Feature Augmentation and Economic Group Importance in Cross-Sectional Return Prediction

Yanzhong(Eric) Huang, Sravya Madireddi, Thara Venu

2025-11-18

1. Motivation

Machine learning methods have shown strong predictive power in asset pricing (Gu, Kelly, & Xiu, 2020), largely due to their ability to capture nonlinear and high-dimensional relationships among firm characteristics. However, while these models achieve high out-of-sample accuracy, the **economic interpretation of features** remains limited.

This project aims to bridge the gap between predictive accuracy and interpretability by investigating improvements of models by adding **peers-based features**. By comparing model performance with and without these features, we can **quantify the economic importance of feature groups**. Specifically, we use the Open Asset Pricing dataset — a large-scale, cleaned collection of firm-level monthly characteristics — to systematically assess how different economic categories of features contribute to model performance.

2. Objectives

1. **Enhance the predictive feature set** by incorporating structural and relational information, including:
   - **Peer group features**, such as peer-averaged characteristics and peer average returns
     - Define peer groups via K-means
     - Define peer groups via industry classification
2. **Train multiple predictive models** to forecast future stock returns using the augmented dataset:
   - Feedforward Neural Network (nonlinear, flexible)
3. **Assess the economic relevance** of feature groups by comparing model performance with and without peer-based features.

3. Data and Feature Engineering

3.1 Dataset

- **Source:** Open Asset Pricing dataset (monthly frequency, CRSP–Compustat merged sample).
- **Observation unit:** Firm-month (identified by `permno`, `date`).
- **Target variable:** Next-month excess return.
- **Feature space:** firm characteristics + peer-based feature.

3.2 Peer-based Feature Augmentation

Group Type	Weighted Method
k-means clustering	Equal weight
K-means clustering	Market-cap weight
K-means clustering	Distance weight (inverse of Euclidean distance)
Industry classification	Equal weight
Industry classification	Market-cap weight
Industry classification	Distance weight (inverse of Euclidean distance)

Example, for k-means clustering with distance weight:

1. Cluster firms into K groups using K-means on standardized feature space.
2. Calculate the peer average of each feature for each firm, weighted by the inverse of the Euclidean distance to other firms in the same cluster.
3. $\text{PeerFeature}_i = \frac{\sum_{j \in Cluster(i), j \neq i} \frac{1}{d_{ij}} \cdot Feature_j}{\sum_{j \in Cluster(i), j \neq i} \frac{1}{d_{ij}}}$

Where d_{ij} is the Euclidean distance between firm i and firm j in the feature space.

4. Modeling Approach

4.1 Algorithms

Feedforward Neural Network

- Multi-layer architecture with dropout and batch normalization
- Flexible representation learning on augmented features
- Using **Optuna** for hyperparameter tuning (layers, units, learning rate, etc.)

4.2 Model Training Setup

- **Evaluation metrics:**
  - Predictive R-squared (cross-sectional)
  - Mean Squared Error (MSE)
  - Rank IC (Spearman correlation between predicted and realized returns)
  - Portfolio backtest (optional, decile spread returns)

5. Feature Group Importance Experiment

To evaluate **economic interpretability**, we will compare model performance with and without specific feature groups:

1. Train the baseline model with **all features**.
2. Adding peer-based features to the baseline model, retrain and evaluate.
3. Compare performance metrics to assess the contribution of peer-based features.

6. Expected Contributions

- **Methodological:** Introduces relational (peer-based) augmentation to traditional firm characteristic datasets.
- **Empirical:** Quantifies the predictive importance of peer-based features in cross-sectional return prediction.
- **Pedagogical:** Demonstrates a replicable, interpretable machine learning pipeline bridging academic factor research and modern ML methods.

7. Timeline

Week	Task
1	Data cleaning and feature grouping
2	Peer clustering and feature augmentation
3	Model setup and baseline training
4	Group ablation experiments
5	Performance comparison and visualization
6	Report writing and presentation preparation

8. Deliverables

- Python notebooks (data processing, modeling, visualization)

- Feature importance and ablation analysis report
- Presentation slides summarizing model results and insights

9. References

- Gu, Kelly, & Xiu (2020). *Empirical Asset Pricing via Machine Learning.* Review of Financial Studies.
- Green, Hand, & Zhang (2017). *The Characteristics that Provide Independent Information about Average U.S. Stock Returns.*
- Bryzgalova, Pelger, & Zhu (2023). *Principal Components of Characteristic Portfolios.*
- Kelly, Pruitt, & Su (2019). *Characteristics Are Covariances: A Unified Model of Risk and Return.*