VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
MODELLING AND DATA ANALYSIS
MASTER'S STUDY PROGRAMME

DIRECTION-OF-CHANGE FORECASTS OF EXCHANGE TRADED FUND RETURNS

Master's thesis

Author: Simas Vencius
VU email address: simas.vencius@mif.stud.vu.lt
Supervisor: Prof. Habil. Dr. Remigijus Leipus

Vilnius

2022

Biržoje prekiaujamų fondų grąžos krypties prognozė

Santrauka

Bendrai sutariama, kad tiksliai prognozuoti vertybinių popierių grąžos lygi yra beveik neįmanoma. Todėl šiame darbe nagrinėsime ne grąžos lygį, bet grąžos kryptį. Darbo tikslas - atlikti biržoje prekiaujamų fondų grąžos krypties prognozę, bei rasti metodą, kuris yra tiksliausias atliekant prognozę. Tikslui įgyvendinti naudojami keletas skirtingų klasifikacinių metodų: logistinė regresija, atraminių vektorių klasifikatorius ir atsitiktinių medžių klasifikatorius. Papildomai yra atliekamas metodų kombinavimas sukuriant ansamblinį modelį.

Darbe analizuojamos 141 biržoje prekiaujamo fondo dieninės grąžos. Atliekant modeliavimą naudojami skirtingo tipo nepriklausomi kintamieji: techniniai indikatoriai, finansiniai indikatoriai ir momento matas. Modeliavimas atliekamas naudojant mokymosi imtį, kuri prasideda 2005 metų gegužės 3 dieną ir baigiasi 2014 metų gruodžio 31 dieną. Modelių palyginimai atliekami naudojant testavimo imtį, kuri prasideda 2015 metų sausio 1 dieną ir baigiasi 2017 metų lapkričio 9 dieną. Modelių palyginimui naudojami Diebold-Mariano, DeLong ir Pesaran ir Timmermann testai. Papildomai, modeliai yra palyginami su optimistine ir pesimistine prognozėmis.

Atliktus skaičiavimus, tiksliausia prognozė yra gauta naudojant atsitiktinių medžių klasifikatorių, kurio bendras vidutinis tikslumas siekia 52.2%. Iš to išplaukia išvada, kad dieninių grąžų kryptis yra prognozuotina. Tą patvirtina ir statistiniai testai rodantys, kad kai kuriems fondams gaunama prognozė testavimo imtyje yra statistiškai reikšminga.

Raktiniai žodžiai : Kryptinis prognozavimas, biržoje prekiaujami fondai, atraminių vektorių klasifikatorius, atsitiktinių medžių klasifikatorius, logistinė regresija, ansamblinis modelis.

Direction-of-change forecasts of exchange traded fund returns

Abstract

It is commonly agreed that the level of financial asset returns is hardly predictable. Hence, in this thesis, instead of focusing on the level, we explore the direction of the return. Therefore, the aim of this thesis is to perform direction-of-change forecasts of exchange traded fund returns and find the method that produces the most accurate forecast results in an out-of-sample environment. For that purpose, we use several classification methods: logistic regression, support vector machines, and random forests. Additionally, a combination of several classification models is considered by constructing ensemble models.

In this thesis, the daily returns of 141 ETFs are considered. For modeling purposes, several types of independent variables are considered: technical indicators, financial market indicators, and measures of moments. Modeling is performed on a train sample ranging from 3 March, 2005, to 31 December, 2014. Model comparison is performed on a test sample that ranges from 1 January, 2015 to 9 November, 2017. For model comparison, Diebold-Mariano, DeLong, and Pesaran and Timmermann tests are used. Additionally, models are compared against benchmarks: optimistic and pessimistic forecasts.

According to the empirical calculations, the following conclusions were made. The most accurate out-of-sample forecasting results are obtained with the random forests method when the overall average accuracy is 52.2%. That implies that the direction of the daily returns is to some degree predictable and based on the statistical tests performed it was shown that for some ETFs prediction is statistically significant in out-of-sample environment.

Key words : Directional predictability, exchange traded funds, support vector machines, random forests, logistic regression, ensemble model.

Content

1 Introduction

There is no doubt that asset return prediction has vital importance for practitioners in terms of constructing profitable portfolios. In the literature related to financial time series forecast, some work has already been performed examining asset return prediction. In some papers (see (1)) results indicate that the level of asset returns is predictable to some extent. On the other hand, there are researches proving that the obtained predictability for asset return levels is usually based on misleadingly defined statistical tests (4). These conclusions are in line with the EMH (efficient market hypothesis), which states that asset price reflect all publicly available information. That means in order to predict the level of asset return, we have to account for all publicly available information about a specific asset, which is technically impossible. In addition to this, Eugene Fama was awarded the Nobel prize for work related to asset return level prediction, where he concluded that it is hardly predictable. Many arguments were provided in favor of the fact that asset return levels are unpredictable, but what about returns direction? To answer this question, we first have to define the asset return direction. Therefore, the following decomposition is introduced:

$$r_t = sign(r_t) \cdot |r_t|, \tag{1}$$

where r_t is log return of the asset, $sign(r_t) = 1$ if $r_t > 0$, and $sign(r_t) = -1$ if $r_t \leq 0$. In this case, we can disregard the level of the return and instead focus on the direction of the return, denoted by $sign(r_t)$. Following this logic, many authors (10; 11; 16) have managed to obtain significant directional predictability[1] evidences. However, while reviewing the articles, several points of improvement were noticed: arbitrary single asset selection, percentages of positive and negative days in the test sample comparison against model out-of-sample sensitivity and specificity measures, potential overestimation when using non-parametric methods, and exclusive focus only on stock or index type of assets. That is why in this thesis the following amendments are introduced:

1. Instead of focusing on a stocks or index, we will work with ETFs[2] (exchange traded funds) returns, which did not get too much attention in terms of directional predictability.

2. Instead of focusing on a single arbitrary selected asset, we will perform the analysis on 141 different ETFs to see if significant directional predictability evidences can be obtained and generalized for ETFs as an asset class.

3. We introduce sensitivity and specificity comparisons against actual data composition rates in the test sample. Such comparison is not present in the majority of the reviewed articles. However, it is important to make sure that the model has not only good overall accuracy but is also able to classify both categories better than the actual class percentage in the data.

4. Independent variable combinations are selected using logistic regression since non-parametric methods are likely to find too many non-linear relationships in the data, which, in the end, causes overestimation. For example, in the research (11) authors managed to obtain 70%-75% accuracy, which indicates potential.

[1] Directional predictability is the predictability of the sign of r_t.

[2] An ETF is a type of security that tracks an index, sector, commodity, or other asset that can be purchased or sold on a stock exchange the same way as a regular stocks. Thus, the price of an ETF's shares will change throughout the trading day as the shares are bought and sold on the market. This is unlike mutual funds, which are not traded on an exchange and are trade only once per day after the market closes. To add, mutual funds are more actively managed, meaning a fund manager makes decisions about how to allocate assets in the fund, whereas ETFs are usually passively managed and can be structured to track anything from the price of an individual commodity to a large and diverse collection of securities or a particular market index.

As implied by EMH, asset price reflect all asset related public information, which can be economical, financial, political, etc. Such a wide variety of dimensions naturally impose non-linear complexity, which requires non-linear methods to be applied. Based on the articles reviewed, we selected non-linear classification methods that proved to be useful. These are support vector machines and random forests. In addition, we consider logistic regression for comparability purposes against non-linear methods and due to its unexpected success in predicting daily stock returns direction in the recent research (2). Thus, the aim of the thesis is to:

1. Investigate if significant directional predictability can be obtained for ETFs daily returns.
2. Find the most accurate classification method for ETFs directional predictability when considering the three methods mentioned.

In this thesis, we explore 141 ETFs traded on the New York Stock Exchange (NYSE) market. Daily data used ranges from 2005 to 2017 and is divided into train and test samples.

Using logistic regression and a stepwise selection process, the best sets of explanatory variables are identified for each ETF in the train sample. All identified sets of variables are applied to each ETF with all three classification methods. For each classification method, the best set of explanatory variables is identified by examining the predictive accuracy on the test sample.

Subsequently, the directional predictability of the three selected best models is compared by employing standardized statistical tests for classification models: Diebold-Mariano and DeLong tests. In addition, sensitivity and specificity measures are compared against actual data composition rates in the test sample. When the best model is selected, its performance is compared against benchmark models, and its directional predictability significance is evaluated with the Pesaran and Timmermann test. Finally, an ensemble model is created, combining all three classification methods with predefined weights. Ensemble model out-of-sample results are compared against the selected best model.

This thesis is structured as follows. In the Section 2 relevant literature review is performed to build a strong background on the methodologies used in the field. Section 3 then provides an overview of the data used in this thesis. Then, in the Section 4 methodology used for building, selecting, and evaluating models is introduced. Afterwards, in the Section 5 empirical results obtained using detailed methodology are presented and discussed. Finally, in the Section 6 conclusions supported by the results are listed.

2 Literature review

In the article (2) directional predictability of daily stock returns was investigated. In order to model returns direction the authors used various statistical classification techniques, such as logistic regression, generalized additive models, neural networks, support vector machines, random forests, and boosted classification trees. Analysis was performed using 30 stocks that were part of the Dow Jones Industrial Average in 1996. To perform the modeling, quite an extensive set of explanatory variables was considered. This set covers measures of moments of the returns distributions, financial market indicators, risk aversion indicators, yield curve measures and technical indicators. For the model's development purposes, data from 1996 to 2003 was used. For model selection purposes, data ranging from 2004 to 2017 was used. When selecting a model, stepwise forward selection was applied. The authors generated a sequence of models by iterating a procedure that started with the empty model in the first step and sequentially added variables to the model until the full set of regressors was used. In this process, for each generated model, the average out-of-sample hit-rate (OOSH) was calculated. This method is also known as last block cross-validation. Then, the model that generates the greatest improvement was selected. Applying the mentioned techniques with the described explanatory variables, it was found that the direction of daily stock return is predictable to an extent that is statistically significant and trading strategies based on these forecasts generate positive return. In terms of prediction accuracy, it was found that logistic regression significantly outperformed other methods.

In the research (5) same problem was considered: the authors forecasted the direction of the returns. In this case, the authors used logistic regression. To perform the logistic regression as explanatory variables, the authors selected expected returns and expected volatility variables. The mentioned variables were obtained by applying the GARCH model. To conduct the empirical calculations daily S&P 500 index data from January 1, 1963 through December 31, 2003 was considered at horizons ranging from h=1 (one day) through h=250 (one year). Each day, the authors computed an out-of-sample one-day through 250-day return direction probability forecast using five-year rolling estimation windows and different logit models for each horizon to allow expected returns to change over time and horizon. Forecasts were performed for daily, weekly, monthly, quarterly, semiannual, and annual returns. According to the findings, direction forecastability appeared to be strongest at intermediate horizons of two or three months.

In the article (6) a bit extended, but still similar methodology was applied as in the paper (5). In this article, the authors considered one-, two-, and three-month returns of the MSCI[3] index for Hong Kong, UK and US. The data ranges from January, 1980 to June, 2004 . Data from January, 1980 to December 1993 was used as the starting estimation sample, which was recursively expanded as more data became available. Meaning, out-of-sample one-step-ahead forecasts were generated for the period from January, 1994 to June, 2004 recursively updating parameters. Same as in the article (5) authors used logistic regression to forecast return sign. As explanatory variables returns, volatility, skewness, and excess kurtosis were selected. In this case, the volatility forecast was obtained by using ARMA (auto regressive moving average) model which was selected by minimizing AIC (Akaike Information Criterion). To compare the results, first of all, a baseline forecast was generated using the cumulative distribution function of the r_t (series of returns). Going further, two different forecasts were obtained. First, by modelling return direction with a linear relationship between the return mean and volatility (non-parametric). Second, modeling return direction by including skewness and excess kurtosis and allowing interaction between volatility and higher-

[3]MSCI is an acronym for Morgan Stanley Capital International. It is an investment research firm that provides stock indexes, portfolio risk and performance analytics, and governance tools to institutional investors and hedge funds.

ordered conditional moments (extended). To evaluate the results, Brier[4] score was used. Results were evaluated on low, medium and high volatility periods. In high and medium volatility periods, the overall baseline method performed better than extended and non-parametric methods. Only in the low volatility period extended and non-parametric methods outperformed baseline. In addition, in the low volatility period, the extended method outperformed the non-parametric. That proves the importance of allowing for higher-ordered conditional moments when forecasting return sign.

The same problem was analyzed in the article (15), but instead of focusing on individual stocks, the authors focused on the Nikkei 225 index, which is a Japanese stock market index. For the purpose of return direction prediction, the authors employed an Artificial Neural Network (ANN) model. In addition, to improve the prediction accuracy of the index, they optimized the ANN model using a Genetic algorithm (GA). As explanatory variables two different sets of variables were used, to compare the predictive power of different factors. Both type 1 and type 2 variables were technical indicators derived from the Nikkei 225 index. Type 1 variables were: momentum, ROC (rate of change), OSCP (price oscillator), CCI (commodity channel index) etc. Type 2 variables were certain period accumulated measures such as: average return in one, two, three, four, five days, five days moving average, PSY (ratio of the number of rising periods over the 12 day period) etc. The ANN model was applied on both types of variables separately, considering 78.6% of the data (from January 23rd, 2007 to October 18th, 2012). The remaining 21.4% of the data (from October 19th, 2012 to December 30th, 2013) was used to evaluate out-of-sample model performance. For forecast evaluation purposes, the hit ratio was calculated, which indicates overall model accuracy. With type 1 variables, 60.87% hit ratio was obtained, and with type 2 the hit ratio improved up to 81.27%.

In the article (14) direction of three different countries' globally traded indices was considered. S&P 500[5] for the United States, FTSE 100[6] for the United Kingdom, and Nikkei 225[7] for Japan were examined. The entire data set covered the period from January 1967 to December 1995, a total of 348 months of observations. The data set was divided into two periods: train period was from January 1967 to December 1990 (288 months of observations), while the test period was from January 1991 to December 1995 (60 months of observations). Two different types of models were considered: classification models and level estimation models. Classification models considered were discriminant analysis, logit, probit, and probabilistic neural network models. Level estimation models covered adaptive exponential smoothing, vector autoregression with Kalman filter, multivariate transfer function, and multilayered feedforward neural network. For the purpose of the modeling, these explanatory variables were considered: short term interest rates, long term interest rates, lagged index returns, consumer price level, and industrial production level. Once the models were developed on the train sample, their performance was tested on the test sample by calculating the hit ratio. The average hit ratio for the group of all four classification models was 61.67% whereas for the group of all four level estimation models it was 56.11%. Models were evaluated on trading strategies as well. It was observed that the classification models are able to generate higher trading profits than the level estimation models.

[4] A Brier score is a way to verify the accuracy of a probability forecast.

[5] The S&P 500 Index, or Standard & Poor's 500 Index, is a market-capitalization-weighted index of 500 leading publicly traded companies in the U.S.

[6] The FTSE 100 Index is a capitalization-weighted index of the 100 most highly capitalized companies traded on the London Stock Exchange.

[7] The Nikkei is short for Japan's Nikkei 225 Stock Average, the leading and most-respected index of Japanese stocks.

3 Data

In this thesis, we consider the daily returns of 141 ETFs. The data ranges from March 3, 2005 up to November 9, 2017. Both financial market indicators[8] and ETF daily returns data was obtained from the open data source - *kaggle.com*. Once data was obtained, it was used to calculate the response variable - direction of the daily ETF returns, and some of the explanatory variables, which are detailed in Section 3.1. It is important to note that our data includes the subprime mortgage crisis period where a significant impact on ETFs return levels is expected. However, is the effect of the crisis transmitted to the directional returns as well? To check this, we introduce the following graphs.

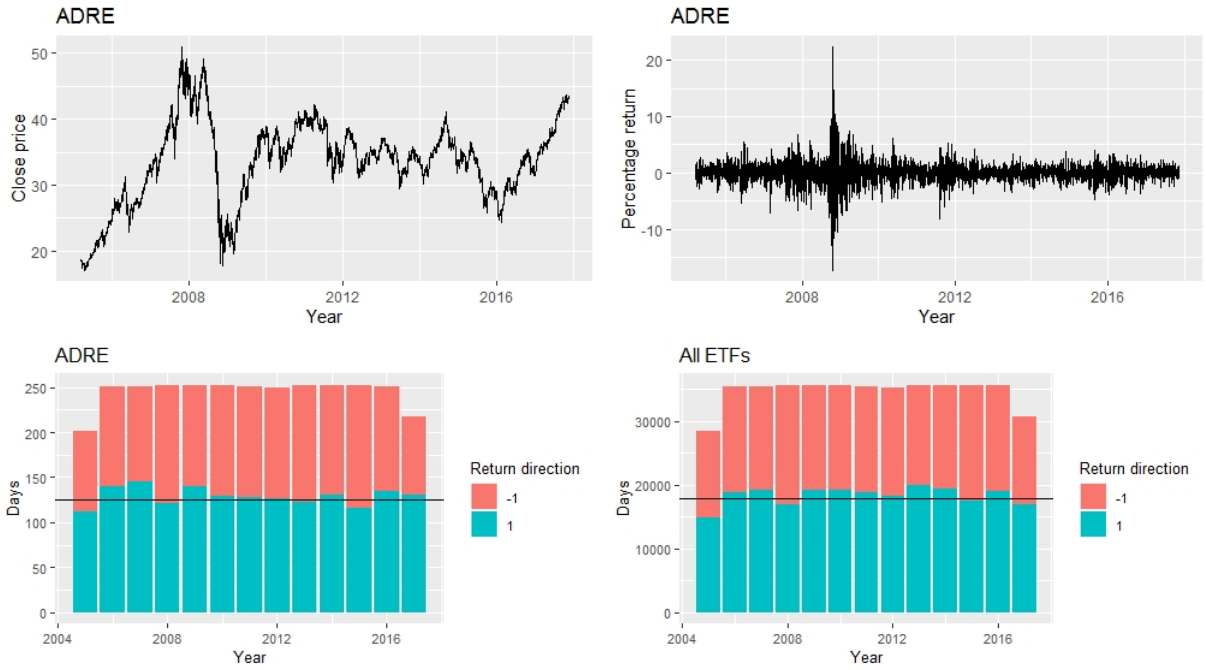

Figure 1: ADRE price, logarithmic returns, binary returns and all ETFs binary returns

In Figure 1, ADRE[9] ETF closing price, logarithmic and binary returns graphs are presented. From the price and logarithmic returns graphs, we can clearly see the crisis effect during 2008, when price sharply decreased and returns were more volatile than usual. However, if we look at the same ETF binary returns graph[10] there is no indication of the 2008 crisis effect. What is even more interesting is that binary returns are roughly equally distributed across the years. In Appendix D Figure 3 similar trend was observed for other ETFs. Going further, all 141 ETFs were merged together to check if the ratio of positive and negative returns will indicate any other insights. In Figure 1 below the right graph shows all ETFs binary returns merged. Similar trend as for ADRE ETF is observed. Binary returns are roughly equally distributed across all years. More precisely

[8] Detailed in section 3.1.

[9] ADRE, the oldest of the emerging market large-cap ETFs, launched at a time when international ETFs were just starting to take hold. As such, the fund tracks a cautious index of just ADRs—a concession that fund issuers don't tend to make anymore, opting instead for local shares.

[10] -1 indicates a negative return and 1 indicates a positive return.

when considering all ETFs it was calculated that on average there are 47.07% of negative and 52.93% of positive return days in the full data set. Going further, ratios of positive and negative returns were more deeply analyzed for each ETF separately.

To do that, binary time series were constructed for each ETF. The series consists of $\{-1, 1\}$ values where -1 indicates negative return direction and 1 indicates positive return direction. Using one sample $t - test$ it is examined whether the binary returns time series mean statistically significantly differs from zero. $t - test$ statistic is calculated as per below:

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}, \tag{2}$$

where $\bar{x}$ is the sample mean, $\mu = 0$, S - sample standard deviation and n - sample size. The hypothesis tested is:

$$\begin{cases} H_0\colon\ \bar{x} = 0, \\ H_1\colon\ \bar{x} \neq 0. \end{cases} \tag{3}$$

Test results were analyzed at the ETF level. At the 1%, 5%, and 10% levels, respectively, for 47, 25, and 16 ETFs (out of 141) we fail to reject null hypothesis and therefore, binary return series mean statistically significantly does not differ from zero. That means for certain ETFs, positive and negative return ratios statistically significantly do not differ, and we can consider the data to be balanced. For other ETFs, it differs to some extent. Thus, that led us to analyze what the ratios between positive and negative returns are for the most extreme differences. The top five ETFs with the biggest differences between positive and negative returns ratios are presented in Table 1. The biggest difference between positive and negative returns is 11% resulting in 44.5% of negative and 55.5% of positive returns. Such results indicate that for some ETFs, data is not balanced and that will have to be accounted for when selecting and interpreting different goodness of fit measures used for model comparison. On the other hand, there are no extreme differences between the two categories, and classification methods can be applied since there are enough observations in both categories.

ETF	Negative	Positive
IWP	44.49%	55.51%
SPYG	44.65%	55.35%
IUSG	44.68%	55.32%
VGT	44.68%	55.32%
IYW	44.81%	55.19%

Table 1: Top 5 biggest differences between positive and negative return ratios

3.1 Explanatory variables

In this section explanatory variables used for modeling are described. The selection of explanatory variables is based on a literature review and an economic hypothesis. Meaning variables that proved to be useful in similar researches and have a plausible economic hypothesis are considered. Variables considered are divided into 3 categories based on their nature: technical indicators, financial market indicators, and measure of moments. **Technical indicators** includes momentum

indicator, A/O oscillator, rate of change, on balance volume, 5-day moving average of ETF return, 12-day moving average of binary ETF return, intraday ETF return, and ETF return. **Financial Market Indicators** covers S&P 500 return, the level and return of VIX (volatility index), and oil return. **Measure of moment** considers high-low variance. Technical and measure of moment variables were calculated manually during the data preparation process. In Appendix A Table 8 formulas used to calculate variables are detailed. Financial market indicators were downloaded from the open data source - *kaggle.com*. For information purposes, a correlation matrix of all explanatory variables is presented in Appendix B Figure 2. It is important to note, that majority of the selected independent variables were proved to be significant predictors for stocks or index return direction prediction in the articles (2) and (15).

4 Methodology

4.1 Classification methods

In this section, we will cover methods to be used for ETFs daily returns direction prediction. Since the return direction can be positive or negative, that leaves us with a binary series as described in the decomposition (1). That means, we have to explore the classification methods that can work well with binary series. During the literature review, exceptional attention was paid to articles with similar research purposes (as in this thesis). The aim was to find out which classification methods could produce the best out-of-sample directional returns predictability results. Therefore, we mainly (but not only) focused on researches that have tested their models in out-of-sample environment. Based on the review performed, several classification methods were identified. In the article (13) the authors successfully used support vector machines (SVM) to predict the Korea composite stock index daily price direction. In the paper (2) accurate daily stock return direction forecasts were obtained using logistic regression (LR) and random forests (RF). Thus, in this thesis, we will consider those three methods. In terms of actual modeling, all explanatory variables are lagged by one day relative to the dependent variable in order to obtain one-step ahead forecast. As SVM and RF are no-linear methods, they are usually calibrated to adapt to the data very well. Therefore, overfitting is highly expected. To overcome this problem, the original data set is divided into train and test samples. Train (TR) sample ranges from 3 March, 2005 to 31 December, 2014. Test (TE) sample ranges from 1 January, 2015 to 9 November, 2017. Each model is developed on a train sample, and model performance is evaluated on a test sample to objectively compare the methods. Previously, in the data section, it was detailed what are the percentages of positive and negative returns in the full data sample. However, it is important to check the same ratios in train and test samples. These are reported in Table 2.

	Positive return	Negative return
Train	53.09%	46.91%
Test	52.37%	47.63%

Table 2: Positive and negative return ratios in test and train samples

Based on the figures reported in Table 2 we can see that both train and test samples are roughly equally distributed in terms of positive and negative returns. That is a good indication, as the models will be both developed and tested on similarly distributed data. Additionally, it is observed that in both samples, there are approximately 53% of positive and 47% of negative returns. That indicates, data is a bit unbalanced and that has to be taken into account when comparing the models. On the other hand, there are enough observations in both categories to identify different patterns between the classes, and therefore, classification methods can be applied. Going further, each non-linear[11] method (SVM and RF) is shortly described.

4.2 Support Vector Machines

Support vector machines are a family of algorithms that have been created for classification purposes. The main idea of support vector machines is to construct a hyperplane as the decision surface such that the margin of separation between positive and negative examples is maximized. Meaning, we strive to maximize the distance between the hyperplane and both classes data points.

[11] Logistic regression is omitted since it is a well-known technique.

To be even more precise, in the binary case, the distance between the hyperplane and two data points (one from each class) that are closest to the hyperplane is maximized. These data points are called "support vectors". In the case of linearly separable data, a hyperplane is just a simple line that separates both classes and maximizes the distance between the line and support vectors. However, if the data is not linearly separable, the task becomes more complex. In that case, a so-called "kernel trick" is used to make the data linearly separable. The underlying concept of the kernel trick is to transform non-linearly separable data into linearly separable. That is supported by Cover's[12] theorem, which states that given a set of training data that is not linearly separable, one can with high probability transform it into a training set that is linearly separable by projecting it into a higher-dimensional space via some non-linear transformation. Therefore, the kernel trick helps to project the data into a higher dimensional space where it becomes more easily separable. This method is also known as "generalized dot product", where the dot product of the two vectors is calculated to check how much they make an effect on each other. When using kernel trick, different kernel functions such as linear, polynomial etc. can be considered. Overall, it can be concluded that for a training set of samples, with input vectors $x_i \in R^d$ and corresponding labels $y_i \in (+1, -1)$, SVM learns how to classify objects into two classes. More detailed explanation of SVM method can be found e.g. in (7).

4.3 Random Forests

In this thesis, we employ Breiman's random forest algorithm. This algorithm is based on growing an ensemble of trees where each tree casts a vote for the most popular class given the input provided. In order to grow the ensembles, random vectors $\theta_1, ..., \theta_k$ are generated to perform the growth, where k is the number of trees. It is important to note, that each vector is independent of the past random vectors, but has the same distribution. Therefore, a k^{th} tree is produced by using the train data set and θ_k vector. This process results in the $h(x, \theta_k)$ classifier where x is an input vector. Once the number of tree classifiers is generated, each tree votes for the most popular class. The output of the random forests is generated by taking the average of the outputs produced by different decision trees. The definition of random forests from (3) is present below.

Definition 4.1 *A random forest is a classifier consisting of a collection of tree-structured classifiers* $h(x, \theta_k), k = 1, ...,$ *where* θ_k *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.*

Clearly, there are many features of random forests that can be explored and discussed further, however we will cover two that are most important in our use case. The first is that random forests ensure that the behavior of each individual tree is not too correlated with the behavior of any other tree generated in the model. That is ensured because of the bagging process and feature randomness. Decision trees are very sensitive to the data they are trained on, and bagging is a process where each individual tree is trained on a randomly sampled data set with replacements. Therefore, each bagging iteration results in a different tree. Feature randomness is a process when at each node split tree is allowed to consider only a subset of random features in contrast to traditional tree, where at each node split we consider every possible feature and pick the one that produces the most separation between the observations. Thus, the feature randomness process ensures that there is more variation between the trees, which implies lower correlation. The second useful feature of random forests is that they do not overfit as more trees are added. The proof of this statement can be found in (3). Having that in mind, we can consider generating more

[12] The theorem is named after the information theorist Thomas M. Cover who stated it in 1965.

trees to see if that improves the performance. On the other hand, the random forests method is a computationally intense process which is highly dependent on the number of trees considered. The more trees, the more intense the calculations are. Therefore, we cannot predefine too many trees given the number of models to be estimated. Taking that into account, we define 200 trees to be estimated, which is still computationally feasible given our modeling scope. Random forests theory was summarized using (3). Thus, for more details, the respective reference can be explored.

4.4 Model selection and comparison

It would be ideal to consider all possible combinations of independent variables for each ETF, but that would require evaluating $2^{13} \cdot 141$ different models[13] which is computationally too intense. Since the former procedure is not feasible, we require a strict model selection procedure to obtain the best performing set of variables. For this purpose, we use forward and backward stepwise elimination procedures. Stepwise procedures employ AIC measure to determine whether additional variables should be added or removed to or from the model. AIC for a model is calculated as per the below formula.

$$AIC = 2K - 2ln(L), \tag{4}$$

where K is the number of independent variables in the model, L is the log-likelihood of the fitted model. The interpretation of the measure is straight forward: the lower the AIC the better the model is. It is important to note that as additional variables are added to the logistic regression, L increases by default. That is why AIC measure introduces a penalty for additional variable inclusion - $2K$. $2K$ is selected as a genuine penalty since we don't want to limit the number of independent variables too much as it might lead to unused potential of machine learning algorithms[14]. However, the penalty is still important to offset the AIC improvement solely based on the additional variable inclusion, no matter that the variable itself might not add any additional explanatory power to the model. The forward stepwise selection procedure starts with an empty model (only with an intercept). Additional variables are included in the model based on AIC improvement until none of the additional variables are able to increase the AIC of the model. The backward stepwise selection procedure starts with a full model, meaning all independent variables are included. Variables are removed from the model based on AIC improvement until there is no variable to eliminate that would lead to AIC improvement. These procedures are performed using logistic regression for each ETF separately. The process results in 282 models (2 per ETF). However, some of the variable combinations in the models are identical. After removing the duplicates, the remaining models are estimated for each ETF using logistic regression, support vector machines, and random forests methods. Going further, when selecting the best models, it would be optimal to perform classical regression diagnostic tests for each model, but the majority of them would not be applicable to machine learning algorithms and it would not be feasible to manually review each model's results, having in mind the number of models considered. Therefore, the quality of the model is determined based on several goodness of fit criteria: hit rate (HR), sensitivity (SE) and specificity (SP) calculated on the test sample. These measures are detailed below.

$$HR_i = \frac{\sum_{i=1}^{TE} I(\hat{y}_{it} = y_{it})}{TE}, \tag{5}$$

[13] 13-number of independent variables, 141-number of ETFs.

[14] For example, random forests are usually likely to find relationships among more variables than logistic regression.

$$SE_i = \frac{\sum_{i=1}^{TE} I(\hat{y}_{it} = y_{it} = 1)}{\sum_{i=1}^{TE} I(y_{it} = 1)}, \tag{6}$$

$$SP_i = \frac{\sum_{i=1}^{TE} I(\hat{y}_{it} = y_{it} = 0)}{\sum_{i=1}^{TE} I(y_{it} = 0)}, \tag{7}$$

where TE is the test sample, $\hat{y}_{it}$ is the $i-th$ ETF return direction prediction for day t, y_{it} is the observed $i-th$ ETF return direction at day t. For each classification method, the best model is selected based on the SE and SP sum. Meaning the model with the highest SE and SP sum per classification method is selected as the best. This procedure results in three models, one for each classification method (LR, SVM and RF). It might be argued that selection can be solely based on HR. However, that approach would not be the best since there are many models that result in relatively "good" HR around $50\% - 53\%$ when $SE = 93\%$ and $SP = 7\%$. Which indicates that positive days are very accurately classified at the cost of very poor negative day classification. To give a more extreme example, let's consider a model that has $SE = 100\%$ and $SP = 0\%$. That means the model is classifying all days as positive, which can be done without any modeling at all. However, such model will still have HR around 50% (depending on data composition). Thus, we strive for a model that maximizes both SE and SP.

When each classification method has the best set of variables selected, we proceed to the model comparison procedure. To compare different models, we use several generalized statistical tests designed for classification methods. These are DeLong and Diebold-Mariano tests. DeLong (8) test is based on a comparison of two classification methods AUC and statistic is calculated as per below:

$$S_D = \frac{AUC_i^{(1)} - AUC_i^{(2)}}{\sqrt{Var(AUC_i^{(1)} - AUC_i^{(2)})}}, \tag{8}$$

where i is $i-th$ ETF, AUC is Area under the ROC Curve, and ROC is receiver operating characteristic curve. Numbers (1) and (2) indicate two competing models. The hypothesis tested is:

$$\begin{cases} H_0\text{: } AUC_i^{(1)} = AUC_i^{(2)}, \\ H_1\text{: } AUC_i^{(1)} > AUC_i^{(2)}. \end{cases} \tag{9}$$

Diebold-Mariano (9) test statistic is calculated as per below:

$$\begin{aligned} S_{DM} &= \frac{\bar{d}}{\sqrt{Var(d_{it})}}, \\ d_{it} &= (y_{it} - \hat{y}_{it}^{(1)})^2 - (y_{it} - \hat{y}_{it}^{(2)})^2, \\ \bar{d} &= TE^{-1} \sum_{t=1}^{TE} ((y_{it} - \hat{y}_{it}^{(1)})^2 - (y_{it} - \hat{y}_{it}^{(2)})^2), \end{aligned} \tag{10}$$

where TE is the test sample, $\hat{y}_{it}^{(1)}$ and $\hat{y}_{it}^{(2)}$ are the $i-th$ ETF return direction predictions for day t of two competing classification models, y_{it} is the observed $i-th$ ETF return direction at day t. The hypothesis tested is:

$$\begin{cases} H_0\text{: There is no difference in the accuracy of two competing forecasts,} \\ H_1\text{: Forecast (1) is more accurate than forecast (2).} \end{cases} \tag{11}$$

These tests are used to compare two different models' results obtained for the same ETF. Since we have many different ETFs, when describing the test we introduce statistic dependence on i, which indicates for which ETF statistic is calculated. That is done only for information purpose to give a better understanding on what granularity level statistic is calculated.

The above-described tests are mainly based on model overall accuracy. However, as mentioned before, it can be the case that a model has a good hit ratio by always predicting the same outcome. That is why we introduced the comparison of SE and SP against the actual percentage of positive return (PR) and negative return (NR) days in the test sample. Therefore, the following comparison is performed:

$$\begin{aligned} \overline{SE} &> \overline{PR}, \\ \overline{SP} &> \overline{NR}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \overline{SE} &= \frac{\sum_{i=1}^{N} SE_i}{N}, \\ \overline{SP} &= \frac{\sum_{i=1}^{N} SP_i}{N}, \\ PR_i &= \frac{\sum_{i=1}^{TE} I(y_{it} = 1)}{TE}, \\ NR_i &= \frac{\sum_{i=1}^{TE} I(y_{it} = 0)}{TE}, \\ \overline{PR} &= \frac{\sum_{i=1}^{N} PR_i}{N}, \\ \overline{NR} &= \frac{\sum_{i=1}^{N} NR_i}{N}, \\ \overline{HR} &= \frac{\sum_{i=1}^{N} HR_i}{N}, \end{aligned} \tag{13}$$

where SE_i and SP_i are from expressions (6) and (7), TE is the test sample, y_{it} is the observed $i-th$ ETF return direction at day t and $i = 1, ..., N$.

A comparison (12) is performed for each classification method in order to make sure that, on average across all ETFs, each method is classifying each class with a higher percentage accuracy than the actual class percentage in the test sample. Finally, by summarizing the results of all the tests, the best classification method is selected.

4.5 Ensemble model

There is a saying that two heads are better than one. It means that it is better to rely on a few expert opinions than on just one. Therefore, in this section, we will combine all three best models (one from each classification method) to perform the forecast. The combined output of the models is determined according to the below expression:

$$Ensemble = \sum_{i=1}^{m} w_i f_i, \tag{14}$$

where w_i is the weight assigned to the classification method i, f_i is the outcome of the classification method i, $i = 1, .., m$, and m is the number of methods used. In terms of weights, we experiment with all possible weight combinations given the below restrictions:

$$\begin{cases} w_1 + w_2 + w_3 = 1, \\ w_i \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8). \end{cases} \tag{15}$$

For each weight combination, $\overline{HR}$, $\overline{SE}$ and $\overline{SP}$ from expression (13) are calculated and compared against the best model. If any indications are obtained that a certain ensemble model could be better than the best single model, DeLong and Diebold-Mariano tests are used for model comparison. After model comparison (if any), we proceed with the best model based on the test results. Otherwise, ensemble models are disregarded.

4.6 Performance of the final model

When the best model is selected, disaggregated forecasting results for individual ETFs are compared against benchmark models with Diebold-Mariano test. Two benchmark models are introduced. First, is the optimist forecast (OF), which always predicts a positive return direction. Second, is the pessimistic forecast (PF), which always predicts a negative return direction. Then, each ETF result obtained with the best model is compared against OF and PF by calculating (10) statistics and testing (11) hypothesis.

In addition, the performance of the model is assessed by performing the Pesaran and Timmermann test on the ETF level. This test checks if there is any evidence of significant directional predictability. Test statistic is calculated as per below:

$$S_{PT} = \frac{\sqrt{TE} \cdot (SE_i + SP_i - 1)}{\sqrt{\frac{\overline{P}_{PR_i}(1-\overline{P}_{PR_i})}{\overline{PR}_i(1-\overline{PR}_i)}}}, \tag{16}$$

where i is $i - th$ ETF, SE_i and SP_i are from expressions (6) and (7), TE is the test sample, PR_i is from expression (13) and $\overline{P}_{PR_i} = \overline{PR}_i \cdot SE_i + (1 - \overline{PR}_i) \cdot (1 - SP_i)$. The hypothesis tested is:

$$\begin{cases} H_0\text{: No directional predictability,} \\ H_1\text{: Direction of the return is to some degree predictable.} \end{cases} \tag{17}$$

Detailed tests and comparison results are reported for the best model at the individual ETF level. Meaning, the best model for each ETF is compared against benchmarks and is examined with the Pesaran and Timmermann test.

5 Results

5.1 Model selection and comparison

As detailed in the methodology section in the first place, for each ETF logistic regression is applied. Two different best combinations of independent variables are identified for each ETF using forward and backward stepwise elimination procedures. This procedure results in $141 * 2 = 282$ models. Obviously, some of the variable combinations are identical. After removing duplicates, 122 unique models are left.

Going further, all 122 models are applied to each ETF with LR, SVM, and RF. That results in $141 * 122 = 17202$ model[15] estimations per classification method. For each model (out of 122), the average HR, SE, and SP are calculated on the test sample. Since we aim for a model that is accurate in classifying both categories, we order models based on $SE+SP$ and for each classification method, select the one that has the highest sensitivity and specificity sum. In Table 3 variables combinations selected for each classification method are detailed. For LR and SVM five, and for RF six, variables are selected. There are a few variables that are present in several models. These are A/O oscillator, ETF return and S&P return. Oil return is selected in all three classification methods. That implies financial market indicators are quite important for ETF return development, because out of only four included variables, two of them are selected multiple times. Three variables out of thirteen were not selected at all. These are momentum, intraday return, and VIX level.

Variables	LR	SVM	RF
Momentum			
A/O oscillator		+	+
Rate of change	+		
On balance volume			+
Intraday return			
5 day moving average of the ETF return			+
12 day moving average of the ETF binary return		+	
High-low variance	+		+
ETF return	+	+	
S&P 500 return	+		+
VIX level			
VIX return		+	
Oil return	+	+	+

Table 3: Classification methods and selected variables

Going further, all three selected models are compared against each other on the test sample. For comparison purposes, we use:

1. DeLong test.
2. Diebold-Mariano test.
3. SE and SP comparison against actual data composition ratios in the test sample.

Firstly, we perform model comparisons with the DeLong and Diebold-Mariano tests. In Table 4 each model comparison against other models is performed at the ETF level. For DeLong and

[15] 141-number of ETFs, 122-number of different independent variable combinations considered.

Diebold-Mariano tests, respectively, (9) and (11) hypotheses are tested. Each cell indicates how many times the forecast of the model stated in the respective row is statistically significantly more accurate than in the respective column at a 10% confidence level. According to the DeLong test, logistic regression is better than other methods in 58 cases. Support vector machines are better than other methods in only 6 cases. Random forests are in the middle, outperforming other methods in 42 cases. In terms of the Diebold-Mariano test, slightly different results are obtained. Logistic regression is still the best, outperforming other methods in 50 cases, but SVM overtook RF twice by being better in 36 cases compared to 17. However, when interpreting the results, it is important to understand the nature of the underlying test and what is being tested. From the Diebold-Mariano test statistic formula, it is clear that the test simply checks whether the hit rates of the competing models are statistically significantly different. Having that in mind, the reasons for the results obtained become more understandable. That is because logistic regression has on average the highest HR and random forests have the lowest, as stated in Table 5. The DeLong test also has a similar nature. However, as stated before, the model might have a high HR by always predicting the majority class. Meaning that model which obtains $HR = 70\%$ when data consists of 70% of positive and 30% of negative days by always forecasting positive day, could falsely be considered as good model. That is why it is important to check SE and SP measures of the models and compare them against the data composition rates.

DeLong				
	LR	SVM	RF	Sum
LR	0	41	17	58
SVM	1	0	5	6
RF	10	32	0	42
Diebold-Marian				
	LR	SVM	RF	Sum
LR	0	13	37	50
SVM	9	0	27	36
RF	7	10	0	17

Table 4: Classification methods comparison with DeLong and Diebold-Mariano tests

Going further, as stated in Section 4.4, we check if the selected models are accurate in classifying both categories. To be more precise, we want a model to classify each category with a higher percentage of accuracy than the actual category percentage in the data as stated in the expression (12). To check that each model's SE and SP measures (averaged across 141 ETFs) are compared against actual data composition rates. Comparison is performed in Table 5.

	HR	SE	SP	Positive days	Negative days
LR	53.57%	73.85%	30.84%	52.37%	47.63%
SVM	53.22%	89.10%	13.11%	52.37%	47.63%
RF	52.20%	53.14%	51.47%	52.37%	47.63%

Table 5: Classification methods comparison

From Table 5 we can see that the highest HR is equal to 53.57% and it is obtained with LR. The lowest HR is 52.2% and it is obtained with RF. When comparing SE and SP against actual

test data composition rates, we can see that both LR and SVM are very accurate in classifying positive days, and the obtained accuracy is respectively 73.85% and 89.10%. However, the same methods are poor performers when it comes to negative day classification, as the obtained accuracy is respectively 30.84% and 13.11%. It is important to mention that actual data composition rates are 52.37% of positive and 47.63% of negative returns. Thus, it is clear that both methods obtained relatively good HR by over predicting the majority class. That is important because, on average, test data has more positive than negative returns, and that makes the HR to be over 50% for both methods. Talking about the RF method, a slightly different situation is observed. $SE = 53.14\%$ and $SP = 51.47\%$ measures are obtained with the RF method. We can see that both categories are classified with a higher percentage accuracy than the actual class percentage in the data[16]. That means the model is actually able to differentiate between the classes and has a much higher discriminatory power than LR and SVM methods. Although LR and SVM have overall better accuracy than RF, the difference between HRs is far smaller than the difference between their discriminatory powers. In classification type of problems, being able to differentiate between the classes is more important than just having a high HR. Therefore, based on test results, we conclude that the best method in this experiment is RF.

5.2 Ensemble model

The next step is to experiment with ensemble models. In this chapter, we use the methodology detailed in Section 4.5. That means we combine the best models identified for each classification method in Section 5.1. The forecast is performed by the following (14) equation. The most important component in the ensemble model is the weight matrix. One way to determine the weight matrix is based on HR obtained by the model. The higher the HR, the higher the weight assigned to the model. However, we already observed, that judging models solely based on HR in some cases might be misleading. Therefore, instead of predetermining a single weight combination, we will experiment with many different weight combinations and test their performance on a test sample. It is natural that if we let $w_i \in \mathbb{R}$ it is impossible to exhaust all possible weight combinations. That is why we introduce (15) restrictions for weights. The first restriction is straight-forward, it simply requires the weights to sum up to 1. The second restriction defines the scope of the weights that will be considered. In this experiment, we will allow all possible weights with one number after the decimal point. This restriction is important since it allows us to define an exhaustive matrix of all possible weight combinations.

When an exhaustive matrix of weights is defined, we apply those weights to the best selected LR, SVM, and RF models as per (14). For each combination of weights, the average HR, SE and SP across 141 ETFs are calculated. Then we check if there are any promising results in terms of HR, SE and SP that could be better than currently obtained with the RF method. All combinations where one of the methods has a weight greater than 0.5 simply mimic that method's behavior and results. Three examples of such cases are included in Table 6 first three lines. In the first line, LR has a weight greater than 0.5, in the second SVM, and in the third RF. We can see that HR, SE and SP results in these cases are identical to the results obtained with respective single methods. Thus, such results are disregarded. In terms of cases when none of the methods has a weight greater than 0.5, the best result is printed in Table 6 last line. That ensemble model consists of LR, SVM, and RF, where the respective weights are 0.2-0.4-0.4. This ensemble model obtained the HR which is greater than the LR by 0.09%. However, SE and SP results are similar to the LR results. The model classifies positive days very accurately at the cost of poor negative

[16] 53.14%>52.37% and 51.47%>47.63%.

days classification. Thus, no promising results were obtained that would be at least as good as those obtained with the RF.

Weights LR-SVM-RF	HR	SE	SP
0.6-0.2-0.2	53.57%	73.85%	30.84%
0.1-0.6-0.3	53.22%	89.10%	13.11%
0.1-0.3-0.6	52.20%	53.14%	51.47%
0.2-0.4-0.4	53.66%	78.09%	26.39%

Table 6: Ensemble models results

5.3 Performance of the final model

At this stage, we have already selected the best three models, compared them against each other, and experimented with ensemble models. Based on the performed analysis, we conclude that the best out-of-sample forecast result was obtained with the RF method, and we selected it as the best for ETF directional daily return prediction. In this section, we will examine the RF method with additional tests on disaggregated ETFs level. Our purpose is to test for how many ETFs significant directional predictability results are obtained in an out-of-sample environment. That includes performing the Pesaran and Timmermann test and comparing the models against benchmark models with the Diebold-Mariano test.

We start with benchmark methods. A popular method in the industry to assess model forecast is to compare it against naive benchmark forecast. In our case, we will consider two benchmarks. First is the optimistic forecast (OF), which always predicts a positive return. It is important to compare our method against this benchmark since historical daily directional returns are slightly positively skewed as indicated in Table 5. Therefore, by always predicting a positive return, we should get a HR>50%. Because of that, it is important to test whether our model is statistically significantly more accurate than OF. For that purpose, we will use the Diebold-Mariano test to check the hypothesis (11). With the same philosophy, we construct pessimistic forecast (PF) and perform the same comparison as for optimistic forecast. As a last assessment, we perform the Pesaran and Timmermann test which checks if significant directional predictability is obtained with the random forest method by calculating statistic (16) and testing hypothesis (17).

Test results are presented in Table 7. This Table is ordered according HR of the random forests method. Top 10 best and worst models according HR are present. A full list with all 141 ETFs can be found in Appendix C Table 9. These tables have each ETF HR, SE, SP measures and Pesaran and Timmermann test results for the random forests method. For optimistic and pessimistic forecasts HR and Diebold-Mariano[17] test results are reported. With regards to Pesaran and Timmermann test, the null hypothesis is rejected at a 10% level in 45 cases out of 141. Meaning that for 45 ETFs, significant directional predictability is present. In terms of the Diebold-Mariano test, the null hypothesis of equal against the alternative of greater random forests method predictive accuracy is rejected at a 10% level in 16 cases for optimistic forecast and in 92 cases for pessimistic forecast. When considering HR we can see that it varies between 85% and 46%. Extreme results of HR above 70% are obtained only in four cases. But what is positively surprising is that SE and SP are above 70% in these cases as well. That indicates that extremely good HR was not obtained by always predicting the majority class and models have good discriminatory power.

[17]Tested against respective ETF random forests model.

Talking about the rest, HR is usually just a few percentage points above 50%. Additionally, it can be observed that positive returns (optimistic forecast HR) are dominating over negative returns (pessimistic forecast HR) in the test sample data. As a results, models are slightly over predicting positive returns (SE) compared to negative returns (SP). But again, what is promising is that the difference between SE and SP is not extreme in the majority of cases. That again proves that models have good discriminatory power. Based on the results obtained in this section, we can conclude that statistically significant directional predictability results are obtained and the direction of daily ETFs returns is to some degree predictable.

ETF	RF				OF		PF	
Ticker	HR	SE	SP	PT	HR	DM	HR	DM
SPYG	85.42%	83.7%	87.62%	19.01*	56.25%	-10.12*	43.75%	-21.79*
QQQ	80.69%	85.29%	75.55%	16.45*	55.69%	-7.83*	44.31%	-17.1*
SLYV	73.75%	76.9%	70.74%	12.81*	51.11%	-6.61*	48.89%	-10.46*
SOXX	73.47%	77.64%	67.73%	12.22*	56.53%	-5.85*	43.47%	-12.33*
IJK	56.73%	53.57%	60.79%	3.85*	54.37%	-0.91	45.63%	-5.67*
VUG	56.03%	62.02%	48.2%	2.77*	53.68%	-0.92	46.32%	-3.37*
VTV	54.65%	55.65%	53.3%	2.4*	51.6%	-0.99	48.4%	-2.92*
FCT	54.51%	15.82%	88.08%	1.52*	46.46%	-3.32*	53.54%	-1.13
FEZ	54.51%	43.2%	67.05%	2.83*	52.01%	-0.98	47.99%	-3.26*
VCR	54.51%	72.16%	33.33%	1.6*	53.81%	-0.36	46.19%	-3.42*
...	...	...	...	...	...	...	...	...
IUSV	48.27%	52.54%	44.41%	-0.82	49.1%	0.34	50.9%	1.05
RWR	48.13%	12.14%	87.43%	-0.18	52.57%	2.14	47.43%	-1.36*
IYR	47.99%	19.49%	82.48%	0.68	54.09%	1.95	45.91%	-2.48*
VBK	47.85%	7.71%	95.18%	1.58*	53.95%	1.9	46.05%	-2.26*
JKE	47.71%	35.37%	61.59%	-0.84	54.51%	2.53	45.49%	-1.45*
LQD	47.57%	25.32%	72.75%	-0.59	53.68%	2.02	46.32%	-0.84
IUSG	47.43%	36.05%	62.03%	-0.53	56.17%	2.37	43.83%	-2.15*
GLD	47.3%	54.67%	40.62%	-1.28	50.49%	1.08	49.51%	0.92
IYW	47.02%	26.54%	72.61%	-0.26	56.45%	2.62	43.55%	-1.69*
XLK	45.91%	21.43%	77.46%	-0.36	56.31%	3.08	43.69%	-2.32*

The symbol * indicates that the null hypothesis is rejected at the 10% level. PT - Pesaran and Timmermann test statistic for random forest method, DM - Diebold-Mariano test statistic when comparing RF against OF or PF.

Table 7: Random forests method results per ETF

6 Conclusion

As mentioned in the introduction, according to the efficient market hypothesis and relevant financial time series researches, asset return levels are generally treated as unpredictable. However, results obtained in this thesis show that the direction of daily ETFs returns can be predicted to a certain degree. Research revealed that among applied classification techniques, the best out-of-sample results were produced with a non-parametric method - random forests. On average[18] random forest model obtained $SE = 53.1\%$ and $SP = 51.5\%$ when on average test sample consists of 52.37% of positive and 47.63% of negative return days. Thus, the selected method is able to classify both categories with a higher percentage of accuracy than the actual category percentage in the data. Talking about ensemble model, none of the predefined weight combinations outperformed the random forests model.

When considering separate ETFs, the best out-of-sample result is obtained with SPYG[19]. When SE is 83.7% and SP is 87.6%, the obtained accuracy for SPYG is 85%. Such extremely good results might indicate overfitting. However, since results were obtained on test sample, which is strictly separated from the train sample, and only 4 ETFs out of 141 got such high accuracy results, we tend to treat them as outliers. For the rest, the obtained accuracy is between 45%-56% which for some of them results in significant[20] directional predictability results.

When considering the thesis conclusion among similar researches it can be concluded that significant directional predictability results obtained with the random forests method are in line with conclusions in researches (12; 17). In referenced articles, similarly, assuming non-linear dependency in asset return directions, significant directional predictability results were obtained.

Since promising results were obtained, further research could focus on accounting for additional non-linear dependence in daily ETFs returns. That includes experiments with additional variables (e.g. yield curves, measures of moments) and other machine learning classification methods (e.g. generalized additive model, boosted trees). Additionally, when testing a model on an out-of-sample data, the model could be re-estimated every step to account for the latest information[21]. Lastly, model prediction based trading strategies could be introduced to mimic the circumstances a trader would face when making model-based decisions in real time.

[18] Across all 141 ETFs.

[19] SPYG is a large-cap growth fund, holding roughly 300 companies selected from the well-known S&P 500 Index based on three growth factors: sales growth, the ratio of earnings change to price, and momentum.

[20] Confirmed with Pesaran and Timmermann test.

[21] This methodology was not employed in this research due to the large number of models examined, which as a result would increase computational intensity significantly.

References

[1] Ang, A., and Bekaert, G. Stock return predictability: Is it there? *Review of Financial Studies 20* (2001), 651–707.

[2] Becker, J., and Leschinski, C. Directional predictability of daily stock returns. Hannover Economic Papers (HEP) 624, Hannover, 2018.

[3] Breiman, L. Random forests. *Machine Learning 45* (2001), 5–32.

[4] Choi, Y., Jacewitz, S., and Park, J. A reexamination of stock return predictability. *Journal of Econometrics 192* (2016), 168–189.

[5] Christoffersen, P., and Diebold, F. Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science 52* (2006), 1273–1287.

[6] Christoffersen, P., Diebold, F., Mariano, R., Tay, A., and Tse, Y. Direction-of-change forecasts based on conditional variance, skewness and kurtosis dynamics: International evidence. *Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, PIER Working Paper Archive 1* (2006), 1–22.

[7] Cortes, C., and Vapnik, V. Support vector network. *Machine Learning 20* (1995), 273–297.

[8] DeLong, E., DeLong, D., and Clarke-Pearson, D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics 44* (1988), 837–45.

[9] Diebold, F., and Mariano, R. Comparing predictive accuracy. *Journal of Business and Economic Statistics 13* (1995), 134–144.

[10] Huang, W., Nakamori, Y., and Wang, S. Forecasting stock market movement direction with support vector machine. *Computers & OR 32* (2005), 2513–2522.

[11] Kara, Y., Boyacioglu, M., and Baykan, O. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert Systems with Applications 38* (2011), 5311–5319.

[12] Khan, M. A. Financial volatility forecasting by nonlinear support vector machine heterogeneous autoregressive model: Evidence from nikkei 225 stock index. *International Journal of Economics and Finance 3* (2011), 138–150.

[13] Kim, K.-J. Financial time series forecasting using support vector machines. *Neurocomputing 55* (2003), 307–319.

[14] Leung, M., Daouk, H., and Chen, A.-S. Forecasting stock indices: A comparison of classification and level estimation models. *International Journal of Forecasting 16* (2000), 173–190.

[15] Qiu, M., and Song, Y. Predicting the direction of stock market index movement using an optimized artificial neural network model. *PLOS ONE 11* (2016), 1–11.

[16] Senol, D., and Ozturan, M. Stock price direction prediction using artificial neural network approach: The case of Turkey. *Journal of Artificial Intelligence 3* (2010), 261–268.

[17] Tay, F., and Cao, L. Application of support vector machines in financial time series forecasting. *Omega 29* (2001), 309–317.

A Appendix

Variables	Formulas
Momentum	$C_t - C_{t-4}$
A/O oscillator	$\frac{H_t - C_{t-1}}{H_t - L_t}$
Rate of change	$\frac{C_t}{C_{t-14}} \cdot 100$
On balance volume	$OBV_{t-1} + B_t \cdot V_t$
Intraday return	$ln(\frac{C_t}{O_{t-1}}) \cdot 100$
5 day moving average of the ETF return	$\frac{\sum_{i=1}^{n} r_{t-i+1}}{5}$
12 day moving average of the ETF binary return	$\frac{\sum_{i=1}^{n} B_{t-i+1}}{12}$
High-low variance	$\frac{ln(\frac{H_t}{L_t})^2}{4 \cdot ln(2)}$
Asset return	$ln(\frac{C_t}{C_{t-1}}) \cdot 100$

Note: here C_t, H_t, L_t, O_t are respectively closing, highest, lowest and open prices at day t of the respective ETF. OBV_t is on balance volume at day t. $B_t = 1$ if $C_t > C_{t-1}$, otherwise $B_t = -1$. r_t is the logarithmic ETF return.

Table 8: Variables and formulas used

B Appendix

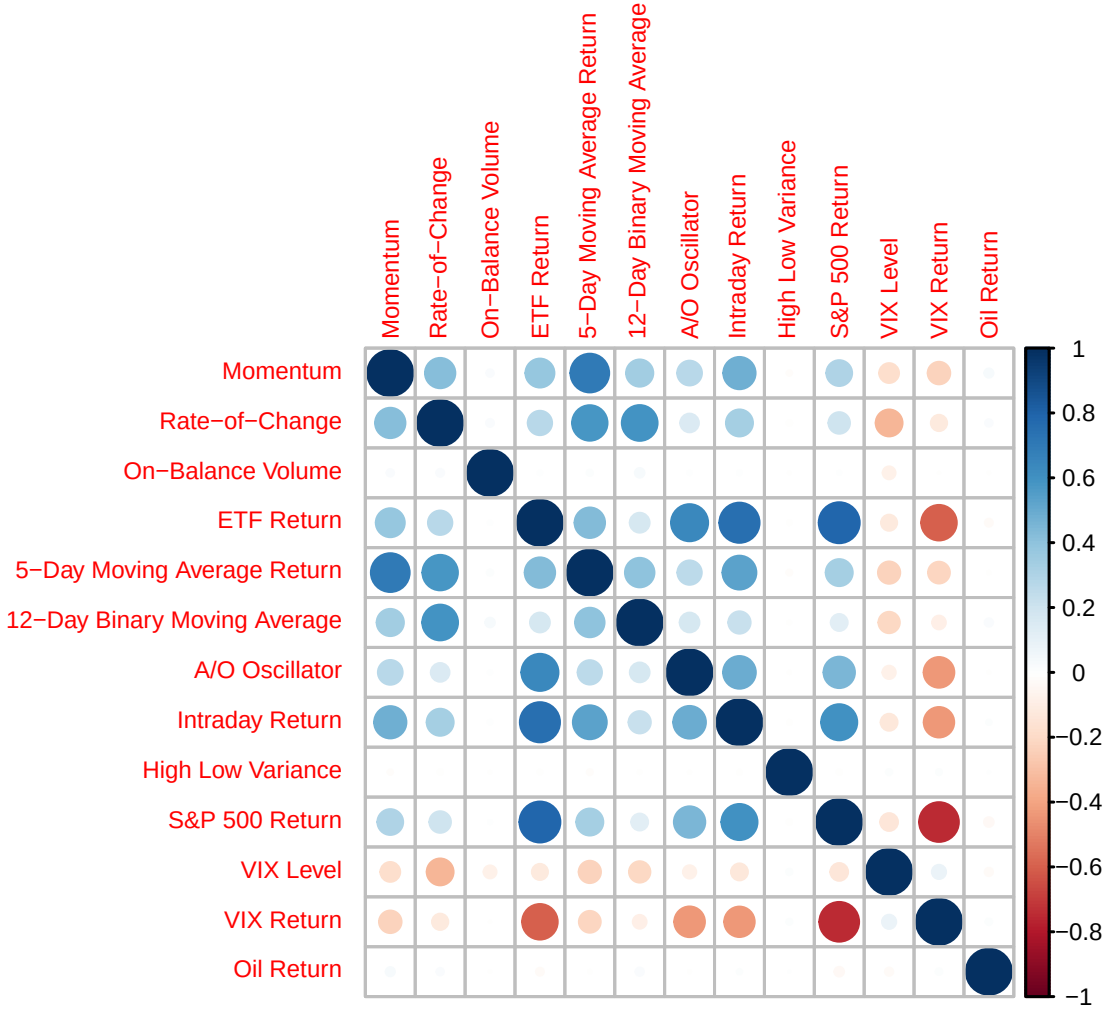

Figure 2: All explanatory variables correlation graph

C Appendix

ETF	RF				OF		PF	
Ticker	HR	SE	SP	PT	HR	DM	HR	DM
SPYG	85.42%	83.7%	87.62%	19.01*	56.25%	-10.12*	43.75%	-21.79*
QQQ	80.69%	85.29%	75.55%	16.45*	55.69%	-7.83*	44.31%	-17.1*
SLYV	73.75%	76.9%	70.74%	12.81*	51.11%	-6.61*	48.89%	-10.46*
SOXX	73.47%	77.64%	67.73%	12.22*	56.53%	-5.85*	43.47%	-12.33*
IJK	56.73%	53.57%	60.79%	3.85*	54.37%	-0.91	45.63%	-5.67*
VUG	56.03%	62.02%	48.2%	2.77*	53.68%	-0.92	46.32%	-3.37*
VTV	54.65%	55.65%	53.3%	2.4*	51.6%	-0.99	48.4%	-2.92*
FCT	54.51%	15.82%	88.08%	1.52*	46.46%	-3.32*	53.54%	-1.13
FEZ	54.51%	43.2%	67.05%	2.83*	52.01%	-0.98	47.99%	-3.26*
VCR	54.51%	72.16%	33.33%	1.6*	53.81%	-0.36	46.19%	-3.42*
EWP	54.37%	46.68%	63.37%	2.73*	52.29%	-1.53*	47.71%	-2.11*
AGG	54.23%	71.73%	36.13%	2.26*	52.01%	-3.01*	47.99%	-4.72*
IJJ	54.23%	69.7%	39.11%	2.48*	50.35%	-2.56*	49.65%	-1.89*
IEV	54.09%	56.44%	51.69%	2.18*	50.62%	-1.99*	49.38%	-1.61*
IWF	54.09%	63.29%	42.94%	1.7*	54.79%	0.39	45.21%	-3.71*
IYF	54.09%	63.81%	43.68%	2.05*	51.73%	-1.03	48.27%	-2.18*
VO	54.09%	66.07%	39.46%	1.54*	53.95%	0.07	46.05%	-2.67*
VXF	54.09%	55.67%	52.25%	2.12*	53.81%	-0.17	46.19%	-4.25*
VGT	53.81%	63.77%	39.41%	0.87	57.42%	1.43	42.58%	-3.98*
EWA	53.68%	61.56%	43.55%	1.4*	51.6%	-0.71	48.4%	-1.63*
EWH	53.68%	48.66%	58.5%	1.93*	51.87%	-0.73	48.13%	-2.35*
DVY	53.54%	50.13%	58.58%	2.34*	53.12%	-0.47	46.88%	-2.6*
IVW	53.54%	60.8%	45.51%	1.71*	55.2%	0.62	44.8%	-4.32*
IYC	53.54%	60.41%	44.58%	1.35*	53.95%	0.45	46.05%	-3.83*
EWN	53.4%	49.87%	57.49%	1.97*	53.68%	0.1	46.32%	-2.59*
IWD	53.4%	65.41%	40.74%	1.7*	51.32%	-0.97	48.68%	-2.04*
IWR	53.26%	68.59%	36.28%	1.38*	52.98%	-0.15	47.02%	-2.36*
IYH	53.26%	53.06%	53.8%	1.84*	54.37%	0.46	45.63%	-5.76*
IYZ	53.12%	45.48%	60.76%	1.7*	49.1%	-1.52*	50.9%	-1.03
SMH	53.12%	57.21%	48.4%	1.5*	56.73%	1.19	43.27%	-3.18*
VV	53.12%	68.31%	35.42%	1.06	53.4%	0.17	46.6%	-2.73*
EWG	52.98%	60.27%	44.73%	1.36*	51.32%	-0.71	48.68%	-1.28*
IWN	52.98%	64.46%	41.62%	1.68*	50.35%	-0.93	49.65%	-0.99
IYJ	52.98%	63.01%	41.95%	1.36*	54.37%	0.48	45.63%	-3.92*
JKL	52.84%	64.5%	40.91%	1.49*	51.18%	-0.74	48.82%	-1.97*
TLT	52.84%	82.17%	18.86%	0.36	53.68%	0.47	46.32%	-6.04*
XLV	52.7%	46.46%	60.29%	1.83*	52.84%	-0.07	47.16%	-3.68*
IJS	52.57%	64.42%	40.86%	1.46*	51.46%	-0.61	48.54%	-1.6*
IWV	52.57%	61.27%	43.02%	1.17	52.29%	-0.11	47.71%	-1.87*
XLY	52.57%	56.62%	47.32%	1.06	53.4%	0.54	46.6%	-3.01*
EWW	52.43%	64.29%	41.24%	1.52*	48.54%	-2.37*	51.46%	-0.48
EZU	52.43%	57.57%	46.15%	1.01	51.32%	-0.21	48.68%	-1.3*
IXC	52.43%	59.45%	44.1%	0.97	50.62%	-0.4	49.38%	-1.38*

RSP	52.43%	58.62%	45.93%	1.23	52.29%	-0.09	47.71%	-2*
VPU	52.43%	46.8%	58.48%	1.42*	54.23%	0.91	45.77%	-3.1*
EWJ	52.29%	45.12%	60.23%	1.45*	52.57%	0.13	47.43%	-2.04*
EWS	52.29%	50.14%	53.31%	0.93	49.79%	-0.76	50.21%	-0.73
IWS	52.29%	62.94%	41.24%	1.15	50.9%	-0.63	49.1%	-1.03
IYY	52.29%	55.44%	48.26%	0.99	52.29%	0.12	47.71%	-2.22*
VAW	52.29%	51.56%	52.23%	1.01	53.26%	0.66	46.74%	-1.67*
VBR	52.29%	57.45%	47.73%	1.4*	51.18%	-0.57	48.82%	-1.78*
IGN	52.16%	62.11%	41.3%	0.93	52.85%	0.28	47.15%	-3.4*
IWO	52.15%	53.52%	51.39%	1.31*	55.2%	0.93	44.8%	-4.74*
VFH	52.15%	67.02%	35.92%	0.83	51.73%	-0.16	48.27%	-1
EWQ	52.01%	45.31%	58.62%	1.06	51.73%	0	48.27%	-1.32*
IVE	52.01%	51.66%	53.2%	1.31*	50.21%	-0.72	49.79%	-1.03
IWB	52.01%	54.74%	49.27%	1.08	52.7%	0.22	47.3%	-2.26*
IYK	52.01%	43.26%	61.79%	1.38*	53.54%	0.57	46.46%	-2.74*
EWZ	51.87%	54.69%	49.14%	1.03	51.73%	-0.11	48.27%	-1.18
IOO	51.87%	55.5%	48.28%	1.01	51.73%	-0.13	48.27%	-1.55*
IXJ	51.87%	55.05%	47.54%	0.7	52.15%	0.39	47.85%	-1.85*
SHY	51.87%	65.57%	40.83%	1.77*	46.32%	-2.72*	53.68%	0.59
VDC	51.87%	50.53%	52.77%	0.88	52.43%	0.47	47.57%	-1.68*
IWM	51.73%	51.84%	51.03%	0.77	52.7%	0.68	47.3%	-1.85*
MDY	51.73%	44.3%	59.59%	1.06	52.29%	0.29	47.71%	-1.84*
XLI	51.73%	46.35%	58.16%	1.22	53.26%	0.58	46.74%	-2.19*
DIA	51.6%	52.15%	50.92%	0.82	54.79%	1.15	45.21%	-3.37*
EEM	51.6%	54.55%	49.28%	1.03	51.87%	-0.06	48.13%	-1.06
EFA	51.6%	43.28%	61.03%	1.18	51.6%	-0.11	48.4%	-1.38*
EWM	51.6%	49.58%	53.31%	0.78	49.79%	-0.64	50.21%	-0.32
IAU	51.6%	66.38%	37.33%	1.04	49.1%	-1.99*	50.9%	-0.19
IYM	51.6%	60.86%	42.24%	0.85	51.73%	-0.06	48.27%	-1.59*
IYT	51.6%	64.1%	37.68%	0.49	52.15%	0.28	47.85%	-1.81*
JKG	51.6%	56.22%	45.67%	0.51	53.54%	0.69	46.46%	-1.85*
PEY	51.6%	43.94%	58.74%	0.73	49.24%	-0.91	50.76%	-0.32
PPH	51.6%	62.09%	42.3%	1.2	50.49%	-0.81	49.51%	-1.1
IBB	51.46%	58.7%	42.86%	0.42	53.4%	1.11	46.6%	-3.09*
IGM	51.46%	53.19%	49.84%	0.81	56.59%	1.79	43.41%	-3.92*
IJH	51.46%	43.64%	60.71%	1.18	53.4%	0.68	46.6%	-3.38*
IXP	51.46%	57.54%	46.28%	1.03	49.65%	-0.79	50.35%	-0.46
IGE	51.32%	77.97%	25.89%	1.21	49.1%	-1.21	50.9%	-0.24
IJT	51.32%	55.01%	47.29%	0.62	53.95%	1.23	46.05%	-2.43*
ADRE	51.25%	51.71%	50.15%	0.5	52.92%	0.65	47.08%	-1.49*
EPP	51.18%	49.6%	54%	0.97	51.46%	-0.13	48.54%	-1.02
EWT	51.18%	51.21%	50.86%	0.56	51.73%	0.19	48.27%	-1.5*
IVV	51.18%	65.35%	35%	0.1	52.84%	0.97	47.16%	-2.05*
IYG	51.18%	51.48%	51.14%	0.7	51.46%	0.08	48.54%	-1.09
XLP	51.18%	47.48%	54.65%	0.57	52.29%	0.58	47.71%	-1.08
EWK	51.04%	58.67%	43.06%	0.47	52.01%	0.53	47.99%	-0.98
OEF	51.04%	56.23%	44.19%	0.11	52.29%	0.84	47.71%	-1.23
VTI	51.04%	61.26%	39.82%	0.3	52.98%	0.81	47.02%	-1.52*

FFA	50.9%	36.29%	66.76%	0.86	51.6%	0.22	48.4%	-1.23
FVD	50.9%	51.06%	50.72%	0.48	52.15%	0.43	47.85%	-1.78*
FXI	50.9%	57.53%	44.38%	0.52	50.62%	-0.21	49.38%	-0.62
EWI	50.76%	42.39%	58.64%	0.28	51.04%	0.22	48.96%	-0.65
ILF	50.76%	59.08%	42.61%	0.46	51.18%	0.05	48.82%	-0.74
JKF	50.76%	71.82%	28.69%	0.15	50.21%	-0.15	49.79%	-0.17
EWC	50.62%	52.97%	47.29%	0.07	51.32%	0.46	48.68%	-0.62
EWD	50.62%	44.66%	56.18%	0.23	50.62%	0.12	49.38%	-0.51
EZA	50.62%	54.08%	47.88%	0.52	51.04%	0	48.96%	-0.96
ICF	50.62%	52.48%	47.93%	0.11	53.12%	1.16	46.88%	-1.62*
VNQ	50.62%	37.31%	65.97%	0.92	53.54%	1.2	46.46%	-1.64*
EWO	50.49%	42.26%	59.71%	0.53	52.84%	0.79	47.16%	-1.21
VB	50.49%	53.16%	47.21%	0.1	52.7%	1.16	47.3%	-1.89*
MFD	50.42%	44.78%	56.74%	0.41	50.56%	-0.04	49.44%	-0.42
IXG	50.35%	48.39%	51.58%	-0.01	51.6%	0.81	48.4%	-0.52
JKD	50.35%	66.15%	32.04%	-0.52	53.68%	1.04	46.32%	-1.8*
OIH	50.35%	79.33%	26.02%	1.69*	45.63%	-2.23*	54.37%	1.42
TIP	50.35%	37.9%	64.18%	0.58	51.6%	0.4	48.4%	-0.8
VHT	50.35%	40.97%	61.59%	0.7	54.51%	1.44	45.49%	-4.12*
SPY	50.07%	61.72%	36.8%	-0.41	53.26%	1.57	46.74%	-1.33*
VDE	50.07%	81.92%	21.69%	1.21	47.57%	-1.7*	52.43%	0.67
XLU	50.07%	42.61%	58.7%	0.35	55.34%	1.9	44.66%	-1.64*
PGJ	49.93%	41.88%	59%	0.24	52.98%	0.86	47.02%	-1.03
XLB	49.93%	40.05%	60.77%	0.22	52.98%	1.25	47.02%	-1.06
BBH	49.79%	23.82%	79.65%	1.12	52.98%	0.97	47.02%	-1.76*
FAM	49.65%	53.93%	46.31%	0.06	51.18%	0.3	48.82%	-0.33
IYE	49.65%	76.32%	24.8%	0.35	47.43%	-1.39*	52.57%	1.11
IDU	49.51%	34.09%	68.01%	0.59	55.34%	2.1	44.66%	-2.25*
IJR	49.38%	36.05%	65.69%	0.49	52.7%	0.9	47.3%	-1.29*
ONEQ	49.38%	56.3%	41.14%	-0.69	56.17%	2.38	43.83%	-3.31*
XLF	49.38%	38.25%	60.28%	-0.4	50.76%	0.53	49.24%	0.08
IEF	49.24%	43.24%	56.13%	-0.17	51.32%	0.77	48.68%	-0.52
EWL	49.1%	40.22%	59.78%	0	50.35%	0.18	49.65%	-0.13
RTH	49.1%	41.48%	58.54%	0	54.51%	2.76	45.49%	-2.04*
EWY	48.96%	61.25%	36.08%	-0.74	51.18%	0.72	48.82%	-0.07
IGV	48.96%	45.97%	53.21%	-0.22	56.73%	3.36	43.27%	-2.51*
IXN	48.96%	41.36%	57.74%	-0.24	57%	2.71	43%	-3.07*
EWU	48.68%	29.87%	68.5%	-0.48	52.01%	1.38	47.99%	-0.21
XLE	48.68%	55.59%	42.47%	-0.53	48.4%	-0.16	51.6%	1.09
IWP	48.4%	44.44%	53.16%	-0.64	56.17%	3.34	43.83%	-2.52*
IUSV	48.27%	52.54%	44.41%	-0.82	49.1%	0.34	50.9%	1.05
RWR	48.13%	12.14%	87.43%	-0.18	52.57%	2.14	47.43%	-1.36*
IYR	47.99%	19.49%	82.48%	0.68	54.09%	1.95	45.91%	-2.48*
VBK	47.85%	7.71%	95.18%	1.58*	53.95%	1.9	46.05%	-2.26*
JKE	47.71%	35.37%	61.59%	-0.84	54.51%	2.53	45.49%	-1.45*
LQD	47.57%	25.32%	72.75%	-0.59	53.68%	2.02	46.32%	-0.84
IUSG	47.43%	36.05%	62.03%	-0.53	56.17%	2.37	43.83%	-2.15*
GLD	47.3%	54.67%	40.62%	-1.28	50.49%	1.08	49.51%	0.92

IYW	47.02%	26.54%	72.61%	-0.26	56.45%	2.62	43.55%	-1.69*
XLK	45.91%	21.43%	77.46%	-0.36	56.31%	3.08	43.69%	-2.32*

Table 9: Random forests method results for all ETFs

D Appendix

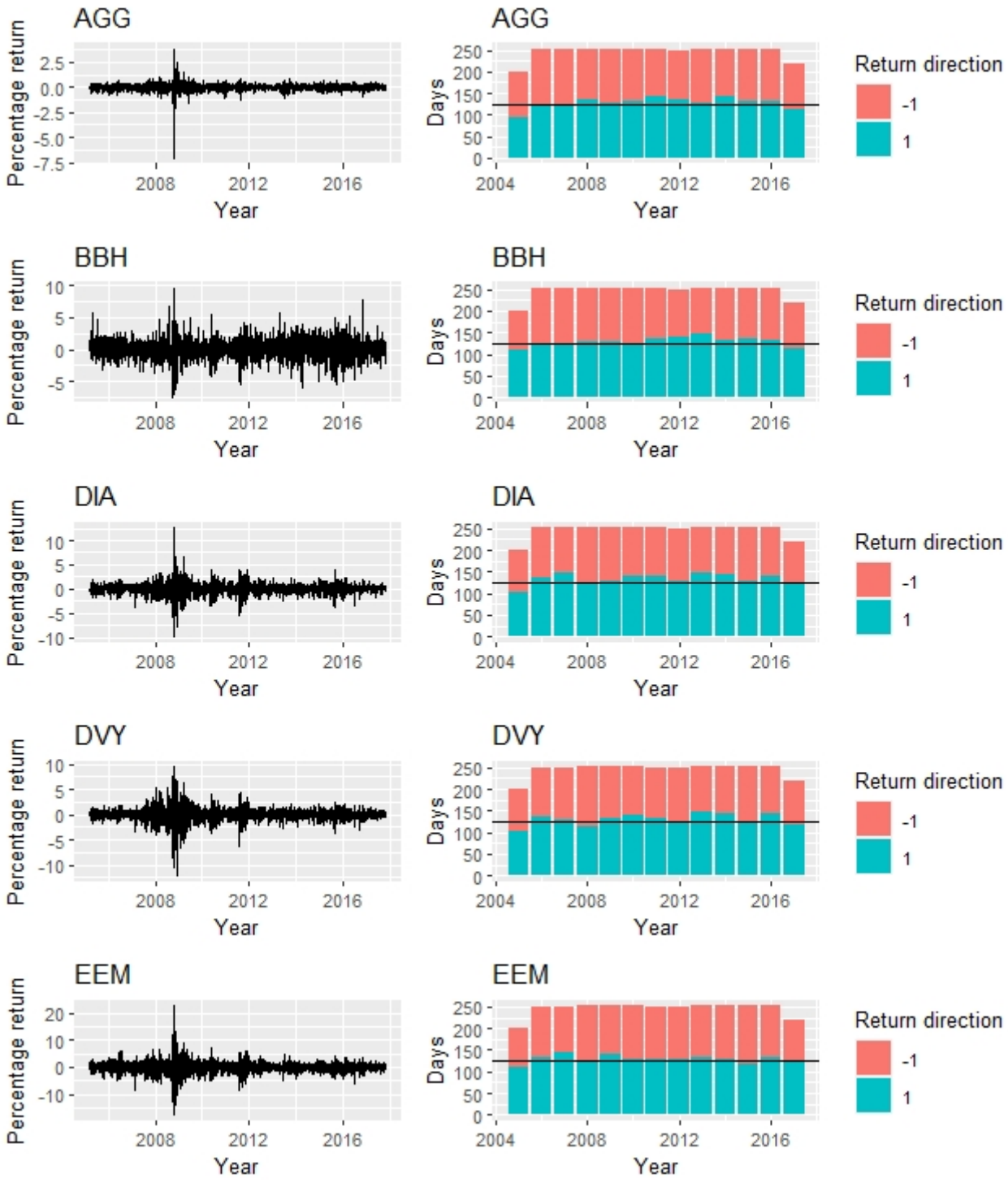

Figure 3: Percentage and binary returns of different ETFs