Predicting Institutional Investment Trends from SEC Form 13F Filings Using Machine Learning

Aditya Saxena
Department of Computer Science
Harvard University
Cambridge, USA
adityasaxena@fas.harvard.edu

***Abstract*—The Securities and Exchange Commission (SEC) Form 13F filings provide insights into the investment strategies of large institutional investors, offering a quarterly snapshot of their holdings. This paper explores the predictive modelling of institutional asset values based on historical Form 13F filings from 2013 to 2024. Machine learning techniques, including Linear Regression, Decision Trees, Random Forests, and XGBoost, are applied to predict the total market value of institutional holdings using engineered financial metrics and firm-level attributes. Extensive feature selection and transformation techniques, including log transformations and scaling, are utilized to optimize model performance. Experimental results demonstrate that tree-based models outperform linear regression, with Random Forest and XGBoost achieving the highest predictive accuracy. Feature importance analysis reveals that an institution's maximum and minimum security values are the most influential predictors. The findings contribute to quantitative finance research by providing an automated approach to analyzing institutional investment patterns and estimating portfolio valuations. The proposed methodology offers practical implications for financial analysts, hedge funds, and market researchers aiming to enhance investment decision-making.**

***Index Terms*—Machine Learning, Institutional Investment, Form 13F, Predictive Modeling, Financial Data Analysis, Random Forest, XGBoost, Portfolio Valuation**

I. Introduction

Institutional investors significantly impact financial markets through capital allocation, market liquidity, and investment trends [1]. The SEC Form 13F filings mandate institutional investors managing over $100 million in assets to disclose their equity holdings quarterly, providing transparency into hedge funds, mutual funds, and pension funds [2]. These filings allow market participants to analyze institutional trading patterns and asset allocations [3].

Recent advancements in machine learning (ML) have revolutionized financial data analysis, enabling predictive modelling techniques to extract patterns from large datasets. Predictive modelling of Form 13F filings has gained traction among researchers, as it offers insights into institutional portfolio valuations and market movements [4]. While traditional econometric models, such as factor-based approaches, have been widely employed, ML techniques particularly tree-based models demonstrate superior predictive performance by capturing complex nonlinear relationships in financial data [5].

This paper develops an ML-based predictive framework to estimate institutional investment values from Form 13F filings. Feature engineering techniques, including log transformations and standardization, are applied to enhance model interpretability. To evaluate predictive accuracy, a comparative analysis is conducted across Linear Regression, Decision Trees, Random Forests, and XGBoost. Results indicate that tree-based models significantly outperform linear regression, with Random Forest and XGBoost achieving the highest predictive accuracy. These findings contribute to the ongoing research on ML applications in portfolio valuation and institutional trading analysis.

II. Literature Review

The application of ML in financial markets has expanded significantly, allowing for predictive analytics in risk assessment, asset valuation, and investment decision-making. Prior research has explored ML-based forecasting models for stock prices, portfolio allocation optimization, and market anomaly detection.

Factor-based approaches such as the Fama–French model remain prevalent in explaining cross-sectional returns for large-cap equities [6]. However, recent research demonstrates the power of machine learning in capturing nonlinear and higher-order interactions among financial variables [7]. In particular, deep learning architectures have shown promise in unearthing complex patterns within large-scale financial datasets [8].

Among ML techniques, tree-based models such as Random Forest and XGBoost have shown exceptional predictive power in financial modelling. Breiman (2001) [9] introduced Random Forest as a robust ensemble learning method, while XGBoost has emerged as an optimized boosting algorithm for financial applications.

Our research builds upon these studies by applying ML models to analyze Form 13F filings. Unlike prior work focused on stock price prediction, our study emphasizes institutional portfolio valuation using ML-driven feature selection and predictive modelling [10]. The results provide insights into the effectiveness of ensemble learning in large-scale financial disclosures.

III. Dataset and Data Preprocessing

A. Dataset Description

This study utilizes the SEC Form 13F dataset, obtained from the EDGAR database, which mandates institutional investors managing over $100 million to disclose their equity holdings quarterly. The dataset provides transparency into the investment strategies of large financial institutions, including hedge funds and mutual funds.

The dataset comprises three primary components:

- **INFOTABLE**: Contains security-level holdings reported in Form 13F filings, detailing attributes such as *VALUE* (market value of the asset/security) and *SSHPRNAMT* (number of shares or principal amount held).
- **COVERPAGE**: Metadata related to Form 13F filings, including the name of the institutional manager and amendment status of the filing.
- **SUMMARYPAGE**: Aggregated statistics summarizing total assets under management (AUM) and total holdings per filing.

Each dataset is linked through the unique *ACCESSION_NUMBER*, assigned by the SEC to every submission, enabling seamless integration of information across the three datasets.

B. Problem Statement

This research uses available financial attributes to predict the total market value of an institutional investor's holdings. Specifically, the objective is to determine whether institutional asset values can be inferred based on security-level attributes, geographic information, and firm-level characteristics. The primary target variable for modelling is **TABLEVALUETOTAL**, representing the total market value of assets reported in a Form 13F filing.

C. Exploratory Data Analysis (EDA)

Prior to model development, extensive data exploration was conducted to understand feature distributions, detect anomalies, and handle missing values [11].

1) Data Statistics: The INFOTABLE dataset contains **3,278,515** rows, reflecting the individual securities held by institutions. The COVERPAGE dataset comprises **10,117** rows, each representing an institutional filer. The categorical variables in the COVERPAGE dataset exhibit significant class imbalances, which may influence modelling outcomes [12], as shown in Figure 1. The SUMMARYPAGE dataset, which aggregates security holdings per filing, contains **8,244** rows.

2) Missing Values and Data Imbalance: Several variables exhibit missing values:

- **NAMEOFISSUER, TITLEOFCLASS, FIGI**: Not critical, as securities can be uniquely identified using *CUSIP*.
- **PUTCALL**: Missing values indicate standard ownership rather than options trading.
- **OTHERMANAGER**: Only applicable when investment discretion is shared with another entity.

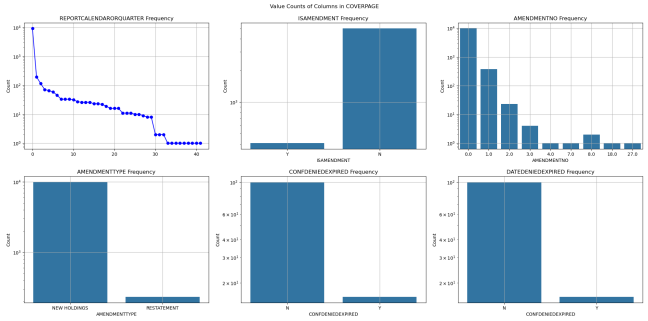

Fig. 1. Frequency distributions of categorical variables in COVERPAGE. Certain variables exhibit significant imbalances, which may affect modelling.

To ensure modelling robustness, missing categorical values were replaced with appropriate labels while missing numerical values were imputed based on their distributions [13].

3) Feature Engineering: To enhance model interpretability and improve predictive performance, the following transformations were applied:

- **Log Transformation**: Due to extreme skewness in financial variables (e.g., *VALUE, SSHPRNAMT, TABLEVALUETOTAL*), log transformations were applied to normalize distributions.
- **Scaling**: Standardization was applied to continuous features to ensure consistent magnitude across variables.
- **Geographic Features**: Institutional filings were classified based on **US vs. non-US** location using state abbreviations.

Institutions exhibit a wide range of portfolio values, with many clustered around the mean but some outliers with significantly higher security values, as observed in Figure 2.

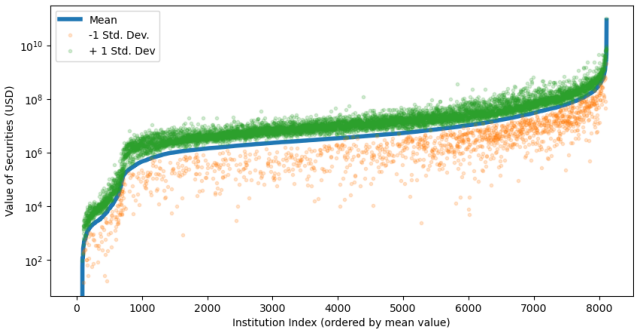

Fig. 2. Distribution of average security values held by institutions. The wide variance suggests a diverse range of portfolio strategies among institutional investors.

4) Correlation Analysis: A correlation heatmap was generated to identify key relationships between variables. Notably, strong positive correlations were observed between:

- **LOG MAX VALUE, LOG MEAN VALUE, and TABLEVALUETOTAL**: Indicates that institutions with high-value securities tend to report larger overall portfolio values.

- **LOG MAX SHAMT and TABLEVALUETOTAL**: Suggests that larger share quantities contribute significantly to total market value.
- **State GDP and Number of Institutional Filers**: Higher economic activity is linked to a greater concentration of institutional investors. A strong positive relationship is observed between state GDP and the number of institutional investors, as visualized in Figure 3.

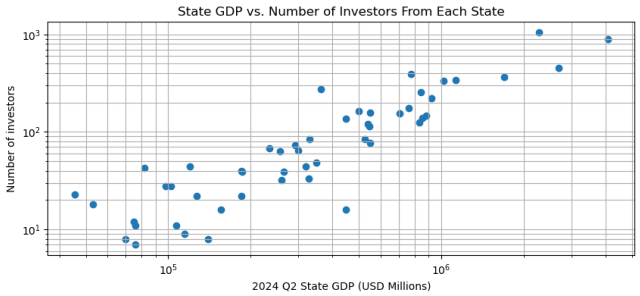

Fig. 3. State GDP vs. number of institutional investors per state. More economically active states attract a higher number of institutional filings.

5) Log-Scale Scatterplots for Key Variables: Scatterplots were generated to visualize trends between financial attributes and total institutional holdings. The application of log scaling revealed clear linear relationships between key variables and **TABLEVALUETOTAL**.

As shown in Figure 4, log transformations reveal clearer linear relationships between security values and institutional holdings, reinforcing their predictive relevance. While security value statistics show stronger correlations with TABLEVALUETOTAL, share amount variables (SHAMT) also exhibit moderate predictive power, as indicated in Figure 5.

6) Final Feature Selection: Based on correlation analysis and exploratory findings, six key predictors were selected for modelling:

- **MAX VALUE, MIN VALUE**: Capture the extreme security values in an institution's portfolio.
- **MEAN VALUE, STD VALUE**: Represent portfolio diversity and risk exposure.
- **MAX SHAMT, STD SHAMT**: Measure share quantity and variability across holdings.

The correlation heatmap in Figure 6 validates the selection of key predictors, as they exhibit strong relationships with TABLEVALUETOTAL.

The next section details the machine learning models employed for predicting institutional asset values based on these preprocessed features.

IV. Methodology

A. Linear Regression

Linear Regression serves as our baseline model due to its simplicity and interoperability [14]. It assumes a linear relationship between the dependent variable y (total asset value) and independent variables X (financial attributes). The model is defined as:

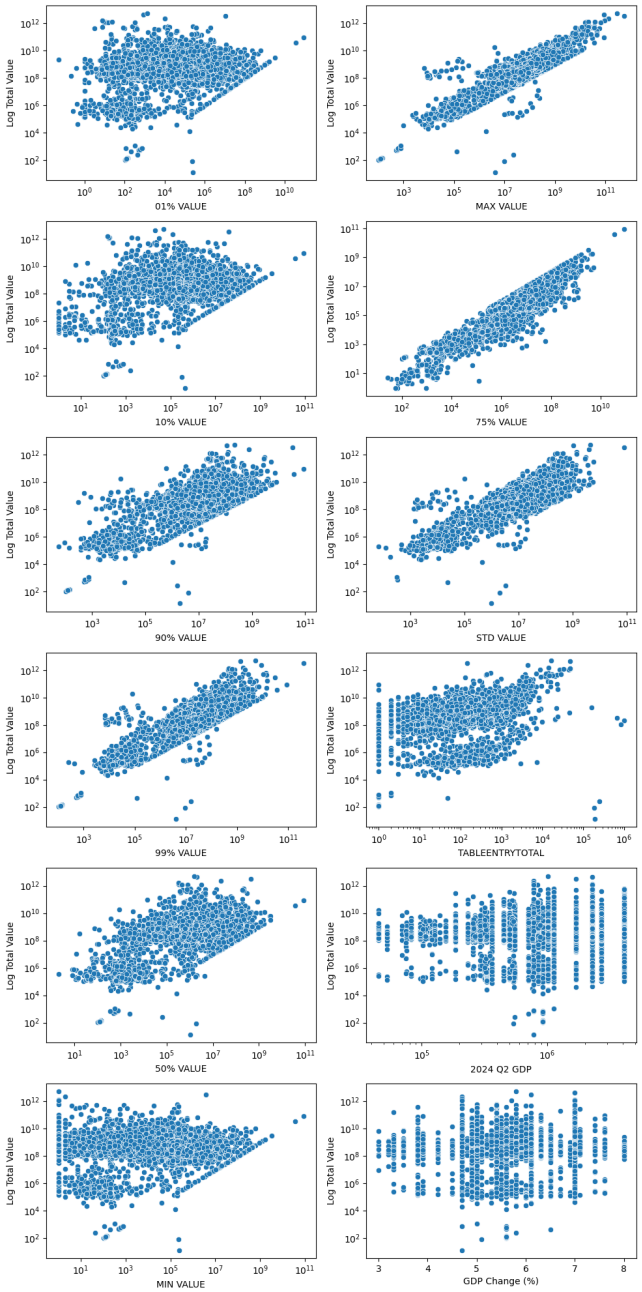

Fig. 4. Scatter plots showing correlations between log-transformed security values and total institutional holdings. Strong linear trends justify the use of log transformations in modelling.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{1}$$

where β_0 is the intercept, β_n are the regression coefficients, and ϵ is the error term. The model is trained by minimizing the residual sum of squares (RSS):

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

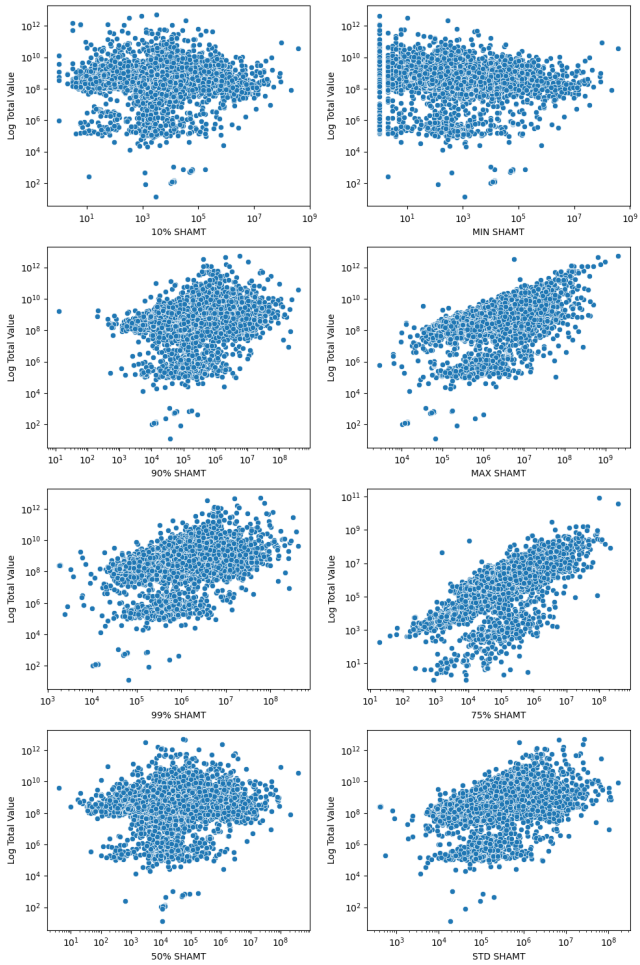

Fig. 5. Scatter plots for log-transformed share amounts and TABLEVALUETOTAL. While SHAMT metrics exhibit some correlation, VALUE-based statistics are stronger predictors.

Although Linear Regression provides a good baseline, its performance was suboptimal due to its inability to capture non-linearity in financial datasets. This limitation is addressed in subsequent models.

B. Decision Tree

Decision Trees provide a non-linear approach by recursively partitioning the dataset based on feature splits that minimize impurity [15]. The model is trained using a recursive binary splitting approach, where at each node, a feature X_j is selected that minimizes the impurity function I, such as Gini Index or Entropy:

$$I_{Gini} = 1 - \sum_{i=1}^{k} p_i^2 \tag{3}$$

$$I_{Entropy} = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{4}$$

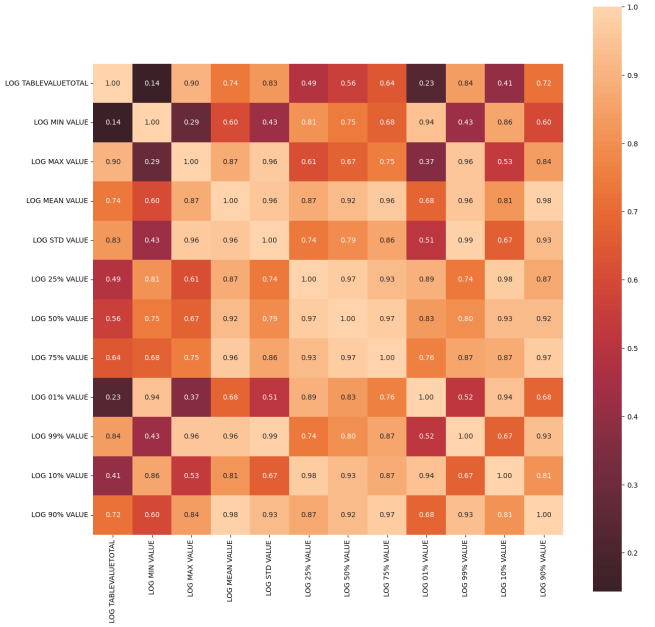

Fig. 6. Heatmap of correlations between log-transformed financial variables. The selected predictors exhibit strong relationships with TABLEVALUETOTAL, validating their inclusion in the model.

A maximum depth of 10 and a minimum sample split of 2 were specified for the implementation. Figure 7 illustrates the feature importance derived from the Decision Tree model.

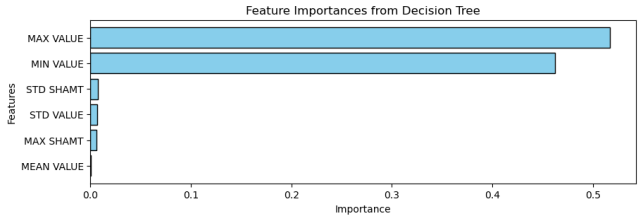

Fig. 7. Feature Importances from the Decision Tree Model. MAX VALUE and MIN VALUE dominate the predictions, highlighting their significance in institutional portfolio estimation.

C. Random Forest

Random Forest enhances Decision Trees by constructing an ensemble of multiple trees and aggregating their predictions to improve robustness. The model applies bootstrap aggregation (bagging), where each tree is trained on a different subset of the data. The prediction is obtained through majority voting for classification or averaging for regression:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(X) \tag{5}$$

where $f_t(X)$ represents an individual decision tree, and T is the total number of trees. The model hyperparameters were tuned using GridSearchCV, where the optimal number of estimators was found to be 100, and a maximum depth of 20 provided the best results [16].

Figure 8 illustrates feature importance in the Random Forest model.

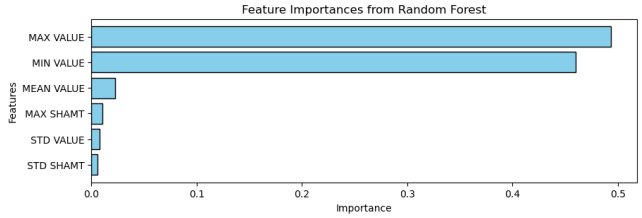

Fig. 8. Feature Importances from Random Forest. Compared to Decision Trees, the model slightly increases the relevance of MEAN VALUE.

D. Extreme Gradient Boosting (XGBoost)

XGBoost [17] is a gradient-boosting algorithm that sequentially builds trees to minimize the loss function. Each tree corrects the errors of its predecessors using the following optimization:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k) \tag{6}$$

where $l(y_i, \hat{y}_i)$ represents the loss function (e.g., squared error for regression), and $\Omega(f_k)$ is the regularization term penalizing complexity.

Key hyperparameters optimized include:

- **Number of Trees (n_estimators)**: 100, 200, 300, 500, and 700.
- **Learning Rate**: 0.01, 0.05, 0.1, 0.2.
- **Max Depth**: 1 and 2 (shallow learners to prevent overfitting).
- **Regularization (L1 and L2)**: Alpha = 0.1, Lambda = 0.01.

The optimal parameters found via GridSearchCV set $n\_estimators = 681$, as seen in Figure 9.

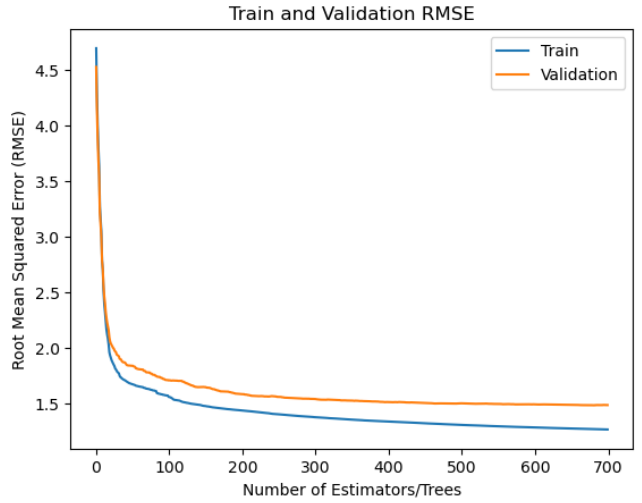

Fig. 9. Training and Validation RMSE for XGBoost. The lowest RMSE was observed at 681 estimators.

Feature importance for XGBoost, depicted in Figure 10, highlights MAX VALUE as the dominant predictor.

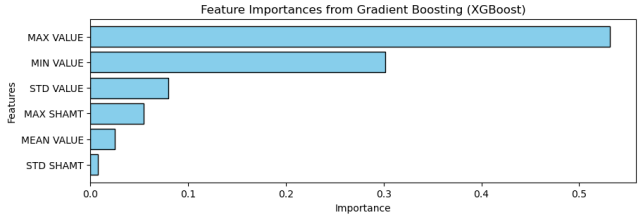

Fig. 10. Feature Importances from XGBoost. MAX VALUE remains the most significant predictor, but STD VALUE gains relevance compared to Random Forest.

V. Results

A. Model Performance Evaluation

The performance of the models was evaluated using the R^2 score and Mean Squared Error (MSE) for both training and test datasets. The Table I presents the results.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	Train R^2	Test R^2	Train MSE	Test MSE
Linear Regression	0.4744	0.5093	13.6274	13.6329
Decision Tree	0.9455	0.8689	1.4137	3.6432
Random Forest	0.9553	0.9000	1.1600	2.7775
XGBoost	0.9349	0.8998	1.6871	2.7829

As expected, Linear Regression performed the worst among all models, with the lowest R^2 scores and the highest MSE values. The Decision Tree model showed significant improvement over Linear Regression, but it was outperformed by ensemble models such as Random Forest and XGBoost.

Random Forest and XGBoost achieved the best predictive performance, with Random Forest slightly outperforming XGBoost in terms of MSE. However, XGBoost is generally considered a more powerful and flexible model due to its regularization capabilities and handling of complex feature interactions.

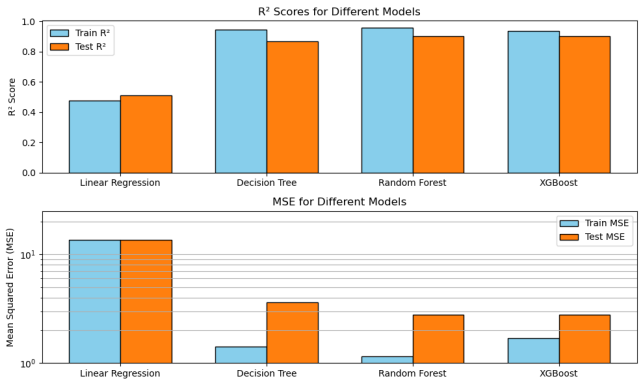

Fig. 11. Comparison of R^2 and MSE across different models. Random Forest and XGBoost demonstrate superior predictive performance compared to baseline models.

B. Feature Importance Analysis

The importance of features across tree-based models (Decision Tree, Random Forest, and XGBoost) was analyzed to identify key predictors of institutional asset values as shown in Table II.

TABLE II
FEATURE IMPORTANCE ACROSS MODELS

Feature	XGBoost	Random Forest	Decision Tree
MAX VALUE	0.5318	0.4937	0.5173
MIN VALUE	0.3015	0.4599	0.4623
STD VALUE	0.0796	0.0076	0.0065
MAX SHAMT	0.0546	0.0104	0.0061
MEAN VALUE	0.0251	0.0225	0.0004
STD SHAMT	0.0075	0.0058	0.0073

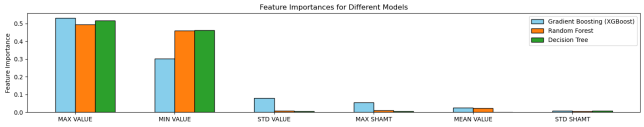

Fig. 12. Feature Importance Across Different Models. MAX VALUE and MIN VALUE remain the dominant predictors across all models.

Across all models, **MAX VALUE** and **MIN VALUE** were the most influential features in predicting institutional holdings. These variables set upper and lower bounds on the total reported value of assets, making them crucial in estimating TABLEVALUETOTAL.

Interestingly, **STD VALUE** (standard deviation of security values) was considered an important predictor in XGBoost but was much less relevant in Random Forest and Decision Tree models. This suggests that XGBoost effectively captures additional variance and non-linear interactions that may not be as emphasized in bagging-based models.

VI. CONCLUSION

This study leveraged the SEC Form 13F dataset to predict institutional asset values using machine learning models, addressing challenges such as high dimensionality, redundant columns, and missing values through extensive feature engineering. The research applied various predictive models, including Linear Regression, Decision Tree, Random Forest, and XGBoost, to assess the effectiveness of ML techniques in estimating institutional portfolio values.

Among the models evaluated, Linear Regression struggled with non-linearity, resulting in poor predictive performance and high errors. The Decision Tree model demonstrated an improvement in capturing complex relationships but suffered from overfitting, limiting its generalization. In contrast, ensemble methods such as Random Forest and XGBoost significantly improved stability and predictive accuracy. While Random Forest marginally outperformed XGBoost in this study, the latter's boosting framework and built-in regularization suggest the potential for further optimization with refined hyperparameter tuning.

The findings emphasize the need for more advanced optimization techniques to enhance predictive performance. Future work should explore hyperparameter tuning strategies such as Bayesian optimization, along with feature engineering refinements, including interaction terms and additional transformations. Further research could also investigate alternative boosting frameworks, such as LightGBM and CatBoost, to assess their efficiency in financial prediction tasks. Additionally, security-level predictions, incorporating CUSIP identifiers, could provide deeper insights into individual asset valuations within institutional portfolios.

Overall, this study highlights the effectiveness of ensemble learning in financial modelling. The results demonstrate that while Random Forest and XGBoost offer strong predictive capabilities, further improvements can be made by refining feature selection strategies and optimizing boosting methods.

REFERENCES

[1] Gompers, Paul A., and Andrew Metrick. 2001. "Institutional Investors and Equity Prices." *The Quarterly Journal of Economics* 116(1): 229–259. https://doi.org/10.1162/003355301556392

[2] Bushee, Brian J. 1998. "The Influence of Institutional Investors on Myopic R&D Investment Behavior." *The Accounting Review* 73(3): 305–333. http://www.jstor.org/stable/248542

[3] Wermers, Russ. 1999. "Mutual Fund Herding and the Impact on Stock Prices." *The Journal of Finance* 54(2): 581–622. https://doi.org/10.1111/0022-1082.00118

[4] Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng. 2005. "On the Industry Concentration of Actively Managed Equity Mutual Funds." *The Journal of Finance* 60(4): 1983–2011. https://doi.org/10.1111/j.1540-6261.2005.00784.x

[5] Jegadeesh, Narasimhan, and Sheridan Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *The Journal of Finance* 48(1): 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

[6] Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *The Review of Financial Studies* 33(5): 2223–2273. https://doi.org/10.1093/rfs/hhz033

[7] Chong, E., C. Han, and F. C. Park. 2017. "Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies." *Expert Systems with Applications* 83: 187–205. https://doi.org/10.1016/j.eswa.2017.04.030

[8] Krauss, Christopher, Xuan Anh Do, and Nicolas Huck. 2017. "Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500." *European Journal of Operational Research* 259(2): 689–702. https://doi.org/10.1016/j.ejor.2016.10.031

[9] Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1): 5–32. https://doi.org/10.1023/A:1010933404324

[10] Chen, Qi, Itay Goldstein, and Wei Jiang. 2010. "Payoff Complementarities and Financial Fragility: Evidence from Mutual Fund Outflows." *Journal of Financial Economics* 97(2): 239–262. https://doi.org/10.1016/j.jfineco.2010.03.016

[11] Tsay, Ruey S. 2010. *Analysis of Financial Time Series*. 3rd ed. Hoboken, NJ: Wiley.

[12] Haixiang, Gao, Li Jia, Song Yang, et al. 2017. "Learning from Class-Imbalanced Data: Review of Methods and Applications." *Expert Systems with Applications* 73: 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

[13] Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley.

[14] Freedman, David A. 2009. *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

[15] Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

[16] Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3

[17] Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29(5): 1189–1232. https://doi.org/10.1214/aos/1013203451