Forecasting Stock Price Using Sentiment Analysis and LSTM Networks

Blake Hillier, Grace Li, Joe Puhalla

31 March 2020

Outline

Introduction

- Forecasting stock prices is a widely known problem many people have attempted to solve through various models.
- We propose a model using macro-economic variables to predict the future price of a stock, one of which is statements from the Federal Reserve about decisions on economic policies.
- Our model is comprised of XLNet to perform sentiment analysis on one macro-economic variable and an LSTM Neural Network to combine all the variables while capturing the effect time has on the future stock price.

Workflow

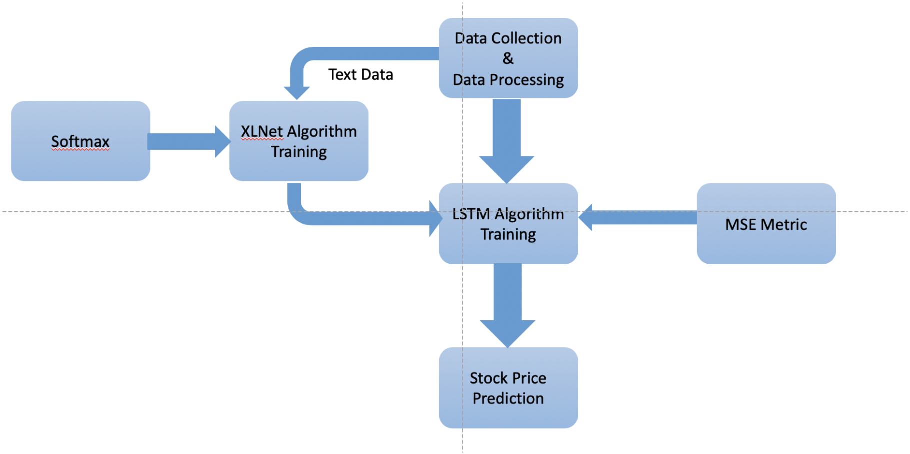


Figure: Workflow

Data Description

	Year Range: 1980 - 2014
Text Data	Federal Reserve issues FOMC statement
Numeric Data	Stock Price
	GDP
	CPI
	Unemployment Rate
	LIBOR
	TNX
Stocks Selection	JPMorgan (Financial Services Sector)
	Microsoft (Technology Sector)
	UnitedHealth Group Inc (Healthcare Sector)

Figure: Data

XLNet

XLNet is an autoregressive pretraining approach for NLP models.

1. Pretraining involves training a model on a generic dataset to understand general patterns within a broad field.
2. Autoregressive pretraining approaches create a conditional probability distribution based on the likelihood function

$$p(x) = \prod_{t=1}^{T} p(x_t | x_{<t})$$

which only sees the relationship between previous text.

XLNet

$$\max_\theta E_{z\sim Z_T}\left[\sum_{t=1}^T \log p_\theta(x_{z_t}|x_{z<t})\right] = E_{z\sim Z_T}\left[\sum_{t=1}^T \log \frac{e^{g_\theta(x_{z<t},z_t)l(x_t)}}{\sum_{x'} e^{g_\theta(x_{z<t},z_t)l(x')}}\right]$$

- Z_T is the set of all permutations of text of length T
- $z \in Z_T, x_{z<t}$ is the sequence of text from 1 to $t-1$
- g_θ transforms x to a sequence of hidden words with the first $t-1$ set of words as additional information

Note: g_θ permutes x and then masks the words

XLNet

In order for g_θ to accomplish this, they split it into two different transforms:

- g_θ which looks at the first $t-1$ words in the permuted order to predict the t^{th} word
- h_θ which simply encodes the first t words in the permuted order

To reduce the complexity, they change the optimization problem to

$$\max_\theta E_{z\sim Z_T}\left[\log_{p_\theta}(x_{z>c}|x_{z\leq t})\right] = E_{z\sim Z_T}\left[\sum_{t=c+1}^{|z|}\log p_\theta(x_{z_t}|x_{z<t})\right]$$

LSTM

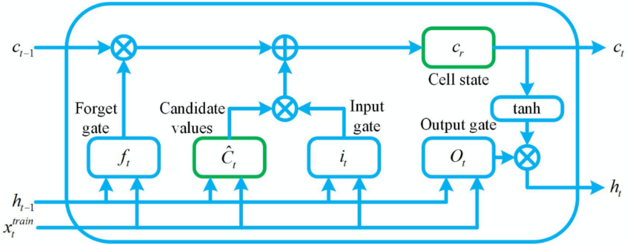


Figure: LSTM Procedure

Cell makes decision by considering current input, previous output and previous memory.Generates new output and alters its memory.

LSTM

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.

- cell: responsible for keeping track of the dependencies between the elements in the input sequence.
- input gate: controls the extent to which a new value flows into the cell.
- forget gate: controls the extent to which a value remains in the cell.
- output gate: controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit.

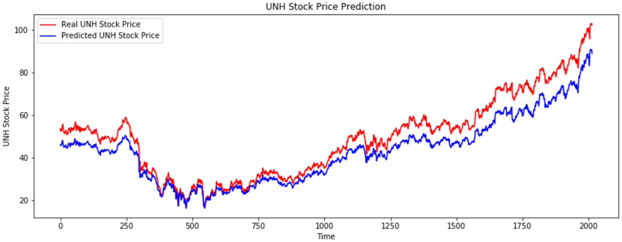
XLNet

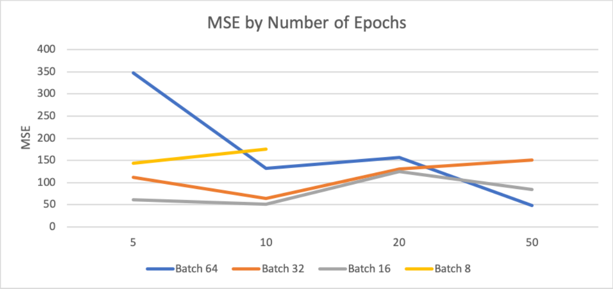
We used pytorch's implementation of XLNet-base for our model.

- Sentiment was assigned to the Fed's Statements by looking at the percent change of the UNH stock
- We used an 80/20 Train/Test split
- Testing was done using a GPU on Colab

After some testing we found a maximum statement length of 128, batch size of 24, and 10 epochs produced the best accuracy of 77.1%.

Prediction


UNH Stock Price Prediction
Real UNH Stock Price
Predicted UNH Stock Price
UNH Stock Price
100
80
60
40
20
0
250
500
750
1000
1250
1500
1750
2000
Time


MSE by Number of Epochs
MSE
400
350
300
250
200
150
100
50
0
5
10
20
50
Batch 64
Batch 32
Batch 16
Batch 8

Main Model

1. We first trained the XLNet on the entire text data, and then predicted the sentiment on the same dataset
2. This was then merged with the input data for the LSTM, and was trained using a portion of the stock data
3. Once trained, we validated it with the last 2014 data points to obtain the MSE: 25.226
4. This is lower than our previous tests with the LSTM, showing the capability of XLNet improving our forecasting accuracy

Main Model

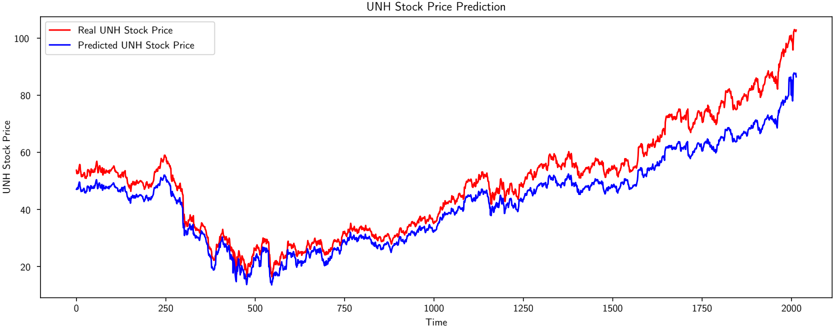


Figure: Forecast with XLNet and LSTM compared with the actual price

Conclusion

- Our model consists of XLNet and an LSTM network
- While our individual results were ok, we showed sentiment analysis using XLNet improved our forecasted results

Future Work

- More macro and microeconomic features
- Use a longer timeframe for data
- Judge final model by simulating a trading strategy