Open Vaccine Competition

Carson Wilde

- **Problem:** COVID vaccine researchers are struggling to predict degradation of RNA sequences. We need to develop intelligent models that can predict the likelihood of decay at any given point in the sequence.

- **Data:** We are given three sequences of data. The RNA sequence, a sequence describing the structure of the RNA, and a sequence predicting the type of loop in the RNA. An example of this data can be seen in this figure.

```
sequence            GGAAAAGUACGACUUGAGUACGGAAAACGUACCAACUCGAUUAAAA...
structure           ......((((((((((.((((.....))))))))))(((((((.....
predicted_loop_type EEEEEESSSSSSSSSSBSSSSSHHHHHSSSSSSSSSSSSSSSSSHHH...
```

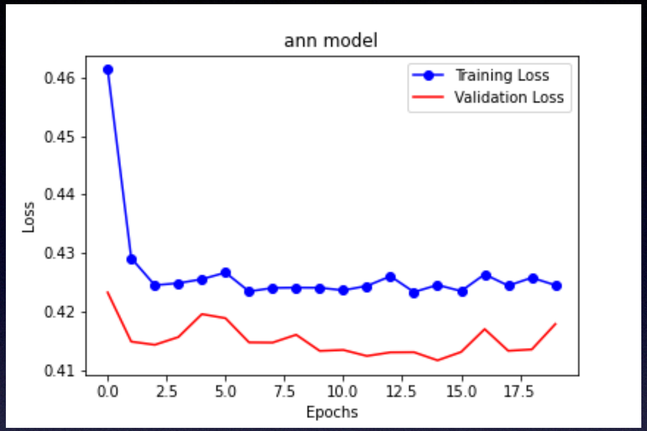
- **Approach:** Two approaches were taken during this project. First, I tried a linear ANN. Second, I attempted to take advantage of the sequential structure of this data and used a recurrent LSTM model.

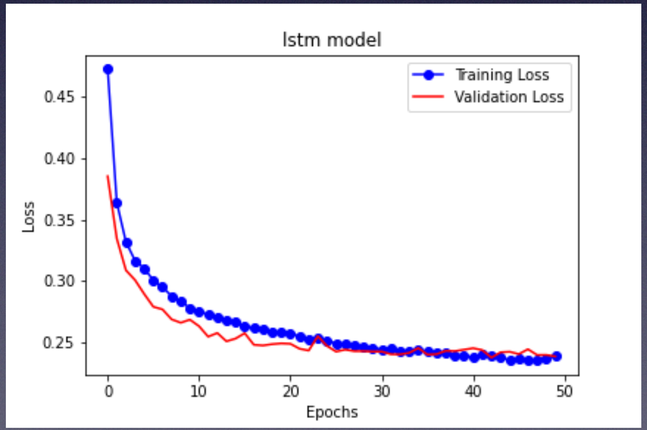
- **Analysis:** As supported by the two graphs to the right, The ANN showed a slight decrease in error initially, but was not able to train significantly after. The LSTM model, however, showed impressive growth, and far outperformed the ANN.

- **Conclusions:** The LSTM model performed well while the ANN refused to train. This demonstrates that a recurrent structure is necessary for this data. This model was able to score 0.28712 on Kaggle's the public leader board.

- **Source Code: https://github.com/Cwilde921/data_science_project**

- **Competition: https://www.kaggle.com/c/stanford-covid-vaccine**





Name	Submitted	Wait time	Execution time	Score
my_submission.csv	5 days ago	0 seconds	6 seconds	0.28712

Complete