Predicting Student Retention by Applying Machine Learning Algorithms

Presenter:

- Dr. Jeremy Monteath-Valdez (Interim Assistant Director of Academic Scheduling)
- Daniel Le (Assistant Director of Strategic Analytics)

Introduction

- Student retention is one crucial key performance indicator in higher education that may signify whether students are satisfied with their institution.
- Colleges can use this indicator to determine if they need to concentrate on establishing new programs for students that help keep them engaged in their classes and involved on campus.
- Reasons why first year students do not return to college include:
  - Lack of financial aid
  - Institutional factors such as nonflexible schedule choices
  - Individual factors such as socio-economic and disadvantage backgrounds

Method

- Datasets include 11,262 students entering Dallas College for the first time in Fall 2019 and 7,921 students entering Dallas College for the first time in Fall 2020. We exclude Dual Credit and Early College High School (ECHS) students.
- Demographic information, admission status, classification, amount of enrollment credits, financial aid information, and first term GPA were collected.
- This research project uses two types of machine learning models to identify significant factors in predicting student retention: logistic regression and random forest models.

Description of Dataset

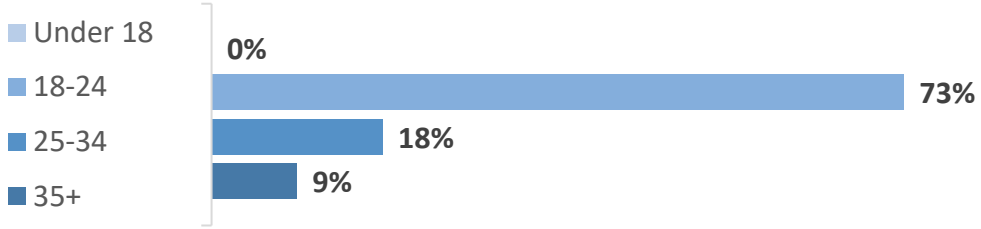
Variable	Values	Description
Y	**0** (Did not return), **1** (Returned)	Response variable indicating whether a student returned to Dallas College for their consecutive second year.
ADMIT_STATUS	**CT** (College/University Transfer), **GED** (General Education Diploma), **HG** (High School Graduate), **HOM** (Home School Graduate)	Admission status the student entered as.
CLASS	**FR** (Freshman), **SO** (Sophomore), **UN** (Unknown)	Classification of the student.
ENRL_CREDS	**1**, **2**, **3**…	The number of enrolled credits the student took in their first semester.
GENDER	**F**, **M**	Gender of the student.
AGE	**16**, **17**, **18**…	Age of the student.
RACE_ETH	**American Indian or Alaskan Native**, **Asian**, **Black/African-American**, **Hispanic**, **International**, **Multiple Races**, **Native Hawaiian or Other Pacific Islander**, **Unknown or Not Reported**, **White**	Race ethnicity of the student.
LOAN_AMOUNT	A non-negative number.	Amount of loans that the student received in thousands.
GIFT_AID_AMOUNT	A non-negative number.	Amount of any gift aid the student received in thousands.
FA2019_GPA or FA2020_GPA	From **0** to **4** inclusive.	Student's term GPA at the end of Fall 2019 or Fall 2020.

Descriptive Statistics of Data Samples

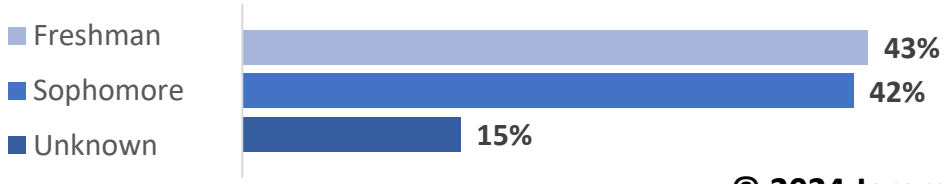
2019

Gender

Age



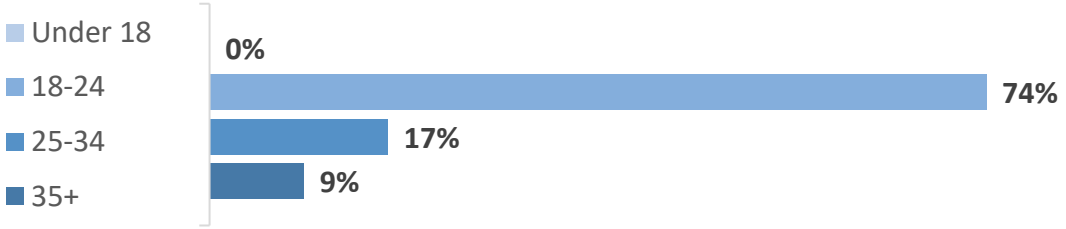
Classification



2020

Gender

Age

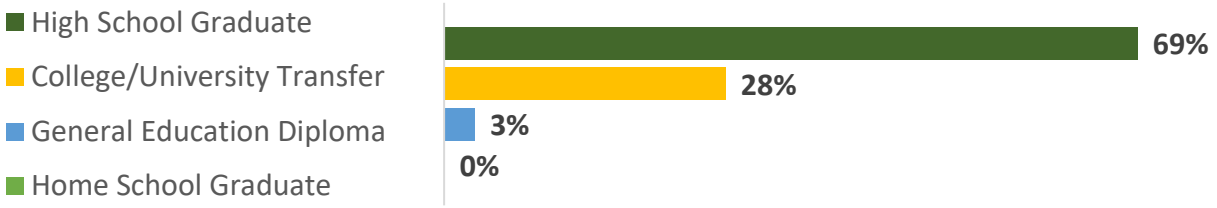


Classification

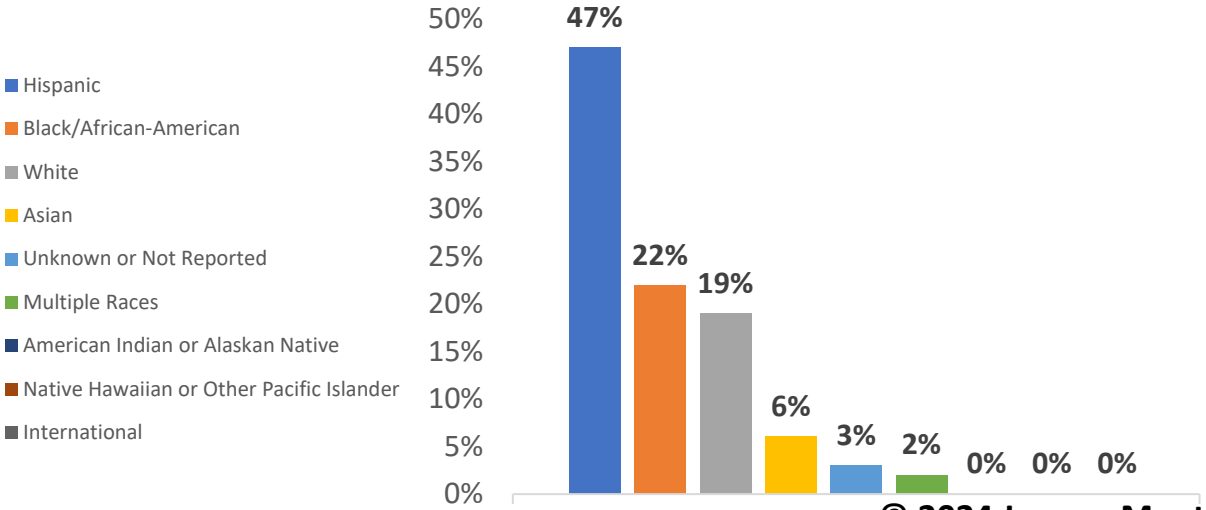
Descriptive Statistics of Data Samples

2019

Admission Status

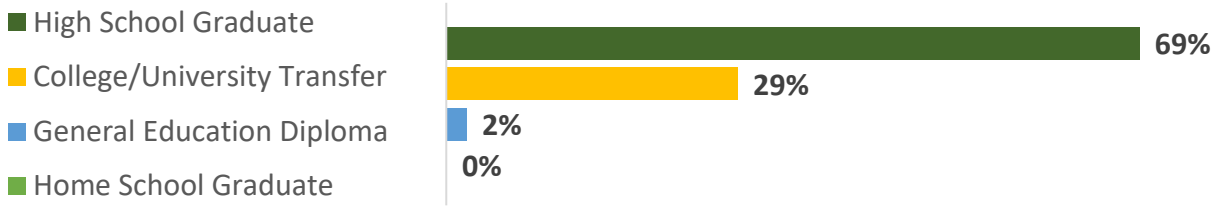


Race/Ethnicity

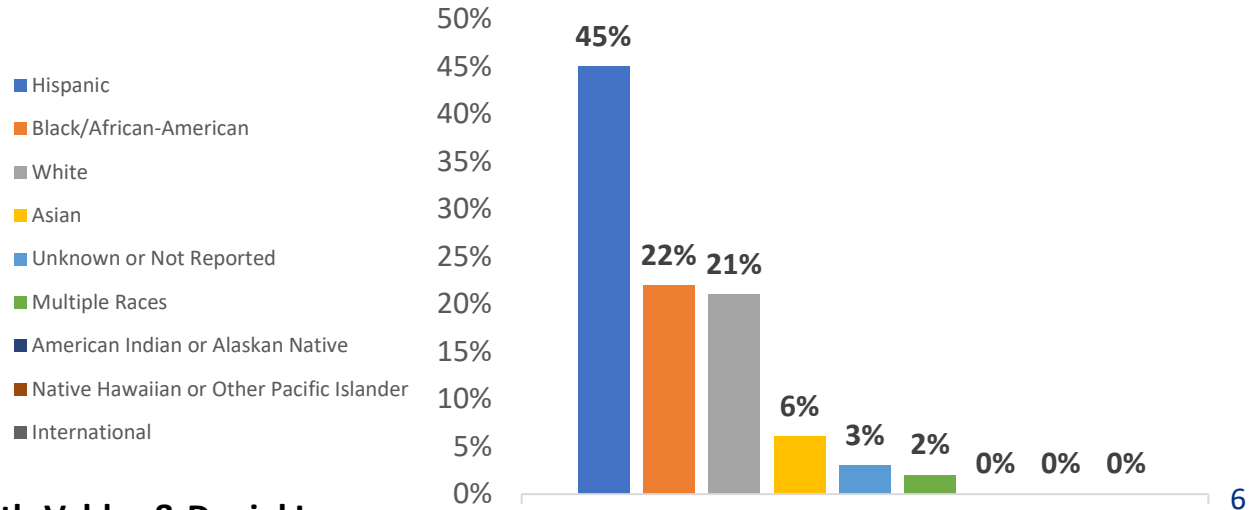


2020

Admission Status



Race/Ethnicity



Descriptive Statistics of Data Samples

2019

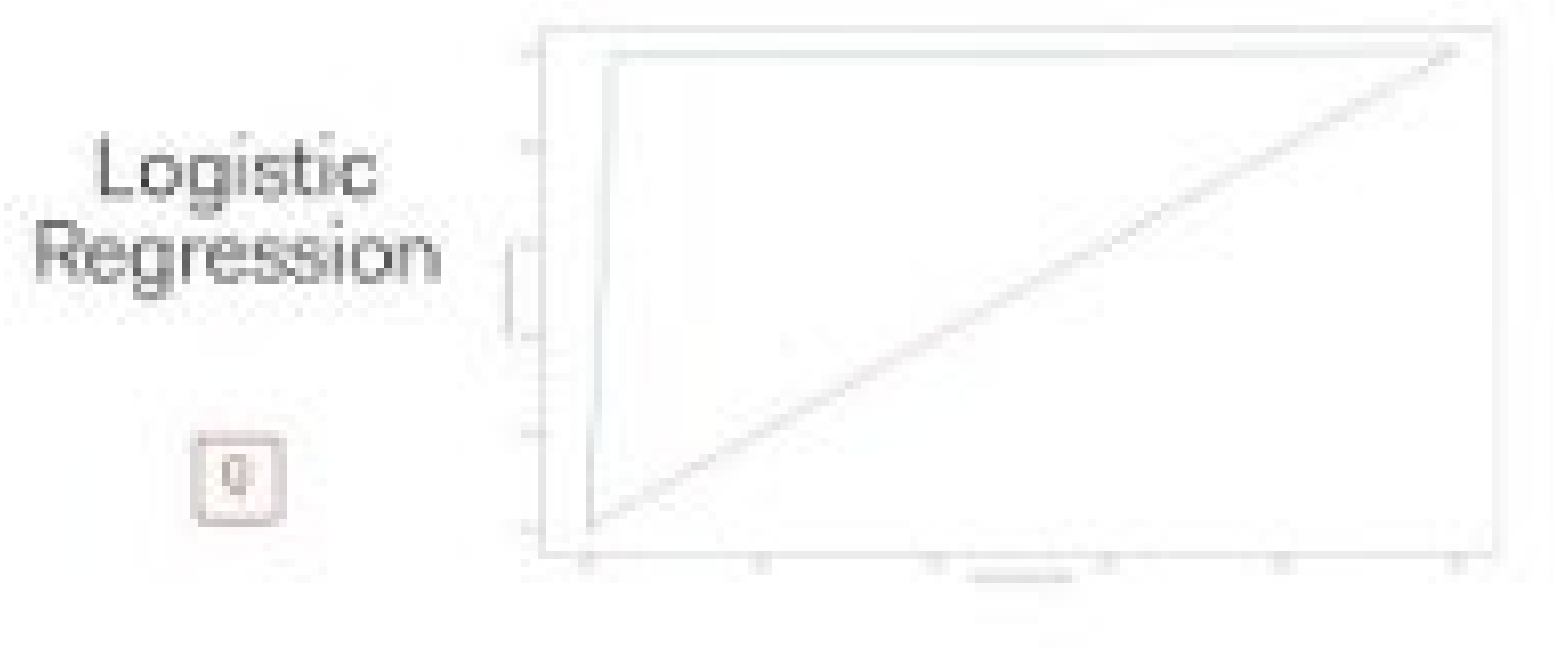
Variable	Label	Mean	Std Dev	Minimum	Maximum
ENRL_CREDS	ENRL_CREDS	9.37	3.62	1.00	24.00
AGE	AGE	24.63	7.41	16.00	87.00
LOAN_AMOUNT	LOAN_AMOUNT	2363.40	3013.28	0.00	18854.00
GIFT_AID_AMOUNT	GIFT_AID_AMOUNT	3057.59	3249.71	0.00	18923.00
FA2019_GPA	FA2019_GPA	2.35	1.39	0.00	4.00

2020

Variable	Label	Mean	Std Dev	Minimum	Maximum
ENRL_CREDS	ENRL_CREDS	9.09	3.80	1.00	21.00
AGE	AGE	23.70	7.21	17.00	81.00
LOAN_AMOUNT	LOAN_AMOUNT	1695.35	2842.75	0.00	19000.00
GIFT_AID_AMOUNT	GIFT_AID_AMOUNT	2759.05	3106.90	0.00	17130.00
FA2020_GPA	FA2020_GPA	2.19	1.53	0.00	4.00

Response (outcome) variable (predictive model):
Retained (1: Student returns to Dallas College for their second consecutive year, otherwise it's 0).

Logistic Regression



Logistic Regression Estimates

2019

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	**-2.6858**	0.3367	63.6166	**< 0.0001**
ADMIT_STATUS_CT	-0.1647	0.3188	0.2669	0.6055
ADMIT_STATUS_GED	0.3363	0.3466	0.9413	0.3319
ADMIT_STATUS_HOM	-1.0651	0.9359	1.2952	0.2551
CLASS_SO	**0.4910**	0.0411	142.5952	**< 0.0001**
CLASS_UN	**1.2024**	0.0619	377.3280	**< 0.0001**
ENRL_CREDS	**0.0695**	0.00910	58.4266	**< 0.0001**
GIFT_AID_AMOUNT	**0.0910**	0.00941	93.4899	**< 0.0001**
FA2019_GPA	**0.4709**	0.0264	317.7590	**< 0.0001**

2020

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	**-1.5212**	0.3468	19.2419	**< 0.0001**
ADMIT_STATUS_CT	-0.5045	0.2872	3.0852	0.0790
ADMIT_STATUS_GED	0.0147	0.3278	0.0020	0.9643
ADMIT_STATUS_HOM	-0.0559	0.8247	0.0046	0.9460
CLASS_SO	**0.8620**	0.0599	206.9904	**< 0.0001**
CLASS_UN	0.1008	0.0932	1.1683	0.2798
ENRL_CREDS	**0.0618**	0.00966	40.9116	**< 0.0001**
GIFT_AID_AMOUNT	**0.0814**	0.0107	57.3981	**< 0.0001**
FA2020_GPA	**0.5190**	0.0259	400.6939	**< 0.0001**
AGE	**-0.0279**	0.00579	21.1860	**< 0.0001**

Logistic Regression Odds Ratio

2019

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
ADMIT_STATUS CT vs BASE_HG	0.347	0.299	0.403
ADMIT_STATUS GED vs BASE_HG	0.573	0.385	0.853
ADMIT_STATUS HOM vs BASE_HG	0.141	0.012	1.624
CLASS SO vs FR	8.884	7.753	10.181
CLASS UN vs FR	18.096	14.699	22.278
ENRL_CREDS	1.072	1.053	1.091
GIFT_AID_AMOUNT	1.095	1.075	1.116
FA2019_GPA	1.602	1.521	1.687

2020

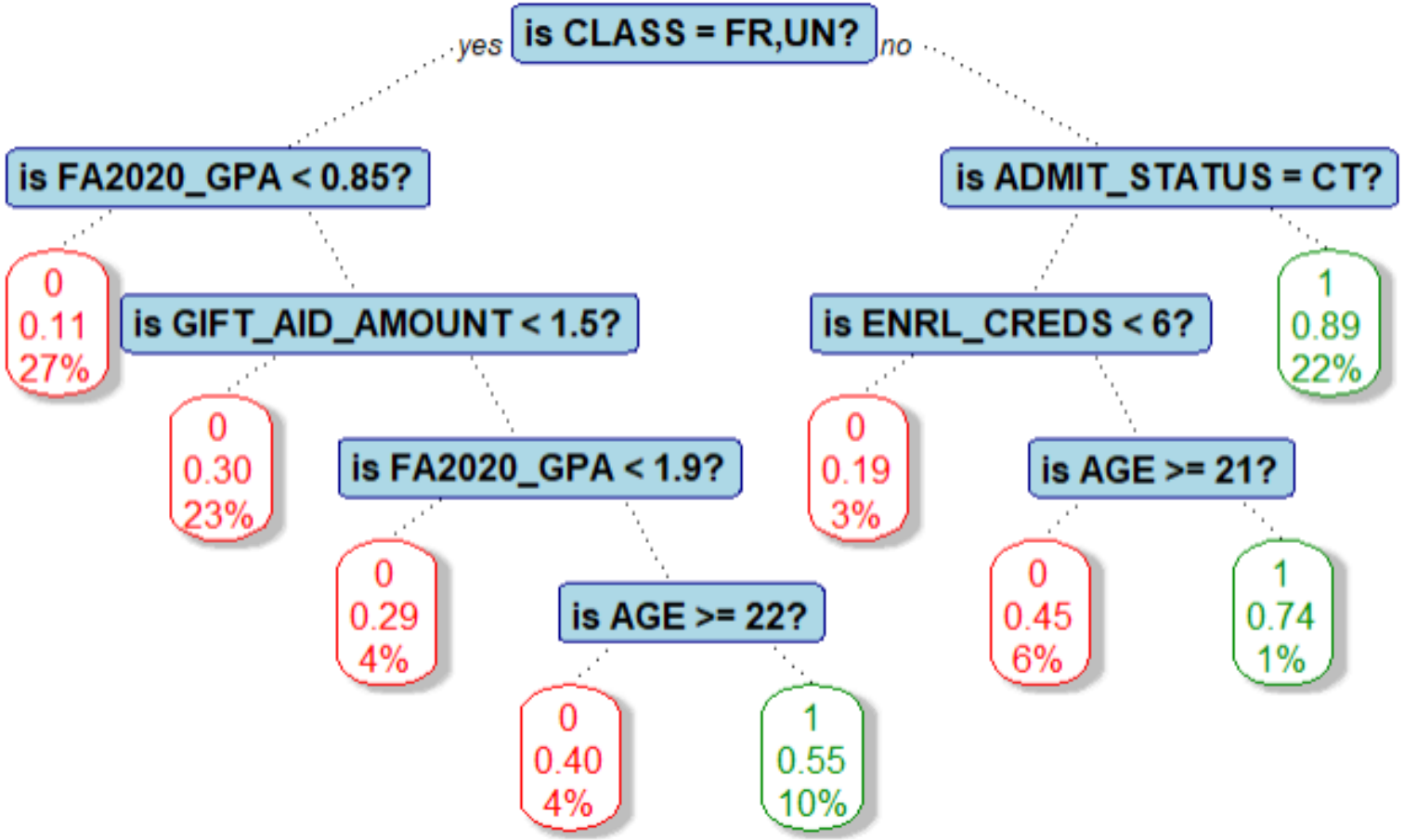
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
ADMIT_STATUS CT vs BASE_HG	0.350	0.289	0.424
ADMIT_STATUS GED vs BASE_HG	0.588	0.366	0.945
ADMIT_STATUS HOM vs BASE_HG	0.548	0.064	4.702
CLASS SO vs FR	6.201	5.247	7.330
CLASS UN vs FR	2.897	2.158	3.889
ENRL_CREDS	1.064	1.044	1.084
AGE	0.973	0.962	0.984
GIFT_AID_AMOUNT	1.085	1.062	1.108
FA2020_GPA	1.680	1.597	1.768

Odds Ratio Interpretation

- Before the pandemic, the odds of returning to Dallas College the following year are:
  - Increased by 7.2% for each additional increase in Fall 2019 enrollment credits.
  - Increased by 9.5% for every $1000 additional increase in Gift Aid.
  - Increased by 60% for each additional increase in Fall 2019 GPA.
  - 788% higher for a sophomore than a freshman.
  - 1709% times higher for a student classified as unknown than a freshman.
- During the pandemic, the odds of returning to Dallas College the following year are:
  - Increased by 6.4% for each additional increase in Fall 2020 enrollment credits.
  - Increased by 8.5% for every $1000 additional increase in Gift Aid.
  - Increased by 68% for each additional increase in Fall 2020 GPA.
  - 520% higher for a sophomore than a freshman.
  - Decreased by 2.7% for every additional increase in age.

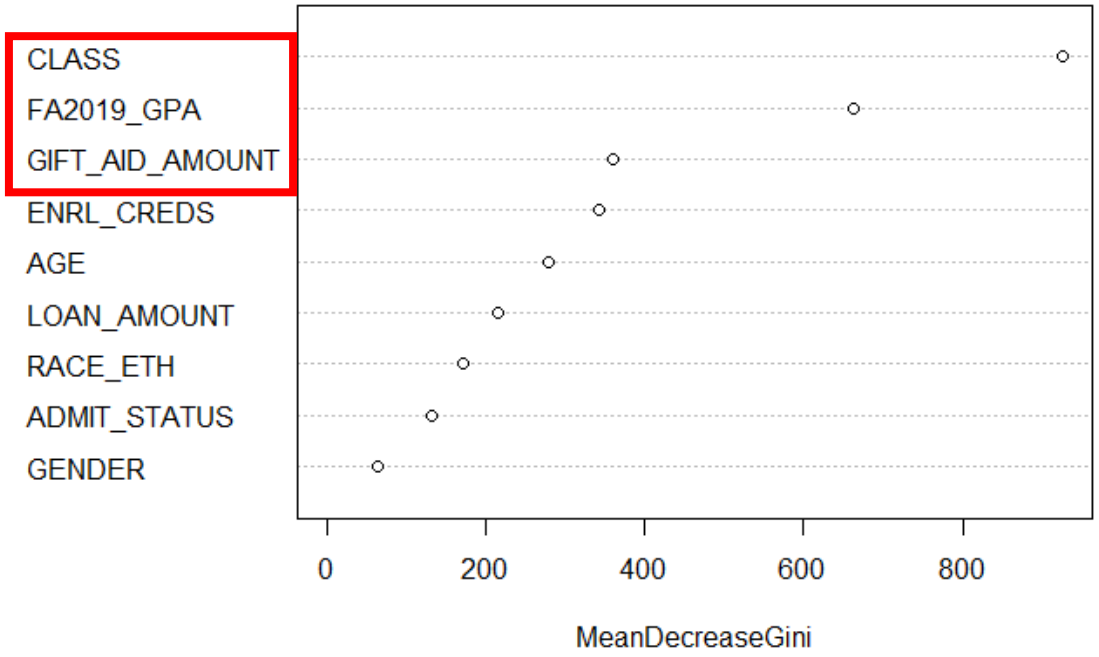
Random Forest

Decision Tree Example

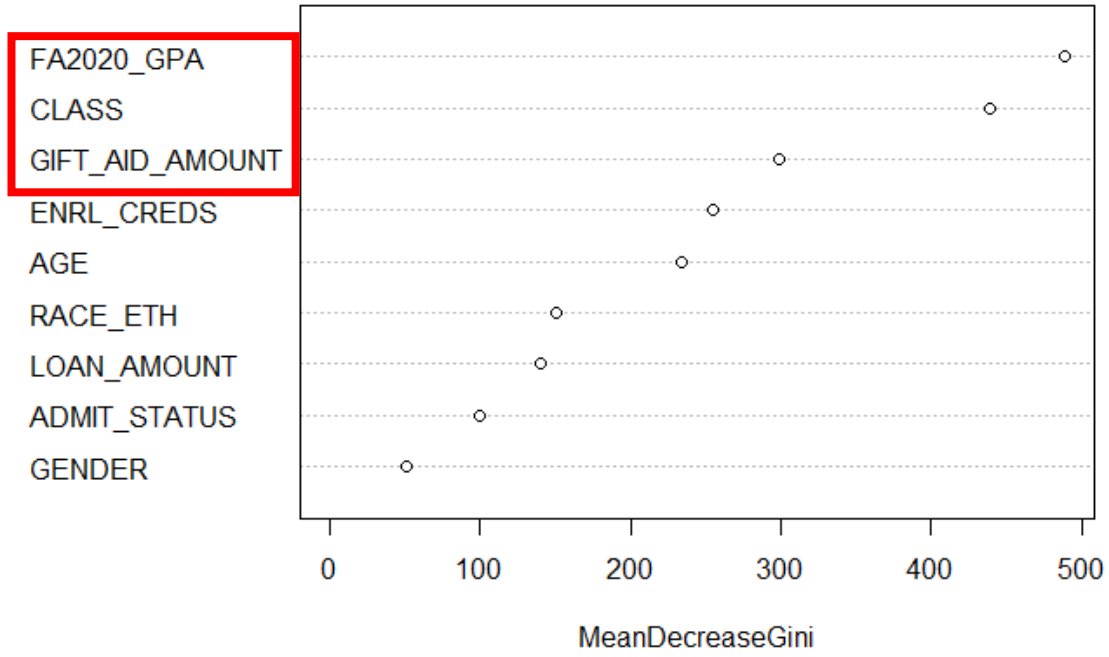


Random Forest Variable Importance Plot

2019 Variable Importance Plot



2020 Variable Importance Plot



Confusion Matrix on Test Set

Logistic Regression

Fall 2019

	0	1	% Correct
0	1450	374	79.5%
1	270	722	72.8%
Overall %	61.1%	38.9%	**77.1%**

Fall 2020

	0	1	% Correct
0	1051	282	78.8%
1	282	366	56.5%
Overall %	67.3%	32.7%	**71.5%**

Random Forest

Fall 2019

	0	1	% Correct
0	1279	306	80.7%
1	268	962	78.2%
Overall %	55.0%	45.0%	**79.6%**

Fall 2020

	0	1	% Correct
0	1014	140	87.9%
1	328	498	60.7%
Overall %	67.8%	32.2%	**76.4%**

Accuracy Rate Calculation:

Testing Accuracy Rate

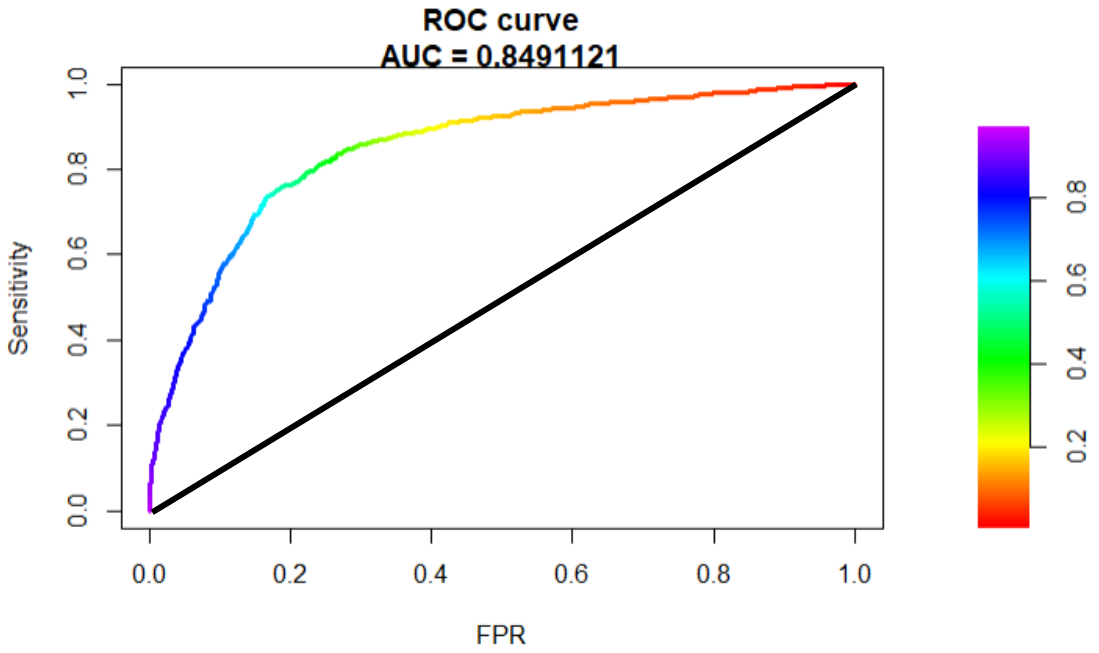
$$= \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

$$= \frac{(498+1014)}{(498+140+1014+328)}$$

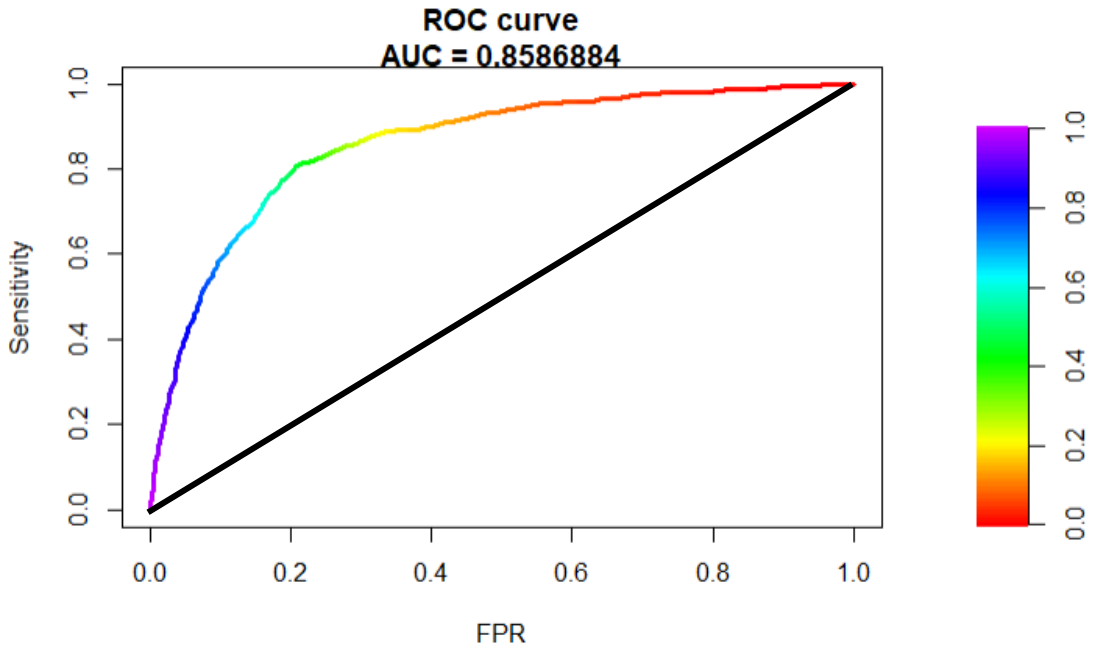
$$= \mathbf{76.4\%}$$

Fall 2019 ROC CURVES

Logistic Regression

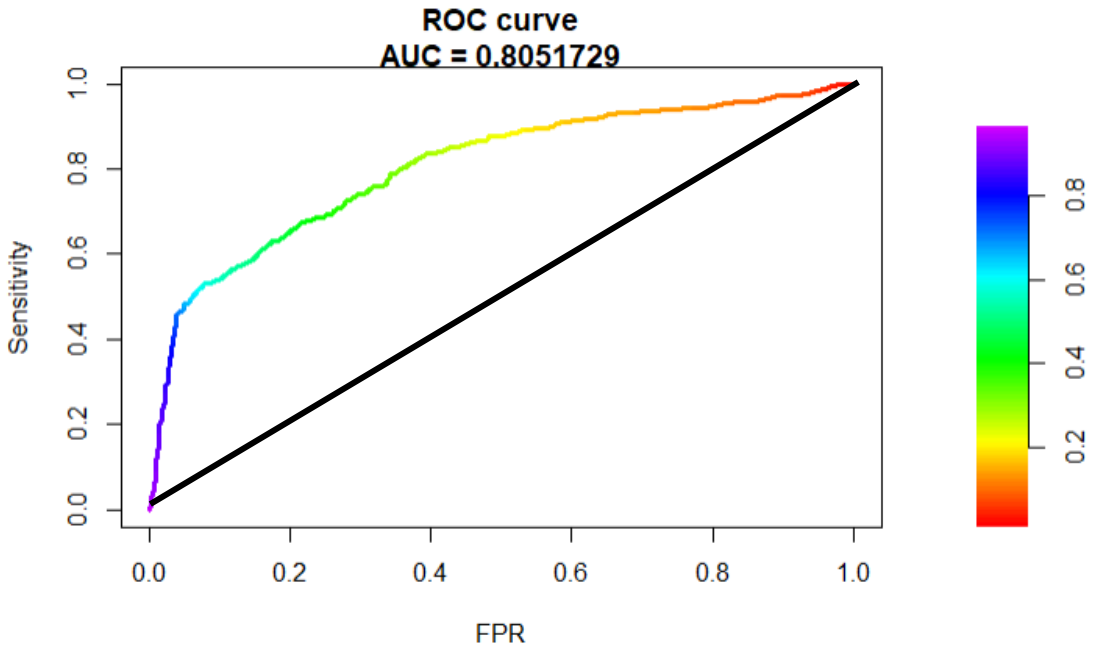


Random Forest

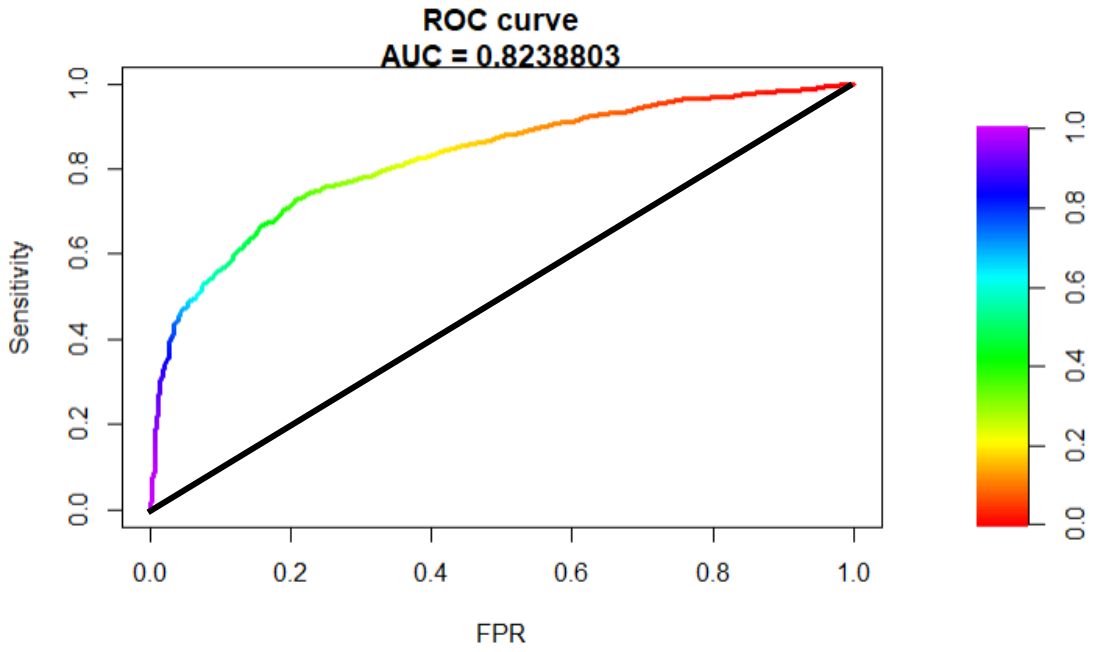


Fall 2020 ROC CURVES

Logistic Regression



Random Forest



Conclusion and Discussion

Conclusion

- We use logistic regression and random forest models to ascertain the main predictors as well as predict student retention rate.
- According to logistic regression, the significant factors in both cohorts were:
  - ✓ Sophomore classification (CLASS_SO),
  - ✓ Number of first term enrolled credits (ENRL_CREDS)
  - ✓ Financial gift aid received (GIFT_AID_AMOUNT)
  - ✓ First Term GPA (FA2019_GPA and FA2020_GPA)
- Two key factors found in one model but not found in the other, according to logistic regression, were:
  - ✓ Unknown classification (Odds were 1709% times higher for a student classified as unknown vs. a freshman, in Fall 2019 model).
  - ✓ Age (Odds were decreased by 2.7% for every additional increase in age, in Fall 2020 model),

Conclusion Cont.

- According to random forest, the top three important variables in predicting student retention for the Fall 2019 cohort were:
  1. Classification (CLASS),
  2. Fall 2019 GPA (FA2019_GPA)
  3. Financial gift aid received (GIFT_AID_AMOUNT)
- The top three important variables in predicting student retention for the Fall 2020 cohort were:
  1. Fall 2020 GPA (FA2020_GPA)
  2. Classification (CLASS),
  3. Financial gift aid received (GIFT_AID_AMOUNT)

Discussion

- Random Forest is more accurate at predicting student retention and had a higher AUC for both cohorts.
- Logistic Regression is better at interpreting the results.
- We can use the information gathered to discuss with other departments (e.g., student success coaches) to see how to better our study and help our students.
- For future studies on student retention prediction, we can use other machine learning techniques such as K-Nearest Neighbors and Neural Networks to compare accuracy and interpretability.

Contact Info

- Feel free to send your question/suggestion/ discussion to:

Jeremy Monteath-Valdez

Jeremy.Valdez@DallasCollege.edu

Daniel Le

dle@DallasCollege.edu

References

- *Logit Regression | SAS Data Analysis Examples*. stats.oarc.ucla.edu/sas/dae/logit-regression. Accessed 4 Apr. 2022.
- Hastie, Trevor, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer, 2009.
- **Logistic Regression**

  https://www.youtube.com/watch?v=1-0zMWp5w8U
- **Visual Guide to Random Forests**

  https://www.youtube.com/watch?v=cIbj0WuK41w
- Dr. Yun, Jonghyun, Data mining lecture notes, UTA Math Department, 2019.

THANK YOU FOR LISTENING!