Radiomics for Disease Characterization: An Outcome Prediction in Cancer Patients

Magnuson, S. J. , Peter, T. K., and Smith, M. A.

Department of Biostatistics

University of Iowa

July 19, 2018

Background- Information

- Lung cancer is the leading cause of cancer-related mortality in the United States
- 234,030 new cases expected in 2018
- 200 CT scans from University of Iowa Hospital Patients
- 410 quantitative imaging biomarkers (Intensity, Shape, Texture) used for analysis
- 5 patient demographics (Lobe, Age, Race, Gender, Packs per Year)
- 45% of cases were benign and 55% of cases were malignant

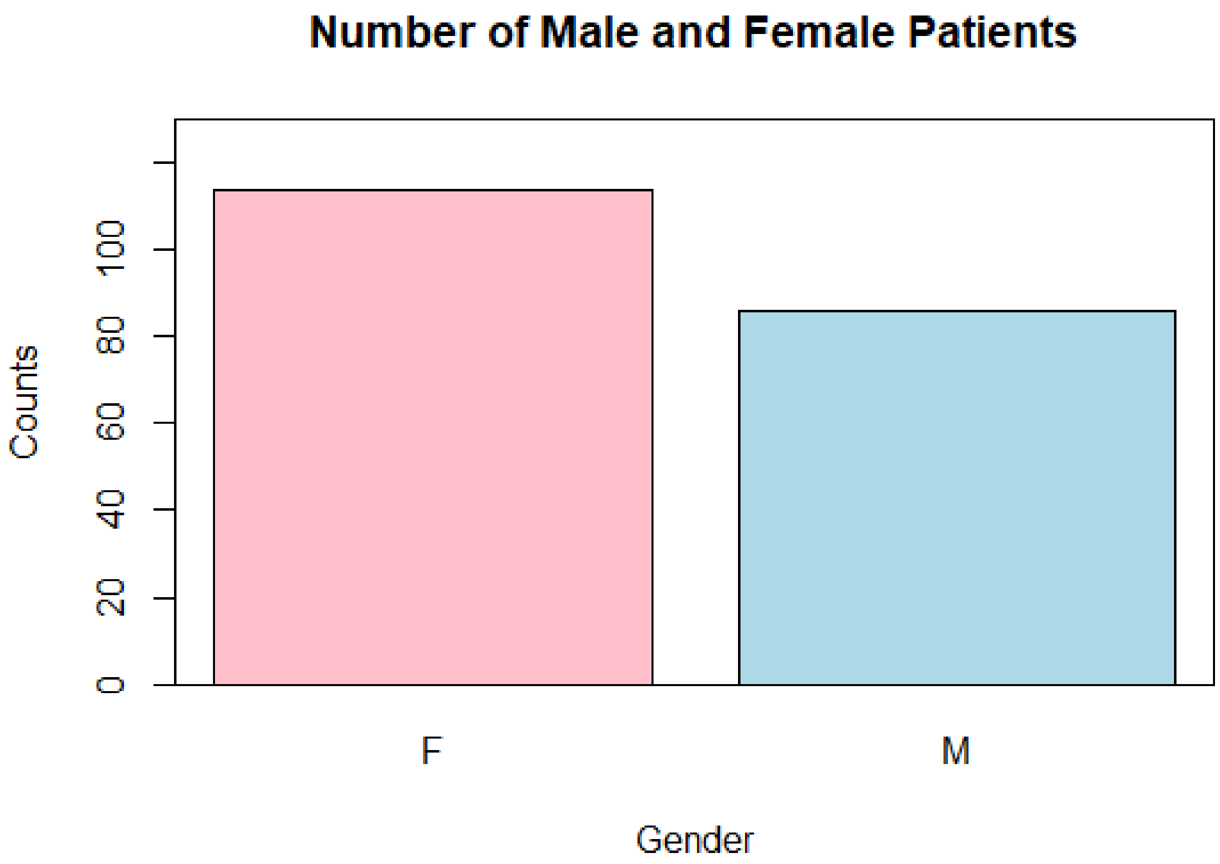
Project Objective

To develop a statistical model to predict lesion malignant/benign status of each patient

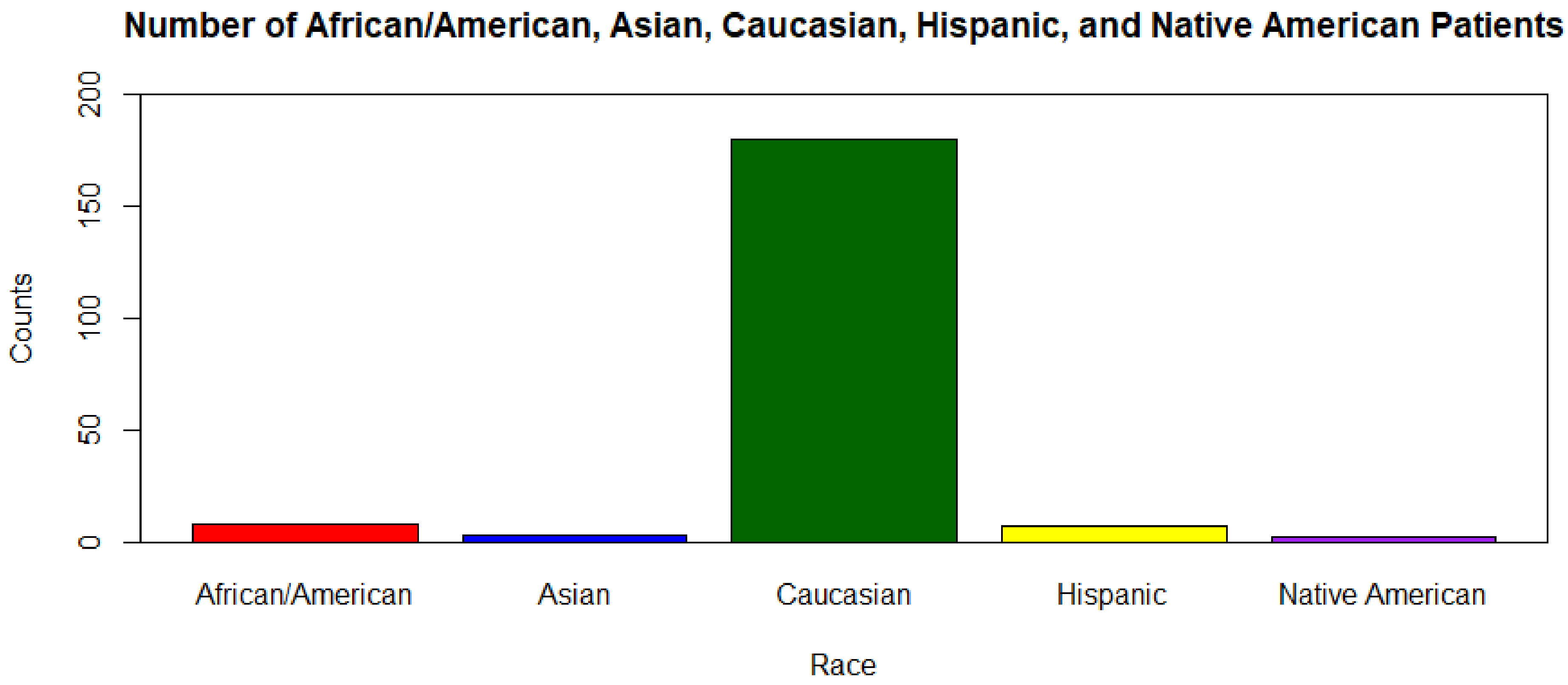
Background – Descriptive Statistics

	Age (years)	Packs Smoked (per year)
Minimum	24	0
Mean	59.88	26.18
Median	60	20
Maximum	90	150

Background – Descriptive Statistics



Background – Descriptive Statistics

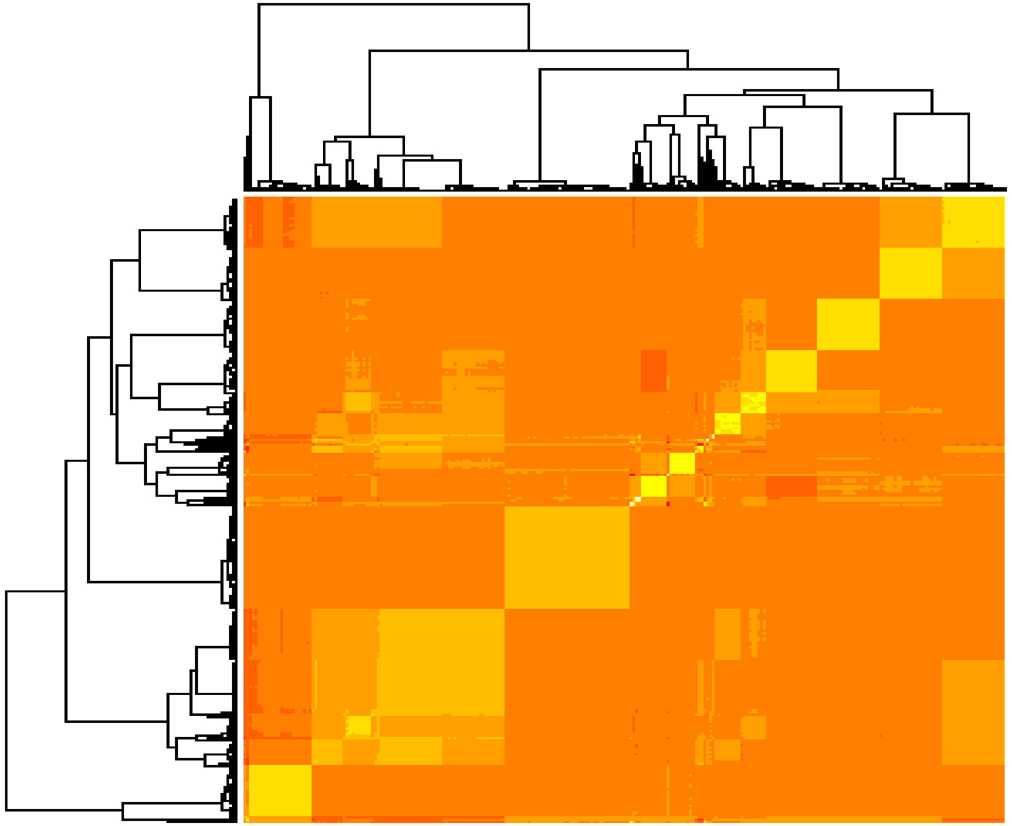


Data-Preprocessing

Filtering Variables

Filtering Variables

Due to the high correlation of predictors, we look for the removal of non-informative/redundant variables to improve model stability and performance



Heat Map

Filtering Variables

Methods for Data-filtering

1. Correlation: remove predictors so that all pairwise correlations are below a specified threshold (0.95)
2. Near Zero Variance: remove variable predictors that are constants

When applied to the full data set, 348 predictors were removed

Model Selection and Assessment

AUC and ROC 🥕

Model Selection and Assessment- AUC

- AUC: area under the receiver operating characteristic (ROC) curve
- Estimates the probability that a randomly selected subject with a malignant lesion will have a greater model predicted probability than a randomly selected subject with a benign lesion
- The closer AUC is to 1.0 (100% specificity and 100% sensitivity), the better the predictive performance
- The closer AUC is to 0.50, the worse the test

Model Selection and Assessment- AUC

Range	Scale
0.97-1.00	Excellent
0.92-0.97	Very Good
0.75-0.92	Good
0.50-0.75	Fair

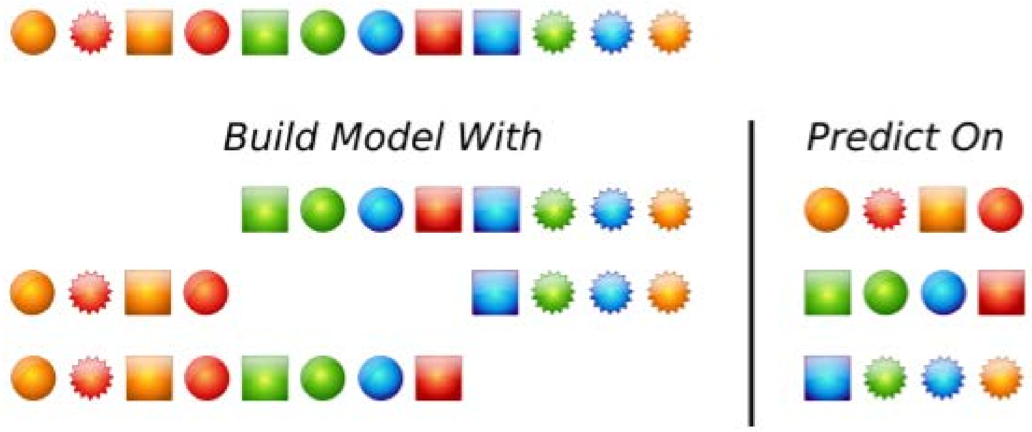
K-Fold Repeated Cross-Validation

Original Data

Fold 1

Fold 2

Fold 3



Cross-Validation Estimate of the Performance Metric, AUC:

$$AUC = \frac{1}{50}\sum_{r=1}^{5}\sum_{k=1}^{10} AUC_{rk}$$

Elastic Net

Model details, filtering vs. non-filtering

Model Details- Elastic Net

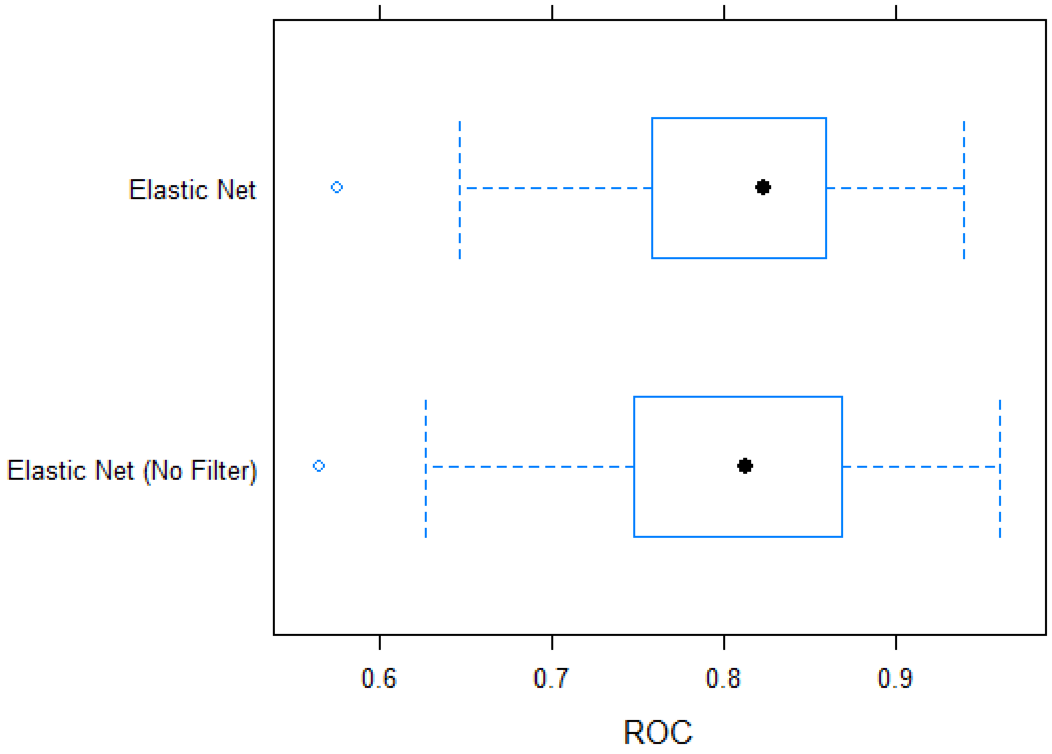
- Logistic regression finds parameters that maximize the binomial likelihood function, $L(p)$
- The parameters can be regularized by adding a penalty to the likelihood function
- There are two types of penalties to add:
  1. Ridge
  2. LASSO (least absolute shrinkage and selection operator)
- Elastic Net combines the two types of penalties

Model Details- Elastic Net

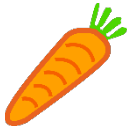
$$\log L(p) - \lambda\left[(1-\alpha)\frac{1}{2}\sum_{j=1}^{P}\beta_j^2 + \alpha\sum_{j=1}^{P}|\beta_j|\right]$$

- λ controls the total amount of penalization
- α is the mixing percentage (when $\alpha = 1$ it is a pure lasso penalty; when $\alpha = 0$ it is a pure ridge-regression-like penalty)
- This enables effective regularization via the ridge-type penalty with the feature selection quality of the LASSO penalty

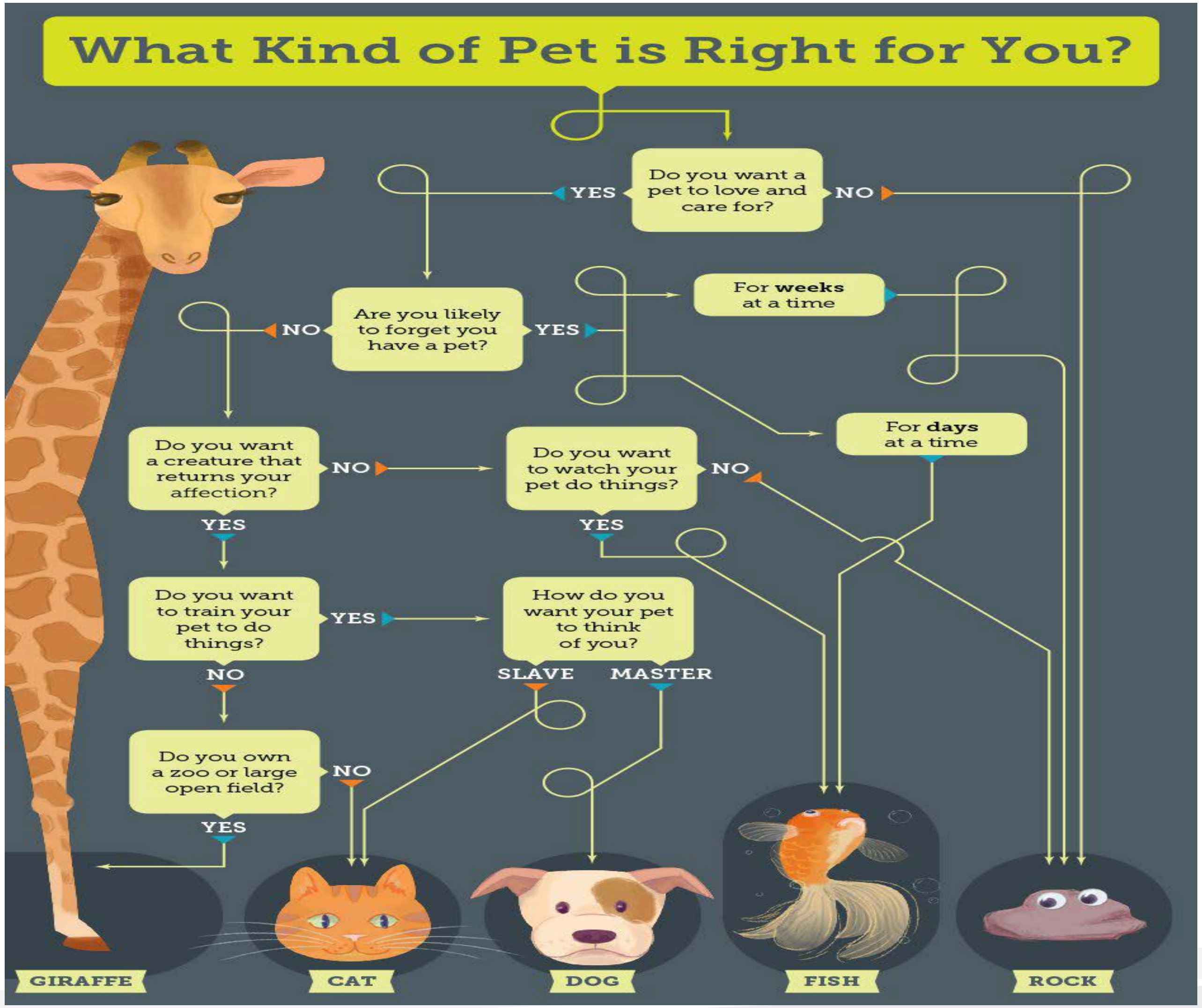
Filtering vs. Non-filtering- Elastic Net



Random Forest

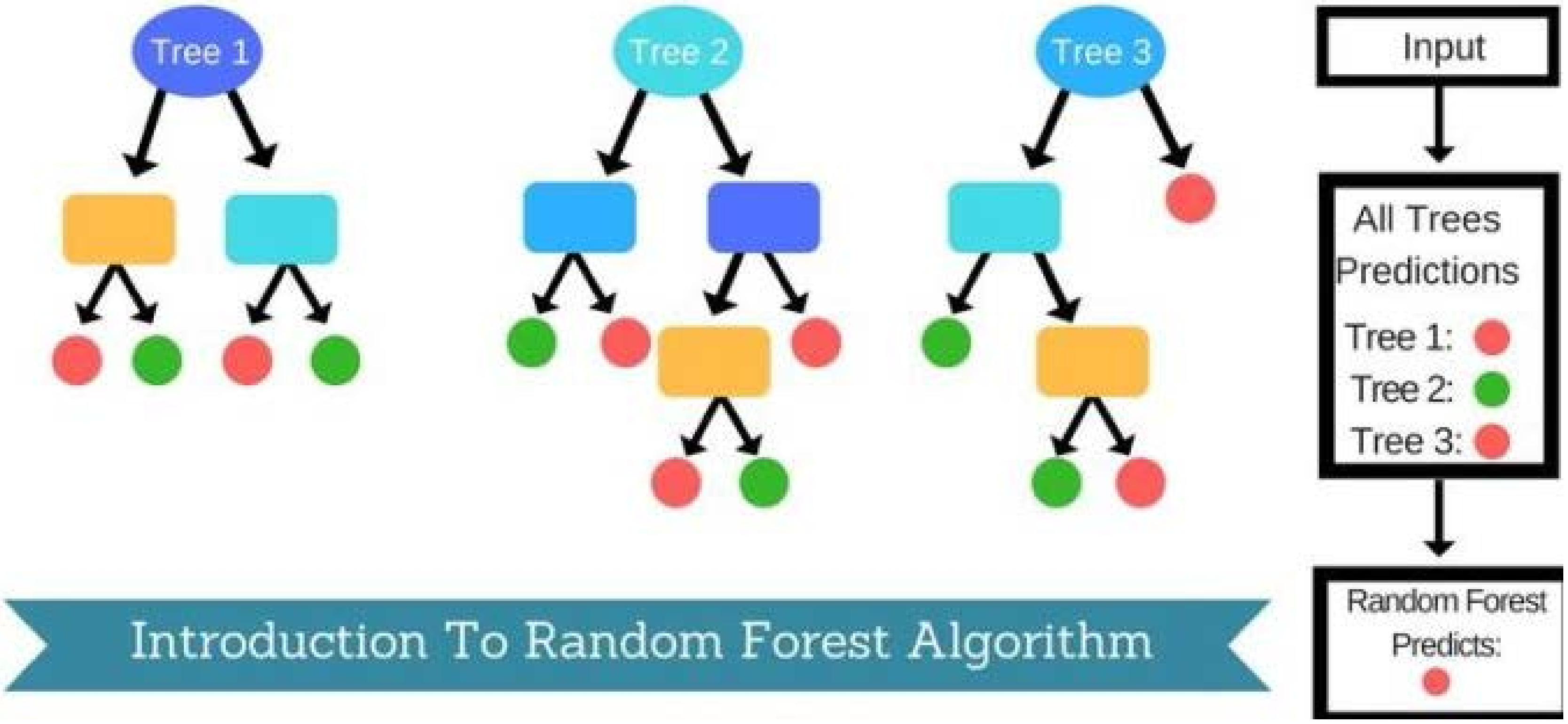
Decision trees, Model Details, filtering vs. non-filtering 

A Forest of Decision Trees



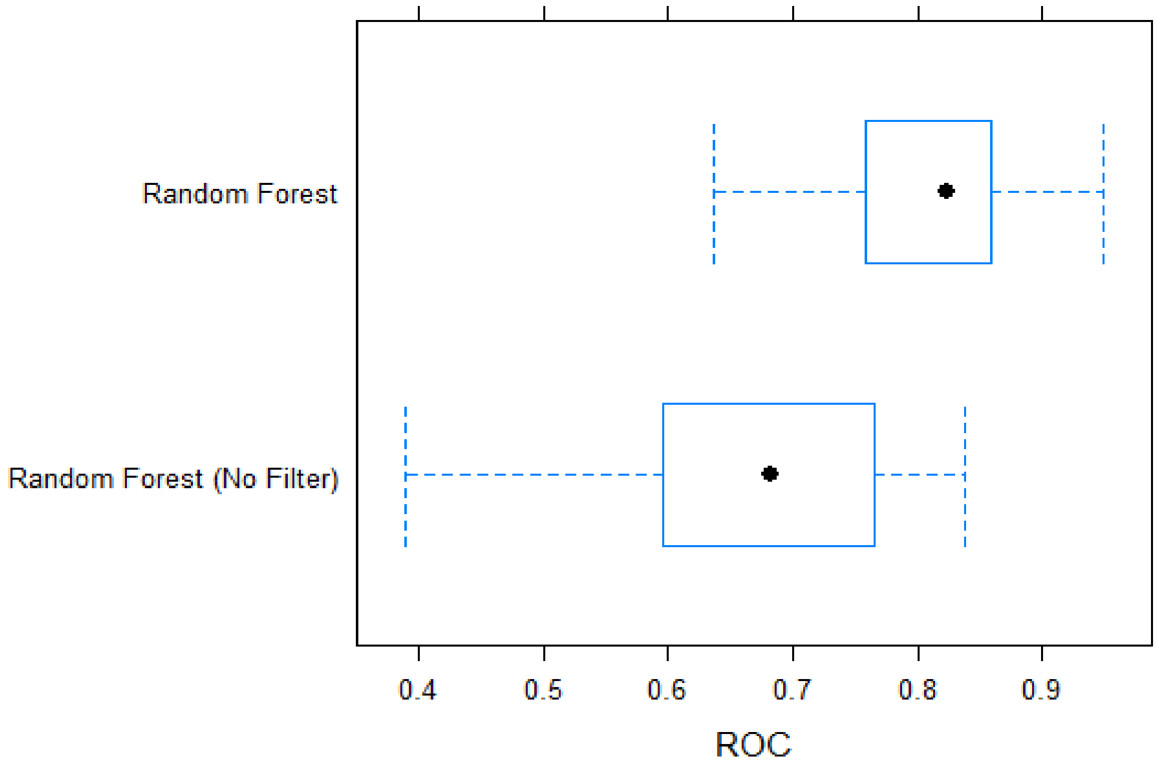
We can apply the same concept of decision making to classifying data.

Random Forest – Model Details



Random forest takes a majority vote over a collection of decision trees to improve accuracy and reduce prediction variability

Filtering vs. Non-filtering- Random Forest



Stochastic Gradient Boosting

Model details, filtering vs. non-filtering

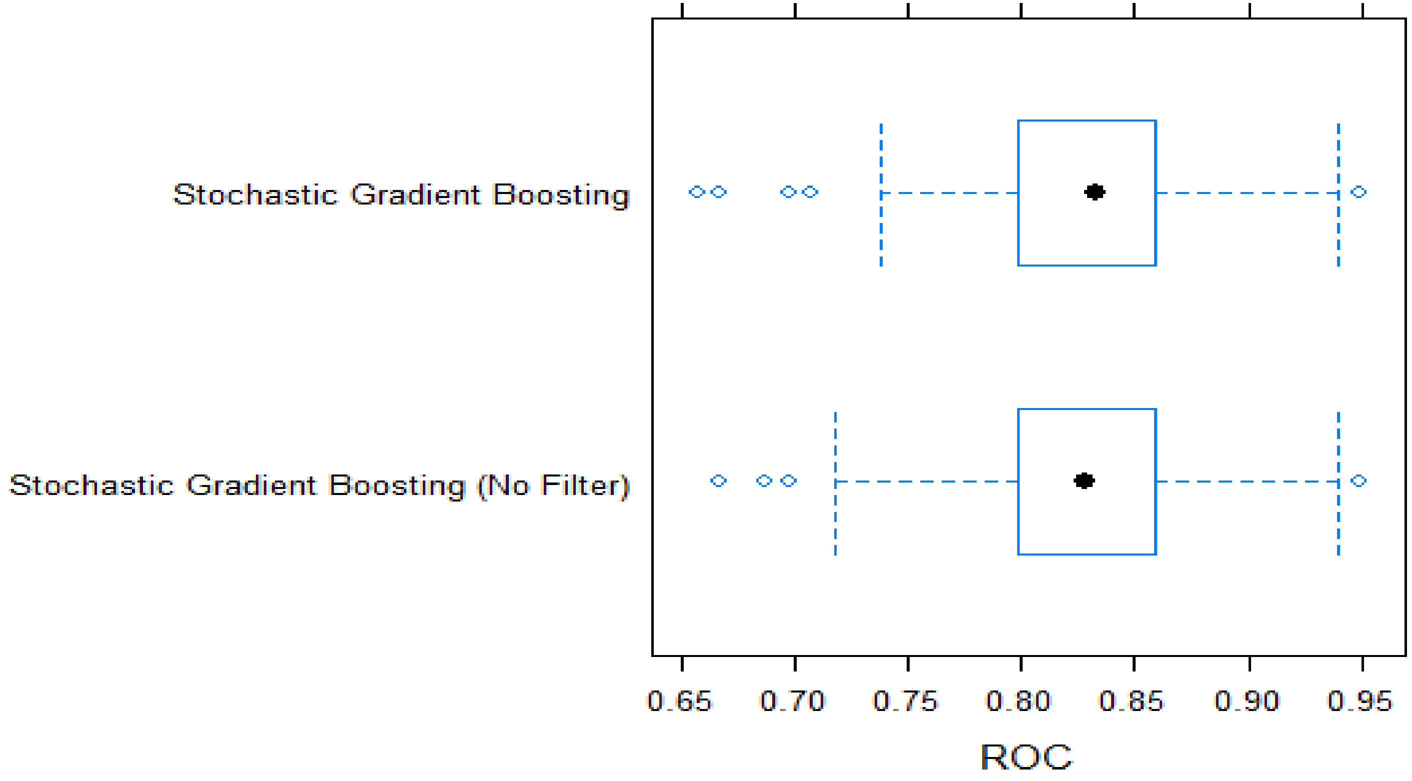
Model Details-Stochastic Gradient Boosting

- Influenced by Learning Theory: a number of weak classifiers are combined to produce an ensemble
- Basic Principles of Boosting:
  1. The algorithm seeks to find an additive model of decision trees to minimize a given loss function
  2. Algorithm initialized with best guess of the response
  3. The gradient (residual) is calculated and a model is fit to the residuals
  4. Current model added to the previous model
  5. Procedure continues for a specified number of iterations

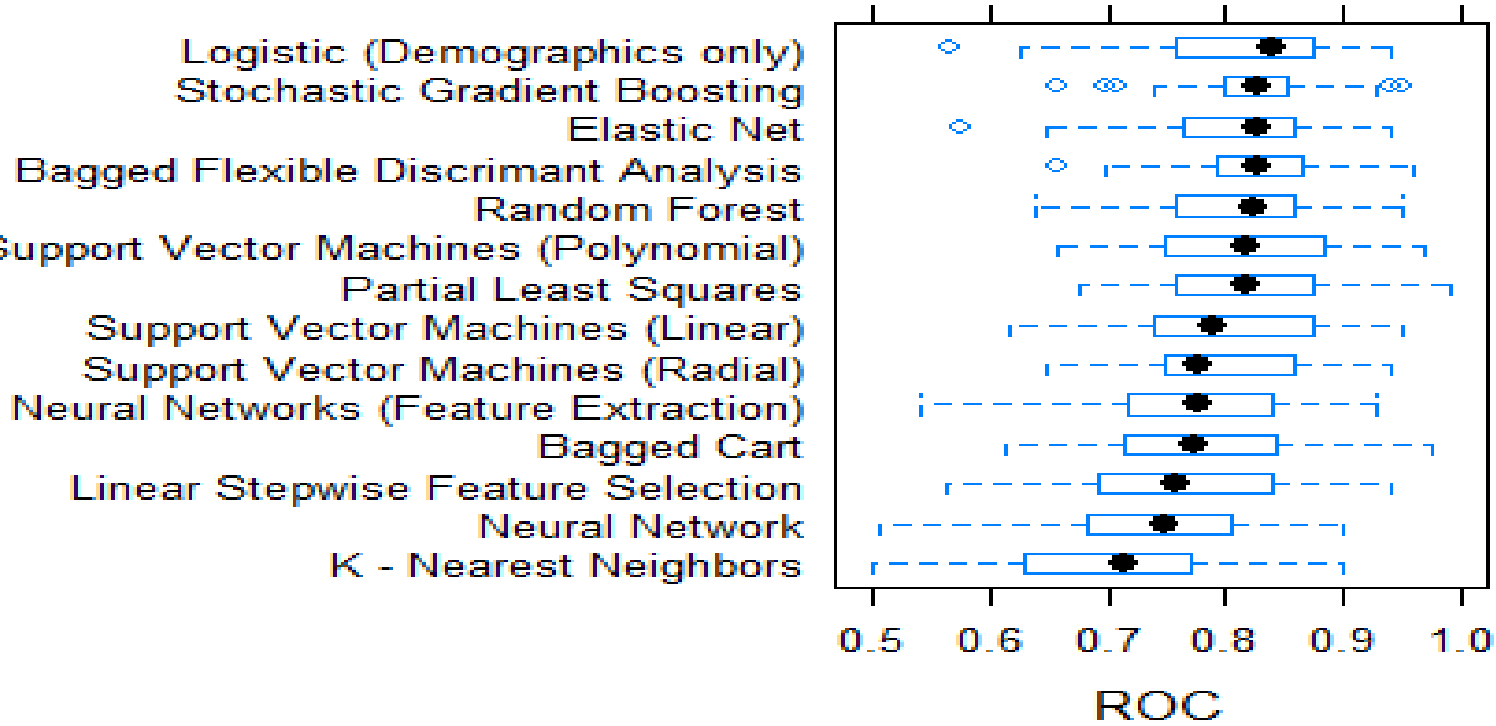
Model Details- Stochastic Gradient Boosting

- Boosting bears similarities to Random Forest and both models give equal predictive performance
- Random Forest and Boosting are constructed differently
- In Random Forest, all trees are created independently and each tree is created to have maximum depth and all trees contribute equally
- In Boosting, the trees are dependent on past trees, have minimum depth, and contribute unequally to the model

Filtering vs. Non-filtering: Stochastic Gradient Boosting



Model Comparison



Index: Method to identify a probability cut point that optimizes the sensitivity and specificity with respect to the prevalence rate and the cost

$$index = \min\left((1 - sens)^2 + r * (1 - spec)^2\right), \text{where}$$

$$r = \frac{(1 - p)}{(cost * p)}$$

and

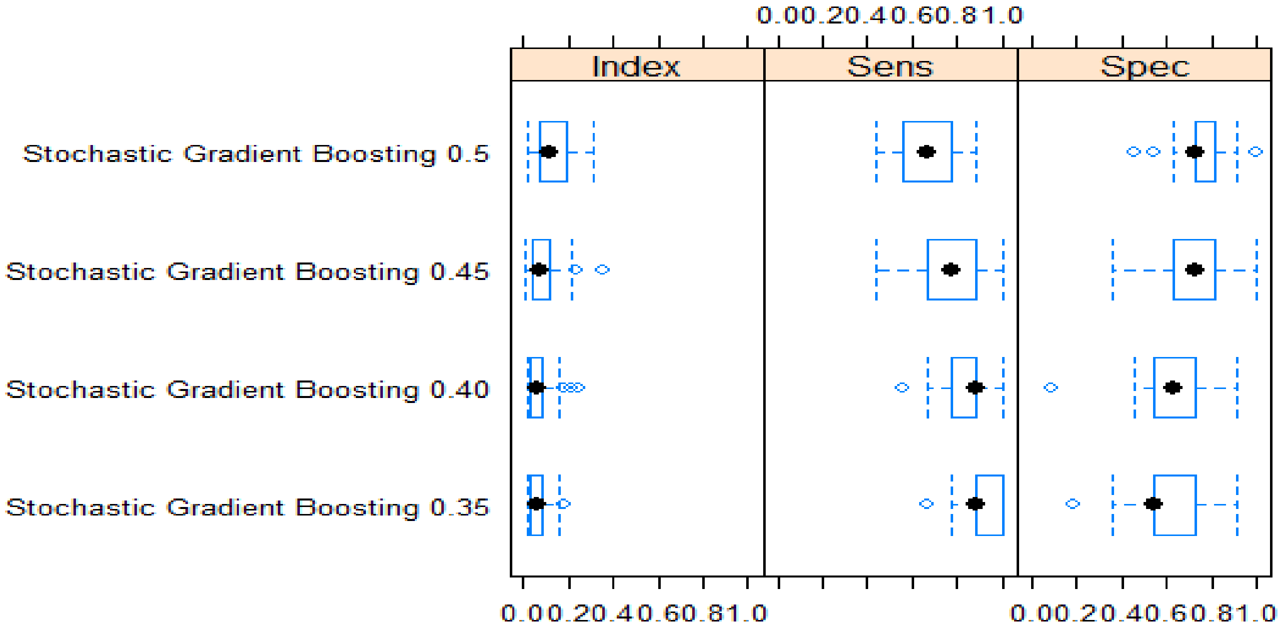
$$p = \text{prevalence} = 0.50$$

and

$$cost = \frac{false\ negative}{false\ positive} = 4.0$$

Index Table: Stochastic Gradient Boosting

Stochastic Gradient Boosting	Index (mean)	Sensitivity (mean)	Specificity (mean)
0.5	0.12	0.70	0.78
0.45	0.09	0.78	0.71
0.40	0.07	0.86	0.63
0.35	0.06	0.90	0.59



Conclusions

Main takeaways, future work

Main Takeaways and Future Work

- The Stochastic Gradient Boosting model had the best performance, considering its high AUC and relatively low variability
- The filtering helped the Random Forest models noticeably
- The logistic regression using only the demographic predictors performed the best
- However, using the biomarkers alone did improve predictive performance
- Plan to explore the index values further
- Plan to explore deep neural networks

Acknowledgments

Dr. Brian J. Smith, Professor, Dept. of Biostatistics University of Iowa

National Heart, Lung, and Blood Institute (NHLBI), grant #HL131467

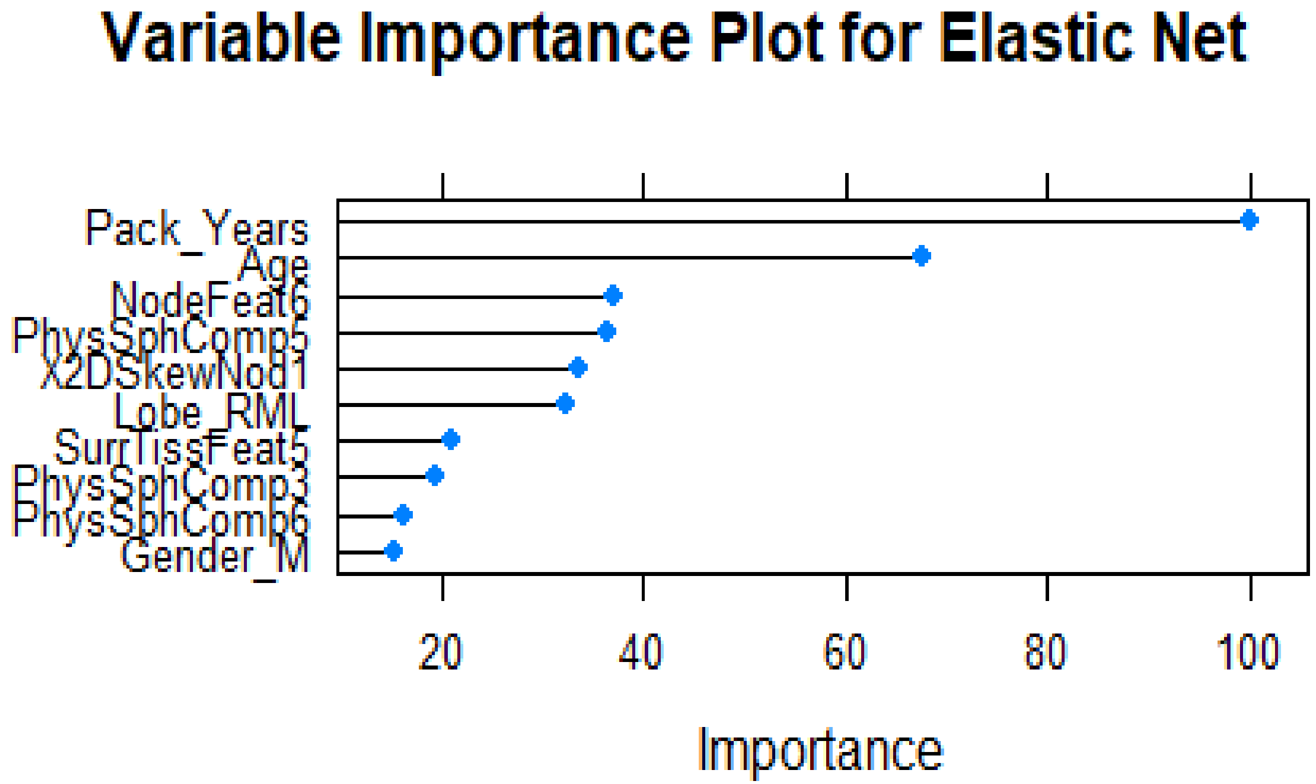
References

- Kuhn, M., & Johnson, K. (2016). *Applied Predictive Modeling*. New York: Springer.
- Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
- Smith, Brian J. (2018) BIOS6720, [PDF]. University of Iowa, Department of Biostatistics

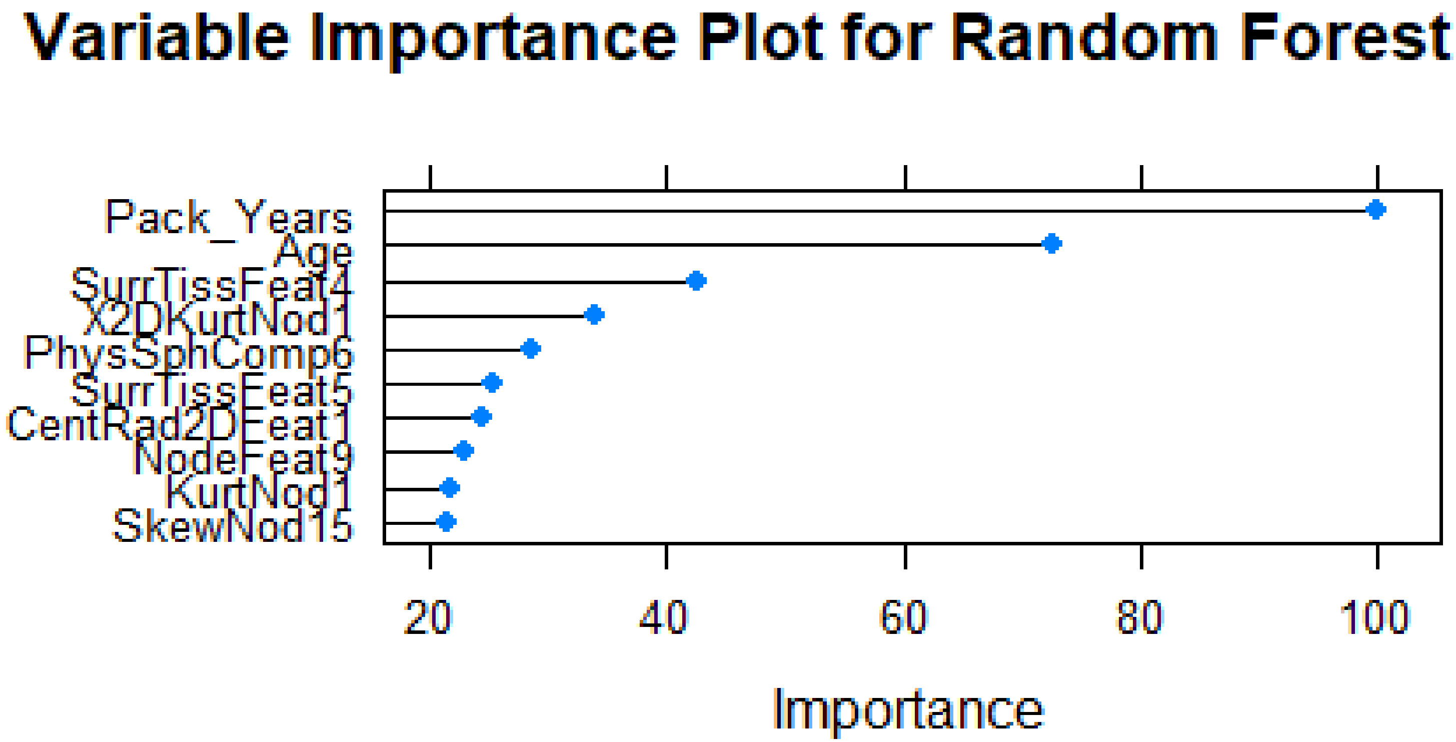
Thank You!

Waits for Audience to Clap

Variable Importance – Elastic Net



Variable Importance – Random Forest



Variable Importance – Logistic