Predicting the South African Bitcoin Market using International Market Data

##1. Introduction This paper examines the relationship between the international Bitcoin market and the South African Bitcoin market. It is hypothesized that the South African market slightly (at a one-minute resolution) lags the international market in terms of price movement. If this is the case and the relationship can be modelled it could lead to improvements in algorithmic trading strategies for Bitcoin traders. It was found that there is a relationship between the international price change, although it is difficult to model to a degree of accuracy that would be useful to algorithmic traders.

This paper used data gathered from Kraken - an American cryptocurrency exchange - to represent the international Bitcoin market and from VALR - a South African cryptocurrency exchange - to represent the South African market. In specific, the per minute pricing history for the rand and dollar markets for Bitcoin were gathered from the two exchanges. Next the international markets were compared to the local markets to test for correlation and finally the relationships were modelled using multiple linear regressions, gradient boosting and neural network algorithms available from SciKit-Learn.

The paper begins by describing the data collection process before detailing the methods of feature extraction used. Next, descriptive statistics are presented and their implications explained, following which further data manipulation is undertaken in order to account for the idiosyncracies of the specific dataset. Next, the methods and results of the modelling process are presented and finally the results are discussed in the conclusion.

##2. Data Collection and Cleaning Data collection was done via email (and WeTransfer) on the VALR side and via download on the Kraken side. The VALR data was delivered as per minute pricing data describing the opening-price, high-price, low-price, close-price and volume per minute. This is known as the OHLC format. No initial cleaning was needed for the VALR data. The VALR data was adjusted from the Pretoria, South Africa timezone to the UTC timezone in order to match the timezone of the Kraken data.

Kraken provides pricing history per minute for the Bitcoin-USD (BTC-USD) market in this Google Drive. Because the aim of this project is to use the international market for Bitcoin and Ethereum to predict the South African market in order to augment algorithmic trading decisions - the per minute OHCL data for both markets was differenced by its first lag. In particular, the closing prices were differenced and converted to a percentage. One of the major benefits of differencing the data is that it centers it around zero. Please see Figures 1 and 2 for a comparison of the differenced and un-differenced BTC-USD closing prices over time. Further, differencing (by percentage) the data brings the scale of the two markets together. Because of the ZAR/USD exchange rate (around R15 per USD over the time period) the ZAR-BTC market has nominal values around 15 times higher than the USD-BTC market. This is problematic if both ZAR lags and USD lags are included in the variables input into models that work with node weightings because it can bias the model towards variables with bigger scales (reference here). This problem is mitigated by the scaling inherent in differencing by percentage. After differencing, the first 5 lags of the closing prices of both the ZAR and USD markets were taken. Similarly, a 100 period moving average and a 5 period moving average were taken for both markets. Note that the last inclusion in the moving averages is the first lag of the closing price.

Finally, the differenced ZAR-BTC and USD-BTC variables were divided into 8 categories dependent on the number of standard deviations that a sample lies away from the mean. Please see Equation 1 for details.

Equation 1: Piecewise categorization of variables by number of standard deviations from the mean

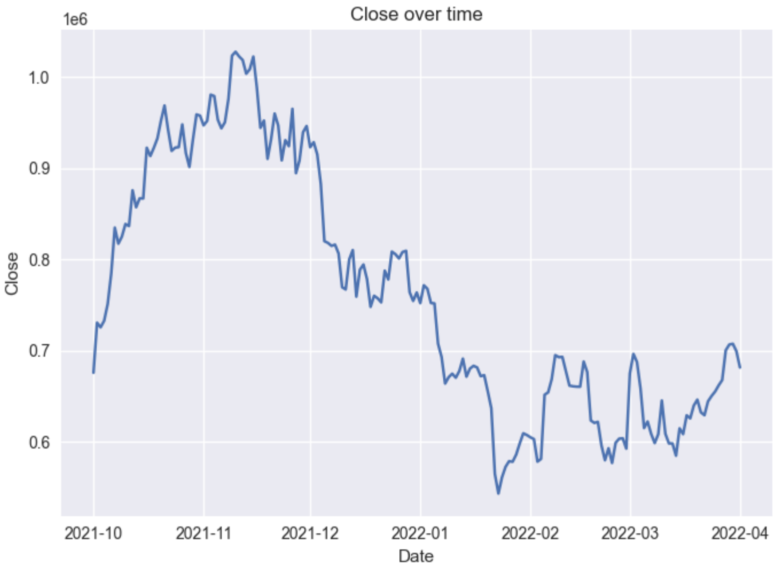

Figure 1: BTC-ZAR over Time

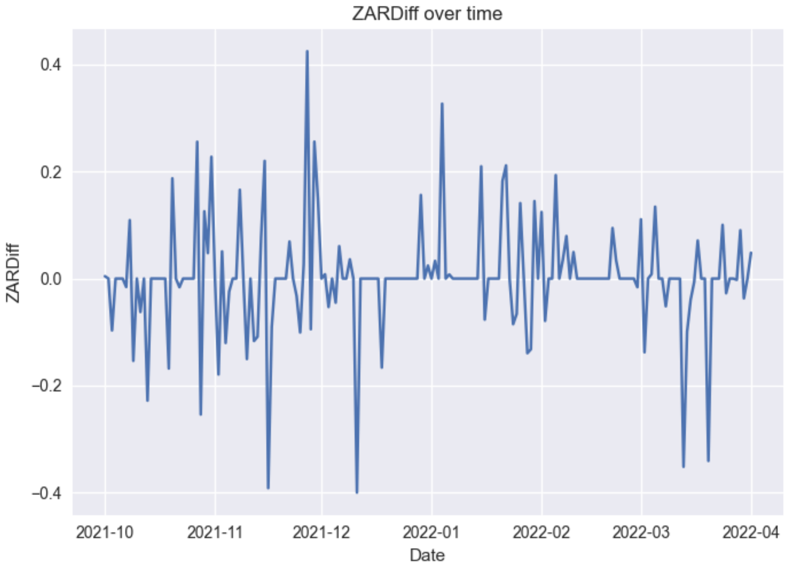

Figure 2: Differenced BTC-ZAR over Time

##3. Initial analysis and variable selection Single linear (OLS) regression was used to test the hypothesis in the most basic way and yielded a positive result. Regressing the un-lagged BTC-ZAR variable on the un-lagged BTC-USD variable reveals a strong positive relationship between the two - please see Figure 3. Next, regressing the un-lagged BTC-ZAR variable on the first lag of the BTC-USD market yields a weaker but comparable positive relationship - please see Figure 4. However, regressing the un-lagged BTC-USD variable on the first lag of the BTC-USD market reveals a very weak positive relationship - please see Figure 5. In fact, even the second lag of the BTC-USD variable is a better predictor of the un-lagged BTC-ZAR market than the first lag of the BTC-USD variable is for the BTC-USD market. Please see Table 1 for coefficients and R-Squared values.

Regression	Coefficient	R-Squared
BTC-ZARDiff on BTC-USDDiff	0,5	0,17
BTC-ZARDiff on BTC-USDDiff_1	0,26	0,05
BTC-USDDiff on BTC-USDDiff_1	0,03	0,0008

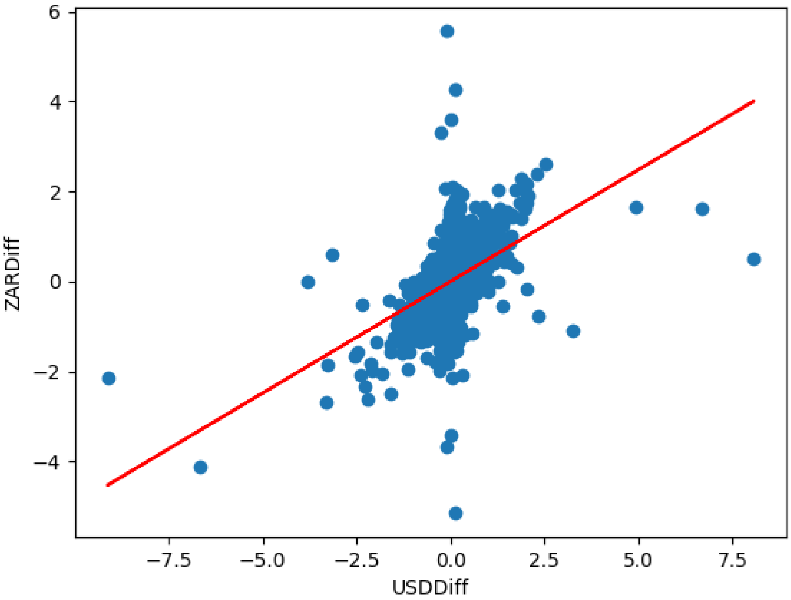

Figure 3: Un-lagged BTC-ZAR vs un-lagged BTC-USD
R-squared: 0.17 , Coefficient:0.5

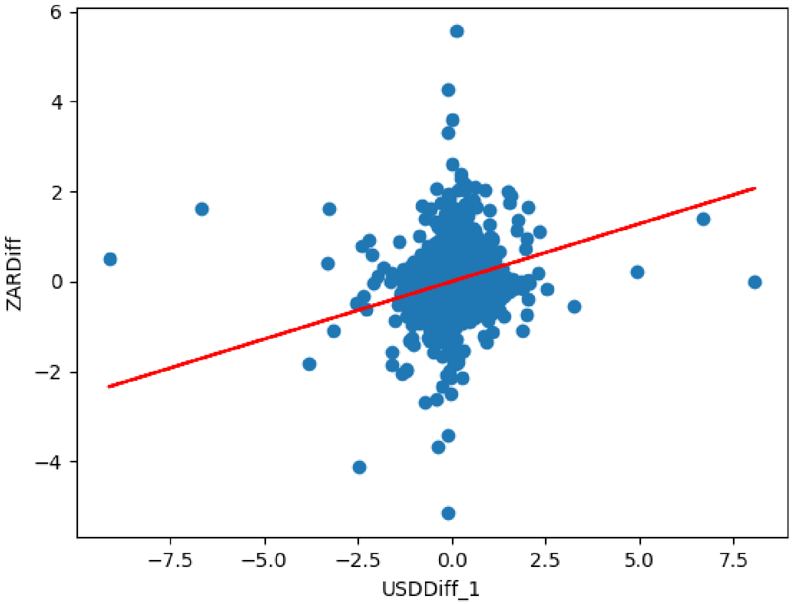

Figure 4: First lag of BTC-ZAR vs un-lagged BTC-USD
R-squared: 0.05 , Coefficient:0.27

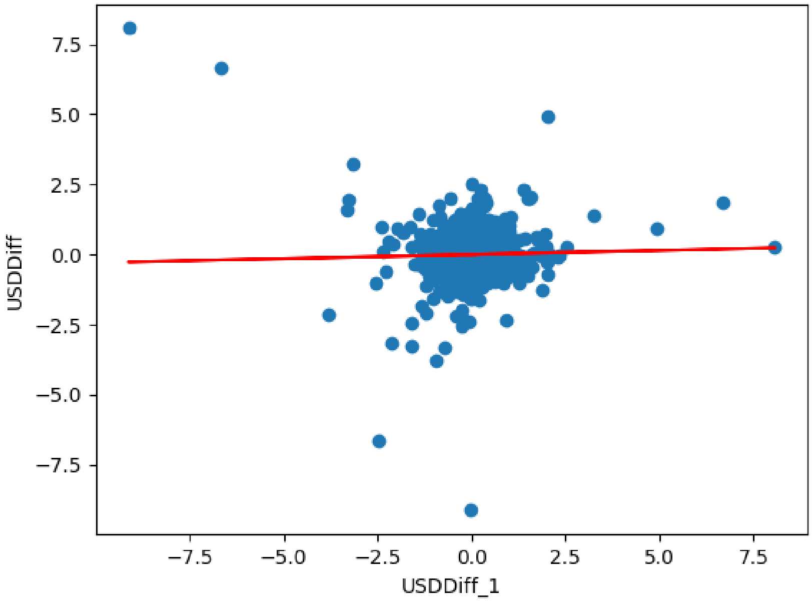

Figure 5: First lag of BTC-USD vs un-lagged BTC-USD
R-squared: 0.0008 , Coefficient:0.03

Running an elastic net regression with a lambda of 0.05 indicated that eight variables are prominent in the prediction of the un-lagged BTC-ZAR variable. Please see Figure 6. A second elastic net regression, including only the eight variables yielded by the first elastic net regression, with a lambda value of 0.01 indicates the five most prominent variables to be USDiff_1, ZARDiff_1, ZARVolume_1, USDDiff_2 and ZARDiff_2; in that order. Please see Figure 7.

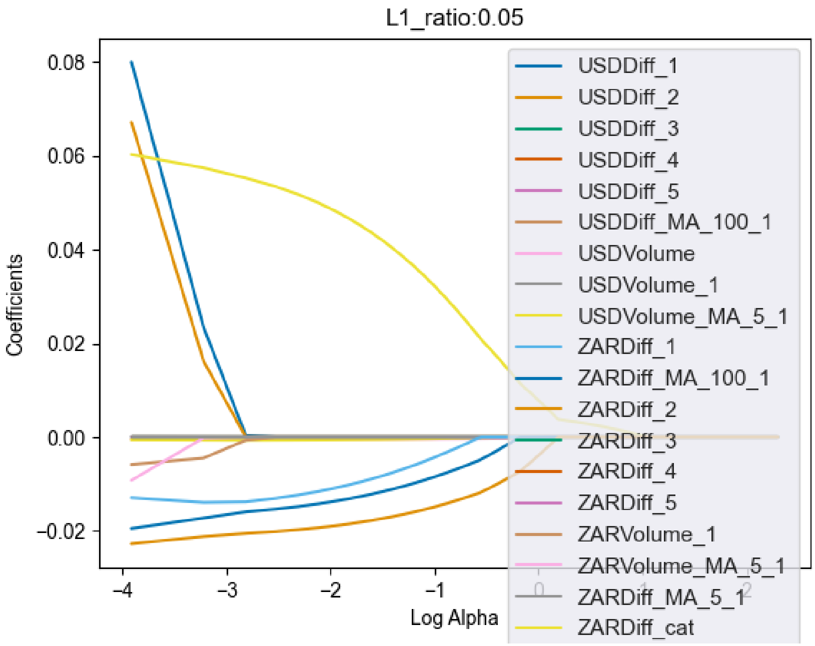

Figure 6: Elastic net regression 1

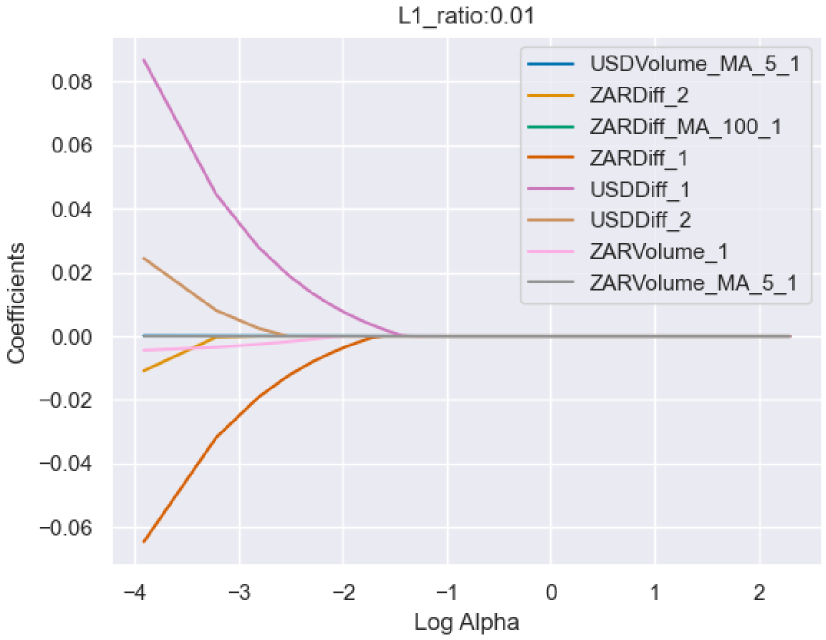

Figure 7: Elastic net regression 2

##4. Machine Learning Analysis 1 and corrections Both the five variable set and the eight variable set were tested across 3 classification algorithms and 3 regression algorithms. The classification algorithms are Logistic Regression, Gradient Boosting and a Neural Network while the regression algorithms are Linear Regression, Gradient Boosting and a Neural Network. The data is arranged into training and test sets using an iterative 5 stage time series approach. The data is split into 5 consecutive time periods of even length. Each algorithm is trained and tested sequentially on the first four time periods and then only tested on the final (fifth) time period. This forces the algorithms to perform prediction instead of interpolation. The results of these tests are displayed in Tables 2,3,4 and 5.

	Logistic Regression	Gradient Boosting	Neural Network
Train Accuracy	0.61	0.71	0.71
Train Precision	0.41	0.76	0.58
Train Recall	0.6	0.098	0.25
Test Accuracy	0.59	0.68	0.68
Test Precision	0.41	0.63	0.56
Test Recall	0.54	0.05	0.18

Table 2: Five variable classification scores

	Logistic Regression	Gradient Boosting	Neural Network
Train Accuracy	0.64	0.71	0.72
Train Precision	0.43	0.78	0.59
Train Recall	0.46	0.1	0.27
Test Accuracy	0.62	0.68	0.68
Test Precision	0.41	0.65	0.55
Test Recall	0.38	0.05	0.24

Table 3: Eight variable classification scores

	Linear Regression	Gradient Boosting	Neural Network
Train MSE	0.013	0.014	0.013
Train MAE	0.066	0.065	0.067
Test MSE	0.008	0.008	0.008
Test MAE	0.051	0.051	0.051

Table 4: Five variable regression scores
Standard error of target variable (ZARDiff) is 0.117

	Linear Regression	Gradient Boosting	Neural Network
Train MSE	0.014	0.014	0.014
Train MAE	0.067	0.065	0.067
Test MSE	0.008	0.008	0.008
Test MAE	0.051	0.051	0.053

Table 5: Eight variable regression scores
Standard error of target variable (ZARDiff) is 0.117

In the classification task no algorithm stands out as superior across all three tasks. What one algorithm gains in terms of accuracy over another it loses in precision or recall. This holds true across the training and test scores and the five and eight variable datasets. In the regression task there is extreme uniformity within the test and training scores across both datasets. Together this indicates that the scores achieved in this round of ML may be the result of a specific distribution of the target variable rather than well-trained algorithms.

Investigation into the distribution of the categorical target variable indicates that this is true. 52% of the categorical data lies in bin 4 (within half a standard deviation below the mean) and a further 16% lies in bin 5 (within half a standard deviation above the mean). Please see Figure 8. Thus, 68% of the data lies within half a standard deviation of the mean. For reference, the mean of the series is 0.000125 and the standard deviation is 0.117. The histogram of the continuous version of the ZAR-BTC variable indicates that a similar distribution exists in that series - see Figure 9.

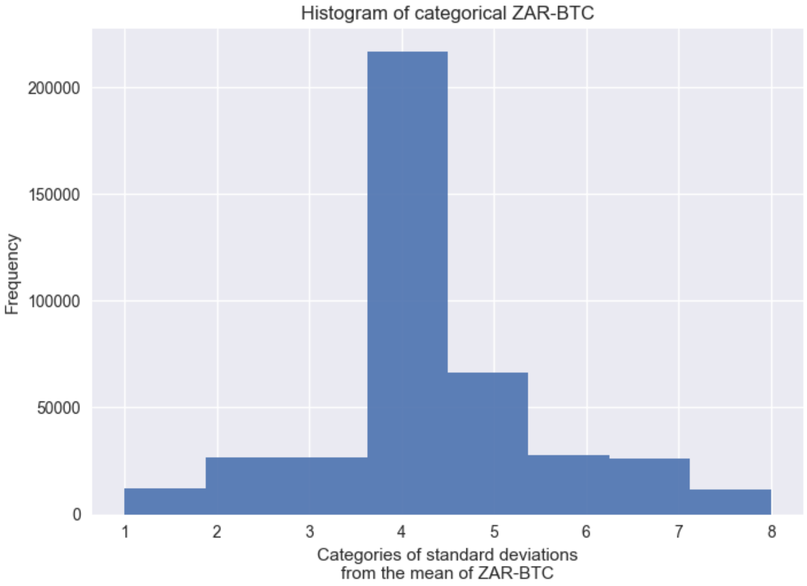

Figure 8: Histogram of the categorical target variable

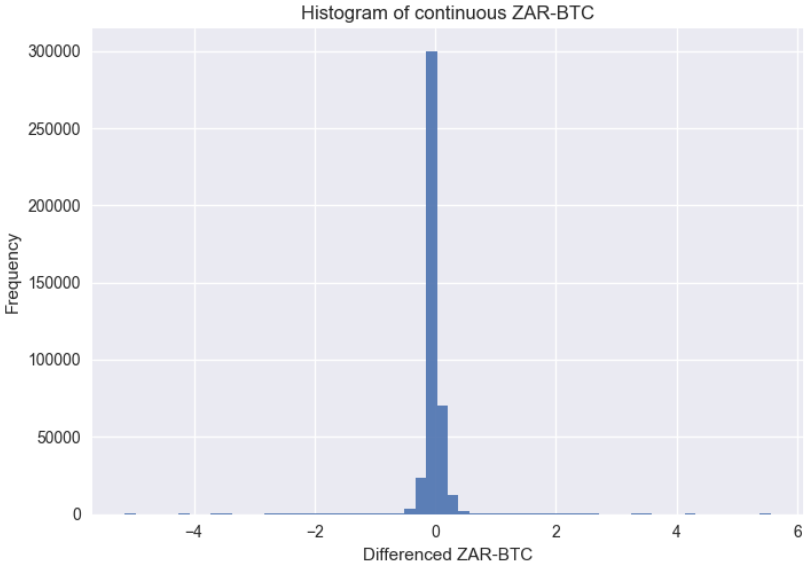

Figure 9: Histogram of the continuous target variable

Category 4 indicates a change in the ZAR-BTC price of between -0,058% and +0,000125% while category 5 indicates a change of between +0,000125% and +0,059%. Trading fees for market takers on VALR are 0.1% (reference). Thus, any trades made within categories 4 and 5 will lose money and are not worth attempting. However, it may be useful to know whether the price is likely to increase or decrease, even within the 0.1% range. Re-categorizing the ZAR-BTC variable into 4 categories - according to Equation 2 - allows for this type of analysis. Over the 28 months analyzed here there were a total of 89306 occasions where the price of Bitcoin changed by more than 0.1% in a minute (about 22% of the samples split almost evenly between increases and decreases) for an average of 3189.5 times per month, or ~101 times per day. However, 56% of the data lies in category 2 and 22% in category 3 and thus prediction models trained on the data are still vulnerable to bias due to the distribution; although, it may be to a lesser extent because the outlying fields contain a greater proportion of the data. This is revealed to be true in the next section.

Equation 2: Piecewise categorization of variables by differences absolutely larger than 0.1%

##5. Machine Learning Analysis 2 and conclusions Running the same sets of five and eight variables through the same classification algorithms as above with the new four category BTC-ZAR variable as the target variable yields results that are similar to the results reported in the Machine Learning Analysis 1 section. They are not depicted here. However, removing the middle section of the data - categories 2 and 3 - yields significantly better results. These are presented in Tables 5 and 6. These results show that increases of more than 0.1% in the South African Bitcoin price are algorithmically separable from decreases of more than 0.1% based only on lags of the South African and international markets. If a model that could differentiate between categories 2 and 3, and 1 and 4 were available - the models presented here could be extremely useful to an algorithmic Bitcoin trader. This is recommended for further research.

	Logistic Regression	Gradient Boosting	Neural Network
Train Accuracy	0.56	0.58	0.56
Train Precision	0.44	0.45	0.43
Train Recall	0.77	0.82	0.77
Test Accuracy	0.75	0.75	0.75
Test Precision	0.75	0.75	0.78
Test Recall	0.76	0.75	0.71

Table 5: Five variable classification scores for no-middle dataset

	Logistic Regression	Gradient Boosting	Neural Network
Train Accuracy	0.56	0.58	0.56
Train Precision	0.44	0.46	0.44
Train Recall	0.78	0.83	0.78
Test Accuracy	0.75	0.75	0.75
Test Precision	0.75	0.75	0.76
Test Recall	0.77	0.75	0.75

Table 6: Eight variable classification scores for no-middle dataset