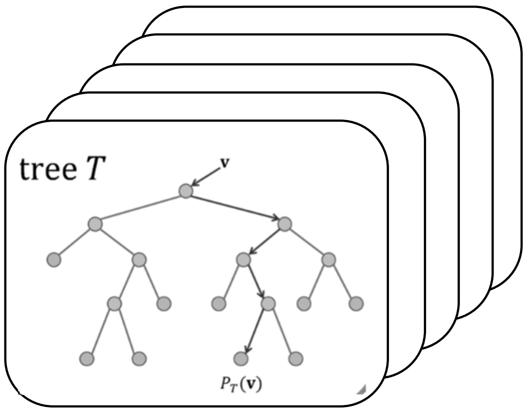
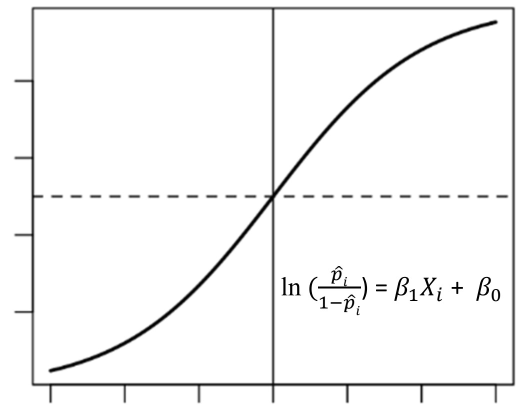
Random Forest vs. Logistic Regression in Predictive Analytics Applications



VS



John Stanley, Director of Institutional Research
Christi Palacat, Institutional Analyst
University of Hawai'i – West O'ahu

UNIVERSITY of HAWAI'I WEST O'AHU

TAIR Conference XLII ● March 1– 4, 2020, San Antonio, TX

Predictive Analytics

- ‘Predictive analytics’ (PA) increasingly prevalent in institutional research (89% investment according to 2018 AIR/NASPA/Educause survey).
- First-year retention probably the most common outcome targeted in PA applications.
- ‘Big data’ environment driving a proliferation of data mining in PA applications.

Today's Objectives

- Overview key differences between classical statistics and data mining, with particular examination of logistic regression and random forest methods.

- Examine results from a U.Hawai'i study that used logistic regression and random forest methods to predict enrollment outcomes.

Relevant Previous Research

Astin, A. (1993). *What matters in college: Four critical years revisited.*

Breiman, L. (2001) Random Forests. *Machine Learning.*

Goenner, C. & Pauls, K. (2006). A predictive model of inquiry to enrollment. *Research in Higher Education.*

He, L., Levine, R., Fan, J., Beemer, J., & Stronach, J. (2017). Random forest as a predictive analytics alternative to regression in institutional research. *Practical Assessment, Research & Evaluation.*

Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in Higher Education.*

Herzog, S. (2006). Estimating student retention and degree completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research.*

Kabacoff, R. (2015). *R in action: Data analysis and graphics with R.*

Pride, B. (2018). Data science: Using data to understand, predict, and improve outcomes. Presented at the 2018 AIR Forum, Orlando, FL.

Review of Approaches

Classical Statistics	Data Mining
Deductive – Provides theory first and then tests it using various statistical tools. Process is cumulative.	Inductive– It explores data first, then extracts a pattern and infers an explanation or a theory. Process is ad hoc.
Formalizes a relationship in the data in the form of a mathematical equation.	Makes heavy use of learning algorithms that can work semi-automatically or automatically.
More concerned about data collection.	Less concerned about data collection.
Statistical methods applied on clean data.	Involves data cleaning (non-numeric data okay, missing data handled internally).
Usually involves working with small datasets or samples of a population (e.g. inference statistics)	Usually involves working with large datasets (i.e., "Big Data").
Needs more user interaction to validate model.	Needs less user interaction action to validate model, therefore possible to automate.
There is no scope for heuristics think.	Makes generous use of heuristics think.

Adapted from: https://www.educba.com/data-mining-vs-statistics/

Review of Methods

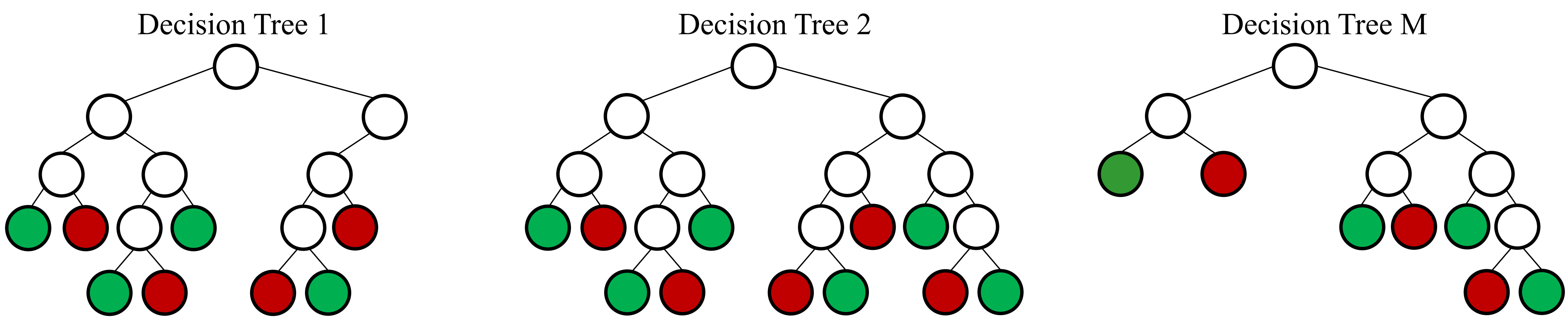
Logistic Regression	Random Forest
Path analysis approach, uses a generalized linear equation to describe the directed dependencies among a set of variables.	Top-down induction based approach to classification and prediction. Averages many decision trees (CARTs) together.
A number of statistical assumptions must be met.	No statistical assumptions; can handle multicollinearity.
Overfitting a concern (rule of ten), as well as outliers.	Robust to overfitting and outliers.
Final model should be parsimonious and balanced.	Final model depends on the strength of the trees in the forest and the correlation between them.
A number of complementary measures can be used to assess goodness of fit (i.e., -2LL, ~R^2, HL).	Random inputs and random features tend to produce better results in RFs (Breiman, 2001).
Logit link function: $\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_1 X_i + \beta_0$	CART Gini impurity algorithm: $\sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$

UNIVERSITY of HAWAI'I WEST O'AHU

Random Forest – bagging and voting

Subsample 1 | Subsample 2 | Subsample M

$$S_1 = \begin{bmatrix} f_{A10} & f_{D10} & f_{M10} & f_{R10} & C_{10} \\ f_{A33} & f_{D33} & f_{M33} & f_{R33} & C_{33} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{A99} & f_{D99} & f_{M99} & f_{R99} & C_{99} \end{bmatrix} S_2 = \begin{bmatrix} f_{B18} & f_{G18} & f_{P18} & f_{Z18} & C_{18} \\ f_{B49} & f_{G49} & f_{P49} & f_{Z49} & C_{49} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{B98} & f_{G98} & f_{P98} & f_{Z98} & C_{98} \end{bmatrix} \ldots S_M = \begin{bmatrix} f_{C22} & f_{F22} & f_{K22} & f_{Q22} & C_{22} \\ f_{C51} & f_{F51} & f_{K51} & f_{Q51} & C_{51} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{C77} & f_{F77} & f_{K77} & f_{Q77} & C_{77} \end{bmatrix}$$



Research Questions

- Does random forest produce better classification accuracy than logistic regression when predicting admission yield at a large R1 university?
- Which method does enrollment management and admissions find easier to interpret?

UNIVERSITY of HAWAI'I WEST O'AHU

Predictive Analytics Approach to Admission Yield

- Identify ‘fence sitter’ non-resident freshmen accepts at peak recruitment season (February 15^{th})
- Develop regression and random forest models to predict enrollment likelihood of future cohort
  - Compare/contrast models’ predictive accuracy, flexibility, interpretability.
- Enrollment likelihood scoring for admitted non-resident freshmen
  - Automated classification and probability score with SPSS (LR) and R (RF); Decile grouping of scored students and “top prospects”
- Reporting of enrollment likelihood via secure online access

UNIVERSITY of HAWAI‘I WEST O‘AHU

Data Description

- Data sources
  - Matriculation system (Banner)
- Student cohorts
  - New first-time freshmen non-resident admits (University of Hawai'i at Manoa)
  - Fall entry 12', 13', 14', 15', 16' for model dev. (training set, N=16,420)
  - Fall entry 17' for model validation (holdout set, N=4,270); 18% baseline yield
- Data elements at February 1
  - Contact: expressed interest, number of applications
  - Geographic: distance, residency, high yield geog region, high yield high school
  - Geodemographic: geog. region by ethnicity, gender, SES
  - Academic: program of study
  - Timing: date of application days/weeks until semester start
  - Financial: FAFSA submitted

UNIVERSITY of HAWAI'I WEST O'AHU

Data Analysis Steps

- Exploratory data analysis
  - Variable selection (bivariate correlation on outcome variable)
  - Variable coding (continuous vs. dummy/binary (LR) vs. columnar form (RF))
  - Missing data imputation
  - Derived variable(s)
    - HSPrep = (HSGPA*12.5)+(ACTM*.69)+(ACTE*.69) (not used today)
- Logistic regression model (SPSS)
  - Preliminary model fit (-2LL test/score, pseudo R2, HL sig.)
  - Refine model fit with forward and backwards elimination of independent variables; choose parsimonious model
  - Check for outliers with diagnostic tools (Std residuals, Cook's D)
  - Check for collinearity (VIF)
  - Check correct classification rate (CCR) for enrollees vs. non-enrollees (i.e., model sensitivity vs. specificity) using baseline probability and Receiver Operating Characteristics (ROC) curve. Make further refinements to cut value.
  - Check for consistency across training sets (stratified sampling)

UNIVERSITY of HAWAI'I WEST O'AHU

Data Analysis Steps (cont.)

- Random Forest (R Studio)
  - Set hyperparameters in Random Forest:
    - Number of trees to grow in the forest. Typical values are around 100-500. More trees sometimes leads to overfitting.
    - Number of variables randomly sampled as candidates at each split for a particular tree. Default is $\sqrt{\#\ of\ variables}$. Check the out-of-bag (OOB) error rate.
    - Sampling can be done with or without replacement (we "set the seed" in order to replicate results).
    - Check correct classification rate (CCR) for enrollees vs. non-enrollees (i.e., model sensitivity vs. specificity) using baseline probability and Receiver Operating Characteristics (ROC) curve. Make further refinements to cut value.

LR Results from SPSS

Logistic Regression Model Accuracy

Enrollment Decision	**Correct Classification %**
Non-Enrolled	80.9
Enrolled	54.5
Overall Accuracy	76.4
Hosmer-Lemeshow	P < .000
Pseudo R^2	.274

First- Time Full-Time Nonresident Freshmen Fall Accepts 12', 13', 14', 15', 16' for model development (training set, N=16,420) ; Fall entry 2017 for model validation (holdout set, N=4,270). Correct classification results are for holdout set. The cut value is .3325. Hosmer-Lemeshow chi-square = 56.565 (p<.000).

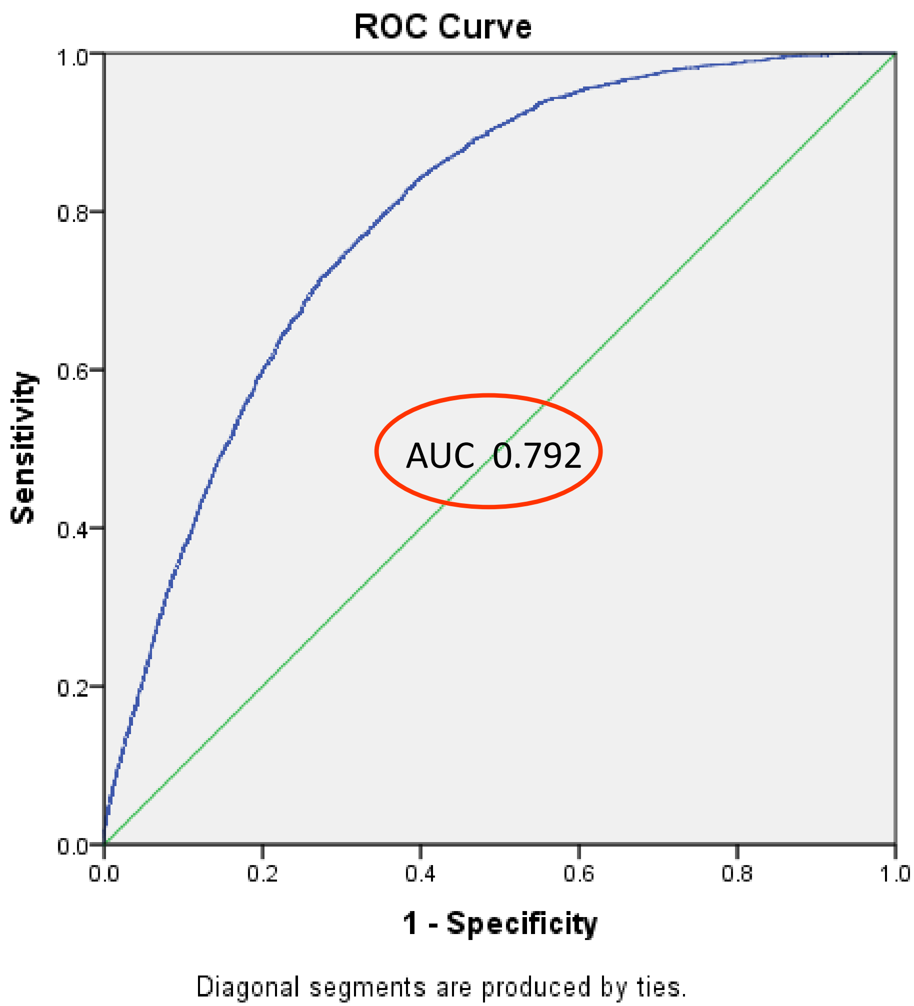
Delta P statistics are calculated using Cruce's formula for categorical variables and Petersen's formula for continuous variables.

Nonresident Freshmen Admissions Yield Predictors (LR)

Variable	**Beta**	**Wald**	**Sig.**	**▼Delta P**	**VIF**
1. No SAT Math Score Reported by Feb 1	-2.937	180.221	0.000	-62%	1.159
2. Completed FAFSA by Feb 1	1.231	554.107	0.000	20%	1.237
3. WUE	1.022	368.327	0.000	17%	1.173
4. High School GPA- Greater than 3.99	-0.904	122.058	0.000	-17%	1.255
5. SAT Writing- Greater than 660	-0.581	53.141	0.000	-11%	1.517
6. Native Hawaiian	0.809	57.059	0.000	10%	1.017
7. High School GPA - Less than 3.00	0.556	59.945	0.000	8%	1.096
8. High School GPA - Between 3.67 and 3.99	-0.456	59.745	0.000	-8%	1.198
9. SAT Writing- Less than 500	0.453	35.176	0.000	7%	1.127
10. Two or more Previous Contacts	0.444	47.012	0.000	6%	1.026
11. Pacific Islander	0.427	6.127	0.013	6%	1.019
12. SAT Writing- Between 590 and 660	-0.262	26.321	0.000	-4%	1.337
13. No High School GPA Reported by Feb 1	0.279	13.596	0.000	4%	1.145
14. SAT Math -Greater than 660	-0.230	7.501	0.006	-4%	1.517
15. Age	0.175	24.210	0.000	3%	1.019
16. Total Grant Amount (per $100)	0.024	301.859	0.000	< 1%	1.281
17. Application Date First Day Instruction Gap	-0.014	10.981	0.001	< 1%	1.038
Constant	-5.602	71.723	0.000		

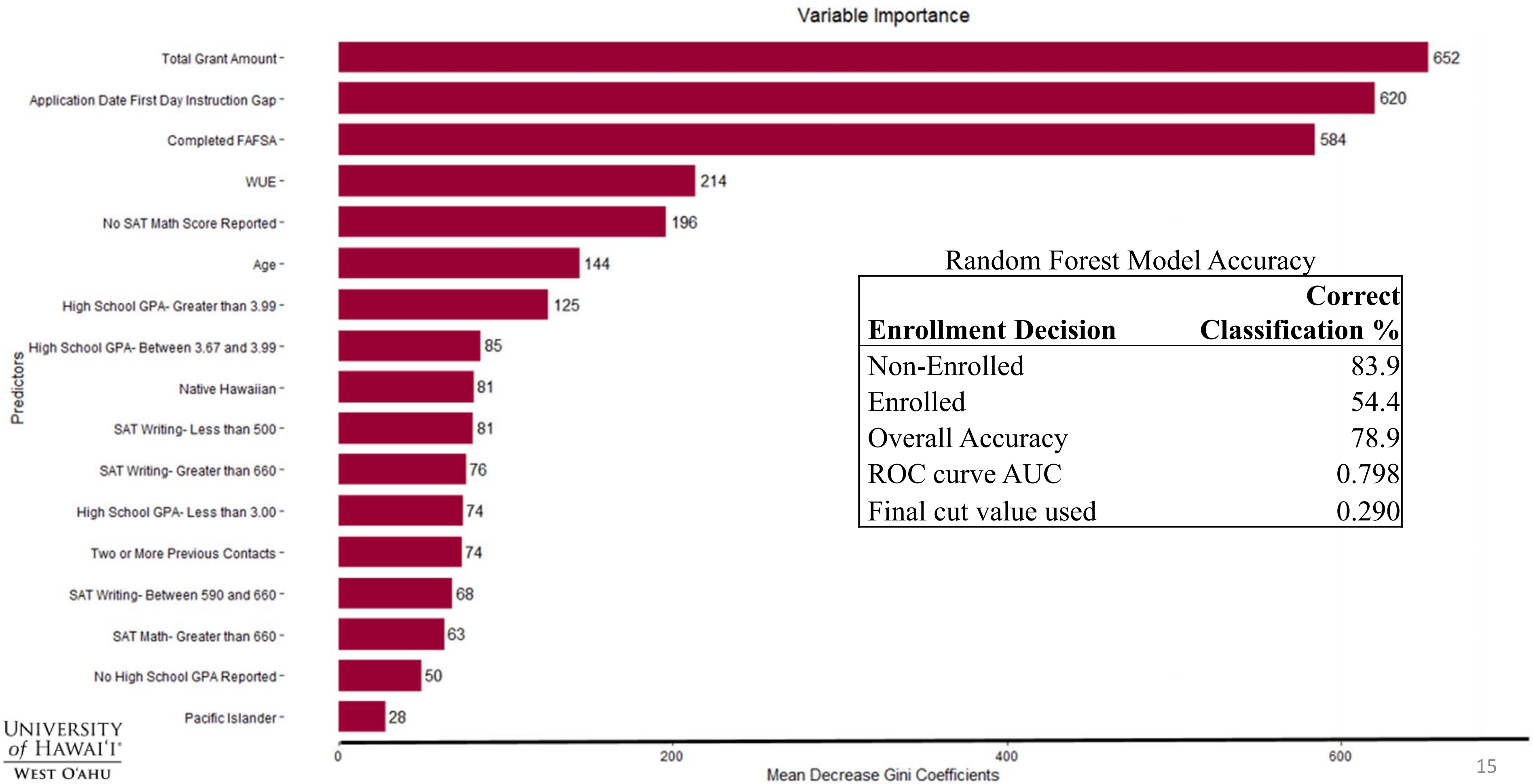
UNIVERSITY of HAWAI'I WEST O'AHU

LR ROC Curve (SPSS)



UNIVERSITY of HAWAI'I
WEST O'AHU

RF Results from R – version 1, identical dataset as LR

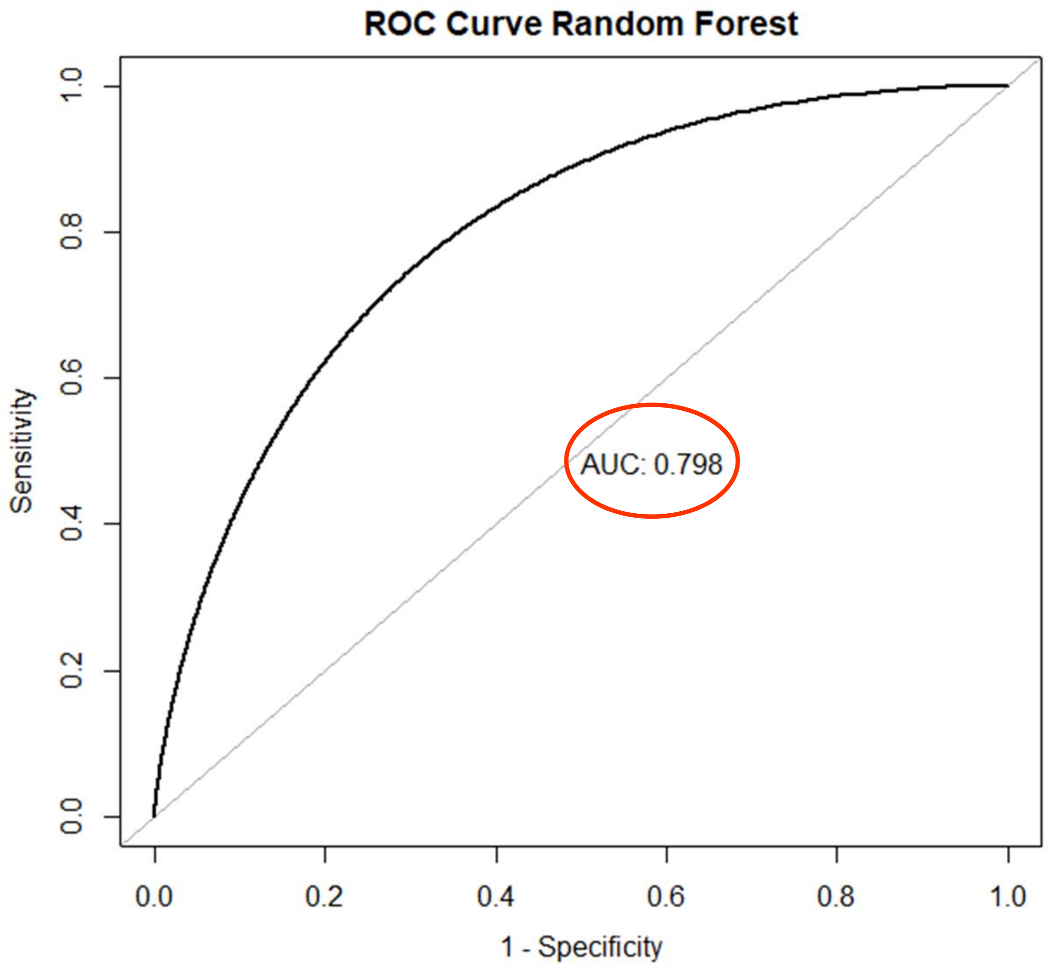


Random Forest Model Accuracy

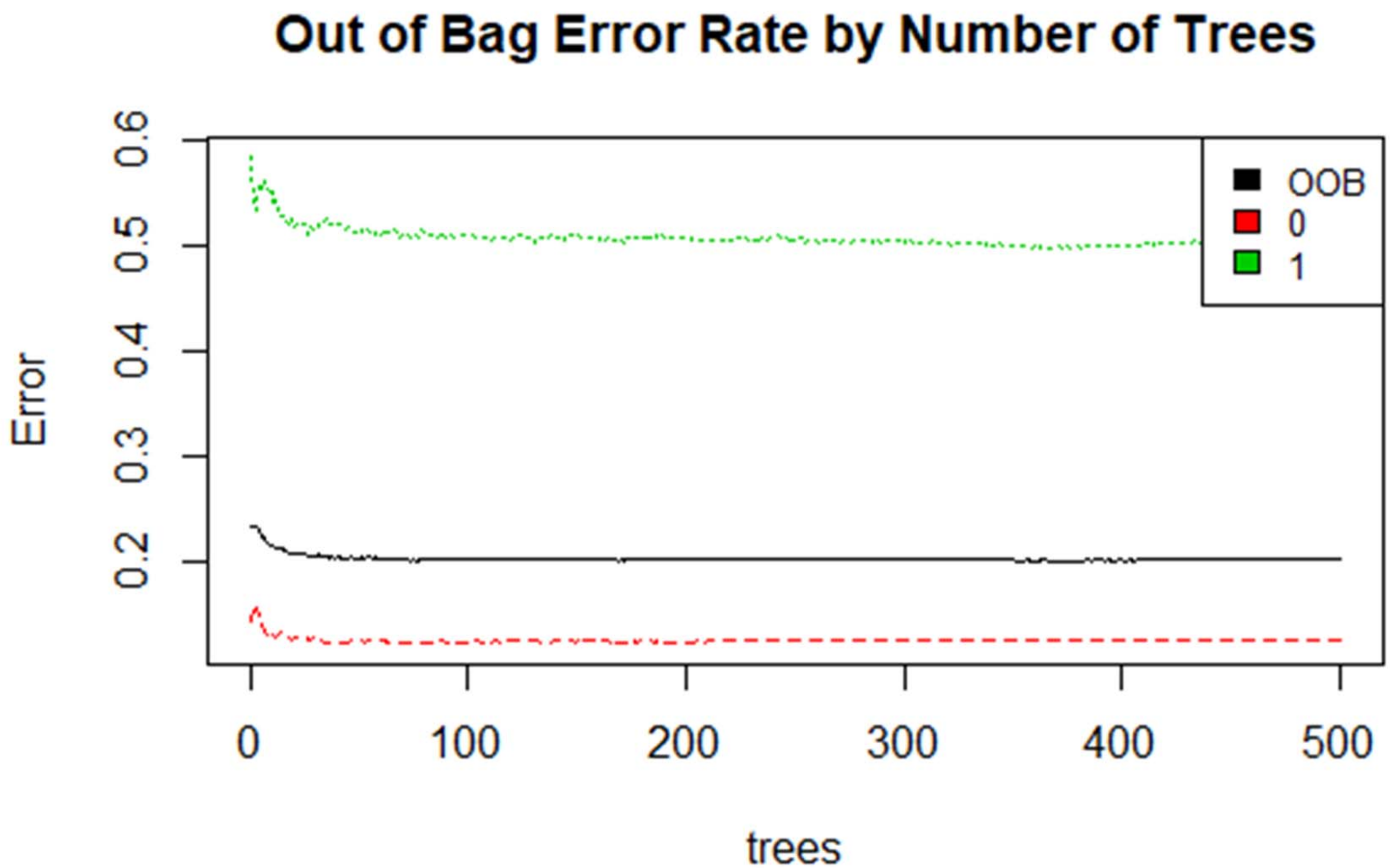
Enrollment Decision	Correct Classification %
Non-Enrolled	83.9
Enrolled	54.4
Overall Accuracy	78.9
ROC curve AUC	0.798
Final cut value used	0.290

University of Hawai'i West O'ahu

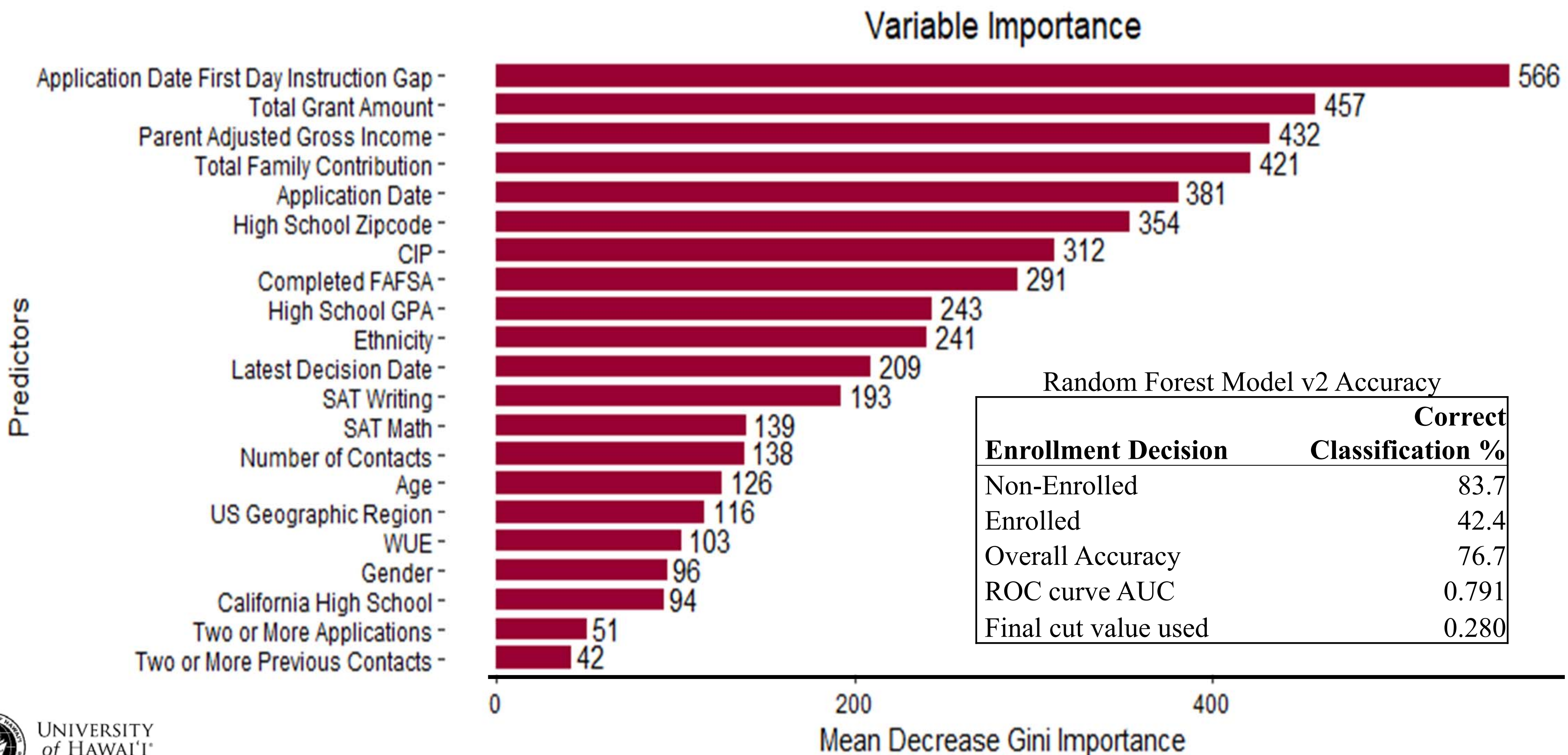
RF ROC Curve (R)



Random Forest Error Rate V1 (R)



RF Results Version 2 – data prepared for RF analysis

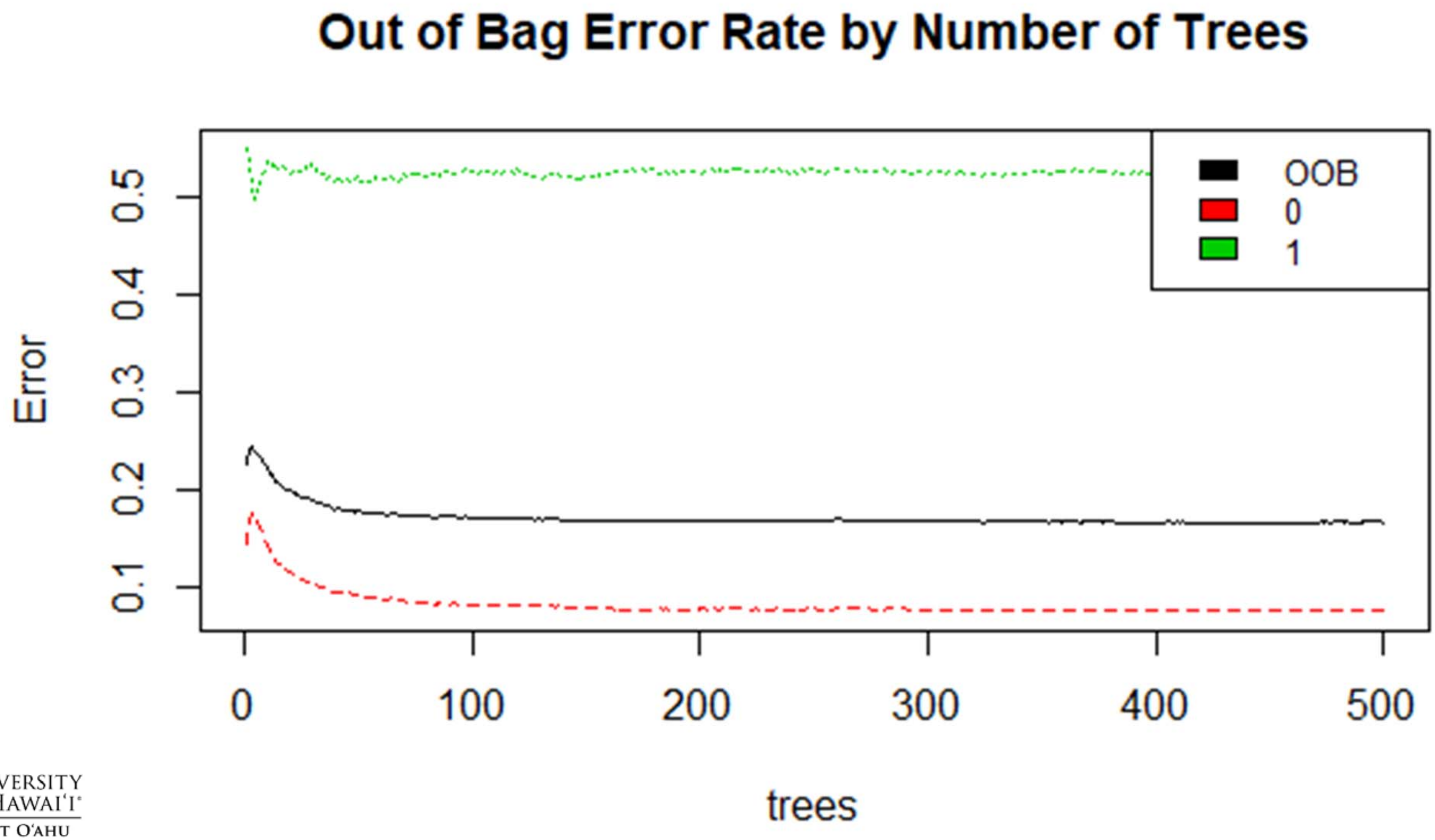


Random Forest Model v2 Accuracy

Enrollment Decision	**Correct Classification %**
Non-Enrolled	83.7
Enrolled	42.4
Overall Accuracy	76.7
ROC curve AUC	0.791
Final cut value used	0.280

UNIVERSITY of HAWAI'I WEST O'AHU

Random Forest Error Rate V2 (R)



Model Accuracy: Random Forest vs Logistic Regression

Correct Classification Rate (%)

Admission Decision	**RF(v1)**	**LR**
Non-Enrolled	83.9	80.9
Enrolled	54.4	54.5
Overall accuracy	78.9	76.4

LR= Logistic Regression; RF= Random Forest

Logistic Regression Syntax (SPSS)

```
**LR baseline run**

LOGISTIC REGRESSION VARIABLES ENR_IRO_IND
  /SELECT=Training1 EQ 1
  /METHOD=ENTER CONTACTS_2_OR_MORE APPL_GAP_DIVBY10 AGE NATIVE_HAWAIIAN PACIFIC_ISLANDER WUE
    COMPLETED_FAFSA_IND GRANTS_AMOUNT_DIV100 SRHSGPA_LESS_3 SRHSGPA_367_399 SRHSGPA_GREATER_399 SRHSGPA_NODATA
    CONV_SATM_GREATER_660 CONV_SATM_NODATA CONV_SATW_LESS_500 CONV_SATW_590_660 CONV_SATW_GREATER_660
  /PRINT=GOODFIT
  /SAVE=ZRESID COOK
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.21).

**Outlier Exclusion Rule**

DATASET ACTIVATE DataSet1.
USE ALL.
COMPUTE filter_$=(COO_1 < .1 & ZRE_1 < 3 & ZRE_1 >  - 3).
VARIABLE LABELS filter_$ 'COO_1 < .1 & ZRE_1 < 3 & ZRE_1 >  - 3 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

**LR final run**

LOGISTIC REGRESSION VARIABLES ENR_IRO_IND
  /SELECT=Training1 EQ 1
  /METHOD=ENTER CONTACTS_2_OR_MORE APPL_GAP_DIVBY10 AGE NATIVE_HAWAIIAN PACIFIC_ISLANDER WUE
    COMPLETED_FAFSA_IND GRANTS_AMOUNT_DIV100 SRHSGPA_LESS_3 SRHSGPA_367_399 SRHSGPA_GREATER_399 SRHSGPA_NODATA
    CONV_SATM_GREATER_660 CONV_SATM_NODATA CONV_SATW_LESS_500 CONV_SATW_590_660 CONV_SATW_GREATER_660
  /PRINT=GOODFIT
  /SAVE=PRED PGROUP
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.3325).

**ROC Curve**

DATASET ACTIVATE DataSet1.
ROC PRE_3 BY ENR_IRO_IND (1)
  /PLOT=CURVE(REFERENCE)
  /PRINT=SE
  /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING=EXCLUDE.
```

UNIVERSITY of HAWAI'I WEST O'AHU

Random Forest Syntax (R Studio)

Source on Save | Run | Source

```r
#Load necessary packages
library(randomForest)
library(pROC)
library(ROCR)
library(ggplot2)

# Nonresident Model/ Classification Table
data.nonres<- (Final_NonResident_Data_Feb_23_filtered)
train.nonres<- data.nonres[1:16090,]
unselected.nonres <- data.nonres[16091:20283,]
train.nonres$ENR_IND<- as.factor(train.nonres$ENR_IRO_IND)
set.seed(9073)
rforest.nonres<-randomForest(train.nonres$ENR_IND~ CONTACTS_2_OR_MORE + APPL_GAP_DIVBY10 + AGE + NATIVE_HAWAIIAN + PACIFIC_ISLANDER + WUE + COMPLETED_FAFSA_IND + GRANTS_A
                             , data=train.nonres, cutoff=c(0.71,0.29), importance=TRUE)
forest.pred.unselected.nonres<-predict (rforest.nonres, unselected.nonres)
forest.perf.unselected.nonres <- table(unselected.nonres$ENR_IRO_IND,forest.pred.unselected.nonres, dnn=c("Observed","Predicted"))

#ROC curve; Area under the curve= 0.798
nonres.rf.pr<-predict(rforest.nonres, unselected.nonres, type='prob')
roc.nonres<- roc(unselected.nonres$ENR_IRO_IND,nonres.rf.pr[,2],smooth=TRUE, plot=TRUE, main= "ROC Curve Random Forest", legacy.axes=TRUE, asp=NA, print.auc=TRUE)

# Bar chart for Mean Decrease Gini
imp<-importance(rforest.nonres)
imp.selected<-imp[,c(4)]
write.csv(imp.selected,"imp.csv",row.names=TRUE)
imp.csv<- read.csv("imp.csv",header=TRUE)
colnames(imp.csv)=c("Variable", "MeanDecreaseGini")
imp.sort<- imp.csv[order(-imp.csv$MeanDecreaseGini),]
imp.sort<-transform(imp.csv, Variable= reorder(Variable,MeanDecreaseGini))
ggplot(imp.sort,aes(x=Variable, y=MeanDecreaseGini,hjust=-0.2,vjust=0.4)) + labs(title="Variable Importance",x= "Predictors", y="Mean Decrease Gini Coefficients")+ theme(
  scale_x_discrete(labels=c("APPL_GAP_DIVBY10"="Application Date First Day Instruction Gap", "GRANTS_AMOUNT_DIV100"="Total Grant Amount","COMPLETED_FAFSA_IND"="Completed
```

Study Limitations

- Little collinearity, randomness, or complexity in variables, so perhaps not the best dataset for Random Forest.
- IVs with low correlation with DV were largely left out of the dataset (since we were approaching this with a regression mindset) but may have otherwise contributed to prediction accuracy in the RF.
- Imbalanced outcome data could affect RF results.

UNIVERSITY of HAWAI'I WEST O'AHU

Extensions of Random Forest in IR

Freshmen Retention Prediction (UH West O'ahu data)

Prediction Model Correct Classification Rate (%)

Retention Outcome	Start of Term (LR)	Start of Term (RF)	End of Term (LR)	End of Term (RF)
Dropouts	61.0	69.5	89.9	91.1
Retainees	61.9	61.9	69.3	58.2
Overall Accuracy	61.6	64.2	75.4	67.9
Pseudo R^2	0.127	N/A	0.398	N/A

LR= Logistic Regression; RF= Random Forest

University of Hawai'i West O'ahu

Enrollment Managers' Reactions

- Logistic Regression
  - Felt that the Delta P statistic was highly intuitive.
  - Liked being able to see the directionality in coefficients.

- Random Forest
  - Finding the cut points for institutional grant aid and total offer amount is operationally useful.
  - Wanted to see a side-by-side comparison of the RF and LR effect scores.

University of Hawai'i West O'ahu

Conclusion

- The random forest model performed at parity with the binomial logistic regression model in terms of prediction accuracy.
- The level of complexity of the data used and the outcome predicted may largely guide the selection of a particular analytical tool.
- Random forest may be ideal candidate for estimating time-to-degree where the dataset is more longitudinal in nature (i.e., more complexity and randomness).
- Conversations with admissions and enrollment management favored the logistic regression analysis as easier to interpret (i.e., goodness of fit stats, Delta P statistic, directionality).

Questions

uhwoiro@hawaii.edu

https://westoahu.hawaii.edu/academics/institutional-research/