Species Distribution Modeling Benchmark Study: Summary and Results

Benton Tripp | btripp@ncsu.edu

1. Overview

1.1. Project Overview

The purpose of this project was to evaluate machine learning models in Species Distribution Modeling (SDM), comparing their performance to traditional models like Maximum Entropy and Inhomogeneous Poisson Process models. Testing was performed using the observation data of 8 distinct bird species, across 4 different US states. Covariates used in model training were derived from bioclimatic variables for each of the sample regions.

1.2. Data

- State Boundary Data
  - Source: ArcGIS Hub
  - Shapefile download with filters set to the 4 states (CO, NC, OR, VT)
- eBird Observation Data
  - Source: eBird Data Access
  - Each of the 4 states listed above, 2016-2019
  - Species included are:
    - Belted Kingfisher
    - Cedar Waxwing
    - Downy Woodpecker
    - Ruddy Duck
    - Sanderling
    - Sandhill Crane
    - Sharp-shinned Hawk
    - Wild Turkey
- Raster Data: raster data was used to produce 39 different covariates available to train the species distribution models. The original data can be found at the following locations:
  - DEM
    - Source: DEM - sciencebase.gov
    - Download by regions (Great Plains, Northeast, Northwest, Southeast, Southwest, and Upper Midwest)
  - Urban Imperviousness
    - Source: Urban Imperviousness - mrlc.gov
    - *NLCD 2019 Percent Developed Imperviousness (CONUS), NLCD 2019 Developed Imperviousness Descriptor (CONUS)*
  - Land Cover
    - Source: Land Cover - mrlc.gov
    - *NLCD 2019 Land Cover (CONUS)*
  - Weather (min/max temperature, avg precipitation)
    - Source: Weather - nacse.org/prism
    - Download weather raster data for “ppt”, “tmax”, “tmin”, 2017-2019 at a 4km resolution and 30-year monthly normals at an 800m resolution
    - URL to download 4km data is: *https://services.nacse.org/prism/data/public/4km/* /

        - URL to download 800m data is *https://services.nacse.org/prism/data/public/normals/800m/* /
    - Hydrography (Water Bodies & Coast)
        - Source: Hydrography Source Page - nationalmap.gov
        - Download Link
    - Vegetation Index
        - Source: USGS Earth Explorer
        - eVIIRS NDVI, 02/23/21-03/08/21 1km; 05/04/21-05/17/21 1km; 09/07/21-09/20/21 1km; 11/30/21-12/13/21 1km

1.3. Pseudo-Absence Selection

In species distribution modeling, the presence of a species in certain locations is often well-recorded, but the absence is typically under-reported or not reported at all. This creates a challenge when trying to understand the complete distribution of a species. To address this, the concept of pseudo-absence data is introduced. Pseudo-absence data are artificially generated points that represent locations where the species is assumed not to be present. In this study, pseudo-absence data points were generated for the eight bird species across the four states.

Initially, the pseudo-absence generation process used was very basic. A buffer was created around each observation point, each with a 5 kilometer radius. These buffers were used to mask the regions of the sample-area from where pseudo-absence points could be sampled (i.e., only the non-buffered zones could be sampled from).

The first attempt of fitting the baseline models using this pseudo-absence data resulted in very inaccurate models. This is not necessarily unexpected, since by its very nature pseudo-absence data is not a perfect representation of the distribution of a species' absence. However, methods do exist to help generate absence data that more accurately represents this distribution. Although these methods aim to represent a species' absence more accurately, they can also introduce bias into the final models, requiring careful application to avoid skewed results. The approach taken to more accurately represent the distribution of a species' absence while minimizing bias in this study was an iterative approach, described as follows:

1. BIOCLIM (by species, state)
    a. For each raster, output a binary raster that is either (1) for binary rasters, equal to the mode of the raster, or (2) for continuous numeric rasters, less than the 10th percentile or greater than the 90th percentile.
    b. Get the sum of all of the rasters for each state.
    c. Identify suitable pseudo-absence sampling regions as (1) points where the BIOCLIM sum is less than the median BIOCLIM sum, and (2) points that are at least 5 kilometers away from an observation point.
3. Sample Pseudo-Absence Points from suitable regions
    a. Sample the same number of points as observations, except with a minimum of 200 points.
    b. Resample (without replacement) 10 times if possible. In some cases when species are more widely distributed in a smaller region, there aren't enough points available; In these cases, resample the maximum allowed times.
    c. Split into training/test sets.
4. Fit LASSO Generalized Linear Models with Updated Pseudo-Absence Points
    a. Iterate through each pseudo-absence resampling data and corresponding presence-only data training set, and fit LASSO models for each (each with the same presence-only training set).
    b. Identify variable importance (use all models, sort by importance).
    c. Predict the probabilities of each resampled point in all of the training and test sets; Select those points that are most likely to be absence points and save those as the final pseudo-absence dataset for modeling (select from training and test sets, corresponding to the number of training/test presence points).

1.4. Modeling the Data

For this study, two modeling methods were used as "baseline" models:

- Inhomogeneous Poisson Process (IPP)
- Maximum Entropy (MaxEnt)

These methods were selected to be used as modeling baselines because they are generally considered the standard for species distribution modeling problems. Despite their similarities they tend to yield different results, and as such were considered separately in this study.

The project primarily focused on evaluating a range of machine learning models, comparing their effectiveness with the model baselines. The methods were chosen for their ability to handle complex nonlinear relationships and interactions between variables, which is often essential in accurately modeling species distributions. The following ML methods were used:

- Logistic Regression
- Classification Tree
- K-Nearest Neighbors (KNN)
- Random Forest
- XGBoost

During the exploratory phase of the analysis, variable importance was measured by fitting LASSO Generalized Linear Models using all of the final derived covariates and covariate interactions. The model coefficients were sorted, and the top 50 were retained for the final modeling of the data. Using these covariates, each model type was fit on the training data for all species, across all of the states. To help minimize potential overfitting, cross-validation was used where possible. In addition, several heuristics were used to help minimize bias or errors:

1. If a fitted model does not converge (this is not applicable for most model types), or results in a prediction error when used for prediction, reduce the number of covariates and re-train the model (eliminating the least "important" variables first).
2. Using the training data, predict species presence. Optimize the prediction threshold (i.e., the cutoff point in the probability where a location is predicted as presence or absence), by maximizing the Sensitivity and Specificity.
3. Measure the Sensitivity and Specificity of the prediction. If Specificity and Sensitivity are a 0/1 pair (i.e., Specificity is 0 and Sensitivity is 1, or vice versa), reduce the number of covariates and re-train the model.
4. Repeat this process until all of the checks are passed successfully.

2. Results

Comparisons of each of the models were made by examining their performance through non-parametric pairwise comparisons and effect size assessments.

The results of the study indicate that ML models generally outperform IPP models in terms of accuracy. However, it is essential to note that these results are not definitive. The limited scope of the study suggests that the conventional approach of relying solely on models like IPP or MaxEnt may not always yield the best outcomes in SDM. Future research, possibly extending to a broader range of species and geographical locations, is necessary to validate and expand upon these findings.

A compilation of all of the visualizations and tables describing the model results can be found in the Appendix.

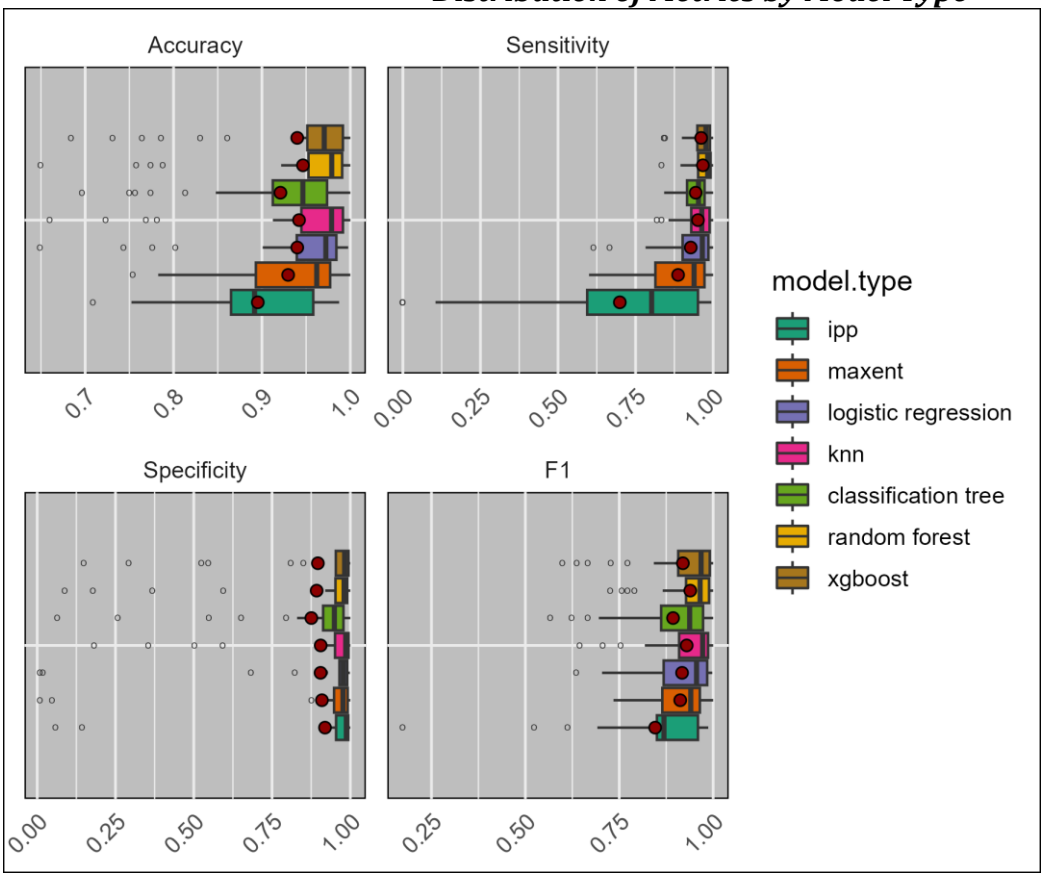
3. Appendix

3.1. Overall Metric Summaries

Average Metrics by Model Type

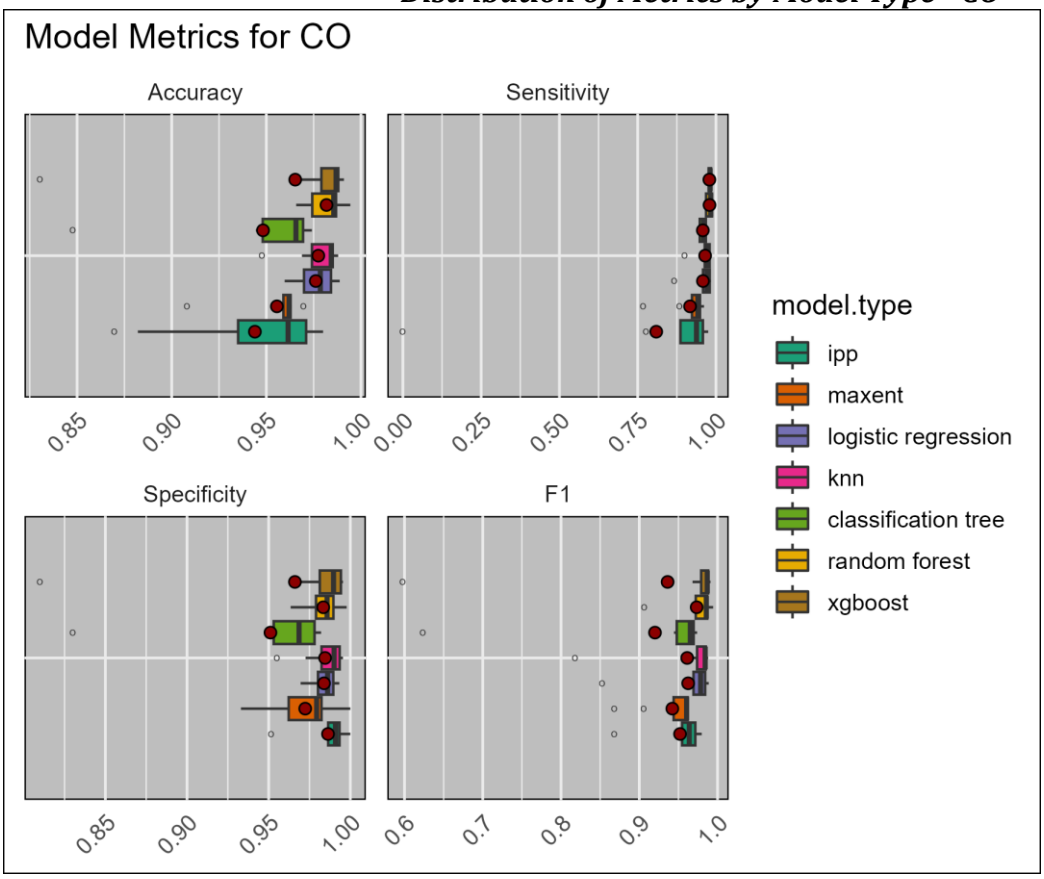
model.type	acc	sens	spec	f1
ipp	0.8953	0.6992	0.9191	0.8456
maxent	0.9297	0.8870	0.9093	0.9124
logistic regression	0.9400	0.9278	0.9050	0.9172
knn	0.9420	0.9509	0.9053	0.9296
classification tree	0.9209	0.9438	0.8756	0.8933
random forest	0.9464	0.9669	0.8925	0.9391
xgboost	0.9399	0.9613	0.8964	0.9194

Distribution of Metrics by Model Type

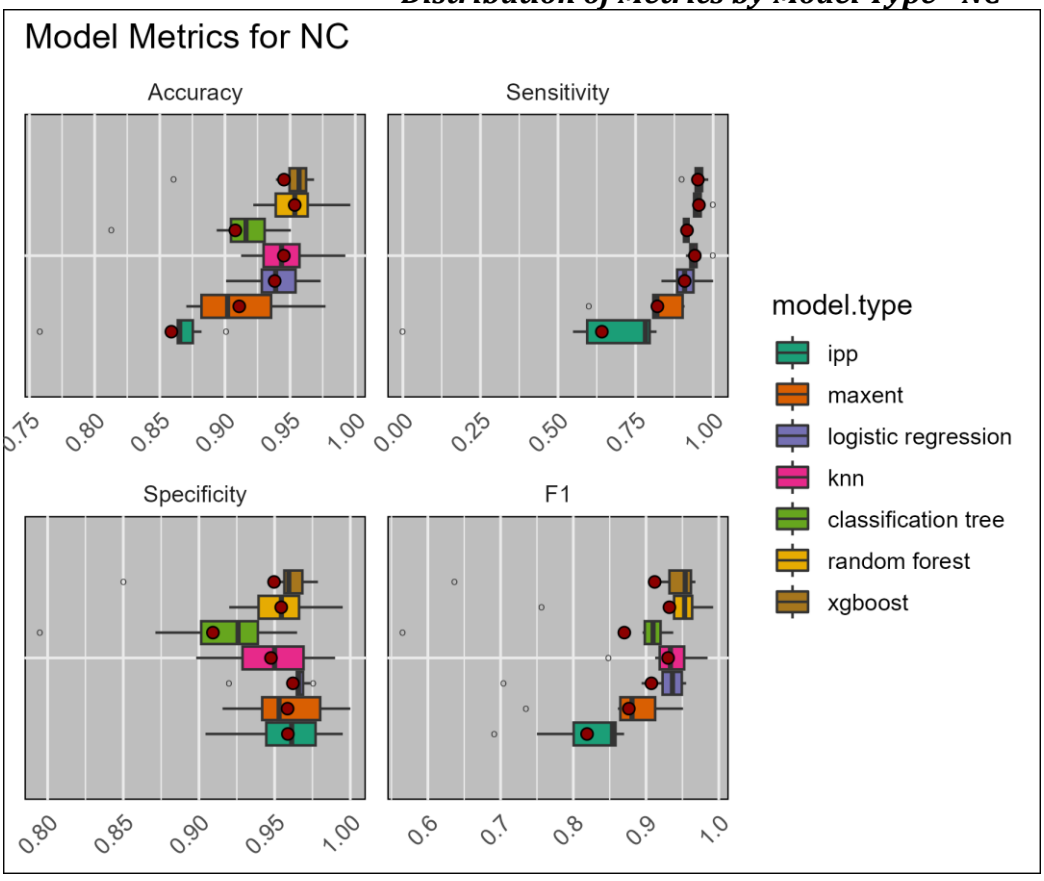


3.2. Metric Summary Plots by State

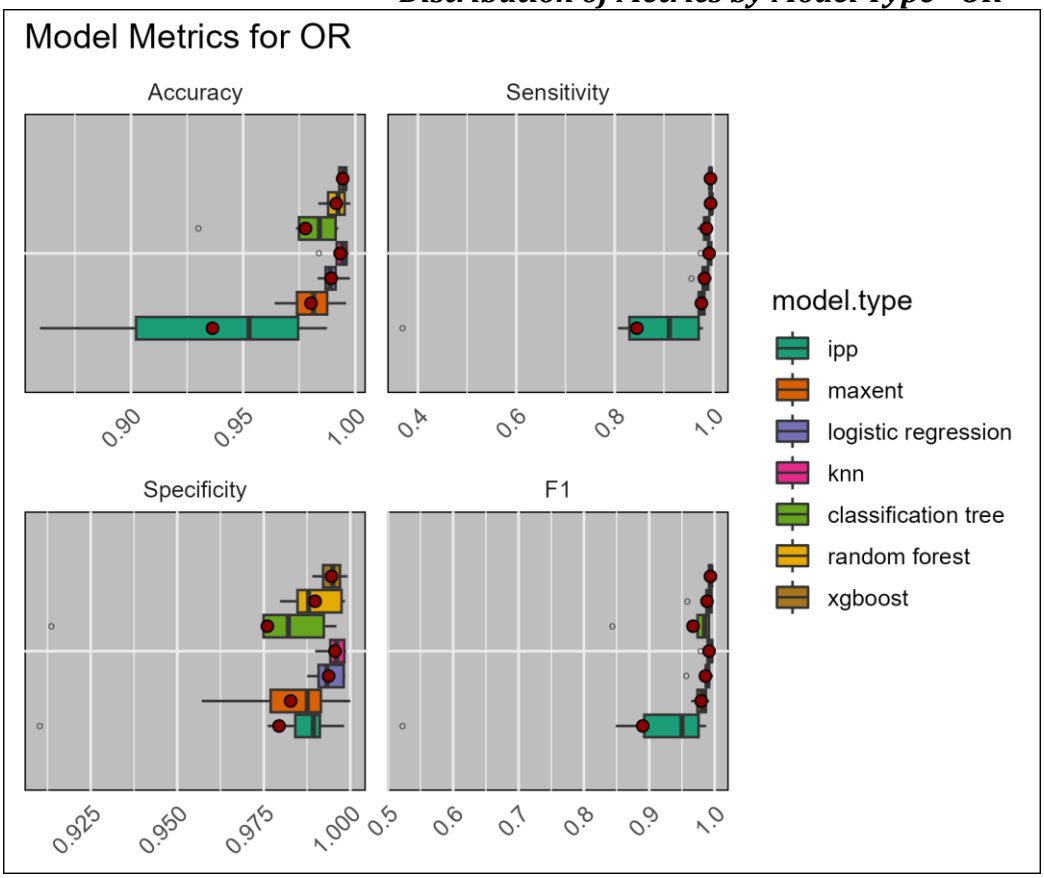
Distribution of Metrics by Model Type - CO



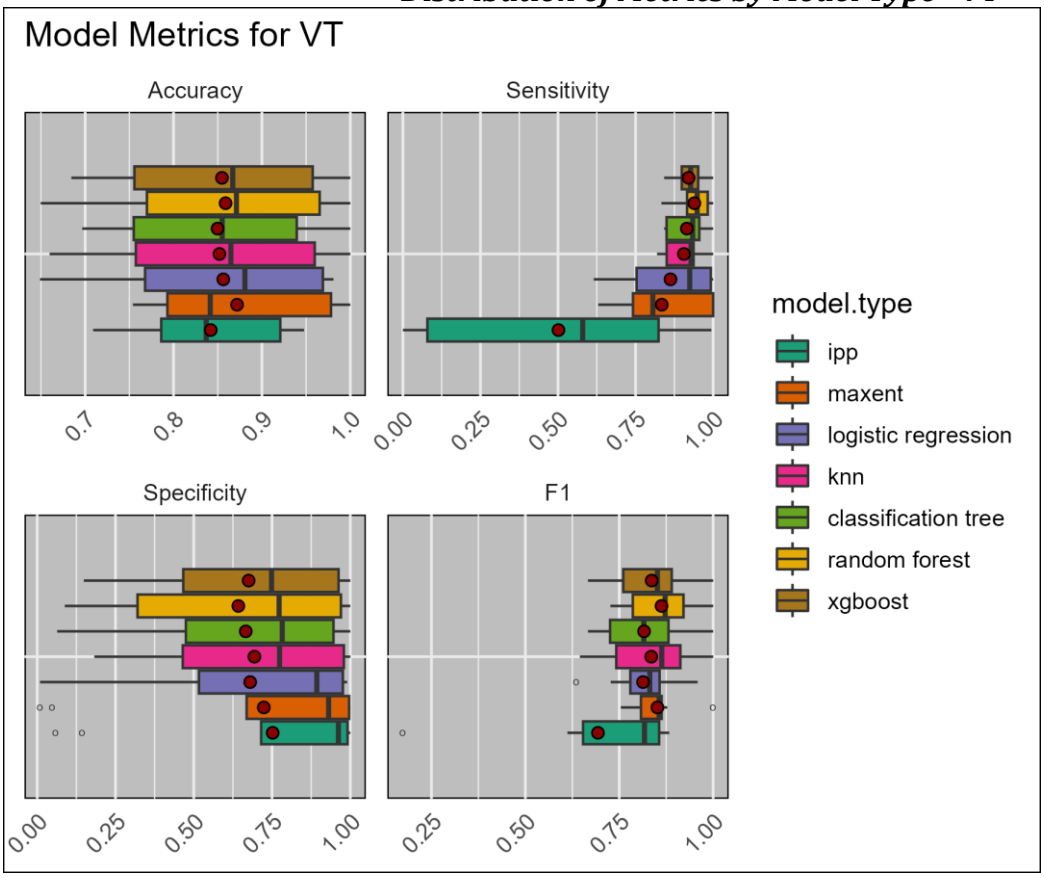
Distribution of Metrics by Model Type - NC



Distribution of Metrics by Model Type - OR

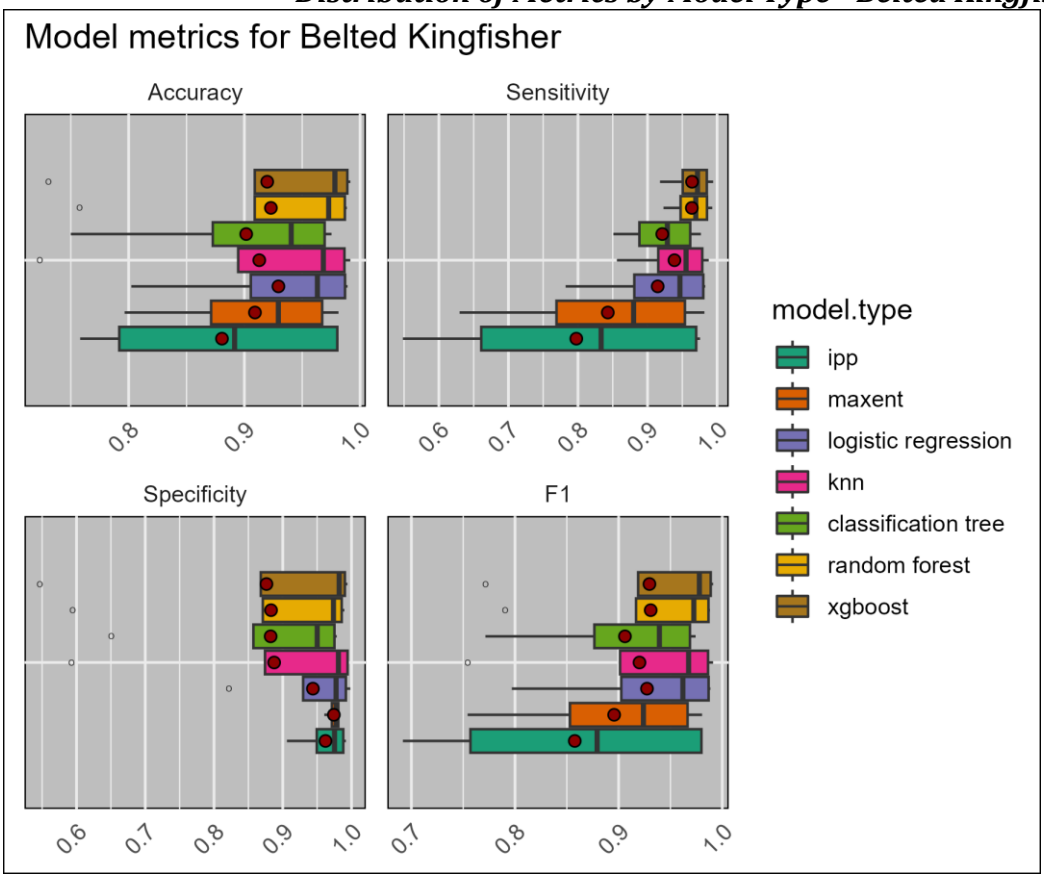


Distribution of Metrics by Model Type - VT

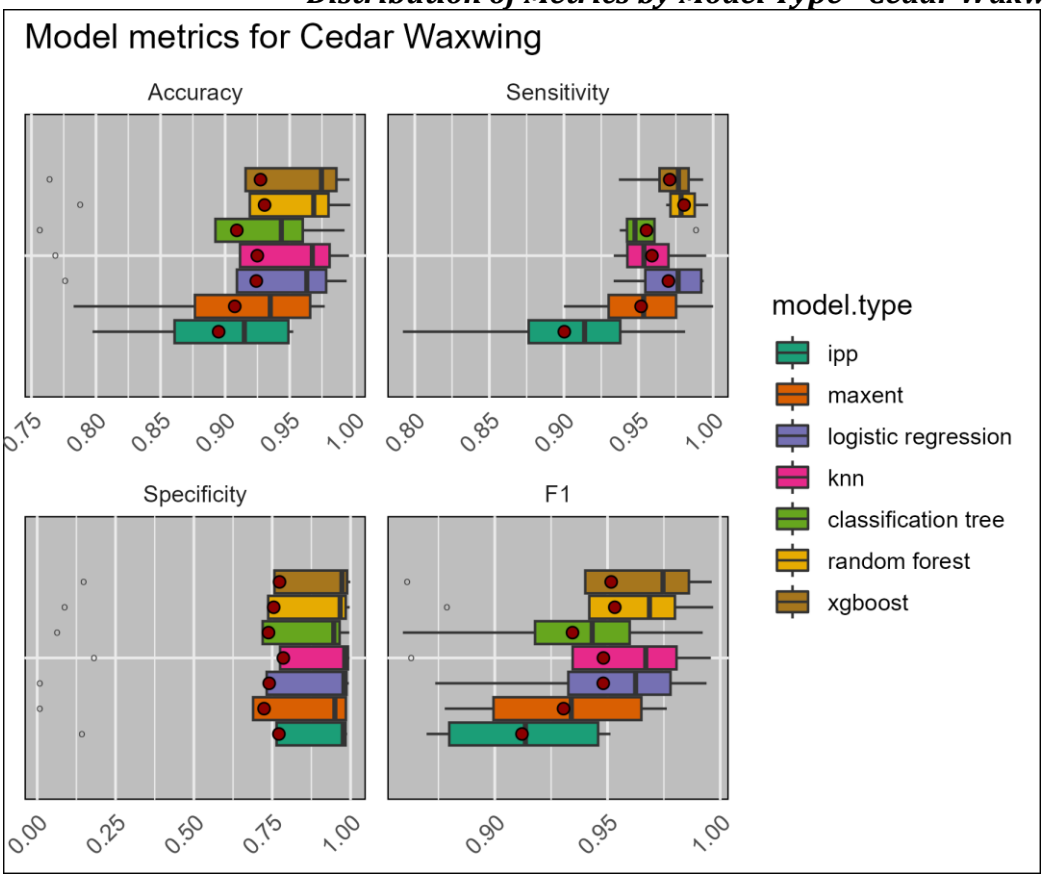


3.3. Metric Summary Plots by Species

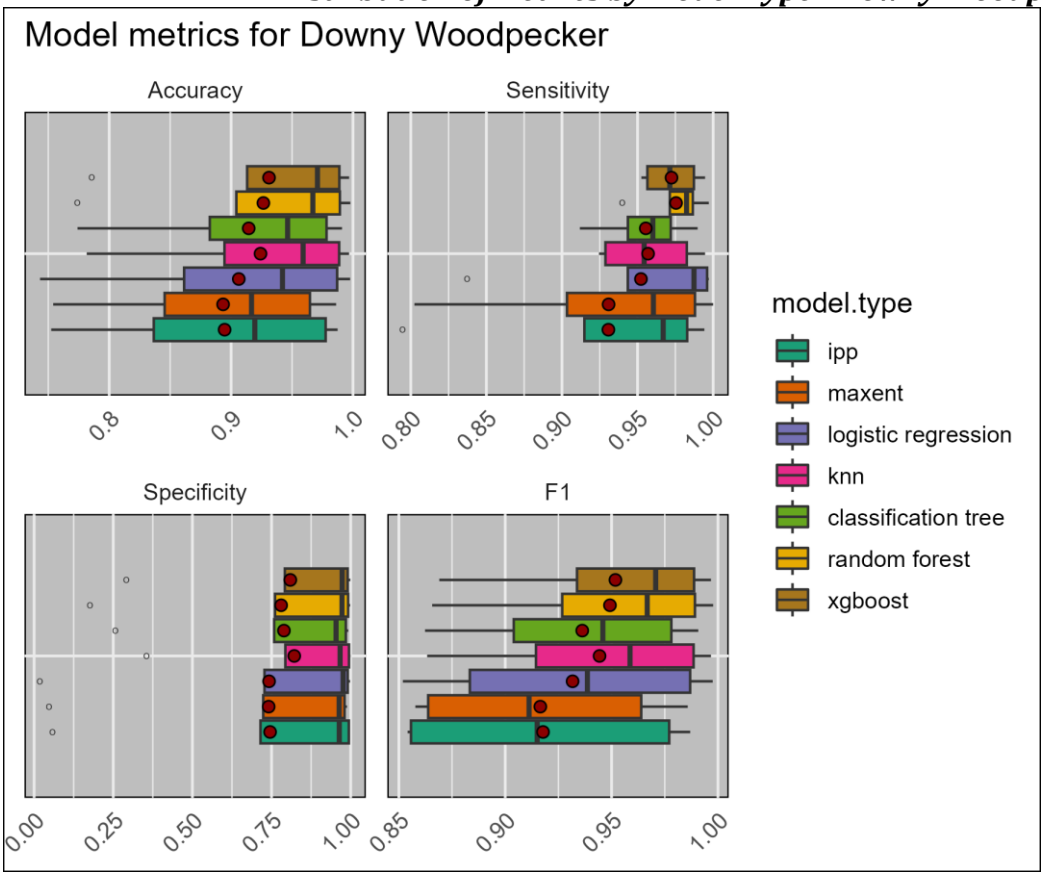
Distribution of Metrics by Model Type - Belted Kingfisher



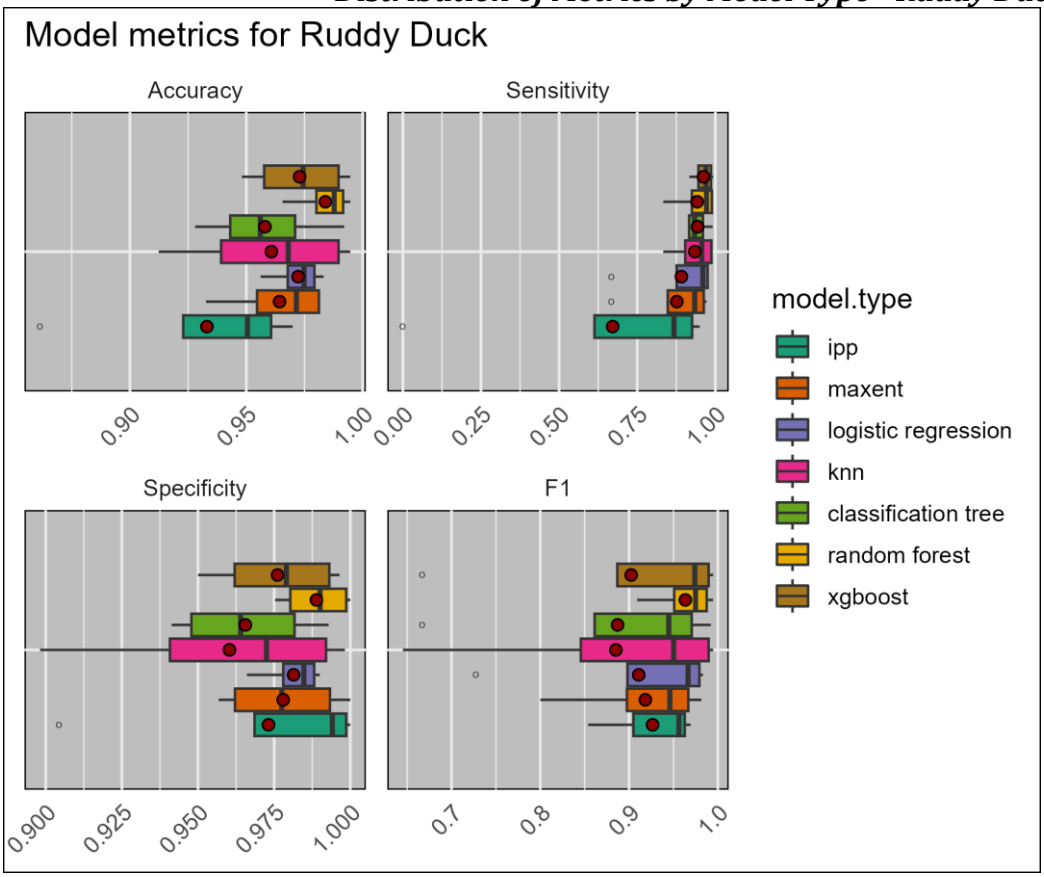
Distribution of Metrics by Model Type - Cedar Waxwing



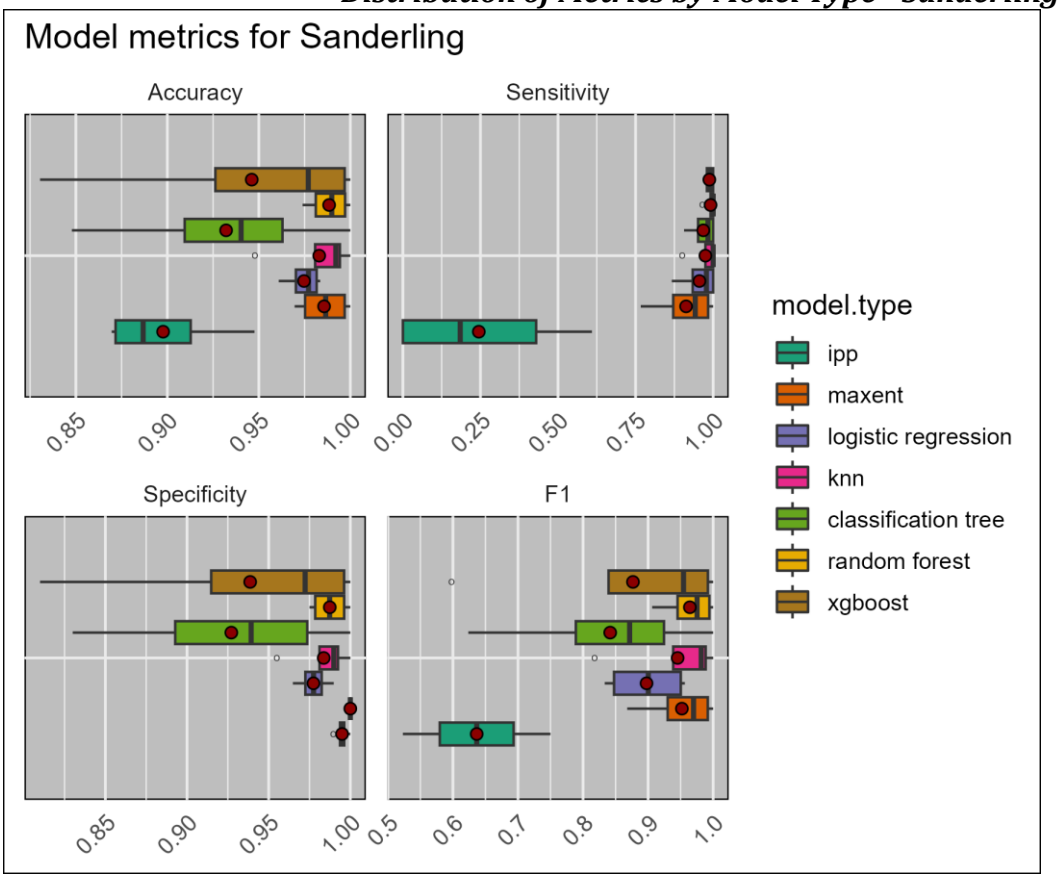
Distribution of Metrics by Model Type - Downy Woodpecker



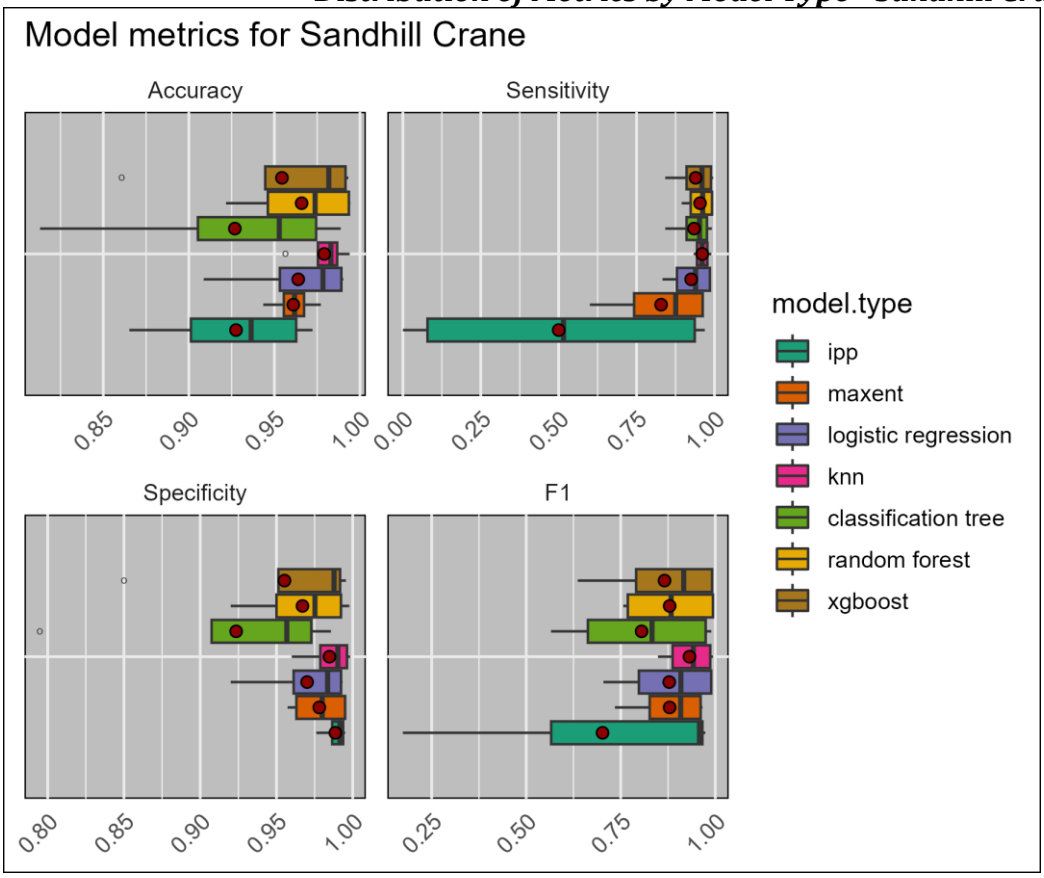
Distribution of Metrics by Model Type - Ruddy Duck



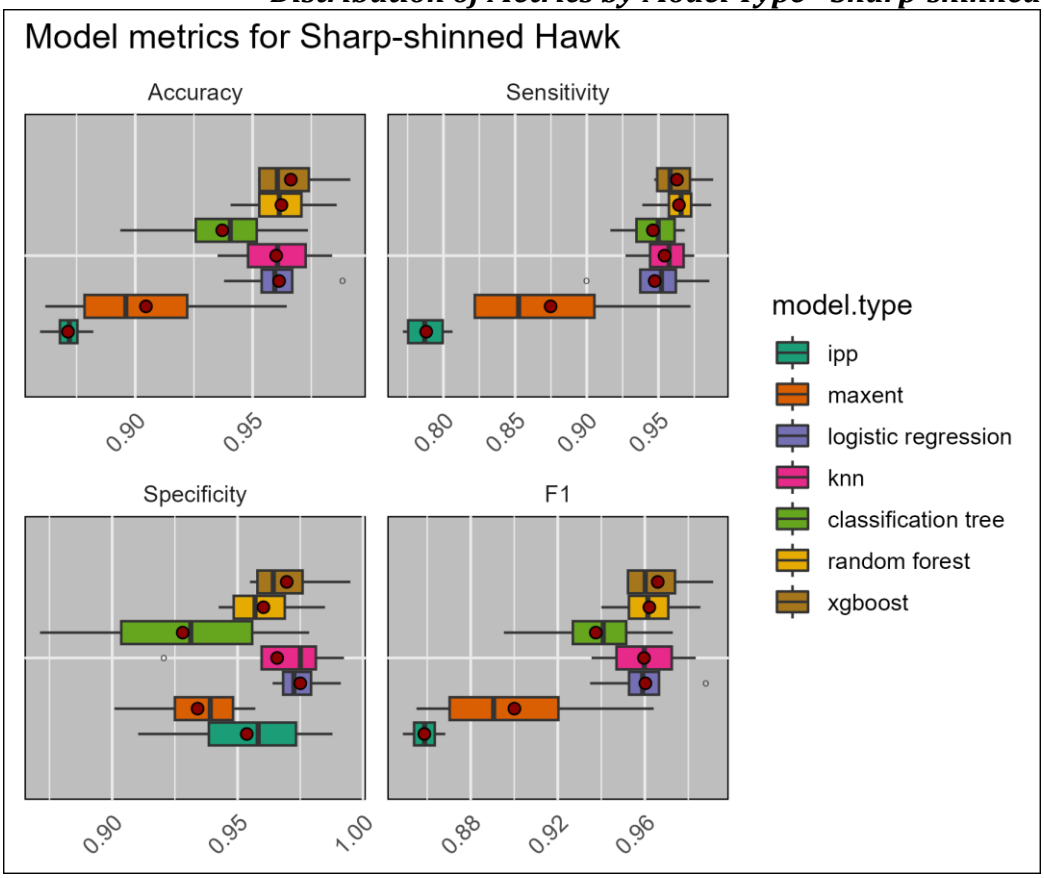
Distribution of Metrics by Model Type - Sanderling



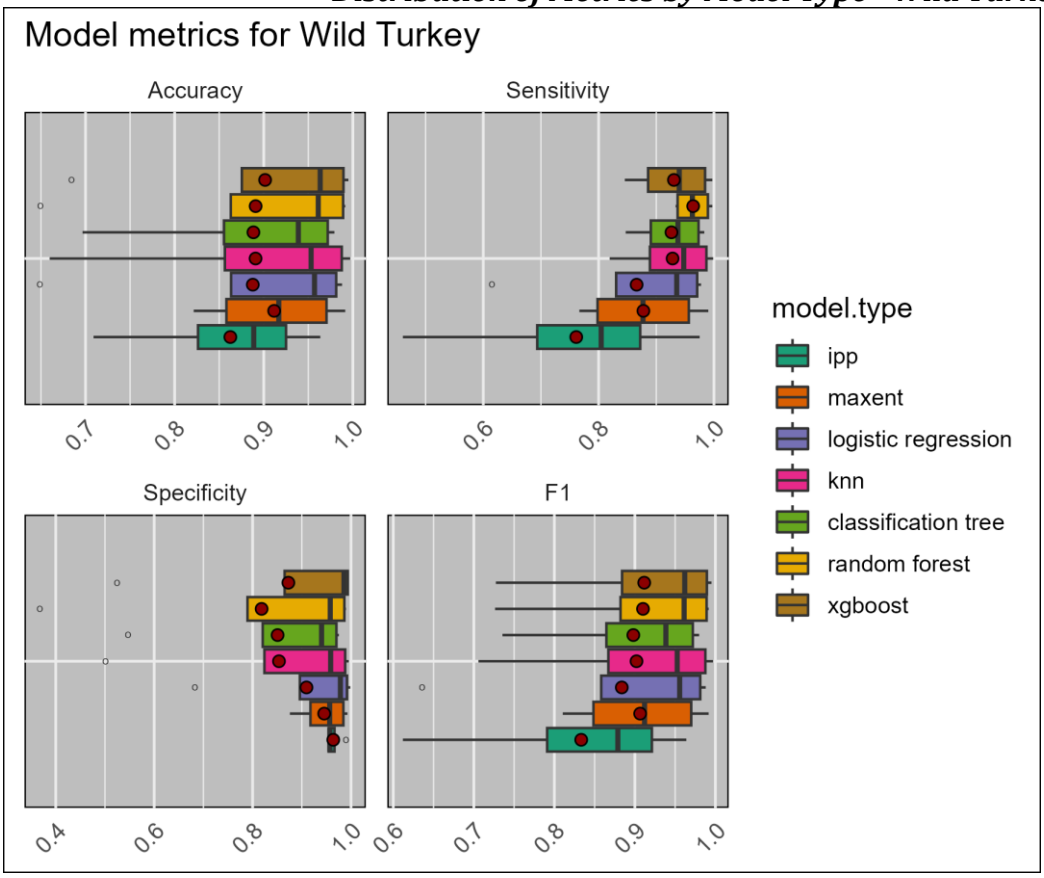
Distribution of Metrics by Model Type - Sandhill Crane



Distribution of Metrics by Model Type - Sharp-shinned Hawk



Distribution of Metrics by Model Type - Wild Turkey

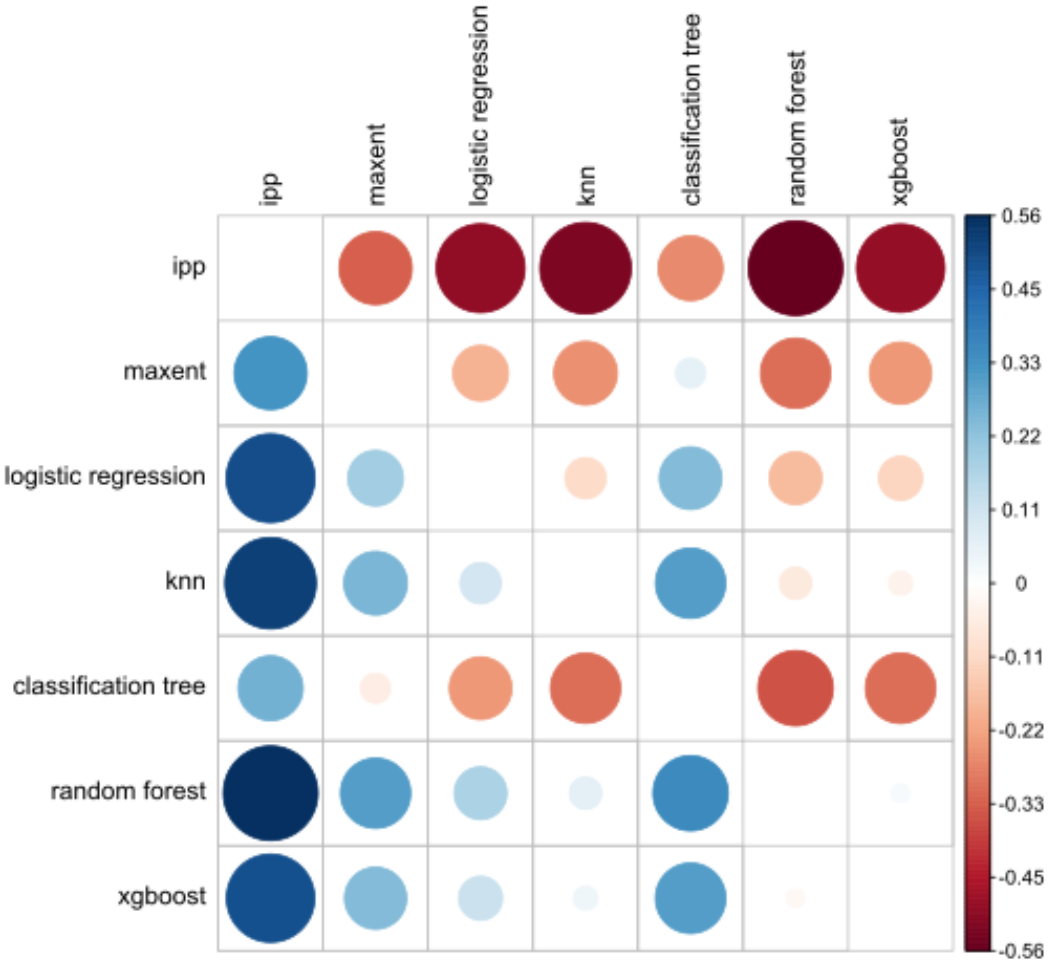


3.4. Model Accuracy Evaluation

Accuracy by Model Type - Significant Pairwise Comparisons and Effect Size

model.1	model.2	delta	lower	upper	variance	p.adj
ipp	knn	-0.520	-0.722	-0.236	0.0156	0.0025
ipp	logistic regression	-0.490	-0.698	-0.206	0.0161	0.0180
ipp	random forest	-0.557	-0.750	-0.275	0.0148	0.0005
ipp	xgboost	-0.484	-0.699	-0.190	0.0172	0.0018

Accuracy by Model Type - Pairwise Comparisons and Effect Size

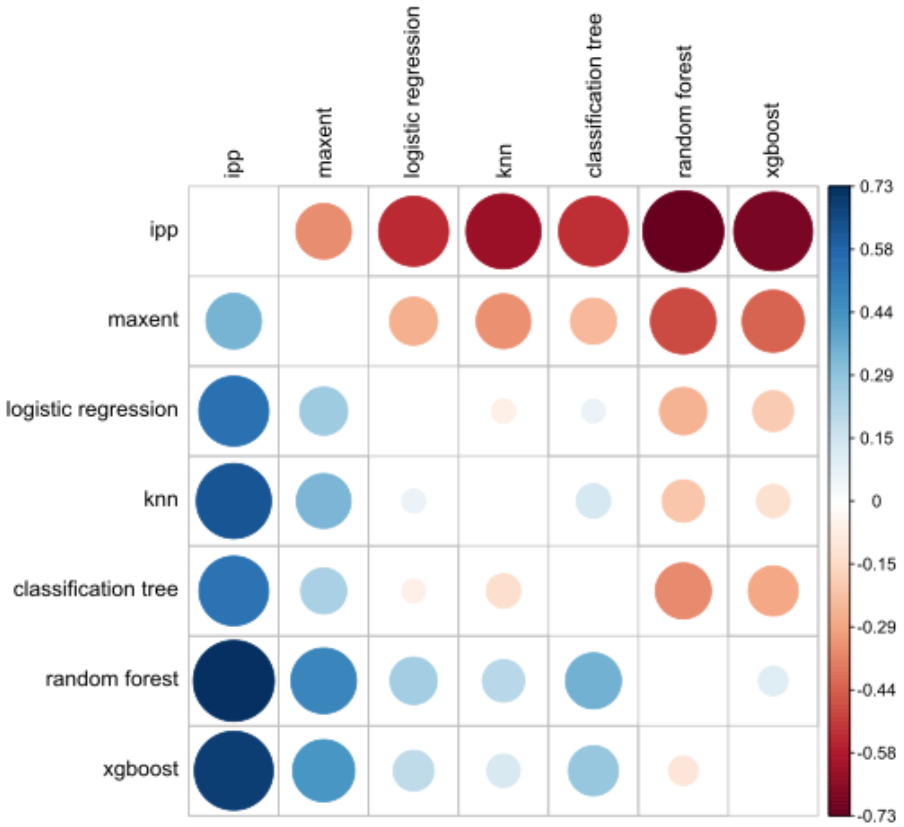


3.5. Model Sensitivity Evaluation

Sensitivity by Model Type - Significant Pairwise Comparisons and Effect Size

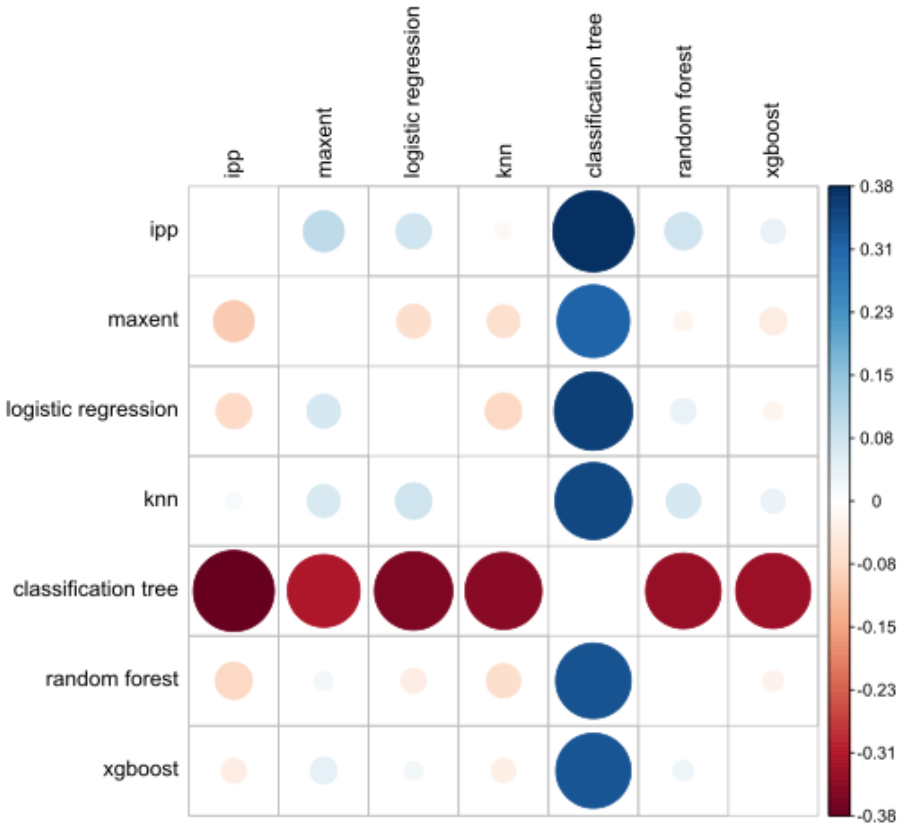
model.1	model.2	delta	lower	upper	variance	p.adj
classification tree	ipp	0.538	0.256	0.736	0.01510	0.00760000
ipp	knn	-0.625	-0.792	-0.370	0.01150	0.00023000
ipp	logistic regression	-0.543	-0.730	-0.280	0.01330	0.00160000
ipp	random forest	-0.729	-0.860	-0.506	0.00782	0.00000041
maxent	random forest	-0.478	-0.686	-0.197	0.01580	0.00600000
ipp	xgboost	-0.691	-0.836	-0.457	0.00905	0.00000560
maxent	xgboost	-0.428	-0.647	-0.143	0.01700	0.03200000

Sensitivity by Model Type - Pairwise Comparisons and Effect Size



3.6. Model Specificity Evaluation

Specificity by Model Type - Pairwise Comparisons and Effect Size



3.7. Model F1 Score Evaluation

F1 Score by Model Type - Significant Pairwise Comparisons and Effect Size

model.1	model.2	delta	lower	upper	variance	p.adj
ipp	knn	-0.449	-0.660	-0.171	0.0158	0.0260
ipp	random forest	-0.502	-0.704	-0.225	0.0151	0.0063
ipp	xgboost	-0.440	-0.651	-0.165	0.0156	0.0140

F1 Score by Model Type - Pairwise Comparisons and Effect Size