Task Report

1. I found that the species code (SPCD) column is unique in the fia_ref_species_table, and that it is also present in the fast fuel tree list – it can thus be mapped to achieve other features like Plant Functional type (PFT), tree genus, and tree species. Therefore I conclude that our task is to infer SPCD for every single tree in the TLS treelist in order to obtain the aforementioned features.

Data sources used are:

1. FIATreeSpeciesCode_pft.csv
2. REF_SPECIES.csv
3. 01_plot_identification.csv
4. 03_tree.csv
5. terrestrial-lidar-scans-tls-and-derived-tree-lists-for-field-sampled-plots-for-uc-climate-actio/TLS_treelist.csv
6. FF_treelist_all.csv

Preprocessing includes:

1. Infer site_name for tls treelist by mapping plot_blk values obtained from plot_identification
2. Convert tls treelist's DBH to meters
3. Convert ff_tree DIA to meters
4. Remove rows from ff_tree where its SPCD occurs fewer than 20 times (essentially filter out rare cases)

Procedure afterwards:

1. Train-test split
2. Hyperparameter tuning using optuna on catboost
3. Train train+valid set with best weights and evaluate on test set
4. Make predictions on TLS treelist, map the corresponding features
5. Plot the predicted PFT, genus, and species against field collected data at a chosen site

2. The model achieved ~0.93 test accuracy with the best weights gained from 20 iterations of hyperparameter tuning. The model performed fairly well on predicting the SPCD in the fast fuel dataset. Since there are no labels on the TLS treelist, we can only infer our accuracy of predictions through examining the predicted vs field distributions of each feature.

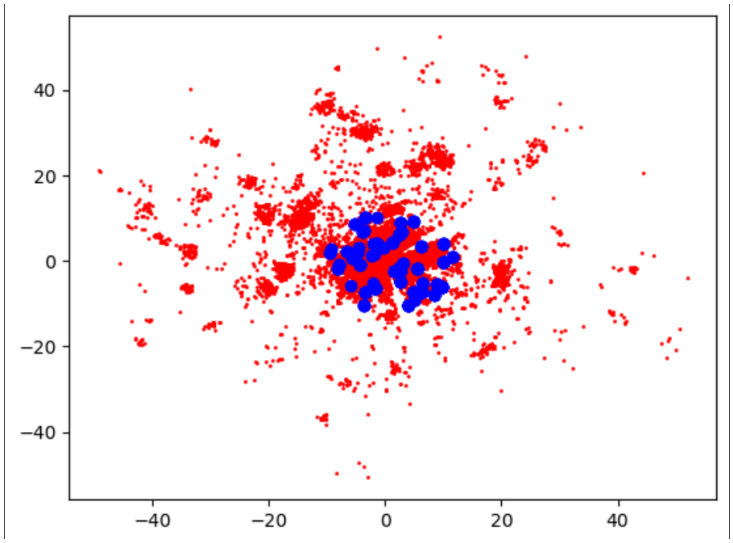
3. The predictions are fairly representative of the actual field data as, in the SHA site, it reflects almost identical distribution of PFT, and have the same top 3 entries of the genus and species compared to the field data distributions. However, it fails to match genus and species of those that are not well-represented in the site.

4. Field-collected data are human-collected data, so I would conclude that they are reliable. However, from the predictions, we find more diverse patterns of tree species at the same site. This could imply that field data may have ignored certain species that are less representative in the area.

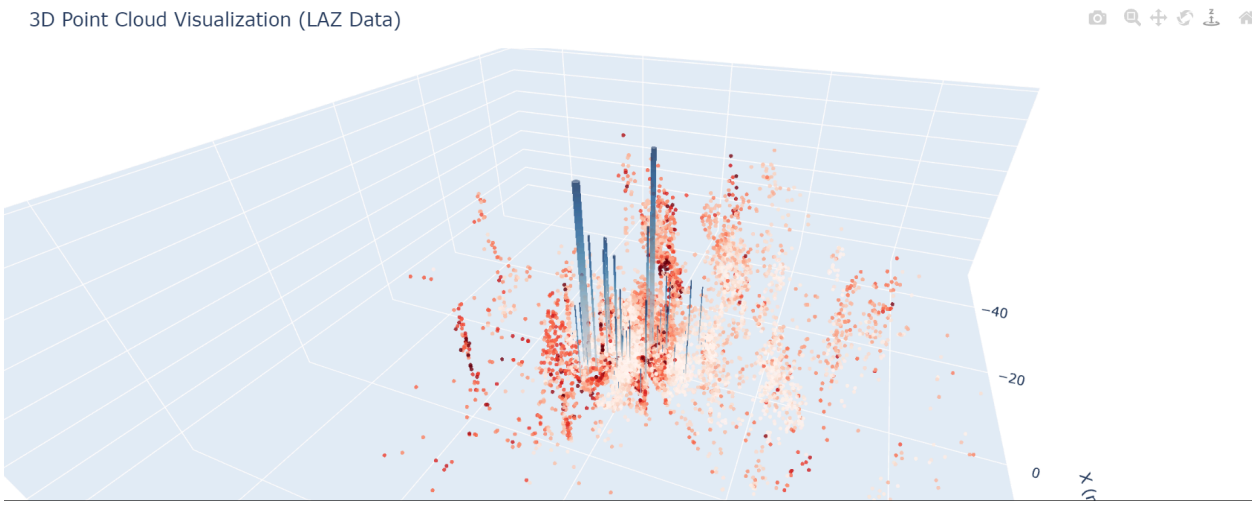
5. Hyperparameter tuning is crucial to the improvements of both validation accuracy and inference quality. There are some limitations to how the ff_tree's DIA may not directly correlate with tls_treelist's DBH, as one is reported at 1.4M height while we have no information about how the other is collected. We assumed that they would have a comparable distribution, but this may be a problem.

6. We have tried using the lidar data as our features but failed due to coordinate mismatch. Specifically, our idea was to locate each tree using the ff_tree dataset's plot_blk and corresponding x and y coordinates; define a cylinder with the radius and height around the coordinate of the tree; any point that falls within this range is considered the tree; and therefore we can train in a supervised manner to infer the SPCD.
However, as we tried to locate the trees, it seems that the coordinates in the tree list do not match the actual trees in the lidar data (the scale seems to be different, xmin, xmax for ff_tree is -11 to 11, while it's around -50 to 50 for the actual lidar data). We have visualized this inconsistency in the next page. This could be due to the fact that the ff_tree dataset is not field-collected. We then inspected the fia dataset and cannot find any information on the coordinates of the trees. Therefore, it would be great if there's actual coordinate of the trees and a method to scale them to the lidar plots, so that our ideal approach may be achieved.

2d plot of sampled x, y coordinates of trees from ff_treelist and lidar points:



3d plot of tls points with our defined cylinders for each tree:



7. For our current model, it would be nice to have information of the trees collected at 1.4M height, so that it matches those in the TLS dataset. This would enhance the reliability and accuracy of our inferences. Further, as explained above, it would be nice if we can have a way to locate each and every single tree, so that we can extract more features for the TLS dataset that may be used for inference, e.g. crown ratio, CBH, etc.