Predicting Earnings and Clustering Performance with Financial and Macroeconomic Data

Yueyao Ren, Peter King, Thomas MacPherson

Overview

This project builds a custom dataset by combining corporate fundamentals from Yahoo Finance (via yfinance) with macroeconomic indicators from the FRED API. The main supervised task is to predict quarterly Earnings-Per-Share (EPS) using features from company earnings reports, enhanced with macroeconomic variables like GDP, inflation, and interest rates. Unsupervised tasks will apply dimensionality reduction and clustering to uncover patterns among firms with similar financial structures. Key challenges include limited historical coverage, heterogeneous financial statements, and constrained time-series modeling. Careful feature engineering (e.g., growth rates, normalization) will ensure comparability and capture temporal effects.

Related Work

Several past studies have attempted to predict EPS or other company fundamentals using ML models. The closest to our proposal (Divo et. al., 2024) designed ML models to predict revenues, income, and equity (but not EPS) from firm-level data. Our project distinguishes itself from this and other past studies in its feature engineering and selection methods, particularly in the use of dimensionality reduction techniques, and including macroeconomic variables.

Part A: Supervised learning

- We will create an aggregated dataset using yfinance and the FRED API as explained above, for both supervised and unsupervised tasks. We will obtain data for ~2500 diverse firms for robustness of the model. If missing data is <15% we will discard those observations; otherwise we will use multivariate imputation or sector-level averages. Our analysis will be restricted to the past five quarterly financial statements obtainable through the yfinance API. Draft dataset available via google drive link.
- We will test different supervised models including linear regression (Ridge and LASSO), ensembles (Random Forest), and deep learning (neural networks). We will enhance the dataset with "manual" feature engineering and use PCA/SVD to address multicollinearity.
- We will use root mean squared error (RMSE) and the coefficient of determination (R^2) to evaluate regression models, and visualize uncertainty using residual and spaghetti plots.

Part B: Unsupervised learning

- General goals are dimensionality reduction (PCA/SVD) and manifold learning (MDS/UMAP) for feature engineering in future stock price prediction models and human cluster visualization, respectively. We will apply clustering (KMeans, DBSCAN) to identify groups of firms with similar financial structures. Finally, we will evaluate these clusters based on ground-truth stock performance compared to a benchmark (S&P500).
- The data preparation for this section will require normalization/standardization across companies and encoding of categorical variables (sector, dividend status, etc.).
- We will use the ratio of between group to within group variance and a silhouette score to assess how well our data clusters, and we will evaluate clusters based on the ground-truth stock performance (homogeneity/completeness/adjusted Rand Index).
- We will develop biplots, silhouette plots, and MDS/UMAP cluster visualizations.

Team Planning:

  - **Melody:** Data acquisition | feature engineering | evaluation | visualization | report
  - **Peter:** Feature engineering | modeling | evaluation | visualization | report
  - **Thomas:** Pipeline | feature engineering | evaluation | visualization | report
- **Timeline:**
  - **Week 1:** Data acquisition | cleaning | pipeline development
  - **Week 2:** Feature engineering | baseline model comparison
  - **Week 3:** Parameter Optimization | Unsupervised learning | Evaluation
  - **Week 4-5:** Finalizing Visualizations | Report Generation

Conclusion:

This proposal balances predictive modeling (EPS forecasting) with structural analysis of companies based on their financial performance. It extends existing literature by employing unsupervised techniques for dimensionality reduction to produce an enhanced feature set for use in supervised techniques. While the potential of this project is great, we recognize that the limitations of free data lead to potential constraints on model performance. Even incremental gains in forecasting accuracy through machine learning have the potential to generate substantial market outperformance relative to traditional analyst benchmarks.

References:

Divo, F., Endress, E., Endler, K., Kersting, K., & Dhami, D.S. (2024). *Forecasting company fundamentals* (arXiv:2411.05791) [Preprint]. arXiv.

Kuryłek, W. (2025). Is the inclusion of a broad set of explanatory variables relevant in EPS forecasting? Evidence from Poland. *Bank i Kredyt*, 56(3), 253-282. https://doi.org/10.5604/01.3001.0055.1459.

Singh, G., Thanaya, I. (2023). Predicting earnings per share using feature-engineered extreme gradient boosting models and constructing alpha trading strategies. *Int. j. inf. tecnol*. 15, 3999–4012. https://doi.org/10.1007/s41870-023-01450-0