PROJECT REPORT : RECESSION

Content Table

Our team

Our team is composed of five members who worked on the following tasks :

- Tim Risse : drafting our presentations, performing basic data analysis, writing the introduction of our paper and structuring of the report
- Daein Cho : analyzing our main findings and carrying out the VIF tests
- Kurumi Shimpo : coding the “test and train split”, coding the multiple logistic regression and writing the literature review
- Wenyuan Wang : working on a html visualization of our dataset
- Romain Fernex : Performing the PCA, coding the multiple logistic regression and the ROC curves, writing the conclusion of the report

Executive Summary:

Structure of the Research Paper:

This research paper is structured in the following way: introduction, literature review, research methodology, data analysis, results, conclusions and discussion and references. The introduction provides background information on the topic and sets the stage for the rest of the paper. The literature review summarizes and critically evaluates existing research on the topic. The research methodology section describes the methods used to collect and analyse data. The data analysis, results, and discussion sections present and interpret the findings of the study. The conclusion summarizes the main findings and their implications.

Research Question:

What are the key macroeconomic indicators that predict downturn in the business cycle in the US?

Introduction:

The topic of this research is the identification of key macroeconomic indicators that predict market recession in the US. Economic recessions are characterized by a decline in economic activity and can have a significant impact on all stakeholders operating within that economy. Market recessions can lead to significant losses in life but also corporate downturn and decreases in stock prices, bond yields, and other financial assets. Therefore, it is important to understand the indicators that signal an impending recession to take proactive steps to mitigate its effects. This research aims to investigate and identify the key macroeconomic indicators that have been shown to predict downturns in business cycles in the US. The question of which macroeconomic indicators predict market recessions in the US is interesting because it has significant implications for both policymakers and investors. Understanding these indicators can help policymakers take proactive steps to mitigate the effects of a recession, while investors can use this information to make informed decisions about their investments. Additionally, studying past recessions and the indicators that preceded them can provide valuable insights into the underlying economic conditions that lead to market downturns.

Literature Review:

The current economic and geopolitics conjuncture has raised concerns all across the world regarding the potential arrival of a new recession. Though these fears have gained in intensity in the recent years, they are in no way new, especially when looking at crises such as the 2008 market crash. Therefore, analysts throughout the world have tried to identify potential predictors for such downturns in the business cycle in order to cushion their impact. Currently, in the US, half of the fifty states have already witnessed a slowing down of their economies which prompted institutions such as the **Saint Louis Federal Reserve Bank (2022)** to warn against the risk of an incoming recession in the coming months. In it latest report the **San Francisco Federal Bank (2022)** backed this warning, this time by highlighting the rise in the unemployment rate which is often regarded as a reliable signal of a downturn. Interestingly, the report also specified that unemployment rate is the most reliable when looking at near term prediction and tends to lose in relevance to the benefit of the market yield curve as the time horizon increases. Apart from these first two indicators, significant attention has also been given to indicators closely linked to the real economy such as consumption and firm's performance indexes. These are indeed often tied to recession, as shocks tend to follow decrease in spendings and investments due to ailing confidence from consumers and investors.

Most of these insights stem from economists who have already identified a wide array of different indicators that perform well in predicting recession and which we have decided to focus on. Firstly, as regards financial indicators, studies have shown that the S&P 500 composite stock price index is especially meaningful (A. Estrella & S.Michkin, 1998) for short term forecasts. Then, following the 2001 US recession, others have instead focused on monetary and labor indicators, such as the aforementioned unemployment rate, and asserted that they are close to financial indicators in terms of prediction quality (Mehdi Mostaghimi, 2006). Lastly we needed to find an appropriate dependent variable that would faithfully tell us when periods of recession took place. To this end we turned to the recession index published by the National Bureau of Economic Research (NBER) which has been used in most studies in this field and has already been used to identify the best recession indicators for different time horizons (Weiling Liu & Emmanuel Moench, 2016)

Research Methodology:

The dataset we are using is sourced from Kaggle and has the name "US macro-economic data [1996- 2020], source: FRED" and contains two datasets: one for recession periods and one for macroeconomic indicators. The data is collected monthly and covers the time range of 1996-2020. The recession periods dataset contains dummy variables indicating whether the economic state is in a normal or recessionary period. The macroeconomic indicators dataset is categorized into several groups including "Output and Income, Labour Market, Real Estate, Consumption and Expenditures,

Money and Credit, Interest and Exchange Rates, Prices, Stock Market". Each of these groups contains several variables such as Real Personal Income, Civilian Employment, New Private Housing Permits, Real Personal Consumption Expenditures, Money Stock, Effective Federal Funds Rate, Finished Goods, and the 500 S&P's Common Stock Price Index.

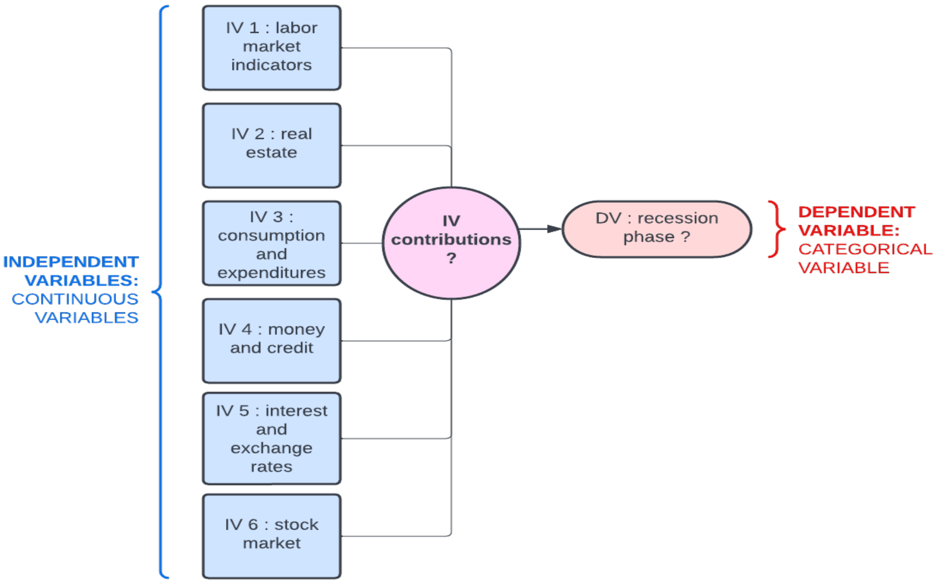
The variables used in the study include the following:

Dependent Variable: State of Economy including "Recession" (1) and "Normal" (0)

Independent Variables: Following the VIF analysis (see below) we retained 8 main indicators

- ISRATIO: Ratio of inventories to sales for all US businesses
- DTCTHFNM: Total consumer loans and leases outstanding
- EXUSUK: US/UK Foreign exchange rate
- SP500: 500 S&Ps Common stock price index (composite)
- NASDAQ: Composite index
- GOLDBAR: gold exchange rate
- P/E: S&P price to earnings ratio
- Dividend Yield: S&P Dividend yield

For the analysis a multiple logistic regression is being used. The multiple logistic regression is a statistical method used to analyse the relationship between multiple independent variables (also known as predictor variables) and a binary dependent variable. The goal of this method is to model the probability of the occurrence of the dependent variable as a function of the independent variables. The logistic regression model uses a logistic function to model the probability of the occurrence of the dependent variable, with the function output ranging between 0 and 1. The independent variables are represented by coefficients (or beta values) in the model, which are estimated through the optimization process. The coefficients indicate the strength and direction of the relationship between the independent variables and the dependent variable.


INDEPENDENT VARIABLES: CONTINUOUS VARIABLES
IV 1 : labor market indicators
IV 2 : real estate
IV 3 : consumption and expenditures
IV 4 : money and credit
IV 5 : interest and exchange rates
IV 6 : stock market
IV contributions ?
DV : recession phase ?
DEPENDENT VARIABLE: CATEGORICAL VARIABLE

Research Design (multiple logistic regression)

Exploratory Data Analysis

Basic observations :

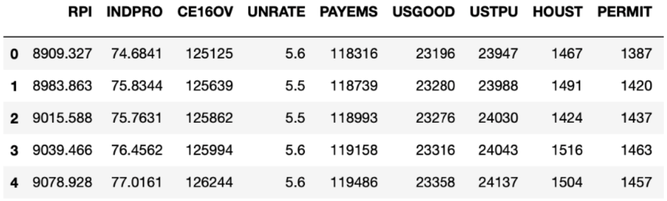
Before conducting an analysis with the data set, several steps must be taken to prepare the data for analysis. One important step is to check for missing values and handle them appropriately. This could include dropping rows or columns with missing value or imputing the missing values with a statistical method. The .describe() function in python is used to generate descriptive statistics of a data set.

It provides various summary statistics of the data such as the mean, standard deviation, minimum, maximum, and quartiles. It is particularly useful for numerical data and it's a good first step to get an overview of the data, understand the distribution of the data and detect any outliers or anomalies in the data. The .info() function in python is used to get a detailed summary of a DataFrame or a Series. It provides information about the data types of the columns, the number of non-missing values and the memory usage. This function is useful for understanding the structure of the data, the types of the variables and the completeness of the data. It is also useful to make sure that the data types of the variables match the expected type.

From both the .describe and .info function we could identify that the data set has high quality and no missing data points, which allowed us to proceed with the data analysis.

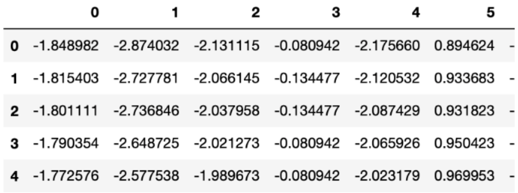
Scaling

After using the .describe() method as well as performing a visual inspection of the data using matplotlib we decided to scale the variables in our dataset in order to make the results easier to interpret. To this end, we relied on sklearn.prepocessing StandardScaler which scales the data to ensure that all variables follow a standard normal distribution. While the dataset still remains sensitive to outliers, scaling the data to the unit variance is vital in order to reduce risks of misclassification error and improve the accuracy rate of our model. Therefore, we limit the risk of a variable contributing more than another due only to a difference in scale. Here is an overview of the dataset before and after the rescaling:



	RPI	INDPRO	CE16OV	UNRATE	PAYEMS	USGOOD	USTPU	HOUST	PERMIT
0	8909.327	74.6841	125125	5.6	118316	23196	23947	1467	1387
1	8983.863	75.8344	125639	5.5	118739	23280	23988	1491	1420
2	9015.588	75.7631	125862	5.5	118993	23276	24030	1424	1437
3	9039.466	76.4562	125994	5.6	119158	23316	24043	1516	1463
4	9078.928	77.0161	126244	5.6	119486	23358	24137	1504	1457

Original Dataset

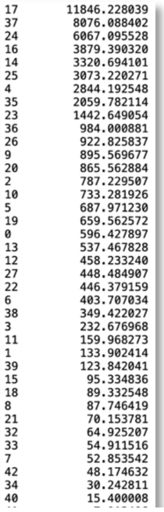


	0	1	2	3	4	5	
0	-1.848982	-2.874032	-2.131115	-0.080942	-2.175660	0.894624	-
1	-1.815403	-2.727781	-2.066145	-0.134477	-2.120532	0.933683	-
2	-1.801111	-2.736846	-2.037958	-0.134477	-2.087429	0.931823	-
3	-1.790354	-2.648725	-2.021273	-0.080942	-2.065926	0.950423	-
4	-1.772576	-2.577538	-1.989673	-0.080942	-2.023179	0.969953	-

Scaled Dataset

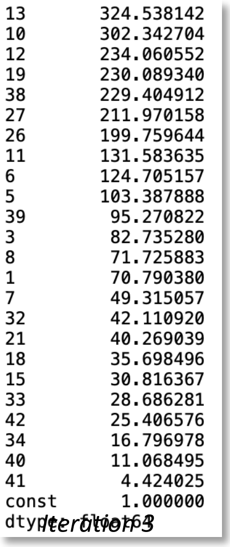
VIF (Multicollinearity test)

To test for multicollinearity we decided to take a stepwise approach using statsmodels variance_inflaction_factor function. We thus performed several VIF analyses and removed variables with extreme VIF scores each time until we were left with variables which all had a VIF score inferior to 5.
After 5 successive iterations of the method we obtained the following results :



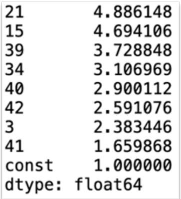
```
17      11846.228039
37       8076.088402
24       6067.095528
16       3879.390320
14       3320.694101
25       3073.220271
4        2844.192548
35       2059.782114
23       1442.649054
36        984.000881
26        922.825837
9         895.569677
20        865.562884
2         787.229507
10        733.281926
5         687.971230
19        659.562572
0         596.427897
13        537.467828
12        458.233240
27        448.484907
22        446.379159
6         403.707034
38        349.422027
3         232.676968
11        159.968273
1         133.902414
39        123.842041
15         95.334836
18         89.332548
8          87.746419
21         70.153781
32         64.925207
33         54.911516
7          52.853542
42         48.174632
34         30.242811
40         15.400008
```

Iteration 1



```
13       324.538142
10       302.342704
12       234.060552
19       230.089340
38       229.404912
27       211.970158
26       199.759644
11       131.583635
6        124.705157
5        103.387888
39        95.270822
3         82.735280
8         71.725883
1         70.790380
7         49.315057
32        42.110920
21        40.269039
18        35.698496
15        30.816367
33        28.686281
42        25.406576
34        16.796978
40        11.068495
41         4.424025
const      1.000000
dtype: float64
```

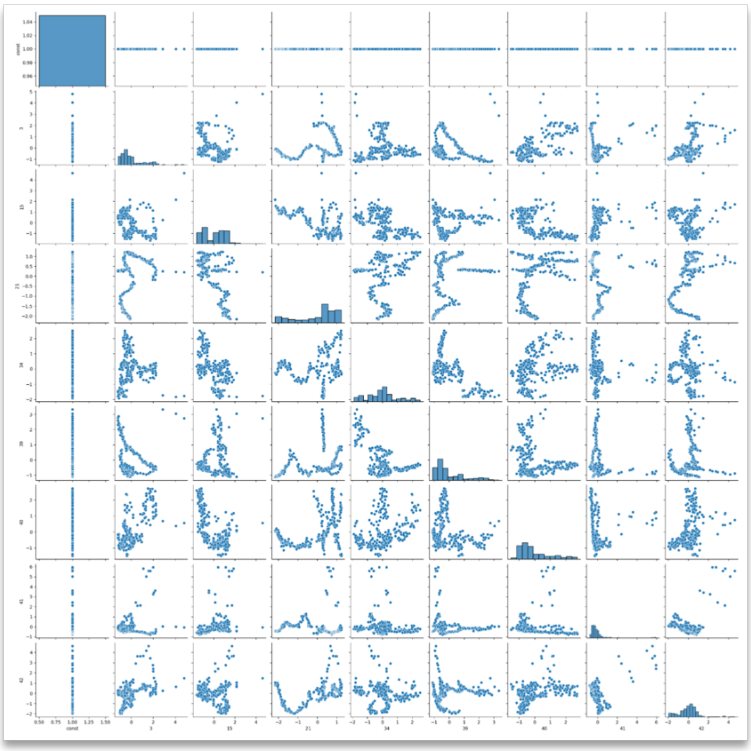
Iteration 3



```
21       4.886148
15       4.694106
39       3.728848
34       3.106969
40       2.900112
42       2.591076
3        2.383446
41       1.659868
const    1.000000
dtype: float64
```

Iteration 5

The numbers correspond to the aforementioned independent variables that we decided to keep in our study. We then performed visual inspection using seaborn's pairplot function to generate a correlation matrix with scatter plots.



Correlation matrix obtained with pairplot

Since no variables exhibited a significant linear relation, we concluded that the VIF analysis was successful and moved on to perform the multivariate logistic regression after splitting our data.

Principal Component Analysis

As the VIF test method resulted in removing too many variables, we decided to rely on unsupervised machine learning methods to tackle this issue. We thus carried out a Principal Component Analysis after dividing our dependent variables in two main categories :

- Financial indicators: S&P PE ratio, S&P dividend yield, US/UK Foreign exchange rate…
- Non-financial indicators : civilian unemployment rate, real personal income, new orders for consumer goods…

Using the PCA, we restricted the number of relevant variables to 8 for each category, giving us a total of 16 dependent variables. This enabled us to improve the model quality, as will be shown below. However this also made interpreting our results more difficult as we cannot know what the variables created by the PCA exactly describe.

Test and Train split

After performing the exploratory data analysis (EDA), we split the data set into a training and testing set using the formula ' train_test_split.' For the training size, we changed the proportion from 80% to 60% of our dataset so as to get more exploitable results for the out-of-sample ROC curve.

This training set is used to generate the logistic regression model which is fitted to the dataset to ensure a sufficient level of in-sample fit. Then, testing set is used to access the quality of the regression model and control for out-of-sample fit. The variables which passed the previous multicollinearity test (or those generated after the PCA in the case of the PCA approach) are used as 'x,' and the dataset containing the values of the dependent variable is used as ‘y’.

The result shows that the dataset is successfully divided into two parts. 60% of the data set goes for the training method of each x and y, and 40% for testing.

In detail, the data sizes for each are listed below :

- Observation in train data: 175
- Observation in test data:118

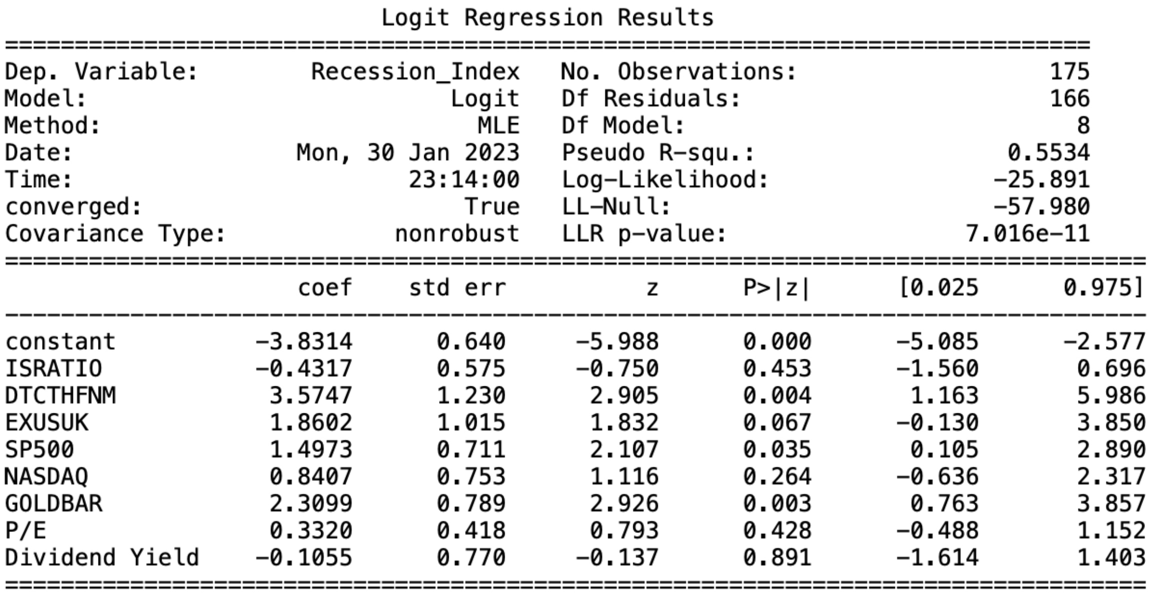
Results:

Multiple logistic regression : process summary

Three steps mainly create the multiple regression processes. Firstly, as explained above, test-train split and VIF testing or PCA are conducted to clean and organise our dataset before creating the data model. Then, as the second step, based on dataset we splitted earlier, we can generate the logistic regression model. This model is fitted it to the training data to ensure a sufficient in-sample fit. After those two steps, we finally use our testing set to assess the quality of the model and control for out-of-sample fit through visualizing the ROC curves for each model.

Results : Multiple Logistic Regression (VIF method)

Consequent to test and train split, the Statsmodel **sm.Logit().fit()** function is applied to find out the ßs and p-values of each independent variable. The results are shown in the chart below which was generated using the **.summary()** function :

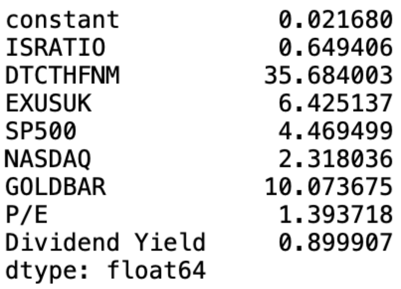


```
                           Logit Regression Results
==============================================================================
Dep. Variable:        Recession_Index   No. Observations:                  175
Model:                          Logit   Df Residuals:                      166
Method:                           MLE   Df Model:                            8
Date:                Mon, 30 Jan 2023   Pseudo R-squ.:                  0.5534
Time:                        23:14:00   Log-Likelihood:                -25.891
converged:                       True   LL-Null:                       -57.980
Covariance Type:            nonrobust   LLR p-value:                 7.016e-11
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
constant          -3.8314      0.640     -5.988      0.000      -5.085      -2.577
ISRATIO           -0.4317      0.575     -0.750      0.453      -1.560       0.696
DTCTHFNM           3.5747      1.230      2.905      0.004       1.163       5.986
EXUSUK             1.8602      1.015      1.832      0.067      -0.130       3.850
SP500              1.4973      0.711      2.107      0.035       0.105       2.890
NASDAQ             0.8407      0.753      1.116      0.264      -0.636       2.317
GOLDBAR            2.3099      0.789      2.926      0.003       0.763       3.857
P/E                0.3320      0.418      0.793      0.428      -0.488       1.152
Dividend Yield    -0.1055      0.770     -0.137      0.891      -1.614       1.403
==================================================================================
```

Results table (with VIF)

Looking at the p-value, which measures the statistical significance of each variable, we identify that only three variables appear to be significant for a 95% significance level: DTCTHFNM, SP500, and GOLDBAR. The table also returns the ß coefficients of the regression model, therefore we need to compute the exponential of each coefficient in order to get the odd ratios.

Using **np.exp(log_reg.params)**, we get the following table:



```
constant           0.021680
ISRATIO            0.649406
DTCTHFNM          35.684003
EXUSUK             6.425137
SP500              4.469499
NASDAQ             2.318036
GOLDBAR           10.073675
P/E                1.393718
Dividend Yield     0.899907
dtype: float64
```

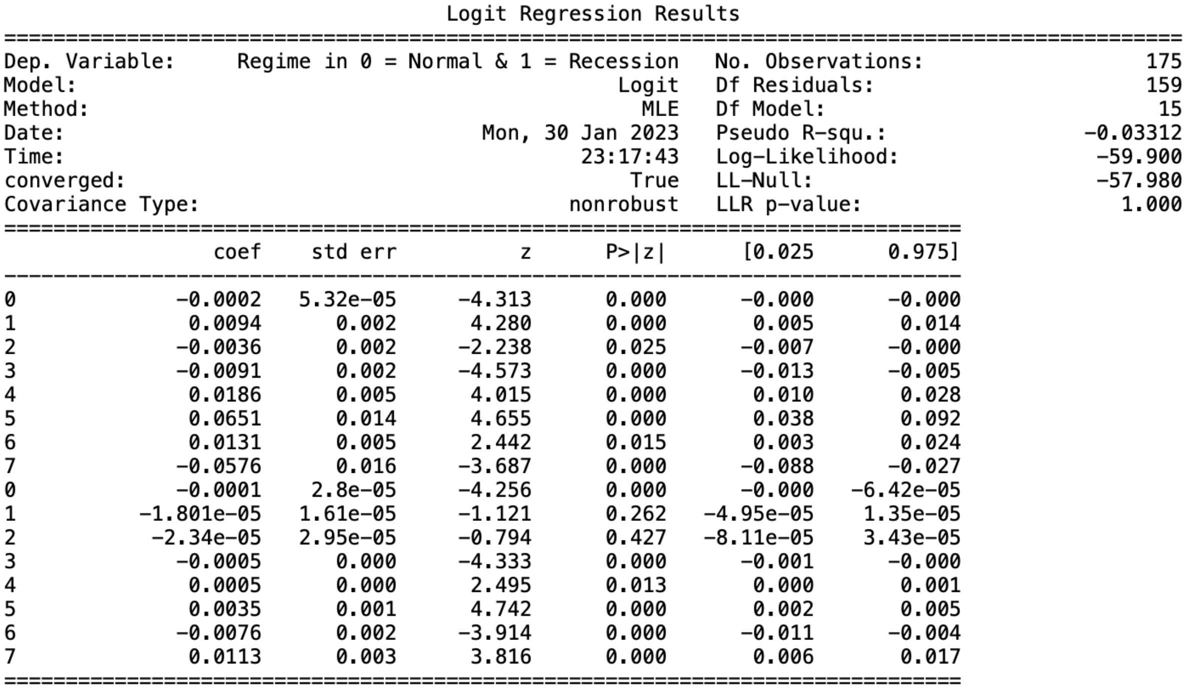
Odds ratios

With the above odd ratios, we can conclude the following things (holding all other variables constant,):

- with one unit increase in DTCTHFNM, the odds of recession increase by 35.684003
- with one unit increase in SP500, the odds of recession increase by 4.469499
- with one unit increase in GOLDBAR, the odds of recession increase by 10.073675

Results : Multiple Logistic Regression (PCA method)

We also applied multiple logistic regression following the dimensionality reduction using PCA. The method of MLR and its results are no different from the previous VIF method. However, the interpretation is very different. The following chart is the result of multiple logistic regression after PCA. The top eight variables are financial group of variables and the bottom eight are non-financial group. Each individual variables does not give us much, information, but, we can infer the characteristics of the two groups by examining the number of significant variables and the sign (positive or negative) of the ß coefficients.



```
                                    Logit Regression Results
================================================================================================
Dep. Variable:     Regime in 0 = Normal & 1 = Recession   No. Observations:                  175
Model:                                            Logit   Df Residuals:                      159
Method:                                             MLE   Df Model:                           15
Date:                                  Mon, 30 Jan 2023   Pseudo R-squ.:                -0.03312
Time:                                          23:17:43   Log-Likelihood:                -59.900
converged:                                         True   LL-Null:                       -57.980
Covariance Type:                              nonrobust   LLR p-value:                     1.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
0             -0.0002   5.32e-05     -4.313      0.000      -0.000      -0.000
1              0.0094      0.002      4.280      0.000       0.005       0.014
2             -0.0036      0.002     -2.238      0.025      -0.007      -0.000
3             -0.0091      0.002     -4.573      0.000      -0.013      -0.005
4              0.0186      0.005      4.015      0.000       0.010       0.028
5              0.0651      0.014      4.655      0.000       0.038       0.092
6              0.0131      0.005      2.442      0.015       0.003       0.024
7             -0.0576      0.016     -3.687      0.000      -0.088      -0.027
0             -0.0001    2.8e-05     -4.256      0.000      -0.000   -6.42e-05
1          -1.801e-05   1.61e-05     -1.121      0.262   -4.95e-05    1.35e-05
2           -2.34e-05   2.95e-05     -0.794      0.427   -8.11e-05    3.43e-05
3             -0.0005      0.000     -4.333      0.000      -0.001      -0.000
4              0.0005      0.000      2.495      0.013       0.000       0.001
5              0.0035      0.001      4.742      0.000       0.002       0.005
6             -0.0076      0.002     -3.914      0.000      -0.011      -0.004
7              0.0113      0.003      3.816      0.000       0.006       0.017
==============================================================================
```

Results table (with PCA)

The characteristics of the two groups are as follows:

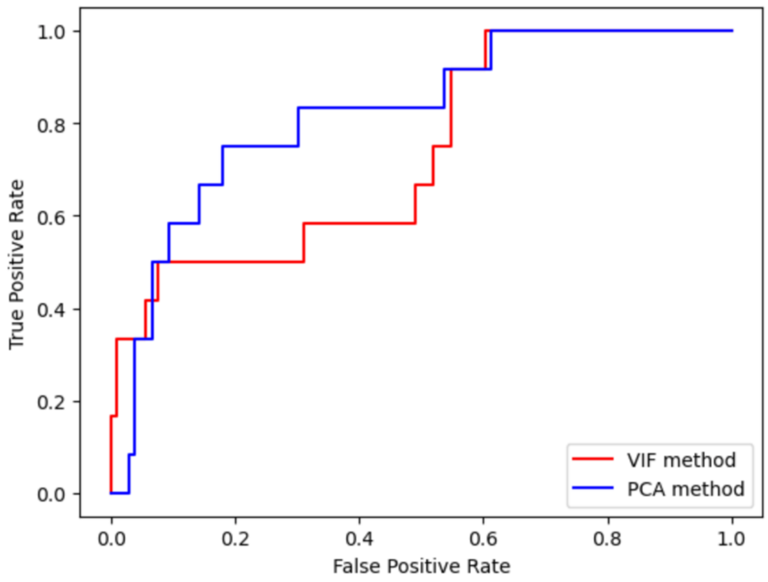
Group	Significant Variables	Positively Correlated
Financial Variables	8 /8	4 / 8
Non-financial Variables	6 / 8	3 / 8

From the above results, we may infer that financial variables are slightly more likely to be relevant predictors of regression compared to non-financial variables. This is consistent with most of the existing scientific literature which tends to place more weight on financial variables. While it is difficult to determine what each variable corresponds to, we observe that there are slightly less variable which contribute positively to the chance of entering a regression.

Finally, the odds ratio measured tend to take values that are much closer to one compared to the previous examples. This may explain why this model turns out to be a more efficient predictor compared to previous model where a few variable have a significant impact on the classification obtained.

Comparison of model performance : Out-of-sample ROC curves

After creating these two models, we tried to compare their respective performance in order to evaluate which one was the most efficient predictor of an incoming recession. To this end, we plotted the out-of-sample Receiver Operating Chatacteristic (ROC) curve for each model. This yielded the following graph :

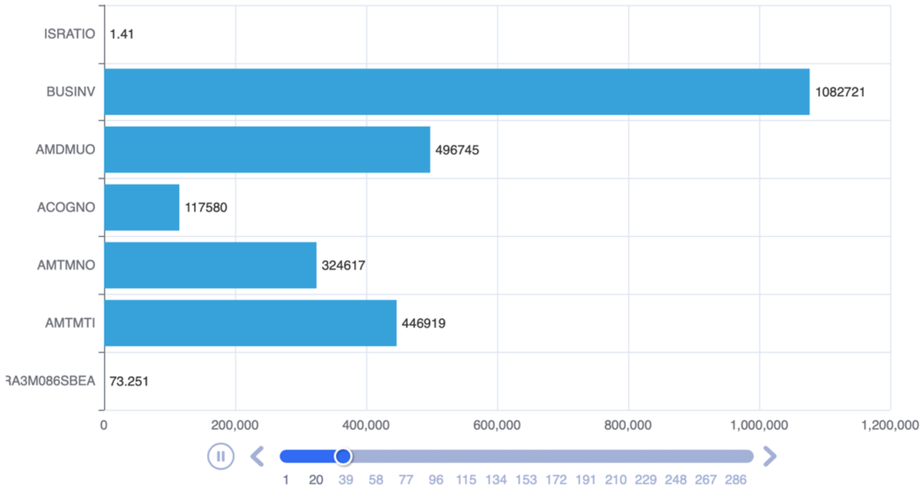


ROC curves obtained with the PCA and VIF methods

We observe that the curve with the highest AUC is the one we obtained with the PCA method. In other words, the model generated after using the PCA tends to perform better than the one obtained with the VIF method. Therefore, we may infer that the PCA based model is the most efficient in terms of out-of-sample fit. That being said, It is worth noting that we made the opposite statement when lowering the size of the test set to only 20% of the source dataset (against 40% in the case of the above graph). Nonetheless, both curves remain quite close to the random guess curve, which means that the model still has a wide margin for improvement.

Additional visualization on html

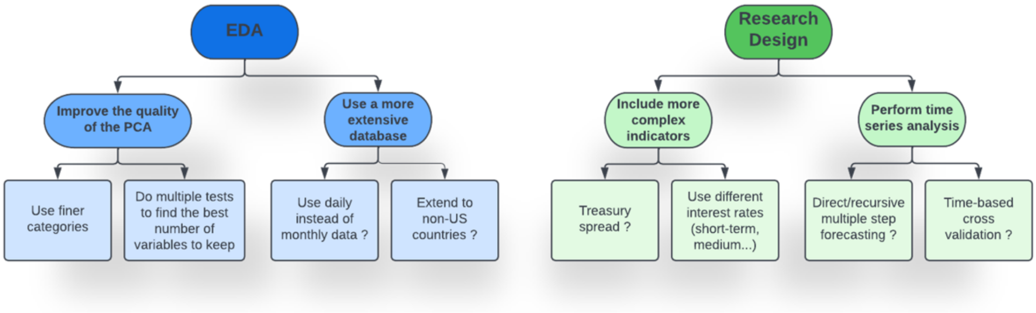
Parallel to our main study, we also worked on ways to visualize our dataset in html in a dynamic and clear way. Thus, we created the following graph which shows the evolution of consumption related indexes over time.



Graphic visualization obtained using python and html

Using this graph, it is easier to visualize which variables tend to change fastest before a recession and are therefore more likely to contribute strongly to the odds of one happening. Ideally we would like to extend this visualization to all categories and work on scaling the data in a more efficiant way in order to render the data even more readable. Further work showing a ranking of leading predictors based on different time horizons might also be interesting, although we unfortunately did not have the time nor experience to do it.

Conclusion : Limitations and Potential Improvements


EDA
Improve the quality of the PCA
Use a more extensive database
Use finer categories
Do multiple tests to find the best number of variables to keep
Use daily instead of monthly data ?
Extend to non-US countries ?
Research Design
Include more complex indicators
Perform time series analysis
Treasury spread ?
Use different interest rates (short-term, medium...)
Direct/recursive multiple step forecasting ?
Time-based cross validation ?

Main leads to improve our study

Exploratory Data Analysis

As mentioned earlier when discussing the principal component analysis approach, we only separated the independent variables in two main categories: financial and non-financial indicators. This made interpreting the results more complicated as we had less visibility on what each variables created by the PCA actually portrayed. A finer categorization might involve segmenting our datasets between employment, consumption, monetary, financial market indicators etc… This would enable us to better understand which specific factors have the highest impact on the probability of a recession happening.

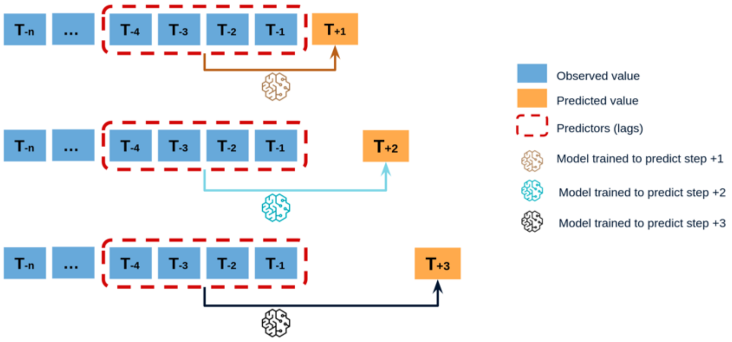
Another important limitation of our study lies in the relatively small database that we had access to. For instance, we had to restrain the size of our training set significantly in order to improve the quality of our ROC curves which suffered from a lack of datapoints in the testing set. This undoubtedly caused a loss of quality in terms of in-sample-fit and altered the quality of our model. Thus, finding a dataset with daily instead of monthly reporting would most likely improve our study significantly.

Research Design and Techniques Used

A core limitation of our study is linked to the dependent variables themselves. Most modern studies that cover this issue tend to include more complex and diverse factors such as interest rate spread or, most importantly, the yield curve. This last indicator has been proven to consistently outperform other indicators when it comes to predicting recession for a 1 to 2 years time horizon (Estrella & Mishkin, 1996). Therefore, a more detailed mathematical model might be needed in order to take these other variables into account in an efficient way.

In the medium term, our model would also gain using with other more advanced classification methods such as adaptive or gradient boosting. Looking at the ROC curve obtained with each of these methods might give us insight regarding which lead would be the most promising to maximize the efficiency of our model based on the existing dataset.

Finally, contrary to our initial objective, our analysis appears to be rather cross-sectional and not longitudinal. This implies that the conclusions derived from our study are more about the potential association between the variables we used rather than about their potential as predictors. While undeniably useful, a better designed analysis would rely primarily on time series analysis' techniques. In the case of our study, direct multi-step forecast strategy appears to be the most promising. Indeed, this method involves creating several separate models for different time horizons. This would enable us to pinpoint which variables are the most relevant predictors in the short, medium or long term.



Description of the direct multi-step forecast method

References:

Newspaper reports

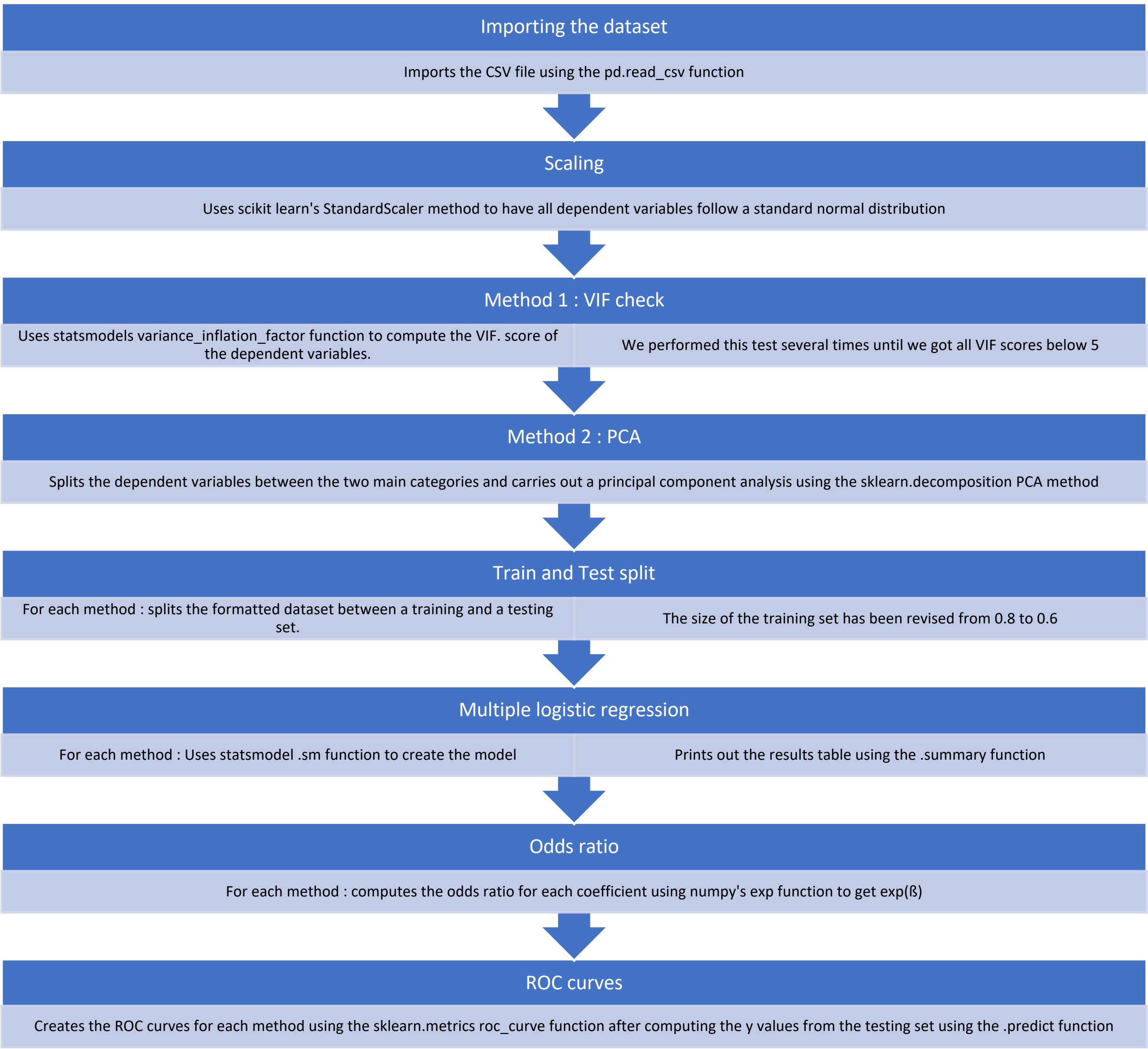
- https://fortune.com/2022/12/12/will-us-europe-recession-2023-economic-outlook-inflation-bcg-carlsson-szlezak-swartz/
- https://www.reuters.com/markets/us/new-fed-research-flags-rising-risk-us-recession-2022-12-30/

Scientific research papers

- Estrella, A., & Mishkin, Predicting US recessions: Financial variables as leading indicators. The Review of Economics and Statistics, 80(1), 45–61, 1998
- Mehdi Mostaghimi, Predicting US 2001 Recession, composite leading economic indicators, structural change and monetary policy, The Singa pore Economic Review, vol. 51, N°3, 343-363, 2006
- W.Liu & Emmanuel Moench, What predicts US recessions ?, International journal of forecasting, n°32, 1138-1150, 2016
- Estrella, A., & Mishkin, An overview of using the yield curve as a forecasting tool, Current Issues in Economics and Finance (2) 7, June 1996

Appendix : code description

The code we used works as follows :



Importing the dataset

Imports the CSV file using the pd.read_csv function

Scaling

Uses scikit learn's StandardScaler method to have all dependent variables follow a standard normal distribution

Method 1 : VIF check

Uses statsmodels variance_inflation_factor function to compute the VIF. score of the dependent variables.

We performed this test several times until we got all VIF scores below 5

Method 2 : PCA

Splits the dependent variables between the two main categories and carries out a principal component analysis using the sklearn.decomposition PCA method

Train and Test split

For each method : splits the formatted dataset between a training and a testing set.

The size of the training set has been revised from 0.8 to 0.6

Multiple logistic regression

For each method : Uses statsmodel .sm function to create the model

Prints out the results table using the .summary function

Odds ratio

For each method : computes the odds ratio for each coefficient using numpy's exp function to get exp(ß)

ROC curves

Creates the ROC curves for each method using the sklearn.metrics roc_curve function after computing the y values from the testing set using the .predict function