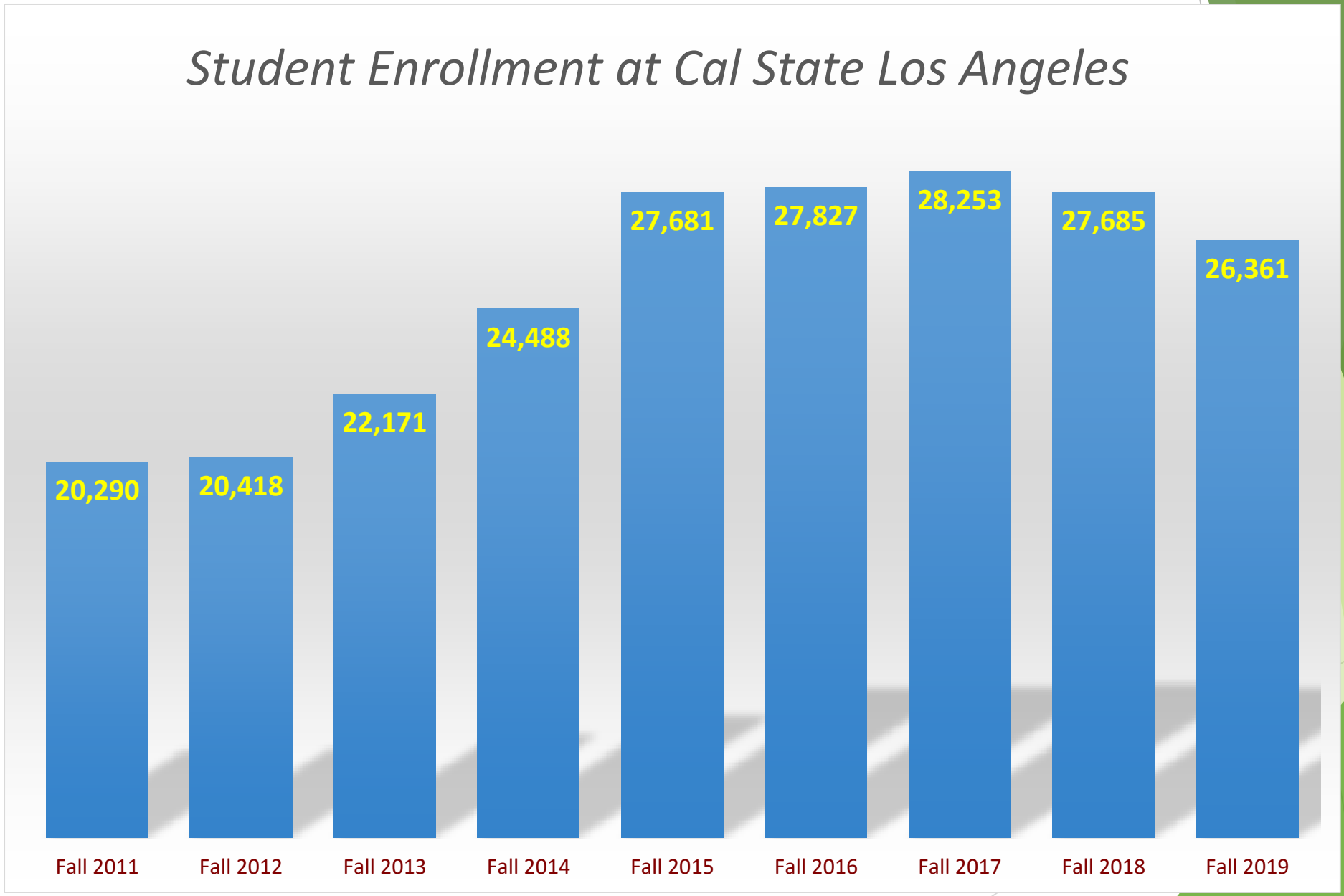
Enrollment Modeling with Machine Learning Algorithms

Su Seon Yang & Sunny Moon

Institutional Effectiveness

California State University, Los Angeles


Student Enrollment at Cal State Los Angeles
20,290
20,418
22,171
24,488
27,681
27,827
28,253
27,685
26,361
Fall 2011
Fall 2012
Fall 2013
Fall 2014
Fall 2015
Fall 2016
Fall 2017
Fall 2018
Fall 2019

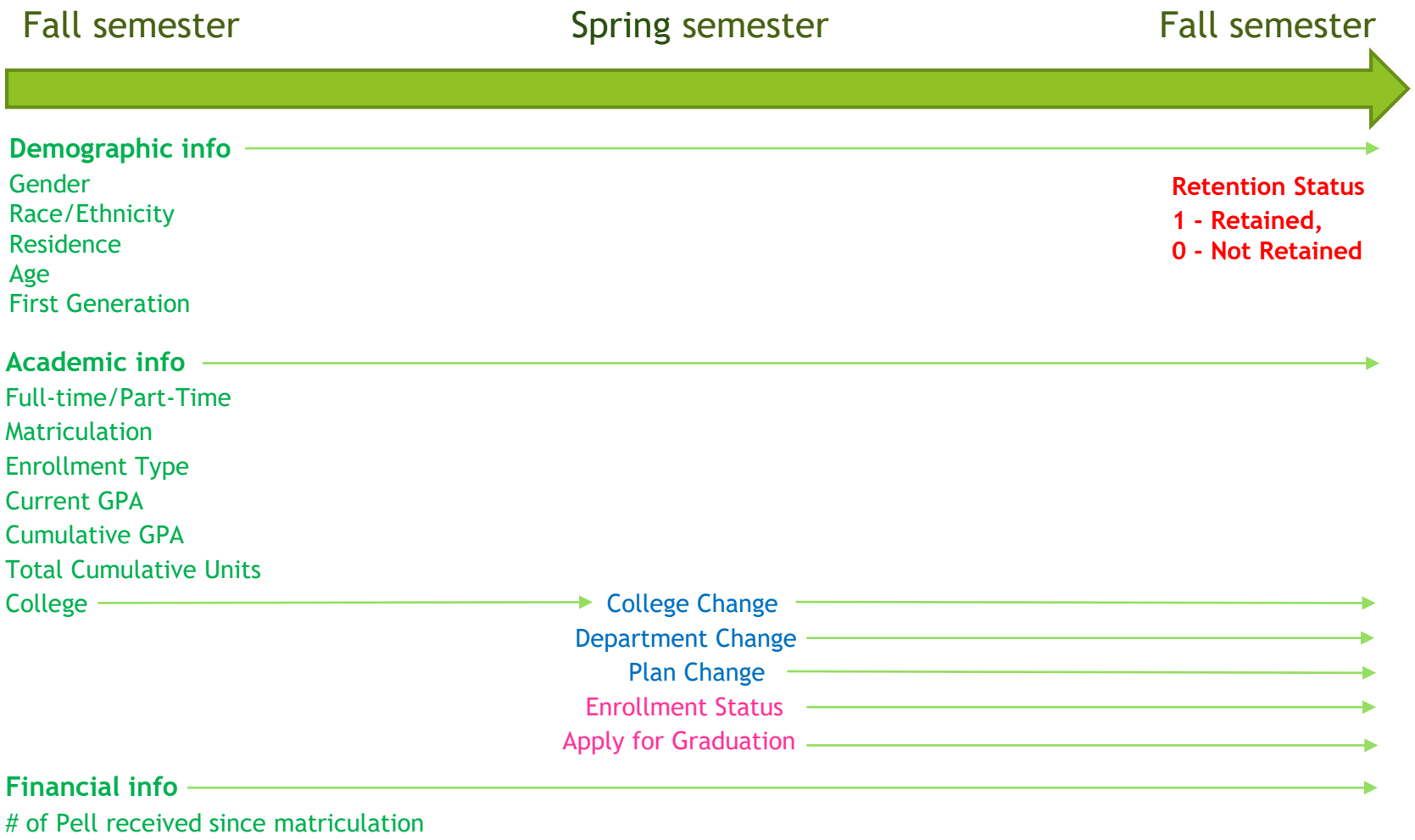
Prediction	
New	FTF
	Transfers
	PB
	Graduate
Returning	UG
	PB
	Graduate
Transitory	UG
	PB
	Graduate
Continuing	UG
	PB
	Graduate

2) New Student Model

1) Continuing Student Model

1) Continuing Student Enrollment Predictive Model

1-1. Design of Continuing Student Enrollment Modeling



1-2. Steps of Continuing Student Enrollment Modeling

- Data used: Fall 2017 students (28,253) and their Spring 2018 information
- Dependent variable: Retention status (1: Yes, 0: No) at Fall 2018
- Data preprocessing:
  - Dummy variable creation for categorical variables
  - Missing data imputation using MICE – 41 Matriculation info are replaced
  - Feature scaling using Min-Max Scalar
  - Oversampling using SMOTE (1: 65.8% / 0: 34.2%)
- Feature (Independent Variable) Selection
  - Univariate selection
  - Recursive feature elimination
  - Boruta
  - In-built feature importance (using Tree-based)
- Predictive Model Development
  - Logistic Regression
  - XGBoost
  - Random Forest
  - Neural Network
- Model Evaluation: Receiver Operating Characteristic (ROC) Curve

1-3. Feature (Independent Variable) Selection

1. Univariate selection

Variable	Score
APPLY_GRADUATION_YES	5558.937264
ENR_SP_NOT-ENROLLED	2751.930063
ENR_TYPE_New transfer	571.435326
FTPT_PART-TIME	316.071119
ENR_TYPE_First-time freshman	226.951868
ENR_TYPE_Continuing UG	138.038118
ENR_TYPE_New GRAD	111.795063
TOT_CUMULATIVE	111.436214
C_PELL	102.566199
RACE_ETH_WHITE	56.696445
ENR_TYPE_Transitory UG	44.406907
AGE	30.244531
COLL_ED	23.383476
RACE_ETH_HISP	19.403433
FIRST_GEN_Unknown	17.647793
COLL_ET	16.093830
CUR_GPA	15.327322
MAJ_CHANGE_NO CHANGE	11.456142
FIRST_GEN_Parent Graduated College	10.389122
DEPT_CHANGE_NO CHANGE	9.070429
ENR_TYPE_Continuing PB	8.672398
COLL_HHS	8.172547
RACE_ETH_BLACK	8.051507
COLL_UN	7.880994
COLL_BE	6.584027
CUM_GPA	5.748966
RACE_ETH_UNK	4.307901
ENR_TYPE_Returning PB	3.456677
RACE_ETH_INTERNATIONAL	2.716553
COLL_CHANGE_NO CHANGE	2.436400
RACE_ETH_PACIF	1.522056
COLL_NSS	0.860046
ENR_TYPE_Returning UG	0.821817
ENR_TYPE_New PB	0.787612
RESIDENCE Resident	0.460874

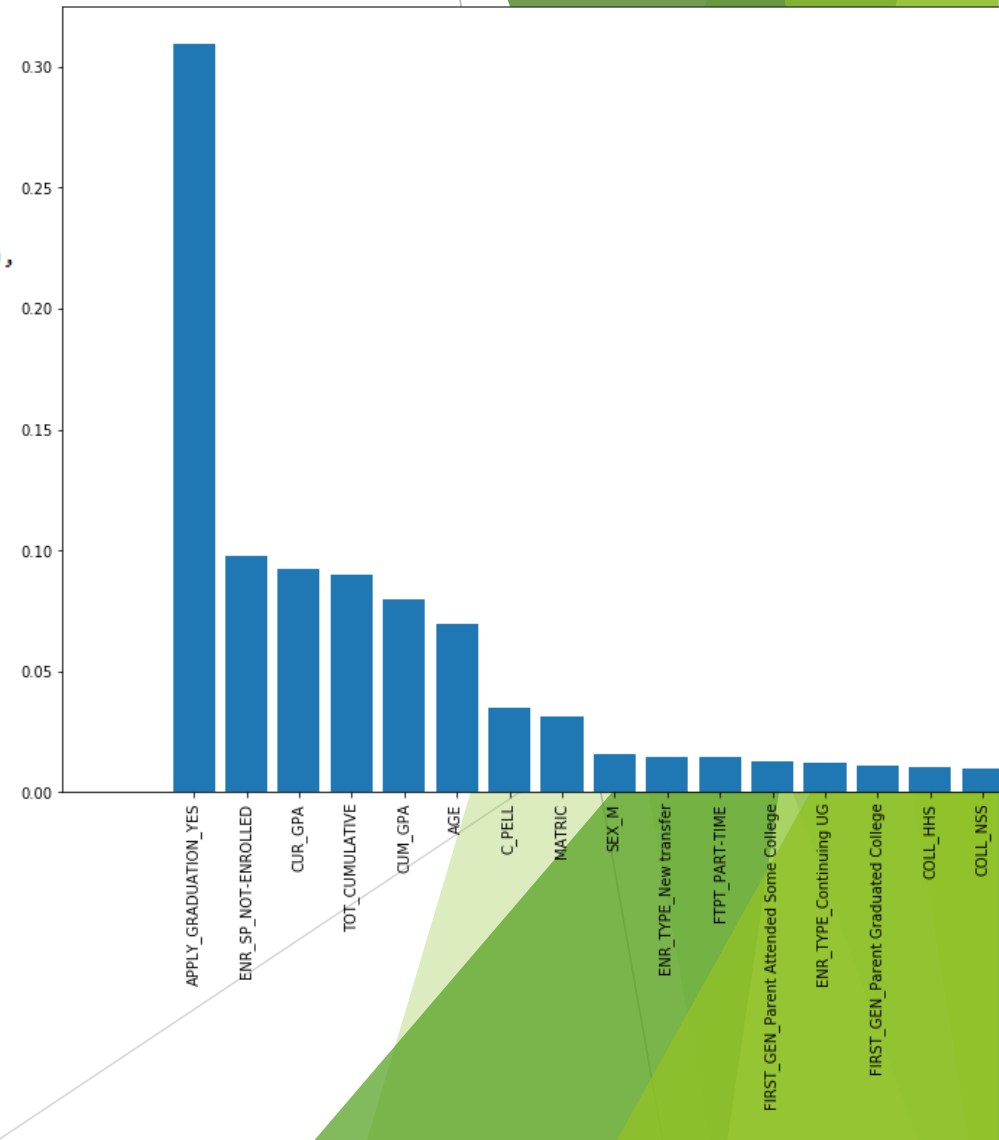
2. RFE

```
(1, 'APPLY_GRADUATION_YES'),
(1, 'CUM_GPA'),
(1, 'CUR_GPA'),
(1, 'ENR_SP_NOT-ENROLLED'),
(1, 'ENR_TYPE_Continuing UG'),
(1, 'ENR_TYPE_New transfer'),
(1, 'ENR_TYPE_Returning UG'),
(1, 'ENR_TYPE_Transitory UG'),
(1, 'TOT_CUMULATIVE'),
(2, 'ENR_TYPE_First-time freshman'),
(3, 'ENR_TYPE_New GRAD'),
(4, 'ENR_TYPE_Returning GRAD'),
(5, 'MAJ_CHANGE_NO CHANGE'),
(6, 'COLL_CHANGE_NO CHANGE'),
(7, 'ENR_TYPE_Continuing PB'),
(8, 'RACE_ETH_PACIF'),
(9, 'COLL_ED'),
(10, 'C_PELL'),
(11, 'DEPT_CHANGE_NO CHANGE'),
(12, 'COLL_UN'),
(13, 'COLL_BE'),
(14, 'ENR_TYPE_New PB'),
(15, 'AGE'),
(16, 'FTPT_PART-TIME'),
(17, 'ENR_TYPE_Returning PB'),
(18, 'RACE_ETH_ASIAN'),
(19, 'RACE_ETH_TWO RACES'),
(20, 'RACE_ETH_HISP'),
(21, 'RACE_ETH_UNK'),
(22, 'RACE_ETH_BLACK'),
(23, 'RACE_ETH_INTERNATIONAL'),
(24, 'COLL_HHS'),
(25, 'MATRIC'),
(26, 'COLL_ET'),
(27, 'SEX_M'),
```

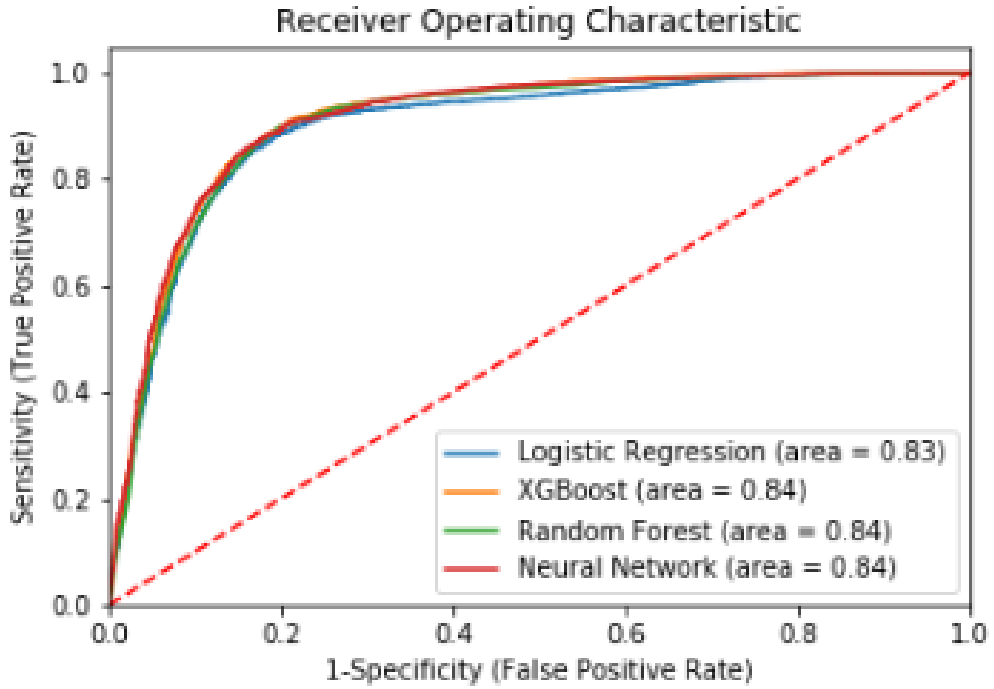
3. Boruta

```
(1, 'AGE'),
(1, 'APPLY_GRADUATION_YES'),
(1, 'CUM_GPA'),
(1, 'CUR_GPA'),
(1, 'C_PELL'),
(1, 'DEPT_CHANGE_NO CHANGE'),
(1, 'ENR_SP_NOT-ENROLLED'),
(1, 'ENR_TYPE_Continuing UG'),
(1, 'ENR_TYPE_First-time freshman'),
(1, 'ENR_TYPE_New GRAD'),
(1, 'ENR_TYPE_New transfer'),
(1, 'FTPT_PART-TIME'),
(1, 'MAJ_CHANGE_NO CHANGE'),
(1, 'MATRIC'),
(1, 'TOT_CUMULATIVE'),
```

4. In-built feature selection



1-4. Model Evaluation for Fall 2018 Prediction



W/ Standardization,
Feature selection by Boruta (k = 15)

1-5. Fall 2019 Prediction Result

- Note that Matriculation plays an important role in this prediction model.
- Out of 27,685 Fall 2018 FTF, 62 students have missing Matriculation. Thus, they are excluded in this prediction.

Metric (Y=1)	Logistic Regression	XG Boost	Random Forest	Neural Network
Precision	0.85	0.86	0.85	0.85
Recall	0.96	0.78	0.92	0.97
F1	0.90	0.82	0.88	0.90
FP rate	0.30	0.23	0.30	0.31

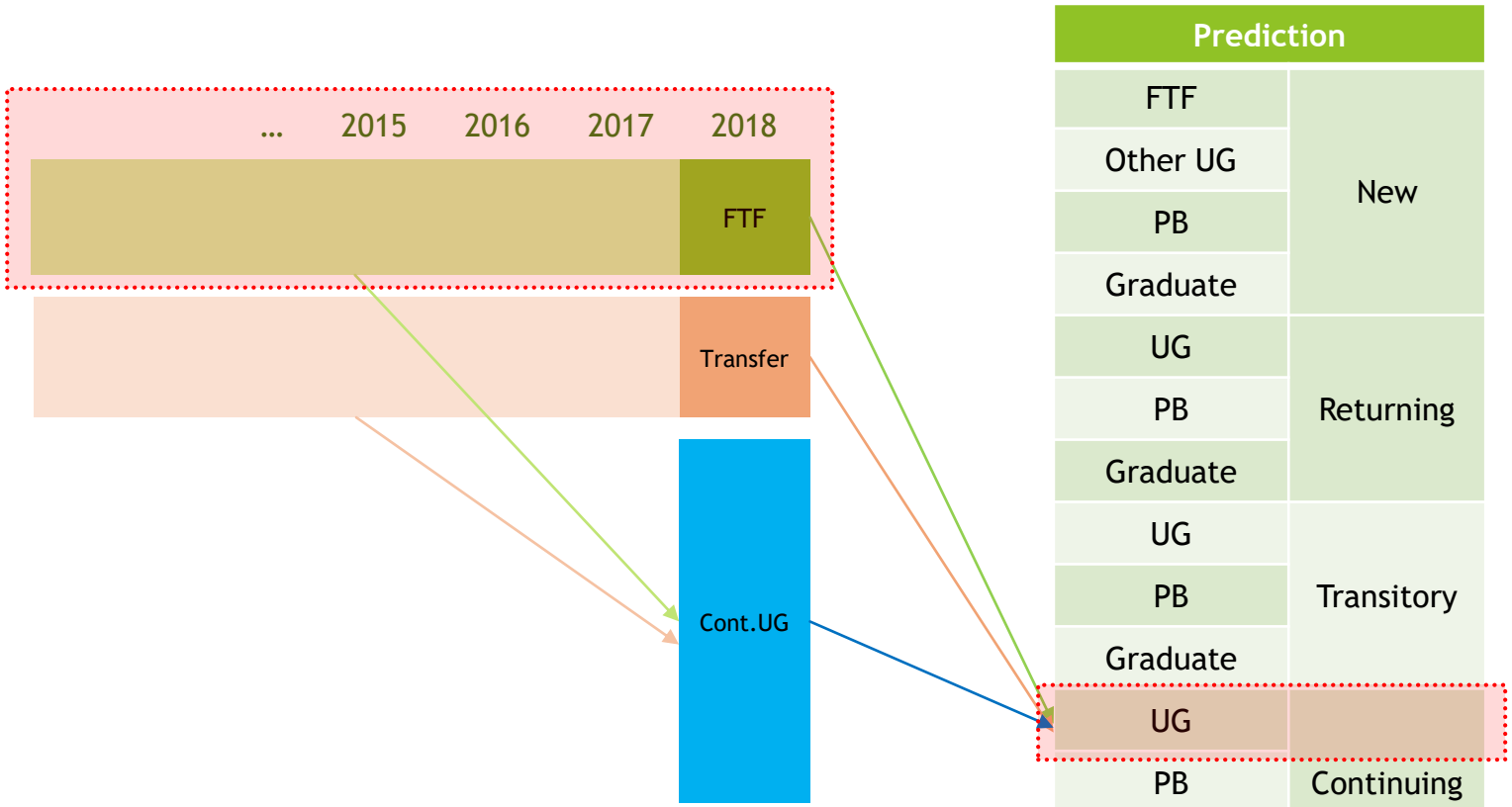
Student Level	Fall 19 Census Data	Logistic Regression	XG Boost	Random Forest	Neural Network
UG	16,069	17,839	15,284	17,204	17,895
PB	458	596	288	552	649
Graduate	1,858	1,571	583	1,495	1,642
Total	18,385	20,006 (108.8%)	16,155 (87.9%)	19,251 (104.7%)	20,186 (109.8%)

1-6. Limitations

- Student groups in continuing-type enrollment model are too broad.
  - It is very hard to determine independent variables, which play an important role over all student groups.

- Next Step
  - Separate student groups into sub-groups: FTF, Transfer, PB and Graduate
  - Add independent variables for each sub-group (ex. FTF)
    - Pre-College: SAT, High School GPA
    - Academic: Unit-load (per 1year), GPA trend, etc.

What if we focus on FTF in Continuing Student Enrollment Model?



Prediction	
FTF	New
Other UG	
PB	
Graduate	
UG	Returning
PB	
Graduate	
UG	Transitory
PB	
Graduate	
UG	
PB	Continuing
Graduate	

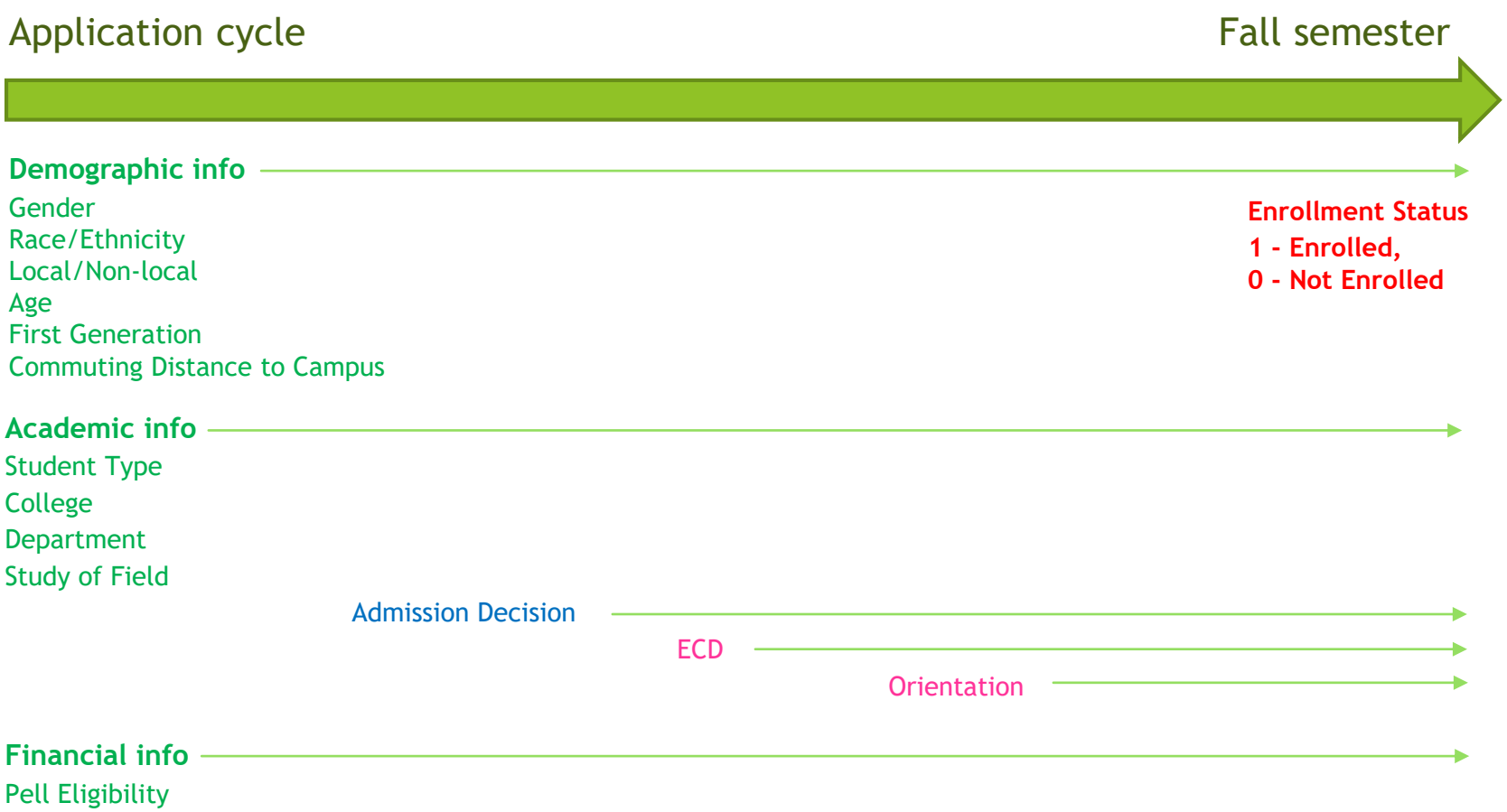
Fall 19 Prediction Result (FTF focus)

- SAT score plays an important role in this prediction model.
- Out of 3,862 Fall 2018 FTF, 6 students have missing SAT score. Thus, they are excluded in this prediction.

Retention Status	XG Boost	Random Forest	Actual Data
1 (Yes)	2,834	3,040 (98.6%)	3,084
0 (No)	1,022	816	772

2) New Student Enrollment Predictive Model

2-1. Design of New Student Enrollment Modeling



2-2. Steps of New Student Enrollment Modeling

- Data used: Fall 2018 Application data (*n* = 67,256)
- Dependent variable: Enrollment status (1: Yes, 0: No) at Fall 2018
- Goal is to *predict as many enrolled students as possible (high sensitivity) while to reduce false-positive rate*.
- Data preprocessing:
  - Dummy variable creation for categorical variables
  - Missing data imputation using MICE
  - Feature scaling using Min-Max Scalar
  - Oversampling using SMOTE (1: 13% / 0: 87%)
- Feature (Independent Variable) Selection
  - Univariate selection
  - Boruta
  - In-built feature importance (using Tree-based)
- Predictive Model Development
  - Logistic Regression
  - XGBoost
  - Random Forest
  - Neural Network
- Model Evaluation
  - Receiver Operating Characteristic (ROC) Curve
  - Confusion Matrix

2-3. Feature (Independent Variable) Selection

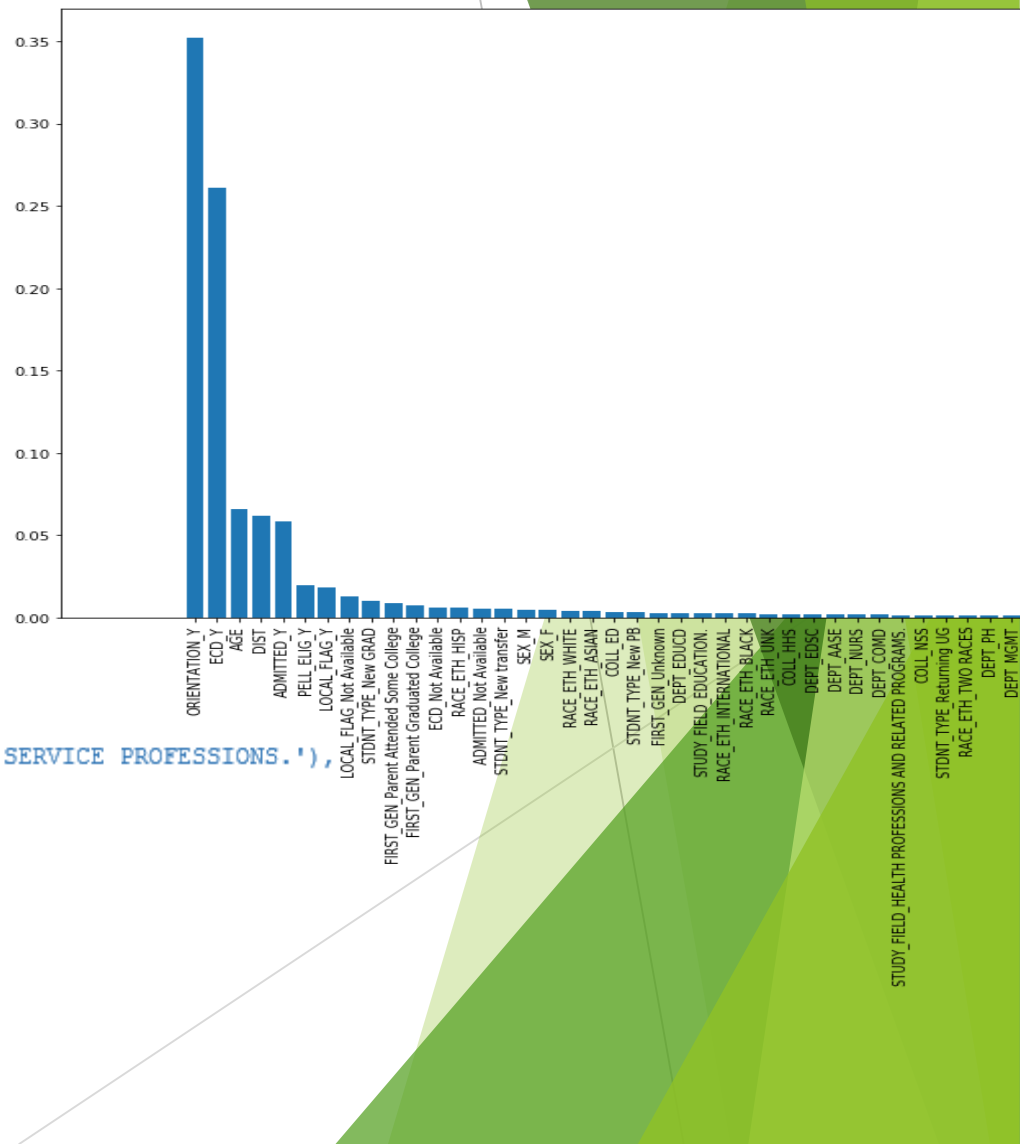
1. Univariate selection

Variable	Score
ORIENTATION_Y	26586.499933
ECD_Y	22110.252919
ADMITTED_Y	3198.119305
LOCAL_FLAG_Y	1797.766336
PELL_ELIG_Y	1573.632203
ADMITTED_Not Available	640.306725
ECD_Not Available	640.306725
LOCAL_FLAG_Not Available	505.903662
COLL_ED	451.014388
STUDY_FIELD_EDUCATION.	428.490543
DEPT_EDUCD	407.006231
STDNT_TYPE_New GRAD	278.500345
DEPT_AASE	194.553631
STDNT_TYPE_New PB	181.214030
STDNT_TYPE_Returning UG	157.291724
FIRST_GEN_Parent Graduated College	83.345275
DEPT_COMD	71.817559
DEPT_EDCI	71.326500
DEPT_CFS	70.473077
RACE_ETH_HISP	67.036327
STDNT_TYPE_New transfer	66.823601
FIRST_GEN_Unknown	61.824399
DIST	59.128706
STUDY_FIELD_PUBLIC ADMINISTRATION AND SOCIAL S...	57.888482
RACE_ETH_BLACK	57.859243
RACE_ETH_WHITE	56.935698
STUDY_FIELD_PSYCHOLOGY.	52.098972
DEPT_PSY	51.761239
DEPT_EDD	49.031463
COLL_NSS	46.596511

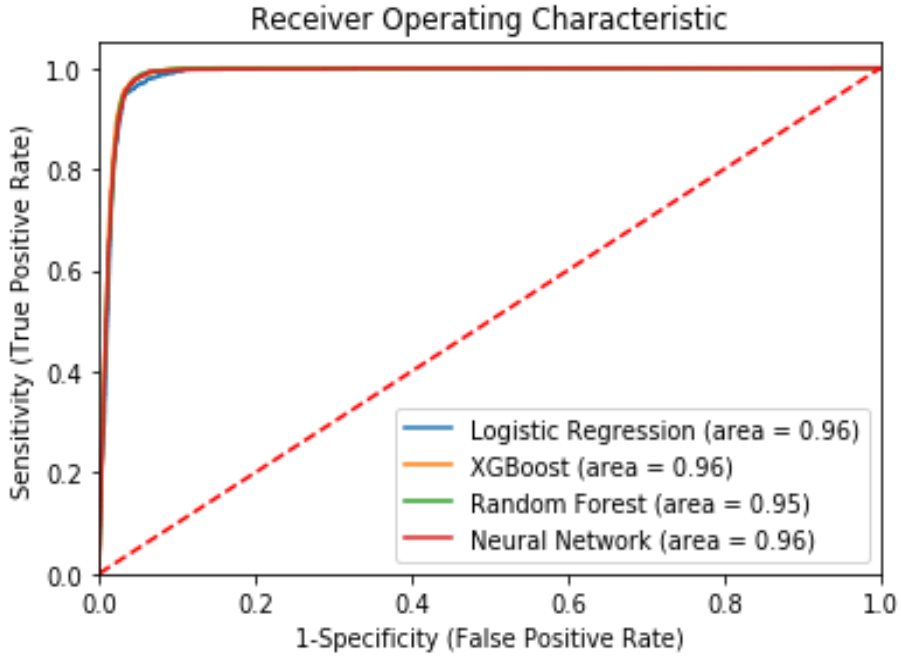
2. Boruta

```
(1, 'ADMITTED_Not Available'),
(1, 'ADMITTED_Y'),
(1, 'AGE'),
(1, 'COLL_ED'),
(1, 'COLL_HHS'),
(1, 'COLL_NSS'),
(1, 'DEPT_AASE'),
(1, 'DEPT_COMD'),
(1, 'DEPT_EDSC'),
(1, 'DEPT_EDUCD'),
(1, 'DIST'),
(1, 'ECD_Not Available'),
(1, 'ECD_Y'),
(1, 'FIRST_GEN_Parent Graduated College'),
(1, 'LOCAL_FLAG_Not Available'),
(1, 'LOCAL_FLAG_Y'),
(1, 'ORIENTATION_Y'),
(1, 'PELL_ELIG_Y'),
(1, 'RACE_ETH_BLACK'),
(1, 'RACE_ETH_HISP'),
(1, 'RACE_ETH_WHITE'),
(1, 'STDNT_TYPE_New GRAD'),
(1, 'STDNT_TYPE_New PB'),
(1, 'STDNT_TYPE_New transfer'),
(1, 'STDNT_TYPE_Returning UG'),
(1, 'STUDY_FIELD_EDUCATION.'),
(1, 'STUDY_FIELD_PUBLIC ADMINISTRATION AND SOCIAL SERVICE PROFESSIONS.'),
```

3. In-built feature selection



2-4. Model Evaluation for Fall 2018 Prediction



W/ Standardization,
Oversampling,
Feature selection by Boruta (k = 27)

Metric (Y=1)	Logistic Regression	XG Boost	Random Forest	Neural Network
Precision	0.79	0.81	0.82	0.81
Recall	*0.96*	*0.96*	*0.93*	*0.95*
F1	0.86	0.88	0.87	0.88
FP rate	*0.04*	*0.03*	*0.03*	*0.03*

2-5. Fall 2019 Prediction Result

Metric (Y=1)	Logistic Regression	XG Boost	Random Forest	Neural Network
Precision	0.86	0.86	0.87	0.86
Recall	*0.87*	*0.88*	*0.87*	*0.88*
F1	0.87	0.87	0.87	0.87
FP rate	*0.02*	*0.02*	*0.02*	*0.02*

Enrollment Status	Student Level	Fall 19 Census Data	XG Boost	Neural Network
New	FTF	2,480	2,794	2,762
	Transfer	1,734	1,948	1,969
	PB	197	116	106
	Graduate	413	111	89
Returning	UG	157	172	171
	PB	10	17	11
	Graduate	58	20	16
Transitory	UG	10	19	22
	PB	2	2	2
	Graduate	0	0	0
Total		5,061	5,199 (102.7%)	5,148 (101.7%)

Comparison and Future Steps

Enrollment Model using Machine Learning Algorithm

- Separate student groups into sub-groups: FTF, Transfer, PB and Graduate
- Add independent variables for each sub-group (ex. FTF)
  - Pre-College: SAT, High School GPA
  - Academic: Unit-load (per 1year), GPA trend, etc.

Traditional Model #1

- Aggregate Model
  - Based on trend of previous year
  - Matriculation Type
  - Currently used

Traditional Model #2

- Aggregate Model: Matriculation Decay
  - Based on trend of previous year
  - Matriculation Type
  - Matriculation Term