



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

ADE

Facultad de Administración
y Dirección de Empresas /UPV

TELECOM ESCUELA
TÉCNICA VLC SUPERIOR
DE INGENIERÍA DE
TELECOMUNICACIÓN

Evaluating Score-Investing Methodologies: A Systematic Review of Tweenvest's Algorithm for Long-Term Stock Investing Using Descriptive Analytics and Predictive Modeling.

Integrated Bachelor's Thesis
Grado en Administración y Dirección de Empresas
Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

Author: Carlos Eduardo Domínguez Martínez

Tutor: Joaquín Martínez Minaya (GADE)

Tutor: Alberto Albiol Colmener (GITST)

External Tutor: José Tatay Sangüesa (Tweenvest)

Academic Year: 2024–2025

Valencia, September 2025

Abstract

This study investigates the effectiveness of a factor-investing methodology developed by Tweenvest, leveraging a proprietary algorithm grounded in fundamental financial analysis. The algorithm scores companies across four key factors: Quality, Growth, Value, and Dividends.

The research aims to evaluate the profitability of investment strategies based on these scores over multiple profitability horizons from 1 month to five years. A comprehensive dataset was constructed, integrating factor scores with additional variables such as sector and geographic region, standardized for currency and timeframes.

Statistical analyses will explore relationships between these factors and returns, identifying optimal investment periods. Subsequently, predictive modeling—including econometric regressions, time series models, and neural networks—will be applied to assess the translation of these factors into market performance.

The study incorporates an interdisciplinary approach, combining financial theory and econometrics with advanced programming, data engineering, and machine learning. This integration bridges Business Management and Telecommunications Engineering, offering insights into the practical application of Tweenvest's scoring algorithm and contributing to the advancement of financial technology analytics.

Relation of the Thesis with ABET Competencies

Table 1: Relation of the Thesis with ABET Competencies

Concept (ABET)	Complies (Y/N)	Where (pages)
1.1 Problem statement and opportunity	Y	1–3
1.2 Constraints (standards, codes, needs, requirements & specifications)	Y	4–6, 9–16
1.3 Setting of goals	Y	6–7
2.1 Creative solution generation (analysis)	Y	33–46
2.2 Evaluation of multiple solutions and decision-making (synthesis)	Y	49–83
3.1 Fulfilment of goals	Y	85–87
3.2 Overall impact and significance (contributions and practical recommendations)	Y	87–89

Dedication

This work is dedicated to my family, specifically: to my mother who has always supported and encouraged me to go one step further and enjoy the journey; to Jorge, the most honorable person I know and that has taught me so much in life about what type of person I want to be; to my grand parents, “Piro y Tita”, who have always given me love and time to take care of me; and finally to María, my partner in life, who has had to wait for me during this long journey that has taken us to the present day.

None of this would have been possible without the help from my tutors. I can’t be more thankful to have had such wonderful and generous people guiding me to this point: Pepe, for always believing and trusting in me with my crazy ideas, that were far more complicated that I could have imagined, because without you this definitely wouldn’t have happened; Gabi, for being a constant inspiration and support with that extraordinary sense of humor; Alberto, for taking your time to listen to me with all of that condensation of ideas that I did not quite understand completely and bringing me light to see the right path, and for being such a wonderful professor that doesn’t give up on what is wrong; and finally, to Joaquín, a teacher, a colleague and a friend, that has taught me that curiosity is the key to everything, I really appreciate all you have done for me and all of your effort, support and time that you have spent in this project.

Finally, I want to encourage everyone who is reading this to always do that extra mile, spend a little more time and effort. It will be worth it. And please, if you are a student and have the opportunity to choose a thesis, don’t look at it as a burden or “*just the last work in college*”, because it can literally change your life. It is the moment to really spend as much time as you want learning something you enjoy, and if you do it right, you can even align it with your future professional career.

Contents

I	Main Report	1
1	Introduction and Objectives	3
1.1	Fundamental Analysis Investing	3
1.1.1	Context	3
1.1.2	Interdisciplinary Nature of the Thesis	3
1.2	Possible Problems With the Scores' Methodology	4
1.2.1	Unquantified Variables	4
1.2.2	Emergent Effects in Complex Systems	5
1.2.3	Linearity and Stationarity	5
1.3	Objectives	6
1.4	Variables for the Analysis	6
1.4.1	Response Variables	6
1.4.2	Explanatory Variables	7
2	Methodology & Theoretical Framework	9
2.1	Tweenvest's Scores	9
2.1.1	Quality	9
2.1.2	Growth	12
2.1.3	Value	14
2.1.4	Dividend	15
2.2	Code practices	16
2.3	Preprocessing	16
2.3.1	Data Consistency	17
2.3.2	Data Transformation	17
2.4	Outlier Detection	19
2.4.1	Inter Quartile Range	19
2.4.2	Single Vector Machine	19
2.4.3	Isolation Forest	20
2.4.4	Local Outlier Factor	21

2.4.5	Multi-Criteria Outlier Detection	22
2.5	Predictive Models	23
2.5.1	Regression Models	23
2.5.2	Time Series	25
2.5.3	Neural Networks	27
2.6	Back-testing	29
2.6.1	Training and Testing	29
2.6.2	Performance	30
2.6.3	Consistency and Overfitting	31
2.6.4	Robustness of Results	31
3	Development	33
3.1	Architecture Enhancement	33
3.1.1	Updating Tweenvest Code	33
3.1.2	Designing new Jobs	36
3.1.3	Telemetry Tracing	37
3.1.4	Dedicated Server Deployment	38
3.2	Dataset Creation	39
3.2.1	Aggregating the data	39
3.2.2	Datasets Construction	40
3.2.3	Preprocessing	41
3.2.4	Outlier Detection	41
3.2.5	Summary	46
4	Results	49
4.1	Descriptive Analysis	49
4.1.1	Variables Distributions	49
4.1.2	Correlations	52
4.1.3	Data clusters	55
4.2	Linear Regression	56
4.2.1	General Models	56
4.2.2	Geographical Effects	61
4.2.3	Balanced Dataset	72
4.3	GAM	73
4.3.1	Limitations and Considerations	73
4.3.2	Results	73
4.3.3	Non-linear relationships	73
4.4	Neural Network	76

4.4.1	Architecture Optimization	77
4.4.2	Binary Classifier	79
4.4.3	Encoder Classifier	81
4.4.4	Regressor	83
5	Conclusion and Discussion	85
5.1	Conclusions	85
5.1.1	Clear Statements	85
5.1.2	Non-Clear Results	87
5.2	Discussion	87
5.2.1	Limitations of the study	87
5.2.2	Future work	88
5.2.3	Recommendations to Tweenvest	89
	References and Bibliography	91
II	Appendix	93
6	Code Listings	95
6.1	Historical Scores Job Implementation	95
6.1.1	Enqueueing Historical Scores Job	95
6.1.2	Calculate Index Scores Data	97
6.1.3	Calculate Stocks Scores Task	98
6.2	Workers Management	99
6.2.1	Tweenvest's Production Server	99
6.2.2	Personal Development Server	101
6.3	Data Export Job Implementation	101
6.4	Custom Python Functions for Data Analysis	105
6.5	Data Preprocessing	106
6.5.1	Data Cleaning	106
6.5.2	Outlier Detection	107
7	Extra Images and Tables	109
7.1	Data	109
7.1.1	Columns and Variables	109
7.1.2	Distributions	109
7.1.3	Correlations	111
7.2	Models	112

7.2.1	Linear Regression	113
7.2.2	GAM	131
7.2.3	NN	132
7.3	2030 Agenda	134
7.4	Executive Summary	134

List of Figures

2.1	Tweenvest's Quality Score	12
2.2	Tweenvest's Growth Score	14
2.3	Tweenvest's Value Score	15
2.4	Tweenvest's Dividend Score	16
2.5	Yeo-Johnson Transformations examples	18
2.6	Visualization of how IF works [11].	21
2.7	MCOD Schema	22
2.8	ICD Schema	23
2.9	pyGAM - Example of GAMs regression splines	25
2.10	Simple Architecture of a NN	27
3.1	General Tweenvest's Architecture	34
3.2	Scores Calculations Code Schema	34
3.3	<i>calculation_date</i> Field Propagation Schema	35
3.4	Historical Jobs First Iteration Schema	36
3.5	Historical Jobs Final Version Schema	37
3.6	First Data Aggregation Schema	39
3.7	Final Data Aggregation Schema	40
3.8	Past Dataset	40
3.9	Future Dataset	40
3.10	Small Data Future - IF (4%) Pair Plot	42
3.11	Small Data Future - SVM (4%) Pair Plot	43
3.12	Small Data Future - LOF (4%) Pair Plot	44
3.13	Small Data Future - MCOD (8%) Pair Plot	45
3.14	Correlations between variables for different outlier detection methods	46
4.1	Big Data Future (IQR) - Scores	50
4.2	Big Data Future (IQR) - Scores Deep	50
4.3	EMMI - Euribor Interest Rates	50
4.4	Big Data Future (IQR) - Dummies	51

4.5	Small Data Future (MCOD) - Dummies	51
4.6	Big Data Future IQR - Profits Boxplot	52
4.7	Big Data Future IQR - Profits	52
4.8	Big Data Past (Clean) - Correlations	53
4.9	Big Data Future (Clean) - Correlations	53
4.10	Big Data Past IF (4%) - Correlations	53
4.11	Big Data Future IF (4%) - Correlations	53
4.12	Big Data Future IF (35%) - Correlations	54
4.13	Small Data Future - MCODE (4%) Pair Plot	55
4.14	Small Data Future - MCODE (4%)	56
4.15	Small Data Future - MCODE 4% - Residuals	57
4.16	Small Data Future - MCODE (4%)	58
4.17	Small Data Future <i>Yeo-Johnson</i> - MCODE (4%)	59
4.18	Small Data Future <i>Yeo-Johnson</i> - MCODE (4%)	59
4.19	Small Data Future (Windowed) - MCODE (4%)	60
4.20	Small Data Future (Windowed) - MCODE (4%)	60
4.21	Geographical Predictability Analysis - <i>Yeo-Johnson</i>	62
4.22	North America <i>Yeo-Johnson</i> - Model Residuals	63
4.23	Europe <i>Yeo-Johnson</i> - Model Residuals	63
4.24	Asia & Oceania <i>Yeo-Johnson</i> - Model Residuals	63
4.25	South America <i>Yeo-Johnson</i> - Model Residuals	64
4.26	Africa & Middle East <i>Yeo-Johnson</i> - Model Residuals	64
4.27	Geographical Predictability Analysis - Windowed Models	68
4.28	North America - Windowed Model Residuals	69
4.29	Europe - Windowed Model Residuals	69
4.30	Asia & Oceania - Windowed Model Residuals	69
4.31	South America - Windowed Model Residuals	69
4.32	Africa & Middle East - Windowed Model Residuals	70
4.33	Small Data Future IF - GAM Performance	73
4.34	Big Data Future IF Balanced - GAM Performance	73
4.35	Small Data Future IF - Dividend 24M Average Splines	74
4.36	Big Data Future IF Balanced - Quality 24M Average Splines	74
4.37	Small Data Future - IF 15% - Growth 12M Average Splines	74
4.38	Big Data Future IF - Growth 12M Average Splines	74
4.39	Small Data Future IF - Growth 24M Average Splines	74
4.40	Big Data Future IF - Growth 24M Average Splines	74
4.41	Small Data Future IF - Quality 24M Average Splines	75

4.42	Big Data Future IF - Quality 24M Average Splines	75
4.43	Small Data Future IF - Value Splines	75
4.44	Big Data Future IF - Value Splines	75
4.45	Small Data Future IF - Value 24M Average Splines	75
4.46	Big Data Future IF Balanced - Value 24M Average Splines	75
4.47	Big Data Future IF Balanced - Region x Value Tensors	76
4.48	Big Data Future IF Balanced - Region x Value 24M Average Tensors	76
4.49	Binary Classifier - (Quality) Hidden Layer Optimization	77
4.50	Encoder Classifier - (Quality) Hidden Layer Optimization	78
4.51	Regressor Classifier - (Quality) Hidden Layer Optimization	78
4.52	Big Data Future IF Balanced - Binary Classifier Performance	79
4.53	Small Data Future IF - Binary SHapley Additive exPlanations (SHAP) Consistency	80
4.54	Big Data Future IF Balanced - Binary SHAP Consistency	81
4.55	Big Data Future IF Balanced - Encoder Classifier Performance	82
4.56	Big Data Future IF Balanced - Encoder SHAP Consistency	83
4.57	Big Data Future IF Balanced - Regressor Performance	84
4.58	Big Data Future IF Balanced - Regressor SHAP Consistency	84
5.1	Twenevest Scores Overview	89
7.1	Number of columns per dataset	109
7.2	Big Data Past (IQR) - Scores	110
7.3	Big Data Past (IQR) - Scores Deep	110
7.4	Small Data Future - Scores	110
7.5	Small Data Future - Scores Deep	110
7.6	Big Data Future - Dummies	110
7.7	Big Data Past - Dummies	110
7.8	Big Data Future - Profits Boxplot	111
7.9	Big Data Future - Profits	111
7.10	Small Data Future USA	111
7.11	Small Data Future USA - MCODE	111
7.12	Small Data Future EU	111
7.13	Small Data Future EU MCODE	111
7.14	Small Data Future LAT	112
7.15	Small Data Future LAT - MCODE	112
7.16	Small Data Future AF	112
7.17	Small Data Future AF - MCODE	112
7.18	Small Data Future AS	112

7.19	Small Data Future AS - MCODE	112
7.20	Small Data Future (Clean) - Performance	113
7.21	Geographical Predictability Analysis - First Model	116
7.22	USA - First Model Residuals	117
7.23	EU - First Model Residuals	117
7.24	Asia & Oceania - First Model Residuals	117
7.25	South America - First Model Residuals	117
7.26	Africa & Middle East - First Model Residuals	118
7.27	Big Data Future IF Balanced - (Windowed) Performance	130
7.28	Big Data Future IF Balanced - (Windowed) Residues	131
7.29	Small Data Future IF - Region x Quality	131
7.30	Big Data Future IF Balanced - Region x Dividend 24M Average	131
7.31	Small Data Future IF - Region x Quality	131
7.32	Big Data Future IF Balanced - Region x Growth 24M Average	131
7.33	Small Data Future IF - Region x Quality	132
7.34	Big Data Future IF Balanced - Region x Quality 24M Average	132
7.35	Small Data Future IF - Encoder SHAP Consistency	132
7.36	Small Data Future IF - Regreesor SHAP Consistency	133

List of Tables

1	Relation of the Thesis with ABET Competencies	3
2.1	Pros and Cons of the Interquartile Rule	19
2.2	Pros and Cons of One-Class SVM	20
2.3	Pros and Cons of the Isolation Forest Method	21
2.4	Pros and Cons of the LOF Method	22
3.1	Tweenvest Server Specifications	38
3.2	Hetzner Server Specifications	38
3.3	(%) Missing values per column for Small Data Future.	41
3.4	Simple Datasets Total Rows	46
3.5	Geographical Datasets Total Rows	47
3.6	Validation Datasets Summary	47
4.1	Small Data Future - MCOD (4%) Consistency	57
4.2	Small Data Future (Windowed) - MCOD (4%) Consistency Summary	61
4.3	North America <i>Yeo-Johnson</i> - Model Consistency	65
4.4	Europe <i>Yeo-Johnson</i> - Model Consistency	65
4.5	Asia & Oceania <i>Yeo-Johnson</i> - Model Consistency	66
4.6	South America <i>Yeo-Johnson</i> - Model Consistency	66
4.7	Africa & Middle East <i>Yeo-Johnson</i> - Model Consistency	67
4.8	North America (Windowed) - Consistency Summary	70
4.9	Europe (Windowed) - Consistency Summary	71
4.10	Asia & Oceania (Windowed) - Consistency Summary	71
4.11	South America (Windowed) - Consistency Summary	72
4.12	Africa & Middle East (Windowed) - Consistency Summary	72
7.1	Small Data Future <i>Yeo-Johnson</i> - MCOD (4%) Consistency	113
7.2	Small Data Future (Windowed) - MCOD (4%) Dividend Consistency	114
7.3	Small Data Future (Windowed) - MCOD (4%) Growth Consistency	114
7.4	Small Data Future (Windowed) - MCOD (4%) Quality Consistency	115

7.5	Small Data Future (Windowed) - MCOD (4%) Value Consistency	115
7.6	North America - First Model Consistency	118
7.7	Europe - First Model Consistency	118
7.8	Asia & Oceania - First Model Consistency	119
7.9	South America - First Model Consistency	119
7.10	Africa & Middle East - First Model Consistency	120
7.11	North America (Windowed) - Dividend Consistency	120
7.12	North America (Windowed) - Growth Consistency	121
7.13	North America (Windowed) - Quality Consistency	121
7.14	North America (Windowed) - Value Consistency	122
7.15	Europe (Windowed) - Dividend Consistency	122
7.16	Europe (Windowed) - Growth Consistency	123
7.17	Europe (Windowed) - Quality Consistency	123
7.18	Europe (Windowed) - Value Consistency	124
7.19	Asia & Oceania (Windowed) - Dividend Consistency	124
7.20	Asia & Oceania (Windowed) - Growth Consistency	125
7.21	Asia & Oceania (Windowed) - Quality Consistency	125
7.22	Asia & Oceania (Windowed) - Value Consistency	126
7.23	South America (Windowed) - Dividend Consistency	126
7.24	South America (Windowed) - Growth Consistency	127
7.25	South America (Windowed) - Quality Consistency	127
7.26	South America (Windowed) - Value Consistency	128
7.27	Africa & Middle East (Windowed) - Dividend Consistency	128
7.28	Africa & Middle East (Windowed) - Growth Consistency	129
7.29	Africa & Middle East (Windowed) - Quality Consistency	129
7.30	Africa & Middle East (Windowed) - Value Consistency	130
7.31	Relation of the Project with the Sustainable Development Goals (SDGs) of the 2030 Agenda	134
7.32	Relation of the Thesis with ABET Competencies	134

Acronyms and Abbreviations

ACF Autocorrelation Function.

API Application Programming Interface.

ARIMA AutoRegressive Integrated Moving Average.

CAGR Compound Annual Growth Rate $\left(\left(\frac{\text{Ending Value}}{\text{Beginning Value}} \right)^{\frac{1}{n}} - 1 \right)$.

CAPEX Capital Expenditures.

Cash Ratio $\frac{\text{Cash and Equivalents}}{\text{Current Liabilities}}$.

CI/CD Continuous Integration/Continuous Deployment.

CROCE Cash Return on Capital Employed.

CROIC Cash Return on Invested Capital.

Cron Cron is a time-based job scheduler in Unix-like computer operating systems. It is used to schedule jobs (commands or shell scripts) to run at specified time intervals. Cron is typically used to automate tasks that need to be performed on a regular basis, such as backups, system maintenance, or data processing.

Current Ratio $\frac{\text{Current Assets}}{\text{Current Liabilities}}$.

DB Database.

Decorator Decorator is a function that takes another function and extends the behavior of the original function without explicitly modifying it. Decorators are typically used to add functionality to functions, such as logging, caching, or authentication.

Dividend Yield $\frac{\text{Dividend per Share}}{\text{Price per Share}} \times 100\%$.

DPS Dividends per Share.

EBIT Earnings Before Interest and Taxes.

EBITDA Earnings Before Interest, Taxes, Depreciation, and Amortization.

EMMI European Money Markets Institute.

EPS Earnings per Share.

EV/EBIT $\frac{\text{Enterprise Value}}{\text{TTM EBIT}}$.

EV/EBITDA $\frac{\text{Enterprise Value}}{\text{TTM EBITDA}}$.

EV/FCF $\frac{\text{Enterprise Value}}{\text{TTM Free Cash Flow}}$.

EV/Sales $\frac{\text{Enterprise Value}}{\text{TTM Revenue}}$.

FCF Free Cash Flow.

Financial Leverage $\frac{\text{Total Assets}}{\text{Total Equity}}$.

FOMO Fear of Missing Out.

GAM Generalized Additive Model.

ICD Island Cases Detection.

IF Isolation Forest.

IQR Interquartile Range.

Job Code script that is executed by the system to perform a task, this could be a daily update of the financial information of the stocks, or a necessary action such as sending emails to get authentication pins.

LOF Local Outlier Factor.

MAE Mean Absolute Error.

Market Cap Market Capitalization.

MCOD Multi-Criteria Outlier Detection.

NN Neural Network.

OCF Ratio $\frac{\text{Operating Cash Flow}}{\text{Current Liabilities}}$.

OCR Optical Character Recognition.

P/B $\frac{\text{Market Cap}}{\text{Tangible Equity}}$.

P/CF $\frac{\text{Market Cap}}{\text{TTM Operating Cash Flow}}$.

P/E $\frac{\text{Market Cap}}{\text{Adjusted TTM Earnings}}$.

P/S $\frac{\text{Market Cap}}{\text{TTM Revenue}}$.

PACF Partial Autocorrelation Function.

Pair Plot Pairwise relationship plot.

Payout Ratio EPS $\frac{\text{DPS}}{\text{EPS}}$.

Payout Ratio FCF $\frac{\text{DPS}}{\text{FCF}}$.

Payout Ratio Owner Earnings $\frac{\text{DPS}}{\text{Owner}}$.

PCA Principal Component Analysis.

PLS Partial Least Squares.

Q-Q Quantile-Quantile.

Quick Ratio $\frac{\text{Current Assets} - \text{Inventory}}{\text{Current Liabilities}}$.

REDIS REDIS is an open-source, in-memory data structure store, used as a database, cache, and message broker. It supports data structures such as strings, hashes, lists, sets, and sorted sets with range queries, transactions, and different levels of on-disk persistence.

RMSE Root Mean Square Error.

ROA Return on Assets.

ROCE Return on Capital Employed.

ROE Return on Equity.

ROIC Return on Invested Capital.

S&P 500 Standard & Poor's 500 Index.

SHAP SHapley Additive exPlanations.

SOE State-Owned Enterprise.

SVM Single Vector Machine.

Unit Test Unit test is a type of software testing where individual units or components of a software are tested. The purpose of unit testing is to validate that each unit of the software performs as expected. Unit tests are typically written by the developers and run automatically as part of the development process.

Part I

Main Report

Chapter 1

Introduction and Objectives

1.1 Fundamental Analysis Investing

1.1.1 Context

Fundamental Analysis is a methodology used to evaluate the intrinsic value of a company, asset, or market by analyzing various economic, financial, qualitative, and quantitative factors. Unlike technical analysis, which focuses on price movements and chart patterns, fundamental analysis seeks to determine an asset's "*true value*" to identify investment opportunities that may be undervalued or overvalued in the market. This intrinsic value is defined by numerous experts and renowned investors —such as Benjamin Graham, Warren Buffett, and Pat Dorsey [1]— as a company's ability to adapt to an ever-changing environment, create value in the market, and establish barriers to entry for competitors; commonly referred to as *economic moats*.

Measuring these moats is challenging due to the difficulty of quantifying certain variables, such as brand strength and market influence. However, over the long term most literature show that these **intangible factors translate into tangible financial data**, which are reflected in a company's reports: balance sheet, income statement, and cash flow statement. When companies are exposed to the public market share, all of this information transforms into the needs of buying or selling the stocks, altering the companies stocks' profitability. So from now on, these **economic moats will be called *alpha*** [2].

This complexity prompted José Tatay, Tweenvest's founder, to develop a series of algorithms to calculate certain scores, based on the fundamental analysis principles, to help users identify possible *alpha* in a company; which is the main objective for all of the long-term investors. These scores play a key role in this thesis, that is why the generative method will be explained further in the next chapter.

1.1.2 Interdisciplinary Nature of the Thesis

This work is the result of a real-world problem posed by *Tweenvest*, a company operating in the field of long-term investing. The challenges are complex and require an approach that integrates both technical and economical expertise.

From the perspective of **Telecommunication Engineering**, the project demands skills in data acquisition, processing, and modeling (machine learning); as well as the design and optimiza-

tion of computational architectures capable of handling large-scale financial datasets. These competencies ensure that the proposed solutions are not only methodologically sound but also technically feasible and efficient in a production environment.

From the perspective of **Business Management**, the work draws on financial theory, econometric modeling and statistics to critically evaluate whether the scoring system truly captures a company's capacity to generate *alpha*. This dimension provides the economic rationale and market context necessary to interpret the results and propose meaningful improvements.

The combination of these two disciplines enables a holistic approach: technical tools and infrastructure are aligned with financial objectives, resulting in solutions that are both operationally robust and economically relevant. This thesis has been carried out under the joint supervision of **Alberto Albiol Colmener** (Telecommunication Engineering), **Joaquín Martínez Minaya** (Business Management), and **José Tatay Sangüesa** (Tweenvest), whose guidance has been essential in bridging the academic and professional perspectives of the work.

1.2 Possible Problems With the Scores' Methodology

Given the central role of these scores in the investment decision-making process, it is essential to assess whether the simplifications and assumptions underlying their design—both those inherited from general financial theory and those specific to Tweenvest's implementation—accurately capture a company's capacity to generate *alpha* over time. This motivates the need for a systematic review of the methodology to evaluate its validity and practical reliability.

1.2.1 Unquantified Variables

Since the current algorithm only includes variables found in a company's financial statements, there is a significant amount of relevant information being left out, for example:

- The **perceived differentiation** of a product is one such element that may not be properly captured by traditional financial analysis. A trusted brand or strong reputation can allow a company to charge premium prices and build customer loyalty, generating extraordinary long-term profits. For instance, brands like Tiffany or Rolex have high perceived value that justifies elevated prices—something that cannot be easily quantified using standard financial metrics such as profit margins or return on capital.
- Additionally, strategies such as **cost leadership** and offering products at lower prices can provide a significant competitive advantage that is not always directly reflected in financial data.
- Companies may also establish **barriers to entry** and **high switching costs** for customers, making it more difficult for them to move to competitors. These strategies may involve investments in technology, patents, or simply building long-term customer and employees relationships.

To properly measure these variables, a more exhaustive analysis of each company would be required, pulling from multiple secondary information sources such as news articles, conference transcripts, customer blogs, competitive product reviews, and more; which could also itself affect the market. This task is commonly known as *social listening* [3]; and is far more difficult to

automate via code, as it would require multi-modal AI techniques –thus, it will fall outside the scope of this work.

1.2.2 Emergent Effects in Complex Systems

In complex systems like financial markets, network effects emerge, adding noise to actual reliable data behind it. For example:

- **Herd behavior** is a network effect where investors tend to follow the actions of others rather than rely on their own analysis. This can amplify market movements, both upward and downward. Herd behavior can lead to speculative bubbles and abrupt corrections when collective expectations shift.
- **Feedback loops** are another effect where market participants' actions reinforce existing market behavior. For example, rising asset prices may attract more investors, which in turn drives prices even higher; such as the Fear of Missing Out (FOMO) effect. This type of positive feedback can cause price escalation that becomes detached from underlying fundamentals. Conversely, negative feedback can occur during a sell-off, where falling prices trigger more selling, further accelerating the decline.
- **Macro-level influences** are another crucial factor where broader structural changes—such as political decisions, economic policies, technological shifts, or industry-wide trends—can significantly impact market behavior. These larger forces often operate beyond the scope of individual company analysis and can create ripple effects throughout the entire market ecosystem.

1.2.3 Linearity and Stationarity

Non-linearity

The relationships between financial variables are often non-linear, meaning that changes in one variable may not result in proportional changes in another.

- **Economies of scale** can create non-linear relationships between production volume and costs. As a company grows, it may experience decreasing marginal costs due to better resource utilization, bulk purchasing discounts, or spreading fixed costs over larger output.
- **Market saturation** can lead to diminishing returns on marketing spend or R&D investments. Initial investments might yield significant returns, but as the market becomes saturated, additional spending may produce smaller incremental benefits.
- **Competitive dynamics** can create threshold effects where small changes in market share or pricing can trigger significant shifts in competitive position or profitability.

Non-stationarity

Traditionally, **standard growth ratios have been employed** as a key component in evaluating a company's performance over time. These ratios, such as revenue growth, earnings growth, or asset growth, provide a normalized measure that facilitates comparison across companies and industries, regardless of size. In the context of score calculation, they serve as a **consistent benchmark**, allowing the algorithm to assess whether a company's growth trajectory exceeds,

matches, or falls below market expectations. The underlying assumption is that in competitive markets, when a new business model or opportunity emerges with above-average margins, entrepreneurs quickly enter to capitalize on it, eventually saturating the opportunity and driving margins back toward average levels.

However, markets are “fluctuating entities”, so **static metrics can become problematic when underlying trends change**. A clear example is the rise of artificial intelligence and the surge in stock prices of companies involved in the production and development of the necessary technologies.

1.3 Objectives

Given the limitations identified mentioned for the current algorithm, the main aim of this work is to **evaluate if the scores provided by Tweenvest represent *alpha* for different time frames, guide users how to interpret them based on different investment strategies, and to propose possible changes to the scores calculation** in order to present the results with more accuracy. To achieve this, we followed this procedure:

1. **System Architecture Enhancement:** Modify Tweenvest’s database architecture to enable historical score storage and retrieval and create the datasets for the analysis.
2. **Data Curation:** Clean and preprocess the collected data to ensure consistency and reliability, handling missing values and outliers appropriately.
3. **Exploratory Analysis:** Process and analyze the data to check the distributions and correlations between variables, looking for early on patterns to later compare and use in the modeling.
4. **Predictive Modeling:** Develop multiple predictive models to better understand the scores and their relations with other variables and time frames.
5. **Validation & Benchmarking:** Back-test the algorithms to check their performance and consistency to see if the scores generate positive *alpha* over the different time frames.

1.4 Variables for the Analysis

To fulfill the objectives, one of the most important variables to consider is the profits of an investment in a company. And for doing so, we are going to look for correlations between the profits of a company with different scores and other categorical variables that we will use in the analysis.

1.4.1 Response Variables

To help users understand how to use the scores depending on their investment horizon, we computed profits for multiple time intervals, denoted as P_i for each horizon: $i \in \{1 \text{ month}; 3 \text{ months}; 6 \text{ months}; 1 \text{ year}; 2 \text{ years}; 5 \text{ years}\}$. These intervals were chosen to capture distinct market perspectives: short-term (1–6 months) to evaluate immediate market reactions and tactical opportunities, medium-term (1–2 years) to reflect cyclical adjustments and strategic positioning, and long-term (5 years) to assess sustained value creation and the durability of a company’s fundamentals.

The profits are defined as:

$$R_i^n = \frac{(P_{i+n} + D_{acc}) - P_i}{P_i} \quad (1.1)$$

Here, R_i^n represents the total return achieved over a period of n months starting from time i . This return accounts for both the price appreciation and the dividends received during the specified time frame. Specifically, P_i denotes the price at time i , while D_{acc} is the sum of all dividends accumulated from time i to $i + n$.

1.4.2 Explanatory Variables

To relate the profits with the scores, we need to use different variables:

- **Main variables**, these are numerical variables from **Tweenvest's Scores**, that are in the range of 0 to 100: S_j where $j \in \{ \text{Quality; Value; Dividend; Growth} \}$. These will be explained in the next section.
- **Dummy variables**, these are boolean variables that allow us to tag the companies by:
 - **Industry**: I_k where $k \in \{ \text{Financials; Healthcare; etc.} \}$
 - **Geographical Region**: G_l where $l \in \{ \text{Europe; North America; etc.} \}$
- **Market Variables**, these are numeric variables that represent the companies "size" using the: **Market Cap** and **Volume**.

Chapter 2

Methodology & Theoretical Framework

This chapter outlines the methodological and theoretical foundations used to evaluate Tweenvest's scoring system. The first part describes in detail the structure and computation of the scores, providing a clear understanding of how each factor—Quality, Growth, Value, and Dividend—is defined and measured. This is followed by the data processing, outlier detection, and predictive modeling techniques applied to assess their relationship with company performance over different investment horizons.

2.1 Tweenvest's Scores

The **long-term advantage** explained earlier in the finance is used in many indexes such as MSCI [4], IBEX-35, S&P 500, etc to achieve higher returns than average. So following this approach, Tweenvest developed a series of scores to **help users identify possible *alpha* in a company**. Upon these scores, we can highlight four –Quality, Growth, Value, and Dividend– that are strictly related to key focus areas that fundamental-based investors consider before making any decision, these areas are: profitability, financial health, predictability, consistent growth, entrance moment, and dividends payed.

2.1.1 Quality

To understand what the Quality score aims to capture, it is useful to consider the three key categories commonly used by successful investors when analyzing financial statements: **profitability**, **financial health**, and **predictability**. Each category encompasses a range of financial ratios that reflect different aspects of a company's performance and by combining these dimensions, we can construct a more comprehensive view of a company's overall financial "quality". To better understand the score, we need to look at each category separately. Starting with **profitability** we need to take into consideration two aspects: profitability margins and performance ratios.

Profitability Margins

These are essential for understanding how a company is managing its costs and generating profits from its revenues.

- **Net margin:** Represents net profit as a percentage of total sales, indicates a company's efficiency in generating profits after accounting for all expenses, taxes, and costs.

- **Operating margin:** Measures operating profit (EBIT) as a percentage of total sales, providing a clear view of the profitability of a company's core operations, excluding interest and taxes. This metric is fundamental for evaluating a company's operational efficiency.
- **EBITDA margin:** Removes the effects of capital structure and accounting policies, offering a clear view of the company's pure operational profitability.
- **Gross margin:** Focuses on revenues after deducting the cost of goods sold, is a key measure of production efficiency and a company's ability to manage its direct costs.

Performance Ratios

These are used to measure the overall performance of the company.

- **Return on Assets (ROA):** Measures how efficiently a company converts its assets into profits. This is especially important in capital-intensive sectors, where efficient asset management can make a significant difference in profitability.
- **Return on Equity (ROE):** Focuses on the profits generated per dollar of equity invested by shareholders. This ratio is crucial for evaluating a company's overall profitability from the shareholders' perspective. It is an especially valuable metric for investors seeking to maximize their returns on equity investment.
- **Return on Invested Capital (ROIC):** Focuses on the return generated by all the funds invested in the company, including both shareholders' equity and debt. It is a comprehensive measure of a company's ability to generate value from all its sources of financing.
- **Return on Capital Employed (ROCE):** Measures the funds used to finance operations, regardless of the source. This ratio is useful for comparing the efficiency of companies with different capital structures, as it focuses on total capital employed rather than just equity.

Also, to not only look at actives, Tweenvest uses the cash generated to calculate: **Cash Return on Invested Capital (CROIC)**, **Cash Return on Capital Employed (CROCE)**, **OCF/Sales**, and **Free Cash Flow (FCF)/Sales**. And lastly it takes account also the Owner's income to calculate: **Owner's Income/Sales**, **Owner's CROIC** and **Owner's CROCE**.

Continuing with the **financial health**, we need to analyze the company's debt state in different aspects: **structure of the debt** (leverage ratios, debt coverage ratios, interest coverage ratios), and **ability to pay** (liquidity ratios).

Leverage Ratios

These ratios assess how much a company relies on debt to finance its assets and operations, and are essential for evaluating financial risk and long-term solvency.

- **Financial Leverage:** Measures the proportion of a company's assets that are financed by shareholder equity. A higher ratio suggests the company is using more debt relative to equity, indicating greater financial risk but also potential return amplification through leverage.

- **Total Debt/Assets:** Indicates what portion of the company's assets is financed through debt. A lower ratio implies a more conservative capital structure, while a higher one may indicate increased risk if the company becomes over-leveraged.
- **Total Debt/Capital:** Measures the share of total capital (debt + equity) that comes from debt. This ratio is useful for understanding how dependent the company is on borrowed funds compared to its overall capital base.
- **Total Debt/Equity:** Compares the company's total debt to its shareholder equity. It provides insight into the balance between debt and equity financing. A high ratio may signal financial risk, but also the potential for higher returns if debt is managed well.

Debt Coverage Ratios

These metrics evaluate a company's ability to cover its debt using its earnings or cash flow, reflecting the sustainability of a company's debt in relation to its operational performance.

- **Net Debt/EBIT:** Shows how many years it would take for a company to repay its net debt using EBIT.
- **Net Debt/EBITDA:** Similar to the above, but adds back depreciation and amortization. This gives a more cash-focused view of a company's ability to handle its debt load, and is especially useful for comparing companies in capital-intensive industries.
- **Net Debt/FCF:** Evaluates how many years of free cash flow would be needed to pay off net debt. Since FCF includes investment needs, this ratio gives a more conservative view of debt sustainability.
- **Net Debt/Owner's Income:** Compares net debt to the income available to equity holders, after all operating and investing costs.

Interest Coverage Ratios

These ratios measure how easily a company can meet its interest payments on outstanding debt, critical for assessing short-term debt service capability.

- **EBIT/Interest:** Indicates how many times a company can cover its interest expenses with its operating income.
- **EBITDA/Interest:** Similar to the above, but adds back depreciation and amortization. This gives a clearer picture of available cash earnings before fixed financial obligations, ideal for heavily asset-based businesses.
- **FCF/Interest:** Since FCF considers investment needs, this is a stringent test of how much real, discretionary cash is available for debt servicing.
- **Owner Earnings/Interest:** Evaluates a company's ability to meet interest payments based on the earnings effectively attributable to shareholders. It accounts for operational cash flow minus necessary capital expenditures.

Liquidity Ratios

These ratios measure a company's ability to meet short-term obligations with its short-term assets. They are essential for evaluating near-term financial health and risk of insolvency.

- **Current Ratio:** Shows whether a company has enough assets to cover its short-term liabilities. A value above 1 is generally considered healthy, though excessively high values may imply inefficiency.
- **Quick Ratio:** A more stringent version of the current ratio that excludes inventory, which may not be easily liquidated. It's useful in assessing true short-term liquidity.
- **Cash Ratio:** The most conservative liquidity metric, focusing only on cash and equivalents. It shows the immediate solvency of a company in a worst-case scenario.
- **OCF Ratio:** Assesses how well the company's operational cash flows can cover its current obligations. This offers a realistic view of liquidity since it is based on actual cash generation rather than accounting figures.

Predictability

And finally, for the quality score we need to look at the company's predictability. This is achieved by trying to fit values related to the company's success —such as Sales— to an exponential curve, using the ordinary least square method, which is supported by large financial literature [4].

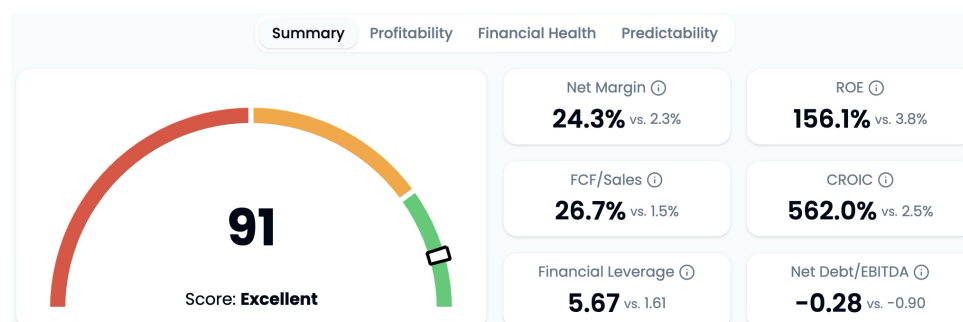


Figure 2.1: Tweenvest's Quality Score

After calculating all of this ratios, Tweenvest compares them to the sectors' median and interpolates each of them to create a single score for the ratio, then it aggregates them all using personalized weights to create the final Quality Score, which is then summarized and shown to the clients as seen in figure 2.1. In the left we have the total score for the factor and in the right some of the most relevant ratios used to calculate it, comparing them to the sector. This format is consistent for all of the factors studied in this work.

2.1.2 Growth

The Growth Score evaluates a company's historical growth across multiple key metrics, and compares them to industry standards. This comprehensive approach ensures a balanced assessment of growth across different aspects of the business:

Revenue & Profitability Growth

- **Sales:** Measures growth in core revenue streams
- **EBITDA:** Captures growth in operational cash-generating ability before non-cash and financing impacts
- **Operating Income:** Reflects growth in profit from core business operations
- **Net Income:** Includes all income sources, showing overall profitability growth

Cash Flow Growth

- **Operating Cash Flow:** Measures growth in cash generation from operations
- **Simple FCF:** A straightforward proxy for available cash after essential investments
- **Levered/Unlevered FCF:** Provide detailed views of free cash flow with and without debt impact
- **Owner Earnings:** Useful for volatile Capital Expenditures (CAPEX) cases, emphasizing cash available to shareholders

Capital Base Expansion

- **Total Assets:** Indicates expansion in overall asset base
- **Equity:** Reflects growth in shareholders' claim on the business
- **Tangible Book Value:** Highlights growth in physical net assets, excluding intangibles
- **Invested Capital:** Captures total capital being put to productive use
- **Capital Employed:** A broader measure of capital supporting business operations

Per-Share Value Growth

- **Diluted EPS:** Tracks per-share earnings growth, accounting for dilution effects
- **Diluted Shares:** Included to track share count changes, ensuring EPS growth is not artificially inflated by buybacks or dilution
- **Ordinary DPS:** Tracks the growth of shareholder payouts, a proxy for confidence in future earnings

To compute the Growth Score, Tweenvest calculates 10-year, 5-year, and 3-year averages and then interpolates the growth rate to industry standards. This approach reinforces the long-term investment philosophy, giving lasting growing companies a better score.

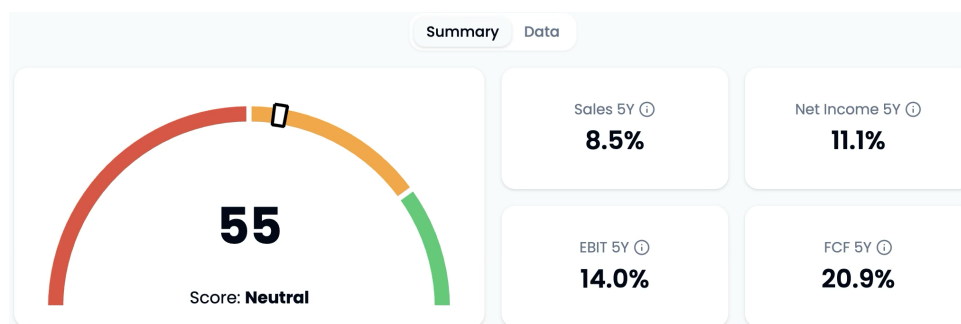


Figure 2.2: Tweenvest's Growth Score

2.1.3 Value

The Value Score measures how attractively a company is priced relative to fundamental multiples such as earnings, cash flow, sales, and dividends. This is critical for investors following value investing principles, where the goal is to buy quality companies for less than their intrinsic worth. The algorithm evaluates a series of valuation multiples—both price-based and enterprise value-based—and dividend yield.

Price-Based Multiples

- **P/E:** Measures how much investors are willing to pay per dollar of earnings
- **P/S:** Useful when earnings are volatile; shows valuation relative to sales
- **P/CF:** Reflects valuation relative to cash-generating ability
- **P/B:** Especially relevant for asset-heavy sectors like banks or industrials

Enterprise Value-Based Multiples

- **EV/Sales:** Indicates how much investors are willing to pay for each unit of revenue, providing a sense of overall company valuation relative to sales.
- **EV/EBITDA:** Shows the value the market assigns to a company compared to its operating profitability before non-cash and financing items, useful for comparing companies with different capital structures.
- **EV/EBIT:** Reflects the market value relative to operating earnings after depreciation and amortization, offering insight into core business profitability.
- **EV/FCF:** Assesses how the market values a company in relation to the actual cash it generates, highlighting the ability to produce free cash flow for investors.

Yield-Based Valuation

- **Dividend Yield:** Measures the annual return on investment from dividends, expressed as a percentage of the current share price.

Each of these ratios is compared to multiple historical statistics and sector benchmarks to create individual scores and then average them.

Note:

- $Market\ Cap = Share\ Price \times Total\ Outstanding\ Shares$
- $Enterprise\ Value = Market\ Cap + Total\ Debt - Cash + Marketable\ Securities$

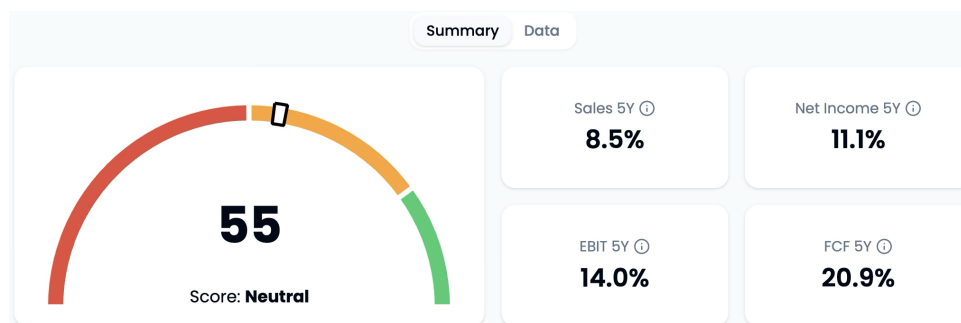


Figure 2.3: Tweenvest's Value Score

2.1.4 Dividend

When analyzing Tweenvest most common investors profile, we see a high tendency to income-focused and long-term investors using mainly dividends as their principal concern. That's why the platform dedicated a full section to this topic. The Dividend Score measures the attractiveness, reliability, and growth potential of a company's dividend payments. It helps investors assess whether the dividend is both rewarding today and sustainable for tomorrow. To do so, the score was built from three primary components:

Safety

This part of the score is used to assess the sustainability of the dividend.

- **Payout Ratio EPS:** Shows if dividends are covered by accounting earnings
- **Payout Ratio FCF:** Shows if dividends are funded by real cash generation
- **Payout Ratio Owner Earnings:** A conservative test of sustainability, excluding CAPEX.

Growth

Measures how consistently and strongly the dividend has grown over time.

- **Ordinary DPS CAGR:** 3-Year, 5-Year, 10-Year growth rates

Yield

This sub-score evaluates the attractiveness of the dividend today relative to the company's historical averages and sector benchmarks on Dividend Yield.

These components are individually scored weighted and interpolated against industry benchmarks to form a composite score.



Figure 2.4: Tweenvest's Dividend Score

Finally, as seen in figure 2.4, the algorithm adjusts this score based on how many years the dividend has been maintained or increased, **rewarding consistency**.

2.2 Code practices

Since the code that needs to be changed is used in a production environment by Tweenvest, it is important to follow some “*good practices*” to ensure the code is easy to understand, maintain, and to avoid introducing new bugs.

- **Understanding and Documenting:** Before starting to work on the code, it is important to understand the codebase and the purpose of the code. For this, it is recommended to use some tools like flux diagrams, code comments, and documentation. As it will be shown later on.
- **Testing:** To ensure proper functionality, every function and class should have unit tests that cover all possible scenarios. These tests are run automatically by the CI/CD pipeline.
- **Pair code review:** After completing the previous steps, the code undergoes a pair code review process, where at least one other team member reviews and approves the changes in *GitHub* before merging into the main branch.
- **Logging:** After each feature is implemented, it is important to analyze the actual performance in the production environment by looking at its logs to check if the feature is working as expected, and to look for any possible optimizations to be made.

2.3 Preprocessing

After implementing the required code changes, the updated Jobs are executed to extract the necessary information from Tweenvest's systems. This process generates the raw dataset, which contains the historical scores alongside other financial and market variables required for the analysis. At this stage, the data is in its initial, unrefined form—directly pulled from various sources and potentially containing inconsistencies, missing values, or irrelevant records. Before it can be used for modeling, the dataset must be carefully inspected and processed to ensure it is coherent, complete, and aligned with the objectives of the study. This preprocessing stage is critical to avoid introducing bias, ensure the reliability of the results, and retain only the data that truly contributes to the evaluation of the scores.

2.3.1 Data Consistency

Since all of the data is coming from a data provider that uses OCR as one of their tools whenever they do not have the data in a structured format, we need to be very restrictive with the data used. Also, as we mentioned before, some of the algorithms for calculating the scores use long-term data such as 10 year growths rates. This is one of the biggest **limitations to the study**: we will be missing the data of companies that stopped existing or that were acquired by other companies, if they did not exist for 10 or more years.

For inspecting the data we will be using Python as the main language, enhanced with: *pandas* [5], *matplotlib* [6], and *seaborn* [7], which are powerful tools for data manipulation and analysis. The focus will be on the following aspects:

- **Missing Values:** We need to check if there are any missing values in the data.
- **Data Distributions:** We need to make sure that the data is distributed in a way that is suitable for the modeling.
- **Variable Correlations:** We also have to see if there are any correlations between the variables. For this, we will use the correlation matrix defined as:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \quad (2.1)$$

where $\text{Cov}(X,Y)$ is the covariance between the variables X and Y , σ_X is the standard deviation of X , and σ_Y is the standard deviation of Y . This matrix will be created with a custom heatmap function (see appendix 6.4).

- **Data Patterns:** We need to check if there are any visible data patterns between the variables. To address this, we will use a custom *Pair Plot* function from *seaborn*, which is a powerful tool for data visualization and exploration depending on the type of variable (see appendix 6.4).

Additionally, we decided to assign a value of 0 to all null Dividend scores, as a null value indicates that the company does not pay dividends. This choice was made to ensure that the Dividend score remains part of the analysis when evaluating its relevance to long-term profitability, rather than being excluded due to missing values.

2.3.2 Data Transformation

To ensure a robust and well-performing model, it is often necessary to transform the data into a format more suitable for learning algorithms, with the obligatory condition of being able to revert it back to its original scale when needed.

This step is particularly important because many components of machine learning models—such as the RBF kernel in Support Vector Machines or the regularization terms (L1 and L2) in linear models—implicitly assume that features are centered around zero and have comparable variances. Without this adjustment, features with significantly larger variances could dominate the optimization process, leading the model to underweight or ignore more informative but lower-variance features.

Standard Scaler

One of the predictive models we will use are neural networks, and for improving the training process of this type of models it is really important to standardize the data. We can easily use the *StandardScaler* from the *scikit-learn* library that transforms a feature x_i with the following formula:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (2.2)$$

where μ is the mean of the feature and σ is its standard deviation. **This transformation is a very common practice**, and it results in a distribution with a mean of 0 and a standard deviation of 1.

Gaussian Transformation

In numerous modeling applications, having normally distributed features is advantageous. **Power transformations** represent a set of parametric, monotonic functions that are an extension of the Box-Cox transformation. They are designed to convert data from various distributions into approximately Gaussian distributions, thereby reducing variance fluctuations and decreasing distribution asymmetry.

We decided to use the **Yeo-Johnson transformation** for the profits response variables, because it allows for negative values and it is reversible.

$$x_i^{(\lambda)} = \begin{cases} \frac{[(x_i+1)^\lambda - 1]}{\lambda}, & \text{if } x_i \geq 0, \lambda \neq 0 \\ \ln(x_i + 1), & \text{if } x_i \geq 0, \lambda = 0 \\ -\frac{[(-x_i+1)^{2-\lambda} - 1]}{2-\lambda}, & \text{if } x_i < 0, \lambda \neq 2 \\ -\ln(-x_i + 1), & \text{if } x_i < 0, \lambda = 2 \end{cases} \quad (2.3)$$

Where λ is a power parameter that helps minimize the skewness of the data. And since we have already deleted extreme outliers, the final distribution should not be too distorted. In figure 2.5, we depict some examples of the transformation for different distributions:

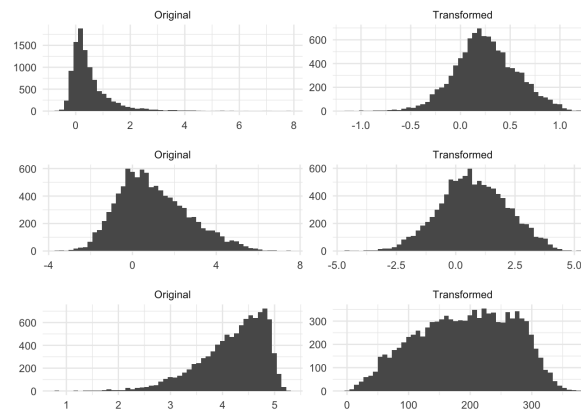


Figure 2.5: Yeo-Johnson Transformations examples

2.4 Outlier Detection

When we curate a database it is also necessary to consider anomalies. Anomalies are data patterns that have different data characteristics from normal instances. The ability to detect anomalies has significant relevance since often provide critical and actionable information in many different contexts. For example, anomalies in credit card transactions could signify fraudulent use of credit cards, or an unusual computer network traffic pattern could stand for an unauthorized access.

2.4.1 Inter Quartile Range

Given a univariate dataset, the Interquartile Rule identifies outliers based on the Interquartile Range (IQR). The steps are as follows:

1. Compute the first quartile (Q_1), which is the 25th percentile of the data.
2. Compute the third quartile (Q_3), which is the 75th percentile of the data.
3. Calculate the interquartile range:

$$\text{IQR} = Q_3 - Q_1$$

4. Define the lower and upper bounds for non-outlier values following the 1.5 rule:

$$\text{Lower bound} = Q_1 - 1.5 \times \text{IQR}$$

$$\text{Upper bound} = Q_3 + 1.5 \times \text{IQR}$$

5. Any data point x is considered an outlier if:

$$x < \text{Lower bound} \quad \text{or} \quad x > \text{Upper bound}$$

As we can see this is a very simple and intuitive method, but it has some limitations:

Pros	Cons
Simple to compute and interpret.	Not suitable for multivariate data with correlated features.
Non-parametric.	May misclassify values in skewed distributions as outliers.
Robust to extreme values, since it relies on percentiles rather than the mean.	Fixed multiplier is arbitrary and may need tuning for each dataset.

Table 2.1: Pros and Cons of the Interquartile Rule

Even though this method is simple to compute and interpret, it is not suitable for multivariate data with correlated features, and that is where the rest of the methods come into play.

2.4.2 Single Vector Machine

The Single Vector Machine (SVM) represent a robust family of unsupervised learning algorithms that can be effectively adapted for anomaly detection tasks. This algorithm establishes itself on the premise that the majority of real-world data is inherently normal. Where the goal is to define

a boundary encapsulating the normal instances in the feature space, thereby creating a region of familiarity.

To capture the bases of the *scikit-learn* implementation [8] used in this work, we can say that the main idea is to find a minimum volume sphere that contain all the training samples. This sphere, described by its center c and its radius r , is obtained by solving the following optimization problem [9]:

$$\min_{r,c} r^2 \quad \text{subject to} \quad \|\Phi(x_i) - c\|^2 \leq r^2 \quad \text{for } i = 1, 2, \dots, n \quad (2.4)$$

This boundary is strategically positioned to maximize the margin around the normal data points, allowing for a clear delineation between what is considered ordinary and what may be deemed unusual. This emphasis on margin maximization is akin to creating a safety buffer around the normal instances, fortifying the model against the influence of potential outliers or anomalies. In summary, this method has the following characteristics:

Pros	Cons
Effective for high-dimensional data and complex boundaries.	Sensitive to the choice of kernel and hyperparameters (e.g., ν , γ).
Works well when the training set contains mostly or only normal instances.	Computationally intensive, especially on large datasets.
Can capture nonlinear patterns using kernels (e.g., RBF kernel).	Difficult to interpret or explain the decision function.
No need for labeled data from the outlier class.	–
Solid theoretical foundation and widely supported in machine learning libraries.	–

Table 2.2: Pros and Cons of One-Class SVM

2.4.3 Isolation Forest

As we are seeing in this small fraction of outlier detection methodologies, most of the existing anomaly detection approaches are based on the premise of normal distributions, then identify anomalies as those that do not conform to the normal profile. But to solve this issue we can use the Isolation Forest algorithm, which is explained in the original paper F. T. Liu et al. [10]. The algorithm constructs multiple isolation binary-trees (iTrees) from the dataset, where anomalies are identified as data points that exhibit shorter average path lengths across these trees.

Lets consider a dataset $X = \{x_1, \dots, x_n\}$ containing n points in d -dimensional space, and a subset $X' \subset X$. An Isolation Tree (iTree) is constructed as follows:

- Each node T in the tree is either:
 - An external node (leaf) with no children, or
 - An internal node with exactly two children (T_l and T_r) and a test condition
- The test condition at each internal node consists of:
 - A randomly selected attribute q

- A split value p
- A test $q < p$ that determines whether a point goes to T_l or T_r

The tree construction process recursively partitions X' by randomly selecting attributes and split values until either:

- The node contains only one instance, or
- All instances at the node have identical values

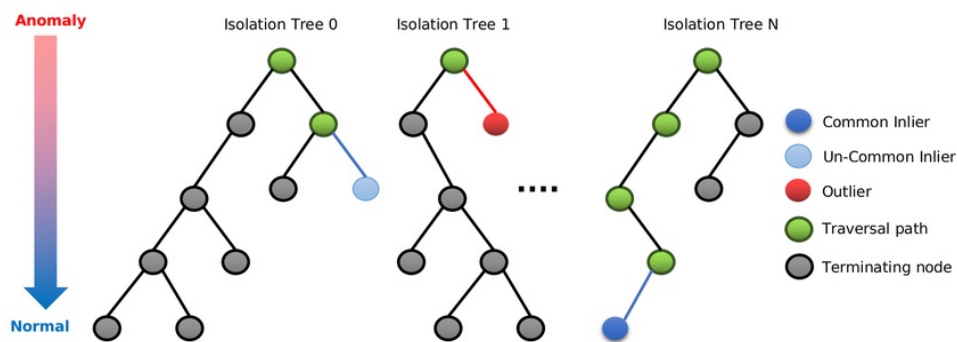


Figure 2.6: Visualization of how IF works [11].

As we can see in figure 2.6, once the iTree is fully constructed, each point $x_i \in X$ is isolated at a leaf node. The path length $h(x_i)$ of a point x_i is defined as the number of edges traversed from the root to its leaf node. Points with shorter path lengths are considered more likely to be anomalies, as they require fewer splits to be isolated from the rest of the data.

This method works very good in most cases, but it has also its own limitations:

Pros	Cons
Efficient and scalable to large datasets due to its tree-based structure.	Less effective for detecting local outliers in densely clustered regions.
Non-parametric: does not assume any data distribution.	Performance can vary with the choice of parameters like number of estimators and subsample size.
Naturally handles high-dimensional data.	Not suitable for detecting subtle anomalies that closely resemble normal data.
Requires little preprocessing (no need for feature scaling).	–
Robust to irrelevant features due to random sub-sampling.	–

Table 2.3: Pros and Cons of the Isolation Forest Method

2.4.4 Local Outlier Factor

Since we are looking at outliers from the perspective of correlations between the different response variables to see which companies do not behave naturally. We also checked on the Local Outlier Factor (LOF) for being based on hyper-plane densities of data. The algorithm

measures the local deviation of the density of a given sample with respect to its neighbors. It is local in that the anomaly score depends on how isolated the object is with respect to the surrounding neighborhood. More precisely, locality is given by k -nearest neighbors, whose distance is used to estimate the local density. By comparing the local density of a sample to the local densities of its neighbors, one can identify samples that have a substantially lower density than their neighbors. These are considered outliers. [12]

Table 2.4: Pros and Cons of the LOF Method

Pros	Cons
Can detect outliers relative to their local neighborhood, allowing it to identify context-specific anomalies.	LOF scores are relative and lack a universal interpretation threshold for outliers.
Effective in datasets with varying densities — unlike global methods, it can recognize outliers near dense clusters.	Sensitivity to parameter selection (e.g., number of neighbors) can affect performance and consistency.
Applicable in any domain where a dissimilarity measure is defined, not limited to vector spaces.	Can detect outliers patterns that could create a bias for models, because they do not belong to the normal distribution.
Works well across domains, such as network intrusion detection or classification tasks.	—
Easily generalizable and adaptable for use in spatial data, temporal data, and network structures.	—

2.4.5 Multi-Criteria Outlier Detection

There is a big controversy in the field about the best method to detect the outliers, since the results can vary a lot depending on the algorithm used.

So to attempt to fix this issue, after reading some aggregative methods such as Abro et al. [13], we created a multi-criteria method that combines the results of the different methods to get a more robust result.

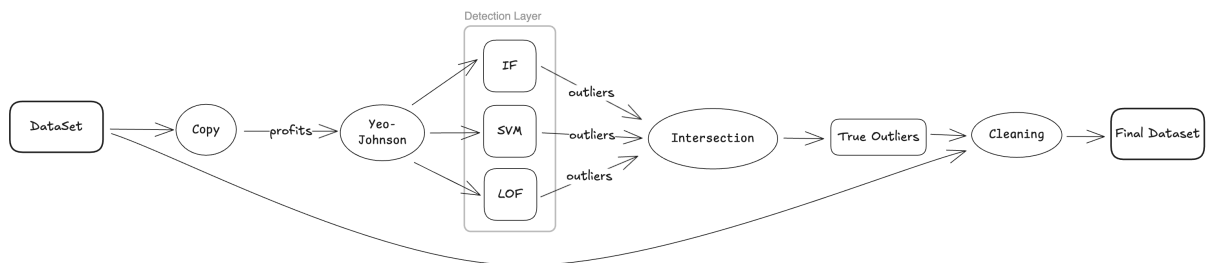


Figure 2.7: MCOD Schema

As shown in the figure 2.7, this starts by homogenizing the variables distribution with a common data transformation so they are more Normal, but to make sure that the final data is not being transformed we copy the original dataset earlier. After this, we use multiple methods to detect the different outliers and then we make the intersection between them to get the common

outliers. By the end of this process we have tagged the original dataset rows, assuring that the data that will be excluded is only formed by “true outliers”.

Island Cases detector

Later on we will see how this method was causing to save some points that should be excluded, causing a worsening in the performance. So after expecting the behavior of each of the individual methods, we saw where the flaw was and decided to create a simple variation of the MCODE method that should only keep the normal data but also saving those islands of data for further future analysis.

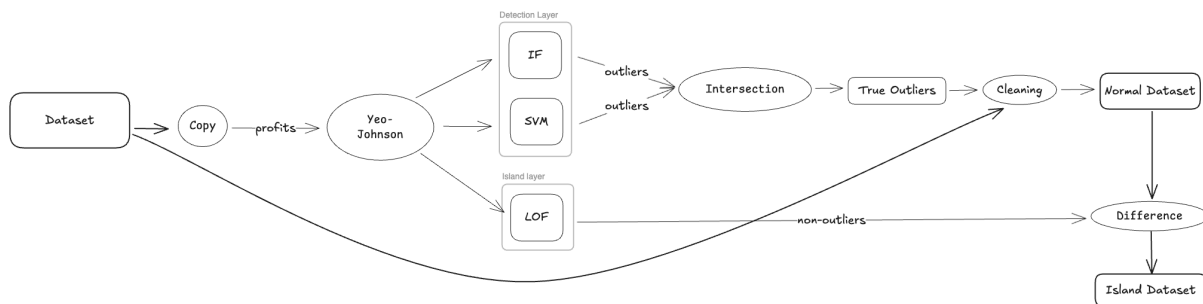


Figure 2.8: ICD Schema

When looking at the figure 2.7, we see that the LOF is saving different data points that are not common with the rest of the methods, this will be our detector of *islands*. Also we create a similar intersection between the other methods that will get the “true outliers”. Later we take the difference between these two new datasets to get the *normal data*, the *extreme but consistent data* and the *true outliers*. We will see how this worked out in the results chapter.

Important: Following the aim of this work, we will not be analyzing the *extreme cases* but we propose it for future work.

2.5 Predictive Models

The main objective of this thesis is to check if the scores have any relation with the future profits of the companies. For doing this in the most generalized way, instead of creating a portfolio following some subjective criteria on the scores, markets and other variables, we used predictive modeling to check for the actual relationships between the response variables and the explanatory ones.

Since the main objective is to check the relations between the profits and the scores, we used regression models to understand it those relationships. Starting with simple linear regression and then generalizing to non-linear models. And finally to test if we can also predict future profits we used Neural Network (NN).

2.5.1 Regression Models

To create the econometrical models, we wanted to be able to understand the actual relationship between the response variables and the explanatory ones. For this we started with the creation

of multiple regression models.

Linear Regression

As for the first models, we created a family of linear regression models ($i \times j$) for each factor i -return and j -score. These models include the interactions between the dummy variables and the j -score used in each model.

$$\mu_{i,j} = \beta_0 + \beta_1 \cdot S_j + \sum_{n=2}^N \beta_n \cdot D_n + S_j \cdot \sum_{m=N+1}^M \beta_m \cdot D_m \quad (2.5)$$

where:

- $R_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma_{i,j}^2)$ is the return for each time horizon i and factor j . So $R_{i,j} = \mu_{i,j} + \varepsilon_{i,j}$ where $\varepsilon_{i,j}$ is the error and $\mu_{i,j}$ is the expected return.
- S_j is the score for each factor j
- $D_{n,m}$ are the dummy variables for the industry and region, where $D_{n,m} := I_k \cup G_l$
- β are the coefficients that weight each variable.

For fitting the models, we also applied a **backward selection** to remove the variables that are not statistically significant, using the p -value criterion **with a 20% tolerance**. This threshold will be used for most of the analysis because in financial data it is harder to get high precisions.

Generalized Additive Models

Since the relationships between the variables are most likely non-linear, we also used the Generalized Additive Models (GAM) for the mentioned ($i \times j$) family to check if they are able to improve the performance, and also see if there are consistent patterns across the different time horizons.

With this analysis we could then let know Tweenvest what are the "best" different scales for each score to better guide its users. Since right now –as it was shown in figures 2.1 to 2.4– they are scaled from [0,50] as a bad score, (50, 80] as a good score, and (80, 100] as a great score. And this has not been tested yet.

According to Servén and Brummitt [14], GAMs are smooth semi-parametric models that can capture non-linear relationships between variables. They take the form:

$$g(\mathbb{E}[y|X]) = \beta_0 + f_1(X_1) + f_2(X_2, X_3) + \dots + f_M(X_N) \quad (2.6)$$

where:

- $X^T = [X_1, X_2, \dots, X_N]$ are the explanatory variables.
- y is the dependent variable, in our case the returns.
- $g()$ is the link function that relates the predictor variables to the expected value of the dependent variable.

- $f_m()$ are feature functions built using penalized B-splines, which automatically model non-linear relationships without requiring manual transformation of variables.

In our case, we used pyGAMs LinearGAM since the model gives a Normal error distribution, and an identity link. Using this library also gives us the possibility to differentiate between the different variable types:

- **s(i) - Smooth Splines:** Standard smooth splines (cubic by default) for continuous variables like scores: $f_m(X_m) = \sum_{j=1}^k \beta_{mj} B_j(X_m)$ where B_j are B-spline basis functions
- **f(i) - Factor Splines:** For categorical variables (sectors and regions): $f_m(X_m) = \sum_{j=1}^k \beta_{mj} I_j(X_m)$ where I_j are indicator functions
- **te(i,j) - Tensor Splines:** For modeling non-linear interactions between variables: $f_m(X_i, X_j) = \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} \beta_{kl} B_k(X_i) B_l(X_j)$

In figure 2.9 we can see an example of how GAMs can show the non-linear relationships between variables.

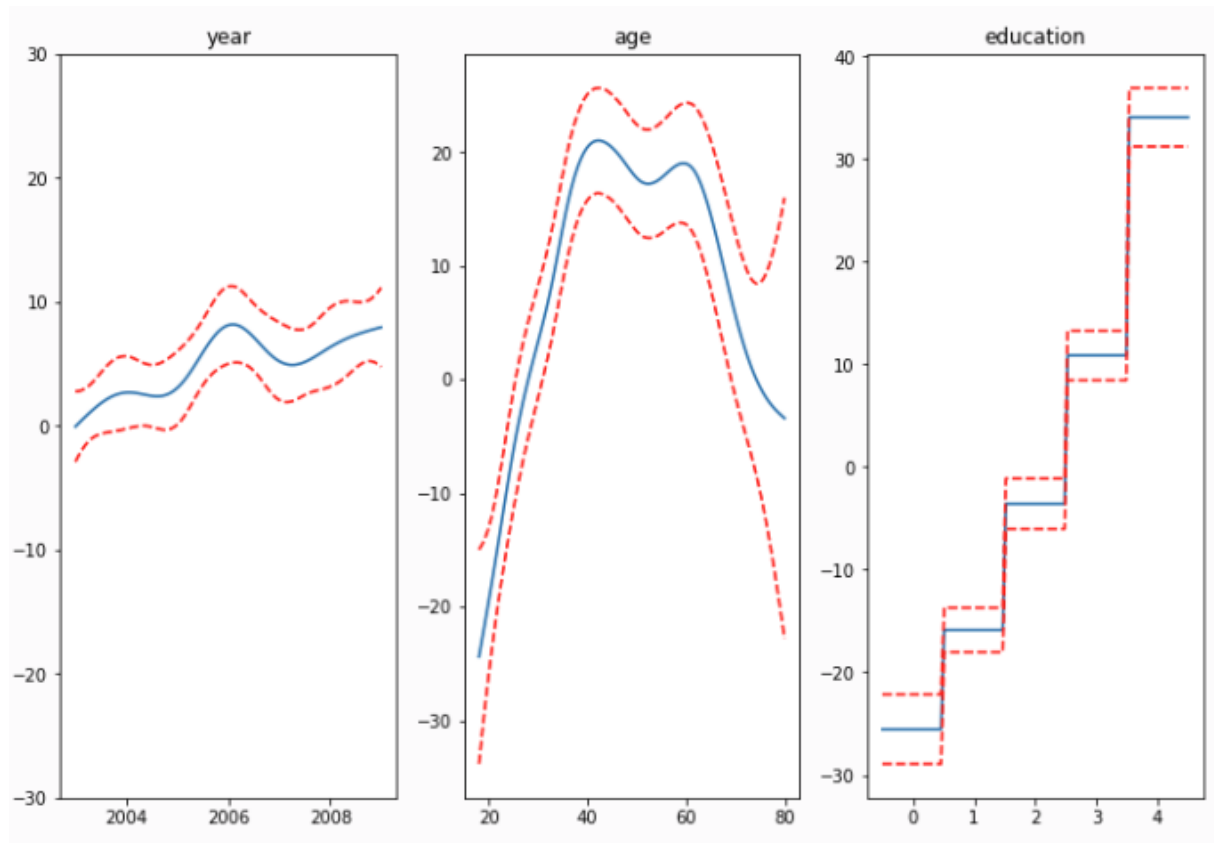


Figure 2.9: pyGAM - Example of GAMs regression splines

2.5.2 Time Series

In the stock market, the prices of the stocks are not independent from each other so as we will see with the descriptive analysis, they are correlated in the time series and they also have memory on past behavior of the company. This is what is known as **Momentum**, so companies that have a good past performance are more likely to have a good future performance, and vice versa. For this reason, we contemplated two possibilities:

ARIMA Models

To take account the profit tendency, we implemented ARIMA models to improve the regression models performance by using the residues of the already fitted model:

$$P_i = \hat{P}_i + \varepsilon_i \quad \longleftrightarrow \quad \varepsilon_i \sim \text{ARIMA}_i(p, d, q) \quad (2.7)$$

The ARIMA model is defined as:

$$\phi_i(B_i)(1 - B_i)^{d_i}\varepsilon_i^t = \theta_i(B_i)\eta_i^t \quad (2.8)$$

where for each i profit:

- ε_i^t is the residual at time t .
- η_i^t is a white noise error term (innovation).
- B_i is the backshift operator, such that $B\varepsilon_i^t = \varepsilon_i^{t-1}$.
- $\phi_i(B) = 1 - \phi_{i1}B - \phi_{i2}B^2 - \dots - \phi_{ip}B^p$ is the autoregressive (AR) polynomial.
- $\theta_i(B) = 1 + \theta_{i1}B + \theta_{i2}B^2 + \dots + \theta_{iq}B^q$ is the moving average (MA) polynomial.
- d_i is the order of differencing.

But this approach supposed a *big issue* that we will explain in the development section.

Windowed Models

So for trying to not fall into the same mistake and for capturing possible memory of the companies behavior in different aspects, we also implemented windowed models that used the last N scores statistical properties to predict future profits.

So for each score j , we calculated the **average of the scores** over the last N periods. Let S_j^t be the score at time t , then for a window of size N :

$$\bar{S}_j^N = \frac{1}{N} \sum_{i=t-N+1}^t S_j^i \quad (2.9)$$

where:

- $\bar{S}_j^N$ is the average score over window N for score j
- N can be 3 months, 6 months, 1 year, or 2 years.
- t is the current time point

For adding also the deviation from the average, we also calculated the **standard deviation of the scores** over the same window:

$$\sigma_j^N = \sqrt{\frac{1}{N} \sum_{i=t-N+1}^t (S_j^i - \bar{S}_j^N)^2} \quad (2.10)$$

These statistical measures are used as additional features in our predictive models to capture the temporal behavior of each score.

2.5.3 Neural Networks

NN are a class of machine learning models inspired by the structure and functioning of the human brain. They consist of an **input layer**, one or more **hidden layers**, and an **output layer**. The input layer receives the explanatory variables or regressors, which are then processed through the hidden layers where the model learns to represent and combine them in increasingly abstract ways. Within these hidden layers, nonlinear activation functions allow the network to capture complex, nonlinear relationships between variables—patterns that traditional linear regression or even more flexible methods such as GAM may fail to detect. Finally, the output layer produces the model's prediction, which in this thesis corresponds to the expected profit or return over a given investment horizon.

The motivation for using NN in this context lies in their ability to approximate highly complex functional relationships without requiring explicit specification of their form. Whereas linear or additive models are constrained by predefined assumptions about the shape of relationships, neural networks can automatically learn intricate dependencies and interactions among variables, making them particularly suitable for modeling financial data where nonlinearity and high-dimensional interactions are common. figure 2.10 illustrates the basic architecture of a neural network.

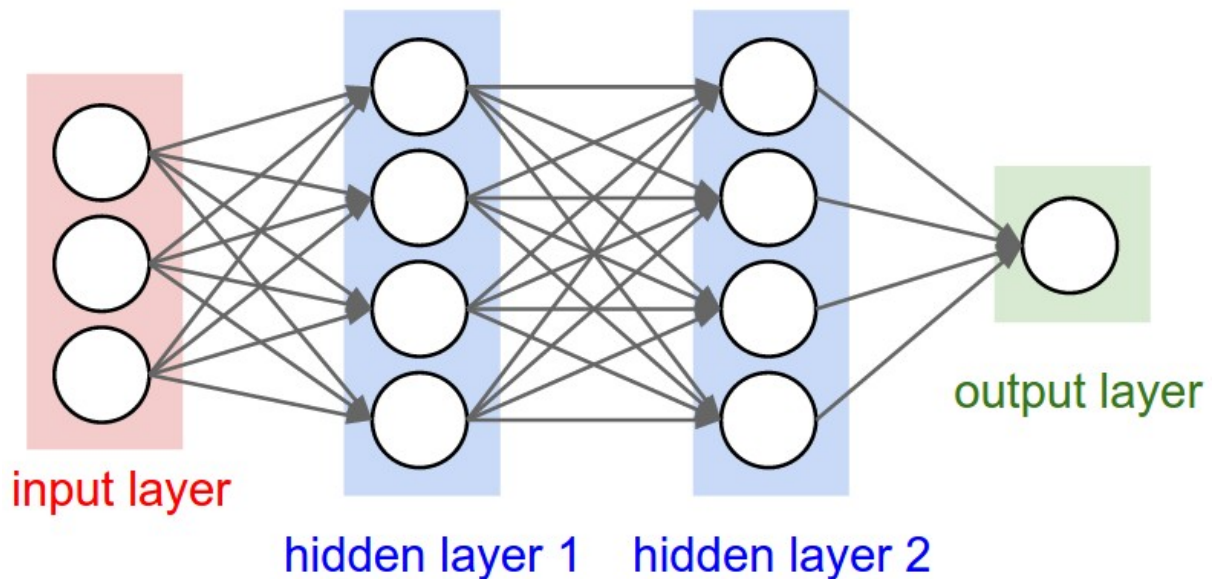


Figure 2.10: Simple Architecture of a NN

Finally, to test if we have enough information to model future profits, we will use two different approaches of different types of NN: **regressors** and **classifiers**.

Regressor Neural Networks

A **regressor neural network** is a type of artificial neural network specifically designed to predict continuous numerical values, rather than discrete categories. In the context of our analysis,

these networks are trained to estimate future profit values based on a set of input features, such as historical scores and other relevant company data. For our implementation, we utilize the **MLPRegressor** from the *scikit-learn* library [15], which provides a flexible and efficient framework for building and training these models.

Classifier Neural Networks

To simplify the implementation, we are going to use two different quantification of the profits:

- **Binary Classification:** In this approach, we partition the profit variable into two classes using the following rule:

$$y = \begin{cases} 0 & \text{if } P \leq 0 \\ 1 & \text{if } P > 0 \end{cases} \quad (2.11)$$

where P denotes the profit for a given time horizon, and y is the resulting binary class label.

- **Multi-class Classification:** Here we created different intervals of profits, using the average inflation rate and the S&P 500 index as references– [16, 17].

$$y = \begin{cases} -3 & \text{if } P_a \leq -0.1 \\ -2 & \text{if } -0.1 < P_a \leq -0.04 \\ -1 & \text{if } -0.04 < P_a \leq 0 \\ 1 & \text{if } 0 < P_a \leq 0.04 \\ 2 & \text{if } 0.04 < P_a \leq 0.1 \\ 3 & \text{if } 0.1 < P_a \end{cases} \quad (2.12)$$

Where P_a is the **annualized profit**, calculated as:

$$P_a = (1 + P)^{\frac{12}{n}} - 1 \quad (2.13)$$

Here, P is the total profit over a period of n months, and P_a is the equivalent annualized return.

And for building these NN we will use the **MLPClassifier** from the *scikit-learn* library [18], which provides a flexible and efficient framework for building and training these models.

Architecture Exploration

Before building the models, we need to explore different architectures to find the best hyper-parameters for the models. In our case we are only going to explore the different hidden layers and the number of neurons in each layer, since the rest of the hyper-parameters are already being optimized by the library. To systematically explore the effect of the architectures, we will use the following configurations:

- **Single hidden layer:** Networks with one hidden layer and varying width:
 - (16), (32), (64), (128), (256), (512)
- **Two hidden layers:** Networks with two hidden layers, both with decreasing and uniform widths:

- (64, 32), (128, 64), (256, 128), (512, 256), (128, 32), (256, 64)
- (32, 32), (64, 64), (128, 128), (256, 256)
- **Three hidden layers:** Networks with three hidden layers, including both bottleneck and uniform structures:
 - (128, 64, 32), (256, 128, 64), (512, 256, 128), (64, 32, 16), (256, 64, 16), (128, 64, 8)
 - (128, 128, 128), (256, 256, 256)
- **Four hidden layers:** Deeper networks with four hidden layers, both with gradually decreasing and large bottleneck structures:
 - (512, 256, 128, 64), (256, 128, 64, 32), (128, 64, 32, 16), (64, 32, 16, 8), (1024, 512, 256, 128)
- **Hard Bottlenecks:** Networks where the number of neurons sharply decreases in each subsequent layer:
 - (256, 64, 16), (128, 32, 8), (512, 128, 32), (1024, 256, 64)

This comprehensive set of architectures allows us to compare the impact of network depth, width and *bottlenecking* on predictive performance, and to identify the optimal configuration for each problem. So we should end up with a set of best architectures for each score and profits, depending on the NN type.

2.6 Back-testing

In order to assess and compare the modeling techniques employed in this thesis, it is important to have a consistent criteria to evaluate them. Firstly we have to make clear that the main intent of this work is not to find the model that best predicts future profits, but to check whether the scores are correlated with the profits, and if so, how.

2.6.1 Training and Testing

Knowing this, as we already explained, we will use the linear models and the GAM to check the correlations and the statistical significance of each variable. Then we will use the NN to actually test the true predictability of the models.

- **Training:** We will use 80% of the stocks of the dataset to train the models, excluding the last 4 months of each stock.
- **Testing:** When it comes to testing, we want to check the NN-models in two different situations:
 - **Unknown stocks:** Here we are using 20% of the stocks left of the dataset, to see if just with the explanatory variables we can predict the future profits of those stocks.
 - **Known stocks:** For these tests we will use 80% of the stocks, that we used to train the models, but we will use the last 4 months of each stock to test if the models are able to predict when they already know past behavior of the stock.

2.6.2 Performance

The first thing we need to see is if the models are able to detect any relationships between the co-variables and the response variable, and how accurate they are:

- **Prediction capability:** we want to see if the models are able to predict the future profits, so we need to quantify the proportion of variance explained by the model. For this we will use:

- **The coefficient of determination**, defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.14)$$

where y_i are the observed values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of the observed values.

- **The adjusted coefficient of determination**, to take into account the number of predictors in the model:

$$R_{adjusted}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \quad (2.15)$$

where n is the number of observations and p is the number of predictors. Since we are also using a NN Classifier, we cannot use the same statistics as for regression.

- Thus, for prediction capability in classification, we will use the **F1 Score** from *sklearn*, defined as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.16)$$

where $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{Recall} = \frac{TP}{TP+FN}$, with TP being true positives, FP false positives, and FN false negatives.

This interprets as how many correct positive predictions the model makes relative to all positive predictions and actual positives, so a higher F1 score indicates better classification performance.

- **Accuracy:** We also need to check on how much error the model is making when adjusted. For this we used:

- **The Mean Absolute Error**, defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.17)$$

- **The Root Mean Square Error**, for penalizing the larger errors or possible excess of outliers:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.18)$$

And again, for classification we decided to use an equivalent to the MAE:

- **Accuracy Score:** Using the *scikit-learn* implementation, the accuracy score measures the proportion of correct predictions relative to the total number of predictions made. To express this metric as an average error rate, we subtract the accuracy score from 1, yielding:

$$\text{Error Rate} = 1 - \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (2.19)$$

- **Log Loss:** For mimicking the penalization for large errors we are using the log loss from *sklearn*:

$$\text{Log Loss} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{i,j} \log(\hat{y}_{i,j}) \quad (2.20)$$

where in this context, $y_{i,j}$ represents the true class label, and $\hat{y}_{i,j}$ represents the model's predicted probability for that class.

2.6.3 Consistency and Overfitting

But a good performance does not mean that the models are explaining the behavior of the variables, they could be overfitting the data. To detect this, we need to check if the models reflect stable, interpretable patterns in the relationships between scores and the *alpha*. For this, we will use the following criteria:

- **Sign of coefficients:** We will check if the sign of the scores coefficients remains stable across different time horizons. This would tell us that the score behaves as expected.
- **Shape of functions:** We will check if the shape of the spline functions of the GAMs remains stable across different time horizons.
- **Number of significant variables:** We will check if the number of significant variables increases with the performance of the models, using the *p-value* criterion.

And for the NN, since we cannot access the internal structure of the models, we are going to use the **SHAP** to check the importance of the variables in the predictions, and then use the same criteria for looking at the consistency of the variables in the models—Lundberg and Lee [19].

2.6.4 Robustness of Results

Even when the models perform well and show good consistency, it is essential to verify that their residuals resemble **white noise**. In the context of regression models—such as GAM and linear regression—this requirement stems from the assumptions underlying valid statistical inference. White noise residuals indicate that the model has successfully captured all systematic structure in the data, leaving only random, uncorrelated noise with constant variance. If residuals exhibit autocorrelation, trends, or heteroskedasticity, it suggests that relevant patterns remain unexplained, meaning the model may be misspecified or omitting important variables.

Therefore, ensuring white noise residuals is a key diagnostic step to confirm that the relationships modeled are both complete and interpretable. This will be assessed through a combination of visual inspection and formal statistical tests Enders [20].

- **Mean Value Test:** The mean of the residuals should be close to 0.
 - **Distribution Plots:** Visualize the distribution of residuals to assess if the mean is centered at zero.
 - **T-test:** Checks if the mean of the residuals is close to 0.
- **Heteroskedasticity Tests:** The variance of the residuals should be constant across all time horizons.
 - **Residuals vs Fitted Plots:** In a model with homoscedasticity (constant variance), the residuals should be randomly scattered around zero with a roughly constant spread.

- **White Test:** Checks for general forms of heteroskedasticity.
- **Breusch-Pagan Test:** Tests whether the variance of the errors depends on the values of the independent variables.
- **Normality Tests:** The distribution of residuals should follow a Normal distribution.
 - **Q-Q Plots:** Visual tool to compare the quantiles of the residuals to a normal distribution. If the data points in the Q-Q plot fall close to a straight diagonal line, it suggests that the data's distribution is similar to the theoretical distribution
 - **Shapiro-Wilk Test:** Statistical test for normality.
 - **Jarque-Bera Test:** Tests whether sample data have the skewness and kurtosis matching a normal distribution.
- **Autocorrelation Test:** The errors should not be correlated with each other.
 - **Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) Plots:** Visualize the autocorrelation and partial autocorrelation of residuals.
 - **Durbin-Watson Test:** Checks for the presence of autocorrelation in the residuals.

Note: These residual diagnostics are not required for neural network (NN) models, as NNs do not rely on the same statistical assumptions as regression models.

Chapter 3

Development

This chapter details the practical implementation of the methodology described earlier, focusing on the steps taken to adapt Tweenvest’s infrastructure, create the datasets, and prepare them for analysis. The development phase bridges the conceptual framework with the operational execution, requiring modifications to the existing system architecture, the design of new data extraction jobs, and the deployment of solutions on dedicated servers. These tasks were carried out in close collaboration with Tweenvest, ensuring that all changes integrated seamlessly with the company’s production environment while maintaining data integrity and traceability.

3.1 Architecture Enhancement

At the current moment of beginning this work, Tweenvest kept a large Database (DB) of most of the needed information for creating the dataset. This included all of the price histories, historical currency multipliers to dollar, dividends payed per stock, etc. but when it came to the scores we had a traceability problem. For optimizing the costs and structure of the DB, Tweenvest chose to only save the latest score for each company and overwriting it each day when calculating the newest one, which led to the **first main task**.

3.1.1 Updating Tweenvest Code

After locating where the score calculators were called, we had to understand the code’s architecture and the relationships between the different database models to see exactly what needs to be changed. Tweenvest uses *Django* as the main backend’s technology to handle a relational database with *PostgreSQL*, this DB was chosen due to the nature of the financial data where information is related to another.

For handling all of the different technologies and environments, Tweenvest uses **Docker** [21] to create a virtual machine that contains all of the needed dependencies to run the platform. This way, they have a consistent environment for developing, testing and deploying the code separately. So as we can see in figure 3.1, each of the components is running in a different container.

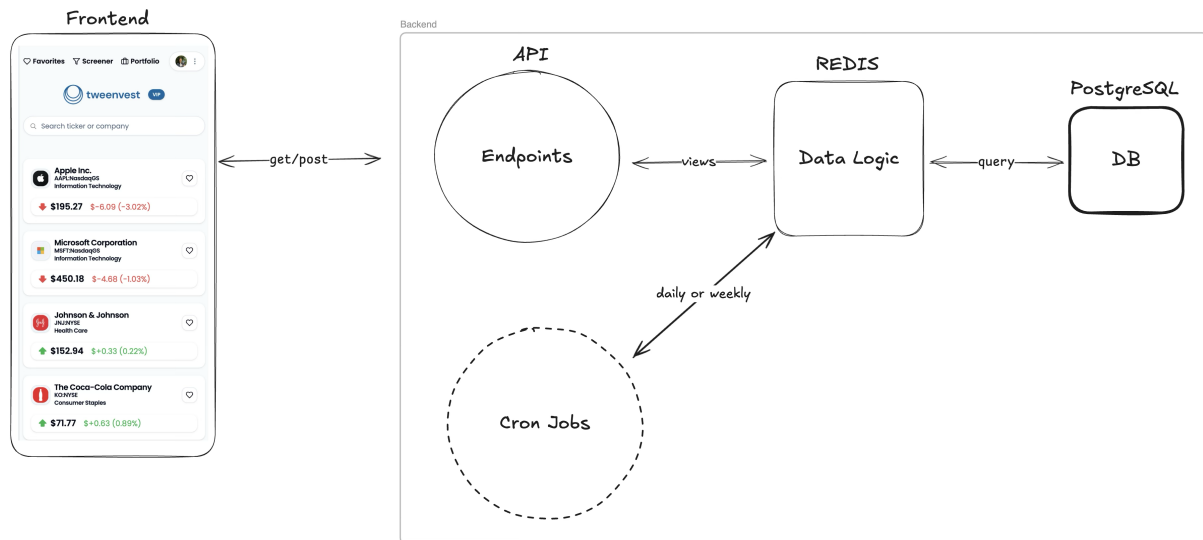


Figure 3.1: General Tweenvest's Architecture

Once we had a clear understanding of the architecture, we could look at the two different actions occurring at the same time:

1. Requests originating from user interactions.
2. Jobs that must be executed daily, weekly or at the moment, such as:
 - Updating all financial information of the stocks.
 - Performing necessary actions like sending emails to deliver authentication pins.

Since we were changing the data logic for calculating and storing historical data of the factor scores, we needed to assure that the systems capabilities do not exceed the platforms needs, and that the changes do not affect the normal calculation of the factor scores, so we had to look at the corresponding code of the factor scores.

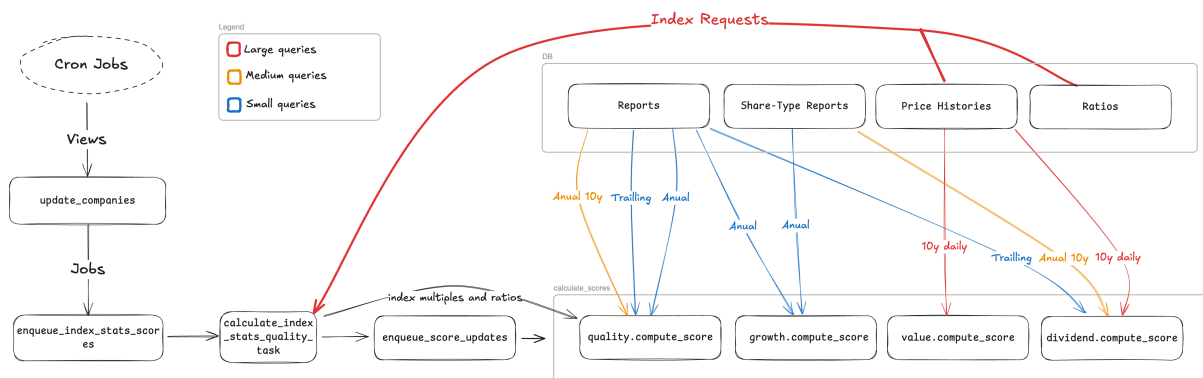


Figure 3.2: Scores Calculations Code Schema

The code integrates several methods and data-flows, so we created the schema shown in figure 3.2 to simplify the logic behind several functions related to the calculations of each factor score. The scheduler triggers a request that is sent to the backend through the *API* views, which then execute a series of functions responsible for calculating the scores for each factor. Seeing the chronology of the events, and the different queries to the DB, we quickly realized that we

had to modify the queries to include an extra filter *–calculation_date–* to get the models related to the score at any time, this way we could set it as “today” whenever we are calculating the present scores, not modifying the normal behavior of the system; while also having the ability to propagate the calculation to a specific past date.

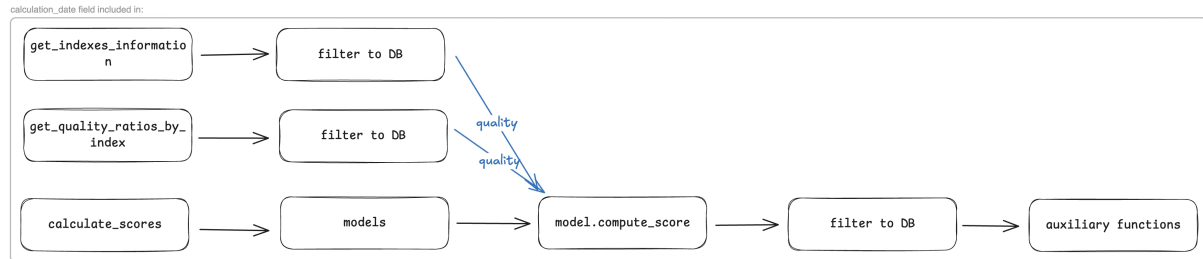


Figure 3.3: *calculation_date* Field Propagation Schema

For not adding redundant code, we had to inspect the propagation schema shown in figure 3.3 to determine the root changes from the main functions: *calculate_scores*, *get_indexes_information* and *get_quality_ratios_by_index*. These are the ones used by the production *Cron* to calculate the factor scores daily, so for a normal function call we set the *calculation_date* to *None* and whenever we want to calculate the scores for a specific date we set it to the desired date.

This helped us create very readable logic to adapt the filters to both scenarios, for example in the value score shown in Listing 3.1:

```

1 if calculation_date is None:
2     calculation_date = date.today()
3     filters = {
4         "stock": stock,
5         "date__gte": calculation_date - relativedelta(years=10),
6     }
7 else:
8     filters = {
9         "stock": stock,
10        "date__gte": calculation_date - relativedelta(years=10),
11        "date__lte": calculation_date,
12    }
13 phs = (
14     PriceHistory.objects.filter(**filters)
15 )

```

Listing 3.1: Value Score Filtering

So as we can observe, the main change for this part of the code was to add a date filter to get only data “lower than or equal” to the *calculation_date*. This may seem a bad approximation, but it has to be done this way due to the nature of the data: most of the financial data come from annual, semestral, and quarterly reports, except to the price history that is updated daily.

Along the way of analyzing the code’s structure we made a small refactoring to simplify the legibility:

- **Renaming functions** whenever they are internal methods or generic auxiliary methods called by different functions.

- **Externalizing mocking functions** that help create fake models and data for *Unit Tests*.
- **Homogenize the code's structure**, field types, to help with readability and keeping a standard order along the backend and frontend.

Another very important part of these updating procedure was to **adapt the Unit Tests** to make sure that the original code still behaved as expected, **and creating new ones** to check that the changes calculated the factor scores properly.

3.1.2 Designing new Jobs

Once all tests have passed and the team had approved the changes, the next step was to create new Jobs to populate the current DB with the historic factor scores while maintaining the servers performance in normal levels. For this, we had to update a basic method, *bulk_create_or_update*, used many times through the whole codebase and add a new conditioning so it can handle conflicts depending if the models are empty or not. After reading the official Django Documentation [22], we saw that in the newest version available they had included the needed fields to its original method *bulk_create*, so **we had to update the Django version** in our *Docker* machine, making sure that it did not introduce any unwanted issues.

After this was implemented, we decided to create three different jobs for enqueueing the tasks and being able to separate the calculations. This way we could restart them in case something broke during the process (see appendix 6.1). Also, we had to manage the server resources, so for assuring the platform's performance, we used the following hierarchy (see Listing 6.4):

1. Seven workers for primary daily jobs from different API data integration
2. One dedicated worker for guaranteeing email sending, and 6 prioritized shared workers
3. Four shared workers with lower priority, except for 1 which is set higher than email.

A crucial part of this code was replicability, equal distribution through different sectors, and mostly to assure that the companies existed during enough time for the factor scores to calculate the 10 years averages.

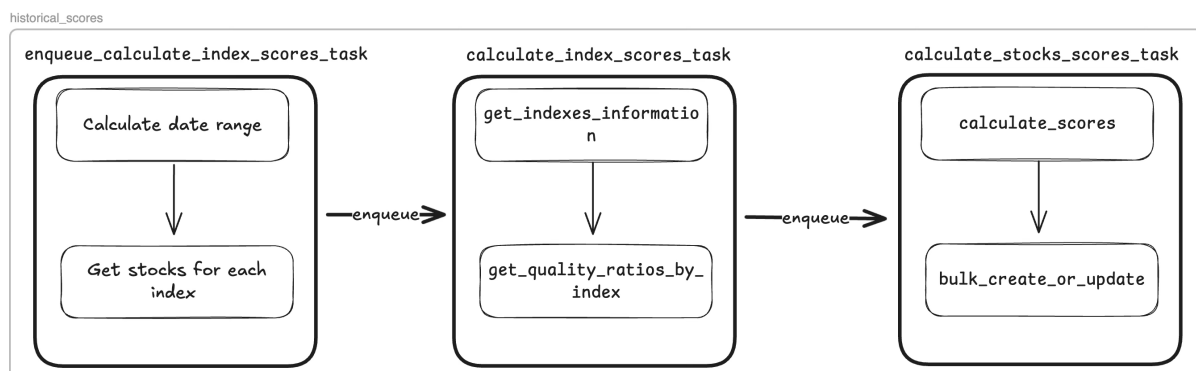


Figure 3.4: Historical Jobs First Iteration Schema

We created the first jobs with the schema shown in figure 3.4, and during the first small test it worked in local, but when tested with the production DB we started to see important issues with how the functions where designed due to the large amount of data being processed:

- Passing incorrect arguments between jobs—such as entire lists of stock objects—led to the REDIS DB reaching its maximum storage capacity.
- Jobs ran out of time, due to the large amount of data being processed.

We then changed the day range logic to enqueue a *calculate_index_scores_task* for each day, so the job did not have to calculate all of the index data for the dates all at once, as shown in figure 3.5. Also, we modified the function *args* to only uses models identifiers "id" to reduce the REDIS memory used. The complete implementation of this job can be found in appendix 6.1.

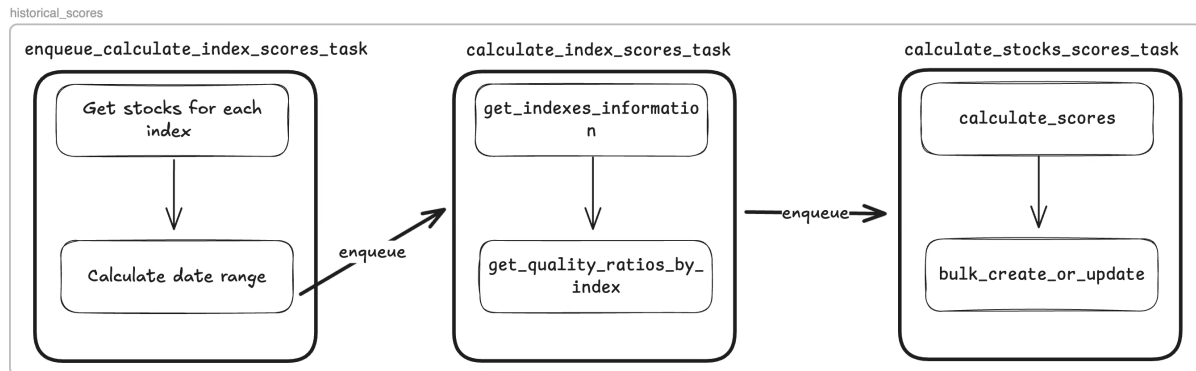


Figure 3.5: Historical Jobs Final Version Schema

3.1.3 Telemetry Tracing

Finally we had the functional jobs for calculating the scores, and ran the first calculation for a dataset of 1,500 stocks. After the first jobs finished we quickly realized that they were taking too much time to process, and saw a disproportionate amount of queries to the DB so we started a process to check on where was the bottleneck using telemetry tracing tools.

To do so, we had to inspect each method and function that was being used through the calculations of the factor scores. Luckily we had just recently implemented *DataDog* which is a software that provides monitoring and analytics for applications and infrastructure, offering real-time metrics, event monitoring, and end-to-end tracing.

To apply the tracer, we simply had to add a specific *Decorator* before each method and set up the environment for being able to trace not only production launched jobs, but also development tests.

Here's an example of how the tracer was implemented in the quality score calculation:

```

1 from opentelemetry import trace
2
3 tracer = trace.get_tracer(__name__)
4
5 @tracer.start_as_current_span("compute_quality_score")
6 def compute_quality_score(
7     stock: "Stock",
8     index_quality_ratios: Dict[str, IndexRatio],
9     calculation_date: date | None,
10 ) -> tuple[dict, dict]:
  
```

```

11     ...
12     return final_score

```

Listing 3.2: Telemetry Tracing Implementation

We then looked at the *DataDog* dashboard and clearly saw an excess of queries to the DB when calculating each score, and tried to reduce them by looking for a simple fault in the logic of the queries. Similar issues were resolved before by adding a *select_related* in the queries. To clarify what this does, here is the definition with a simple use-case:

“Returns a QuerySet that will follow foreign-key relationships, selecting additional related-object data when it executes its query. This is a performance booster which results in a single more complex query but means later use of foreign-key relationships will not require database queries.”
– Documentation [23].

In Django, *select_related()* and *prefetch_related()* are designed to stop the deluge of database queries that are caused by accessing related objects. *select_related()* “follows” foreign-key relationships, selecting additional related-object data when it executes its query. *prefetch_related()* does a separate lookup for each relationship and does the “joining” in Python. So one uses *select_related* when the object that you’re going to be selecting is a single object, so *OneToOneField* or a *ForeignKey*.

After numerous attempts to improve performance, we were forced to reconsider our approach and began analyzing the problem from a different perspective. It quickly became evident that we were operating on a small virtual machine (Table 3.1) shared with the regular platform workload. Tweenvest’s servers were not designed for high computational demand; rather, they were optimized for routine operations, where standard score calculations typically take between 3 to 4 hours. This setup works well under normal conditions, as calculations are performed after market hours and do not impact the user experience. However, it proved inadequate for our needs—specifically, processing data over five or more years within a reasonable timeframe, even when computing just one score per month.

Table 3.1: Tweenvest Server Specifications

VCPUS	RAM
2	4 GB

3.1.4 Dedicated Server Deployment

We made the decision to establish a dedicated server for this thesis, equipped with enhanced computational power to facilitate the calculations and data fitting. After evaluating several providers, we opted for Hetzner servers due to their exceptional quality-price ratio. The final setup consists of a dedicated server with the following specifications:

Table 3.2: Hetzner Server Specifications

VCPUS	RAM	SSD Storage	Extra Volume
16	32 GB	320 GB	480 GB

Deploying the project in a scalable and reproducible environment involved several structured

steps, encompassing server provisioning, secure remote access configuration using SSH, and initializing the containerized application deployment via Docker, with the application codebase managed through GitHub.

After all of the setup was done, we saw an amazing **x100 performance improvement**, going from 15-minute calculation times for index statistics to 8 seconds. Finally, we could launch the factor scores historical calculations, so we created two subsets of 1,500 random companies each with the factor scores calculated for the 1st day of the month for the following periods: 2015-2020 and 2020-2025, which took a total of 3 hours.

Now that we had filled the DB with the necessary factor scores for the study, we ran a big calculation to get all of the possible factor scores for the +100,000 companies, which took several days.

3.2 Dataset Creation

3.2.1 Aggregating the data

For being able to continue, we needed to create the final dataset with all the necessary data for later on doing the analysis and developing the predictive models, so we created a simple algorithm with this logic:

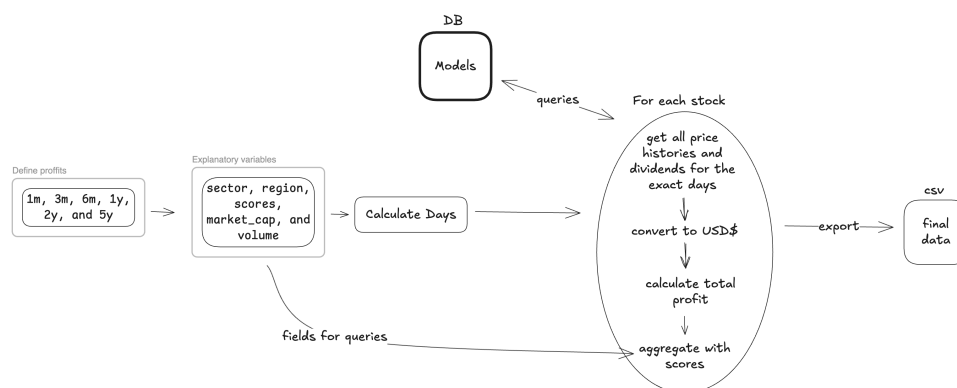


Figure 3.6: First Data Aggregation Schema

Once we had the final csv file we noticed that we had too many empty values for profits, the problem was that some of the days calculated were empty for a specific company, so there was no price history for them. Since companies belong to multiple countries, different holidays could be causing another major loss in the information. To fix this in a general way **we implemented an internal method** for estimating price histories when they did not exist for the wanted date. The complete implementation of this data export and price estimation process can be found in appendix 6.3.

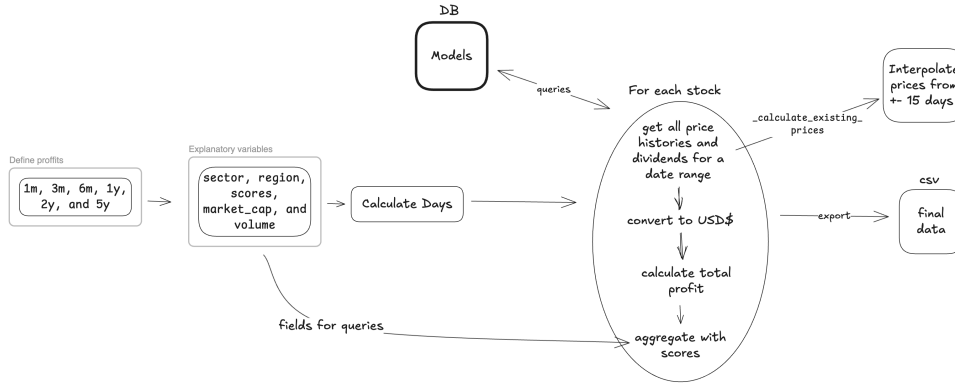


Figure 3.7: Final Data Aggregation Schema

By looking at the figure 3.7 we can see that we are interpolating those price histories. Since we are only looking for long term profitabilities, we can **approximate the price** on a day x by interpolating the price on $x - y$ and $x + z$, weighted by how close they are to the original day x , with a maximum distance from x of 15 days. This significantly reduced the amount of missing values.

3.2.2 Datasets Construction

For this study, we have attacked the objectives from multiple perspectives. First we are making a descriptive analysis to understand the data, and then we are proposing different models to see if there are any relationships between the scores factors and long term profitabilities. That's why when creating the datasets, we made two different approaches to have more information for the analysis:

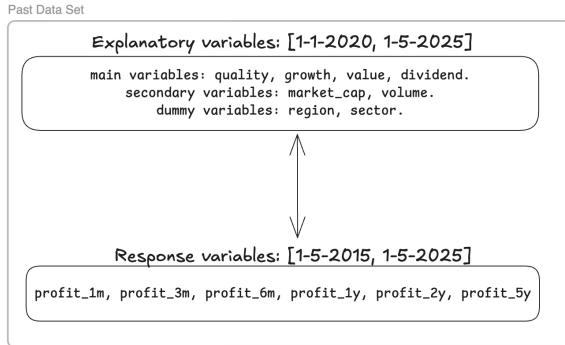


Figure 3.8: Past Dataset

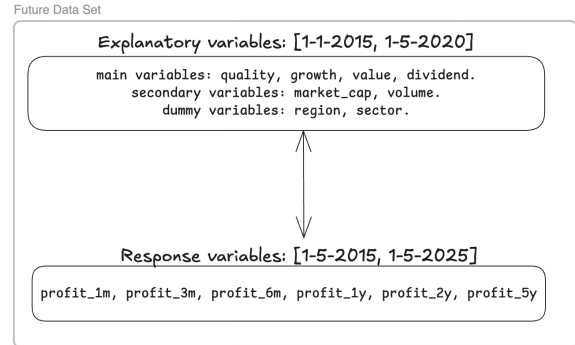


Figure 3.9: Future Dataset

For the past dataset, we calculated the scores for the range 2020-2025 and then related them to the profits obtained until the specified date. For example, the date 1-1-2021 would be linked to the 2 years' profitability obtained until that date:

$$Profit_{1-1-2021}(\%) = \frac{P_{1-1-2021} - P_{1-1-2019} + D_{acc}}{P_{1-1-2019}} \quad (3.1)$$

Which interprets to **how earlier profits have influenced on the explanatory variables**. This will give us valuable information in the descriptive analysis.

For the model fitting, we want to know if the current scores are good predictors of future

profits. So we calculated the scores for the range 2015-2020 and related them to the profits obtained from that date to the future. For example, the date 1-1-2021 would be linked to the 2 years' profitability that will be obtained in the future:

$$Profit_{1-1-2021}(\%) = \frac{P_{1-1-2023} - P_{1-1-2021} + D_{acc}}{P_{1-1-2021}} \quad (3.2)$$

As this approach is what we actually want to use for the predictive models, we will be using the future dataset structure for most of the analysis.

3.2.3 Preprocessing

Even with all of the protocols created to have the most reliable and filled datasets, we still had to follow the methodologies explained in the previous chapter to analyze the data and see if there were any problems not solved.

Cleaning the Data

To begin, when looking at the data structure we saw that **about 50% of growth scores were empty**. This turned on many alerts from problems with the scores calculation algorithm, because compared to the rest of the variables there was a significant difference, since the **normal absence was only around 8%**.

Table 3.3: (%) Missing values per column for Small Data Future.

quality	growth	value	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
13.4	54.0	7.4	2.6	3.1	3.2	3.2	3.3	6.0

But after digging into the data we saw that it was due to the fact that many companies stopped sending reports or selling but kept existing, so we decided to delete them from our dataset because they were not behaving as a "normal company", and this could alterate the models results. Additionally, as noted earlier, we replaced all empty values in the dividend scores with 0, since Tweenvest's algorithm omits a score when a company does not pay dividends. That's why we do not see any missing values in the dividend scores.

3.2.4 Outlier Detection

In this section we will present the performance of the different outlier detection methods. But the main analysis of the final datasets will be done in the results chapter.

Inter Quartile Range

This method is useful for big datasets that need to be filtered quickly and where the meaning of the outliers is not multivariate. But as it will be explained in the results chapter, we saw correlations between the different profits, so we needed to use multivariate outlier detection methods, otherwise we could be creating a bias in the data for the model fitting.

Isolation Forest

We started of using this technique due to its speed and accuracy for non-linear multivariate clusters.

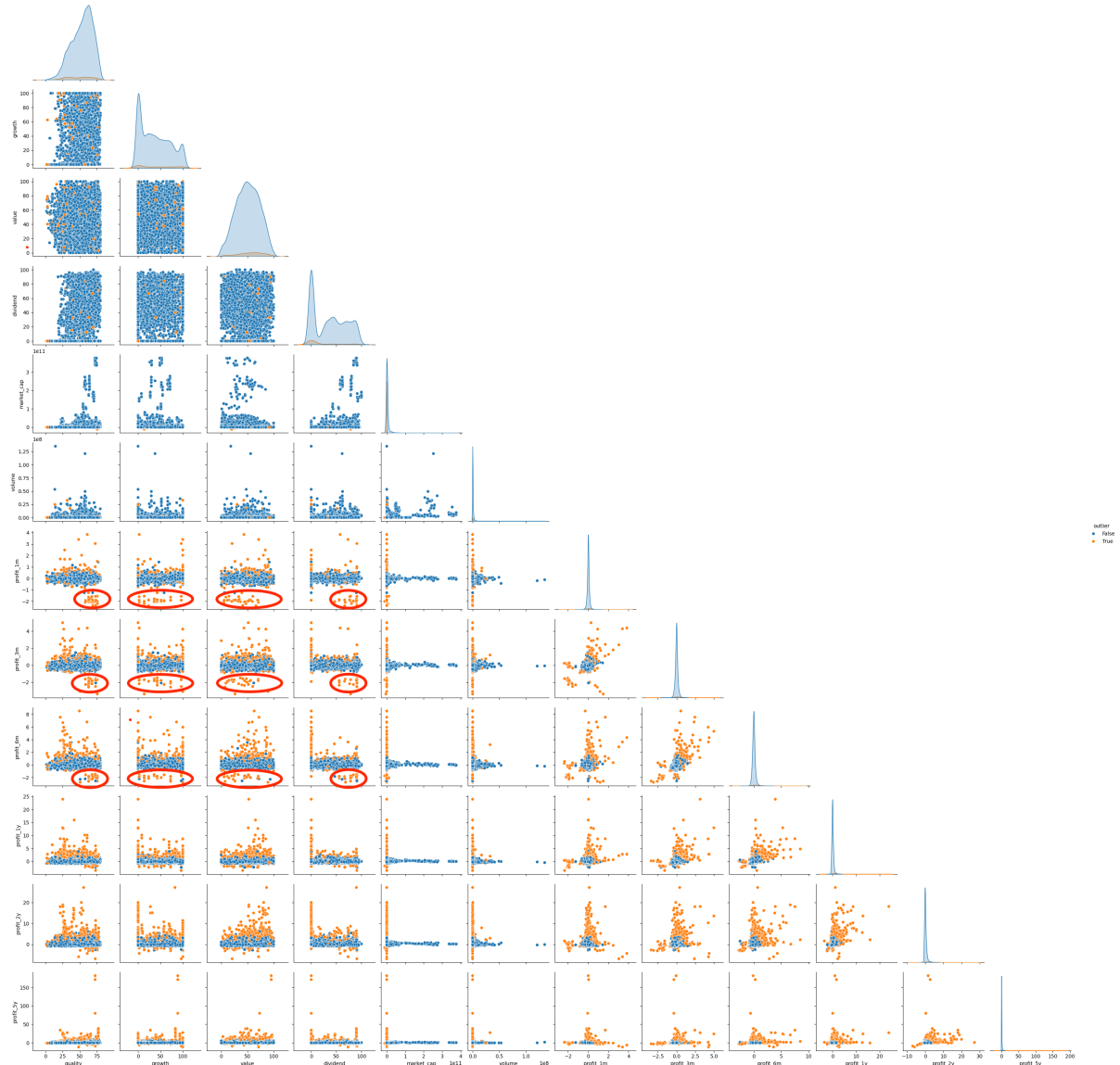


Figure 3.10: Small Data Future - IF (4%) Pair Plot

This Pair Plot shows how the IF method is detecting outliers in the profits data. In this case it is detecting as outliers much of the data that seems to form a structure in the profits. Also, it is deleting some interesting clusters when comparing the profits with the scores, and here we might lose some important information– this is highlighted with the red circles.

Single Vector Machine

When using the SVM we saw a similar detection compared to the IF method, which is very interesting since their approach is completely different.



Figure 3.11: Small Data Future - SVM (4%) Pair Plot

But in this case we also see a slight improvement with recuperating the structures in some of the profits— this is circled with the green color.

Local Outlier Factor

Surprisingly, with the LOF method we see a good improvement for both the profits and the scores clusters– which is highlighted with the green circles.

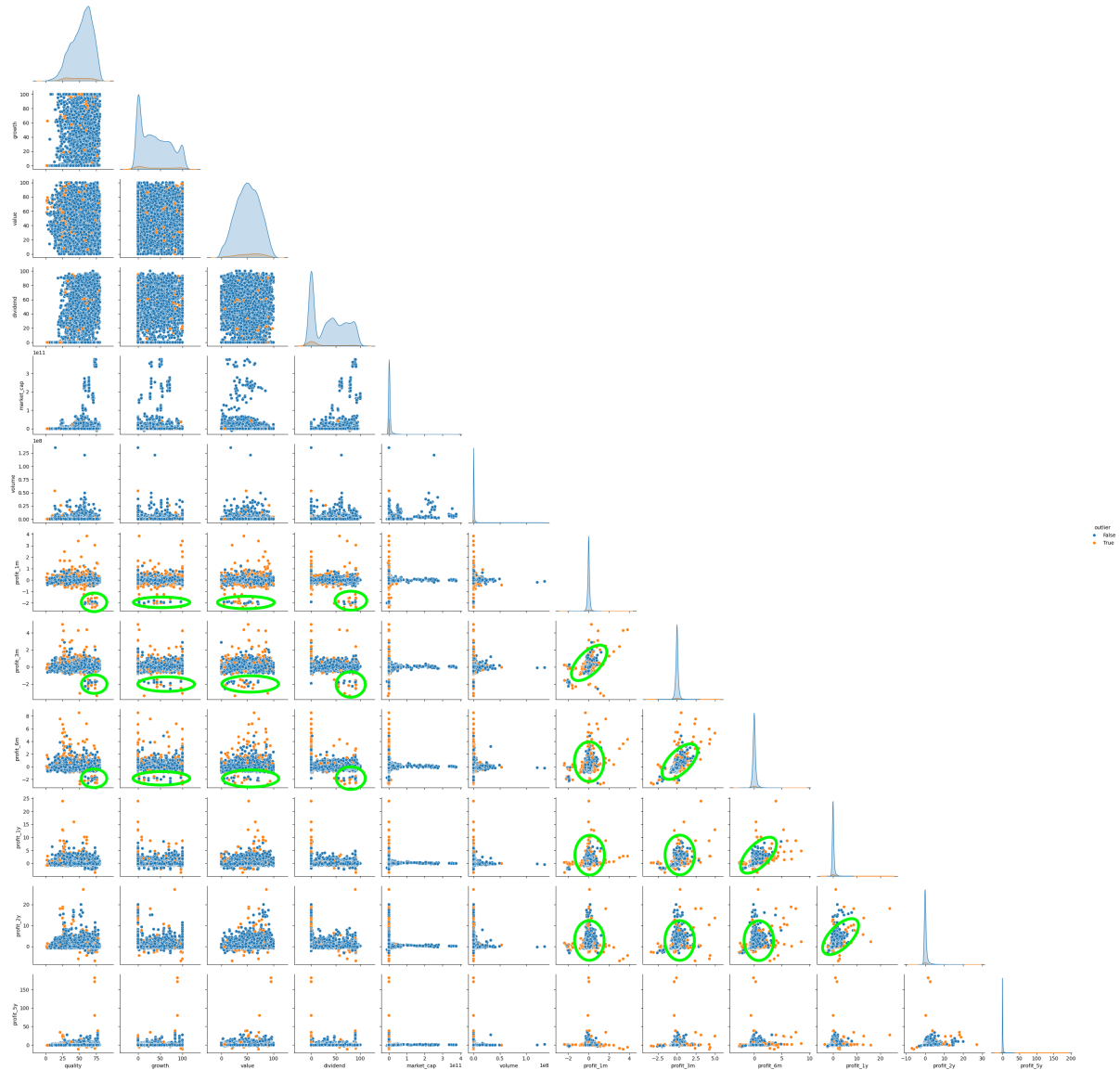


Figure 3.12: Small Data Future - LOF (4%) Pair Plot

We knew we were going in the right direction, but with this method we were missing some outliers that were detected before in the score-profits clusters.

Multi-Criteria Outlier Detection

That's when we realized that if each method was good at detecting some outliers but bad at others, so it made a lot of sense to use them together. Following this approach and after multiple tests, we found that the best results came when increasing the overall tolerance and then selecting the intersection of outliers between them, so with a 20% tolerance we only excluded around 8% of the data.

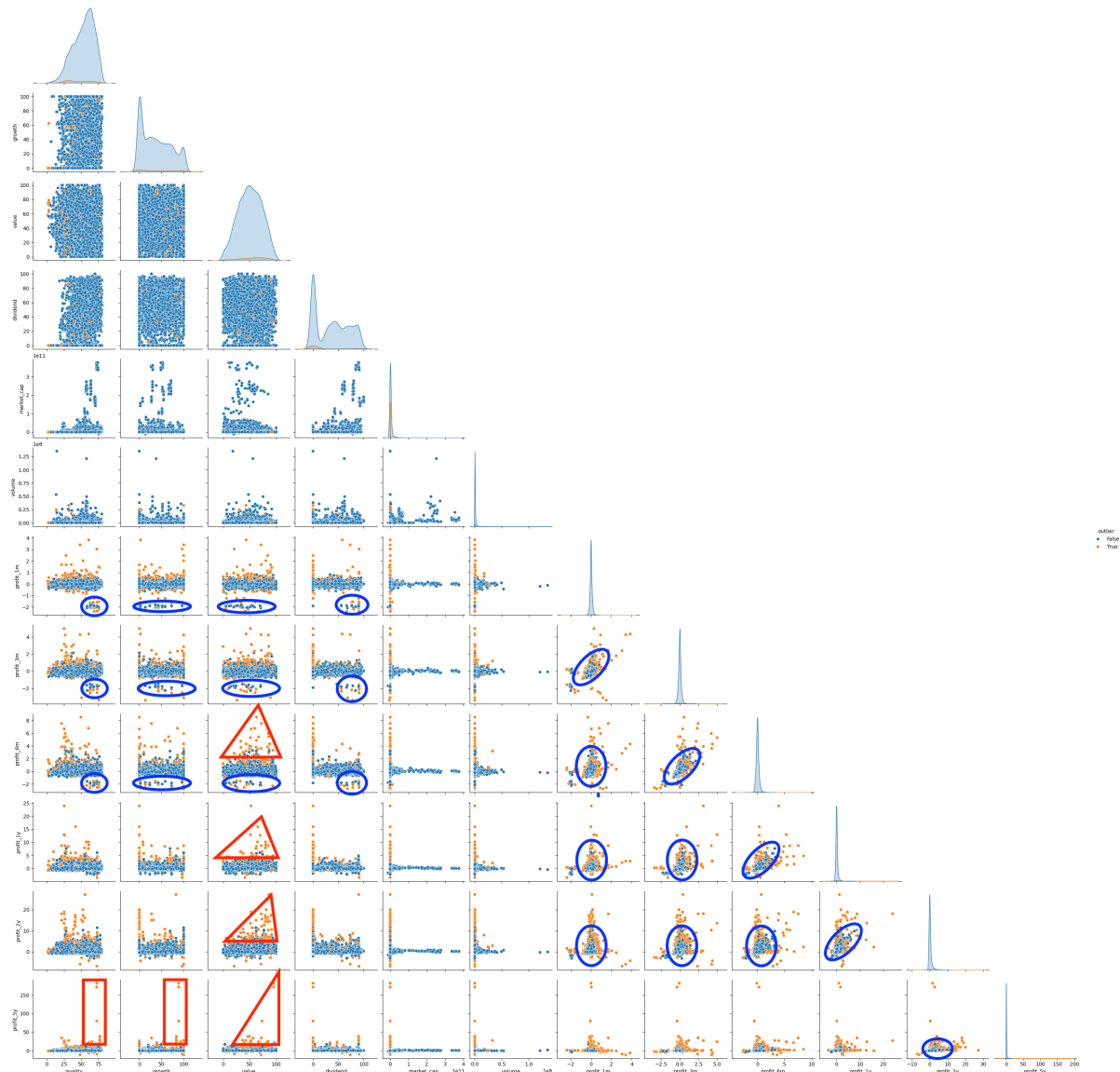


Figure 3.13: Small Data Future - MCOD (8%) Pair Plot

As shown in the Pair Plot, this method is able to maintain most of the structures of the profits and also keep the small clusters for the scores, which is highlighted with the blue circles. But of course, it is not perfect and could be missing important information for extreme cases –highlighted with the red shapes.

Islands of Outliers

After the previous results we started working with the models, but we quickly noticed that those clusters that were not being deleted, on purpose, were affecting the results. So we decided to take a big look at the data again to see what was happening:

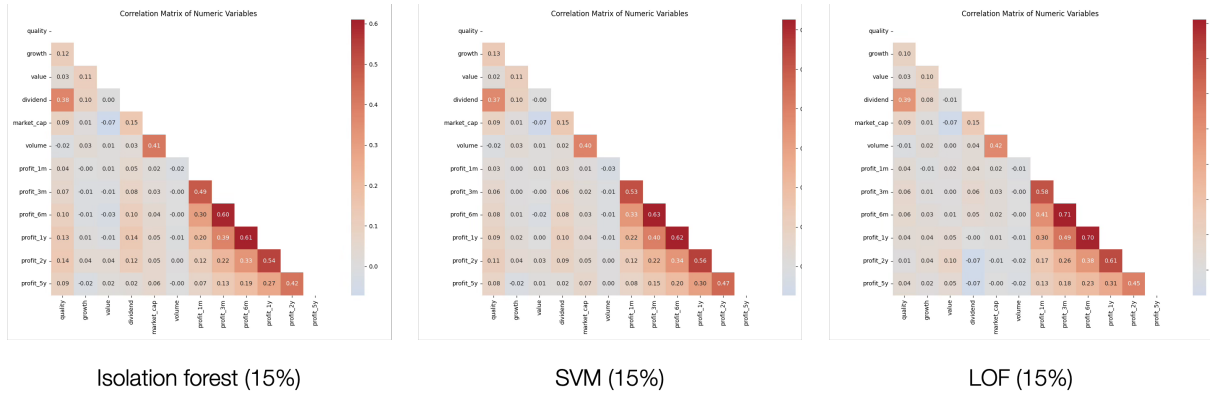


Figure 3.14: Correlations between variables for different outlier detection methods

As we can see, the IF and SVM methods are showing some correlations between the profits and the scores, but the LOF is not showing any correlations, even negative ones. And that is what made us think about the clusters of anomalies that were not being deleted by the LOF method, so we built the ICD, mentioned in the previous section.

3.2.5 Summary

To summarize how we have treated the data, we ended up with different datasets groups, each with 17 columns for the explained variables, the date and also the company's name and stock ticker (see appendix figure 7.1).

Simple Datasets

For the first part of the analysis that consisted of the descriptive analysis, we started by using the Big Datasets, that contain all of the Tweenvest's data for the past 10 years. This way we could see the real distributions and correlations, as we will explain in the next chapter. Also we created the 2 subsets of Small Data with only 1,500 stocks for quicker calculations, making sure that the sectors were balanced, and using a **4% removal rate** for the MCOD method.

Table 3.4: Simple Datasets Total Rows

Dataset	Raw	Cleaned	No Outliers
Big Data Past[†]	16,119,852	9,991,895	9,492,300
Big Data Future[†]	31,933,137	14,518,624	13,792,692
Small Data Past	83,243	48,338	46,415
Small Data Future	65,166	26,214	25,220

[†] In the *Big Data Past* and *Big Data Future* datasets, we removed the outliers using the IF method for saving computational time.

Geographical Datasets

For the second part of the analysis, that consists of exploring the regression models, while advancing with it we had to inspect how the geographical regions were affecting the results, so we created one dataset per region applying the same sector balancing. Here we kept using the **4% removal rate** with the MCODE.

Table 3.5: Geographical Datasets Total Rows

Dataset	Raw	Cleaned	No Outliers
Small Data Future USA	60,503	21,240	20,463
Small Data Future EU	62,725	25,217	24,245
Small Data Future AS	65,772	28,085	27,156
Small Data Future LAT	48,412	20,610	19,851
Small Data Future AF	58,341	21,182	20,335

Validation Datasets

Once we had seen the nature of the data and the simple models results, we realized how much the outliers and countries were affecting the results. So for improving the consistency we used the IF method with higher tolerances and additionally, we balanced the data by regions and sectors for the Big Dataset. However, these steps led to a substantial reduction in the number of rows.

Table 3.6: Validation Datasets Summary

Dataset	Cleaned	IF	Balanced
Small Data Future	26,214	22,282	–
Big Data Future	14,518,624	9,437,105	110,203

These restrictions resulted on a surprisingly higher simple correlations, better performances and a more consistent model fitting as we will see in the results chapter.

Conclusion

Finally, we have to clarify that outliers are anomalies in the *normal* behavior of the data, so even though we are excluding them for obtaining a general set of models, we could be missing valuable information such as the possible structures shown with the red structures in the figure 3.13. That is why we propose for future work a separate analysis for the outliers, to see if really good profits (or bad ones) have any characteristics, such as average high or low scores.

Chapter 4

Results

This chapter presents the outcomes of the analysis and modeling procedures applied to the curated datasets. The results are organized into two main parts. First, a descriptive analysis explores the underlying characteristics of the data, examining distributions, correlations, and clustering patterns to identify preliminary relationships between the scores and company performance. Second, the predictive modeling results evaluate the extent to which the scores –and other explanatory variables– can be used to anticipate future profitability over different investment horizons.

4.1 Descriptive Analysis

Before jumping into the model fitting, it is necessary to understand the data that we are working with. For doing so, we will be looking at the variables distributions, correlations and possible structures. The analysis begins with the complete dataset to capture the overall patterns, followed by an evaluation of smaller subsets to verify whether they exhibit similar behavior. Additionally, because a small number of outliers in the *Big Data* obscure the profit distribution plots, we will be using the IQR method to remove them.

4.1.1 Variables Distributions

Scores

When looking at the scores in the figure 4.1, we see that there are several scores that saturate. So for a better visualization we created a temporal dataset where the dividend and growth scores saturating at 0 were deleted, this is shown in the figure 4.2. Starting with the **quality score**, we can see that its distribution shows a **clear negative skewness**, meaning that there are more companies closer to having a good quality score. This could be due to the fact that we had deleted companies that did not behave as a “normal company”, so we will later analyze the distributions of the profits to see if this was the case. Also, there is a **secondary peak around the 30-40 scores**, which could be related to the time-frame of the data.

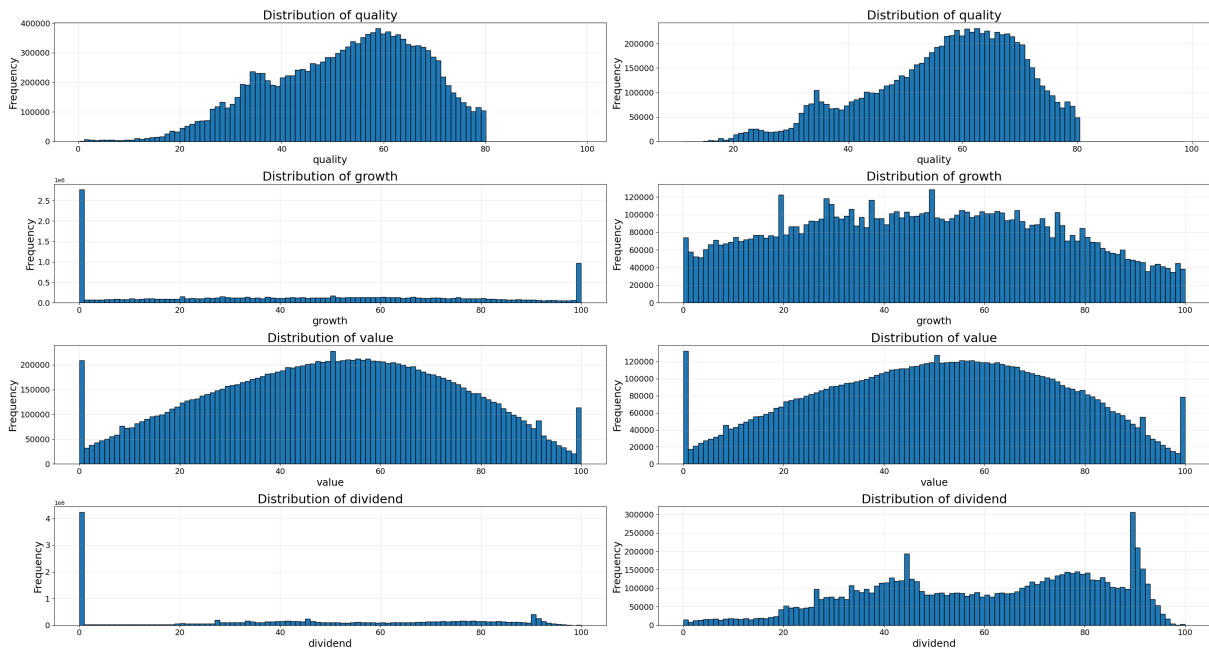


Figure 4.1: Big Data Future (IQR) - Scores

Figure 4.2: Big Data Future (IQR) - Scores Deep

Comparing the interest rates between 2015–2020 and 2020–2025, we can see that the first period was at a historical low, so there could be more companies with higher debt rates than normal. Some central banks, particularly in Europe, experimented with negative interest rate policies on deposit facilities during this period (figure 4.3), [24]. This possibility should be studied in more detail, but knowing that this does not happen for the dataset between 2020–2025 where interest rates are higher, we have a clue that this could be what is causing it (see appendix figure 7.3).

CHART EURIBOR 1 MONTH

Historical Euribor rates

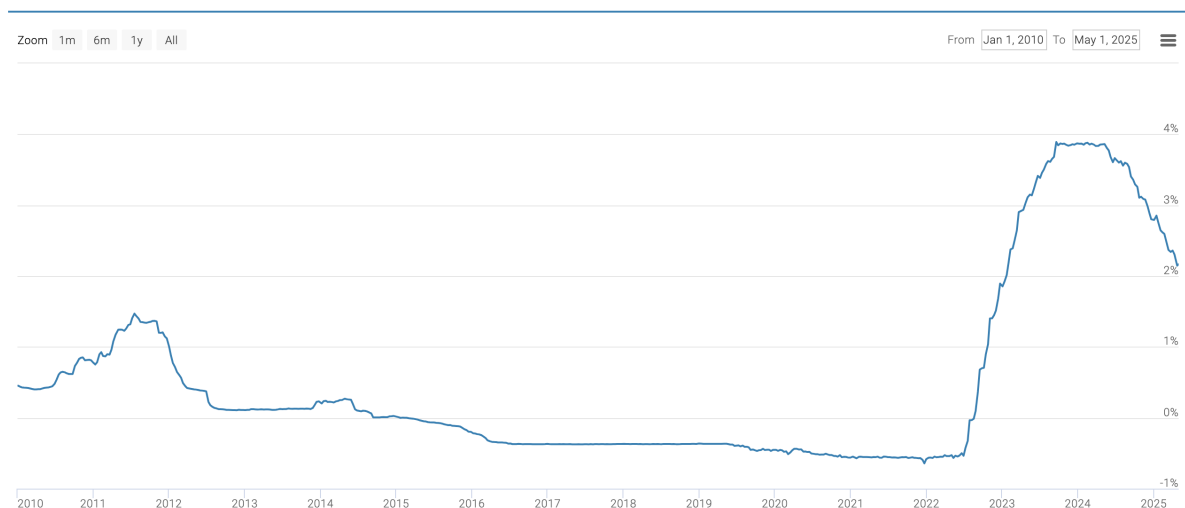


Figure 4.3: EMMI - Euribor Interest Rates

Finishing with quality, there is a “defect” in the dataset. We do not have 10 year averages for the companies that are in the 2015–2020 period –Tweenvest only has data after 2012– so the algorithm gives 0 to some of its criteria, making the maximum quality be 80. But since this is

happening to all of the companies, **it is not a problem for the model fitting**, and we could easily rescale it to 100. And since we see a descent slope after 65, we can rescale it to 100, being in this case approximately **81 the descent point**.

Continuing with **growth score**, we notice that there are many companies saturating at 0 and 100, and the same happens with Dividend. But looking at the deeper plots, we can see that it is closer to a uniform distribution with a **descent slope after 65**, which means that most companies have a similar growth rate, but fewer have higher ones.

When looking at the **value score**, we can see that the distribution almost has an average of 50, and there is a continuum of values trough the whole range decreasing as the score diverges from that average, so it seems that it is a balanced scale for distributing the companies.

Lastly for the **dividend score**, as mentioned before, there are many companies that do not pay dividends. But aside from that, we have an irregular distribution for the score, with a **peak around 85–92**, meaning that there are several companies that care about their dividend policy.

Note: We get very similar distributions for the Small Data datasets, but with lower resolutions of the scores distributions due to the sampling process (see appendix figures 7.4 and 7.5).

Dummies

We have already explored the scores, so now need to examine whether the distribution of the dummy variables could introduce bias into the models.. We started looking at the sectors distributions for the *Big Data Future* to see what is full picture of the market, and then we will look at the *Small Data Future* dataset since it will be the most used one in the analysis.



Figure 4.4: Big Data Future (IQR) - Dummies



Figure 4.5: Small Data Future (MCO) - Dummies

As it was expected, in figure 4.5 we see a **uniform distribution for the sectors** because this condition was forced in the data creation process, but in the full dataset we have a different distribution, which could show some interesting information about how the market behaves. This made us also check the distributions for the past datasets (see appendix figure 7.7) where we detected a noticeable difference. This **temporal variation in sector density** presents an interesting opportunity for future research. When it comes to the regions variables, we have a different result, showing a **clear predomination of the Asiatic region**, which happens for all of the datasets.

This could create issues for the model fitting, since it is known that in countries like China have more volatility and lower returns [25]. Also, their auditors often lack true independence, especially when auditing State-Owned Enterprise (SOE) or firms with strong political ties. The Chinese government tightly controls domestic accounting firms and imposes restrictions on foreign auditors. Even international firms like *PwC-China* operate under local joint ventures, often with limited autonomy, as reviewed in [26]. So we will have to check if the geographical variables are enough to take this effect into account, or if we need a different approach for modeling these effects.

Profits

Finally, when analyzing the profit distributions across different time horizons, we observe a large number of outliers (appendix figures 7.8 and 7.9), which could obscure underlying patterns. So for this reason we are seeing the datasets with the IQR method applied in figures 4.6 and 4.7.

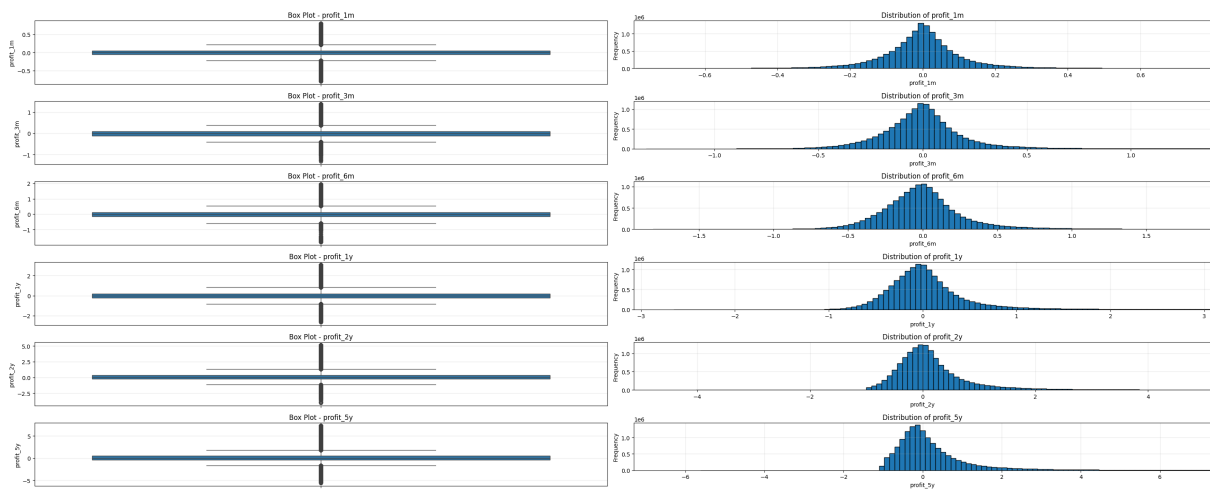


Figure 4.6: Big Data Future IQR - Profits Boxplot

Figure 4.7: Big Data Future IQR - Profits

Interestingly, we notice that the **short-term profit** distributions (1-month to 6-month) are more symmetric and tend to approximate to **normal distributions**, which is consistent with the efficient market hypothesis, suggesting that short-term price movements are largely random. However, as we extend to **longer time horizons**, the **distributions become more positively skewed**, which aligns with the general expectation that most companies do not have good returns, except of the market winners that tend to generate extraordinary positive returns due to the mentioned *economic moats*. This is particularly evident in the longer-term profit distributions (2-year and 5-year), where we see a **longer tail extending toward higher positive returns**. Lastly, in the figure 4.6 we see a tendency to having **less negative returns** compared to the positive ones **the longer the time-frame**. Logically, companies with bad returns do not last long.

4.1.2 Correlations

For analyzing the correlations between the variables, we will start firstly with just the cleaned big datasets, and then we will use also the datasets without the outliers using the multi-variate approach explained in previous chapters, since these are the ones that will be used for the model fitting in most of the analysis. We have to make clear that this is the last section where **we will be using the past datasets**, to see if past profits are being reflected in current scores.

Clean Datasets

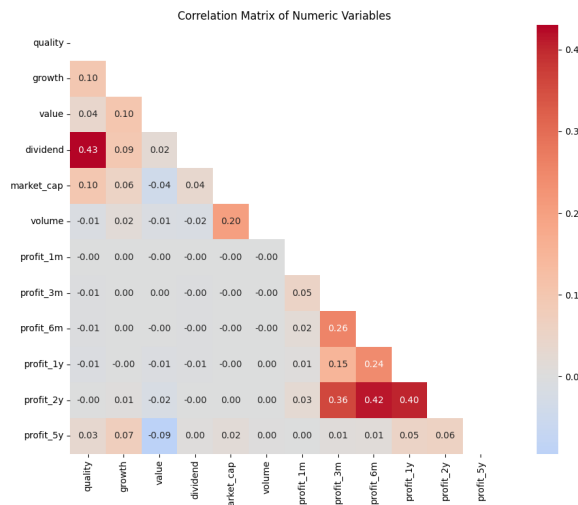


Figure 4.8: Big Data Past (Clean) - Correlations

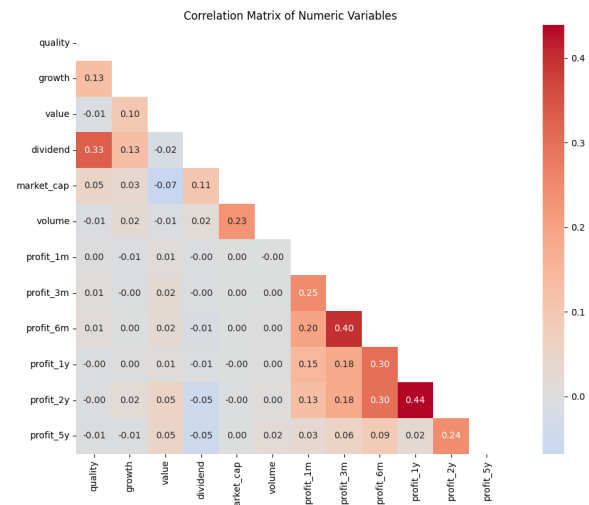


Figure 4.9: Big Data Future (Clean) - Correlations

We can appreciate three main expected correlations:

- **Companies with good quality** –that is so, they have consistent growth, good liquidity and healthy debt– **are more likely to give dividends.**
- There is a strong relationship between the market cap of a stock and the volume of the shares traded.
- **Companies have inertia**, or how it is mostly called *Momentum*, meaning that the profit of a company is positively correlated with the profit of the previous period.

Outliers Datasets

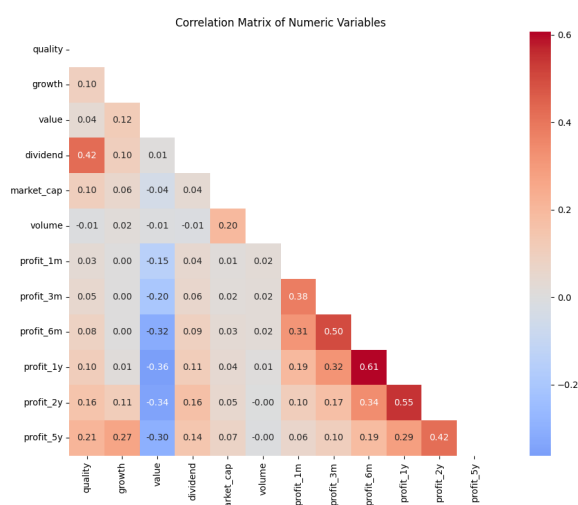


Figure 4.10: Big Data Past IF (4%) - Correlations

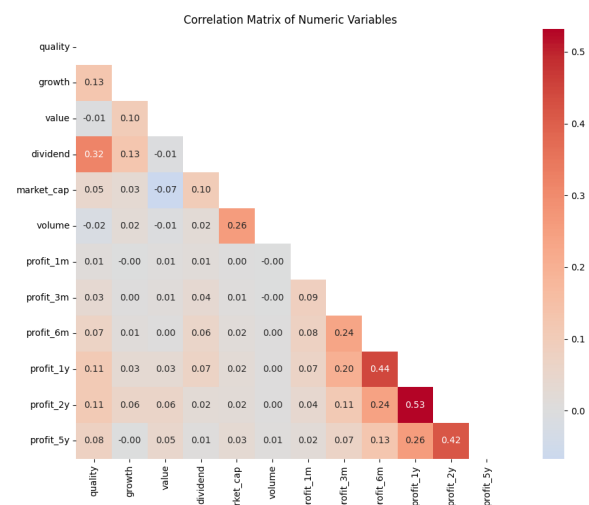


Figure 4.11: Big Data Future IF (4%) - Correlations

For getting a bigger picture we only used a small outlier removal. And as expected, the correlations already seen before any removal stay similar with a slight increase in the *Momentum*,

showing more importance of it than other variables. But actually, there are **new correlations seen** for both datasets:

- **Accumulated profits affect the scores.** In the past dataset we can see that long term past profits translate into higher growth, quality and dividend scores, these were expected to increase together due to their correlation. But also, we see that **accumulated stock performance makes the value score decrease**. This could be explained by two effects:
 - **Bull markets.** As explained in the second chapter, a company can experience returns in the stock market that cannot be followed by financial growth. This makes their value multiples to raise, making their value score to decrease.
 - **Unknown companies.** There is a possibility that the company is not well known, so the market price is lower than it could while their finances are good. This hypothesis could be supported if Value score is correlated to future market profits, which we will see soon.
- **Long term stock returns affect the company's growth.** This is a very interesting result, since it explains the reason why companies try to go to the market to raise capital and help the company grow.
- **Quality is correlated to future profits.** This could be explained by the fact that a company with good quality is more likely to have good future profits, but also by the fact that the market is efficient and the price of a company will reflect all of the information available.
- There is a **mild correlation between most scores**, which could be explain by the fact that a company that does well in one aspect, is more likely to do well in the others. This correlation will not affect the models because they are separated by score.

Although we have to make clear that these correlations do not stand along all of the geographical datasets (see appendix figures from 7.10 to 7.19), so we cannot be sure if these correlations will be properly explained by the linear models.

Validation Dataset

Also, as we mentioned in the previous chapter, in figure 4.12 we see a huge increase in the previous correlations when we removed more outliers. And a new inverse correlation for *Market Cap* and *Value*. It is interesting that this relationship is not as strong as we could expect.

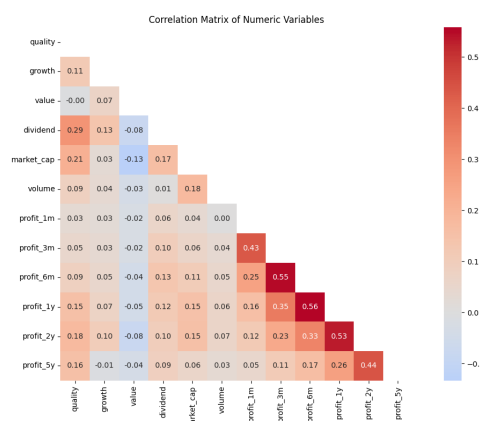


Figure 4.12: Big Data Future IF (35%) - Correlations

4.1.3 Data clusters

Before ending the descriptive analysis, we are going to try to compare these results with the Pair Plot of the datasets, to look for clusters of data when comparing many variables at once.

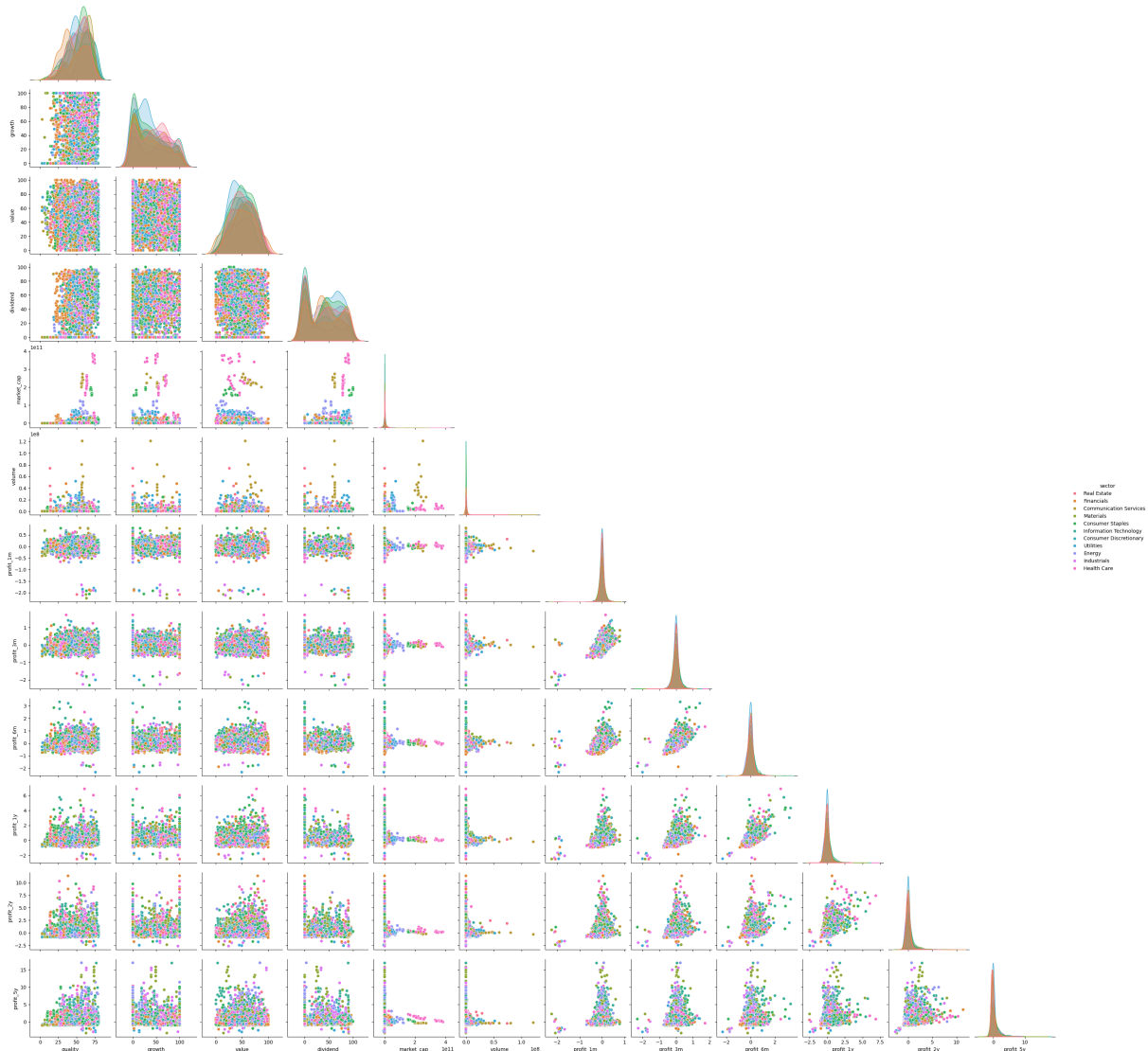


Figure 4.13: Small Data Future - MCOB (4%) Pair Plot

Here it is harder to eye-see the data structure that could form clusters since there are many data-points, but there are some exceptions:

- **Profits.** We can see linear correlations between some of the nearby profits, but for longer horizons the relationships are no longer linear. Which is expected since the further the horizon, more variables take part in the calculation of the profit.
- **Market Cap.** Companies with bigger market caps tend to have higher quality and dividend scores. Bigger companies should have a more stable financial situation, otherwise they would not be as big in the long term. Also there are some sectors that tend to have higher market caps than others, like *Health Care* or *Consumer Staples*.
- **Sector.** It seems that there could be a pattern seen in the sectors distribution for each scores, but it is not clear.

Although, these results are not as clear and they also variate depending on the dataset. Showing early results of non-linear relationships with the dummy variables.

4.2 Linear Regression

Having understood the data and the different datasets created, we now proceed to explain the process of refining the methodologies and the results obtained.

4.2.1 General Models

First of all, we started with a family of simple linear regression models with just the plain explanatory variables and not excluding the outliers. This, of course, resulted in a very poor model that could not predict anything (see appendix figure 7.20). However, after applying the first outlier removal for the initial series of datasets, we started to see some results:

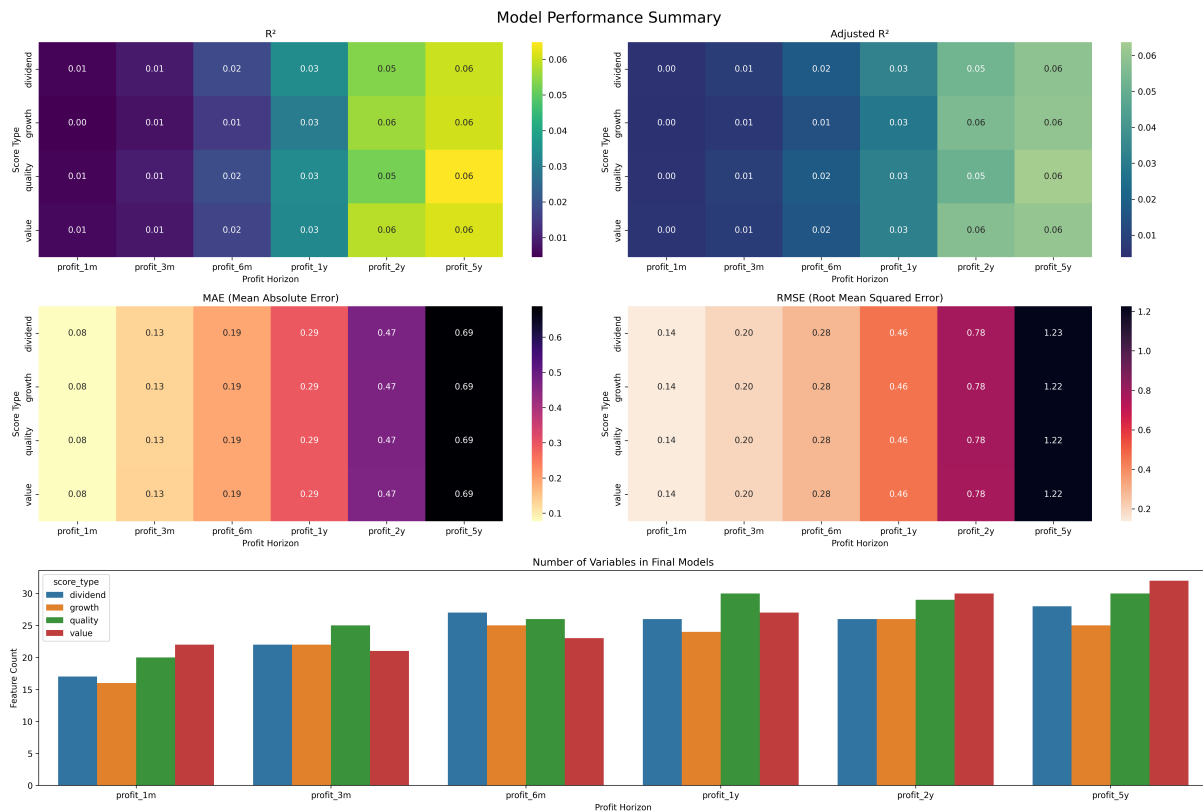


Figure 4.14: Small Data Future - MCOB (4%)

In figure 4.14 we observe a consistent pattern across all models: the further into the future we try to predict profits, the better the model performs in terms of capturing trends, but the prediction error also increases. This is expected, as **long-term forecasts naturally involve greater variance**; however, stable companies tend to exhibit more predictable behavior over time. Additionally, we notice that the RMSE is nearly twice as large as the MAE, which suggests the **presence of numerous outliers in the data**. Also it is worth mentioning that the number of variables increase at the same time as the models' performance improves, which could be a **sign of overfitting**.

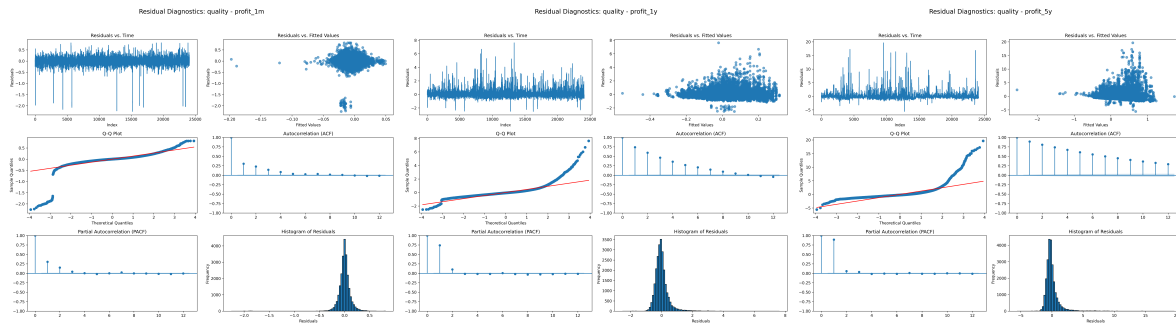


Figure 4.15: Small Data Future - MCOB 4% - Residuals

Examining the residuals, they all show a similar pattern for all of the scores, so from now on we will show just the quality residuals –as seen in figure 4.15– for 1m, 1y and 5y profits. Often, we detect the same issues:

- **Normality.** The Q-Q and the distribution plot reveal that the residuals are not normally distributed, although there is a section where normality appears to hold.
- **Heteroskedasticity.** The residuals vs fitted values plot indicate that the variance is not constant across all time horizons.
- **Autocorrelation.** The ACF and PACF plots also show that the residuals are not independent of each other, which reinforces the momentum correlation seen.
- **Outliers.** When looking at the residuals against time, we see big spikes that show that there are many outliers in the data.

Finally we did not even see any consistency with the scores since its effect (sign) kept changing and the coefficients themselves did not have statistical significance.

Table 4.1: Small Data Future - MCOB (4%) Consistency

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	-0.0013	–	-0.0004	-0.0026	–	0.0003
	Sign	–	0	–	–	0	+
	t-value	-6.38	–	-3.22	-6.92	–	6.44
Growth	Coef.	0.0018	–	0.0007	0.0047	0.0001	–
	Sign	+	0	+	+	+	0
	t-value	11.27	–	5.75	13.86	1.62	–
Quality	Coef.	-0.0038	-0.0059	-0.0010	-0.0076	-0.0003	0.0002
	Sign	–	–	–	–	–	+
	t-value	-6.93	-4.54	-4.29	-8.61	-1.87	2.47
Value	Coef.	0.0015	-0.0054	–	–	–	-0.0002
	Sign	+	–	0	0	0	–
	t-value	4.64	-4.36	–	–	–	-1.81

The problem with ARIMA

Before jumping into trying to improve the results by refining the data or adding new variables, the first main issue we tried to fix was the autocorrelations seen in the residues. So we used an automatic ARIMA fitter from *PyPI* Smith et al. [27] to see if it improved the predictability and also fixed the issue all at once.

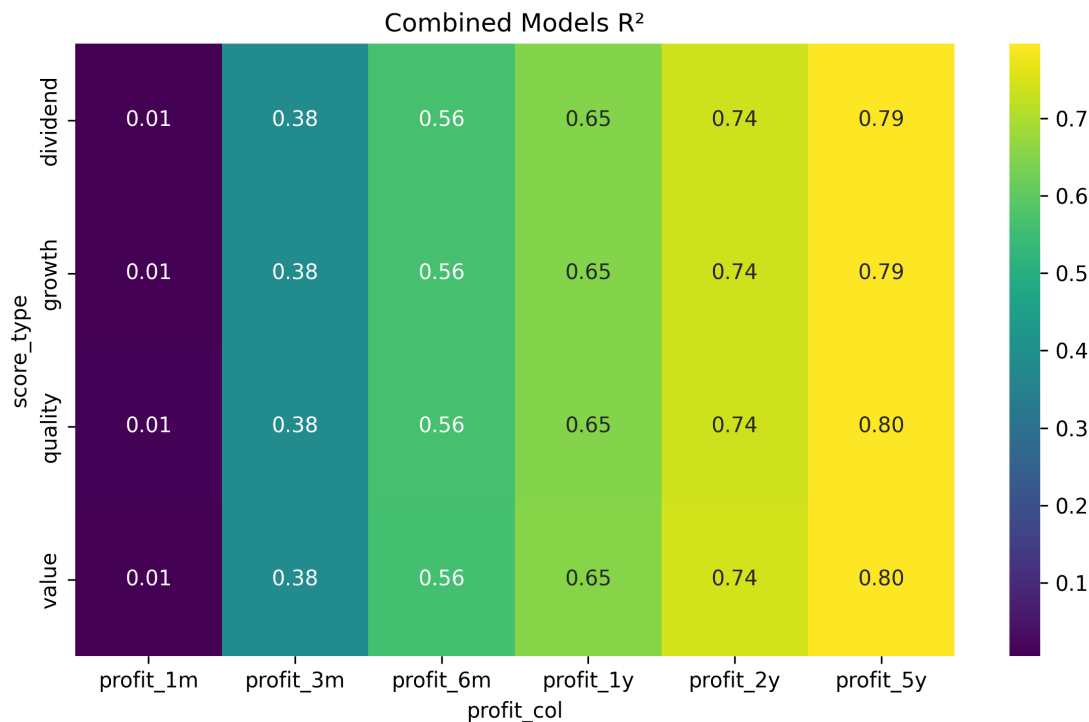


Figure 4.16: Small Data Future - MCOD (4%)

And of course, this worked amazingly: it solved completely the autocorrelations of the residues and it improved a lot the results of the model (figure 4.16). But this came with a major issue that made us abandon the idea of using ARIMA models; **we had to use future profits as inputs for the ARIMA model**, which only could make sense for an ad-hoc model as a way to explain past behavior, but not for a model that is supposed to predict future behavior.

Artificial Normality

So before continuing with the autocorrelations issue, we tried to solve the normality and heteroskedasticity issues. And since it is well known that regression models are not very robust to non-normal data, fixing that should also improve the models' performance. We used the mentioned *Yeo-Johnson* transformation and got some **improvements with the predictability and the RMSE for long term models (5y)** but showed complete noise for the short term (1m to 6m), which suggest that this transformation helped coping with the tails of the 5y-profits distribution and also helped showing the real noise from the short term models. In general, we had a **worsening of the accuracy** in most of the models, except in the long term.

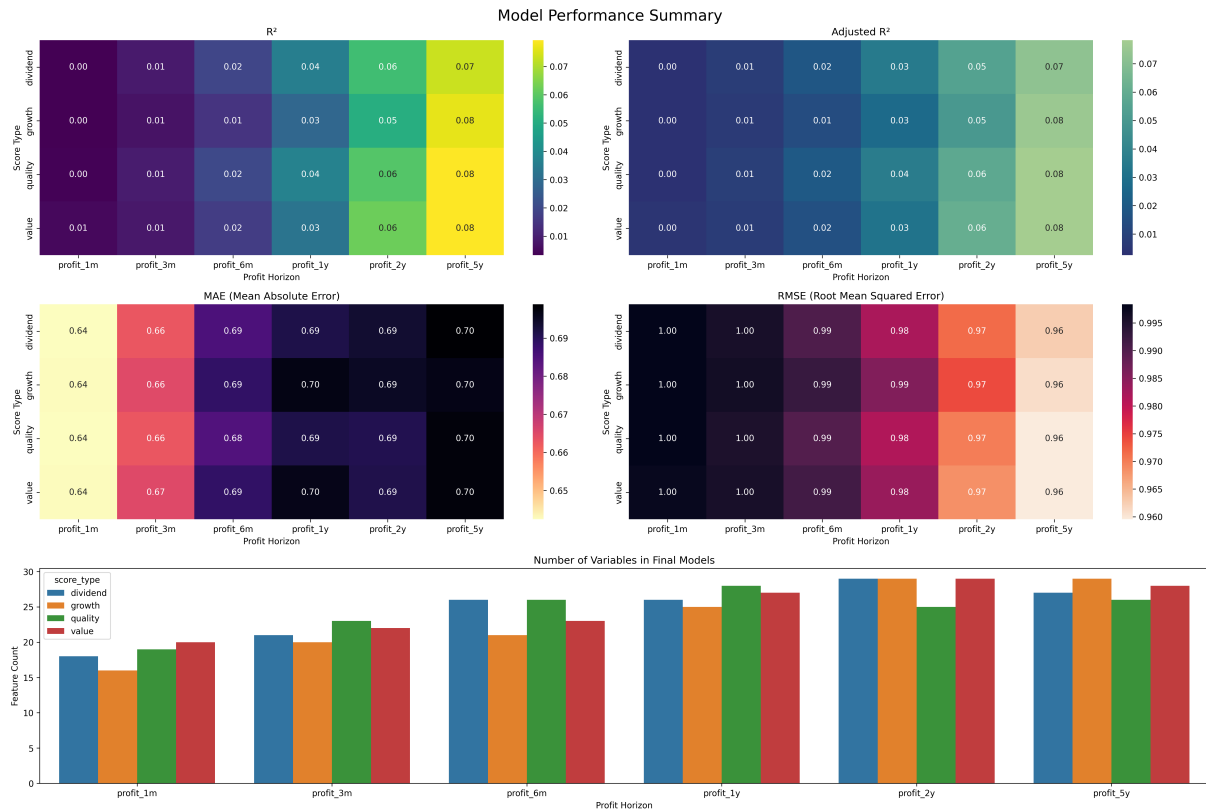


Figure 4.17: Small Data Future Yeo-Johnson - MCOD (4%)

When looking at the residuals in figure 4.18, we can see that the **normality had improved a lot** for the mid and long term models –which aligns with the performance changes–, but we still have issues in the tails of the Q-Q plot and the same peaks in the time series plots due to the outliers.

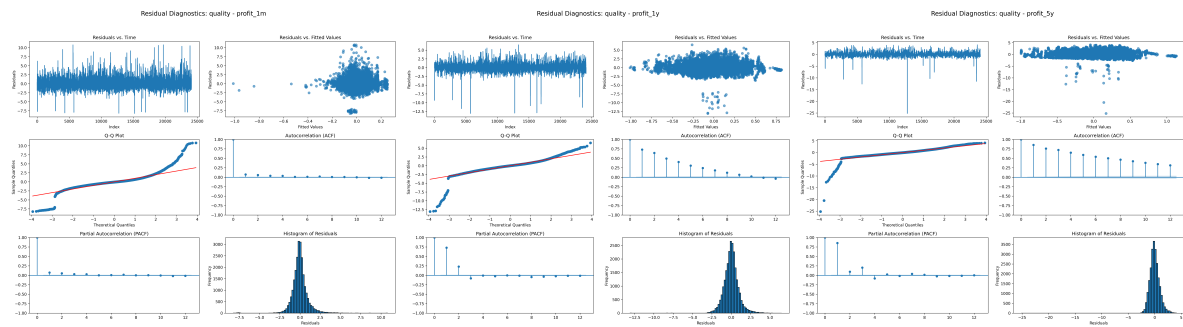


Figure 4.18: Small Data Future Yeo-Johnson - MCOD (4%)

Also, after this transformation we still did not get any consistency and statistical significance (see appendix Table 7.1). But since it improved both the predictability and the normality we applied this transformation for the rest of the models.

Windowed Models

Before examining if the regions' non-linear effects were responsible for some of the issues, we sought to improve the autocorrelations through a different approach. Specifically, we employed a new **family of windowed models** that rely on different statistics to assess whether the market

exhibits memory of past financial performance.

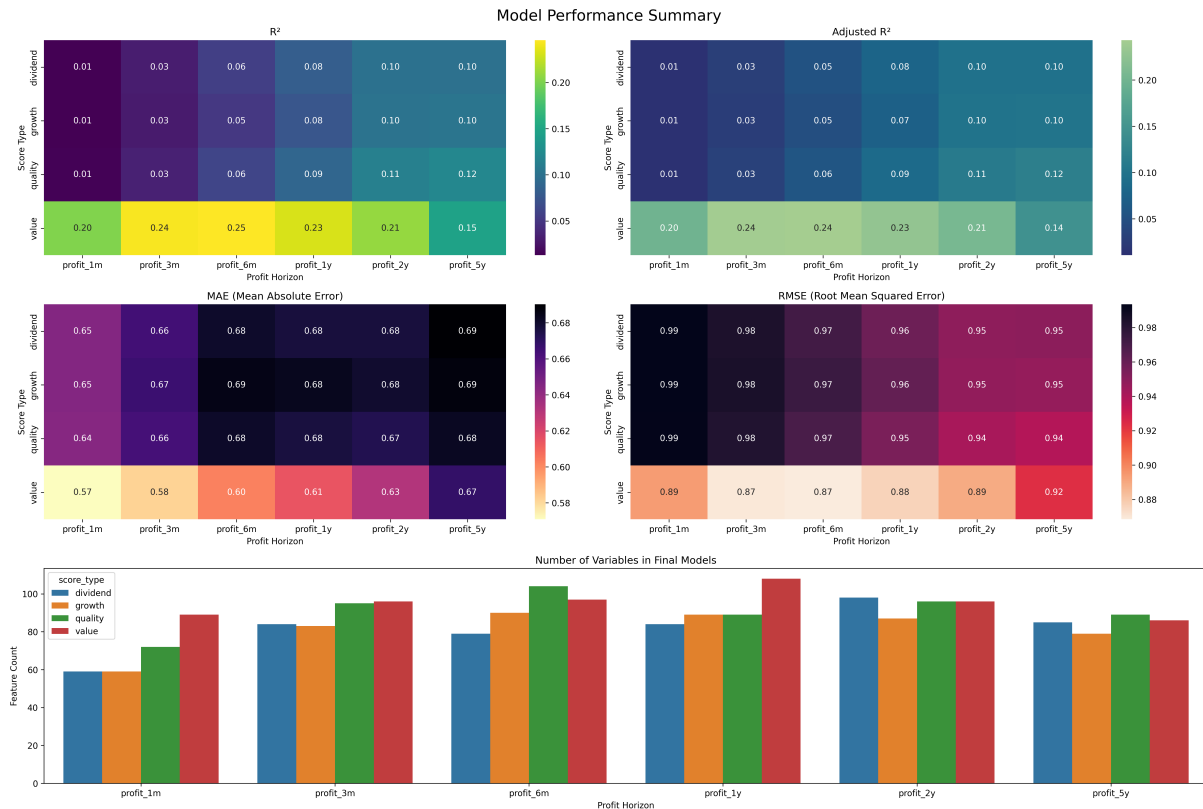


Figure 4.19: Small Data Future (Windowed) - MCODE (4%)

In figure 4.19 it appears to be a **general improvement into all of the models' predictability and accuracy**, which could be caused by overfitting but it appears not to be the case since the amount of significant variables does not increase with the performance. Surprisingly, we encounter a new pattern that will be happening in all of the models that uses these windows: **increase in the Value's models performance, with a peak around 6m-profits**. To understand this, we have to go back to the definition of the Value Score seen in chapter 2, where it was clear that the score contained the price movements within it. So we have figured out a way to implement some of the price's momentum into the models including also financial data into it. This showed us that **the market has memory of a company's price multiples**.

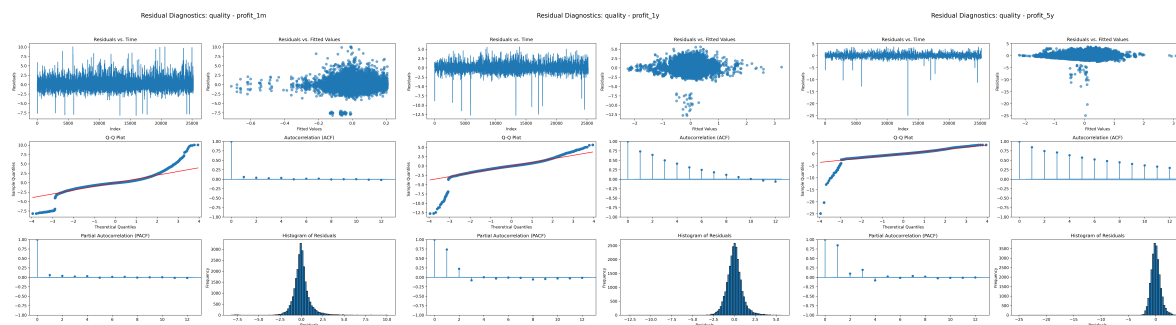


Figure 4.20: Small Data Future (Windowed) - MCODE (4%)

The residues kept having the same issues as the previous models, but **we finally get consistency and significance**. In the table 4.2 we made a summary with relevant results, but for the full data see appendix tables 7.2 to 7.5.

Table 4.2: Small Data Future (Windowed) - MCODE (4%) Consistency Summary

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend_24m_avg	Coef.	0.0041	0.0073	0.0154	0.0172	0.0268	0.0328
	Sign	+	+	+	+	+	+
	t-value	5.09	6.97	7.27	13.30	13.01	11.94
Growth_12m_avg	Coef.	-0.005013	-0.009417	-0.01313	-0.014288	-0.014954	-0.015208
	Sign	-	-	-	-	-	-
	t-value	-3.17	-3.61	-5.27	-5.88	-9.01	-5.38
Growth_24m_avg	Coef.	0.006224	0.011802	0.016234	0.021646	0.02099	0.013468
	Sign	+	+	+	+	+	+
	t-value	3.94	6.77	9.26	9.69	13.09	6.02
Quality_24m_avg	Coef.	0.041583	0.05991	0.108683	0.135561	0.189188	0.072937
	Sign	+	+	+	+	+	+
	t-value	8.63	8.53	18.98	14.57	15.59	9.03
Value	Coef.	0.0654	0.0555	0.0332	0.0198	0.0162	0.0142
	Sign	+	+	+	+	+	+
	t-value	26.78	32.30	19.56	11.05	14.31	12.81

Examining the results, we observe that a company's **24-month averages for Dividend and Quality are positively correlated with overall returns**, where we have highlight the effect of the **Quality score**; this aligns with the theoretical link between these scores and long-term value creation. The relationship with the Growth score is less intuitive but still relevant: **long term growth (24-months averages) is correlated with the profits**. However, temporary slowdowns in short-term growth (12 months averages) can negatively influence market sentiment, often leading to a decline in share price relative to fundamentals. This compression could result in more attractive value multiples—reflected in **Value scores which show a positive correlation** with future profitability.

These results are very promising and indicate a clear trend; however, we cannot be certain of their validity yet because the residuals do not meet all of the model's assumptions.

4.2.2 Geographical Effects

Although all modifications implemented so far have led to improvements, the overall performance remains relatively low. This outcome is common when attempting to fit a linear model to data that exhibits inherently non-linear relationships. To investigate this further, we examined whether these non-linearities could be attributed to regional effects. In contrast, sector-related effects should already be accounted for within Tweenvest's algorithm. Additionally, we assessed whether the presence of outliers could be linked to specific regions.

Simple Models

Following the same approach as for the general datasets, we started with the simple model and then we kept iterating with the same steps as before. So for not being redundant we are only showing the significant results in the section, and the full data is in the appendix figures 7.21 to 7.26 and tables 7.6 to 7.10. When implementing this first model, we obtained similar poor results in performances, but we started to see some behavior that will be constant in all of the models:

- **North America** has the best predictabilities and accuracies.
- **Asia & Oceania** show more outliers than the rest.
- **Different Coefficient behavior** between regions, which confirms the non-linearities.

But once we applied the *Yeo-Johnson* transformation, **the results started to be relevant**. In figure 4.21 we see **higher overall predictabilities** than the ones we had using the original dataset. This also indicates the presence of non-linearities with the geographical regions.

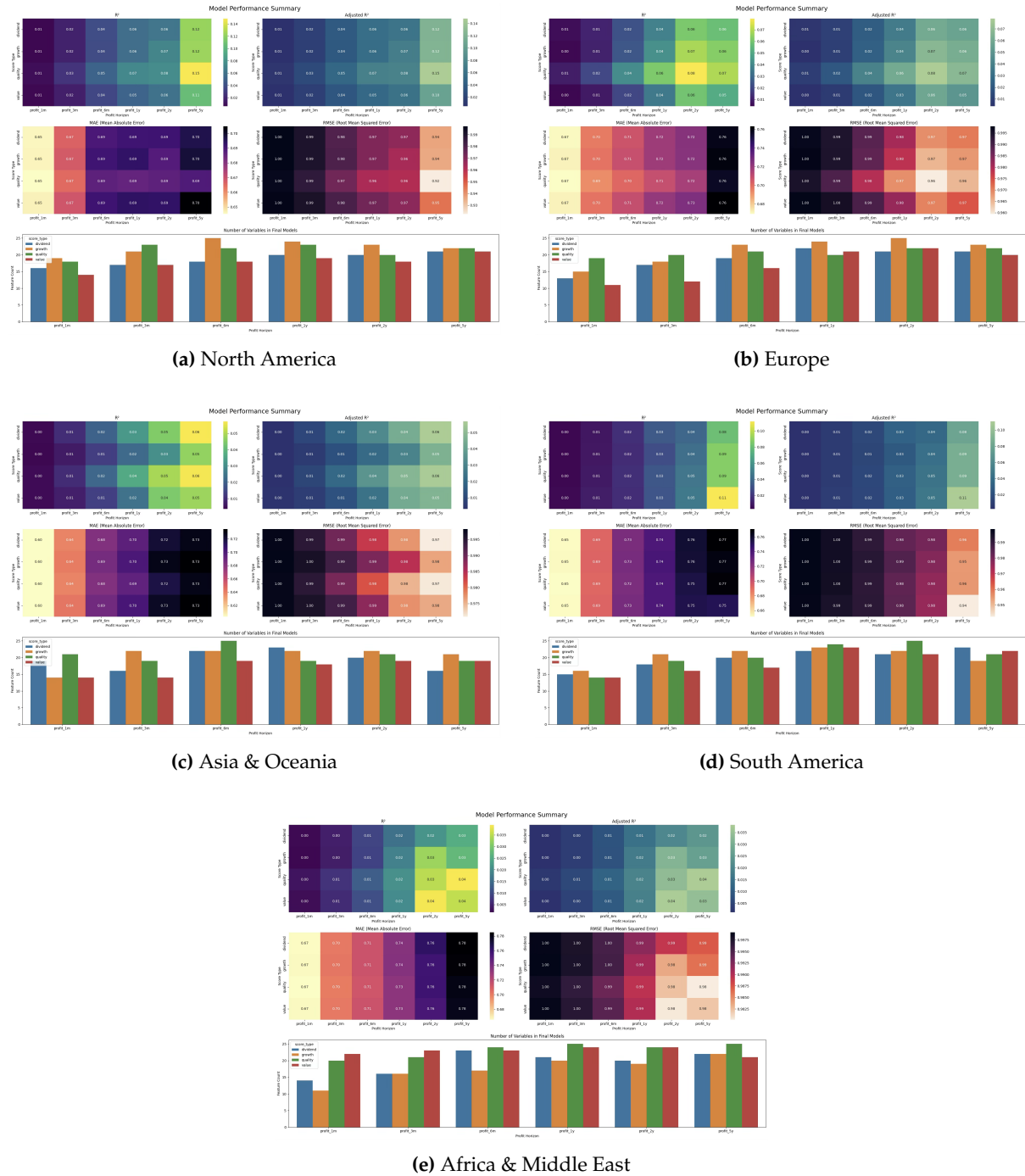


Figure 4.21: Geographical Predictability Analysis - *Yeo-Johnson*

In this case, the **residues' normality improved significantly**, which is along the improvement in performance a good sign of progress. But we still have heteroskedasticity and, of course, the autocorrelations issue for most of the regions and models. Although we do not have perfect models, looking at the Q-Q plots and more from 4.24, we can be certain that **most of the outliers are coming from Asia & Oceania**; which is also the region with more data-points in the complete dataset.

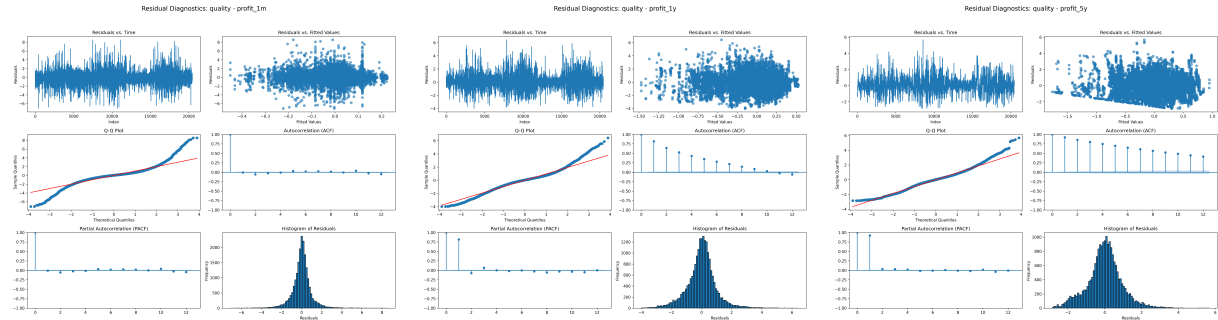


Figure 4.22: North America Yeo-Johnson - Model Residuals

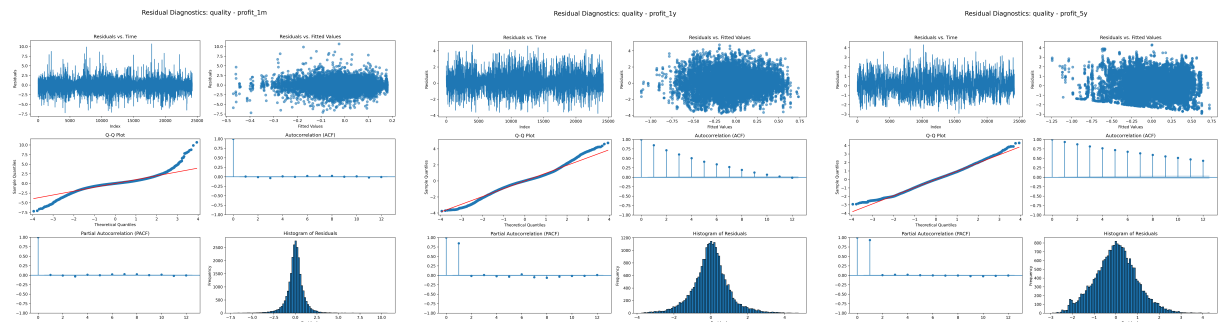


Figure 4.23: Europe Yeo-Johnson - Model Residuals

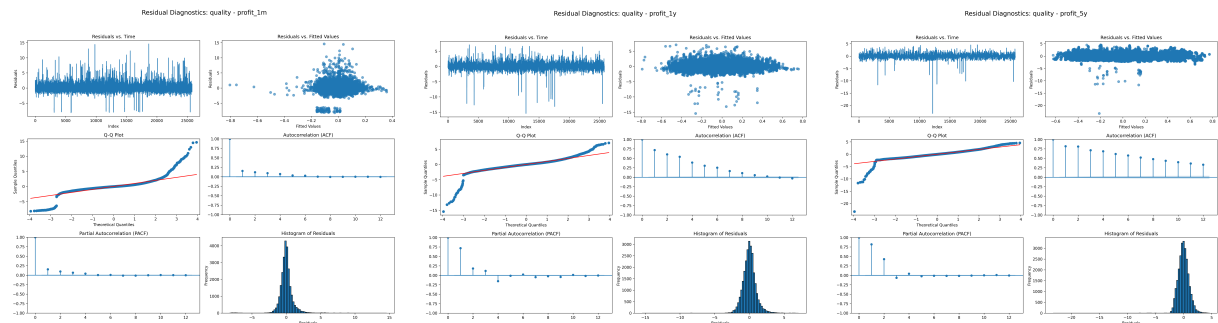


Figure 4.24: Asia & Oceania Yeo-Johnson - Model Residuals

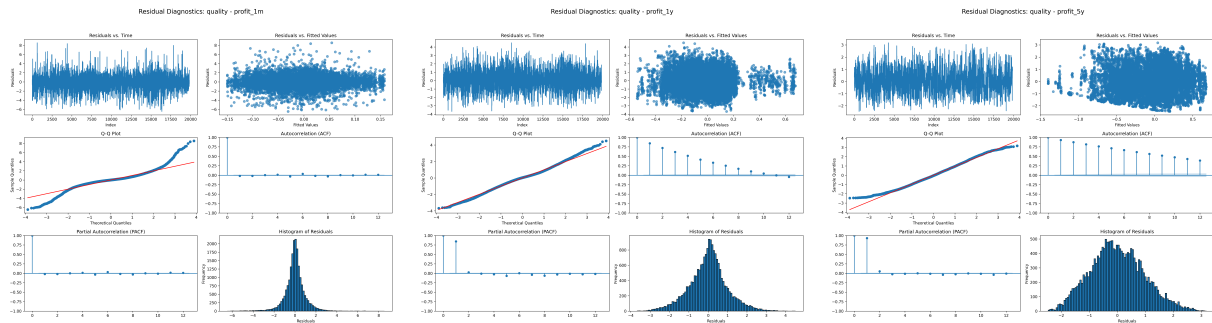


Figure 4.25: South America *Yeo-Johnson* - Model Residuals

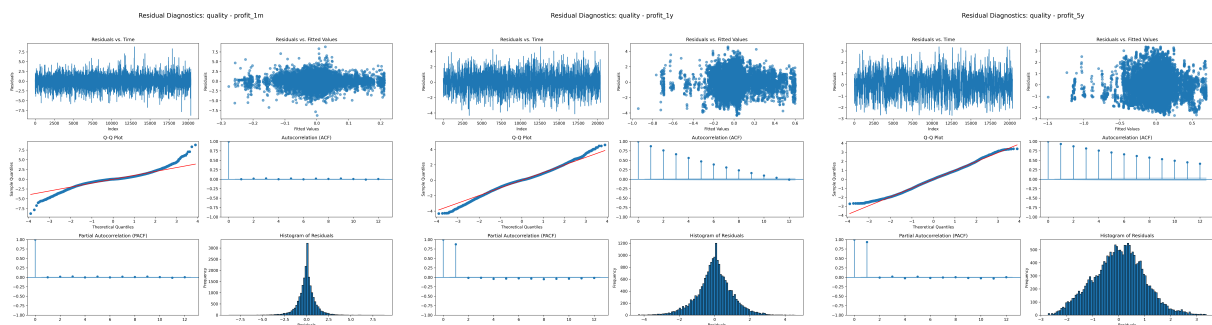


Figure 4.26: Africa & Middle East *Yeo-Johnson* - Model Residuals

Also, with these more stable residues; we can study in detail the **coefficients' results**. But before starting going in depth with them, we have to note that they **behave similarly than before applying the transformation**; which is a good sign of consistency and robustness. Now we will explore all of the regions individually and try to check if the patterns presented could be supported by economic theory.

Starting with **North America** in table 4.3, we highlight the following:

- **Dividend** seems to have a positive correlation with the profits.
- **Value** appears to have a small positive impact on the profits, and since the american market is very volatile it could help investors decide when to enter it.

Table 4.3: North America Yeo-Johnson - Model Consistency

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	–	0.003705	0.001228	0.001495	0.000676	0.000867
	Sign	0	+	+	+	+	+
	t-value	–	15.100	6.625	8.088	4.972	6.361
Growth	Coef.	0.000722	-0.000367	-0.001513	-0.000498	-0.000654	-0.002101
	Sign	+	–	–	–	–	–
	t-value	3.096	-2.106	-4.907	-1.868	-3.320	-4.838
Quality	Coef.	0.003300	–	0.001479	–	-0.001368	–
	Sign	+	0	+	0	–	0
	t-value	7.982	–	6.710	–	-2.353	–
Value	Coef.	-0.002220	0.001054	0.000481	0.001541	0.001650	0.001226
	Sign	–	+	+	+	+	+
	t-value	-8.957	6.194	2.244	7.961	5.649	4.921

Europe in table 4.4, shows behaviors that are more common to a **stable market that prioritizes financial health**, which could be related to the EU investment conservative policies:

- **Dividend and Quality** show a clear positive impact, with statistical significance. We have previously observed this behavior in the earlier models and even with simple correlations.
- **Growth** could have a negative impact, but it is not statistically significant so we can't be sure.
- **Value** could have a positive impact, but it is not constant across all time horizons.

Table 4.4: Europe Yeo-Johnson - Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	0.000424	0.002823	0.000729	0.000595	0.001202	0.000445
	Sign	+	+	+	+	+	+
	t-value	5.046673	16.154619	6.037506	3.758153	7.856079	3.920517
Growth	Coef.	-0.000197	-0.003022	-0.001026	-0.000779	-0.001040	-0.000397
	Sign	–	–	–	–	–	–
	t-value	-1.644037	-11.376062	-4.547723	-2.890478	-2.929000	-3.077300
Quality	Coef.	0.001838	0.005873	0.004024	0.004615	0.006613	0.002814
	Sign	+	+	+	+	+	+
	t-value	5.386076	15.618397	11.439039	14.357360	18.654053	7.928337
Value	Coef.	–	0.002201	–	0.001557	0.002308	–
	Sign	0	+	0	+	+	0
	t-value	–	7.980292	–	7.571191	10.664820	–

For the regions of **Asia & Oceania** in table 4.5, it seems that **all factors are positive and significant**. This result could mean that in a growing market, investors makes more rational decisions also because unmeasurable effects –such as brand strength– are not as relevant as in North America or Europe causing less effects like FOMO. However, these markets exhibit lower stability and high volatility, necessitating careful consideration.

Table 4.5: Asia & Oceania Yeo-Johnson - Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	0.002706	0.001996	0.002103	0.002697	0.001314	0.000921
	Sign	+	+	+	+	+	+
	t-value	16.305686	15.822070	14.489915	17.023119	12.235443	5.669478
Growth	Coef.	0.000821	0.000922	0.000713	0.001069	0.000602	0.000293
	Sign	+	+	+	+	+	+
	t-value	4.280952	4.795172	3.697878	5.763015	3.114600	2.687339
Quality	Coef.	0.007622	0.008248	0.007712	0.009832	0.003962	0.002321
	Sign	+	+	+	+	+	+
	t-value	22.085857	25.980872	8.327695	25.002458	12.139435	6.139051
Value	Coef.	0.001930	0.003812	0.000879	0.003423	0.001265	0.000900
	Sign	+	+	+	+	+	+
	t-value	10.168522	22.223979	4.377369	21.304016	9.347055	6.903840

South America in table 4.6, shows a very similar behavior to **Asia & Oceania**, but with less consistency across time horizons.:

- **Dividend** is slightly positive and significant for the mid and long term.
- **Growth** and **Quality** are positive and significant for all time horizons.
- **Value** is positive and significant except for 3M and 1Y.

Table 4.6: South America Yeo-Johnson - Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	0.000589	-0.001633	0.001206	0.000545	0.000703	0.000570
	Sign	+	-	+	+	+	+
	t-value	2.993752	-4.461563	8.253964	2.915853	5.427420	4.236314
Growth	Coef.	0.003605	0.004210	0.001985	0.006252	0.001384	0.000641
	Sign	+	+	+	+	+	+
	t-value	10.886228	16.229706	5.589288	17.793972	3.952738	3.416957
Quality	Coef.	0.009461	0.005873	0.005789	0.014780	0.003928	0.002168
	Sign	+	+	+	+	+	+
	t-value	10.912310	10.844574	12.511430	8.968703	9.048040	6.211781
Value	Coef.	0.005235	-	0.002801	-	0.002318	0.002174
	Sign	+	0	+	0	+	+
	t-value	9.745774	-	11.726903	-	10.919653	6.196305

Finally, for the regions of **Africa & Middle East** in table 4.7, we get the same results. Showing that it is very unpredictable with these models, and we can only highlight that **Quality** is **positive and significant for all time horizons** and Dividend seems to have a small negative impact; but since the effects are so small it could be overfitting.

Table 4.7: Africa & Middle East *Yeo-Johnson* - Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	-0.000866	-0.004017	-0.000262	-0.001265	-0.000145	–
	Sign	–	–	–	–	–	0
	t-value	-7.56	-7.76	-4.84	-6.82	-4.30	–
Growth	Coef.	–	–	–	0.000679	–	–
	Sign	0	0	0	+	0	0
	t-value	–	–	–	2.80	–	–
Quality	Coef.	0.002288	0.007478	0.001800	0.005933	0.000984	0.000373
	Sign	+	+	+	+	+	+
	t-value	7.99	10.52	6.65	9.34	7.50	5.22
Value	Coef.	0.000324	0.005205	–	0.001033	–	–
	Sign	+	+	0	+	0	0
	t-value	2.34	5.91	–	3.27	–	–

Windowed Models

Continuing with the methodology, we applied the windowed models to all of the regions datasets, and we ended up with the most robust results so far. Looking at the performances in figure 4.27, we can see the similar results as before:

- **Predictability:** **North America** has the best predictabilities and accuracies. On the other side, **Asia & Oceania** and **Africa & Middle East** have the worst performances; the first due to their outliers and the second probably because of its financial and political instability.
- **Relevant Scores:** across all regions, we can conclude that the **Value** and **Quality** models outperform the others.

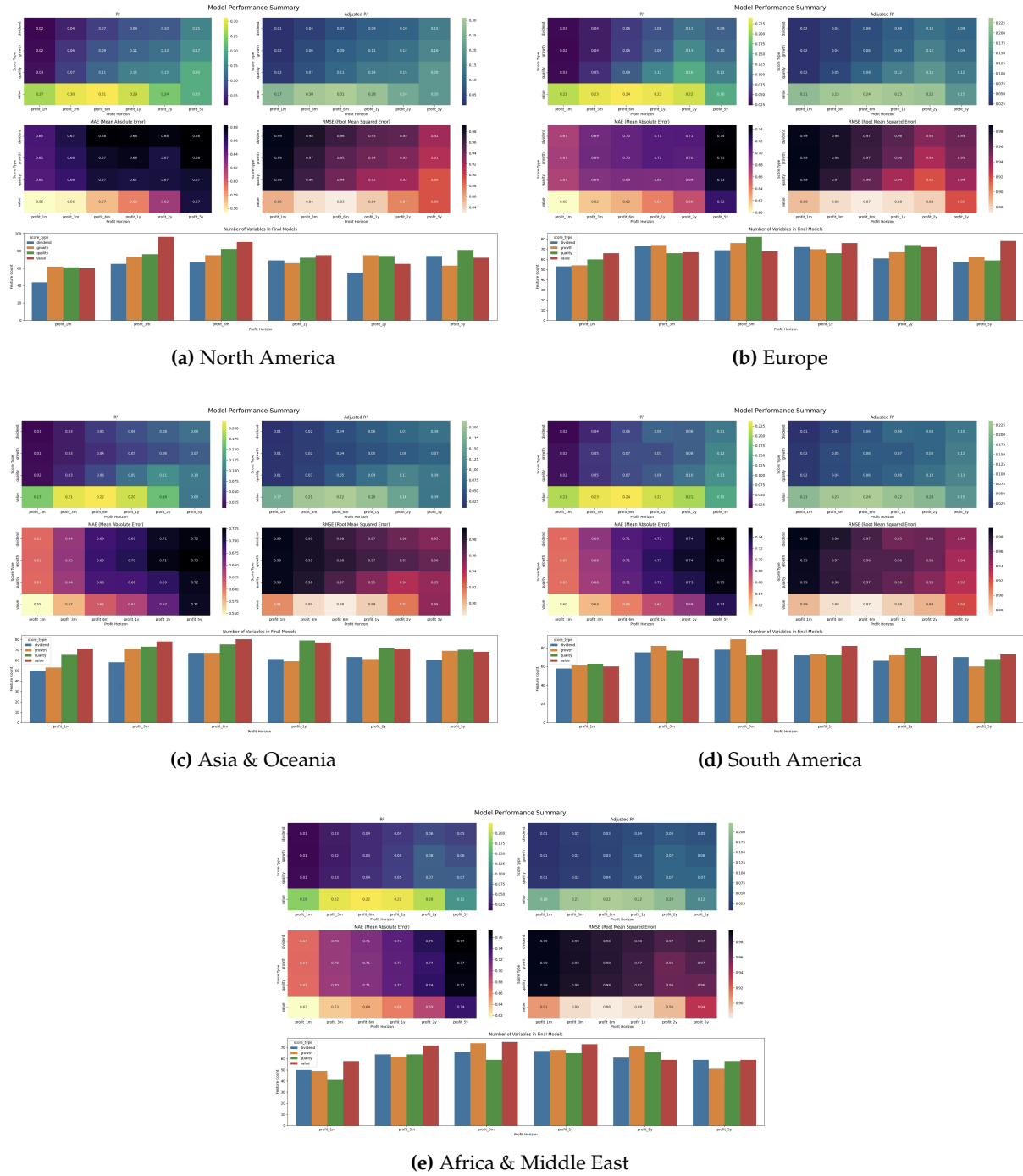


Figure 4.27: Geographical Predictability Analysis - Windowed Models

Before analyzing the coefficients, we must check on the residuals. Finally, we observe that not only is normality good, but we also start to observe some homoscedasticity in figures 4.28 to 4.32. Looking closely, North America and Europe are show similar patterns in terms of outliers and with a slight underestimation on the long term (RMSE).

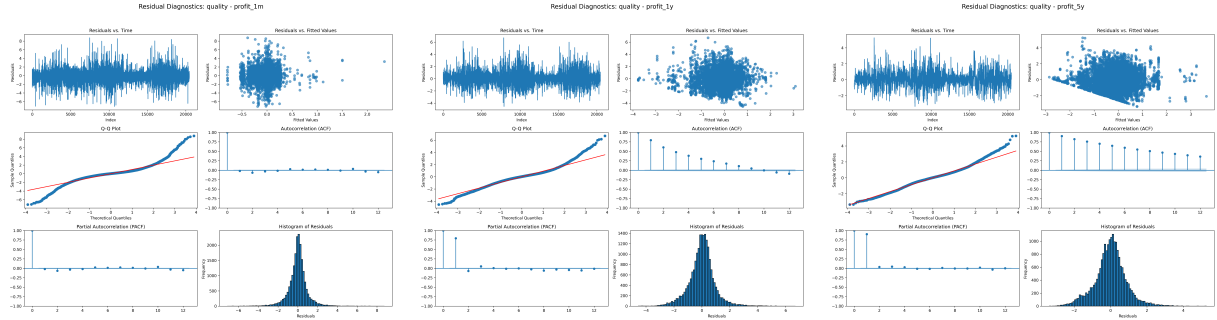


Figure 4.28: North America - Windowed Model Residuals

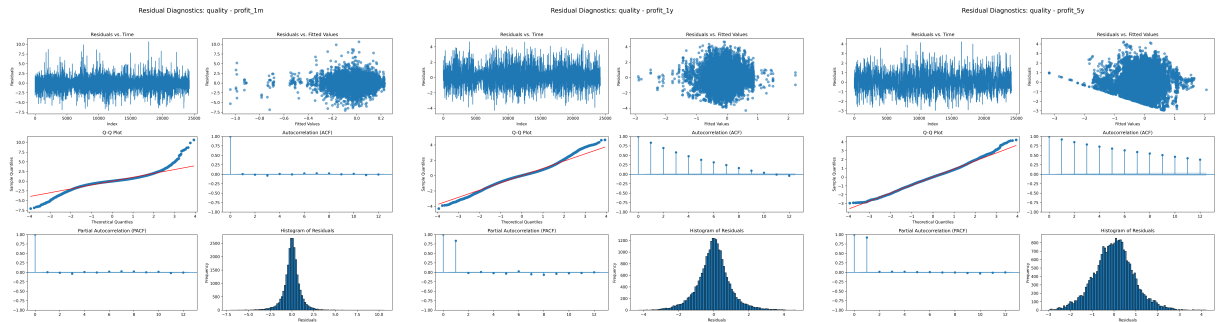


Figure 4.29: Europe - Windowed Model Residuals

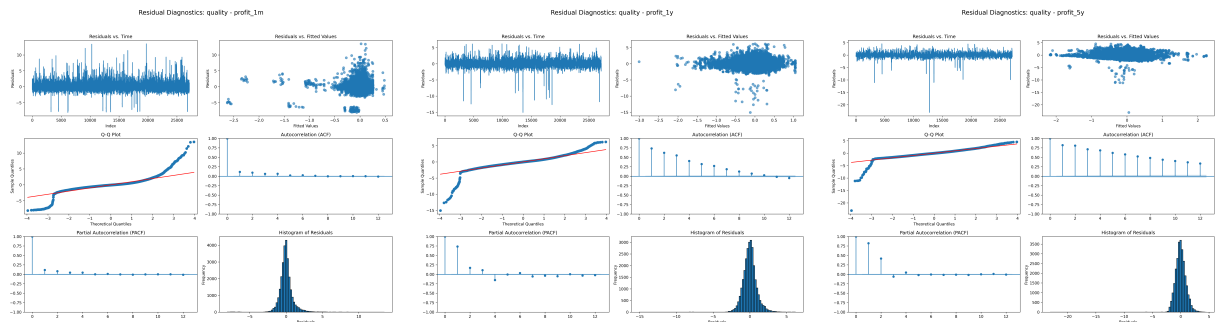


Figure 4.30: Asia & Oceania - Windowed Model Residuals

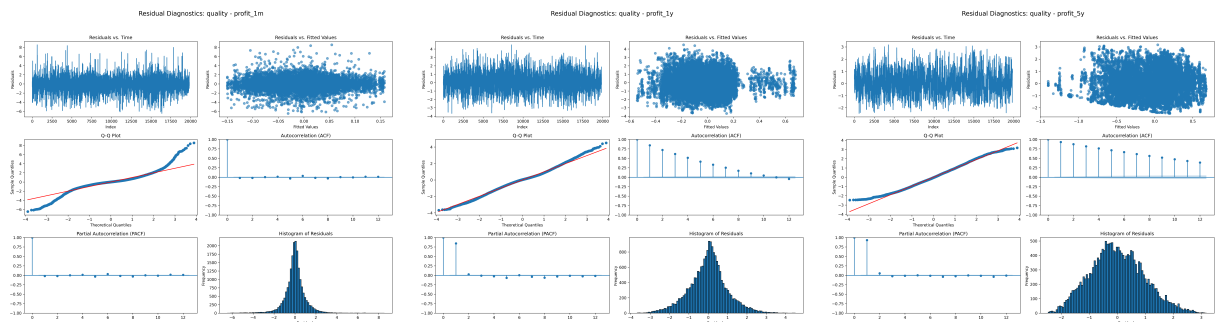


Figure 4.31: South America - Windowed Model Residuals

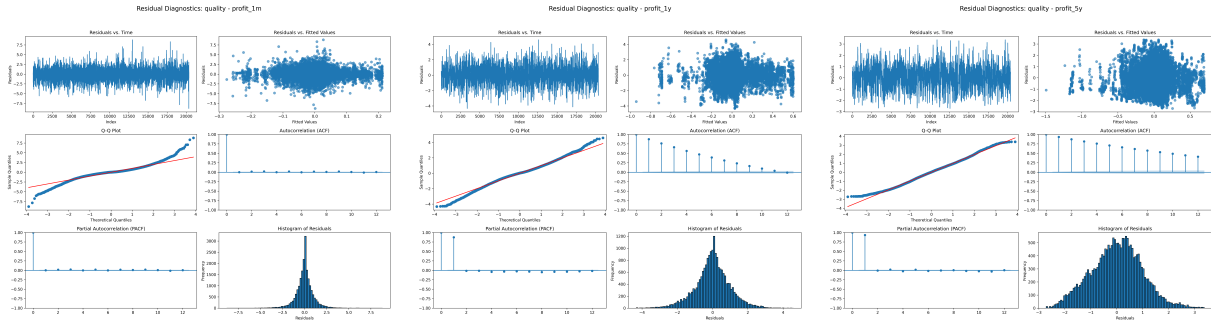


Figure 4.32: Africa & Middle East - Windowed Model Residuals

For analyzing the coefficients, we created a summary table per region (full data in appendix figures 7.11 to 7.30). Starting with North America and Europe in figures 4.8 to 4.9, we see a common behavior between them and previous results, where only **long term Value** is **negatively correlated to profits**, and stable scores increase overall returns.

Table 4.8: North America (Windowed) - Consistency Summary

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Growth_24m_avg	Coef.	0.003565	0.007495	0.013033	0.015419	0.006597	0.001522
	Sign	+	+	+	+	+	+
	t-value	6.300	9.227	13.719	15.737	5.889	4.469
Quality_24m_avg	Coef.	0.019811	0.042821	0.054187	0.081220	0.065399	0.017921
	Sign	+	+	+	+	+	+
	t-value	7.656	10.017	12.563	17.276	9.129	2.908
Value	Coef.	0.046496	0.035204	0.020768	0.012553	0.008890	0.006598
	Sign	+	+	+	+	+	+
	t-value	55.124	45.616	25.571	17.930	16.524	15.597
Value_24m_avg	Coef.	-0.003114	-0.016167	-0.011902	-0.012283	-0.028632	-0.027899
	Sign	-	-	-	-	-	-
	t-value	-5.754	-5.969	-10.147	-12.061	-29.415	-26.523

Table 4.9: Europe (Windowed) - Consistency Summary

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend_24m_avg	Coef.	0.006838	0.012544	–	0.011382	0.009546	0.010074
	Sign	+	+	0	+	+	+
	t-value	9.487	13.741	–	14.584	13.924	11.231
Growth_24m_avg	Coef.	0.002266	0.006223	0.009021	0.010648	0.009251	0.005465
	Sign	+	+	+	+	+	+
	t-value	4.977	8.715	9.367	10.281	12.791	7.113
Quality_24m_avg	Coef.	0.019378	0.088280	0.073143	0.086730	0.099170	0.046953
	Sign	+	+	+	+	+	+
	t-value	10.140	19.679	12.576	18.025	18.959	10.833
Value	Coef.	0.041385	0.007550	0.017503	0.012756	0.009658	0.033926
	Sign	+	+	+	+	+	+
	t-value	26.063	12.232	23.779	17.162	16.896	26.338
Value_24m_avg	Coef.	-0.003632	-0.019976	-0.006693	-0.014679	-0.027589	-0.006000
	Sign	–	–	–	–	–	–
	t-value	-4.981	-20.202	-8.512	-17.659	-22.505	-9.887

Surprisingly, when looking at figures 4.10 to 4.12 they all behave almost the exact same way as the first two regions and also it appears that short-term growth is negatively correlated. This similarity in patterns is a very good sign of stable behavior.

Table 4.10: Asia & Oceania (Windowed) - Consistency Summary

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend_24m_avg	Coef.	0.003419	0.007207	0.008895	0.006618	0.008068	0.008521
	Sign	+	+	+	+	+	+
	t-value	4.896	9.449	11.231	9.421	10.853	11.601
Growth_12m_avg	Coef.	-0.003753	-0.004498	-0.005803	-0.005715	-0.010915	-0.014371
	Sign	–	–	–	–	–	–
	t-value	-3.629	-4.332	-5.528	-5.221	-10.270	-8.390
Growth_24m_avg	Coef.	0.002967	0.004537	0.006634	0.006872	0.010789	0.011809
	Sign	+	+	+	+	+	+
	t-value	5.117	7.372	9.865	10.318	13.006	11.118
Quality	Coef.	-0.023861	-0.019494	-0.020660	-0.014892	-0.019225	-0.008036
	Sign	–	–	–	–	–	–
	t-value	-4.569	-9.036	-8.059	-5.597	-6.275	-3.893
Quality_24m_avg	Coef.	0.010252	0.022449	0.039548	0.076738	0.078367	0.125694
	Sign	+	+	+	+	+	+
	t-value	4.614	10.375	14.268	19.246	20.388	14.657
Value	Coef.	0.042605	0.026478	0.012090	0.012708	0.008177	0.004550
	Sign	+	+	+	+	+	+
	t-value	33.253	26.853	16.222	17.911	12.156	7.702
Value_24m_avg	Coef.	–	–	–	-0.010340	-0.028875	-0.021639
	Sign	0	0	0	–	–	–
	t-value	–	–	–	-14.361	-30.132	-13.536

Table 4.11: South America (Windowed) - Consistency Summary

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Growth_12m_avg	Coef.	-0.007731	-0.017594	-0.034042	-0.029328	-0.019617	-0.017643
	Sign	–	–	–	–	–	–
	t-value	-5.135	-9.873	-7.215	-14.635	-12.264	-11.736
Growth_24m_avg	Coef.	0.007296	0.013563	0.023194	0.018842	0.017345	0.014027
	Sign	+	+	+	+	+	+
	t-value	8.429	13.641	9.135	16.584	16.198	14.528
Quality_24m_avg	Coef.	–	0.012276	0.042725	0.110538	0.151205	0.104161
	Sign	0	+	+	+	+	+
	t-value	–	3.320	16.045	9.841	14.207	12.109
Value	Coef.	0.033626	0.026984	0.015424	0.011372	0.008776	0.005103
	Sign	+	+	+	+	+	+
	t-value	48.474	38.593	24.273	19.666	15.566	9.407
Value_24m_avg	Coef.	-0.001221	-0.002786	-0.002022	-0.013676	-0.026588	-0.013843
	Sign	–	–	–	–	–	–
	t-value	-1.557	-3.581	-2.360	-12.981	-15.264	-11.707

Table 4.12: Africa & Middle East (Windowed) - Consistency Summary

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Growth_12m_avg	Coef.	-0.005066	-0.012923	-0.015521	-0.018503	-0.018322	-0.002617
	Sign	–	–	–	–	–	–
	t-value	-4.987	-7.110	-7.855	-10.096	-6.795	-3.501
Growth_24m_avg	Coef.	0.005516	0.012733	0.017077	0.020452	0.017803	0.006444
	Sign	+	+	+	+	+	+
	t-value	5.413	8.572	10.828	10.821	7.900	8.692
Quality_24m_avg	Coef.	0.011701	0.010906	0.017645	0.029879	0.022980	0.046570
	Sign	+	+	+	+	+	+
	t-value	9.041	5.147	6.630	7.643	5.528	13.094
Value	Coef.	0.036820	0.026390	0.015372	0.010815	0.008244	0.005772
	Sign	+	+	+	+	+	+
	t-value	42.388	39.243	23.029	17.932	18.351	15.446
Value_24m_avg	Coef.	–	-0.005729	-0.003473	-0.015642	-0.029044	-0.013761
	Sign	0	–	–	–	–	–
	t-value	–	-5.946	-4.516	-15.917	-32.465	-15.460

4.2.3 Balanced Dataset

Previous analysis indicate the importance of non-linear relationships from regions, and the amount of outliers in this data. So before continuing with the models, we created a new dataset with less outliers in total and balancing the regions, and confirmed that **it improved the performances and residues**, see appendix figures 7.27 and 7.28, but it worsened the effect of the *momentum*, showing higher ACF values.

4.3 GAM

Observing the consistent improvements in the results suggests that the decisions made regarding the treatment and structure of the data have been effective. The next logical step is to incorporate models capable of capturing the non-linear relationships between variables, thereby enhancing performance. At the same time, and in line with the objectives established at the outset of this work, we aim to assist Tweenvest in defining recommended score ranges that will help investors better interpret and apply them. This could become straightforward once the spline functions for each variable are analyzed and understood.

4.3.1 Limitations and Considerations

One big limitation of GAMs is **their tendency to overfitting**, partly due to the nature of the spline functions, which are constructed from polynomial and trigonometric components. To address this, we adopt the following precautions:

- The endings of the spline curves may lack statistical interpretation because of the mathematical form of the fitted functions; also known as Runge's phenomenon [28] and Gibbs phenomenon [29], for polynomial and trigonometric functions respectively. For this reason, **we focus our interpretation on overall trends** rather than on boundary values.
- To minimize the risk of misinterpretation caused by overfitting —and to maintain consistency with the findings from the linear models— we concentrate primarily on variables that have demonstrated significance across multiple datasets.

4.3.2 Results

As we could have expected, the GAM models have **improved the overall predictability**. Seeing in 4.34 for the first time **performances over the 40% mark** combined with better accuracies than ever. Unfortunately, these results still are not good enough to be used in a real test with non-trained data.

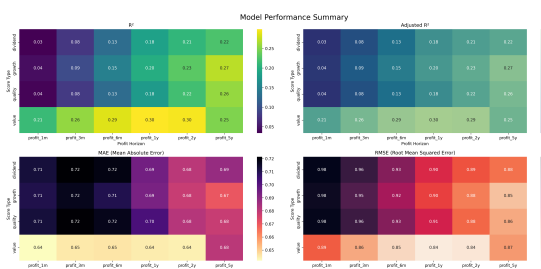


Figure 4.33: Small Data Future IF - GAM Performance

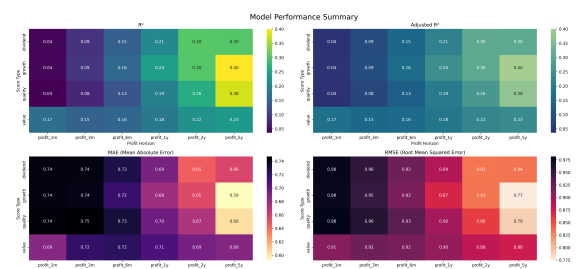


Figure 4.34: Big Data Future IF Balanced - GAM Performance

4.3.3 Non-linear relationships

Here we will be exploring the results of the GAM models to check if they are consistent with earlier findings or not.

Simple Splines

We begin by examining the simple splines, which relate each explanatory variable to a profit time-frame. Starting with Dividend in figures 4.35 to 4.36, we observe better and clearer results for the non-balanced dataset, but both display a generally positive slope across most profit horizons, which is consistent with the linear regression results.

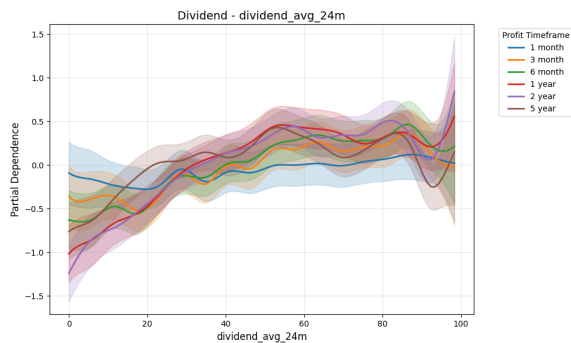


Figure 4.35: Small Data Future IF - Dividend 24M Average Splines

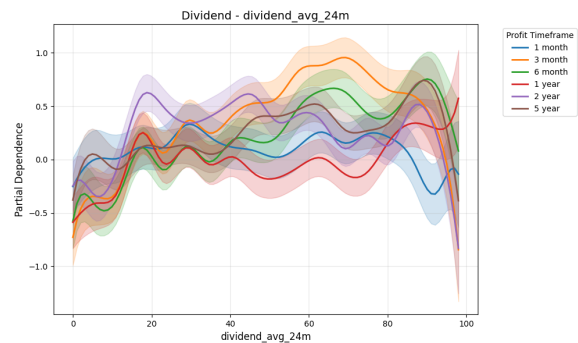


Figure 4.36: Big Data Future IF Balanced - Quality 24M Average Splines

When it comes to both, 12M and 24M Growth's splines in figures 4.37 to 4.38 and 4.39 to 4.40, we have inconsistent results. For each dataset we have different general tendencies, so we can't be sure that the behavior seen before is actually true.

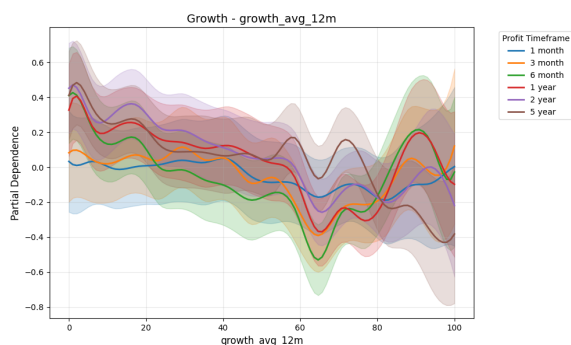


Figure 4.37: Small Data Future - IF 15% - Growth 12M Average Splines

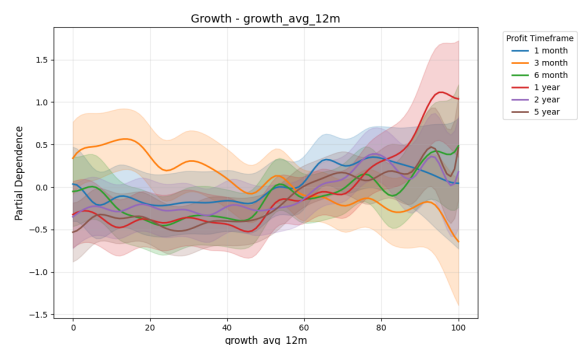


Figure 4.38: Big Data Future IF - Growth 12M Average Splines

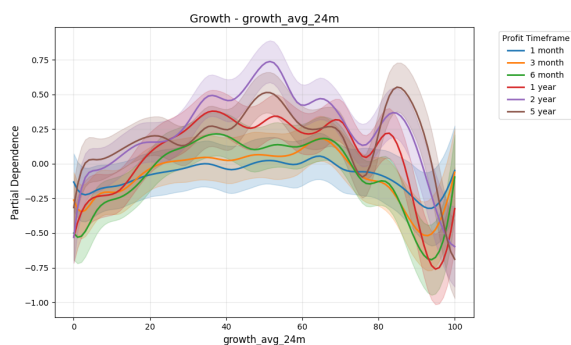


Figure 4.39: Small Data Future IF - Growth 24M Average Splines

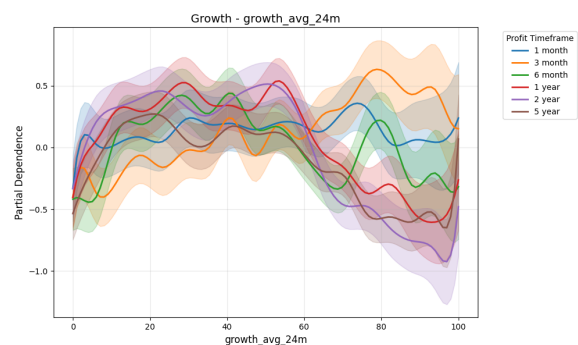


Figure 4.40: Big Data Future IF - Growth 24M Average Splines

For **Quality**, we see a **very slight positive relationship** in figure 4.41, but it is harder in the figure 4.42 since the extreme values are less stable. But it seems that around 50-60, which then rescales to 60-75, we have a small switch in tendency. Notice that **the partial dependence with profits is the second highest one of all**. So it does seem that the results accompany earlier findings.

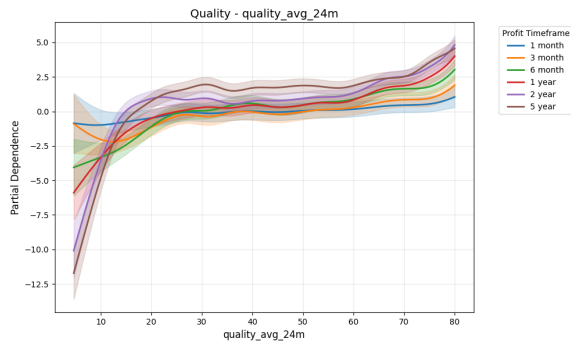


Figure 4.41: Small Data Future IF - Quality 24M Average Splines

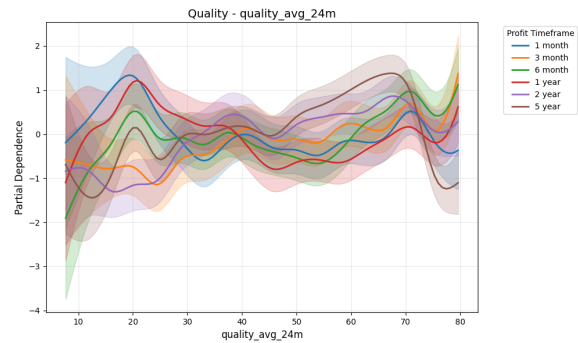


Figure 4.42: Big Data Future IF - Quality 24M Average Splines

Looking at the Value splines in figures 4.43 and 4.46, we observe **consistent and clear alignment with the Value scores across both datasets**, showing the same behavior as the linear models and **producing the strongest impact on profits so far**. Notably, there is a breaking point around a score of 50, beyond which profits begin to improve; as seen in the descriptive analysis, this corresponds to the same change point identified in the distribution.

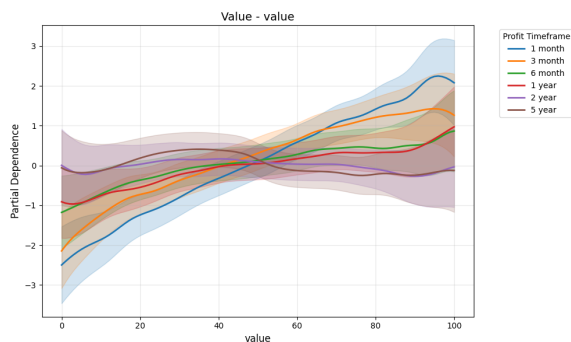


Figure 4.43: Small Data Future IF - Value Splines

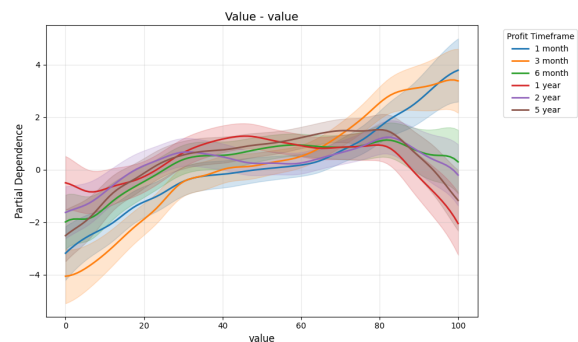


Figure 4.44: Big Data Future IF - Value Splines

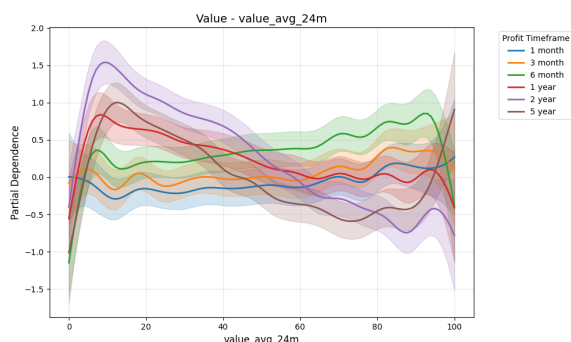


Figure 4.45: Small Data Future IF - Value 24M Average Splines

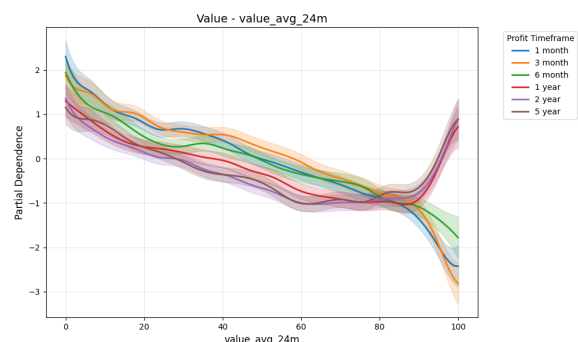


Figure 4.46: Big Data Future IF Balanced - Value 24M Average Splines

Tensor Splines

Working with splines always comes with the risk of overfitting. So following the criteria to avoid false interpretations, we looked at the tensor splines from the variables before analyzed: region and scores; where we could only guess a **slight indicator of possible consistence with the value score** and just for the balanced dataset in figures 4.47 to 4.48. But in general, the lack of interpretable and consistent results reveals its tendency to overfitting.

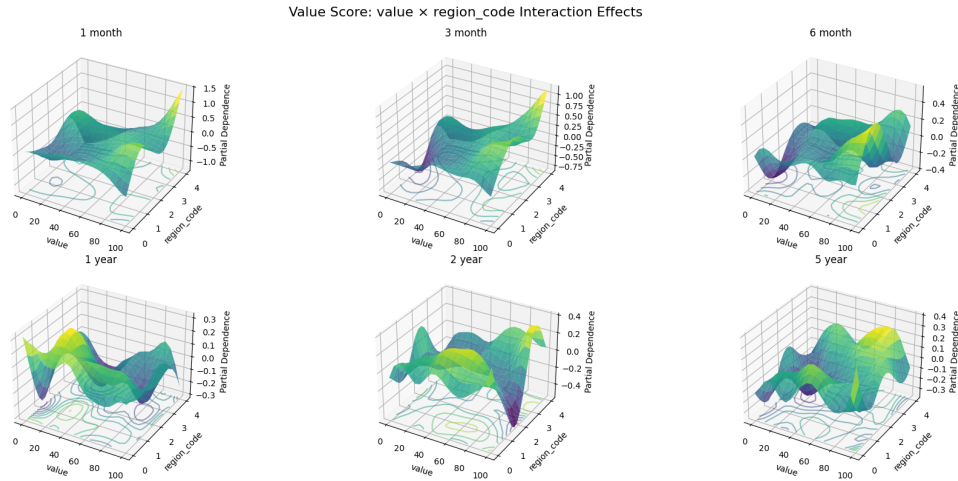


Figure 4.47: Big Data Future IF Balanced - Region x Value Tensors

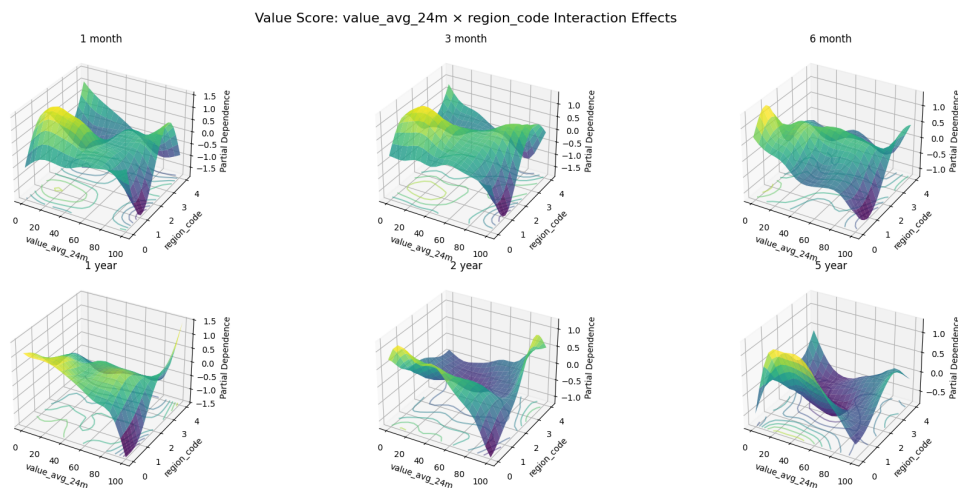


Figure 4.48: Big Data Future IF Balanced - Region x Value 24M Average Tensors

Note: The rest of the splines are in appendix figures 7.29 to 7.34, where you can also find the encoding used for the regions.

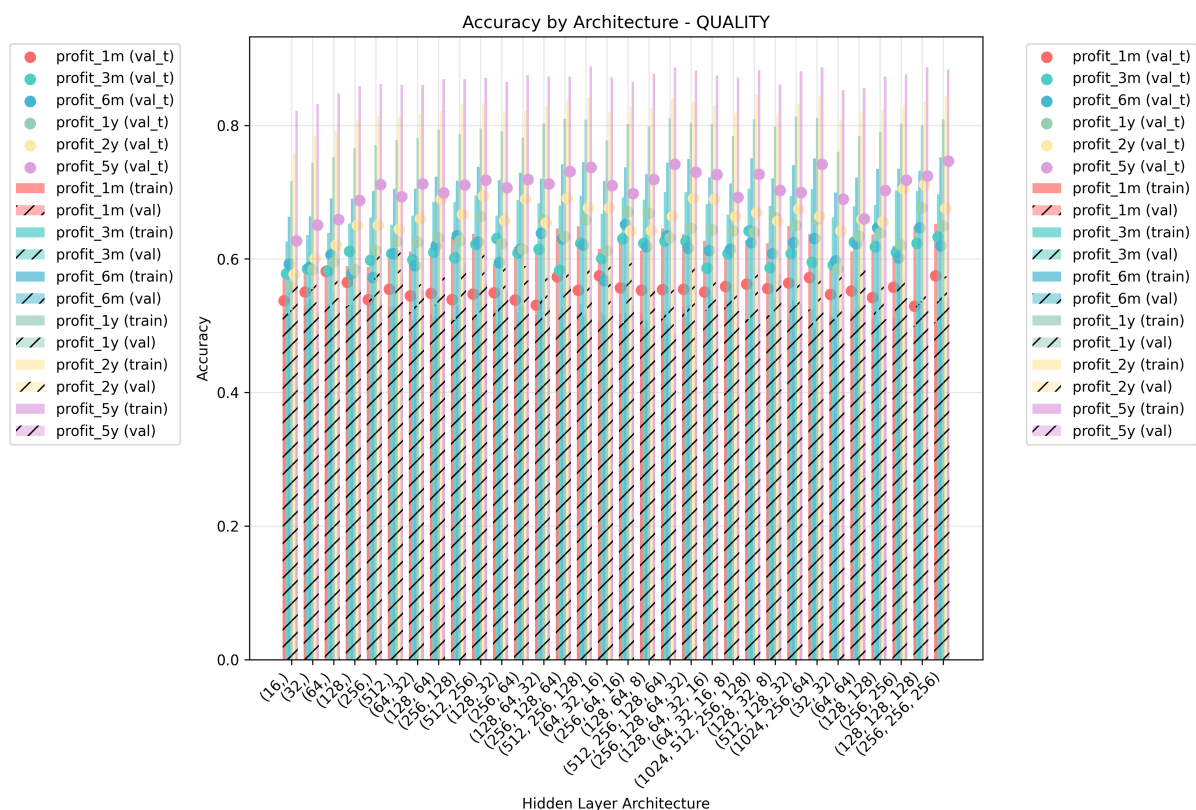
4.4 Neural Network

We have checked that both, linear regression models and the GAM, show consistency between them and a significant improvement when introducing non-linear functions. But even though the results are better, we have to remember that we were only using these models to identify patterns, and not for real testing with non-trained data due to their low accuracies and performances.

Now we are going to check if we can actually predict when not knowing the data. For this we are going to use different NNs, starting from a simple binary classifiers, continuing with encoders and finally regressors.

4.4.1 Architecture Optimization

Before running the simulations for each model, it is always recommended to search for the most optimized group of hyper-parameters to obtain a better performance. In this case, the packages we used from *scikit-learn* already had pre-built optimizations, so we only looked for the best hidden layers and number of neurons per each one. But surprisingly, we found in figures 4.49 to 4.51 that most of the architectures obtained similar results, for the 3 different types of NNs. And because the results were not going to change much, we decided to use the same architecture for all of the models, in this case, a standard (64,64).



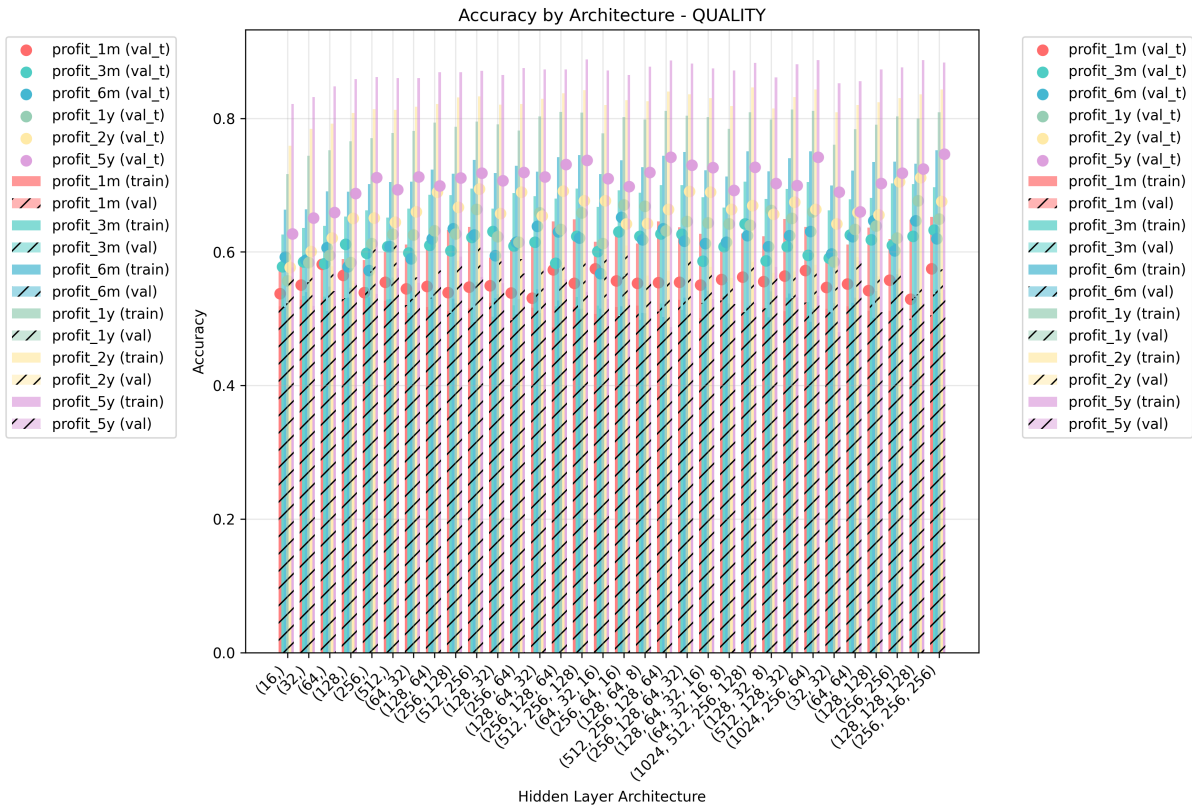


Figure 4.50: Encoder Classifier - (Quality) Hidden Layer Optimization

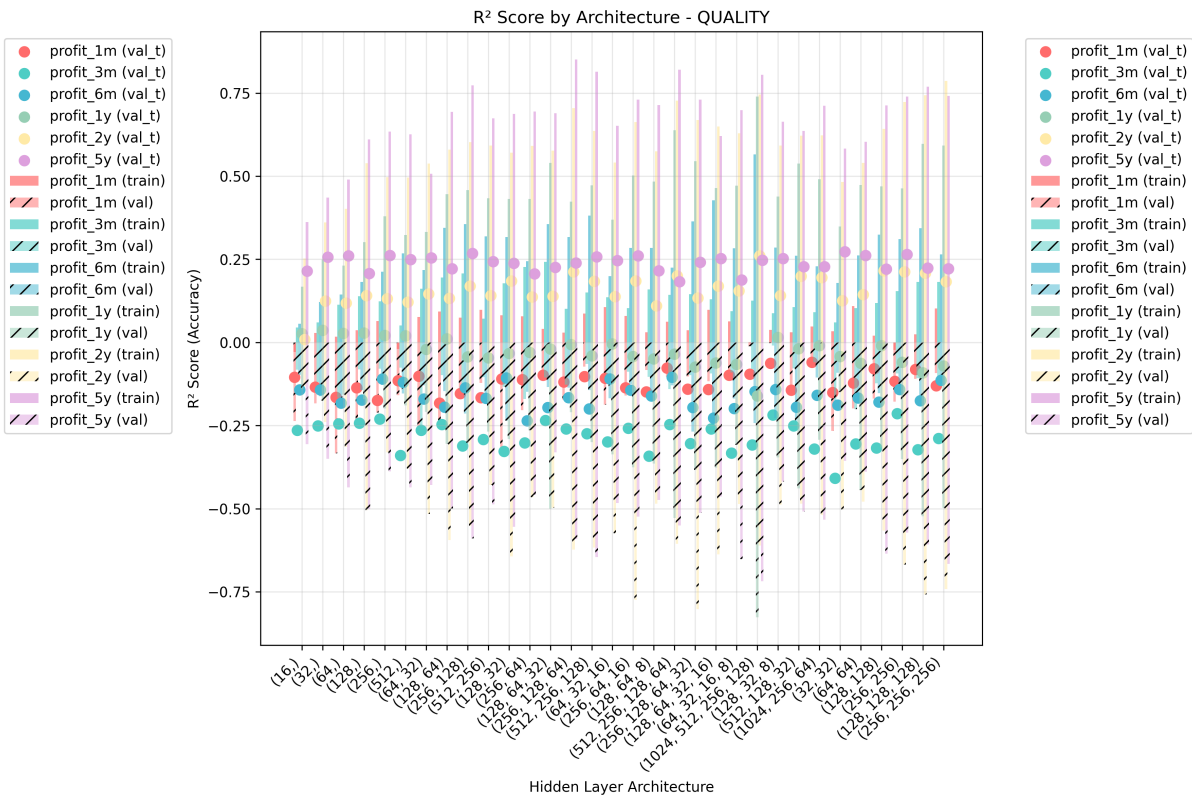


Figure 4.51: Regressor Classifier - (Quality) Hidden Layer Optimization

The plots reveal a consistent pattern observed across all models: **long-term models consistently**

achieve higher performance. Furthermore, as this is the first instance in which validation sets have been used, two additional insights emerge. First, **knowing past behavior of a stock produces higher performances.** Second, **performance decreases for companies that are entirely new to the model.** But this will be discussed individually in the following sections.

4.4.2 Binary Classifier

We started with the most simple NN possible, a model that tries to predict whether a company will have positive or negative profits. So, to properly interpret the results, we have to clarify that since we have transformed a regression predictive problem into a binary classification problem, a complete random guess would get a 50% predictability and a 50% error rate.

Performance

Knowing this, we can see that the results in figure 4.52 are quite good. With a huge increase in predictability for the training set, and for the first time in this work **we were able to predict more than random for validating datasets**, this includes known and unknown companies. Achieving performances over 60% and an error rates under 40%, for both validating datasets. Also it is interesting that some of the past behavior still occurs:

- Value scores perform better at short horizons and have the best performances in most cases.
- Quality scores outperform all of the models for known stocks, but in this case on a validation set.

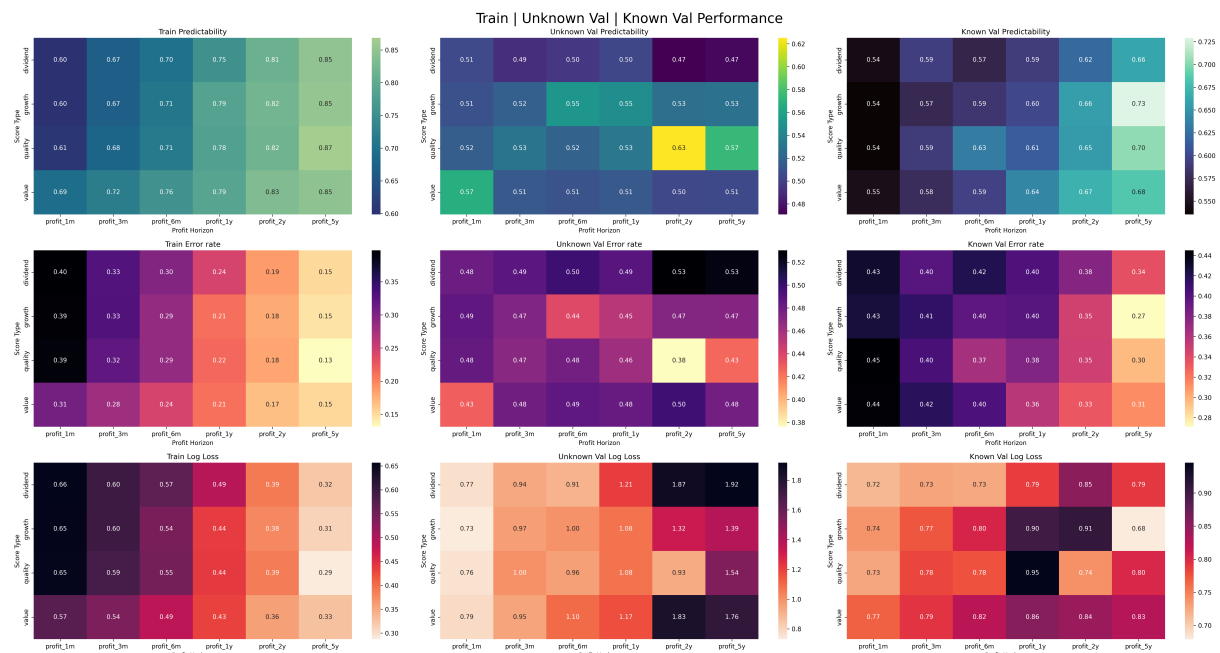


Figure 4.52: Big Data Future IF Balanced - Binary Classifier Performance

SHAP-Consistency

As previously discussed, SHAP is a numerical technique that enables us to quantify the contribution of each variable to the predictions made by the NNs. Applying the same criteria used for

the linear regression and GAM models, we restrict our analysis to variables that were previously identified as significant. Interestingly, the results seen in figures 4.54 and 4.53 mirror those obtained from both the linear regression and GAM models, displaying the same patterns for the variables. These relationships are very clear in the SHAP plots, where: **Dividend, Growth, and Quality 24-month averages, as well as the most recent Value score, show a clear positive impact on profits, whereas Value 24-month averages exhibit a negative impact.**

Neural Network SHAP Dependency Analysis by Variable and Profit Horizon

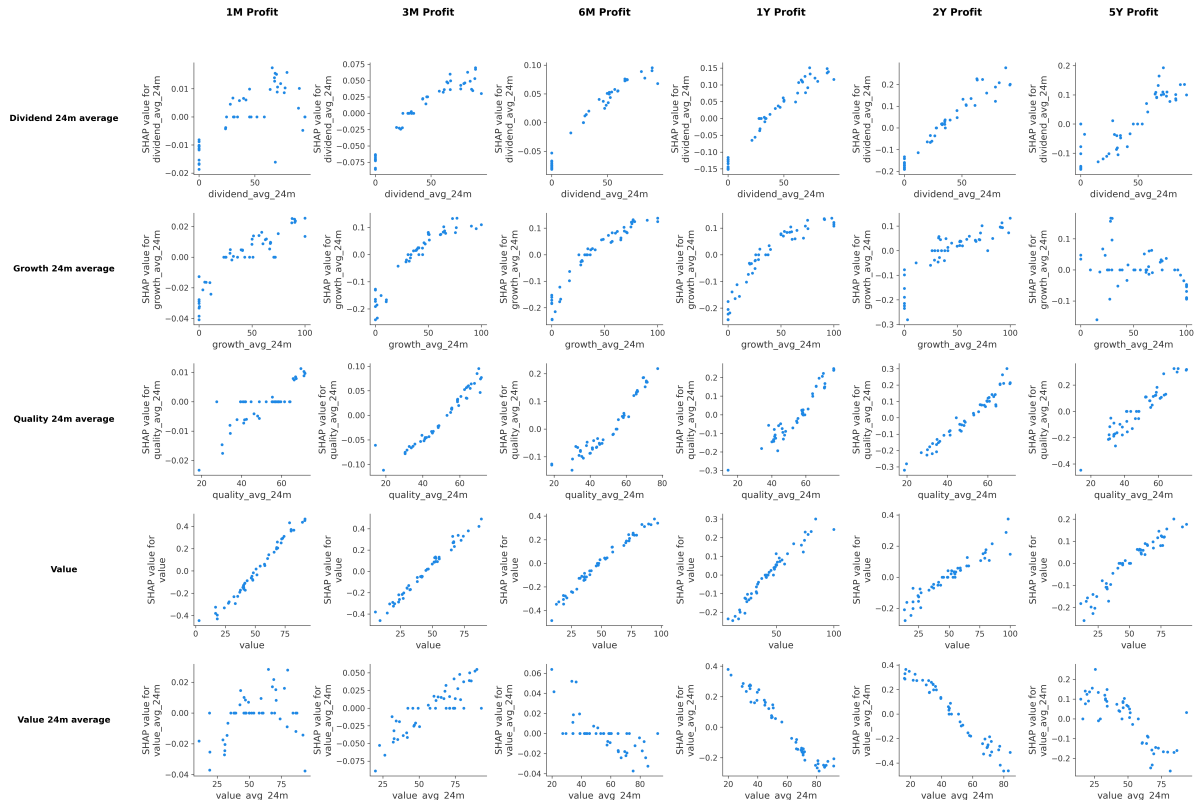


Figure 4.53: Small Data Future IF - Binary SHAP Consistency

Neural Network SHAP Dependency Analysis by Variable and Profit Horizon

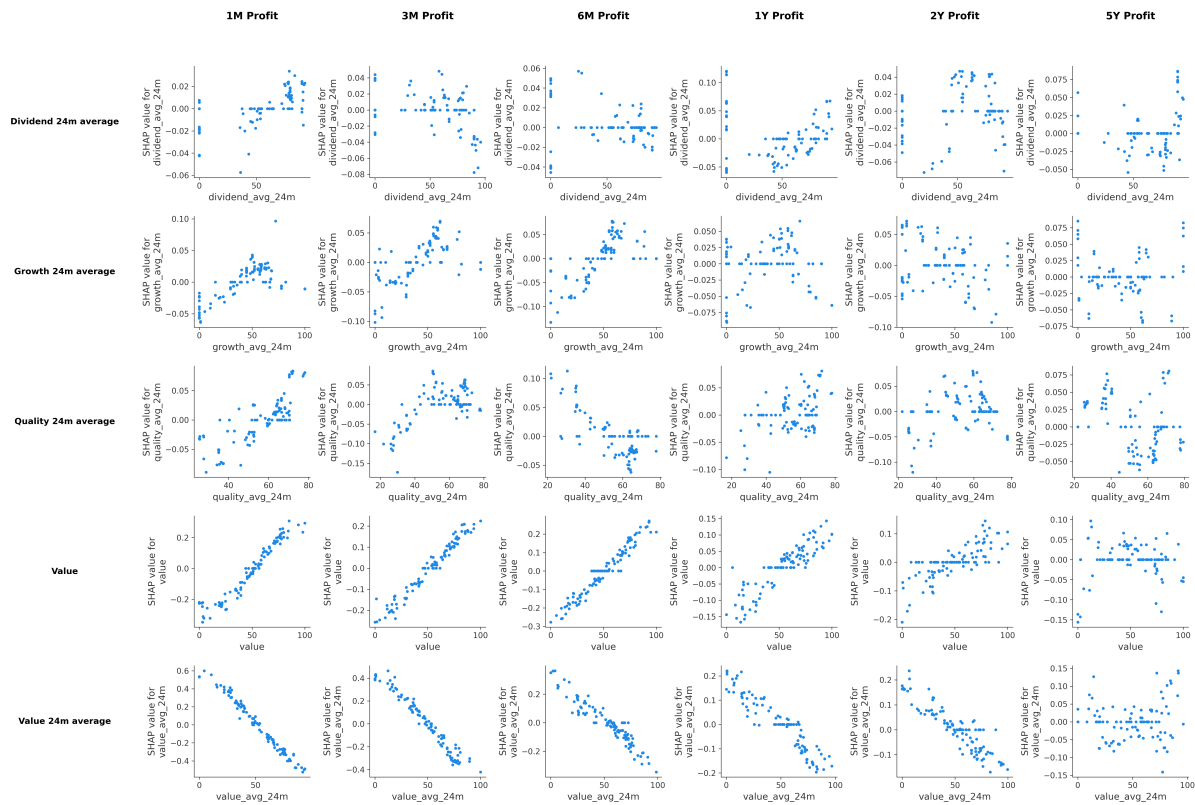


Figure 4.54: Big Data Future IF Balanced - Binary SHAP Consistency

4.4.3 Encoder Classifier

Continuing with a more sophisticated approach, we applied an investor-oriented encoding of the profits. To recall the coding scheme, the profits were classified into six distinct categories:

- **Very Bad:** -infinity to -10%
- **Bad:** -10% to -4%
- **Quite Bad:** -4% to 0%
- **Neutral:** 0% to 4%, mimicking keeping the money in the bank.
- **Normal:** 4% to 10%, this is what we would expect from investing in a diversified index like the S&P 500.
- **Very Good:** 10% to infinity,

It is important to emphasize the need for a balanced dataset when training a classifier. Annualizing the profits was a crucial step to mitigate bias, since longer investment horizons naturally exhibit more pronounced positive tails, as discussed in the descriptive analysis.

Performance

Having six classes means that complete noise should obtain a 16% of performance and 64% of error rate. But the results in figure 4.55 are remarkably good, Tweenvest's Scores once again

allowed us to achieve predictive performance above random chance:

- For unknown stocks the worst predictability is 15% and the best is 46%, with error rates of 85% and 51% respectively.
- For known stocks the worst predictability is 38% and the best is 51%, with error rates of 62% and 47% respectively.
- Again Value models are the best predictors overall.
- Quality models are still the second best predictors for both known and unknown stocks.

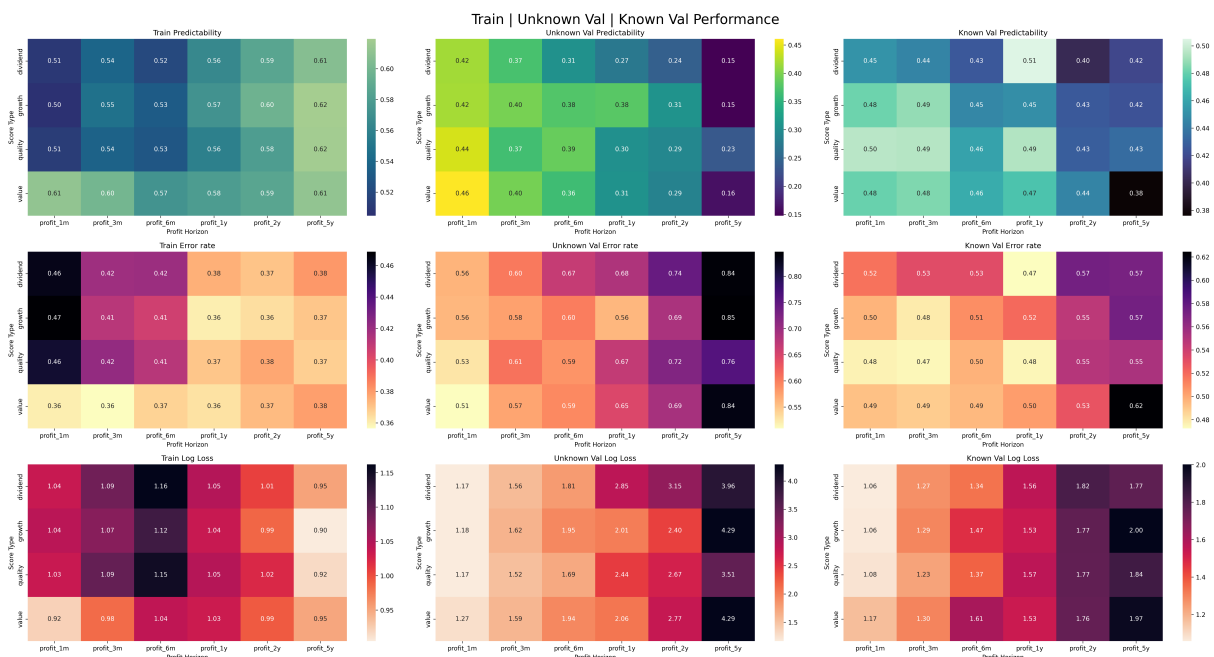


Figure 4.55: Big Data Future IF Balanced - Encoder Classifier Performance

In this case, the model has higher predictabilities than the expected error rates for the known stocks. But we have to make clear that, even though these results are very good for our purposes, **they are far away to be used as a real investment automation** since their log-loss is higher than we should have for a safe decision.

SHAP-Consistency

If we look at the results of the SHAP plots in figure 7.35 and appendix figure 4.56, we can see that these results are similar to the binary classifier, with less consistency except for the Value scores. This could be due to ability of the model to overfit the data, losing some of the interpretability; but as we have also seen, the model does perform well for unknown stocks so we cannot assure it is due to the overfitting.



Figure 4.56: Big Data Future IF Balanced - Encoder SHAP Consistency

4.4.4 Regressor

Finally, after reaching such good performances with the classifiers we aimed to determine if we could improve the results of the models by using a regressor NN.

Performance

Unfortunately, the results obtained when attempting to predict exact profit values are poor. This outcome is not unexpected, as forecasting a continuous variable such as stock returns with *simple models*—including standard NNs—is inherently challenging. Indeed, there is an entire field, known as *quantitative finance*, devoted to this problem, employing sophisticated models that incorporate differential equations and other advanced techniques in algorithmic trading to achieve even a modest 1% advantage over the market.

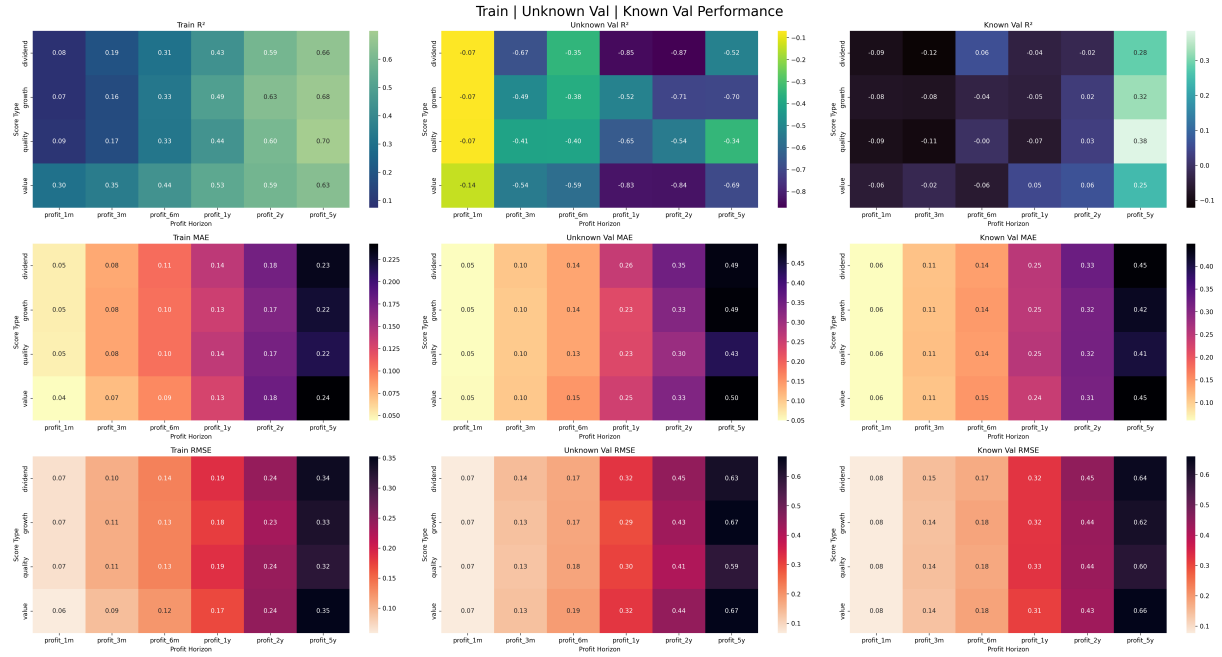


Figure 4.57: Big Data Future IF Balanced - Regressor Performance

SHAP-Consistency

Knowing this, we can see in figures 7.36 and appendix figure 4.58 that the SHAP plots reveal complete noise.

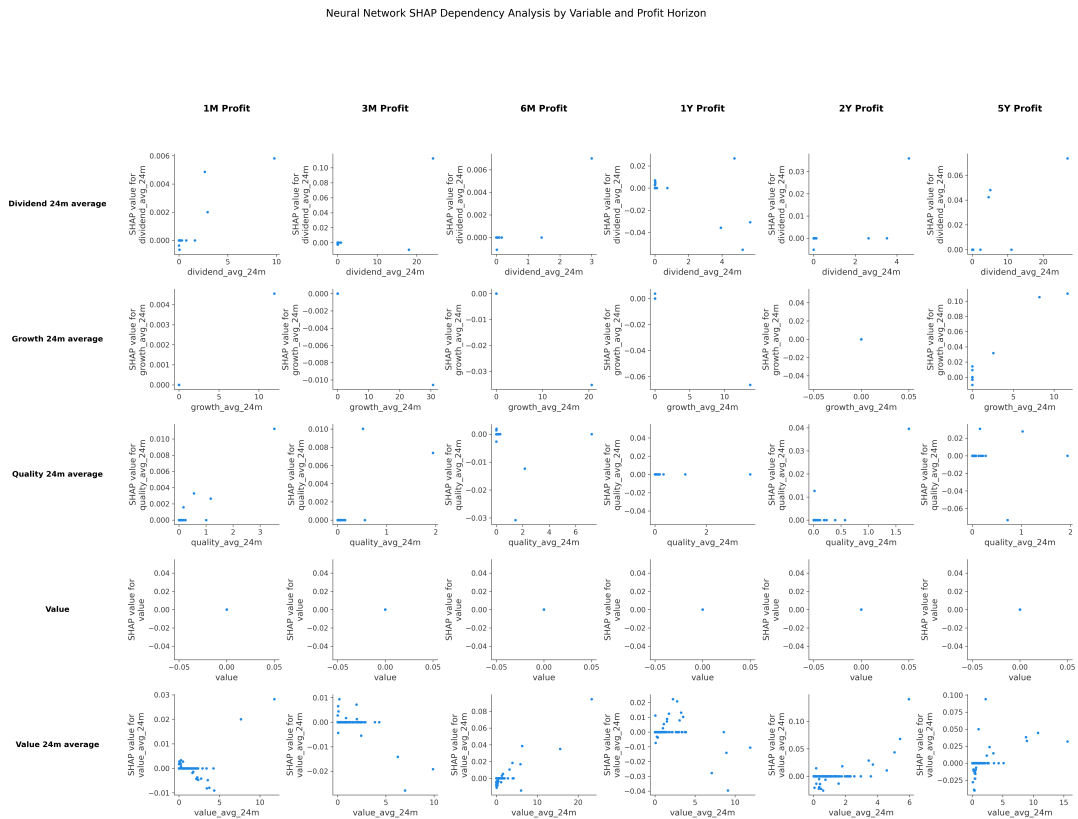


Figure 4.58: Big Data Future IF Balanced - Regressor SHAP Consistency

Chapter 5

Conclusion and Discussion

Throughout this project, we have explored many different avenues. We began by gaining an in-depth understanding of Tweenvest’s internal operations and the methodology behind their score calculations. At that stage, it became necessary to design a method for computing and storing historical scores within a production environment. To achieve this, we worked with Jobs, Crons, APIs, telemetry, data pipelines, and server deployment. The result was a robust work environment capable of calculating and integrating all required data, which also served as a platform for running all model-fitting simulations.

From there, we moved on to understanding and interpreting the data. We identified key patterns and developed a comprehensive outlier detection system. We then explored multiple approaches using various models and datasets to investigate potential correlations between the explanatory variables and future stock profits. Finally, when analyzing the results, we observed a series of recurring behaviors. In this closing chapter, we will summarize our conclusions, discuss the study’s limitations, and provide recommendations for both Tweenvest and its users.

5.1 Conclusions

The main goal of this work was to **evaluate if the Factor Scores represented *alpha* for different time frames, guide users** how to interpret them, **and propose improvements**.

5.1.1 Clear Statements

During the analysis we have identified a series of common patterns with enough consistency, significance, and coherence to be taken into account.

Investment Strategies

Firstly, we want to emphasize some well-known facts from the financial literature that we have confirmed with multiple methods in this work:

- **Short-term or Trading.** On the short-term, the market is mostly random brownian movements, making this type of investment very difficult since it cannot be analyzed from the financial perspective. We have seen this in both descriptive analysis of the distributions and the models capabilities to predict.

- **Long-term or Investing.** Long-term investing has been shown to be closely linked to economic growth and, on average, **yields higher returns** than short or medium-term strategies. This pattern is also evident in our profit distribution analysis and in the models' ability to generate more accurate predictions over longer horizons.
- **Variance and Risk:** As we have just said, most of the models could only predict something for longer horizons, but these **long-term predictions came with higher variances** for most of the models. When trying to predict future profits, it is more probable to miss by a lot on the exact number, but also it is more probable to obtain better returns.
- **Market Variability:** it is trivial that each market has its own characteristics. So before making any investment decision, it is important to understand the market since **the analysis of a company can differ a lot depending on the region, sector, and other factors**. We have seen this in the amount of outliers, predictabilities and the different effects of the variables per region.

Profits

Secondly, one of the clearest findings is that **profits exhibit inertia**, meaning they depend on prior performance. This is evident in:

- **Simple correlations.** Past profits show a stronger correlation with future profits than would be expected solely from the portion of overlapping data. For example, although an annual return includes 50% of the data from the corresponding semestral returns, the correlation between them is even higher. This pattern was observed consistently across all datasets and time frames.
- **Pair Plots.** As discussed in chapter 4, short-term profit horizons display clearly visible linear relationships in the plots.
- **ACFs and PACFs.** The autocorrelation functions show a clear pattern of high dependency for all of the profits.
- **ARIMA models.** The ARIMA models demonstrated a remarkable improvement in both predictability and accuracy, which was consistent with the autocorrelation patterns previously observed. However, these models could not be applied in our final analysis.

Scores

Interpreting and understanding the scores and their relationships is crucial for the users to make the best decisions. First of all, we have checked that **the four scores have a mild correlation** between them, so we should expect that if the company is doing well in one of them, it could do well in the others. But what is very interesting is that **Quality and Dividend are very correlated**, companies that do well financially tend to have a good dividend yield.

In terms of their effect on the future profits, we have strong evidence that on the data we have, **average scores have a significant effect on the future profits**, highlighting:

- **Quality's 24-months average** has the **second highest positive effect** to future profits. This relationship happens **for most of the datasets, models and timeframes**.
- **Value's 24-months average** is the only score **with a negative effect** to future profits. This relationship occurs **in all of the datasets, models and timeframes**.

- **Value's last score** has the **highest positive effect** to future profits. This relationship occurs in **all of the datasets, models and timeframes**.
- **Growth's 24-month average** may have a **modest positive effect** on future profits. While this effect is neither as strong nor as statistically significant as those observed for other variables, it has appeared consistently across multiple analyses.

Furthermore, our analysis indicates that **accumulated past profits influence current scores**, leading to:

- **Decreases in the Value score.** When a company is in a bull market, its multiples will increase, making the Value score to decrease. So when interpreting the Value score, it is interesting to see how has the company been performing; to therefore use this score for interpreting or to ignore it. This happens for companies with a very strong growth in popularity and brand strength such as Apple, NVIDIA, etc.
- **Increases in the Growth score.** If a company rises a lot in the stock market it could mean that the market has anticipated future financial growth.

5.1.2 Non-Clear Results

Handling with real-world data is always a challenge, and more when the nature of it is mostly random. We have tried to apply different modelings, outliers thresholds and techniques, statistical tests and coherent protocols to not fall into the trap of *confirmation bias*. Doing so has allowed us to get some very valuable insights; but there were some cases where we could not get a clear answer to our questions.

Interpretation Intervals

Tweenvest scores are not a numerical methodology to use as a single indicator to invest, each company and market has its own characteristics, and needs to be meticulously analyzed. But for simplifying the process, the platform includes a series of interpretation intervals to guide the users on when to start taking into account the scores. These intervals are reasonable enough to be used as a starting point, but we could not identify a clear pattern between them to assign a specific meaning other than the distributions of the scores that show when the scores start to discriminate more companies.

White Noise Test

When analyzing the models' residues, we have mostly used the visual inspection because it was almost impossible to obtain any insights from the statistical tests. To be clear, the only tests that passed for the best models –GAMs and NN-Regressors– were: the mean T-test and the Breusch-Pagan Test, with a 25% of tolerance.

5.2 Discussion

5.2.1 Limitations of the study

All of the results presented here have been obtained with a couple of limitations:

- **Data Availability.** We have only used the data available in the database, which was limited to 2012-2025. Were we only tested 2020-2025 profits based on 2015-2020 financial data.
- **Outliers.** Even taking all the cautions possible, we had to leave out many outliers from the data analysis due to their effect on the results. This could be a problem for the inference, making the results not to be reliable for all of the real companies.
- **Residues.** The linear regression models are only valid when the data is linear, and the residues follow the white noise assumption. We have gotten close to this by the final datasets and models, but we could not achieve complete homoscedasticity, and it was also impossible to not have autocorrelation due to the nature of the data.
- **Overfitting.** As we have explained, the GAM models tend to overfit the data, this makes the predictability to be higher when training but useless for validation. But luckily, we were able to obtain good interpretations from the splines.

5.2.2 Future work

There has been multiple occasions where we have found an interesting hypothesis to investigate further, but due to the scope of this project, the time and proper knowledge of the subject, we had to leave them aside.

Hyper-parameter Optimization

In the process of improving the MCODE and its equivalent ICD, we identified a potential opportunity to better distinguish between random extreme cases and consistently extreme behaviors. In this work, the hyperparameters were approximated through a combination of trial runs and a conceptual understanding of each method's underlying principles:

1. **SVM** creates geometrical hyper-dimensional frontiers to form a safe hyper-volume of data clusters. So the intuition tells us that this method should probably have a **higher tolerance** than the rest.
2. **IF** looks for the most isolated points following a random forest approach. It is good for detecting those points that seem to make not much sense, so its **tolerance should not be as high**.
3. **LOF** searches for clusters of high density points, not taking account the distance between the clusters. So this method is very good for detecting extreme phenomena that could be consistent. Its **tolerance could be between the other methods**.

But this is just an idea, and for future developments of these type of techniques it should be study in depth and tested with rigorous tests.

Extreme-cases Modeling

In financial data, achieving a truly *normal* distribution is uncommon. To simplify this project, outliers were excluded from the main analysis; however, these extreme cases often contain valuable information about companies that perform exceptionally well or poorly. We propose a dedicated analysis of the removed outliers to investigate potential relationships with the scores. For instance, companies with severe stock market losses may systematically exhibit low Quality scores, while top performers might be associated with high Growth scores.

Other Factors

In this work, we chose to only analyze four of the available factors: Quality, Growth, Value, and Dividend; we chose these because they are the most relevant in the platform, since they also have their own separate sections. But there are still many other already developed by tweenvest that should be analyzed, such as we see in figure 5.1.

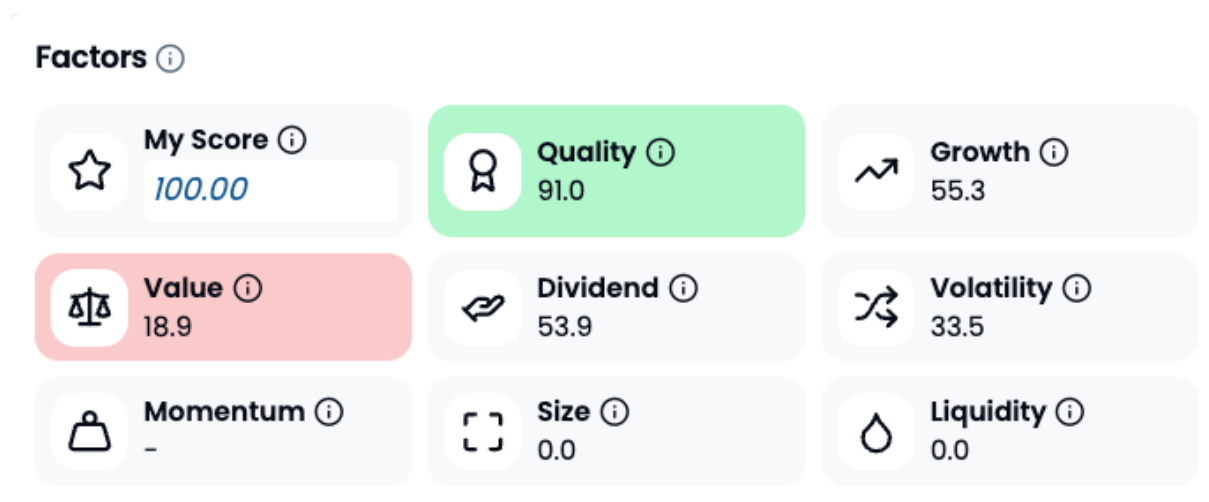


Figure 5.1: Tweenvest Scores Overview

Multivariate Analysis

To obtain more interpretable results, we have analyzed the models separately by scores and profits. However, it would be valuable to perform a multivariate analysis that incorporates all factors along with the other explanatory variables, in order to assess their combined contribution to future profits. Methods such as Principal Component Analysis (PCA) or Partial Least Squares (PLS) could be applied for this purpose.

Bayesian Approach

Throughout this work we have observed that **patterns and distributions can change over time**: factor effects drift, and relationships differ by market and period. This weakens the assumptions behind static models, so a Bayesian approach could be especially interesting here because it is designed to **learn under non-stationarity**.

5.2.3 Recommendations to Tweenvest

Finally, to conclude the work, we want to give some recommendations for future development of the platform:

- **Historical Scores.** We have seen that most of the scores alone are not enough to get clear insights. But if we combine them with the historical profits, we can get a better understanding of the company and the future profits. So we propose to, first **calculate all of the past scores** and then include the possibility to either see their time series, or just **include the 24 months average** along the current score.

- **Momentum development.** It is undeniable that **momentum is a very important factor** to predict future profits, actually the most relevant we have seen so far. So we propose to take back a look at it and include it again in the platform.
- **User Guide.** The factor score methodology is very interesting and definitely gives Tweenvest a true differentiator in the market, but investors tend to be very skeptical. It could be interesting to have a section for users to dive into the methodology so they can understand it better and trust the scores more. Also, adding a small guide in each section for interpretation could give a lot of Value and increase their use.

References and Bibliography

- 1 Dorsey, P., & Mansueto, J. (2011). *The five rules for successful stock investing*.
- 2 Sharpe, W. F. (1964, September). *Capital asset prices: A theory of market equilibrium under conditions of risk* (Vol. 19).
- 3 Wang, X., Xiang, Z., Xu, W., & Yuan, P. (2022). The causal relationship between social media sentiment and stock return: Experimental evidence from an online message forum. *Economics Letters*, 216, 110598. <https://doi.org/https://doi.org/10.1016/j.econlet.2022.110598>
- 4 MSCI. (2025). Fundamental data methodology. <https://www-cdn.msci.com/documents/10199/0b41b3a3-9da8-46f6-fce7-1dc1177e4a23>
- 5 pandas development team, T. (2025). Pandas documentation [URL: <https://pandas.pydata.org/>].
- 6 Hunter, J. D., Droettboom, M., Caswell, T. A., Firing, E., Dale, D., et al. (2025). Matplotlib: Python plotting package [URL: <https://matplotlib.org/stable/>].
- 7 Waskom, M. L. (2025). Seaborn: Statistical data visualization [URL: <https://seaborn.pydata.org/index.html>].
- 8 scikit-learn developers. (2025a). Oneclasssvm [URL: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html>].
- 9 Zineb, N., Honeine, P., & Richard, C. (2012). On simple one-class classification methods. *IEEE International Symposium on Information Theory Proceedings*. https://www.researchgate.net/publication/261248728_On_simple_one-class_classification_methods
- 10 Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1), 3. <https://www.lamda.nju.edu.cn/publication/tkdd11.pdf>
- 11 Dairi, A., Harrou, F., & Sun, Y. (2022). Efficient driver drunk detection by sensors: A manifold learning-based anomaly detector. *IEEE Access*, PP, 1–1. <https://doi.org/10.1109/ACCESS.2022.3221145>
- 12 Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: Identifying density-based local outliers. *ACM sigmod record*. <https://dl.acm.org/doi/pdf/10.1145/335191.335388>
- 13 Abro, A. A., Taşci, E., & Uğur, A. (2020). A stacking-based ensemble learning method for outlier detection. *Balkan Journal of Electrical & Computer Engineering*, 8(2). <https://dergipark.org.tr/en/download/article-file/1106223>
- 14 Servén, D., & Brummitt, C. (2018). Generalized additive models [URL: https://pygam.readthedocs.io/en/latest/notebooks/tour_of_pygam.html].
- 15 scikit-learn developers. (2025b). Mlpregressor [URL: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html#sklearn.neural_network.MLPRegressor].
- 16 Bank, W. (2025). Global inflation rates [URL: <https://www.worldbank.org/en/research/brief/inflation-database>].
- 17 S&P Dow Jones Indices. (2024). S&p 500 [URL: <https://www.spglobal.com/spdji/en/indices/equity/sp-500/#overview>].

- 18 scikit-learn developers. (2025c). Mlpclassifier [URL: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html].
- 19 Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*. <https://arxiv.org/abs/1705.07874>
- 20 Enders, W. (1948). *Applied econometric time series* [URL: <http://sggscclibrary.saraswatilib.com/UploadImg/637617646888135985.pdf>].
- 21 Docker, I. (2025). Docker compose documentation [URL: <https://docs.docker.com/compose/>].
- 22 Documentation, D. (2025a). Bulk_create method [URL: <https://docs.djangoproject.com/en/5.2/ref/models/querysets/#bulk-create>].
- 23 Documentation, D. (2025b). Select_related method [URL: https://docs.djangoproject.com/en/5.2/ref/models/querysets/#django.db.models.query.QuerySet.select_related].
- 24 European Money Markets Institute (EMMI). (2025). Euribor 1 month interest rate [URL: <https://www.global-rates.com/en/interest-rates/euribor/1/euribor-interest-1-month/>].
- 25 Chen, X., Li, B., Singh, T., & Worthington, A. (2024). Economic policy uncertainty and the predictability of stock returns: Impact of china in asia. *China Accounting and Finance Review*, 26(5), 645–679. <https://doi.org/10.1108/CAFR-11-2023-0144>
- 26 Liu, Q., & Tian, G. (2012). Controlling shareholder, expropriations and firm's leverage decision: Evidence from chinese non-tradable share reform [Special Section: Contemporary corporate finance research on South America]. *Journal of Corporate Finance*, 18(4), 782–803. <https://www.sciencedirect.com/science/article/pii/S0929119912000533>
- 27 Smith, A., et al. (2025). Pmdarima: Arima estimators for python [URL: <https://pypi.org/project/pmdarima/>].
- 28 Runge, C. (1901). Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten. *Zeitschrift für Mathematik und Physik*, 46, 224–243.
- 29 Carslaw, H. S. (1930). *Introduction to the theory of fourier's series and integrals* (Third). Dover Publications Inc.

Part II

Appendix

Chapter 6

Code Listings

6.1 Historical Scores Job Implementation

The following code snippets shows the complete implementation of the historical scores job that was used to populate the database with historical factor scores data:

6.1.1 Enqueueing Historical Scores Job

```
1 @job("historical_scores", timeout="3h")
2 def enqueue_calculate_index_scores_task(
3     start_date: datetime.date,
4     end_date: datetime.date,
5     index_names: list[str] | None = None,
6     region_names: list[str] | None = None,
7     total_stocks: int | None = None,
8     seed: float | None = None,
9     daily: bool = False,
10 ):
11     logger.info(
12         "[H1] Starting enqueue_year_tasks from %s to %s",
13         start_date,
14         end_date,
15     )
16
17     days = get_range_from_dates(
18         start_date,
19         end_date,
20         delta=relativedelta(days=1) if daily else relativedelta(
21             months=1),
22     )
23     days.reverse()
24
25     if not index_names:
26         index_names = list(
27             Index.objects.exclude(name__in=[Index.INDEX_ALL, "Other"])
28             .only("name")
29             .values_list("name", flat=True))
```

```

30     logger.info("[H1] Enqueueing tasks for %s indexes", len(index_
        names))
31
32     stocks_per_index = (
33         int(total_stocks / len(index_names)) if total_stocks else
        None
34     )
35
36     stock_count = 0
37     total_stocks_dict = {}
38     for index_name in index_names:
39         stock_ids = get_stock_ids_for_index(
40             index_name,
41             stocks_per_index,
42             start_date,
43             end_date,
44             seed,
45             region_names,
46         )
47         if not stock_ids:
48             raise ValueError(
49                 f"[H1] No stocks found for index '{index_name}' after
                    {start_date}"
50             )
51         total_stocks_dict[index_name] = stock_ids
52         for day in days:
53             calculate_index_scores_task.delay(
54                 index_name=index_name,
55                 day=day,
56                 stock_ids=stock_ids,
57             )
58
59         stock_count += len(stock_ids)
60
61     logger.info(
62         "[H1] Enqueued days %s for stocks:%s .", len(days), stock_
            count
63     )
64
65     return {
66         "stocks_processed": stock_count,
67         "days": len(days),
68         "total_stocks": total_stocks_dict,
69     }

```

Listing 6.1: Enqueueing Historical Scores Job

This job function is responsible for:

- Accepting date ranges and configuration parameters for historical score calculation
- Generating a list of dates to process (either daily or monthly intervals)
- Retrieving stock IDs for each index based on the specified criteria
- Enqueueing individual calculation tasks for each day and index combination
- Providing detailed logging and return statistics about the processing

6.1.2 Calculate Index Scores Data

```

1
2 @job("historical_scores", timeout="45m")
3 def calculate_index_scores_task(
4     index_name: str,
5     day: datetime.date,
6     stock_ids: list[int],
7 ):
8     start_time = time.perf_counter()
9     logger.info(
10         "[H2] Calculating index",
11         extra={
12             "index": index_name,
13             "day": day,
14             "stocks": len(stock_ids),
15         },
16     )
17
18     # ----- Index calculation ----- #
19     index_stats = get_indexes_information(
20         index_names=[index_name], calculation_date=day
21     )
22     index_quality = get_quality_ratios_by_index(
23         index_names=[index_name], calculation_date=day
24     )
25     stats = index_stats.get(index_name)
26     quality = index_quality.get(index_name)
27
28     # ----- #
29     if (
30         not stats
31         or not quality
32         or all(value is None for value in quality.values())
33         or all(
34             value is np.nan
35             for stat in stats.values()
36             for value in stat.values()
37         )
38     ):
39         raise RuntimeError(
40             f"[H2] Missing index data for {index_name} on {day},
41             stats: {stats}, quality: {quality}"
42         )
43
44     calculate_stocks_scores_task.delay(
45         day=day,
46         stats=stats,
47         quality=quality,
48         stock_ids=stock_ids,
49     )
50     return {
51         "index_calculated": index_name,
52         "total_time": time.perf_counter() - start_time,
53     }

```

Listing 6.2: Calculate Index Scores Data

This job function is responsible for:

- Calculating all the necessary index stats data
- Enqueuing the calculation of the stocks scores for the given index and date
- Providing detailed logging and return statistics about the processing

6.1.3 Calculate Stocks Scores Task

```

1 @tracer.start_as_current_span("calculate_stocks_scores_task")
2 @job("historical_scores", timeout="4h")
3 def calculate_stocks_scores_task(
4     day: datetime.date,
5     stats: dict,
6     quality: dict,
7     stock_ids: list[int],
8 ):
9     """
10    Job 3: For each stock, calculate scores and bulk save.
11    """
12    start_time = time.perf_counter()
13    logger.info(
14        "[H3] Calculating scores",
15        extra={"day": day, "stocks": len(stock_ids)},
16    )
17
18    if not stock_ids:
19        return {
20            "stocks_processed": 0,
21        }
22    # Select related for trying to avoid multiple queries to the DB
23    stocks = Stock.objects.filter(id__in=stock_ids).select_related(
24        "share_type",
25        "share_type__company",
26        "share_type__company__industry",
27        "share_type__company__industry__industry_group",
28        "share_type__company__industry__industry_group__sector",
29    )
30    factor_scores = []
31    for stock in stocks:
32        score = calculate_scores(
33            stock,
34            stats,
35            quality,
36            calculation_date=day,
37            bulk_create_or_update=True,
38        )
39        if not score:
40            logger.warning(
41                "[H3] No score for calculated",
42                extra={"stock_id": stock.pk, "day": day, "score":
43                    score},
44            )
45            continue # Not as important as the others, need to check
46                      # errors in DataDog
47        factor_scores.append(score)

```

```

46
47     bulk_create_or_update(
48         model=FactorScore,
49         object_list=factor_scores,
50         unique_fields=["stock", "date"],
51         update_fields=["quality", "growth", "value", "dividend"],
52     )
53
54     return {
55         "stocks_processed": len(stock_ids),
56         "total_time": time.perf_counter() - start_time,
57         "time_per_stock": (time.perf_counter() - start_time) / len(
58             stock_ids),
59     }

```

Listing 6.3: Calculate Stocks Scores Task

This job function is responsible for:

- Calculating the factor scores for each stock in the given index and date
- Bulk saving the factor scores for the given index and date
- Providing detailed logging and return statistics about the processing

6.2 Workers Management

In this section we have 2 different configurations for the workers management, since we had to switch servers.

6.2.1 Tweenvest's Production Server

```

1 [program:worker1]
2 command=python manage.py rqworker internal_tasks ciq_priority fmp_
   priority ciq indices_partial indices_full historical_scores fmp_
   emails --with-scheduler
3 directory=/app
4 autostart=true
5 autorestart=true
6 redirect_stderr=true
7 stdout_logfile=/dev/fd/1
8 stdout_logfile_maxbytes=0
9 environment = SERVICE_NAME="tweenvest-worker"
10
11 [program:worker2]
12 command=python manage.py rqworker internal_tasks ciq_priority fmp_
   priority ciq indices_partial indices_full historical_scores fmp_
   emails --with-scheduler
13 autostart=true
14 directory=/app
15 autorestart=true
16 redirect_stderr=true
17 stdout_logfile=/dev/fd/1
18 stdout_logfile_maxbytes=0
19 environment = SERVICE_NAME="tweenvest-worker"

```

```
20
21 [program:worker3]
22 command=python manage.py rqworker internal_tasks ciq_priority fmp_
    priority ciq indices_partial fmp emails historical_scores --with-
    scheduler
23 autostart=true
24 directory=/app
25 autorestart=true
26 redirect_stderr=true
27 stdout_logfile=/dev/fd/1
28 stdout_logfile_maxbytes=0
29 environment = SERVICE_NAME="tweenvest-worker"
30
31 [program:worker4]
32 command=python manage.py rqworker ciq_priority fmp_priority ciq
    indices_partial fmp emails historical_scores --with-scheduler
33 autostart=true
34 directory=/app
35 autorestart=true
36 redirect_stderr=true
37 stdout_logfile=/dev/fd/1
38 stdout_logfile_maxbytes=0
39 environment = SERVICE_NAME="tweenvest-worker"
40
41 [program:worker5]
42 command=python manage.py rqworker fmp_priority ciq_priority ciq fmp
    emails --with-scheduler
43 autostart=true
44 directory=/app
45 autorestart=true
46 redirect_stderr=true
47 stdout_logfile=/dev/fd/1
48 stdout_logfile_maxbytes=0
49 environment = SERVICE_NAME="tweenvest-worker"
50
51 [program:worker6]
52 command=python manage.py rqworker emails --with-scheduler
53 autostart=true
54 directory=/app
55 autorestart=true
56 redirect_stderr=true
57 stdout_logfile=/dev/fd/1
58 stdout_logfile_maxbytes=0
59 environment = SERVICE_NAME="tweenvest-worker"
60
61 [program:worker7]
62 command=python manage.py rqworker ciq_priority ciq emails summaries
    --with-scheduler
63 directory=/app
64 autostart=true
65 autorestart=true
66 redirect_stderr=true
67 stdout_logfile=/dev/fd/1
68 stdout_logfile_maxbytes=0
69 environment = SERVICE_NAME="tweenvest-worker"
70
71 [program:worker8]
```

```

72 command=python manage.py rqworker ciq_priority ciq emails summaries
   --with-scheduler
73 directory=/app
74 autostart=true
75 autorestart=true
76 redirect_stderr=true
77 stdout_logfile=/dev/fd/1
78 stdout_logfile_maxbytes=0
79 environment = SERVICE_NAME="tweenvest-worker"
80
81 [program:worker9]
82 command=python manage.py rqworker historical_scores ciq_priority
   summaries ciq emails --with-scheduler
83 directory=/app
84 autostart=true
85 autorestart=true
86 redirect_stderr=true
87 stdout_logfile=/dev/fd/1
88 stdout_logfile_maxbytes=0
89 environment = SERVICE_NAME="tweenvest-worker"
90
91 [program:worker10]
92 command=python manage.py rqworker summaries ciq_priority fmp_priority
   ciq fmp emails --with-scheduler
93 directory=/app
94 autostart=true
95 autorestart=true
96 redirect_stderr=true
97 stdout_logfile=/dev/fd/1
98 stdout_logfile_maxbytes=0
99 environment = SERVICE_NAME="tweenvest-worker"

```

Listing 6.4: Production Workers Management

6.2.2 Personal Development Server

```

1 [program:worker1]
2 command=python manage.py rqworker historical_scores
3 process_name=%(program_name)s_%(process_num)02d
4 numprocs=16
5 directory=/app
6 autostart=true
7 autorestart=true
8 redirect_stderr=true
9 stdout_logfile=/dev/fd/1
10 stdout_logfile_maxbytes=0
11 environment = SERVICE_NAME="tweenvest-worker"

```

Listing 6.5: Hetzner Workers Management

6.3 Data Export Job Implementation

The following code shows the complete implementation of the data export job that was used to create the final dataset, including the price estimation for missing values due to weekends and

holidays:

```

1 def export_factors_and_pricehistory_task(
2     stocks_id: list[int],
3     start_date: datetime.date,
4     end_date: datetime.date,
5     export_name: str,
6     daily: bool = False,
7 ):
8     """
9     Export FactorScore, PriceHistory y rentabilidades a CSV sin
10    cargar todo en memoria.
11    """
12    start_time = time.perf_counter()
13    print("Starting CSV export...")
14
15    periods = {
16        "profit_1m": relativedelta(months=1),
17        "profit_3m": relativedelta(months=3),
18        "profit_6m": relativedelta(months=6),
19        "profit_1y": relativedelta(years=1),
20        "profit_2y": relativedelta(years=2),
21        "profit_5y": relativedelta(years=5),
22    }
23
24    factor_fields = [
25        "stock__ticker",
26        "stock__share_type__company__name",
27        "stock__share_type__company__industry__industry_group__sector__name",
28        "stock__share_type__company__country__region",
29        "date",
30        "quality",
31        "growth",
32        "value",
33        "dividend",
34    ]
35
36    price_fields = ["market_cap_usd", "volume"]
37
38    days = get_range_from_dates(
39        start_date,
40        end_date,
41        delta=relativedelta(days=1) if daily else relativedelta(
42            months=1),
43    )
44    days.reverse()
45
46    profitability_fields = list(periods.keys())
47    export_fields = factor_fields + price_fields + profitability_fields
48
49    missing_prices = 0
50    missing_factors = 0
51    fs_path = f"api/factors/data_exports/{export_name}.csv"
52    with open(fs_path, mode="w", newline="", encoding="utf-8") as f:
53        writer = csv.DictWriter(f, fieldnames=export_fields)
54        writer.writeheader()
55
56        for stock_id in tqdm(stocks_id, desc="Stocks"):

```

```

53
54     price_qs = PriceHistory.objects.filter(
55         stock__id=stock_id,
56         date__range=(
57             start_date,
58             end_date + relativedelta(years=5),
59         ),
60     ).values("date", "close_price", "market_cap", "volume", "
        fx_mult")
61     price_lookup_stock = {
62         (stock_id, ph["date"]): ph
63         for ph in price_qs
64         if ph["fx_mult"] is not None
65     }
66
67     dividend_qs_stock = DividendHistory.objects.filter(
68         stock__id=stock_id,
69         ex_dividend_date__range=(start_date, end_date),
70     ).values("ex_dividend_date", "adjusted_dividend", "fx_
        mult")
71     dividend_lookup = {
72         (stock_id, dh["ex_dividend_date"]): (
73             float(dh["adjusted_dividend"]) * float(dh["fx_
74                 mult"])
75         )
76         for dh in dividend_qs_stock
77         if dh["fx_mult"] is not None
78     }
79     fs_qs_stock = (
80         FactorScore.objects.filter(stock_id=stock_id, date__
81             in=days)
82         .select_related(
83             "stock__share_type__company__industry__industry_
84                 group__sector",
85             "stock__share_type__company__country",
86         )
87         .values(*factor_fields)
88     )
89
90     fs_lookup = {(stock_id, fs["date"]): fs for fs in fs_qs_
91         stock}
92
93     for day in days:
94         factor = fs_lookup.get((stock_id, day))
95         if not factor:
96             missing_factors += 1
97             continue
98         day = factor["date"]
99         price_day = day
100         while price_day.weekday() >= 5:
101             price_day -= datetime.timedelta(days=1)
102
103         price = price_lookup_stock.get((stock_id, price_day))
104         if not price or price["close_price"] is None:
105             price = _calculate_existing_prices(
106                 price_day, price_lookup_stock, stock_id
107             )
108         if not price:

```

```

105         missing_prices += 1
106         continue
107
108     try:
109         market_cap_usd = (
110             float(price.get("market_cap", 0) or 0)
111             * price["fx_mult"]
112         )
113     except Exception:
114         market_cap_usd = None
115
116     try:
117         volume = (
118             float(price.get("volume", 0) or 0) * price["
119             fx_mult"]
120         )
121     except Exception:
122         volume = None
123
124     if price["close_price"] is None or price["fx_mult"]
125         is None:
126         continue
127     close_now_usd = float(price["close_price"]) * price["
128     fx_mult"]
129
130     profitabilities = {}
131     for field, delta in periods.items():
132         future_date = day + delta
133         while future_date.weekday() >= 5:
134             future_date -= datetime.timedelta(days=1)
135
136         future_price = price_lookup_stock.get(
137             (stock_id, future_date)
138         )
139         if not future_price or future_price["close_price"]
140             is None:
141             future_price = _calculate_existing_prices(
142                 future_date, price_lookup_stock, stock_id
143             )
144         if not future_price:
145             profitabilities[field] = None
146             missing_prices += 1
147             continue
148
149         close_future_usd = float(
150             future_price["close_price"]
151         ) * float(future_price["fx_mult"])
152         if close_future_usd == 0:
153             profitabilities[field] = None
154             missing_prices += 1
155             continue
156
157     # Dividend sum
158     dividend_sum = sum(
159         v
160         for (s_id, ex_div_date), v in dividend_lookup
161             .items()
162         if s_id == stock_id
163         and day <= ex_div_date < future_date
164     )

```

```

158         profitabilities[field] = (
159             close_future_usd + dividend_sum - close_now_
160             ) / close_now_usd
161
162         row = {**factor}
163         row["market_cap_usd"] = market_cap_usd
164         row["volume"] = volume
165         row.update(profitabilities)
166         writer.writerow(row)
167
168     elapsed = time.perf_counter() - start_time
169     print(f"Exportado a {fs_path} en {elapsed:.2f}s")
170     print("MISSING PRICES:", missing_prices)
171     print("MISSING FACTOR SCORES:", missing_factors)

```

Listing 6.6: Data Export Job Implementation

This job function is responsible for:

- Exporting factor scores, price history and profitability data to CSV without loading everything into memory
- Handling missing price data due to weekends and holidays by estimating prices from nearby dates
- Converting all monetary values to USD using appropriate exchange rates
- Calculating profitability for different time periods (1m, 3m, 6m, 1y, 2y, 5y)
- Including dividend payments in profitability calculations
- Tracking and reporting missing data points for both prices and factor scores

6.4 Custom Python Functions for Data Analysis

```

1 def correlation_matrix(
2     df: pd.DataFrame, columns: list[str], file_path: str | None =
3     None
4 ) -> None:
5     """
6     Plot a correlation matrix for the DataFrame showing only the
7     lower triangle.
8     """
9     df_columns = df[columns]
10    corr_matrix = df_columns.corr()
11
12    # Create a mask for the lower triangle
13    mask = np.triu(np.ones_like(corr_matrix, dtype=bool))
14
15    # Plot using seaborn for better visualization
16    plt.figure(figsize=(12, 8))
17    sns.heatmap(
18        corr_matrix,
19        mask=mask,
20        annot=True,

```



```

19     cmap="coolwarm",
20     center=0,
21     fmt=".2f",
22     square=True,
23 )
24 plt.title("Correlation Matrix of Numeric Variables")
25 plt.tight_layout()
26
27 if file_path:
28     plt.savefig(file_path)
29     plt.show()
30 else:
31     plt.show()

```

Listing 6.7: Correlation Matrix Plot Function

```

1 def pairplot(
2     df: pd.DataFrame,
3     file_path: str | None = None,
4     kde: bool = False,
5     hue: str = "region",
6     hue_order: list[str] | None = None,
7     palette: dict[str, str] | None = None,
8 ) -> None:
9
10     if len(df) > 10000:
11         df = df.sample(10000, random_state=42)
12
13     g = sns.pairplot(
14         df,
15         hue=hue,
16         diag_kind="kde",
17         hue_order=hue_order,
18         palette=palette,
19         corner=True,
20     )
21     if kde:
22         g.map_lower(sns.kdeplot, levels=4, color=".2")
23
24     if file_path:
25         plt.savefig(file_path)
26         plt.close()
27     else:
28         plt.show()

```

Listing 6.8: Custom Pairplot Function

6.5 Data Preprocessing

6.5.1 Data Cleaning

```

1 import sys
2 from pathlib import Path
3
4 # Add the parent directory (code/) to sys.path

```

```

5     sys.path.append(str(Path().resolve().parent))
6     from utils import load_data, plot_numeric_distributions
7     import pandas as pd
8
9     # %% [markdown]
10    # # Loading the data
11
12    # %%
13    file_name = "Small Data future AF"
14    data_path = f"../data/raw/{file_name}.csv"
15
16    # Read the CSV file
17    df = load_data(data_path)
18
19    # %% [markdown]
20    # ## Cleaning and renaming
21
22    # %%
23    df = df.rename(
24        columns={
25            "stock__share_type__company__industry__industry_group__
                sector__name": (
26                "sector"
27            ),
28            "stock__share_type__company__country__region": "region",
29            "market_cap_usd": "market_cap",
30            "stock__share_type__company__name": "company",
31            "stock__ticker": "ticker",
32        }
33    )
34
35    df["dividend"] = df["dividend"].fillna(0.0)
36    df["date"] = pd.to_datetime(df["date"])
37    df_no_nan = df.dropna()
38
39
40    # %% [markdown]
41    # # Analyzing
42
43    # %%
44    export_path = f"../data/cleaned/basic/plots/{file_name}.png"
45    plot_numeric_distributions(df, file_path=export_path)
46
47    # %% [markdown]
48    # # Export cleaned data
49
50    # %%
51    # Export the cleaned dataset to CSV
52    cleaned_data_path = f"../data/cleaned/basic/{file_name}.csv"
53    df.to_csv(cleaned_data_path, index=False)
54
55    print(f"Cleaned data exported to: {cleaned_data_path}")

```

Listing 6.9: Basic Data Cleaning

6.5.2 Outlier Detection

Listing 6.10: Outlier Detection

Chapter 7

Extra Images and Tables

7.1 Data

7.1.1 Columns and Variables

Response Variables	Explanatory Variables				Time	Extra
Profit 1 month	Scores	Dummies		Size	Date	Company Name
Profit 3 months	Quality	<i>Region</i>	<i>Sector</i>	Market cap		Stock Ticker
Profit 6 months	Growth	USA and Canada	Communication Services	Volume		
Profit 1 year	Value	Europe	Consumer Discretionary			
Profit 2 years	Dividend	Latin America	Consumer Staples			
Profit 5 years		Asia and Oceania	Energy			
		Africa	Financials			
			Health Care			
			Industrials			
			Information Techonology			
			Materials			
			Real State			
			Utilities			
6	4	1	1	2	1	2

Figure 7.1: Number of columns per dataset

7.1.2 Distributions

Here we have all of the not shown images for further analysis, if needed.

Scores

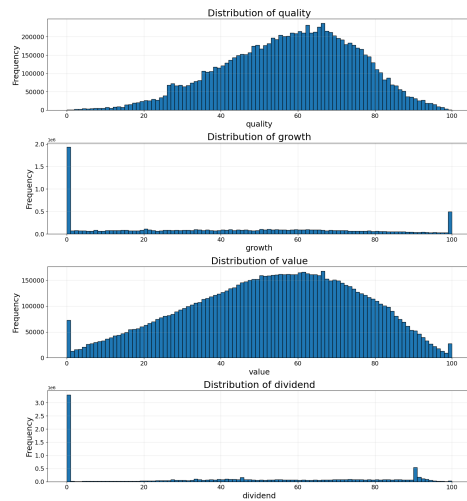


Figure 7.2: Big Data Past (IQR) - Scores

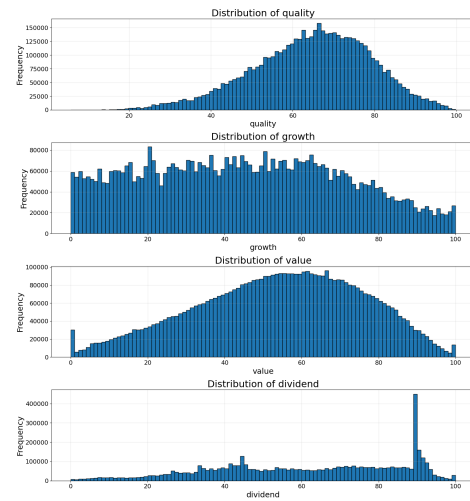


Figure 7.3: Big Data Past (IQR) - Scores Deep

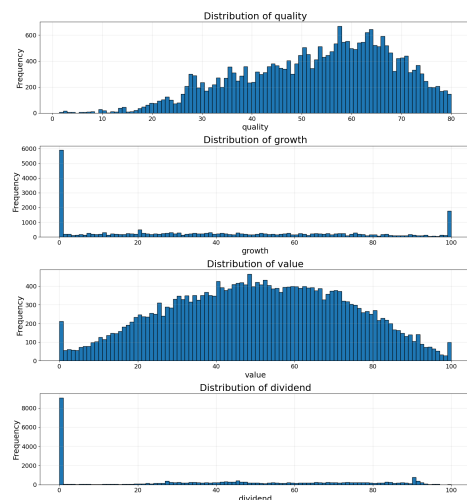


Figure 7.4: Small Data Future - Scores

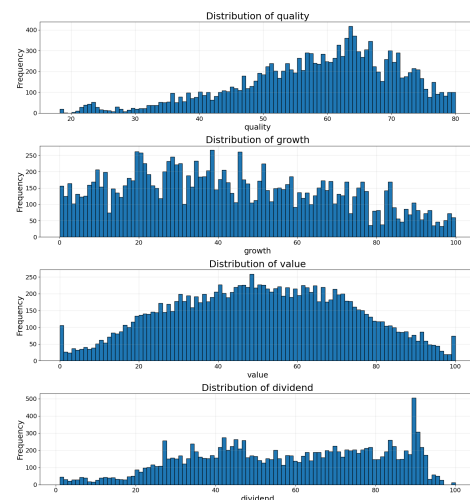


Figure 7.5: Small Data Future - Scores Deep

Dummies



Figure 7.6: Big Data Future - Dummies

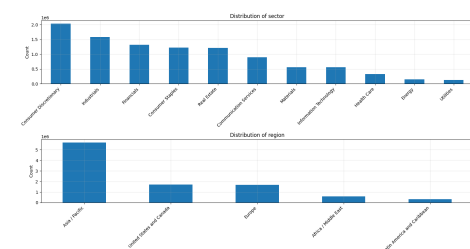


Figure 7.7: Big Data Past - Dummies

Profits

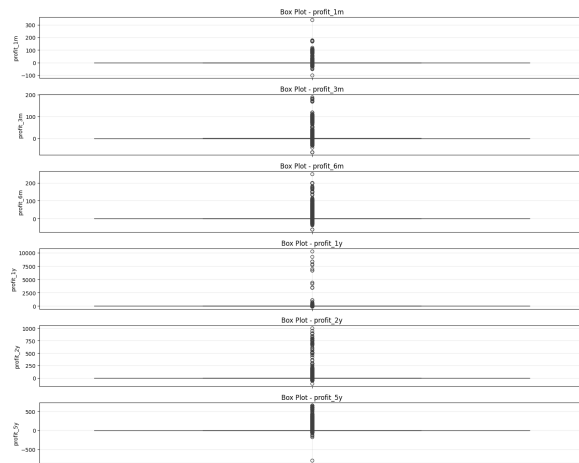


Figure 7.8: Big Data Future - Profits Boxplot

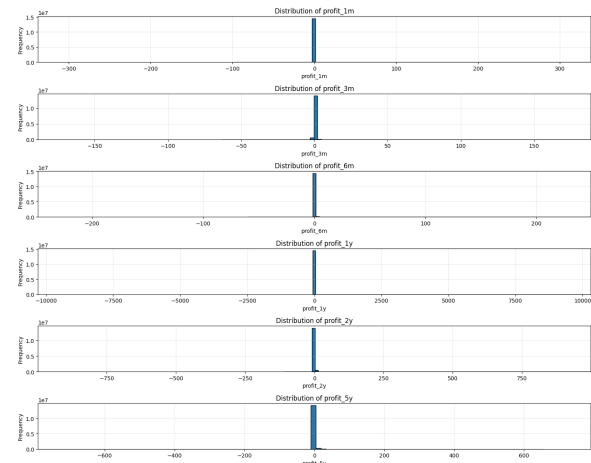


Figure 7.9: Big Data Future - Profits

7.1.3 Correlations

USA and Canada

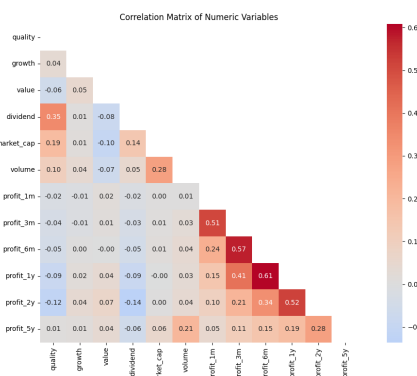


Figure 7.10: Small Data Future USA

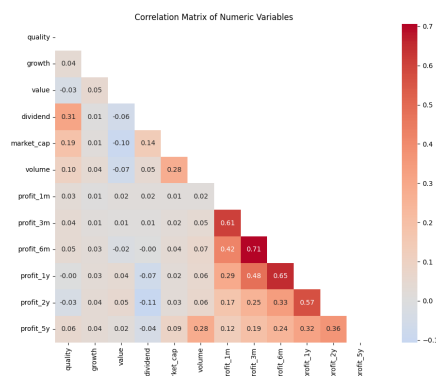


Figure 7.11: Small Data Future USA - MCODE

Europe

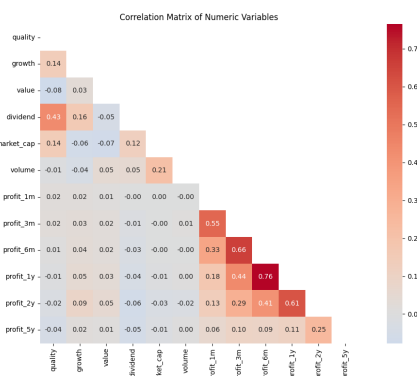


Figure 7.12: Small Data Future EU

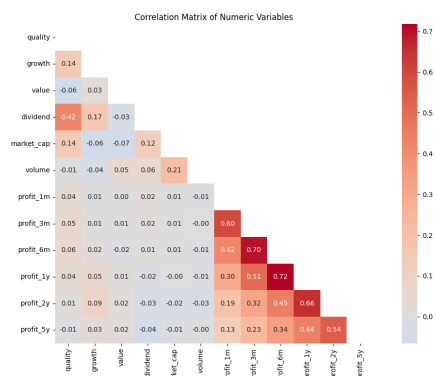


Figure 7.13: Small Data Future EU MCODE

Latin America

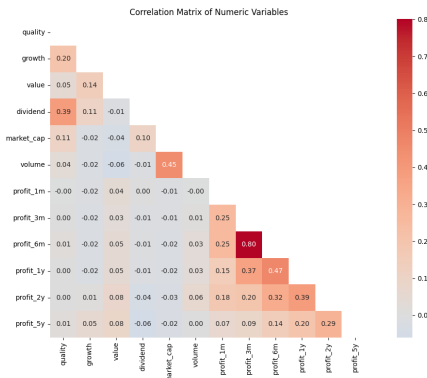


Figure 7.14: Small Data Future LAT

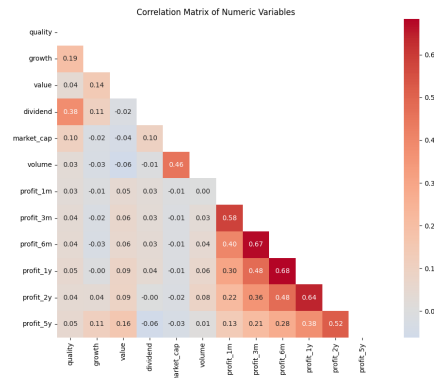


Figure 7.15: Small Data Future LAT - MCODE

Africa

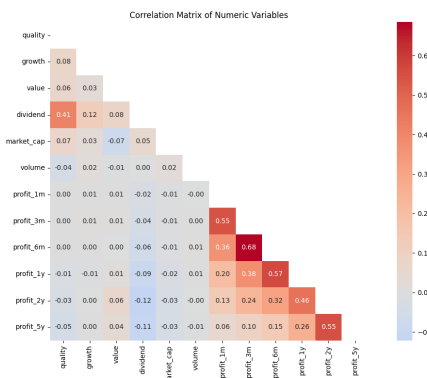


Figure 7.16: Small Data Future AF

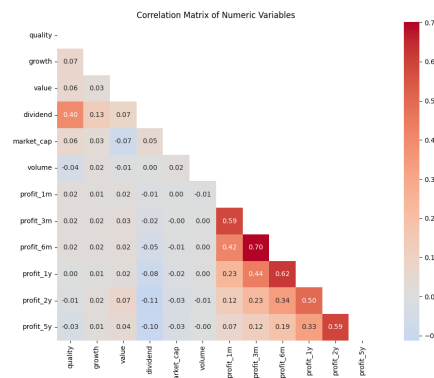


Figure 7.17: Small Data Future AF - MCODE

Asia and Oceania

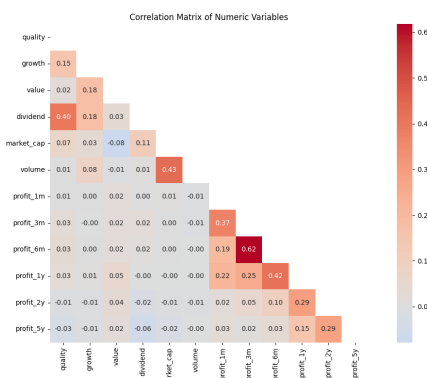


Figure 7.18: Small Data Future AS

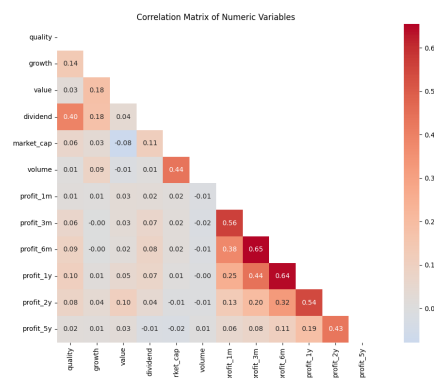


Figure 7.19: Small Data Future AS - MCODE

7.2 Models

7.2.1 Linear Regression

First Analysis

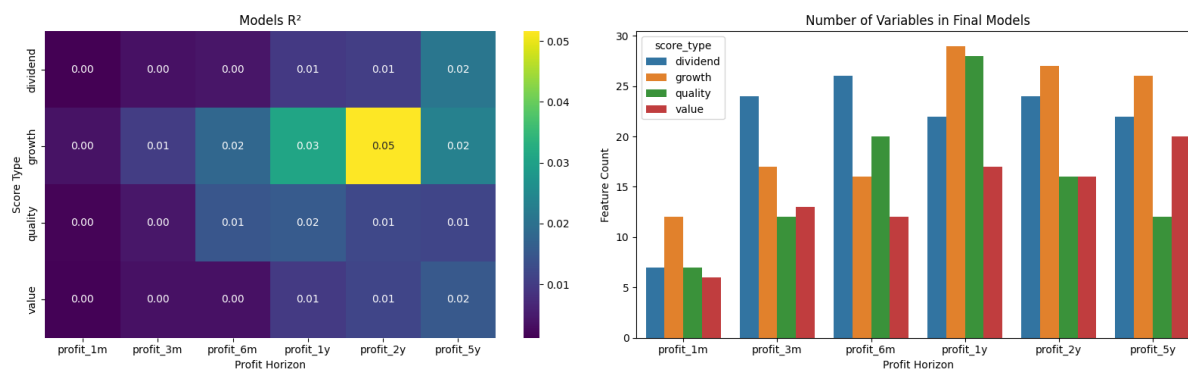


Figure 7.20: Small Data Future (Clean) - Performance

Table 7.1: Small Data Future *Yeo-Johnson* - MCO (4%) Consistency

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	—	—	—	—	-0.0011	0.0013
	Sign	0	0	0	0	—	+
	t-value	—	—	—	—	-2.06	2.65
Growth	Coef.	0.0031	—	—	0.0045	—	0.0008
	Sign	+	0	0	+	0	+
	t-value	7.35	—	—	7.79	—	3.11
Quality	Coef.	-0.0040	-0.0020	—	-0.0055	—	—
	Sign	—	—	0	—	0	0
	t-value	-3.95	-2.43	—	-5.53	—	—
Value	Coef.	—	-0.0023	—	-0.0021	—	—
	Sign	0	—	0	—	0	0
	t-value	—	-2.90	—	-1.92	—	—

Table 7.2: Small Data Future (Windowed) - MCOD (4%) Dividend Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	–	-0.011255	-0.005307	–	0.00692	–
	Sign	0	–	–	0	+	0
	t-value	–	-4.52	-3.64	–	3.63	–
Avg 3M	Coef.	-0.001144	0.006685	-0.008074	-0.014068	-0.008895	–
	Sign	–	+	–	–	–	0
	t-value	-1.54	2.52	-2.96	-10.32	-4.51	–
Avg 6M	Coef.	–	–	0.00638	–	–	-0.010871
	Sign	0	0	+	0	0	–
	t-value	–	–	2.27	–	–	-4.37
Avg 12M	Coef.	–	–	-0.004749	–	-0.022233	-0.018156
	Sign	0	0	–	0	–	–
	t-value	–	–	-1.92	–	-9.25	-6.05
Avg 24M	Coef.	0.00405	0.007274	0.015401	0.017175	0.026821	0.032814
	Sign	+	+	+	+	+	+
	t-value	5.09	6.97	7.27	13.30	13.01	11.94

Table 7.3: Small Data Future (Windowed) - MCOD (4%) Growth Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Growth	Coef.	–	-0.003412	–	–	–	–
	Sign	0	–	0	0	0	0
	t-value	–	-2.31	–	–	–	–
Avg 3M	Coef.	–	-0.006505	-0.013862	-0.001784	–	–
	Sign	0	–	–	–	0	0
	t-value	–	-1.96	-6.16	-1.74	–	–
Avg 6M	Coef.	–	0.01075	0.015124	–	–	0.003423
	Sign	0	+	+	0	0	+
	t-value	–	3.12	5.08	–	–	2.21
Avg 12M	Coef.	-0.005013	-0.009417	-0.01313	-0.014288	-0.014954	-0.015208
	Sign	–	–	–	–	–	–
	t-value	-3.17	-3.61	-5.27	-5.88	-9.01	-5.38
Avg 24M	Coef.	0.006224	0.011802	0.016234	0.021646	0.02099	0.013468
	Sign	+	+	+	+	+	+
	t-value	3.94	6.77	9.26	9.69	13.09	6.02

Table 7.4: Small Data Future (Windowed) - MCOB (4%) Quality Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Quality	Coef.	-0.026349	-0.033214	-0.044733	–	–	–
	Sign	–	–	–	0	0	0
	t-value	-4.12	-4.99	-8.72	–	–	–
Avg 3M	Coef.	0.026577	0.019478	–	-0.027661	–	–
	Sign	+	+	0	–	0	0
	t-value	2.40	1.56	–	-3.80	–	–
Avg 6M	Coef.	-0.037059	-0.074902	-0.060701	-0.11063	-0.035347	-0.029798
	Sign	–	–	–	–	–	–
	t-value	-3.90	-4.15	-8.03	-9.55	-4.13	-3.47
Avg 24M	Coef.	0.041583	0.05991	0.108683	0.135561	0.189188	0.072937
	Sign	+	+	+	+	+	+
	t-value	8.63	8.53	18.98	14.57	15.59	9.03

Table 7.5: Small Data Future (Windowed) - MCOB (4%) Value Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Value	Coef.	0.06537	0.055281	0.033066	0.020026	0.016281	0.015605
	Sign	+	+	+	+	+	+
	t-value	26.73	32.18	21.63	11.18	14.33	20.82
Avg 3M	Coef.	-0.077709	-0.024767	0.018085	0.00612	–	–
	Sign	–	–	+	+	0	0
	t-value	-24.06	-9.77	10.46	2.91	–	–
Avg 6M	Coef.	0.014356	-0.02986	-0.03336	0.023383	0.010252	–
	Sign	+	–	–	+	+	0
	t-value	6.43	-9.34	-12.64	7.09	5.01	–
Avg 12M	Coef.	-0.010381	-0.01438	-0.035998	-0.047533	0.020324	0.017316
	Sign	–	–	–	–	+	+
	t-value	-5.09	-5.38	-15.95	-14.20	7.23	6.62
Avg 24M	Coef.	–	–	–	-0.020006	-0.066239	-0.049306
	Sign	0	0	0	–	–	–
	t-value	–	–	–	-10.55	-25.46	-18.19

Geographical Analysis

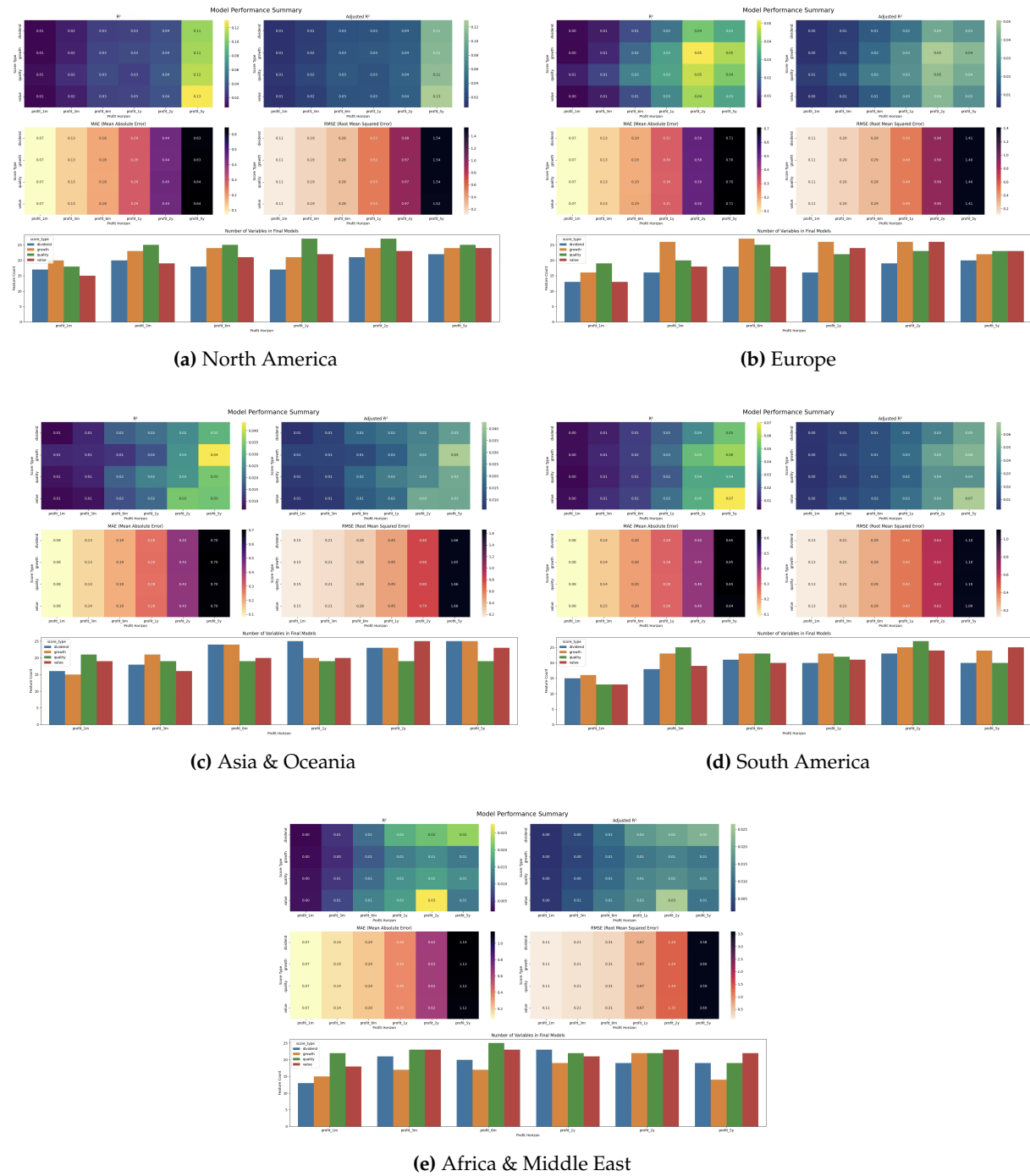


Figure 7.21: Geographical Predictability Analysis - First Model

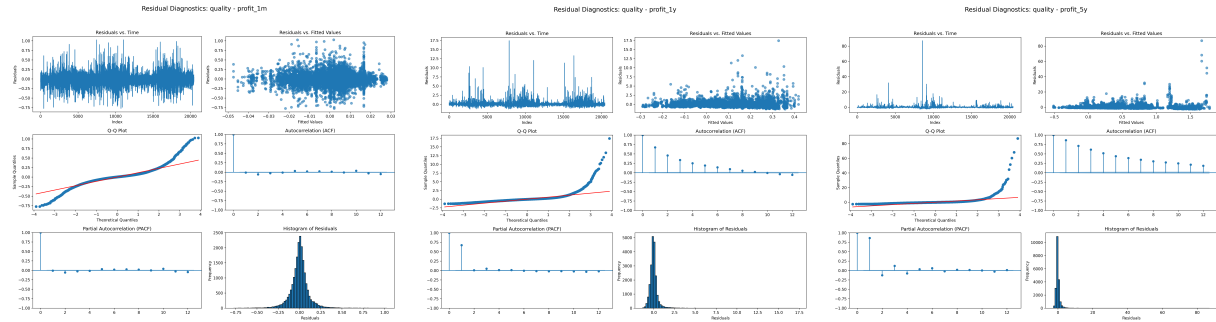


Figure 7.22: USA - First Model Residuals

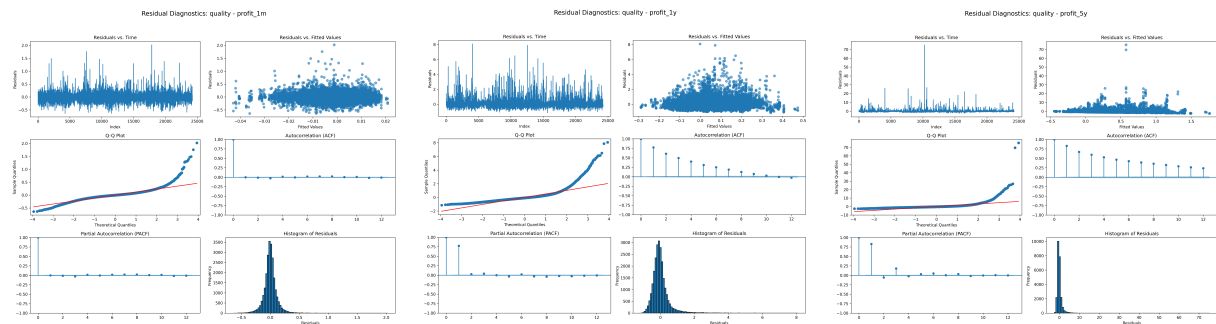


Figure 7.23: EU - First Model Residuals

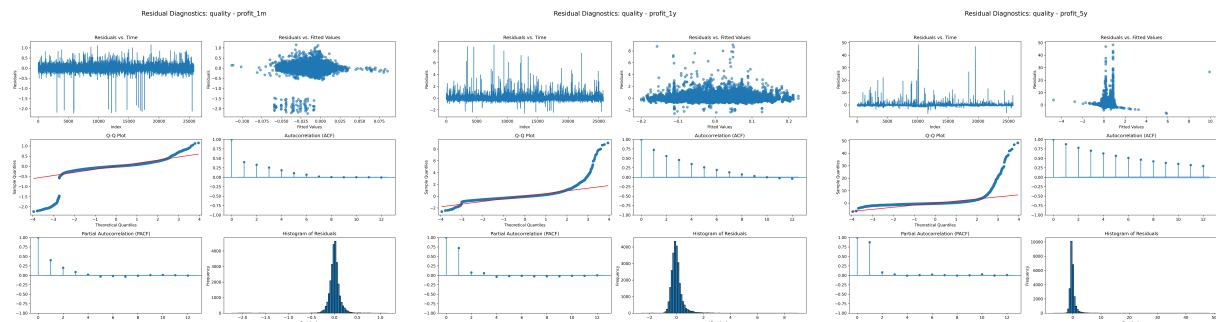


Figure 7.24: Asia & Oceania - First Model Residuals

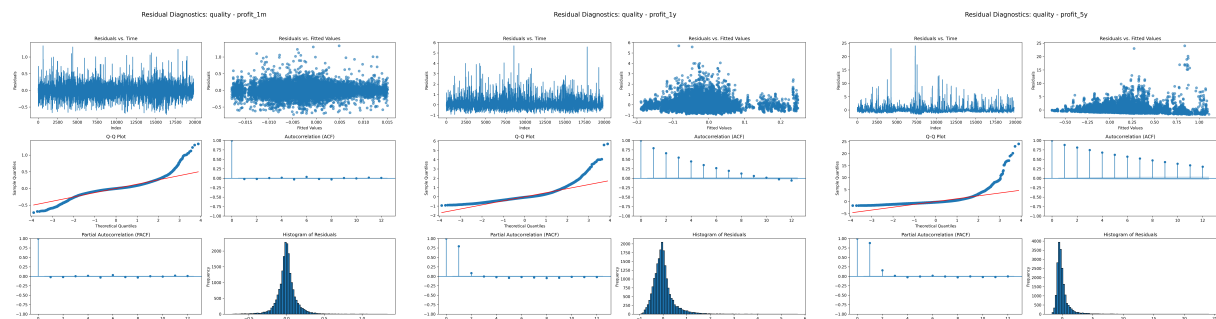


Figure 7.25: South America - First Model Residuals

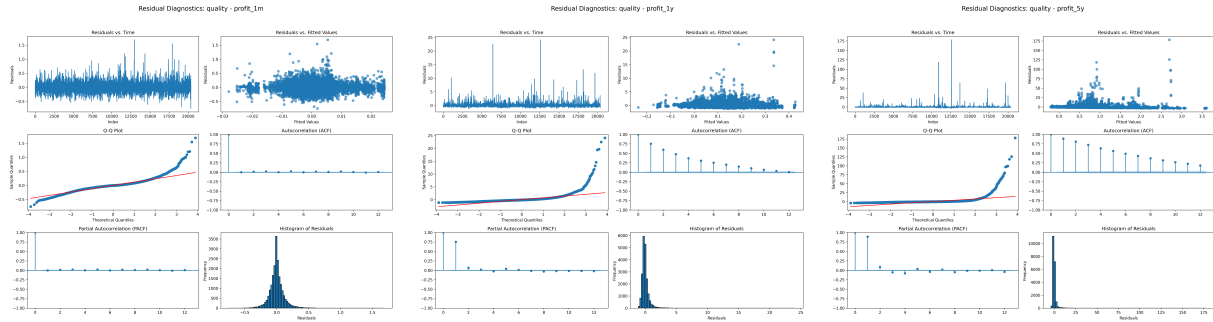


Figure 7.26: Africa & Middle East - First Model Residuals

Table 7.6: North America - First Model Consistency

Score	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	—	—	0.000620	—	-0.001933	0.000089
	Sign	0	0	+	0	—	+
	t-value	—	—	2.00	—	-11.60	2.22
Growth	Coef.	-0.000038	-0.002244	—	0.002143	-0.000109	-0.000258
	Sign	—	—	0	+	—	—
	t-value	-1.98	-6.68	—	10.38	-2.76	-4.03
Quality	Coef.	-0.003752	-0.019694	—	-0.015246	-0.000893	-0.001066
	Sign	—	—	0	—	—	—
	t-value	-6.16	-13.76	—	-13.64	-5.01	-4.15
Value	Coef.	0.000133	0.001205	0.007211	0.006137	0.000503	0.000722
	Sign	+	+	+	+	+	+
	t-value	6.21	8.95	9.13	14.40	7.44	6.46

Table 7.7: Europe - First Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	0.000034	-0.000409	0.000137	—	0.000080	-0.000181
	Sign	+	—	+	0	+	—
	t-value	3.491	-2.095	3.913	—	3.512	-1.472
Growth	Coef.	-0.000021	-0.000177	-0.000513	-0.000916	-0.002708	—
	Sign	—	—	—	—	—	0
	t-value	-1.523	-3.250	-4.876	-5.045	-8.186	—
Quality	Coef.	0.000162	0.000458	0.000901	0.001214	0.002756	—
	Sign	+	+	+	+	+	0
	t-value	4.417	6.704	7.969	13.772	7.535	—
Value	Coef.	0.000027	0.000258	0.000119	0.001028	0.003683	0.004361
	Sign	+	+	+	+	+	+
	t-value	2.411	6.577	2.095	6.954	9.689	10.399

Table 7.8: Asia & Oceania - First Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	0.000906	-0.000652	0.000605	0.000978	0.000266	0.000131
	Sign	+	-	+	+	+	+
	t-value	10.78	-1.80	12.43	7.65	10.64	6.10
Growth	Coef.	0.000261	0.008546	0.000222	0.001628	0.000161	0.000051
	Sign	+	+	+	+	+	+
	t-value	4.11	19.48	3.91	10.77	4.52	2.66
Quality	Coef.	0.002736	-	0.001551	0.004908	0.000830	0.000357
	Sign	+	0	+	+	+	+
	t-value	17.80	-	16.76	16.61	12.11	6.59
Value	Coef.	0.000977	0.010772	0.000344	0.002790	0.000233	0.000093
	Sign	+	+	+	+	+	+
	t-value	13.84	22.46	7.40	11.84	6.97	3.35

Table 7.9: South America - First Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	0.000066	-0.003331	0.000270	0.000239	0.000723	0.000122
	Sign	+	-	+	+	+	+
	t-value	4.06	-15.62	6.25	3.62	5.07	4.55
Growth	Coef.	0.000070	0.004234	0.000569	0.001376	0.005562	0.000297
	Sign	+	+	+	+	+	+
	t-value	2.97	10.08	5.79	9.71	20.23	4.17
Quality	Coef.	0.000224	0.004259	0.001693	0.002852	0.011861	0.001159
	Sign	+	+	+	+	+	+
	t-value	6.20	13.46	9.44	13.92	11.24	4.62
Value	Coef.	0.000212	-	0.000940	0.002553	0.000768	0.000398
	Sign	+	0	+	+	+	+
	t-value	7.69	-	10.71	12.29	2.23	8.50

Table 7.10: Africa & Middle East - First Model Consistency

Score	Metric	profit_1m	profit_3m	profit_6m	profit_1y	profit_2y	profit_5y
Dividend	Coef.	-0.000866	-0.004017	-0.000262	-0.001265	-0.000145	-
	Sign	-	-	-	-	-	0
	t-value	-7.56	-7.76	-4.84	-6.82	-4.30	-
Growth	Coef.	-	-	-	0.000679	-	-
	Sign	0	0	0	+	0	0
	t-value	-	-	-	2.80	-	-
Quality	Coef.	0.002288	0.007478	0.001800	0.005933	0.000984	0.000373
	Sign	+	+	+	+	+	+
	t-value	7.99	10.52	6.65	9.34	7.50	5.22
Value	Coef.	0.000324	0.005205	-	0.001033	-	-
	Sign	+	+	0	+	0	0
	t-value	2.34	5.91	-	3.27	-	-

Table 7.11: North America (Windowed) - Dividend Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	-0.003533	-	-	-	-	0.002736
	Sign	-	0	0	0	0	+
	t-value	-2.557	-	-	-	-	3.043
Avg 3M	Coef.	0.004725	0.006878	0.006331	-	-	-
	Sign	+	+	+	0	0	0
	t-value	2.764	4.968	2.568	-	-	-
Avg 6M	Coef.	-	-	0.010553	0.020918	0.006700	0.009805
	Sign	0	0	+	+	+	+
	t-value	-	-	3.756	6.128	5.563	2.816
Avg 12M	Coef.	-0.002796	-0.015889	-0.021812	-0.022473	-0.004578	-0.024517
	Sign	-	-	-	-	-	-
	t-value	-1.875	-6.835	-7.153	-6.589	-3.735	-5.823
Avg 24M	Coef.	0.002135	0.009623	0.005791	0.003355	-	0.017524
	Sign	+	+	+	+	0	+
	t-value	1.914	6.165	3.857	2.626	-	7.630

Table 7.12: North America (Windowed) - Growth Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Growth	Coef.	-0.003847	–	0.004537	–	–	–
	Sign	–	0	+	0	0	0
	t-value	-4.932	–	2.601	–	–	–
Avg 3M	Coef.	–	-0.004490	-0.006014	–	-0.002181	–
	Sign	0	–	–	0	–	0
	t-value	–	-6.548	-5.463	–	-2.179	–
Avg 6M	Coef.	0.001922	–	–	–	0.002828	–
	Sign	+	0	0	0	+	0
	t-value	2.446	–	–	–	1.996	–
Avg 12M	Coef.	–	-0.002837	-0.010174	-0.016151	-0.007067	–
	Sign	0	–	–	–	–	0
	t-value	–	-2.577	-8.399	-16.542	-5.147	–
Avg 24M	Coef.	0.003565	0.007495	0.013033	0.015419	0.006597	0.001522
	Sign	+	+	+	+	+	+
	t-value	6.300	9.227	13.719	15.737	5.889	4.469

Table 7.13: North America (Windowed) - Quality Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Quality	Coef.	-0.044063	-0.029268	-0.043875	–	–	-0.015212
	Sign	–	–	–	0	0	–
	t-value	-8.338	-7.730	-5.369	–	–	-2.553
Avg 3M	Coef.	0.037309	-0.021267	-0.027318	-0.018541	-0.007487	–
	Sign	+	–	–	–	–	0
	t-value	4.942	-3.205	-3.857	-3.920	-1.643	–
Avg 6M	Coef.	-0.012506	0.026198	0.012333	–	–	-0.073538
	Sign	–	+	+	0	0	–
	t-value	-2.200	3.244	1.671	–	–	-7.024
Avg 12M	Coef.	–	-0.017565	-0.013525	-0.047918	-0.035989	0.070625
	Sign	0	–	–	–	–	+
	t-value	–	-2.349	-2.054	-7.885	-4.195	5.399
Avg 24M	Coef.	0.019811	0.042821	0.054187	0.081220	0.065399	0.017921
	Sign	+	+	+	+	+	+
	t-value	7.656	10.017	12.563	17.276	9.129	2.908

Table 7.14: North America (Windowed) - Value Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Value	Coef.	0.046496	0.035204	0.020768	0.012553	0.008890	0.006598
	Sign	+	+	+	+	+	+
	t-value	55.124	45.616	25.571	17.930	16.524	15.597
Avg 3M	Coef.	-0.055977	-0.010366	0.007323	0.001726	–	–
	Sign	–	–	+	+	0	0
	t-value	-45.578	-7.308	7.432	1.584	–	–
Avg 6M	Coef.	0.010512	-0.030808	-0.020960	0.005571	0.006391	–
	Sign	+	–	–	+	+	0
	t-value	11.652	-17.408	-14.481	5.213	6.054	–
Avg 12M	Coef.	–	0.019451	–	-0.013875	0.008084	0.013861
	Sign	0	+	0	–	+	+
	t-value	–	6.267	–	-11.950	6.524	12.748
Avg 24M	Coef.	-0.003114	-0.016167	-0.011902	-0.012283	-0.028632	-0.027899
	Sign	–	–	–	–	–	–
	t-value	-5.754	-5.969	-10.147	-12.061	-29.415	-26.523

Table 7.15: Europe (Windowed) - Dividend Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	-0.002319	–	–	0.002447	–	–
	Sign	–	0	0	+	0	0
	t-value	-4.627	–	–	4.727	–	–
Avg 3M	Coef.	–	–	-0.004990	–	-0.001809	-0.001703
	Sign	0	0	–	0	–	–
	t-value	–	–	-7.015	–	-3.330	-2.839
Avg 6M	Coef.	–	-0.011594	–	–	0.003648	–
	Sign	0	–	0	0	+	0
	t-value	–	-11.971	–	–	2.709	–
Avg 12M	Coef.	-0.003771	-0.005616	–	–	-0.005270	-0.008001
	Sign	–	–	0	0	–	–
	t-value	-4.862	-5.298	–	–	-5.813	-6.342
Avg 24M	Coef.	0.006838	0.012544	–	0.011382	0.009546	0.010074
	Sign	+	+	0	+	+	+
	t-value	9.487	13.741	–	14.584	13.924	11.231

Table 7.16: Europe (Windowed) - Growth Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Growth	Coef.	-0.003834	–	–	-0.001480	-0.001226	-0.004033
	Sign	–	0	0	–	–	–
	t-value	-4.728	–	–	-2.964	-3.101	-5.267
Avg 3M	Coef.	0.005981	–	–	–	0.007625	–
	Sign	+	0	0	0	+	0
	t-value	4.451	–	–	–	4.221	–
Avg 6M	Coef.	-0.004325	–	–	0.002594	–	-0.004810
	Sign	–	0	0	+	0	–
	t-value	-4.204	–	–	2.585	–	-2.468
Avg 12M	Coef.	–	-0.012226	–	-0.008831	-0.004896	-0.009216
	Sign	0	–	0	–	–	–
	t-value	–	-9.180	–	-10.960	-3.653	-11.507
Avg 24M	Coef.	0.002266	0.006223	0.009021	0.010648	0.009251	0.005465
	Sign	+	+	+	+	+	+
	t-value	4.977	8.715	9.367	10.281	12.791	7.113

Table 7.17: Europe (Windowed) - Quality Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Quality	Coef.	-0.012149	–	–	–	-0.017072	-0.010249
	Sign	–	0	0	0	–	–
	t-value	-3.717	–	–	–	-6.212	-5.986
Avg 3M	Coef.	0.011110	–	–	–	–	–
	Sign	+	0	0	0	0	0
	t-value	2.037	–	–	–	–	–
Avg 6M	Coef.	-0.015817	0.011698	–	-0.031290	-0.020334	–
	Sign	–	+	0	–	–	0
	t-value	-3.508	1.499	–	-6.743	-4.466	–
Avg 12M	Coef.	–	-0.040263	-0.047057	-0.057587	-0.032370	-0.036177
	Sign	0	–	–	–	–	–
	t-value	–	-6.852	-5.470	-8.507	-6.638	-7.096
Avg 24M	Coef.	0.019378	0.088280	0.073143	0.086730	0.099170	0.046953
	Sign	+	+	+	+	+	+
	t-value	10.140	19.679	12.576	18.025	18.959	10.833

Table 7.18: Europe (Windowed) - Value Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Value	Coef.	0.041385	0.007550	0.017503	0.012756	0.009658	0.033926
	Sign	+	+	+	+	+	+
	t-value	26.063	12.232	23.779	17.162	16.896	26.338
Avg 3M	Coef.	-0.050866	–	0.008328	–	–	-0.015267
	Sign	–	0	+	0	0	–
	t-value	-24.956	–	9.516	–	–	-9.707
Avg 6M	Coef.	0.012620	0.007815	-0.011331	0.014378	0.006061	-0.018148
	Sign	+	+	–	+	+	–
	t-value	9.974	6.317	-10.855	10.532	6.007	-21.376
Avg 12M	Coef.	-0.003519	0.002363	-0.014963	-0.019574	0.006115	–
	Sign	–	+	–	–	+	0
	t-value	-3.676	2.055	-12.965	-15.114	5.640	–
Avg 24M	Coef.	-0.003632	-0.019976	-0.006693	-0.014679	-0.027589	-0.006000
	Sign	–	–	–	–	–	–
	t-value	-4.981	-20.202	-8.512	-17.659	-22.505	-9.887

Table 7.19: Asia & Oceania (Windowed) - Dividend Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	-0.001047	–	–	–	–	–
	Sign	–	0	0	0	0	0
	t-value	-2.342	–	–	–	–	–
Avg 3M	Coef.	–	-0.003958	-0.003135	-0.001382	-0.001445	-0.000730
	Sign	0	–	–	–	–	–
	t-value	–	-4.388	-3.825	-1.811	-1.928	-1.759
Avg 6M	Coef.	0.002904	0.008625	0.006047	0.002837	0.004974	–
	Sign	+	+	+	+	+	0
	t-value	3.185	6.414	4.529	2.478	4.087	–
Avg 12M	Coef.	-0.003844	-0.009180	-0.008636	-0.004538	-0.007463	-0.004333
	Sign	–	–	–	–	–	–
	t-value	-3.475	-7.315	-6.433	-4.063	-6.493	-5.249
Avg 24M	Coef.	0.003419	0.007207	0.008895	0.006618	0.008068	0.008521
	Sign	+	+	+	+	+	+
	t-value	4.896	9.449	11.231	9.421	10.853	11.601

Table 7.20: Asia & Oceania (Windowed) - Growth Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Growth	Coef.	–	–	–	–	–	0.000874
	Sign	0	0	0	0	0	+
	t-value	–	–	–	–	–	1.423
Avg 3M	Coef.	-0.004340	-0.005294	-0.003929	-0.002820	–	-0.001962
	Sign	–	–	–	–	0	–
	t-value	-4.951	-6.552	-4.761	-3.759	–	-1.774
Avg 6M	Coef.	0.005920	0.006249	0.004275	0.003018	0.002020	0.005674
	Sign	+	+	+	+	+	+
	t-value	4.554	5.029	3.525	2.558	2.869	3.681
Avg 12M	Coef.	-0.003753	-0.004498	-0.005803	-0.005715	-0.010915	-0.014371
	Sign	–	–	–	–	–	–
	t-value	-3.629	-4.332	-5.528	-5.221	-10.270	-8.390
Avg 24M	Coef.	0.002967	0.004537	0.006634	0.006872	0.010789	0.011809
	Sign	+	+	+	+	+	+
	t-value	5.117	7.372	9.865	10.318	13.006	11.118

Table 7.21: Asia & Oceania (Windowed) - Quality Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Quality	Coef.	-0.023861	-0.019494	-0.020660	-0.014892	-0.019225	-0.008036
	Sign	–	–	–	–	–	–
	t-value	-4.569	-9.036	-8.059	-5.597	-6.275	-3.893
Avg 3M	Coef.	0.017262	–	–	–	–	–
	Sign	+	0	0	0	0	0
	t-value	3.093	–	–	–	–	–
Avg 6M	Coef.	–	–	-0.011317	-0.018661	-0.007820	-0.006274
	Sign	0	0	–	–	–	–
	t-value	–	–	-3.311	-3.740	-2.117	-1.535
Avg 12M	Coef.	–	–	–	-0.032597	-0.038348	-0.098705
	Sign	0	0	0	–	–	–
	t-value	–	–	–	-6.352	-6.646	-10.163
Avg 24M	Coef.	0.010252	0.022449	0.039548	0.076738	0.078367	0.125694
	Sign	+	+	+	+	+	+
	t-value	4.614	10.375	14.268	19.246	20.388	14.657

Table 7.22: Asia & Oceania (Windowed) - Value Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Value	Coef.	0.033626	0.026984	0.015424	0.011372	0.008776	0.005103
	Sign	+	+	+	+	+	+
	t-value	48.474	38.593	24.273	19.666	15.566	9.407
Avg 3M	Coef.	-0.042091	-0.007557	0.010292	0.003684	0.003353	0.001833
	Sign	-	-	+	+	+	+
	t-value	-36.923	-7.661	10.450	3.835	4.642	1.962
Avg 6M	Coef.	0.009323	-0.022025	-0.017362	0.006815	-	-0.001703
	Sign	+	-	-	+	0	-
	t-value	7.812	-20.955	-15.936	6.837	-	-1.607
Avg 12M	Coef.	-0.002614	0.001280	-0.012359	-0.015557	0.009052	0.008761
	Sign	-	+	-	-	+	+
	t-value	-2.758	1.324	-11.081	-14.332	5.138	6.696
Avg 24M	Coef.	-0.001221	-0.002786	-0.002022	-0.013676	-0.026588	-0.013843
	Sign	-	-	-	-	-	-
	t-value	-1.557	-3.581	-2.360	-12.981	-15.264	-11.707

Table 7.23: South America (Windowed) - Dividend Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	-0.001830	-0.003625	-0.001051	-	-	0.003446
	Sign	-	-	-	0	0	+
	t-value	-3.124	-4.890	-1.597	-	-	5.154
Avg 3M	Coef.	0.008949	0.009207	0.004809	0.000788	0.001211	-
	Sign	+	+	+	+	+	0
	t-value	3.753	7.182	3.215	1.370	1.897	-
Avg 6M	Coef.	-0.008871	-	0.002759	-	-	0.001650
	Sign	-	0	+	0	0	+
	t-value	-3.755	-	2.089	-	-	1.948
Avg 12M	Coef.	-	-0.007342	-0.009263	0.006750	0.001493	0.003820
	Sign	0	-	-	+	+	+
	t-value	-	-6.984	-7.199	4.508	1.673	1.997
Avg 24M	Coef.	0.002592	0.003323	0.005186	-0.004974	-	-0.010637
	Sign	+	+	+	-	0	-
	t-value	5.938	4.174	6.004	-3.020	-	-5.447

Table 7.24: South America (Windowed) - Growth Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Growth	Coef.	-0.001278	-0.005084	–	0.001146	0.003180	–
	Sign	–	–	0	+	+	0
	t-value	-2.443	-4.312	–	1.573	3.750	–
Avg 3M	Coef.	–	0.003927	-0.007610	-0.001924	-0.002252	–
	Sign	0	+	–	–	–	0
	t-value	–	2.592	-4.395	-1.455	-1.861	–
Avg 6M	Coef.	0.003673	0.007569	0.021172	0.015871	0.009344	0.008857
	Sign	+	+	+	+	+	+
	t-value	2.956	4.282	5.159	8.657	6.404	8.723
Avg 12M	Coef.	-0.007731	-0.017594	-0.034042	-0.029328	-0.019617	-0.017643
	Sign	–	–	–	–	–	–
	t-value	-5.135	-9.873	-7.215	-14.635	-12.264	-11.736
Avg 24M	Coef.	0.007296	0.013563	0.023194	0.018842	0.017345	0.014027
	Sign	+	+	+	+	+	+
	t-value	8.429	13.641	9.135	16.584	16.198	14.528

Table 7.25: South America (Windowed) - Quality Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Quality	Coef.	–	-0.004672	-0.014771	-0.028249	-0.027696	-0.015662
	Sign	0	–	–	–	–	–
	t-value	–	-1.958	-5.764	-3.054	-3.373	-6.776
Avg 6M	Coef.	-0.016699	-0.015467	-0.019944	0.038669	0.006875	–
	Sign	–	–	–	+	+	0
	t-value	-3.945	-3.320	-6.194	2.884	1.564	–
Avg 12M	Coef.	0.020411	0.014073	–	-0.110023	-0.112309	-0.079526
	Sign	+	+	0	–	–	–
	t-value	4.789	2.395	–	-6.575	-10.378	-9.365
Avg 24M	Coef.	–	0.012276	0.042725	0.110538	0.151205	0.104161
	Sign	0	+	+	+	+	+
	t-value	–	3.320	16.045	9.841	14.207	12.109

Table 7.26: South America (Windowed) - Value Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Value	Coef.	0.042605	0.026478	0.012090	0.012708	0.008177	0.004550
	Sign	+	+	+	+	+	+
	t-value	33.253	26.853	16.222	17.911	12.156	7.702
Avg 3M	Coef.	-0.052202	-0.006456	0.015578	0.002115	0.002001	0.001679
	Sign	-	-	+	+	+	+
	t-value	-29.413	-4.363	11.437	2.012	2.116	2.286
Avg 6M	Coef.	0.007253	-0.030704	-0.019211	0.010127	-0.001531	-
	Sign	+	-	-	+	-	0
	t-value	6.550	-15.720	-14.370	8.278	-1.443	-
Avg 12M	Coef.	-	0.006943	-0.014393	-0.018728	0.011924	0.009159
	Sign	0	+	-	-	+	+
	t-value	-	5.232	-17.956	-16.920	10.815	8.761
Avg 24M	Coef.	-	-	-	-0.010340	-0.028875	-0.021639
	Sign	0	0	0	-	-	-
	t-value	-	-	-	-14.361	-30.132	-13.536

Table 7.27: Africa & Middle East (Windowed) - Dividend Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Dividend	Coef.	-0.003627	-0.002118	-	-0.002984	-0.002635	-
	Sign	-	-	0	-	-	0
	t-value	-5.567	-2.714	-	-4.860	-5.411	-
Avg 3M	Coef.	0.004216	0.002480	-	-	-	-
	Sign	+	+	0	0	0	0
	t-value	3.400	1.665	-	-	-	-
Avg 6M	Coef.	-0.002936	-0.003450	-0.005748	0.002934	-	-
	Sign	-	-	-	+	0	0
	t-value	-2.229	-2.113	-4.820	2.479	-	-
Avg 12M	Coef.	0.002174	0.003028	0.002659	-0.012563	-0.006813	-0.004529
	Sign	+	+	+	-	-	-
	t-value	2.977	3.453	1.642	-8.344	-7.888	-5.409
Avg 24M	Coef.	-	-	0.002864	0.012339	0.008672	0.004068
	Sign	0	0	+	+	+	+
	t-value	-	-	3.154	13.037	12.215	4.875

Table 7.28: Africa & Middle East (Windowed) - Growth Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Growth	Coef.	–	-0.001912	-0.001453	–	0.000940	–
	Sign	0	–	–	0	+	0
	t-value	–	-3.774	-1.994	–	1.825	–
Avg 3M	Coef.	–	–	-0.003749	–	–	–
	Sign	0	0	–	0	0	0
	t-value	–	–	-2.558	–	–	–
Avg 6M	Coef.	–	0.002948	0.004849	–	0.002432	–
	Sign	0	+	+	0	+	0
	t-value	–	2.629	2.610	–	1.958	–
Avg 12M	Coef.	-0.005066	-0.012923	-0.015521	-0.018503	-0.018322	-0.002617
	Sign	–	–	–	–	–	–
	t-value	-4.987	-7.110	-7.855	-10.096	-6.795	-3.501
Avg 24M	Coef.	0.005516	0.012733	0.017077	0.020452	0.017803	0.006444
	Sign	+	+	+	+	+	+
	t-value	5.413	8.572	10.828	10.821	7.900	8.692

Table 7.29: Africa & Middle East (Windowed) - Quality Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Quality	Coef.	-0.008892	0.005270	–	-0.006157	-0.014295	-0.017649
	Sign	–	+	0	–	–	–
	t-value	-6.969	1.752	–	-2.033	-3.370	-5.340
Avg 3M	Coef.	–	-0.020470	–	0.007158	–	–
	Sign	0	–	0	+	0	0
	t-value	–	-3.925	–	1.587	–	–
Avg 6M	Coef.	–	0.009254	-0.012054	–	0.006732	–
	Sign	0	+	–	0	+	0
	t-value	–	2.096	-4.518	–	1.641	–
Avg 12M	Coef.	–	–	–	-0.024250	-0.012727	-0.025676
	Sign	0	0	0	–	–	–
	t-value	–	–	–	-4.994	-2.232	-6.435
Avg 24M	Coef.	0.011701	0.010906	0.017645	0.029879	0.022980	0.046570
	Sign	+	+	+	+	+	+
	t-value	9.041	5.147	6.630	7.643	5.528	13.094

Table 7.30: Africa & Middle East (Windowed) - Value Consistency

Variable	Metric	1M	3M	6M	1Y	2Y	5Y
Value	Coef.	0.036820	0.026390	0.015372	0.010815	0.008244	0.005772
	Sign	+	+	+	+	+	+
	t-value	42.388	39.243	23.029	17.932	18.351	15.446
Avg 3M	Coef.	-0.044231	-0.008629	0.008019	0.001898	—	—
	Sign	—	—	+	+	0	0
	t-value	-34.762	-6.960	8.092	2.055	—	—
Avg 6M	Coef.	0.007667	-0.018441	-0.013302	0.012251	0.003187	—
	Sign	+	—	—	+	+	0
	t-value	7.018	-13.701	-10.194	9.491	3.957	—
Avg 12M	Coef.	-0.004391	—	-0.014410	-0.018220	0.009018	0.004212
	Sign	—	0	—	—	+	+
	t-value	-6.825	—	-13.395	-13.741	8.556	4.667
Avg 24M	Coef.	—	-0.005729	-0.003473	-0.015642	-0.029044	-0.013761
	Sign	0	—	—	—	—	—
	t-value	—	-5.946	-4.516	-15.917	-32.465	-15.460

Validation Dataset

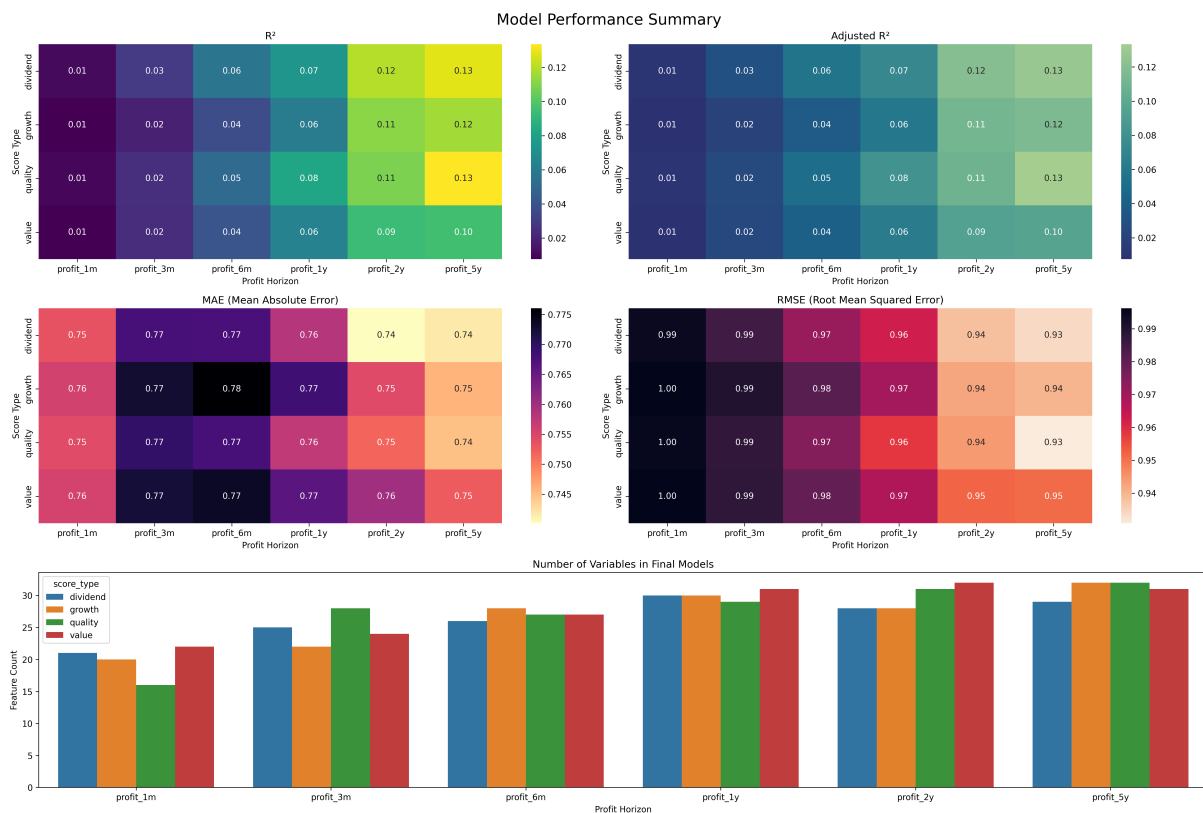


Figure 7.27: Big Data Future IF Balanced - (Windowed) Performance

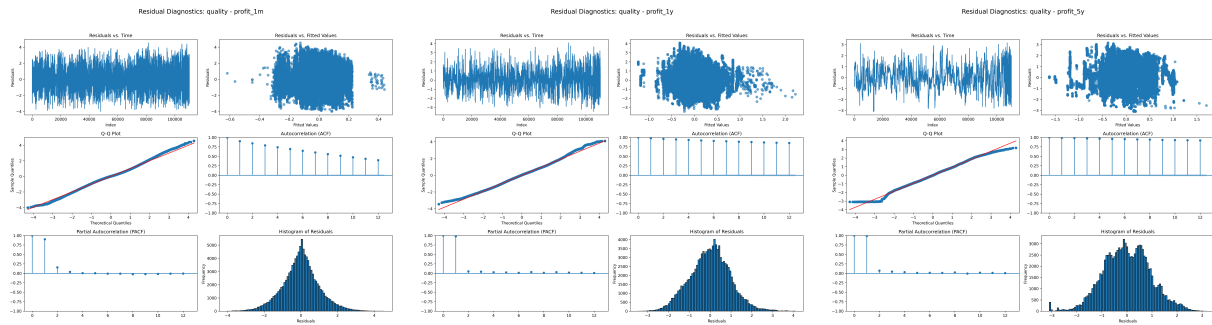


Figure 7.28: Big Data Future IF Balanced - (Windowd) Residues

7.2.2 GAM

Tensor Splines

Encoding of the regions:

- Africa & Middle East: 0
- Asia & Oceania: 1
- Europe: 2
- South America: 3
- North America: 4

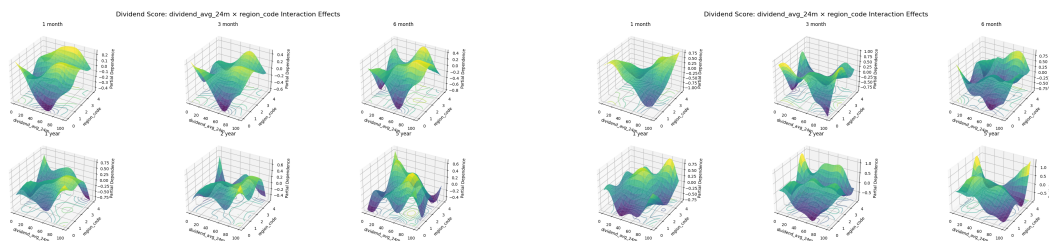


Figure 7.29: Small Data Future IF - Region x Qual-ity

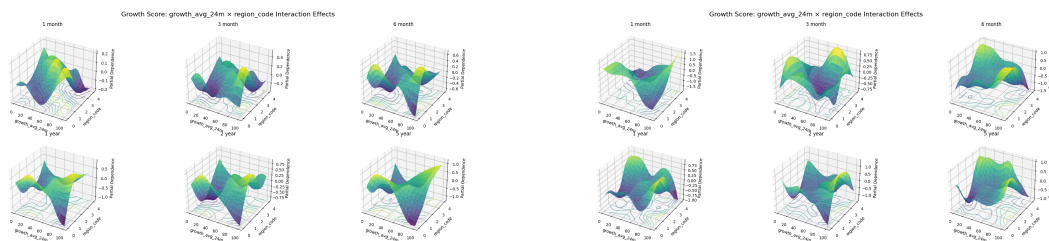


Figure 7.31: Small Data Future IF - Region x Qual-ity

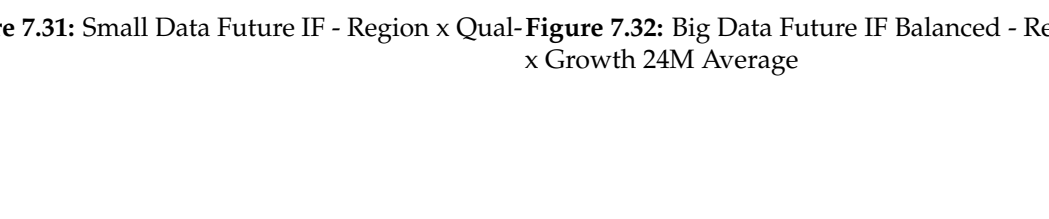


Figure 7.32: Big Data Future IF Balanced - Region x Growth 24M Average

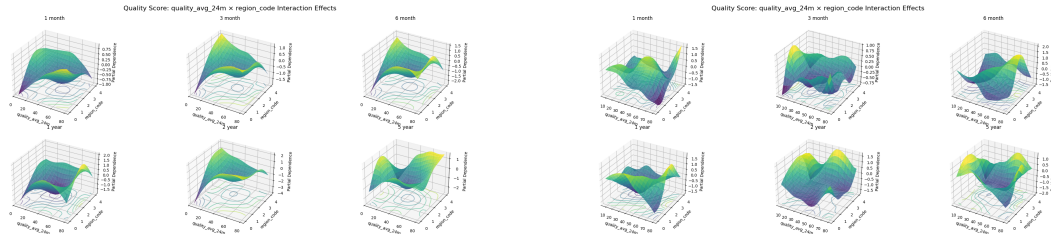


Figure 7.33: Small Data Future IF - Region x Quality 24M Average
Figure 7.34: Big Data Future IF Balanced - Region x Quality 24M Average

7.2.3 NN

Encoder Classifier

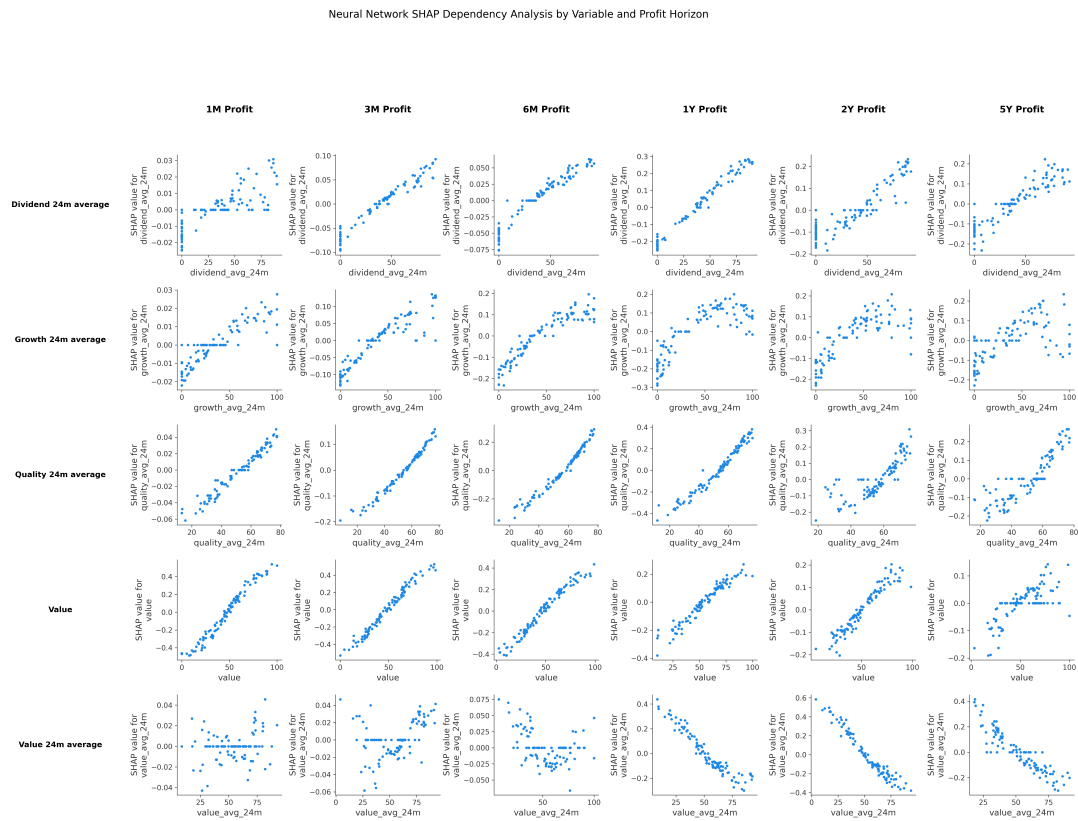


Figure 7.35: Small Data Future IF - Encoder SHAP Consistency

Regressor

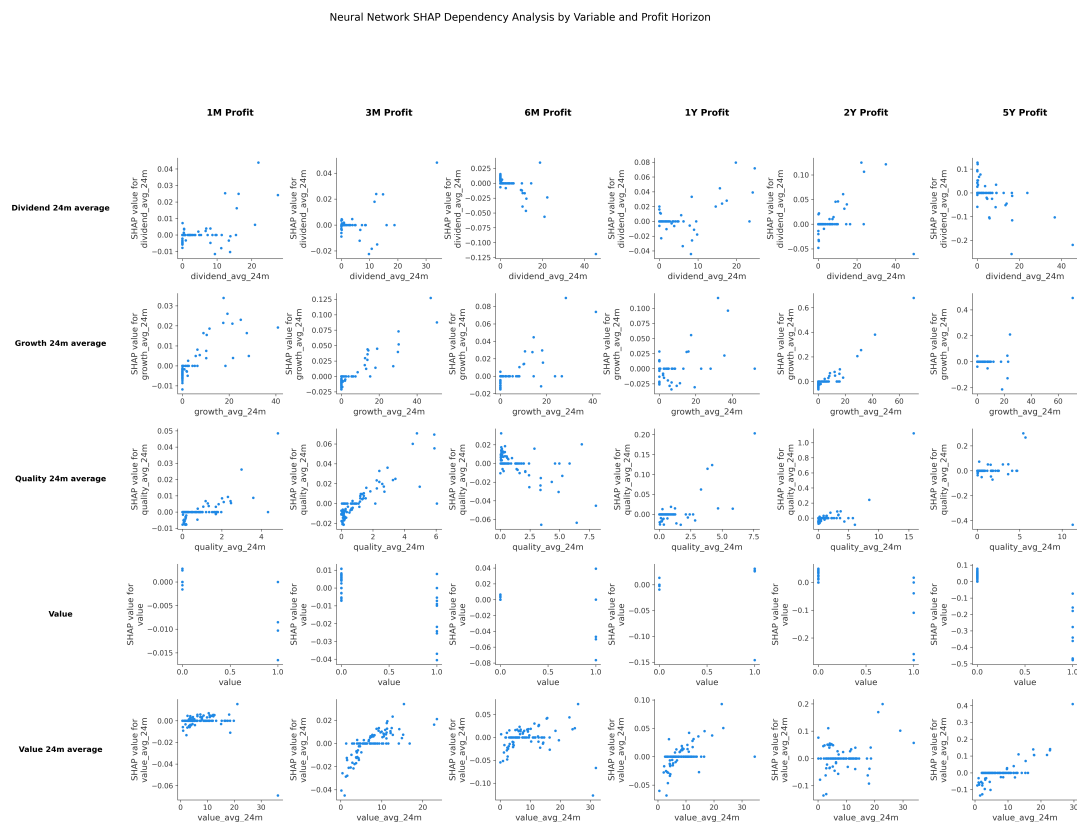


Figure 7.36: Small Data Future IF - Regreesor SHAP Consistency

7.3 2030 Agenda

Table 7.31: Relation of the Project with the Sustainable Development Goals (SDGs) of the 2030 Agenda

Sustainable Development Goal (SDG)	Degree of relation
SDG 1. No poverty	Not applicable
SDG 2. Zero hunger	Not applicable
SDG 3. Good health and well-being	Not applicable
SDG 4. Quality education	Medium
SDG 5. Gender equality	Not applicable
SDG 6. Clean water and sanitation	Not applicable
SDG 7. Affordable and clean energy	Not applicable
SDG 8. Decent work and economic growth	High
SDG 9. Industry, innovation and infrastructure	High
SDG 10. Reduced inequalities	Not applicable
SDG 11. Sustainable cities and communities	Not applicable
SDG 12. Responsible consumption and production	High
SDG 13. Climate action	Not applicable
SDG 14. Life below water	Not applicable
SDG 15. Life on land	Not applicable
SDG 16. Peace, justice and strong institutions	Not applicable
SDG 17. Partnerships for the goals	Not applicable

7.4 Executive Summary

Table 7.32: Relation of the Thesis with ABET Competencies

Concept (ABET)	Complies (Y/N)	Where (pages)
1.1 Problem statement and opportunity	Y	1–3
1.2 Constraints (standards, codes, needs, requirements & specifications)	Y	4–6, 9–16
1.3 Setting of goals	Y	6–7
2.1 Creative solution generation (analysis)	Y	33–46
2.2 Evaluation of multiple solutions and decision-making (synthesis)	Y	49–83
3.1 Fulfilment of goals	Y	85–87
3.2 Overall impact and significance (contributions and practical recommendations)	Y	87–89