Classification Model for Predicting Bear, Bull or Correction Market

Team #: 35

Data Analytics Business - MGT-6203

Team members:

Seychelle Shauna Khan; skhan414
Hareesh Kumar Ghanta; hghanta3
Kumar Subramanyan; ksubramanyan3
Cesar Henrique Ceneviz; cceneviz3
Praminda Mahesh Imaduwa-Gamage; pmwi3

11/20/2022

Overview

Problem Statement

The purpose of this analysis is to identify which analytics model is a best predictor of a stock market Bear Market, Market Correction, or Bull Market based on the analysis of six leading market indicators.

Background Information on chosen project topic

With the current economic, geopolitical, and inflationary landscape the United States and the world is facing, Group 35 decided to select a project which is timely and deeply impactful to everyone on the team. The U.S. stock market is currently in a Bear Market (Reuters.com, 2022) which has an impact for everyone in our group's retirement savings, and investments plans, along with a broader impact regarding the decisions our firms make towards hiring and developing new business initiatives. The repercussions of the Bear Market can already be seen in the U.S. tech industry layoffs which since July, 2022 has laid off 26,474 workers (CrunchBase, 2022), with the expectation to flow into other employment sectors.

In the current bleak and uncertain situation facing the economy and the stock market, we question if we could have predicted and prepared ourselves for this Bear Market or even the market Correction which preceded it. If so, which would be the best analytical model to predict it all three of the different market categories (Bear, Correction, and Bull).

We want to see if there are signals that can help predicting the market Correction or the bear market. As per Forbes, when a stock index falls more than 10% from a recent high, it is said to have entered a “Market Correction”, and if it falls more than 20%, then it defined as a Bear Market. (Forbes.com, 2022)

Primary Research Question

What is the best analytics model to predict a Bear Market, Market Correction, or Bull Market using a combination of leading economic indicators?

Data Wrangling

Overview of Data Sources

To answer the Primary Research Question, multiple datasets (Appendix 1) needed to be consolidated into a single dataset so that the interaction of various predictors could be compared when predicting a Market Correction of Bear Market. The following datasets were merged to consolidate the predictor variables into a single dataset (Appendix 2):

Data Source 1: U.S. Unemployment Rate

This data source contains the monthly U.S. Unemployment Rate tracked by the U.S. government. The data contains the data from 1948 up until Nov 2021. We will supplement this table with additional Unemployment data until Sept 2022. The data also contains unemployment data for different age groups, and for men, and women.

Data Source 2: U.S. M2 Money Supply, Personal Saving Rate, Real Disposable Personal Income

This data source contains the monthly U.S. M2 Money Supply, Personal Saving Rate, Real Disposable Personal Income from 1981 to Oct 2021. We will supplement this table with additional information until

Sept 2022. M2 is a measure of the U.S. market money supply that includes cash, checking deposits, and easily-convertible to money.

Data Source 3: Federal Reserve Funds Rate

This data source contains the monthly U.S. Federal Reserve Funds Rate (FFR), which is the interest rate at which depository institutions trade federal funds (balances held at Federal Reserve Banks). The data contains both a date for each month, and the Feds Interest Rate for that month.

Data Source 4: S&P500 Stock Market Data on Market Correction from 1928 to 2022

The data contains the dates when the S&P500 had a market Correction or Bear Market over the last 94 years. The data source also contains the highest Closing Stock Price of the S&P500 Index, PEAK DATE, before the market started its path towards a Correction or Bear market; along with the TROUGH DATE, which was the final bottoming off price of the Correction/Bear market before the stock price started to rebound. The Data source also contains the total stock price % loss indicating if it was categorized as a Correction (10%+), or Bear Market (20+), along with the quantity of days from the PEAK to the TROUGH.

Data Source 4 was necessary for adding the dependent variable to the consolidated dataset. Our plan for the analysis of the data and models for this project is outlined in Appendix 3.

Data Enrichment

Each Data Source experienced transformation before it was merged with the others. This usually involved editing their date columns into a single, standardized format. The following steps were taken:

1. The date column for each dataset was then split into 3 separate columns, day (DD), month (MM) and year (YYYY). If a dataset's date column only contained the month and year, then the column would be split into a month column and a year column. The year column was re-formatted so that it only included the last 2 digits of the year.
2. For datasets with a day column (Data Sources 1, 2 and 3), the datasets were merged on the 3 columns, DD, MM, YYYY. Through the use of an inner join, this produced an initial dataset with 490 rows for data from Jan 1st, 1981, to Jan 10th, 2021.
3. Data Source 4 was parsed in order to determine which dates marked the beginning of a bull market, bear market, or market correction.
4. The dates when a change occurred were the only ones which appeared in this dataset, so the merge was done using a left join between the month and year columns of the consolidated dataset, and the month and year columns of Data Source 4. There were many blank rows in the 'Scenario' column of this consolidated dataset, which was expected.
5. The empty rows were filled with the Scenario values of the previous change date.
6. Further analysis revealed that the tcs column contained 144 blanks, while the reer column contained 156 blanks. Not all of these blanks overlapped; omission of blanks would, thus, result in a much smaller dataset. To avoid this issue, the tcs and reer columns were removed.

Description of each of the consolidated data table's column names:

Independent Variables:

- FEDFUNDS: U.S. Federal Reserve Fund Rate (Interest Rate) – referred to as FFR in our analysis

- psr: U.S. Personal Saving Rate
- m2: U.S. M2 Money Supply
- dspic: U.S. Real Disposable Personal Income
- pce: Personal Consumption Expenditures
- reer: Real Broad Effective Exchange Rate
- ir: Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity
- ffer: Federal Funds Effective Rate
- tcs: Total Construction Spending
- indpro: Industrial Production: Total Index
- ccpi: Core Consumer Price Index
- unrate: U.S. Unemployment rate per month
- unrate_men: U.S. Unemployment rate for men per month
- unrate_women: U.S. Unemployment rate for women per month
- unrate_16_to_17: U.S. Unemployment rate for 16 to 17 age group
- unrate_18_to_19: U.S. Unemployment rate for 18 to 19 age group
- unrate_20_to_24: U.S. Unemployment rate for 20 to 24 age group
- unrate_25_to_34: U.S. Unemployment rate for 25 to 34 age group
- unrate_35_to_44: U.S. Unemployment rate for 35 to 44 age group
- unrate_45_to_54: U.S. Unemployment rate for 45 to 54 age group
- unrate_55_over: U.S. Unemployment rate for 55 and over age group

Dependent Variable:

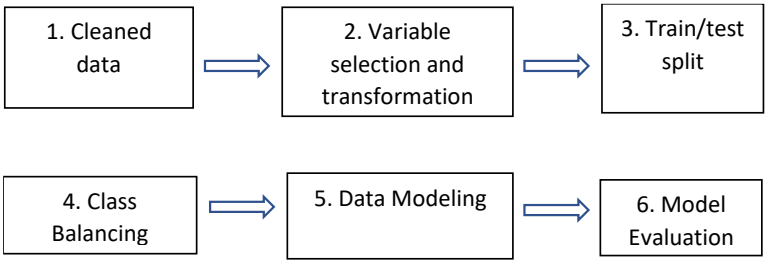
- Scenario: If market is in a Bull Market, Market Correction, or Bear Market

Methodology

Data Analysis and Modeling

The following flow chart summarizes our modelling approach.

Figure 1: Data modeling approach



In this modeling approach, the real broad effective exchange rate and total construction spending data columns were ignored as 29% and 32% data, respectively, are missing. All time-related data (year, month, date) were also ignored. Total ten predictors including the federal reserve fund rate, personal saving rate, M2 money supply, real disposable personal income, market yield on U.S. treasury securities at 10-year constant maturity, federal funds effective rate, industrial production: total index,

core consumer price index, and unemployment rate per month were selected. All these predictors are numerical time series data.

The scenario variable with three classes (bear, correction, and bull) was considered as the target variable. Multicollinearity analysis of the general unemployment rate, unemployment rate of men, unemployment rate of women, and unemployment rates of different age groups indicates that these timeseries are strongly correlated (adjusted. $R^2 > 0.80$). Hence, in data modeling, only the general unemployment rate variable was considered.

We conducted a basic correlation analysis between variables to understand the correlation between variables. Visual inspection of autocorrelation in the timeseries was performed by plotting the autocorrelation function of timeseries and lag-plots before and after the transformation of variables. Both original and transformed variables were subjected to Phillips-Perron unit root test to evaluate stationarity of timeseries.

The selected variables were transformed with second order differencing to minimize autocorrelation and non-stationarity in timeseries. These transformed data were randomly separated into train and test sets in 70:30 ratio. The train dataset was used for training multiple models while the test dataset was used for evaluating the model performance.

Our data consists of imbalance classes with correction as the major (345 instances) while bear (72) and bull (73) as minor classes. The SMOTE (Chawla, N.V. et al., 2002) oversampling method available in smotefamily, (Siriseriwan W, 2019), R-package was utilized to reduce the class imbalance in both train and test datasets.

We experimented with multinomial logistic regression, K-nearest neighbors (KNN), support vector machines (SVM), naïve Bayes, and random forest algorithms available in nnet (Venables WN, Ripley BD, 2002) caret (Kuhn M,2022), and e1071 (Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, 2022), and ranger (Wright MN, Ziegler A, 2017), R-packages. Table 1 provides a summary of functions and parameters used in each model.

Table 1: R-functions and parameters

Model	**Library, function**	**Parameters**
Naïve Bayes	e1071, naiveBayes	Default
SVM	e1071, svm	Default
Random forest	ranger, ranger	num.trees = 100, min.node.size = 10
KNN	caret, knn3	Default, k = 5
Multinomial logistic regression	nnet, multinorm	Default

The performance of these algorithms was evaluated with accuracy, macro measures of precision, recall, and F1-score. These measures were calculated using the *confusionMatrix* function available in the caret R-package.

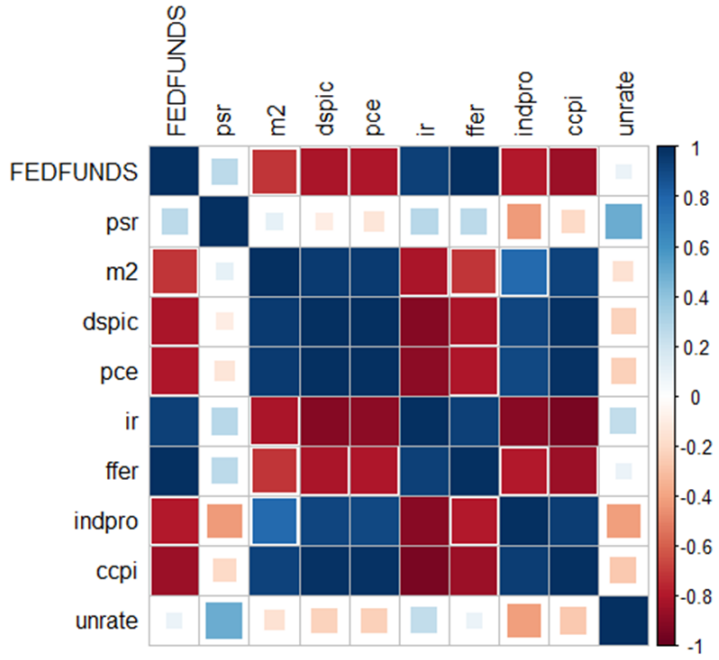
In this study, we did not conduct exhaustive hyperparameter search, variable selection, or sophisticated training performance evaluation with cross-validation as having timeseries data as predictors are experimental. Our focus was limited to finding a baseline model. Hence, we followed fundamental data modeling and model evaluation techniques.

Results and Discussion

a. Correlation Analysis

Two correlation plots were created and examined, as shown below.

Figure 2: Correlation plot for Fed Funds data, US M2 Money Supply data and Unemployment rate

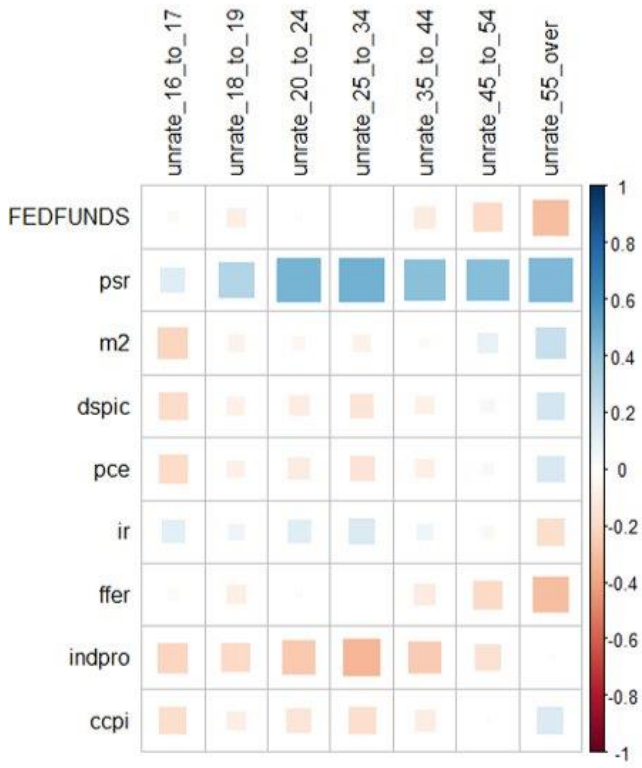


The diagram above shows correlations between various predictor variables from the Federal Reserve Funds rate data, Unemployment rate data and US Money Market Supply data.

FEDFUNDS, ffer and ir have strong, positive correlations, with Pearson correlation coefficients in the range [0.9,1]. FEDFUNDS, ffer and ir should be included in the models, although one of them may have to be discarded.

FEDFUNDS also has strong, negative correlation with m2, dspic, pce, indpro and ccpi, with correlation values falling in the range of [-0.7,-0.9]. Similarly, m2, dspic, pce,indpro and ccpi all have strong, positive correlations with each other [0.7,0.95]. Though it would be best to remove these predictors from the model, it would be useful to see their effects on the prediction.

Figure 3: Correlation plot between Fed Funds data, US M2 Money Supply data and Unemployment rates



The plot above shows correlation between unemployment for various age groups, against the predictors from the Federal Reserve Funds rate data and US Money Market Supply data.

The unemployment rates for different age groups are weakly and positively correlated to psr with correlation rates in the range [0.2,0.5]. The correlations between the other predictors and the unemployment rates are weaker [0,0.25]. The unemployment rates for ages 20 – 34 have the highest correlations with psr, with values approximating 0.47.

Additionally, the unemployment rates show negative correlation with the indpro predictor, with values in the range [-0.15, -0.30]. It is interesting to note that for m2, dspic, pce and ccpi, the unemployment rate for the age group 16 to 17 shows negative correlation, while the unemployment rate for the age group 55 and over shows positive correlation of equal magnitude.

b. Autocorrelation and Non-stationarity of time series

We initiated by exploring the time series data (Github Group 35, n.d.); we identified that there is an autocorrelation in time series data. We will explore techniques that we can use the time series data in a classification model as predictors.

Our goal is to reduce autocorrelation in time series using a suitable lag or differencing so that time series can be used as predictor variables in a classification model. Here we present an investigation of autocorrelation, non-stationarity, and differencing technique using two predictor variables for reducing autocorrelation and non-stationarity in time series.

We looked for non-stationarity in the time series using the Phillips-Perron unit root test. The Phillips-Perron test consists of evaluating whether time series was first order trend stationary with null hypothesis that it had a unit root and was not stationary. (Pillips & Perron, 1988)

Additionally, since the time series that we will be comparing below are non-stationary (i.e., increasing over time) we will need to transform the original time series data using second-order differencing. (Stationarity and Differencing, n.d.)

Figure 4 (a) shows federal reserve funds rate from January 1981 to January 2021 that has been declining nonlinearly. This appears to be a nonstationary process in which the mean and the variance do not remain constant long-term. The autocorrelation function in figure 4 (b) suggests that there exists a strong autocorrelation in this time series. The US federal reserve funding rate (FFR) against lag-5 of FFR shown in figure 4 (c) is an example for the existence of strong autocorrelation in the time series.

Figure 4: US federal reserve funding rate (FFR) Analysis

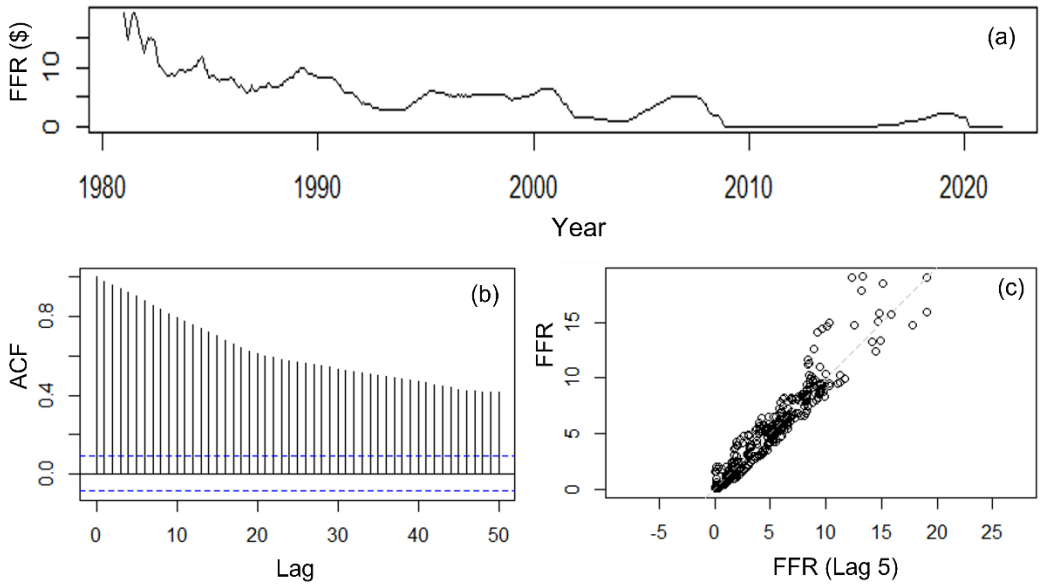


Figure 4: (a) US federal reserve funding rate (FFR) from Jan 1981 – Jan 2021.
(b) Autocorrelation function of federal funds rate (FFR). (c) Lag-5 plot of FFR

In order to obtain a stationary time series, the second-order differencing of the FFR (DD FFR) was computed as shown in figure 5 (a). The DD FFR exhibits nearly white noise characteristics with mean 0.006. Autocorrelation function in figure 5 (b) shows extremely poor autocorrelation. The Lag plot shown in figure 5 (c) exhibits no correlation between DD FFR time series and its lag -1.

Figure 5: Second-order differencing of federal reserve funds rate (DD FFR) Analysis

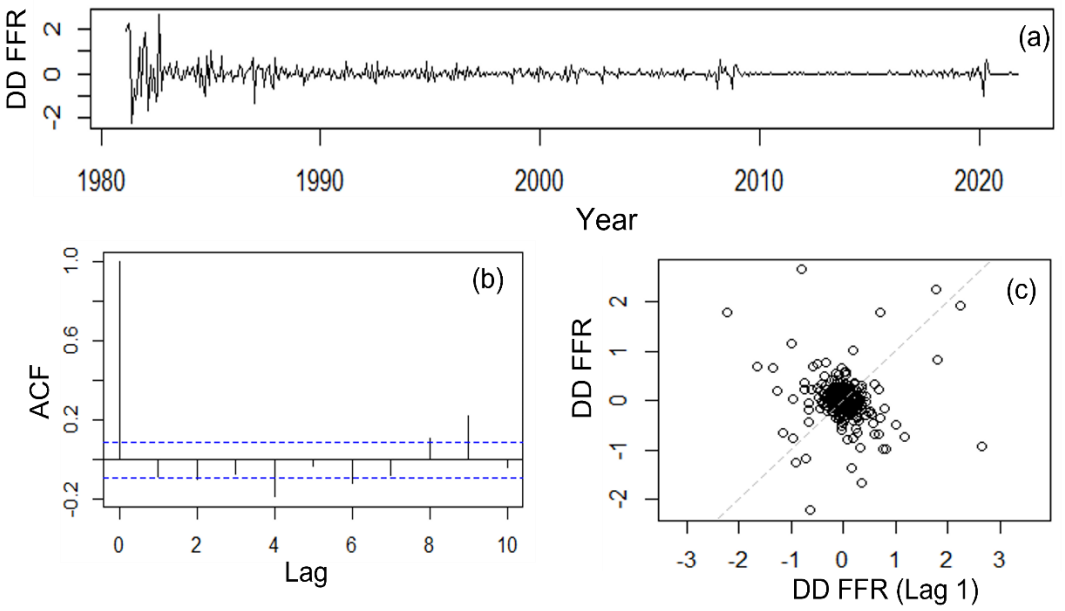


Figure 5: (a) Second-order differencing of federal reserve funds rate (DD FFR).
(b) Autocorrelation function of DD FFR. (c) Lag-1 plot of DD FFR

The unemployment rate is another predictor variable that is investigated in this project. Figure 6 (a) shows Unemployment rate of men in US from January 1981 to January 2021. The time series appears to be composed of seasonal and trend components. The autocorrelation function in figure 6 (b) suggests that there exists a strong autocorrelation in this time series. The unemployment rate of men against lag-5 of it is shown in figure 6 (c), which shows strong presence of autocorrelation in the time series.

Figure 6: Unemployment Rate for Men Analysis

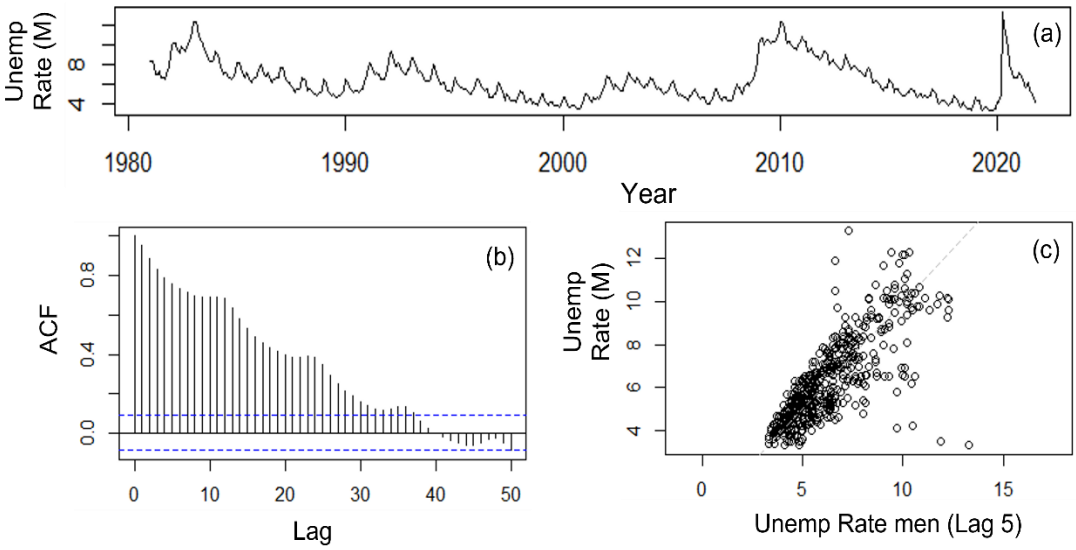


Figure 6 (a) Unemployment rate of men since Jan 1981 – Jan 2021. (b) Autocorrelation function of unemployment rate of men (c) Lag-5 plot of unemployment rate of men.

To obtain a stationary time series, the second-order differencing of the unemployment rate of men (DD unemployment rate (M)) was computed as shown in figure 7 (a). The second-order differencing of unemployment rate exhibits white noise characteristics with mean 0.001. Autocorrelation function in figure 7 (b) shows extremely poor autocorrelation. The lag plot shown in figure 47(c) exhibits no correlation between DD unemployment rate and its lag -1.

Figure 7: Second-order Differencing of Unemployment Rate of Men Analysis

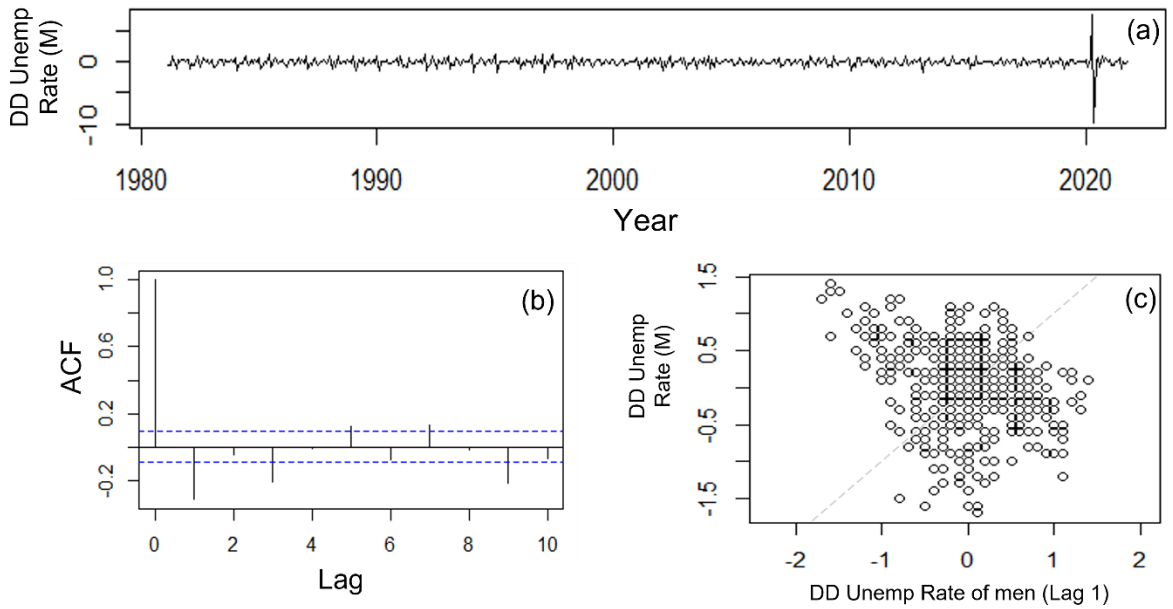


Figure 7: (a) Second-order differencing of unemployment rate of men (DD Unemployment Rate (M)). (b) Autocorrelation function of DD Unemployment rate (M). (c) Lag-1 plot of DD Unemployment rate (M)

For the Phillips-Perron test, values of 1 indicate rejection of the unit-root null hypothesis in favor of the alternative. Values of 0 indicate failure to reject the unit-root null hypothesis. When test statistics are outside critical values, Phillips-Perron test returns maximum (0.999) or minimum (0.001) p-values. (Mathworks, n.d.)

In both predictors discussed above, Phillips-Perron unit root test was carried out with original and the transformed time series. Both transformed time series passed the test for stationarity with p-value < 0.01.

c. Performance of Classification

Table 2: Model Performance

Model	Accuracy	Macro Precision	Macro Recall	Macro F1

Naïve Bayes	48%	63%	47%	42%
SVM	41%	37%	40%	35%
Random Forest	38%	36%	38%	33%
KNN	29%	31%	30%	29%
Logistic Reg	24%	18%	24%	20%

The performance measures of the tested classifiers are listed in table 2. The naïve Bayes model exhibited the highest accuracy (48%) among all the tested models. A better way to evaluate a classification model is to look at precision, recall, and F1 score. The naïve Bayes classifier outperformed other classifiers in all measures: precision (63%), recall (47%) and F1 (42%).

SVM and random forest algorithms exhibited close performances. The accuracy, precision, recall, and F1 were found to be similar for both classifiers. The accuracy of classification with KNN algorithm was better than that of the multinomial logistic regression. However, the logistic model exhibited the worst performance in precision, recall and F-1.

Table-3 shows the precision, recall and F1-score of the naïve Bayes classifier across bear, correction and bull scenarios and their average of value (macro measure). The classifier performs well having few false positives when predicting bear (70% precision) and bull (80%) scenarios, but it makes more errors when predicting the majority class, the correction (39%).

The recall score of 12% on the bull scenario suggests that although the model predicts bull class with high precision, it performs poorly on predicting other two classes with many false negatives. The highest recall score (87%) was reported for the correction scenario; although the model performs poorly on predicting the correction scenario it makes few errors on predicting the negative classes.

Table 3: Performance of Naïve Bayes classifier across classes

	Bear	**Correction**	**Bull**	**Macro Value**
Precision	70%	39%	80%	63%
Recall	42%	87%	12%	47%
F1	53%	53%	21%	42%

In summary, with 53% F1-score on both bear and correction classes, the model performs slightly above average on predicting both bear and correction and their respective negative classes simultaneously but performs below average on predicting the negative classes when predicting the bull scenario.

Challenges

- Missing data: The columns total consumption index (tcs) and real broad effective exchange rate (reer) had to be removed entirely; removing only the missing data rows would make the dataset too small. Imputation is an alternative option, but the time constraints of this project did not allow for it. Furthermore, imputation comes with its own set of risks and there was the potential for this method to negatively impact the accuracy of the model.

- Small dataset: Despite efforts to reduce shrinkage of the dataset, there were only 490 rows of data available for analysis. To preserve this, two columns (reer and tcs) had to be removed. The models would have been more accurate if there were more data points to train and test the models.
- Unbalanced dataset: The majority of the Scenario values were Corrections. Though oversampling methods such as SMOTE can aid in correcting this, SMOTE does not take into account neighboring examples that can be from other classes. Additionally, SMOTE is not very good for high-dimensional data. Under sampling is another option, but this would further reduce the number of rows available for modeling.

Conclusion

Through our project we analysed five different models to identify the best model for predicting and classifying when the S&P500 stock index will enter a Market Correction, Bear Market, or Bull Market. The five models were Multinomial Logistics Regression, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Naïve Bayes Model, and Random Forest. We utilized 70% of our data to train each of the models and utilized 30% of the data to evaluate each of the model's performance.

From our analysing the data from each of our 5 models it shows us that the Naïve Bayes model is the best for predicting each of the three types of markets, Market Correction, Bear Market, or Bull Market. From performance matrix in table 2 we saw that the Naïve Bayes had a better performance in accuracy, precision, recall, and F1 score. SVM, and Random Forest didn't perform as well as Naïve Bayes, however they performed similar to each other. While KNN and Logistic Regression were the two worst performing models, in that order.

In conclusion, having established from our analysis that the Naïve Bayes model is the best for predicting each of the three types of market conditions with the economic and market indicators that we analysed, our team can continue to use and analyse the market with the Naïve Bayes model. Which with the deep impact that a Market Correction or Bear Market have on companies, the job market, and retirements funds, and the current economic, geopolitical, and inflationary landscape the United States and the world is facing, being able to predict the next market scenario is crucial.

References

CrunchBase. (2022, Nov). Retrieved from CrunchBase: https://news.crunchbase.com/economy/tech-layoffs-2022-meta-amzn-twtr/

Forbes.com. (2020, March). Retrieved from Markets: https://www.forbes.com/sites/katinastefanova/2020/03/12/2020-bear-market-demystified-in-bullet-points/?sh=2f1653438856

Forbes.com. (2022). Retrieved from https://www.forbes.com/advisor/au/investing/what-is-market-correction/

Reuters.com. (2022). Retrieved from Markets: https://www.reuters.com/markets/europe/bear-market-beckons-us-stocks-2022-descent-deepens-2022-06-13/

Mathworks. (n.d.). Retrieved from PPTEST: https://www.mathworks.com/help/econ/pptest.html

Pillips, P. C., & Perron, P. (1988). Testing for a Unit Root in Time Series Regression. Niometrika. Retrieved from PPTest.

Stationarity and Differencing. (n.d.). Retrieved from OTexts: https://otexts.com/fpp2/stationarity.html

Github Group 35. (n.d.). Retrieved from https://github.gatech.edu/MGT-6203-Fall-2022-Canvas/Team-35/tree/main/Code

Chawla, N.V. et al., (2002). Retrieved from SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp. 321–357

Siriseriwan W (2019). _smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE_. R package version 1.3.1.

Venables WN, Ripley BD (2002). _Modern Applied Statistics with S_, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/

Kuhn M (2022). _caret: Classification and Regression Training_. R package version 6.0-93, https://github.com/topepo/caret/

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2022). _e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien_. R package version 1.7-12.

Wright MN, Ziegler A (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." _Journal of Statistical Software_, *77*(1), 1-17. doi:10.18637/jss.v077.i01 https://doi.org/10.18637/jss.v077.i01

Appendix

Appendix 1: Data Sources

No.	Title	Link
1	S&P500 Contraction and Bear Market Data	https://www.yardeni.com/pub/sp500corrbeartables.pdf
2	S&P 500 (^GSPC) Index – Yahoo Finance	https://finance.yahoo.com/quote/%5EGSPC/history?period1=-1325635200&period2=1665187200&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true
3	U.S. Inflation Rate	https://www.kaggle.com/datasets/neelgajare/usa-cpi-inflation-from-19132022
4	U.S. Unemployment Rate	https://www.kaggle.com/datasets/axeltorbenson/unemployment-data-19482021
5	U.S. M2 Money Supply, Personal Saving Rate, Real Disposable Personal Income	https://www.kaggle.com/code/calven22/forecasting-inflation-with-arima-and-lstm/data
6	Federal Reserve Funds Rate	https://fred.stlouisfed.org/series/fedfunds

Appendix 2: Consolidated Data for Analysis

	DATE	FEDFUNDS	YYYY	MM	DD	psr	m2	dspic	pce	reer	ir	ffer	tcs	indpro	ccpi	unrate	unrate_men	unrate_women	unrate_16_to_17	unrate_18_to_19	unrate_20_to_24	unrate_25_to_34	unrate_35_to_44	unrate_45_to_54	unrate_55_over	Scenario
2	01/01/1981	19.08	81	1	1	10.9	1612.9	4980.4	1870	NA	12.56857143	19.08451613	NA	51.1668	85.4	8.2	8.3	8	22.2	19	13.3	8	5.5	4.6	3.9	Bear
3	01/02/1981	15.93	81	2	1	10.8	1608.125	4965	1884.2	NA	13.19444444	15.93428571	NA	50.9509	85.9	8	8.3	7.7	23.1	19.3	13	7.7	5.2	4.6	4.2	Bear
4	01/03/1981	14.7	81	3	1	10.8	1629.4	4979	1902.9	NA	13.11590909	14.70387097	NA	51.2066	86.4	7.7	7.9	7.4	23.2	17.6	12.6	7.2	5.2	4.3	4.1	Bear
5	01/04/1981	15.72	81	4	1	10.9	1665.575	4965.1	1904.4	NA	13.67952381	15.719	NA	50.9711	87	7	6.9	7.1	21.2	16.1	11.6	6.6	4.9	3.5	3.5	Bear
6	01/05/1981	18.52	81	5	1	11	1655.15	4974.8	1913.8	NA	14.0995	18.51774194	NA	51.2645	87.8	7.1	6.9	7.5	20.5	17.3	12.4	6.9	4.2	3.8	3.3	Bear
7	01/06/1981	19.1	81	6	1	10.8	1664.5	5001.9	1934.5	NA	13.47227273	19.09966667	NA	51.5247	88.6	7.7	7.3	8.3	26.4	18.7	12.6	6.8	4.5	3.9	3.1	Bear
8	01/07/1981	19.04	81	7	1	12.3	1685.225	5080.8	1942.1	NA	14.28090909	19.03580645	NA	51.8727	89.8	7.3	6.7	8.2	19.1	17	11.4	7	4.4	4.1	3.3	Bear
9	01/08/1981	17.82	81	8	1	12	1693.28	5095.9	1966.6	NA	14.93714286	17.81774194	NA	51.8156	90.7	7.2	6.6	8	17.4	15.7	11.5	7.1	4.6	4.3	3.4	Bear
10	01/09/1981	15.87	81	9	1	12.4	1706.15	5087.2	1965.5	NA	15.32380952	15.874	NA	51.5692	91.8	7.3	6.5	8.4	20.2	19.6	11.9	6.8	5	3.7	3.5	Bear
11	01/10/1981	15.08	81	10	1	13	1725.45	5093.8	1963.9	NA	15.14809524	15.08	NA	51.1851	92.1	7.5	6.9	8.3	20.4	19.9	12	7.3	5	4.2	3.4	Bear
12	01/11/1981	13.31	81	11	1	13.2	1738.54	5096.9	1970.6	NA	13.39277778	13.307	NA	50.5914	92.5	7.9	7.8	8.2	23.4	20.7	12.2	7.6	5.6	4.5	3.6	Bear

Append 3: Project Timeline and Planning

Phase	Description	Team's Milestone Dates	Class Milestone Dates
1	Team Formation Sheet	Team members submitted (complete)	9/18 11:59 pm
2	Project Proposal	Project objectives and plan determined. • Datasets chosen – 9/29 • Models chosen – 10/3 • Training procedures chosen – 10/4	10/9 11:59 pm
3	Project Progress Presentation Video (4-5 min)	Various models have been explored and are in the process of being analyzed. • Models created – 10/15 • Models cross-validated – 10/20 • Results in the process of being analyzed – 10/22 • Progress presentation and report created - 10/28	10/30 11:59 pm
	Project Progress Report (4-5 pages)	• Report created – 10/28	10/30 11:59 pm
4	Final Presentation Video (10-12 minutes)	Analysis complete; research questions answered • Analysis of results complete – 11/01 • Most accurate model determined – 11/01 • Documentation of results and which models met our expectations – 11/05 • Best predictors noted based on classification models – 11/07 • Final presentation created – 11/15	11/16 11:59 pm
	Final Report (8-10 pages)	• Report created – 11/18	11/20 11:59 pm
	Final Video Presentation Slides, Code, Data, etc.	• Materials organized for presentation – 11/18	11/20 11:59 pm
5	Peer Review: Out-of-Group Final Video Presentation	• Video presentation created – 11/22	11/23 11:59 pm
	Peer Review: Within-Group Performance Evaluation	• Evaluation completed – 11/22	11/23 11:59 pm