Disclaimer: I am not a financial institution nor do I represent one. All that I am providing is educational material: Do not take this information as professional investment advice.

Cryptocurrency Capstone I

Introduction

With all the hype that has been generated about cryptocurrencies in 2017, there has been nothing that has skyrocketed more than the fear of missing out on what could be one of the hottest investment vehicles ever available. To help translate this hype into practical insight that you can use to aid your investments, four separate cryptocurrencies will be looked at:

- Bitcoin
- Ethereum
- Litecoin
- Zcash

For the cryptocurrencies, mentioned above, the following will be investigated:

- Is there a price correlation between cryptocurrencies trading on the market? Does one cryptocurrency move with the price of another?
- Are we able to predict future values of cryptocurrencies? If so, how accurately?
- And, last but not least, can we develop an investment strategy to make some money? If so, how accurately?

Client

Clients include individuals and/or financial institutions interested in investing in cryptocurrencies. Based on the analysis performed in this report, individuals will be able to make better informed decisions when investing their money in the cryptocurrency market.

Data Collection & Wrangling

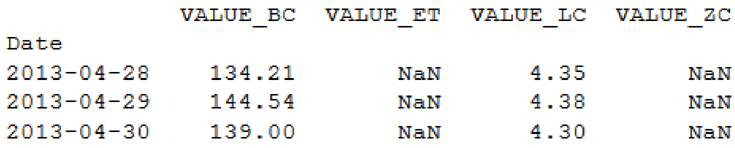
This section is broken out into three parts and will describe the methodologies used to collect and transform data for the analyses performed in this report. All data for this report was compiled from CoinMarketCap, which is a website that provides historical and present data for all cryptocurrencies trading on a public exchange. The data was analyzed over a time period ranging from 4/28/2013 to 11/28/2017. The technology stack used to transform and perform analysis on the data is Python.

Part I: Collecting Files & Converting to DataFrame

Files for each cryptocurrency were extracted as individual Comma Separated Values (CSV) files. Each of the files were stored locally, picked up using the ‘glob’ function, and converted to a Pandas DataFrame to facilitate analysis. The pseudocode and sample output are shown below.

Pseudocode

```
a) Collect  CSV files locally using ‘glob’ function
b) Convert each file to a Pandas DataFrame
     •  Dates used for index
     •  Cryptocurrency values used as column  (one value column for each
        cryptocurrency)
     •   Relabel column names
            – VALUE_BC = Value of Bitcoin
            – VALUE_ET = Value of Ethereum
            – VALUE_LC = Value of Litecoin
            – VALUE_ZC = Value of Zcash
c) Join all individual DataFrames into single DataFrame with ‘concatenate’ in
   Pandas
```



```
            VALUE_BC  VALUE_ET  VALUE_LC  VALUE_ZC
Date
2013-04-28    134.21       NaN      4.35       NaN
2013-04-29    144.54       NaN      4.38       NaN
2013-04-30    139.00       NaN      4.30       NaN
```

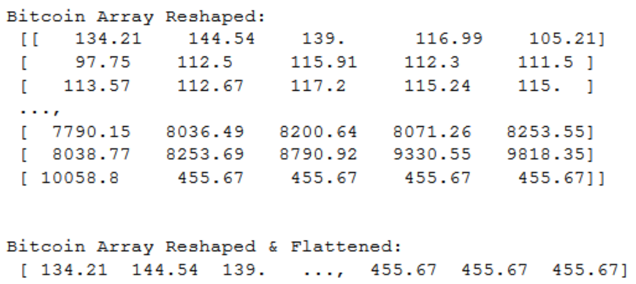
```
d) NaN’s  were left alone at this stage, since when plotting DataFrames Pandas is
   smart enough to ignore them
```

Part II: Reshaping Data for Machine Learning Algorithm

The values from each cryptocurrency column, stored in the DataFrame in part I, were then extracted and passed to a user defined function that was used to reshape the values from each respective column into a rectangular matrix. The user defined function ‘reshape_array()’ allows data to be reshaped into any specified shape of n rows by m columns. This function will be extremely helpful to reshape data when analyzing the optimal number of days needed to make future cryptocurrency value predictions with the machine learning algorithms implemented. The pseudocode and sample output are shown below.

Pseudocode

```
a) Extract columns in DataFrame above,  as individual  NumPy arrays
b) Develop reshaping function called ‘reshape_array()’
    • Goal:  Reshape array into a matrix with specified ‘number of columns’
    • If specified ‘number of columns’ makes matrix un-rectangular, elements
      will be added to make it rectangular.
    • To fill extra elements, the median value of array is taken
    • Function returns a reshaped and flattened array
c) Results of ‘reshape_array()’ function with 5 columns
```



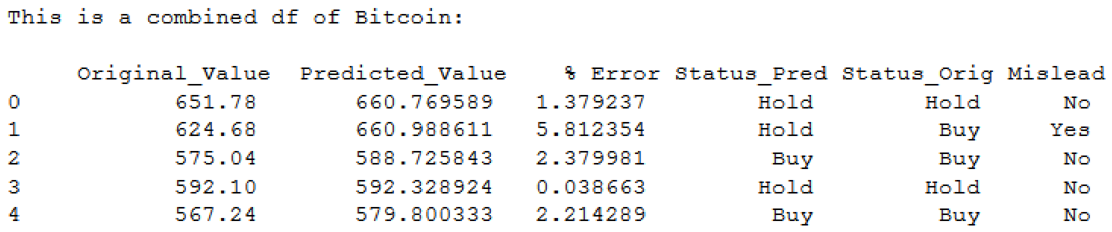
```
Bitcoin Array Reshaped:
 [[   134.21    144.54    139.      116.99    105.21]
 [    97.75    112.5     115.91    112.3     111.5 ]
 [   113.57    112.67    117.2     115.24    115.  ]
 ...,
 [  7790.15   8036.49   8200.64   8071.26   8253.55]
 [  8038.77   8253.69   8790.92   9330.55   9818.35]
 [ 10058.8     455.67    455.67    455.67    455.67]]


Bitcoin Array Reshaped & Flattened:
 [ 134.21  144.54  139.    ...,   455.67  455.67  455.67]
```

PART III: Creating DataFrame to Hold Machine Learning Outcomes

Another user defined function, ‘pred_compare_df()’ , was generated to create a new DataFrame to store original and predicted values along with keeping track of investment or trading moves. The pseudocode and sample output are shown below.

Pseudocode

```
a) Develop function ‘pred_compare_df()’ to create a DataFrame and hold the
   following:
    • Predictions from Machine Learning algorithm
    • Original or true values of cryptocurrency
    • Percent change between original and prediction values
    • Whether the cryptocurrency should be held or bought for investing
```



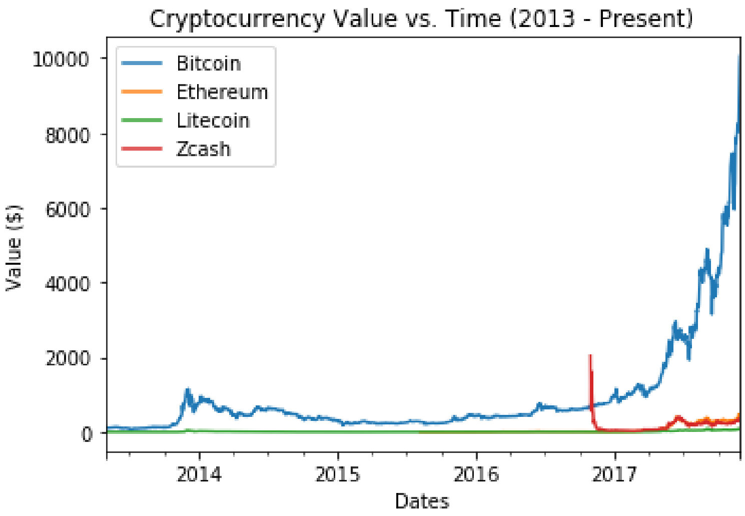
```
This is a combined df of Bitcoin:

     Original_Value  Predicted_Value    % Error Status_Pred Status_Orig Mislead
0            651.78       660.769589   1.379237        Hold        Hold      No
1            624.68       660.988611   5.812354        Hold         Buy     Yes
2            575.04       588.725843   2.379981         Buy         Buy      No
3            592.10       592.328924   0.038663        Hold        Hold      No
4            567.24       579.800333   2.214289         Buy         Buy      No
```

Now with the data in tip top shape we can go ahead and start exploring it to see if some interesting insights can be draw about the exciting cryptocurrency market.

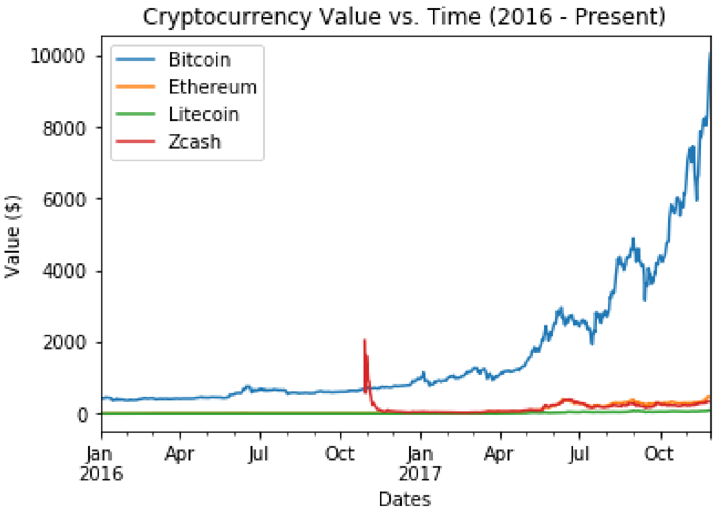
Disclaimer: I am not a financial institution nor do I represent one. All that I am providing is educational material: Do not take this information as professional investment advice.

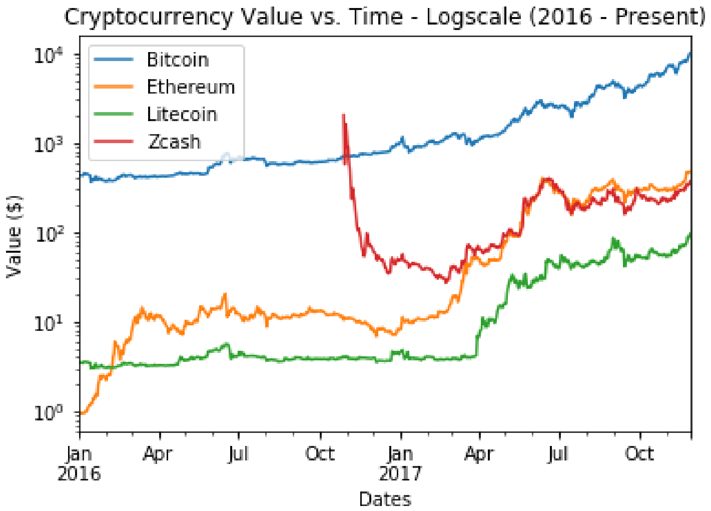
Exploratory Data Analysis

Plotting the value of four separate cryptocurrencies (Bitcoin, Ethereum, Litecoin, and Zcash) over time, we can clearly see the steep ascent Bitcoin's value has taken but it's difficult to generalize about the other 3 cryptocurrencies. The difficulty in discerning the general trend for Ethereum, Litecoin, and Zcash stems from the differing scales of each cryptocurrency value and the times they came onto the market. Taking a closer look and shifting the timeline (x-axis) we may be able to gain better insight about the general trend each cryptocurrency is taking on.

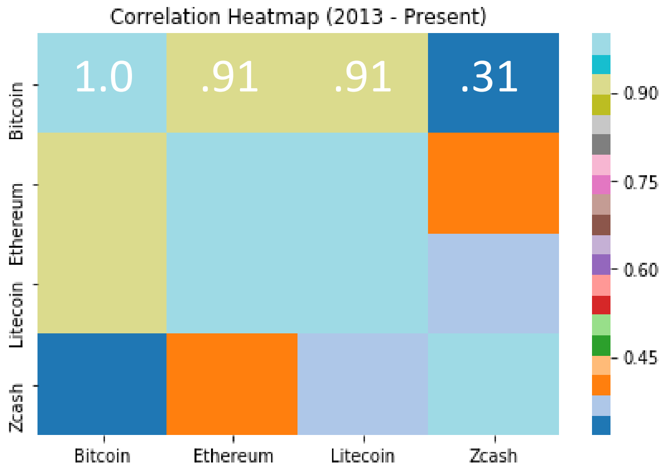


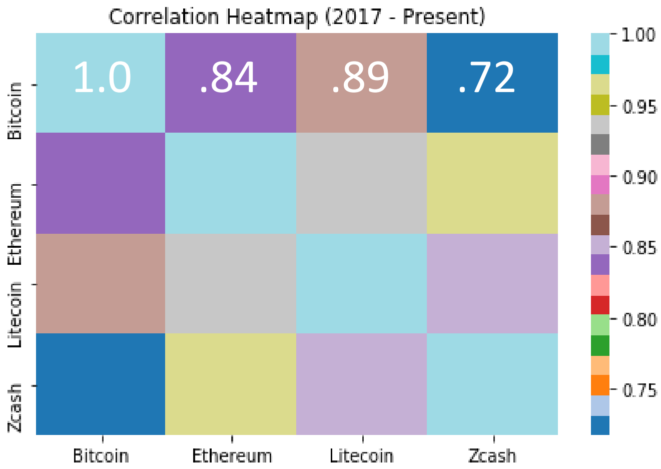
Shifting our timeline scale (x-axis) to start at 2016, we still can't gather much with the exception that Bitcoin's peaks and valleys are more clearly defined (graph on left). This time around we shift our value scale (y-axis) with a logarithmic transformation, and bingo, the trends become much clearer (graph on right). Amazingly, at a glance, all cryptocurrencies plotted look to have some correlation between one another.





Running a correlation test (using Pearson's R) for all cryptocurrencies we can assess how strong the linear relationships are between cryptocurrencies. A strong relationship would have a value of 1, while a weak relationship would show up having a value of 0. To visually highlight the relationship between all variables, heat maps have been generated below.





Looking at the heat map for all data (2013-Present) we can see that Bitcoin, Ethereum and Litecoin are strongly correlated (correlation is approximately .9) while Zcash sits alone with a correlation of approximately .3. This is interesting enough, but things get even more interesting as we shift our timescale to a fairer point where all cryptocurrencies actually had a chance to enter the market, the year 2017. Looking at the heat map generated from 2017 to present (heat map on the right) we can see that all cryptocurrencies, overall, have an even stronger relationship with one another. What's nice to see is that Zcash is getting in on the action with approximately a .7 correlation. With this information it would be very exciting to see how well we could use multiple cryptocurrencies to predict the value of another, but at this time these predictions still need to be looked into and will come as a future update to this project. Instead we will investigate how well we can predict a single cryptocurrency using its historical data.

Machine Learning

The goal of this section is to test two separate machine learning algorithms to see which one can best predict the future values of a given cryptocurrency. Linear Regression and Ridge Regression are the two machine learning algorithms that will used for this analysis.

For each of the machine learning algorithms or models implemented, the Root Mean Square Error (RMSE) will be used as the evaluation metric to measure the accuracy of the cryptocurrency predictions. To optimize each of the models we will vary the look back period (i.e., number of historical days being used to make a single day prediction) and select the model with the lowest RMSE. The RMSE is a descriptive number that tells us on average how far the models predictions are from the true cryptocurrency values.

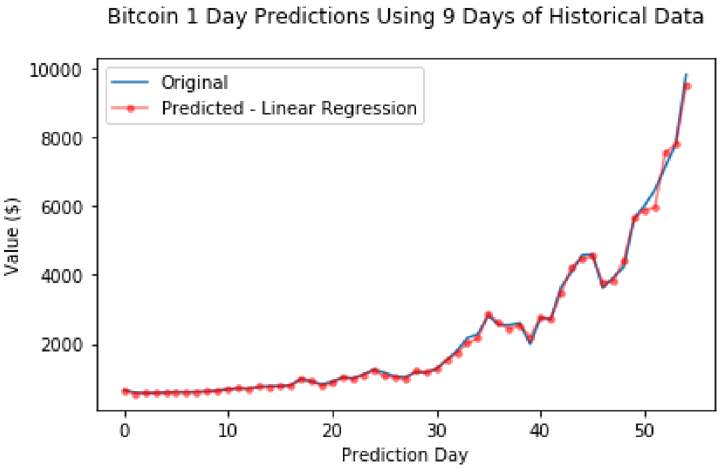
Linear Regression

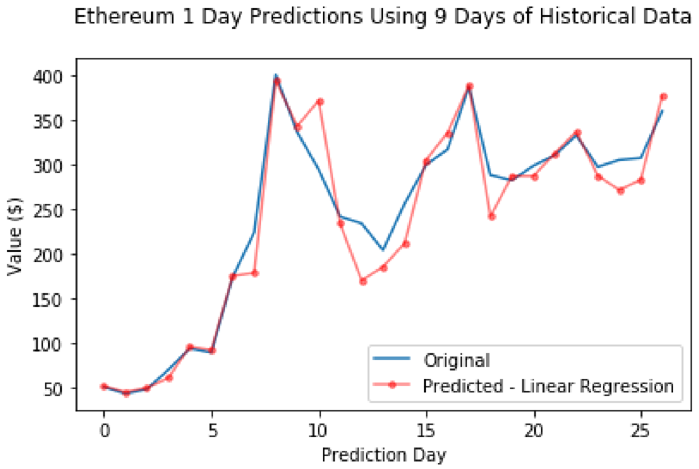
Here we will be using a multivariate Linear Regression analysis to develop a model that will be able to predict the value of each of the four cryptocurrencies selected. What makes this a multivariate problem is the use of more than one feature variable (i.e., the lookback period which can span multiple days) to predict the desired target variable (i.e., the value of the cryptocurrency). After varying the look back period, it was found that a look back period of 9 days provided the optimal model with a RMSE of approximately $170. The results of the analysis are summarized in the table below.

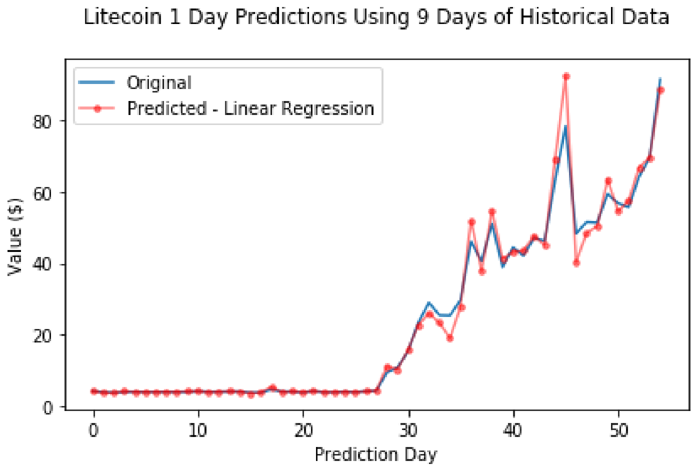
RMSE Evaluation Using Linear Regression

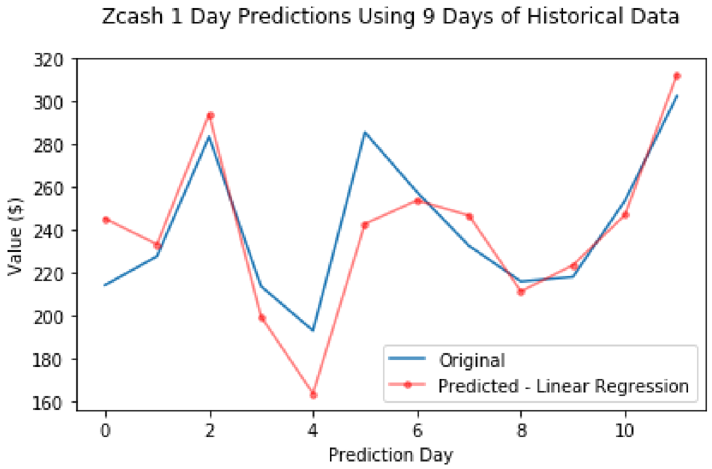
Look Back Period	**Bitcoin**	**Ethereum**	**Litecoin**	**Zcash**	**∑ RMSE's**
(Days)	($)	($)	($)	($)	($)
1	136.6	15.7	2.49	42.3	197.16
2	148.1	17.2	2.8	48.1	216.18
3	137.8	16.8	3.07	17.7	175.33
4	163.6	15	2.96	20.7	202.2
5	152.7	20.6	2.92	19	195.25
6	150.7	16.8	2.86	14.7	185.02
7	143.5	25.3	1.98	24.1	194.78
8	193.5	21.7	4.17	22.4	241.77
9	**121.2**	**26.9**	**2.97**	**19.1**	**170.22**
10	170.7	19.4	3.75	22.9	216.73
20	203.2	40.8	3.78	61.3	309.05
30	186.2	16.6	3.78	61.3	267.84

Using the optimized model, value predictions were generated for each of the four cryptocurrencies. Predictions were made using a holdout or a test set of data that was composed of 30 percent split (i.e., 70 percent of data was dedicated for training and 30 percent for testing). Below are four plots that were generated to show the model's predictions against the actual values of each cryptocurrency.


Bitcoin 1 Day Predictions Using 9 Days of Historical Data


Ethereum 1 Day Predictions Using 9 Days of Historical Data


Litecoin 1 Day Predictions Using 9 Days of Historical Data


Zcash 1 Day Predictions Using 9 Days of Historical Data

Taking a look at all four plots, we can easily see that the linear regression model is doing a very good job in predicting and capturing the overall profile for each of the cryptocurrencies analyzed.

Ridge Regression

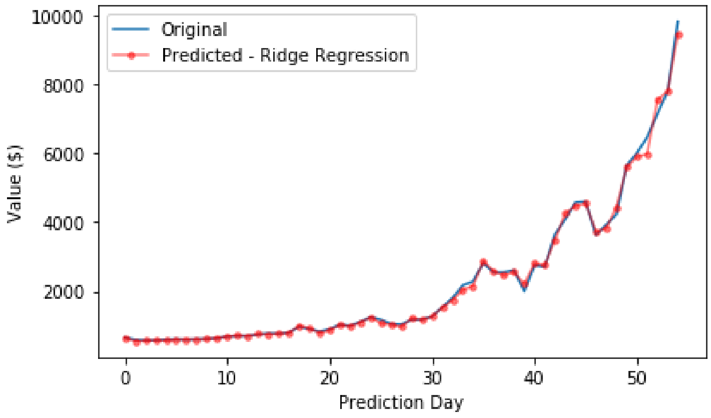
Expanding on the multivariate linear regression model previously developed, we will use Ridge Regression to help combat the potential problem of overfitting. Overfitting is the result of a model capturing the trend of the training data so well that is unable to expand and capture the overall trend when unseen data is thrown at it. Ridge Regression is a regularization method used to help tackle this problem of overfitting by shrinking the models weighted coefficients by adding a penalty term to the traditional linear regression model we looked at in the previous section. After implementing GridSearchCV in SKlearn to find the optimal hyperparameter, alpha, for each cryptocurrency it was again found that a look back period of 9 days generated the optimal Ridge Regression model. The table below summarizes the results of the Ridge Regression model selection.

RMSE Evaluation Using Ridge Regression

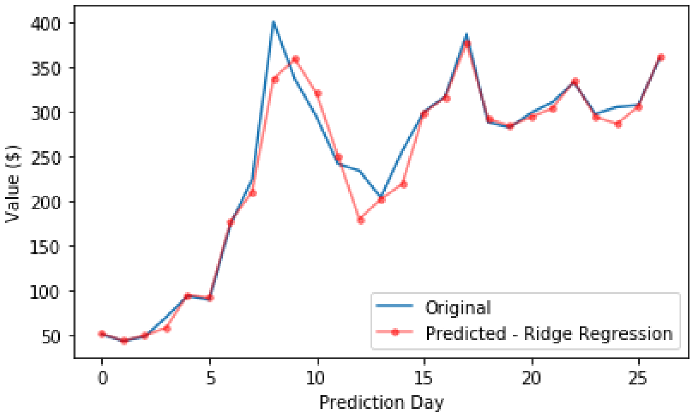
Look Back Period	Bitcoin	Ethereum	Litecoin	Zcash	∑ RMSE's
(Days)	($)	($)	($)	($)	($)
1	136.69	15.69	2.49	42.34	197.21
2	148.08	15.19	2.8	47.94	214.01
3	137.82	16.75	2.11	17.72	174.4
4	163.5	14.69	2.96	20.7	201.85
5	153.61	24.9	2.92	18.14	199.57
6	152.48	16.6	2.31	14.62	186.01
7	146.59	25.74	2.08	23.5	197.91
8	197.53	30.3	4.33	22.41	254.57
9	**120.29**	**19.84**	**2.07**	**18.11**	**160.31**
10	172.08	19.84	3.74	19.66	215.32
20	225.04	30.59	4.08	25.95	285.65
30	150.16	19.49	2.73	31.87	204.25

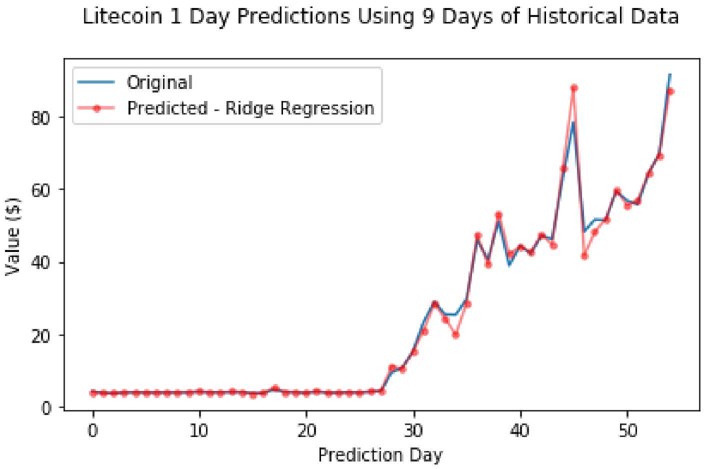
Once again the optimized model was used to generate value predictions for each of the four cryptocurrencies selected. Predictions were made using a holdout or a test set of data that was composed of 30 percent split (i.e., 70 percent of data was dedicated for training and 30 percent for testing). Below are four plots that were generated to show the model's predictions against the actual values of each cryptocurrency.

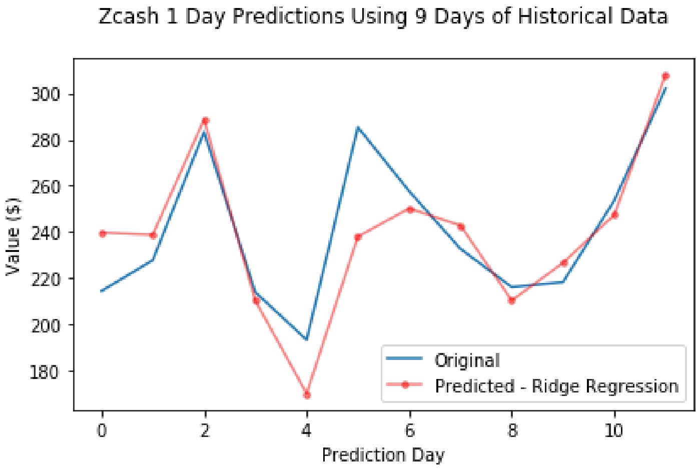
Bitcoin 1 Day Predictions Using 9 Days of Historical Data



Ethereum 1 Day Predictions Using 9 Days of Historical Data




Litecoin 1 Day Predictions Using 9 Days of Historical Data


Zcash 1 Day Predictions Using 9 Days of Historical Data

The optimal Ridge Regression provides a slight improvement over the previously developed Linear Regression model. This slight improvement is shown in the lowering of the RMSE by $10, and otherwise would be very difficult to discern just by looking at the prediction profiles.

Trading Strategy Implementation with Each Model

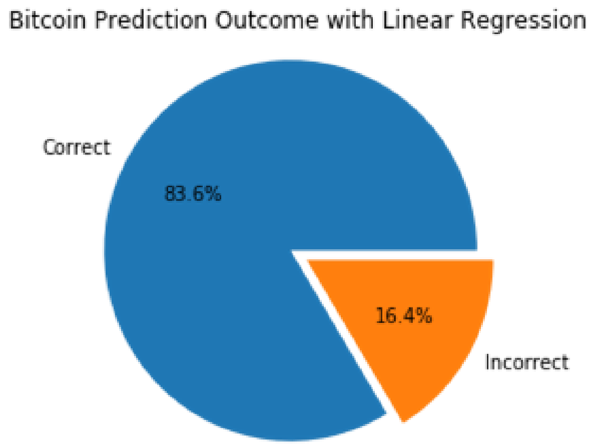
Taking this analysis a step further, a long-term investing strategy was investigated to see how accurately the models could guide us on a day to day basis. Because we are looking at a long-term investing strategy our trading actions will only encompass a "Buy" or a "Hold". To determine the appropriate trading actions, the difference in price over consecutive days will be looked at. The table below summarizes when a particular trading action will be triggered.

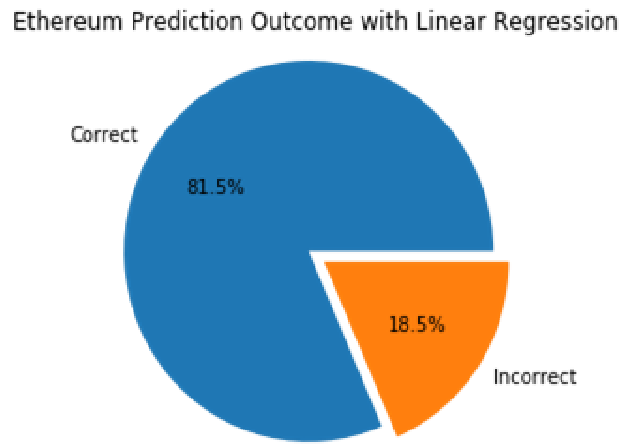
Trading Action Summary

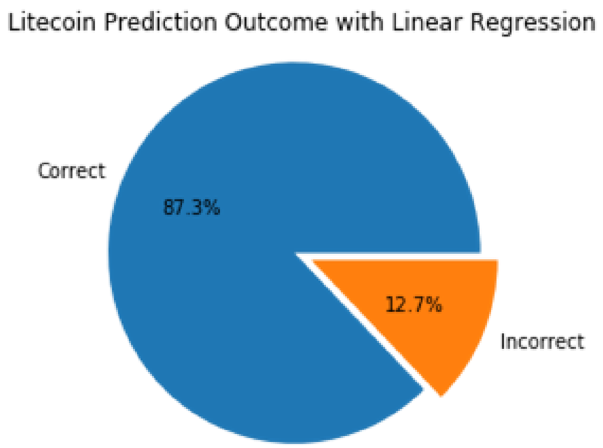
Comparison of Consecutive Days	Action
Negative (Price Decreasing)	Buy
Positive (Price Increasing)	Hold

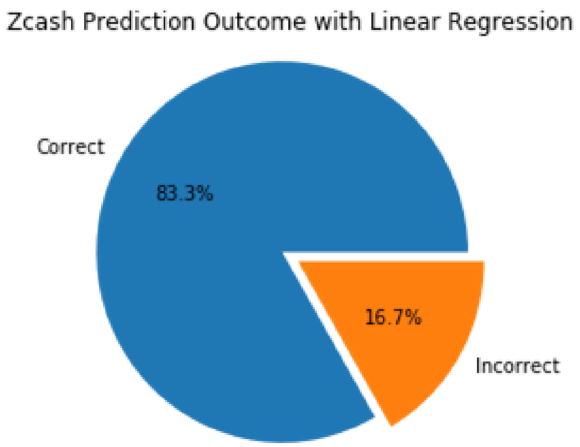
Linear Regression Trading Strategy Assessment

Using the previously developed Linear Regression model, and its predictions, we were able to assess the performance we would get from implementing the suggested long-term trading strategy. Below are four pie charts summarizing the number of times, as a fraction, the model would have guided you to make the 'correct' move versus the 'incorrect' move. A 'correct' move would consist of 'buying' the cryptocurrency when its value was decreasing or 'holding' it when its price was increasing.





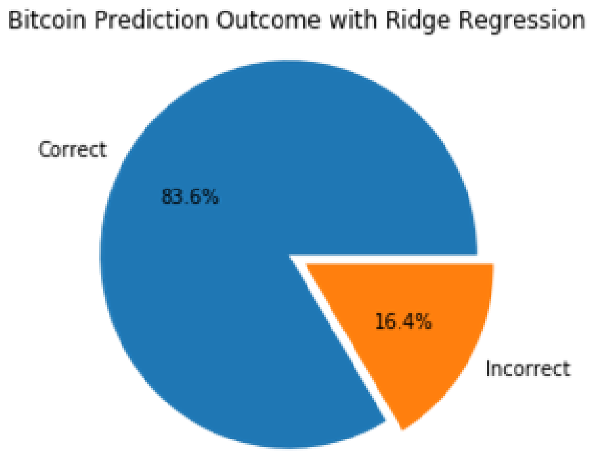


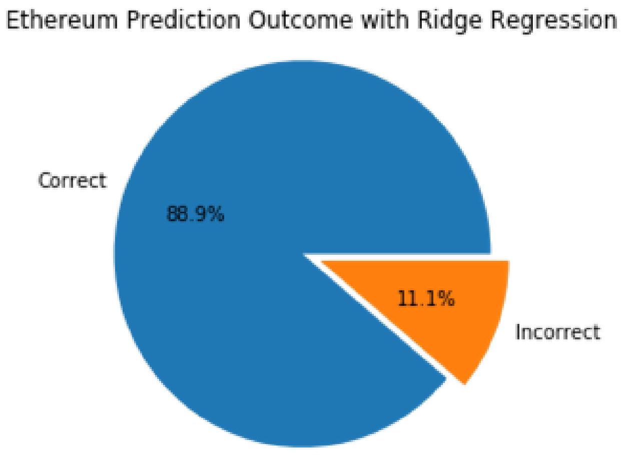


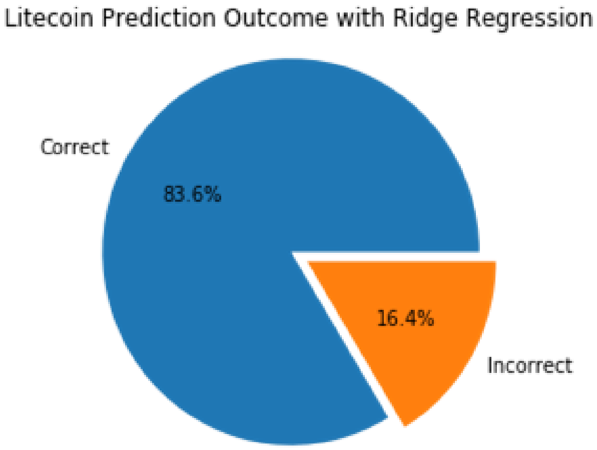
Looking at the pie chart assessments for each of the cryptocurrencies we can see that on average the optimized Linear Regression model would guide us to make the 'correct' move approximately 84 percent of the time.

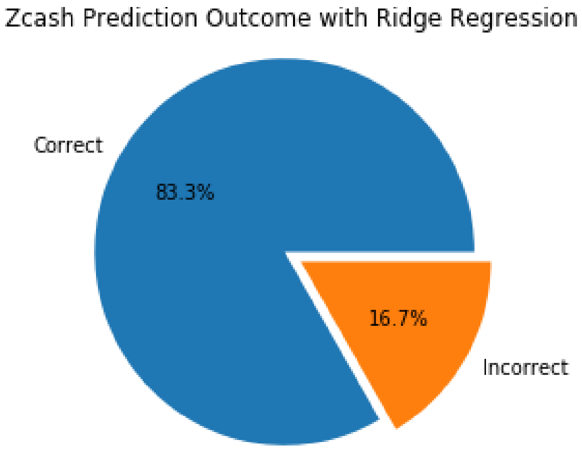
Ridge Regression Trading Strategy Assessment (Optimal)

Using the optimal Ridge Regression model to guide us in our trading moves, right off the bat, we would expect to get slightly better performance due to the $10 RMSE improvement that was gained over the optimal Linear Regression model. Below are four pie charts that can be used to quantify and assess the trading performance yielded by the optimized Ridge Regression model.









Looking at the pie chart assessments for each of the cryptocurrencies we can see that on average the optimized Ridge Regression model would guide us to make the 'correct' move approximately 85 percent of the time. This gives the Ridge Regression model approximately a 1% advantage in trading prediction power, validating our earlier intuition.

Takeaways

After analyzing four different top contending cryptocurrencies (Bitcoin, Ethereum, Litecoin, and Zcash) it is clear that a strong correlation is present amongst all four of them. How strong you ask? According to the Pearson correlation, all four of them have an average correlation of .86 from the year 2017 onward. In comparing the performance of both machine learning algorithms implemented: Linear Regression and Ridge Regression both showed to have some serious prediction power, but Ridge Regression proved to earn itself ultimate bragging rights. The optimal Ridge Regression model gathered nine days of historical data to make single day predictions that varied no more than outlined in the table below.

Optimal Prediction Model Summary

Cryptocurrency	Average Price Variance
Bitcoin	$120
Ethereum	$20
Litecoin	$2
Zcash	$18

Running a long-term trading assessment using Ridge Regression allowed it to boast its bragging rights even further, delivering you the right moves, on average, approximately 85% of the time across a portfolio holding all four cryptocurrencies analyzed. **Just remember that this is educational material and it should not be used as professional investment advice.**