Finding a Dataset

1 Prelude

In our initial discussions we decided that we wanted to investigate a model related to time series. There were many interesting ways we could take this, we'd all previously done projects related to finance so this was an obvious starting point to motivate our work.

2 Dataset Exploration

Initially, we considered looking at stock price data and thought it would be interesting to see if we could predict future stock prices. After some more investigation into this, we decided against it for several reasons. First and foremost was the lack of variables that are involved in stock price datasets. The large majority of datasets we found only contained information for 4 different variables for each day: the opening price of the stock, the closing price, the highest price of the stock, and the lowest price of the stock. One reason we didn't think this would be good is because it didn't provide much in the way of contextual information about why these prices were what they were. Also, we didn't think this would be very appropriate for a project which was centred around exploring a time series by a neural network approach. We looked at others attempting to do work involving stock prices and saw that on the whole it hadn't gone down very well, due to the unpredictable and random nature of stocks many models were virtually useless in terms of prediction and we didn't think this would make for a very interesting project. Human behaviour effects stocks which is a lot trickier to get data for. This however, would be quite interesting to model via a neural network approach.

The next area we considered was the interest rates in the UK. Our main obstacle for this was combining a vast amount of datasets. We first considered what sort of factors might impact the interest rate and looked at datasets for things such as the unemployment rate, inflation, GDP, growth, and credit. We found that a lot of the datasets provided data for different time periods, some were given quarterly, some monthly, and some daily. We didn't see this as a massive problem as we could extrapolate some of these datasets and cut out some of the data from others so that they were all done on the same time frame, ideally monthly. However, we still had the problem of combining these datasets correctly and thought it would require too much effort and nuance just to set up the data before we'd even got into the meat of the project. Furthermore, since the data was only collected monthly, we'd require datasets that date back very far in order to compile training and test sets of an acceptable size to then conduct our work involving neural networks. We believe that given more time, this would have been interesting to look into due to how similar it would be to situations we'd encounter if in the future we had to do some data science work involving massively parallel processing.

We then decided to look at other prediction projects that people had undertaken on Kaggle relating to time series and found an interesting dataset involving air pollution in Beijing. This dataset contained a good amount of relevant features such as temperature and dew concentration and had been collected hourly over a five-year period. These characteristics remedied problems some of us had seen earlier in Time Series module so we felt confident in performing EDA analysis and implementing standard time series models like VAR to compare with the neural network approach. Also, the dataset had useful features that pertained to what we were investigating and would allow us to compare if the added complexity was necessary to forecast time series or if it simply added noise. Additionally, all of the data was in one file, we wouldn't have to spend time tediously and meticulously manipulating and combining datasets to be in the form that we wanted and could crack on with the work we were more interested in. The modestly large amount of data also meant that we could build a good idea with how our models would perform on scale. Furthermore, there were plenty of resources we could use for inspiration that involved the massively parallel computing techniques that we were interested in that we could draw inspiration from and that could advise us on our journey. In a 'real' world setting we might encounter more difficulty when it comes generating and processing a similar dataset. The dataset we have chosen has all the necessary information in one place with no missing values, however in the real world we're likely to see several missing values potentially due to problems with the machinery measuring the features, or human error when combining into a dataset. We'd have to compile the data from a variety of measuring equipment into one complete dataset, similar to what was discussed earlier for the interest rates, which proved difficulty.

3 Splitting the data

Now we had to consider how to do the training and test splits for this data that were appropriate for our respective models. We deiced predicting the future would be the best option so left out the last 10% of data. The standard time series models will likely struggle with so we will experiment with different splits and decide on the best option. Whereas, this should prove no difficulty for the LSTM which predicts sequentially. In training the LSTMs, we feed in a sequence of 10 records to predict the pollution of the 11th which somewhat has a ”Rolling Window” effect. So for the more standard VAR and ARIMA we performed Rolling Window CV which is the most appropriate for predicting the future. Unfortunately, LSTMs don’t feature cross validation techniques, so we instead focus on finetuning to speed up the lengthy process while maintaining good outcomes. We also looked into performing expanding window, but this would accumulate a lot of data on a large scale and since the goal is a model that could be generalised to much larger datasets for massively parallel modelling we believe that the Rolling Window is more appropriate.