National College of Ireland

# Tracking Error of Passive Equity Funds. Data Analysis Using Morningstar Financial Data

MSc Research Project
Data Analytics

Stefano Leone
Student ID: 20198019

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Mr. Stefano Leone |
| **Student ID:** | 20198019 |
| **Programme:** | Data Analytics **Year:** 1 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Jorge Basilio |
| **Submission Due Date:** | 2021-09-23 |
| **Project Title:** | Tracking Error of Passive Equity Funds. Data Analysis Using Morningstar Financial Data |
| **Word Count:** | 6590 **Page Count** 19 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Stefano Leone

**Date:** 2021-09-21

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Tracking Error of Passive Equity Funds. Data Analysis Using Morningstar Financial Data

Stefano Leone
20198019

**Abstract**

This data analytics project examines the financial data publicly available on Morningstar for index mutual funds and Exchange-Traded Funds (ETFs), aiming to identify the key aspects that make them successful in replicating the return of the indices they track. The gap between fund's and index's returns is called "tracking error", a metric frequently investigated in the past using small samples of funds along with their indices, and that often resulted in contradictory conclusions regarding the impact of several fund aspects such as management fees, exit fees, number of portfolio securities, asset size, and fund age.

The innovative elements of this analysis regard the larger financial data – retrieved using the Morningstar Rest API – and the Extract, Transform, Load (ETL) process designed to scrape the funds information along with their prices and the prices related to their indices, which allow to calculate the tracking error metric that is eventually used by machine learning models for the identification of its driving factors.

# 1 Introduction

Mutual funds are open-ended investment products that allow investors to have a financial exposure on a diversified and professionally-managed portfolio. Mutual funds were introduced in the 1890 in the United States and were exclusively actively-managed, meaning that they aimed to overperform the market by leveraging the managers' ability to select the optimal securities at the right time. The introduction of index mutual funds in the 1970s was a financial revolution initially criticized as it conflicted with the common belief that active mutual funds would constantly guarantee higher returns due to the investment managers' competence. When in 1975 John Bogle started the first index mutual fund called "Vanguard 500 Index Fund", which tracked the S&P 500 Index, only few investors were interested in merely obtaining the average return, despite lower fees and volatility. However, its higher long-term performance compared to active funds started to draw a significant attention in the following years to the argumentation about the benefits of active portfolio management, which is still an open debate.

Index portfolios managed 6.72 trillion dollars in 2017, that is, approximately 35% of the total net asset under mutual fund management[1] and the index funds market share grew in pair with Exchange-Traded Funds (ETFs), a more innovative type of investment product that is traded on stock exchanges for a price that fluctuates through each dealing day and does not always match its Net Asset Value (NAV) per share, as opposed to traditional mutual funds.

ETFs were introduced in the 1990s, when State Street Global Investors launched the "S&P 500 Trust ETF" (tracking the S&P 500 Index), which quickly became popular due to

high liquidity, potential tax efficiency, transparency, and stock-like features, and is still one of the most actively traded ETFs today. From 2008 to 2019 the number of ETFs worldwide grew from 1617 to 6940, while the value of assets held in ETF products during the same period also grew from $716 billion to over $6 trillion[2], witnessing a success of passive funds that could have never been imagined when Bogle launched the "Vanguard 500 Index Fund".

Initially, all ETFs invested directly in the securities of the index they tracked (*physical* structure), but as they became more popular among investors, a new type of replication method was introduced: this marked the inception of *synthetic* ETFs, which use derivatives – contracts stipulated with counterparties – to replicate the total return of their index without directly holding any of its underlying securities. In addition, synthetic ETFs have a high exposure to counterparty risk due to their higher use of derivatives[3].

As passive mutual funds and ETFs intend to replicate the index portfolio (and therefore the index return), their efficiency should be assessed not based on their overall return but based on the tracking error (TE) – gap between fund's and index's returns.

For this reason, the current project aims to analyse the tracking error of all passive equity funds for which there is financial information and prices on Morningstar, and to identify the fund aspects that contribute the most to the tracking error.

The upcoming sections are organized in the following way:

- Related work: description of similar research papers that have inspired this study and illustrate their findings, but also their weaknesses in the approach or in their data quality and size;
- Research methodology: explanation of the chosen CRISP-DM approach and provision of some context to each of its steps;
- Design specification: illustration of how the data has been collected and manipulated before it was ready to be analysed, along with the planning of Exploratory Data Analysis and the machine learning models;
- Implementation: exploration of the cleaned data with correlation matrix and factorplots, followed by the description of the tuned prediction models;
- Evaluation: interpretation of the models' results and comparison between them;
- Conclusion and future work: culmination of the analysis into relevant information regarding the causes of the passive funds' tracking error, and final comments on potential further developments of this study.

# 2 Related Work

Passive mutual funds and ETFs have been the object of several research papers, mainly focused on small samples of funds and with a financial perspective that involved the use of financial ratios as Sharpe, Treynor, Sortino, etc[4]; this projects aims, on the other hand, to analyse a larger sample of funds and with a data analytics perspective. All the following studies represent a useful starting point to find how passive funds have been investigated in the past and see what elements seem to cause the tracking error.

Frino and Gallagher (2001) closely studied the tracking error of index funds when almost half of them (94 out of 199) were tracking the S&P 500 Index, and concluded that this measure is both positively and significantly correlated with the dividend payments[5]. The researchers

remarked that their study has a survivorship bias problem as terminated funds were excluded from the data collected from Morningstar.

Kostovetsky (2003) developed a simple one-period model to compare ETFs and index funds, discovering that the differences between them are mainly management fees, shareholder transaction fees, and taxation efficiency. The researcher also mentioned that the main reason ETFs have lower expense ratios is that they are not in charge of shareholder accounting. The task of keeping track of shareholder transactions is a large percentage of the expense ratio; for ETFs, these tasks are performed by the brokerage house of the shareholder[6]. However, the study does not measure the tracking error due to lack of true benchmarks to use for comparison.

Blitz, Huji, and Swinkels (2010) sourced the data from the Thomson Financial Datastream and the Morningstar websites for funds listed in Europe, filtering only funds that track one of the broadly diversified equity market indexes, and finding more than 40 passive funds for the study. The researchers noticed that the performance differences between funds that track different benchmark indexes cannot be explained by expense ratios, and that the performance of index funds and ETFs listed in Europe exhibited an annual tracking error that spans from 0.5% to 1.5%, much higher than the US passive funds[7].

Dhabolkar and Reddy (2019) investigated the tracking capabilities of ETFs and index funds in India using a sample of 30 funds obtained from the website of "Association of Mutual Funds in India" (AMFI) and respective fund houses, implementing a regression analysis that allowed them to notice that index mutual funds have a significantly higher tracking error than ETFs[8]. Tracking error was calculated from the fund inception date, but the study does not analyse its main causes, proposing to investigate the tracking error's factors in a future work.

As confirmation that tracking error analysis has been conducted with limited amount of financial data, Chen and Frijns (2017) concentrated their studies on regression methods and cointegration analysis of three New Zealand ETFs, using three different tracking error measures:

- The first based on the absolute differences between the returns on the ETF ($r_t^{ETF}$) and its underlying index ($r_t^{IND}$), where T is the timeframe that the tracking error is measured over:
  $$TE_1 = \frac{\sum_{t=1}^{T} |r_t^{ETF} - r_t^{IND}|}{T}$$
- The second based on the standard deviation of the price differences:
  $$TE_2 = \sqrt{\frac{\sum_{t=1}^{T} (r_t^{ETF} - r_t^{IND})^2}{T}}$$
- The third based on the standard error of the residuals of a linear regression of ETF returns on their corresponding underlying index returns:
  $r_t^{ETF} = \alpha + \beta\, r_t^{IND} + \varepsilon_t$
  where $TE_3 = S.E.(\varepsilon_t)$

Having noticed that most of the variation of the ETF is not explained by the variation in the index, they conclude that the daily tracking performance of these ETFs is mediocre, probably due to the low liquidity and the effect of market microstructure noise[9].

Singh and Kaur (2016) investigated a sample of 12 Indian ETFs, exploring the factors that affect the tracking error, and observing that volume positively affects the tracking ability,

whereas volatility and fund age do not. Tracking error has been calculated using the same three formulas exhibited in the work published by Chen and Frijns, and the expense ratio has been found to be non-significant, which is in contrast with the majority of previous researches[10].

Wong and Shum (2010) opted for a different approach to analyse ETFs, considering the performances of 15 ETFs from 7 countries during bullish and bearish markets and finding a higher volatility in bullish markets. The researchers claimed that no other work has examined before the fund performances under different market conditions, and their study included the analysis of daily prices from 1999 and 2007 and financial data retrieved from Thomson Datastream and Yahoo! Finance[11].

Waller, Nanigan, and Finke (2018) focused their research on the impact of redemption fees (also called *exit* fees) to the overall U.S. fund returns, finding that the ones with redemption fees outperform their counterparts by 1.0% to 1.4% a year, likely due to changes in portfolio composition rather than a preference for exit fees by high-quality managers. Another insight found in the research paper is that redemption fees prevent short-term trading by investors and allow fund managers to reduce the cash reserves[12]. As changes in the portfolio composition are one of the factors that can differentiate the passive funds' returns from their indices', redemption fees will be included in the list of potential key drivers of tracking error for this current study.

Filip (2020) analysed the risk drivers of 82 equity funds that operate in Poland, collecting the data for the period 2000-2015 to observe whether fund size, age, family size, and investment policy influenced the investment risk. The risk measures used for the study were standard deviation, continuous semi-deviation (downside risk), and tracking error, leading to the conclusion that fund age and family size are strongly negatively correlated to the tracking error, as older funds that belong to large families perform better in replicating the index returns[13].

The finding regarding the fund age is opposite to the result observed by Rompotis (2011), who examined the tracking error from 2002 and 2007 for 50 iShares ETFs, noting that older funds exhibit higher tracking error values, mainly due to the increased tendency of fund managers to engage in more aggressive strategies in their attempt to increase the overall returns[14].

Meinhardt, Mueller, and Schoene (2015) studied more closely the tracking errors for ETFs, comparing physical and synthetic replication structures from a sample of fixed-income and equity ETFs listed at the Frankfurt Stock Exchange. They observed that the high tracking error measures were not related to their asset classes, but the TE was smaller for fixed-income ETFs rather than equity ETFs[15].

Fassas (2015) tested the statistical significance of the difference in tracking error between 10 pairs of synthetic and physical European ETFs that follow the same index and are denominated in the same currency, concluding that both physical and synthetic ETFs deliver the same average daily return – with physical funds exhibiting a higher degree of co-movement with the index returns[16].

As Morningstar shares the ETFs replication methods, this current work will review the findings of the latest two mentioned research papers to see whether they can be confirmed or not. As seen in this section, the passive funds' tracking error is a measure that allowed multiple researchers to assess the capability of index funds and ETFs to efficiently track their respective

indices. However, all studies published so far have been conducted on small samples of fund data and with few financial aspects to analyse.

This project intends to fill the existing gap.

# 3 Research Methodology

The financial data for index mutual funds, ETFs, and the related indices is collected from Morningstar and stored in a database using an ETL (Extract, Transform, Load) process that aims to include the passive funds with inception date before 1st January 2019, their fees information, and their prices available also for the index they track. The tracking error between a given passive fund and its related index is then calculated comparing the price difference for each dealing date with records stored in the database.

The calculated value represents the dependent variable in the consolidated dataset that includes both index mutual funds and ETFs, and that allows to investigate the clean data with the Exploratory Data Analysis (EDA). The dependent variable will be used in five predictive machine learning models to establish the aspects that allow passive funds to track more efficiently their indices.

The methodology applied to the study is the CRISP-DM and it includes the following steps:

## 3.1 Business Understanding

Setting the purpose of the study after describing the inception of index mutual funds and ETFs, explaining the value of finding the correct approach to analyse the passive funds' performance with the tracking error metric. As the investments in funds have gradually shifted from active to passive management in the last two decades, this work intends to calculate the tracking error of all passive equity funds available on Morningstar and conduct an analysis of what are the key factors. The critical aspect of the analysis is indeed the identification of all potential aspects such as fees, inception date, asset size, portfolio securities, and other aspects that can be a key indicator of the tracking error.

## 3.2 Data Understanding

Morningstar provides with its Rest API the financial data required for this research, along with passive funds' and indices' prices. The scraping procedure starts from finding all the passive fund identifiers and then use them to scrape the financial information. The last scraping step involves the collection of prices for each pair of passive fund and tracked index (also called *benchmark*), which will allow to calculate the tracking error.

The data includes financial information regarding the following aspects:

- Fund type: distinction between mutual funds and Exchange-Traded Funds;
- Indexing approach: specifies whether the index is tracked by physically investing in the same securities (*physical full* replication) or by investing in derivatives that have a total return similar to the index (*synthetic* replication). There is also a less frequent indexing approach – called *physical sample* replication – that involves investing

directly only in a subsample of the index portfolio, in order to reduce the transactions and hence the overall fund fees;

- Number of stock holdings: indication of how diversified the portfolio is. This field is heavily affected by the indexing approach, given that physical full replication often implies a higher overall stock holding, whereas synthetic replication often involves the investment in a single derivative contract that imitates the index returns;
- Equity style: determines the type of equity securities in which the index (and therefore the passive funds) invests, with *value* style indicating stocks of undervalued companies, *growth* style indicating stocks of companies with future potential to overperform the market, and *blend* being a mix of the two styles above;
- Management fees: charges that cover the cost of investment management for activities such as market research and investment selection. Passive funds have much lower management fees compared to active funds, given the different investment strategy;
- Exit fees: charges triggered when the investors redeem the fund's shares they owned;
- Fund age: it represents how old a fund is, based on the number of days with prices and not based on the actual years since inception;
- Asset size: calculated in a common currency (EUR) for all funds, it has been demonstrated in previous studies that larger funds have in general lower fees, which result in lower tracking error values.

## 3.3 Data Preparation

The available Morningstar data needs to be collected and cleaned in a controlled process that can be easily replicated. For this reason, it is necessary to implement an ETL (Extract, Transform, Load) process, which automatically scrapes the records from the Morningstar website into CSV files, loads the data in two database tables for funds information and prices, and calculates the tracking error for each passive fund. The funds data along with the calculated tracking error is then consolidated in a clean CSV file that is used for the Exploratory Data Analysis and the tuning of predictive machine learning models.

## 3.4 Modelling

Using the clean data generated by the ETL process, it is possible to explore its fields and enrich the dataset with derived features so that the data is more suitable for machine learning models. The models are built with the *sklearn* Python package and aim to predict the long-term tracking error based on the several fund aspects included in the dataset.

The following algorithms are implemented for the prediction task:

- Ridge Regression: the regression method that introduces, on top of the regular multiple linear regression model (also called *least squares*), a regularization parameter called *lambda* ($\lambda$). This parameter reduces the variance of the estimates to avoid *overfitting*[17], which is the excessive model adaptation to the train data that leads to poor performance on the unseen values of the test set;

- Lasso Regression: method similar to Ridge, but with a different function that can remove the redundant independent variables from the model, hence making the prediction easier to interpret and more accurate than Ridge regression when there are several unnecessary variables (high-dimensional regression problems)[18];
- Decision Tree for regression: it follows the principle of "*divide-et-impera*" (i.e. "divide-and-conquer")[19] as it breaks down the data into subsets based on the independent variables that allow to reduce the standard deviation value by the highest figure. It can be explained and visualized clearly to the audience thanks to the possibility of plotting its tree-like structure;
- Random Forest: ensemble method that combines several decision trees with a random subset of features that run in parallel using the bagging method (random sampling with replacement). The Random Forest was proposed by Leo Breiman[20] and calculates the mean of all decision trees' predictions, which becomes the final predicted value;
- XGBoost (Extreme Gradient Boosting): ensemble method that combines several decision trees that are built in sequence to gradually reduce the errors (optimized gradient boosting). This model was developed by Tianqi Chen and Carlos Guestrin in 2016 as a project at University of Washington[21] and immediately became popular due to its high prediction effectiveness and computational efficiency.

### 3.5 Evaluation

Regression models are usually assessed using one of the two following metrics:

- Mean Absolute Error (MAE): calculated as the mean absolute difference between actual values ($y_i$ – in this case, the actual fund's tracking error) and predicted values ($\hat{y}_i$):

  $$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}$$

- Root Mean Square Error (RMSE): calculated as the square root of the mean squared difference between actual values ($y_i$) and predicted values ($\hat{y}_i$):

  $$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}$$

Both metrics represent the model's prediction error and can have a value between zero and infinity, but MAE was chosen as the evaluation metric for this study as it is easier to interpret and more common in the machine learning industry.

## 4 Design Specification

### 4.1 ETL Process

This work includes the application of an ETL (Extract, Transform, Load) process that scrapes the data for mutual funds, ETFs, their prices, and their benchmarks' prices, storing the records in a database that allows to calculate the tracking error.

The ETL process used in this study comprises the following types of tasks:

- SQL Server: runs a SQL statement or stored procedure from a script file that can be used for truncating, creating, altering, or dropping tables/views[22];
- Execute Process: runs a custom application (like Python) specified in the executable variable that is directed to a script file determined in the argument variable[23];
- Data Flow: encapsulates the data flow engine that moves data between sources and destinations, and lets the user transform, clean, and modify data as it is moved[24].

This tool can collect financial data and prices for thousands of funds, representing the type of instrument never built by previous researchers and that have prevented them from supporting an analysis with a large sample.

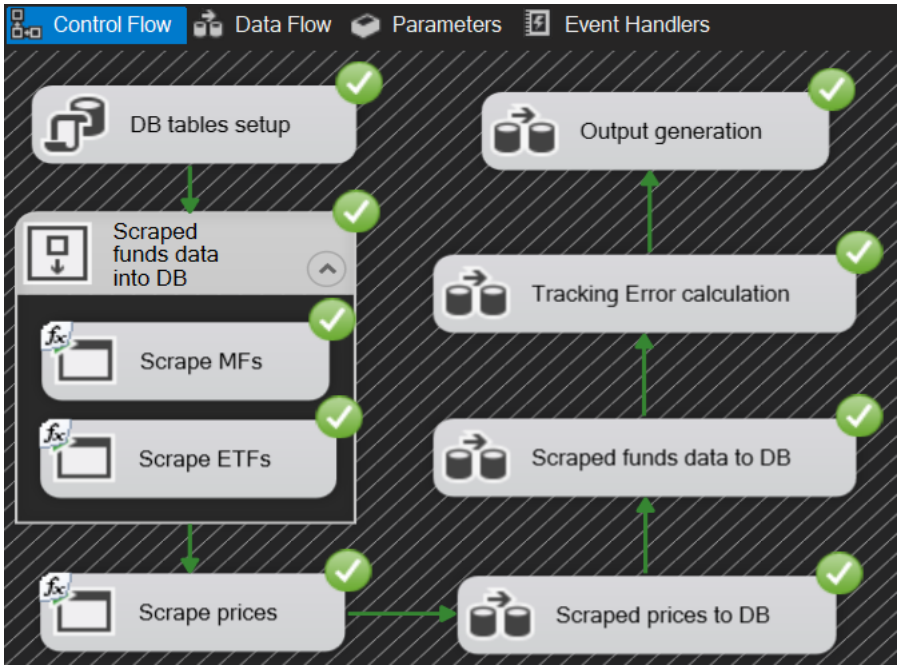


**Figure 1: ETL process workflow with SQL Server, Execute Process, and Data Flow tasks**

The ETL process workflow displayed above comprises the following steps:

1) Running a SQL Server task that removes the existing records for the database tables, so that the updated data can be scraped and inserted into the database;

2) Running in parallel two Execute Process tasks for Python scripts that collect the Morningstar identifiers for passive mutual funds and ETFs, then scrape the financial information using the Morningstar Rest API, and save the output in two CSV files;

3) Running an Execute Process task for a Python script that scrapes prices displayed on Morningstar in JSON format for the fund and index identifiers already retrieved, then saves the output in a CSV file. As Morningstar includes prices for all calendar days, which are displayed in timestamp rather than actual dates, the script converts them in dates and filters out all the week-end days as they are not dealing days;

4) Performing the first Data Flow task that uses the fund's and index's daily prices from the CSV file to calculate the squared difference of each pair of prices, used for the calculation of the tracking error, and then stores the 8,272,626 records in the *prices_te* database table;

5) Performing a Data Flow task that filters only fund identifiers for which there are prices in the database, retains the fees data that was recently reported on Morningstar (after 2019), directs the funds launched after the beginning of 2019 and the ones without reported fees to the "invalid_funds.csv", and eventually inserts the 2,744 clean records in the funds table. The cleaning Data Flow task finds that there are 956 funds that are not valid for the scope of this research:

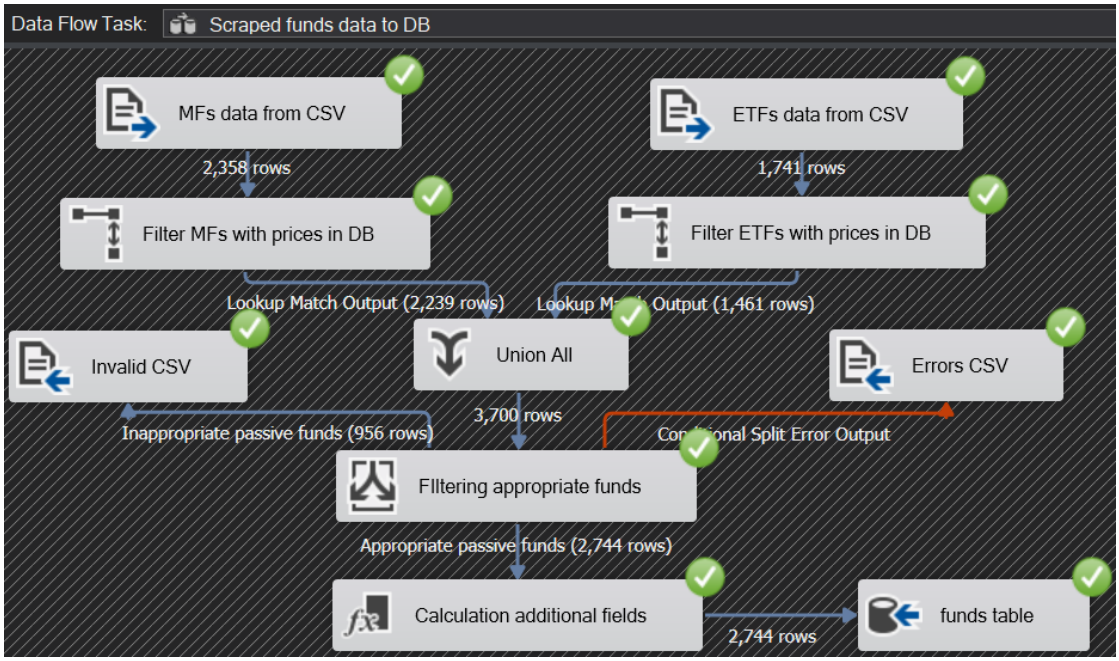


**Figure 2: Funds' data cleaning task in the ETL process**

6) Performing a Data Flow task that calculates the daily price difference for each passive fund compared to the index they track, using this calculated field to extract the tracking error measure to be saved in the *funds_te* table using the following method:
   - Standard deviation of the difference between passive fund ($R_P$) and index ($R_I$) prices:

$$TE = \sqrt{\frac{\sum_{i=1}^{N}(R_P - R_I)^2}{N}}$$

7) Performing the last Data Flow task, which retrieves the funds data from the *funds* database table and performs the lookup to collect the tracking error from the *funds_tracking_error* database table and then saves the consolidated dataset in a CSV file.

## 4.2 Design Integrations on Cleaned Data

The data cleaned by the ETL is analysed with an Exploratory Data Analysis (EDA) process to find insightful information about the correlation values and the status of passive funds; then it is handled by tuned machine learning algorithms to see if there are findings in line or in contrast with previous research papers.

Both processes are carried out in the Python environment called *Jupyter Notebook*, an interactive web framework that embeds live code, output values, and visualizations, which co-

exist harmoniously. This environment's feature makes it ideal for maintaining and presenting a work that needs to be continuously modified and improved to obtain information regarding the driving factors of tracking error. The other significant advantage is that it allows the notebooks' readers to see how the entire workflows (EDA visualizations, Machine Learning pipelines, etc.) have been organized.

# 5 Implementation

## 5.1 Exploratory Data Analysis (EDA)

The first part of the EDA involves the correlation analysis to gain insights on the relationship between the key fields:


**Figure 3: Correlation matrix of the key fields**

The following correlation values are worth to mention:

- Tracking error and management fees: 0.75
  A moderate positive correlation was expected, given that previous research papers suggested management fees are a key factor for passive funds. However, the correlation analysis shows that there is no other field in the dataset with a similar strong correlation with tracking error;
- ETFs and physical replication: -0.64
  ETFs are more likely to use derivatives to track the index return, or invest in a subsample of the index portfolio, while passive mutual funds often perform the full replication method, mirroring the entire index portfolio. Additionally, given that ETFs have lower management fees than index mutual funds, they exhibit a lower tracking error;
- Tracking error and fund age (expressed as *days_with_prices*): 0.45
  The effect of fund age on tracking error has been studied in similar projects and on smaller fund samples, but with controversial results. The correlation coefficient is moderately positive, revealing that funds with longer tenure tend to have a higher tracking error compared to funds recently launched. This seems to be partially caused by the fact that older funds charge higher fees to their investors.

However, correlations are not sufficient to explain the impact of some fields to the tracking error measure: factorplots allow to display the tracking error values based on the categories of one field, with records broken down by the categories of another column.

The graph below shows the impact of the indexing approach and the equity style:

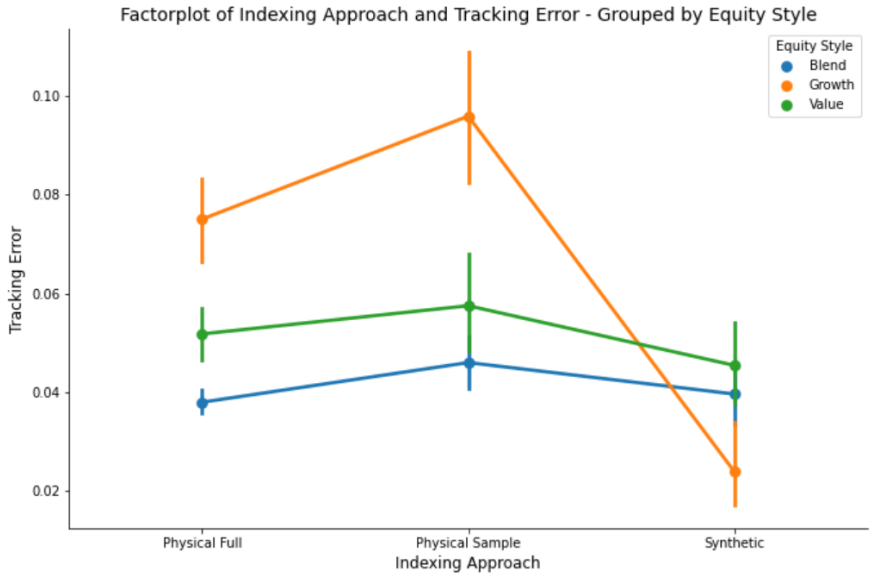


**Figure 4: Factorplot of Indexing Approach and Equity Style Impact to Tracking Error**

The indexing approach seems to have a clear impact on tracking error, especially considering the abnormal trend of funds with growth equity style: they exhibit a larger TE when they follow the physical replication compared to funds with value and blend styles, but lower tracking error when they try to replicate the index returns via derivatives (synthetic replication).

The following graph shows the impact of stocks in portfolio and fund type:

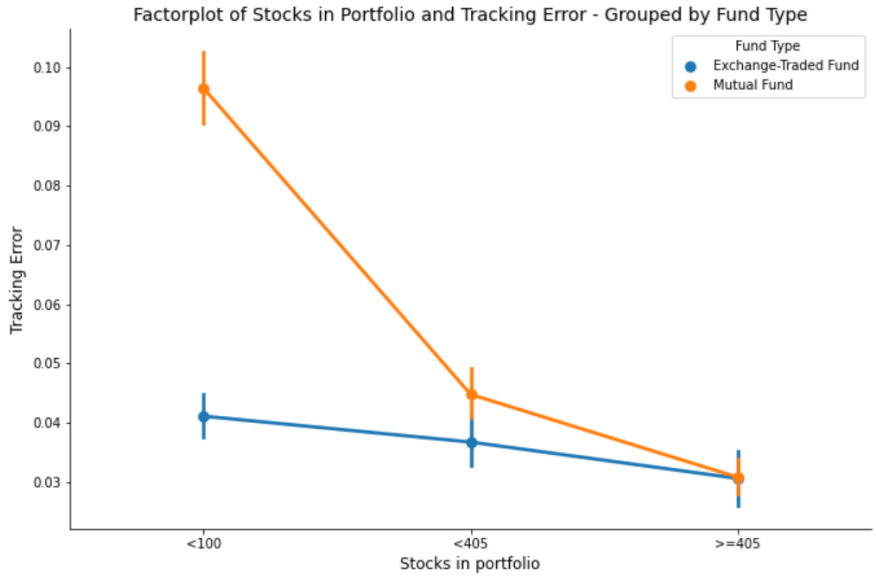


**Figure 5: Factorplot of Stocks in portfolio and Fund Type Impact to Tracking Error**

The higher the number of stocks in passive funds' portfolio, the lower the tracking error, but this negative trend is more prominent for mutual funds rather than ETFs: for funds with limited stocks in the portfolio, mutual funds show a much higher TE compared to ETFs, while, for funds with many stocks in the portfolio, mutual funds exhibit the same tracking error as ETFs.

Before building the predictive models, the dataset was manipulated by converting some categorical features into numeric and by creating additional fields – a process called *feature engineering* – that make the data suitable to be used for the machine learning algorithms.

Feature engineering involved the following changes:

- Use the *country_exposure* column (countries in which the fund invests, along with the percentage of each asset allocation) to retrieve the country name with the highest investment (*first_country_exposure*), and a second field with number of countries (*n_countries_exposure*);
- Add the Boolean field *monosector*, which evaluates whether the fund has all of its assets allocated to a single sector;
- Add the categorical column *highest_sector_name*, which displays the name of the sector with the highest investment;
- Add the field *physical_replication* that contains the numeric data of the string column *indexing_approach*, with the "Physical Full" values converted to 1, "Physical Sample" converted to 0.5 (partial physical replication), and the "Synthetic Replication" values converted to 0 (no direct investment in any underlying security of the indices);
- Add the field *dividends_in_year* that contains the numeric data of the string column *dividend_frequency* ("Annually" values converted to 1, etc.);
- Converting the *fund_type* string field into the numeric *exchange_traded_fund*, with value 1 for passive ETFs and value 0 for index mutual funds;
- Converting the string columns *domicile*, *region_exposure* (list of each world region in which the fund invests, along with the percentage of asset allocation), *first_country_exposure*, and *highest_sector_name* (previously calculated) in dummy numeric fields.

To accommodate the use of machine learning algorithms, the dataset has also been standardized using the "StandardScaler" *sklearn* class. Standardization rescales the numeric features so that they fit a standard normal distribution that has a mean equal to zero ($\mu=0$) and a standard deviation equal to one ($\sigma=1$), and it is important when the data is highly varying in magnitudes, units, and range. If the numeric features are not rescaled, some machine learning models may have higher residuals (i.e. worse predictions) because they fail to factor in the variance of the numeric features[25]. For example, this is necessary to balance the value ranges when there are fields with very large values (such as *asset_size_eur*, with magnitude of millions/billions) and other ones with values that range from 0 to 1 (such as *exchange_traded_fund* and *physical_replication*).

The image below displays the distribution of tracking error values, comparing the original/unscaled data with the standardized one:

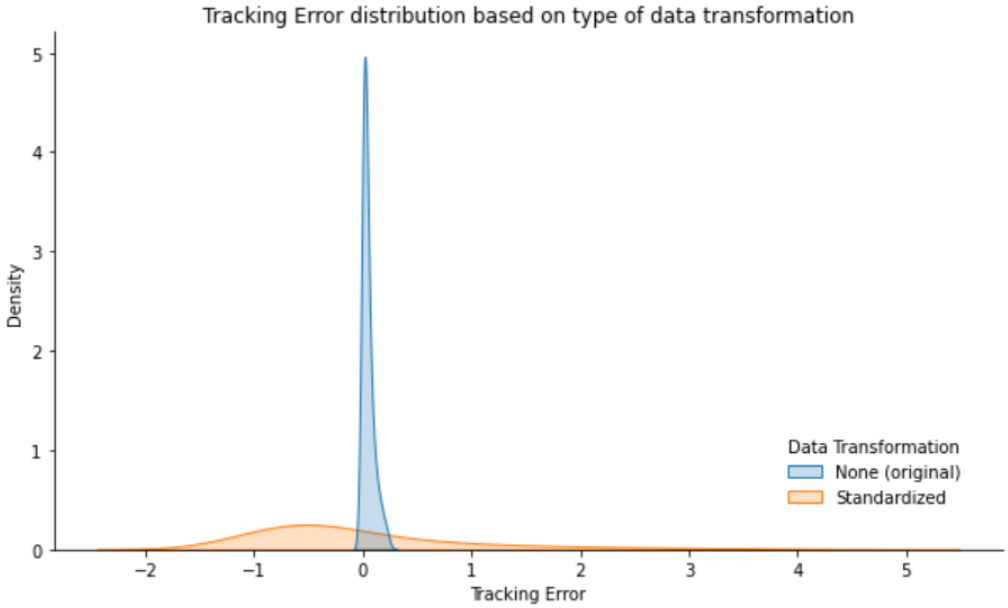


**Figure 6: Distribution of Tracking Error after standardization**

## 5.2 Machine Learning models

After analysing the dataset and adjusting the fields to accommodate the predictive algorithms, the data has been split to train-test sets of 80%-20% respectively, and *k*-fold Cross-Validation (CV) was also applied to further split the train set in multiple folds, so that the models avoid overfitting the predictions on the whole train set and scoring poorly on the unseen test set. The *k*-fold value for the machine learning models used in this study is 10, which is considered the default value[26].

Additionally, multiple linear regression (MLR) has been applied to the fund data, before any feature engineering was performed, so that it acts as a benchmark, allowing to estimate how much the following models have learned to predict the tracking error compared to the standard regression algorithm's score (called *baseline*):

- Ridge Regression: the model finds *management_fees* and *days_with_prices* to be the aspects with the highest positive coefficients of importance (i.e. their values increase together with TE), while *holdings_stock* and *exchange_traded_fund* are the ones with the highest negative coefficients of importance (i.e. their values decrease as TE increases). Features that display the asset allocation in specific sectors such as technology, healthcare, and real estate also have some impact on the tracking error according to this model. The following horizontal bar chart shows the coefficients of importance for the most relevant fields in the Ridge Regression model:

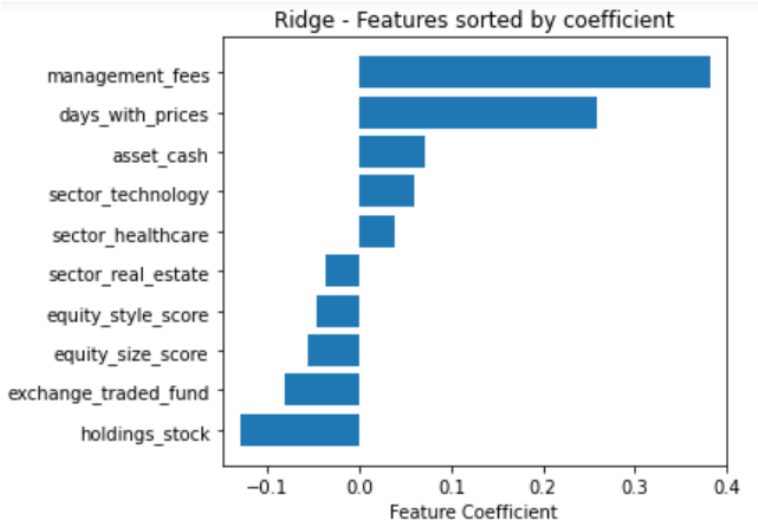


**Figure 7: Feature importance in the Ridge Regression model**

- Lasso Regression: the image below shows that it has similar coefficients of importance for the fields mentioned in the Ridge algorithm, but this model removed all the sector-related features and kept the *physical_replication* field – which negatively affects the tracking error in a marginal manner. It is also worth to note that, while the MAE scores of Ridge and Lasso are the same on the test set, they both perform slightly better on the test set than the train one, although the difference in the predictions error is minimal and seem to be purely incidental;

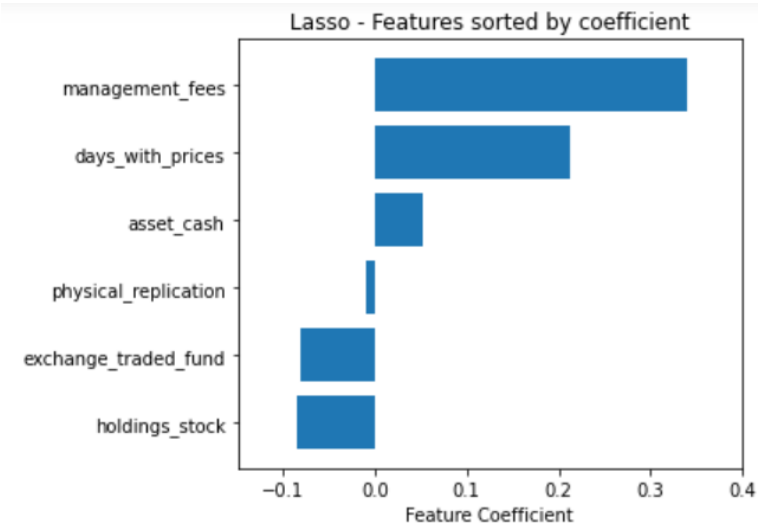


**Figure 8: Feature importance in the Lasso Regression model**

- Decision Tree: the "*divide-et-impera*" logic behind this tree-based model can be easily communicated in the typical plot that shows the top-down decision flow, with the root node at the top representing the entire data, followed by various nodes where the splits occur based on the variable capable to reduce the variance by the highest amount, and the final predictions are displayed in the tree leaves. The following graph shows the pattern used by the model for predicting the tracking error:

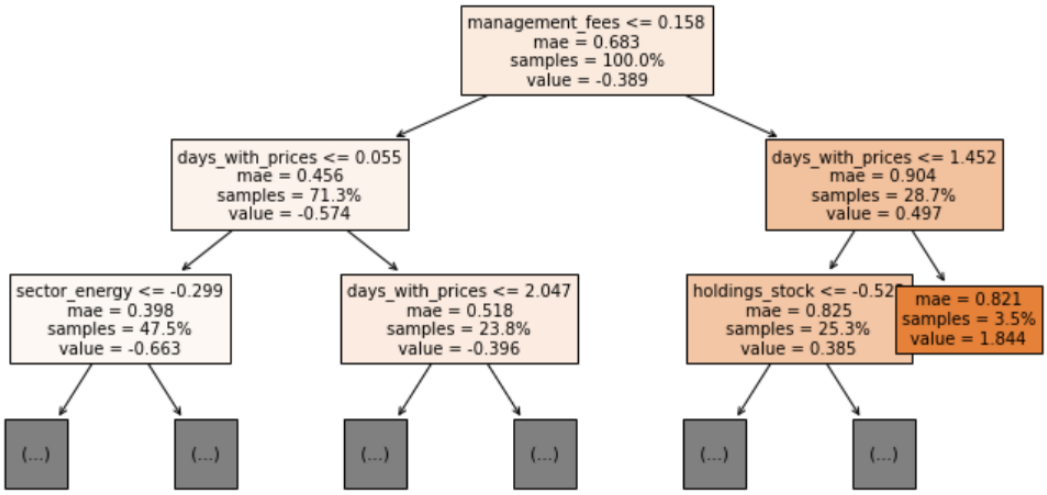


**Figure 9: Decision Tree plot with the most important variables used to predict the Tracking Error**

The key variables used by the Decision Tree to split the records and forecast the tracking error from the subset are the same ones already selected by the regression models, except for *exchange_traded_fund*, which does not appear to be significant;

- Random Forest: the ensemble bagging model uses a similar pattern to split the records compared to the Decision Tree in terms of features importance, but with better prediction scores on both train and test sets, as evident in the following figure:

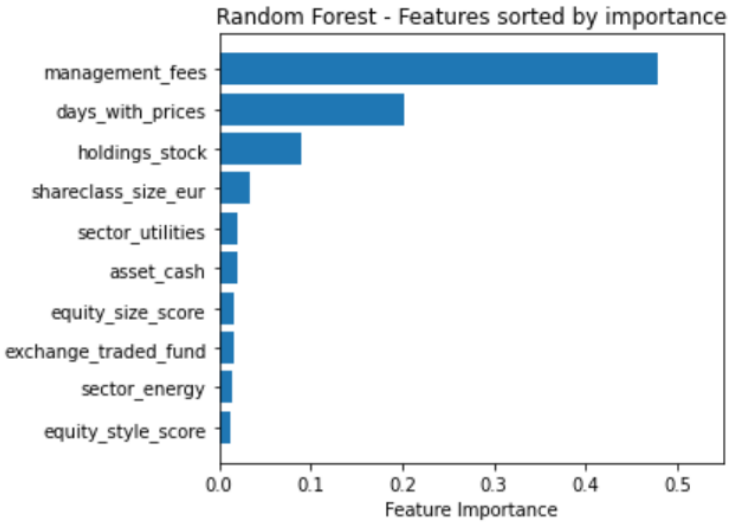


**Figure 10: Feature importance in the Random Forest model**

- XGBoost: the ensemble boosting model shows a very different rank of features relevancy, given that the importance distribution is no more front-loaded as in the Random Forest – although management fees still appear to be the factor with the highest influence on the tracking error. The other relevant factors are fund type, age, stocks in portfolio, and exit fees, followed by several sector-based features displayed in the picture below:

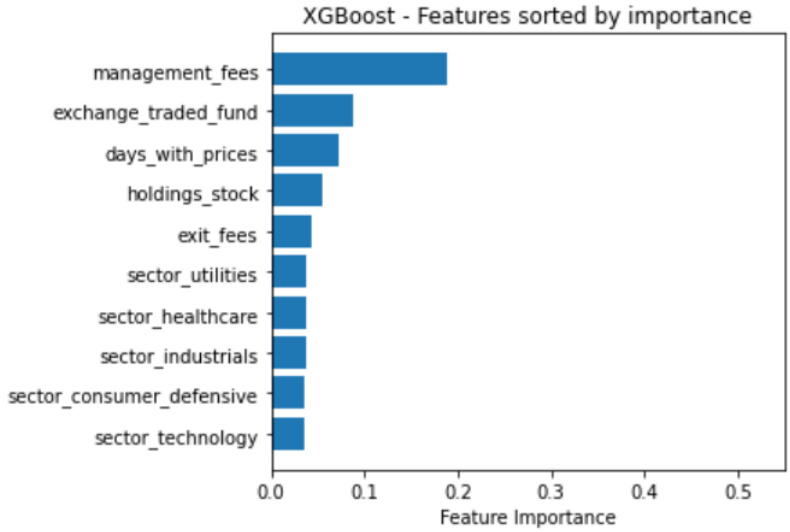


**Figure 11: Feature importance in the XGBoost model**

Whilst the Ridge and Lasso models applied similar regression coefficients to the same independent variables, Random Forest and XGBoost have different feature importance ranks as result of their different approaches: the bagging algorithm (Random Forest) runs in parallel 100 random trees with only a random subset of the original list of independent variables, and then calculates the average tree; the boosting algorithm (XGBoost) runs in sequence 100 trees in which the new ones predict the residuals of previous trees, so that they are added together in a single model.

# 6 Evaluation

This study has found evidence that tree-based machine learning models predict more accurately the tracking error compared to the regression algorithms. XGBoost is the model with the lowest mean absolute error (MAE) and therefore is considered ideal for assessing the tracking error's key factors.

All the models outperformed the basic multiple linear regression (MLR) algorithm and its 0.642 MAE on the test set, but exhibit prediction scores that are worth to be described:

- Ridge Regression: it performs worse than the baseline on train data but has a better MAE score of 0.578 on test data. The algorithm has moderate overfitting on the train set, but substantially lower than the inefficient MLR baseline;
- Lasso Regression: like the Ridge model, it has a higher MAE score on the train set and a lower MAE on the test set compared to the MLR baseline. It is however slightly less accurate at predicting the tracking error on the train data rather than the test data, which is an unusual occurrence that seems to be incidental;
- Decision Tree for regression: with the first tree-based model the prediction errors are significantly reduced. The algorithm performs much better than the regression-based models on both train and test (MAE: 0.530) sets;
- Random Forest: it performs better than the plain Decision Tree, with a lower MAE score on train (0.380 vs 0.467) and test set (0.467 vs 0.530), and has a similar level of overfitting;

- XGBoost: the ensemble boosting model shows a remarkable precision on the train data, with an extremely low MAE of 0.055, but it is partially caused by overfitting, given that the MAE score on the test set is only slightly better than Random Forest (0.429 vs 0.467). All the different tuning parameters that have been experimented on the XGBoost algorithm have proved to be ineffective at reducing the overfitting.

The following table summarizes the MAE scores for all models on both train and test sets.

**Table 1: Models Mean Absolute Errors (MAE) on train and test sets**

| Model name | MAE on Train set | MAE on Test set |
|---|---|---|
| Baseline - MLR | 0.552791 | 0.642070 |
| Ridge | 0.564508 | 0.577969 |
| Lasso | 0.585616 | 0.575591 |
| Decision Tree | 0.467266 | 0.529534 |
| Random Forest | 0.379642 | 0.467182 |
| XGBoost | 0.055666 | 0.428573 |

As the 10-fold Cross-Validation did not prevent the XGBoost from heavily overfitting the train data, a second set of machine learning models has been tuned after filtering only 33 relevant fields out of the original 79 ones. This action did not intend to provide further information regarding the independent variables to be used for the prediction, but it only aimed to compare the results of all five models to see if the MAE results and overfitting on train set have changed: the overall prediction scores were marginally worse than before but exhibited a significant reduction of overfitting on train data, especially for the two most accurate models (Random Forest and XGBoost).

The following line charts allow to appreciate the overfitting of all models and the MAE score comparison with the benchmark:

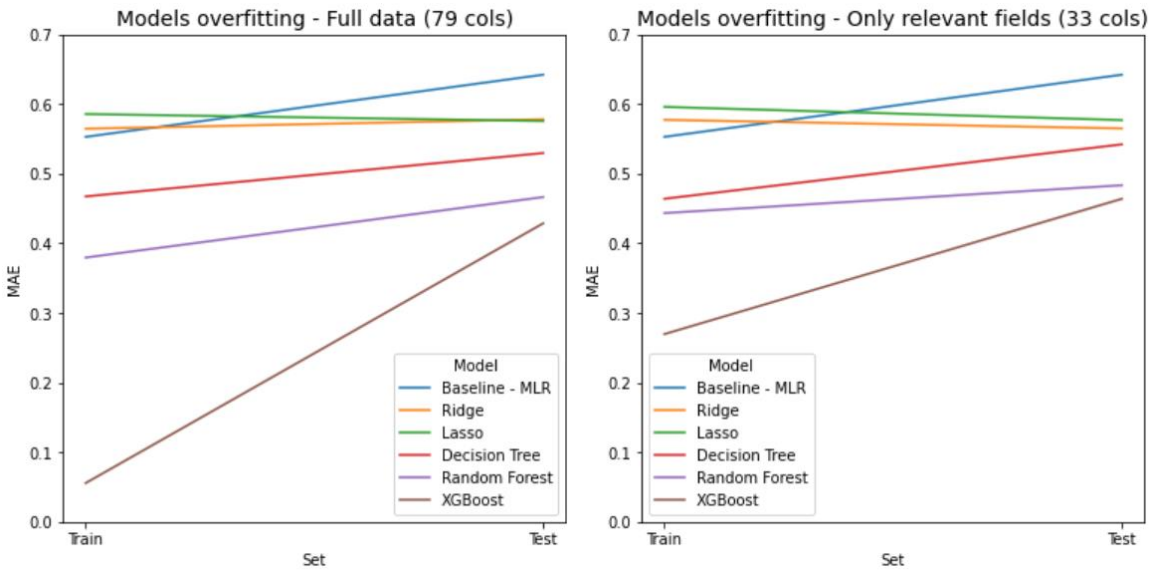


**Figure 12: Line chart with the MAE models, comparing the metric between train and test sets**

# 7 Conclusion and Future Work

## 7.1 Conclusion

A data analysis research with the purpose of examining the key factors of the passive funds' tracking error is very valuable, especially if it includes a larger sample of data compared to the other research papers published so far. Morningstar is the most reliable financial data provider and, using their Rest API, this study is based on the data collected for all available passive funds and their related indices.

To collect the financial data that can be used for the analysis, the records for funds data and their prices are cleaned and stored in multiple database tables using an ETL process that allows to generate a consolidated dataset in which the tracking error for each index mutual fund or ETF is calculated along with the relevant fund data for net asset value, fees, age, etc.

Once the data is collected and cleaned, the exploratory data analysis is carried out to obtain an overview of the current passive funds' sector, and is followed by the implementation of regression- and tree-based machine learning models that aim to obtain valuable information regarding the prediction of the tracking error for index mutual funds and ETFs.

The models' results indicate that the driving factors of tracking error are the following:

- Management fees: it is the only fund aspect that has been repeatedly considered in previous research papers to be a key element in the estimation of tracking error. So it is not a surprise that also this work considers this independent variable to be the main indicator of how well a passive fund tracks its index's return;
- Fund type: there have been studies that demonstrated the higher ETFs efficiency compared to mutual funds regarding fees and tracking error, but not all research papers have agreed on this aspect. This study has found evidence of the better performance of ETFs, especially when they hold a limited number of stocks in their portfolio, a situation where mutual funds exhibit a significantly larger TE;
- Fund age: despite this element has not been often studied in previous works, there have been some contradicting results, with the older funds being considered more stable and therefore causing lower TE, but also being considered riskier by other studies that assume an increased propension of fund managers to engage in more aggressive strategies. The current work has noticed a positive relationship between fund age (expressed as *days_with_prices*) and TE, leaning towards the latter finding expressed in previous studies;
- Stocks in portfolio: diversification is a key aspect of funds, even if it has not been frequently considered in previous studies. In passive funds, the portfolio diversity can have a dual consequence: the higher the number of stocks, the more likely the fund can try to perfectly replicate the index's (increased returns' replication accuracy), but also the higher the likelihood of requiring to buy/sell securities that have been added/removed from the index (increased fees). According to this work, portfolio diversity has an indirect relationship with the tracking error, which is more evident for index mutual funds: both fund types show a lower tracking error when they hold several stocks in their portfolio, but mutual funds with less than 100 stocks exhibit a TE value almost 3 times higher than when they hold more than 405 stocks;

- Exit fees: although it is not possible to monitor the changes in portfolio composition over time (due to lack of information regarding the underlying securities and their trades), redemption fees permit to recognize the funds that can prevent short-term trading by investors and allow fund managers to reduce the cash reserves, leading to a lower TE.

## 7.2 Future Work

There are extensions for future research resulting from this work:

- Instead of calculating the tracking error for all passive funds from their inception date, a useful future study would limit the tracking error analysis to a specific period of bull market and a period of bear market to see which fund aspects made them more efficient at tracking their indices during periods of high volatility;
- Another possibility is to focus the research only on indices that are tracked by both mutual funds and ETFs, so that a deeper analysis on the consequences of the passive fund type to the tracking error can be carried out.

# References


[1] Mingo-Lopez, D., Matallin-Saez, J., and Soler-Dominguez, A. (2020) 'Cash Management and Performance of Index Mutual Funds', *Academia Revista Latinoamericana de Administración*, 33(1), pp. 549-565.

[2] Clements, R. (2019) 'Post-Crisis Financialization Through Product Innovation: Assessing Complexity, Growth & Design in Exchange Traded Funds', *Virginia Law & Business Review*, pp. 8-20.

[3] Crisóstomo, R. and Sanchez-Seco, J. M. (2018) 'ETFs and Financial Stability: A Compendium of Possible Risk Sources', *CNMV Bulletin*, 4(1), pp. 71-82.

[4] Rompotis, G. (2011) 'The Performance of Actively Managed Exchange-Traded Funds', *The Journal of Index Investing Spring*, 1(4), pp. 53-65.

[5] Frino, A. and Gallagher, D. (2001) 'Tracking S&P 500 Index Funds', *The Journal of Portfolio Management*, 28(1), pp. 44-55.

[6] Kostovetsky, L. (2003) 'Index Mutual Funds and Exchange-Traded Funds', *The Journal of Portfolio Management*, 29(4), pp. 80-92.

[7] Blitz, D., Huji, J., and Swinkels, L. (2010) 'The Performance of European Index Funds and Exchange-Traded Funds', *European Financial Management*, 18(4), pp. 649-662.

[8] Dhabolkar, P. and Reddy, Y. (2019) 'Evaluating the Tracking Performance of Index Mutual Funds and Exchange Traded Funds in India', *IUP Journal of Financial Risk Management*, 16(1), pp. 37-49.

[9] Chen, J., Chen, Y., and Frijns, B. (2017) 'Evaluating the tracking performance and tracking error of New Zealand exchange traded funds', *Pacific Accounting Review*, 29(3), pp. 443-462.

[10] Singh, J. and Kaur, P. (2017) 'Tracking Efficiency of Exchange Traded Funds (ETFs): Empirical Evidence from Indian Equity ETFs', *Paradigm*, 20(1), pp. 176-190.

[11] Wong, K. and Shum, W. (2010) 'Exchange-traded funds in bullish and bearish markets', *Applied Economics Letters*, 17(16), pp. 1615-1624.

[12] Waller, W., Nanigan, D., and Finke, M. (2018) 'Redemption Fees: Reward for Punishment', *Journal of Financial Services Providers*, 72(2), pp. 49-68.

[13] Filip, D. (2020) 'Are Fund Attributes Risk Drivers? Evidence for the Polish Mutual Funds', *Journal for Economic Forecasting, Institute for Economic Forecasting*, 1(1), pp. 22-36.

[14] Rompotis, G. (2011) 'Predictable patterns in ETFs' return and tracking error', *Studies in Economics and Finance*, 28(1), pp. 14-35.

[15] Meinhardt, C., Mueller, S., and Schoene, S. (2015) 'Physical and Synthetic Exchange-Traded Funds: The Good, the Bad, or the Ugly?', *The Journal of Investing Summer*, 24(2), pp. 35-44.

[16] Fassas, A. (2015) 'Tracking Ability of ETFs: Physical versus Synthetic Replication', *The Journal of Index Investing Fall*, 5(2), pp. 9-20.

[17] Maronna, R. (2011) 'Robust Ridge Regression for High-Dimensional Data', *Technometrics*, 53(1), pp. 44-53.

[18] Genovese, C. R., Jianshun, J. J., and Wasserman, L. (2012) 'A Comparison of the Lasso and Marginal Regression', *Journal of Machine Learning Research*, 13(68), pp. 2107-2143.

[19] Dinov, I. D. (2018) 'Decision Tree Divide and Conquer Classification', *Data Science and Predictive Analytics*, pp. 307-343.

[20] Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5-32.

[21] Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[22] Microsoft (2017) *Execute SQL Task*. Available at: https://docs.microsoft.com/en-us/sql/integration-services/control-flow/execute-sql-task?view=sql-server-ver15/ [Accessed 10 August 2021].

[23] Microsoft (2017) *Execute Process Task*. Available at: https://docs.microsoft.com/en-us/sql/integration-services/control-flow/execute-process-task?view=sql-server-ver15/ [Accessed 10 August 2021].

[24] Microsoft (2017) *Data Flow Task*. Available at: https://docs.microsoft.com/en-us/sql/integration-services/control-flow/data-flow-task?view=sql-server-ver15/ [Accessed 10 August 2021].

[25] Paper, D. (2020) *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*. 1st edn. Apress.

[26] StackExchange (2020) *Why is 10 considered the default value for k-fold cross-validation?* Available at: https://datascience.stackexchange.com/questions/75789/why-is-10-considered-the-default-value-for-k-fold-cross-validation/
[Accessed 10 August 2021].