Andrew Thomas Kent
1 Point Street
Providence, RI 02903

Katie Rouse
XR Trading
550 W Jackson Blvd STE 1000,
Chicago, IL 60661

February 23, 2022

Dear Katie,

This letter is to explain a little of what my code does, how I attacked this challenge and also my results. I'll go over the major points briefly, and I would be happy to go into greater detail if your team would like.

1. **Processing Data**

   (a) *Cleaning up and Rearranging Data*

   First I noticed that there were repeat rows in the data, so I deleted them and then decided to rearrange the data based off of the ordering of the number of days after February 1st. The simple thinking behind this is that SO_2 levels would be most highly correlated between adjacent days more than any other variable. This argument is less strong when there are a lot of date gaps in the data.

   (b) *Scaling Data and Checking Stationarity*

   In order to properly predict future SO_2 values using my model I would need to make sure the data for SO_2 was stationary and properly scaled - I did a couple tests after scaling the maximum value to 1 and the minimum value to 0 and found it was acceptably stationary.

   (c) *Splitting Data*

Knowing that I would eventually use a neural network to predict future values, I needed to have testing and training data which both had values for SO_2. What I decided to do was to split up *TRAINING_DATA* into two parts: a (sub-training) set with 80% I would use for training the NN and the remaining 20% (a sub-validating) I would use for validating the results of the NN.

2. **Choosing the Model**

   (a) *Using a CNN-LSTM w/ Self Attn*

   Knowing the amazing results from DeepMind recently in predicting rainfall, I decided to try and use a NN to predict future SO_2 values, after looking for a little bit I decided to use a model a paper I found which forecasts strawberry yields based on satellite imagery, soil based parameters and temperature data. I thought this would do a good job of predicting SO_2 values because the model from the paper looked to do a good job at separating relevant cues in the data.

   (b) *More Specifics*

   Of course there are a lot more things I included in the code than what I can speak about in a short paragraph, so I'll just elect to talk about two of the more important things I did.

      i. *Hyper-Parameters Used*

      Firstly, the model is giant compared with the size of the data set - so to ensure I wasn't over-fitting to the sub-training set I made the dropout rates for the $LSTM$ cells very high ($\approx .9$). I also kept the window size for the $LSTM$ cells relatively low (≈ 3), this was because the data is sparse and ordered on date.

      ii. *Multiple Predictions Averaged*

      Just to give my predictions the best shot at getting good values for SO_2 I decided to average over evenly spaced predictions during the training process, I am also able to see each of the different stages which helped me find better hyper-parameters to use. These different predictions can be seen in the plot below.

3. **Results**

   (a) *Accuracy of Predictions: R^2*

   I was able to get $R^2 \approx .1$, which is not exactly close to what I originally was shooting to get. The way I calculated this was by training my NN on the sun-training data and then making predictions on the sub-validation data. Those predictions I then compared with the true SO_2 values from those days to find the

R^2 value. What I found was that the model did pretty well with the data which was available - but it clearly could have done a lot better if I had been given minute-by-minute data (or something close to that fidelity).

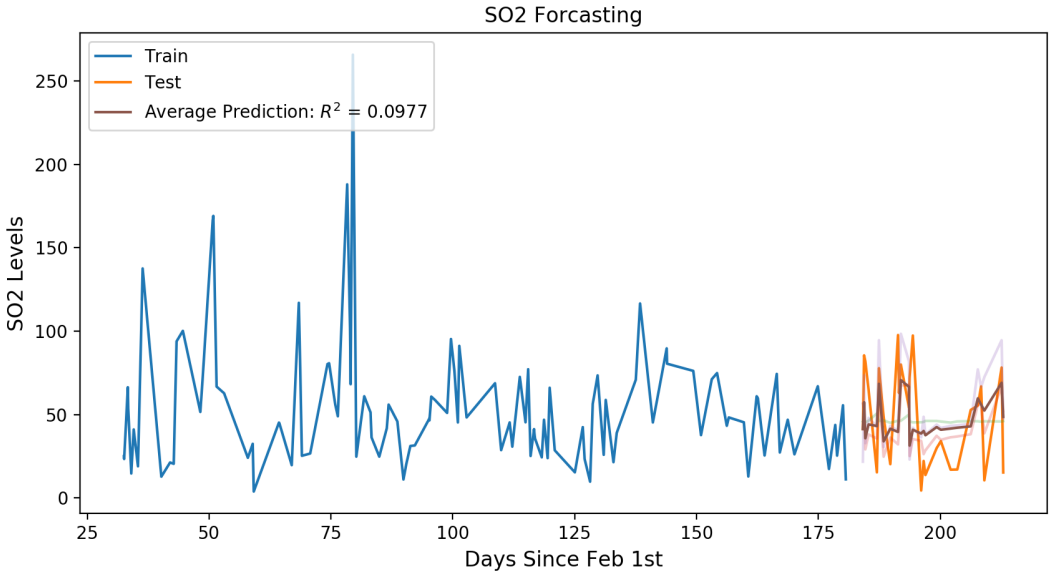

Figure 1: SO_2 Predictions on Sub-Testing Data

(b) *Future Predictions*

After running the NN to fit on the sub-training data-set, I ran the $TESTING\_DATA$ through the NN to find the predicted values. Those values were stored onto a *.csv* and then I just transferred them to the original *.txt* data-set which was sent to me. One additional comment about the predictions, the way my NN works is it takes in a few rows of data corresponding to the window size and it predicts the next SO_2 value - so when looking at the predictions you'll notice that the first three rows are zeros which corresponds to the NN not being able to make predictions before it "sees" three rows of data to predict the fourth row's SO_2 value.

The predictions are included in the folder I attached to the email, also I included *.png*'s of the residuals and the training and validation loss in the folder as well

Sincerely,

Andrew Thomas Kent