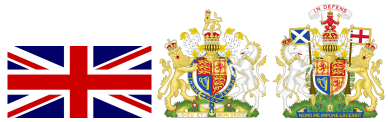


Classical Machine Learning vs. Deep Learning Second Elizabethan Age Financial Portraiture Post-Europe: Forecasting the GBP/USD Exchange Rate in the Era of Brexit

Alp Kutlualp*
Stanford University
alpkalp@stanford.edu

Abstract

The world's oldest currency still in use, the British Pound, is a symbol of financial stability. In the past three decades, it fluctuated the most over one single issue: Britain's relationship with Europe. The day after the Brexit vote, on June 24^{th}, 2016, the British Pound moved as much as 10% intraday. This week, Britain is heading to another crucial election to resolve Brexit - its 5^{th} vote, 3^{rd} General Election, with its 3^{rd} Prime Minister in 5 years - in what many describe as *the election of our lifetime*. As the British people go to the polls on Dec 12^{th}, 2019, we select among machine learning and deep learning models to forecast the GBP/USD exchange rate in high-frequency space, using tick data the day after the Brexit vote in 2016. We predict the GBP/USD exchange rate in 30 second intervals using ARIMA, MLP, CNN, LSTM, Bidirectional LSTM, CNN-LSTM and ConvLSTM models. We show that the classical machine learning ARIMA model outperform the Deep Learning models. Among the Deep Learning Models, we show that Multi-Layer Perceptrons and Bidirectional LSTM models outperform other models. The ConvLSTM hybrid model outperforms its constituents.

1 Introduction

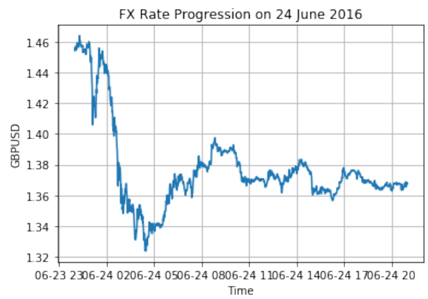
"We have not successfully rolled back the frontiers of the state in Britain, only to see them re-imposed at a European level with a European super-state exercising a new dominance from Brussels." (Margaret Thatcher, Bruges Speech, 1988)

The British Pound (ISO Code: GBP) is the world's oldest currency still in use. Britain's longest reigning Sovereign, Queen Elizabeth II, was featured on the Pound for the first time in 1960, and ruled over a series of political and economic transformations of the United Kingdom, each influencing the value of the Pound. Taking inspiration from the monarch it features on the currency, the Pound is an acclaimed symbol of financial stability. Representing the challenges Britain has faced, the value of the Pound and its volatility has been a reliable barometer of Britain's political and economic state.

*SCPD Student. This project is shared with CS230. MLP, CNN and LSTM are part of the work for CS 229, and Stacked LSTMs, Bidirectional LSTM, and the hybrid models, CNN-LSTM and ConvLSTM, are part of the work CS230. The work done for ARIMA modeling is shared by both projects as it serves as a crucial time series benchmark. **Disclaimer:** This paper is for academic purposes only and is NOT investment advice.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Of all the issues that impacted Britain and the value of the British Pound in recent history, one issue has stood alone at the center: its relationship with Europe. In the last three decades, the Pound sterling displayed most volatility over debates over Europe, while Margaret Thatcher's visionary warnings over the dangers of the single currency as well as a political European union of Europe that went beyond the economic single market. The largest price moves in the Pound sterling would take place in September 1992, when Britain would leave the European Exchange Rate Mechanism, and in June 2016 when the most number of votes cast in British electoral history would vote to leave the European Union in the Brexit referendum. On June 24^{th} 2016, GBP/USD exchange rate would move as much as 10% intraday and ending the day with 8% decline.



With Brexit unresolved more than 3 years after the vote, Britain is heading to its 5^{th} public vote, 3^{rd}General Election with its 3^{rd} Prime Minister in 5 years, in what many call *the election of our lifetime* on December 12^{th}, 2019.

In this paper, we apply deep learning to a *different* type of image dataset, to the financial image of a country represented by the value of its currency. In keeping up with the electronification of financial markets, we do not look at the traditional daily exchange rate time series data. We rather analyze exciting alternative financial data, namely, intraday high-frequency GBP/USD exchange rate tick data. With the prevalence of high-frequency trading in financial markets, reliable time series models and forecasts are important for financial risk managers to understand the intraday fluctuations that cannot be captured from variation in classical end-of-day daily time series data. The Brexit-related volatility in the British Pound provides an excellent case-study to analyze the performance of machine learning and deep learning models in the high-frequency space.

Clearly stated, the input to our algorithms is the average GBP/USD exchange rate in 30 second intervals, and the output is the 1 step-ahead time series prediction of the GBP/USD exchange rate in the next 30 second interval . We are using both classical and state-of-the-art machine learning and deep learning algorithms, namely Autoregresive Integrated Moving Average (ARIMA) models, Multi-Layer Perceptrons (MLPs), vanilla Long Short Term Memory (LSTM) models, stacked LSTMs, Bidirectional LSTMs, Convolutional Neural Networks (CNNs), and hybrid models, namely CNN-LSTMs and Conv-LSTMs. This paper is novel in its application of machine learning and deep learning to one of the most idiosyncratic and topical financial time series of the past 3 years, the British Pound in high-frequency space, in its application of hybrid deep learning architecture and comparisons at the one-step-ahead forecast level.

As the British public go into Election Day with which party they will vote for in mind, we will go into the results announcements with the best performing model among those listed above to forecast the GBP/USD exchange rate.

2 Related work

The jury is still out on the performance comparison between classical machine learning versus deep learning methods on time series in the literature. Given that most financial time series data is small data (252 trading days in a year, one can have 2520 observations in 10 years), deep Learning's home turf on very large datasets is not easily available here. For this reason, it is not surprising to see Gu et al.'s conclusion that less deep networks outperform deeper ones (2018). On the other hand Ryll and Seidens (2019) note that machine learning methods do better than traditional stochastic methods used in finance, and that more complicated Recurrent Neural Network (RNN) models outperform MLP - a contradiction to Gu et al.'s finding. In the high-frequency space exchange rate data, Ganesh et al. (2018) restrict their analysis MLPs only and Rundo (2019) looks at a variation of stacked LSTMs. There are very limited studies that compare different models for high-frequency space on the same data, and when they do, they do not explore synergies between the different models. Zhou et al. (2018) use LSTMs and CNNs but do not explore hybrid models, such as CNN-LSTM and ConvLSTM. Brownlee (2019) provides use-cases of both vanilla and hybrid deep learning architectures in non-financial settings, which we leverage in our application to tick data. Ntakaris et al (2019) argues that deep learning outperforms classical time series models, such as ARIMA. There is a lot of disagreement within the literature and the state-of-the-art is yet to be found - or is not disclosed. There would be few incentives for creators of successful prediction models to share them given the revenue opportunities involved with accurate forecasts. Our study comprehensively compares a variety of classical machine learning and deep learning models on the same dataset.

3 Dataset and Features

3.1 Keep Calm: Alternative Alternative High Frequency Financial Time Series Data

We are using the GBP/USD intraday high-frequency tick data from June 24^{th}, 2016, the day after the Brexit vote as our dataset. The source of the dataset is Dukascopy, an online Swiss broker. The data is available for free upon registration at **https://www.dukascopy.com/swiss/english/marketwatch/historical**. Our dataset has 254,111 observations at the millisecond level. Here is a snapshot below:

Gmt time	Ask	Bid	AskVolume	BidVolume
24.06.2016 00:00:00.236	1.45604	1.45583	1000000	1870000
24.06.2016 00:00:00.391	1.45604	1.45582	1000000	2250000
24.06.2016 00:00:00.507	1.45604	1.45596	1000000	1000000
24.06.2016 00:00:00.611	1.45604	1.45571	1000000	1870000
24.06.2016 00:00:01.120	1.45604	1.45572	1000000	1870000

3.2 And Carry On: Pre-Processing Unevenly Spaced High-Frequency Tick Data

The traditional train/validation/test split methodology does not apply to time series prediction tasks for two reasons: 1) Time series data is strictly sequential and order is important; the data cannot be shuffled. 2) Forecasts of time series models start to become quickly unreliable after a few time steps into the future, by either following the trend of the most recent observations or flattening off. Hence one-step ahead forecasts or n-step ahead forecasts, for very small n, are more valuable for test performance evaluation. We hence use walk-forward testing in our analysis. The model is trained on 70% of the dataset and then predicts the one-step ahead forecast. Then, iteratively, the model adds the new observation into the training dataset and predicts the next, *unseen*, one-step-ahead forecast and repeats this for the entire test set.

Given the nature of the high-frequency data, there is not necessarily a quote for every millisecond. This is not a problem faced in traditional time series where there is an observation for every day, month or quarter. We have hence pre-processed our data such that we sample the Open, High, Low and Close (OHLC) of every 30 second time interval and take its average. The average OHLC enables us to avoid outliers, while also pooling the various data points in the interval. The features are the lag(s) of the GBP/USD high-frequency data at 30 second intervals, itself. We are looking for how preceding values and movements in the time series is forecasting the value at the next time step.

4 Methods

We tested the following algorithms to forecast GBP/USD time series in the high-frequency space:

1. Autoregressive Integrated Moving Average (ARIMA): ARIMA models are the archetypal time series analysis models that uses Maximum Likelihood Estimation (MLE) to determine the relationship between a variable, Y_t and its lags, eg. Y_{t-1}, Y_{t-2}, and its forecast error and the lags of the forecast errors, eg. ϵ_t, ϵ_{t-1}, ϵ_{t-2}. The Autoregressive (AR) component refers to the p number of lags of the variable Y_t, Integrated (I) component refers to the presence of a trend in the series and hence whether it needs to be differenced, (MA) component refers to the q number of lags for the forecast errors. ARIMA (1,1,1) can be written as:

$Y_t - Y_{t-1}=\phi_1 \ (Y_{t-1} - Y_{t-2}) + \epsilon_t+\theta_1 \ (\epsilon_{t-1})$

2. Vanilla Long Short Term Memory Model (Vanilla LSTM): LSTMs are a special type of Recurrent Neural Networks (RNNs), which are a type of neural network method used for sequence data, which takes care of the vanishing gradient problem in RNNs. LSTM uses a memory unit, the current input, previous output and forget gate output and memory blocks to predict sequences, enabling the model to remember a significant number of values.

3. Stacked LSTM: Stacked LSTMs are multi-layer LSTMs, enabling us to go deeper in our models.

4. Bidirectional LSTM: Drawing inspiration from text processing, bidirectional LSTMs enable us to look at both the preceding and the following values to make time series predictions

5. Convolutional Neural Networks (CNNs): Renowned for their success in image recognition, we employ CNNs in the time series context. We map a sliding sequence segment to the next value in the time series. (eg. [1.3099. 1.3087. 1.3095] mapped to the 4th value in the sequence [1.3091])

6. CNN-LSTM Hybrid Model: The time series input is first read into the CNN and the output of the CNN is then fed into the LSTM for training and making predictions.

7. ConvLSTM Hybrid Model The reading of the input in a convolutional layer takes place inside the LSTM layer in this hybrid model.

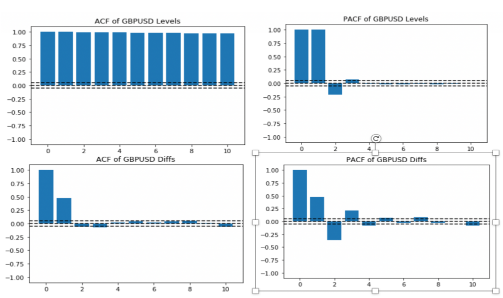
5 Experiments/Results/Discussion

5.1 Loss Function and Evaluation Metrics For our loss function, we are using Mean Squared Error (MSE) in training our models. To compare model performances, we look at the standard Root Mean Squared Error (RMSE). Since the exchange rate fluctuates and magnitudes of error are relative to the true value, we believe a better measure would be the Mean Absolute Percentage Error.

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \text{ , RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \text{ , MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|$$

5.2 Hyperparameter Selection

For the ARIMA(*p,d,q*) model, we chose *p*=1, *d*=1, *q*=1 for our hyperparameters, using the Box-Jenkins method.we determine *d=1* by the non-declining Autocorrelation Function (ACF) and the value of the first lag of the Partial Autocorrelation (PACF) function close to 1. After differencing, we determine *p*=1 by the number of most significant lags in the PACF and *q*=1 by the number of most significant lags in the ACF. While we could have used more lags, we wanted to avoid overfitting and chose 1 lag for both the differenced time series and its forecast errors.

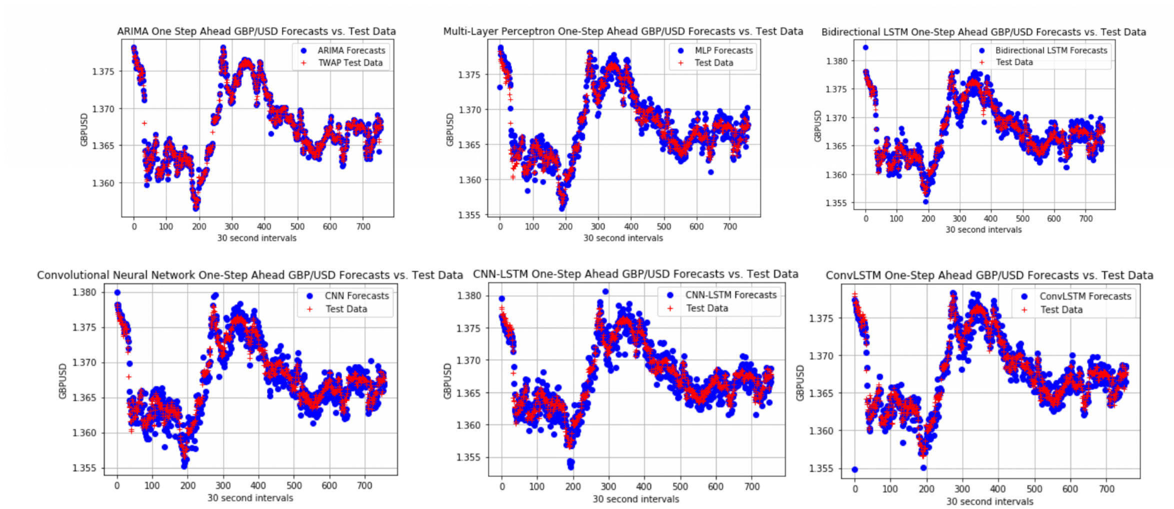


Given the empirical nature of deep learning work, we tried batch sizes of 32, 64, 128, 256 and 512 for the MLPs, LSTMs, CNN, CNN-LSTMs, ConvLSTMs. Batch size of 512 yielded better results in terms of reduced MSE. We attempted 50 neurons and 100 neurons per layer, and decided on 50 upon better performance.

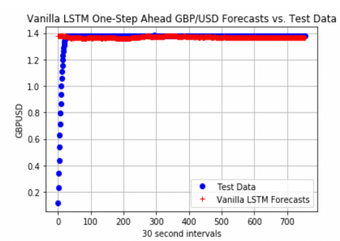
5.3 Test Results:

ARIMA (1,1,1) model clearly outperforms its neural network peers in the forecasts on test data. Visually, the blue forecasts and red test data are clearly closest for the ARIMA model. This is also confirmed by the lowest RMSE and MAPE values we obtain for ARIMA.

Among the deep learning models, we observe that MLP is the top performer, with Bidirectional LSTM coming second. The hybrid ConvLSTM outperforms its components, standalone CNN and LSTM models. MLP's strong performance echoes two papers we reviewed in the literature, but is surprising given LSTM's superiority in modeling sequences. To our surprise, the vanilla LSTM performs very poorly and so did our stacked LSTMs in our experiments. We would have expected a more parsimonious model (vanilla LSTM) to perform closer to ARIMA. We excluded stacked LSTMs from our graphs and both vanilla and stacked LSTMs from the table.



Model	RMSE ($)	MAPE (%)
ARIMA (1,1,1)	0.0004627	0.0226326
MLP	0.0011496	0.0639537
CNN	0.0015240	0.0870874
CNN LSTM	0.0016940	0.0970920
Conv LSTM	0.0015676	0.0771198
Bidirectional LSTM	0.0013143	0.0742361



6 Conclusion/Future Work: God Save the Queen (and Her Prime Ministers)

In conclusion, we note the classical ARIMA(1,1,1) machine learning model has outperformed its neural network counterparts by a landslide. Among the neural network models, the MLP model, simplest by design, has outperformed the other neural nets, followed by Bidirectional LSTM, while ConvLSTM performed better than its standalone components.

For future directions, we have the following suggestions: 1) We would like to study a sliding window to train our model as opposed to an expanding set. 2) We would also like to combine a week's worth of intraday data leading up to a risk event. This would require pre-processing any overnight gap risks in thin, illiquid hours where we may not have many quotes. 3) We would also like to build a model to combine ARIMA with a successful deep learning architectures, such as MLP and ConvLSTM. to see if the combination can beat ARIMA's standalone performance.

Perhaps, in one of the most polarizing issues in the world's oldest democracy and one of the most respected, it is not surprising that either the most traditional (ARIMA), or the models that are bidirectional (Bidirectional LSTM) or hybrid (ConvLSTM) have performed the best in explaining the variation in fluctuations of the financial expression of its image - mathematically underscoring the need for cooperation across the political spectrum embedded in the country's tradition at a time of national interest. God Save the Queen.

7 Contributions & Acknowledgements

Alp Kutlualp completed this project individually. Alp is inspired by Margaret Thatcher, Britain's longest serving Prime Minister of the 20^{th} century, who forecasted these problems and provided their solutions for Britain - and the wider world.

References

[1] Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276.

[2] Gu et al. (2018). Empirical Asset Pricing via Machine Learning. University of Chicago Booth School of Business.

[3] Ryll, Lukas; Seidens, Sebastian (2019) Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey. arXiv: 1906.07786.

[4] Brownlee, Jason (2019). Deep Learning for Time Series Forecasting

[5] Zhou et al (2018). Stock Market Prediction on High-Frequency Data Using Generative Adversarial Nets

[6] Ntakaris et al (2019). Feature Engineering for Mid-Price Prediction with Deep Learning. arXiv:1904.05384v3

[7] Ganesh et al. (2018) Deep Neural Networks in High Frequency Trading. arXiv:1809.01506v2

[8] The Financial Times (Nov 2019) Sterling rise underscores optimism over Tory win forecast

[9] Rundo, Francesco (2019). Deep LSTM with Reinforcement Learning Layer for Financial Trend Prediction in FX High Frequency Trading Systems. Applied Sciences

[10] Chollet, Francois & Others (2015). Keras