COMS 4995 - Project Proposal

Team Members:

David Davila-Garcia
Yuyi Tang
Ziyao Tang
Rain Wei
Shrinjay Kaushik

1. Introduction

In the dynamic and complex financial environment, comprehending the relationships between pivotal economic indicators and investment yields is essential for devising knowledgeable investment strategies. In sight of this consideration, our team has decided to utilize the JPMaQS Quantamental Indicators dataset, applying a variety of techniques including but not limited to regression analyses and time series, to explore the relationships between Quantamental indicators and the selected target variable DU05YXR_VT10 (return on fixed receiver position).

The objectives of this project are:

1. Explore and illuminate the correlations between the target variable and Quantamental indicators to unravel the hidden patterns and correlations within the data.
2. Be familiar with the entire data processing workflow, including preprocessing, exploratory data analysis, model training and evaluation, result analysis, and visualization.

The boundaries and scope of the project are the prediction of the target variable and the evaluation of relevant methods and algorithms. It will be largely empirical, and more advanced theoretical analysis will not be the focus of this project.

2. Background/Significance

JPMaQS is a quantitative fundamental dataset developed by JP Morgan along with Macrosynergy for training algorithmic trading models [4]. As JP Morgan noted on the website, the quantamental indicators they produced in the dataset are the building blocks for macro quantamental algorithmic trading [4]. Prior studies conducted by Macrosynergy suggested a correlation between a macroeconomic cycle, which is a type of macroeconomic indicator, and asset class returns. Macrosynergy has concluded this research with the following results: macroeconomic cycle length correlates negatively with equity index future returns and positively with FX forward returns [5].

Knowing that macroeconomic cycle lengths can impact asset class returns, our project aims to perform a similar statistical analysis to explore if other macroeconomic indicators, such as growth trends and customer information trends, can predict the return on the fixed receiver position.

This dataset provides clean and structured market data for the developers to train the model. Our project aims to conduct regressions and time series analysis to explore the correlations between quantamental indicators such as growth trends and customer information trends and our target variable DU05YXR_VT10 (return on fixed receiver position). Our result can benefit market participants and algorithm traders as it can explain how the macroeconomic indicators potentially affect the target returns. [6] Quantitative traders can leverage our research to build more accurate trading algorithms. Regarding social benefits, our research results can be used to more accurately predict the market and reduce the risk of the financial crisis [6].

3. Dataset Description

Dataset Link

This dataset was curated by Macrosynergy and JP Morgan, consisting of quantamental ("quantitative-fundamental") indicators to develop trading decision support algorithms. Past quantamental datasets have suffered from missingness, particularly for date variables, "value errors, undocumented distortions, and structural breaks" [source: Kaggle Dataset Website]; as such, models trained on these datasets are subject to information leakage and other forms of biases. This dataset was curated to address the issues of missingness, with dates spanning Jan 2000 to Dec 2022. This specific dataset from Kaggle is a subset of a much larger dataset and contains only 24 quantamental indicators, compared to 5000+ indicators for the entire dataset. In total, the dataset contains 3,350,271 rows and seven features.

Features:

- real_date: the date that data was recorded (year-month-day)
- cid: specific currency (ex: USD)
- xcat: 25 quantamental indicators
    - Our target variable for this project is the quantamental indicator DU05YXR_VT10
        - This indicator estimates the sensitivity of an investment's price to changes in interest rates.
- value:
    - Specifically, the value for DU05YXR_VT10 represents the estimated returns for a given investment, assuming risk equals 10%.
- grading: "denoting a grade of the observation, giving a metric of real-time information quality" [1]*
- eop_lag: "referring to days elapsed since the end of the observation period" [1]*

- mop_lag: “referring to the number of days elapsed since the mean observation period” [1]*

No missing values and errors exist, and the dataset is publicly accessible and available. We will do further analysis to determine whether there are outliers in the dataset.

4. Hypothesis

Process of Hypothesis Testing:

1. Defining the Hypothesis.
2. Setting the decision-making criteria
3. Calculating the statistics (z-test, p-test, p-value, etc.) for the selected sample
4. Based on the statistical tests' results, we will either accept the null hypothesis or reject it and accept the alternative hypothesis.

Null Hypothesis (H0):
There is no statistically significant difference in the impact of quantamental indicators on return on fixed receiver position (DU05YXR_VT10) in the panel dataset.
Alternative Hypothesis (H1):
There is a statistically significant difference in the impact of quantamental indicators on return on fixed receiver position (DU05YXR_VT10) in the panel dataset.

With the help of this hypothesis, we can make conclusions on whether or not the quantamental indicators can help make risk-free or at least low-risk algorithmic trading decisions.

We expect that there will be a statistically significant difference in the impact of quantamental indicators on return on fixed receiver position (DU05YXR_VT10) in the panel dataset based on our literature search. We expect that predictors such as “real GDP growth rate” will statistically correlate with the target variable.

5. Proposed Methods

- **Data Preprocessing & Exploratory Data Analysis:**
  We will load the dataset and examine the predictors and the response variables. If necessary, we will clean and preprocess the dataset by replacing null values and checking the data types of each variable. Since this is time-series data, we will create compound keys combining time and ‘cid’ as the dataframe index to make data retrieval more efficient.

  We will then conduct exploratory research using the dataset, exploring the time dimensional and cross-sectional information. This includes plotting each time-series

macro trend in the dataset, such as inflation rates, private credit expansion, IRS yields, etc., computing mean and standard deviation of features, and examining correlations between features. During the explorative analysis stage, we may create new features or signals by doing operations with existing ones. For example, we may create a column of compound interest rates or moving averages. We may also delete features if they correlate highly with another feature. We may downsample the data by reducing the time series frequencies if it is deemed helpful. We will also standardize the features for training and testing purposes. Lastly, we will split the dataset into training (60%), validation (20%), and test sets (20%). We acknowledge that the data splitting method for time-series data may differ from other datasets. We will explore how to do this correctly.

- **Machine learning algorithms:**
  Our ML models use prediction loss minimization and mean squared error (MSE) as the objective function. Since the target response is continuous, we will use **ordinary least squares (OLS)** as a baseline method. This method will create a linear model that uses the linear combination of macro factors on a given day or over a rolling window to predict the return on fixed receiver position as of a given day. We may add **l1 and l2 regularizers** to create sparse models or models with better generalization power. We will also use **ARIMA models**, such as autoregressive models and moving-average models, to capture the autocorrelation in data. In addition, we will use **neural networks**, including LSTM and RNN, which may better incorporate the effects of long-term dependencies.
- **Hyperparameter tuning and evaluation:**
  We will use the validation set for hyperparameter tuning. There are many hyperparameters in each model mentioned above, e.g., penalty rate in regularized least squares and learning rate in neural networks. We will create a pipeline and perform a grid search of different hyperparameter values. We will choose the sets of hyperparameters that demonstrate the highest prediction accuracies. We will then evaluate our models on the test set and report relevant results by comparing the test accuracies of different models. There may be evaluation metrics unique to each model above. For example, with OLS, we can assess the F statistics of each predictor to see whether it has a linear association with the response variable. We will perform these analyses if we see fit (this may also be part of our exploratory data analysis).
- **Risk minimization:**
  Some risks associated with our proposed project and methods include:
  - The poor predictive power of predictors. we may find after doing preliminary analysis that the predictors have low predictive power. We will mitigate this by combining different features after doing a fundamental analysis of the meaning of these features. It will require us to learn more about the macroeconomic factors and the fundamental relationship between these factors and our target variable - return on fixed receiver position. This process may take some time, so we will start this stage early to ensure enough time.
  - Poor performance of proposed models: in this case, we will broaden the scope of models and explore other time-series models.

6. Tentative Timeline

- Data cleaning and preprocessing (due Sun 10/15)
- Explorative analysis (due Sun 10/29)
- OLS model and regularized least squares (due Sun 11/5)
- Deliverable #2 (due 11/7)
- ARIMA model (due Sun 11/12)
- Neural networks (due Sun 11/19)
- Hyperparameter tuning and evaluation (due Sun 11/26)
- Other explorations and report drafting (due Sun 12/3)
- Deliverable #3 - Finalize and submission of report and code (due 12/5)

7. References

[1]* https://www.kaggle.com/code/macrosynergy/jpmaqs-with-statsmodels
[2] https://otexts.com/fpp2/bootstrap.html
[3] https://www.kaggle.com/datasets/macrosynergy/fixed-income-returns-and-macro-trends/data
[4] https://www.jpmorgan.com/insights/global-research/markets/macrosynergy-quantamental-system
[5] https://research.macrosynergy.com/macroeconomic-cycles-and-asset-class-returns/
[6] https://research.macrosynergy.com/
[7] https://www.kaggle.com/code/macrosynergy/panel-regression-with-jpmaqs-python

NOTES:
* Could not find a better data description other than this site. Documentation on Kaggle is poor.