Documentation

Tanmay Shukla, Hong Ye, Miaomiao Fu, Mingliang Ge, Xiaoqing Xia

Abstract

Financial crises have significant implications for the lives of billions of people. Bad economic decisions during the financial recession can cause people's assets to shrink. Our goal is to predict the likelihood of another recession in the future using the two recessions we have experienced. Also, it is to help us make more accurate financial decisions when we know exactly when the recession will occur and how long it will last, such as investing in stocks at the lowest point and avoiding entering the stock market during the economic downturn. Spotting their warning signs early can facilitate the timely activation of countermeasures to prevent them. Many people have tried to come up with models for this challenging task.

An economic downturn affects people's lives in many ways: higher unemployment, reduced economic activity, reductions in income and wealth, exchange rate, the decline in GDP and more significant uncertainty about future jobs and income.

The challenges are:

1. The number of variables is vast, making it difficult to determine the significant ones.
2. The countries are different, making it difficult to find a “one fits all” solution. It is necessary to find commonalities among countries where recessions occur to find the main contributing factors and the proportion of each factor in determining the likelihood of a recession occurring.
3. Replicating the paper’s accuracy (AUC, a trade-off between valid positive rate and false positive) and determining the most important indicators.
4. Predicting a recession requires accurate timing and duration. We need to determine these key points with past data to obtain more precise information about the next recession.
5. Also, it can be difficult to translate complex early warning models into simple indicators that can help prevent financial crises.

Data Collection

1. Data scraping API Keys (R)

Web scraping is the technique for converting unstructured data to structured data. We grabbed the following four data lists from the FRED website via the API key, as they were too complex and heavy, so R Studio helped us to present them directly in software and clean them up further. Because Fred contains financial data, we get almost all the features from this website. From the EIA website, we get the electricity feature.

FRED
(1). 10-Year Treasury Constant Maturity Minus 3-Month Treasury Constant Maturity
(2). Job Openings (Monthly), Seasonally Adjusted
(3). Business Tendency Surveys for Manufacturing: Confidence Indicators: Composite Indicators: OECD Indicator for the United States (Federal Reserve Bank of St. Louis)
(4). NBER based Recession Indicators for the United States from the Period following the Peak through the Trough (USREC)

EIA (U.S. Energy Information Administration)
(1). Electric Power Monthly

2. Other Data download from the website (R & Excel)

We download the data from Nasdaq, Yahoo, Pitchbook and OECD..

NASDAQ
(1). NASDAQ Composite Index (COMP) (1970-2022)

YAHOO
(1). Dow Jones Industrial Average (^DJI) (1970-2022)

Pitchbook
(1). Crunchbase Inc. (Company Financial Information) (2016-2021)

OECD (Organisation for Economic Co-operation and Development)
(1). Monthly GDP for Project

Show 10 entries Search:

	date	GDP (YoY % Change)	GDP (Continuously Compounding)	gdp_change_lagged	rate_generated_change	rate_ge
1	1982	-1.04192567962547	-6.26282406213718	-6.23030788692134	0.161448433853319	
2	1982.08333333333	-2.19034230684082	-6.29568142470234	-6.26282406213718	-0.0394875371449598	
3	1982.16666666667	-1.79902560135352	1.82340022768841	-6.29568142470234	0.0872962075551481	
4	1982.25	-1.40576598888233	1.82063377379151	1.82340022768841	-0.0259380183278248	
5	1982.33333333333	-1.01054896347055	1.81787570168055	1.82063377379151	-0.0480948059057968	
6	1982.41666666667	-1.52976132555177	-1.53015617607863	1.81787570168055	-0.115798974945491	
7	1982.5	-2.04486138841827	-1.5321098157564	-1.53015617607863	0.0241554350162266	
8	1982.58333333333	-2.5558978151039	-1.53406845046788	-1.5321098157564	0.138307510344643	
9	1982.66666666667	-2.18768719998561	0.159759216354161	-1.53406845046788	0.0444818386835286	
10	1982.75	-1.81678246607878	0.159737950011873	0.159759216354161	-0.00245687816236058	

Showing 1 to 10 of 454 entries Previous 1 2 3 4 5 … 46 Next

Data Cleaning

1. Python (Preliminary Cleaning)

We put all the downloaded datasets and API key data into Python and found some missing values in the official statistics, two of which are important indicators for our recession study: 10-Year and 3-month Treasury Constant Maturities. These data have blanks and NAs, but we did not simply remove them from consideration due to their importance. Instead, we filled in the blanks and replaced NA by taking the average between the corresponding known data cells so that the entire 10-Year and 3-month Treasury Constant Maturities can become complete data chains for our use.

2. Excel (Replace Missing)

Our dataset contains 250 features and 150 countries in a time span from 1970 to 2021.

After feature selection, we have 27 key features. Among these key features, there are three features that contain some value missing. To make the model perfect, we use the financial model to predict the values for three financial data. Then we replace the missing value with the values from the financial models.

3. SQL (Get Extra Values)

We do most of our data processing in Python as well as Excel, so we don't use SQL very much. We collected data from a private company, Crunchbase Inc., for changes between 2016 and 2021

and imported it into SQL. Since the data we obtained about Crunchbase Inc. is confidential and paid for, the first-hand information we obtained is presented in the form of a PDF report. We manually organized the data into SQL and dumped it into a .csv file. Since then, all of our datasets have been converted to .csv files and are ready to be put into R Studio for the subsequent regression model analysis and time series.

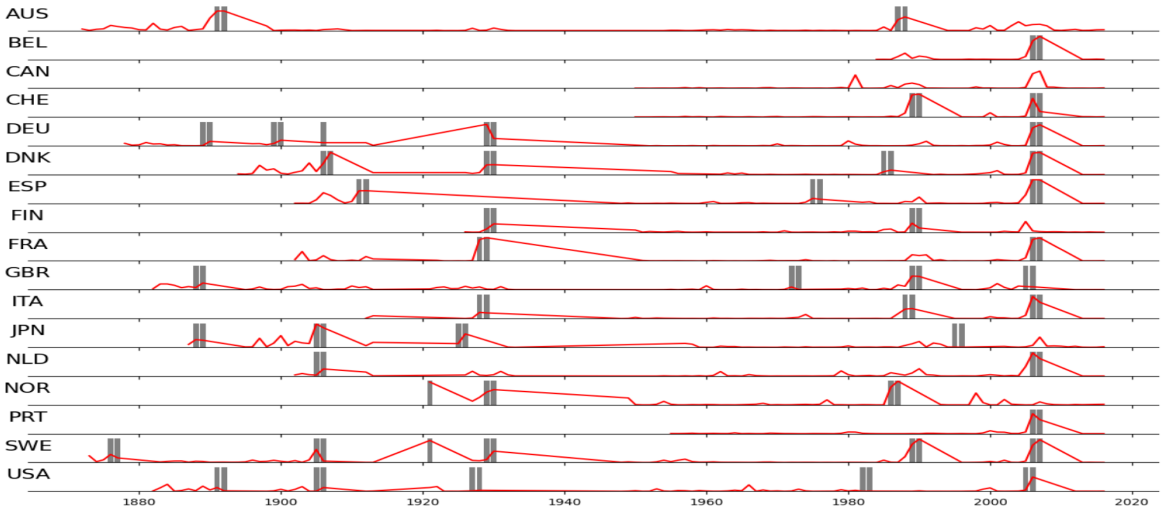
Method

1. Regression Model

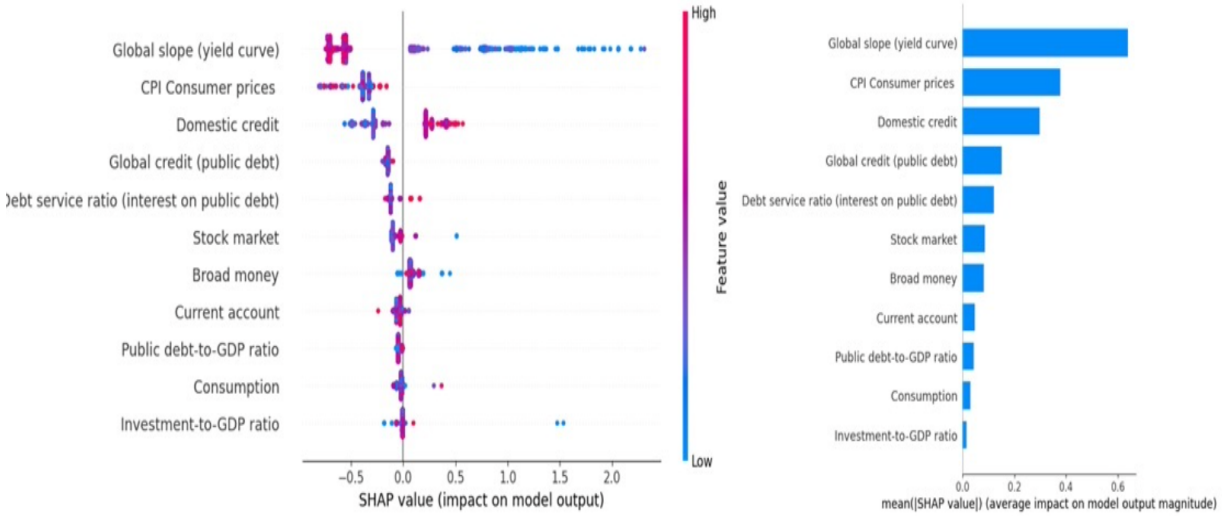
The main purpose of the project is to find the key features of recession in the United States. Before going through the different methods, the heatmap shows the different levels of importance of features for different countries. Different colors means different levels. From the heat map, there are many variables on rows and many countries on columns. From the heatmap, different countries exhibit different intensity for different variables that cause the financial crisis. For instance, in the column of China, the NFC debt takes 166.3% of GDP which is highest and represented as red. NFC debt is the most significant feature that can lead to recession in China.

Variable	*Argentina*	*Brazil*	*Chile*	*China*	*Colombia*	*Czech Republic*	*Hungary*	*India*	*Indonesia*	*Malaysia*	*Mexico*	*Philippines*	*Poland*	*Russia*	*South Africa*	*South Korea*	*Thailand*	*Turkey*
Economic indicators	**-8.1**	**-3.0**	**-3.7**	**6.5**	**-5.0**	**2.2**	**-1.1**	**-0.1**	**1.4**	**5.2**	**-3.3**	**-1.8**	**3.1**	**7.2**	**-7.1**	**7.9**	**3.5**	**-3.8**
Current account deficit (as % of GDP)	2.0	-1.0	-0.2	1.1	-3.3	0.2	-1.3	1.3	-1.4	3.5	1.5	0.6	3.1	2.3	-1.3	3.7	4.0	-3.4
Inflation (% y-o-y)	46.6	3.0	3.0	3.6	3.1	3.3	3.5	6.5	2.3	-0.5	3.3	2.2	3.4	3.1	3.4	0.5	-0.6	11.5
Economic growth (%)	-8.2	-3.7	-6.6	1.8	-5.3	-3.9	-3.0	-5.9	-0.3	-3.8	-7.5	-5.6	-1.0	-1.9	-6.1	0.0	-4.8	2.0
Budget balance (as % of GDP)	-10.2	-15.3	-9.9	-6.5	-8.2	-6.8	-6.9	-8.2	-6.4	-6.0	-3.6	-8.7	-8.0	-4.6	-16.8	-5.9	-6.0	-6.0
Competitiveness (value)	4.0	4.1	4.7	5.0	4.3	4.8	4.3	4.6	4.7	5.2	4.4	4.4	4.6	4.6	4.3	5.1	4.7	4.4
Political risk (scale 0-10)	2.5	3.0	2.5	1.9	2.4	1.5	1.4	2.1	1.7	1.6	3.1	2.1	1.9	1.5	2.4	1.5	2.4	2.4
Security risk (scale 0-10)	1.3	1.6	1.9	1.7	2.9	0.7	0.7	2.4	2.1	1.2	2.0	2.6	1.4	1.5	1.7	1.3	1.9	2.6
FX reserves import cover (months)	8.8	27.2	8.3	18.4	15.6	10.9	3.4	15.3	10.3	6.4	5.7	11.6	5.6	21.9	8.1	10.3	13.0	3.5
Debt vulnerability indicators	**-5.2**	**-0.6**	**-5.9**	**1.2**	**0.7**	**-0.8**	**-6.5**	**4.4**	**5.3**	**-0.4**	**2.1**	**3.1**	**1.2**	**2.7**	**-0.1**	**-0.9**	**3.6**	**-4.0**
Total external debt (as % of GDP)	74.9	49.0	85.2	15.1	56.6	85.3	145.2	26.2	38.7	71.0	45.5	25.0	61.2	33.8	61.3	31.1	36.8	67.2
Government debt (as % of GDP)	95.5	93.1	35.8	63.0	62.2	39.4	75.7	80.2	36.4	59.0	42.9	45.7	57.5	18.0	73.7	45.9	38.2	43.6
Household debt (as % of GDP)	5.6	33.8	50.0	59.8	31.2	33.4	20.4	14.1	17.2	77.1	16.9	17.2	35.0	22.5	36.7	100.6	76.2	18.0
NFC debt (as % of GDP)	15.7	48.4	121.6	166.3	38.9	59.3	72.3	50.6	23.9	77.9	30.4	33.1	44.8	90.6	43.8	110.2	53.8	77.5
NFC debt in foreign currency (as % of GDP)	9.5	20.4	40.8	6.5	12.3	28.7	34.4	7.5	9.4	13.8	23.6	NA	14.5	22.8	20.0	18.0	8.6	35.2
Gov. debt in foreign currency (as % of GDP)	68.3	6.0	10.1	0.8	21.0	3.5	12.9	2.3	10.2	2.1	8.3	NA	11.5	3.2	6.9	1.1	0.0	24.4
Financial market index	**-2.5**	**0.3**	**-2.8**	**6.5**	**-1.3**	**-1.6**	**-1.8**	**0.2**	**-2.7**	**0.1**	**1.5**	**-0.4**	**-0.5**	**-0.2**	**-1.4**	**4.7**	**0.7**	**1.2**
Volatility	2.7	20.6	13.8	4.4	16.9	12.5	12.7	6.3	11.3	5.0	19.3	4.4	12.1	18.7	16.8	7.8	5.4	13.2
Market Beta	0.0	-0.5	-0.5	-0.3	-0.6	-0.4	-0.3	-0.4	-0.5	-0.3	-0.6	-0.5	-0.3	-0.6	-0.4	-0.6	-0.2	-0.3
Liquidity	0.1	1.1	0.3	4.3	0.2	0.4	0.4	1.1	0.2	0.4	2.1	0.2	0.6	1.1	1.1	1.7	0.5	1.7
Hot money indicator	-0.6	-1.8	1.7	0.6	0.0	0.5	-0.4	1.1	2.3	-0.7	0.2	0.6	-1.1	0.1	1.1	-1.7	-2.1	-0.9

Firstly, we use the regression model to predict the recession and figure out the different significant features for 150 countries that can contribute to the financial crisis. The figure below is the graph output of our prediction model. Each row represents one country. The red line represents the predicted financial crisis and the gray bar represents the actual crisis.



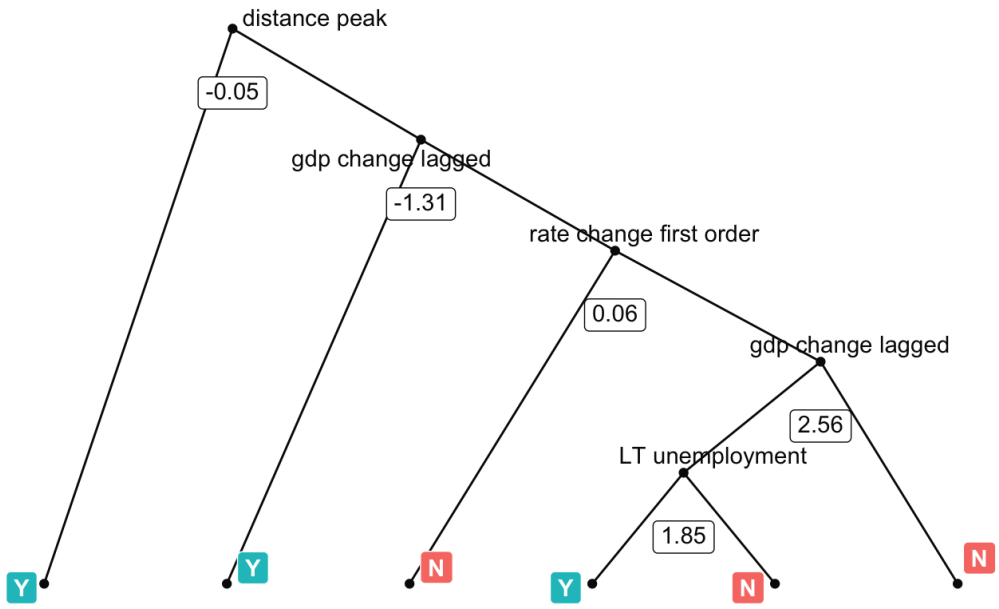
Secondly, we use the feature selection method to obtain the key features that can cause financial crises which only work in the United States. From all the features, we have 27 key features for our model. To persuade the significant features of the model by using feature selection, we use the SHAP value, which is used to assign each feature an importance value for a particular prediction. SHAP values calculate the magnitude and negative and positive direction of a feature's effect on a prediction model by temporarily removing it from the model. We calculate two graphs: The left figure is the calculated SHAP value; the right figure is the SHAP Feature Significance.



2. Machine Learning (Decision Tree)

Then we list the factors that are important to the US based on these features: predict financial crisis and recession, logistic regression, random forest decision trees and XGboost. We will know which feature affects the most.

After we found all the key factors that determine recession, predict financial crisis and recession, logistic regression, random forest decision trees, and XGboost gave us the weight of each factor in our determination of recession. Distance peak is the first key factor we consider; if the distance peak is less than -0.05 in a short period of time, then based on the model we obtain, we can determine that the tube time is in recession. On the contrary, if it is not, then it will move to another key factor: the size of GDP Change lagged to determine whether it is in recession or not, and so on until the last factor in the chart: LT Unemployment. If all decisions are adverse, then it will be defined as “There is no potential recession”. However, if we get an affirmative answer to any of the previous decisions, the conclusion of “out of potential recession” will be obtained directly. A clear diagram is presented below;



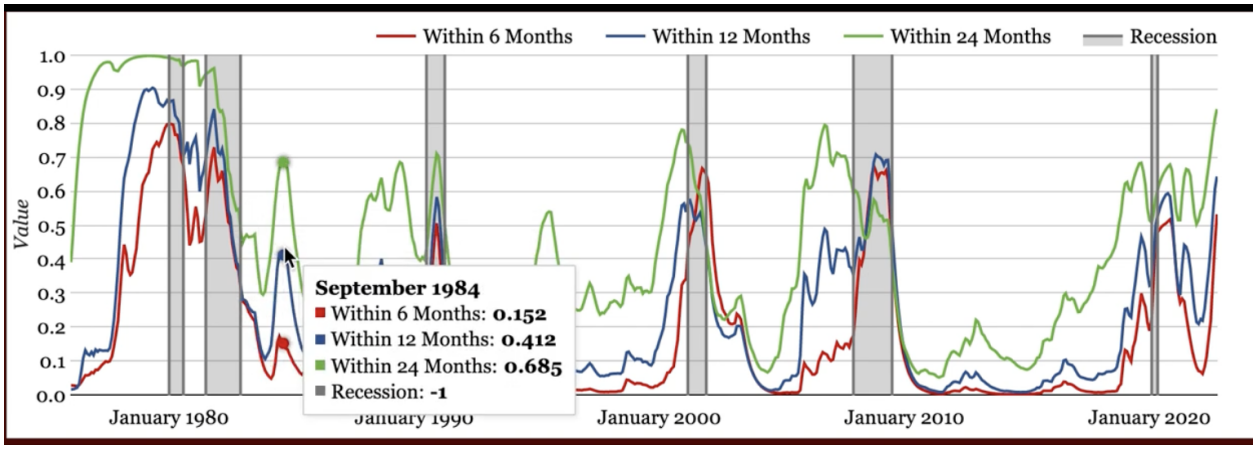
3. Time Series Analysis

After machine learning modeling, we already get the probability of recession. However, it’s never enough to get this, the important factor is time is still missing, so we still don’t know when the recession will come. Here comes the time series. After we get those features, we use LSTM, exponential smoothed recurrent neural networks (α-RNNs) method for forecasting. We find the dataset from different time duration. Next step is to go back and clean the dataset so that it makes the dataset monthly. After that, we add one dataset from time to the LSTM method to predict the

recession time.

The figure below is a recession prediction under our time series model. Firstly, we can predict every recession only with the probability, it also has the exact time from this graph. There are six recessions that have happened before, and we marked it with shadow in the graph.

Secondly, we divided the forcast timetable into three different timelines. We can see that we divided the prediction timetable into within 6 months, within 12 months, and within 24 months. From the graph, we can see that the red line is within 6 months, blue line is within 12 months, the green line is within 24 months. But if you divide the data into different time intervals,you will get a different value of probability. Take the first recession in the graph for example,we can see the probability of recession within 6 months is the lowest which means that the probability of recession from that point is lowest and the probability of recession within 24 months is high.



Conclusion

With the data we have now and the models we have built, we have very good evidence to conclude that we are in a recession. However, there are very many limitations to our study. Firstly, there are so many datasets, so we need to drop many datasets, do better cleaning and design more precise models. We need to use more computational power to train the dataset. Also, the more tests will help us in the selection of features. Secondly, because the change in the financial market is so fast, if we want to make the prediction more precisely we need to update the data frequently.