Data Extraction and Predictive Modelling with Artificial Intelligence

Tutor: Alina Rosu

Program: Master in Finance

Authors: Nil Ferrara Marzo & Joan Prous Conde

Academic Year: 2024/2025

Keywords: Predictive modelling, data extraction, machine learning, SVR, OLS

Contents

Abstract

This thesis focuses on how artificial intelligence (AI) can be used in the world of corporate finance and asset management. It starts by looking at the current state of the art of AI in these disciplines, both as a data extraction tool and as a predictive model engine.

Regarding the field of data extraction, traditional methods currently used by financial institutions are compared to emerging AI techniques, such as Natural Language Processing (NLP). These innovative methodologies are explored in collaboration with Captide, a US-based start-up that is achieving great progress in the field.

When it comes to predictive modelling, 28 AI models are empirically tested to forecast stock returns of U.S.-listed companies. Through this experimental framework, the thesis seeks to answer a critical question: **can cost-effective AI models available to the masses reliably predict stock returns?**

In the course of this quest, the best-performing models are selected and benchmarked against a standard Ordinary Least Squares regression.

ABBREVIATIONS AND SYMBOLS

- **AI:** Artificial Intelligence
- **API:** Application Programming Interface
- **ARD:** Automatic Relevance Determination
- **ARPU:** Average Revenue Per User
- **CAPM:** Capital Asset Pricing Model
- **CCA:** Canonical Correlation Analysis
- **COGS:** Cost of Goods Sold
- **FF3:** Fama-French Three-Factor Model
- **FF5:** Fama-French Five-Factor Model
- **GICS:** Global Industry Classification Standard
- **HML:** High Minus Low
- **LTM:** Last Twelve Months
- **MD&A:** Management Discussion and Analysis
- **Mkt-RF or MRP:** Market Risk Premium
- **ML:** Machine Learning
- **NER:** Named Entity Recognition
- **NLP:** Natural Language Processing
- **NOSH:** Number of Shares Outstanding
- **OCR:** Optical Character Recognition
- **OLS:** Ordinary Least Squares
- **PE Ratio:** Price-to-Earnings Ratio
- **REVPAR:** Revenue Per Available Room
- **R^2:** Coefficient of Determination
- **SG&A:** Selling, General, and Administrative Expenses
- **SVR:** Support Vector Regression

1. INTRODUCTION

Predicting stock returns is perhaps the most debated subject in finance. If successful, it can provide valuable insights for investment strategies and potentially unlimited profits. In this thesis, we aim to explore whether Machine Learning models can help in this pursuit.

Traditional methods like Ordinary Least Squares (OLS) regression struggle with high-dimensional data, a challenge that AI is particularly suited to handle. However, if markets are truly efficient —as Eugene Fama (1970) argues— then future prices already reflect all available information, making return predictions essentially impossible. [10]

Assuming markets are not fully efficient, there are three broad approaches financial research has taken to find out return patterns:

1.1 Factor-based Models

The first popular method sorts stocks into groups based on factor loadings. For example, small-cap stocks tend to outperform large-cap stocks, and value stocks (companies with high book-to-market ratios) tend to outperform growth stocks (companies with low book-to-market ratios). This idea led to the development of multi-factor models, such as the three- and five-factor models introduced by Fama and French (1993, 2015), as well as the profitability factor proposed by Novy-Marx (2013). These analyses are done in cross-section, meaning: at the same point in time, stocks that have a higher exposure (beta) to the small-minus-big (SMB) factor have on average a higher expected return. [8] [9] [17]

1.2 Firm Characteristics Approach

An alternative perspective suggests that stock characteristics, rather than factor loadings, drive expected returns. Sheridan Titman (1997) and colleagues argue that it is not the sensitivity to factors like the small-minus-big (SMB) portfolio that matters, but rather the financial fundamentals of the company are similar to the average small company or not. For example, two companies may both have high SMB loadings, but one might be a tiny startup with exponential growth, while the other is a small manufacturer with stagnant growth, illustrating how the fundamental characteristic can matter more than the factor loading itself. [7]

1.3 Aggregate Market Prediction

A third approach predicts aggregate market returns rather than individual stock performance. John Cochrane (2005) shows that certain financial ratios, such as the dividend-price ratio, can help forecast aggregate market returns with moderate short term predictive power ($R^2 = 15\%$), which increases significantly at longer horizons (60% at five years). Other valuation metrics, such as the price-to-earnings (P/E) ratio, have been studied with similar conclusions. More recently, Kelly, Malamud, and Zhou (2024) applied AI to predict aggregate stock returns using a model with over 10,000 parameters. One of their highlights: models with too many parameters can suffer from overfitting and may even produce negative R^2 values. [6] [14]

In this thesis, we want to check if AI can predict stock prices better than traditional methods, and clearly explain its strengths and weaknesses.

Our road map to this objective starts by reviewing how AI is currently used in corporate finance, especially for extracting data from financial statements. To do this, we partnered with Captide, a San Francisco-based start-up specializing in quarterly earnings data extraction with AI.

After this, we move into predictive modeling, where we build 28 AI models to predict stock returns.

Finally, we compare our AI models to a traditional method (OLS regression) to see which of those predicts best, and identify which variables are most important for predicting stock prices.

2. STATE OF THE ART OF AI IN FINANCE

Artificial Intelligence is transforming every industry that relies on data-driven decisions, and finance is no exception. By processing large amounts of information, AI boosts market efficiency, reduces overhead and frees professionals from monotonous tasks. But how exactly can financial institutions leverage AI to gain a competitive edge? Which specific technologies make this advancement possible? The following chapter will analyze the possible applications of AI in finance —capabilities barely imagined just a decade ago.

The main 4 applications developed during the past decade are:

- **Data extraction from financial filings:** Automated text-processing methods help analyze large numbers of financial statements with a fraction of the current financial analysts' input. [2]
- **Predictive modelling for asset prices:** Machine learning techniques, including neural networks can take into account historical market data, news sentiment, and economic indicators to forecast asset price movements. [20] [22]
- **Fraud detection:** AI can identify suspicious patterns in real-time transaction data, helping institutions comply with legal regulations. [3]
- **Robo-advisory services:** Automated portfolio management platforms —also called "robo-advisors"— provide financial advice based on the user's risk profile and market conditions. The main advantage of roboadvisors are the lower fees (e.g. 0.15% for Vanguard [22] and 0.25% for Wealthfront [23]) compared to human wealth managers (1-2% industry average); while exhibiting similar performance. [3]

In this thesis, we will dive deeper into data extraction and predictive modelling, two applications closely linked to each other, and with a high impact on the day-to-day of many corporate finance professionals.

2.1 Data Extraction: Past & Future

2.1.1 Looking Back: Traditional Data Retrieval in Finance

Since the early 2000s, the typical approach to collecting financial information has been based on retrieving data through Application Programming Interfaces (APIs) from large providers such as Bloomberg.

This process is time-consuming. For instance, the typical day for an equity research analyst usually begins with logging onto these platforms to search for company-specific data (e.g., revenue, earnings, or sector performance), followed by manual reviews of large spreadsheets, reports, and news to gather further insights. Once all the numbers are

compiled, hours are spent writing analyses and developing client-facing presentations. While these methods are reliable, they are also labor-intensive: analysts must constantly switch between data sources and verify each figure's accuracy manually.
What will these types of job look like in the future?

2.1.2 Looking Ahead: Disruptions Driven by NLP and AI

Over the past five years, a major change has emerged in the way financial data can be extracted, mainly due to the exponential growth in Natural Language Processing (NLP). As of 2025, many startups are working towards handling unstructured or atypical data formats in financial documents, especially when tables contain multiple subcolumns or nested indices, since these layouts were impossible to read automatically before. [5]

Recent literature attributes these improvements to three processes:

1. **Named Entity Recognition (NER):** Models trained to locate and interpret numbers, such as revenue or profit. This is crucial to quickly capture the key financial metrics of a company.

2. **Optical Character Recognition (OCR):** Technology that reads text from PDFs and images, enabling AI to extract useful information from charts.

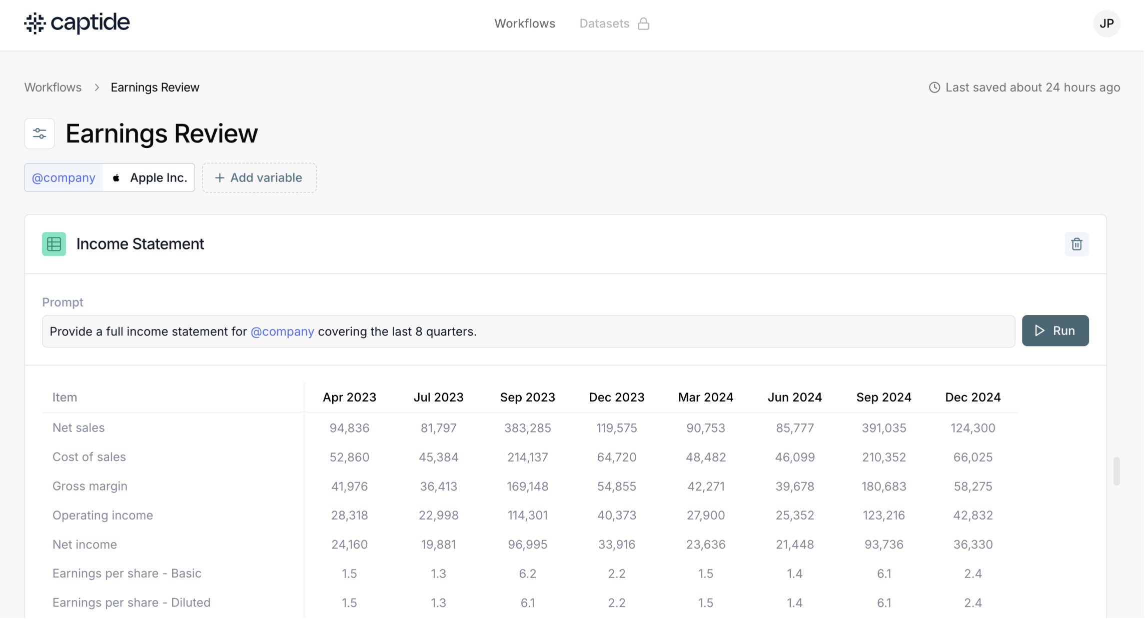
3. **Dependency Parsing:** After text and numbers are identified, dependency parsing creates a logical map of how words relate to each other. For example, it may clarify that "revenue increased" refers to a particular quarter's positive trend.

With these tools in hand, tomorrow's equity research analyst could almost feel like they have "superpowers". Instead of going through documents and comparing spreadsheets for hours, any analyst will rely on software that retrieves entire financial filings in just minutes. Complex tables will get automatically parsed, and each finding will be easily traceable for verification.

2.2 Case Study: Captide's technology as a breakthrough

To explore this topic more deeply, we have partnered with Captide, a startup based in San Francisco that focuses on automating the extraction of financial data from corporate reports, press releases and earnings calls. Their platform offers a chatbot interface[1], meaning users can start using it without the steep learning curve often required by today's more complex platforms (like Bloomberg).

[1] See Figure 1.



Earnings Review

Income Statement

Prompt

Provide a full income statement for @company covering the last 8 quarters.

Item	Apr 2023	Jul 2023	Sep 2023	Dec 2023	Mar 2024	Jun 2024	Sep 2024	Dec 2024
Net sales	94,836	81,797	383,285	119,575	90,753	85,777	391,035	124,300
Cost of sales	52,860	45,384	214,137	64,720	48,482	46,099	210,352	66,025
Gross margin	41,976	36,413	169,148	54,855	42,271	39,678	180,683	58,275
Operating income	28,318	22,998	114,301	40,373	27,900	25,352	123,216	42,832
Net income	24,160	19,881	96,995	33,916	23,636	21,448	93,736	36,330
Earnings per share - Basic	1.5	1.3	6.2	2.2	1.5	1.4	6.1	2.4
Earnings per share - Diluted	1.5	1.3	6.1	2.2	1.5	1.4	6.1	2.4

Figure 1: Captide's chatbot displaying Apple's latest 8-quarter income statement

2.2.1 Captide's Backend Technology

In order to explain the technologies in which this thesis is based, Captide's workflow is shown below in 6 steps:

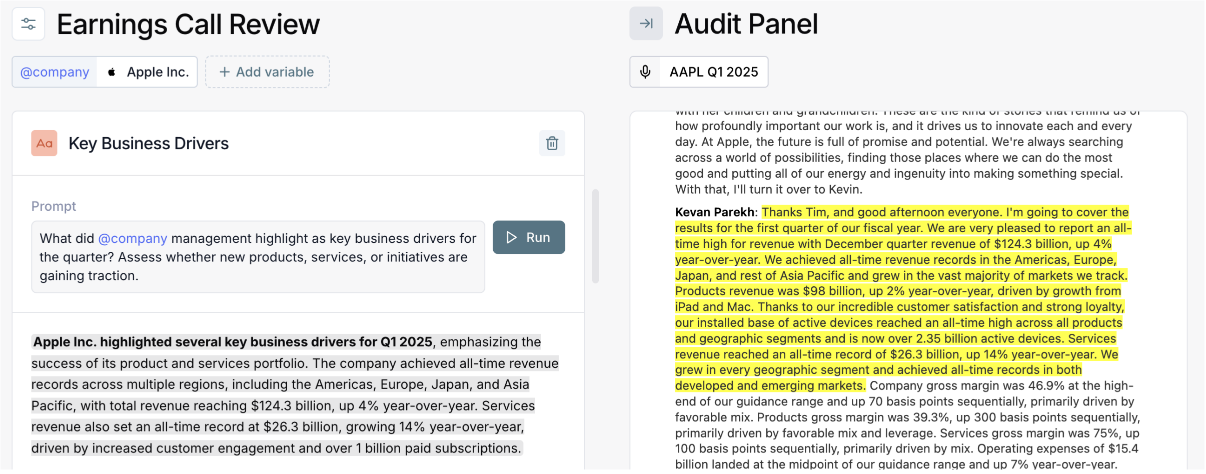
1. **Document Retrieval:** Captide starts by finding and downloading relevant documents like 10-K, 10-Q, and 8-K filings.
2. **Data Preprocessing:** Once the documents are retrieved, it organizes them into smaller, meaningful pieces. For example, a table might be broken into sections like "Revenue Breakdown" or "Segment Performance", while text sections like MD&A may be labeled with dates to make them easier to find in later stages [4].
3. **Vectorization and Query Matching:** Captide converts each piece of the document into a format that can be compared mathematically (in a nutshell, every word is converted into a vector). Then, it matches these pieces with the user's question to find the most relevant information. Using a process called cosine similarity, the best matches are calculated.
4. **Chunk Evaluation and Filtering:** Captide checks the quality of the retrieved "puzzle pieces" to make sure they are relevant. In this step, the results are analyzed and all the irrelevant objects are eliminated.
5. **Fallback Mechanisms:** If the system lacks enough information in the first search, it tries other approaches. These include searching for different types of documents (like switching from 10-K to 8-K) or even using web searches to fill in the gaps.
6. **Synthesis and Delivery:** Finally, Captide combines all the relevant information into an easy-to-use format which the user choses. This could be a structured table

(e.g., CSV file) or a written summary, depending on what the user needs.

2.2.2 Captide's Edge: Unique Differentiators

With established giants like Bloomberg dominating the market, what sets Captide apart?

1. **Faster Real-Time Data Coverage:** Captide extracts granular data from the most recent earnings calls and documents. This data is often available with a time lag on other platforms. Since Captide can provide it in a few seconds, analysts can discover opportunities faster and create strategies that generate alpha.
2. **Sentiment Analysis:** Captide's vision is to not only provide numbers; it also aims to interpret tone and sentiment in management discussions (MD&A sections). This will enable analysts to identify shifts in strategy or confidence, such as increased caution during economic uncertainty or optimism following strong earnings.
3. **Verification:** an "Instant Audit" feature helps analysts double-check the information in each document[2]. Hovering around the chatbot's responses (left part of the screen), the user can see where the content comes from (right part of the screen).



What did @company management highlight as key business drivers for the quarter? Assess whether new products, services, or initiatives are gaining traction.

Apple Inc. highlighted several key business drivers for Q1 2025, emphasizing the success of its product and services portfolio. The company achieved all-time revenue records across multiple regions, including the Americas, Europe, Japan, and Asia Pacific, with total revenue reaching $124.3 billion, up 4% year-over-year. Services revenue also set an all-time record at $26.3 billion, growing 14% year-over-year, driven by increased customer engagement and over 1 billion paid subscriptions.

with her children and grandchildren. These are the kind of stories that remind us of how profoundly important our work is, and it drives us to innovate each and every day. At Apple, the future is full of promise and potential. We're always searching across a world of possibilities, finding those places where we can do the most good and putting all of our energy and ingenuity into making something special. With that, I'll turn it over to Kevin.

Kevan Parekh: Thanks Tim, and good afternoon everyone. I'm going to cover the results for the first quarter of our fiscal year. We are very pleased to report an all-time high for revenue with December quarter revenue of $124.3 billion, up 4% year-over-year. We achieved all-time revenue records in the Americas, Europe, Japan, and rest of Asia Pacific and grew in the vast majority of markets we track. Products revenue was $98 billion, up 2% year-over-year, driven by growth from iPad and Mac. Thanks to our incredible customer satisfaction and strong loyalty, our installed base of active devices reached an all-time high across all products and geographic segments and is now over 2.35 billion active devices. Services revenue reached an all-time record of $26.3 billion, up 14% year-over-year. We grew in every geographic segment and achieved all-time records in both developed and emerging markets. Company gross margin was 46.9% at the high-end of our guidance range and up 70 basis points sequentially, primarily driven by favorable mix. Products gross margin was 39.3%, up 300 basis points sequentially, primarily driven by favorable mix and leverage. Services gross margin was 75%, up 100 basis points sequentially, primarily driven by mix. Operating expenses of $15.4 billion landed at the midpoint of our guidance range and up 7% year-over-year.

Figure 2: Captide Sentiment Analysis and Instant Audit features

2.2.3 The Cost Factor: Analyzing the expenses behind Captide

Under the current parsing model, each query to retrieve data from a 10-K or 10-Q file has a fixed cost of approximately EUR0.06 —regardless of how much information is extracted. For instance, obtaining only Apple's Q2 2024 revenue figure (a single data point) has the same EUR0.06 cost as extracting all other structured information from Apple's 10-Q for that quarter (such as net income, earnings per share and all additional metrics).

[2] See Figure 2.

After the intial query, updating data quarterly has a cost of EUR0.01 per company. Assuming there are around 5500 public companies in the US as of 2025, the cost magnitude of maintaining the data infrastructure is almost negligible!

Because the cost is both low and trending downward, real-time data extraction across a large universe of firms becomes more accessible for all market participants, increasing efficiency and transparency.

2.3 Predictive Modelling

After reviewing data extraction, the next logical step is to use the information towards some specific task. That is where data extraction and predictive modelling link together. Although the latter has been around for decades, Big Data and AI have made it more accurate and flexible than ever, so it now plays a key role in many industries.

2.3.1 Recent Innovations in Predictive Modelling

Simply put, predictive modelling finds patterns in historical data so it can anticipate future outputs from fresh inputs. How do alternative asset managers apply them? Which models do they use?

This firms have used machine learning for decades, mainly in statistical arbitrage and factor investing. Firms like AQR Capital Management and Two Sigma rely on models such as Ridge and Bayesian Ridge for risk management. Citadel and others use decision tree ensembles like Random Forest Regressor and Gradient Boosting. Recently, Man AHL has explored reinforcement learning for portfolio optimization, while Gaussian Processeshave improved volatility forecasting. However, due to regulatory constraints, many asset managers still prefer simpler, more transparent models for risk assessment.

Finally, deep learning models like Transformers are becoming very popular. These process data by understanding patterns in text and numbers. Although used in academic research and asset management, they are still less accessible to amateur investors.

2.3.2 Usual steps to build a predictive model

The procedure to build a predictive model work as follows:

- **Building/training the model:** Around two-thirds of the historical data are used to let AI figure out the patterns. With this, the model is created and consolidated.
- **Testing/validating the model:** The remaining historical data can be used to cross-check the model's accuracy. Firstly, only the inputs are provided and the model is asked for predictions; these are later compared with the actual historical values. Further analysis can be conducted in this phase, such as determining if there are any irrelevant inputs that can be eliminated.

With these steps, a model will be created, knowing how reliable its predictions are.

As any mathematical tool, the applications of this technique are very diverse, even if we only focus on the financial or managerial world. For example, predictive modelling could be used to forecast sales, logistics issues, the performance of an ad, etc. Of all these implementations, we will focus in this thesis on what is probably the most notorious financial application: predicting stock returns.

A model with such objective will be built in chapter 3.

3. APPLYING AI – PREDICTIVE MODELLING FOR STOCK PERFORMANCE

The key question in the realm of stock return prediction is: which inputs will best help AI make the most accurate predictions?

To answer this, we will review past research to identify relevant financial metrics and then introduce several new variables to see if they improve the model's accuracy and whether they are genuinely relevant.

3.1 Feature Selection: Key Metrics for Financial Valuation

Starting from the well-known Fama and French five-factor model as a baseline, as well as looking at more recent research papers that have added interesting points of view to it, a list of five key metrics can be created to explain the expected returns of individual stocks.

Although the most accurate procedure would be to average all these metrics over the year, we use December 31 values for simplicity. This approach reflects a common real-world practice of using year-end data to forecast returns for the following year.

The five key metrics for valuation, as per the FF5 model, are:

1. **Market risk premium (to explain the market factor):**
   It explains the difference in average returns between stocks and risk-free bonds. Although this is clearly a backward-looking metric, it yields information about the recent behavior of the stock market as a whole, which may be key for forward-looking models. [9]

2. **Market capitalization (to explain the size factor):**
   As we know from the three-factor model of Fama and French (1993, 2015), small-cap companies tend to regularly outperform the overall market. Any indicator regarding size would be suitable for capturing the effect of this factor, but we have decided to use market capitalization due to its relevance and ease of access. [9]

3. **Book-to-market equity ratio (to explain the value factor):**
   Also given by the three-factor model itself, this metric accounts for the fact that high-value firms tend to regularly outperform the overall market. In this case, the most relevant indicator is the book-to-market equity ratio, which will be computed and used for the model. [9]

4. **Gross profits-to-assets (to explain the profitability factor):**
   The same authors of the three-factor model, Eugene Fama and Kenneth French,

developed a five-factor model in 2014, which proved to perform better than the three-factor approach. One of the two newly added variables was the profitability factor. [8] [13] [17]

Although Fama and French realized that a company's profitability would be perfectly explained by its operating profitability. In their paper, it is defined as: "annual revenues minus cost of goods sold, interest expense, and SG&A, all divided by book equity at the end of the previous fiscal year", but they considered gross profits-to-assets to be a good proxy. Hence, we decided to use LTM profits-to-assets for simplicity. [8]

5. **Investment (to explain how conservative/aggressive a firm is):**
The second of the two newly added factors in the five-factor model portrays how much investment the company is undergoing. Fama and French colloquially referred to this term as how conservative or aggressive the firm was. In line with the Fama and French approach, we capture this metric by dividing the average total assets of the last fiscal year by the average total assets of the previous year. [8] [13]

3.2 Additional Metrics to be Included

The above-mentioned key factors for valuation are enough to predict individual stock returns (on average) if we believe the five-factor model to be the "true valuation model". However, AI can detect irrelevant factors and discard them automatically, so we feed the model as many variables as possible to not miss on the upside (given that the downside is limited).

However, the variables that are almost identic are discarded. For instance, "Investment" and "CapEx" essentially convey the same information, so only one of those is included.

The additional metrics that were used in this study were:

- P/E ratio
- LTM revenue growth
- LTM net income growth
- Debt-to-assets ratio
- Cash ratio
- Dividends issued as a percentage of share price
- LTM inflation rate
- Yield-to-maturity of 5-year US Treasury bond

- Share price
- LTM stock return

3.2.1 Excluded Metrics

We would like to highlight a few other metrics that could strengthen our model but were excluded due to limited resources.

- **News tone:** Previous studies have analyzed thousands of news articles, searching for specific keywords or expressions to determine whether the media coverage was positive or negative. Although the most notable research has appeared in politics (Gentzkow, M. and Shapiro, J., 2010), due to the media's impact on voters, the same could be done for individual companies [12]. By assessing the tone of articles mentioning a company, one could infer whether overall news coverage is positive or negative, potentially influencing investors. However, performing this analysis thoroughly is a substantial research task in its own right, so we have chosen to exclude it.

- **Industry-specific metrics:** Industry metrics such as Netflix's subscriber growth or Walmart's revenue per available square meter (RevPAM) are powerful forecasting tools because they measure firms' performance relative to closest competitors, which indeed gives hints about future valuation. However, using these specialized metrics requires a much larger dataset. Assigning a standardized set of metrics to every category of firm is really complex and increases the amount of data needed exponentially.

- **Industry classifications:** Although we include these classifications in our dataset[3], they are not used in the predictive section because the AI models used cannot directly process categorical data. Converting these classifications into numerical form would also require thousands of rows per sub-industry, potentially expanding the necessary database by a factor of 10–20x. We specifically encourage future researchers to build on our work by comparing a stock's P/E ratio to the average P/E ratio within its industry.

3.3 Database Collection

In order to train and deploy our AI model, we built a database where each row represents one year of financial data for each company studied. This structure allowed us to track

[3] The present study uses the Global Industry Classification Standard (GICS) to categorize the stock market into 11 sectors, 25 industry groups, 74 industries, and 163 sub-industries.

how companies perform over time and compare them with each other. Below is a heading of the dataset:

Company Name	**Ticker**	**Year**	**Industry**	**Share Price**	**...**
Apple	AAPL	2020	Consumer Electronics	129.7516022	...
Microsoft	MSFT	2021	Software - Infrastructure	327.8180237	...
...	...	...	...	...	...

Table 1: Partial structure of the dataset

The following subsections explain all the specificities and steps to complete the study:

3.3.1 Data Scope

The preset experiment is limited to U.S.-listed companies. While adding other markets would have increased the dataset size, always a good thing when training AI models, the US stock market already provides a great balance between the size of inputs and the ease of accessibility to the necessary financial information providers.

Clearly, however, this geographical constraint becomes a limitation for out-of-sample testing for the proposed AI model. We would be able to draw more meaningful conclusions if the model was tested on European or other foreign markets equities, which we strongly recommend future researchers to include in their scope.

A good remark to make is, some foreign companies indeed trade in the US market, and for those companies, all data was standardized in U.S. dollars to avoid currency-related inconsistencies.

3.3.2 Data Sources

The database was constructed using both Yahoo Finance and Alpha Vantage APIs, accessed via Python. On the one hand, Yahoo Finance provided stock market data, such as dividends, market capitalization, and industry classifications, whereas Alpha Vantage supplied financial statement data, such as debt and cash levels.

Premium subscriptions were required for both platforms to ensure access to the most extensive historical data available.

3.3.3 Dataset Timeframe

The study covers an 18-year span, from the end of 2005 to the end of 2023. During this time frame, it is essential to include any company that was publicly traded for at least a year to avoid survivorship bias. That includes not only firms still listed in 2023 but also those that delisted, went bankrupt, merged, or otherwise disappeared from the market.

If our sample only looked at those firms trading in 2023, the worst performing businesses would fall off the radar, artificially inflating the results towards an overly optimistic market performance. In practice, this meant building a complete list of all ticker symbols traded at any point in those 18 years —making sure to capture the "survivors" and the "non-survivors".

However, it is also worth noting that from 2005 to 2008, the information is not as comprehensive, so there may be some inevitable gaps in the data during that initial period.

3.3.4 Raw Data Fetching

Given that the AI capabilities of Captide are not available at large scale yet (only accessible company by company), the data extraction process was performed via API through three main steps:

1. **Identifying Tickers for Public Companies:** A comprehensive list of tickers representing 17,977 companies that were publicly traded in the United States from 2008 to 2023 was obtained from the WRDS Database. First, the tickers listed on the NYSE for each year were downloaded. Then, they were filtered by the specified year range and deduplicated so that each ticker appeared only once. The complete list is available in CSV format in the Appendix.

2. **Extracting Company Data:** Company data was gathered using Python code which used functions to fetch financial variables for each stock ticker. This process iterated through every company and collected data for each available year, using Yahoo Finance and Alpha Vantage as sources:

    - Yahoo Finance provided stock market data, including share prices, market capitalization, and dividends.
    - Alpha Vantage was used for detailed financial data, such as income statements and balance sheets.

    Data points were aligned to the end of each year, and any unavailable information was recorded as "NA" to maintain consistency.

3. **Appending Macro-Level Data:** Macro-level variables were sourced from the following:

    - Inflation rates: Bureau of Labor Statistics
    - U.S. Treasury yields: Federal Reserve Bank of St. Louis

    After completing data collection, the company-specific and macro-level data were merged on the Ticker and Year columns, creating a unified dataset in CSV format.

3.3.5 Pre-Cleaning

The code was adapted to pre-clean the data to ensure accuracy and usability for calculations. The following steps were applied:

1. Missing or invalid values (e.g., infinite values or results from division by zero) were replaced with NaN and subsequently removed.

2. Columns were converted to numeric formats to facilitate proper computations.

After pre-cleaning, the dataset contained a total of 26,616 rows.

3.3.6 Derivations

The remaining financial and economic metrics were calculated using the following predefined formulas:

- **Debt-to-Assets Ratio:**

$$\text{Debt-to-Assets Ratio} = \frac{\text{Long Term Debt} + \text{Short Term Debt}}{\text{Total Assets}}$$

- **Gross Profits to Assets:**

$$\text{Gross Profits to Assets} = \frac{\text{Revenue} - \text{COGS}}{\text{Total Assets}}$$

- **PE Ratio:**

$$\text{PE Ratio} = \frac{\text{Market Capitalization}}{\text{Net Income}}$$

- **Investment Growth:**

$$\text{Investment Growth} = \frac{\text{Total Assets}_t - \text{Total Assets}_{t-1}}{\text{Total Assets}_{t-1}} \times 100$$

- **Revenue Growth:**

$$\text{Revenue Growth} = \frac{\text{Revenue}_t - \text{Revenue}_{t-1}}{\text{Revenue}_{t-1}} \times 100$$

- **Next Year Returns:**

$$\text{Next Year Returns} = \text{Returns}_{t+1}$$

- **Dividend Payout:**

$$\text{Dividend Payout} = \frac{\text{Dividends}}{\text{Share Price}}$$

3.3.7 Dataset Summary: A Snapshot of the Data

After cleaning, the database provides a snapshot of the U.S. stock market. This subsection verifies whether it reliably represents the broader market.

Table 3.3.7 summarizes the market capitalization distribution of our dataset for 2022, compared with approximate averages for the U.S. public equity market over the last decade, as reported by Morningstar [16].

	Our Dataset (2022)		U.S. Market Avg.	
Metric	**Count**	**% of Total**	**Count**	**% of Total**
Large-Cap (incl. Mega)	544	*19.4%*	600	*10.9%*
Mid-Cap	758	*27.0%*	1500	*27.3%*
Small-Cap (incl. Micro)	1504	*53.6%*	3400	*61.8%*
Total	**2804**	*100.0%*	**5500**	*100.0%*

Our dataset includes 2804 companies in 2022, with approximately 20% large-cap, 27% mid-cap, and 53% small-cap or micro-cap. In contrast, the broader U.S. market, with an average of 5500 companies, has 11% large-cap, 27% mid-cap, and 62% small-cap. While our dataset is slightly biased toward large-caps due to data availability, it still captures a broad range of U.S. equity performance.

Figure 3 represents a snapshot of the database for the year 2019, illustrating the market capitalization (size) and returns (colour) of companies across sectors. For reference, the S&P 500 gained 31.5% that year.

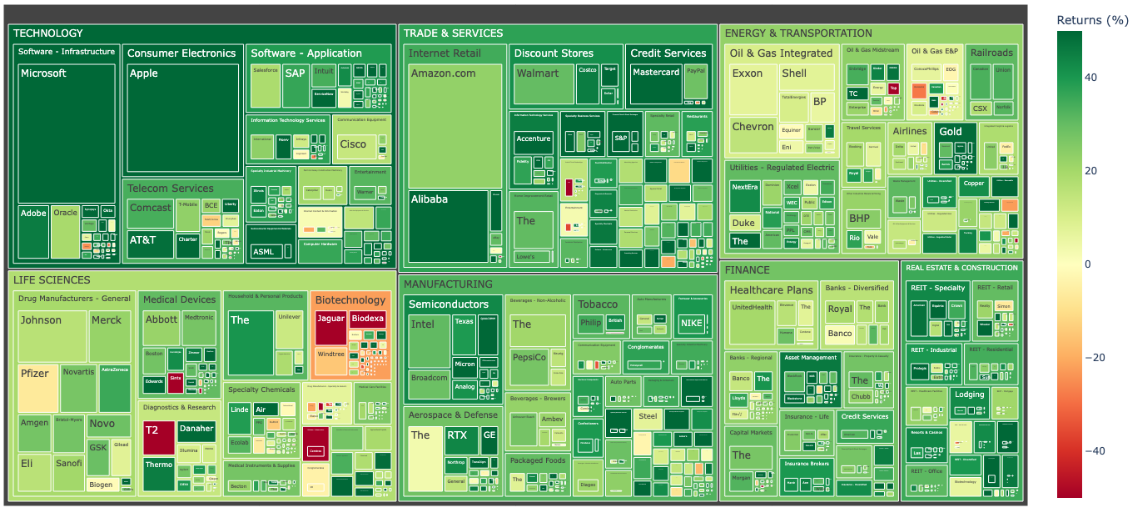


Figure 3: Treemap of U.S. companies grouped by sector, sized by market capitalization, and coloured by stock performance (2019).

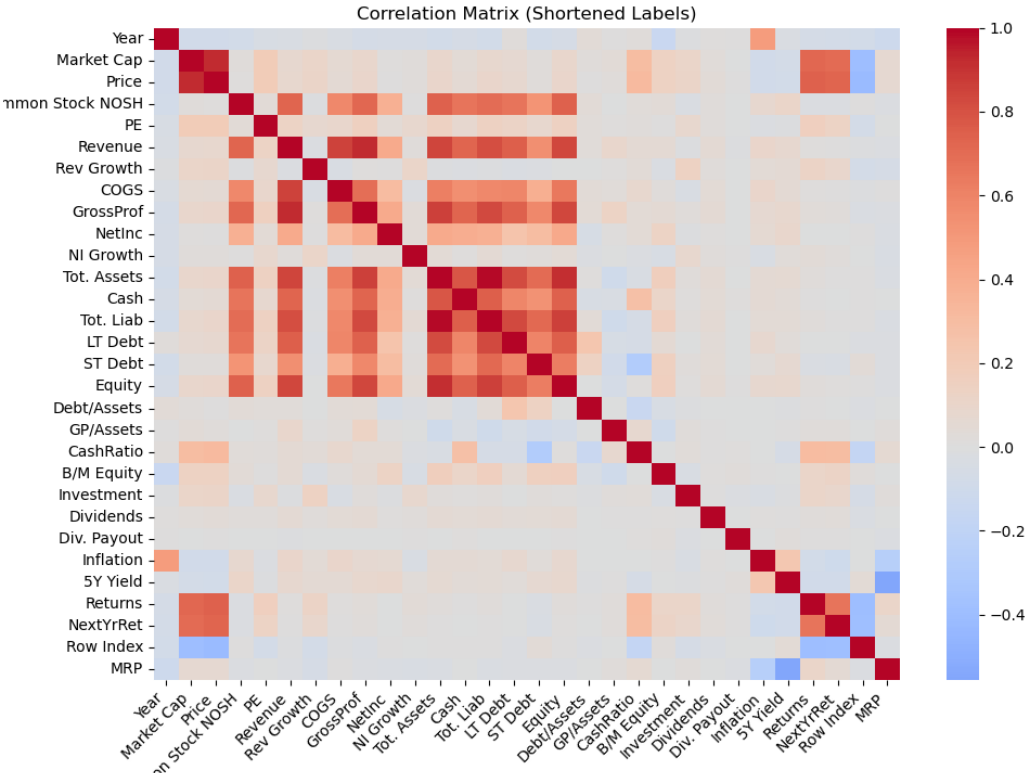


Figure 4: Correlation Matrix showcasing the experiment's variables

Also, a correlation matrix was performed to check the relationships among variables.

- **Strongly Correlated "Scale" Variables:** Market Cap, Price, Revenue, Net Income, Total Assets, Liabilities, and Equity tend to be positively correlated. Larger firms often have higher revenues, assets, and market capitalization, which in turn boosts share price and other "size" metrics.

- **Strongly Correlated Debt Variables:** Long Term Debt, Short Term Debt, Total Liabilities, and Debt to Assets show strong intercorrelations, as larger companies typically have higher debt in absolute terms, with these liabilities moving together with firm size.

Although some of these variables could have actually been removed to reduce redundancy, they were retained to test whether the AI models could handle extra "noise" features that standard regression models often lack.

3.4 Raw AI: Which models are being used?

Once the dataset was ready[4] the AI models were developed with Python. This process involved multiple steps:

[4] Python codes available in Appendix A.0

3.4.1 Technology used

When it comes to machine learning technology, there are countless models and mathematical procedures that work best in specific conditions. From this pool of possibilities, the following 28 models were selected for the experiment, which are described below in really simple terms: [18]

Category 1: Linear Models Linear models assume a linear relationship between inputs and the target variable, and therefore they are mostly regression-based.

- **Ridge** – It shrinks big coefficients in the regression to avoid overfitting.
- **Lasso** – Similar to Ridge, but brings coefficients directly to zero instead of just reducing them partially.
- **Elastic Net** – A combination of Lasso and Ridge regression: it both shrinks and/or eliminates some of the features.
- **Lars (Least Angle Regression)** – Builds the model step by step, adding only one feature at a time.
- **Lasso Lars** – Same concept as Lars, but also removes unnecessary features using Lasso's L1 regularization.
- **Orthogonal Matching Pursuit** – Another model that takes features one by one (like Lars) but uses a "greedy" method to always choose the most useful feature at each step.
- **Bayesian Ridge** – Similar to Ridge, but adds uncertainty to the coefficients. In very plain terms, Bayesian Ridge says: "I believe this is the effect of this feature, but I am just X% sure".
- **ARD Regression (Automatic Relevance Determination)** – A "smarter" version of Bayesian Ridge. It decides how important each feature is by measuring its variance, and ignores the ones that do not have a relevant impact.
- **SGD Regressor** – Learns step by step using small pieces of the data at a time, making it a big asset for large datasets in terms of efficiency.
- **Passive Aggressive Regressor** – Used when data arrives only bit by bit: If a prediction is wrong, it radically updates itself to correct the erros in the following rounds.

- **Theil Sen Regressor** – Instead of using all datapoints to fit the regression line, Theil Sen tries many possible lines and picks the median slope. It is a great model when there are lots of outliers.

- **RANSAC Regressor (Random Sample Consensus)** – RANSAC ignores outliers completely by testing many small subsets of the data and choosing the model that fits the majority best.

Category 2: Kernel-Based Models These models transform the data into a new space where patterns are easier to find. In other words, they change the dimensionality of the data to look for nonlinear patterns.

- **Kernel Ridge** – A Ridge regression model that uses kernel functions (curves instead of straight lines) to find more complex patterns.

Category 3: Support Vector Machines (SVM) These models try to find the best boundary between data points in a high-dimensional space.

- **SVR (Support Vector Regression)** – Predicts numbers (and not categories) by focusing only on the most important points (support vectors) and allowing some tolerance.

- **Nu SVR** – Similar to SVR, but gives more control over how many support vectors are used using a hyperparameter (ν).

- **Linear SVR** – A simpler and faster version of SVR that works with straight-line relationships and is powerful for large datasets.

Category 4: Nearest Neighbors As the name indicates, the model predicts output based on affinity to the closest data points available.

- **K Neighbors Regressor** – Predicts the target value by averaging the values of the k-nearest neighbors (with many values of k being tested).

Category 5: Gaussian Processes Gaussian Processes (GP) model uncertainty distributions to make several probabilistic predictions.

- **Gaussian Process Regressor** – It does not just guess a number, but rather gives a range and a confidence interval. It works better on small datasets

Category 6: Cross Decomposition Models These models find relationships between two sets of variables.

- **PLS Regression (Partial Least Squares Regression)** – Simplifies high-dimensional data and finds patterns between inputs and outputs, even when features are highly related among each other (multicollinear).
- **PLS Canonical (Partial Least Squares Canonical Correlation Analysis)** – A version of PLS that looks for the strongest correlations between two sets of data.
- **CCA (Canonical Correlation Analysis)** – Finds common structure between two datasets by creating new variables that are as correlated as possible (projections).

Category 7: Decision Trees & Ensemble Methods As its name indicates, Decision Trees split data into branches and combine many models to see which "path" is best.

- **Decision Tree Regressor** – Splits data into yes/no questions, forming a tree that leads to a prediction.
- **Bagging Regressor** – Builds many trees on random parts of the data and averages their results.
- **Random Forest Regressor** – Almost identical to the Bagging Regression, but also picks random features at each split, making it even more stable.
- **Extra Trees Regressor** – Goes further by picking splits randomly, making the forest more diverse and reducing variability.
- **Ada Boost Regressor** – Sequentially trains weak models, each focusing more on the data points that were hard to predict.
- **Gradient Boosting Regressor** – Similar to AdaBoost, but uses gradients (the respective errors) to improve the model at each step.

Category 8: Neural Networks These models are inspired by how the human brain works, and can learn very complex patterns from data.

- **MLPRegressor (Multi-Layer Perceptron)** – A basic neural network with layers of "neurons" that learn to predict numbers by adjusting weights in training during thousands of iterations.

All of these models have an open-source environment in the Scikit-learn package in Python, which was used to perform the analyses.

3.4.2 Procedure

The first step was to compare the 28 above-mentioned AI models.

To do such thing, the dataset was divided into training data and testing data. The training data (around 67% of the rows) was used to build the AI model, while the testing data (the other 33% of the rows) was used to predict the target variable using only the input features. Each prediction was then compared to the actual value. This allowed to compute a coefficient of determination used to compare all models and identify the best-performing one.

More precisely, through the code "1_Best_Model.py", which can be found in Appendix A.1., this process was repeated multiple times for each AI model, giving a range of the R^2 coefficients. Figure 5 was the output of such analysis, with a selection of the 13 models with the best predictive performance:

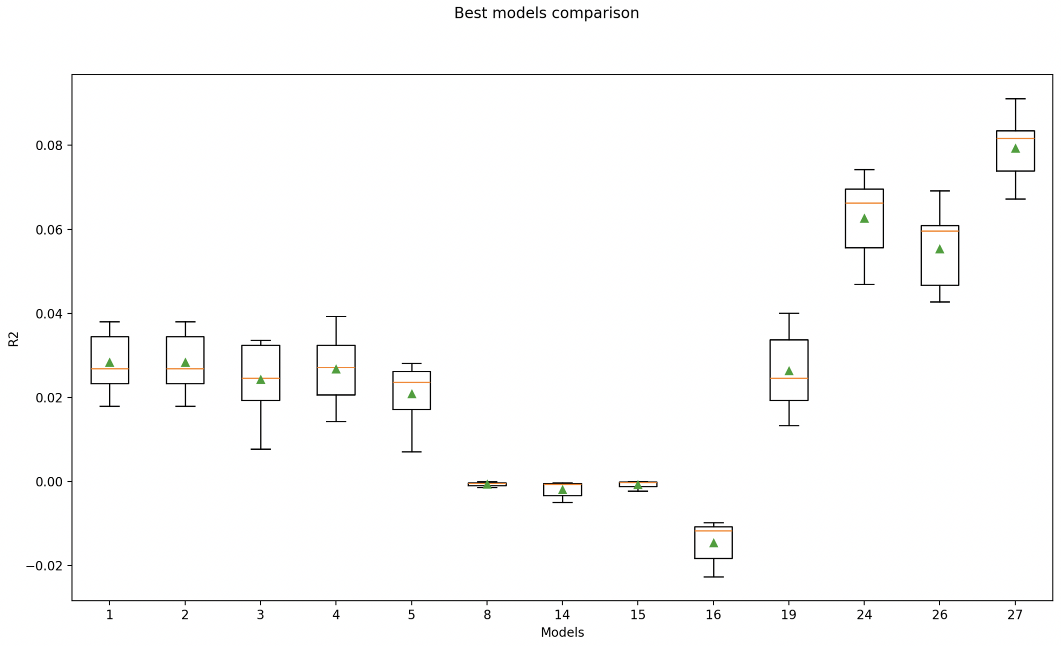


Figure 5: R^2 coefficients for raw data predictive models.

Two key conclusions were drawn from the analysis:

- First, Decision Trees (Category 7) outperformed all other model categories by a significant margin, showing the highest predictive power for our dataset.

- Second, inside that category, model number 27, Gradient Boosting Regressor, delivered the best results. This method works by building many small decision trees, where each new tree focuses on fixing the errors made by the previous ones. For example, if one tree slightly mispredicts the return of a certain stock, the next tree will focus on correcting that specific mistake, improving accuracy step by step. [11]

Furthermore, some parameters inside the model could be customised. In the code “2_Parameter_Optimization.py”, found in Appendix A.2., a great variety of different set-ups were tested and compared with the target.

Once the best parameters were found, the next step was to measure the relevance of all independent variables, so that negligible ones could be erased.

The results[5] are shown in Figure 6.

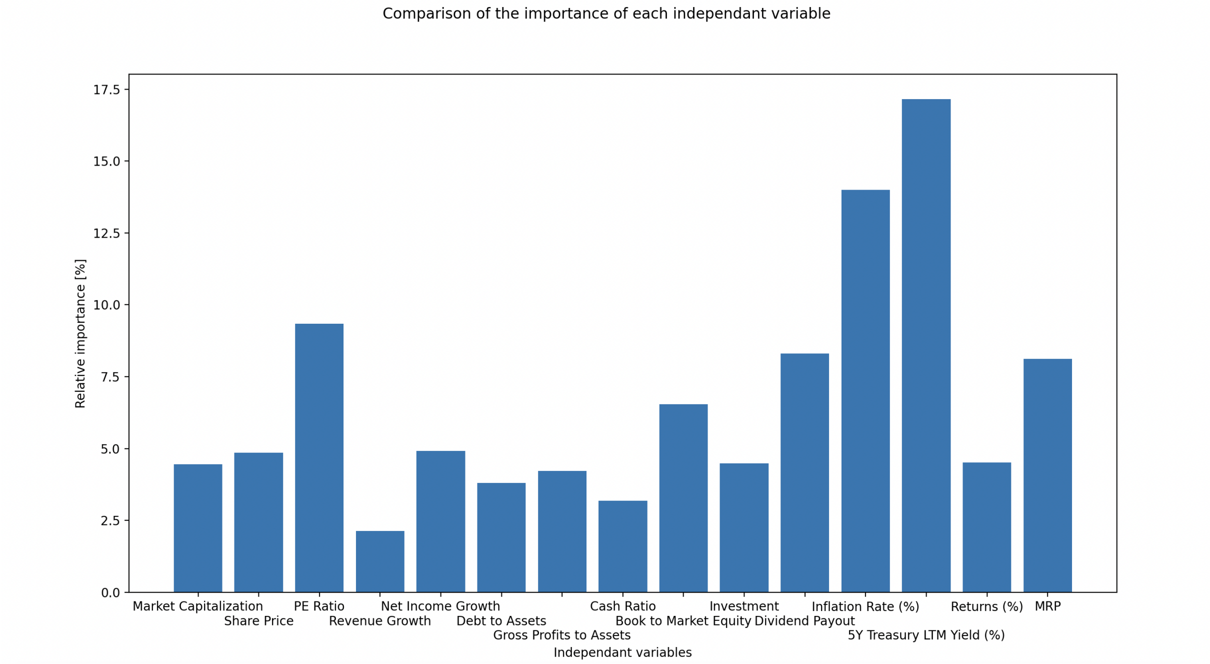


Figure 6: Input variables’ importance.

The numerical data behind the chart indicates that no input variable has an importance lower than 2%, meaning they are all relevant for the model.

Taking a deeper look, while the macroeconomic variables show high importance, it is surprising to see the five dimensions included in the FF5 model by Fama and French (Market Capitalization, Book-to-Market Equity, Gross Profits to Assets, Investment and MRP) not considered more relevant than most of the other variables.

How can this be explained? Firstly, one could argue that this discrepancy comes from the fact that FF5 is looking in cross-section, while this AI model is forward-looking, so the FF5 factors are relevant for explaining difference between the expected return of companies at a given point in time but not useful when it comes to predictability. But in this case, the explanation lies in structural problem: the R^2 coefficient turns to be very poor, showing that the study is actually not significant.

This final computation of the R^2 coefficient was done using the code “4_Final_Test.py”, found in the Appendix A.4., and yielded the following results:

[5] See the code `\3_Importance.py"` used to obtain this figure in Appendix A.3.

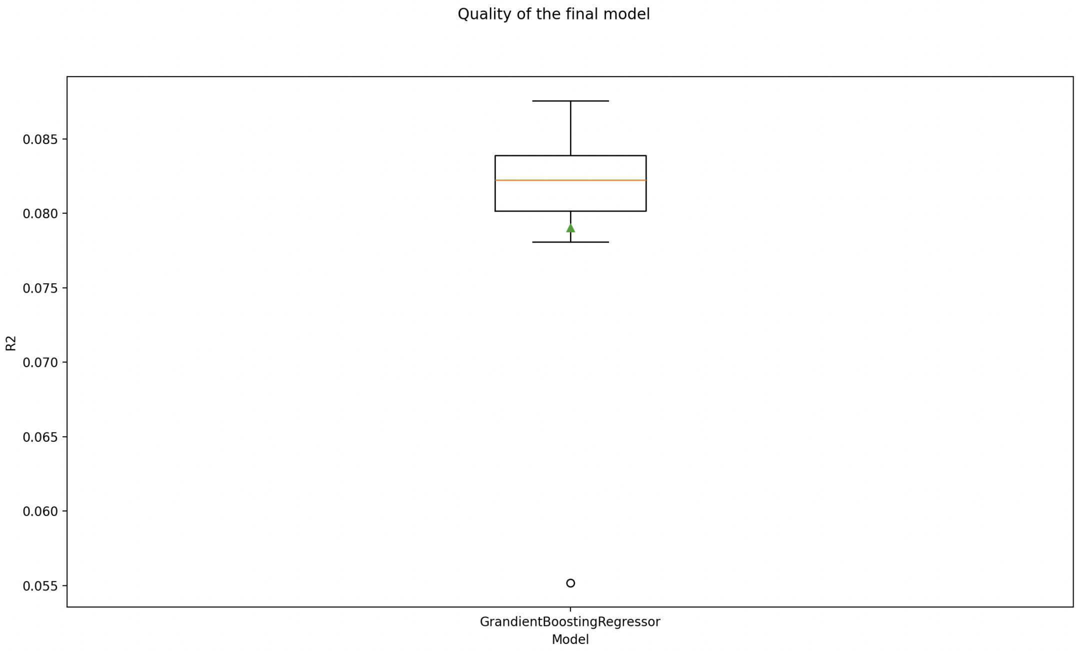


Figure 7: R^2 coefficient for the raw model.

After being tested several times, the R^2 coefficient of the final model varies between 0.55% and 0.85%, with an average R^2 of 0.8%.

These weak results, which are graphically showcased in Figure 7 led us to try new tools and data. This led to a second model, more consistent but with some limitations. We explain it in the next section: the “Logarithmic Model”.

3.5 Logarithmic Model

The process to obtain this second model is quite similar to the one already explained. The major difference lies in both the processing of the dataset before it is fetched to the model and the results that the model yields. Hence, we divide this chapter into these two sections.

3.5.1 Pre-Processing: Refining Data for the Log Model

After looking at the disappointing results of the first model, we played around with the inputs to see if there was a way to make them more “understandable” for the AI models. For example, it is proven that many AI tools work better with standardized data (zero mean and unit variance), so this is one of the things we tried, along with logarithmic scalability. These structured pre-processing also favours cleanliness and readiness for analysis. [19]

All these modifications to the original data are kept in the code “Filter.py”, which can be found in Appendix A.5., with which the original dataset file is transformed into other dataset files, each with different alterations.

Among all these altered dataset files, the one which yielded the best results included the following changes:

- First, logarithmic transformations were applied to selected columns to reduce skewness and handle wide value ranges. The formula used was:

  $$\log(x) = \ln(1 + x)$$

  To ensure valid transformations, non-positive values were removed beforehand, and return-related columns were adjusted by adding 100 to make all values positive.

- Secondly, all columns were normalized to have a mean of 0 and a standard deviation of 1 using the formula:

  $$z = \frac{x - \mu}{\sigma}$$

- Third, rows with potential outliers were identified and removed. Two conditions were applied: for columns such as "Revenue Growth" and "Returns (%)", rows were flagged if their absolute value fell outside the range of the 5th or 95th percentile.

This preprocessing pipeline enhanced the quality and usability of the data, but greatly reduced the amount of data, obtaining a dataset of roughly 8.500 rows. We will see and explain in a few paragraphs that this yielded outstanding results, while being at the same time a trade-off for having scraped out many companies.

3.5.2 Test and Analysis

Using this altered dataset with logarithmic and normalized columns, by following the previously mentioned python codes with minor adaptations, the model that performs best is called SVR (Support Vector Regression), this time yielding an R^2 coefficient between 60% and 66%, an outstanding improvement compared to the "Raw Model". See Figure 8 for the corresponding graphical representation.

Regarding the model itself, SVR works by mapping data into a high-dimensional space. It can find non-linear patterns through a smart trick (called the "kernel trick") that lets it work with complex data shapes without doing heavy calculations. [1]

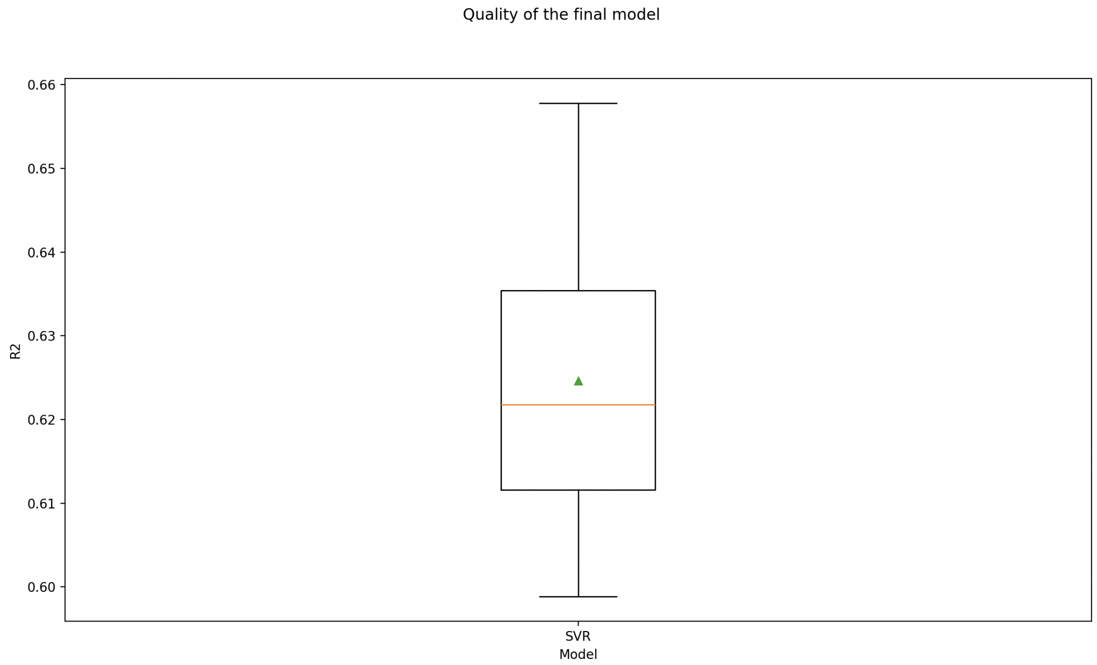


Figure 8: R^2 coefficients for the logarithmic predictive model.

One limitation of SVR is its lack of transparency. Since it is based on SVR, a non-tree model, we cannot measure the importance of each variable. This means we do not know which features contribute most to the predictions.

Alternatively, the same model (SVR) was run with only the FF5-related variables as inputs (market capitalization, book-to-market equity, gross profits-to-assets, investment and MRP). This way, the relevance of those specific variables can be roughly approximated. Given that an R^2 coefficient of around 54% was obtained (see Figure 9) the dimensions highlighted by Fama and French prove to be very relevant for explaining and predicting stock returns. In fact, they explained almost all variability in the model.

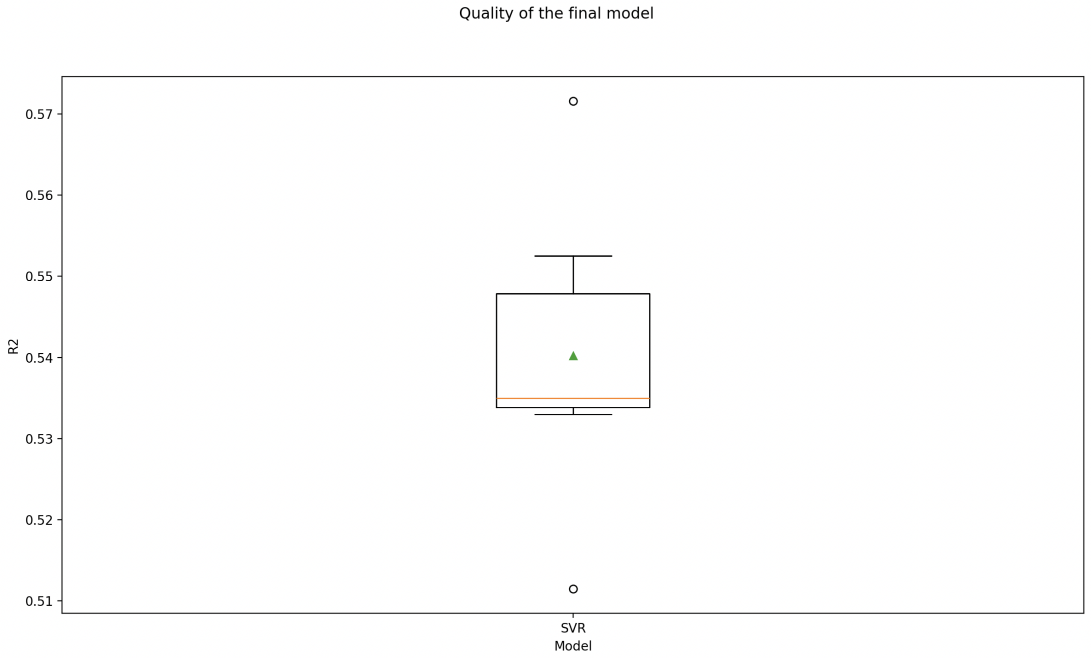


Figure 9: R^2 coefficients for the SVR model with only FF5-related inputs.

Let us recapitulate and finish this section by being mindful again of the trade-off that we had to do to achieve this result: this R^2 coefficient comes with the limitation of only considering companies that fell inside the range of certain percentiles for many categories, leaving out of this model all the rest, which amounted to around 75% of firms. In practical terms, this means that the model can only be applied to a limited set of companies.

For example, if an investor used this model to predict the stock returns of company ABC, there would be two possible outcomes:

- If ABC fell within the percentiles the model was trained on, the prediction would likely be reliable.
- If ABC fell outside those percentiles, the model would indicate that providing a reliable prediction is not possible.

Although the model is less transparent, we believe its better performance outweighs its drawback. It is preferable to have an accurate model for a limited amount of companies, rather than one that covers a wide range of companies but produces unreliable results.

4. BONUS SECTION: BENCHMARKING WITH REGULAR OLS REGRESSION

After developing and applying the 28 AI models to our dataset, it made sense to benchmark it to another model in order to check its accuracy.

What models could we choose? The FF5 model is typically used for explaining differences between companies, and not for prediction. Therefore, a simple OLS regression was used instead, with time fixed effects to account for macroeconomic factors.

4.1 OLS Regression Model Procedure

In essence, the procedure of this experiment was exactly the same as with the AI models, but this time using a basic OLS regression.

First, the 70% of rows were chosen randomly in order to compute the regression parameters. More specifically, a linear regression was performed in which the dependent variables were the returns at time t+1 (following year's returns) and the independent variables were all remaining numeric columns and excluding `Next Year Returns (%)`).

The resulting regression output provided coefficient estimates, standard errors, t-values, p-values, and confidence intervals. The next step was to apply those coefficient estimates to real data (the 30% of rows remaining).

4.2 Model Diagnostics and Discussion

Overall, the model explained approximately 56.7% of the variability in the returns ($R^2 = 0.567$), and the F-statistic was 317.6 (p-value = 0.00), indicating that the model is statistically significant.

Table 2: OLS Regression Results

OLS Regression Results

Dep. Variable:	Next Year Returns (%)	R-squared:	0.567
Model:	OLS	Adj. R-squared:	0.565
Method:	Least Squares	F-statistic:	285.4
Date:	Wed, 19 Mar 2025	Prob (F-statistic):	0.00
Time:	16:31:38	Log-Likelihood:	-5912.9
No. Observations:	5902	AIC:	1.188e+04
Df Residuals:	5874	BIC:	1.207e+04
Df Model:	27	Covariance Type:	nonrobust

Table 3: OLS Regression Results & Feature coefficients

	coef	**std err**	**t**	$P > \|t\|$	[0.025, 0.975]
const	4.0180	5.717	0.703	0.482	[-7.188, 15.225]
Year	-0.0019	0.003	-0.668	0.504	[-0.007, 0.004]
Market Capitalization	0.1935	0.022	8.614	0.000	[0.149, 0.238]
Share Price	0.4778	0.023	20.841	0.000	[0.433, 0.523]
Common Stock NOSH	0.0358	0.014	2.512	0.012	[0.008, 0.064]
PE Ratio	0.0018	0.009	0.202	0.840	[-0.016, 0.019]
Revenue (in USD)	-0.0088	0.038	-0.234	0.815	[-0.083, 0.065]
Revenue Growth	0.0017	0.009	0.191	0.849	[-0.016, 0.019]
COGS (in USD)	0.0222	0.019	1.155	0.248	[-0.015, 0.060]
Gross Profits	0.0079	0.028	0.285	0.776	[-0.046, 0.062]
Net Income (in USD)	0.0217	0.011	1.992	0.046	[0.000, 0.043]
Net Income Growth	0.0123	0.009	1.410	0.159	[-0.005, 0.029]
Total Assets (in USD)	-0.1650	0.087	-1.898	0.058	[-0.335, 0.005]
Cash & Equivalents (in USD)	-0.0450	0.017	-2.629	0.009	[-0.079, -0.011]
Total Liabilities (in USD)	0.0466	0.067	0.699	0.484	[-0.084, 0.177]
Long Term Debt (in USD)	-0.0025	0.019	-0.133	0.894	[-0.039, 0.034]
Short Term Debt (in USD)	0.0357	0.014	2.586	0.010	[0.009, 0.063]
Shareholder Equity (in USD)	0.0194	0.031	0.633	0.526	[-0.041, 0.079]
Debt to Assets	-0.0016	0.010	-0.152	0.879	[-0.022, 0.019]
Gross Profits to Assets	0.0041	0.010	0.414	0.679	[-0.015, 0.024]
Cash Ratio	0.0963	0.012	8.205	0.000	[0.073, 0.119]
Book to Market Equity	0.0499	0.009	5.316	0.000	[0.031, 0.068]
Investment	0.0040	0.009	0.452	0.651	[-0.013, 0.021]
Dividends	0.0046	0.008	0.559	0.576	[-0.011, 0.021]
Dividend Payout	-0.0438	0.062	-0.703	0.482	[-0.166, 0.078]
Inflation Rate (%)	-0.0410	0.010	-3.956	0.000	[-0.061, -0.021]
5Y Treasury LTM Yield (%)	-0.0266	0.011	-2.447	0.014	[-0.048, -0.005]
Row Index	-1.254e-05	1.24e-06	-10.102	0.000	[-1.5e-05, -1.01e-05]
MRP	-0.0016	0.001	-2.125	0.034	[-0.003, -0.000]

What about the significance of each variable?

- Several predictors (e.g. PE Ratio, Revenue Growth, Cash & Equivalents) show statistically significant relationships (p-values below 0.05) with future returns.
- Some predictors (e.g. Common Stock NOSH, COGS) are not statistically significant, therefore they could be considered as noise. However, intuitively, they should indeed carry at least some information since they measure the operational and financial efficiency of a company.
- The high condition number (2.25×10^{15}) indicates that some variables are highly correlated (as demonstrated through the correlation matrix), potentially inflating standard errors.

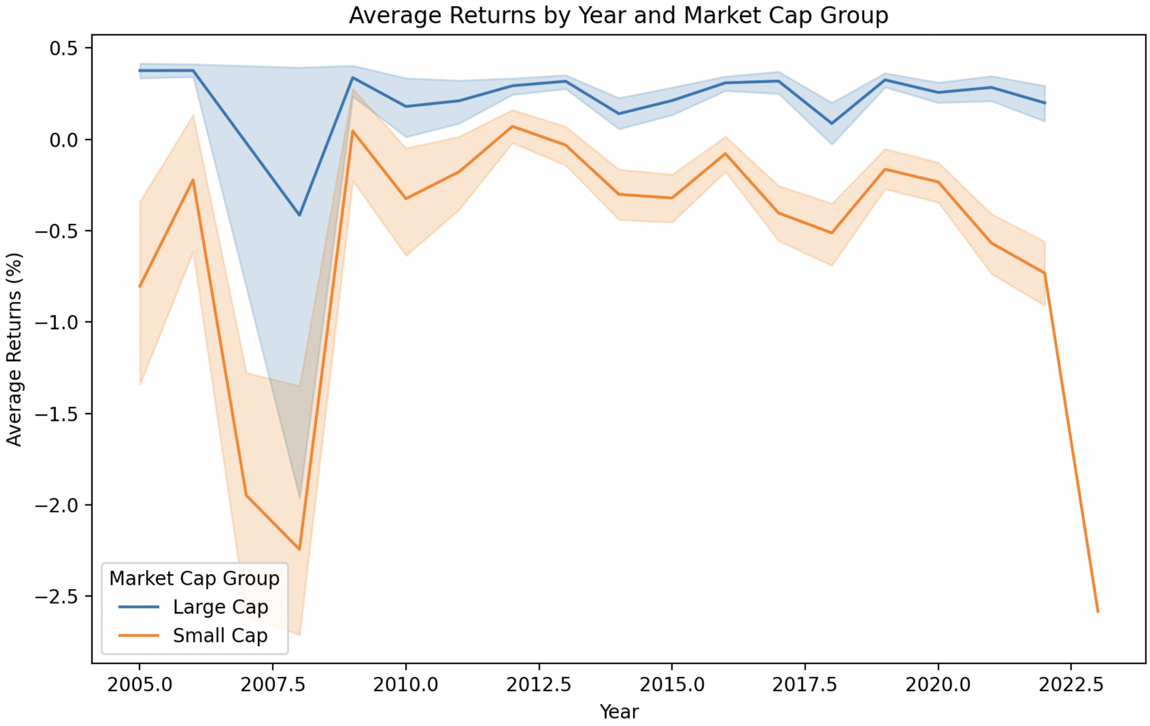


Figure 10: Difference between Large Cap & Small Cap returns

A key question remained: Why do Share Price and Market Capitalization demonstrate the highest predictive power among all variables in this case?

To investigate this, let us consider how different classes of stocks have performed over the most recent market cycle. In particular, we can compare returns of small-cap versus large-cap stocks. As illustrated by the log-normalized returns in Figure 10, large-cap stocks have demonstrated better performance over the past few years. Such outperformance can make it "easier" for a regression model to predict returns. The clearest examples are the Magnificient 7 stocks (Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia and Tesla) which have outperformed the market by a big stretch, and had great stock return autocorrelation.

But why is Share Price also very relevant? A point worth noting is that Share Price is actually highly correlated with Market Capitalization, so if one is relevant, the other one will be too. The initial correlation matrix highlights the substantial difference in stock prices between large-cap and small-cap companies. In fact, small-cap firms often have such low share prices that they are commonly referred to as "penny stocks". Hence, companies with higher valuations (large-cap) also tend to have higher share prices, creating an observable correlation between these two features.

5. POTENTIAL NEXT STEPS – COMPATIBILITY WITH CAPTIDE

After successfully implementing the AI models and benchmarking them with the classical OLS regressions, let's recapitulate on all the steps that could be taken to make this thesis more relevant:

1. **Out-of-Sample Testing:** Evaluate the model's performance on European and/or foreign market data to achieve more robust results.

2. **Larger Dataset and Additional Variables:** Expand the dataset with more observations and potentially introduce new variables as predictors. For instance, future researchers could take advantage of new technologies to include industry-specific metrics (such as REVPAR in hotels or ARPU in telecommunications). Also, leveraging news tone could be a great addition.

3. **Advanced Models:** Add new models to capture complex relationships within the data. For instance, a great surge in Transformer models is happening across the Big Tech sector. While these models demand greater computational resources and expertise, they may uncover patterns that simpler models overlook.

4. **Automate Prediction Workflow:** Build an automated data pipeline (for instance, together with the emerging Startup Captide) that processes data in real time. That way, the models could adapt to market conditions faster, having a huge competitive edge against the rest of market players.

AI will definitely play a major role in finance in the upcoming years. Our experimental model gave good results, but even better performance is needed for AI to truly lead in prediction.

For Captide's case in particular, already having a powerful product regarding data extraction, we see great synergies in offering predictive models for stock returns, as any AI predictive model would require inputs (data extraction) and AI expertise, both already mastered by Captide.

6. ENVIRONMENTAL IMPACT & BUDGET

The direct environmental impact of this thesis has been close to zero. Minimal material has been consumed and, although the code processing of some scripts took several hours, the amount of energy consumed by these computers has also been negligible.

However, it is worth noting the environmental impact of large-scale computational models and data centers themselves. For example, in the United States, data centers alone contributed approximately 31.5 million tons of CO_2-equivalent greenhouse gas emissions, representing 0.5% of the country's total emissions. [15] [21]

In terms of the budget, the total cost of this thesis has amounted to $149.97. This total cost includes:

- One month of a Gold account for Yahoo Finance ($49.99/month).
- Two months of a low-tier Premium account for Alpha Vantage ($49.99/month).
- Premium account for Captide ($200/month, reduced to $0/month due to partnership).

Our opinion is that, in academic works like this, one should try to minimize financial costs and should never rely on paying for expensive products to enhance the outcome. However, in this case Yahoo Finance and Alpha Vantage were crucial to obtain the financial dataset.

7. CONCLUSIONS

In this thesis, we investigated whether cost-effective, widely accessible AI models can reliably predict stock returns. The findings present an unconclusive answer to this question. On one hand, the AI approaches employed do outperform classical OLS regressions in terms of R^2 values, with the logarithmic model reaching predictability ratios of 60% to 66% compared to just over 50% for OLS. On the other hand, these improvements alone are not enough to justify real money investments on the model's predictions at this point.

Despite many variables showing statistical significance, none of the factors showed extremely high predictability. Although some variables showed marginally greater strength, their predictive value was likely overstated due to specific market conditions (from 2008 through 2023, interest rates were unusually low, favoring certain types of companies). The timeframe of the dataset is relevant, but its results may be limited to a certain macroeconomic framework.

In that same regard, the decision to apply logarithmic transformations and exclude firms outside the 5th to 95th percentile substantially enhanced predictability by removing noise. However, this also means the model only works well for companies very simillar to the average (roughly just over 30% of the total sample). Thus, a trade-off appears: more accurate models may have narrower applicability.

When it comes to the practical implications of the thesis, although the model's predictive capabilities approached the 70% benchmark (an industry threshold for strong forecasting models) the current 60%-66% predictability still falls slightly short of what we would consider "reliable alpha generation". Nevertheless, the insights obtained could be highly valuable for financial professionals, specially in terms of risk management.

Finally, this thesis highlights the growing synergies between data extraction tools and predictive analytics. As a partner company, Captide can combine its data collection ability with machine learning techniques to exploit market ineffícienies. This combination has immense potential and can offer valuable insights for finance professionals and researchers.

Table 4: Summary: OLS Regression vs. AI Models

Model	Dataset	R^2	Applicability	Speed
OLS Regression	Raw Dataset	0.8%	High	High
	Log Dataset	56%	Limited	High
Best-performing AI Model	Raw Dataset	1.8%	High	Medium
	Log Dataset	64%	Limited	Medium

8. ACKNOWLEDGEMENTS

We would like to express our gratitude to Alina Rosu, our tutor, for her dedication, support and guidance throughout this project. Her ideas and insights into the benchmarks of the financial industry were key in shaping our work and understanding of the field.

We also extend our thanks to Miquel Trafí (CEO at Captide.co) and Maurits Brinkman (CTO at Captide.co) for their significant contributions to our thesis. Their support included demonstrating how their platform back-end operates and granting us full access to the beta version of their product and its features. Their expertise and generosity enhanced the scope and understanding of the topic.

REFERENCES

[1] Awad, M; Khanna, R. *Efficient Learning Machines.* Chapter: Support Vector Regression (pp. 67-80). Apress Open (2015).

[2] Bahrammirzaee, A. *A Comparative Survey of Artificial Intelligence Applications in Finance: Neural Networks, Expert Systems, and Hybrid Intelligent Systems.* Neural Computing & Applications, 19, 1165–1195 (2010).

[3] Belanche, D; Casaló, L; Flavián, C. *AI-Driven Advice: Theoretical Foundations and Research Directions for Robo-Advisors.* Journal of Service Research (2020).

[4] Brinkman, M.; Trafí, M.; Noguer, M. *CRAG Framework for Multi-Document Financial Analysis.* Captide and AIFI, November 2024.

[5] Chen, Y; Li, X. *Automated Information Extraction from Financial Filings Using NLP.* In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[6] Cochrane, J. H. *Asset Pricing: Revised Edition.* Princeton University Press (2005).

[7] Daniel, K.; Titman, S. *Evidence on the Characteristics of Cross Sectional Variation in Stock Returns.* The Journal of Finance (March 1997).

[8] Fama, E.; French, K. *A Five-Factor Asset Pricing Model.* University of Chicago and Dartmouth College (Draft of September 2014).

[9] Fama, E.; French, K. *Common risk factors in the returns of stocks and bonds.* University of Chicago (September 1992).

[10] Fama, E.; French, K. *Efficient Capital Markets: A Review of Theory and Empirical Work.* Journal of Finance (May 1970).

[11] Friedman, J. *Greedy function approximation: a gradient boosting machine.* Stanford University (2001).

[12] Gentzkow, M.; Shapiro, J. *What drives media slant? Evidence from U.S. daily newspapers.* Econometrica (Vol. 78, No. 1, January 2010).

[13] Hou, K.; Xue, C.; Zhang, L. *Digesting Anomalies: An Investment Approach.* Oxford University Press (September 2014).

[14] Kelly, B. T.; Malamud, S.; Zhou, K. *The Virtue of Complexity in Return Prediction.* Swiss Finance Institute (2024).

[15] Masanet, E., Shehabi, A., Lei, N., Smith, S., Koomey, J. *Recalibrating global data center energy-use estimates.* Science, 367(6481), 984–986 (2020).

[16] Morningstar. *Tools and resources for financial analysis.*
Available at: https://tools.morningstar.es/ [Accessed: 2023].

[17] Novy-Marx, R. *The other side of value: The gross profitability premium.* Journal of Financial Economics (June 2012).

[18] Scikit-learn developers. *Machine learning in Python.* Scikit-learn. Retrieved March 13, 2025, from `https://scikit-learn.org/stable/index.html`.

[19] Scikit-learn developers. *Preprocessing data.* Scikit-learn (Website as checked on December 2024).

[20] Tsaih, R; Lin, H; Chen, M. *Forecasting S&P 500 Stock Index Futures with a Hybrid AI System.* Decision Support Systems, 23(2), 161–174 (1998).

[21] U.S. Environmental Protection Agency. *Inventory of U.S. greenhouse gas emissions and sinks 1990–2018.* Report, Washington, D.C. (2020).

[22] Vanguard. *Vanguard Digital Advisor.* Retrieved March 13, 2025, from `https://investor.vanguard.com/advice/robo-advisor`.

[23] Wealthfront. *Robo-advisor investing.* Retrieved March 13, 2025, from `https://www.wealthfront.com/robo-advisor-investing`.

APPENDIX

A. Python codes

Find HERE the python codes used for this experiment. Use the following index to find the exact name of the file you are looking for:

A.0. 0_Database_Making.py

A.1. 1_Best_Model.py

A.2. 2_Parameter_Optimization.py

A.3. 3_Importance.py

A.4. 4_Final_Test.py

A.5. Filter.py

A.6. OLS.py

B. Databases

Find HERE the databases not derived from an API used for this experiment. Use the following index to find the exact name of the file you are looking for:

B.0. Tickers_List.csv

B.1. FF5_factors_annual.csv

B.2. 2024_Industry_Classification_GICS.xlsx