Error Reporting

I think it would be useful to have the data structured like for the incidences in the **test** set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample (yr, county) | Actual Cases | Error Between Predicted and Actual Cases - Baseline | Error Between Predicted and Actual Cases – MLP | Error between Predicted and Actual Cases – LSTM Model (can do by particular model or across all model types, or just best model type. I would recommend best model type). | Error between Predicted and Actual Cases – xLSTM Model (can do by particular model or across all model types, or just best model type. I would recommend best model type). |

Which gives you a dataset for all test cases.

What we are looking for is – is there a statistically significant difference between the predictions from one model to the next by sample input? For instance, do the models usually predict Kern county accurately but have huge outliers of different magnitude for Pasadena county predictions? This gives you a better idea of model performance.

Here are my recommendations.

1. Pick a high-performing year. Graph the actual cases against the predictions for the best model of each type by county, and include this in the paper. This will let readers get an idea of variation for each model in each county.

2. Perform a test of statistically significant difference between each model type prediction, using the best models, by year and county. I would recommend a Wilcoxon signed rank test (<https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test>, <https://www.sciencedirect.com/topics/medicine-and-dentistry/wilcoxon-signed-ranks-test>) if the differences between the different models are not normally distributed. Basically, you’ll have a distribution of input pairs (samples of county and year) for testing statistical significant differences between each model.

For example, testing if there is a statistically significant difference between the LSTM and xLSTM models would have you give as input to the Wilcoxon signed rank test a list of errors (predicted vs. actual cases) of the LSTM model for each year and county and a then a list of errors (predicted vs . Actual cases) of the xLSTM model for each year and county). The output of the test less than or equal to 0.01 or 0.05 usually indicates a statistically significant difference. I believe sci-kit learn offers this.