Final Project Presentation

Finance with Big Data

Luca D'Ambrosio, Filippo De Min, Filip Juren

December 6, 2024

Overview

1. **Research Question**
   - **1.1 Initial Motivation**
   - **1.2 New Framework**
2. **Data**
3. **Methodology**
   - **3.1 Predictive Strategy**
   - **3.2 Models**
   - **3.3 Performance Measurement and Robustness Checks**
   - **3.4 Summary**
4. **Results**
   - **4.1 Overall Performance**
   - **4.2 Portfolio Evaluation**
   - **4.3 Variable Importance**
5. **Discussion And Limitations**
6. **References**
7. **Appendix**

Table of Contents

1. **Research Question**
   - 1.1 **Initial Motivation**
   - 1.2 **New Framework**
2. Data
3. Methodology
   - 3.1 Predictive Strategy
   - 3.2 Models
   - 3.3 Performance Measurement and Robustness Checks
   - 3.4 Summary
4. Results
   - 4.1 Overall Performance
   - 4.2 Portfolio Evaluation
   - 4.3 Variable Importance
5. Discussion And Limitations
6. References
7. Appendix

1. First Idea

- Our initial goal for this project was to connect together the papers:
  - ”Empirical Asset Pricing Via ML” by Gu, Kelly and Xiu (2019)
  - ”Is there a Replication crisis in Finance?” by Kelly et. al.

- We wanted to take advantage of the large dataset available from the latter using the methodologies of the former

2. Motivation

Our research motivation is two-fold:

- Can factors from the "factor zoo" provide predictive power in a non-linear machine learning algorithm? Shouldn't factors' usefulness be decided by the model themselves?
- Can factors be used for empirical asset pricing as an alternative of Gu et al.' extensive 900 firm' specific predictors?

3. Adversial Results

- Using Gu et al.' models, together with their hyperparameters range, we would always find OOS R-Squared measures close to 0, sometimes negative, indicating that on average our model predicted worse than the sample mean.

4. Second idea

- When we tried to move from a regression setting to a classification setting (predicting if stocks would go up or down), we noticed that our algorithm was achieving close to perfect accuracy (95%). This was happpening because the majority of the returns were positive and hence our model learned to classify the majority of observations as positive.
- When analyzing the model' outputted probabilities, we noticed that for observations for which the model was very confident about an upward movement, it was rarely wrong. Instead, when the predicted probability was closer to 50%, it made more errors.
- This let us understand that we could leverage on the certainty levels of the predictions to select the best stocks each month and test the results empirically by simulating portfolios.

Table of Contents

1. Research Question
    - 1.1 Initial Motivation
    - 1.2 New Framework
2. **Data**
3. Methodology
    - 3.1 Predictive Strategy
    - 3.2 Models
    - 3.3 Performance Measurement and Robustness Checks
    - 3.4 Summary
4. Results
    - 4.1 Overall Performance
    - 4.2 Portfolio Evaluation
    - 4.3 Variable Importance
5. Discussion And Limitations
6. References
7. Appendix

2.1 Data Collection

We extracted the ”Global Factor Data” from the Wharton database Through the following query:

```
SELECT id, eom, excntry, gvkey, permno, size_grp, me, {vars}
FROM contrib.global_factor
WHERE common = 1
  AND exch_main = 1 # prominent exhanges
  AND primary_sec = 1 # if 2 or more stocks, chose the main one
  AND obs_main = 1 # one obs. x month
  AND excntry = ’USA’ #
  AND date > ’{date}’ # > 1995
```

Filters were provided by the authors as well as the **vars** list, which includes all 153 factors used in their paper. We subset the data to only include U.S. Stocks after 1995, and later excluded smaller stocks due to computational constraints.

2.2 Dataset Description

Variables

1. Identifiers
2. Accounting Features
3. Market Based Features
4. Other indexes

Total Observations: 320,812
Unique Companies: 4,988
Number of Features: 166
[1]

- Identifiers include the CRSP stock id, the date, the Fama-French industry id etc.
- The Accounting features include Equity, Assets, Various growth measures etc.
- The Market Based features include lagged cumulative returns, dividends etc.
- The indexes are constructed by the authors based on the literature.

[1] For any doubt or further descriptions refer to the **Official Documentation**

Table of Contents

1. Research Question
   - 1.1 Initial Motivation
   - 1.2 New Framework
2. Data
3. **Methodology**
   - **3.1 Predictive Strategy**
   - **3.2 Models**
   - **3.3 Performance Measurement and Robustness Checks**
   - **3.4 Summary**
4. Results
   - 4.1 Overall Performance
   - 4.2 Portfolio Evaluation
   - 4.3 Variable Importance
5. Discussion And Limitations
6. References
7. Appendix

3.1.1 Training, Validation and Test sets

- We build our training, validation and test sets as follows:
  - **Training**: 1995 - 2017
  - **Validation**: 2018 - 2021
  - **Test**: 2022 - 2023

- We decided on this split to mimic Empirical Asset Pricing' splits despite having fewer years of training data.
- Ideally we would need to retrain the model fully to evaluate a subsequent year. For computation constraints, we test our model on the subsequent 2 years.

3.1.2 - Binary Outcome Variable

- We build our binary target variable in the following way:
  - **1**: If next month's returns are greater than 20%
  - **0**: Otherwise
- Our target variable is indeed highly imbalanced with a frequency of 1s ranging from 4% to 5% across sets. Overall model performance is expected to be weak but it doesn't matter with this strategy.

3.1.2 - Binary Outcome Variable

- The reasoning behind this approach is to set a harsh threshold for the model to train it to predict larger returns

- However, in an attempt to validate this procedure, we will provide a Regression setting, where we predict stock returns and select the top performing ones according to the algorithm.

- Going one step further, we also train a model in a regression setting using 5-fold cross validation.

3.2.1 Models - XgBoost

- The paper by Gu et. al. uses **Gradient Boosting regression trees (GBRT)**. However, in our classification setting, we opt for **XgBoost** instead, as it represents the literature's consensus on the best model for classification on tabular data.
- We tune the hyperparameters on the validation set and test the performance on the test set:

Finetuned or relevant Parameters (others are set to default)

- **learning_rate**: 0.01
- **n_estimators**: 1443
- **max_depth**: 2
- **subsample**: 0.6
- **colsample_bytree**: 0.8
- **colsample_bylevel**: 0.8
- **colsample_bynode**: 0.8
- **objective**: ”binary:logistic” (cross-entropy)

Note: The number of estimators is the result of early stopping and was not specified through a grid search. It represents the number of iterations until no reduction in validation loss was recorded.

3.2.2 Models - Neural Networks

Architecture

Feed-forward networks with **1 to 5 hidden layers** using pyramidal reduction:

- ◇ from NN1: [128] to NN5: [128, 64, 32, 16, 8]
- ◇ Each model includes **Batch Normalization** and **Dropout (0.2)** to prevent overfitting.

Model Evaluation and Analysis:

- **Out-of-Sample R²:** Measures the proportion of variance in the test set that is predictable from the model, indicating how well the model generalizes to unseen data
- **Out-of-Sample Accuracy**: Evaluate the accuracy of the model on the test set; with target variables y greater than 0.2.

Differences from Gu et al. (2020):

- **Feature Dimensionality:** Less predictors.
- **Focus:** Trained for a classification task

3.3.1 Performance Measurement

1. For each observation in the unseen test set, the model predicts the probability that the stock' will do +20% return next month.

2. For each month we select the top 2 % outputted probabilites (around 20 stocks).

3. We initialize a portfolio of value 100 and each month we invest in equal weights the budget among the top-selected stocks.

3.3.1 Performance Measurement

4. At the end of the month, we sell all the stocks, we compute the new budget for our portfolio and reinvest it all in the new top 2% stocks selected for that month.

5. We iterate this procedure over the 2-year period of the test set.

6. As a benchmark we also plot the evolution of Fama-French 3 factors' model using the same methodology and the average return of the pool of stocks in our dataset.

3.3.1 - Performance Measurement

Summary of the strategy for the classification setting

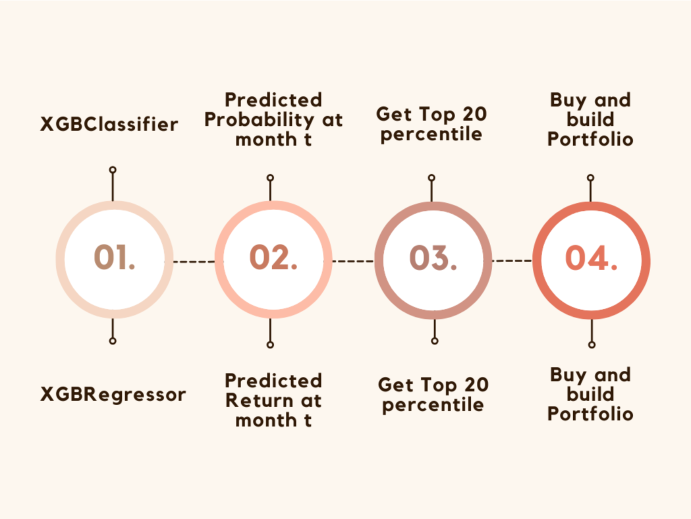
Strategy

- **Signal**: If this signal is strong (high $\boldsymbol{\hat{P}(return > 0.2)}$), we buy the stock.
- **Buy**: Each month we buy the top 2% stocks suggested by the model (*equally weighted*)
- **Sell**: We sell at the end of each month and use that budget to buy the newly suggested stocks.

3.3.2 Robustness Checks

- Concerned by the short time-spawn of our initial results, we train a second model on year priors to 2020 and evaluate it on a test year grouping years 2021-2022. We then retrain a model in a similar fashion and test it on 2019 and 2020.
- We re-explore our initially unsuccessful XgBoost model in a regression setting over 2022-2023.
- We go one-step further and re-train the XgBoost model in a regression setting but we include 5-fold cross validation during training.

3.4 Summary of our strategy



XGBClassifier	Predicted Probability at month t	Get Top 20 percentile	Buy and build Portfolio
01.	02.	03.	04.
XGBRegressor	Predicted Return at month t	Get Top 20 percentile	Buy and build Portfolio

Table of Contents

1. Research Question
   - 1.1 Initial Motivation
   - 1.2 New Framework
2. Data
3. Methodology
   - 3.1 Predictive Strategy
   - 3.2 Models
   - 3.3 Performance Measurement and Robustness Checks
   - 3.4 Summary
4. **Results**
   - **4.1 Overall Performance**
   - **4.2 Portfolio Evaluation**
   - **4.3 Variable Importance**
5. Discussion And Limitations
6. References
7. Appendix

4.1.1 Overall Model Performance

XgBoost Metrics

Metric	Train	Validation	Test
Precision	0.23	0.20	0.14
Accuracy	0.91	0.91	0.85
F1-score	0.31	0.27	0.20

INTERPRETATION

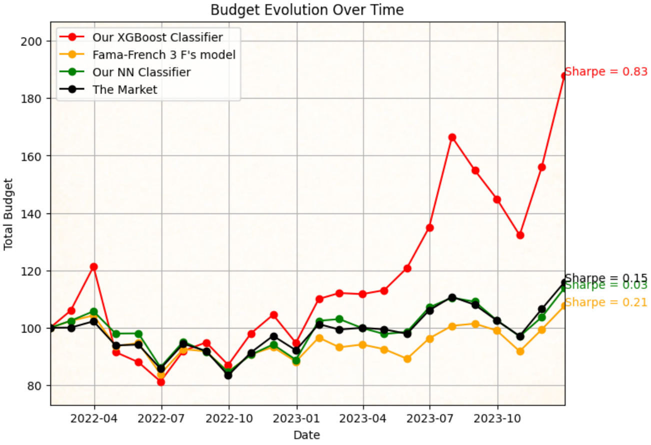
- If we look at the disparity between the F1 score and accuracy, it follows that the model tends to predict with all zeroes and get high accuracy, but low precision. This is due to the very high threshold that we have set.
- With our Neural Network attempt we obtained similar though less performant metrics.

4.1.2 Overall porfolio performance

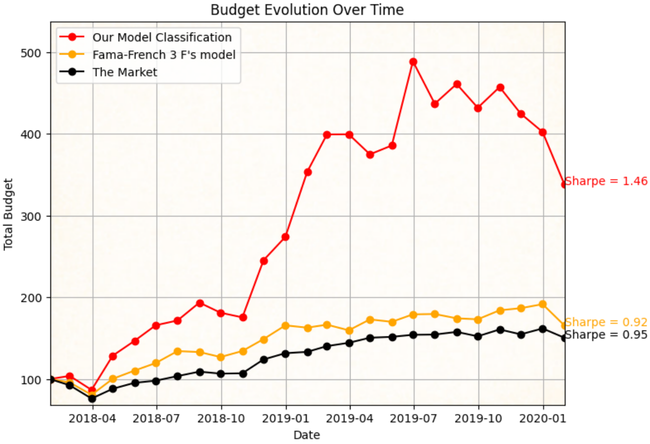
Statistic	Market	Top 1000	Bottom 1000
Mean Returns %	0.77	3.45	- 1.47
Min Returns %	-56.37	-51.41	-28.01
Max Returns %	149.37	149.37	28.96
Standard Deviation %	11.44	18.74	5.83

The stocks predicted by our model display average monthly return which beats the market by almost **3 percentage points**.
We also notice that the 1000 stocks for which the predicted probability is the lower display mean returns **2 percentage points** below the market.

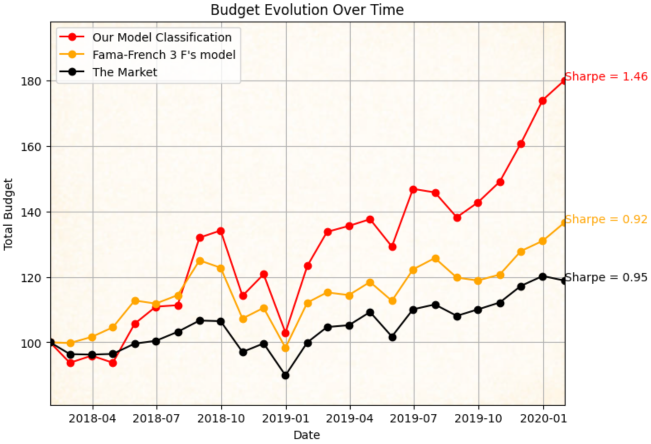
4.2.1 Classification Setting over 2022-2023



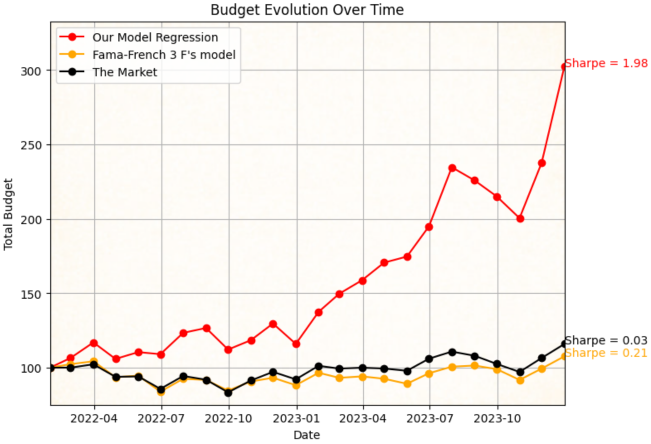
4.2.2 Classification Setting over 2020-2021



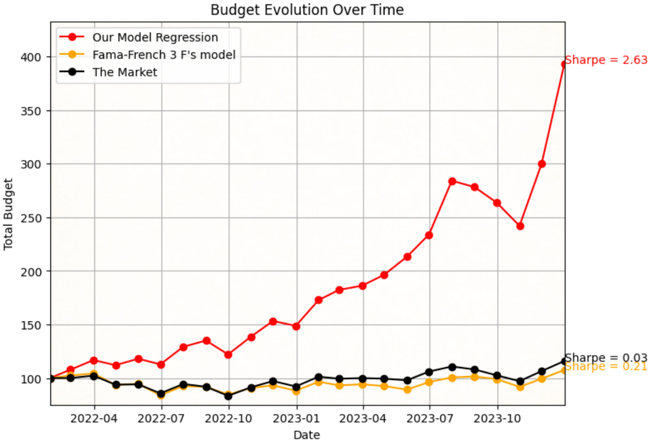
4.2.3 Classification Setting over 2018-2019



4.2.4 Regression Setting over 2022-2023



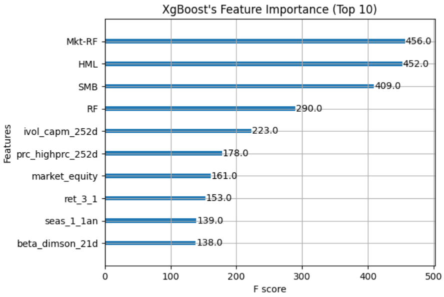
4.2.5 Regression Setting cross-validated over 2022-2023

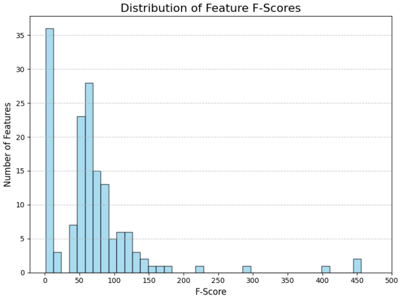


Portfolio Results

- Generated portfolios display great results with Sharpe ratio ranging from 0.8 to 2.6.
- Results are consistent through a 6-year evaluation period.
- Results are even better in a regression setting, suggesting to move away from a classification setting and its threshold value problem.
- Cross validating further improves results, consistent with machine learning' theory.

4.3 Variable Importance





Variable Importance Analysis:

- Interestingly, XgBoost detects itself Fama-French factors as the most important ones, which are also the ones that receive more support from the literature.
- The F score for most features is greater than 50 meaning that they appear in at least 50 decision nodes of the 1443 trees of the model. The 37 features receiving close to 0 for the F score are the dummies created for the industry sector.

Table of Contents

1. Research Question
   - 1.1 Initial Motivation
   - 1.2 New Framework
2. Data
3. Methodology
   - 3.1 Predictive Strategy
   - 3.2 Models
   - 3.3 Performance Measurement and Robustness Checks
   - 3.4 Summary
4. Results
   - 4.1 Overall Performance
   - 4.2 Portfolio Evaluation
   - 4.3 Variable Importance
5. **Discussion And Limitations**
6. References
7. Appendix

Discussion

- Overall, we are satisfied with our portfolio's performance across the test years.
- We managed to consistently beat the market and the baseline model (FF3)
- XgBoost seems to yield better results compared to Neural Networks, which displays returns close to FF3
- Given variable importance measures, we found that almost all factors play a role in the return prediction, albeit many of them being discarded by the literature.
- Altough, we are not able to predict returns, we are able to detect good stocks and to trade successfully.

Limitations

1. The results obtained seem a bit **unrealistic**, but after a thorough inspection of the code and the methods, we are unable to detect any error. However, this procedure might indeed be successful due to the leverage on big data, extensive research in creating factors and advanced machine learning techniques.

2. Certain **firm features** might not be available to the public in during the current month (which we would need to compute the best stocks)

3. **Transaction costs** might hinder substantially the trading strategy.

Table of Contents

1. Research Question
   1.1 Initial Motivation
   1.2 New Framework
2. Data
3. Methodology
   3.1 Predictive Strategy
   3.2 Models
   3.3 Performance Measurement and Robustness Checks
   3.4 Summary
4. Results
   4.1 Overall Performance
   4.2 Portfolio Evaluation
   4.3 Variable Importance
5. Discussion And Limitations
6. **References**
7. Appendix

References

1. JENSEN, T.I., KELLY, B. and PEDERSEN, L.H. (2023), Is There a Replication Crisis in Finance?. J Finance, 78: 2465-2518. https://doi.org/10.1111/jofi.13249
2. Shihao Gu, Bryan Kelly, Dacheng Xiu, Empirical Asset Pricing via Machine Learning, The Review of Financial Studies, Volume 33, Issue 5, May 2020, Pages 2223–2273, https://doi.org/10.1093/rfs/hhaa009

Table of Contents

1. Research Question
   - 1.1 Initial Motivation
   - 1.2 New Framework
2. Data
3. Methodology
   - 3.1 Predictive Strategy
   - 3.2 Models
   - 3.3 Performance Measurement and Robustness Checks
   - 3.4 Summary
4. Results
   - 4.1 Overall Performance
   - 4.2 Portfolio Evaluation
   - 4.3 Variable Importance
5. Discussion And Limitations
6. References
7. **Appendix**

A1 - Data Pre-Processing Pipeline

A. General Cleaning

Remove small companies, dates in 2024, add monthly return and industry sector.

B. Creation of dependent variable

We construct $return_{t+1}$ by shifting the return column, carefully inserting NaNs for the last observation of every stock (time wise) since we don't have any more periods after. We manually create NaNs also when a company does not have next month's return

C. Adding Fama French Factors

We download the historic US monthly time series of FF3 from Professor Kenneth French's website and merge them to our dataset.

A1 Data Pre-Processing Pipeline

D. Dummies

Get dummies for categorical variables: firm size and industry sector.

E. Missing Values

Drop features with too many missing values. Impute remaining missing values with stock-level median. If an observation has a lot of variables for which no value is recorded drop the observation is dropped.

F. Outliers

Drop the 99th percentile and the 1st percentile

G. Feature Normalization

Normalize the Covariates using Mix-Max Scaling