# Predicting Ethereum Movement

## 41204: Machine Learning

Shreyash Agarwal, Satya Biswal, Jonathan Caldwell,
Srikanth Geedipalli, Christopher Hemm, Zach Peschke

# Project Goal & Challenges

- **Project Goal:**
  - Predict intraday patterns in the price of ether (currency for Ethereum).
  - Due to volatility of ether, look to classify as a simple up or down movement.
  - Use the model to create potential trading strategy.

- **Challenges:**
  - Price tends to be very volatile which makes predictions difficult.
  - Limited history - relatively new crypto, with data going back to only 2016.
  - Lack of fundamental underlying indicators.

# Background on Ethereum

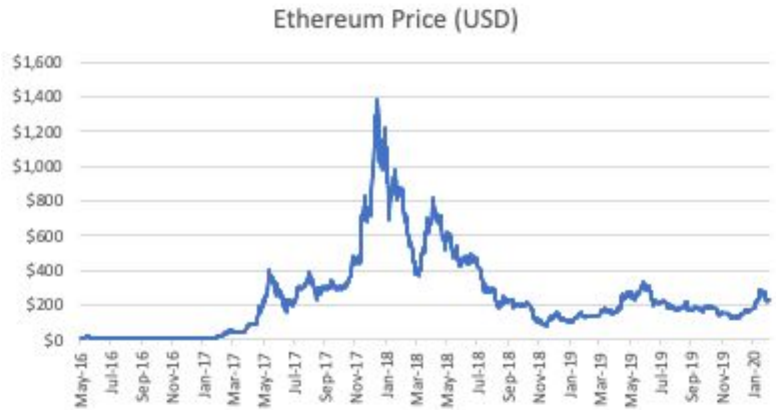
- **Background:**
  - 2nd largest cryptocurrency by market cap behind bitcoin (ranged b/w $14-30B in 2020)
  - Enables SmartContracts and Distributed Applications.
  - Developed to be more than just a currency through the use of its contracts and virtual machines.

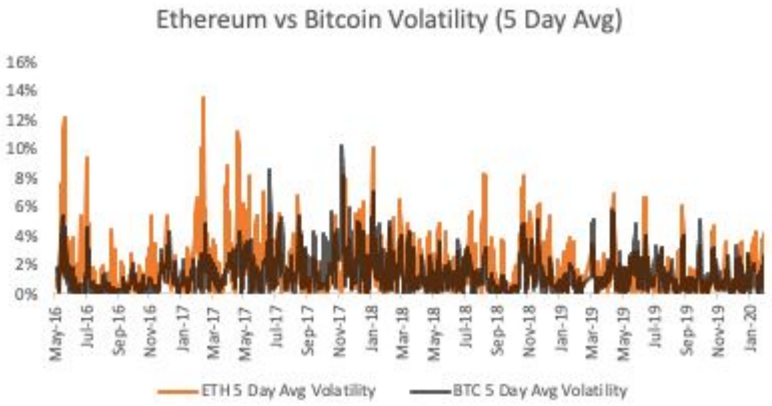
- **History:**
  - Initial release of currency on July 30, 2015.
  - Suffered from security issues in June 2016, resulting in $50M of ether being taken by a hacker.
  - Reached a peak price of ~$1,386 in January 2018, but has since settled around ~$200 USD price over the past year.

# Challenges

- Cryptocurrencies as a whole tend to be difficult to predict due to their lack of underlying fundamentals.

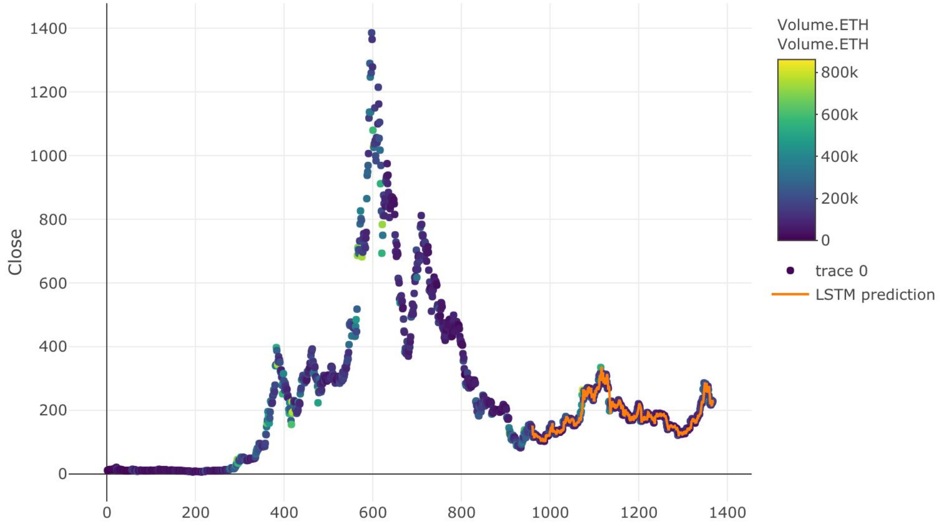
- Ethereum itself has actually been more volatile than bitcoin, with a steep price rise and fall in late 2017-mid 2018, as seen in the two charts below.



# Dataset Info

- **OHLCV Data** (**o**pen, **h**igh, **l**ow, **c**lose, **v**olume)
  - Coinbase is used for OHLC
  - Volume is exchange dependent, we'll use several
- **Technical Indicators (TIs)**
  - Are often lagged transformation of OHLCV
  - Ex: SMA, EMA, BBands, RSI
- **70:30 train/test split, data since May '16:**
  - Purple - Train
  - Orange - Test




**Challenges:**
- Training on high volatility (must train chronologically)
- Newer asset means smaller train/test
- Calibration of TIs means more data may be excluded

CHICAGO BOOTH

# Data Cleaning

**We focus efforts on prediction of intraday returns**

- Apply differencing functions to stationarize data

**Standardize daily returns to ~ N(0,1)**

**Apply one period lag (1 day or 1 hr, depending on data granularity)**

- Preserves the prediction efficacy, no confounding due to “peeping”

| Close |
|---|
| 11.25 |
| 11.93 |
| 12.34 |
| 12.41 |
| 14.00 |

| closeDiff |
|---|
| 0.00 |
| 0.68 |
| 0.41 |
| 0.07 |
| 1.59 |

| closeDiffScaled |
|---|
| -0.007426532 |
| 0.024329331 |
| 0.011720385 |
| -0.004157546 |
| 0.066826147 |

# Exploratory Data Analysis

**Approach:**

- Due to the lack of underlying fundamentals driving price, we felt the best approach would depend on ethereum data itself (e.g., OHLC data).

- The assumption was that demand for ethereum itself would overwhelmingly drive price, so looking at the OHLC data and using a trended approach would be most predictive.

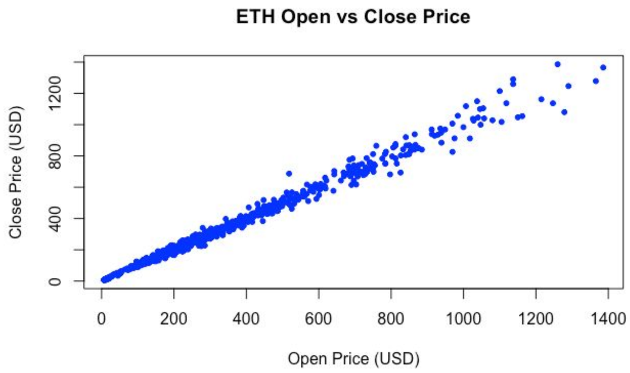
- An additional assumption with our data and model was that the market is efficient and that the exchange used (Coinbase for our data) would not make a significant difference. This will be tested in further iterations.

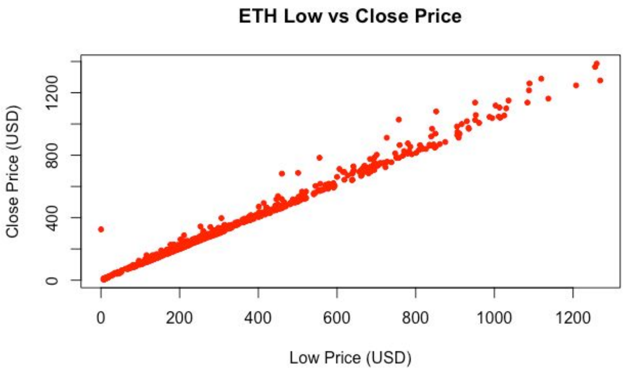
- Based on these, the EDA focused primarily on looking at the volatility of the price itself to ensure the data still contain predictive power.

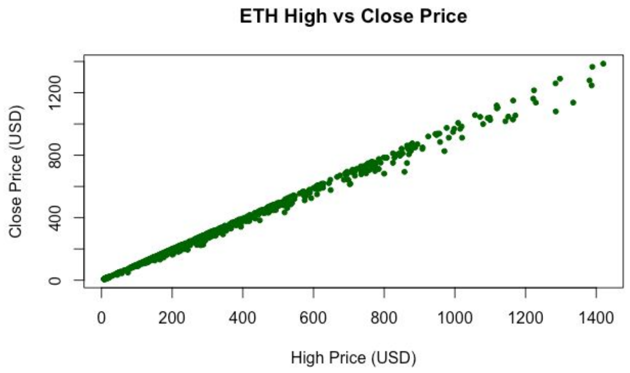
# Exploratory Data Analysis

- Reviewing the OHLC data, there were some data points where the close price was significantly different, but overall the majority of data points were relatively steady.

- Based on this, it does appear that trending the price data can return potential value in the model.

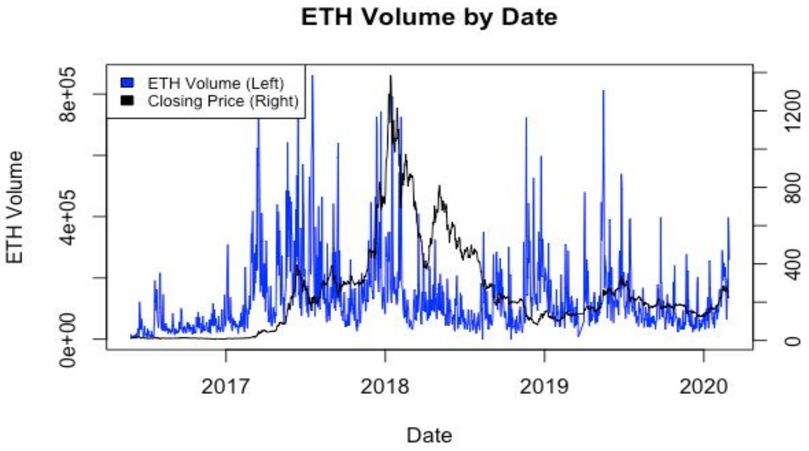


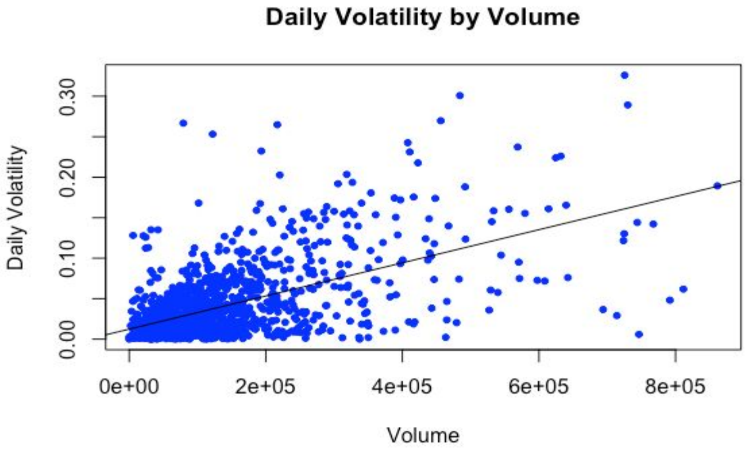



# Exploratory Data Analysis

- The other component that needed to be considered was the volume of ethereum traded and how that potentially impacted price.
- The data showed that there was a statistically significant relationship between the volume traded and the volatility of price, so this was another factor that would need to be considered in model development.




CHICAGO BOOTH

# Exploratory Data Analysis

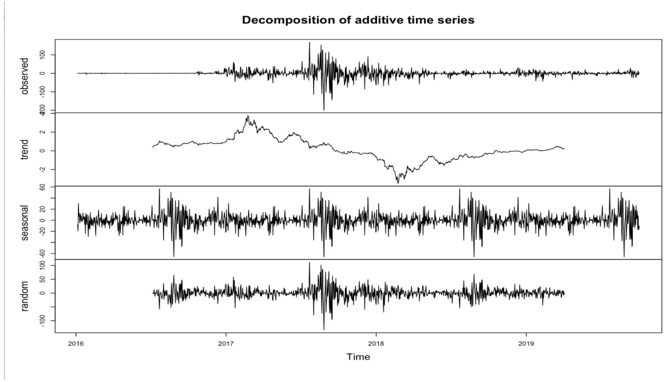
**Time series decomposition:**

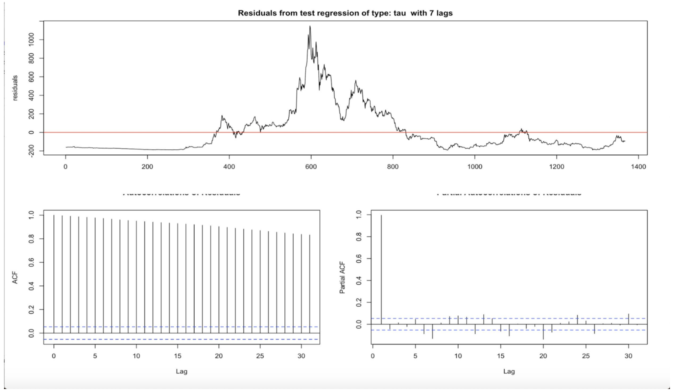
- Decomposition suggested that differenced data was relatively stationarized, exhibiting no meaningful trends

**Unit root test:**

- Performed Kwiatkowski-Phillips-Schmidt-Shin test (KPSS), which indicated a differencing of 1 was sufficient to stationarize data

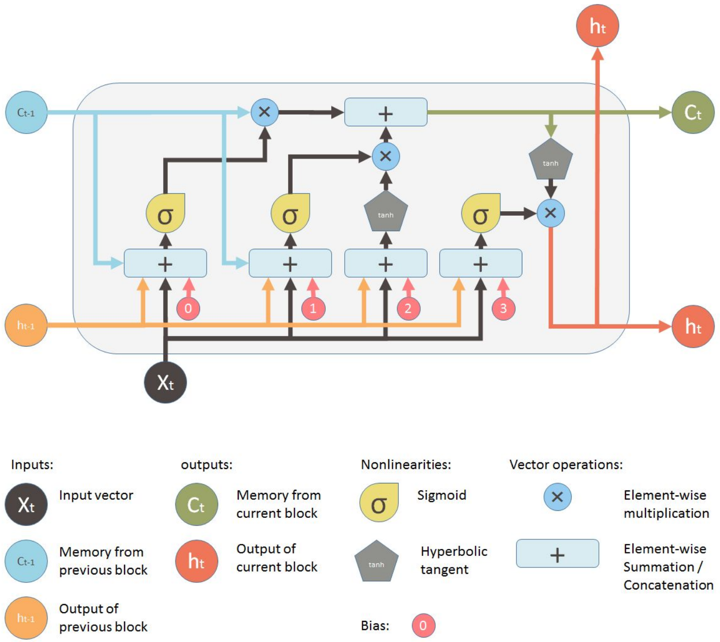
**Conclusion:** both graphical decomposition and unit root test implied that differencing the data by an order of 1 was sufficient for time series analysis and forecasting, notably for the ARIMA model



# Preliminary Models - LSTM

- **Long Short Term Memory Network (LSTM)**
  - RNN, excels at handling long-term dependencies
- **Place differenced, scaled Returns in 3D array necessary for LSTM analysis**
  - We build NN using only returns initially
  - Ensures a minimum viable product for predictions
- **Ran conventional Keras LSTM NN with 5 hidden layers, two dropout**



# Preliminary Models - LSTM (cont.)

**Results:**

- **Model efficacy primarily quantified through accuracy of predictions**
- **We focus on sign, rather than magnitude**
- **409-day out of sample predictions -- 1/20/19~3/5/20**
- **57% predictive accuracy**
    - **> 50% with 95% CI**
- **Still near margin, but preliminary results are promising**

```
          Reference
Prediction   0   1
         0 119  79
         1  97 114

               Accuracy : 0.57
                 95% CI : (0.52, 0.618)
    No Information Rate : 0.528
    P-Value [Acc > NIR] : 0.0509

                  Kappa : 0.141

 Mcnemar's Test P-Value : 0.2000

            Sensitivity : 0.551
            Specificity : 0.591
         Pos Pred Value : 0.601
         Neg Pred Value : 0.540
             Prevalence : 0.528
         Detection Rate : 0.291
   Detection Prevalence : 0.484
      Balanced Accuracy : 0.571
```

# Preliminary Models - LSTM (cont.)

**Alternative Results Quantification: Construction of Signal for Naive Trading Strategy**

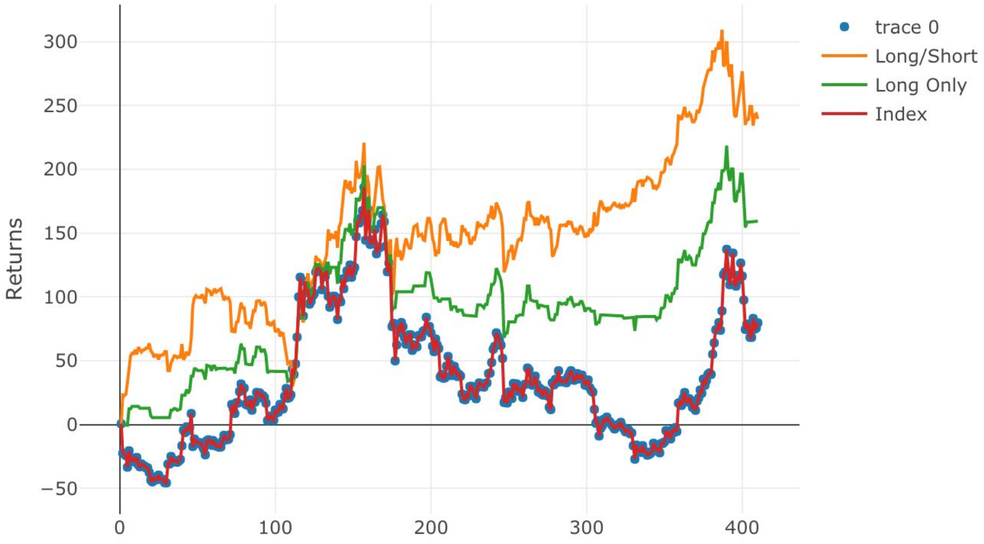
**Represents binary intraday movement classifier**

- Long/Short
- Long Only
- Index (Control)

**“Traded” on all Test data**

**Observations:**

- L/S ability to capitalize on index decline
- Long Only abstinence advantage
- Both have unique strategy features



# Preliminary Models - Supervised Models

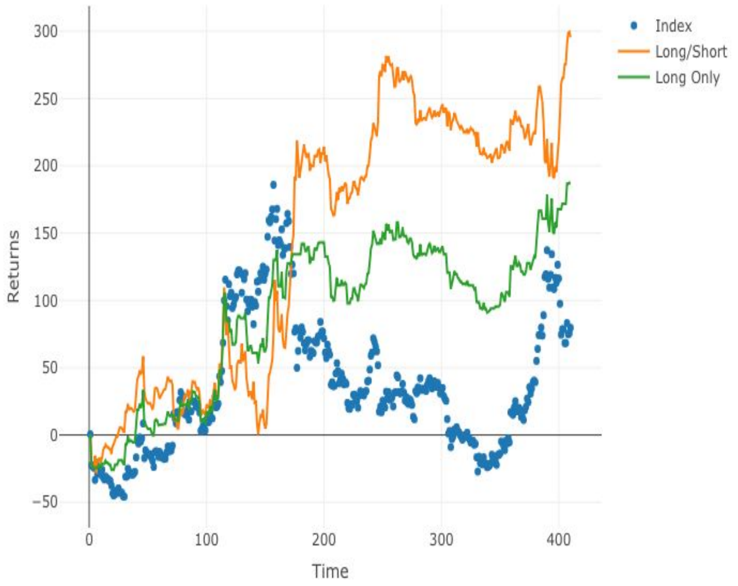
- **Supervised models used: ARIMA, Gradient Boosting and Random Forests**
  - **ARIMA (autoregressive integrated moving average) :** model that is commonly used for time series analysis and forecasting, used on data that is stationarized
- **Predictive methodology**
  - **ARIMA:** used a difference of 1 and a lag of 0 based on exploratory analysis and the lowest AIC value
  - **Boosting and Random Forest**: predict close price difference based on a 5-day exponential moving average
    - Boosting: 100,000 trees
    - Random Forest: 10,000 trees
  - **Outputs:** confusion matrix predicting positive/negative price movements, returns based on long, long/short trade strategy
  - **Note:** unlike LTSM model, data is not scaled and unscaled, as variable units are consistent (no variables are resultantly overweighted)

# Supervised models - Gradient Boosting

- **Confusion matrix results**
  - 53.7% prediction accuracy
  - Not >50% with 95% confidence

- **Returns comparison from long and long/short trading strategy**
  - Both trading strategies outperform buy and hold (control) after ~400 days
  - Performs well predicting high volatility movements, allowing for strong returns

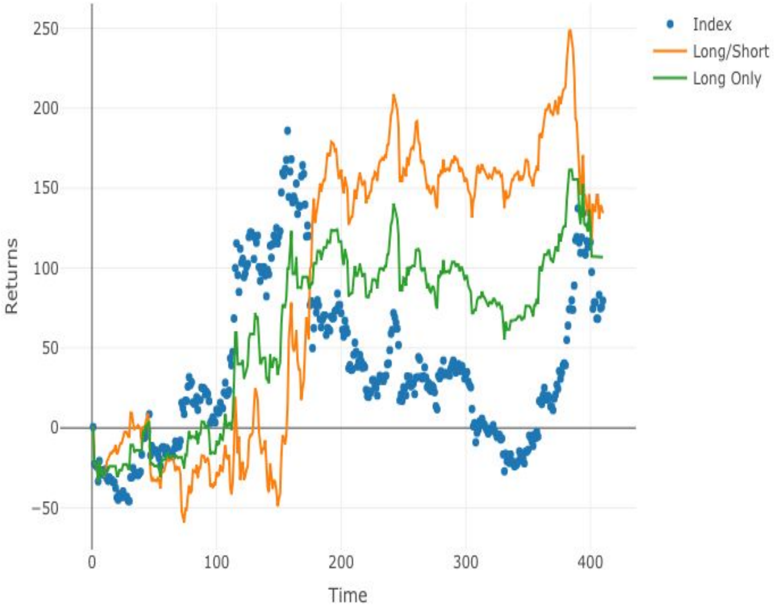
```
          Reference
Prediction   0   1
         0  72  56
         1 134 148

               Accuracy : 0.537
                 95% CI : (0.487, 0.586)
    No Information Rate : 0.502
    P-Value [Acc > NIR] : 0.0911

                  Kappa : 0.075

 Mcnemar's Test P-Value : 0.0000000232

            Sensitivity : 0.350
            Specificity : 0.725
         Pos Pred Value : 0.562
         Neg Pred Value : 0.525
             Prevalence : 0.502
         Detection Rate : 0.176
   Detection Prevalence : 0.312
      Balanced Accuracy : 0.538

       'Positive' Class : 0
```



# Supervised models - Random Forest

- **Confusion matrix results**
  - 52.4% prediction accuracy
  - Not >50% with 95% confidence
- **Returns comparison from long and long/short trading strategy**
  - Both trading strategies outperform buy and hold (control) slightly after 400 days
  - Does not predict high volatility movements as well as boosting

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  69  58
         1 137 146

               Accuracy : 0.524
                 95% CI : (0.475, 0.574)
    No Information Rate : 0.502
    P-Value [Acc > NIR] : 0.201

                  Kappa : 0.051

 Mcnemar's Test P-Value : 0.0000000233

            Sensitivity : 0.335
            Specificity : 0.716
         Pos Pred Value : 0.543
         Neg Pred Value : 0.516
             Prevalence : 0.502
         Detection Rate : 0.168
   Detection Prevalence : 0.310
      Balanced Accuracy : 0.525

       'Positive' Class : 0
```



# Supervised models - ARIMA

- **Confusion matrix results**
  - 52.2% prediction accuracy
  - Does not predict >50% with 95% confidence
- **Returns comparison from long and long/short trading strategy**
  - Both trading strategies do not significantly outperform control after 400 days

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 158 148
         1  48  56

               Accuracy : 0.522
                 95% CI : (0.472, 0.571)
    No Information Rate : 0.502
    P-Value [Acc > NIR] : 0.229

                  Kappa : 0.042

 Mcnemar's Test P-Value : 0.00000000000153

            Sensitivity : 0.767
            Specificity : 0.275
         Pos Pred Value : 0.516
         Neg Pred Value : 0.538
             Prevalence : 0.502
         Detection Rate : 0.385
   Detection Prevalence : 0.746
      Balanced Accuracy : 0.521

       'Positive' Class : 0
```

# Next Steps

- **Additional Data (for incorporation into most promising model):**
  - Technical Indicators
  - Blockchain Data
- **Implement Hidden Markov Model**
- **Model magnitude of returns in addition to the sign of the returns.**
- **Based on best model, determine how prediction signal can be utilized optimally for a trading strategy**