Team Lambda Finalized Project Proposal "Predicting Investor Stock Purchases"

**Team Lead:** Amita Patil  
**Recorder:** Hunter Huberdeau  
**Spokespersons:** Eliel Polanco & Hayden Realmuto

**Background & Question**

**Defined Research Question:**  
Can we predict which stocks investors like Warren Buffett are likely to buy next using historical data and company-level financial indicators?

**Motivation and Need:**  
In financial markets, understanding and anticipating the actions of influential investors such as Warren Buffett can provide a valuable edge to both individual investors and financial professionals. Buffett is known for value investing—identifying fundamentally strong but undervalued companies. Our question aims to replicate Buffett's behavior using machine learning, offering a novel decision-support tool to inform investment strategy. This fills a niche at the intersection of behavioral finance and predictive analytics.

**Why It's Worth the Effort:**  
Investors already track Buffett's moves through SEC filings and media reports. However, these sources are retrospective. Our model attempts to go a step further by building a predictive system that suggests potential future investments based on past patterns. If successful, our work could empower smaller investors with insights that are typically available only in hindsight. It also demonstrates how machine learning can enhance traditional financial analysis.

**Novelty:**  
While others have attempted stock price predictions or portfolio performance modeling, our focus is specifically on emulating the *buy decision-making process* of a known value investor using explainable ML models. We believe this angle—forecasting stock selections based on known investor behavior—is relatively unique and original within the academic and applied finance space.

**Hypothesis:**  
Historical financial metrics such as low P/E ratio, strong revenue growth, and high dividend yield are significant predictors of the types of stocks Warren Buffett is likely to purchase.

**Prediction:**  
Stocks that are financially strong, undervalued, and exhibit steady growth are more likely to be added to Buffett's portfolio in a future quarter.

**Data & Methods**

Our primary dataset was compiled from a private financial data provider that aggregates Bloomberg-level stock fundamentals. This includes metrics like P/E ratio, dividend yield, debt-to-equity, margin ratios, and various growth indicators on a quarterly basis. A second dataset was created by scraping investor activity from platforms such as Dataroma, which tracks Buffett’s quarterly buys, sells, and position changes. We also use macroeconomic data sourced from publicly available indices that capture inflation, interest rates, GDP, and consumer sentiment.

The data was acquired in CSV and Excel format. We performed basic cleaning including formatting dates, removing nulls, converting percentages to floats, and aligning quarterly timestamps. The datasets were merged using a combination of stock ticker symbols and quarterly dates. This merged dataset allows us to see stock-level fundamentals and macro conditions at the time an investment decision was made.

**Data Dictionary:**

* **Ticker** – Stock symbol (string)
* **Name** – Company name (string)
* **Quarter, Year** – Fiscal timestamp (string, int)
* **Activity** – Buy/Add/Sell (string)
* **Shares Added** – Number of shares added in a quarter (int)
* **% Change to Portfolio** – Percent change to position weight (float)
* **Label** – Target variable: 1 = stock bought, 0 = not bought (binary)
* **P/E Ratio** – Price-to-Earnings ratio (float)
* **Dividend Yield** – Dividend as a percentage of share price (float)
* **Debt-to-Equity** – Leverage metric (float)
* **Revenue Growth** – YoY revenue change (float)
* **Gross Margin** – Percentage of revenue remaining after COGS (float)
* **Macro Variables** – Federal interest rate, CPI, GDP growth, etc. (float)

Key predictors in our analysis include financial ratios (P/E, D/E), profitability metrics (margins), and macroeconomic context (interest rate, CPI). These predictors will inform our classification models.

**Analysis Plan**

**How We Will Know if the Question Is Answered:**  
We will evaluate the model based on how accurately it predicts whether a given stock was purchased in a given quarter. High performance on classification metrics (accuracy, precision, recall, AUC) on unseen data will confirm that the model captures Buffett's decision patterns. If the most important features identified by the model align with Buffett’s known value investing principles, our hypothesis will also be supported.

**Technical Steps:**

* Finalize dataset and complete missing data imputation where applicable.
* Perform Exploratory Data Analysis (EDA) to identify outliers, distributions, and feature correlations.
* Normalize or standardize numerical variables.
* Encode categorical features such as Activity.
* Split data into training and testing sets (80/20 split).

**Modeling Plan:**

* **Baseline Model**: Logistic Regression (interpretable, good benchmark)
* **Model 2**: Random Forest Classifier (handles non-linearity and provides feature importance)
* **Model 3**: Gradient Boosted Trees (XGBoost or LightGBM for enhanced performance)

We will evaluate using confusion matrices, F1-scores, ROC-AUC curves, and SHAP values for interpretability.

**Language:** Python  
**Libraries:** pandas, scikit-learn, matplotlib, seaborn, xgboost, shap

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