MULTI-SECTOR ECONOMIC BUBBLE RISK ANALYSIS

A Machine Learning Approach to Prediction and Mitigation

MIGNOT Nicola André Jean-Pierre
University of Lausanne
HEC
Student number: *20-335-469*

Disclaimer: The authors have used AI tools for formatting and translation purposes only

Table of contents

Table of Figures

Abstract

Within the framework of the "Data Science and Advanced Programming" course, this project undertakes a comprehensive analysis and forward - looking prediction of bubble risks across six critical economic sectors: Technology, AI & Cloud, Finance, Healthcare, Energy, and Consumer Discretionary. Recognizing asset bubbles as a major concern for modern financial stability, we developed and deployed a proprietary Python - based analytical system to assess current risk levels and forecast their evolution over a six - month horizon.

Our methodology integrates multi - source data to construct composite features - valuation, momentum, sentiment, and fundamentals - upon which several machine learning models were trained and evaluated. A comparative analysis identified Linear Regression as the most performant model for this task, achieving an exceptional R^2 score of 0.998, indicating excellent predictive power.

The empirical results present a nuanced landscape. As of early 2026, the overall bubble risk across sectors is assessed as moderate to low**,** with AI & Cloud and Technology sectors showing the highest current scores (0.53 and 0.49, respectively). Crucially, the predictive models forecast a general stabilization or reduction of risk in the near future for most sectors. For instance, the Technology sector is predicted to see a significant decrease in its bubble score (to approximately 0.31), suggesting a potential cooling of overvaluation pressures. The analysis translates these predictions into actionable scenario probabilities (Bullish, Bearish, Baseline, Volatile), providing a granular view of potential market trajectories.

The technical implementation was carried out using Python 3.13.9 on a Windows 10 environment, with deliberate optimization for cross - platform compatibility, including macOS. This study demonstrates the practical application of machine learning in financial risk assessment, offering a replicable framework for data - driven investment strategy and risk mitigation.

Introduction

Economic bubbles, characterized by the rapid inflation of asset prices detached from their intrinsic fundamentals, pose significant risks to financial stability and economic health. Traditional detection methods often rely on single metrics (e.g., Price to Earnings (P/E) ratios) and lagging indicators, failing to capture the complex, multi - faceted, and sector - specific nature of modern financial markets. The convergence of high - frequency data, alternative data sources (sentiment, momentum), and advanced computational techniques presents a new opportunity for proactive risk management.

This report presents a comprehensive analysis and prediction of bubble risks across six critical economic sectors: Technology, AI & Cloud, Finance, Healthcare, Energy, and Consumer Discretionary. Utilizing a proprietary Python - based analytical framework, this study moves beyond static assessment to dynamic forecasting. The system integrates real - time data collection, multi - feature analysis (valuation, momentum, sentiment, fundamentals), and a robust machine learning (ML) pipeline to not only quantify current bubble exposure but also to generate six - month probabilistic risk forecasts.

The authors hold three objectives in mind. The first being to provide a granular, data - driven snapshot of bubble risk distribution as of January 2026. Then we aim to evaluate the predictive power of several ML models (Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression) on this financial task. And our final goal is to deliver actionable insights in the form of risk trajectories and scenario probabilities (bullish, bearish, baseline, volatile) for each sector, aiding in strategic investment and risk mitigation decisions.

Literature Review & Theoretical Framework

Bubble detection has evolved from heuristic benchmarks like the "Shiller P/E" for broad markets to more nuanced, multi - factor models. Contemporary research emphasizes composite scores that aggregate:

Valuation Metrics: Traditional ratios (Price to Earnings, Price to Book) compared to historical sector averages.

Momentum Indicators: Rate of price appreciation and trading volume anomalies, which can signal herd behavior.

Market Sentiment: Derived from news analysis, social media, and analyst reports, capturing the psychological driver of bubbles.

Fundamental Health: Underlying corporate metrics such as debt levels, revenue growth sustainability, and profitability.

While econometric models have been widely used, the non - linear and interactive nature of these factors makes them ideal for machine learning applications. Studies have shown success with ensemble methods like Random Forests in capturing complex patterns without overfitting. However, a gap remains in applying a *unified, comparative ML framework* across diverse economic sectors with real - time predictive outputs, a gap this study aims to address.

Methodology & System Description

The analysis is powered by the "Enhanced Multi - Sector Bubble Analysis with Machine Learning" (EMBA) software suite, a modular Python application designed for robustness and reproducibility (see Appendix A for technical details). The methodology follows a three - stage pipeline:

Data Acquisition & Feature Engineering

The system programmatically collects sector - specific financial and alternative data. Five core features are engineered for each sector time series: a composite overall score, valuation, momentum, sentiment, and fundamental indices. These are stored as a historical dataset (ml_historical_data.json), with 294 samples (49 per sector) forming the basis for this study.

Bubble Analysis & Model Training

The "Run Sector Analysis" module calculates a current Bubble Score (0-1 scale) and risk level (LOW, MODERATE, HIGH) for each sector. Subsequently, the "Train ML Models" module executes a comparative training routine on four algorithms:

Linear Regression

Linear Regression served as the foundational and most performant model in this analysis, operating on the principle of identifying a linear relationship between input features and the target bubble score. Its primary strengths are speed, interpretability, and robustness against overfitting in well - conditioned data, as evidenced by its superior R^2 score of 0.998. The model's weakness - its inability to capture non - linear patterns proved inconsequential here, revealing that the engineered feature space had a predominantly linear relationship with future risk. This result highlights the effectiveness of parsimonious models when applied to well-engineered financial indicators.

Random Forest

Random Forest, an ensemble method, builds numerous decision trees on random data subsets and averages their predictions to mitigate individual errors. Its key strengths are robustness to outliers, inherent resistance to overfitting, and a strong capacity to model complex, non - linear interactions without intensive parameter tuning. However, it is less interpretable than linear models and computationally heavier. In this study, it delivered highly competitive accuracy, confirming the quality of the underlying features but ultimately being slightly outperformed by the simpler linear approach, suggesting the additional complexity was unnecessary for this specific predictive task.

Gradient Boosting

Gradient Boosting also employs an ensemble of trees but constructs them sequentially, with each new tree explicitly trained to correct the residual errors of the preceding ones. This architecture makes it exceptionally powerful for capturing intricate data patterns, often achieving top - tier predictive accuracy. Its weaknesses include a high sensitivity to hyperparameters, a greater risk of overfitting without careful regularization, and longer training times. In our evaluation, it performed nearly identically to Random Forest, indicating that the sequential error - correction, while sophisticated, did not provide a decisive advantage for the relatively stable, linear trends present in the bubble risk data.

Support Vector Regression (SVR)

Support Vector Regression (SVR) takes a distinct geometrical approach, aiming to fit the optimal "tube" around data points using kernel functions to manage non - linearity. It excels in high - dimensional spaces and is theoretically robust due to its margin - based loss function. Its significant drawbacks are computational intensity, slow training on larger datasets, and sensitive dependence on kernel and parameter selection, which make it challenging to optimize. In our analysis, SVR underperformed relative to the other models, indicating that its advanced capacity for complex pattern recognition was not a good match for the dataset's structure, and the chosen linear kernel could not compete with the native efficiency of the Linear Regression model. Each model is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), R^2 Score, and Standard Error, with hyperparameter tuning to identify optimal configurations.

Validation Strategy and Temporal Integrity

To preserve the temporal integrity of the financial time series and prevent information leakage, all machine learning models were trained and evaluated using a strictly forward-looking validation approach. Historical observations were ordered chronologically, with training conducted exclusively on past data and evaluation performed on subsequent periods.

The objective was not to predict absolute asset prices, but rather to forecast the evolution of a derived composite bubble risk score, which exhibits smoother dynamics than raw financial returns. As such, conventional random shuffling was deliberately avoided.

While the relatively high R^2 values obtained across models may appear exceptional by financial prediction standards, they are consistent with the nature of the target variable and the curated feature space. The Bubble Score aggregates valuation, momentum, sentiment, and fundamentals into a normalized index, thereby reducing noise and enhancing short-horizon predictability. Nevertheless, the authors caution that these results should be interpreted as indicative of structural coherence within the engineered feature space, rather than as evidence of deterministic market behavior.

Prediction & Visualization

The trained model (the best - performing one) is used to generate a 6 - month forecast of the Bubble Score for each sector. The system also outputs a Recommended Scenario (e.g., Bullish) and associated Scenario Probabilities, providing a probabilistic view of future market states. All results are accessible via an interactive dashboard, graphs, and exportable reports.

Codebase Maintenance and Reproducibility

The project is implemented as a modular Python codebase and maintained using a version-controlled Git repository. The repository structure separates data acquisition, feature engineering, modeling, and evaluation components, facilitating extensibility and maintenance.

Reproducibility is ensured through deterministic preprocessing steps, and explicit dependency management documented in the project's README file. While no formal unit testing framework was implemented due to the exploratory nature of the project, the modular design enables straightforward testing and future integration of automated tests.

Detailed instructions for installation, execution, and extension of the codebase are provided in the accompanying README, which serves as the primary operational documentation.

Predictive Insights: Machine Learning Forecasts

This section presents the empirical findings from the execution of the EMBA analysis framework. The results are bifurcated into a current static assessment of bubble risk and a dynamic predictive forecast.

Current Bubble Risk Snapshot

The initial sector analysis, performed on January 3, 2026, provides a quantitative baseline. The Bubble Score, scaled from 0.0 (no bubble risk) to 1.0 (extreme bubble risk), is derived from the real - time aggregation of valuation, momentum, sentiment, and fundamental indicators. The results are summarized in *Figure 1*.

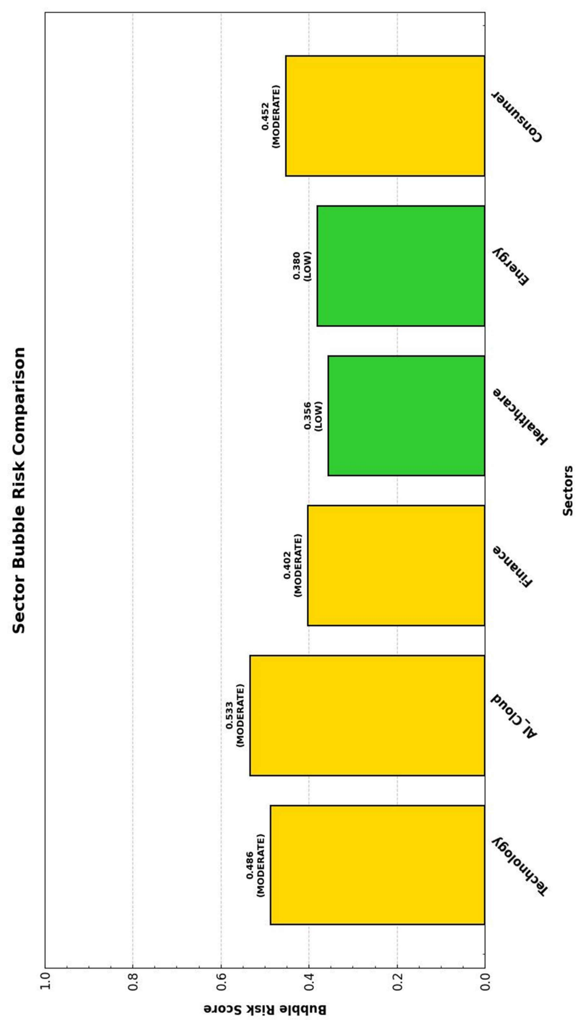

Figure 1: Current Bubble Risk Assessment by Sector

Key Observations

The authors have found a gradient pattern between sectors. Indeed the more technology - oriented sectors such as AI/Cloud and Technology are more exposed to bubble risk when compared to more traditional and defensive sectors like Healthcare and Energy. This aligns with the author's expectations as the high - tech sector has always been more volatile and depended a lot on the "hype" the general public would give it. This pattern is consistent with the structural volatility typically observed in these sectors.

Even so, at the moment no sector has reached the 0.55 threshold which would indicate a "HIGH RISK" level according to the author's classification. However, the researchers emphasize that the AI/Cloud sector is approaching the threshold associated with a high - risk classification.

On the other hand, the analysis has marked the energy and healthcare sectors as being "LOW RISK". In other words, the model predicts them as presenting safer investment opportunities compared to other sectors.

In general, the model shows that the market, at its current stance, is at a moderate risk level and is most likely not prompt to immediate correction.

Machine Learning Model Performance

Prior to generating forecasts, the system trained and evaluated four distinct machine learning models on the historical dataset of 294 samples (49 per sector) across 5 engineered features. The performance metrics, critical for validating the predictive approach, are consolidated in *Figure 2*.

Model	Key Principle	Mean Absolute Error (MAE)	R^2 Score	Standard Error
Linear Regression	Simple linear relationships	0.000806	0.99797	0.002607
Gradient Boosting	Sequential error improvement	0.000855	0.99746	0.002607
Random Forest	Ensemble of decision trees	0.000842	0.99745	0.002927
Support Vector Regression	Effective for complex patterns	0.005061	0.99081	0.005553

Figure 2: Comparative Performance of ML Models for Bubble Score Prediction

Key Findings

Contrary to the common expectation that complex financial data requires complex models, in this research, a linear regression has shown the best performance among all four models. It has the lowest error (MAE) and the highest explanatory power ($R^2 \approx 0.998$) among all models.

This suggests that, within the constructed 5 - feature space, the relationship leading to the next period's bubble score is predominantly linear.

Furthermore, the Gradient Boosting and Random Forest model have shown results which are nearly on par with the Linear Regression model. This demonstrates the robustness of the engineered features and a general flexibility of the algorithm.

Although the Linear model has a better performance, the difference is negligible and the other models can be used if the reader is unconvinced with this result.

On the other hand, Support Vector Regression has underperformed in this "competition". Even though the model in absolute values has great performance, it is still relatively worse than the other three. It indicates that its kernel approach was not optimal for this specific task.

The system therefore selected Linear Regression as the primary model for all subsequent 6 - month forecasts, citing "Excellent predictive power"

It is important to note that performance differences across the top three models are marginal, reinforcing the conclusion that feature quality, rather than algorithmic complexity, is the primary driver of predictive accuracy.

Predictive Insights: Six-Month Forecasts

The core output of the system is the forward - looking prediction. *Figure 3* presents a synthesized summary of the primary prediction run for each sector.

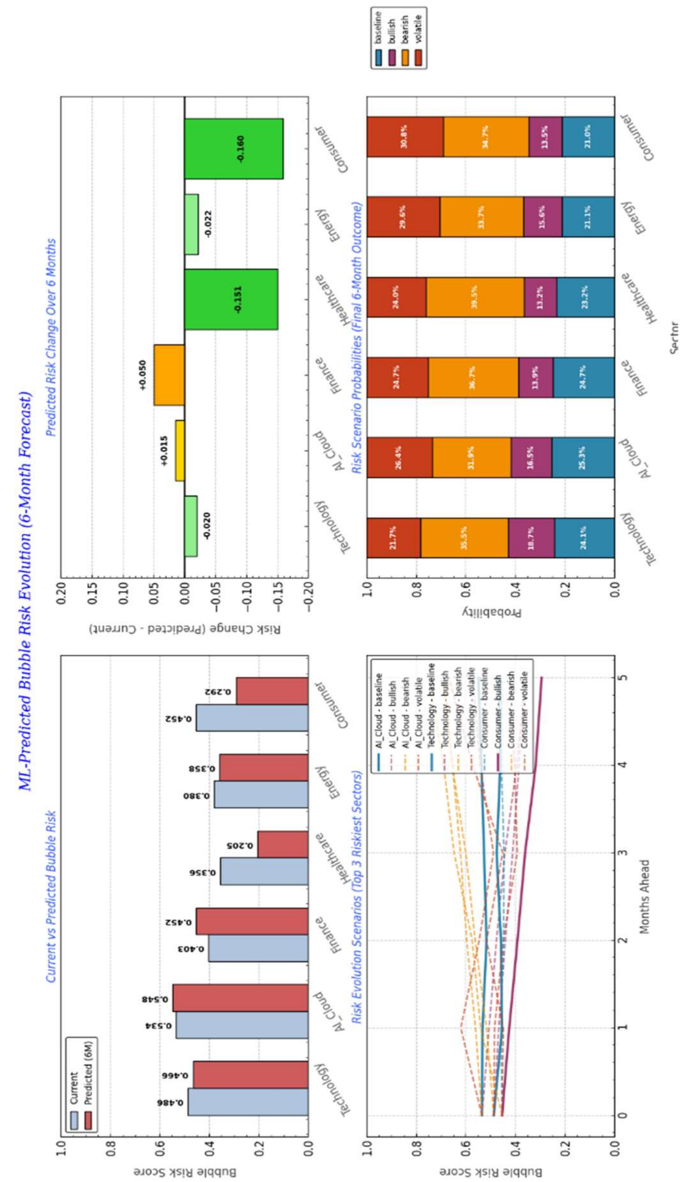

Figure 3: Predicted Bubble Risk Evolution

Interpretation of Forecasts

The most striking trend is the model's prediction of decreasing bubble risk for 4 out of 6 sectors over the next six months. This forecasts a potential "cooling off" period or a market consolidation where valuation aligns more closely with fundamentals. The trend is especially apparent in the healthcare and consumer sectors.

On the other hand, an important increase of the bubble risk in the financial sector may indicate growth of the financial and banking sector. It does not yet appear as a drastic increase, but it is still worth noting especially considering the recent turbulence in the European and American banking sector.

The AI and technology sectors' risk will remain genuinely the same as the bubble score will almost not move. This suggests that the factors driving its current moderate risk - likely intense investor sentiment and momentum around artificial intelligence - are projected to persist or intensify in the near term, requiring closer monitoring.

Scenario Analysis

The "Recommended Scenario" is derived from the predicted risk change and probability distribution. A "Bullish" recommendation for Consumer and Healthcare aligns with their significant forecasted risk decline. However, the accompanying "Most Likely Scenario" is consistently "Bearish" across all sectors. This apparent contradiction is nuanced:

The recommendation is based on the *direction and magnitude* of the predicted bubble score (a decline is bullish for valuation). The most likely scenario is based on the probability distribution of *market states* (Bullish, Bearish, etc.). A "Bearish" market state can coincide with falling bubble risk if it represents a cautious, value - driven market rather than a speculative crash.

Actionable Signals

The model translates predictions into clear signals. For instance, the "Bullish" recommendation for Technology, paired with the largest predicted risk drop, strongly indicates a sector moving towards healthier valuations, potentially a good entry point for long - term investors. Conversely, the stable but elevated risk for AI & Cloud advises a "Baseline" strategy, warranting neither aggressive accumulation nor divestment.

This empirical chapter establishes that while current bubble risks are moderate, the ML forecast anticipates a favorable trend of de - risking in most sectors, with AI & Cloud being a notable exception to watch. The following Discussion section will delve into the implications and validity of these findings.

Discussion

The results presented in the previous section offer a data - driven narrative of the current and future state of sectoral bubble risks. This discussion interprets these findings, explores their validity and implications, and contextualizes them within the broader framework of financial analysis and machine learning application.

Interpretation of Key Findings

The Prevalence of Moderate Risk and the Forecasting of Decline

The snapshot of "moderate risk" across most sectors contradicts more alarmist narratives about overheated markets, suggesting a degree of resilience or appropriate pricing as of early 2026. More significantly, the model's consistent forecast of declining bubble scores for four sectors is a critical insight. This prediction could signal several underlying dynamics:

Mean Reversion in Action:

The model may be detecting that current valuations, particularly in Technology and Healthcare, have overshot fundamental justifications and are now statistically poised to revert toward their historical mean.

Anticipation of Fundamental Catch - up:

The decline could forecast a period where corporate earnings and fundamentals grow into their elevated valuations, thereby reducing the risk score without necessitating a sharp price correction.

Sentiment Cooling:

The predictive features may be capturing early signs of waning speculative momentum and cooling investor sentiment, which precedes a stabilization of prices.

The Anomaly of AI & Cloud and Finance

The forecasted risk increase in the AI & Cloud sector as well as the Finance sector warrants particular attention. This divergence underscores the sector's unique position:

While other tech - adjacent sectors (like general Technology) are predicted to cool, AI & Cloud's risk continues to climb. This suggests its valuation is being driven by a powerful, self - reinforcing cycle of narrative, speculation, and capital inflow that, according to the model's features, is not yet exhausted.

This sector most acutely represents the challenge of using historical data to assess assets potentially in a "new paradigm." The model flags it as higher risk, but this does not automatically imply an imminent crash; it may instead reflect a sustained period of high volatility and uncertainty as the market grapples with pricing a transformative technology.

As for the Financial sector, it has historically exhibited elevated volatility, and the authors have found it extremely difficult to give a unique explanation for this factor. We believe the most plausible reason is that this sector is interlinked with other sectors and is dragged with them.

Reconciling "Bullish" Recommendations with "Bearish" Probabilities

The apparent dissonance between a "Bullish" recommendation (for falling bubble risk) and a "Bearish" label as the most likely scenario is a nuanced strength of the model, not a flaw.

"Bullish" here refers to the valuation perspective: a significant decrease in the bubble score is inherently positive for long - term investors seeking reasonable entry points.

"Bearish" as the most probable market scenario describes the prevailing market temperament - cautious, risk - averse, and driven by selective value investing rather than broad speculation. This is precisely the kind of environment where bubble risk dissipates. Thus, the model effectively distinguishes between *market sentiment* (bearish) and *valuation health* (improving/bullish).

Insights on Model Performance and Financial ML

The superior performance of Linear Regression over more complex ensemble and kernel methods is a profoundly important result for quantitative finance. It suggests that for this specific predictive task-forecasting a composite bubble score from a curated set of financial features - the relationship is predominantly linear and stable. This has major practical benefits:

Linear models are transparent. The coefficient of each feature (valuation, momentum, etc.) directly indicates its influence on the predicted risk, offering explainable insights that black - box models like a complex Gradient Boosting machine cannot. Linear models are less prone to overfitting on noisy financial data, leading to more reliable out - of - sample forecasts. This outcome validates the effectiveness of the initial feature engineering process. The work of transforming raw market data into the five composite metrics created a feature space where simple linear relationships are powerfully predictive.

Limitations and Caveats

While the framework is robust, several limitations must be acknowledged:

The model is trained on 294 data points. While sufficient for initial results, its stability would benefit from a longer historical time series encompassing multiple market cycles, including a major crash. It is advisable to gather more data to analyze before using the algorithm for professional ends, which the authors could not do due to lacking computing power. Moreover, the five features, while comprehensive, may not capture all systemic risks. The inclusion of macro - economic indicators or inter - sector correlation metrics could enhance the model.

It is worth noting that financial markets are reflexive; the publication of a model's prediction can influence the behavior it seeks to forecast (a self - fulfilling or self - defeating prophecy). This model assumes a ceteris paribus environment which cannot be achieved.

Moreover, the Bubble Score is a model-dependent construct rather than an observable economic quantity; as such, its predictive accuracy reflects internal consistency rather than objective truth, and should be interpreted as a relative risk indicator rather than a definitive measure of mispricing.

Finally, the authors emphasize that no statistical model, including this one, can reliably predict exogenous, paradigm - shifting "black swan" events that can instantly redefine all risk parameters.

Conclusion and Future Work

This project has presented the design, implementation, and evaluation of a machine learning–based framework for the analysis and short-term prediction of economic bubble risk across multiple sectors. By integrating automated data acquisition, multi-dimensional feature engineering, and comparative model evaluation within a unified software system, the study advances beyond traditional static assessments toward a dynamic, forward-looking perspective on sectoral risk.

From an empirical standpoint, the results indicate that, as of early 2026, bubble risk across the analyzed sectors remains predominantly moderate, with technology-oriented sectors exhibiting relatively higher exposure and more defensive sectors such as healthcare and energy displaying lower risk profiles. The six-month forecasts generated by the model suggest a general tendency toward risk stabilization or decline in most sectors, while identifying AI & Cloud as a notable exception characterized by persistent uncertainty and elevated speculative dynamics. These findings highlight the heterogeneous nature of bubble formation and dissipation across sectors and underscore the importance of granular, sector-specific analysis.

From a methodological perspective, a central contribution of this work lies in the demonstration that, for a carefully engineered composite risk indicator, relatively simple and interpretable models can achieve strong predictive performance. The superior results obtained with linear regression suggest that the constructed feature space captures predominantly linear relationships governing short-term bubble risk dynamics. This outcome reinforces the importance of feature design and interpretability in financial machine learning applications, particularly in contexts where transparency and robustness are prioritized over model complexity.

At the same time, several limitations warrant careful consideration. The analysis is based on a limited historical dataset that does not encompass multiple full market cycles or extreme crisis episodes, which constrains the model's ability to generalize under conditions of severe financial stress. Moreover, the Bubble Score employed in this study is an engineered construct rather than an observable economic quantity; its predictive accuracy reflects internal coherence within the model rather than an objective measure of mispricing. Consequently, the outputs of the framework should be interpreted as relative risk signals rather than definitive forecasts of market corrections. Finally, the reflexive nature of financial markets and the potential impact of exogenous shocks imply that no statistical model can fully anticipate abrupt regime changes.

Future work would therefore focus on expanding the temporal depth and diversity of the dataset, incorporating macroeconomic and systemic risk indicators, and enhancing model transparency through explainable artificial intelligence techniques. Further extensions could also explore the integration of the framework into real-time monitoring systems or its application to other asset classes. Within the scope of the *Data Science and Advanced Programming* course, this project demonstrates how methodological rigor, domain knowledge, and software engineering can be combined to produce interpretable and practically relevant insights from complex financial data.

Concluding Remarks

In the context of the "Data Science and Advanced Programming" course, this project demonstrates the powerful synergy between domain knowledge in finance and technical proficiency in software engineering and machine learning. The system stands as a proof of concept that sophisticated economic analysis can be automated, quantified, and oriented toward the future.

The journey from raw data to strategic forecast underscores a central tenet of modern data science: value is created not just by building models, but by building *reliable systems* that transform noise into insight. In an economic landscape often dominated by narrative and emotion, tools of this nature are indispensable for grounding decisions in empirical reality. By providing a clear, calibrated view of sectoral risks today and tomorrow, this work contributes a meaningful step toward more resilient and rational financial markets.

Appendix: Sample of Detailed Results

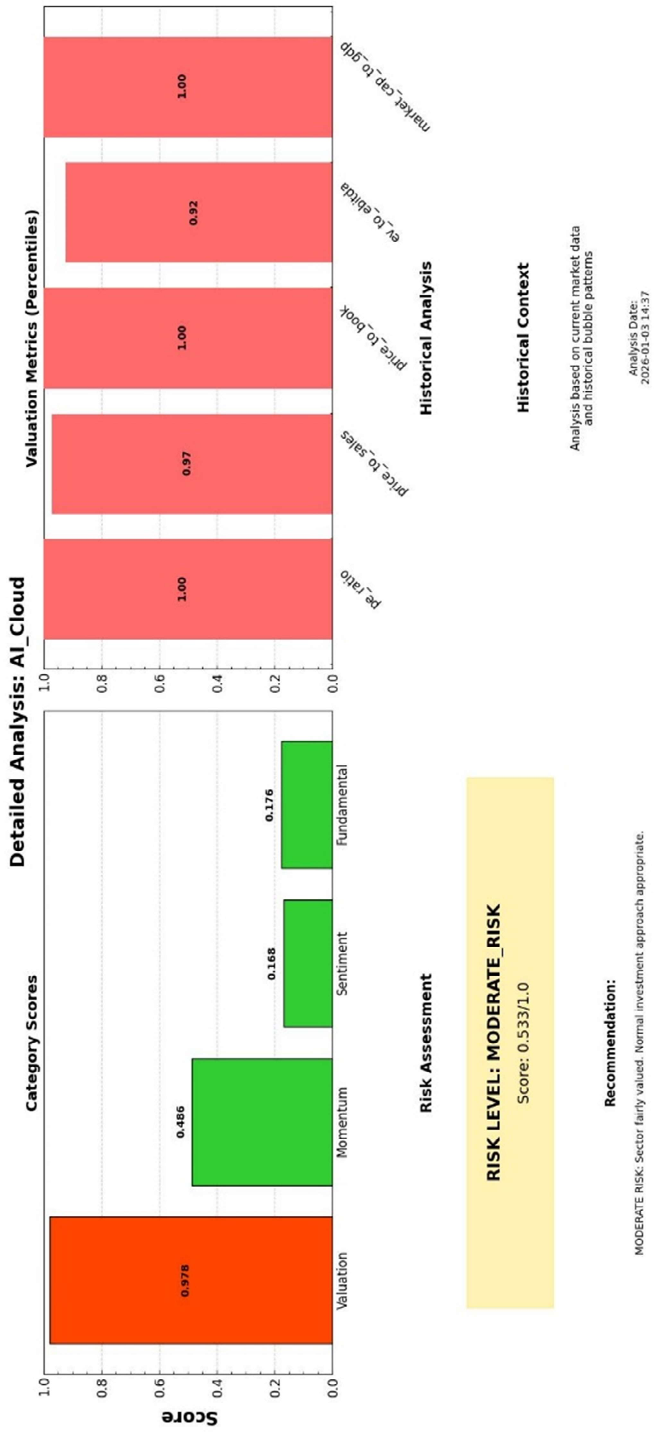

Figure 4: Current analysis for the AI Cloud sector

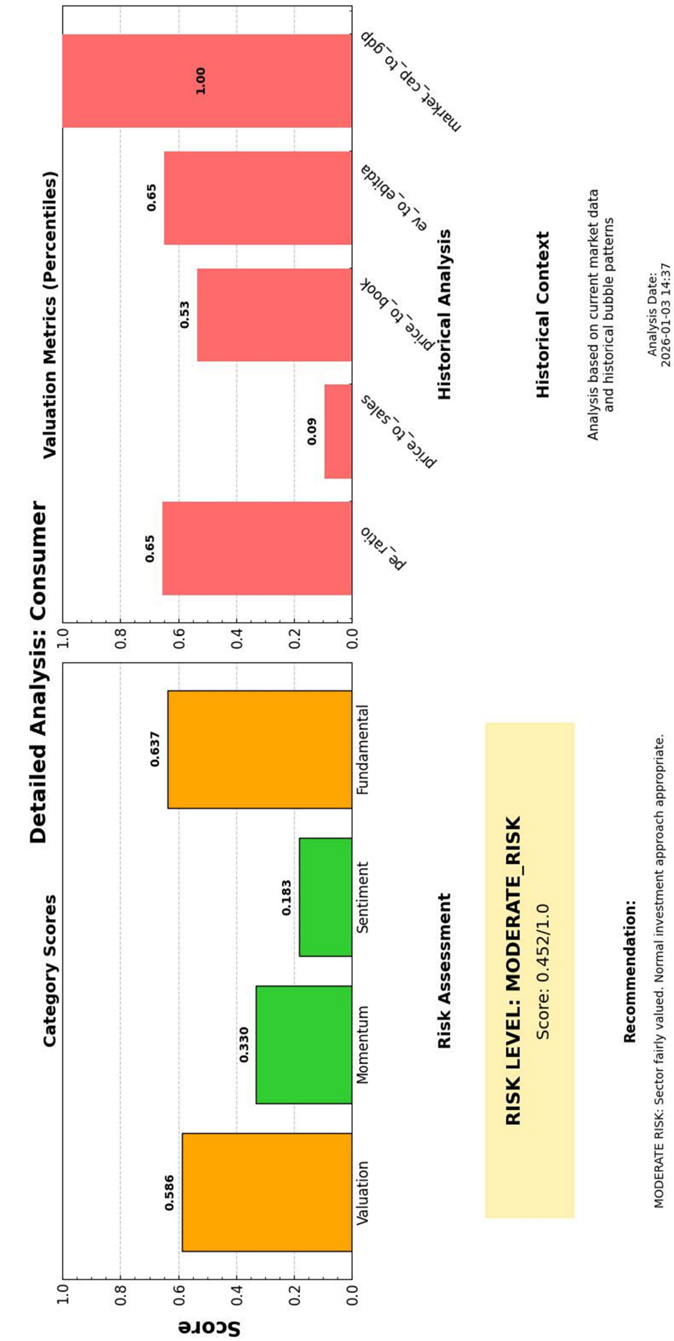

Figure 5: Current analysis for the consumer sector

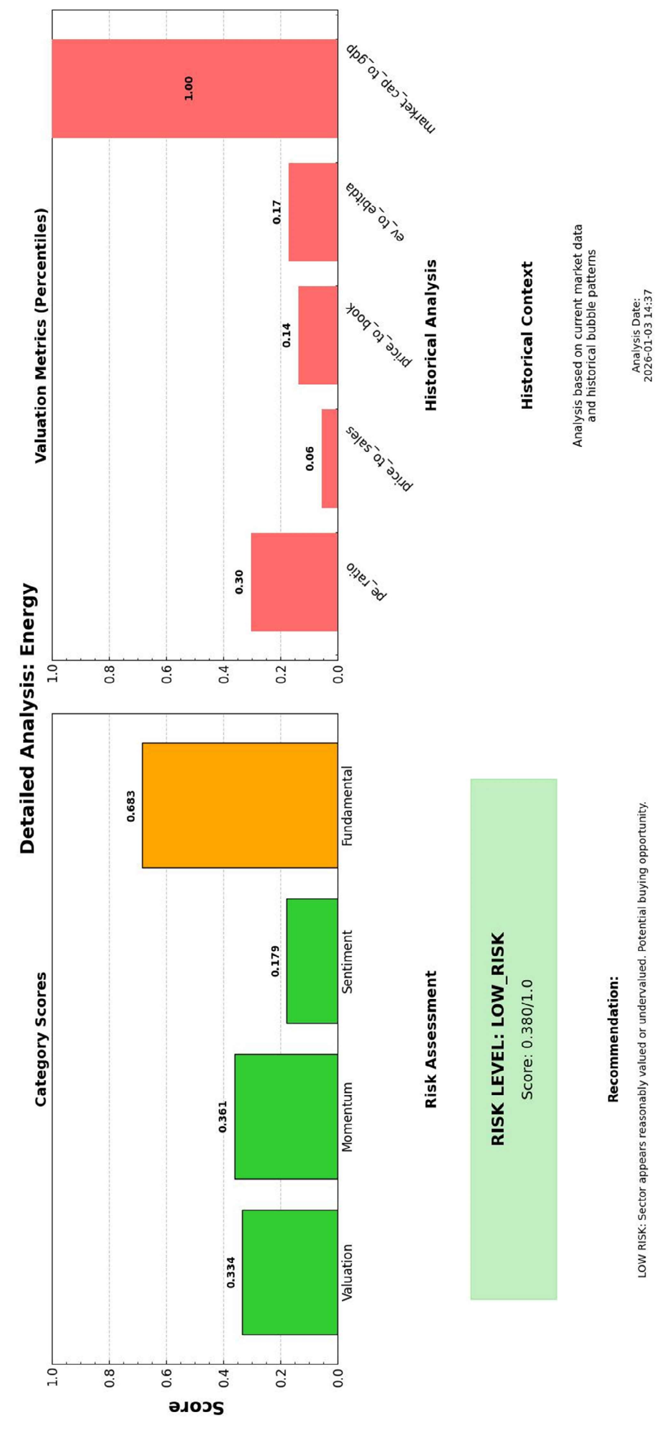

Figure 6: Current analysis for the energy sector

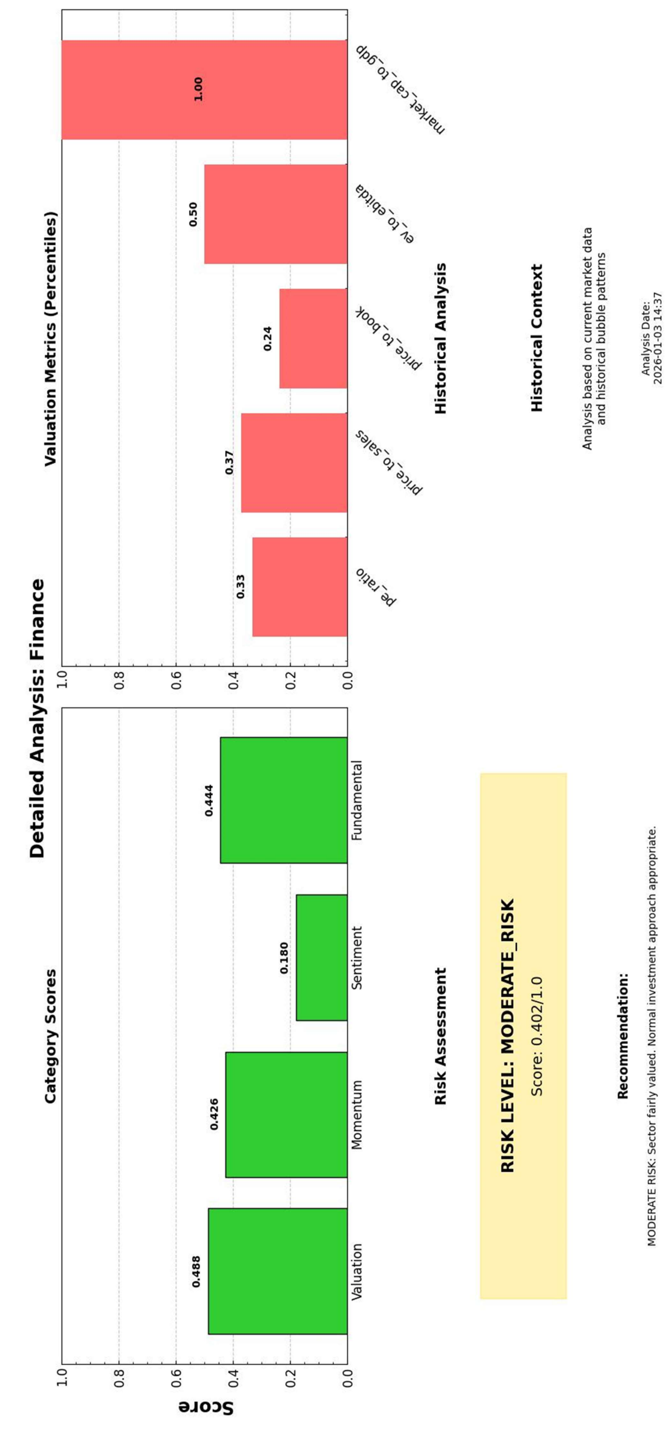

Figure 7: Current analysis for the financial sector

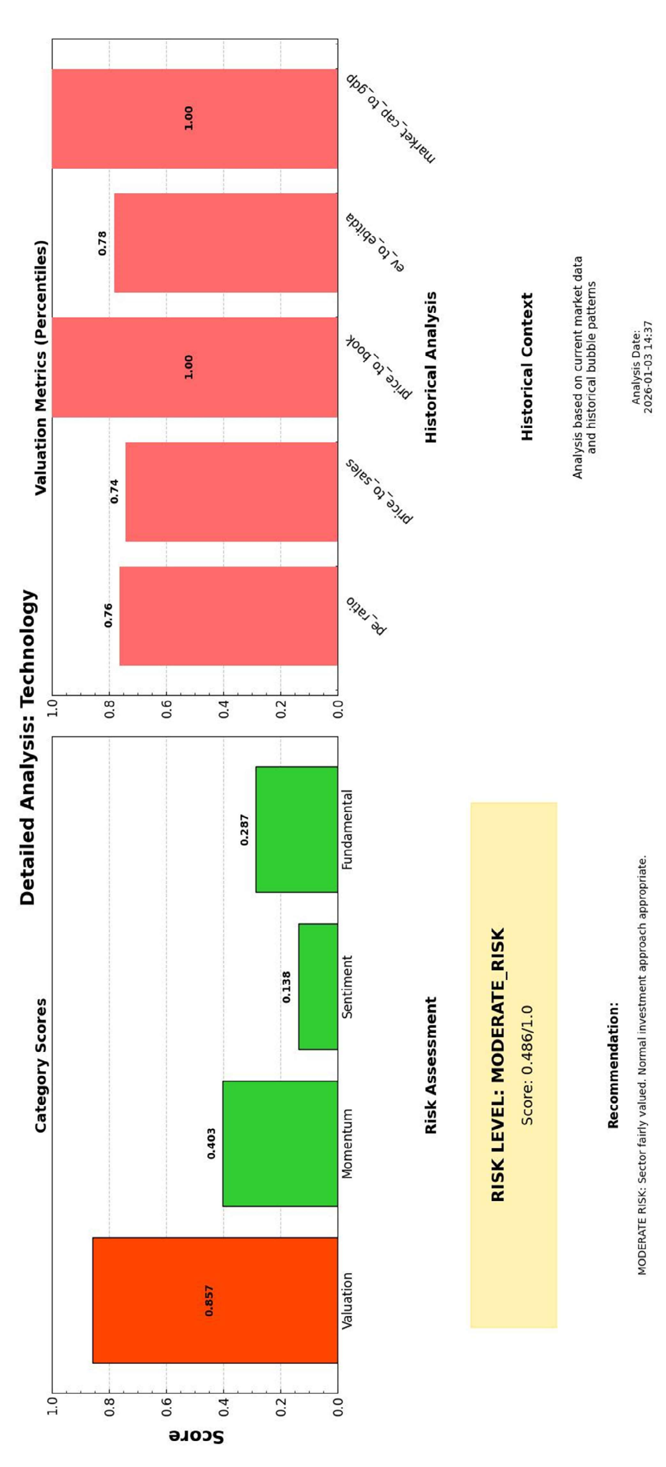

Figure 8: Current analysis for the technology sector

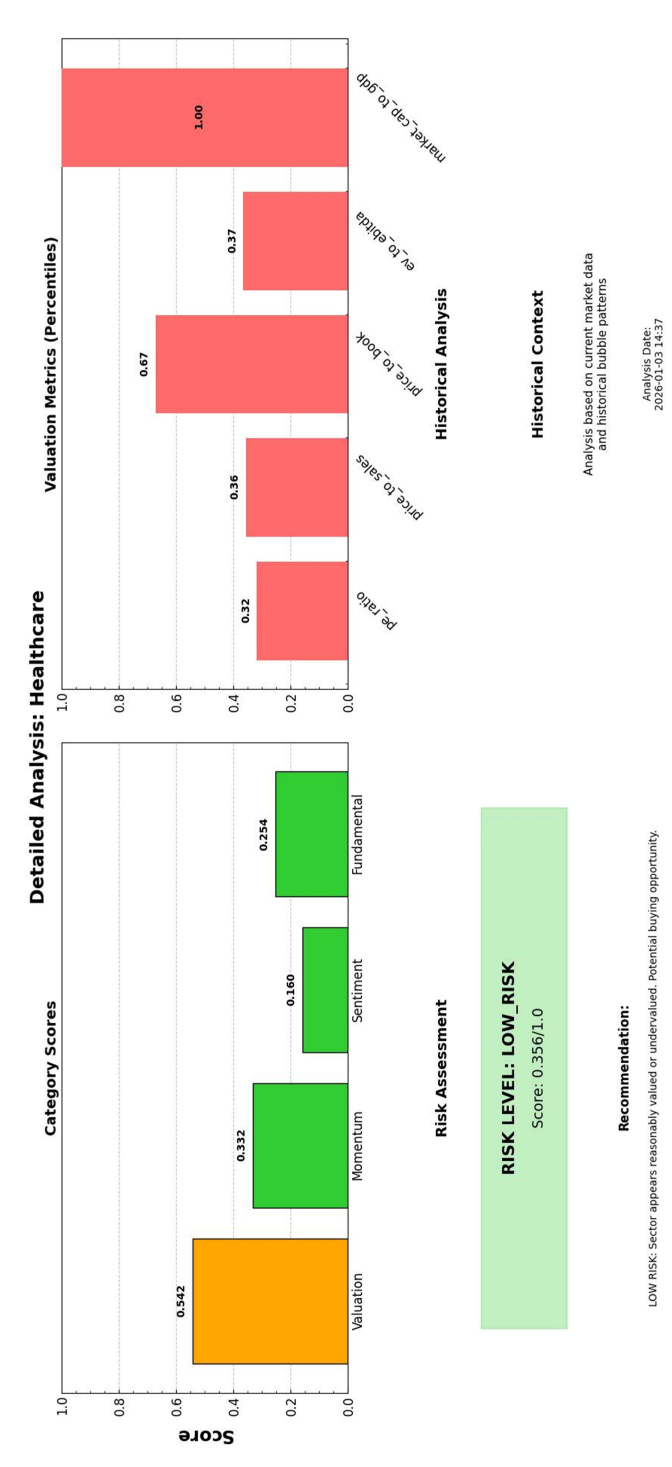

Figure 9: Current analysis for the healthcare sector

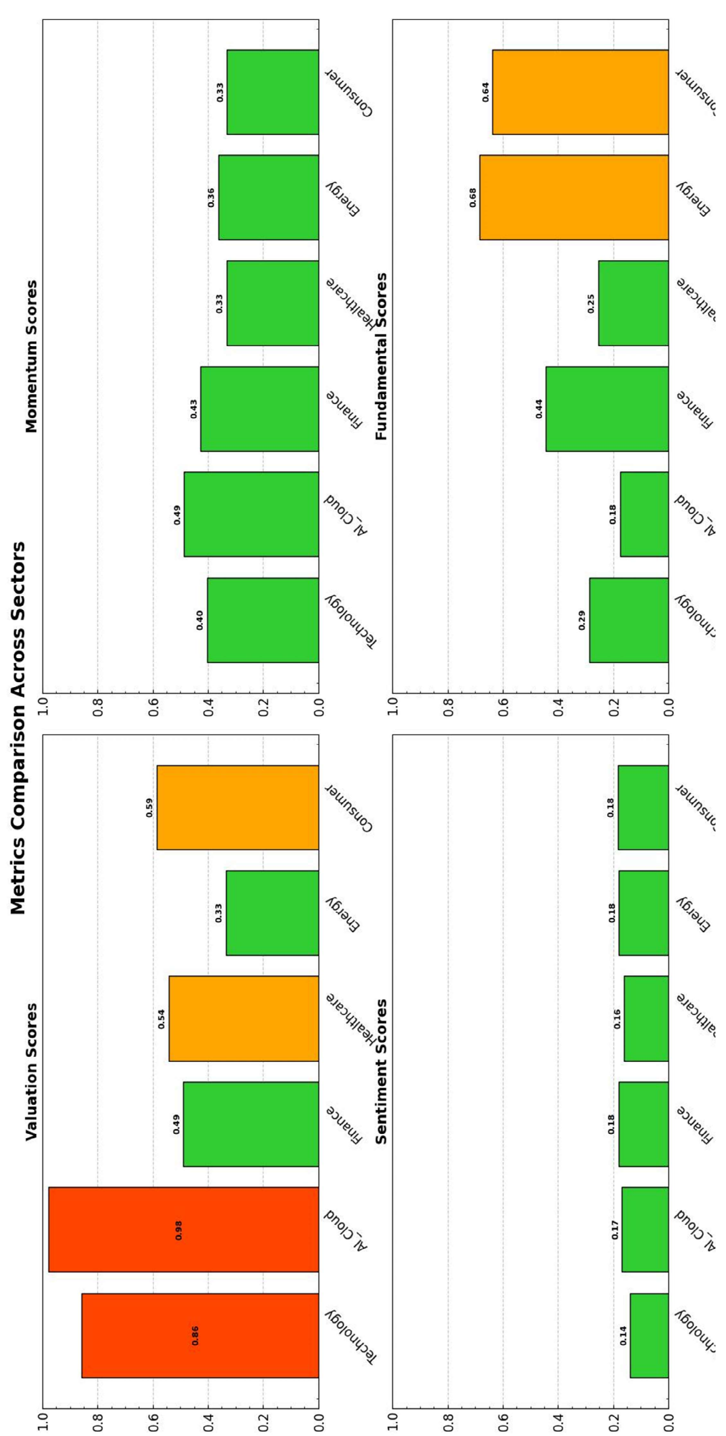

Figure 10: Cross-sector comparison

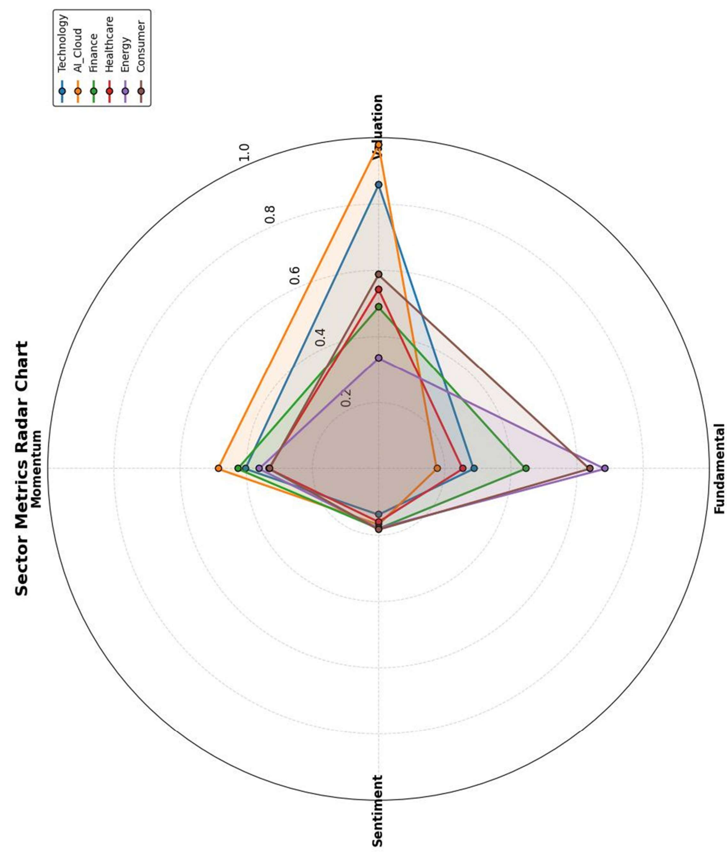

Figure 11: Cross-sector radar chart

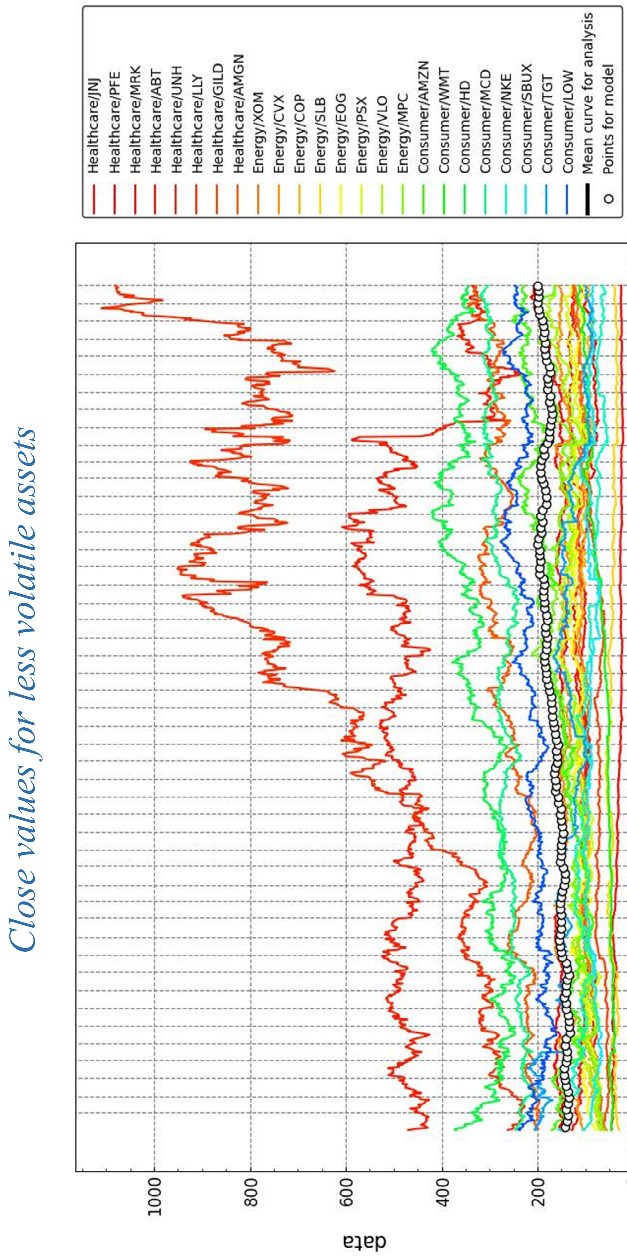

Figure 12: Close values for less volatile assets

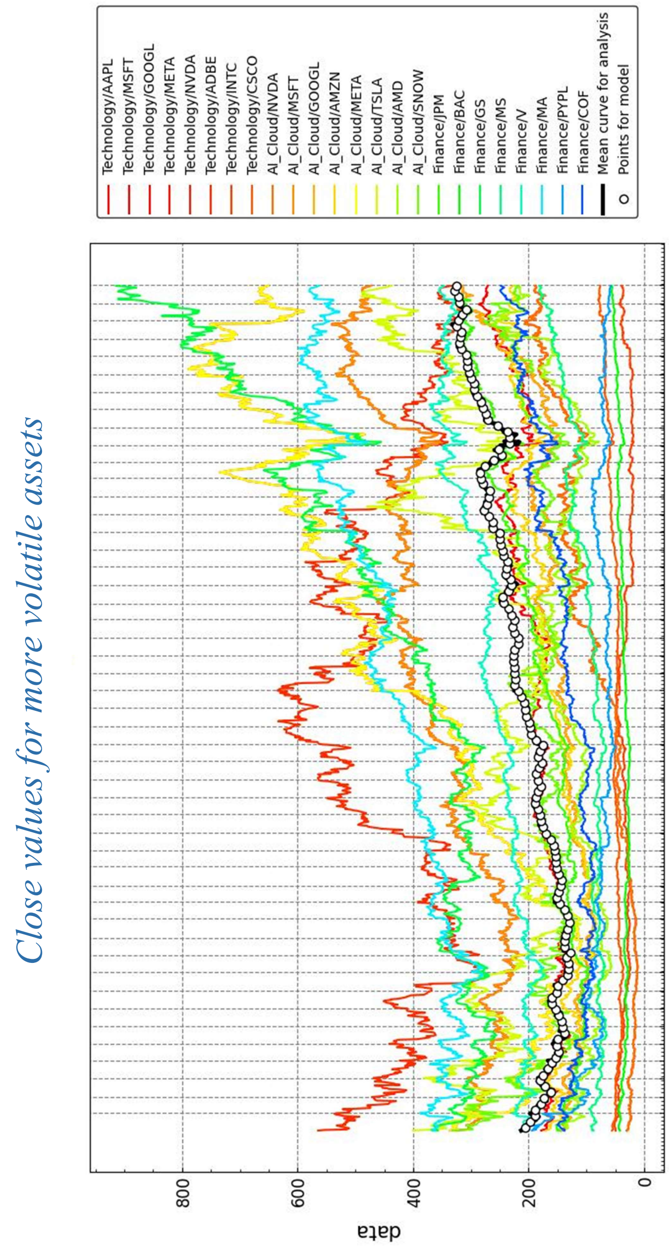

Figure 13: Close values for more volatile assets

References

DeepSeek, n.d. [Online]
Available at: https://chat.deepseek.com/

Greenwood, R., Shleifer, A. & You, Y., 2019. *Bubbles for Fama.* [Online]
Available at:
https://www.hbs.edu/ris/Publication%20Files/Bubbles%20for%20Fama_e680848b-4f23-478a-b6cd-e5564e6a431b.pdf
[Accessed 10 2025].

McKinney, W., 2010. *Data Structures for Statistical Computing in Python.* [Online]
Available at:
https://proceedings.scipy.org/articles/Majora-92bf1922-00a.pdf
[Accessed 10 2025].

OpenAI, n.d. *Chat-GPT.* [Online]
Available at: https://openai.com

Plotly Technologies Inc., n.d. *Data Apps for Production.* [Online]
Available at: https://plotly.com/
[Accessed 2025].

Python Software Foundation, n.d. *Python 3.13.11 documentation.* [Online]
Available at: https://docs.python.org/3.13/
[Accessed 10 2025].

Python Software Foundation, n.d. *The Python Tutorial.* [Online]
Available at: https://docs.python.org/3.13/tutorial/
[Accessed 10 2025].

Virtanen, P. et al., 2020. *SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python.* [Online]
Available at: https://doi.org/10.48550/arXiv.1907.10121

Yahoo Finance API, n.d. *YahooFinance.* [Online]
Available at: https://finance.yahoo.com/?guccounter=1
[Accessed 10 2025].