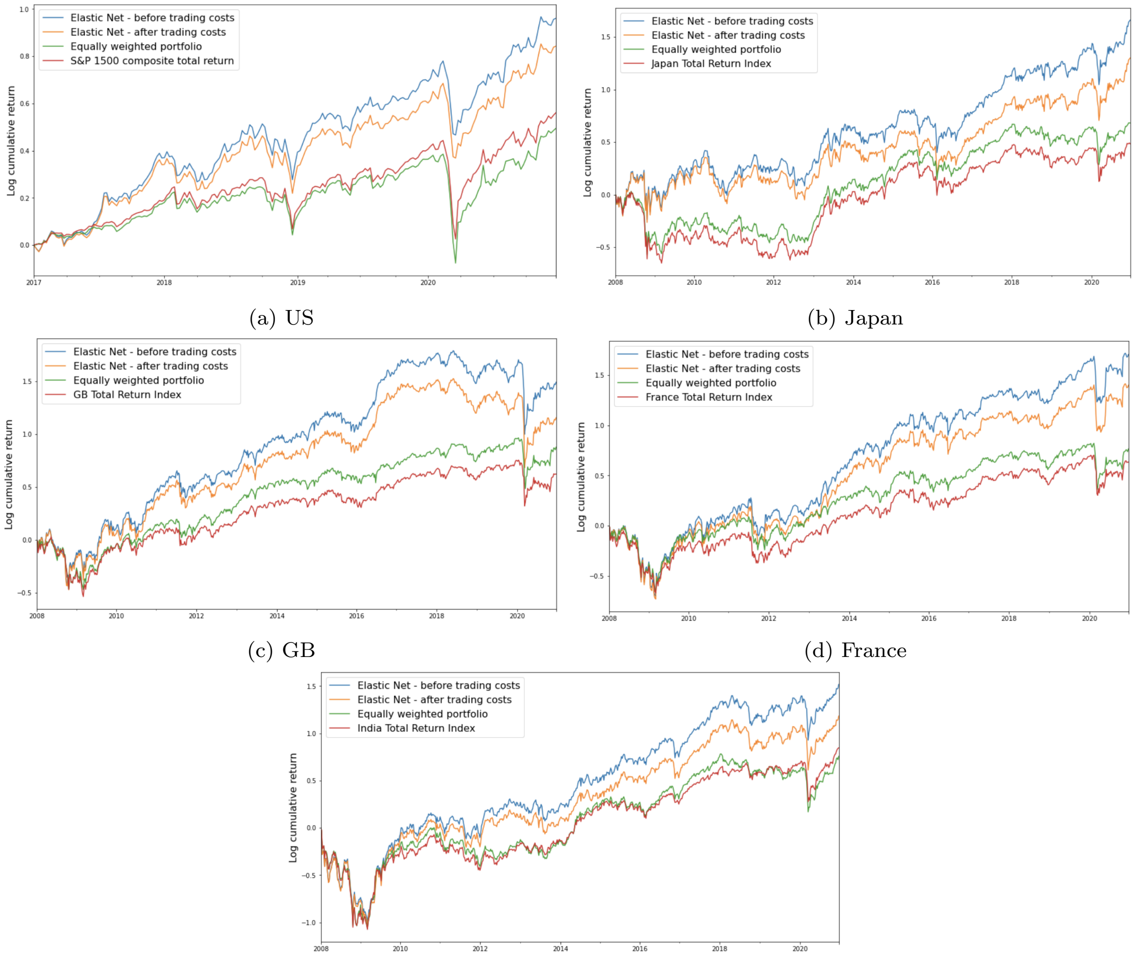
Table 6: Trading metrics for **GB**, full sample (01-01-2008 to 31-12-2020).

Model	Type	Return	Volatility	Sharpe	Max drawdown	Jensen's alpha
Refinitiv France Total Return Index	Before fees	7.4%	21.4%	0.31	-3.2%	0.00%
Equally-weighted buy-and-hold	Before fees	8.7%	22.2%	0.35	-2.7%	1.7%
Elastic Net	Before costs	18.6%	28.2%	0.59	-2.6%	10.6%
	After costs	15.9%	28.2%	0.50	-2.7%	8.0%

Table 7: Trading metrics for **France**, full sample (01-01-2008 to 31-12-2020).

Model	Type	Return	Volatility	Sharpe	Max drawdown	Jensen's alpha
Refinitiv India Total Return Index	Before costs	9.8%	23.9%	0.37	-7.42%	0.00%
Equally-weighted buy-and-hold	Before costs	9.0%	23.9%	0.34	-6.6%	0.10%
Elastic Net	Before costs	16.8%	27.7%	0.54	-7.9%	7.1%
	After costs	13.9%	27.7%	0.45	-8.0%	4.4%

Table 8: Trading metrics for **India**, full sample (01-01-2008 to 31-12-2020).



(a) US (b) Japan

(c) GB (d) France

(e) India

Figure 5: Log of cumulative return of the Elastic Net model for the out-of-sample period.

4.3 Robustness checks

Here we perform two different robustness checks. First we compare the performance of the Elastic Net model to a random strategy. We then assess the model's stability by checking its performance for alternative parameter configurations.

4.3.1 Performance compared to random strategies

Here we compare our model performance against a random strategy. The random strategy uses a similar trading strategy. The only difference is that now, a random set of industries are selected for the long and short portfolio each week. This shows how a model without any predictive power would perform using the same trading strategy. The analysis is conducted on the out-of-sample period for all five countries. We present the average annualised return. For the random strategy, metrics are presented based on 10 000 independent realizations. The standard deviation (std.

dev.) is defined as the standard deviation of the average returns of all realizations. In table 9, *Elastic Net outperformance* shows the outperformance of the Elastic Net model over the random strategy, in standard deviations of the cumulative returns of the random strategies. The results presented are after transaction costs.

Metric	**US**	**Japan**	**GB**	**France**	**India**
Average return, Elastic Net	26.0%	15.0%	13.0%	15.9%	13.9%
Average return, Random Strategy	11.9%	5.1%	6.0%	6.0%	6.9%
Average return, Equally Weighted	15.5%	7.8%	8.9%	8.7%	9.0%
Average return, Total Return Benchmark	17.3%	6.4%	6.8%	7.4%	9.8%
Std. dev. average return, Random	5.1%	3.0%	3.8%	3.7%	3.2%
90%-quantile average return, Random	18.6%	8.9%	10.8%	10.7%	11.2%
95%-quantile average return, Random	20.5%	9.8%	12.5%	12.1%	12.3%
Elastic Net outperformance (in std. dev. of random cumulative return)	2.76	3.34	1.86	2.63	2.15

Table 9: Performance of random strategies, compared to the Elastic Net model and the benchmarks.

We see that for all countries, the Elastic Net model outperforms the random strategy at the 95%-level or higher, suggesting that our model can accurately predict the returns of the industry indices.

4.3.2 Parameter stability

Here we will assess the model's stability by considering its performance for alternative parameter configurations. We look at alternative values for the regularization strength, λ, and the L1-weight ratio, α. The analysis is conducted on the out-of-sample period for all five countries. Table 10 shows the annualised weekly after-cost Jensen's alpha for the out-of-sample period of all five countries, given different regularization strengths, λ. The other parameters are kept constant. The main specification, $\lambda = 0.01$, is bolded. Table 11 shows Jensen's alpha for the out-of-sample period of all five countries, given different L1-weight ratios, α. The other parameters are kept constant. The main specification, $\alpha = 0.2$, is bolded.

Overall, the model performance is relatively stable to parameter-changes, with similar results for many of the alternative model specifications. This suggests that the strong model performance is not the result of coincidental parameter selection. On the contrary, this is an indication that our overall approach shows promise, and that the exact model specification is of less importance.

Parameter	US	Japan	GB	France	India
$\lambda = 0.0001$	9.1%	3.0%	1.4%	6.8%	6.9%
$\lambda = 0.001$	8.2%	3.1%	3.2%	7.2%	7.4%
$\boldsymbol{\lambda = 0.01}$	**11.0%**	**8.0%**	**5.8%**	**8.0%**	**4.4%**
$\lambda = 0.1$	7.0%	5.4%	3.3%	4.6%	7.4%
$\lambda = 1$	-6.5%	0.7%	3.3%	-0.5%	4.4%

Table 10: Annualised weekly Jensen's alpha given different λ-parameters, for the out-of-sample period.

Parameter	US	Japan	GB	France	India
$\alpha = 0.1$	7.6%	7.8%	5.8%	7.6%	4.0%
$\boldsymbol{\alpha = 0.2}$	**11.0%**	**8.0%**	**5.8%**	**8.0%**	**4.4%**
$\alpha = 0.3$	12.9%	8.0%	4.2%	9.0%	3.4%
$\alpha = 0.4$	12.2%	8.3%	2.5%	7.2%	3.3%
$\alpha = 0.5$	10.0%	8.5%	1.7%	6.6%	3.2%
$\alpha = 0.6$	6.2%	4.8%	4.9%	6.4%	2.0%
$\alpha = 0.7$	4.0%	3.9%	5.6%	6.3%	2.9%
$\alpha = 0.8$	3.1%	3.9%	4.4%	4.0%	2.0%
$\alpha = 0.9$	-0.1%	3.6%	4.3%	6.3%	1.8%

Table 11: Annualised weekly Jensen's alpha given different α-parameters, for the out-of-sample period.

4.4 Model deep-dive

In this section we look closer at the out-of-sample performance of the Elastic Net model. First we look at the model performance under different market conditions. Second, we look at the predictive strengths of the model. Third, we look more closely at the GT predictors, and the importance of having a dynamic model. Lastly, we consider the cross-national correlation of the model's performance.

4.4.1 When is the model performing well?

We now consider the performance of the Elastic Net under different market conditions. Table 12 shows the model's after-cost performance during bull and bear markets. We define a bull market week as a week where the specific country's total return index achieves a positive return, and a bear market week when it does not. We also show how often the Elastic Net model outperforms the total return index, where a value of 58% for a bear market means the model outperformed the index in 58% of all bear market weeks. All values presented are weekly.

Country	Type	Number of weeks	Total return index, mean return	Elastic Net, mean return	Elastic Net, outperformance share
US	Bear market	72	-2.04%	-1.73%	58.3%
	Bull market	136	1.55%	1.59%	49.3%
Japan	Bear market	308	-2.31%	-1.99%	56.8%
	Bull market	369	2.15%	2.16%	47.4%
GB	Bear market	301	-1.89%	-1.57%	56.5%
	Bull market	376	1.74%	1.68%	48.1%
France	Bear market	304	-2.23%	-2.01%	56.6%
	Bull market	373	2.07%	2.15%	47.2%
India	Bear market	290	-2.43%	-2.22%	53.5%
	Bull market	387	2.13%	2.10%	51.4%

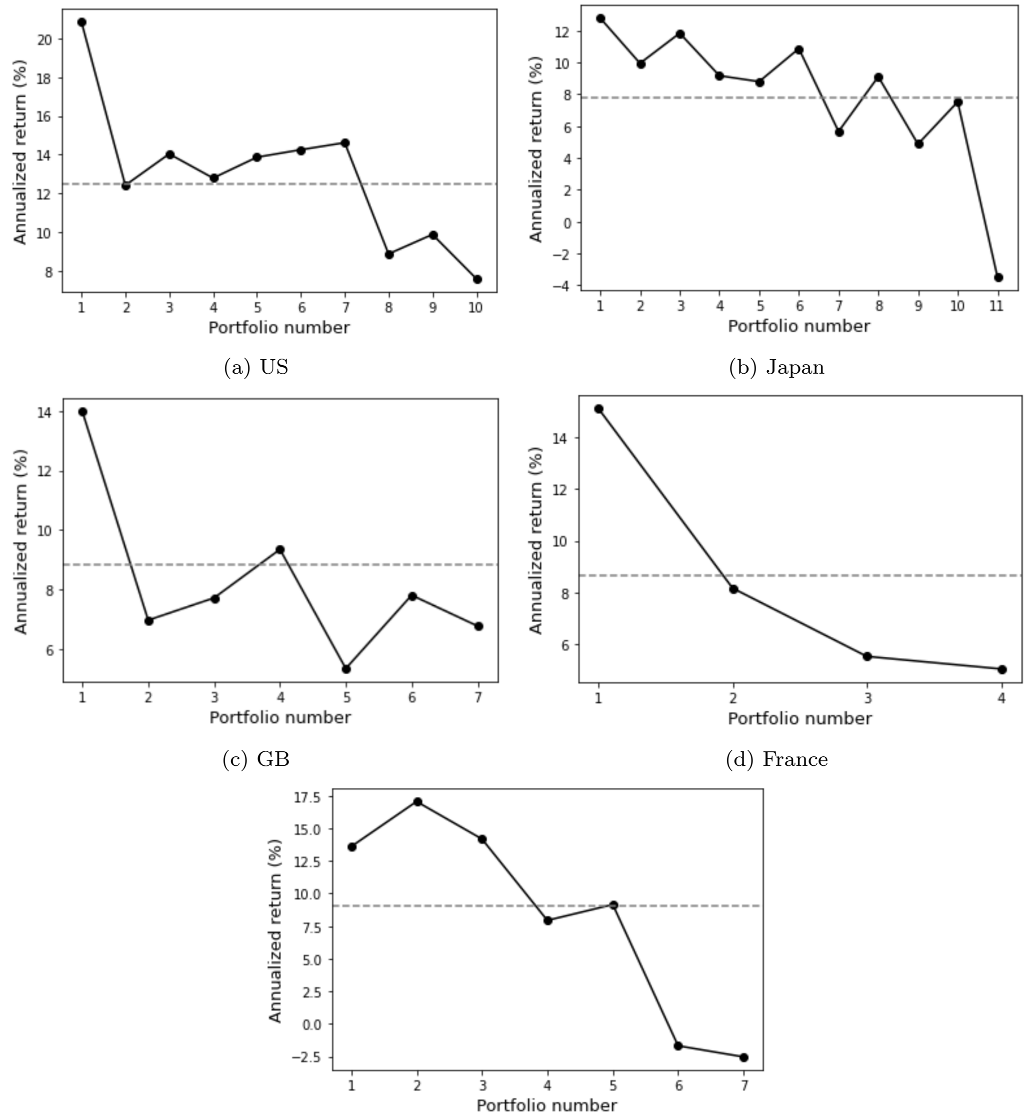
Table 12: Model's after-cost performance during bull- and bear weeks.

We see that during bull market weeks, the Elastic Net model performs similar to the benchmark. However, during bear market weeks the outperformance is more frequent and larger in magnitude. We can thus conclude that the model's main strength is to reduce losses during bad weeks, which seems to indicate that the GT predictors contain more predictive power in weeks where the total stock market falls. This is in line with the findings of Da et al. (2014) and Choi & Varian (2009), who find that Google searches are more accurate predictors during market declines.

4.4.2 Predictive strength

In this subsection we look at how well the Elastic Net model is able to rank the industry indices according to their next week return. We do this by creating a set of portfolios containing the industries in ranked order based on their predicted return. For instance, for the US, there are ten portfolios. Each week, the 10% of industries with the highest next-week predicted return are put into the first portfolio, the next 10% of industries in the second portfolio, and so on. The industries are equally weighted, and all returns are stated before transaction costs. The number of indices in each portfolio varies across countries, following table 2. Figure 6 shows the annualised average return of the ranked portfolios for all five countries.

For a model without any predictive power, we would expect to see no pattern in the portfolio returns. Instead, we see a declining trend from the first to the last portfolio. For most of the countries, this trend is very strong, suggesting that our model is able



(a) US

(b) Japan

(c) GB

(d) France

(e) India

Figure 6: Average before-costs return of ranked portfolios, for the out-of-sample period.

to rank the industry returns well. We note that the model is better at predicting the first and last portfolio, whereas for the middle portfolios the declining trend is not always as strong. However, since we only use the top and bottom portfolio in our trading strategy, the predictive power of the middle portfolios are of less interest.

4.4.3 The Google Trends predictors

Since we use a forward rolling approach, the model is refitted every week, which means the model coefficients are free to change over time. It should be emphasized that we use a large set of predictors, and that each time the Elastic Net model is fitted, it performs feature selection. Therefore, our model not only changes the coefficients of each predictor over time, but also selects a new subset of predictors to be used each week. This addresses one of the biggest limitations with the existing body of GT research, namely that most approaches either use a single search term or a static list of multiple search terms which stay constant over the full sample period. Da et al. (2014) is one of the few earlier studies who actively deal with this. In their approach, they re-select the thirty most statistically significant keywords from a larger pool of search terms every six months. We believe such an adaptable approach is important to stay up to date with changing market trends. Figure 7 and 8 show how the Elastic Net coefficients change over time for two selected predictors. To allow for comparison between coefficients, in the following analysis they are standardized as $\beta_i^* = \frac{S_x}{S_y}\beta$, where β^* is the standardized coefficient of predictor i, and S_x and S_y is the standard deviation of the predictor and regressand, respectively.

Figure 7: Coefficient of search category *Air Travel* for the industry Airlines.

Figure 8: Coefficient of search category *Property Development* for the industry Real Estate Management & Development.

We see that the magnitude of the coefficients are changing significantly over time. This supports the hypothesis that while a GT category might be important in one period, this is not necessarily the case for the next period. Further, we note that the sign of the coefficient is also changing over time. Let's consider figure 7, showing the coefficient of the search category *Air Travel* over time. A positive coefficient means that an unexpected increase in search activity for air travel-related searches predicts higher next-week returns of the airline industry, whereas a negative coefficient would mean the opposite. Since the $ASVI$ of a GT predictor can be interpreted as the public's abnormal attention towards a topic, it makes sense that the coefficient can change. In one period, the increased attention towards air travel can be positively loaded, perhaps after a loosening in travel-restrictions. However, in another period, it might be negatively loaded, for instance after increased focus on environmental issues. Both these findings emphasize the importance of a flexible model that can frequently reevaluate the predictor coefficients.

To strengthen this hypothesis, we show the 52-week rolling mean coefficients for 40 additional selected predictors, shown in figure 9. We see that the predictor coefficients behave similarly to those presented in figure 7 and 8 - they change over time in terms of both magnitude and sign.

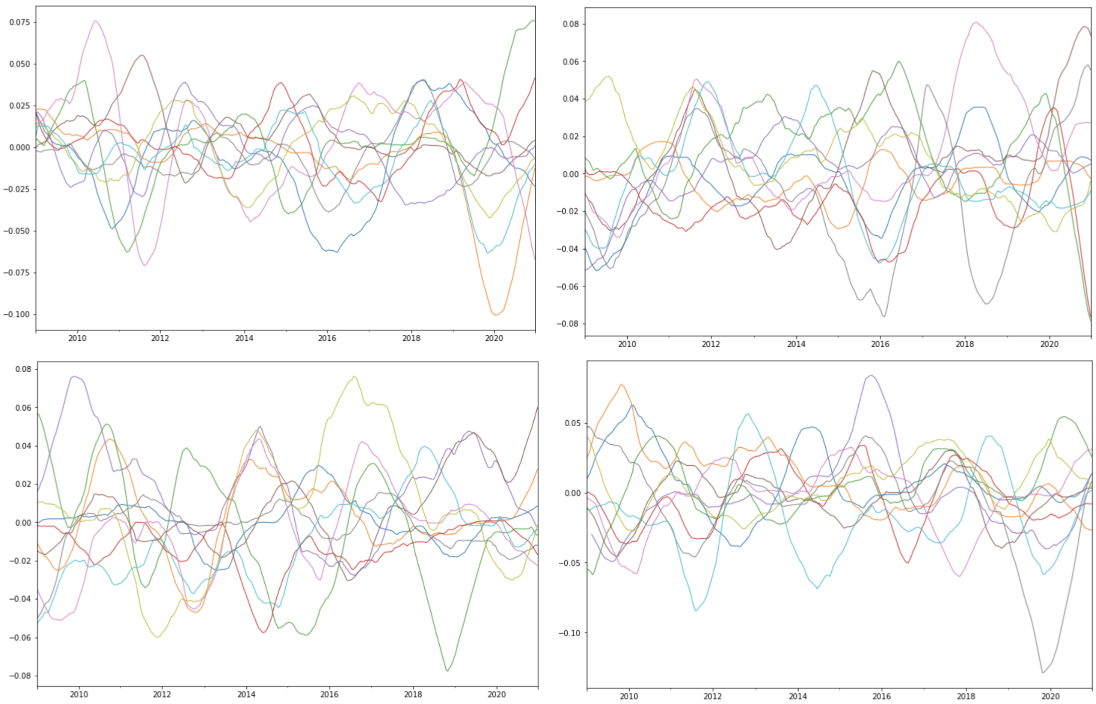


Figure 9: Rolling 52-week mean of 40 randomly selected coefficients, in four separate plots.

4.4.4 Geographical diversification

We now look at the correlation of the model's performance across countries. Table 13 shows the correlation matrix of weekly portfolio returns, and table 14 shows the correlation matrix of weekly abnormal returns. We define the weekly abnormal return for a country as the weekly return of the Elastic Net model minus the same-week return of the country's specific benchmark.

	US	Japan	GB	France	India
US	1.000				
Japan	0.550	1.000			
GB	0.598	0.305	1.000		
France	0.590	0.539	0.550	1.000	
India	0.419	0.507	0.507	0.608	1.000

Table 13: Correlation of portfolio returns across countries.

We see that the cross-national correlation of abnormal returns is relatively weak for all countries. Given the out-of-sample results presented earlier, we know that all models perform relatively well on their own. The weak correlation presented in table

	US	Japan	GB	France	India
US	1.000				
Japan	0.172	1.000			
GB	0.130	-0.148	1.000		
France	0.116	0.136	-0.106	1.000	
India	-0.129	-0.0548	0.0545	0.0818	1.000

Table 14: Correlation of abnormal portfolio returns across countries.

14 suggest that our approach can be further improved by considering a multinational trading strategy, potentially achieving greater risk-adjusted return by reducing the country-specific idiosyncratic risk. Additionally, since GT categories are language neutral, one can easily extend our approach to other countries, opening up for even better multinational diversification.

5 Conclusion

In this paper we analyse the usefulness of Google search activity for stock prediction. Since Google made its search data available in 2006 through the service Google Trends (GT), many researchers have have investigated its usefulness for predicting stock returns with mixed results. However, most existing studies that use GT data use individual *search terms*, as opposed to the less used *categories*, a measure aggregated from several related search terms. Additionally, most relevant studies are restricted to the US market or the author's home market. In this study we use GT categories to predict weekly returns of industry stock indices in the world's five biggest stock markets where GT data is available: the US, Great Britain, Japan, France and India. The language neutrality of GT categories enables us to analyse all five countries using the same approach.

We develop several models that use a large set of GT categories to predict the next week returns of several industry indices. Since we work with many predictors per sample, the models we consider are all selected to deal with high-dimensional data. The predictions are used in a long-short trading strategy, which significantly outperforms the total US stock market after also accounting for transaction costs. The best performing model, an Elastic Net model, is subsequently shown to achieve similar results on the out-of-sample data for all five countries. Furthermore, we find that the Elastic Net model achieves most of its excess returns during weeks where the overall stock market drops. We also find that over time, the relationship between a search category and industry returns can change both in magnitude and direction.

We address this by re-fitting the prediction model on a weekly basis, which allows the model coefficients to adapt to changing market trends. Lastly, we find that the abnormal returns of our strategies are only weakly correlated across countries, suggesting that our approach is also suitable for more sophisticated multinational strategies.

There is currently no consensus in the literature about whether or not Google Trends data can be used to predict stock returns. The overall conclusion of this study is that Google Trends categories can indeed be useful predictors for industry stock indices. Finally, we suggest that future research further explore the benefits of GT categories, which are language-neutral and seem to be a less noisy measure of public interest compared to single search terms. We also encourage researchers to include complementary input variables and to look at shorter or longer time horizons for both the input and target variables.

References

Aouadi, A., Arouri, M. & Teulon, F. (2013), 'Investor attention and stock market activity: Evidence from france.', *Economic Modelling* **35**, 674–681.

Baele, L., Bekaert, G. & Inghelbrecht, K. (2010), 'The determinants of stock and bond return comovements.', *Rev. Financ. Stud.* **23**, 2374–2428.

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

Bank, M., Larch, M. & Peter, G. (2011), 'Google search volume and its influence on liquidity and returns of german stocks', *Swiss Society for Financial Market Research* .

Barber, B. M. & Odean, T. (2008), 'All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors', *The Journal of Finance* **21**(2), 785–818.

Barreira, N., Godinho, P. & Melo, P. (2013), 'Nowcasting unemployment rate and new car sales in south-western europe with google trends', *NETNOMICS: Economic Research and Electronic Networking* **14**(3), 129–165.

Bijl, L., Kringhaug, G., Molnár, P. & Sandvik, E. (2016), 'Google searches and stock returns.', *International Review of Financial Analysis* **45**, 150–156.

Bortoli, C. & Combes, S. (2015), 'Contribution from google trends for forecasting the short-term economic outlook in france: limited avenues', *Institut National de la Statistique et des Éstudes Économiques. Available online: https://www. insee. fr/en/statistiques/1408911* .

Borup, D. & Schütte, E. C. M. (2020), 'In search of a job: Forecasting employment growth using google trends', *Journal of Business & Economic Statistics* pp. 1–15.

Broadstock, D. C. & Zhang, D. (2019), 'Social-media and intraday stock returns: The pricing power of sentiment', *Finance Research Letters* **30**, 116–123.

Brown, G. & Cliff, M. (2004), 'Investor sentiment and the near-term stock market', *Journal of Empirical Finance* **11**, 1–27.

Buncic, D. & Melecky, M. (2014), 'Equilibrium credit: The reference point for macroprudential supervisors', *Journal of Banking & Finance* **41**, 135–154.

Challet, D. & Bel Hadj Ayed, A. (2013), 'Predicting financial markets with google trends and not so random keywords'. https://ssrn.com/abstract=2310621 [Accessed: December 1, 2020].

Chemmanur, T. & Yan, A. (2019), 'Advertising, attention, and stock returns.', *Quart. J. of Fin.* **09**.

Chen, J., Tang, G., Yao, J. & Zhou, G. (2019), 'Investor attention and stock returns', *Available at SSRN 3194387* .

Choi, H. & Varian, H. (2009), 'Predicting initial claims for unemployment benefits', *Google Inc* **1**, 1–5.

Choi, H. & Varian, H. (2012), 'Predicting the present with google trends.', *Economic Record* **88**, 2–9.

Clarke, R., de Silva, H., Sapra, S. & Thorley, S. (2008), 'Long–short extensions: How much is enough?', *Financial Analysts Journal* **64**(1), 16–30.

Curme, C., Preis, T., Stanley, H. & Moat, H. (2014), 'Quantifying the semantics of search behavior before stock market moves.', *Proceedings of the National Academy of Sciences* **111**, 11600–11605.

Da, Z., Engelberg, J. & Gao, P. (2014), 'The sum of all fears investor sentiment and asset prices.', *Review of Financial Studies* **28**(1), 1–32.

De Long, J., Shleifer, A., Summers, L. & Waldmann, R. (1990), 'Noise trader risk in financial markets.', *Journal of Political Economy* **98**, 703–738.

Dimpfl, T. & Jank, S. (2015), 'Can internet search queries help to predict stock market volatility?', *Eur Financial Management* **22**, 171–192.

Dondio, P. (2013), Stock market prediction without sentiment analysis: using a web-traffic based classifier and user-level analysis, *in* '2013 46th Hawaii International Conference on System Sciences', IEEE, pp. 3137–3146.

Duz Tan, S. & Tas, O. (2020), 'Social media sentiment in international stock returns and trading activity', *Journal of Behavioral Finance* pp. 1–14.

Dzielinski, M. (2012), 'Measuring economic uncertainty and its impact on the stock market.', *Finance Research Letters* **9**, 167–175.

Fink, C. & Johann, T. (2012), 'May i have your attention, please: The market microstructure of investor attention.', *SSRN Journal* .

Garveya, R. & Wu, F. (2009), 'Intraday time and order execution quality dimensions', *Journal of Financial Markets* **12**(2), 203–228.

Gervais, S., Kaniel, R. & Mingelgrin, D. (2001), 'The high-volume return premium.', *The Journal of Finance* **56**, 887–919.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. & Brilliant, L. (2009), 'Detecting influenza epidemics using search engine query data.', *Nature* **457**, 1012–1014.

Goddard, J., Kita, A. & Wang, Q. (2015), 'Investor attention and fx market volatility.', *Journal of International Financial Markets, Institutions and Money* **38**, 79–96.

Grullon, G., Kanatas, G. & Weston, J. (2003), 'Advertising, breadth of ownership, and liquidity.', *Rev. Financ. Stud.* **17**, 439–461.

Gwilym, O. A., Kita, A. & Wang, Q. (2014), 'Speculate against speculative demand.', *International Review of Financial Analysis* **34**, 212–221.

Hagströmer, B. (2021), 'Bias in the effective bid-ask spread', *Journal of Financial Economics* .

Hirshleifer, D. & Shumway, T. (2003), 'Good day sunshine: Stock returns and the weather', *The Journal of Finance* **58**(3), 1009–1032.

Hou, K., Xiong, W. & Peng, L. (2009), 'A tale of two anomalies: The implications of investor attention for price and earnings momentum.', *SSRN Journal.* .

Hu, D., Jones, C. M., Zhang, V. & Zhang, X. (2021), 'The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery', *Available at SSRN 3807655* .

Hu, H., Tang, L., Zhang, S. & Wang, H. (2018), 'Predicting the direction of stock markets using optimized neural networks with google trends', *Neurocomputing* **285**, 188–195.

Jiang, W. (2016), 'Stock market valuation using internet search volumes: Us-china comparison', *Portfolio and Security Analysis Commons* .

Joseph, K., Babajide Wintoki, M. & Zhang (2010), 'Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search.', *International Journal of Forecasting* **27**, 1116–1127.

Kenneth R. French (2021). https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html [Accessed: april 1, 2021].

Keynes, J. M. (1936), 'The general theory of employment, interest, and money', *Macmillan Publishers* .

Kim, D. & Lee, B. J. (2019), 'Shorting costs and profitability of long–short strategies', *Available at SSRN 3519706* .

Kim, N., Lučivjanská, K., Molnár, P. & Villa, R. (2019), 'Google searches and stock market activity: Evidence from norway.', *Finance Research Letters* **28**, 208–220.

Lazer, R., Lev, B. & Livnat, J. (2001), 'Internet traffic and portfolio returns', *Financial Analysts Journal* **57**(3), 30–40.

Lou, D. (2014), 'Attracting investor attention through advertising.', *Rev. Financ. Stud.* **27**, 1797–1829.

Lyocsa, S. & Molnar, P. (2020), 'Stock market oscillations during the corona crash: The role of fear and uncertainty', *Finance Research Letters* **36**.

Molnar, P. & Basta, M. (2017), 'Google searches and gasoline prices', *14th International Conference on the European Energy Market* .

Nadler, J. & Cicilline, D. N. (2020), 'Investingation of competition in digital markets'. https://fm.cnbc.com/applications/cnbc.com/resources/editorialfiles/2020/10/06/investigation_of_competition_in_digital_markets_majority_staff_report_and_recommendations.pdf [Accessed: December 12, 2020].

Nazemi, A. & Fabozzi, F. J. (2018), 'Macroeconomic variable selection for creditor recovery rates', *Journal of Banking & Finance* **89**, 14–25.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Preis, T., Moat, H. & Stanley, H. (2013), 'Quantifying trading behavior in financial markets using google trends.', *Scientific Reports* **3**.

Preis, T., Reith, D. & Stanley, H. (2010), 'Complex dynamics of our economic life on different scales: insights from search engine query data.', *Phil. Trans. R. Soc. A.* **368**, 5707–5719.

Rajgopal, S., Kotha, S. & Venkatachalam, M. (2000), 'The relevance of web traffic for stock prices of internet firms', *Available at SSRN 207989* .

Seasholes, M. S. & Wu, G. (2007), 'Predictable behavior, profits, and attention', *Journal of Empirical Finance* **14**(5), 590–610.

Sibley, S., Wang, Y., Xing, Y. & Zhang, X. (2016), 'The information content of the sentiment index.', *Journal of Banking Finance* **62**, 164–179.

Stambaugh, R., Yu, J. & Yuan, Y. (2012), 'The short of it: Investor sentiment and anomalies.', *Journal of Financial Economics* **104**, 288–302.

Sul, H. K., Dennis, A. R. & Yuan, L. (2017), 'Trading on twitter: Using social media sentiment to predict stock returns', *Decision Sciences* **48**(3), 454–488.

Tantaopas, P., Padungsaksawasdi, C. & Treepongkaruna, S. (2016), 'Attention effect via internet search intensity in asia-pacific stock markets', *Pac.-Basin Finance J.* **38**, 107–124.

Tian, S., Yu, Y. & Guo, H. (2015), 'Variable selection and corporate bankruptcy forecasts', *Journal of Banking & Finance* **52**, 89–100.

Vaughan, L. & Chen, Y. (2015), 'Data mining from web search queries: A comparison of google trends and baidu index', *Journal of the Association for Information Science and Technology* **66**(1), 13–22.

Vlastakis, N. & Markellos, R. (2012), 'Information demand and stock market volatility.', *Journal of Banking Finance* **36**, 1808–1821.

Vosen, S. & Schmidt, T. (2011), 'Forecasting private consumption: survey-based indicators vs. google trends', *Journal of forecasting* **30**(6), 565–578.

Vozlyublennaia, N. (2014), 'Investor attention, index performance, and return predictability', *Journal of Banking & Finance* **41**, 17–35.

Yuan, Y. (2012), 'Market-wide attention, trading, and stock returns', *SSRN Electronic Journal* **116**(3).

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the royal statistical society: series B (statistical methodology)* **67**(2), 301–320.

Appendix

A Model parameters

All models use the implementation of the python library Sklearn (v. 0.24.2). Here we specify parameters that are changed from the default parameters of Sklearn. For OLS and Random Forest, only default parameters are used.

A.1 Ridge regression

For Ridge, the only parameter changed is the regularization penalty.

Regularization penalty, λ	0.01

Table 15: Ridge non-default parameters

A.2 Lasso

For Lasso, the important non-default parameter is the regularization penalty. The remaining non-default parameters were changed to achieve quicker convergence.

Regularization penalty, λ	0.0001
Tolerance	0.0001
Maximum iterations	100 000
Precompute Gram matrix	True
Feature selection	'Random'

Table 16: Lasso non-default parameters

A.3 Elastic Net

For Elastic Net, the important non-default parameters are the regularization penalty and the L1-weight ratio. The remaining non-default parameters were changed to achieve quicker convergence.

Regularization penalty, λ	0.01
L1-weight ratio	0.2
Tolerance	0.001
Maximum iterations	10 000
Precompute Gram matrix	True
Feature selection	'Random'

Table 17: Elastic Net non-default parameters

B Industry statistics

The following table shows how many weeks the different industry indices are included in our portfolio using the Elastic Net model on the US market, for the full sample period.

Table 18: Industry overview.

Industry	**Weeks in long portfolio**	**Weeks in short portfolio**
Internet & Direct Marketing Retail	183	102
Real Estate Management & Development	153	156
Construction Materials	129	128
Wireless Telecommunication Services	126	144
Airlines	124	119
Technology Hardware, Storage & Peripherals	116	76
Energy Equipment & Services	113	153
Interactive Media & Services	107	74
Biotechnology	105	83
Equity Real Estate Investment Trusts (REITs)	104	116
Independent Power and Renewable Electricity Producers	103	132
Marine	100	97
Automobiles	98	108
Water Utilities	95	87

Industry	**Weeks in long portfolio**	**Weeks in short portfolio**
Metals & Mining	92	132
Health Care Technology	91	91
Construction & Engineering	83	88
Consumer Finance	80	76
Semiconductors & Semiconductor Equipment	78	63
Diversified Financial Services	78	76
Personal Products	77	82
Tobacco	76	71
Paper & Forest Products	75	75
Banks	73	70
Multiline Retail	69	93
Road & Rail	68	51
Household Durables	67	59
Building Products	66	51
Software	61	22
Diversified Consumer Services	61	87
Leisure Products	58	75
Household Products	57	54
Distributors	56	49
Diversified Telecommunication Services	55	65
Trading Companies & Distributors	54	45
Entertainment	54	37
Multi-Utilities	53	46
Life Sciences Tools & Services	53	39
Oil, Gas & Consumable Fuels	51	67
Electric Utilities	48	73
Gas Utilities	48	45
Thrifts & Mortgage Finance	43	90

Industry	Weeks in long portfolio	Weeks in short portfolio
Capital Markets	43	54
Health Care Providers & Services	41	47
Beverages	40	27
Communications Equipment	35	57
Specialty Retail	32	25
Food & Staples Retailing	32	39
Industrial Conglomerates	30	39
Pharmaceuticals	29	36
Aerospace & Defense	29	17
Food Products	29	29
Electrical Equipment	26	26
Hotels, Restaurants & Leisure	22	17
Containers & Packaging	21	20
Machinery	20	19
Media	16	14
Health Care Equipment & Supplies	15	21
Chemicals	12	10
Insurance	7	8
IT Services	7	6
Commercial Services & Supplies	4	7

C Search categories

Table 19: Search categories.

Search category
Finance
Home & Garden
Business & Industrial
Shopping
Books & Literature
Advertising & Marketing
Office Services
Real Estate
Computer Hardware
Software
Movies
Music & Audio
Banking
Insurance
Beauty & Fitness
Health
Autos & Vehicles
Construction & Maintenance
Manufacturing
Transportation & Logistics
Travel
Apparel
Food & Drink
Mass Merchants & Department Stores
Education
Enterprise Technology
Consumer Electronics
Environmental Issues
Search Engine Optimization & Marketing
Vehicle Parts & Accessories
Fitness
Office Supplies
Real Estate Agencies
ISPs
Home Storage & Shelving

Search category
Investing
Grocery & Food Retailers
Tobacco Products
Clothing Accessories
Homemaking & Interior Decor
Vehicle Maintenance
Face & Body Care
Hair Care
Cosmetology & Beauty Professionals
Human Resources
Home Improvement
Hotels & Accommodations
Fashion & Style
Air Travel
Car Rental & Taxi Services
Cruises & Charters
Radio
TV & Video Equipment
Energy & Utilities
Weight Loss
Cosmetic Surgery
Vision Care
Pharmacy
Health Insurance
Medical Devices & Equipment
Health Foundations & Medical Research
Health Education & Medical Training
Pharmaceuticals & Biotech
Medical Facilities & Services
Sporting Goods
Gardening & Landscaping
Home Furnishings
Home Appliances
Photo & Video Sharing
Restaurants
Alcoholic Beverages
Credit & Lending
Industrial Materials & Equipment
Chemicals Industry
Freight & Trucking
Packaging
Weddings

Search category
Desktop Computers
Laptops & Notebooks
Computer Peripherals
Document & Printing Services
E-Commerce Services
Customer Relationship Management (CRM)
Data Management
Networking Equipment
Import & Export
Aerospace & Defense
TV Shows & Programs
TV Networks & Stations
Audio Equipment
Gadgets & Portable Electronics
Apartments & Residential Rentals
Service Providers
Phone Service Providers
Communications Equipment
Mobile Phones
Beer
Wine
Liquor
Newspapers
Nursing
Health Conditions
Property Management
Toys
Electronics & Electrical
Mental Health
Vehicle Wheels & Tires
Ecology & Environment
Scientific Institutions
Property Inspections & Appraisals
Home Insurance
Home Financing
Auto Insurance
Auto Financing
Vehicle Shopping
Architecture
Search Engines
Computer Drives & Storage

Search category
Business & Productivity Software
Alternative & Natural Medicine
Cable & Satellite Providers
Social Networks
Undergarments
Shopping Portals & Search Engines
Textiles & Nonwovens
Events & Listings
Camera & Photo Equipment
Metals & Mining
Entertainment Industry
Online Media
Retirement & Pension
Food Production
Aging & Geriatrics
Men's Health
Pediatrics
Drugs & Medications
Women's Health
Building Materials & Supplies
Civil Engineering
Construction Consulting & Contracting
Renewable & Alternative Energy
Electricity
Oil & Gas
Waste Management
Factory Automation
Aviation
Distribution & Logistics
Maritime Transport
Rail Transport
Space Technology
Defense Industry
Agrochemicals
Cleaning Agents
Coatings & Adhesives
Dyes & Pigments
Plastics & Polymers
Property Development
Luxury Goods
Footwear
Public Storage

Search category
Bus & Rail
Computer Components
Fire & Security Services
Computer Servers
Electronic Components
Electromechanical Devices
Power Supplies
Test & Measurement
Aquaculture
Agricultural Equipment
Crops & Seed
Forestry
Livestock
Business News
Technology News
Educational Software
Credit Cards
Currencies & Foreign Exchange
Vehicle Brands
Doors & Windows
HVAC & Climate Control
Generators
Heavy Machinery
Retail Trade
Social Network Apps & Add-Ons
Animal Products & Services
Game Systems & Consoles
Coffee & Tea
Fast Food
Poker & Casino Games
Public Health
Bed & Bath
Cleaning Supplies & Services
Construction & Power Tools
Kitchen & Dining
Yard & Patio
Nuclear Energy
Hospitality Industry
Food Service
Legal Services
Photo & Image Sharing
Water Supply & Treatment

Search category
Video Sharing
Athletic Apparel
Casual Apparel
Formal Wear
Men's Clothing
Outerwear
Luggage & Travel Accessories
Specialty Travel
Real Estate Listings
Timeshares & Vacation Properties
Recording Industry
Film & TV Industry
Boats & Watercraft
Mail & Package Delivery
Consulting
Financial Markets
Economy News
Mobile & Wireless Accessories
Printing & Publishing
Commercial & Investment Real Estate
Radio Equipment
Car Electronics
Automotive Industry
Bicycles & Accessories
Electronic Accessories
Company News
Journalism & News Industry
Commercial Vehicles
Cargo Trucks & Trailers
Engine & Transmission
Auto Exterior
Auto Interior
Eyeglasses & Contacts
Apparel Services
Airport Parking & Transportation
Health News
Climate Change & Global Warming
Health Policy
Vehicle Specs, Reviews & Comparisons
Fuel Economy & Gas Prices
Vehicle Fuels & Lubricants
Live Sporting Events

NTNU

Norwegian University of
Science and Technology