Stock Bubble Classifier

Colin Salama

What is a Bubble and Why does it matter?

GME

Say you have a large portion of your net worth invested into a single stock. If this stock suddenly spiked from $50 to $350, your net worth would also spike about 7x.

What goes up must come down

This is not always true in a financial market, but when it is true, your 7x net worth is facing a lot of risk.

Bubble

A bubble occurs when you see the stock price grow significantly higher than the fundamental value. Bubbles often end with a significant crash.

GameStop

Agenda

- Goals
- Data
- Classification Method
- Conclusion / Moving Forward

Goal

Build a classifier to accurately and instantly predict when a stock is in a bubble

Traditional economic models rely on months before and after the bubble in order to determine that it was in a bubble. This is too slow for investment decisions.

Learn about the most important features in determining a bubble

These learnings could be used in other analyses.

Data: Features

- **Stock price data** taken from the SimFin API for all stocks prior to January 2020.
- **Quarterly Income Statements, Balance Sheets, and Cash Flow Statements** legally must be released by all public companies, and SimFin API also provides access to this.
- Features:
  - Revenue
  - Net Income
  - Net Change in Cash Flow
  - Total Assets
  - Total Liabilities
  - Monthly Returns
  - Standard Deviation of Returns (Volatility)
- Missing data was **imputed with the mean value** for the column.

Data: Target

Labeling a bubble is a complex problem in economics without a consensus.

After originally attempting a more rigorous method that was very computationally expensive, I ended up opting for a more simple method. When monthly returns are less than -2 multiplied by the standard deviation of returns, a bubble is labeled.

We expect to have an imbalanced dataset here because bubbles are less common than normal price movement.

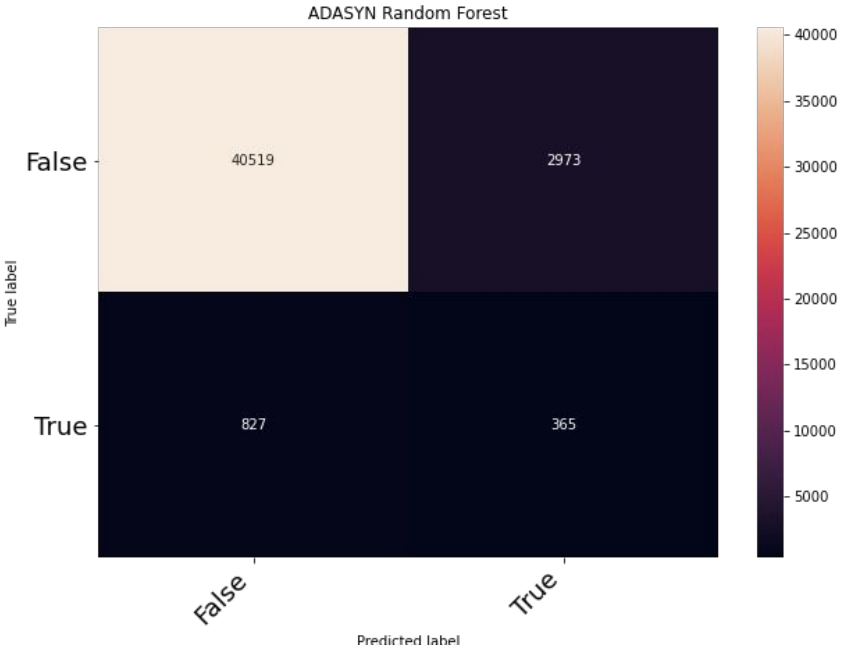
5,906 stock/time-period combinations were labeled as a crash, and 217,511 were not. This is **2.6%** of the dataset. ADASYN Oversampling is used in future models to account for imbalances.

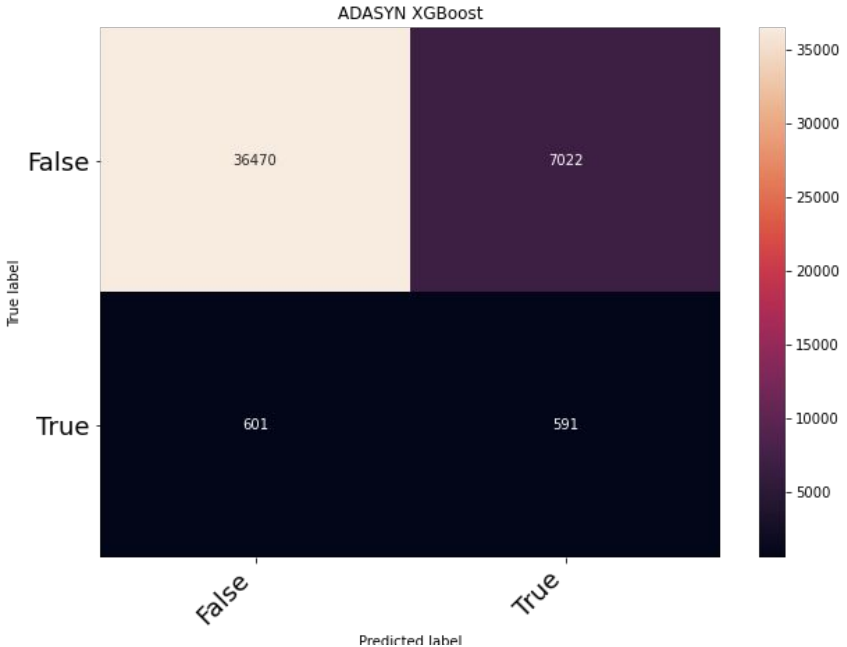
Classification Model

The dataset is initially split into training, validation, and testing datasets. A variety of models were tested on the validation dataset in order to find the best model.

	Logistic Regression	Naive Bayes	LinearSVC	KNN (n = 5)	Random Forest	XGBoost
ROC AUC	56.2%	50.8%	63.4%	62.5%	**61.6%**	**66.7%**
Accuracy	77.4%	9.1%	40.7%	54.4%	**91.7%**	**82.9%**
F1 Score	7.4%	5.3%	7.3%	7.7%	**16.0%**	**13.4%**

Confusion Matrices





Random Forest vs. XGBoost

Both models performed well on the validation dataset, so we will also look at the F-Beta with Beta = 0.5 to put greater weight on precisely labeling a bubble.

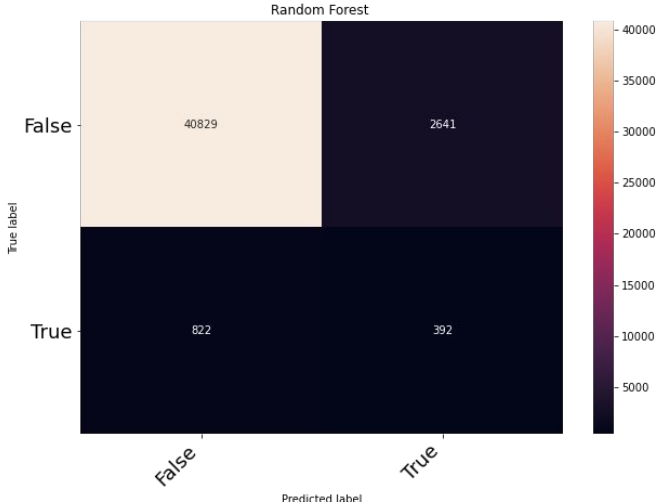
Because Random Forest performed better, we will move forward with this model.

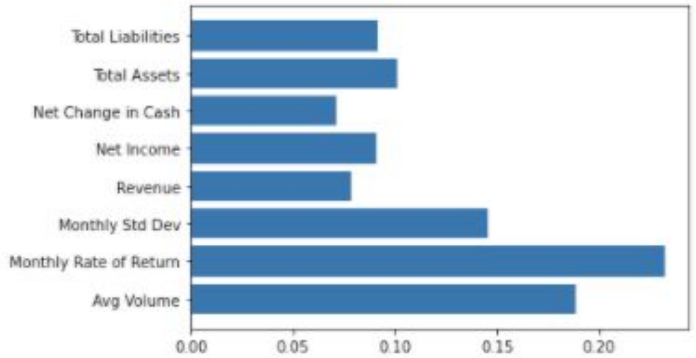
	Random Forest	XGBoost
F-Beta	12.5%	9.3%

Random Forest

On the testing dataset, Random Forest had an F-Beta score of 14.7%.

Looking at Random Forest feature importance, Monthly Returns and Avg Volume were the most important features.





Conclusion / Moving Forward

Although this model was valuable in determining whether there would be a crash, the relatively low F-Beta scores would currently prevent investment decisions based on this model. We need more data to improve this.

Stocks will often be swayed based on news (such as COVID or a presidential election). If there was some way to incorporate news factors, the model would be improved.

Questions?