Building Footprint Identification with Airborne LiDAR: A Final Project for FOR 796

Lucas K Johnson

Abstract

This is a test.

1. Introduction

2. Methods

2.1. Building and LiDAR Data

Rasterized building footprint data served as the response data in this study Microsoft (2018). The raw raster contains counts of the number of buildings intersecting each pixel. This data was converted to a Boolean raster where ones represent any buildings present, and zeroes represent no buildings present. This data was chosen for it's availability across the entire country, it's high accuracy (> 99% positive predictive value; Pourpeikari Heris et al. (2020)), and it's 30m resolution.

The raw LiDAR data originates from a single LiDAR acquisition covering portions of Erie, Genesee, and Livingston counties in western New York made available by the New York Stat GIS Program Office (New York Office of Information Technology Services 2019). The raw LiDAR data was then height-normalized and converted into a set of 39 predictors chosen for their prevalence in models of forest structure (Hawbaker et al. 2010; Huang et al. 2019; Pflugmacher et al. 2014). LiDAR data was selected due to it's known ability to characterize three dimensional height-profiles at high-resolution. More specifically, these forest-structure predictors were chosen for practicality. It would be a great benefit to those mapping forest structure with LiDAR if the same predictors could also be leveraged for building better forest masks.

The LiDAR predictors, in raster stack form, were overlaid with the Boolean building raster to create stack of data where each pixel contained a set of 39 predictors and one building indicator response variable. A stratified random sample was conducted on the raster stack, with the building indicator providing the levels of stratification. 3,500 observations (pixel locations) were selected from each stratum resulting in 7,000 observations for model training and testing. This final dataset was converted to a 7000x40 (rows, columns) data frame. The lidR (Roussel and Auty 2020; Jean-Romain et al. 2020) and raster (Hijmans 2021) packages were used for height-normalization and dataset generation.

Additionally, principle components were derived from the final dataset to remove multicollinearity from the 39 predictors (R Core Team 2021). This alternative dataset consisted of the first seven principle components, as they accounted for $\geq$ 95% of the information in the predictors. This alternative dataset was a 7000x7 data frame.

2.2. Models

Three candidate classification models were fit to a random 70% (calibration data; n = 3500) of the observations, with the remaining 30% reserved for model performance assessment (holdout data; n = 1500).

The first candidate model was a simple logistic regression model (R Core Team 2021), and was trained on the principle components variant of the calibration data. The second candidate model was a random forest (RF herafter) trained with the ranger R package (Wright and Ziegler 2017) and the calibration dataset. The third candidate model was a stochastic gradient boosting machine (LGB hereafter) trained with the LightGBM R package (Ke et al. 2021) and the calibration dataset. The hyperparemeters for both the RF and LGB models were selected using a grid search where each combination of hyperparameters were compared against eachother using the cross-entropy loss function (CEL; Equation (1)) computed from a random five-fold cross-validation with the calibration dataset. CEL is computed as follows:

$$\text{CEL} = \sum_{i=1}^{n} - \log\left(\hat{y_i}\right) \tag{1}$$

Where n is the number of observations in the fold, and $\hat{y_i}$ is the predicted probability of the true class.

Each of the models was assessed against the holdout dataset and compared to one another using overall accuracy, specificity, sensitivity, and AUC. Additionally ROC curves were plotted for each model's results on the holdout set. The caret and pROC R packages were used to compute these accuracy metrics (Kuhn 2021; Robin et al. 2011).

3. Results

The RF model performed the best across all accuray metrics (Table 1). The LGB model was more specific but less sensitive than the Logisitc model, and only scored marginally better in AUC and Overall Accuracy. However, all three candidate models performed quite well with all AUC values ≥ 0.87, and all overall accuracies ≥ 0.79. The ROC curves plotted in Figure 1 display similar patterns.

4. Discussion

- RF was superior to all the others
- Logistic still rather good. Makes you wonder if the marginal benefits of ML here are worth the effort, time, understandability

Table 1: Model accuracy metrics computed against 30% holdout partition (n = 1500).

Model	AUC	Overall Accuracy	Sensitivity	Specificity
Logistic	0.87	0.79	0.83	0.75
LGB	0.88	0.80	0.81	0.78
RF	0.96	0.90	0.92	0.89

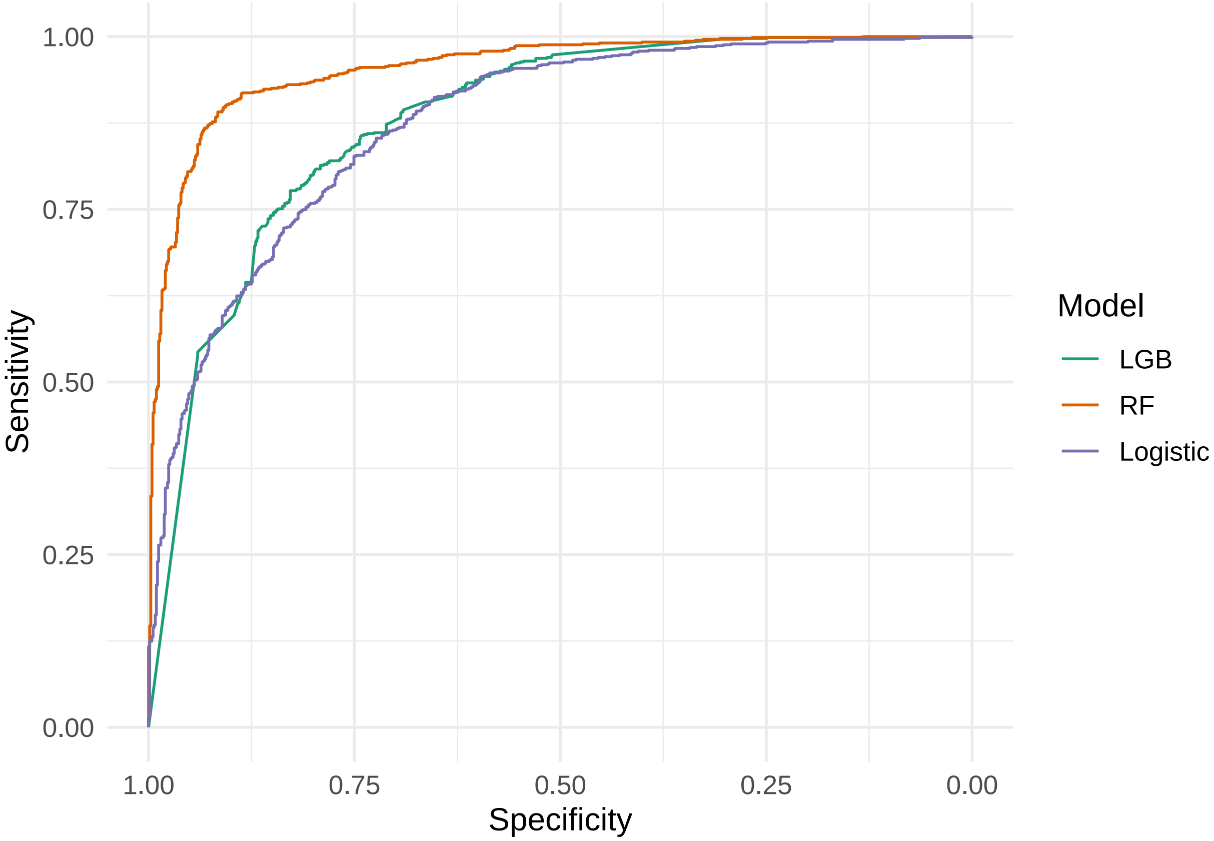

Figure 1: ROC curves showing sensitivity against specificity for each model.

- What could have been done better?
    - ensembling methods?
    - wider ranges of hyperparameters
    - larger inclusion than just 95% of the information for principle components and logistic regression.
    - Timing. Response from 2018. LiDAR from 2019
- What might these models be used for
    - Since the predictors herein are primarily used for predicting forest structure - these models make a relatively easy way to extract man-made structures away from LiDAR-based maps of forest structure like biomass.
    - There are questions about transferrability of these data from this region and this lidar coverages to others from different locations and times.

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Hawbaker, Todd J., Terje Gobakken, Adrian Lesak, Eric Trømborg, Kirk Contrucci, and Volker Radeloff. 2010. "Light Detection and Ranging-Based Measures of Mixed Hardwood Forest Structure." *Forest Science* 56 (3): 313–26. https://doi.org/10.1093/forestscience/56.3.313.

Henry, Lionel, and Hadley Wickham. 2021. *Tidyselect: Select from a Set of Strings*. https://CRAN.R-project.org/package=tidyselect.

Hijmans, Robert J. 2021. *Raster: Geographic Data Analysis and Modeling*. https://CRAN.R-project.org/package=raster.

Huang, Wenli, Katelyn Dolan, Anu Swatantran, Kristofer Johnson, Hao Tang, Jarlath O'Neil-Dunne, Ralph Dubayah, and George Hurtt. 2019. "High-Resolution Mapping of Aboveground Biomass for Forest Carbon Monitoring System in the Tri-State Region of Maryland, Pennsylvania and Delaware, USA." *Environmental Research Letters* 14 (9): 095002. https://doi.org/10.1088/1748-9326/ab2917.

Jean-Romain, Roussel, David Auty, Nicholas C. Coops, Piotr Tompalski, Tristan R. H. Goodbody, Andrew Sánchez Meador, Jean-François Bourdon, Florian de Boissieu, and Alexis Achim. 2020. "lidR: An r Package for Analysis of Airborne Laser Scanning (ALS) Data." *Remote Sensing of Environment* 251: 112061. https://doi.org/10.1016/j.rse.2020.112061.

Ke, Guolin, Damien Soukhavong, James Lamb, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2021. *Lightgbm: Light Gradient Boosting Machine*. https://CRAN.R-project.org/package=

lightgbm.

Kuhn, Max. 2021. *Caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret.

Microsoft. 2018. "US Building Footprints." Microsoft.

New York Office of Information Technology Services. 2019. "LIDAR collection (QL2) for Erie, Genesee, and Livingston Counties New York Lidar; Classified Point Cloud."

Pflugmacher, Dirk, Warren B. Cohen, Robert E. Kennedy, and Zhiqiang Yang. 2014. "Using Landsat-Derived Disturbance and Recovery History and Lidar to Map Forest Biomass Dynamics." *Remote Sensing of Environment* 151: 124–37. https://doi.org/10.1016/j.rse.2013.05.033.

Pourpeikari Heris, Mehdi, Nathan Foks, Kenneth J Bagstad, and Austin Troy. 2020. "A National Dataset of Rasterized Building Footprints for the u.s." U.S. Geological Survey. https://doi.org/10.5066/P9J2Y1WG.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for r and s+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12: 77.

Roussel, Jean-Romain, and David Auty. 2020. *Airborne LiDAR Data Manipulation and Visualization for Forestry Applications.* https://cran.r-project.org/package=lidR.

Team, Bing Maps. 2018. "Computer Generated Building Footprints for the United States." Microsoft.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-

Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wright, Marvin N., and Andreas Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r." *Journal of Statistical Software, Articles* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide.* Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook.* Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.*