Capstone Project Proposal

Name: Ji Hoon Park

Objective

[Describe the overall objective of the project. If necessary, explain the context and motivation.]

In October 2024, Goldman Sachs predicted that the S&P 500, a key stock market index, would only grow about 3% per year on average over the next 10 years. This is much lower than the 13% average growth it had over the previous decade. They based this prediction on the fact that stock prices had gotten very high, and most of the market's gains were coming from just a few big tech companies. At the time, people were very excited about artificial intelligence, and tech stocks were driving the market up, but Goldman Sachs warned that this excitement might not last forever.

I was very shocked by this prediction because I am an investor of the S&P myself. I then read the Global Strategy Paper published by the firm and soon figured out that the forecast is not surprising at all. The predictive model they have made was perfectly rational, solidifying their bearish expectation of the stock market.

Inspired by Goldman's forecast, I will be building a predictive model to forecast the 10-year U.S. Treasury yield on a monthly basis for the next decade. The model will use five key variables: the 10-Year Breakeven Inflation Rate, GDP growth, Nonfarm Payroll Growth, Credit Spreads, and the S&P 500 Annual Returns. These variables will form a well-rounded foundation for understanding and predicting Treasury yields. I will explain further in my project report.

Impact

[In a few sentences, delineate the potential impact of the project]

The 10 year Treasury yield is a benchmark for other key rates including the mortgage rate and corporate bond yields. Creating this model will benefit individual investors because they will be able to reasonably expect returns and refer to it as a future benchmark. Furthermore, the model will help average investors on how to think about interest rates and markets in a rational manner.

Dataset(s)

[List your data sources with links to them. If you have already uploaded them to your capstone repository on GitHub, please mention the location. In addition, briefly discuss the datatypes and the reliability of the data.]

I will collect past data of the five variables from the Federal Reserve Economic Data (FRED) database, World Bank database, and financial APIs like S&P Global and Yahoo Finance.

10-Year Breakeven Inflation Rate:

- Source: FRED
- Data Type: Monthly time series, percentage.
- Reliability: High

GDP Growth:

- Source: World Bank, FRED
- Data Type: Quarterly and Annual percentage changes
- Reliability: High

Nonfarm Payroll Growth:

- Source: FRED
- Data Type: Monthly time series, job growth in thousands
- Reliability: High

Credit Spreads:

- Source: FRED
- Data Type: Monthly time series, basis points
- Reliability: High

S&P 500 Annual Return:

- Source: Calculated from S&P Global or financial APIs like Yahoo Finance
- Data Type: Monthly time series, percentage
- Reliability: High

Approach

[Talk about how you plan on approaching this capstone through several steps. List the steps below.]

1. **Data Collection and Preprocessing**
    - Collect datasets from FRED, S&P Global, or other reliable sources.
    - Align time series data by converting quarterly data (e.g., GDP growth) to monthly data as much as possible
    - When working with variables that are reported less frequently
        - Use Linear or Spline interpolation

2. **Feature Engineering**
    - Create lagged variables to account for delayed effects (e.g., 1-month, 3-month lags).
    - Generate rolling averages or smoothed trends for each variable.
    - Add derived features such as the yield curve spread (10-Year minus 2-Year Treasury yields)
3. **Exploratory Data Analysis (EDA)**
    - Visualize historical relationships between the variables and the 10-year Treasury yield.
    - Perform correlation analysis and stationarity tests.

Elevator Pitch & Github Repo

4. **Modeling**
    - Build predictive models using a combination of techniques:
        - **Multiple Linear Regression:** Simple and interpretable baseline
        - **Random Forest:** Non-linear relationships and feature importance
        - **Time-Series Models:** Multivariate or univariate sequential forecasting.
    - Evaluate model performance using cross-validation and metrics like R^2 and RMSE.
5. **Backtesting and Validation**
    - Test the model on historical data to assess accuracy
    - Compare predictions with actual yield movements during out-of-sample periods

Mock Presentation & Final Github Repo

6. **Report and Presentation**
    - Summarize findings and actionable insights in a comprehensive report.
    - Create clear visualizations and a dashboard for forecasting results.

Poster Session & Video Presentation

Timeline

[Edit the following example timeline:
This is a rough timeline for this project:
- (3 Weeks) Data Automation
- Collect data from sources
- Align datasets by timestamp and convert quarterly data to monthly using interpolation.

- (3 Weeks) Feature Importance
- Create lagged variables, rolling averages, and derived metrics (e.g., yield curve spread).
- Perform exploratory data analysis to visualize relationships and identify key drivers of the 10-year Treasury yield.

- (4 Weeks) Modeling

- Build and test models, including Multiple Linear Regression, Random Forest, and Time-Series techniques (e.g., ARIMA, VAR).

- (2 Weeks) Combine satellite images with covariate features

- Integrate preprocessed variables and features into a single dataset for modeling.
- Test for multicollinearity and other potential data issues.

- (1 Weeks) Compiling results

- Compare predictions with actual yield movements during out-of-sample periods.
- Summarize findings with tables, graphs, and performance metrics.

- (1 Week) Writing up the report

- Write a comprehensive report summarizing objectives, methods, results, and insights.
- Incorporate visualizations and key takeaways for clarity.

- (1 Week) Poster and Final Presentation

- Design and print the poster
- Practice and refine the final presentation

Possible Issues

[List some of the prospective challenges and issues and discuss how you envision overcome them]

1. **Data Gaps or Missing Values:** Some datasets may have incomplete records due to reporting delays
2. **Overfitting in Models:** Models may fit the training data too closely, capturing noise instead of actual patterns
3. **Non-Stationary Time Series:** Economic and market data often have trends or seasonality, which can make it difficult for models to identify consistent relationships
4. **Interpreting Results for Investors:** Complex models and technical outputs may be difficult for non-experts to understand and apply effectively. I will try the best in my report to explain the variables and why these variables are used to predict the yields