**DATA SCIENCE PROJECT 2 (MAST90107)   
Semester 2 2021**

**GROUP 17**

## Meeting No. 8

**Date & Time:** 15th Sept 2021, 10AM - 11AM

**Agenda:** Change of tasks and priorities

**Attendees:** Kartika Waluyo, Roberto Bonelli, Vrinda Rajanahally

**Discussion:**

1. Re-routing tasks and switching responsibilities
2. Running xgboost in parallel
3. Structure of the report
4. Creating a structure for 2 presentations

* CSL-WeHi: aim to make it 45 minutes long
* Unimelb: aim to make it about 20 mins long

**Resources/Next steps:**

1. Run the XGBoost model in parrallel - <https://topepo.github.io/caret/parallel-processing.html>
2. Guide to running models in parallel <https://privefl.github.io/blog/a-guide-to-parallelism-in-r/>
3. Run the XGBoost model with the least Nrounds and highest ETA for 50 genes at random.
4. Come up with an introduction for the report
5. Put down a list of all things done (sequentially) during this semester so far.

Meeting No. 9

**Date & Time:** 16th Sept 2021, 10AM - 11AM

**Agenda:** Presentation Outline, Results of XGBoost

**Attendees:** Brendan Ansell, Roberto Bonelli, Kartika Waluyo, Vrinda Rajendar Rajanahally

**Discussion:**

1. Report Outline
2. Presentation Outline
3. XGboost - results, parallelisation and genes chosen for training

**Next steps:**

1. Make a Google Doc folder with all docs and share with team
2. START REPORT!
3. Might end up doing 2 presentations for the clients - after 19th Nov. 2021
4. 23rd Nov, Tuesday - presentation for WeHi (TENTATIVE)

25th Nov, Thursday - presentation for CSL (TENTATIVE)

1. COMPARE THE CORRELATION RESULTS and prediction capabilities WITH THE BASU ET. AL. RESULTS
2. Train the XGBoost model on 50 random genes
3. Find the best parameters to use across both the modelling methods - find the parameters that give the best results - Avg out the loss across all parameters and select the combination with the least avg. loss.
4. Another method: Take prediction of NN, prediction of XGBoost and then combine the prediction together - to check if this works better by evaluating the correlation.

XGBoost lacks in predictability for many outcomes, but it can handle predicting one gene very well. NN works better when it comes to predicting many genes, and potentially across all tissues.