Final Report: Predicting US Diabetes Prevalence

Model Metrics

By Aisling Casey – July 29th, 2021

An LSTM deep learning network, accepting sequences 4 time points (years) in length, was the final model selected. It was created using the TensorFlow and Keras libraries.

Model Parameters

Architecture

- The model was created using the Sequential class
- The first layer was an LSTM layer, which accepted arrays of shape (4, 25)
- The final layer was a dense layer that returned one value.
- Because this was a regression problem, there was no activation function in the final layer.

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_3 (LSTM)                (None, 50)                15200
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 51
=================================================================
Total params: 15,251
Trainable params: 15,251
Non-trainable params: 0
```

Figure 1- Summary of the model architecture

Hyperparameters

- LSTM parameters: defaults, such as activation='tanh' 'and recurrent_activation='sigmoid'
- Optimizer: 'adam' with defaults, i.e. learning rate=0.001
- Loss function: mean absolute error (MAE)

Weights

The model weights are available in an hdf5 file located in the project directory under 'models'. The file is called 'best_model_LSTM_4.hdf5'.

Model Performance

Using the 2017 data as a validation set, the model selected yielded a mean absolute error of 0.78%, compared to a mean absolute error of 0.88% when using 2016's prevalence values as the prediction for 2017. Its performance across population quartiles, as compared to using 2016's value as a prediction, is shown in figure 2.

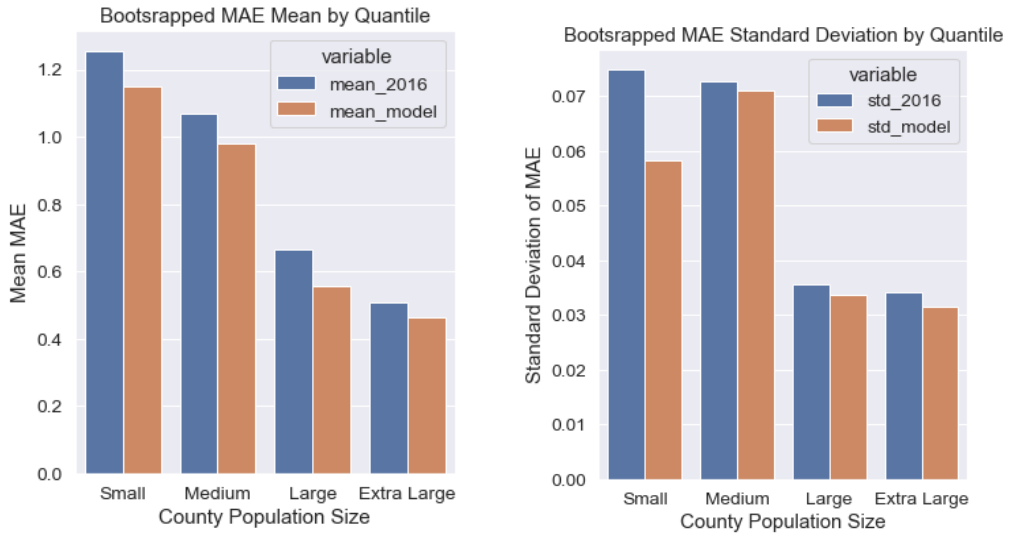

Figure 2 – For both the final model and using 2016 as a prediction for 2017, mean and standard deviation of bootstrapped MAE for each population quartile

To get a sense of whether the model is at risk of overfitting the data, 50 LSTM n=4 models were created and their MAEs recorded, as were 50 LSTM n=5 models for comparison. The results are shown in figure 3. In the case of the LSTM n=5 models, the distribution of MAEs covers a larger range and is not at all normally distributed. The LSTM n=4 models yield a smaller range and a more normal looking distribution. Thus an n=4 model is less likely to overfit the data.

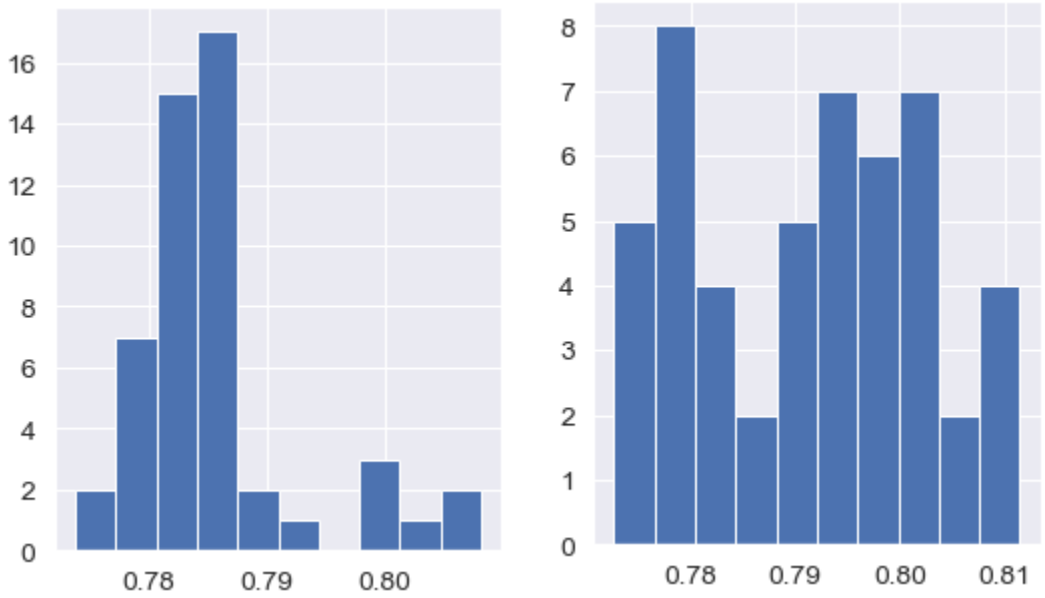

Figure 3 –Histogram of MAEs from 50 LSTM n=4 models (left) and 50 n=5 models (right)