A Machine Learning Approach to Equity Bubble Detection and Financial Crash Prediction

Hongkai Yu
Advisor: Dr. Jonathan Graves
Vancouver School of Economics,
University of British Columbia

Introduction: bubbles and market crashes cause losses, and predicting them can be valuable

Tulip Mania, 1637 → Dot-com bubble, 1990s → US housing, 2010s → Bubble? 2020 –

"In the summer of 1982, large American banks lost close to all their past earnings (cumulatively), about everything they ever made in the history of American banking—everything."

—Nassim N. Taleb

Research purpose: predicting bubbles and financial crashes through machine learning

- Policymakers: Early Warning Systems
- Investors: Risk Management tools

Literature: previous empirical studies are limited by their theoretical assumptions or feature selection

<u>Too many theoretical assumptions</u>
Rational, risk neutral agents
Complete information
Dividend as the fundamental value
Stationary time series
Brownian motion for asset prices
…

<u>Too few features included in ML studies</u>

Only the price index (Moser, 2019)
Price index, bond yield, exchange rate (Chatzis et al., 2018)

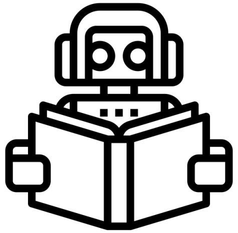
<u>Critiques</u>
Joint Hypothesis Problem
No bubbles (Fama & French, 2014)
Alternative explanations (Gürkaynak, 2008)
Investors not rational (Daniel, 2005)
Stationary tests inconclusive (Evans, 1991)
Normal distributions unreliable (Taleb, 2009)

<u>Research methodology:</u>
model-free machine learning, features are selected based on theories but not rely on theories

Data: indicators of trading, macroeconomics, and market fundamentals; the definition of bubbles

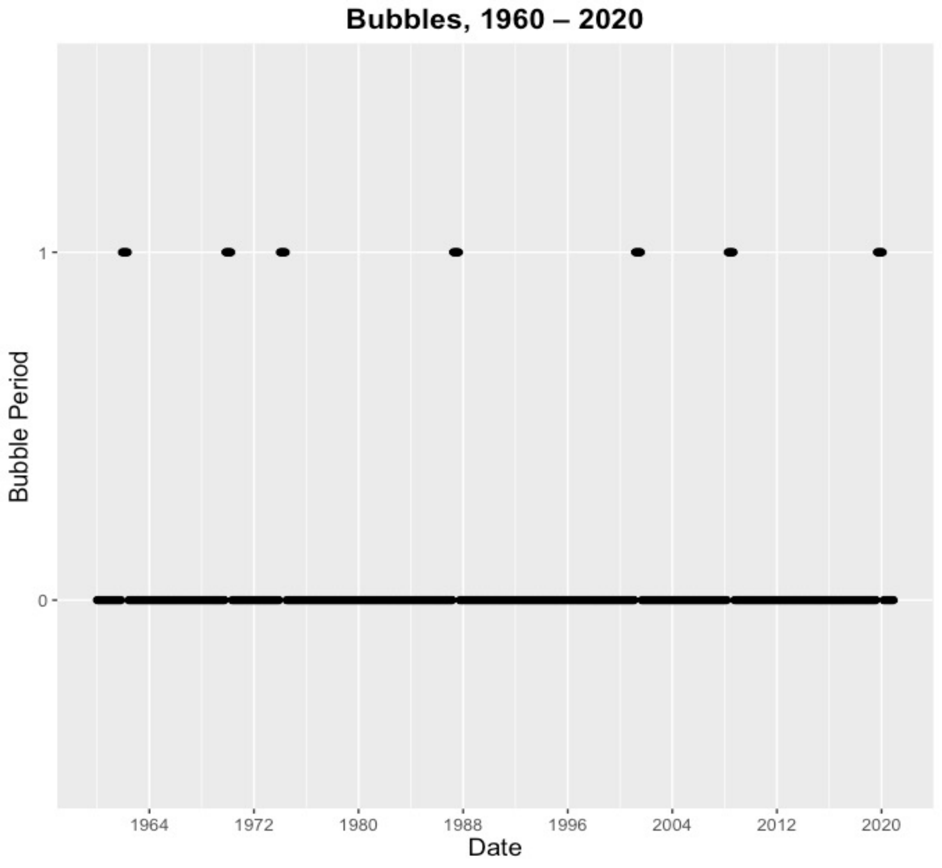
X

Shiller P/E ratio
Market-cap-to-GDP
Consumer confidence
S&P 500 return (1, 3, 6, 12, 60 months)
T-bill yield
Inflation
GDP growth rate



machine learning

Y (Bubble): a market crash (1% quantile return) will happen in the next 6 months



Models: three models are adopted based on the nature of the problem and previous studies

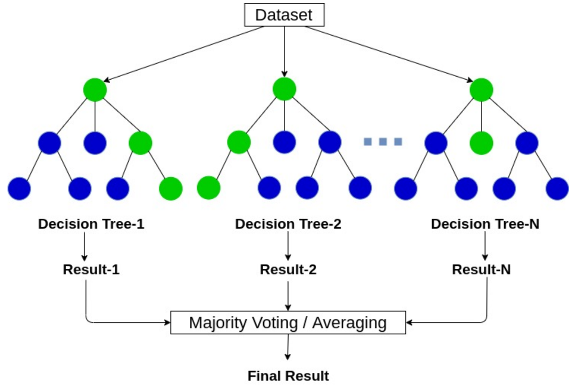
Logistic Regression

Predicting log-odds by linear models
Baseline predictor for binary classification

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

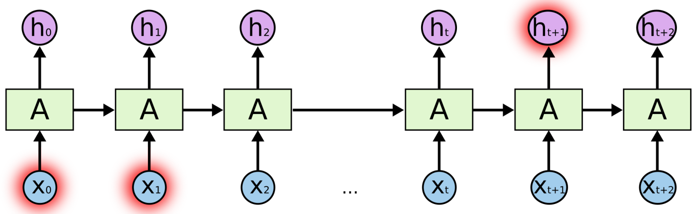
Random Forests

Detecting non-linear patterns with decision trees
Ensemble methods are promising (Lin et al., 2012)



RNN + BiLSTM

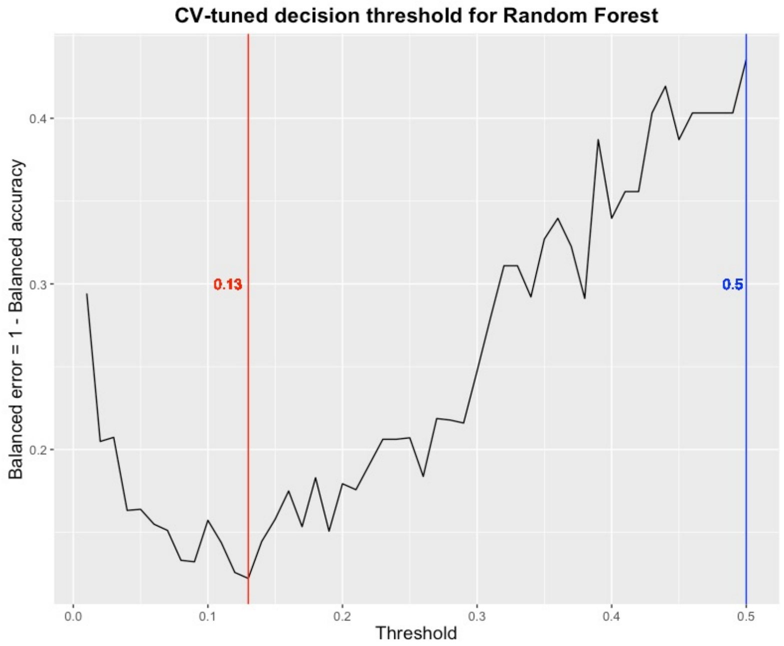
Good with financial time series (Namini et al., 2019)
Similar studies (Bashchenko & Marchal, 2020)



Models: the imbalanced nature of the dataset requires special consideration for classification

Problem: Even a naïve classifier $f(x) = 0$ can achieve a (useless) high overall accuracy

Category	Non-bubble	bubble
Count	690	42



CV-tuned decision thresholds

Re-sampling for balancing weights

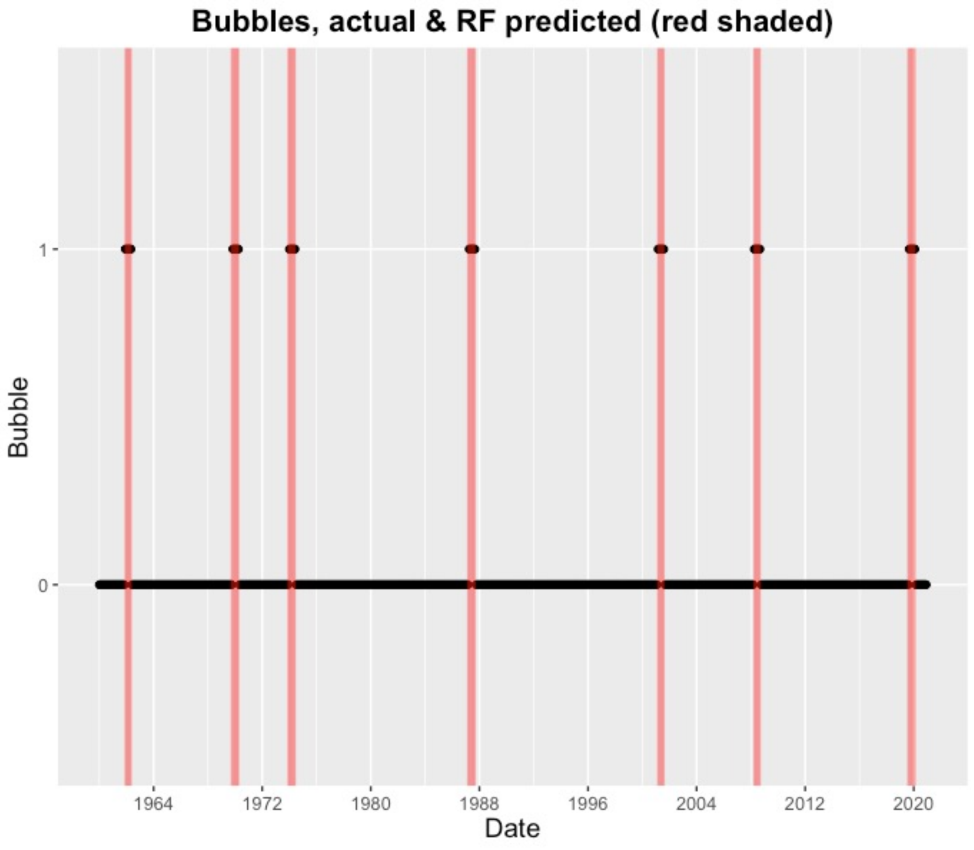
Asymmetric loss (Lin et al., 2018)

Model Evaluation:

$$Balanced\ Accuracy = (Senstivity + Specificity)/2$$

Result: the Random Forest model with a CV-tuned threshold has the highest balanced accuracy

Performance on test data	Sensitivity	Specificity	Balanced Accuracy
Logistic regression, threshold as prevalence	72.72%	72.06%	72.39%
Logistic regression, re-weighted data	72.72%	73.52%	73.12%
Random Forest, CV-tuned threshold	100%	94.85%	97.43%
Random Forest, re-weighted data	45.46%	100%	72.72%
RNN + BiLSTM, focal Loss	0%	100%	50%

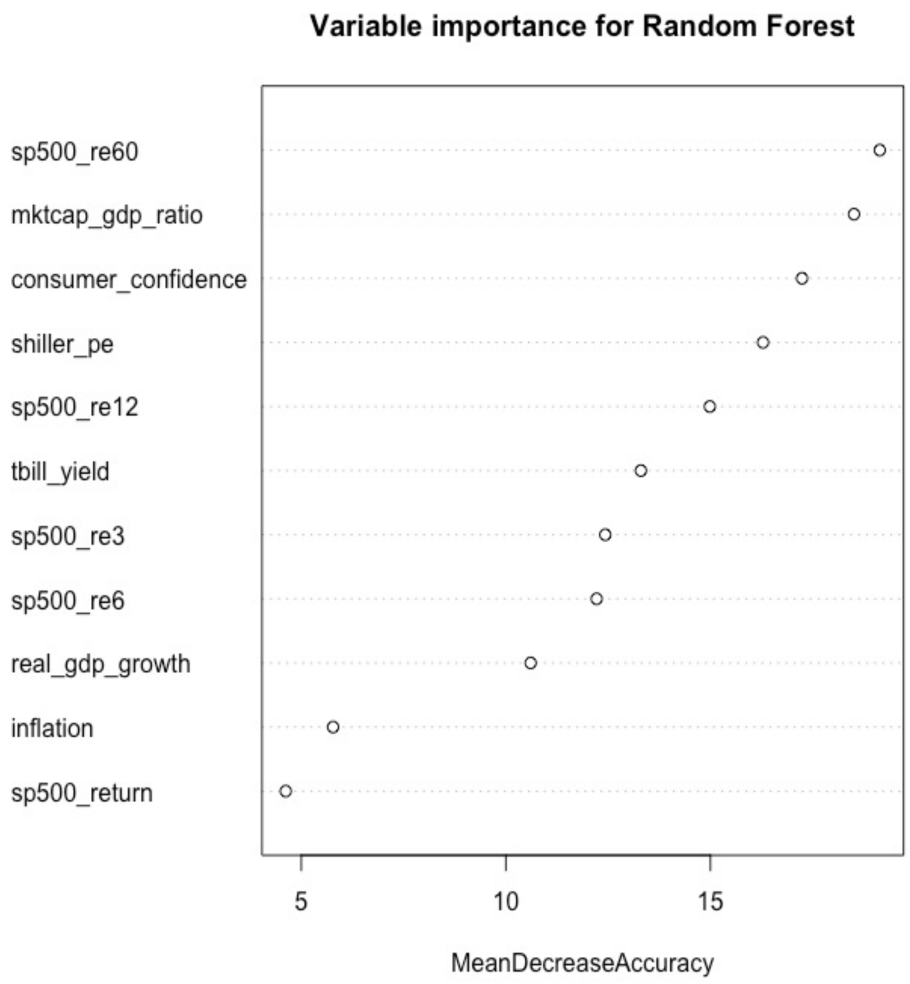


Insights: the most important factors for prediction are fundamental indictors and long-term trends

Random Forest,
CV-tuned threshold (left)

Logistic regression,
re-weighted dataset (right)

Short-term index returns have low predictive power



	Dependent variable:
	bubble
real_gdp_growth	-0.007
	(0.035)
inflation	0.099^{***}
	(0.028)
tbill_yield	0.237^{***}
	(0.073)
shiller_pe	-0.658^{***}
	(0.058)
consumer_confidence	-1.182^{***}
	(0.147)
mktcap_gdp_ratio	216.592^{***}
	(18.361)
sp500_return	-0.105^{***}
	(0.033)
sp500_re3	0.018
	(0.022)
sp500_re6	-0.021
	(0.020)
sp500_re12	-0.143^{***}
	(0.016)
sp500_re60	0.075^{***}
	(0.006)
Constant	106.314^{***}
	(13.641)
Observations	1,112
Log Likelihood	-410.346
Akaike Inf. Crit.	844.693
Note:	*p<0.1; **p<0.05; ***p<0.01

Conclusion: a high-performing predictor is found, and it predicts that we are likely in a bubble now

Select features suggested by theories

Implement 3 machine learning models

Mitigate the problem of the imbalanced data set

Get a high-performing predictor

It predicts that we are likely in a bubble now (April 2021) and market crashes are likely in the next six months
*not financial advice

Research contributions: New methodology and strong performance

100% sensitivity, 94.85% specificity vs. 59%, 90% (Chatzis et al., 2018)