Review of G11's Project – Forecasting GNI per Capita

Reviewer ID: R9

To make predictions regarding GNI per capita by machine learning models is the issue discussed in Group 11's project. Techniques like K-nearest neighbors, Recurrent neural networks and Tree-based methods were applied during the process. It is found that Recurrent neural networks outperformed other models with 705 root-mean-square deviation (RMSD) on validation set if prediction accuracy is the main concern. However, one might prefer to use Tree-based methods like Random forest towards this task if the interpretability of the result is more important.

One of the most attractive highlights in Group 11's project is missing values handling. Instead of simply removing all observations with any missing entries or merely setting missing values to the mean of observed values, K-nearest neighbors was applied. Since data are really scarce sometimes, especially for this kind of economics-related data which is expensive to collect, the aforementioned implementation does help maintain the size of the dataset.

Another impressive part is the regularization technique, Dropout, used in RNN. The key idea is to randomly drop units from the neural network during training thus prevents units from co-adapting too much. This state-of-art technique significantly reduces overfitting and gives major improvements over other regularization methods. We see that there is no significant difference between the RNN performances of training and validation sets in this project.

Although there is no doubt that this project is awesome with considerable workload, there are still some deficiencies as far as I am concerned. The very first one is about the datasets. From the chart "properties of data sets" in the appendix, it seems that different models were trained using different datasets. RNN was trained on the dataset where all missing values were removed while Random forest was trained on the dataset where all missing values were imputed by KNN. Even, Single decision tree was trained with missing values preserved. Details that explicitly explain this approach are supposed to be there; otherwise one could criticize that these three models would not be comparable since they were not trained on one certain dataset. Also, to make the model with KNN imputation statistically meaningful, one could always impose a requirement. For instance, the imputed features must have missing data for less than 30% (or some other threshold values) of total observations.

Moreover, one of the explanations, the training RMSE of the training and validation sets are different because the model used on the validation set is trained using all data not labeled as validation data, might be problematic. First, the model used on validation set should be the same as the one used on training set. The difference between the model performances of training and validation sets should not be issued like that. Second, how was the model trained using data without labels given the condition that all the methods used are supervised? The authors should explain more clearly here.

To go further, it would be reasonable to try methods based on linear regression. If the linear models are comparable with those non-linear models above, then linear models are preferred since they are not computational expensive. Plus, RNN is likely to perform better if the tuning parameters could be tuned further with the help of GPU and Cloud.