Data

In this section, I will give an overview of the data used in this research. First, I will introduce the input variables, a.k.a., features, and my reasons for selecting them. The features include macroeconomic data, market return trends, and market fundamental indicators. Then, I will explain my definitions of equity bubbles and financial crashes. The range of the data is from January 1960 to December 2020 in the US. Unless otherwise specified, the frequency of the variable is monthly. The raw data are retrieved from Federal Reserve Economic Data (FRED) and \emph{multpl.com}.

Input variables

GDP growth rate. The real Gross Domestic Product growth rate measures the productivity growth of a nation. It is one of the most important macroeconomic metrics. Given its importance, sophisticated machine learning models might gain insights from the growth pattern for bubble detection. Notice that the frequency of GDP is quarterly rather than monthly. To make the data frequency consistent, I replicate the quarterly GDP growth rate to all the corresponding months.

Inflation. The growth rate of the Consumer Price Index measures the price inflation in a nation. Similar to the GDP growth rate, it is one of the most important macroeconomic indicators. In the context of financial markets, previous research finds that there is an inverse relationship between inflation and lower share prices. \parencite{inflation}

T-bill yield. The 10-year Treasury constant maturity rate is a key benchmark in the debt market. It captures the long-term expected yield of a risk-free investment. From the perspective of an investor, the debt investment and the equity investment are substitutions. It has been found that bond returns tend to be higher than stock returns when the market uncertainty increases. \prencite{bond}

S\&P 500 returns. The Standard \& Poor is one of the most widely used stock market indexes. It measures the stock returns of the largest 500 companies listed in the US. In this research, the output variables, equity bubbles and financial crashes are measured by S\&P 500 return patterns. As a general practice in time-series forecasting, it is natural to use historical return trends to forecast future data. To measure the price returns in different periods, I include 1-, 3-, 6-, 12-, 60-month returns of S\&P 500.

Consumer confidence. As noted by \textcite{psychology}, the psychological factors of investors have a significant influence on investment decisions. I use the data of consumer opinion surveys as a proxy for the public sentiment of the general business environment. While the capital market sentiment does not always align with consumer confidence, this is the most easily accessible and complete data.

Shiller P/E ratio. Invented by Robert J. Shiller, \parencite{shiller2015IE} the Shiller P/E is also known as the cyclically adjusted price-to-earnings ratio. It measures stock prices as a ratio to the average inflation-adjusted earnings. \parencite{shillerPE } If the Shiller P/E ratio is too high, it suggests that the stock prices could be overvalued based on the corporate earnings.

Market capitalization-to-GDP ratio. Warren Buffet believes that “the market value of all publicly traded securities as a percentage of the country's business” is a useful metric for investment decisions. \parencite{buffet} Given the investment success of Warren Buffet, the effectiveness of this metric seems worthwhile to explore. Since collecting data for all publicly traded securities is cumbersome, I calculate the ratio between the S&P 500 index and the nominal GDP as a proxy. Similar to the real GDP growth rate, the frequency of the nominal GDP is quarterly. I replicate the GDP data to match the monthly frequency.

Table~\ref{tab:features} presents a summary of input variables.

Output variable

I follow the definition of \textcite{Chat2018} and define a market crash as an extreme instance of negative market return (one percentile). Given this definition, I define the bubble as a pre-stage of market crashes. The operational definition of a bubble is a binary variable that takes the value of 1 if and only if: 1) there will be a market crash in the next six months; 2) the current period does not see a market crash. The biggest benefit of this definition is that it is entirely based on observable data. It does not involve the theoretical definition of the fundamental value of the stock. As discussed in the background section, definitions of fundamental values are subjected to the assumptions of economic models, and it would limit the power of bubble detection models. Of course, this definition is not suited for a theoretical debate of bubbles; however, it offers practical value for equity market practitioners.

The one percentile and six-month period may seem arbitrary. In the discussion section, I will alter the percentile and period length for a robustness check of the model.

Table~\ref{tab:imbalance} shows the distribution of output data. Figure ~\ref{fig: bubble\_def} shows the bubble categories through time. We can see that the bubble definition captures the pre-stages of most notable market crashes, including the dot-com crash around 2001 and the US housing market crash in 2008. This suggests that our bubble definition is in line with reality.

Finally, as a typical practice for machine learning problems, I split the full data into the training data and the testing data. The training data is 80\% of the full data, while the testing data is the remaining 20\%.