Udacity 2018: Capstone Project

Forecasting time series with machine learning

with applications to currency exchange rates

Author: Oscar Javier Hernandez

July 29, 2018

1 Domain background

A set of data that is indexed by time is known as a time series. They appear in many different fields, such as statistics, physics, finance, economics, biology, or even business [1]. Because of their wide applicability, it is important to generate accurate forecasts of time series data. These forecasts are generated using specific mathematical models or algorithms which are trained on a subset of the past values of a given time series. For the purpose of simplifying future discussions, we will adopt the following notation for a time series, denoted $X(t)$ or X_t, as

$$\{X(t); t = 0, 1, ...\}. \tag{1}$$

Where t denotes the time-index of the series. One of the simplest models for a time series is the ARIMA (Auto regressive integrated moving average) model. This model is denoted as ARIMA(p, q, d), and assumes that the time series X_t has the form

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{p} \phi_i L^i \left[(1-L)^d\right] X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \tag{2}$$

where $\{\phi_i | i = 1, ..., p\}$, $\{\theta_i | i = 1, ..., q\}$ are model parameters and L is the lag operator defined as $LX_t = X_{t-1}$. The term ϵ_t denotes the error terms, assumed to be independent, identically distributed random variables sampled from a zero-mean, normal distribution. The value μ denotes the average of this model. ARIMA models can be applied to make forecasts of stationary time series (defined as a time series whose mean, variance and auto correlation does not change over time), or to a time series that can be transformed into a stationary time series. However, there are other state-of-the-art machine learning methods that can be used to model time series methods. Which will the main goal of this project.

One important type of financial time series is the exchange rate between different currencies (Fig. 1). An exchange rate, is the rate at which one currency will be exchanged for another. There are many factors that can influence this rate, such as balance of payments, interest rate levels, inflation levels and other economical factors which are beyond the scope of this project [3].

Because of the relevance of exchange rates to the global financial trade, it is very important to make accurate forecasts of how exchange rates will change in the future. The ultimate goal of this capstone project will be to analyze basic time series and apply machine learning methods to make forecasts about currency exchange rates.

My personal motivation for working on time series is to use it to take advantage of the best exchange rates. As someone who lives abroad in a country that uses a different currency, I often need to transfer money to and from my different bank accounts. These transfers are subject to fluctuating currency exchange rates. Without a way to predict what the exchange rate will be for the time of the transfer, I end up losing money in these transfers. Therefore, I am interested in developing a way to forecast the exchange rates so that I can minimize the loses during these transfers.

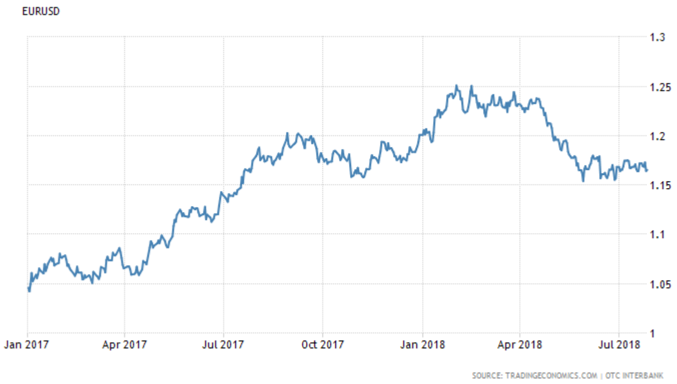

Figure 1: The exchange rate from EUR to USD from Jan 2017 to Jul 2018.

2 Problem statement

The main objective of this project will be to use classical and more recent machine learning techiques to make forecasts of different time series with the goal of applying the best methods to predict currency exchange rates. The simplest model that we will use is the ARIMA model, defined in the previous section as the baseline model, along with a linear regression model. The ARIMA model has been shown to be adequate in estimating the exchange rates of certain currencies Ref. [2]. We will then use different neural networks architectures such as the feed forward and recurrent networks, as in Ref. [5, 4, 6], to make predictions of time series and use our baseline models and root mean square differences to quantify and compare the performance of our different architectures.

3 Data sets and inputs

In order to get a good understanding of forecasting methods we will use a number of different data sets that have more regular patterns for testing the different forecasting methods before applying the algorithm to the financial data sets. In all cases, the data sets will contain the date t along with the value of the feature X_t. The data sets that we will use for the project are the following,

1. Sunspot data
   - A data set containing the number of sunspots from 1771-1869 `http://bit.ly/1yFUqWY`
   - A data set containing the number of sunspots from 1970-1995 `http://bit.ly/1x9kgSm`
   - This data set is well known, and will be used to test the different forecasting models.
2. Monthly U.S air passenger miles January 1960 through December 1977
   - `http://bit.ly/1HHh9ry`

- Another well known data set having both seasonal and linear trends. It will be used for bench marking.

3. Total annual rainfall in inches, for London England
    - The time range is between 1813-1912
    - `http://bit.ly/1gok63g`
4. Mean daily temperature (Celsius), of the Fisher river near Dallas
    - The date range is between Jan 01,1988 until Dec 31, 1991.
    - `http://bit.ly/18f14sz`
    - This data set will be used for testing the algorithms.
5. The Kaggle EURUSD Data set
    - The data contains Forex EURUSD currency rates in 15-minute slices (OHLC - Open High Low Close, and Volume). BID price only.
    - The data ranges from Jan 1, 2010 until Jan 1, 2017. `https://www.kaggle.com/meehau/EURUSD/home`
6. Historical EURCAD data set
    - The data contains historical data, BID price only.
    - The data ranges from Oct. 24, 2004 until July 21, 2018.
    - The data was downloaded online from `https://www.dukascopy.com/swiss/english/marketwatch/historical/`
7. Exchange rate of the Australian dollar
    - The date range is from 1969-1995
    - `http://bit.ly/1x9kgSm`

4 Solution statement

The solutions that we purpose for this project is to use a feed forward artificial neural networks [5], along with a recurrent network [6] to make time series predictions. If time is available, we also propose to use a generative adverserial network (GAN) which has shown some promise for time series models in the medical field [8]. We will train these neural network architectures using the data sets described in the previous sections and make predictions. We will use the evaluation metrics described in later sections to compare the performance of these solutions against the benchmark models in the next section.

5 Benchmark model

For the benchmark models, we will use two well known models, a simple linear regression model and an ARIMA model. The linear regression model will assume the following simple short term relation between points in the time series,

$$X_t = a \cdot t + b. \tag{3}$$

The parameters to be fit are slope, a, and the intercept, b. The **sklearn** package in python can be used to fit lines to the data.

The second model that we will use to benchmark our machine learning algorithms will be the ARIMA(p, q, d) model defined in the introduction. The **stats** package in python contains numerical routines that will fit an ARIMA model to a time series [7].

Both the linear and the ARIMA model will be our basic models that will be used to compare the performance of our neural network based machine learning models using the evaluation metrics in the next section. This will allow us to determine if the neural network machine learning methods are in fact better than the base models.

6 Evaluation metrics

There are several metrics that we can use to evaluate the predictions of our models Ref. [1], however, for our project we will focus on three commonly used metrics. We will first define some terminology, the forecast error, e_t, is given by

$$e_t = X_t - F_t, \tag{4}$$

where X_t is the value of the time series at time step t, and F_t is the forecasted value at the same time step. The three metrics that we will use for our project are

1. The Mean Absolute Percentage Error (MAPE)
   - MAPE $= \frac{1}{N} \sum\limits_{t=1}^{N} \left| \frac{e_t}{X_t} \right| \cdot 100$
2. The Mean square error (MSE)
   - MSE $= \frac{1}{N} \sum\limits_{t=1}^{N} e_t^2$
3. The Mean absolute error (MAE)
   - MAE $= \frac{1}{N} \sum\limits_{t=1}^{N} |e_t|$

In these three cases, the smaller the value of the MAPE, MSE, and MAE, then the better the model.

7 Project design

The project design will consist of the following steps

1. Data preprocessing
   - First we can perform transformations on the data to remove seasonal patterns, linear trends for each of the data sets that we will consider.

- We can also perform different smoothing operators, like the running average, or exponential smoothing.

2. Splitting the data

   - We will split the data into testing and training.
   - The amount will be (80%) training and (20%) testing, but we will evaluate the models using different amounts of training data.

3. Training

   - The models that we will train are: Linear regression, ARIMA, and at least two different neural network architectures described in the previous sections.
   - Once the data has been split into training and testing, then we will use the training data set to determine the model fit parameters, and make predictions on the testing set.
   - The model parameters will be varied during this phase, until the objective function is minimized. The objective function will be either the MAPE, MSE or MAE metric.
   - For each of the trained models, we will compare the value of the metrics defined in the previous section.

4. Forecasts

   - Once the training phase is complete, then the model will be used to make predictions of the future.
   - We will also attempt to quantify the uncertainty associated with the forecasts of the various models.

References

[1] R. Adhikari and R. K. Agrawal, An Introductory Study on Time Series Modeling and Forecasting, 2013, [arXiv:1302.6613] https://arxiv.org/abs/1302.6613.

[2] Research Journal of Finance and Accounting www.iiste.org ISSN 2222-1697 (Paper) ISSN 2222-2847 (Online) Vol.7, No.12, 2016 http://iiste.org/Journals/index.php/RJFA/article/viewFile/31511/32351.

[3] P. J. Patel, N. J. Patel and A. R. Patel, IJAIEM 3, 3, 2014. http://www.ijaiem.org/volume3issue3/IJAIEM-2014-03-05-013.pdf

[4] B. Oancea, S. Cristian Ciucu, Proceedings of the CKS 2013, [arXiv:1401.1333] https://arxiv.org/abs/1401.1333.

[5] T. D. Chaudhuri and I. Ghosh, Journal of Insurance and Financial Management, Vol. 1, Issue 5, PP. 92-123, 2016, [arXiv:1607.02093] https://arxiv.org/abs/1607.02093.

[6] Pant, N. (2017, September 07). A Guide For Time Series Prediction Using Recurrent Neural Networks (LSTMs). Retrieved from https://blog.statsbot.co/time-series-prediction-using-recurrent-neural-networks-lstms-807fa6ca7f.

[7] Vincent, T. (2018). ARIMA Time Series Data Forecasting and Visualization in Python — DigitalOcean. [online] digitalocean.com. Available at: https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3 [Accessed 29 Jul. 2018].

[8] C. Esteban, S. L. Hyland and G Rätsch, https://github.com/ratschlab/RGAN [arXiv:1706.02633].