A comparison between the latest Prophet forecasting model and the ARIMA model: An evaluation of BTC/ZAR predictability

27th October 2017

Fiona Ganie (GNXFIO001)

GNXFIO001@myuct.ac.za

Abstract

The Autoregressive Integrated Moving Average (ARIMA) model is one of the most widely used time series models that has gained popularity in the exchange rate market due to its ease of implementation and tractability. With the evolution of computational power, soft computing techniques have since been used in exchange rate markets. The ARIMA model has been used as the standard benchmark model against which these complex methods have been compared. Although they have produced significantly better results than the ARIMA model, these models lack interpretability and building them is a challenging task. Prophet is a sophisticated model that differs from traditional time series models in that it can be utilised by non-experts who have little knowledge about the statistical intricacies involved in the model, however have domain knowledge about the data generating process. Prophet is also robust to outliers and can handle missing values in the time series without the need for interpolation. This paper compares the out-of-sample forecasts of Prophet and ARIMA over varying forecast horizons, using a Mincer-Zarnowitz approach for forecast evaluation. A complete dataset of the prices of Bitcoin/Rand as well as a dataset which has missing prices and outliers present is used

in the analysis. It is found that the ARIMA model produces the most efficient and unbiased forecasts over all forecast horizons, and that the ARIMA model is more robust when faced with missing values and outliers present in the data.
Keywords: Prophet, ARIMA, Mincer-Zarnowitz, Box-Jenkins, Bitcoin, Model Confidence Set

1 Introduction

The Autoregressive Integrated Moving Average (ARIMA) model is one of the most widely used time series models that have attracted attention in financial market forecasting (Khashei, Bijari, and Ardali 2009). Although the Random Walk model has typically been applied to foreign exchange markets and has produced superior results, some researchers have contended that foreign exchange markets are not efficient and believe that future prices depend on current and past events (Abu-Mostafa and Atiya 1996). This has led to the application of the ARIMA model to exchange rate problems, where it has since gained popularity due to its ease of implementation and tractability. The ARIMA model is easy to use and yields good forecasts over short forecast horizons, however the linearity of the ARIMA model fails to adequately capture the non-linearity inherent in exchange rate data (Zhang and Hu 1998).

With the evolution of computational power, non-linear, soft computing techniques have been proposed as a solution. The ARIMA model has been used as the standard benchmark model against which these more complex methods have been compared. Although they have produced significantly better results than the ARIMA model, these models lack interpretability and building them is a challenging task. The ARIMA model is tractable and less computationally expensive. It has been used as the building block for more advanced models and has provided the inspiration for hybrid versions of the model which have been used in exchange rate forecasting.

In December 2016, Facebook open-sourced their forecasting model Prophet. Prophet differs from traditional time series model such as the ARIMA in that it can produce high quality forecasts in a straightforward way. Prophet is a sophisticated model that provides informative results, however is config-

urable and easy to use. It can be utilised by non-experts who have little knowledge about the statistical intricacies involved in the model, however have domain knowledge about the data generating process. This knowledge can easily be incorporated into the model through its intuitively adjustable parameters (Taylor and Letham 2017). Furthermore, Prophet is robust to outliers and can handle missing values in the time series without the need for interpolation (Taylor and Letham 2017). This allows analysts without knowledge on how to pre-process the data, to utilise the model without sacrificing predictive accuracy. If Prophet produces forecasts that are as good as the ARIMA model, it can be compared to more complex forecasting methods that are less tractable and flexible.

This paper broadly aims to compare the forecasts produced by the traditional ARIMA model and Prophet through an evaluation of the exchange rate between Bitcoin and the Rand (BTC/ZAR), using the Model Confidence Set and a Mincer-Zarnowitz approach of measuring forecast accuracy. More specifically, this paper aims at:

1. Comparing the out-of-sample forecasts of ARIMA and Prophet over varying forecast horizons.

2. Comparing the out-of-sample forecast performance with missing values and outliers present in the data.

Section 2 will present the key findings from the major works in which comparisons have been made to the ARIMA model in exchange rate forecasting. Section 3 will provide a brief explanation and motivation for the dataset used in this paper. Section 4 explains and compares the formulation of the ARIMA and Prophet model. Section 5 will describe the methodology used in selecting the ARIMA and Prophet models, the Model Confidence Set and the

Mincer-Zarnowitz approach of measuring forecast accuracy. The results from applying the ARIMA and Prophet model to the dataset is presented in Section 6. Finally, section 7 discusses the results and its financial implications in a statistical context.

2 Background

2.1 Forecasting Exchange Rates with the ARIMA model

ARIMA models have commonly been used in financial forecasting and are popular for observing stock prices and exchange rates due to its power and statistical properties (C.-S. Lin, Chiu, and Lin 2012). The ARIMA model is attractive as it is tractable and produces good short-term forecasts when more than 100 observations are used (Tseng et al. 2001). They have frequently been used as a benchmark to compare new forecasting techniques that have emerged over time and have yielded satisfactory results when predicting exchange rates. Nwankwo (2014) forecasted the exchange rate between the Naira and dollar, using Akaike's Information Criterion (AIC) as a measure of performance. Diagnostic testing revealed that the ARIMA(1,0,0) model was the best fit for the data.

Although the ARIMA model has the advantage of ease of implementation and flexibility, it fails to capture the non-linearity and volatility present in exchange rate data. Over time, the ARIMA model has evolved to cater for a wider variety of data and to compensate for some of its shortcomings. Some of the most popular versions of the ARIMA model that has been implemented in exchange rate forecasting is the Seasonal ARIMA (SARIMA) and Fractional ARIMA (ARFIMA) model.

The SARIMA model was introduced to capture the periodic behaviour of

data and extends the ARIMA class by including a term for seasonal differencing. Etuk, Wokoma, and Moffat (2013) modelled the Naira/CFA Franc exchange rate which exhibited monthly seasonality using an additive SARIMA model, to demonstrate that it can be a useful fit for exchange rate data which displays seasonality. Their results showed that the SARIMA model adequately described the variation in the exchange rate series.

The ARFIMA model generalises the ARIMA model in that the degree required to make the data stationary can assume any real value, and is no longer restricted to the integer domain. The ARFIMA model has the ability to capture the dependence between observations that are widely spread apart in time (Cheung 1993). This makes the model parsimonious since it can capture long memory in data as well as short term dynamics (Cheung 1993). Cheung (1993) fitted the ARFIMA model to examine five exchange rates and found that there was strong evidence of long memory in the exchange rate time series.

2.2 Other techniques used to forecast exchange rates

Generalised autoregressive conditional heteroskedasticity (GARCH) models were later developed due to the failure of ARIMA models to capture the volatility in financial markets (Anastasakis and Mort 2009). Fahimifard et al. (2009) applied a GARCH model to the Rial/USD and Rial/EUR to compare the forecasts yielded by the GARCH and ARIMA model over varying forecast horizons. Their results illustrated that the GARCH model outperformed the ARIMA model when using the Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Deviation (MAD) as a performance criteria.

Over time, financial forecasting methods have moved away from linear

models like ARIMA and GARCH, to soft computing techniques. These complex techniques are non-linear and can fit complex time series more easily (Castillo and Melin 2002). Unlike the ARIMA model, soft computing techniques do not impose structural assumptions on the model apriori (Castillo and Melin 2002). Some of the most commonly used artificial intelligence methods used to forecast exchange rate data are Neural Networks and Fuzzy Logistic Systems.

Artificial Neural Networks (ANNs) mimic the structure of the brain and have had many successful applications in forecasting exchange rates. They are data driven, can adapt to non-stationary environments and can approximate any continuous function (Khashei and Bijari 2011). In a study done by Fahimifard et al. (2009), the ANN was found as an effective way to improve the forecasts of exchange rates. Superior results were produced when compared to the ARIMA and GARCH model using the RMSE, MSE and MAD as a measure of performance.

Fuzzy Logistic Systems (FLSs) were initially developed to solve problems involving linguistic terms and have successfully been used in financial forecasting (Khashei, Bijari, and Ardali 2009). Fuzzy logic tries to imitate human reasoning and the decision-making process and allows for finer rather than discrete decisions to be provided. Santos, Costa, and Santos Coelho (2007) investigated how well FLSs and ANNs perform compared to the traditional ARMA and GARCH model. He examined the forecasts of Brazilian exchange rate returns by considering different frequencies of the series and comparing their one-step ahead forecasts. By analysing accuracy statistics, he found that FLSs and ANNs achieved higher returns based on the forecasts they produced. Similar results were found by Khashei, Bijari, and Ardali (2009) when he analysed the predictive capabilities of FLSs, ANNs, the traditional

ARIMA model and a Fuzzy ARIMA model.

Although ANNs have been broadly applied in financial forecasting, the process of building them is a complex task and there is no consistent method of design compared to the traditional Box-Jenkins ARIMA model. The performance of ANNs is sensitive to many modelling factors such as the number of input nodes included and the size of the training sample chosen (Zhang and Hu 1998). Like ANNs, there is no systematic approach for designing FLSs and they are only understandable when simple. Although FLSs has the advantage over ARIMA models in that they can be applied to data with few observations available, it yields acceptable rather than accurate results and are more suitable for problems which do not require high accuracy.

The traditional ARIMA model produces less superior forecasts than its hybrid forms and other complex non-linear techniques, however its forecasts are still satisfactory. It is a simple model that is easy to implement and has a consistent method of model design and selection. ARIMA models are more robust and efficient than complex structural models in relation to short-run forecasting. The fact that they have been used as the foundation for more advanced models and have commonly been used as a benchmark for comparison, justifies it as a good starting point to compare it to Facebook's forecasting method, Prophet, that was recently released.

2.3 Forecasting with Prophet

The techniques that have been considered for exchange rate forecasting thus far, require the analyst to have vocational knowledge about time series. Prophet differs from traditional time series models in that it is flexible and can be customised by a large number of non-experts who have little knowledge about time series, however have domain knowledge about the data

generating process. Prophet allows for a large number of forecasts to be produced across a variety of problems and consists of a robust evaluation system that allows for a large number of forecasts be evaluated and compared. This is Facebook's definition of forecasting at scale.

Prophet's Bayesian approach to forecasting allows the analyst to incorporate their expert knowledge into the model building process and has produced significantly improved forecasts compared to the ARIMA model. Taylor and Letham (2017) forecasted the number of events on Facebook using Prophet. The time series was impacted by holidays, had strong multi-period seasonality, and a piecewise trend. The forecasts produced by Prophet were compared to common forecasting techniques such as exponential smoothing, ARIMA, the seasonal naive, and the naive model. While exponential smoothing and the seasonal naive model were quite robust, the ARIMA forecasts were fragile. No model besides Prophet accounted for the dips around holidays and the upward trend of the time series towards later observations. If Prophet produces forecasts that are as good as the ARIMA model when forecasting BTC/ZAR, it can be compared to more complex forecasting methods that are less tractable such as the hybrid models and machine learning techniques seen earlier.

3 Data

The dataset used in this study consists of the daily closing prices of BTC/ZAR over the period 24 January 2016 to 17 July 2017. This comprises of a total of 541 trading days and was the chosen time period for analysis due to constraints in obtaining data over a longer period. The data was obtained from Bitcoincharts.

Bitcoin is of specific interest as it presents an interesting parallel to traditional exchange rate markets. The cryptocurrency is built on a decentralised system and as a result its value cannot be directly influenced by a central authority (Fantazzini et al. 2016). In addition, the Bitcoin market has attracted attention worldwide and is currently the leading cryptocurrency, with awareness and adoption of the currency growing over time (Fantazzini et al. 2016). Its novelty makes it a highly volatile hence speculative market and provides an opportunity for forecasting.

In this study, the analysis of BTC/ZAR forecasts are based on the daily continuous log returns. The daily closing prices are transformed into returns by taking the log difference at each time t and is calculated as follows:

$$r_t = ln(\frac{p_t}{p_{t-1}})$$

where p_t represents the closing price and r_t represents the log return for time t = 1,2,.,T.

4 Forecasting Methods

4.1 The ARIMA model

The ARIMA model expresses the process $\{y_t\}$ as a function of the weighted average of past values of the process and lagged values of the residuals. The weighted average of the past p values of the process represents an autoregressive (AR) process of order p. It feeds back past values of the process into the current value, inducing correlation between all lags of the process. The weighted average of the q lagged residuals represents a moving average (MA) process of order q. The purpose of mixing the MA process with the AR process is to reduce the large number of past values required by AR processes

and to control for the autocorrelation which it creates between lagged values of the process. The combination of the AR(p) and MA(q) process results in a more parsimonious model, and forms a stationary autoregressive moving average (ARMA(p,q)) process defined as:

$$y_t = c + \sum_{i=1}^{p} \theta_i y_{t-i} + \sum_{i=1}^{q} \phi_i e_{t-i} + e_t$$

where c is a constant and $\{e_t\}$ is a white noise process with zero mean and variance σ^2.

The ARIMA(p,d,q) model generalises the ARMA model in that it includes both stationary and non-stationary processes. The parameter d is the degree of differencing required to render the process stationary. If d is equal to zero the process is stationary and equivalent to an ARMA model, and if d is strictly positive the process requires differencing to make it stationary. The ARIMA model can be defined succinctly using the backward shift operator B, which shifts the process back by one unit of time, and is defined as $By_t = y_{t-1}$. The ARIMA model has the form (Hyndman and Athanasopoulos 2014):

$$(1 - \sum_{i=1}^{p} \theta_i B^i)(1 - B)^d y_t = c + (1 + \sum_{i=1}^{q} \phi_i B^i) e_t$$

where c is a constant and $\{e_t\}$ is a white noise process with zero mean and variance σ^2.

4.2 The Prophet model

Prophet is similar to a Generalized Additive Model (GAM) - an additive regression model that consists of non-linear and linear regression functions applied to predictor variables (Taylor and Letham 2017). The decomposable

model is of the form:

$$y(t) = g(t) + s(t) + h(t) + e_t$$

where the components of the model represent the growth, seasonality and holiday respectively, and e_t is white noise.

Prophet, like the GAM, frames the forecasting problem as a curve fitting exercise and uses backfitting to find the regression functions. This allows for the model to be fitted quickly and missing values and large outliers to be handled elegantly. The regression model also provides model flexibility and allows the analyst to interactively change model parameters (Taylor and Letham 2017). The growth component of Prophet may be modelled as a linear or non-linear function of time. Linear growth is modelled by a piecewise constant function while non-linear growth is modelled similar to population growths which use a logistic growth model (Taylor and Letham 2017). A time-varying upper limit may be specified for logistic growth, at which point the forecasts will saturate. This carrying capacity allows the analyst to incorporate their prior knowledge about the maximum obtainable growth level such as the total market or population size into the model. Prophet accounts for changes in the trajectory of this trend by automatically detecting and selecting changepoints in the data at which the growth rate is allowed to change. These changepoints have a Laplace prior distribution placed on them and its scale parameter may be used to adjust the flexibility of the trend and to choose how aggressively the model should follow historical trend changes. Analysts may also adjust the number of potential changepoints included or manually specify their location. This allows non-experts with knowledge about events that may affect the growth rate to use the parameter as a knob to either increase or decrease the number of changepoints included (Taylor

and Letham 2017). It also allows for the analyst to add changepoints which the automatic selection procedure may have missed or remove changepoints when the model is overfitting historical trends (Taylor and Letham 2017).

The decomposable form of the model allows for multiple seasonality components with different periods to be added to the model. Seasonality components are modelled by a Fourier series and have a Normal prior distribution placed on its parameters (Taylor and Letham 2017). The spread parameter can be adjusted by analysts to smooth the seasonality and change how much of historical seasonality is projected into the future (Taylor and Letham 2017).

The analyst may also provide a list of important events and holidays which have impacted the time series in the past or which they know might impact it in the future. The list could include the name, date, and country in which they have taken place or are expected to take place (Taylor and Letham 2017). By specifying the country of occurrence, separate lists can be populated for global events and holidays, and country-specific events and holidays. The union of the global and country-specific lists can then be used for forecasting. Like seasonality, a Normal prior distribution is placed on the parameters of the holiday component, and the scale parameter can be adjusted by analysts to smooth the holidays (Taylor and Letham 2017).

Prophet's ability to forecast at scale enables it to model a wide variety of data and may be able to adequately fit exchange rate data. Its non-linear components could capture the non-linearity present in exchange rates. In contrast, the ARIMA model is a linear function of previous observations and lagged residuals and fails to capture the non-linearity inherent in exchange rate data. Prophet performs well on data with strong multiple "human scale" seasonalities and historical trend changes. This differs to the ARIMA model

which requires the data to be de-trended and the variance stabilised before the model can be fitted. Hence Prophet could model the weekly seasonality of closing prices due to low trading activity which occurs around the weekend and high trading activity which occurs mid-week.

Choosing the correct combination of parameters for the ARIMA model is a challenging task due to the array of possible choices. Although the auto.arima function in R may be used to automatically select an ARIMA model that best fits the data, completely automatic forecasting methods are too brittle and do not allow for useful assumptions to be incorporated into the model. Prophet makes use of a semi-automatic forecasting technique that keeps the analyst-in-the-loop. Its default settings are said to generate forecasts that are as accurate as those produced by skilled forecasters. If the forecasts produced are unsatisfactory, they can be improved by the analyst by configuring the model through its easily interpretable parameters. Furthermore, non-experts who have domain knowledge about factors that affect Bitcoin or if the dates of events which could impact the price of Bitcoin are known, it may be incorporated into the model by the analyst. Prophet can produce forecasts over irregular time intervals and allows for missing values in the time series without the need for interpolation (Taylor and Letham 2017). The ARIMA model on the other hand requires large outliers to be removed and handles missing values by interpolation. If Prophet produces forecasts that are as good as the ARIMA model when forecasting BTC/ZAR, it can be compared to more complex forecasting methods that are less tractable such as the hybrid models and machine learning techniques seen earlier.

5 Methodology

In this paper, the out-of-sample forecasts of ARIMA and Prophet are compared over varying forecast horizons. The data is first split into a training set and test set. The training set starts on the 24th January 2016 and ends on the 16th January 2017, while the test set starts on the 17th January 2017 and ends on the 17th July 2017. The Box-Jenkins methodology is then used to select the correct ARIMA model while Prophet's semi-automatic procedure is used for model selection. The selected ARIMA and Prophet model are then fitted to the training set and rolling window forecasts are made 1 day, 30 days and 90 days ahead. This allows us to examine the forecast horizon effect. The Model Confidence Set and Mincer-Zarnowitz test is then used to evaluate the forecasts produced by ARIMA and Prophet against the test set. Since there is no consensus on which accuracy statistic best measures the performance of forecasting techniques, the most popular criteria, the Root Mean Squared Error (RMSE) is employed. The results obtained from the ARIMA model will be used as a benchmark for comparison. This process is then repeated based on data which has outliers and missing values present.

5.1 The Box-Jenkins Methodology

Box, Jenkins, and Reinsel (1970) proposed a set of guidelines that can be followed when selecting ARIMA models. It consists of a four-stage iterative process in which:

1. The process is either transformed or differenced to de-trend and stabilise the variance of the data.

2. The autocorrelation and partial autocorrelation plots are used to determine the order of p and q.

3. The parameters of the model are then estimated.

4. A diagnostic check is performed to ensure that the residuals are a white noise process.

If the residuals are not white noise, steps 2-4 are repeated until a satisfactory model is identified. On the contrary, if the diagnostic check reveals that the residuals are random, the developed model will be the final model used for forecasting.

5.2 Semi-Automatic Selection

When a large number of forecasts are produced, manually identifying problematic forecasts becomes a time consuming and difficult task. Prophet provides a semi-automated forecast evaluation system that selects the best model which fits the data. When there are large forecast errors, the forecasts are flagged so that the analyst can explore the cause of the errors, identify and remove potential outliers and either adjust the model or choose a more appropriate model (Taylor and Letham 2017). This keeps the analyst-in-the-loop. Prophet has the following default settings:

- The trend is set to be linear.
- The width of the uncertainty intervals is set to 80%.
- Weekly and yearly seasonality are automatically detected and included in the model if present
- The smoothing parameter for holidays and seasonality is set at 10 while the smoothing parameter for the trend is set to 0.05.
- The number of potential changepoints is set to 25.

This paper makes use of Prophet's default settings to fit the data, however weekly seasonality is included in the model. This allows us to account for our prior knowledge about the closing prices of Bitcoin which tend to be lower around the weekend however higher mid-week due to fluctuations in trading activity. Furthermore, a linear trend is appropriate since Bitcoin does not have an upper limit on its closing prices.

5.3 Rolling Window Forecasts

Rolling window forecasts are useful in evaluating the robustness of a forecasting method. In a rolling window forecast, the forecasts are made h-steps at a time and the actual observation rather than the predicted value is used for the next prediction in the forecast horizon. In this way, a poor forecast will not have negative consequences on future forecasts to be made since the observed values will be used to correct itself for the remaining forecasts (Zhang and Hu 1998). In this paper, the ARIMA and Prophet model is first fitted to the training set and the returns are forecasted h-days ahead for h = 1, 30, 90. The size of the training set is then increased by one observation and the models are refitted. The next h-day ahead return is forecasted and this process is repeated until all forecasts have been made into the test set. The h-day ahead forecasts are then compared to the h-day ahead observed values in the test set and the forecast errors are calculated to determine the RMSE.

5.4 Model Confidence Set

The Model Confidence Set (MCS) procedure developed by Hansen, Lunde, and Nason (2011) is a sequential procedure which starts with an initial set of models M_0 of size m. The worse model is then eliminated at each step until

the null hypothesis of equal predictive ability (EPA) fails to be rejected for all the models belonging to the set at a given confidence level $1-\alpha$. This smaller set of models M^* is the superior set of models (SSM) with the best case occurring when the SSM consists of a single best model.

The null hypothesis for EPA for a set of models M is given by:

$$H_0 : \mu_{i,j} = 0 \quad \forall \;\; i,j = 1,...,m$$

where $\mu_{i,j} = E[d_{i,j}]$ represents the expected loss differential, $d_{i,j,t} = l_{i,t} - l_{j,t}$ represents the loss differential between models i and j at time t, and $l_{i,t}$ and $l_{j,t}$ are the loss functions associated with model i and j at time t respectively. This paper makes use of the squared forecast error as a loss function, however any arbitrary loss function that satisfies the weak stationarity conditions described in Hansen, Lunde, and Nason (2011) may be used. The null hypothesis is tested by constructing the following test statistics:

$$t_{i,j} = \frac{\bar{d_{i,j}}}{\sqrt{\hat{var}[\,\bar{d}_{i,j}]}}$$

$$T_R = \max_{i,j \in M} \;\; |t_{i,j}|$$

where $\bar{d}_{i,j} = m^{-1} \sum_{t=1}^{m} d_{i,j,t}$ measures the sample average loss differential between model i and j and $\hat{var}[\,\bar{d}_{i,j}]$ is estimated through bootstrapping as described in Hansen, Lunde, and Nason (2011).

This paper uses a block–bootstrap procedure of 5000 resamples as in Hansen, Lunde, and Nason (2011). The block length p is chosen by fitting an AR(p) model to the $d_{i,j}$ terms and determining the maximum number of significant AR parameters obtained. While $t_{i,j}$ has a t-distribution, T_R has a non-standard distribution under the null hypothesis hence is estimated

by implementing a bootstrapping procedure similar to that used to estimate $\hat{var}[\bar{d}_{i,j}]$.

The worst model is then eliminated according to an elimination rule which is coherent with the calculated t-statistic and is defined as:

$$e_R = \arg max_i \left(\sup_{j \in M} \frac{\bar{d_{i,j}}}{\sqrt{\hat{var}[\bar{d}_{i,j}]}} \right)$$

5.5 Mincer-Zarnowitz Approach to Forecast Evaluation

The Mincer-Zarnowitz approach to evaluating forecast accuracy is commonly used and can be useful when comparing the forecasts produced by Prophet and the ARIMA model. Mincer and Zarnowitz (1969) proposed an absolute measure to evaluate forecast accuracy that considers the distance between the actual and predicted values. To analyse the absolute errors produced by the forecasts, the observed values r_t are regressed against the predicted values $\hat{r}_t$, i.e.

$$r_t = \alpha + \beta \cdot \hat{r}_t + e_t$$

where α represents the mean distance between the observed and predicted values while β represents the correlation between the forecasted errors and predicted values. When $\alpha = 0$ it implies that the forecasts are unbiased and do not systematically overestimate or underestimate the data. When $\beta = 1$ it implies that the forecasts are efficient and uncorrelated with the forecasted errors. A joint hypothesis test of $H_0 : \alpha = 0 \cup \beta = 1$ is performed to check the efficiency and bias of the forecasts and the model which produces the best results generates superior forecasts.

5.6 Evaluation of forecasts based on data with missing values and outliers

The original dataset is modified in order to examine how robust the ARIMA and Prophet model are when faced with missing values and outliers. 20 values are randomly removed from the training set and 5 outliers are randomly inserted in place of existing values. Both models are then fitted to this "dirty" dataset and the forecasts are evaluated in the same way as the "clean" dataset.

6 Results

6.1 Evaluation of clean dataset over varying forecast horizons

By following the Box-Jenkins methodology, an ARIMA(0,0,1) model was found to be the best fit for the clean training set. Making use of a mean model would result in a returns forecast of zero in a rolling window context. This is due to the white noise behaviour of log returns. Hence a decision was made to employ the ARIMA(1,0,0) model instead. The ARIMA(1,0,0) model has a marginally higher AIC and an analysis of the p-value indicated that the AR parameter is statistically significant in the explanation of the movement of BTC/ZAR. The Prophet model fitted to the clean training set included a linear trend and weekly seasonality with all other parameters set to the default values as described in section 5.2. The models chosen manually are consistent with the models selected by the automatic-selection procedure for ARIMA and Prophet.

Table 6.1 displays the MZ results obtained by regressing the test set

returns against the forecasted returns, the RMSE, and the models that are included in the MCS. It is evident that the ARIMA model ranks ahead of the Prophet model over all forecast horizons. The MCS p-values indicates that both models lie within the 90% confidence interval at the 1 day and 3 month forecast horizon, while the Prophet model is eliminated from the MCS at the 1 month forecast horizon. This suggests that the ARIMA and Prophet model have equal predictive accuracy at the 1 day and 3 month forecast horizon, and that the ARIMA model produces superior forecasts at the 1 month forecast horizon. Furthermore, the MCS p-values indicates that while ARIMA model lies within the 90% confidence interval with certainty across all forecast horizons, the Prophet model lies within the 90% confidence interval 52.44% of the time at the 1 day forecast horizon and 45.78% of the time at the 3 month forecast horizon.

Table 6.1: MCS, MZ-test and RMSE results based on clean data

Forecast Horizon	Model	Model Rank	MCS p-values	α	β	MZ-test p-values	RMSE
1	ARIMA	1	1	0.009546	-2.005624	0.02545	0.04406
	Prophet	2	0.5244	0.013702	-1.709744	0.007152	0.04425
30	ARIMA	1	1	0.031320	-12.736757	0.06136	0.04459
	Prophet	2	0.023 *	0.017169	-2.621914	0.0003773	0.04556
90	ARIMA	1	1	0.006012	-0.938687	0.8430658	0.04231
	Prophet	2	0.4578	0.006953	-0.615435	0.4086752	0.04259

The intercept estimate represented by α is small and positive for ARIMA and Prophet across all forecast horizons, with the smallest estimates obtained at the longest forecast horizon. This suggests that both models systematically underestimates the returns, but to a smaller degree at long forecast horizons. The intercept estimate for ARIMA is closer to zero com-

pared to Prophet at the 1 day and 3 month forecast horizon, indicating that the mean forecasts yielded by ARIMA is closer to the observed mean than Prophet. The slope estimate represented by β is negative for both models across all forecast horizons, indicating that the observed returns are actually the opposite of the suggested forecast and that both models fail to capture the explosive behaviour of BTC/ZAR.

The MZ p-values at the 1 day and 1 month forecast horizon show that the joint hypothesis of a unity slope and zero intercept is rejected at the 10% significance level for both models. This suggests that the forecasts generated by ARIMA and Prophet are biased and/or inefficient. Both models only produce unbiased and/or efficient forecasts at the longest forecast horizon, with the ARIMA model yielding a considerably larger p-value compared to Prophet. The RMSE produced by the ARIMA model is marginally smaller than Prophet across all forecasts horizons, with the smallest RMSE calculated at the longest forecast horizon. These results agree with the results obtained from the MCS and the MZ-test.

6.2 Evaluation of data with missing values and outliers present

By following the Box-Jenkins methodology, an ARIMA(0,0,3) model was found to be the best fit for the dirty training set. A decision was made to employ the ARIMA(3,0,0) model instead. The motivation behind this choice is the same as that provided in the case of the clean data. The ARIMA(3,0,0) model has a marginally higher AIC and an analysis of the p-value indicated that the AR parameters are statistically significant in the explanation of the movement of BTC/ZAR. The Prophet model fitted to the clean training set included a linear trend and weekly seasonality with all other parameters set

to the default values as described in section 5.2. The models chosen manually are consistent with the models selected by the automatic-selection procedure for ARIMA and Prophet.

Table 6.2 displays the MZ results obtained by regressing the test set returns against the forecasted returns, the RMSE and the models that are included in the MCS. We observe that the presence of outliers and missing values in the data has no effect on the rank of the ARIMA and Prophet model. Contrary to the results from the clean data, the Prophet model lies within the 90% confidence interval at the 1 day forecast horizon only. This suggests that the ARIMA and Prophet model only have equal predictive accuracy at the shortest forecast horizon. The MCS p-values indicates that while the ARIMA model always lies within the 90% confidence interval with certainty, the Prophet model only lies within the confidence interval at the 1 day forecast horizon 41.3% of the time. This differs to the results from the clean data in which the Prophet model lies within the 90% confidence interval 52.44% of the time at the 1 day forecast horizon.

Table 6.2: MCS, MZ-test and RMSE results based on dirty data

Forecast Horizon	Model	Model Rank	MCS p-values	α	β	MZ-test p-values	RMSE
1	ARIMA	1	1	0.011895	-0.8832	0.00003182	0.045530
	Prophet	2	0.413	0.014802	-0.6761	0.00000532	0.046210
30	ARIMA	1	1	0.114100	-12.9853	0.1075	0.044758
	Prophet	2	0.0016*	0.014400	-0.6516	0.00000148	0.048180
90	ARIMA	1	1	-0.226700	27.0635	0.01294	0.042302
	Prophet	2	0.0336*	-0.009291	0.6438	0.000764	0.045061

We notice that a positive intercept and negative slope is still estimated at the 1 day and 1 month forecast horizon when the data is dirty. The

intercept estimates are larger than the case of the clean data, suggesting that the biasness of the forecasts yielded by both models increases when there are outliers and missing values present in the data. Contrary to the results from the clean data, Prophet's intercept estimate lies closer to zero at the 1 month and 3 month forecast horizon. This indicates that the mean forecasted returns yielded by Prophet is closer to the mean observed returns when outliers and missing values are present in the data. The negative intercept and positive slope estimated for the 3 month horizon contrasts with the positive intercept and negative slope estimated in the case of the clean data. This shows that although the forecasts produced by both models are systematically overestimated at long forecast horizons when the data is dirty, they are no longer the opposite of the observed returns.

The p-values obtained from the F-test for the joint hypothesis of a unity slope and zero intercept is approximately zero for ARIMA and Prophet across all forecast horizons, with the p-value for Prophet lying closer to zero than ARIMA. The null hypothesis is thus rejected at the 10% significance level, indicating that the forecasts generated by Prophet and ARIMA are always biased and/or inefficient. This contrasts with the clean data which generated unbiased and/or efficient forecasts at the longest forecast horizon. The RMSE produced by both models agrees with the MZ-test and MCS results, with ARIMA producing a marginally smaller RMSE than Prophet across all forecast horizons, and the RMSE being larger than the case of the clean data.

7 Discussion and Conclusions

The results show a clear difference in the predictive accuracy of the ARIMA and Prophet model in addition to the forecast horizon effect. The ARIMA

model ranks ahead of Prophet when forecasting the returns of BTC/ZAR across all forecast horizons. These results hold true irrespective of whether there are missing values and outliers present in the data. Nevertheless, it is reassuring to see that both models have equal predictive accuracy when forecasting at short and long forecast horizons when the data is clean and at short forecast horizons when the data is dirty. Furthermore, the Mincer-Zarnowitz p-values for both models are significant at the same forecast horizons, suggesting that the efficiency and biasness of the return forecasts yielded by ARIMA and Prophet are synchronized at each forecast horizon. Hence, if the analyst is engaged in short-term or long-term trading, they could choose to employ the Prophet model instead of the ARIMA model. This would allow them to produce return forecasts that are as good as the ARIMA model, without having any expert knowledge about model building and selection. It would also provide them with the flexibility to customise the model to suit their needs and incorporate their domain knowledge about BTC/ZAR through the model's intuitively adjustable parameters. If the analyst is interested in generating unbiased and/or efficient forecasts with the least amount of forecast error, they could utilise either model, however should only forecast BTC/ZAR returns over long forecast horizons with data that do not include any outliers or missing values.

The results from the dirty data show that the ARIMA and Prophet model are sensitive to the presence of outliers and missing values, with both models producing unbiased and/or inefficient forecasts and higher RMSEs across all forecast horizons. This is disappointing as Prophet is said to be robust when faced with outliers and can handle missing values without the need for interpolation unlike the ARIMA model. Since both models have equal predictive accuracy at the shortest forecast horizon, an analyst engaged in short-term

trading who has no knowledge on how to pre-process BTC/ZAR data before forecasting, could apply the Prophet model as it would provide them with flexibility and ease of implementation without sacrificing any predictive accuracy.

Although the Prophet model was eliminated from the Model Confidence Set at other forecast horizons, these results might change as more data and information about events which might impact the price of BTC/ZAR becomes available and is incorporated into the Prophet model. Furthermore, the RMSE shows that the difference between the forecast errors of ARIMA and Prophet is marginal, even in the case where both models don't have equal predictive accuracy. Hence, if the analyst is willing to sacrifice a small amount of predictive accuracy in return for the tractability and ease of implementation which Prophet provides, then they might favour Prophet over the ARIMA model.

The attractive features of Prophet were not fully utilised in this paper due to the stationarity of log returns. Different results may emerge when forecasting the direction of returns or closing prices. Furthermore, the reader should exercise caution when generalising these results to other exchange rate data or to exchange rates which involve fiat currencies as results may differ. As a point of further research, more complex techniques such as Neural Networks could be used as a benchmark for comparison when evaluating the forecasts produced by Prophet.

References

Abu-Mostafa, Yaser S, and Amir F Atiya. 1996. "Introduction to Financial Forecasting." *Applied Intelligence* 6 (3). Springer: 205–13.

Anastasakis, Leonidas, and Neil Mort. 2009. "Exchange Rate Forecasting Using a Combined Parametric and Nonparametric Self-Organising Modelling Approach." *Expert Systems with Applications* 36 (10). Elsevier: 12001–11.

Box, George EP, Gwilym M Jenkins, and G Reinsel. 1970. "Forecasting and Control." *Time Series Analysis* 3: 75.

Castillo, Oscar, and Patricia Melin. 2002. "Hybrid Intelligent Systems for Time Series Prediction Using Neural Networks, Fuzzy Logic, and Fractal Theory." *IEEE Transactions on Neural Networks* 13 (6). IEEE: 1395–1408.

Cheung, Yin-Wong. 1993. "Long Memory in Foreign-Exchange Rates." *Journal of Business & Economic Statistics* 11 (1). Taylor & Francis: 93–101.

Etuk, Ette Harrison, Dagogo SA Wokoma, and Imoh Udo Moffat. 2013. "Additive Sarima Modelling of Monthly Nigerian Naira-Cfa Franc Exchange Rates." *European Journal of Statistics and Probability* 1 (1): 1–12.

Fahimifard, SM, Masuod Homayounifar, M Sabouhi, and AR Moghaddamnia. 2009. "Comparison of Anfis, Ann, Garch and Arima Techniques to Exchange Rate Forecasting." *Journal of Applied Sciences* 9 (20): 3641–51.

Fantazzini, Dean, Erik Nigmatullin, Vera Sukhanovskaya, and Sergey Ivliev. 2016. "Everything You Always Wanted to Know About Bitcoin Modelling but Were Afraid to Ask."

Hansen, Peter R, Asger Lunde, and James M Nason. 2011. "The Model Confidence Set." *Econometrica* 79 (2). Wiley Online Library: 453–97.

Hyndman, Rob J, and George Athanasopoulos. 2014. *Forecasting: Principles and Practice*. OTexts.

Khashei, Mehdi, and Hehdi Bijari. 2011. "Exchange Rate Forecasting

Better with Hybrid Artificial Neural Networks Models." *Journal of Mathematical and Computational Science* 1 (1). Science & Knowledge Publishing Corporation Limited (SCIK): 103.

Khashei, Mehdi, Mehdi Bijari, and Gholam Ali Raissi Ardali. 2009. "Improvement of Auto-Regressive Integrated Moving Average Models Using Fuzzy Logic and Artificial Neural Networks (Anns)." *Neurocomputing* 72 (4). Elsevier: 956–67.

Lin, Chiun-Sin, Sheng-Hsiung Chiu, and Tzu-Yu Lin. 2012. "Empirical Mode Decomposition–based Least Squares Support Vector Regression for Foreign Exchange Rate Forecasting." *Economic Modelling* 29 (6). Elsevier: 2583–90.

Mincer, Jacob A, and Victor Zarnowitz. 1969. "The Evaluation of Economic Forecasts." In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, 3–46. NBER.

Nwankwo, Steve C. 2014. "Autoregressive Integrated Moving Average (Arima) Model for Exchange Rate (Naira to Dollar)." *Academic Journal of Interdisciplinary Studies* 3 (4): 429.

Santos, Andr Alves Portela, Newton Carneiro Affonso da Costa, and Leandro dos Santos Coelho. 2007. "Computational Intelligence Approaches and Linear Models in Case Studies of Forecasting Exchange Rates." *Expert Systems with Applications* 33 (4). Elsevier: 816–23.

Taylor, Sean J, and Benjamin Letham. 2017. "Forecasting at Scale."

Tseng, Fang-Mei, Gwo-Hshiung Tzeng, Hsiao-Cheng Yu, and Benjamin JC Yuan. 2001. "Fuzzy Arima Model for Forecasting the Foreign Exchange Market." *Fuzzy Sets and Systems* 118 (1). Elsevier: 9–19.

Zhang, Gioqinang, and Michael Y Hu. 1998. "Neural Network Forecasting of the British Pound/Us Dollar Exchange Rate." *Omega* 26 (4). Elsevier:

495–506.