Time Series Analysis and Forecasting

Dušan Stević SW10-2016

Problem

Comparison between algorithms for time series forecasting. The main focus of the project is to study performance of statistical models (AR, MA, ARIMA) and machine learning models (LSTM).

Motivation

- A time series is a series of data points indexed in time order.
- Time series forecasting is the use of a model to predict future values based on previously observed values.
- The motivation for this study is to present performance evaluation for state-of-the-art algorithms for time series forecasting.

Objectives

- Data collection (Data acquisition)
- Data preprocessing
- Develop models
- Evaluate algorithmic performance

Data collection and preprocessing

- Historical stock prices
- Stock market API
- Handling NA values
- Feature engineering and scaling

Detecting stationarity in time series

Rolling (moving) statistics

- Additive time series
- Multiplicative time series
- Visual Test
- Rolling Mean
- Rolling Standard Deviation

Augmented Dickey Fuller Test

- Additive time series
- Statistical Tests
- p-value threshold 5%
- (H0): Time series is non-stationary. It has some time dependent structure.
- (H1): Time series is stationary

Kwiatkowski–Phillips–Schmidt–Shin Test

- Multiplicative time series
- Statistical Tests
- p-value threshold 5%
- (H0): Time series is non-stationary. It has some time dependent structure.
- (H1): Time series is stationary

Techniques for making time series stationary

- Differencing (Time Shift Transformation)
- Decomposition (Time series = Trend + Seasonality + Noise)

Differencing

Given a set of observation on the time series:

$$x0, x1, x2, x3, \ldots . xn$$

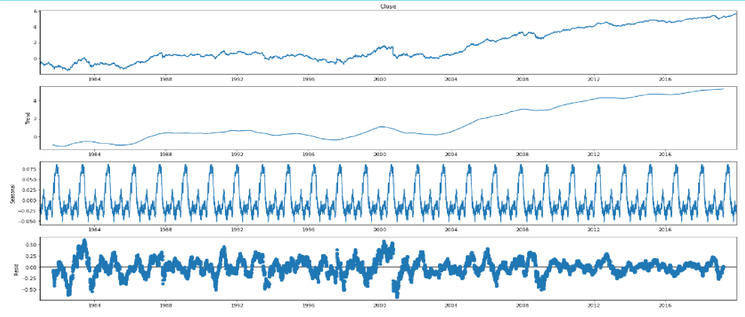
The shifted values will be:

$$null, x0, x1, x2, \ldots . xn$$ ⟵ - basically all xi's shifted by 1 pos to right

Thus, the time series with time shifted values are:

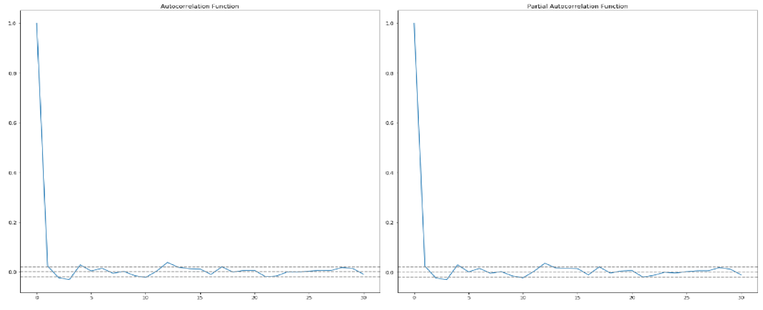
$$null, (x1 - x0), (x2 - x1), (x3 - x2), (x4 - x3), \ldots . (xn - x_{n-1})$$

Decomposition



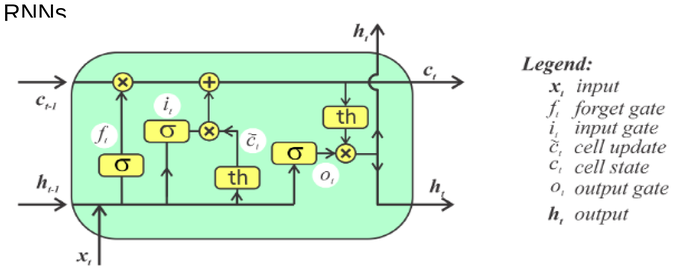
Statistical Models

- The Autoregressive AR(p) Part of ARIMA : describes relationship between a current observation and previous observations.
- The Integrated I(d) Part of ARIMA: describes number of lags (shifts) until time series becomes stationary.
- The Moving Average MA(q) Part of ARIMA: describes relationship between an observation and a residual error from previous observations.



Machine learning models

- Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture.
- Capable of handling sequential data (time series)
- A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.
- LSTMs were developed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs.



Tuning Models

- Grid search method (AR, MA, ARIMA)
- Trial and error method (LSTM)

Evaluation

- ARIMA Root Mean Square Error (RMSE): 23.2218
- LSTM Root Mean Square Error (RMSE): 9.4728