Incrementally Learning Predictive Models of the S&P500 Index Based on Macroeconomic Indicators

Name: Ran Arino

Student ID: 153073200

School: Seneca Polytechnic

Department: Student at School of Software Design & Data Science

email: rarino@myseneca.ca

Course Name: Predictive Analysis

Course ID: BDM500NAA.05379.2237

Professor: Dr. Tamanna Eini Keleshteri

Submission Date: December 12, 2023

Table of Contents

Introduction

Building robust predictive models of the stock index could be a primary objective for investors and market participants with backgrounds in mathematics, finance, economics, and computer science. One of the critical factors to predict the stock index would be a myriad of economic indicators or statistics released from the government or institutions. Comprehensive stock analyses with those economic data will be widely known as macroeconomic analysis or fundamental analysis. Especially, the majority of market participants track the S&P500 with the US economy in the world, and most of them will strive to recognize the deeper relationships between the performance of stocks and the trends of the economy.

In this project, the primary objective is to develop incrementally-learning predictive models for the S&P500 index from a macroeconomic approach and to assess those models from various perspectives. These models leverage economic indicators sourced from the Federal Reserve Economic Data (FRED), Federal Reserve Bank of St. Louis. Also, this project strives to explore various insights from multiple models by different conditions. The intended outcome is to develop a reliable predictive model that can anticipate the S&P500 index's performance for the upcoming one to six months, based on the newly released economic indicators on a monthly basis. This model will provide valuable insights that can guide investment decisions and risk management strategies. Here are research questions to achieve the outcomes.

- How will the future performance of the S&P 500 index be changed by different conditions?
    - Moving averages against the target values.
    - Scopes – number of the most recent data to be considered for parameter updates.
    - Number of months that the target values are predicted ahead.
    - Models: linear regression, logistic regression, classification and regression tree (CART).
- Which economic indicators are likely to significantly affect the future performance of the S&P500 index, and how those impacts have been changed over time?
- How accurately can each model predict the performance of the S&P500 index from one to six months ahead based on the latest economic indicator?

Given the project targets, the intended audience includes stakeholders in finance, economics, and stock market fields who are interested in the future performance of the major US stock index;

more specifically, individual traders, and institutional traders in mutual funds, investment banks, pension funds, and insurance companies. These individuals require accurate and timely predictions to make data-driven decisions about their investment strategies. However, considering that the dataset focuses on the economic data and stock market index, the project might face several challenges listed below:

- *Distribution of the economic data*:
  All feature values attempt to convert to a normal distribution, but the histogram of several economic indicators might not change to a clear bell-shaped curve. For example, one of the macroeconomic data, the unemployment rate, generally fluctuates around 3-5% during a stable economy and spikes only in a recession or depression. It is expected that the shape of the histogram is highly skewed toward the right side, even if those values are transformed to the logarithmic scale and applied to the standardization.
- *Lag of Economic Data Release*:
  Every data consists of the same month but note that all economic data are arranged according to the actual statistical outcome of each month, rather than the release date. In other words, there is a time lag (around 1 month) between the real-time data, such as the S&P500 and the 10-year Treasury yields, and the economic data, such as the Consumer Price Index and Unemployment Rate. Therefore, for example, when we completely acquire the whole economic data in September, the real time data will be close to the end of October.
- *Market Volatility*:
  Market prices are determined by human decisions, so their emotional actions could cause unpredictable actions. Therefore, the volatility or randomness in the market makes it challenging to explain the market behaviors (Thakkar & Chaudhari, 2021).
- *Complexity of the World*:
  The stock market is susceptible to a myriad of influences, including complex economic interdependencies, unexpected economic disruptions, and erupting geopolitical tensions (Sundar et al., 2023). These factors could make it difficult for the development of accurate predictive models.

Datasets

The following list items show all the column data of the original data and the metadata of each, such as data types, brief descriptions, and resources:

- **Date**: string -> showing the end dates of each month; the format of “yyyy-mm-dd”.
- **Year**: int -> Year of date by integer type.
- **Month**: int -> Month of date by integer type.
- **SP500**[*1*2]: float -> S&P 500 Index (ticker code: ^GSPC) [(1)].
- **MY10Y**[*2]: float -> Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity [(2)].
- **CPI**[*1*3]: float -> Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [(3)].
- **CSENT**[*1*3]: float -> Consumer Sentiment [(4)].
- **IPM**[*1*3]: float -> Industrial Production in Manufacturing [(5)].
- **HOUSE**[*1*3] float -> New One Family Houses Sold [(6)].
- **UNEMP** [*3] -> float -> Unemployment rate [(7)].
- **LRIR**: float -> Long-term Real Interest rate, calculated by MY10Y – %YoY CPI

*1: The data will be converted to the Year-over-Year(YoY) percent growth rate – the percent changes compared to the same month of the previous year.

*2: Data in each row shows the closing value on the last trading day of each month; it will be generally applied if the original data is daily base.

*3: Data in each row is adjusted by the actual result of each month, rather than the released month.

Data Preprocessing

(1): Dropping Incomplete Data

If one or more attributes are missing, the rows are dropped out. *Figure 1* shows the results of applying the data drops.

	Date	Year	Month	MY10Y	CPI	CSENT	IPM	HOUSE	UNEMP	SP500
265	1972-02-29	1972	2	6.04	41.4	92.8	36.2194	711.0	5.7	106.570000
268	1972-05-31	1972	5	6.05	41.6	88.6	36.9454	677.0	5.7	109.529999
271	1972-08-31	1972	8	6.42	41.9	95.2	37.5429	773.0	5.6	111.089996
274	1972-11-30	1972	11	6.28	42.4	90.7	38.8702	735.0	5.3	116.669998
277	1973-02-28	1973	2	6.64	43.0	81.9	40.3702	737.0	5.0	111.680000
280	1973-05-31	1973	5	6.93	43.9	77.0	40.6467	660.0	4.9	104.949997
283	1973-08-31	1973	8	7.25	45.0	72.0	40.7324	566.0	4.8	104.250000
286	1973-11-30	1973	11	6.69	45.9	76.5	41.7624	547.0	4.8	95.959999
289	1974-02-28	1974	2	7.01	47.3	61.8	41.2533	539.0	5.2	96.220001
292	1974-05-31	1974	5	7.52	48.6	72.1	41.3365	590.0	5.1	87.279999
295	1974-08-31	1974	8	8.11	49.9	64.4	40.9558	492.0	5.5	72.150002
298	1974-11-30	1974	11	7.64	51.5	59.5	39.4681	450.0	6.6	69.970001

***Figure 1**: Dropping all rows with incomplete attributes.*

(2): Removing the Incomplete Time Series of Data

If the dataset is grouped by the “Year” column and visualized to show the amount of data for each year, we can find that the data for the first couple of months are incomplete, shown in *Figure 2*. In the project, various moving averages or the Year-over-Year (YoY) percentage growth are applied to the dataset. In other words, the consistency of the data is critical in terms of the time periods. Therefore, those data will be removed, *Figure 3* shows its result.

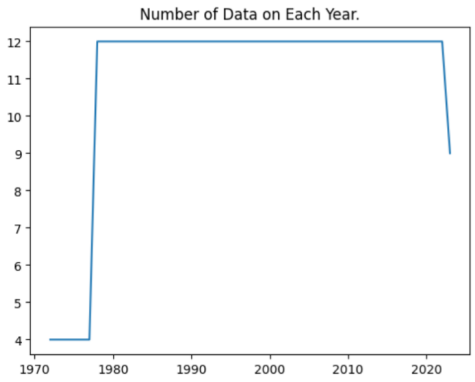

***Figure 2**: Number of data grouped by “Year” column*

	Date	Year	Month	MY10Y	CPI	CSENT	IPM	HOUSE	UNEMP	SP500
0	1978-01-31	1978	1	7.94	62.7	83.7	43.7471	795.0	6.4	89.250000
1	1978-02-28	1978	2	8.04	63.0	84.3	43.9139	791.0	6.3	87.040001
2	1978-03-31	1978	3	8.15	63.4	78.8	44.6899	814.0	6.3	89.209999
3	1978-04-30	1978	4	8.24	63.9	81.6	45.3905	864.0	6.1	96.830002
4	1978-05-31	1978	5	8.42	64.5	82.9	45.6235	857.0	6.0	97.239998

Figure 3*: Result of dealing incomplete time periods of data.*

(3): Attributes Modifications

Figure 4 shows the result of the data frame after the following modification processes.

- Copy the current “SP500” column to “SP500_Price”.
- Create categorical data “SP500_Rise”, which implies whether the S&P500 rises (1) or falls (0) compared to the same month of the previous year on the price base.
- Converting all features and a target value of the “SP500” column to the %YoY scale.
- Dropping the missing data generated by applying %YoY.

	Date	Year	Month	MY10Y	CPI	CSENT	IPM	HOUSE	UNEMP	SP500	SP500_Price	SP500_Rise
0	1979-01-31	1979	1	8.95	9.250399	-13.859020	7.862464	-5.157233	5.9	11.966387	99.930000	1.0
1	1979-02-28	1979	2	9.17	9.841270	-12.336892	7.786828	-8.596713	5.9	10.615806	96.279999	1.0
2	1979-03-31	1979	3	9.11	10.252366	-13.197970	6.418676	-2.579853	5.8	13.877365	101.589996	1.0
3	1979-04-30	1979	4	9.35	10.485133	-19.117647	2.997984	-13.425926	5.8	5.091398	101.760002	1.0
4	1979-05-31	1979	5	9.06	10.697674	-17.852835	3.917937	-15.169195	5.6	1.892230	99.080002	1.0

Figure 4*: Data frame after modifying attributes*

Here are comparisons of other metadata between the original dataset and the dataset after data cleaning.

- Dimensions: (888, 10) -> (538, 12)
- Total Missing Values: 820 -> 0
- Date Ranges: Jan 1950 ~ Sep 2023 -> Jan 1979 ~ Oct 2023

Data Observations

(1): Fundamental Statistics

In *Figure 5*, There are two basic statistics from the pd.Dataframe.describe() method, particularly focusing on numerical data. The first table is from the original dataset, and the second table from the cleared dataset.

- To compare the two tables, the greatest differences would be scaling changes, which caused smaller data distributions.
- More specifically, the standard deviation was drastically revised to a lower value from the original data in the consecutively growing prices (SP500) and the cumulative data (CPI and HOUSE).
- Therefore, the fluctuation of each feature is not the same, but the dataset was partially standardized.

	MY10Y	CPI	CSENT	IPM	HOUSE	UNEMP	SP500
count	743.000000	886.000000	642.000000	622.000000	730.000000	886.000000	887.000000
mean	5.872759	122.482098	85.549688	74.287014	655.771233	5.731603	781.078714
std	2.990855	83.914413	12.899156	23.722896	208.186081	1.711548	1041.129126
min	0.550000	23.510000	50.000000	35.254500	270.000000	2.500000	17.049999
25%	3.855000	34.750000	76.025000	50.395325	515.000000	4.400000	88.895000
50%	5.620000	110.600000	89.200000	82.272800	628.500000	5.500000	249.220001
75%	7.630000	193.675000	94.975000	98.345950	765.000000	6.800000	1215.405029
max	15.840000	307.619000	112.000000	106.420200	1389.000000	14.700000	4766.180176

	MY10Y	CPI	CSENT	IPM	HOUSE	UNEMP	SP500
count	538.000000	538.000000	538.000000	538.000000	538.000000	538.000000	538.000000
mean	5.769591	3.521535	0.436273	1.879105	1.551672	6.110409	10.051391
std	3.386562	2.755466	13.486383	4.901427	19.954754	1.782045	15.899805
min	0.550000	-1.958761	-41.520468	-19.513084	-50.534759	3.400000	-44.756241
25%	2.942500	1.901664	-6.547973	-0.287418	-10.868578	4.800000	2.071541
50%	5.070000	2.844395	0.299743	2.346995	2.734806	5.700000	11.647982
75%	7.997500	4.129159	6.687060	4.720490	13.591408	7.200000	19.604311
max	15.840000	14.592275	47.582205	20.819818	88.200590	14.700000	53.714506

Figure 5*: Basics Statistics from original (top) and cleaned data frame (bottom)*

(2): Correlation Matrix

Figure 6 shows the correlation matrix, focusing on only numerical data.

- Relatively stronger correlations with the S&P500 index were consumer sentiment (CSENT), then industrial production (IPM) and housing sales (HOUSE) will follow; they are all positively correlated. Based on only the correlation, the consumer sentiment could have the strongest impact to the S&P500 index.
- The weakest groups are the 10-year Treasury yields, the Consumer Price Index, and unemployment rate.
- Focusing on the inter-relationships among economic indicators, the strongest (positive) correlation shows between the 10-year Treasury yields and Consumer Price Index.
- Considering that both economic indicators show a lower correlation with the target variable, and both indicators have a relatively high association with each other, we can create the derived feature and conduct the feature reduction.

	MY10Y	CPI	CSENT	IPM	HOUSE	UNEMP	SP500
MY10Y	1.000000	0.592226	0.131251	0.159912	-0.106573	0.322408	0.040220
CPI	0.592226	1.000000	-0.229705	0.023940	-0.362021	0.025027	-0.019945
CSENT	0.131251	-0.229705	1.000000	0.297866	0.404014	0.175331	0.419360
IPM	0.159912	0.023940	0.297866	1.000000	0.133360	-0.187744	0.397629
HOUSE	-0.106573	-0.362021	0.404014	0.133360	1.000000	0.126207	0.378720
UNEMP	0.322408	0.025027	0.175331	-0.187744	0.126207	1.000000	0.029060
SP500	0.040220	-0.019945	0.419360	0.397629	0.378720	0.029060	1.000000

***Figure 6**: Correlation Matrix*

(3): Feature Reduction / Selection

A new feature will be derived from two economic indicators: the 10-year Treasury yields and Consumer Price Index. Here is the basic information:

- Symbol: LRIR

- Name: Longer-term Real Interest Rate
- Calculation: The 10-year Treasury yields – Year-over-Year (YoY) Consumer Price Index changes

The real interest rate is calculated by the nominal rate minus the inflation rate (Wolla, 2012). In this project, the 10-year Treasury yields (MY10Y) are regarded as the nominal interest rate; according to an article from IMF, the yields of 10-year Treasury Notes reflect the real Treasury yields, future economic growth, and inflation break-even rate, future inflation rate among investors, then the sum of both elements provides the nominal rate (Adrian, 2021). In addition, considering that the YoY percent change in the consumer price index (CPI) is widely known as the popular measure of the inflation rate, the subtraction of the nominal rate (MY10Y) by the inflation rate (CPI) is the longer-term real interest rate (LRIR); its new derived data is added to the dataset instead of removing MY10Y and CPI.

Figure 7 shows the new correlation matrix after dropping two attributes and adding a derived data. Although the correlation between LRIR and SP500 is still low, the complexity of the model would be mitigated by reducing the features.

	CSENT	IPM	HOUSE	UNEMP	LRIR	SP500
CSENT	1.000000	0.297866	0.404014	0.175331	0.380726	0.419360
IPM	0.297866	1.000000	0.133360	-0.187744	0.168055	0.397629
HOUSE	0.404014	0.133360	1.000000	0.126207	0.224958	0.378720
UNEMP	0.175331	-0.187744	0.126207	1.000000	0.361453	0.029060
LRIR	0.380726	0.168055	0.224958	0.361453	1.000000	0.067550
SP500	0.419360	0.397629	0.378720	0.029060	0.067550	1.000000

***Figure 7**: Correlation matrix after feature selection*

(4): Histograms (Skewness):

Figure 8 displays the histogram of all features.

- CSENT and HOUSE are slightly positive skewness, which imply that distribution is slightly skewed toward the right side.
- LRIR and SP500 are negative skewness, which means a mild left-skewed distribution.

- IPM shows a strong left skewness, although the shape of the distribution is close to the normal distribution.
- On the other hand, UNEMP shows a strong right skewness; the distribution shape is substantially different from a normal distribution.
- In order to handle imbalanced data, the log scale is applied to UNEMP data; *Figure 9* shows the histogram of log scaled UNEMP data; the distribution shape is not close to the normal distribution, but the skewness was mitigated.
- Since two economic data, UNEMP and LRIR, are not applied by the YoY growth changes (the original math unit is already a percentage), the shape of those distributions might not be close to normal distribution. This is one of the expected concerns and challenges that was mentioned in the introduction part.

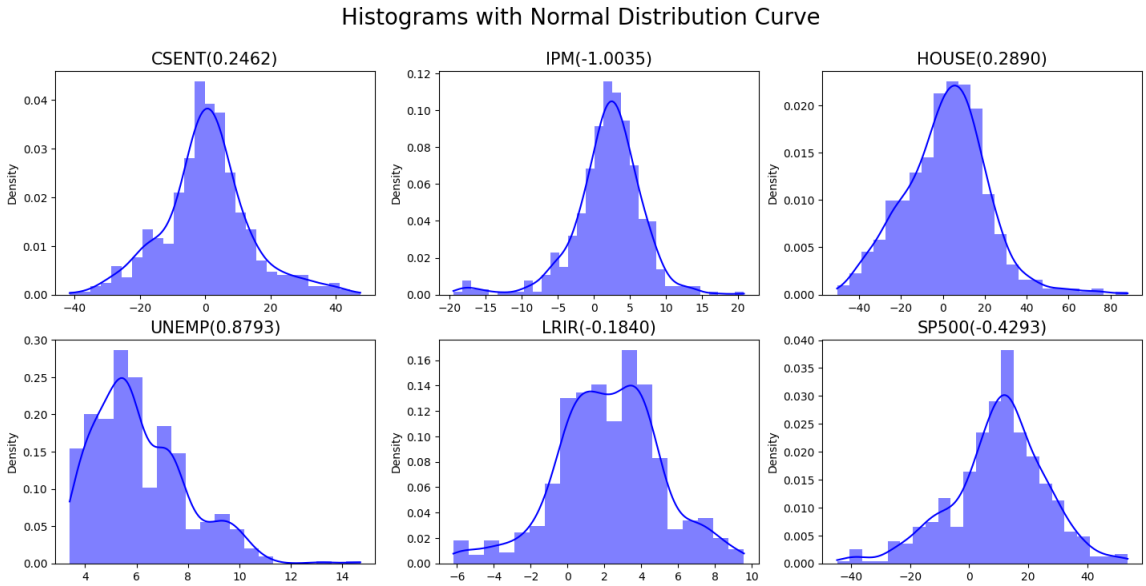

***Figure 8**: Histograms of all features with skewness*

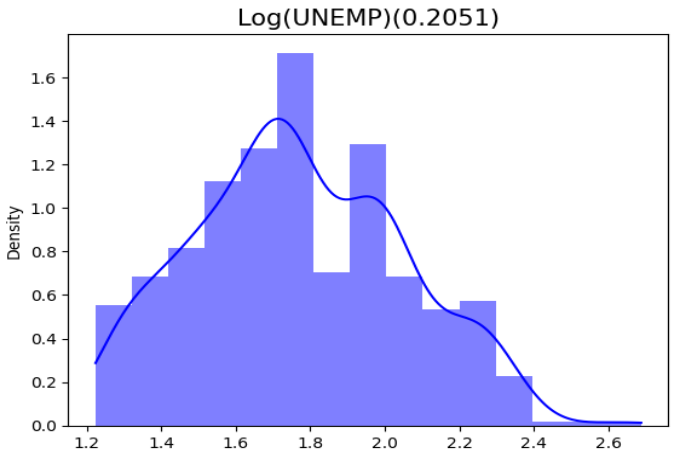

***Figure 9**: Histogram of Log(UNEMP)*

(5): Correlation plots

Figure 10 shows the scatter plots between two variables and the histogram of each data.

- There is a linear relationship (despite wider bands) between the S&P500 index and three economic indicators (consumer sentiment, Industrial production, and housing sales).
- Focusing on the consumer sentiment (CSENT), almost all features are positively correlated with CSENT, which means that its sentiment data could be an important factor of other economic data.
- A possible negative correlation could be found between unemployment rate (UNEMP) and the industrial production in manufacturing (IPM). This relationship will be straightforward; the more people are working, the higher industrial productions.

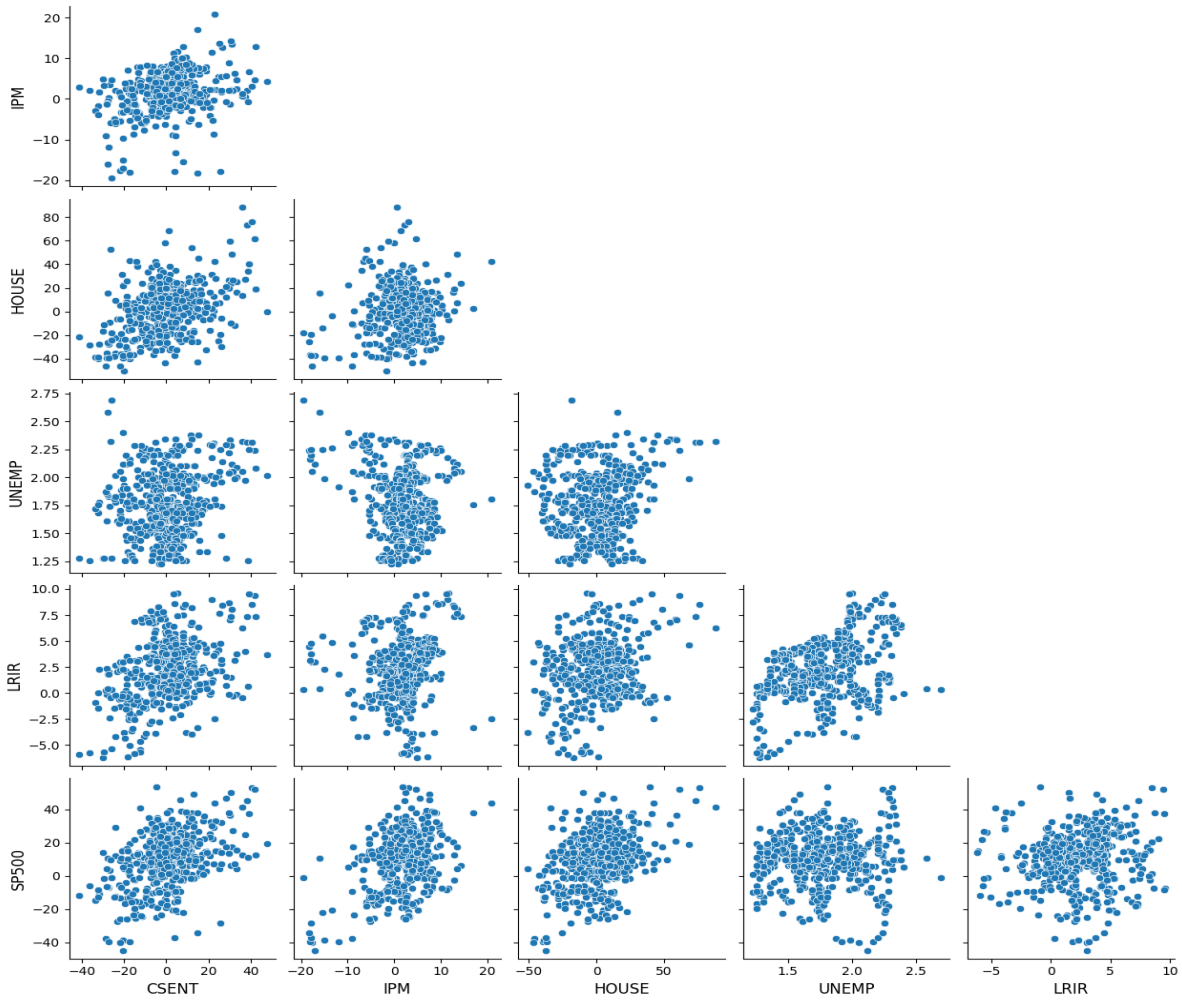

***Figure 10**: Correlation plots among numerical features*

(6): Trend and relationships among features

Figure 11 shows a comparison of line charts between S&P500 and all economic indicators.

- Consumer Sentiment (CSENT):

  We can observe a lot of periods when the S&P500 and consumer sentiment moved together. If we focus on the huge bottom of the S&P500, such as in 2010 and 2022, consumer sentiment is likely to move in advance before raising the stock index.

- Industrial Production (IPM):

  Similar to consumer sentiment, industrial production in manufacturing has been correlated to the S&P500 index historically. Once production begins to rebound, the stock is likely to show

a strong surge in the short periods.

- Housing Sold (HOUSE):
  The housing sales also correlated to the S&P500 index. Once people decide to purchase their house, they are likely to renew furniture or home appliances. Therefore, the increase in consumer consumption in broad sectors might push the S&P500 index higher.

- Unemployment Rate (UNEMP):
  We can see huge volatility in 2022, which was caused by the city shutdown and restriction of travel to tackle the global pandemic.
  Several peaks in unemployment rate tend to correspond to the bottom of the S&P500; therefore, they are likely to be correlated inversely. The current unemployment rate is almost the lowest level historically – below 4%.

- Long-term Real Interest Rate (LRIR):
  Focusing more on the areas around the S&P500 bottoms, the real interest rate is likely to move forward to the S&P500, especially in 2011-2022 and 2022-2023.
  In other words, if we use this indicator as the prediction of the future development of the S&P500, it could improve the model performance.

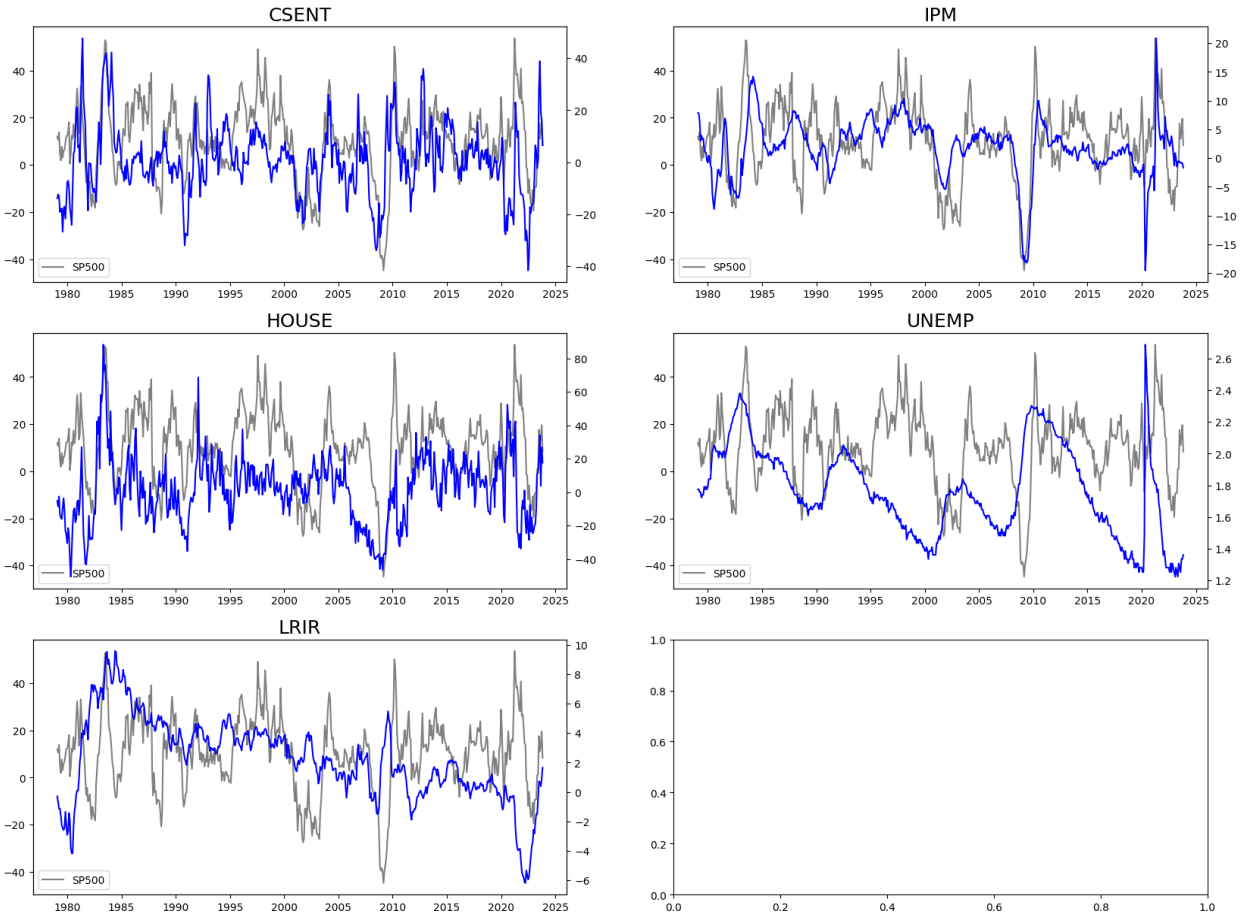

Figure 11*: Relationships between the S&P500 and economic indicators*

(7): Histogram of Economic Data Based on Categorical Target Variable

Figure 12 shows the distribution of five economic indicators stratified by whether the S&P500 index rises or falls compared to the previous year.

- Consumer Sentiment (CSENT):

  When the YoY growth of the consumer sentiment is higher than 10%, the S&P500 is likely to increase compared to the previous year; otherwise, the probability of falling rises.

- Industrial Production (IPM):

  Focusing on the cases when industrial production plunged from the previous year (let's say 5% or more decline), the S&P500 index is also likely to fall compared to a year ago.

- Housing Sales (HOUSE):

  When the YoY percent growth of housing sales is less than 20%, the S&P500 index is likely to decline compared to the previous year; otherwise, the S&P500 rose from the previous year in almost all cases.

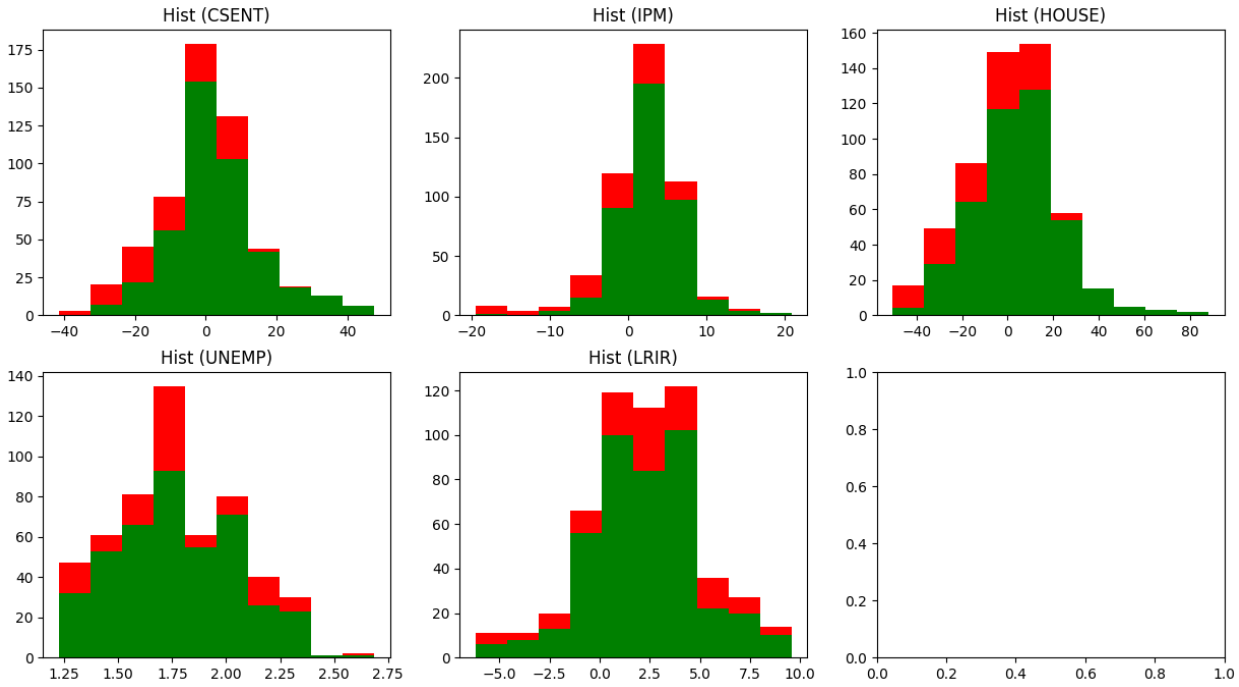

***Figure 12**: Distribution of economic indicators based on categorical value*

(8): Correlation Plots Stratified by Categorical Value

Figure 13 displays the correlation plots based on whether the S&P500 rises (green) or falls (red) compared to the same month of the previous year.

- All economic indicators seem to have a similar distribution regardless of the outcome of the target variables; if two distributions are clearly separated from each other, we can infer the strong linear relationship between an economic indicator and the S&P500 index.
- Several indicators show interesting insights into the growth of the S&P500 from the previous year. Also, those observations could be helpful to implement the robust tree algorithm.
    - When IPM or CSENT declined by over 10% or 20% from a year before, respectively, the S&P500 is likely to be below the level of the previous year.

- When the YoY growth of HOUSE is above 25% regardless of what kinds of economic indicators as the other axis, the S&P500 is likely to rise from the previous year.
- When the IPM and HOUSE decline simultaneously, it will have a negative impact on the S&P500 index.

- Thus, three economic indicators – CSENT, IMP, and HOUSE –may have a significant impact on whether the S&P500 rises or falls.

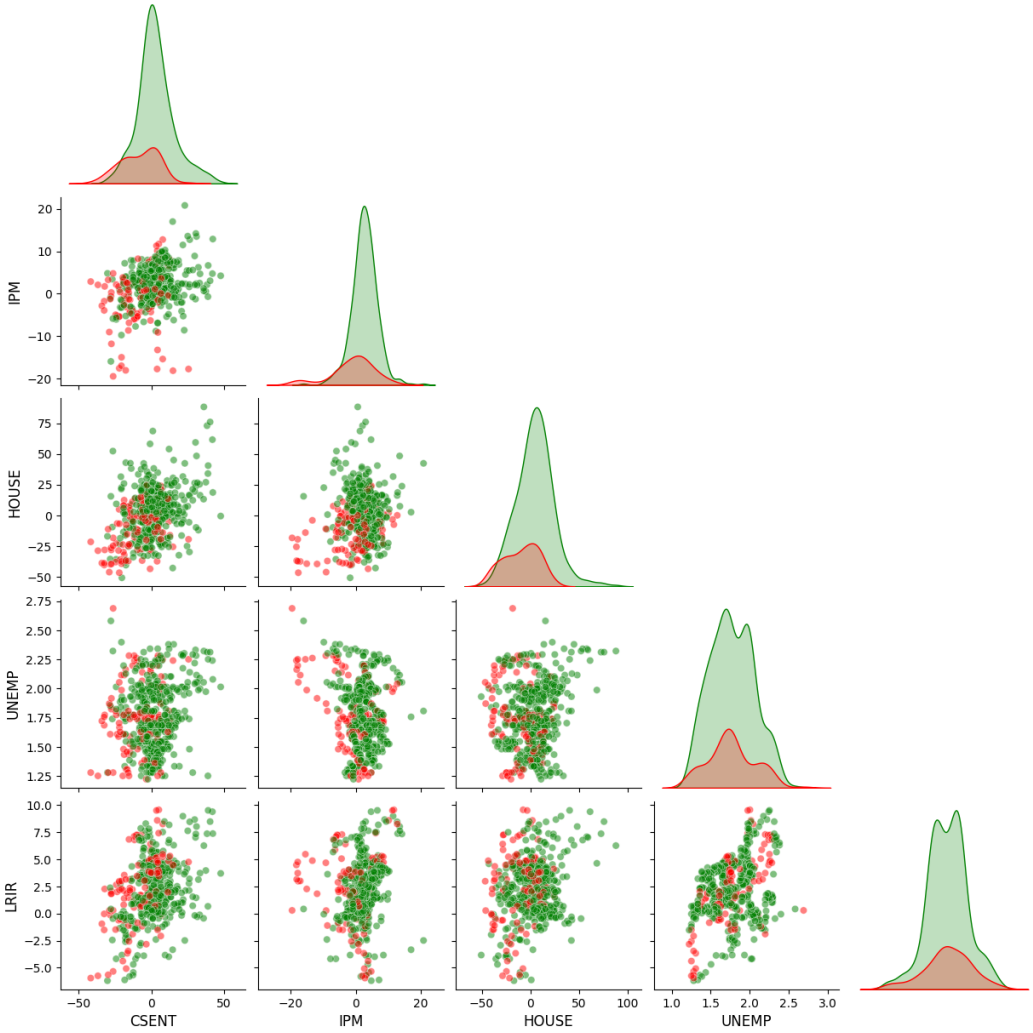

***Figure 13**: Correlation plots based on categorical values*

(9): Potential Challenges

- Due to employing the YoY percentage growth, after the actual price or index significantly falls or rises, the data shows a strong reversal; those volatilities affect the model performance.

- Some economic indicators (e.g., the real interest rate and unemployment rate) show weak relationships in the longer run, but they are likely to have a significant impact at certain situations or periods; it would be difficult for the model to consider these features.
- In the end, the dataset can be applied to both time series analysis and CART (Classification and Regression Trees), since the dataset has a categorical variable for a target and Year/Month attributes.

Methodology

The primary method of this project is to develop various machine learning models and assess their predictive performance. To generate a variety of models, three types of moving averages will create different target variables. Also, for the purpose of analyzing the impacts of each economic indicator from coefficients of the regression, all independent variables are standardized based on the rolling means and standard deviations. In terms of the model training phases, different lengths of scopes (only for linear and logistic regression) and various future durations will be taken into consideration; each condition is conducted for the purpose of updating the parameters or the model itself on different considerations.

(1): Applying Moving Averages and Standardization

Firstly three types of moving averages – from one to three months – will be applied to the original target values to consider the broad trend of the S&P500 index. *Equation 1* shows the formula of moving averages In this case, one month of moving average indicates the normal values without applying any smoothing technique. The reason why there are multiple options for moving averages to be applied is to compare multiple models' performance of each. The S&P500 index are subject to the adjustment of those moving averages.

$$y_i^{ma} = \frac{1}{k} \sum_{j=i-k+1}^{i} y_j \ , \ where\ k\ is\ number\ of\ months,\ i >= k$$

***Equation 1**: Moving Average*

Secondly, all independent variables are standardized by mean and standard deviation over certain periods. Here are the detailed steps of the standardization in this project;

1. Standardizing the first 10 years of data based on the mean and standard deviation over its periods, shown in *Equation 2*. In this project, the dataset is the monthly base, so $t = 120$.

$$X_{ss_i} = \frac{X_i - \mu}{\sigma},\ where\ \mu = \frac{1}{t}\sum_{i=1}^{t} X_i\ and\ \sigma = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(X_i - \mu)^2},\ i < t$$

***Equation 2**: Standard deviation for the first 120 of data*

2. Standardizing the rest of data by the mean and standard deviation over the most recent 10 years of data (120 datasets) from the newly added dataset, shown in *Equation 3*. This rolling standardization is implemented, instead of the normal standardization, in order to appropriately handle the huge market volatility (Bris, 2018). For example, if the data is standardized by the entire period, all standardized values are affected by huge spikes or declines regardless of when those excessive volatiles happened. However, if the mean and standard deviation are modified by a defined range of time, the data will be standardized by the recent data adaptively while removing the older data.

$$X_{ss_i} = \frac{X_i - \mu_i}{\sigma_i},\ where\ \mu_i = \frac{1}{t}\sum_{j=t-i+1}^{t} X_j,\ \sigma_i = \sqrt{\frac{1}{t}\sum_{j=t-i+1}^{t}(X_i - \mu_i)^2},\ i >= t$$

***Equation 3**: Rolling standard deviation*

After those two data modification phases, the dataset for model training will be finalized. Due to applying the moving averages and shifting the target vector according to a defined future duration (how many months the model predicts the target ahead), the first several observations in 1979 will be NaN values. Hence, for consistency again, all data in 1979 was removed, and the completed data shown below.

- Dimensions: (526, 12)
- Date Ranges: Jan 1980 ~ Oct 2023

(2): Model Development

In this phase, each dataset is primarily trained by three models: multiple linear regression, logistic regression, and classification and regression tree (CART), utilizing scikit-learn libraries in Python. Also, in each adaptive or incremental learning phase, the model attempts to predict target values based on different scopes (how much data will be considered to update regression coefficients for the next step) and different future time periods (how many months in the future the model predicts based on the latest series of economic indicators).

(2)-1: Multiple Linear Regression

First of all, the model trains the first (the oldest) 10-year of the data based on the normal equation shown in *Equation 4* for the purpose of defining the initial parameter vector of θ_{fp}. In this project, fp shows a certain month(s) of the future that the model attempts to predict based on the newly added data, and $fp \in \{1, 2, 3, 4, 5, 6\}$ by default. For example, if $fp = 4$, the first four vectors of parameters will be the ones generated by the initial training.

$$\theta_{fp} = \left(X_t^T X_t\right)^{-1} X_t^T y_t \ where \ t = 120$$

***Equation 4**: Initial Model Training; Normal Equation*

Once the initial parameters are defined, the model starts incremental learning by using the rest of the datasets. The detailed steps are shown below. Note that, from the following formulas, i always shows the incremental steps and which is always equivalent to the subtraction of the current t by the initially defined t, which is 120.

1. *Equation 5* shows that the model predicts the target values in fp months ($\widehat{y}_{t+fp}$) based on the newly added data (X_t) at the time t and the corresponding vector of parameters (θ_i).

$$\widehat{y}_{i+fp} = X_t^T \theta_i$$

***Equation 5**: Predicted Value*

2. As shown in *Equation 6*, the residuals or errors (ε_{i+fp}) are calculated by the subtraction of the actual target value by the predicted value.

$$\varepsilon_{i+fp} = y_{t+fp} - \widehat{y}_{i+fp}$$

***Equation 6**: Residuals*

3. As shown in *Equation 7*, the parameters will be updated based on the gradient of the cost function with elastic net regularization for a certain future month(s) ahead. The elastic regularization balances between the ridge and lasso regularization for the purpose of mitigating the risk of overfitting (Alzoubi et al., 2022). In this case, n_{sc} shows the number of subset of data, which is determined by the number of scopes – how much recent data will be

considered to adjust the parameter for next step (including the newly added data at the time t). Also, the parameter updates are conducted by a given number of iterations (s), which is set to 100 by default. The hyperparameters α (default is 0.1) and λ (default is 0.5) shows the degree of l1 and l2 norm regularizations and the balancing between those two of them, respectively. Also, η is the learning step of each parameter update (default is 0.01).

$$J(\theta) = \frac{1}{2n_{sc}} \sum_{t'=t-n_{sc}}^{t} (y_{t'+fp} - X_{t'}^{T}\theta)^2 + \alpha(\lambda \sum_{k=1}^{l} |\theta_k| + \frac{1-\lambda}{2} \sum_{k=1}^{l} \theta_k^2)$$

$$\nabla J(\theta) = -\frac{2}{n_{sc}}(X^T(y - X\theta))^T + \alpha(\lambda \cdot sgn(\theta) + (1-\lambda) \cdot \theta)$$

$$\theta_{i+fp}^{(s+1)} = \theta_i^{(s)} - \eta \nabla J(\theta_i), \; s = 0, 1, 2, \ldots 99$$

***Equation 7**: Parameter updates at each step*

As we can see in *Equation 7*, one of the unique operation in this project is to update the vector of parameters θ after two or more steps, instead of the next step. It enables the model to forecast the growth of the S&P500 two or more months ahead based on the set of the latest economic indicators. Also, This incremental learning of the multiple linear regression model takes care of the following assumption; the trend of the stock market changes over time, and the most recent data indicates more reliable behavior than the past data. Therefore, the default options of scopes (n_{sc}) are set to [1, 3, 6, 9, 12]. It means that at least five different models will be developed from those different scopes; we can expect that the more recent data are focused on, the higher predictive performance by taking the risk of overfitting to the recent data; the model might learn unexpected behavior in the market. Therefore, the model will take care of all defined scopes and aggregate them to pursue generalization. In other words, for this ensemble learning by different scopes, the aggregated result is expected to balance between the generalized model toward the past data and the adaptive model toward the most recent data.

(2)-2: Logistic Regression

The second model is logistic regression. Similar to the multiple linear regression, the first 10-year of data will be used for the training in the scikit-learn's LogisticRegression() class. *Equation 8* shows

the logistic function, considering that θ is a vector of parameters and X is the feature matrix; both matrices are included the bias term in the number of parameters n.

$$\sigma(X\theta) = \frac{1}{1+exp(-X\theta)},\ where\ X \in \Re^{t\times n},\ \theta \in \Re^{n\times 1},\ t = 120$$

***Equation 8**: Logistic (sigmoid) function*

Equation 9 shows the cost function of the logistic regression. Based on this cost function, a vector of parameters will be updated and gradually optimized. Since the number of the class labels is imbalanced in the dataset, the logistic regression applies the weight w_c for each class *c*, referred from scikit-learn document (Scikit-Learn Developers(1), 2023); where $m_{c \in \{0,\ 1\}}$ is the number of observations on each class label *c* within the first 10-year of data, n is the number of parameters, and y_{t+fp} is the binary target label after defined months.

$$J(\theta) = -\frac{1}{m}\sum_{t=1}^{m}[w_c(y_{t+fp}log(\sigma(X_t\theta)) + (1 - y_{t+fp})log(X_t\theta))] + \lambda\sum_{j=1}^{n}\theta_j^2$$

$$w_c = m/2m_c\ ,\ where\ c \in \{0,\ 1\}$$

***Equation 9**: Cost function of the logistic regression and balanced weights*

After initially defining the parameters, the incremental learning phases begin. In this phase, a vector of parameters is continuously updated based on the newly added vector of features as shown the following steps:

1. calculate the probability of classifying into the label "1" ($\widehat{y}_{i+fp}$) toward certain months ahead by applying the dot product of newly added feature vector ($X_t \in \Re^{1\times n}$) at the time t and the corresponding vector of parameters ($\theta_i \in \Re^{n\times 1}$) to the sigmoid function; therefore, the formula is similar to *Equation 8*. Note that again the incremental step i is always equivalent to the subtraction of the current time step t by the firstly defined t for the initial model learning.

2. calculate logistic loss at each incremental step i shown in *Equation 10*. In this case, w_c is the same definition shown in *Equation 9*.

$$L(y_{t+fp}, \widehat{y}_{i+fp}) = - w_c[y_{t+fp} log(\widehat{y}_{i+fp}) + (1 - y_{t+fp}) log(1 - \widehat{y}_{i+fp})]$$

***Equation 10**: Logistic loss with weight*

3. As shown in *Equation 11*, calculate the gradient of the cost function with the elastic net regularization at each incremental step i. Although w_c is a vector of weights corresponding to each actual target label defined in *Equation 9*, the values are continuously updated because the new data is added in each incremental step. Also, $X \in \Re^{n_{sc} \times k}$ and $y \in \Re^{n_{sc} \times 1}$, where n_{sc} is the number of observations based on defined scopes. The hyperparameters, η, λ , and α, are the same values as the multiple linear regression.

$$\nabla J(\theta) = - \frac{1}{n_{sc}} X^T (w_c \odot (y - \sigma(X\theta))) + \alpha(\lambda sgn\theta + (1 - \lambda)\theta)$$

$$\theta_{i+fp}^{(s+1)} = \theta_i^{(s)} - \eta \nabla J(\theta_i), \; where \; s = \{0, 1, 2, \ldots 99\}$$

***Equation 11**: Gradient of cost function in logistic regression*

(2)-3: Classification and Regression Tree (CART)

Lastly, the CART is applied to the model dataset. In the CART model, the incremental or adaptive learning cannot be implemented, like we conducted in the linear and logistic regression. Hence, we implemented the offline learning; the model trains all the historical data once the newly added economic data is released.

The core theory of the CART is to decide variables and thresholds in order to split the data into binary trees by focusing on a certain criterion (Hastie, et al. 2009). In other words, the order of the tree nodes shows what variables can provide more effective thresholds to minimize certain criteria in both nodes after splitting. In this project, since the YoY growth of the S&P500, which is numerical data, is defined as a target value, the tree attempts to decide independent variables and their threshold so that the sum of the square of errors minimizes. Equation 12 shows the mathematical concepts for its minimization (Hastie, et al. 2009).

$$S_1(d,\ v) = \{x \mid x_d \leq v\},\ S_2(d,\ v) = \{x \mid x_d < v\}$$

$$min_{d,\,v}\left[\sum_{x_i \in S_1(d,\,v)} (y_{i+fp} - \hat{y}_{S_1})^2 + \sum_{x_i \in S_2(d,\,v)} (y_{i+fp} - \hat{y}_{S_2})^2\right]$$

***Equation 12**: optimization of the tree's criterion; splitting process*

In the *Equation 12*, d is an selected independent variable and s is an defined threshold from its variable, After selecting two subsets of data based on the pair of d and s, the algorithm calculates the sum of squared predicted errors and finds the optimal d and s. In the second line of formula, y_{i+fp} shows the actual target value with the consideration of the predicted months ahead (fp) based on the time series data of i. Hence, the algorithm attempts to find the optimal independent variable and its threshold to predict the future growth rate of the S&P500 index. In addition, for the purpose of addressing the risk of overfitting, the algorithm automatically stops the process if the depth of trees reaches five layers. It means that until reaching the five layer, these splitting processes are conducted.

(3): Model Evaluation

In terms of the multiple linear regression and Classification and Regression (CART), the following three variables are continually stored at each step of this incremental learning; the parameter (θ), the predicted value ($\hat{y}$), and the predicted error (ε). Those results are mainly used for the following four evaluation strategies;

- Fundamental regression measures; Root Mean Square Error (RMSE), Standard Error of Estimate (SE), Coefficient of Determination (R2), and the Adjusted R2 (Adj-R2).
- Backward Elimination: Assessing the impact of each economic indicator on RMSE and Adj-R2 for all models (only for linear regression).
- Coefficient of each independent variable: Evaluating the impact of each economic indicator over time.
- Visual Comparison of the Prediction and Actual Development: showing the predicted and actual price movements and comparing with each other
- Plotting the Predicted Error: Showing the distribution of the predicted errors to check their homoscedasticity and normality.

Based on the backward elimination results, the most influential feature (an economic indicator) will be chosen. Also, the backward elimination results may indicate how the most influential feature will differ if we focus on the different conditions in terms of the various months of moving averages and a variety of scopes of updating the parameters. In addition, mainly focusing on the RMSE and Adj-R2, we will observe how each model can accurately predict the YoY growth of the target prices or how each model can explain the fluctuation of the actual target movements. Finally, considering all four evaluation strategies, the best sets of model parameters will be chosen; how many months of moving average should be adapted to the target value and how many scopes should be considered to update the parameter at each step.

As for the Logistic Regression, the following five evaluation measures will be taken into consideration; accuracy, precision, recall, f1 score, and Area Under the ROC Curve(AUC). From the accuracy, we will observe how the model can correctly predict whether the S&P500 rises or falls from the year before. However, the dataset is is imbalanced; the number of the label "0", which shows the decline of the S&P500 from a year before, is significantly smaller than the label "1". So, from the precision, recall and f1 score, we will check whether the models do not skew toward either side of the labels; for example, if the model always predicts "1" despite showing higher accuracy, it would not be the good model. In other words, precision and recall should be high and balanced between them. Furthermore, we will observe the AUC. Considering that the ROC will be closer to the perpendicular line as the model performance improves, the AUC should be expected to be higher for the good model. In addition, the backward elimination test will be conducted in the logistic regression. The detail processes are similar to what we saw in the evaluation phases of the multiple linear regression. The only difference is to apply the revised linear formula, either one of parameters is automatically set zero, to the logistic function. Also, we will observe how coefficient of each independent variable changed over time.

Discussion

By following the Methodology section while focusing on the research questions, each model will be strictly evaluated in this phase.

(1) Multiple Linear Regression

The result of the fundamental regression measures is shown in *Figure 14*. First of all, different colors shows three types of moving averages. Also, each data point is the averaged measures generated from four five different scopes; 1-, 3-, 6-, 9-, and 12-month. There are two notable relationships from this figure; (1) predicting the target with longer moving averages improves the regression performance – lower predictive errors and higher R-square; (2) predicting longer terms of the future decreases the predictive performance. Overall, the model performance showed that the latest updated economic indicators no longer predict the S&P500 index in two months or longer, considering that more than 10% of the root mean square error and less than 0.6 of the R square.

Next, we will observe the result of backward elimination (BE). In this experiment, after each coefficient will be set zero automatically one by one, the model predicts the target value, and the regression metrics are calculated again. In other words, these tests attempt to examine how the model performance will be changed if the impact of a certain independent variable is lost. *Figure 15* shows the BE result with respect to different scopes. The figure shows that, overall, the unemployment and long-term real interest rate from inflation rate and the US Treasury yields are likely to affect greater impacts on the model performance compared to other economic indicators; it is because losing the impact of those indicators may deteriorate the model performance – the increase in RMSE and the decrease in R-square However, we could not specify which economic indicator would affect significant impact on the S&P500 index from this figure.

Figure 16 shows the impact of each economic indicator on the S&P500 YoY growth on average; data is calculated from results from all multiple linear models with different scopes, moving averages, and prediction ranges. Since the model employs incremental learning, all coefficients are continuously updated. Also, all independent variables are standardized, so the y-axis shows how strong the impact of each economic indicator is for the S&P500. If we focus on the recent trend, the consumer sentiment data (CSENT) and housing prices (HOUSE) could have positive impacts on the S&P500. Although the rise in unemployment rate(UNEMP) is likely to affect positive impact, but its

degree is decreases recently. Therefore, this result may be helpful for investors or traders to focus on which economic indicators are likely to cause larger volatilities in the S&P500 and facilitate determining what the best timing when they buy or sell its stock index.

Lastly, we will focus on the comparison between the actual and predicted growth over time and the distribution of the predicted errors. For example, Figure 17 shows one of the results among a significant number of multiple linear models. The green line at the time t is based on the economic indicator at the time $t - 2$, in other words, two months ago (technically, creating a prediction around 30-40 days before due to a time lag of releasing economic indicators). Also, the prediction is the average of all predicted values from different scopes; the white bands show the plus and minus three standard deviations from those results. The predicted line is very close to the actual line by adjusting the parameters incrementally but strictly speaking, the model is sometimes behind the actual data; that is not the prediction. Thus, although the model may be able to predict the actual values during the stable situation, huge volatility in short periods cannot be predicted in advance, and the model seems to struggle with its parameter adjustments. If we focus on *Figure 18*, we can recognize how difficult the model prediction forecasts the market volatility. The scatter plot shows the predicted errors are normally random and constant between -10 and 10%; however, once the huge spikes or declines happen, the errors fluctuate up and down.

***Figure 14**: Regression measures in linear regression*

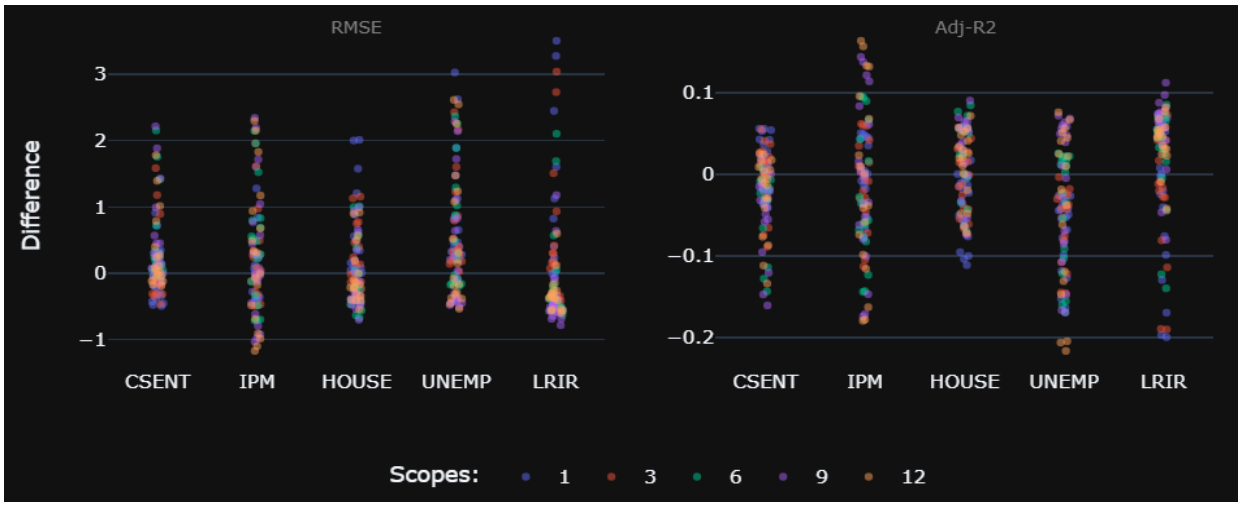

***Figure 15**: Backward elimination in linear regression*

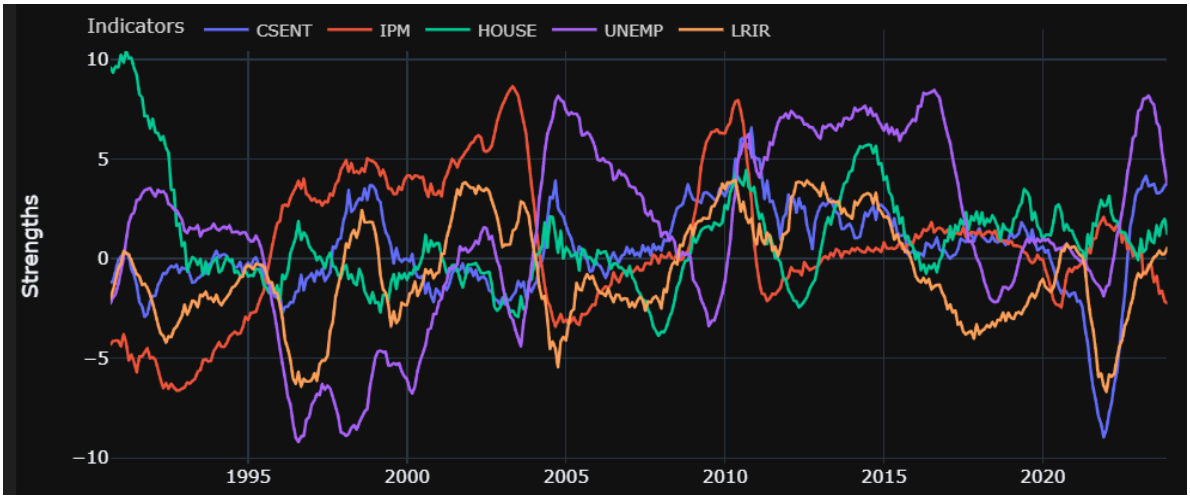

***Figure 16**: Coefficients of each economic indicator in linear regression*

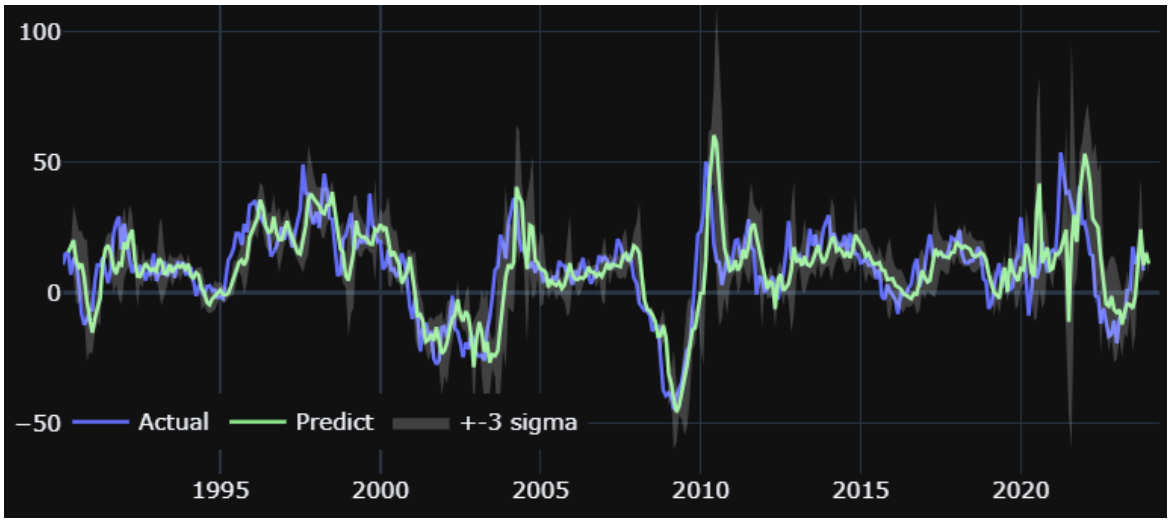

***Figure 17**: Comparison between actual and predicted S&P500 YoY growth(%) over time in linear regression*

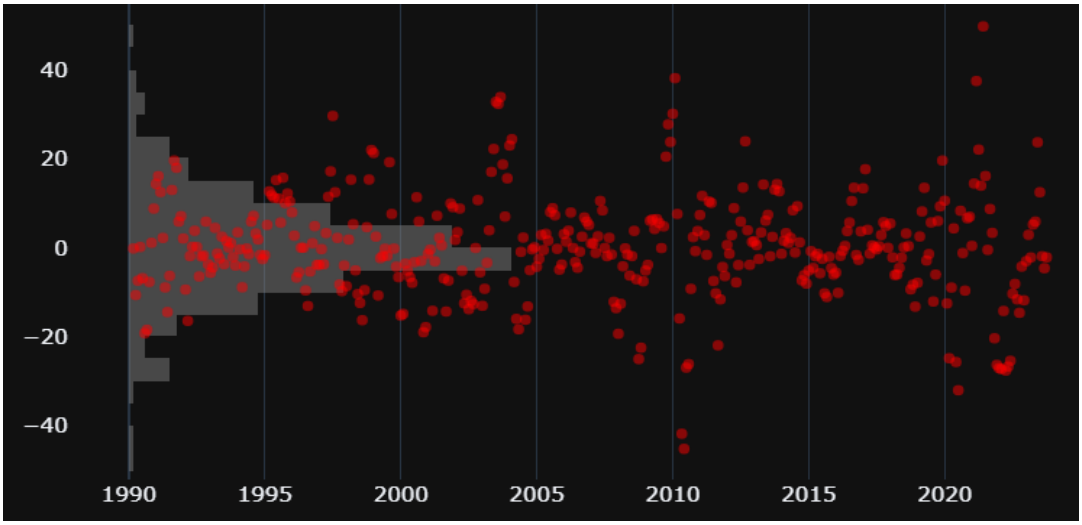

***Figure 18**: Distribution of the predicted errors in linear regression*

(2) Logistic Regression

Figure 19 shows the basic classification metrics of logistic regression. As we saw in the multiple linear regression models, the larger moving average increases the model performance, but the forecasts in longer months ahead decrease its performance. Overall, precision, recall, and f1 score showed relatively higher scores. Considering that precisions are higher than recalls, the model be conservative when it forecasts to rise the S&P500 compared to the previous year. Accuracy and area under the ROC curve (AUC) show moderately high performance. Although all metrics seem to be

useful in real situations due to relatively higher scores, we need to consider how the class labels were defined. As we observed in the data creation phase, these class labels were set by whether the S&P500 prices are above or below compared to the prices in the same month of the previous year. In other words, as long as the prices are above the previous year, the class labels are set "1" regardless of causing a huge decline during the periods. Therefore, if the S&P500 has already achieved huge gains, the model would not help forecast further rises or declines in the months ahead based on the current price level.

Figure 20 shows the backward elimination results with respect to different scopes. In terms of the logistic regression, the long-term real interest rate (LRIR) based on the inflation rate and the US Treasury yields has higher impacts on the prediction; the test results showed the removal of the LRIR variable, the model performance could deteriorate due to decreasing accuracy and f1 score. Also, the consumer sentiment data (CSENT) also may have a relatively high impact. On the other hand, housing sales (HOUSE) and unemployment rate (UNEMP) may not affect the classification performance, rather the removal of those indicators could improve the performance.

Figure 21 shows the factor analysis based on the coefficients of independent variables at each time step. The figure implies that similar results in the multiple linear regression. Currently, the rise in unemployment rate could have positive impacts on the S&P500 growth despite showing the downtrend recently. Therefore, in the coming future, the housing sales (HOUSE) and consumer sentiment data (CSENT) could highly positive correlated to the development in the S&P500 growth.

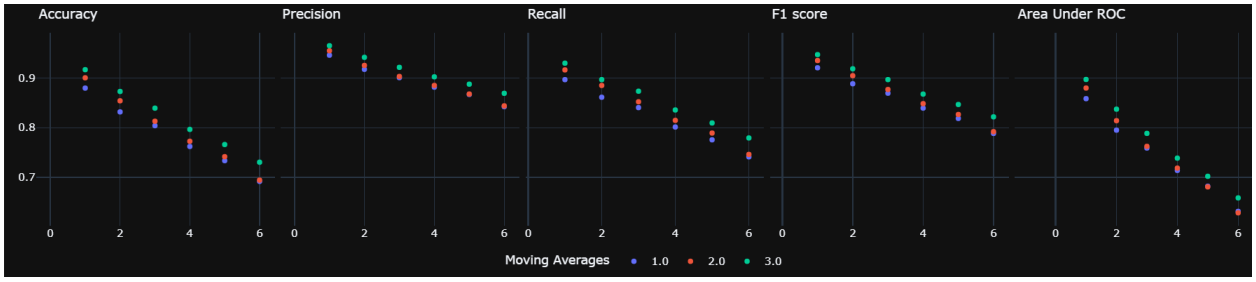

***Figure 19**: Classification metrics in logistic regression*

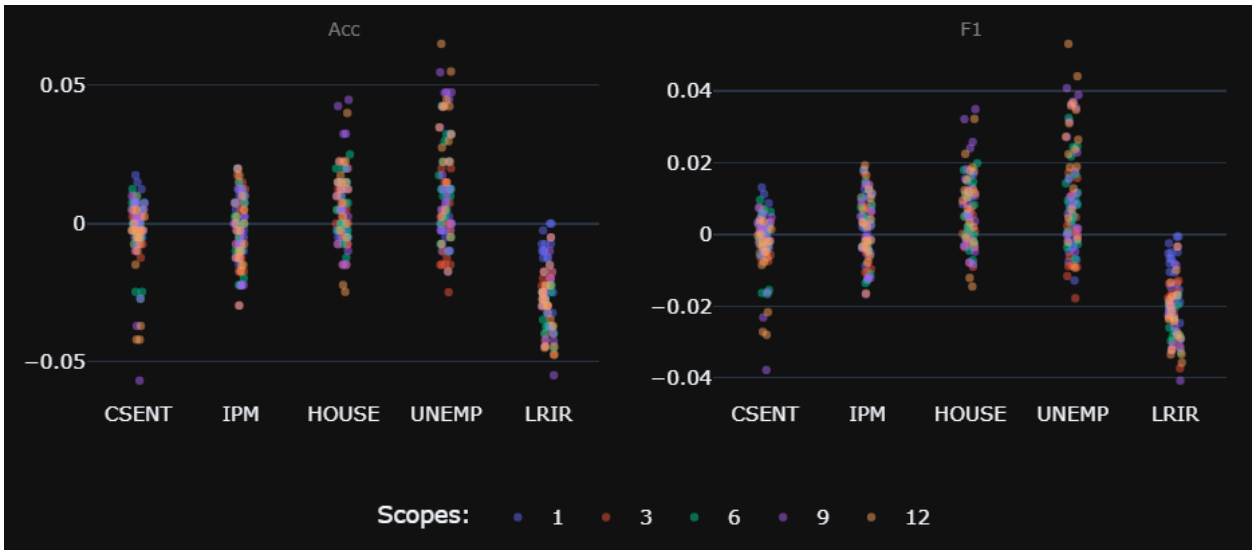

***Figure 20**: Backward elimination in logistic regression*

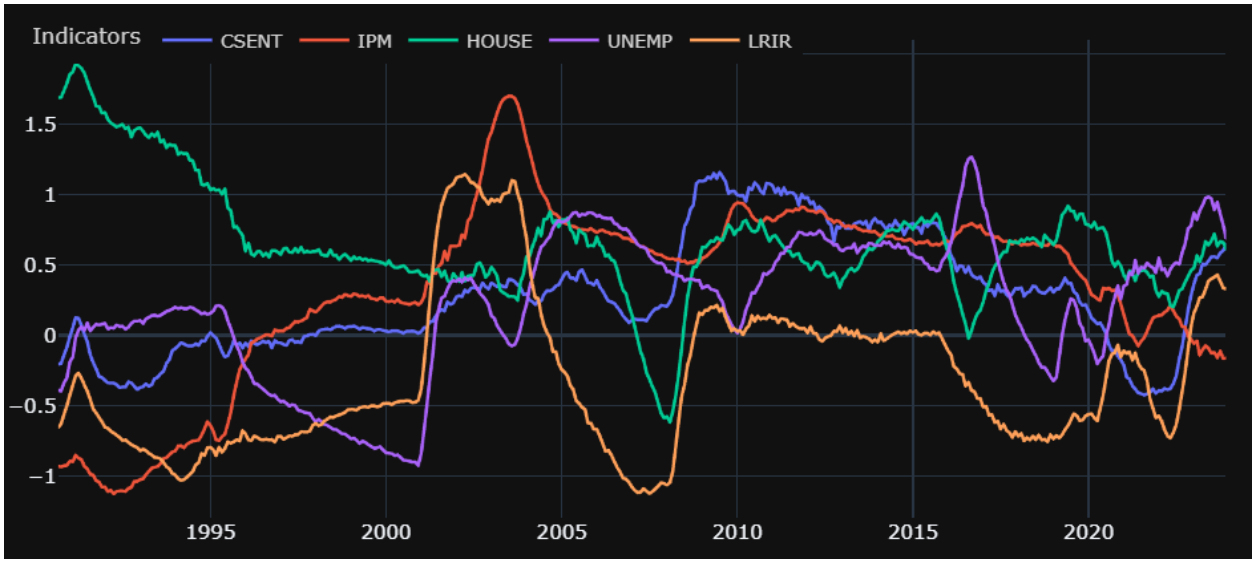

***Figure 21**: Coefficients of independent variables over time*

(3) Classification and Regression Tree (CART)

Figure 22 shows the basic regression metrics in CART. Comparing those results with the same outcome in the multiple linear regression, all metrics deteriorate; especially, R2 decreased by around 0.5 points from the highest scores in the linear regression. It could be caused by the nature of the CART model; predicting the average of values in the bottom nodes. Also, the CART model cannot implement the similar incremental or adaptive algorithm with the linear regression and logistic

regression, so the model may not be able to highlight the recent relationships between economic indicators and the S&P 500 index.

Figure 23 shows the feature importances of each economic indicator generated from "DecisionTreeRegressor().feature_importances_". According to the scikit-learn document, those values are calculated as "Gini importance" (Scikit-Learn Developers(2), 2023). In the general classification tree algorithm, this measure shows how much an economic indicator contributes to homogenizing the class labels in the tree nodes – a defined threshold in a particular economic indicator contributes to separating the class labels so that either one of the labels is majority or dominant in two nodes. However, this project uses the numerical variables as a target; therefore, the feature importance shows how a defined threshold from an economic data contributes to minimize the mean square of the predicted error. From the *Figure 23*, the industrial production in manufacturing (IPM) showed the largest contribution in minimizing the predicted errors, especially since around 2010. To interpret the model result, *Figure 24* shows the top parts of the tree diagram from one of the CART model. The diagram indicates, for example, when the IPM is less than or equal to -2.16 and HOUSE is less than 0.46, the average growth of the S&P500 was -21.26% compared to the same month of the previous year among 30 samples. This insight could be useful for one of the risk management strategies in the stock market.

Figure 25 shows the comparison of the line plots between the actual and the predicted values; as we saw in the linear regression, the green line (prediction) at all time steps is technically drawn around 30-40 days before the blue line (actual). Since the predicted values in the CART are based on the average values on each node, larger fluctuations occurred in the green line. Also, there are large divergences between two lines, so the CART model is not useful to forecast the YoY growth of the S&P500 index. From the perspective of the predicted error distributions in *Figure 26*, although the shape of histogram is close to the normal distribution, the errors are not constant, especially in the past 10 years; the errors show a clear cyclical pattern, which implies that the model violates the homoscedasticity. Therefore, the model was highly likely to underfit the data, or the model learns insufficient data.

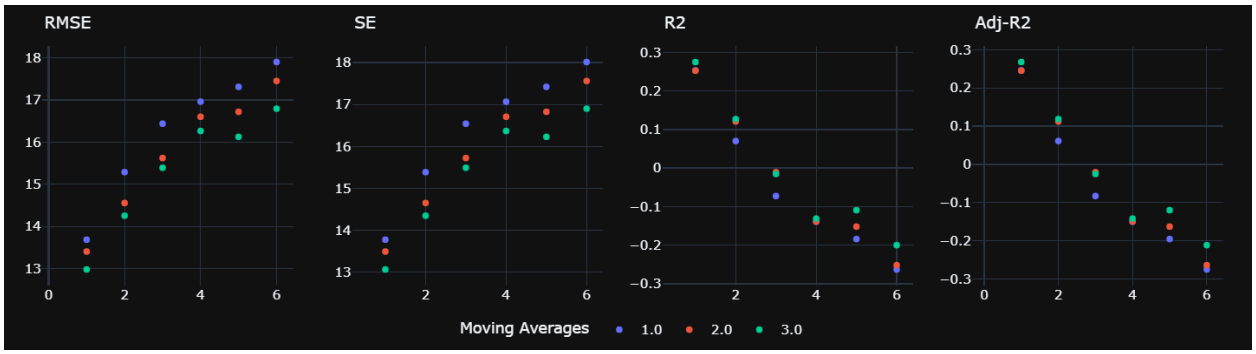

***Figure 22**: Regression metrics in the CART*

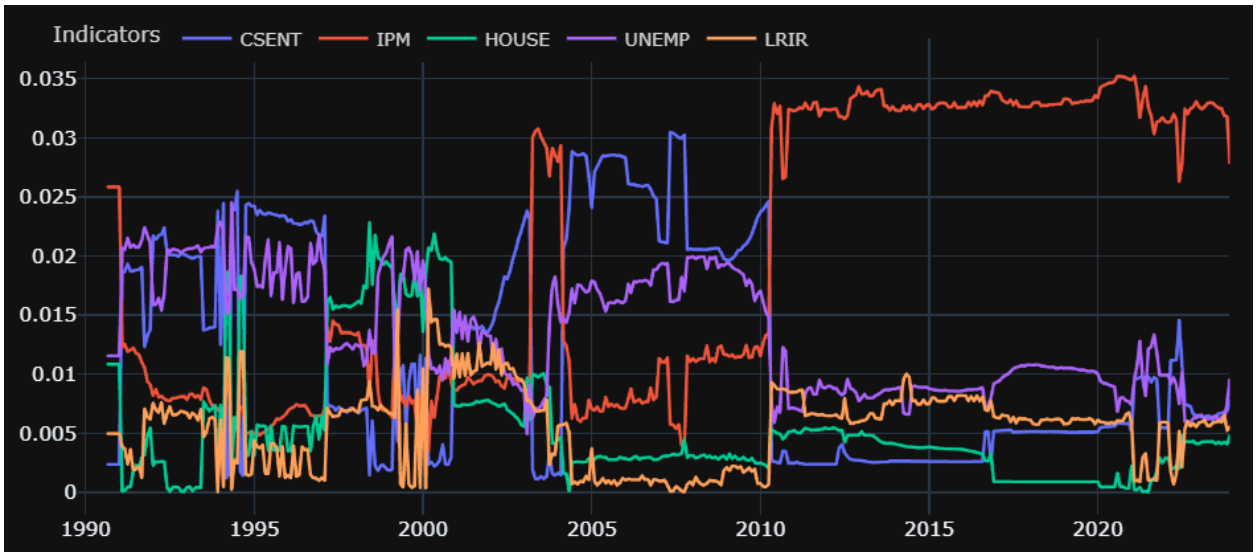

***Figure 23**: Feature importance at each time step in the CART*

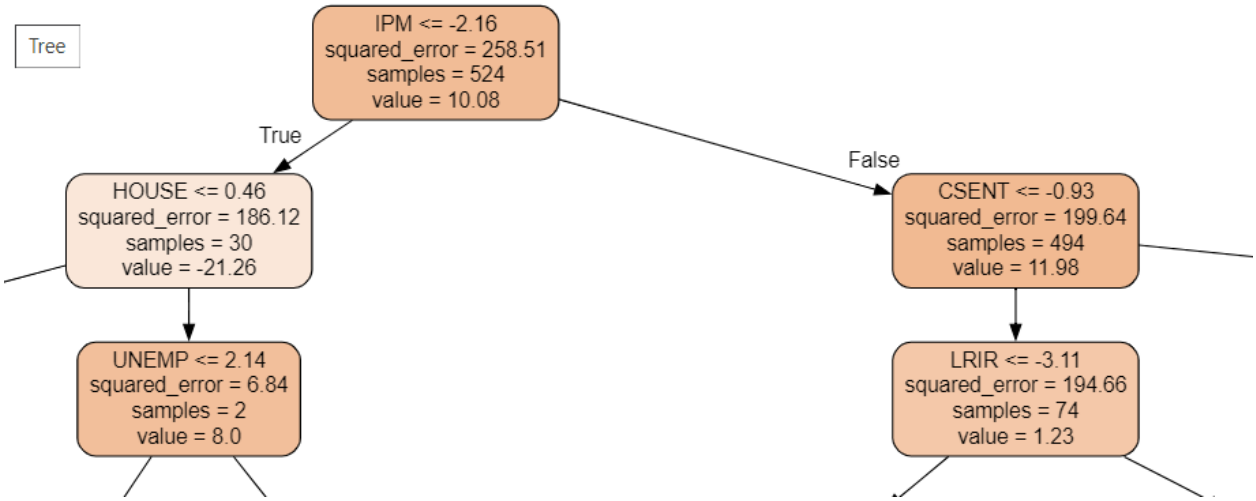

***Figure 24**: Top parts of tree from one of the model results*

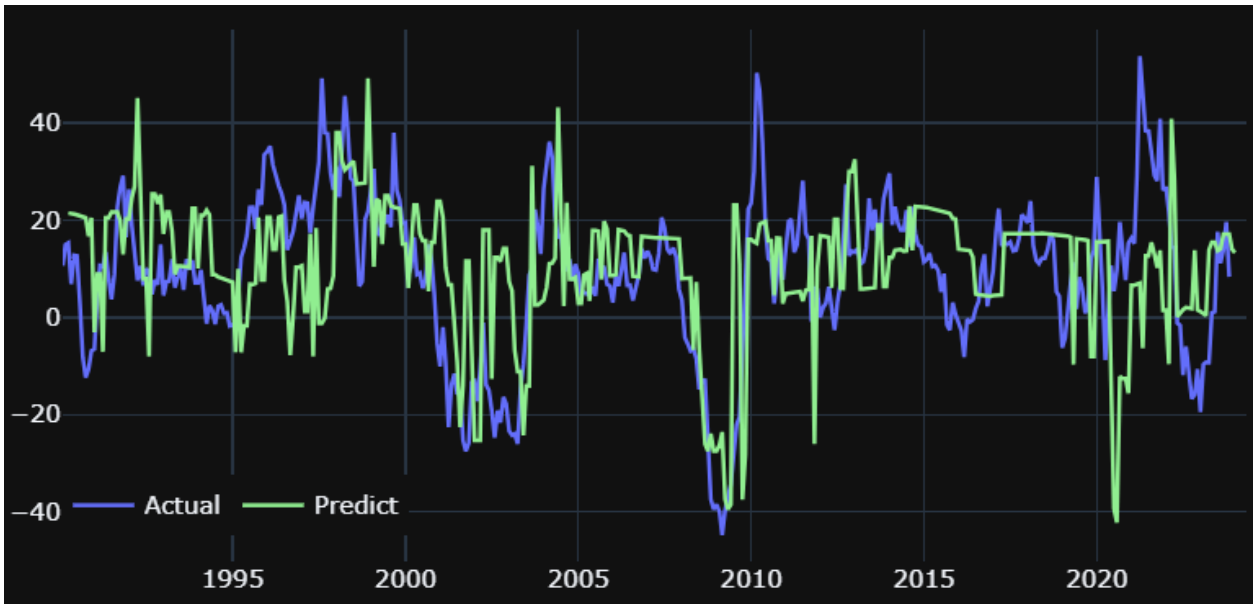

***Figure 25**: Comparison between actual and predicted S&P500 YoY growth(%) over time in the CART*

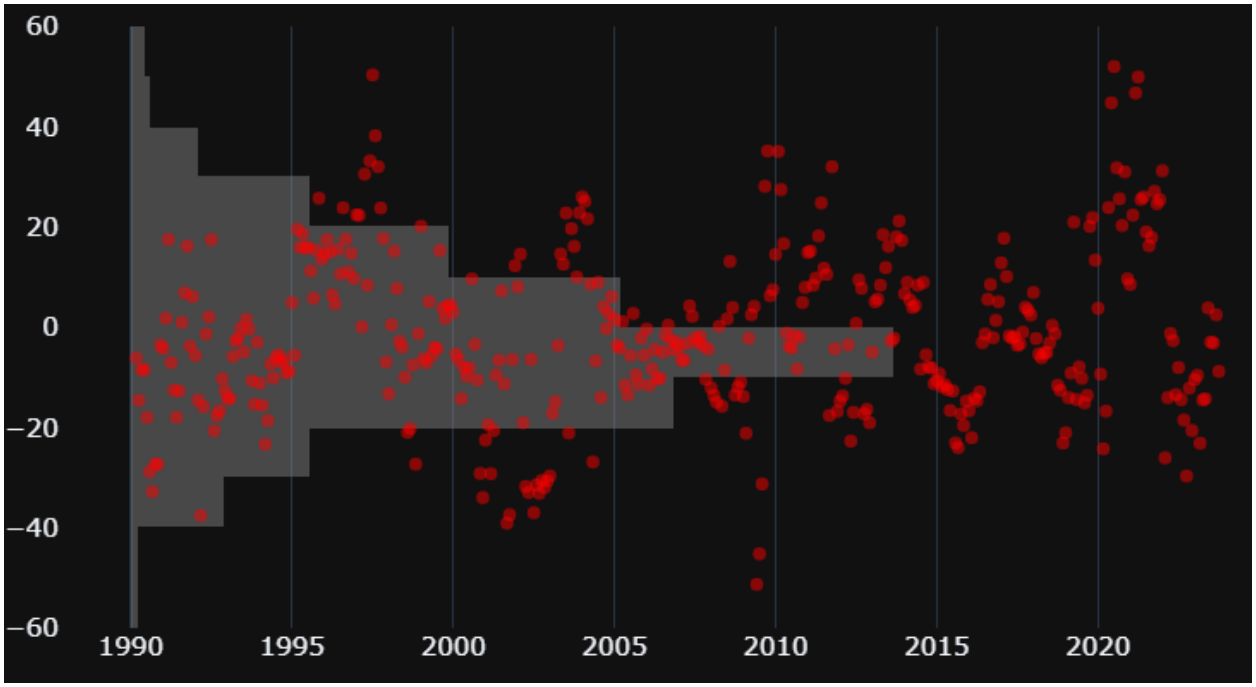

***Figure 26**: Distribution of the predicted errors in the CART*

(4) Finalizing the Best Model

Considering all results and concerns in each mode, the best model could be the ensemble linear regression model of predicting less than three months ahead (future prediction; FP) and focusing on the recent one and three months of data (scope; SC). The lists below are summaries of the prediction outcomes.

- From *Figure 27*, the RMSE (7.0-9.5%) and the R square (0.62-0.77) could be acceptable despite insufficient scores as the sophisticated predictive model.
- From *Figure 28*, although large divergences occur during huge volatilities (getting wider the bands), the model can forecast the S&P500 growth on the stable market around two months ahead (technically 30-40 days ahead, due to the time lag of releasing economic data).
- From *Figure 29*, the distribution of the predicted errors would meet normality and randomness; however, it does not meet the constancy at several short periods of time due to the volatile events in the market.

Furthermore,the multiple linear model was selected as the best model due to the following concerns in other models;

- From the definition of the class labels in this project, the logistic model may not be useful to forecast whether the S&P500 growth increases or decreases from the current level in a couple of months; the model can only predict above or below the level in the same month of the year before. Hence, using logistic regression may not provide investors or traders with a useful predictive model for making buy or sell decisions on the S&P500 index.
- Although the CART may provide valuable insight if we strictly observe every condition of the nodes and average performance, the model no longer have the ability to forecast the YoY growth on the S&P500 index overall because of worse predictive performance; in order to capture more complex relationship between economic indicators and the S&P500 index, the model may be required to learn more data.

MA	FP	SC	RMSE	SE	R2	Adj-R2
1.0	1.5	2.0	9.570140	9.641932	0.624525	0.619812
2.0	1.5	2.0	8.266870	8.328892	0.697965	0.694173
3.0	1.5	2.0	7.065381	7.118390	0.768854	0.765952

***Figure 27**: Regression results on the ensemble linear regression; averaged values on each Moving Average(MA)*

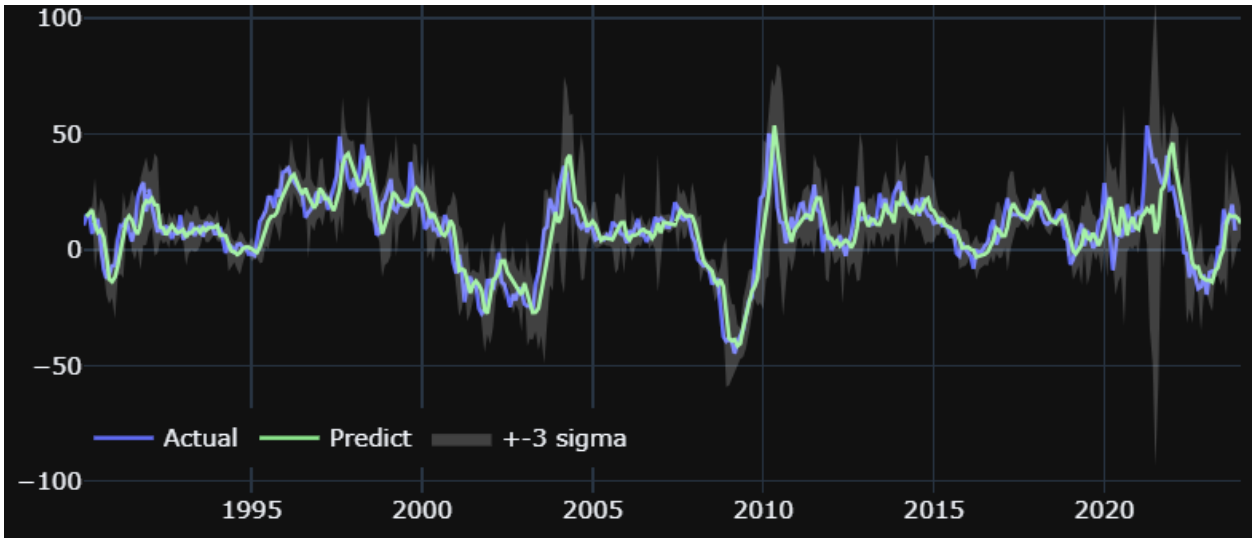

***Figure 28**: Comparison between actual (no-moving average) and predicted values in ensemble linear regression*

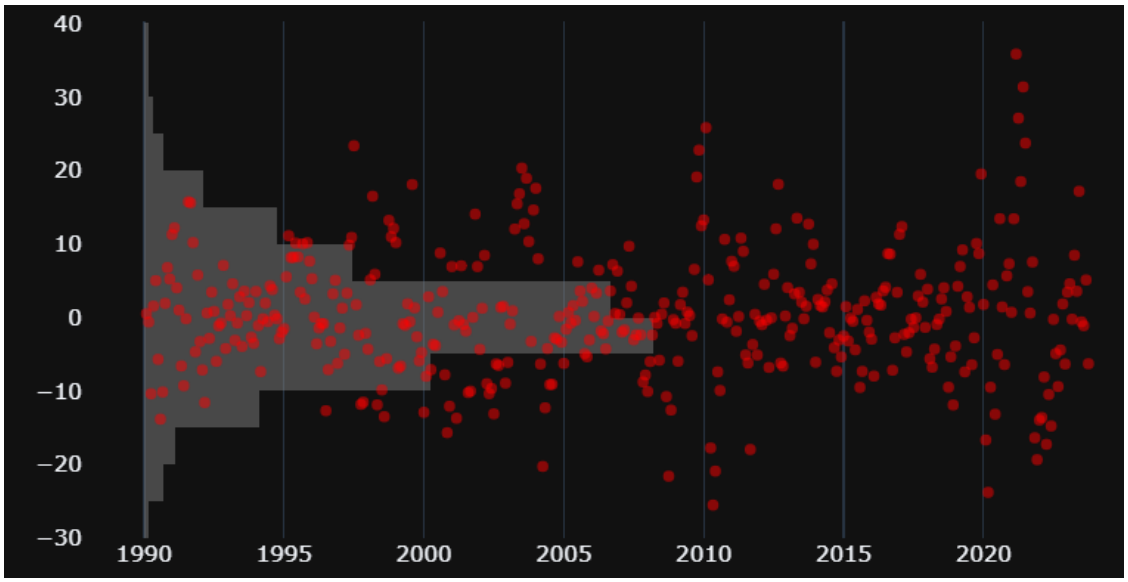

***Figure 29**: Predicted error distributions in ensemble linear regression*

Conclusion

This project aimed to develop predictive models for the S&P500 index using a macroeconomic approach, focusing on economic indicators from the Federal Reserve Economic Data (FRED) of the Federal Reserve Bank of St. Louis. The primary objective was to create reliable models that can forecast the S&P500 index's performance one to three months ahead, utilizing monthly economic indicators. Here are findings from the primary questions in this project;

- Model performance on different conditions:
    - The longer the moving averages are applied to the YoY growth of the S&P500, the predictive performance improves toward all three models.
    - The smaller the scopes, the fewer months of the recent data, are considered to adjust parameters adaptively at each time step, the predictive performance improves toward all three models.
    - The further into the future (three months or longer) the model attempts to predict, the model's predictive performance deteriorates.
- Here are the possible most significant impact of economic indicators on the S&P500 on each different model;
    - Linear regression & logistic regression:
      focusing on the recent 12 months, the rise in the unemployment has been likely to affect significant impact on the rise in the YoY growth of the S&P500.
    - CART:
      focusing on the recent 10 years, the industrial production in manufacturing (IPM) has contributed to separate two nodes so that either one of class labels is likely to be more dominant in those two groups. Therefore, IPM may not affect significant impact on the S&P500, but this data may be useful to predict the YoY growth of the stock index based on the classified group.
- Here are how each model accurately predict the YoY growth of the S&P500, based on different conditions from moving averages and future predictions, but taking averages from different scopes:

- Linear regression:
  RMSE (6.37–17.16), SE (6.41–17.29), R2 (-0.16–0.82), Adj-R2 (-0.18–0.82)
- Logistic regression:
  Accuracy (0.69–0.92), Precision (0.84–0.96), Recall (0.74–0.93), F1 score (0.79–0.95), Area Under Curve(0.6-0.9)
- Classification and regression tree:
  RMSE (13.39–17.78), SE (13.47–17.89), R2 (-0.27–0.24), Adj-R2 (-0.28–0.23)

Overall, all models may not be useful in the real business situations, such as risk management and investment decisions. Therefore, this project could conclude that the most recently released macroeconomic indicators no longer predict the YoY growth of the S&P500, especially three or more months ahead. However, beside from the primary targets of this project, the incremental learning approach provided crucial insights; the coefficients of each independent variable indicated what and when economic indicators would have significant impact on the S&P500 index. These results may provide investors and traders with what economic data should be focused on at certain time. Hence, the coefficient of each independent variable at each time step could be useful in the real situations.

Finally, in order to build more robust adaptive financial model, there are two possible approaches for the future research;

- More specific incremental learnings:
  Considering that all macroeconomic indicators are released on different dates, the model's parameters should be updated one by one right after releasing the new macroeconomic data. If its strategy works, the financial model will be close to what stakeholders in the financial market do in real life. Also, the model can attempt to learn the daily data, which could enable the model to be more adaptive toward the volatile market.
- Switching the models:
  This project confirmed that the linear model can forecast the development of the S&P500 on the YoY percent change base during a stable situation. Therefore, once the economy faces a sudden shock or recession (the model keeps monitoring certain data to recognize such emergency states), then the parameter or the independent data itself may be able to switch to another one. In other words, multiple models will be implemented as an ensemble model, and each model is in charge of predictions at certain situations or periods.

References

Adrian, T., Goel, R., Malik, S., & Natalucci, F. (2021, April 22). Understanding the Rise in Long-Term Rates. IMF. https://www.imf.org/en/Blogs/Articles/2021/04/22/blog-understanding-the-rise-in-long-term-rates

Alzoubi, H. M., Sahawneh, N., AlHamad, A. Q., Malik, U., Majid, A. and Atta, A. (2022) Analysis Of Cost Prediction In Medical Insurance Using Modern Regression Models. *2022 International Conference on Cyber Resilience (ICCR)*, pp. 1-10, doi: 10.1109/ICCR56254.2022.9995926.

Bris, D. L. (2018). What is a market crash?. *The Economic History Review, 71*(2), 480-505. https://doi.org/10.1111/ehr.12540

Hastie, T., Tibshirani, R., and Friedman, J. "Elements of StatisticalLearning", Springer, 2009.

Scikit-Learn Developers(1). (2023). LogisticRegression. *Scikit Learn*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Scikit-Learn Developers(2). (2023). DecisionTreeRegressor. *Scikit Learn*. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Sundar, S., Dhyani, M.B., & Chhajer, D.P. (2023). Factors Affecting Stock Market Movements: An Investors Perspective. *European Economic Letters*.

Thakkar, A., & Chaudhari, K. (2021). A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Systems with Applications*, 177, 114800–. https://doi.org/10.1016/j.eswa.2021.114800

Wolla, S. A. (2012, May 14). Getting real about interest rates. Economic Lowdown Podcast Series. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/education/economic-lowdown-podcast-series/episode-14-getting-real-about-interest-rates

For the dataset:

(1) Yahoo Finance; https://finance.yahoo.com/quote/%5EGSPC/

(2) Board of Governors of the Federal Reserve System (US), Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis [DGS10], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DGS10

(3) U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPIAUCSL

(4) University of Michigan, University of Michigan: Consumer Sentiment [UMCSENT], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UMCSENT

(5) Board of Governors of the Federal Reserve System (US), Industrial Production: Manufacturing (NAICS) [IPMAN], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/IPMAN

(6) U.S. Census Bureau and U.S. Department of Housing and Urban Development, New One Family Houses Sold: United States [HSN1F], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/HSN1F

(7) U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE