Weekly Progress Report: Nov 6 - Nov 10

Contributors: Ryan Choe, Rakin Hussain, Sid Taneja

Progress/Accomplishments:

- Our team has completed the data preparation process
    - The process of data preparation included making a copy of the train and test data frames for each of the six countries. These new copies were then named “train_imputed” and “test_imputed” so that way we could have data that remained untouched with respect to imputing missing values and also copies that had been touched. This was done in preparation for developing models since some are robust to missingness (ie. ARIMA and AutoARIMA) and using any imputed data would introduce further bias
- We have also completed the feature engineering process of our project, implementing a couple of things. First, we added lag features to our model after looking and analyzing ACF and PACF plots of our response variable (new cases). The plots showed a clear spike in auto-correlation in lags 7 and 14, so we added those columns to our dataset for each country. In addition, we added a 14 day rolling mean of new cases, and this column shows a clear correlation with the response variable. Finally, we added a ‘weekend’ column to the dataset that shows whether the day in the time-series was a weekend or not. Again, this column shows a clear correlation with the response variable.
- In addition, all members of the team have begun to develop models now that our data has been cleaned and prepared. We have been developing each model individually as instructed, and as of Nov 10th, we have ARIMA, Auto-ARIMA, and XGBoost models completed individually. As explained more below, we plan to complete more models and then discuss collectively as a group over the weekend and early next week.

Problems/Challenges:

- The last remnants of data preparation included some tedious challenges:
    - For each country’s train and test dataframes, it was necessary to read and evaluate whether or not strategies such as filling missing values with zeros, backward fill or forward fill were the most appropriate. Furthermore, the number of missing values for certain variables drastically varied and had to be imputed in separate groups depending on the quantity of missing entries. Multiplying this by six countries and there were 12 data frames that required this process. Creating the copies of imputed values on the test data

frames was not as tedious since almost all simply required the forward fill technique.

- In terms of models, our team faced a few challenges:
    - ARIMA Model: determining the parameter values for p, d, and q required extensive research and some trial, but we were able to understand what needed to be done eventually.
    - Auto-ARIMA Model: This mostly went smoothly, except the package downloaded (pmdarima) was causing some issues. Stack Overflow helped solve this issue.
    - XGBoost Model: Scaling the data was something that we oversought when preparing the data. As this was the first multivariate model that any of us developed, we scaled the data using Standard Scaler before proceeding with the model.

Plan/Next Steps:

- Over the weekend, our team will continue to add and develop models individually and check in over the weekend and early next week to discuss whose initial model was the best. After comparing whose model is the best, the other two members will learn from the third member what/why their model does better. Then the other two members will improve their initial model and get a better/more accurate model and hopefully have an accuracy as close to the third member.
- On Monday, our team will meet and discuss our individual models. We will choose the best models after discussion, and try to replicate the best model individually.
- Our team will also create and rehearse our final presentation.