MINOR PROJECT

A COMPARISON OF ARIMA AND LSTM IN TIME SERIES FORCASTING

Sima Siami-Namini
Department of Applied Economics
Texas Tech University
Email: sima.siami-namini@ttu.edu

Neda Tavakoli
Department of Computer Science
Georgia Institute of Technology
Email: neda.tavakoli@gatech.edu

Akbar Siami Namin
Department of Computer Science
Texas Tech University
Email: akbar.namin@ttu.edu

Shashank Sinha (1810002)
Subject code : PHL8902

Supervisor : Dr. Anurag Sahay

DATA COLLECTION

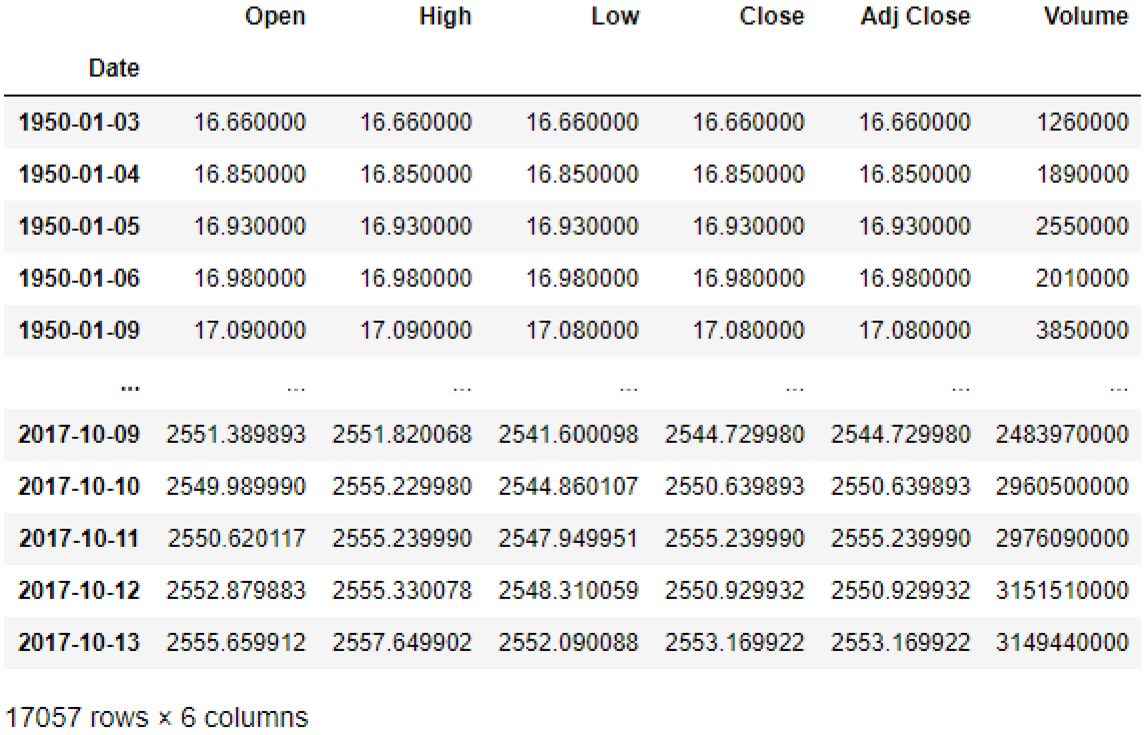
For this project I have used seven datasets which are following :

- S&P 500 historical data (1733 * 6)
- "LeBron James" Google Trends (170 * 2)
- "Coldbrew" Google Trends (170 * 2)
- "Kentucky Derby" Google Trends (170 * 2)
- "Gilmore Girls" Google Trends (170 * 2)
- "Olympics" Google Trends (170 * 2)
- "Zika Virus" Google Trends (170 * 2)

TECHNIQUES USED

- time series analysis and forecasting
- pandas and numpy for datetime, data preparation
- statsmodel and keras for ARIMA and LSTM modeling
- sklearn for metrics & scaling, and matplotlib for plots

S&P 500 DATASET



	Open	High	Low	Close	Adj Close	Volume
Date						
1950-01-03	16.660000	16.660000	16.660000	16.660000	16.660000	1260000
1950-01-04	16.850000	16.850000	16.850000	16.850000	16.850000	1890000
1950-01-05	16.930000	16.930000	16.930000	16.930000	16.930000	2550000
1950-01-06	16.980000	16.980000	16.980000	16.980000	16.980000	2010000
1950-01-09	17.090000	17.090000	17.080000	17.080000	17.080000	3850000
...	...	...	...	...	...	...
2017-10-09	2551.389893	2551.820068	2541.600098	2544.729980	2544.729980	2483970000
2017-10-10	2549.989990	2555.229980	2544.860107	2550.639893	2550.639893	2960500000
2017-10-11	2550.620117	2555.239990	2547.949951	2555.239990	2555.239990	2976090000
2017-10-12	2552.879883	2555.330078	2548.310059	2550.929932	2550.929932	3151510000
2017-10-13	2555.659912	2557.649902	2552.090088	2553.169922	2553.169922	3149440000

17057 rows × 6 columns

MODELS USED

ARIMA

An autoregressive integrated moving average model is a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables. The model's goal is to predict future values of the dataset by examining the differences between values in the series, instead of going through actual values.

LSTM

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency (Vanishing gradient) problem. Remembering information for long periods of time is practically their default behavior.

Autoregressive Integrated Moving Average (ARIMA)

- ▷ The function arima_model creates an ARIMA model rolling forecast for a given input time series. The various steps in the function include:

1. log transforming data
2. creating train/test splits
3. creating an ARIMA model for the train set
4. forecasting the first value in the test set, followed by adding that value to the training set and remodeling, forecasting the next value in the test series, adding that second value to the train set, and so on.
5. inverse transforming the data
6. creating plots and generating error metrics.

Long short term memory (LSTM)

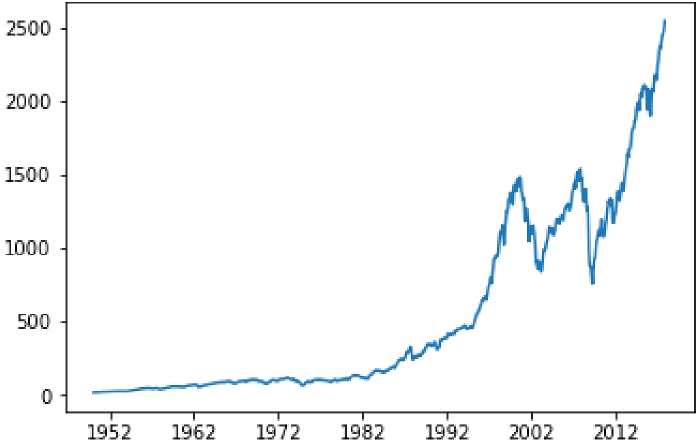
The LSTM calculations use three different functions for creating a dataset that can be modeled. Each function is described in more detail below.

1. Creating the dataset : This function is used to create the datasets required for training and testing LSTM neural nets. It accepts a time series, the number of previous periods the user would like to model, the train / test split fractions, and whether to perform differencing or log transforms on the data to make it stationary. It will also scale all data between 0 and 1 for input into the LSTM.
2. Inverse transformations : This inverse transform function simply reverses any transformations performed when generating the dataset. Inversing the transformations allows for the model predictions to be based on the same scale as the original dataset for more intuitive interpretation of results. Both the model creation and inverse transformation functions are automatically called within the LSTM function below.
3. LSTM modelling : Calling the LSTM model only requires a time series dataset, the number of desired look-back periods, the train/test split, whether to log transform or difference the data, and parameters for training such as number of nodes and epochs. Within the function, it creates the train and test datasets—both features and targets—and then trains an LSTM model, followed by forecasting the out-of-sample data. The predictions from the model, as well as the actual target values are then inverse transformed using the function above, and plots are generated along with error metrics.

S&P 500 HISTORICAL DATA

The ARIMA model performed slightly better than the LSTM model, but the improvements were mostly negligible. Gaussing filtering the data improved accuracy for both models.

S&P 500 historical closing price data from 1950 through October 2017 (obtained through Yahoo Finance). The data is averaged monthly and plotted in the graph. This trend shows increasing values over time, along with some steep increases and decreases.

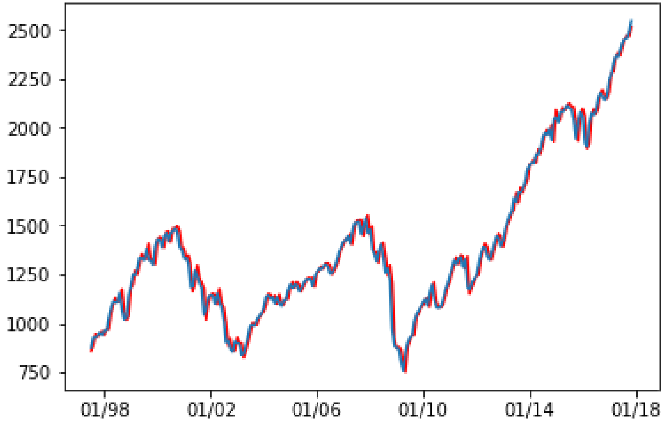


Outcomes:

- ARIMA RMSE (no filter): 45.50
- LSTM RMSE (no filter): 46.67
- ARIMA RMSE (filtered): 24.57
- LSTM RMSE (filtered): 24.75

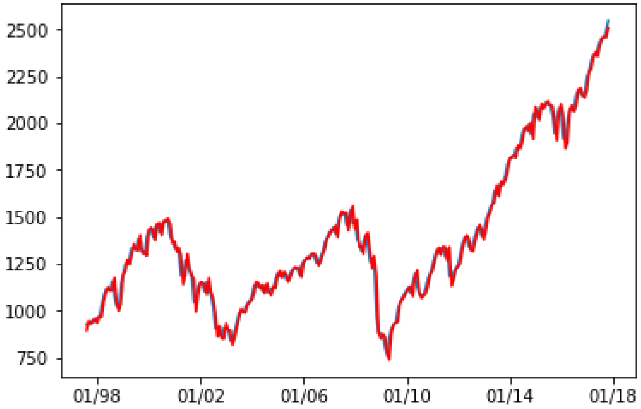
S&P 500 ARIMA model

The ARIMA predictions below have an average RMSE value of 45.5 (with the range of actual values being from ~800 to 2500). One differencing period and log transformation was used.



S&P 500 LSTM model

The LSTM model below also uses log transformation and differencing, with 5 training epochs to update model weights. It has a very similar average RMSE error of 46.7.



RESULT

Data Set	ARIMA (test RMSE)	LSTM (test RMSE)	Gaussian ARIMA (RMSE)	Gaussian LSTM (RMSE)
S&P 500	45.495	46.671	24.570	24.751
LeBron James	20.009	27.175	9.816	13.743
Cold Brew	9.157	8.597	3.025	4.543
Kentucky Perby	30.935	30.224	15.232	15.232
Gilmore Girls	12.026	13.885	6.517	7.592
Olympics	17.420	18.506	9.146	11.078
Zika Virus	16.771	16.040	8.543	8.586

CONCLUSION

Overall, the results demonstrate that both ARIMA and LSTM are quality algorithms for forecasting time series data. In general, the ARIMA model provided slightly lower errors, but also can suffer from convergence errors for series with sharp gradients. The LSTM series can train and make predictions on any series—though the accuracy must be evaluated.

Further, Gaussian filtering of the dataset improved predictions in every case, even when comparing the filtered predictions to the original, un-filtered dataset.