NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science

Bachelor's Programme 'HSE and University of London Double Degree Programme in Data Science and Business Analytics'

Software Project Report

Time Series Forecasting

Fulfilled by the Student:

Kurlovich Elizaveta

Checked by the Project Supervisor:

Lukyanchenko Peter Pavlovich

Department of Computer Science

Moscow
2022

Contents

1 Introduction

1.1 Abstract

> Since we're all rich with Bitcoins
> ... we ought to put some of this
> unearned wealth to good use
>
> *Hal Finney*

Time series forecasting is the best choice to work with quantifiable data, particularly the stock exchanges. The digital age has greatly motivated people to trade on cryptocurrency exchanges, where the main topic of interest is the first cryptocurrency developed– Bitcoin (BTC), which is paving the way as a disruptive technology gradually declaring competition to immortalized financial payment systems that have been in place for many decades. Cryptocurrency's rate is not constant but continuously changing. Bitcoin prices are influenced by the factors such as the supply of Bitcoin and the market's demand for it, news developments, the cost of producing a bitcoin through the mining process. One of the most prevalent solutions to this problem is forecasting. Time series forecasting may provide the most innovative opportunity to predict Bitcoin exchange rates for investors and academic researchers alike. This paper researches the existing methods and models to forecast data and further develops an algorithm for trading with accurately sending 'Buy' and 'Sell' signals. The research is done on Bitcoin exchange rate to US dollars (BTC-USD) per-hour data of Close Price over 97 days period from 2021-03-20 to 2021-06-27. The methodology of the project is based on five time series forecasting models: ARMA, ARIMA, ARIMAX and Prophet (Facebook Inc.). As a result, models are predicting the data, representing the trend of Bitcoin market prices.

1.2 Project tasks and objectives

The target of this research is to develop a trading algorithm and forecasting model and conduct an empirical analysis. Project objectives:

1. Data preparation – data is defined for the required time and processed for analysis.
2. Forecasting models research – four models are implemented according to found metrics, the best model is determined by quality metrics.
3. Trading algorithm development – based on the best model, the trading algorithm that predicts the future price based on historical data is developed.

2 Theory

2.1 Key terms and definitions

- Time series is a set of observations obtained with equal intervals between them.
- Forecasting model is trained and tested framework that makes predictions based on past and present data.
- Cryptocurrency is a virtual currency with authenticated transactions based on decentralized system secured by cryptography.
- Bitcoin is a decentralized digital currency that can be sent person-to-person on the Bitcoin network without the need of a central server and mediators.
- Stationary time series is time series which distribution properties are stable all over the time.
- Autoregressive model (AR) is a representation of a random process. The autoregressive model specifies that the output depends on the previous inputs and a stochastic term therefore making it a stochastic differential equation.
- Moving Average model (MA) is a stationary model. The moving average is a linear regression of the current value against interferences the in the series.

3 Project Implementation

3.1 Selected models

This research is based on four forecasting models: ARMA, ARIMA, ARIMAX, Prophet (Facebook Inc.).
Autoregressive model (AR(p)) - a time series model, where previous-step observations are an input data for regression model predicting value on the next time period. AR(p) is an autoregression model of order p:

$$y_i = c + \sum_{i=1}^{p} \varphi_i y_{t-1} + \varepsilon_t \tag{1}$$

Moving Average model (MA(q)) - a time series model, where previous-step noise components are an input data for regression model predicting value on the next time period. MA(q) is a moving-average model of order q:

$$y_i = c + \sum_{i=1}^{q} \theta_i y_{t-1} + \varepsilon_t \tag{2}$$

The basic elements of AR and MA models can be combined to produce a great variety of models. ARMA (Autoregressive Moving Average) model used to describe slightly stationary time series in terms of two polynomials, one of them is moving average, second- autoregression. ARMA(p, q) is a sum of autoregression model of order p and moving-average model of order q:

$$y_i = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i y_{i-t} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{3}$$

If non-stationarity is added to a mixed ARMA model, then the general ARIMA model is obtained. ARIMA (Autoregressive Integrated Moving Average) uses previous time series data plus an error to forecast future values on non-stationary data. ARIMA (p, d, q) is a differentiated d times ARMA model:

$$\triangle^D y_t = \alpha + \varepsilon_t + \sum_{i=1}^{p} \phi_i \triangle^D y_{t-i} + \sum_{j=1}^{q} \theta_i \varepsilon_{t-j} \tag{4}$$

An extended version of the ARIMA model is the ARIMAX (Autoregressive Integrated Moving Average Extended). It includes also other independent (predictor) variables. X- exogenous factor. This model allows to take advantage of autocorrelation that may be present in residuals of the regression to improve the accuracy of a forecast. ARIMAX (p, d, q) refers to differentiated d times ARMA model with exogenous data Xt:

$$\triangle^D y_t = \sum_{i=1}^{p} \phi_i \triangle^D y_{t-i} + \sum_{j=1}^{q} \theta_i \varepsilon_{t-j} + \sum_{m=1}^{M} \beta_m X_m, t + \varepsilon_t \tag{5}$$

Prophet model is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. The algorithm of this model and the installation methods are described on Prophet website[1]. This model is extremely useful with depending on season time series. Prophet can be considered as a non-linear regression model of the form:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t, \tag{6}$$

where g(t) describes a growth term, s(t) describes the various seasonal patterns, h(t) captures the holiday effects.

3.2 Data

The dataset of Bitcoin exchange rate to US dollars (BTC-USD) is collected from API of Polygon. The dataset is read in three parts: from 2021/02/20 to 2021/04/25, from 2021/04/26 to 2021/05/26, from 2021/05/27 to 2021/06/27. The frequency is one hour. It total, the collected

[1]https://facebook.github.io/prophet/

dataset forms 97 days, which is 2328 hours. The date and time were represented in the form of timestamp, which was converted into convenient date-time form. Consequently, the dataset is as follows:

Num	Volume	Open	Close	Timestamp
1	4108.668109	55866.00	56254.8600	2021-02-20 00:00:00
2	3022.427323	56142.00	56235.7800	2021-02-20 01:00:00
3	2126.892304	56240.00	55741.8800	2021-02-20 02:00:00
...	...	...	...	...
2328	1183.992554	38542.80	38556.8800	2021-05-27 23:00:00

The particular research analyzes and develops the prediction models based on per-hour Close Price data. The graph below represents the Close Price rate over the time:

Even though the graph above represents that the data is non-stationary, it is necessary to apply an accurate Augmented Dickey-Fuller (ADF) test, which is a statistical test that is built to test whether univariate time series data is stationary or not.

```
Augmneted Dickey_fuller Statistic: -1.035085
p-value: 0.740254
critical values at different levels:
        1%: -3.433
        5%: -2.863
        10%: -2.567
```

This test is based on a hypothesis and shows the degree of probability with which it can be accepted. The results obtained suggest that ADF is greater than any critical value at different levels. Moreover, p-value is greater than significance level = 0.05, so there is a fail to reject the null hypothesis H0 at three different confidences: at 90%, at 95%, at 99%. Thus, the data is definitely non-stationary.

3.3 Quality metrics

The next part of this research is devoted to evaluating models and identifying the best one with a use of quality metrics (time series forecast error metrics). There are two types of error metrics used in this research – scale-dependent metrics and percentage-error metrics.

Quality metrics	
Scale dependent metrics	Percentage-error metrics
Mean Squared Error (MSE)	Mean Absolute Percentage Error (MAPE)
Mean Absolute Error (MAE)	Symmetric Mean Absolute Percentage Error (sMAPE)

Implementation of these four quality metrics gives the following results:

	sMAPE	MAE	MSE	MAPE
ARMA	1.486	1053.020	1811089.849	0.030
ARIMA	1.517	1067.157	1865992.187	0.030
Prophet	2.824	2012.159	6023289.414	0.057

Figure 1: Quality metrics

It can be concluded that ARMA model, that has the lowest indications for the most of error metrics, is the best model to choose for further predictions.

3.4 Technologies and Instruments

- Python 3.8.3 (MacOS)
- Google Collab
- GitHub
- Python libraries:
  - Pandas
  - NumPy
  - Matplotlib
  - Seaborn
  - Plolty.express
  - Datetime
- Python modules:
  - Prophet

4 Conclusion

Finally, the project 'Time series forecasting', which goal is to predict Bitcoin close price using time series forecasting models is done. The main tasks are fulfilled, namely: data is processed for analysis and classified; three out of four models (ARMA, ARIMA, Prophet) are implemented and the best model is determined by one of the most appropriate quality metrics. The several issues to struggle with are finding an appropriate exogeneous factor of ARIMAX model suitable for given per-hour Bitcoin close price data, and negative profit of best chosen by quality metrics ARIMA model. The only one ARMA model got the positive profit.

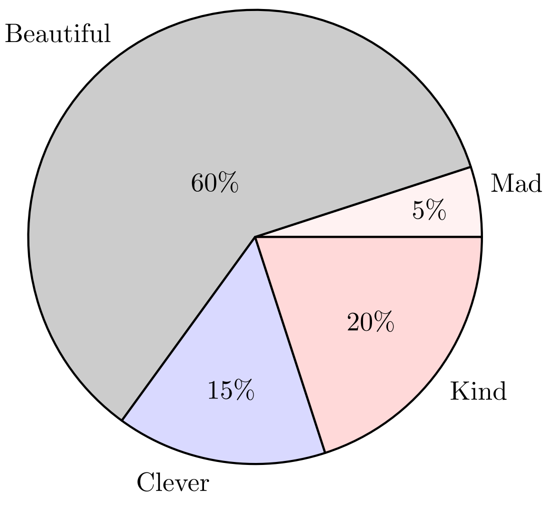
5 List of references

References

[1] Peter J. Brockwell, Richard A. Davis (2006) *Time Series: Theory and Methods.*

[2] Makridakis, Wheelwright, and Hyndman (2004) *Prediction, Methods, and Applications.*, 3rd ed.

[3] Chris Chatfield (2000) *Time series forecasting.*



$$S(\omega) = \frac{\alpha g^2}{\omega^5}\, e^{[-0.74\{\frac{\omega U_\omega 19.5}{g}\}^{-4}]} \tag{7}$$