SDS M3 Group Assignment

Andreas Jørgensen, Cathrine Olsen, Louise Christoffersen
Mette Møller

9. December 2020

1 Introduction

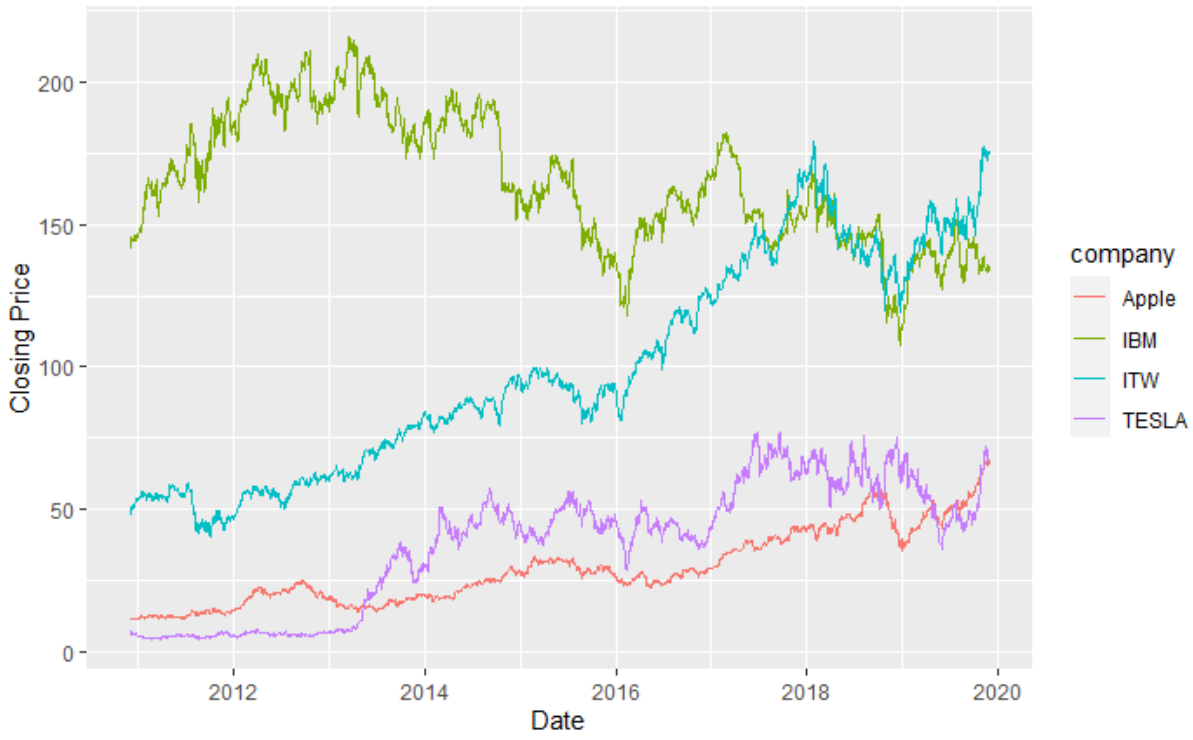
The aim of this assignment is to build a supervised machine learning model and a deep learning model capable of predicting the price of a stock tomorrow. A lot of people has tried to build such a model capable of predicting prices with a high accuracy but without great success. Therefore we did not expect to achieve a model capable of very precise predictions - and if we did we probably would have kept it to ourselves!

Earlier stock investment was mainly accessible for professional stock brokers but since the release of liberty bonds during the First World War securities became available for the general population. Today it is common for ordinary people to hold stocks and the development in stock prices is therefore of interest to the public.

5 different stocks are chosen for this assignment since stock prices often correlate in some way. Apple is chosen as the main stock of interest and the models are constructed to predict the price of Apple stocks tomorrow. The price is predicted based on the price of Apple stocks today along with the price of 3 other stocks - ITW, IBM and Tesla. Stock prices over 9 years are fetched for each stock.

1.1 EDA

The plot below shows the development in each of the stocks in the 9 year period. Some of the stocks seem to be positively correlated and hence move in the same direction eg. Apple and ITW while others seem to be negatively correlated.



A variance-covariance matrix is shown below. As expected the correlation between Apple and ITW is strong and positive whereas Apple is negatively correlated with IBM. Overall there seem to be a rather high correlation between Apple and the other stocks, and therefore it is expected that they to some extend are capable of explaining some of the movements in the Apple stock.

	Apple	**IBM**	**ITW**	**Tesla**
Apple	1.0000000	-0.6571646	0.9106984	0.7844832
IBM	-0.6571646	1.0000000	-0.6600532	-0.6353711
ITW	0.9106984	-0.6600532	1.0000000	0.8768346
Tesla	0.7844832	-0.6353711	0.8768346	1.0000000

2 Supervised Machine Learning

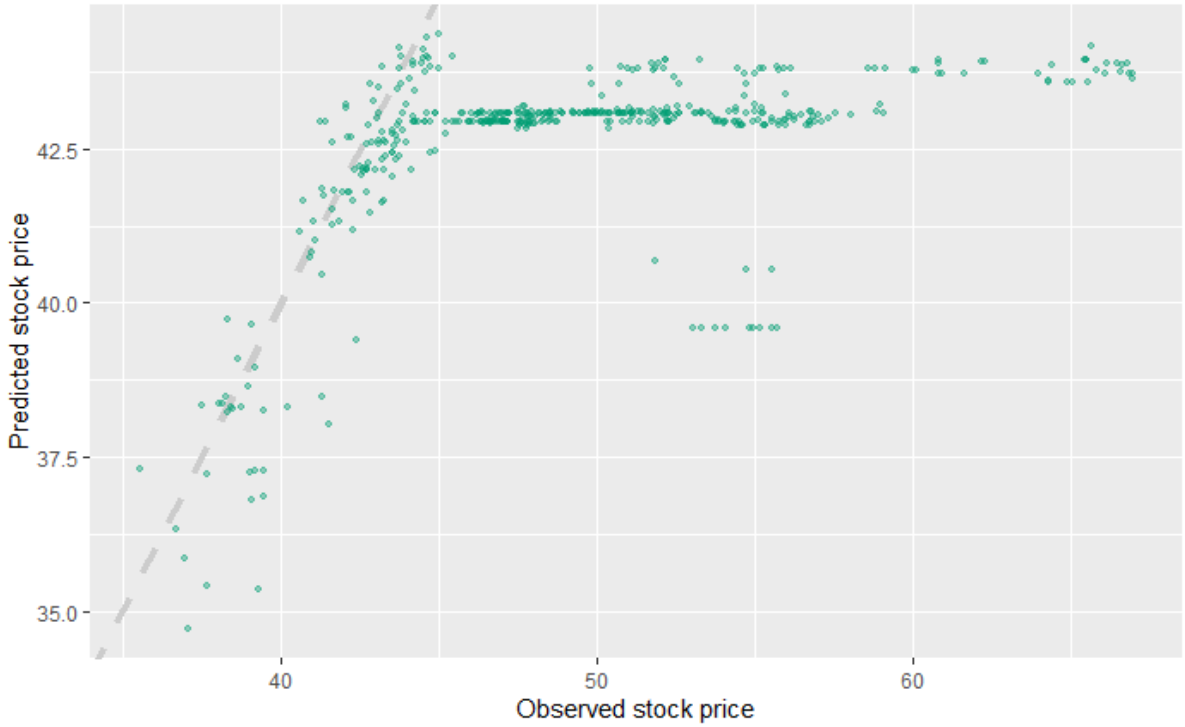
For the Supervised Machine learning part a multitude of different models have been tried and tested, but for brevity only the final and best model is shown here. The data is split into to parts 80 percent for training and 20 percent for testing, this split takes the sequence of the data into account. Additionally, the data has been scaled to insure accuracy. All the models are trained, tested and hypertuned in the same way. The models used are elastic-net, random forest and an Xgboost model. The random forest variant ended up being the better model and was chosen as the final model. The RMSE of training for all the models are shown below.

	RMSE
Random forest	0.00571030
Xgboost	0.01044366
Elastic net	0.01184030

To better show how the final model performs it is tested on the data that was set aside earlier. This result has been scaled back to dollars for ease of inference. Below is a table that shows how much the random forest is off on average when used to predict the testing data.

	RMSE scaled	**Dollars**
Random forest	0.2779173	8.838922

Lastly is a plot of the predicted stock prices compared to the observed/true stock prices. This plot shows that the model is performing well in at first and with the lower prices but struggles to predict the later and higher prices.



3 Deep learning

Ordinary supervised machine learning models as presented in the section above have no “memory” which might explain why the models have a hard time predicting the true values. In order to fix this problem deep learning models are created. These have the advantage of being able to ”remember”.

Four different models are created and fitted: An LSTM , GRU, Bidirectional LSTM and Bidirectional GRU. The results are presented below.

Model	**Loss**	**MSE**
GRU	0.09850515	0.09850515
LSTM	0.1305248	0.1305248
Bi_GRU	0.09871213	0.09871213
Bi_LSTM	0.06418348	0.06418348

As seen by the evaluation of the mse and loss function the Bidirectional LSTM model is the best which is why this is fitted as the final model. The results are presented below:

Model	Loss	MSE	Dollars
Final model	0.003782385	0.003782385	1.95599

The evaluation on the unscaled test data is plotted below and it is seen that the model performs rather well with the exception being the last period where the predictions seem to be around 10 dollars off.