Preprocessing & Feature Engineering Report

Team Lambda

Team Lead: Hayden Realmuto

Recorder: Eliel Polanco

Spokesperson: Amita Patil, Hunter Huberdeau

***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.

***Method:***

Our prior plan focused on cleaning and standardizing financial fundamentals and economic indicators. Based on EDA results, we found many skewed variables and some sector imbalances. We revised our preprocessing to include log transformations, missing values, Winsorization, quarter cyclic encoding, and categorical conversions. We also applied PCA to explore dimensionality and ran a supervised Elastic Net model using cross-validation.

**What methods did you use to perform preprocessing? Why did you make the choices that you did?**

The preprocessing phase was undertaken to prepare the dataset for machine learning by addressing issues such as missing values, outliers, skewed distributions, and categorical variables. Columns with more than 30% missing values, such as Inv Gr and Work Cap Gr, were removed unless they demonstrated strong predictive power. For columns with moderate missingness (10–30%) in key financial indicators like ROE, P/FCF, and FCF Gr, the Iterative Imputer (MICE) method was applied under the assumption that the data was missing at random (MAR), allowing the retention of important signals. Columns with low levels of missingness were imputed using the mean for numeric variables and the mode for categorical variables, such as Sector. To address outliers and skewed distributions, Winsorization was applied to features with extreme outliers, such as P/E, Debt/Equity, and Profit Margin, by capping values at the 1st and 99th percentiles. Additionally, log transformations using the log1p function were employed to normalize skewed features, making them more suitable for machine learning algorithms that assume normality.

Categorical variables were processed to enhance model performance and prevent overfitting. Rare categories in the Sector variable were grouped into an "Other" category to reduce noise, while logical columns (e.g., TRUE/FALSE) were converted into numeric values (1/0). One-hot encoding was applied to categorical variables like Sector, and the Quarter variable was encoded cyclically using sine and cosine transformations to capture seasonality trends without introducing artificial linearity. Feature reduction was performed by removing highly correlated or redundant features, such as PPI YOY Index (highly correlated with CPI YOY Index) and Debt/EBITDA, based on insights from a correlation heatmap. Duplicate rows were also identified and removed to maintain data integrity and prevent overrepresentation of certain data points, ensuring unbiased modeling. These preprocessing steps collectively enhanced the dataset's quality and ensured its readiness for machine learning applications.

**What methods did you use to perform feature engineering? Why did you make the choices that you did?**

To enhance the dataset's predictive power and align with Buffett’s value-investing principles, several feature engineering methods were employed to create meaningful variables and optimize the dataset for machine learning. A binary feature, COVID\_Period, was introduced to indicate whether a record fell within the 2020–2021 timeframe, capturing the market shock caused by the pandemic and its sector-specific impacts. Additionally, a new ratio, Profit\_Per\_Leverage, was created by dividing ROE by (Debt/Equity + ε) to reflect efficient capital usage, emphasizing strong fundamentals and controlled leverage, which are central to Buffett’s philosophy. The P/E ratio was categorized into interpretable investment groups (e.g., Negative, Low, Good, High, Craziness) and encoded as a numeric variable (PE\_Category\_Num) to enable ordinal modeling. To account for seasonality, sine and cosine transformations of the quarter (Quarter\_sin and Quarter\_cos) were added, capturing cyclic investment behavior without introducing artificial linearity. Furthermore, a lagged version of the COVID\_Period feature (COVID\_Period\_Lag1) was included to model temporal spillover effects of the pandemic on subsequent quarters.

Principal Component Analysis (PCA) was applied to explore the structure of high-dimensional financial variables, revealing latent clusters and identifying redundancy among predictors. While primarily exploratory, PCA provided insights into how valuation metrics like P/E, ROE, and FCF growth aligned, aiding in feature selection. Finally, a baseline Elastic Net regression model was trained using 80/20 stratified sampling to rank feature importance. This approach balanced regularization and feature selection, preventing overfitting and guiding the final selection of predictors before deploying more complex models like XGBoost. These feature engineering choices were made to enrich the dataset with variables that not only aligned with Buffett’s investment philosophy but also enhanced the dataset’s suitability for predictive modeling.

**Handling Imbalanced Target Variable**

The target variable is highly imbalanced, with only about 1.2% of observations representing "stocks Buffett bought" (Label = 1). This imbalance can bias the model toward always predicting the majority class (Label = 0). While we have not yet implemented specific techniques to address this imbalance, we plan to use the XGBoost technique in future iterations. By leveraging its scale\_pos\_weight parameter, we aim to balance the importance of the minority class during training.

Additionally, we may explore other strategies such as SMOTE (Synthetic Minority Oversampling Technique), undersampling, or adjusting the decision threshold to improve recall for the minority class. These approaches will ensure that the model is better equipped to handle the imbalance and accurately predict Buffett's stock purchases.

**Visualizations**

Figure 1 Correlation heatmap of numeric features

**A screenshot of a graph

AI-generated content may be incorrect.**

PPI YOY Index: Highly correlated with CPI YOY Index (positive correlation close to 1). Both represent inflation metrics, and retaining CPI YOY Index is sufficient as it is more commonly used in economic analysis.

Debt/EBITDA: Strong positive correlation with Debt/Equity. Both measure financial leverage, but Debt/Equity is more widely used and interpretable in financial analysis.

FCF Gr\_2: Highly correlated with FCF Gr. Both represent free cash flow growth, and retaining FCF Gr is sufficient to capture this information.

EHUPUS Index : Moderately correlated with GDP CYOY Index. Both represent economic growth trends, but GDP CYOY Index is more standard and interpretable for macroeconomic analysis.

COVID\_Period\_Lag1: Highly correlated with COVID\_Period. Both represent temporal information related to the COVID-19 period, and retaining COVID\_Period is sufficient

In the histogram figures shown below, the features P/E, P/CF, P/B, Debt/Equity, FCF Growth, Profit Margin, and Market Value % exhibit significant right skewness, with most values concentrated near zero and a few extreme outliers on the higher end. To address this skewness and improve the distribution of these variables, we will consider handling the outliers or applying appropriate transformations.

Figure 2: Histograms of numeric features before preprocessing (winsorization and log transformation)

A screenshot of a graph

AI-generated content may be incorrect.

Figure 2 represents the histograms of all numeric features obtained during exploratory data analysis. This data collected was before preprocessing (winsorization and log transformation). The features P/E, P/CF, P/B, Debt/Equity, FCF Growth, Profit Margin, and Market Value % exhibit significant right skewness, with most values concentrated near zero and a few extreme outliers on the higher end. To address this skewness and improve the distribution of these variables, we will consider handling the outliers or applying appropriate transformations.

Figure 3: Data after winsorization

A screenshot of a graph

AI-generated content may be incorrect.

Figure 4: Data after log transformation

A screenshot of a graph

AI-generated content may be incorrect.

Figure 3 and Figure 4 are a transformation of Figure 2 and represent the numeric features after winsorization (Figure 3) and log- transformation (Figure 4). Scale and bin size were increased to add increased effectiveness in our visualizations.

Figure 5: Boxplots of features after log transformation)

A graph with different colored boxes

AI-generated content may be incorrect.

Boxplots were used to identify and confirm the reduction of outliers after Winsorization and log transformations. After log transformation, features such as Profit Margin, Op Margin, FCF Gr, EBITDA Gr, and Net Inc Gr displayed some skewness that was not solved post-processing. More analysis will be needed for processing skewness.

***Results:***

Table 1: PCA result summary

A computer screen shot of numbers

AI-generated content may be incorrect.

The Principal Component Analysis (PCA) revealed that the first few principal components (PCs) account for a significant portion of the variability in the dataset. Specifically, PC1 explains approximately 8.5% of the total variance, followed by PC2 at around 6.8%, and PC3 at 5.3%. Collectively, the first 10 components account for nearly 48.6% of the variance, indicating that dimensionality reduction using PCA can effectively capture much of the data’s structure with fewer features. These principal components are derived from linear combinations of original financial and macroeconomic indicators (e.g., P/E, ROE, Debt/Equity, Free Cash Flow Growth, VIX Index), and their loadings reflect the underlying drivers of variance across investor behaviors and stock characteristics. By focusing on the top components, the dataset becomes more manageable for modeling, while still retaining the most important patterns and relationships.

Figure 7: Cross Validation (CV) results of Elastic Net model

A graph with red dots

AI-generated content may be incorrect.

The graph above represents the cross-validation results of our Elastic Net model, used to tune the regularization parameter lambda (λ). On the x-axis, we see the log values of λ, which control the strength of regularization: smaller λ values (to the right) allow the model to fit more closely to the training data, while larger λ values (to the left) enforce more regularization, simplifying the model. The y-axis shows the binomial deviance, a measure of model error, where lower values indicate better performance. Each red dot corresponds to the mean cross-validated deviance for a particular λ, and the vertical bars represent the standard error at each point. The two vertical dashed lines highlight key λ values: the first marks the λ that results in the minimum deviance (best performance), while the second shows the largest λ within one standard error of that minimum (offering a simpler model with nearly equivalent performance). For our analysis, the leftmost line indicates the most accurate model, while the rightmost line suggests a more regularized version that could generalize better with fewer features. This balance between model complexity and predictive power is critical for ensuring our stock purchase predictions remain both accurate and interpretable.

***Discussion & Next Steps***

Our findings to date reinforce the validity of our original hypothesis: that Warren Buffett’s investment decisions can be predicted using a combination of value-based financial indicators and macroeconomic conditions. The EDA highlighted the dominance of certain sectors and financial ratios, which helped guide our preprocessing decisions and the creation of meaningful features like PE\_Category\_Num, sector dummies, and cyclical time variables. PCA results confirmed that the most influential components of variance were driven by valuation and profitability measures, while the Elastic Net model validated these same variables as key predictors by optimizing performance at lambda ≈ 0.023.

These outcomes align with the assumptions we made in our initial proposal and support our belief that the Buffett Buy Score and engineered variables capture investor intent effectively. The results so far suggest that our preprocessing pipeline is robust and that we are on track to generate a predictive model with practical implications for retail investors and analysts alike.

Next, we will implement additional supervised learning models, including logistic regression and random forests, to benchmark performance. Utilizing these benchmarks will allow for identification of the most influential variables through SHAP value analysis and feature importance rankings to further refine our variable set. We also plan to test time-lagged and interaction features for seasonality and compound effects. The goal is to compare model performance and interpretability to determine the best approach for anticipating investor stock selections in a real-world setting.