**Hyperparameter Tuning and Model Evaluation Report**

**Predicting Investor Stock Picks**

Team Lead: Hunter Huberdeau

Recorder: Hayden Realmuto

Spokesperson: Amita Patil, Eliel Polanco

**Recap Background & Question:**

Understanding and anticipating the investment decisions of influential figures has long been a subject of interest in financial markets. Warren Buffett, renowned for his disciplined value investing approach, focuses on identifying fundamentally strong yet undervalued companies (Hagstrom, 2013). Investors and financial professionals closely track his moves through SEC filings and media reports, but these sources provide insights only in retrospect. Team Lambda aspires to bridge that gap by leveraging machine-learning to predict which stocks Buffett is likely to buy next, based on historical data and company-level financial indicators. By deploying predictive modeling and analysis, we aim to create a decision-support tool that empowers individual investors with insights traditionally available only after the fact.

The intersection of behavioral finance and predictive analytics presents a compelling opportunity to model Buffett’s investment patterns. While previous research has explored stock price prediction and portfolio performance modeling, few studies have specifically attempted to emulate the buy decision-making process of a known value investor using explainable machine learning models (Fischer & Krauss, 2018). Our approach is novel in that it does not merely assess market trends but seeks to identify the financial characteristics that make a stock appealing to Buffett before he purchases.

We hypothesize that historical financial metrics such as a low price-to-earnings (P/E) ratio, strong revenue growth, and high dividend yield are significant predictors of his stock selections. This is because Buffett prioritizes companies that exhibit financial strength, are undervalued relative to their intrinsic worth, and demonstrate steady, long-term growth potential (Buffett & Cunningham, 2020). If successful, we predict this work could enhance traditional financial analysis by providing predictive insights into high-profile investment strategies, offering a practical tool for investors who seek to align their decisions with Buffett’s time-tested approach.

**Addressing Class Imbalance**

Initially, the dataset exhibited a significant class imbalance, with approximately 38,700 negative samples (controls) compared to only 472 positive samples (Buffett Buys). This imbalance caused the model to favor the negative class, leading to poor model performance across all results particularly precision and recall for identifying Buffett Buys.

To address this issue, we consulted with our professor and implemented a 1:1 matching strategy using Nearest Neighbors (NN). Each positive sample was paired with its closest negative sample based on the following key features:

* **Year**
* **Quarter\_Num**
* **Market Value (%)**
* **PE\_Category\_Num**

This approach preserved the structure of the data while balancing the classes, resulting in improved model calibration and performance.

**Distance Summary Analysis**

The distance summary provides statistical insights into the proximity between each positive sample (Buffett Buy) and its nearest negative sample (Control). Below is an explanation of the key metrics:

|  |  |
| --- | --- |
| **Distance Summary (Matching Quality):** | |
| **Statistic** | **Value** |
| Count | 472 |
| Mean | 0.0038 |
| Std | 0.0205 |
| Min | 0 |
| 25% | 0.0001 |
| Median | 0.0005 |
| 75% | 0.0013 |
| Max | 0.3657 |

A graph with purple lines

AI-generated content may be incorrect.

Figure 1: Matching Distances (Buffett Buy vs Control)

Most matches are close: The mean and median distances are very small, indicating that most positive samples are well-matched with negative samples in the feature space.

Outliers Exist: The maximum distance (0.365736) is significantly larger than the mean, suggesting a few outliers where the matching is less optimal.

Low Variability: The small standard deviation (0.020519) shows that the distances are generally consistent, with only a few exceptions.

This is a significant improvement over our previous data without matching. Overall, the matching quality between Buffett Buy samples and the control group is strong, with most pairs closely aligned in the feature space, as shown by the low mean, median, and standard deviation in the distance summary. While a few outliers exist, indicated by the higher maximum distance, the overall consistency suggests that the matching process was effective and reliable for much of the data.

**Train and Test Split:**

The dataset was split into features (X) and the target (y), where X contains all columns except the Label column, and y contains the Label. A train-test split was performed, allocating 80% of the data for training and 20% for testing, while maintaining the class distribution using stratification.

Three models were trained on the ‘X\_train’ and ‘y\_train’ datasets and saved as ‘.pkl’ files for future use. The models trained include Logistic Regression, Random Forest, and XGBoost.

**Threshold Optimization**: Optimizes the decision threshold based on the F1 score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report at Best Threshold (0.28)** | | | | |
| **Metric** | **Class 0** | **Class 1** | **Macro Avg** | **Weighted Avg** |
| **Precision** | 0.75 | 0.67 | 0.71 | 0.71 |
| **Recall** | 0.62 | 0.79 | 0.7 | 0.7 |
| **F1-Score** | 0.68 | 0.73 | 0.7 | 0.7 |
| **Support** | 95 | 94 | 189 | 189 |
| **Accuracy** | **0.7** | **-** | - | - |

A graph showing different colored lines

AI-generated content may be incorrect.

Figure 2:threshold based on F1 score

By lowering the classification threshold to **0.28** (from the default 0.5), we achieved:

* **Increased Recall for Class 1 (Buffett Buy)** from previous values to **0.79**, meaning the model is **identifying more true Buffett Buy cases** — directly supporting our goal of **not missing key investment opportunities**.
* An overall **F1-Score of 0.73 for Class 1**, showing **stronger balance between identifying and correctly labeling Buffett Buys**.

**Hyperparameter Tuning:**

* **Overfitting Control**:
  + Cross-validation (cv=3) was used during hyperparameter tuning to ensure the model generalizes well to unseen data.
  + The scoring metric chosen was roc\_auc, which balances sensitivity and specificity.
  + Early stopping was implemented for XGBoost to prevent overfitting during training.
* **Improving Model Fit**:
  + RandomizedSearchCV was used to efficiently explore a wide range of hyperparameters.
  + Hyperparameters tuned:
    - **XGBoost**: n\_estimators, max\_depth, learning\_rate, subsample, colsample\_bytree.
    - **Random Forest**: n\_estimators, max\_depth, min\_samples\_split, min\_samples\_leaf, max\_features.
    - **Logistic Regression**: C, penalty, solver.

**SHAP Interpretability:**

This SHAP summary bar plot shows the average impact of each feature on the model's predictions. Debt/Equity and P/E are the most influential features, with the highest mean SHAP values, indicating their strong contribution to the model's output. The "Sum of 28 other features" represents the combined impact of less significant features.

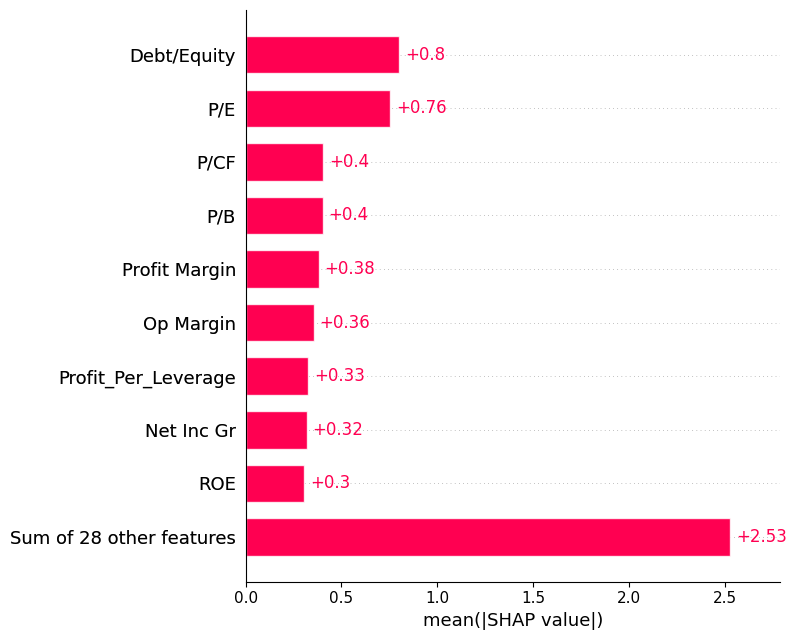


Figure 3:Summary bar plot

A graph of different colored lines

AI-generated content may be incorrect.

Figure 4:SHAP beeswarm plot

This SHAP beeswarm plot shows the impact of various features on the model's predictions. Features like Debt/Equity, P/E, and Profit Margin have the most significant influence, with red indicating high feature values and blue indicating low feature values. These variables are in line with the analytical literature describing Warren Buffett’s investment style.

-----------------------------------------------------------------------------------------------------

* **XGBoost:**

**Model Calibration and Trade-offs**  
After hyperparameter tuning, **XGBoost's ROC AUC Score increased from 0.799 to 0.816**, indicating that the model now **separates the two classes more effectively** across all thresholds. This is a strong indicator of improved model quality.

*Table Reference: "ROC AUC Score" row in the XGBoost Comparison Table*

**Recall vs. Precision Shifts**

* **Class 0 Recall improved** from **0.77 to 0.80**, showing the model became **better at identifying negative/control cases** after tuning.
* However, **Class 1 Recall dropped slightly** from **0.68 to 0.65**, meaning **some Buffett Buys are now missed more frequently**.
* Despite that, **Class 1 Precision increased** from **0.74 to 0.76**, indicating **fewer false positives** for Buffett Buys.

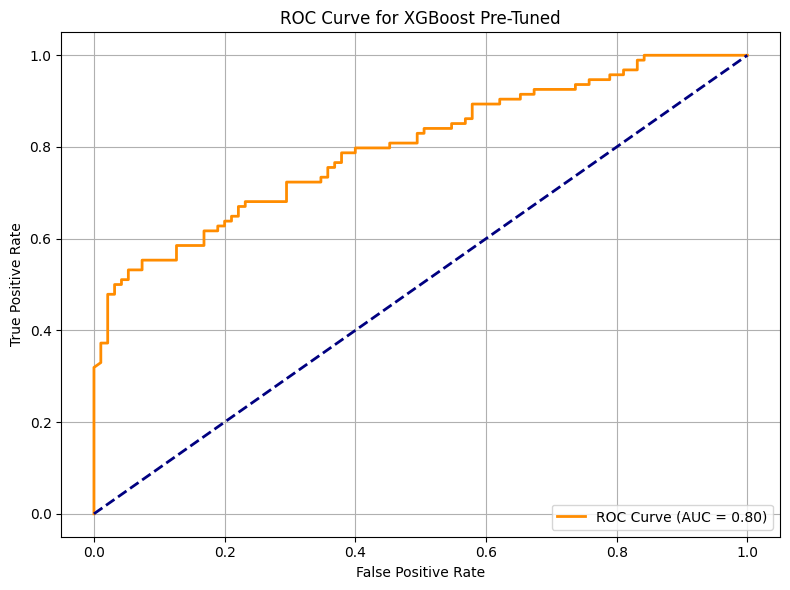
*Table Reference: Rows "Precision (Class 1)" and "Recall (Class 1)"*

This trade-off is often expected in model tuning and can be adjusted based on the relative cost of false positives vs. false negatives.

**Consistency in Accuracy and Macro Metrics**

* **Accuracy remained constant at 0.72**, suggesting overall classification capability is stable.
* **Macro F1-score stayed at 0.72**, meaning the **overall balance between precision and recall did not degrade**.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Pre-Tuned** | **Post-Tuned** |
| **Precision (Class 0)** | 0.71 | 0.7 |
| **Recall (Class 0)** | 0.77 | 0.8 |
| **F1-Score (Class 0)** | 0.74 | 0.75 |
| **Precision (Class 1)** | 0.74 | 0.76 |
| **Recall (Class 1)** | 0.68 | 0.65 |
| **F1-Score (Class 1)** | 0.71 | 0.7 |
| **Accuracy** | 0.72 | 0.72 |
| **Macro Avg F1** | 0.72 | 0.72 |
| **ROC AUC Score** | 0.799 | **0.816** |

 A graph of a curve

AI-generated content may be incorrect. A graph of a graph

AI-generated content may be incorrect. A graph of a graph

AI-generated content may be incorrect.

--------------------------------------------------------------------------------------------------------

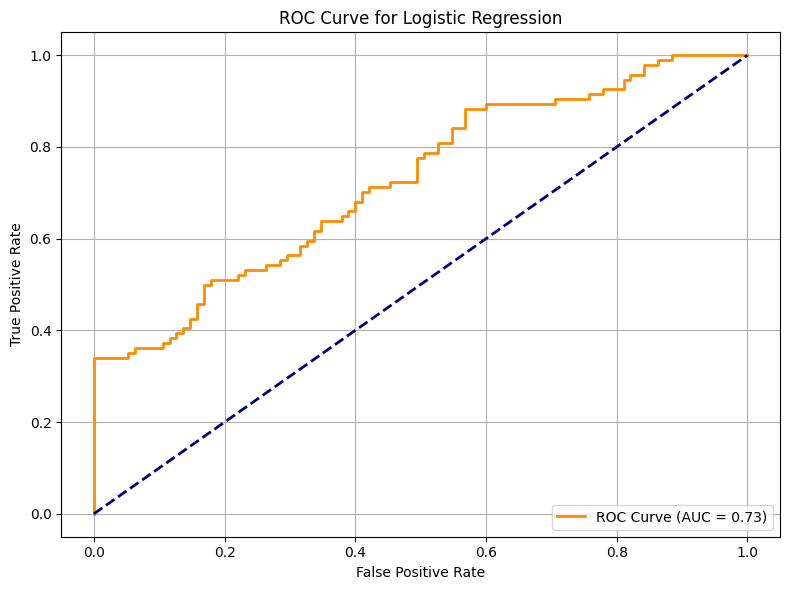
* **Logistic Regression:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Pre-Tuned** | **Post-Tuned** | **Change** |
| **Precision (Class 0)** | 0.56 | 0.64 | **0.08** |
| **Recall (Class 0)** | 0.51 | 0.63 | **0.12** |
| **F1-Score (Class 0)** | 0.53 | 0.63 | **0.1** |
| **Precision (Class 1)** | 0.54 | 0.63 | **0.09** |
| **Recall (Class 1)** | 0.6 | 0.64 | **0.04** |
| **F1-Score (Class 1)** | 0.57 | 0.63 | **0.06** |
| **Accuracy** | 0.55 | 0.63 | **0.08** |
| **Macro Avg F1** | 0.55 | 0.63 | **0.08** |
| **ROC AUC Score** | 0.581 | **0.726** | **0.145** |

**Substantial improvement across all metrics**, especially in ROC AUC, which increased by nearly **0.15**—a significant jump for classification.

**Balanced precision and recall for both classes** after tuning, indicating better generalization and fairness in classification.

A graph of a logistic curve

AI-generated content may be incorrect. 

A graph with a line graph

AI-generated content may be incorrect. A graph of a logistic regression line

AI-generated content may be incorrect.

---------------------------------------------------------------------------------------------------------------

* **Random forest :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Pre-Tuned** | **Post-Tuned** | **Change** |
| **Precision (Class 0)** | 0.71 | 0.71 | No Change |
| **Recall (Class 0)** | 0.81 | 0.77 | −0.04 |
| **F1-Score (Class 0)** | 0.76 | 0.74 | −0.02 |
| **Precision (Class 1)** | 0.78 | 0.74 | −0.04 |
| **Recall (Class 1)** | 0.67 | 0.68 | 0.01 |
| **F1-Score (Class 1)** | 0.72 | 0.71 | −0.01 |
| **Accuracy** | 0.74 | 0.72 | −0.02 |
| **Macro Avg F1** | 0.74 | 0.72 | −0.02 |
| **ROC AUC Score** | 0.825 | 0.822 | −0.003 |

**Performance stayed relatively stable**, with only **minor decreases in most metrics**.

**Slight dip in ROC AUC and F1-scores**, but **Recall for Class 1 (Buffett Buy)** improved by a small margin (+0.01).

Tuning didn’t significantly improve results and may have slightly **reduced generalization**.

A graph of a curve

AI-generated content may be incorrect. A graph of a curve

AI-generated content may be incorrect.

A graph showing a curve

AI-generated content may be incorrect. A graph of a curve

AI-generated content may be incorrect.

**Discussion & Next Steps:**

* 1. XGBoost is the top performer post-tuning, especially with threshold optimization.
  2. SHAP confirmed model decisions align with Buffett-style investing.
  3. Logistic Regression saw the most improvement, showing strong potential despite simplicity.
  4. the performance of the Random Forest model did not significantly improve — and in some cases, it slightly decreased.

**Appendix**

**GitHub repo:** <https://github.com/amita29patil/team_lambda>

**Data Dictionary**

|  |  |
| --- | --- |
| Column Name | Description |
| Market Value (%) | The stock's weight in the portfolio, indicating its relative importance. |
| P/E | Price-to-Earnings ratio, a valuation metric showing how much investors pay per unit of earnings. |
| P/CF | Price-to-Cash Flow ratio, a valuation metric showing the stock price relative to cash flow. |
| P/B | Price-to-Book ratio, a valuation metric comparing the stock price to book value. |
| Debt/Equity | Financial leverage ratio, showing the proportion of debt relative to equity. |
| ROE | Return on Equity, a profitability metric indicating how efficiently equity generates returns. |
| ROA | Return on Assets, a profitability metric showing how efficiently assets generate returns. |
| Profit Margin | The percentage of revenue that turns into profit after expenses. |
| FCF Gr | Free Cash Flow Growth, indicating the growth rate of free cash flow over time. |
| EBITDA Gr | EBITDA Growth, showing the growth rate of earnings before interest, taxes, depreciation, and amortization. |
| Net Inc Gr | Net Income Growth, indicating the growth rate of net income over time. |
| GDP CYOY Index | Year-over-Year GDP growth, reflecting economic growth trends. |
| CPI YOY Index | Year-over-Year Consumer Price Index, a measure of inflation. |
| PPI YOY Index | Year-over-Year Producer Price Index, another measure of inflation. |
| VIX Index | Volatility Index, reflecting market uncertainty and investor sentiment. |
| Sector | The industry or sector to which the stock belongs. |
| Year | The year of the data record, used for time-based analysis. |
| Quarter | The quarter of the year (Q1, Q2, Q3, Q4), capturing seasonality. |
| Label | The target variable indicating whether the stock was purchased (1) or not (0). |

**Codebook**

|  |  |  |
| --- | --- | --- |
| Variable | Description | Type |
| Label | Indicates whether the stock was purchased (1) or not (0). | Label |
| Sector | The sector classification of the stock. | Categorical |
| Quarter | Fiscal quarter (Q1, Q2, Q3, Q4). | Categorical |
| Year | Year of the data record. | Categorical |
| Sector (Grouped) | Grouped industry classification. | Categorical |
| Market Value (%) | Percentage of the portfolio held in this stock. | Numeric |
| P/E | Price-to-Earnings ratio. | Numeric |
| P/CF | Price-to-Cash Flow ratio. | Numeric |
| P/B | Price-to-Book ratio. | Numeric |
| Debt/Equity | Ratio of debt to shareholder equity. | Numeric |
| ROE | Return on Equity. | Numeric |
| ROA | Return on Assets. | Numeric |
| Profit Margin | Percentage of revenue that becomes profit. | Numeric |
| FCF Gr | Free Cash Flow Growth. | Numeric |
| EBITDA Gr | EBITDA Growth. | Numeric |
| Net Inc Gr | Net Income Growth. | Numeric |
| GDP CYOY Index | GDP Year-over-Year Index. | Numeric |
| CPI YOY Index | Consumer Price Index Year-over-Year. | Numeric |
| VIX Index | Volatility Index. | Numeric |
| Div Yld | Dividend Yield. | Numeric |
| Div Pay Ratio | Dividend Payout Ratio. | Numeric |
| Free CF | Free Cash Flow. | Numeric |
|  |  |  |