



Predicting Bitcoin's Big Days using Machine Learning

Mark Scott
Ran Dong



Goals of this Analysis

- Make insights into the role of bitcoin (and other virtual currencies) in the global market
- Apply machine learning techniques to predict the price movement of bitcoin the following day
- Make money!
- Create an informative up-to-date app that assists the speculator in making decisions

Pipeline and Modeling Techniques

- R (VIM) and MySQL
- Time Series Modeling (quantmod, tseries, forecast)
- Support Vector Machines
- Random Forest
- Shiny App (dygraphs/Streaming twitter wordcloud)

Outline

- Introduction (where are we now?)
- Classification Problem
- EDA + Feature Extraction + Time Series
- Classification Problem
- Random Forest and SVM
- Arbitrage Opportunities
- Future Directions (Twitter Sentiment Analysis)

Since the Web Scraping Project

Bitcoin Price History

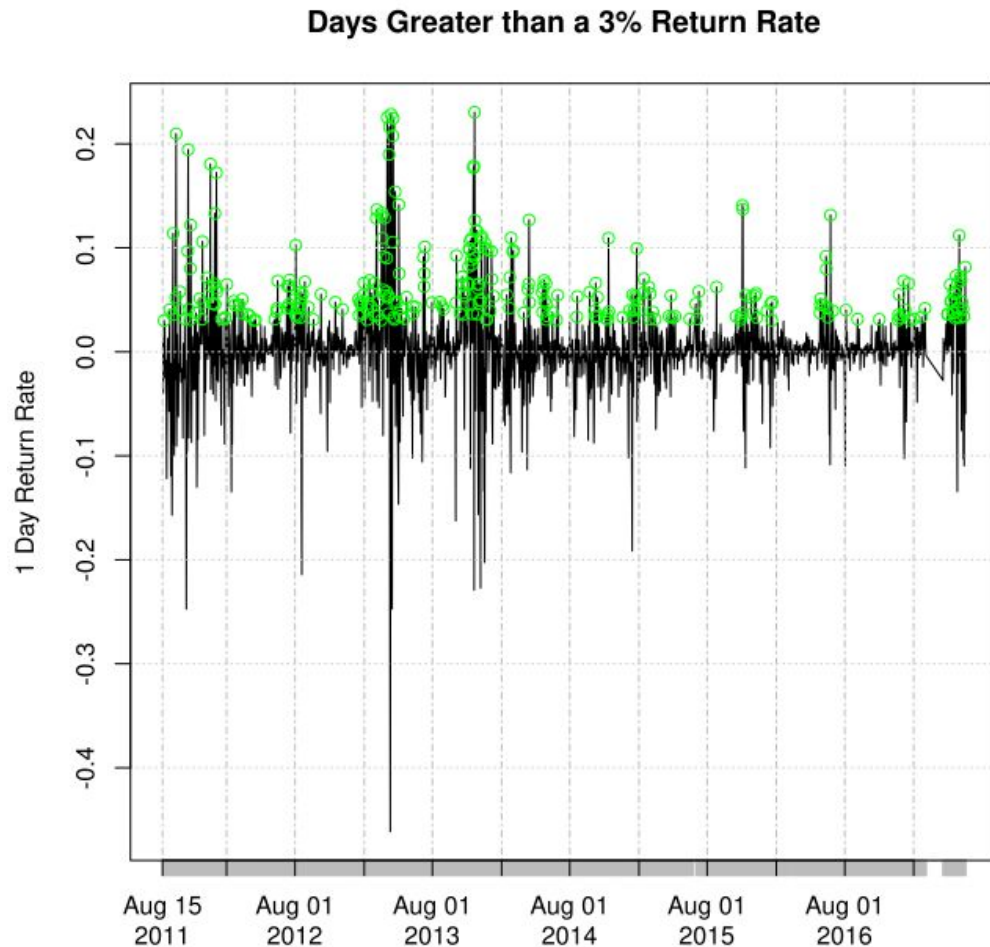


Bitcoin Price Since 1/1/17



The Problem

- How to forecast the price?
 - This is very difficult
- How about forecasting the return rate?
 - More feasible because it is a stationary time series using the Augmented Dickey-Fuller Test
- **Predict if the bitcoin return rate will be greater than a certain threshold (3%) tomorrow**

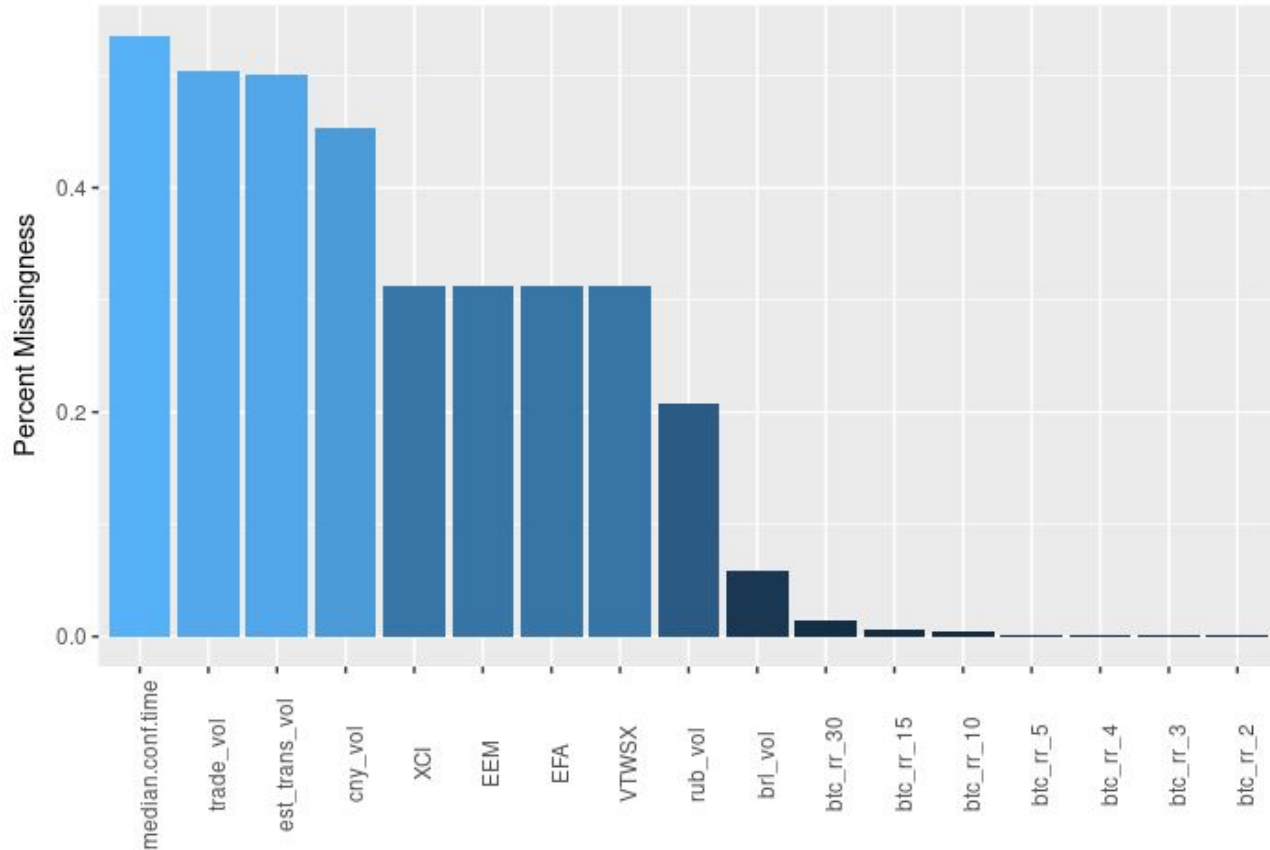


Features

1 Day Return Rate	2 Day Return Rate	3 Day Return Rate	4 Day Return Rate
5 Day Return Rate	10 Day Return Rate	15 Day Return Rate	30 Day Return Rate
Maximum 1 Day Return Rate	Intra Day Fluctuation	Daily US Transactions	US Volume
Russian Volume	Chinese Volume	Brazilian Volume	XCI (Arca Computer Tech Index)
EEM (iShares Emerging Markets)	EFA (Developed Market Equities Large/Mid Cap)	VTWSX (Vanguard Total World Stock Index)	

Missingness

- Missingness highly dependent on time
- Bitcoin never stops trading
- Markets close on weekends
- Start from early 2012
- Impute the rest with KNN and na.interp from forecast package



Removing Trend from Time Series Features w/ auto.arima

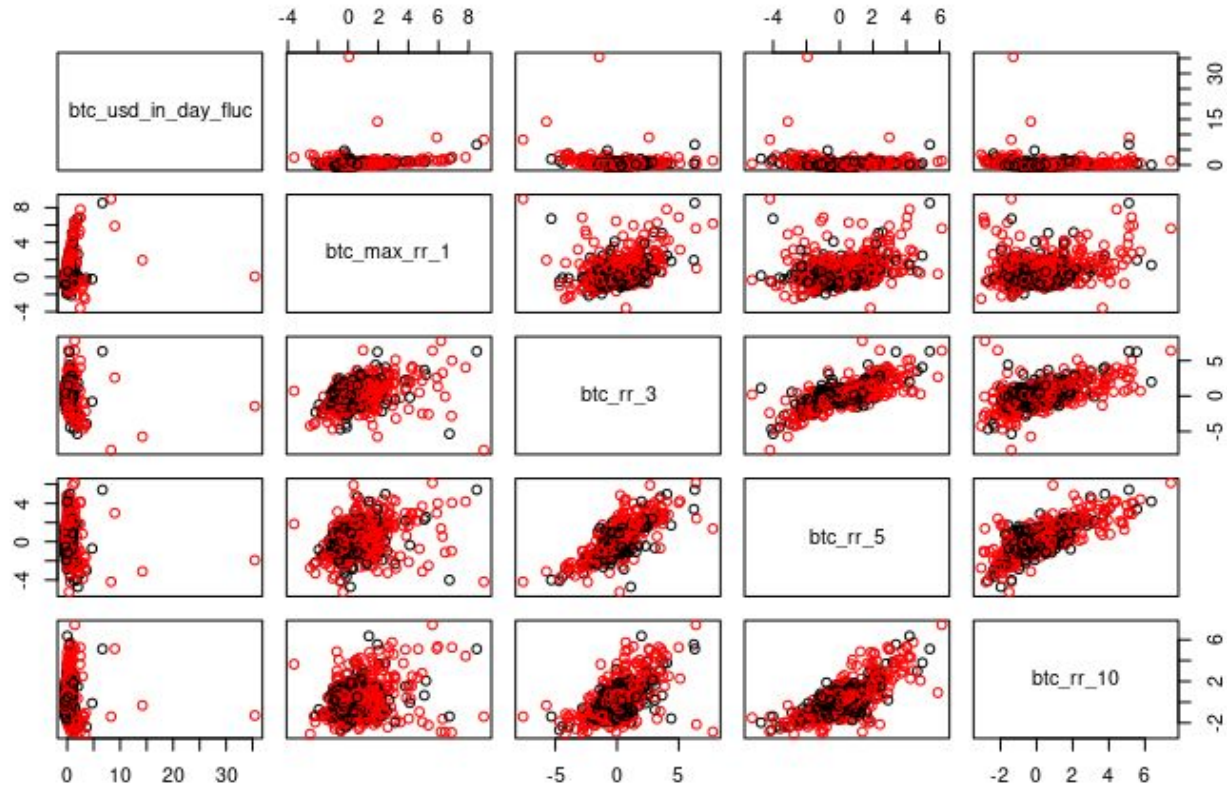
NYSE ARCA Computer Tech Index (XCI)



XCI residuals



SVM motivation

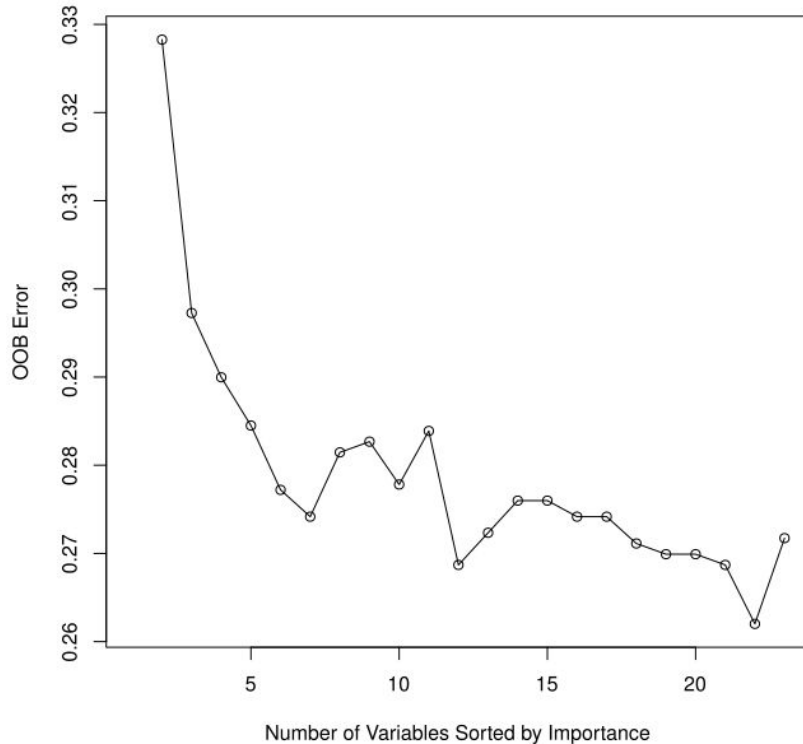


Machine Learning Recipe

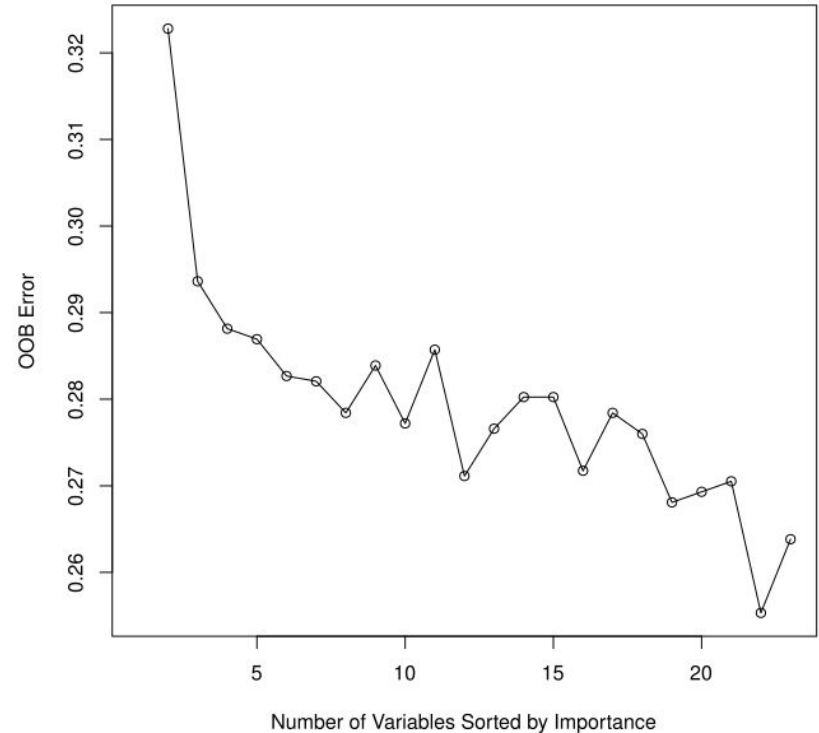
1. Impute with KNN and `na.interp()` function from forecast package
2. Split the data into 80/20 train to test
3. Find the OOB error for Random Forest by moving variables one by one according to Accuracy Feature Importance (repeat multiple times)
4. Using the optimal number of features, cross validate the number of variables at each split and the number of trees
5. Train the optimal Random Forest and make predictions
6. Train the SVM with the linear kernel and make predictions

Random Forest Feature Selection

Random Forest OOB Error vs. # of Variables (Accuracy Metric)

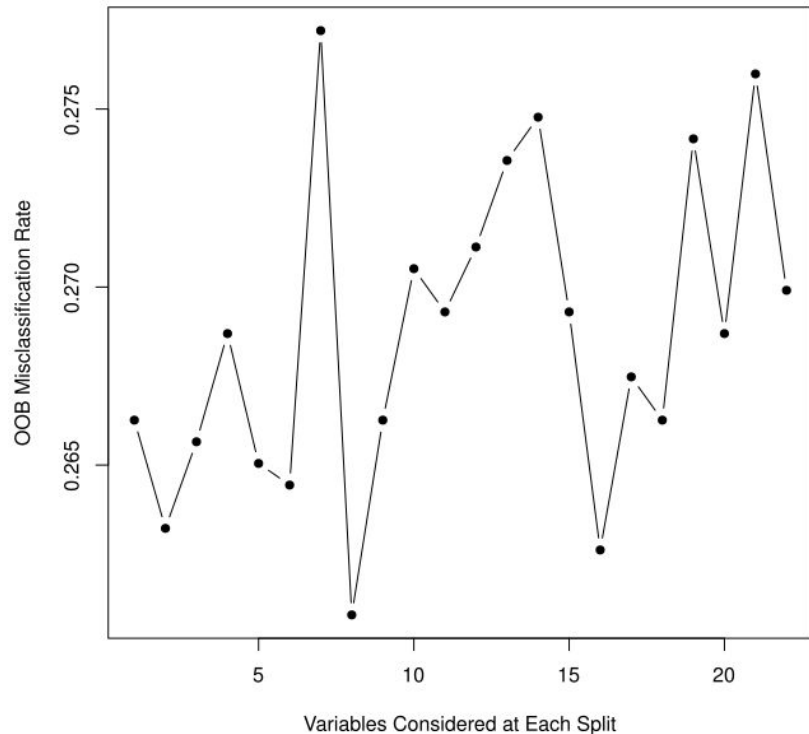


Random Forest OOB Error vs. # of Variables (Gini Metric)

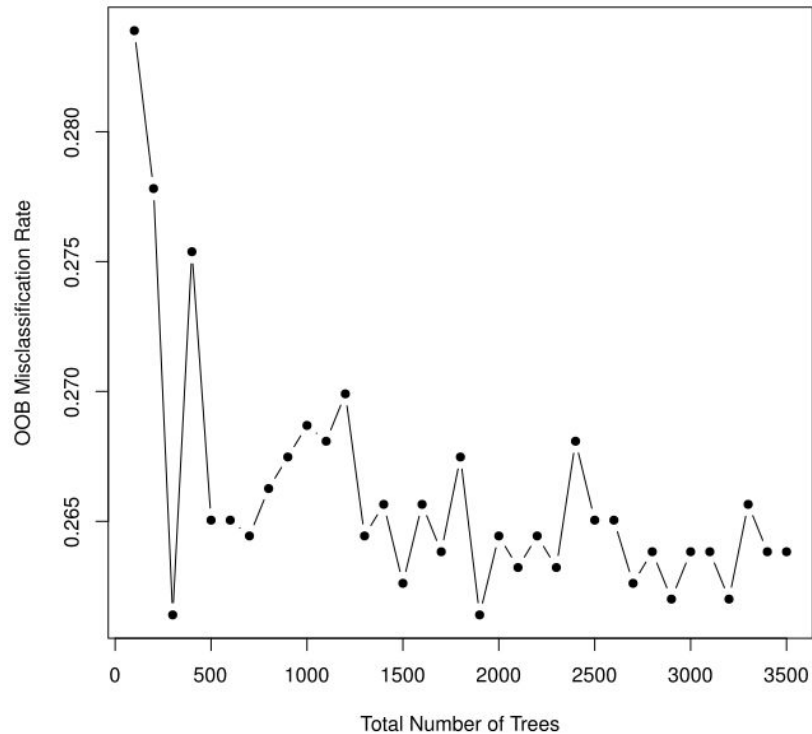


Random Forest Cross Validation

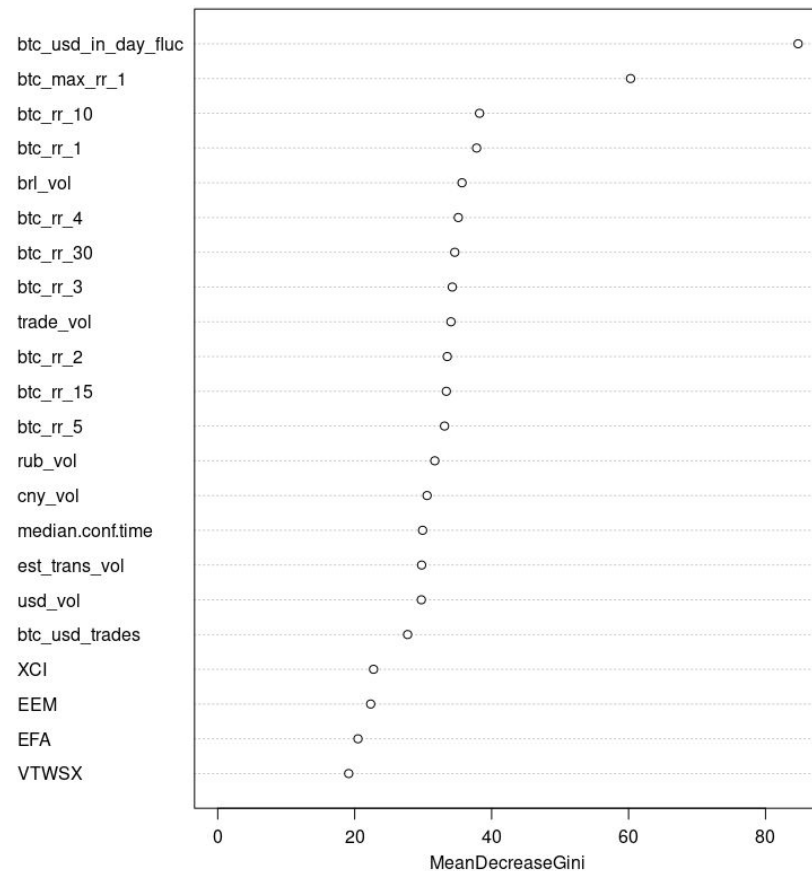
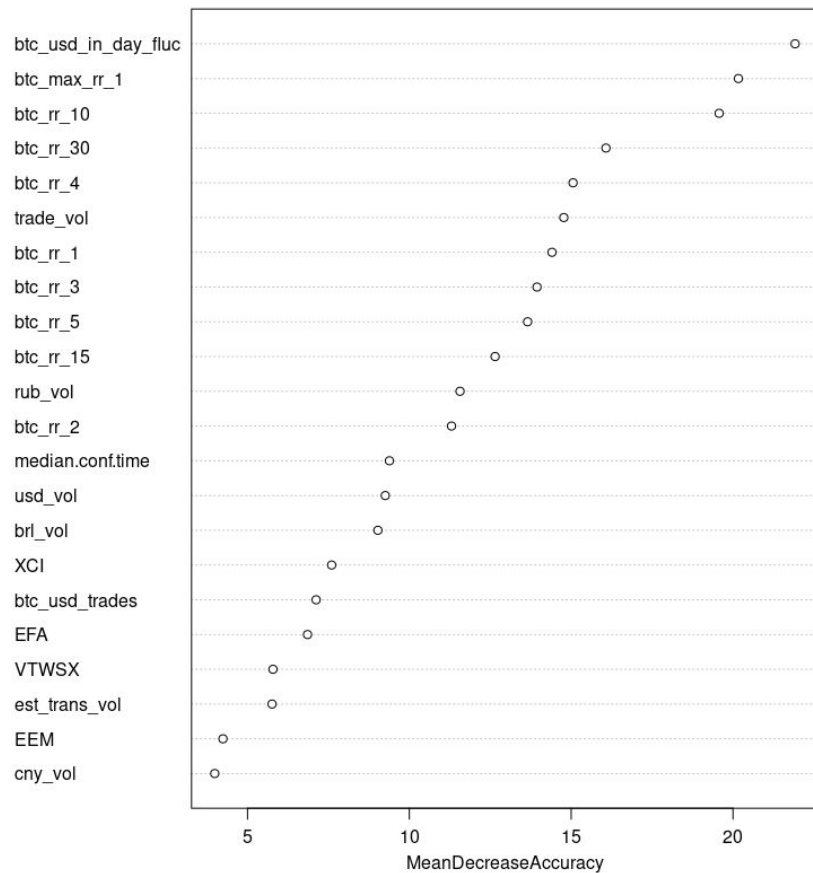
Random Forest OOB Error Rates
by # of Variables



Random Forest OOB Error Rates
by # of Trees



rf.optimal



SVM Results

Radial Kernel

Predicted	True	Frequency
consider	consider	310
buy	consider	15
consider	buy	63
buy	buy	24
Accuracy	Sensitivity	Specificity
81.1%	27.6%	61.5%

Linear Kernel

Predicted	True	Frequency
consider	consider	323
buy	consider	2
consider	buy	78
buy	buy	9
Accuracy	Sensitivity	Specificity
80.6%	10.3%	81.8%

Random Forest Results

Predicted	True	Frequency
consider	consider	883
buy	consider	287
consider	buy	167
buy	buy	308
Accuracy	Sensitivity	Specificity
72.4%	64.8%	51.8%

Predicted	True	Frequency
consider	consider	312
buy	consider	13
consider	buy	56
buy	buy	31
Accuracy	Sensitivity	Specificity
83.2%	35.6%	72.1%

Arbitrage Opportunities (shiny app)



Bitcoin USA: \$2600
Bitcoin China: \$19450
Bitcoin Exchange rate = 7.48
USA/CNY: 6.9

$$7.48/6.9 = 1.08$$

\$1 USD Dollar is turned into
\$1.08

Future To Dos

- PCA on the various return rates and better handle missingness
- Get better data/streaming data
- Migrate away from daily data and process trade by trade
- A full-fledged real time app that gives users the ability to decide whether or not to buy bitcoin given the SVM predictions and current sentiment analysis from Twitter
 - Sends out notifications
- Better understand the discrepancy in exchange rates
- Expand analysis to other cryptocurrencies
- Cryptocurrency portfolio management
- Country case studies (eg: Venezuela, Great Britain)

Thank you!

- TA's
- Instructors
- Fellow Students
- NYC Data Science Academy