Fire-Ready Forests Data Challenge Report

Team: Pranav Rajaram, Atherv Vidhate, Zubin Sannakkayala, Abhinav Chinnam

Our overall goal for this project was to build a model that could predict Plant Functional Type (PFT), Genus, and Species distributions of tree plots based on Terrestrial Laser Scanning (TLS) data.

---

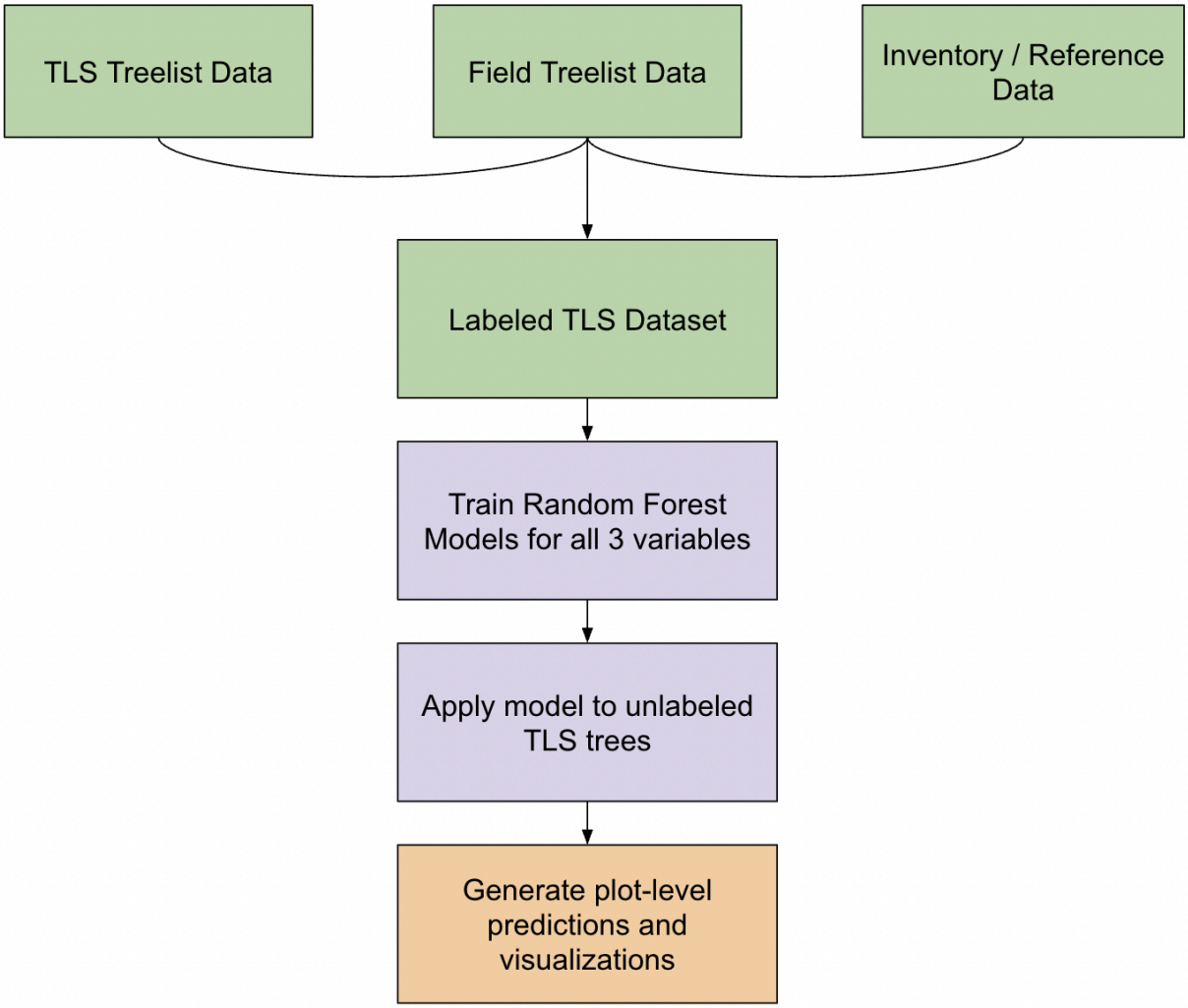
1. Can you describe your full pipeline? Which data did you use? What kind of data preprocessing tasks did you perform? Can you visualize this pipeline through a flowchart?

Our pipeline begins with the integration of multiple datasets: TLS-derived tree measurements, field-collected tree labels (including PFT, Genus, and Species), and mapping tables such as plot inventories and species reference data. The key preprocessing step was to merge these datasets by matching trees using plot_blk, TreeID, and inventory_id, enabling us to label a subset of TLS observations with classifications collected in the field.

275 of the 1,382 TLS trees had a valid match in the field dataset, which became our labeled training set. We engineered a new feature called crown volume (calculated as Radius² × Height), modeling the crown of a tree as a cylinder. We observed a significant amount of class imbalance, particularly in the PFT column where most entries were "Evergreen conifer." To address this, we applied oversampling using RandomOverSampler from the imblearn package before training our classifiers.

Each classification task (PFT, Genus, Species) was handled by training a separate Random Forest model. The features in the model included a tree's diameter, radius, height, basal area, and crown volume. Once trained, these models were applied to the remaining TLS-only trees for prediction and analysis. The final output was an interactive tool that allows users to visualize the distribution of PFT, Species, and Genus at 112 different plots.

A flowchart of our overall pipeline is shown below.



2. How well did your model perform across these classification tasks?

The model performed best on the PFT classification task, achieving an F1 score of approximately 0.96. However, this high score reflects the dominance of a single class (Evergreen conifer) in the training data. For Genus and Species predictions, the models achieved more modest F1 scores of around 0.50, which is still a promising result considering the limited and imbalanced training set. These scores suggest that while the models can detect patterns in the structural features of trees, more data and diversity are needed for precise taxonomic predictions.

3. Were the predicted distributions representative of the actual field data?

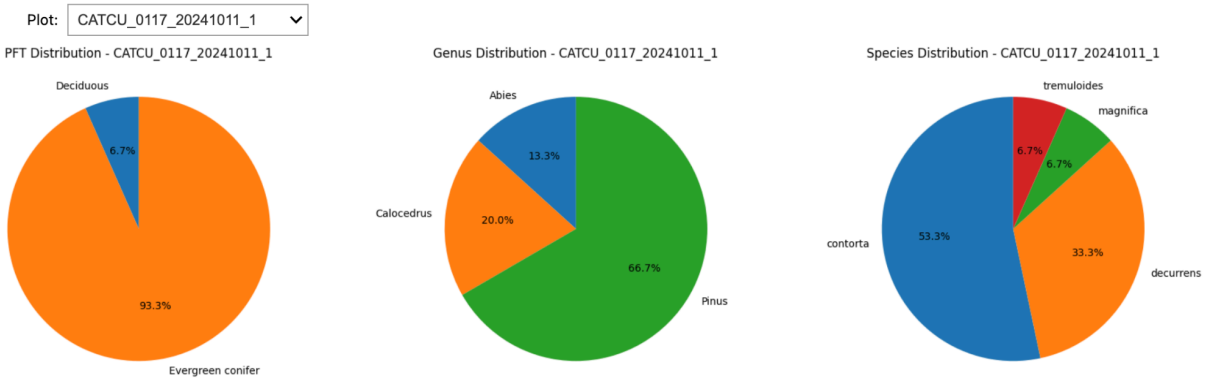
Yes and no. The predicted distributions generally mirrored the imbalances present in the field data. For example, without oversampling, the PFT model predicted almost exclusively Evergreen conifers. After oversampling, the predictions became more balanced, with some Deciduous trees appearing in the outputs. This suggests that the model learned from the field data patterns but was also highly influenced by its skewed class distributions. In summary, the predictions reflect the field data well but definitely also inherit its biases.

4. Do you think that the field-collected data was representative of the surrounding site?

No, the field-collected data was not fully representative. Out of over 1,300 TLS-observed trees, only 275 had matched field labels, and within that subset, class diversity was limited. Certain Genus and Species categories were underrepresented or missing entirely, which limited the training data's ability to reflect the full ecological diversity of the TLS sites. This lack of representativeness likely introduced bias into both model performance and the interpretation of model outputs.

5. What worked well, and what were the limitations?

Several aspects of the pipeline worked well. The model was able to effectively predict PFT with high accuracy. We also thought our feature engineering added meaningful information to support classification. The modular structure of the notebook allowed us to easily adapt the pipeline for PFT, Genus, and Species prediction tasks, and we thought the final interactive tool was a clean and easy way to view the results of our modeling.



However, there were also notable limitations. The small size and imbalance of the training dataset significantly hindered the generalizability of the models. The oversampling method helped, but synthetic balancing cannot fully replace diverse and representative data. Furthermore, relying solely on structural TLS features made it difficult to distinguish between fine-grained taxonomic labels like species.

6. What strategies or techniques could improve your current pipeline (e.g., feature engineering, additional data sources, advanced models)?

There are several clear paths for improving the pipeline. First, we could explore more advanced models such as XGBoost or LightGBM, which often perform better with limited data. Hierarchical modeling (ex: predict PFT first, then Genus and Species conditional on that) could help incorporate biological relationships into the process. On the feature side, we could add variables like canopy base height, crown ratio, or shape asymmetry, and also experiment with Cross Validation to better tune the hyperparameters in the Random Forest. Dimensionality reduction techniques such as PCA might also help in generalizing across trees. Finally, experimenting with synthetic augmentation beyond simple oversampling (like SMOTE) may help address the imbalance more effectively.

7. What new types of data could be included to enhance your model's predictive power?

Additional data sources could greatly enhance model performance. Environmental variables such as elevation, slope, and soil type could add important ecological context to the data. Hyperspectral imagery or NDVI data would provide information about foliage composition and health that structural data can't capture. Time-series data showing growth patterns or seasonal changes could also aid in species-level identification. Together, these additions could create a richer feature set for more accurate classification, allowing us to build on our progress.