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

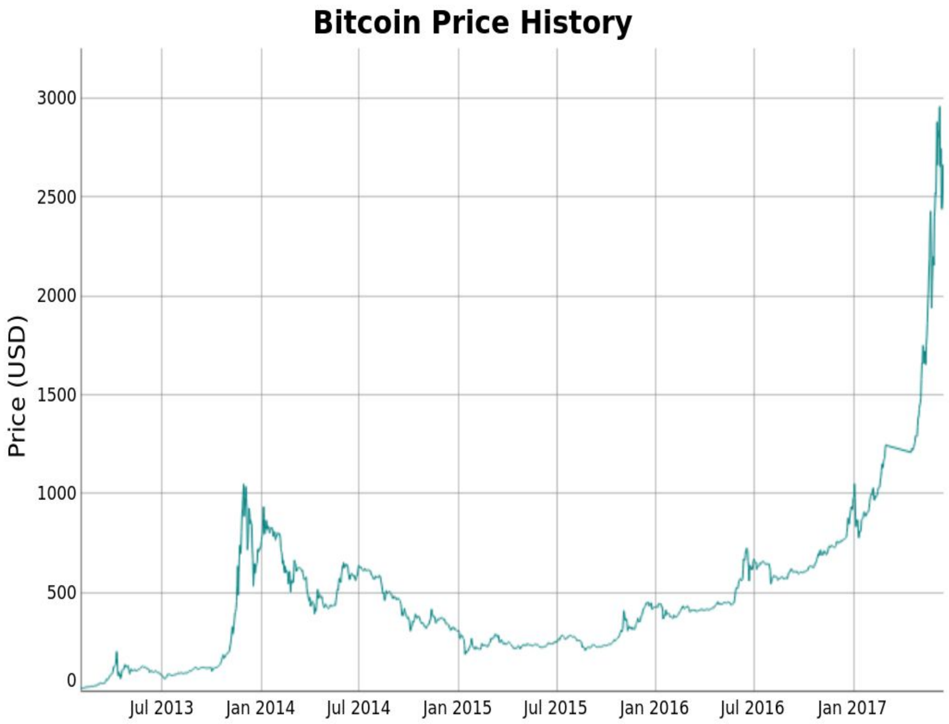
Pipeline and Modeling Techniques

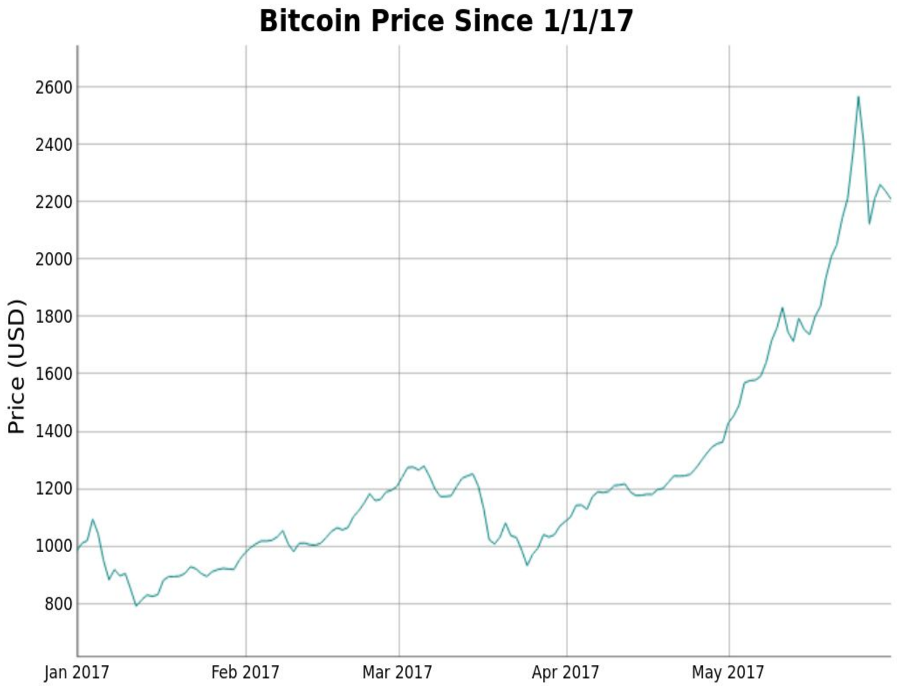
- R and MySQL
- Missingness (VIM)
- Time Series Modeling (quantmod, tseries, forecast)
- Support Vector Machines
- Random Forest (randomForest) and SVM (e1071)
- Shiny App (dygraphs/twitteR)

Outline

- Introduction (where are we now?)
- Classification Problem
- EDA + Feature Extraction + Time Series
- Random Forest and SVM
- Arbitrage Opportunities
- Shiny App (Visualizing Arbitrage, Real Time Tweets)
- Future Directions

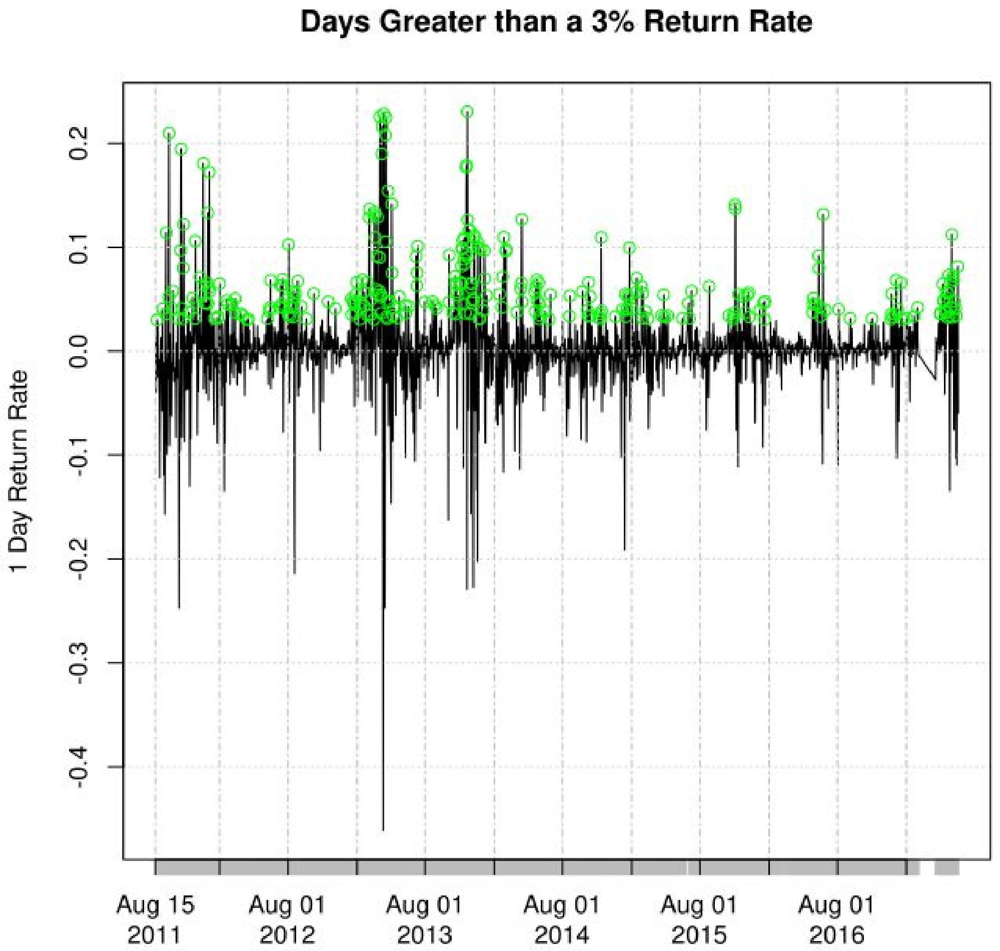
Since the Web Scraping Project





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**

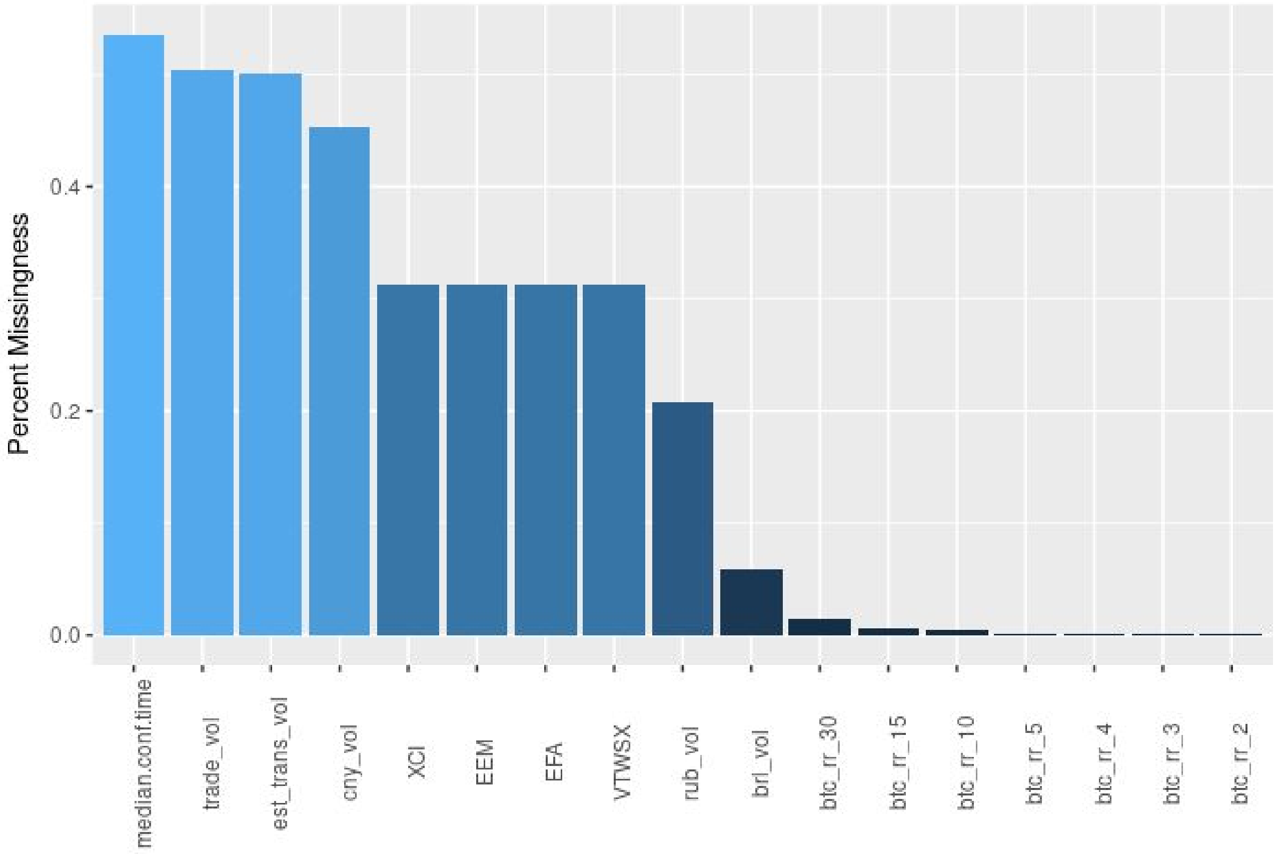


Features

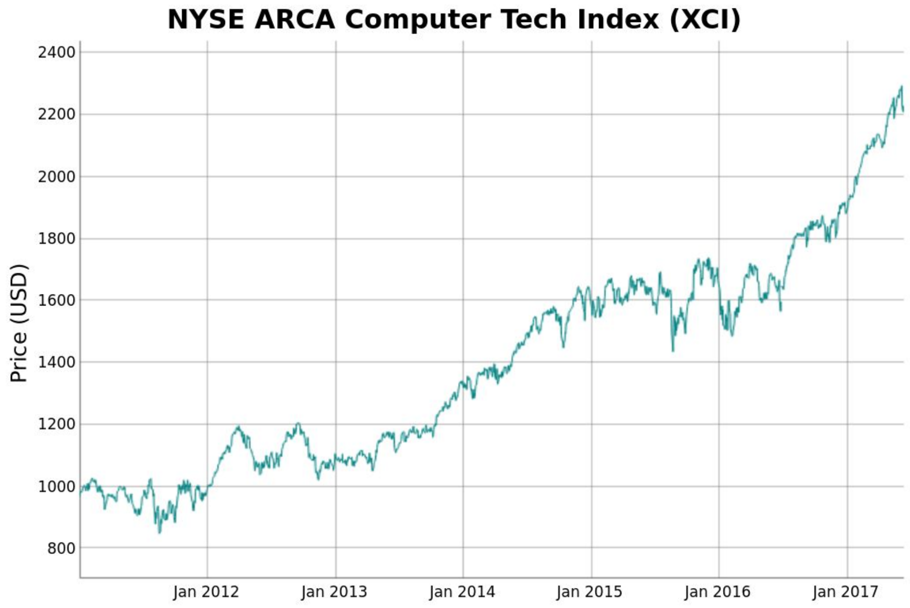
1 Day Return Rate	2 Day Return Rate	3 Day Return Rate	4 Day Return Rate
5 Day Return Rate	10 Day Return Rate	15 Day Return Rate	30 Day Return Rate
Maximum 1 Day Return Rate	Intra Day Fluctuation	Daily US Transactions	Median Confirmation Time
Russian, Chinese, Brazilian, US Volume	Total Transactions	Total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

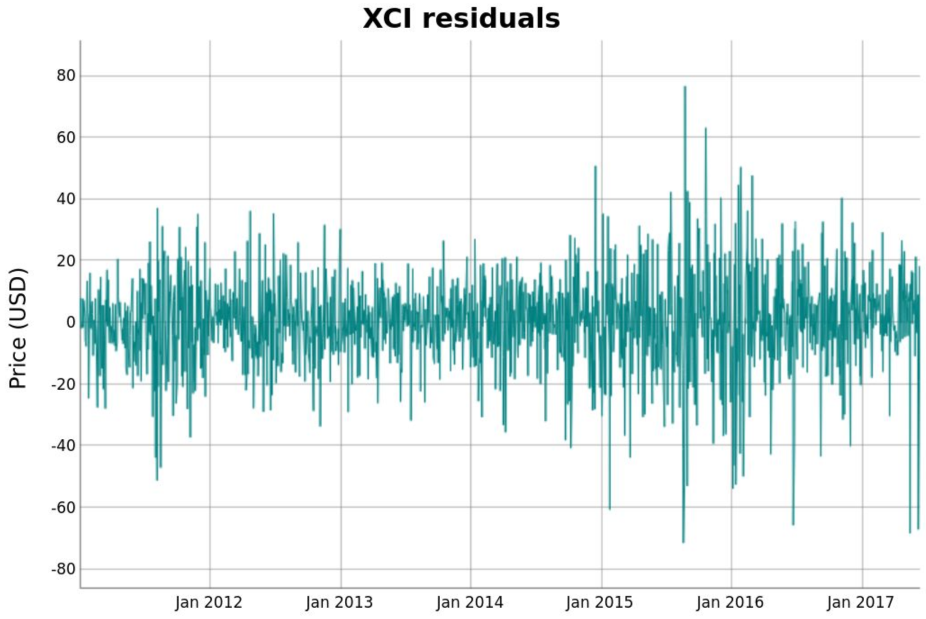
Missingness

- 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

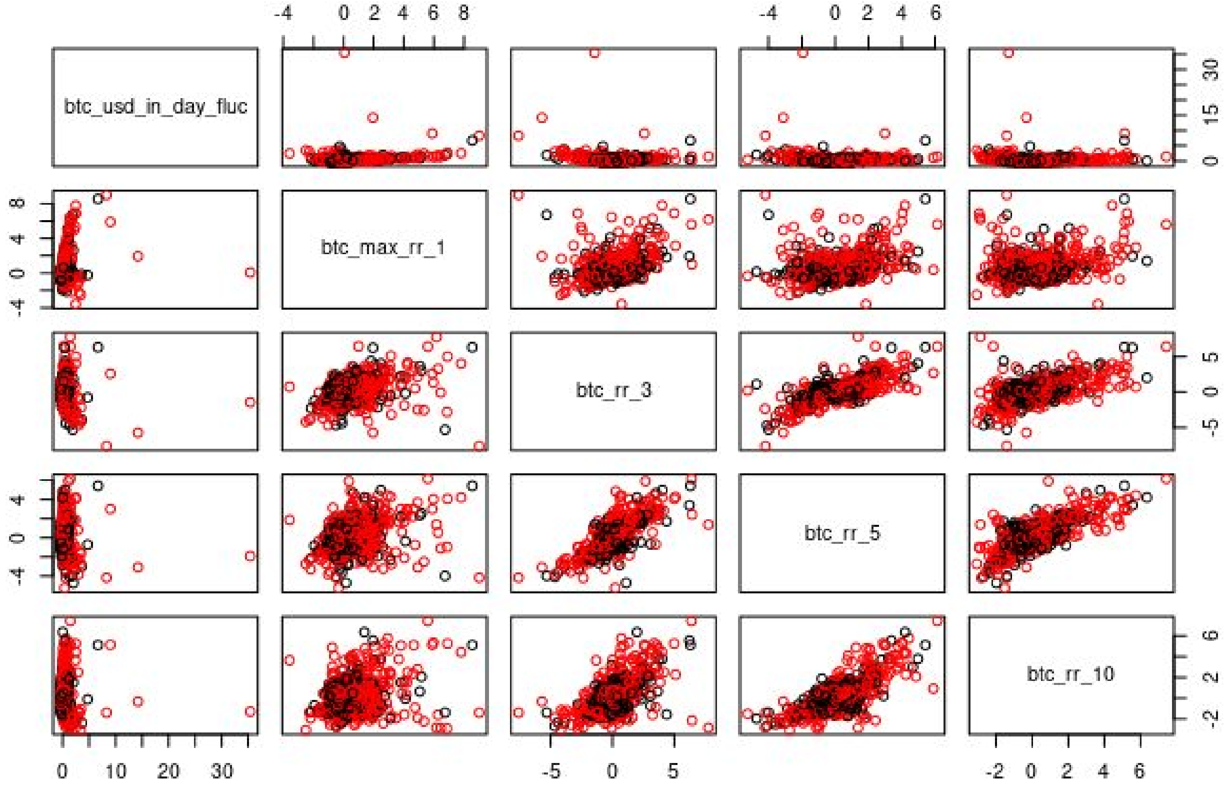


Removing Trend from Time Series Features w/ auto.arima





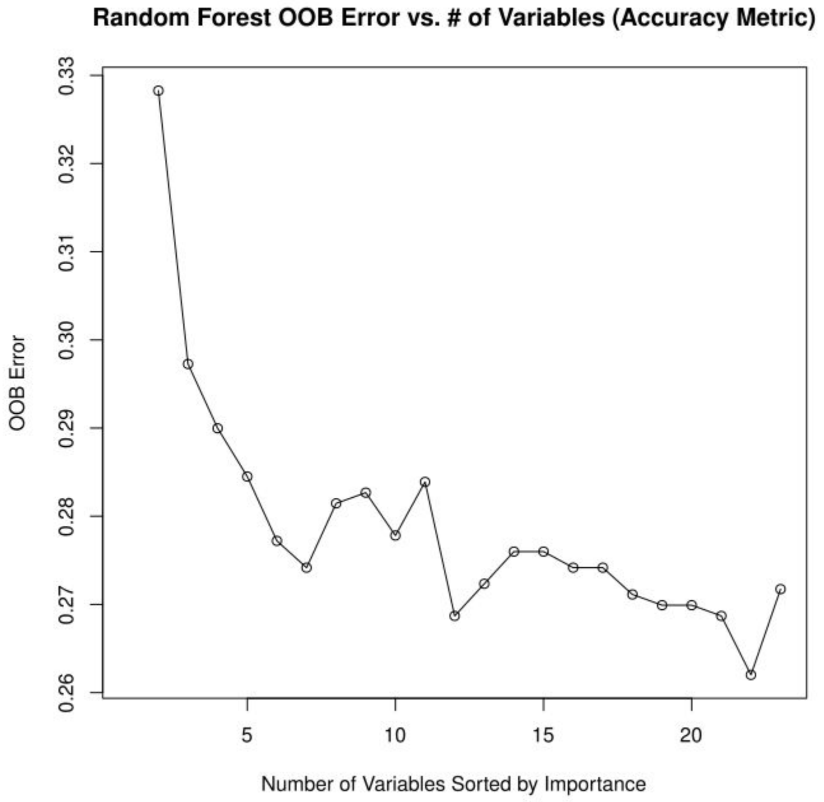
SVM motivation

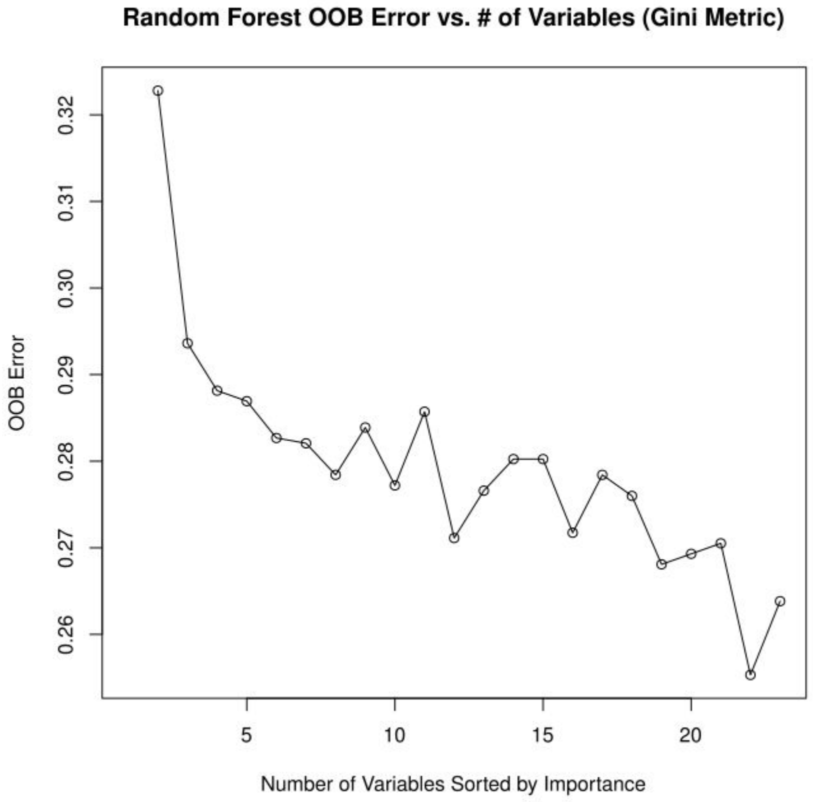


Machine Learning Recipe

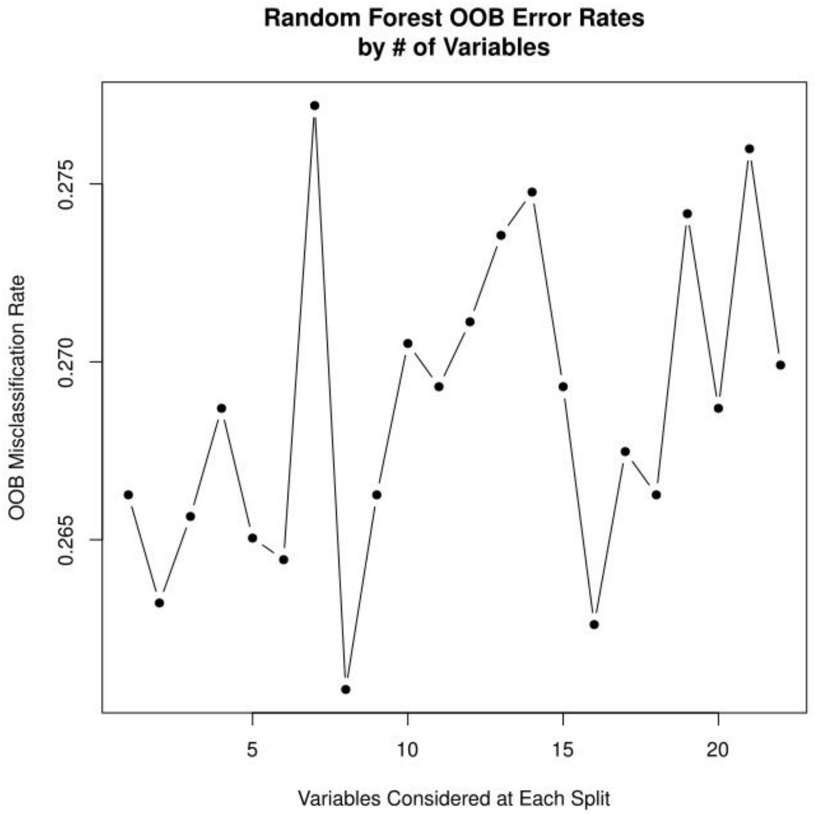
1. Impute with KNN and na.interp() function from forecast package
2. Split the data into 80/20 train to test
3. Find the OOB error for Random Forest by moving variables one by one according to Accuracy Feature Importance (repeat multiple times)
4. Using the optimal number of features, cross validate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

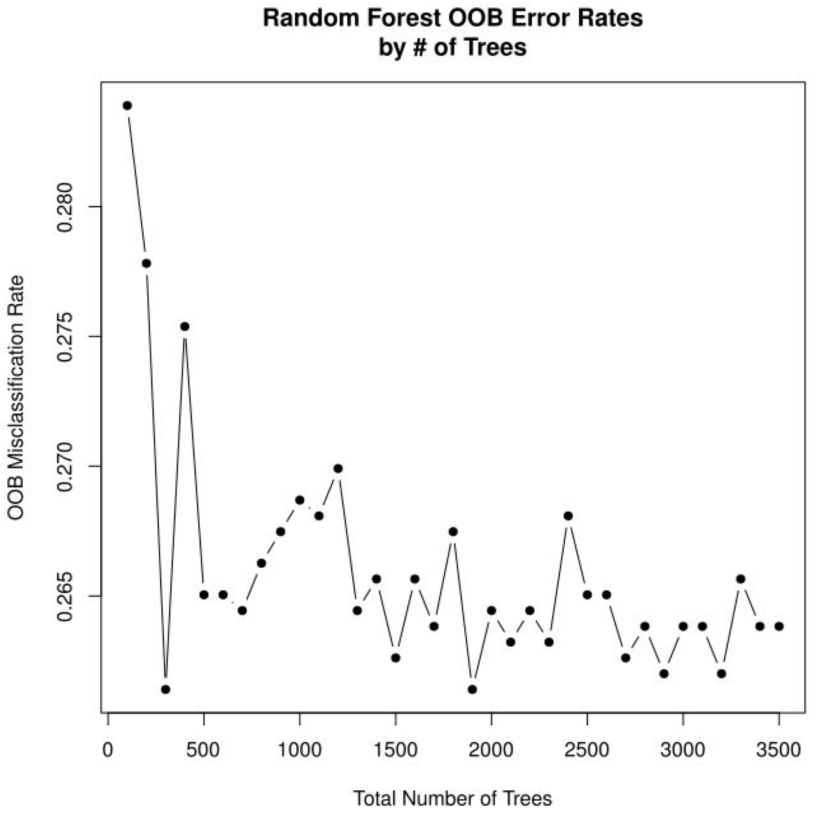
Random Forest Feature Selection



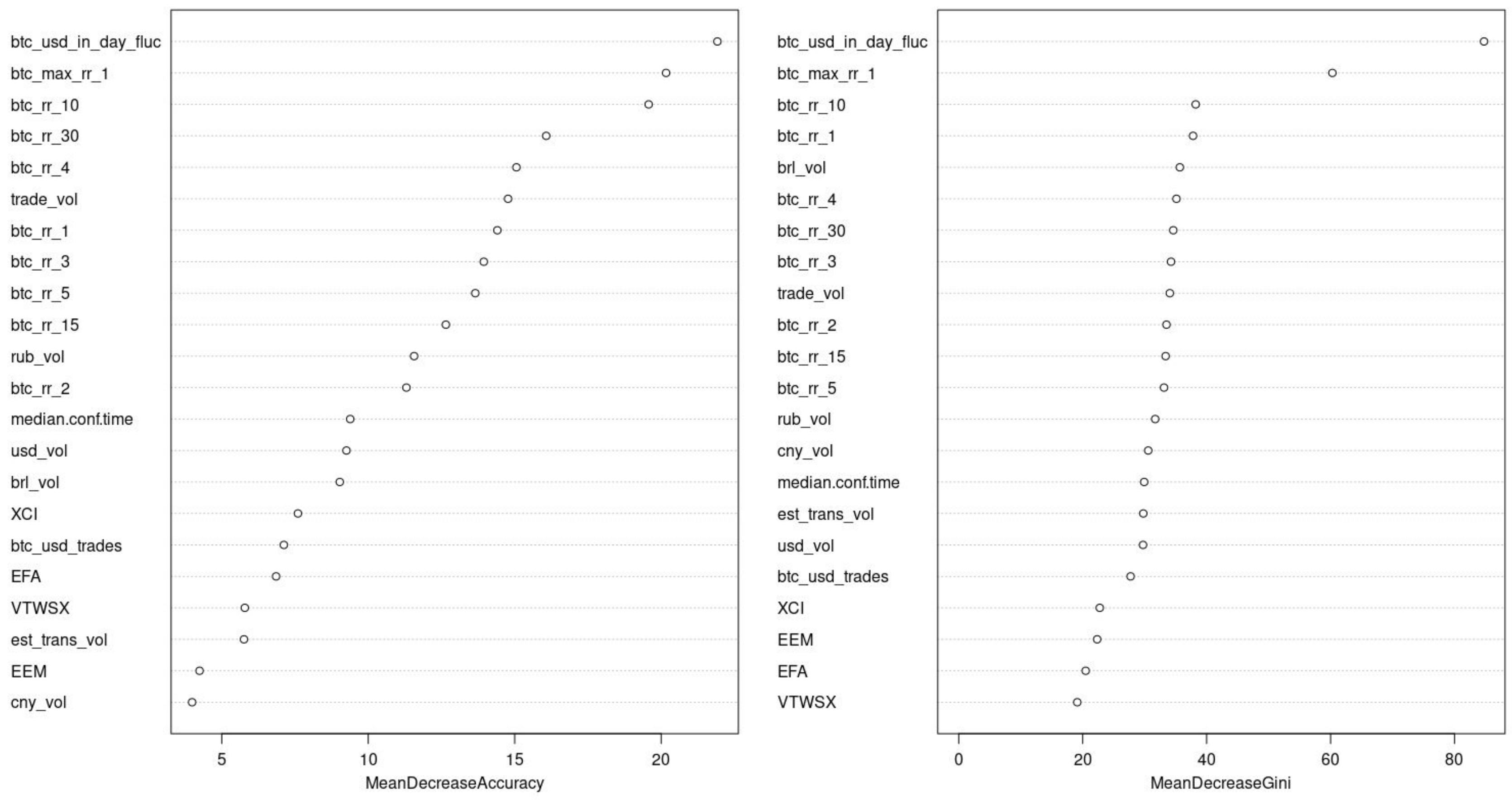


Random Forest Cross Validation





rf.optimal


btc_usd_in_day_fluc
btc_max_rr_1
btc_rr_10
btc_rr_30
btc_rr_4
trade_vol
btc_rr_1
btc_rr_3
btc_rr_5
btc_rr_15
rub_vol
btc_rr_2
median.conf.time
usd_vol
brl_vol
XCI
btc_usd_trades
EFA
VTWSX
est_trans_vol
EEM
cny_vol
5
10
15
20
MeanDecreaseAccuracy
btc_usd_in_day_fluc
btc_max_rr_1
btc_rr_10
btc_rr_1
brl_vol
btc_rr_4
btc_rr_30
btc_rr_3
trade_vol
btc_rr_2
btc_rr_15
btc_rr_5
rub_vol
cny_vol
median.conf.time
est_trans_vol
usd_vol
btc_usd_trades
XCI
EEM
EFA
VTWSX
0
20
40
60
80
MeanDecreaseGini

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

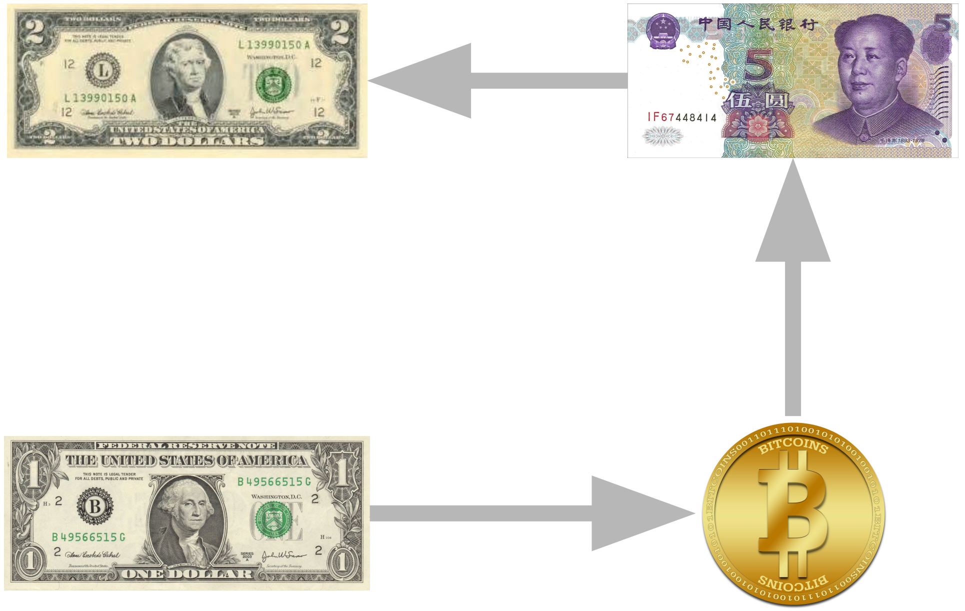
Training Performance

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

Test Performance

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 turns into $1.08

Future To Dos

- PCA on the various return rates and better handle missingness
- Get better data/streaming data
- 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
- 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