**Abstract**

In this research, I aim to detect equity bubbles and predict financial crashes in the S\&P 500 index. I implement machine learning models with special consideration to the imbalanced data problem. Features of the models are selected based on economic theories and market fundamentals. I find that the Random Forests model with CV-tuned decision threshold performs the best, with a 93.0\% balance accuracy rate. The Recurrent Neural Networks with Bidirectional Long-term Short Memory and focal loss function gives unsatisfactory results, possibly due to insufficient data. Features with the highest predictive power are market fundamental indicators, long-term market returns, and the psychology factor. The model is limited to predicting long-term (more than 3 months) and extreme market downturns.

**Introduction**

Bubbles and market crashes are important themes of financial markets. Asset bubbles describe the situation where asset prices significantly deviate from their fundamental values. Notable historical bubble includes the Dutch tulip mania in 1637, the dot-com bubble in the 1990s, and the US housing bubble in 2000s. Investors who are unaware of the potential risks of bubbles realize huge losses when bubbles burst and the market crashes. Some people believe the market crash in 1982 wiped out the cumulative profits of the history of American banking history \parencite{nnt2009}.

More recently, despite the general economic downturn caused by the COVID-19 pandemic, stock markets are performing exceedingly well. The S&P 500 index was up by 18.23\%, while the tech-heavy NASDAQ gained 43.6\% in 2020. With the historical lesson of bubbles, we may wonder if we are experiencing one right now.

Given the catastrophic consequences caused by bubbles, it would be meaningful if we can identify bubbles and predict potential crashes. An accurate bubble detection tool may serve as an Early Warning System for policymakers or a risk management tool for investors. This is the primary motivation of this study. While bubbles and market crashes are broad phenomena, in this research, I focus on bubble detection in the stock markets. Specifically, the scope of the research is limited to the prediction of the S\&P 500 market index. The main research goal is to come up with a method that can reliably detect equity bubbles and forecast market crashes in the S\&P 500 index.

If we can realize the main research goal, we can answer certain further questions. If the prediction method is transparent enough, we can identify important factors that are the most important for prediction. On the technical side, we can gain insights into the type of empirical methods that work the best for equity bubble detection problems. Also, we can predict if we are in an equity bubble right now using the prediction method.

There is much previous research trying to achieve this goal. Some researchers, based on the economic theory such as rational agents and fundamental value as discounted dividends, came up with statistical tests like variance bound test for bubble detection (Shiller, variance test). More recently, researchers have applied machine learning methods, which are powerful prediction tools for detecting complex patterns, to the study of predicting market crashes \parencite{Chat2018}. However, as we will see in the literature review section, previous studies are limited in two ways. First, the theory-based studies are highly reliant on their theoretical assumptions and suffer from problems such as joint hypothesis testing. Second, on the other hand, model-free machine learning studies typically only include short-term trading variables and have unsatisfactory prediction power.

I attempt to solve these limitations in this paper. To avoid the theory reliance problem, I adopted the definition of bubble purely from observable data. As readers may have noticed, I have used the term “bubble” interchangeably with potential market crashes. That is precisely my definition of bubbles. While it may not be suitable for theoretical debates, e.g., “do bubbles really exist”, it captures most of the practical value of bubble-related discussions for practitioners like policymakers and investors. With this definition, the bubble detection problem is transformed into a supervised machine learning classification problem, where we try to predict the output, i.e., bubble, given the input variables. One of the distinguishing features of this research is that I include more input variables than many previous studies to boost the prediction power of models. I select variables that are suggested by economic theories as input variables for bubble detection, including macroeconomic data, long-term trading trends, and fundamental indicators. Notice that the validity of my research would not be dependent on the correctness of theories, because they are merely serving as feature-selection tools.

I implemented three machine learning models as candidate classifiers: logistic regression (logit), Random Forests (RF), and Recurrent Neural Networks with Bi-directional Long Term Short Memory layers (RNN-BiLSTM). One unique challenge of this classification problem is that the dataset is imbalanced ding imbalanced, meaning that there are far more non-bubble instances than bubbles. I have adopted special techniques to mitigate this problem, for example, re-sampling for re-weighting, tuning decision thresholds, and using asymmetric loss functions.

The result is satisfactory. My best performing model, RF with CV-tuned decision thresholds (RF-CV), achieved a balanced accuracy of 93.0\%. To the best of my knowledge, it is the highest-performing machine learning model for predicting market crashes. I believe this can be used as a meaningful tool for financial market practitioners.

The rest of the paper is organized as follows. The background section provides an overview of previous research and my motivation in this research. The data section lists all the input variables, their theoretical justification of relevance, and my construction of the output variable “bubble”. The model section introduces the modelling process including the consideration for the imbalanced data set. The result section presents the results and their implications. The discussion section diagnoses models, provides a robustness analysis, and discusses the limitation of the research. Finally, the conclusion section summarizes the key findings from this research.

**Conclusion**

I start the research aiming to detect equity bubbles and predict financial crashes. I find that previous empirical studies are limited by their theoretical assumptions or feature selection. I choose the model-free machine learning approach. The features are selected based on theories but do not rely on theories. Based on the nature of the problem and previous research, I use logistic regression, Random Forests, and Recurrent Neural Networks as models. To mitigate the imbalanced data problem, I change the decision thresholds, re-sample the data for weighting, and adopt an asymmetric loss function.

The key findings of the research are as follows. The Random Forest with CV-tuned performs the best, with a balanced accuracy rate of 93.0\%. Perhaps due to insufficient data, the RNN model does not give a satisfactory result. Other models perform decently well. It is found that long-term market returns and market fundamental indicators are the most important factors, whereas short-term market returns and macroeconomic data have weak predictive power.

The research has two main contributions. First, I used a new methodology that takes advantage of both the predictive power machine learning algorithms and the feature selection insights from economics theories. Second, I build a high-performing predictor (RF-CV) for equity bubble detection and market crash prediction.

The research is limited that the model cannot be generalized to short-term and mild market crash prediction. From a practical perspective, the model has a low precision rate so that it is prone to send false alarms. Besides, it is suspected that a positive correlation between the training data and the testing data makes the model overstate its testing performance. Future research is needed to overcome the correlation problem.