Predicting Student Starts

Mike Kirkpatrick

May 17, 2018

1. Introduction

This report serves as a broad overview of the analyses. Ommissions are noted where appropriate and further details can be incorporated upon request.

Within the student lifecycle (see figure 1), potential students are first exposed to marketing and recruiting efforts. Then they are admitted to the University. A student is considered to have "started" once they take their first class. **The goal of this analysis was to predict which student would start at National University based on only information available within 7 days of matriculation**. The input data were therefore limited to marketing data, data obained from the students' application and past ineractions with NU. Student transfer data (e.g., external GPA, tranfer units, etc.) were not included because students have up to 90 days to submit this information. That is, transfer data is not available within 7 days of admission so such data could not be used as inputs in a production model.

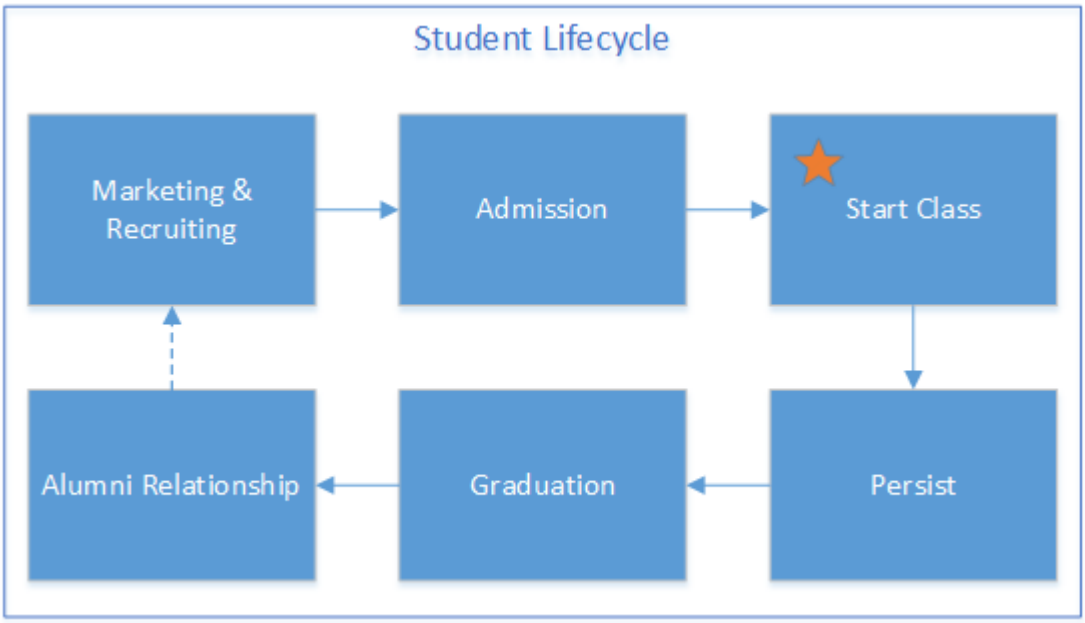


Figure 1. Student Lifecycle

2. Methods

2.1 The Data

A total of 30 variables were included. These variables are listed in the table below. Descriptive statistics for each variable can be found in Appendix A The outcome variables were eventStartYN and daysToEvent. The 28 input variables consisted of only data available within 7 days of student matriculation. A total of 42485 students were included. These students had matriculation dates between 2015-01-01 and 2017-05-20.

Variable	Description

Variable	Description
eventStartYN	Whether or not (Y/N) the student started or did not start
daysToEvent	number of days between matriculation and event (i.e., Start or Non-start)
ACAD_CAREER	Academic career level of student
ACAD_PLAN_TYPE	Academic plan type of student
ADM_APPL_METHOD	Application method of student
admitFyQtr	Fiscal Year Quarter of student matriculation date (Q1 = July-Sept)
age	Student age at matriculation
appFeeStatus	Application fee status of student at matriculation
campus	Campus that received the student application
cancelSum	Total number of previous program cancelations at NU by student
completionSum	Total number of previous program completions at NU by student
coursesScheduled	Total courses scheduled within 7 days of student matriculation
daysToFirstSchedClass	Number of days between student matriculaiton date and first scheduled class (default=380)
discSum	Total number of previous program discontinuations at NU by student
ENTITY_GROUP	Entity Group of matriculated student
ETHNICITY	Student reported ethnicity
FIN_AID_INTEREST	Whether or not (Y/N) the student indicated interest in finalcial aid on application
flagBLK	Whether or not (Y/N) the student had any enrollment blocking “Business Office Lock” service indicator within 7 days of matriculation
flagBlockEnrl	Whether or not (Y/N) the student had any enrollment blocking service indicators within 7 days of matriculation
flagCRD	Whether or not (Y/N) the student had an enrollment blocking “Credential Office Lock” service indicator within 7 days of matriculation
flagHold	Whether or not (Y/N) the student had any enrollment blocking “Hold” service indicator within 7 days of matriculation
flagLock	Whether or not (Y/N) the student had any enrollment blocking “Lock” service indicator within 7 days of matriculation
flagREC	Whether or not (Y/N) the student had an enrollment blocking “Registrar’s Lock” service indicator within 7 days of matriculation
marital	Student reported marital status
military	Student reported military status (categorized)
nonCA	Whether or not (Y/N) the student indicated they resided outside of CA

Variable	Description
NU_LEAD_SOURCE_ID	Lead source of student (categorized)
nuCoursesTakenSum	Total number of previous courses taken by student at NU
sex	Student reported gender
startYNSum	Total number of previous program starts at NU by student

2.2 Algorithms

The outcome that we were trying to predict is whether or not students would start once admitted. The outcome values (i.e., “Start” and “Not Start”) are distinct classes. Therefore, classification algorithms were chosen to accomplish this task.

Four algorithms were trained and evaluated. The algorithms included were Cox Proportional-Hazards Regression, Logistic Regression, Extreme Gradient Boost (XGboost) and Random Forest (see table).

Approach	Algorithm	Description
Statistical	Cox Proportional-Hazards Regression	Survival model that estimates the probability of a binary event occuring over time
Machine Learning	Logistic Regression	Linear model that estimates the probability of a binary event occuring
	XGBoost	Tree-based ensemble model that implements gradient boosting to predict binary, multi-class and continuous outcomes
	Random Forest	Tree-based ensemble model that implements bootstrap aggregation (i.e., “bagging”) and random feature subsetting to predict to binary, multi-class and continuous outcomes

2.3 Model Training and Evaluation

The complete data set was split into two data sets: training and testing. The training data set contained 70% of the total observations (N=29740) and the testing data set contained 30% of the total observations (N=12745). The training and testing data sets are considered “unbalanced” with regard to the class variable (i.e., eventStartYN) since the proportion of students that start versus not start is not equal (starts = 65.3%, non-starts = 34.7%). A “balanced” copy of the training data set was created so that the proportion of starts was equal to non-starts. This was done by simply selecting a random sample of starts. This process is called class balancing and it often produces favorable results when the goal is to predict the minority class. Thus, three data sets were created: Unbalanced Training, Balanced Training, and Test.

All 4 algorithms were trained on each of the training data sets, thus producing a total of 8 models (see table below). For the machine learning models, 10-fold cross-validation was implemented during model training, which is not feasible for the Hazard algorithm. After model training, all 8 of the models were evaluated with the test data set. The Hazard model requres a time point for prediction. The max time value for the data set was used, 380 days, which represents the students have the total time possible to start their program.

The test data are fed through the trained models to produce the models' predictions. These predictions are then compared to the observed values (i.e., eventStartYN) in the test data. Predicted versus observed values are compared and various accuracy metrics were calculated for comparison purposes. It's important to note that all of the trained models were evaluated on the same test data set. It is also important to note that the trained models have not seen ANY of the observation in the test dataset, thus mimicking how the models would perform once put into a production environment.

Algorithm	Training Data	Model Name	Test Data
Cox Proportional Hazard Regression	Balanced	Haz Bal	Test
Cox Proportional Hazard Regression	Unbalanced	Haz Unbal	Test
Logistic Regression	Balanced	LR Bal	Test
Logistic Regression	Unbalanced	LR Unbal	Test
Random Forest	Balanced	RF Bal	Test
Random Forest	Unbalanced	RF Unbal	Test
XGBoost	Balanced	XGB Bal	Test
XGBoost	Unbalanced	XGB Unbal	Test

3. Results

3.1 Modeling Results

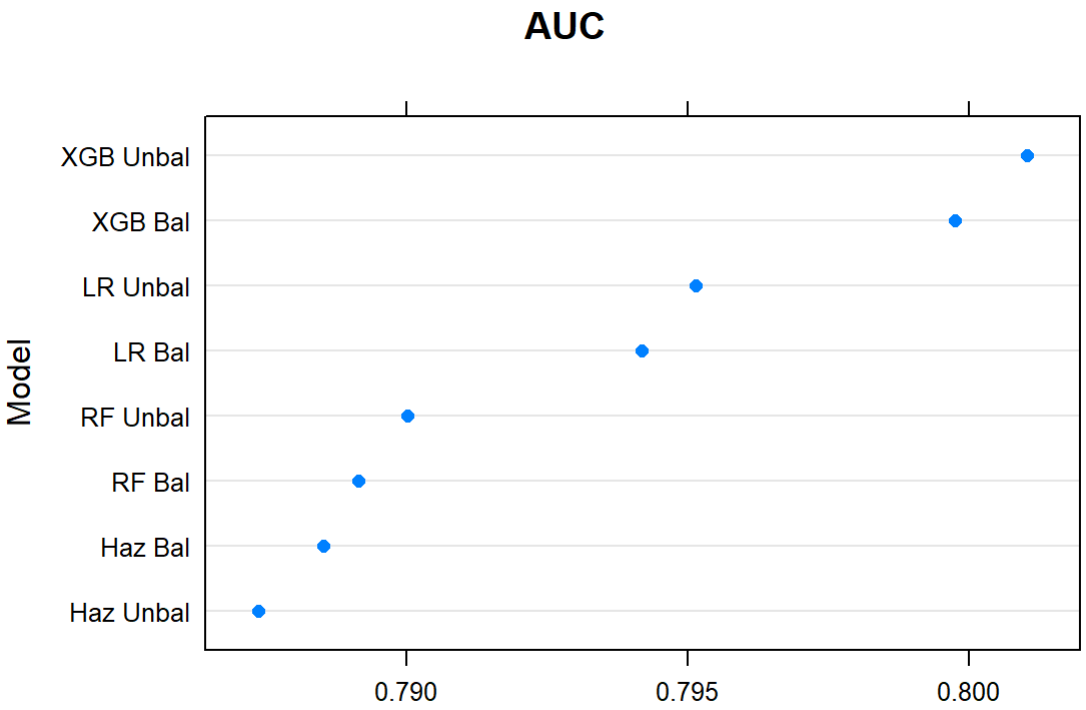
The results for each model are displayed in the table below, followed by an interpretation of the results for each evaluation measure. Confusion Matrixes for every model are in Appendix B. The The measures used for evaluating each model are:

1. AUC:
    - "Area Under the Curve" is the primary measure used to evaluate the models since it rewards the model for correct classifications and penalizes the model for incorrect classifications (this is in contrast to Accuracy, which does not penalize for incorrect predictions). This measure ranges from 0 to 1 and compares the true-positive rate against the false positive rate.
2. Accuracy.Non-Starts
    - Accuracy for predicting students that would not start is the secondary measure used to evaluate the models since the goal is to identify students that would not start (these students would require intervention). This measure ranges from 0 to 1 and is the percentage of non-start students that the model correctly classified.
3. Accuracy.Overall
    - Overall Accuracy is the tertiary measure used to evaluate the models because it does not penalize for incorrect classifications and it does not empasize the importance of predicting non-start students. This measure ranges from 0 to 1 and is the percentage of start and non-starts that the model correctly classified.
4. Average of Measures
    - Average of Measures is composite measure that, as the name suggests, is the average of the 3 preceding measures. This measure was included to simultaneously compare the performance of

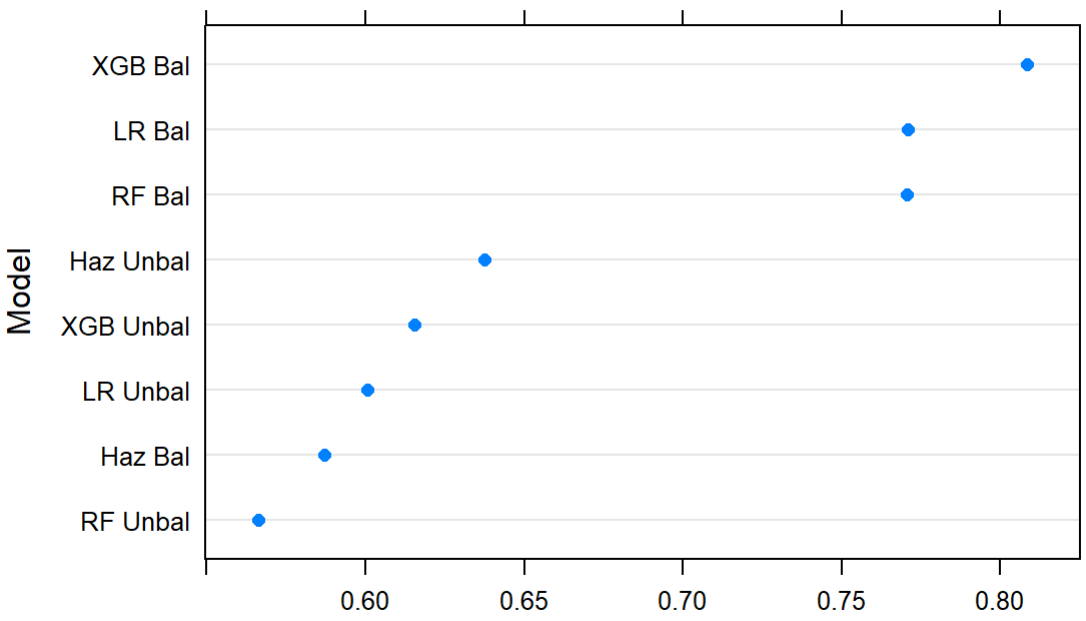
each model across all measures.

Model	AUC	Accuracy.Non.Starts	Accuracy.Overall	Average.of.Measures
Haz Bal	0.789	0.587	0.724	0.700
Haz Unbal	0.787	0.637	0.716	0.714
LR Unbal	0.795	0.601	0.732	0.709
LR Bal	0.794	0.771	0.720	0.762
RF Bal	0.789	0.770	0.715	0.758
RF Unbal	0.790	0.566	0.724	0.693
XGB Bal	0.800	0.808	0.719	0.776
XGB Unbal	0.801	0.616	0.733	0.717

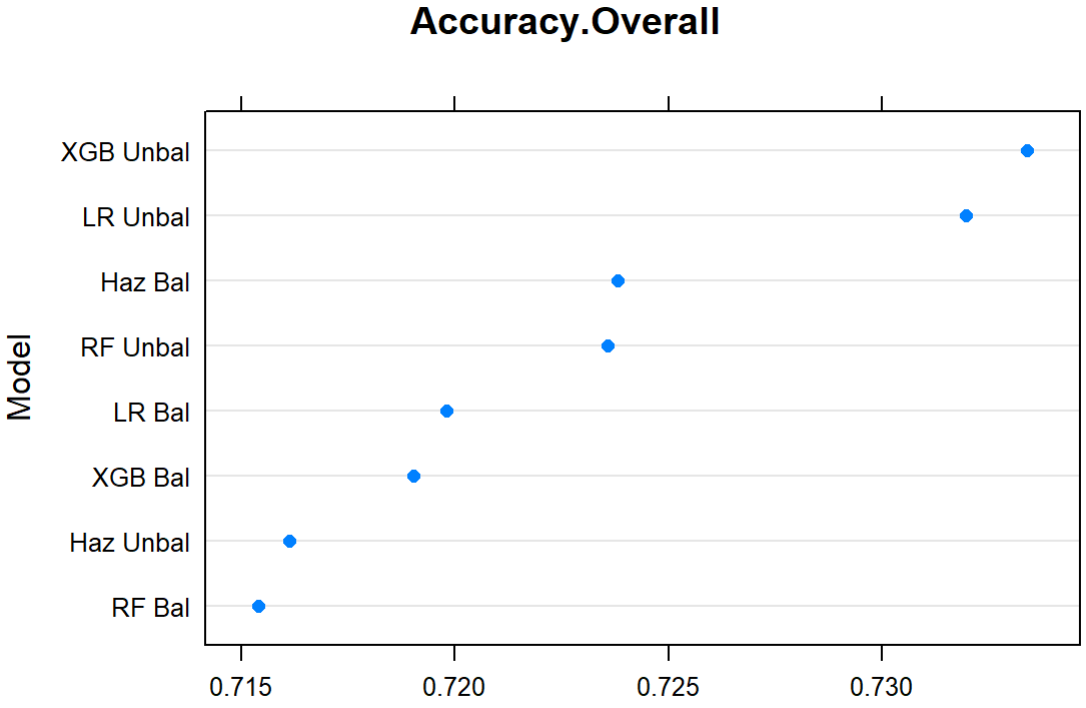
All 8 models had considerably high AUC values. XGBoost consistently had the highest AUC, followed by Logistic Regression, Random Forest and finally Cox-Proportional Hazards Regression. The AUC values were very similar across all models. In fact, all AUC values were within 1% of each other, ranging from 0.787 to 0.801. Therefore, AUC is not the best measure for distinguishes which model performed best.



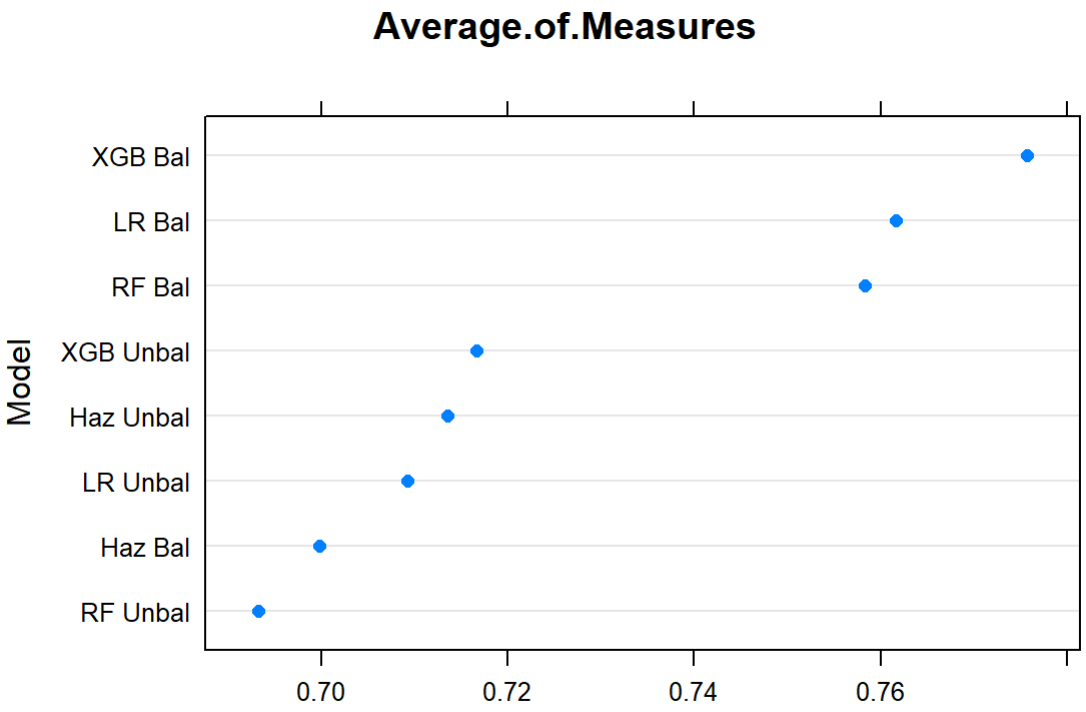
Accuracy for non-starts is the measure that differentiates model performance very well. Values ranged from 0.566 to 0.808 The machine learning algorithms that were trained on the balanced data sets performed much better than all other models. XGBoost on the balanced data set had the highest AUC (0.808) followed by Logistic Regresstion (0.771) and Random Forest (0.77). All of the unbalanced datasets did considerably worse.



All 8 models had respectively high accuracy values, ranging from 0.715 to 0.733. XGBoost and Logistic Regression trained on unbalanced data sets were the only models to exceed 73%.



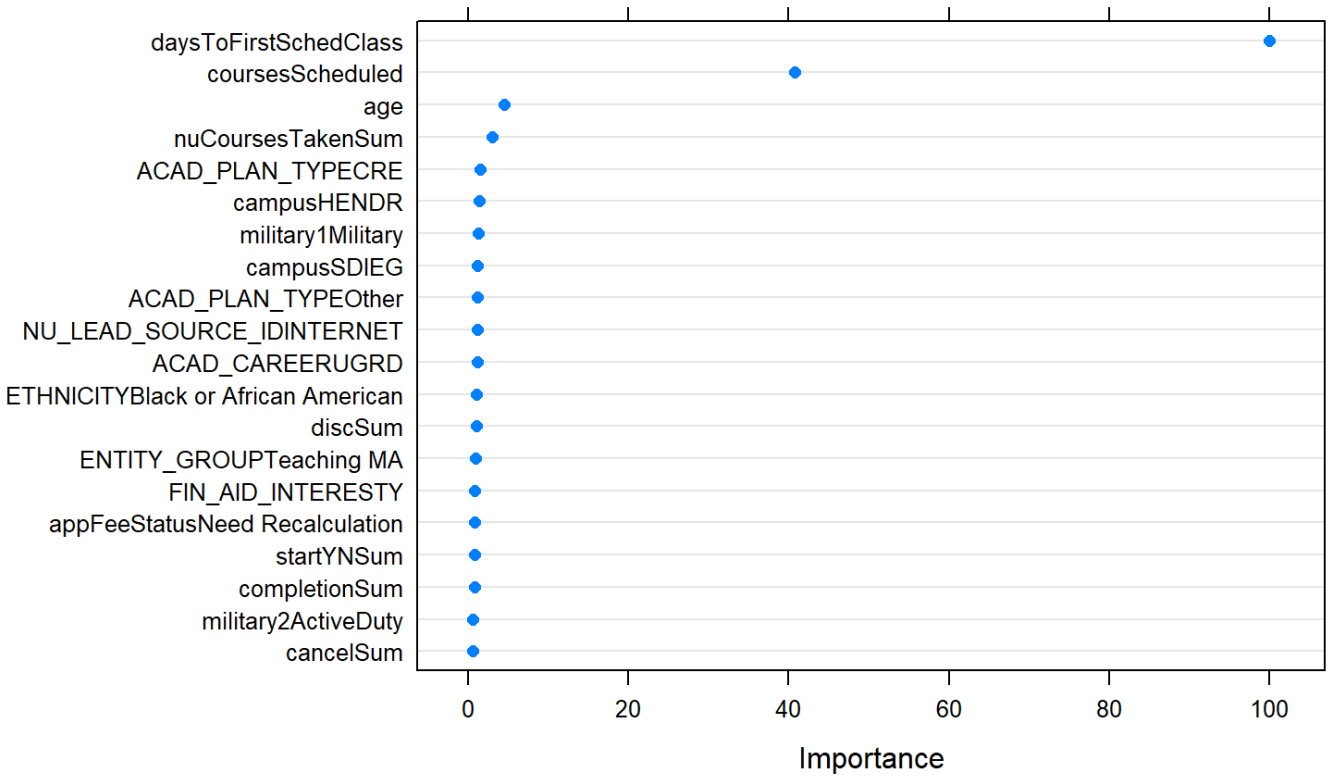
Finally, the average values of AUC, Accuracy.Non.Start and Accuracy.Overall is plotted below. Values ranged from 0.693 to 0.776 and the machine learning models that were trained on the balanced data sets performed the best. Yet again, XGBoost outperformed all other models followed by logistic regression.



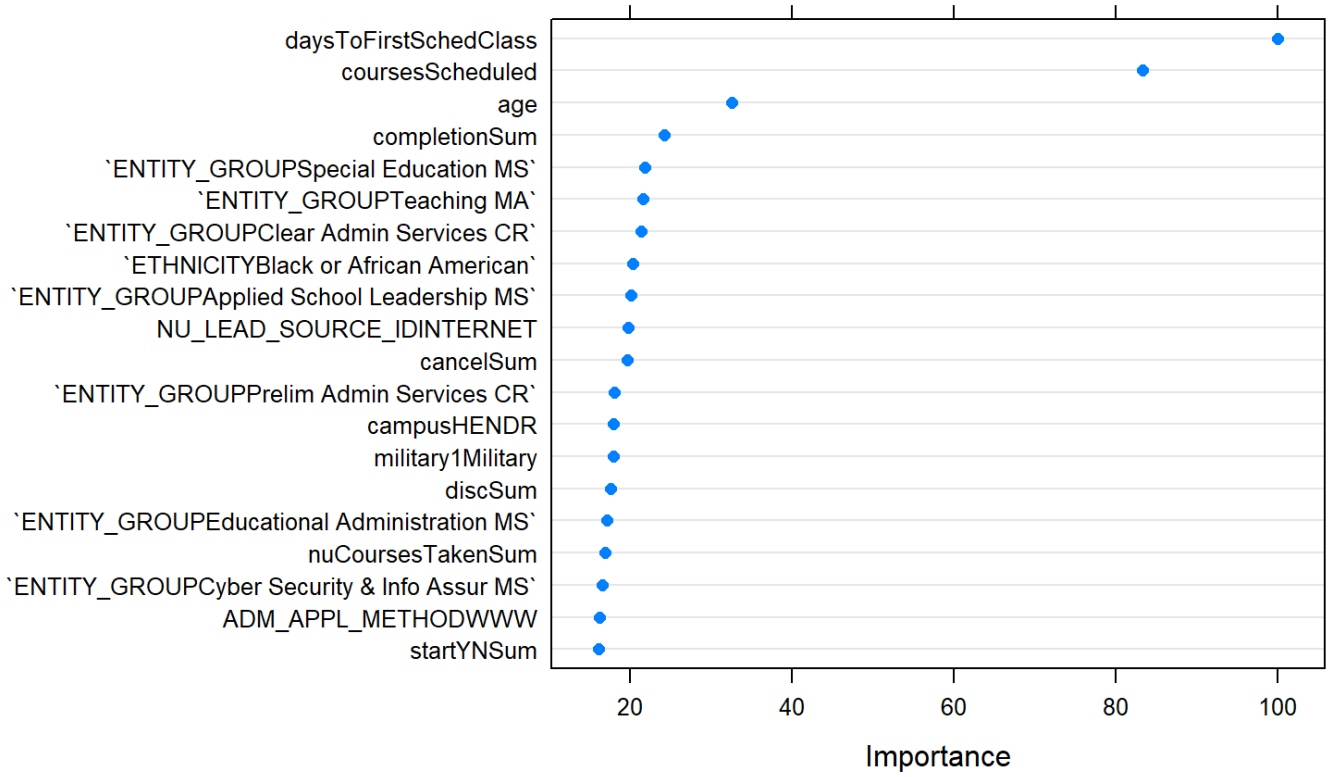
3.2 Variable Importance

Variable importance was calculated for the best performing model from each of the 4 algorithms. The top 20 most important variables for each model are plotted below. The relative importance of variables was incredibly consistent across all 4 models. Across every model, daysToFirstSchedClass was by far the most important variable. The second most important variable across every model was coursesScheduled. The third most important variable across all 4 models was age. However, age was relatively not as important as daysToFirstSchedClass or coursesScheduled. Looking across all 4 models, it is clear that daysToFirstSchedClass and coursesScheduled were the driving variables behind each models' predictive power. These 3 variables are interpreted in the next section.

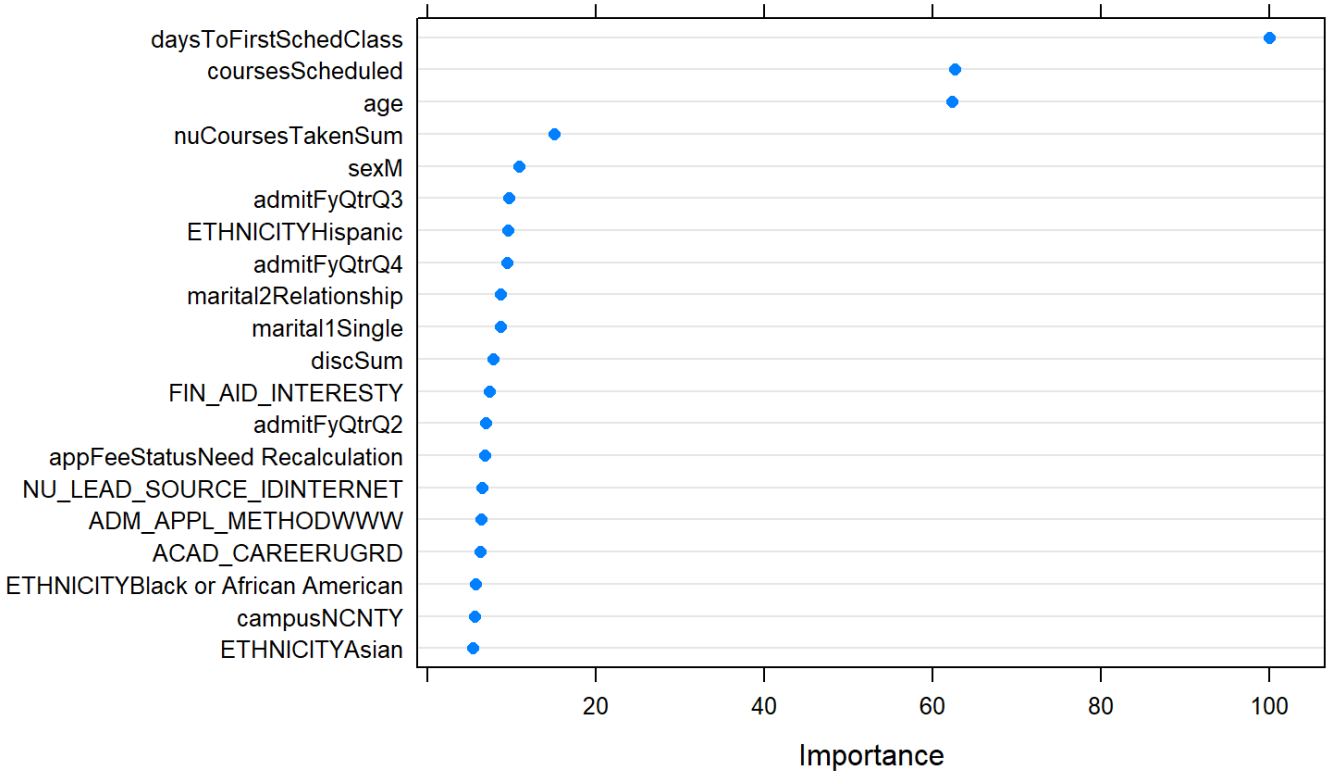
Relative Variable Importance for Balanced XGBoost



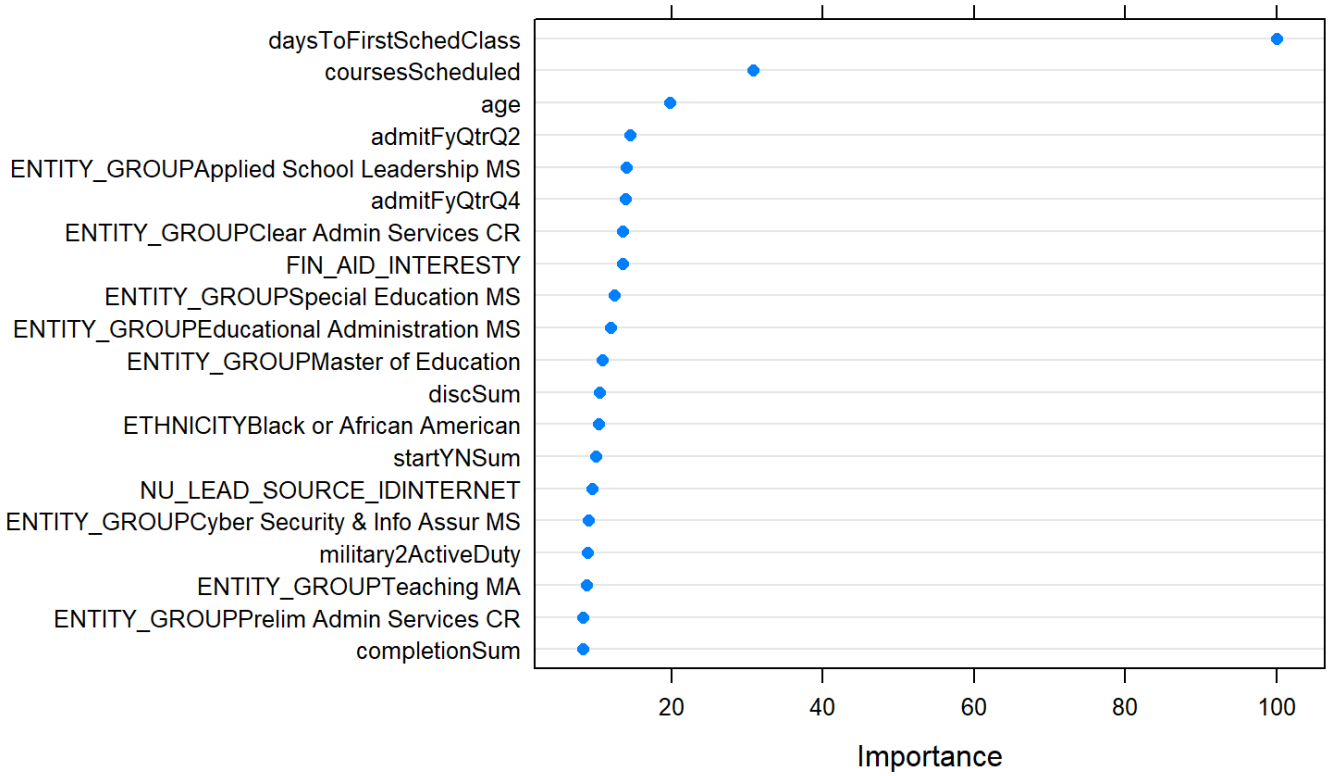
Relative Variable Importance for Balanced Logistic Regression



Relative Variable Importance for Balanced Random Forest



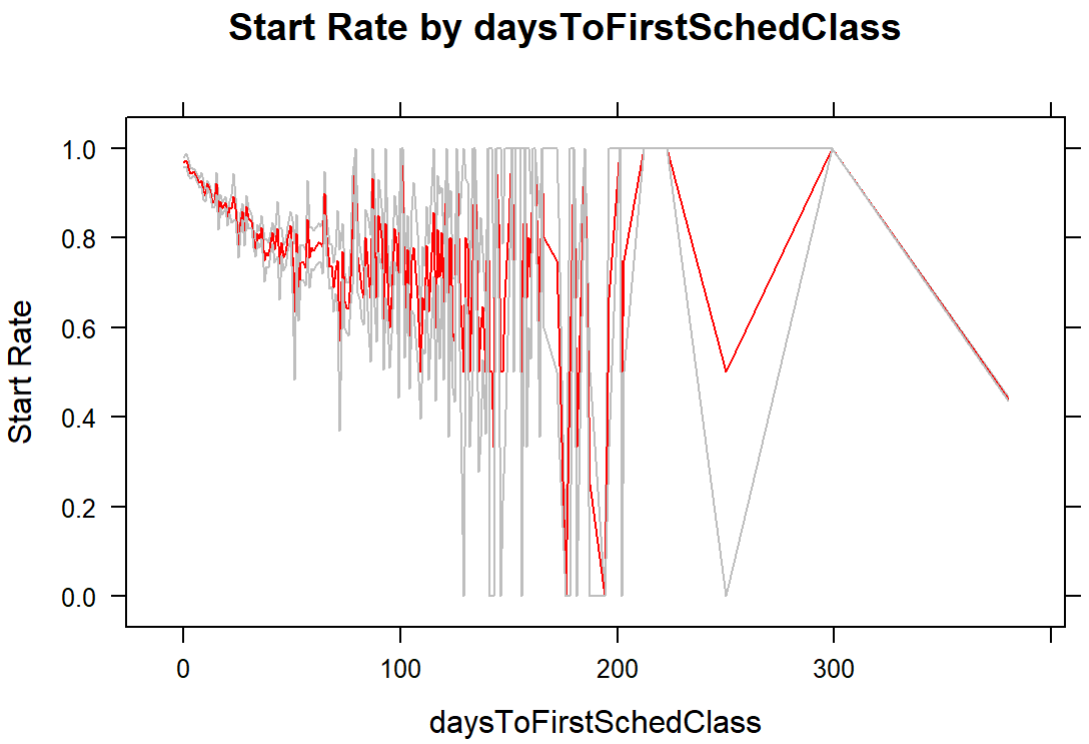
Variable Importance for Balanced Cox-Proportional Hazards Regression



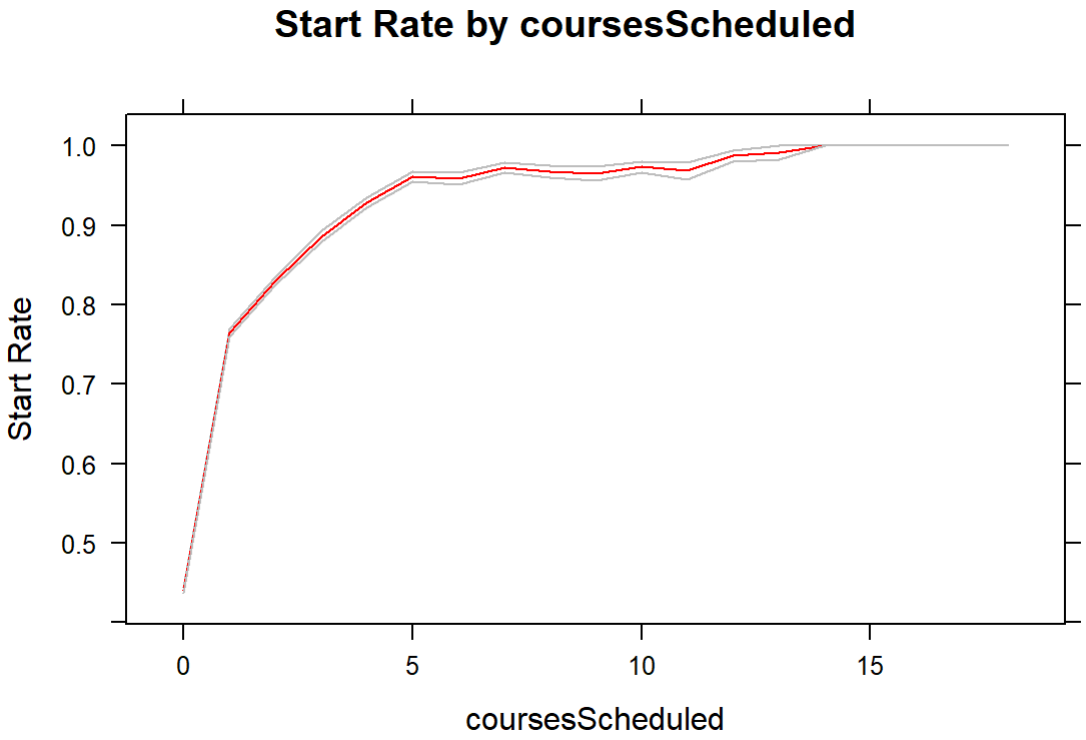
3.3 Independent Effects

The average start rate for daysToFirstSchedClass, coursesScheduled and age are plotted below. These 3 variables were the most important variables for all of the top algorithms. The red line in each plot is the average start rate and the gray lines are the standard errors. It is desireable to have small standard errors. That is, the closer the gray line is to the red line, the less error/variance there is.

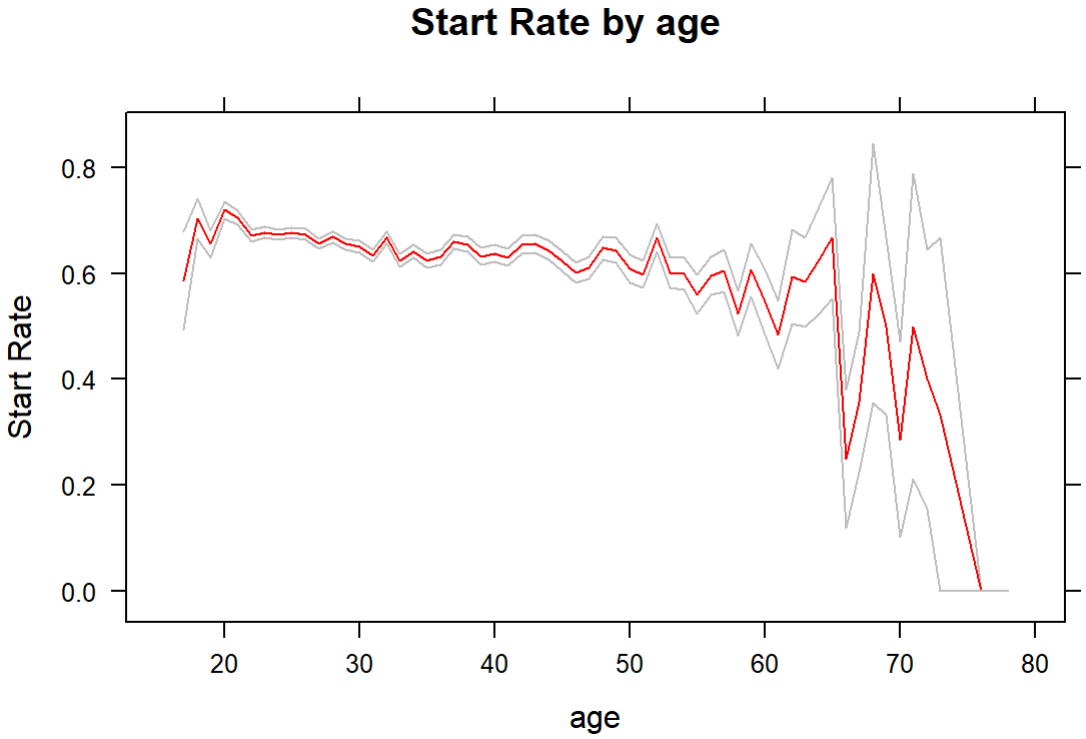
The variable daysToFirstSchedClass was by far the most important variable across all models. This variable represents how many days until the students' first scheduled class, measured 7 days after matriculation. If a student did not have a class scheduled at this time, the value of 380 was substituted. Although the plot is somewhat noisy, there is more-or-less a negative relationship. That is, the further away the students first scheduled class is, the less likely they are to start.



The variable coursesScheduled was the second most important variable across all models. This variable represents how many courses the student had scheduled as of 7 days after matriculation. If a student did not have any classes scheduled at this time, the value of 0 was substituted. The relationship between start rate and coursesScheduled is positive and very consistent (small standard errors). That is, the more courses a student has scheduled, the more likely they are to start.



The variable age was the third most important variable across all models, although, it's influence was not as strong as daysToFirstSchedClass or coursesScheduled. This variable represents how old (in years) the student was as of 7 days after matriculation. The relationship between start rate and age is somewhat parabolic. That is, as age increases from roughly 18 to 22, the start rate increases. Then, from roughly 22 and greater, the start rate has a steady decline. Overall, the older a student is, the less likely they are to start.



4. Conclusions

The goal of this analysis was to predict which student would start at National University based on only information available within 7 days of matriculation. Overall, all models did surprisingly well. All had high AUC values around 0.80 and accuracy levels above 0.70.

The modeling results indicate that XGBoost was superior across all evaluation measures. XGBoost was the best performing algorithm in every instance. Logistic Regression was the second best performing algorithm across every evaluation measure. Logistic Regression consistenly performed only slightly worse than XGBoost. **Despite XGBoost outperforming all other algorithms, Logistic Regression would be chosen as the model to put into a production environment because Logistic Regression provides detailed explanations about its predictions whereas XGBoost does not supply as detailed explanations**. Specifically, the Logistic Regression model trained on the balanced data set would be the final model since it had very high AUC while also having exceptionally high accuracy when predicting non-start students.

Variable importance indicates that daysToFirstSchedClass and coursesScheduled are by the most influencial variables for predicting which students will start. Should we want to put this model into production, it would be worthwhile to investigate how much the other variables impact model performance and potentially exclude some of them. It also might be beneficial to see if interaction terms improve model performance.

Appendix

Appendix A: Descriptive Statistics

```
 eventStartYN    daysToEvent    ACAD_CAREER  ACAD_PLAN_TYPE
Min.   :0.000   Min.   :  0.0   GRAD:21995   0MAJ :33669
1st Qu.:0.000   1st Qu.: 18.0   UGRD:20490   CER  : 1132
Median :1.000   Median : 40.0                CRE  : 3919
Mean   :0.652   Mean   :107.8                Other: 3765
3rd Qu.:1.000   3rd Qu.:129.0
Max.   :1.000   Max.   :380.0

ADM_APPL_METHOD admitFyQtr      age                    appFeeStatus
HRD:10909       Q1:10268   Min.   : 0.00   Calculated        : 7551
WWW:31576       Q2: 7132   1st Qu.:25.00   Need Recalculation:33262
                Q3:13462   Median :31.00   Pending           : 1672
                Q4:11623   Mean   :32.72
                           3rd Qu.:38.00
                           Max.   :81.00

    campus        cancelSum       completionSum     coursesScheduled
SDIEG  : 8874   Min.   :0.00000   Min.   :0.00000   Min.   : 0.000
MILIT  : 6921   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 0.000
NCNTY  : 5379   Median :0.00000   Median :0.00000   Median : 1.000
INGWD  : 3514   Mean   :0.02933   Mean   :0.09914   Mean   : 1.663
SACRA  : 2539   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.: 2.000
CMESA  : 2530   Max.   :5.00000   Max.   :4.00000   Max.   :22.000
(Other):12728
daysToFirstSchedClass    discSum                      ENTITY_GROUP
Min.   :  0.0         Min.   :0.0000   0 All Other Programs: 5106
1st Qu.: 20.0         1st Qu.:0.0000   Public Health BS    : 4480
Median : 95.0         Median :0.0000   Non-Degree          : 3751
Mean   :198.4         Mean   :0.2857   Master of Education : 3078
3rd Qu.:380.0         3rd Qu.:0.0000   Special Education MS: 1817
Max.   :380.0         Max.   :6.0000   BBA                 : 1316
                                       (Other)             :22937
                    ETHNICITY     FIN_AID_INTEREST flagBLK
0White                   :16378   N: 9632          N:41432
Hispanic                 :10740   Y:32853          Y: 1053
Black or African American: 4845
Unknown                  : 4333
Asian                    : 3542
Two or more races        : 1911
(Other)                  :  736
flagBlockEnrl flagCRD   flagHold  flagLock  flagREC
N:37915       N:42260   N:41848   N:38491   N:39677
Y: 4570       Y:  225   Y:  637   Y: 3994   Y: 2808




             marital              military     nonCA
0Unknown          : 3884   0NonMilitary:30984   N:36659
1Single           :17685   1Military   : 6131   Y: 5826
2Relationship     :17212   2ActiveDuty : 5370
3EndedRelationship: 3704
```

```
           NU_LEAD_SOURCE_ID nuCoursesTakenSum sex         startYNSum
 0REFERRAL           :27378   Min.   :  0.000   F:25217   Min.   :0.0000
 INTERNET            : 8382   1st Qu.:  0.000   M:16875   1st Qu.:0.0000
 EVENT/OUTREACH      : 2029   Median :  0.000   U:  393   Median :0.0000
 ELECTRONIC MARKETING: 1393   Mean   :  4.157             Mean   :0.3291
 Other               : 1370   3rd Qu.:  0.000             3rd Qu.:0.0000
 SOCIAL MEDIA        :  783   Max.   :101.000             Max.   :7.0000
 (Other)             : 1150
```

Appendix B: Confusion Matrixes

Hazard Balanced

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3215 1259
         1 2261 6010

               Accuracy : 0.7238
                 95% CI : (0.716, 0.7316)
    No Information Rate : 0.5703
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4235
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8268
            Specificity : 0.5871
         Pos Pred Value : 0.7266
         Neg Pred Value : 0.7186
             Prevalence : 0.5703
         Detection Rate : 0.4716
   Detection Prevalence : 0.6490
      Balanced Accuracy : 0.7070

       'Positive' Class : 1
```

Hazard Unbalanced

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1986 2488
         1 1130 7141

               Accuracy : 0.7161
                 95% CI : (0.7082, 0.7239)
    No Information Rate : 0.7555
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3303
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7416
            Specificity : 0.6374
         Pos Pred Value : 0.8634
         Neg Pred Value : 0.4439
             Prevalence : 0.7555
         Detection Rate : 0.5603
   Detection Prevalence : 0.6490
      Balanced Accuracy : 0.6895

       'Positive' Class : 1
```

Logistic Regression Balanced

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 3449 2546
         Y 1025 5725

               Accuracy : 0.7198
                 95% CI : (0.7119, 0.7276)
    No Information Rate : 0.649
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4296
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6922
            Specificity : 0.7709
         Pos Pred Value : 0.8481
         Neg Pred Value : 0.5753
             Prevalence : 0.6490
         Detection Rate : 0.4492
   Detection Prevalence : 0.5296
      Balanced Accuracy : 0.7315

       'Positive' Class : Y
```

Logistic Regression Unbalanced

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 2687 1629
         Y 1787 6642

               Accuracy : 0.732
                 95% CI : (0.7242, 0.7396)
    No Information Rate : 0.649
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4069
 Mcnemar's Test P-Value : 0.007227

            Sensitivity : 0.8030
            Specificity : 0.6006
         Pos Pred Value : 0.7880
         Neg Pred Value : 0.6226
             Prevalence : 0.6490
         Detection Rate : 0.5211
   Detection Prevalence : 0.6614
      Balanced Accuracy : 0.7018

       'Positive' Class : Y
```

Random Forest Balanced

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 3447 2600
         Y 1027 5671

               Accuracy : 0.7154
                 95% CI : (0.7075, 0.7232)
    No Information Rate : 0.649
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.422
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6856
            Specificity : 0.7705
         Pos Pred Value : 0.8467
         Neg Pred Value : 0.5700
             Prevalence : 0.6490
         Detection Rate : 0.4450
   Detection Prevalence : 0.5255
      Balanced Accuracy : 0.7281

       'Positive' Class : Y
```

Random Forest Unbalanced

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 2534 1583
         Y 1940 6688

               Accuracy : 0.7236
                 95% CI : (0.7157, 0.7313)
    No Information Rate : 0.649
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.382
 Mcnemar's Test P-Value : 2e-09

            Sensitivity : 0.8086
            Specificity : 0.5664
         Pos Pred Value : 0.7752
         Neg Pred Value : 0.6155
             Prevalence : 0.6490
         Detection Rate : 0.5248
   Detection Prevalence : 0.6770
      Balanced Accuracy : 0.6875

       'Positive' Class : Y
```

XGBoost Balanced

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 3616 2723
         Y  858 5548

               Accuracy : 0.719
                 95% CI : (0.7111, 0.7268)
    No Information Rate : 0.649
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4372
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6708
            Specificity : 0.8082
         Pos Pred Value : 0.8661
         Neg Pred Value : 0.5704
             Prevalence : 0.6490
         Detection Rate : 0.4353
   Detection Prevalence : 0.5026
      Balanced Accuracy : 0.7395

       'Positive' Class : Y
```

XGBoost Unbalanced

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 2754 1678
         Y 1720 6593

               Accuracy : 0.7334
                 95% CI : (0.7256, 0.741)
    No Information Rate : 0.649
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4136
 Mcnemar's Test P-Value : 0.4818

            Sensitivity : 0.7971
            Specificity : 0.6156
         Pos Pred Value : 0.7931
         Neg Pred Value : 0.6214
             Prevalence : 0.6490
         Detection Rate : 0.5173
   Detection Prevalence : 0.6523
      Balanced Accuracy : 0.7063

       'Positive' Class : Y
```