Machine Learning Models for International Student Success in Texas Higher Education

Greg Argueta, Maitreyi Pillai, and Sonali Singh

University of Texas at Dallas

EPPS 6323 Knowledge Mining: Final Report

Instructor: Dr. Karl Ho

05/01/2024

Abstract

The rapid growth of the international student population in U.S. postsecondary institutions, increasing by over 84% between 2006 and 2017, underscores the need for enhanced predictive capabilities regarding their educational outcomes. This study leverages machine learning techniques to perform predictions of international student college graduation based on student-level factors utilizing administrative data collected by Texas Higher Education Coordinating Board (THECB). Some of these factors are in line with variables commonly used in causal research, such as total courses taken and age. Key predictors include demographic factors (usually included as control variables in causal research frameworks), academic preparedness, and integration into campus life. Traditional statistical methods and innovative machine learning models, such as random forests, were employed to analyze data on over 83,000 students, revealing that variables like age, academic load, and socioeconomic factors significantly influence graduation outcomes. This study refines the predictive accuracy of international student graduation rates and highlights the critical role of academic and non-academic support in enhancing student success.

Introduction

From the 2006-07 academic year to the 2016-17 academic year, the international student population is the United States (US) postsecondary education system has grown more that 84% to over $1 million students (Alsakran, 2018). Prior descriptive studies have identified differences in persistence and gradations rates between domestic and international students (Alsakran, 2018).

Theoretical models have also been proposed which consider preexisting student characteristics such as the country of origin, financial support, gender, institutional status (new or transfer student), and TOEFL score (Kwai, 2010). Model features also included common persistence metrics consisting of attempted credit hours per semester (momentum), grade point averages (GPA), and campus community engagement and participation. While these methods thoroughly illustrate the lower success rates of international students, they have not adequately explained or predicted outcomes.

This paper's objective is to take what is known about international student bachelor's degree attainment and utilize new and established variables as predictors in a machine learning framework. Machine learning methods have fewer limitations than traditional causal techniques and higher predictive potential. This unique approach is expected to produce better predictions of international student college success.

Literature Review

Existing studies have examined how transferring between higher education institutions impacts persistence and found that students who are willing and able to stay at the same institution are generally more likely to enroll in the next academic year (Alsakran, 2018). The degree attainment of undergraduate international students is shaped by an intricate interplay of individual student characteristics and institutional factors. Although the existing literature does not focus specifically on international students in Texas, valuable insights can be drawn from broader studies on degree attainment.

These studies focused on community college students, and persistence patterns were similar across race classifications of US students. International students have the additional burden of maintaining their international visa requirements which may impact their ability to persist in college. Alsakran (2018) demonstrated that the international students who transferred to different community colleges were less likely to persistence than those who stayed at the same institution. Alsakran and Slate (2017) found that international student persistence was inconsistent across Texas community colleges in a descriptive study covering the 2010-11 to 2012-13 academic years. In this study, the authors found that the gap between community college persistence rates for international students was as high as 35.5%. The large difference may be an indication of the type of support, advising, or services available to international students at some institutions and not others.

Student characteristics such as academic preparedness, socioeconomic background, and parental education levels are significant predictors of degree completion (Gansemer-Topf et al., 2018; Arellano, 2019). Academic preparedness not only encompasses prior educational achievements but also the student's ability to adapt to the academic rigors of higher education. Socioeconomic background influences access to resources that can support educational success, while parental education often correlates with higher aspirations and support for postsecondary education.

Institutional factors also play a pivotal role in student success. These include the size and selectivity of the institution, which often reflect the quality of educational offerings and the level of student support available. Institutions that maintain high standards for entry tend to provide environments that foster academic excellence. Moreover, the presence of robust undergraduate research programs and the quality of faculty and curriculum development are crucial for nurturing student potential (Flores, 2000; Oseguera, 2005).

Non-academic factors such as social and academic integration into the campus community significantly affect degree attainment. Students who are well-integrated socially and academically are more likely to persist through their college years. This integration helps in building a supportive network that enhances the student experience and aids in overcoming academic and personal challenges (Gansemer-Topf et al., 2018). Furthermore, the perception of the campus climate, particularly in terms of inclusivity and discrimination, can significantly impact student retention and success. Positive perceptions of the campus environment are associated with higher graduation rates, particularly among minority students (Brown et al., 2005).

The realm of predictive analytics in higher education is pivotal, with institutions increasingly leveraging machine learning models to inform student support services and enhance graduation rates. Ojha (2017) underscores the importance of a multi-model approach in predicting graduation delays, identifying key pre-university characteristics and advocating for model transparency in predictive analytics, a call echoed by Bird et al. (2018). They argue for transparency in predictive modelling, highlighting the balance between model complexity and thoughtful sample construction in enhancing performance.

Kovacic (2010) and Attewell et al. (2010) explore the critical impact of student demographics and the educational environment on student success, advocating for a nuanced approach to understanding their interplay. These studies collectively signal the transformative potential of predictive analytics in resource allocation and intervention targeting within higher education. Yet, they also bring to light ethical considerations around demographic profiling and bias reinforcement.

The literature collectively suggests a growing appreciation for predictive analytics in identifying at-risk students, particularly in STEM fields, as delineated by the National Academies of Sciences, Engineering, and Medicine (2016) and studies by Bernacki and Raković. Their innovative approaches to predicting student success underscore a paradigm shift towards using behavioral data and digital log data for educational interventions, reflecting broader social science research trends (Krumm et al., 2014).

This synthesis elucidates the multi-dimensional aspects of degree attainment among international students in Texas. It highlights the contributions of existing research, identifies gaps in knowledge, and substantiates the imperative for further study. By examining the intersection of student characteristics and institutional factors, it provides a contextual backdrop for future research aimed at enhancing the educational outcomes of this unique student population.

Research Question

The research question which guided the data analysis described below was: "How do student characteristics and institutional factors influence the degree attainment of undergraduate international students in Texas?". The data used to develop the primary data set for analysis was collected by the Texas Higher Education Coordinating Board (THECB) and contains student-level information for all postsecondary institutions across Texas. The major data sets used contained student status and demographic information, degrees awarded, course schedules, and faculty information for courses taken.

Development of the final dataset was completed and constructed with the following elements to answer the research question. Awarded degree information was included as the dependent or outcome variable for training supervised models. Student demographic data was included to inform the student characteristics component of the research question. Similarly, faculty and courses taken were integrated to provide insight into institutional factors influencing degree attainment by international students.

Because of Texas' population and major investments in higher education in previous decades, all data sets exceeded one million records with many of the course-related data sets approaching three million records. The selection criteria for this study required that the data sets be reduced to include only international students in undergraduate programs. These criteria reduced the data set to slightly more than 83000 records. As an additional note regarding the analysis dataset, the focus was graduation outcomes for the 2021 school year, with the assessment period including student and course information for the prior six years. This was done to follow the standard practice of assessing graduation rates in the sixth year after initial enrollment.

Exploratory Data Analysis

A preliminary working data set containing 166 variables or features was used to assess impacts on international student graduation outcomes. Degree completion is the outcome variable for this initial exploratory data analysis. Initial exploration set out to understand average treatment effects, which might identify underlying reasons for variation in student attainment of undergraduate degrees. An ordinary least squares (OLS) model using the preliminary data set was utilized to gain an initial understanding of the success rates for international students attempting to acquire undergraduate degrees.

The outcome (degree completion), a character variable, contained entries for each degree that a student earned. If a student had not earned a degree, the data field was blank. This variable was translated into a binary outcome by coding earned degrees with 1 and unearned degrees with 0.

Two models were developed for the preliminary exploration of the data. The first contained seven features (athletic scholarships, veteran scholarships, white ethnicity among international students, non-white ethnicities, sex of the student, full-time or part-time students, rank of the professor who taught the course, and the tenure status of the professor). Results from this model indicated that both athletic and veteran scholarships were statistically significant. It is important to note that the number of scholarships awarded was low relative to the overall number of students in the sample, but when awarded, the scholarship amounts relatively high. The effect of the white ethnicity classification returned a "not applicable" (NA) in the summary results. Contrastingly, non-white ethnicities showed positive and statistically significant impacts. The part-time or full-time status of a student was positive but not significant. The sex of the student was positive and statistically significant; however, a closer look at the data is warranted to interpret this result appropriately. The position of the faculty and the tenure status also had a positive and significant coefficient.

The second model used five features (athletic scholarships, non-white ethnicities, sex of the student, rank of the professor who taught the course, and the tenure status of the professor). Athletic scholarships remained statistically significant in this model although only with 95% confidence rather than the 99% confidence seen in the first model. Sex was positive but not significant. The results for non-white ethnicities, faculty position, and faculty tenure status remained positive and statistically significant as observed in the first model.

Meaningful independent variables were generally identified, and many were categorical or ordinal variables which limited the number and type of models that could be applied to the data.

The rpart tree model and randomForest R models were developed for the machine learning component of this this study and were configured for classification. The unit of analysis was the student-semester-course.

The independent variables selected for preliminary assessment include the age and sex of the student, binary variables for if the student's ethnicity was either Asian, Black, Hispanic, or White, the student's major, the courses taken, and faculty indicators for rank and tenure status. With these variables, the rpart model ran on the training dataset of 50,000 observations (total observations are currently 83,331). Prediction using the trained rpart model was not performed because the primary reason for the single tree model was to assist in the predictor process which informed the random forest model applied for prediction purposes.

Challenges were experienced during the model development process including the R version at the site where data was accessible having x11 font limitation that prevented the display of tree model decision tree output. The challenge did not prevent the use of the model for variable identification, but it did cause the team to seek out other methods for graphically representing the tree model's output.

Data and Methods

The outcome measured in this study will be a binary variable identifying either bachelor's degree completed or not completed. This structure is well suited to machine learning classification models. Tree and forest-based classification methods were used to identify the primary contributing features that lead international student degree attainment. Although the planned methods primarily focused on supervised learning techniques, unsupervised clustering methods were also attempted but did not add context to pre-existing and experiential student characteristics.

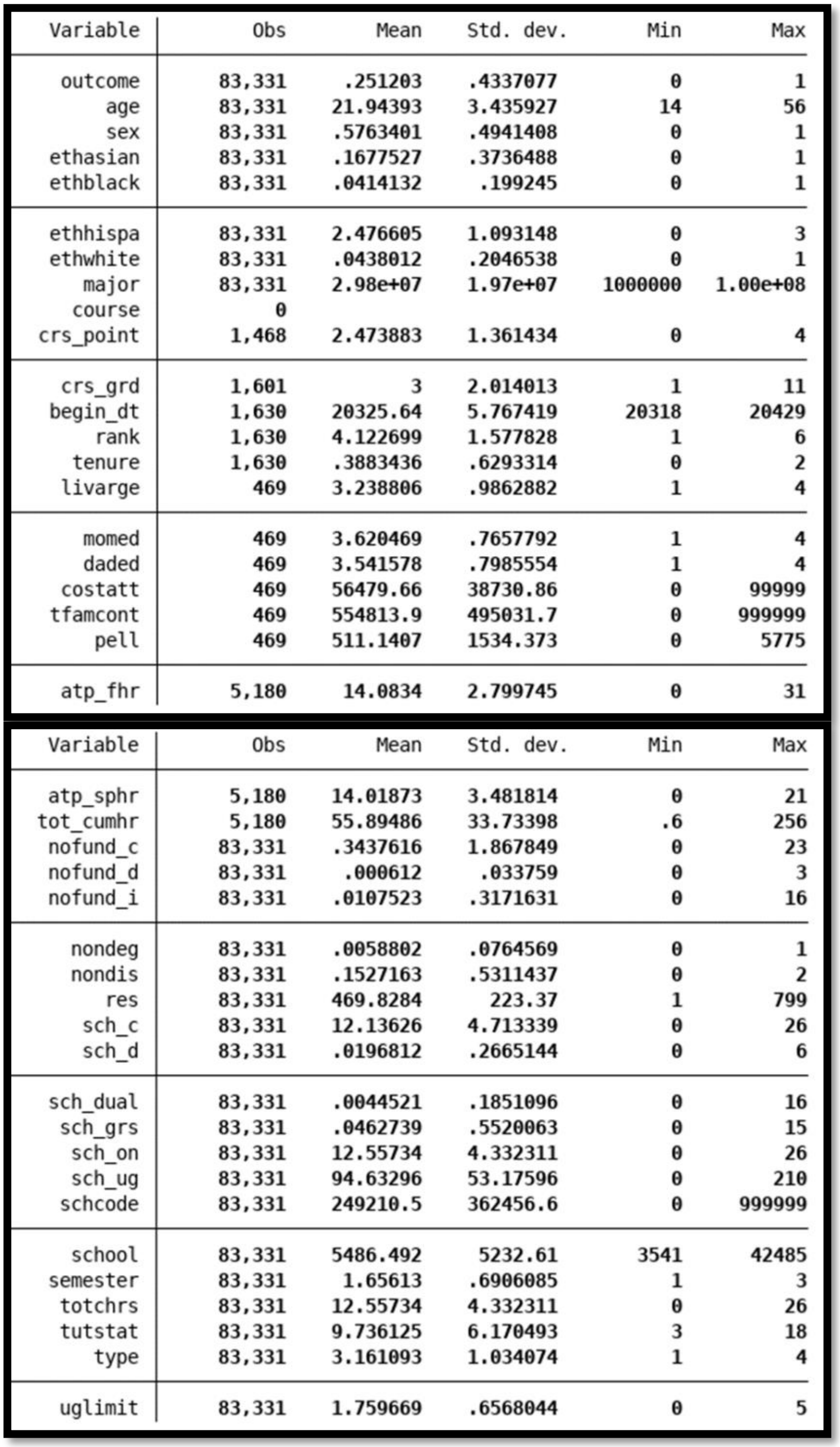
The Texas Schools Project (TSP) at the University of Texas at Dallas (UT Dallas), in connection with the Education Research Center maintains a rich set of administrative data from Texas primary and secondary schools, higher education institutions, and employers. The amount of data available for research was unparalleled when first developed and has made Texas a focal point for education research. This study will utilize TSP data in the evaluation of international student outcomes in Texas to take advantage of the large data set available within the state.

After initial review of the initial analysis, the data structure required redevelopment because a more comprehensive data table was identified as a base for merging with other data sets. The base data table was originally a THECB financial aid file, but the file would only include international students who received financial aid. Since this would represent a selection bias in the study data, the data set was rebuilt with a THECB enrollment file to ensure the full set of international students were included in the data set.

Degree attainment and financial aid data were then merged with the enrollment file to begin the build-out of the 2021 international student landscape across higher education institutions in Texas. Semester files were then developed for each fall and spring for 2021 and the prior five years. These files included courses taken, and attributes of the faculty teaching those courses. This represented twelve semester specific files over the six-year period and these files were merged to the primary data of 2021 enrollment and graduation information.

Lastly, to answer the research question, a binary variable was again created based on if a degree was awarded in 2021 to an international student. A 1 indicated that a student had attained a degree and 0 indicated that no degree was earned in the year. This binary dependent variable guided the data methods to be applied and led to a concentration on classification models. Descriptive statistics for the variables selected for predictive modelling are shown in Figure 1. Figure 2 contains variable descriptions as well as an identifier for each model in which each variable was included.

Figure 1 – Descriptive Statistics



```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     outcome |     83,331     .251203    .4337077          0          1
         age |     83,331    21.94393    3.435927         14         56
         sex |     83,331    .5763401    .4941408          0          1
    ethasian |     83,331    .1677527    .3736488          0          1
    ethblack |     83,331    .0414132     .199245          0          1
-------------+---------------------------------------------------------
    ethhispa |     83,331    2.476605    1.093148          0          3
    ethwhite |     83,331    .0438012    .2046538          0          1
       major |     83,331    2.98e+07    1.97e+07    1000000   1.00e+08
      course |          0
   crs_point |      1,468    2.473883    1.361434          0          4
-------------+---------------------------------------------------------
     crs_grd |      1,601           3    2.014013          1         11
    begin_dt |      1,630    20325.64    5.767419      20318      20429
        rank |      1,630    4.122699    1.577828          1          6
      tenure |      1,630    .3883436    .6293314          0          2
     livarge |        469    3.238806    .9862882          1          4
-------------+---------------------------------------------------------
       momed |        469    3.620469    .7657792          1          4
       daded |        469    3.541578    .7985554          1          4
     costatt |        469    56479.66    38730.86          0      99999
    tfamcont |        469    554813.9    495031.7          0     999999
        pell |        469    511.1407    1534.373          0       5775
-------------+---------------------------------------------------------
     atp_fhr |      5,180     14.0834    2.799745          0         31

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    atp_sphr |      5,180    14.01873    3.481814          0         21
   tot_cumhr |      5,180    55.89486    33.73398         .6        256
    nofund_c |     83,331    .3437616    1.867849          0         23
    nofund_d |     83,331     .000612     .033759          0          3
    nofund_i |     83,331    .0107523    .3171631          0         16
-------------+---------------------------------------------------------
      nondeg |     83,331    .0058802    .0764569          0          1
      nondis |     83,331    .1527163    .5311437          0          2
         res |     83,331    469.8284      223.37          1        799
       sch_c |     83,331    12.13626    4.713339          0         26
       sch_d |     83,331    .0196812    .2665144          0          6
-------------+---------------------------------------------------------
    sch_dual |     83,331    .0044521    .1851096          0         16
     sch_grs |     83,331    .0462739    .5520063          0         15
      sch_on |     83,331    12.55734    4.332311          0         26
      sch_ug |     83,331    94.63296    53.17596          0        210
     schcode |     83,331    249210.5    362456.6          0     999999
-------------+---------------------------------------------------------
      school |     83,331    5486.492     5232.61       3541      42485
    semester |     83,331     1.65613    .6906085          1          3
     totchrs |     83,331    12.55734    4.332311          0         26
     tutstat |     83,331    9.736125    6.170493          3         18
        type |     83,331    3.161093    1.034074          1          4
-------------+---------------------------------------------------------
     uglimit |     83,331    1.759669    .6568044          0          5
```

Figure 2 – Description of Variables

Variable	Variable Description	Model Notes
outcome	Earned a degree - binary variable	25 and 42 Variable Model
age	Age	25 and 42 Variable Model
sex	Gender - ‘0’ is female, ‘1’ is male	25 and 42 Variable Model
ethasian	Asian	25 and 42 Variable Model
ethblack	Black or African-American	25 and 42 Variable Model
ethhispa	Ethnic Origin-Hispanic or Latino or not	25 and 42 Variable Model
ethwhite	White	25 and 42 Variable Model
major	Major Area of Concentration	25 and 42 Variable Model
course	Course Number	42 Variable Model Only
crs_point	Grade Points	42 Variable Model Only
crs_grd	Course Grade	42 Variable Model Only
begin_dt	Begin Date	42 Variable Model Only
rank	Rank of the faculty member	42 Variable Model Only
tenure	tenure	42 Variable Model Only
livarge	Living Arrangement	42 Variable Model Only
momed	Mother's Highest Grade Level	42 Variable Model Only
daded	Father's Highest Grade Level	42 Variable Model Only
costatt	Cost of Attendance	42 Variable Model Only
tfamcont	Total Family Contribution	42 Variable Model Only

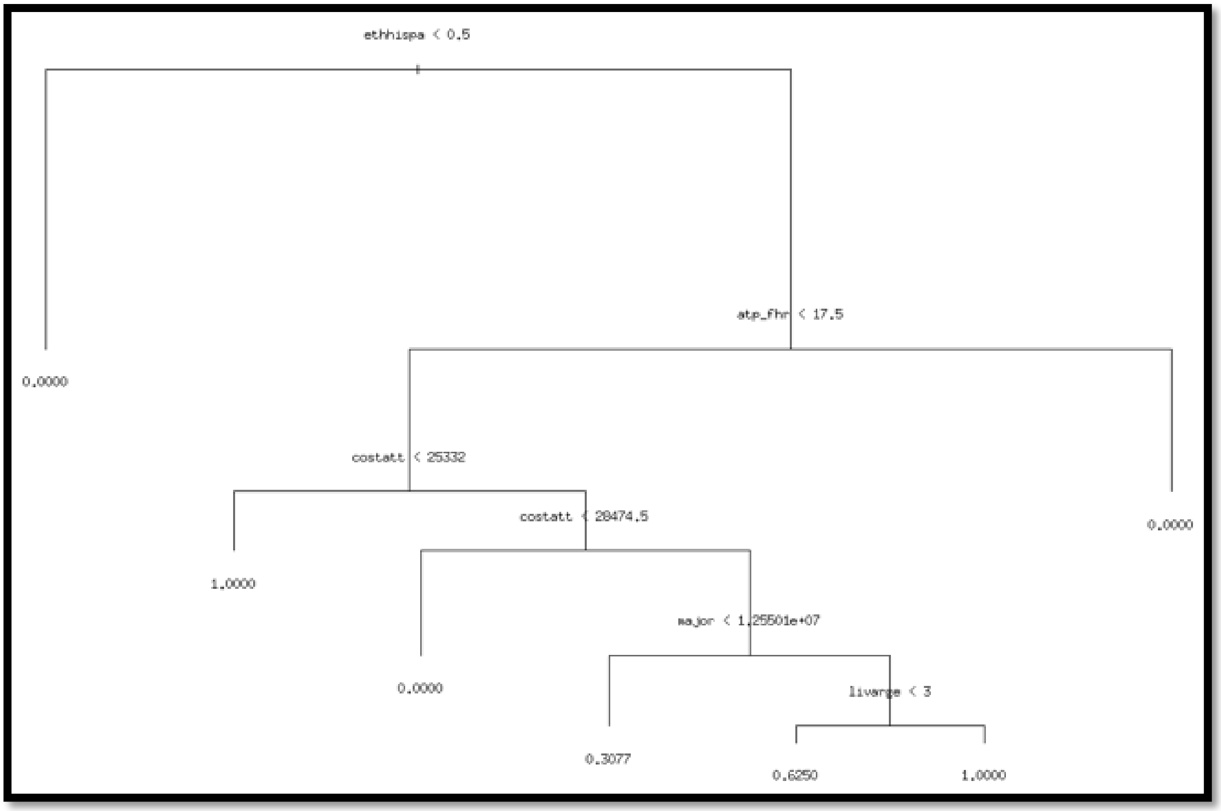
pell	Federal Pell Grant	42 Variable Model Only
atp_fhr	Attempted Hoursin the Fall Semester	42 Variable Model Only
atp_sphr	Attempted Hoursin the Spring Semester	42 Variable Model Only
tot_cumhr	Total Cumulative AttemptedHours	42 Variable Model Only
nofund_c	SCH-collegiate not fund	25 and 42 Variable Model
nofund_d	SCH-develop not fund	25 and 42 Variable Model
nofund_i	SCH-inter-institute not fund	25 and 42 Variable Model
nondeg	Non-Degree-Seeking Student	25 and 42 Variable Model
nondis	Non-disclosure	25 and 42 Variable Model
res	Residence	25 and 42 Variable Model
sch_c	SCH-collegiate fund	25 and 42 Variable Model
sch_d	SCH-develop fund	25 and 42 Variable Model
sch_dual	SCH-dual credit	25 and 42 Variable Model
sch_grs	Graduate SCH of Seniors	25 and 42 Variable Model
sch_on	SCH Load	25 and 42 Variable Model
sch_ug	SCH-undergrad degree program	25 and 42 Variable Model
schcode	High School Code	25 and 42 Variable Model
school	Institution Code - FICE	25 and 42 Variable Model
semester	Semester	42 Variable Model Only
totchrs	Total SCH	25 and 42 Variable Model
tutstat	Tuition Status	25 and 42 Variable Model

type	Classification	42 Variable Model Only
uglimit	Student Affected by UG funding limit	25 and 42 Variable Model

Basic Tree Model

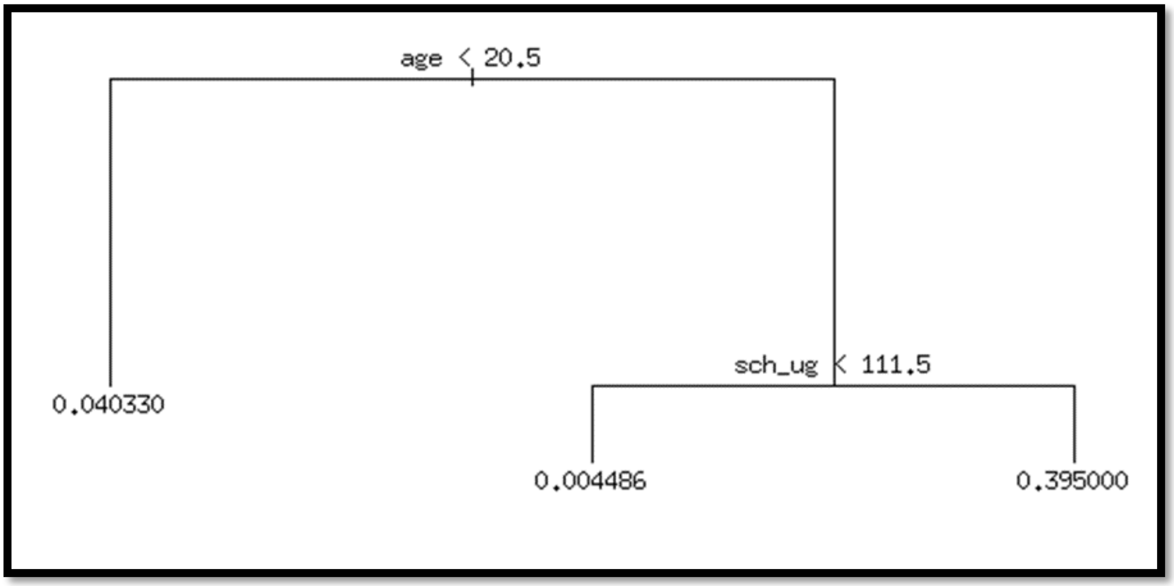
The rpart tree function in R experienced difficulty when using data with high missingness for building single tree models. The following graphs represent two tree models after breaking the data set into a 42-variable model (Figure 3) and a 25-variable model (Figure 4) based on predictors that were expected to aid in predicting international student graduation.

Figure 3 – 42 Variable Model (with High Missingness)



Ethhisp is the first split which is a variable indicating if a student is classified as Hispanic or not. This split shows that all student who graduated in the 42 variable data set was Hispanic. The second split is atp_fhr which represents the number of hours attempted in the fall semester and indicates that all graduates attempted less than 17.5 credit hours in fall. The following two splits were based on the costatt variable which is a metric for the cost of attendance at the higher education institution. For students who graduated in 2021, nearly all higher education institutions attended had a cost of attendance higher than $28,474.50. A student's selected major was the fifth split in the model, although the number identifying the major is a classification variable not interpreted as a number. Regardless, majors starting with 1255 or less make-up 30.77% of graduates in 2021. The sixth and final variable is livarge for the living arrangements of the student. Students with living arrangements less than three represented 62.5% of graduates within this dataset. Living arrangements less than three in the data indicated that students either living with their parents or on campus were more likely to graduate than those leaving in other situations. As seen in the 42 variable tree model, some of the splits did not make sense from an external validity perspective. This was due to the significant missingness in the data which leads to significant observation exclusion in the tree assessment. In order to develop a more stable classification tree with no missingness, a second dataset was developed with only the 25 variables that contained complete data. The results from that reduced but complete model are shown in Figure 4.

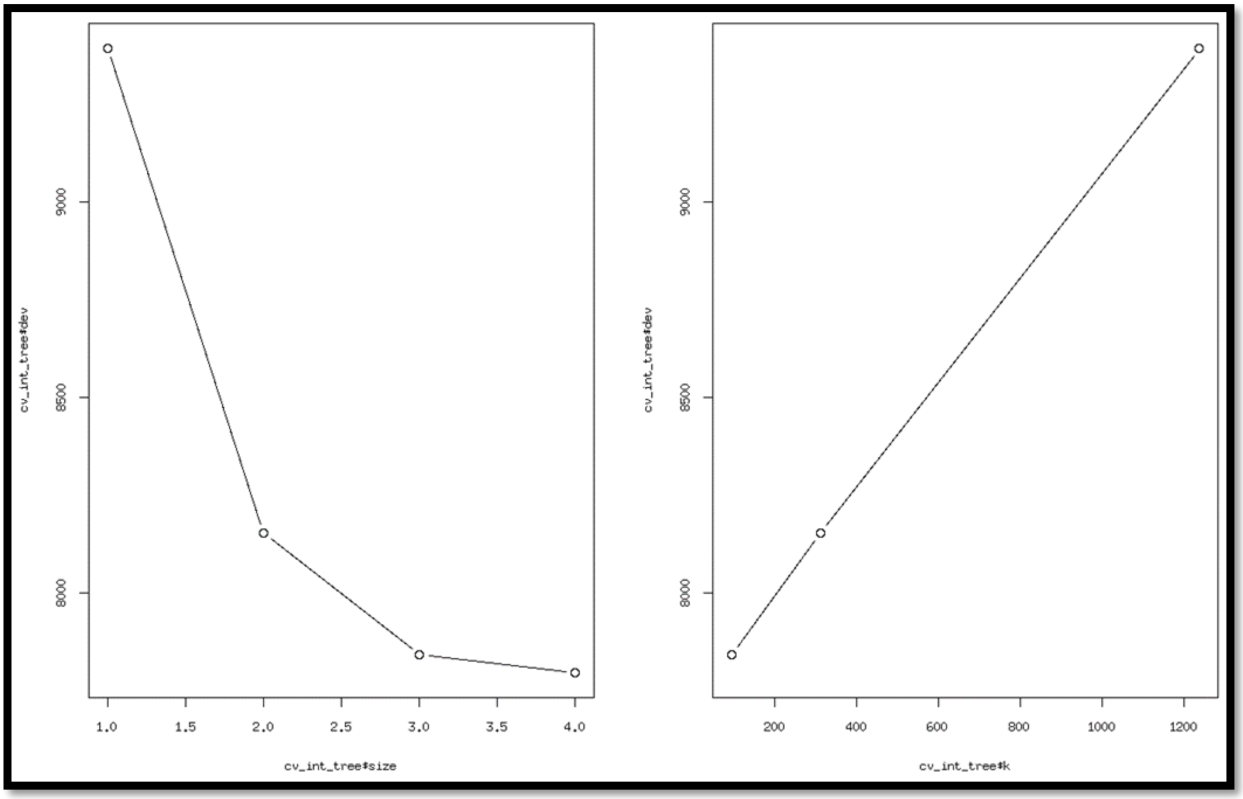
Figure 4 – 25 Variable Model (with No Missingness)



The 25-variable tree was much shallower with age representing the first split and showed that most graduates are older than 20.5 years of age. The second split of sch_ug was the variable for the semester credit hours taken in undergraduate programs and indicated that most graduates have completed more than 111.5 semester credit hours.

Robustness checks in the form of cross-validations were performed on the single tree model and depicted in Figure 5. In the cross-validation plots, the y-axis was a measure of prediction deviance from the actual student outcome. Deviance in the single tree model decreased with the number of tree splits to an optimal level of 4 splits. As the data was separated into a greater number of folds (k), the deviance increases. This is likely a sign of a testing set that includes too few observations to make accurate assessments against or that the model contains too few variables to make accurate predictions on consistent basis with individual cases.

Figure 5 – 25 Variable Cross Validation Assessment



Random Forest Models

Using the same 25-variable data set to avoid missingness problems, the data for the random forest model was separated into a training data set representing 66% of the original data set and testing data representing the remainder of the observations. The variables for this model were plot in Figure 6 by their importance to the forest's ability to make predictions about student graduation outcomes. Age was an understandably strong predictor because most college graduates reach a certain age prior to graduation. The following variable was sch_ug which is the number of semester credit hours earned in total by an undergraduate student. Again, most graduation criteria require students to have a minimum number of hours taken, making it intuitive why this variable was a strong predictor of outcomes. The third most important predictor was the major a student had selected. The reason why this variable is highly important is less clear and would most likely require a causal analysis to understand. It's possible that certain majors require less credit hours to complete or are less expensive, making them associated with other variables that are more easily understood.

The last significant important variable was res representing residency status. Although all students in this data set were classified as international, some will have met the residency criteria specific to tuition and fee charges. This variable could be important because it would significantly reduce the cost of attendance for international students. School and sex were less important than those listed above but represent interesting areas for potential future causal research.

Figure 6 – 25 Variable Importance Plot – Random Forest

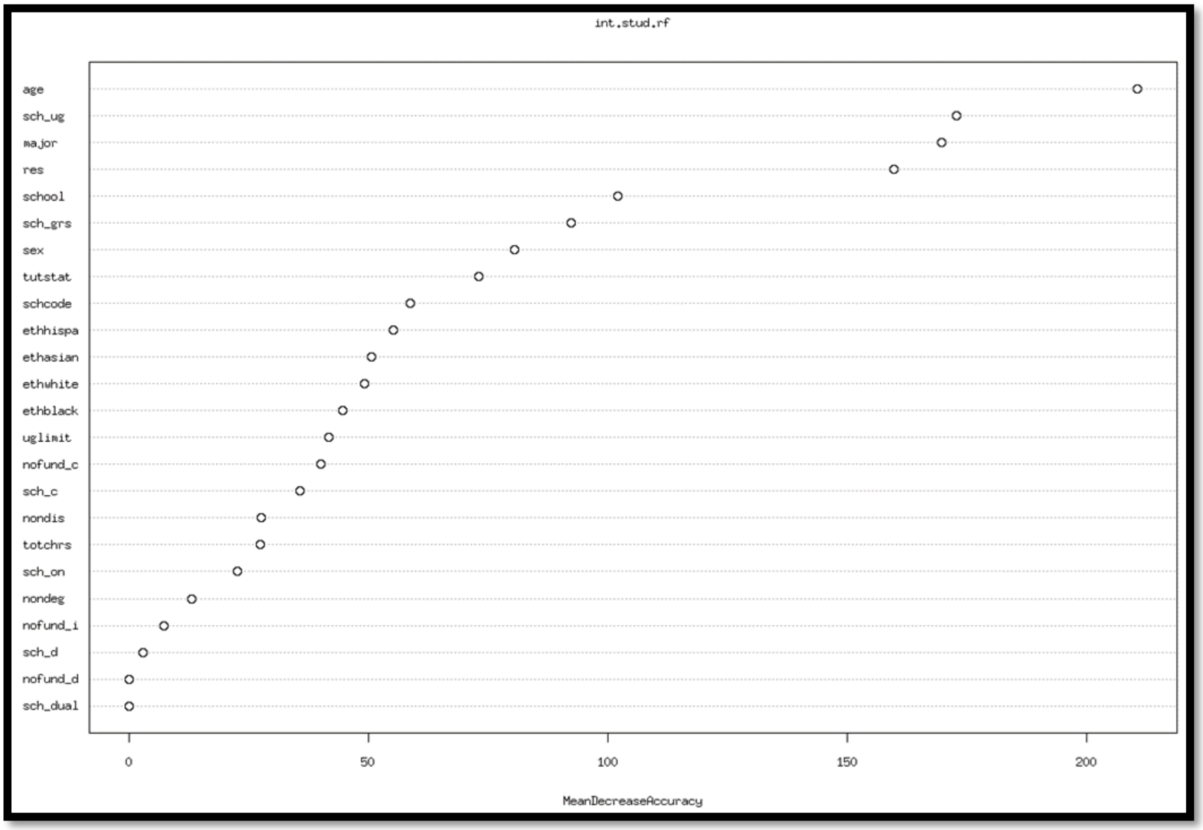
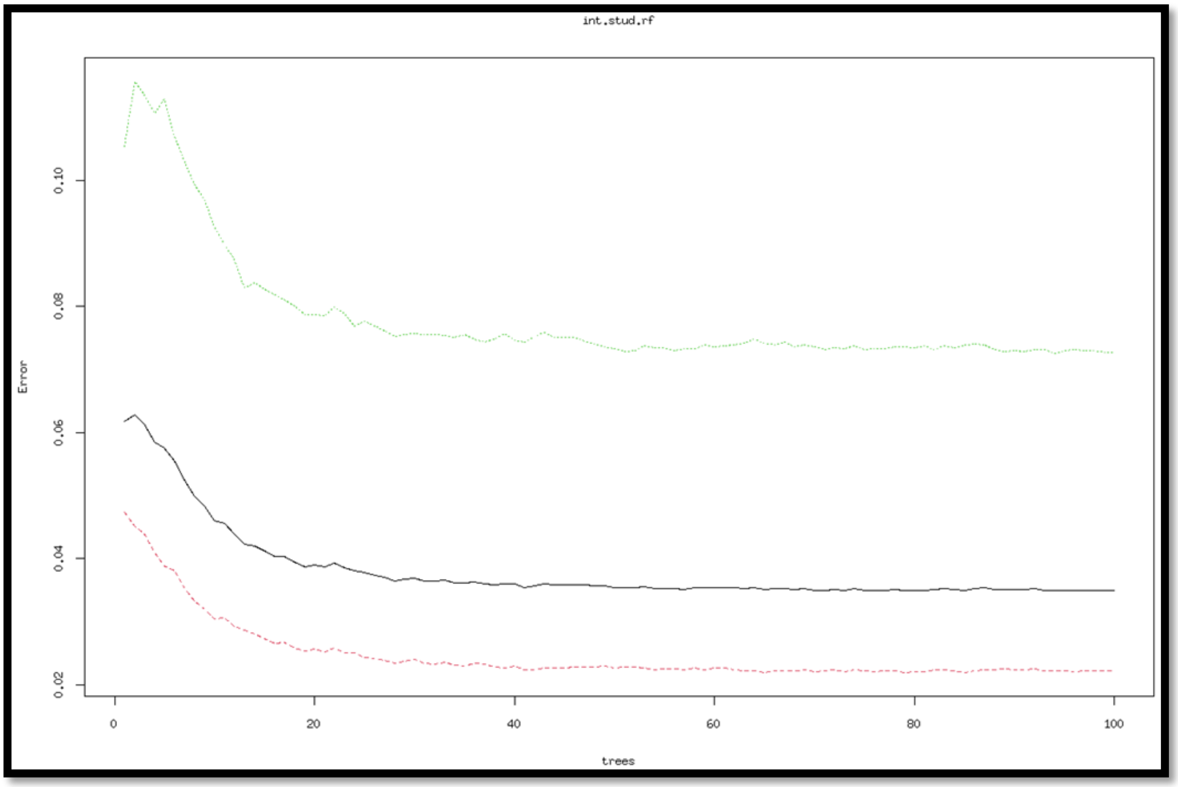


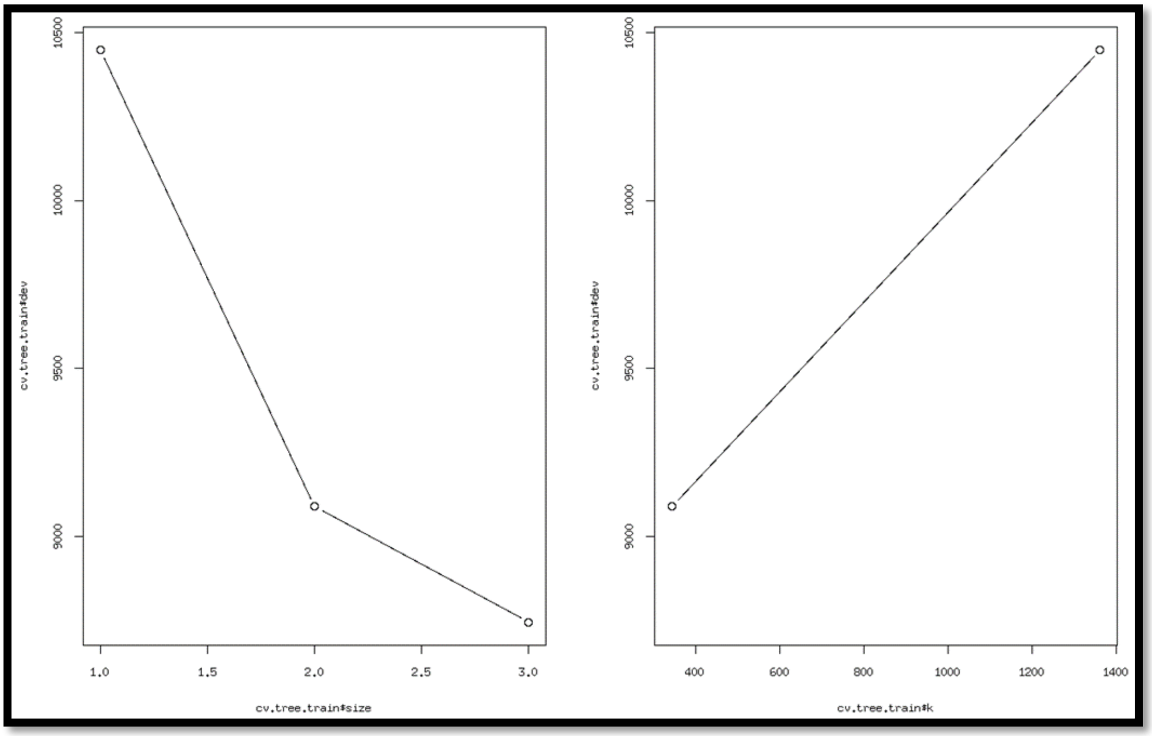
Figure 7 is a plot which is intended to assess the number of trees necessary within the random forest model to produce predictions with the lowest possible error rate. Generally, a model would be limited to the minimum number of trees necessary to produce strong predictions to reduce the amount of time and resources consumed during the model's running process. In this case, random forest predictions of student graduation steadily improve to approximately thirty trees within the forest and smaller improvement until around forty trees when prediction no longer improves by adding more trees. At forty trees, the model settles to a relatively strong error rate of approximately 3.5% which is supported by the out-of-bag error rate estimates shown in Figure 12.

Figure 7 – 25 Variable Assessment of Forest Size – Random Forest



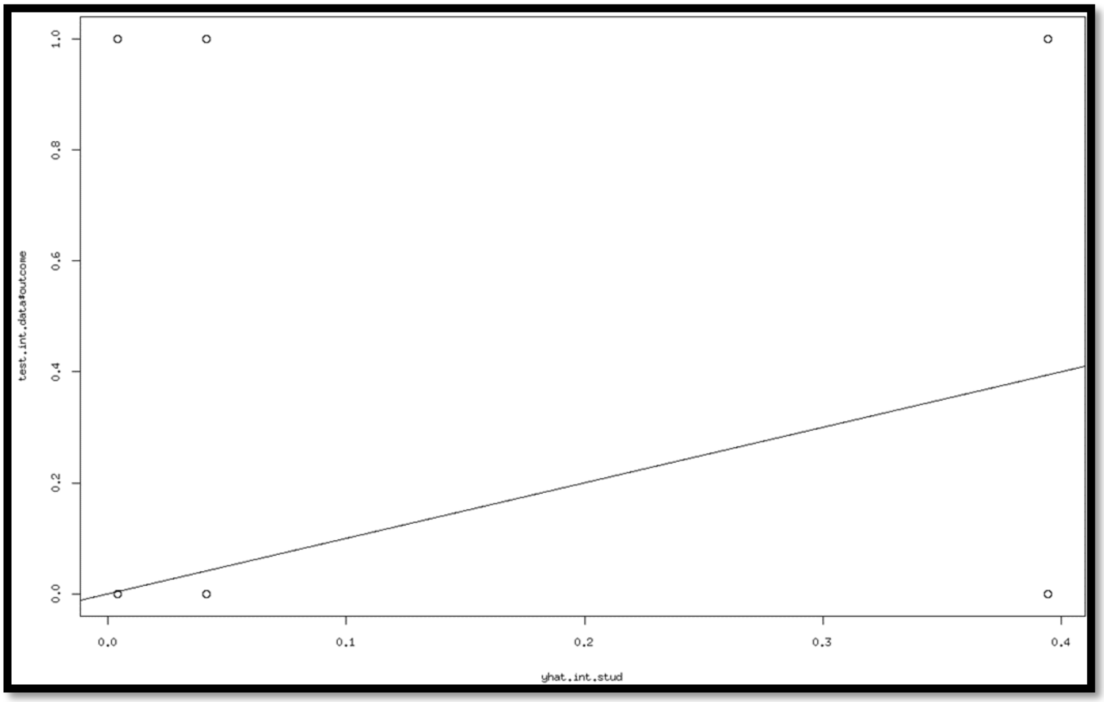
Similar to the single tree model described previously, the random forest cross-validation plots in Figure 8 indicate that trees within the model produce the best predictions with relatively few splits with only three in this case. This is slightly smaller than the single tree model, and again, the deviance is generally lower when the k number of folds used in the cross-validation are lower to ensure that the test data sets are sufficiently large to make accurate predictions.

Figure 8 – 25 Variable Random Forest Cross Validation Assessment



The random forest prediction plot shown in Figure 9 is not as useful as might be expected because the outcome variable is binary, and the model was designed for classification rather than producing an estimate of a continuous variable. That said, it is clear from the plot that true graduation outcomes (y-axis), and the graduation predictions (x-axis) were progressing through the averages at approximately the same rate indicating that the predictions were relatively accurate.

Figure 9 – Random Forest Prediction Plot



The confusion matrix outcome in Figure 10 shows that the random forest model was accurate in its predictions more than 96% of the time. In addition, the model was relatively consistent in its ability to predict both graduations and non-graduations. Sensitivity, or the true positives in this matrix, represented those students that did not graduation and predictions were accurate more than 97% of the time. For true negatives, or specificity, which represented students that did graduate, this model experienced a better than 92% accuracy rate.

Figure 10-1 – Random Forest Confusion Matrix and R Output

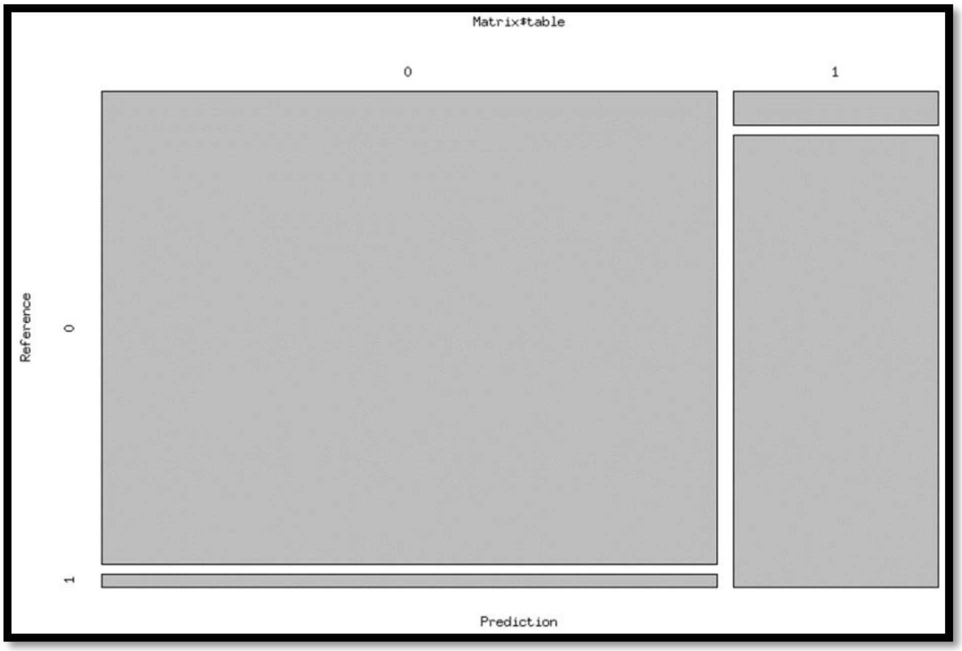
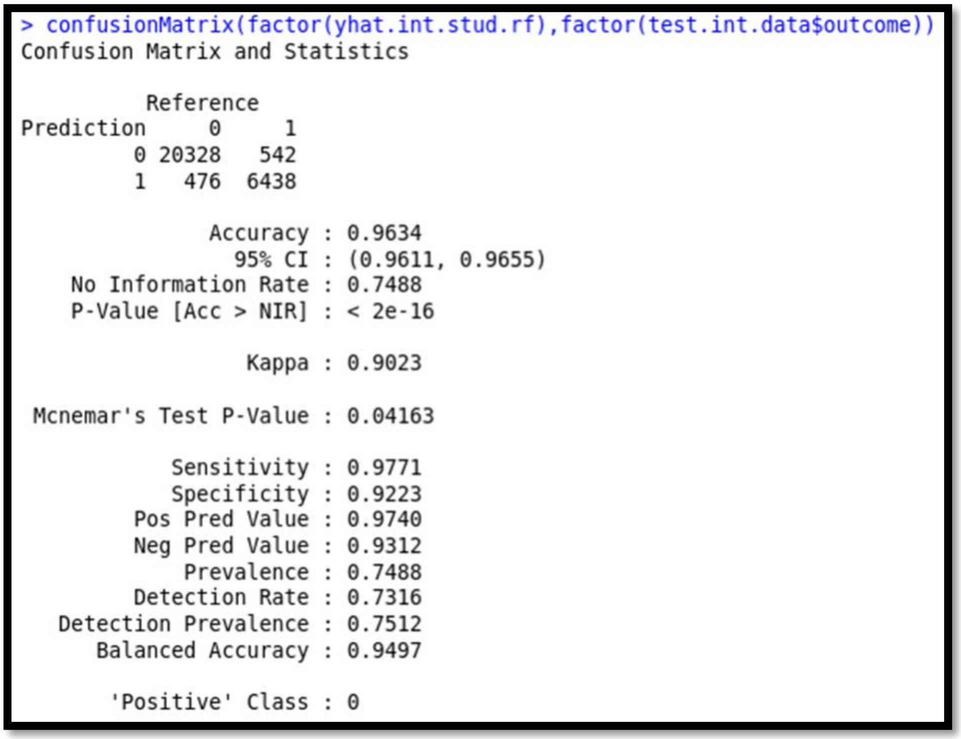


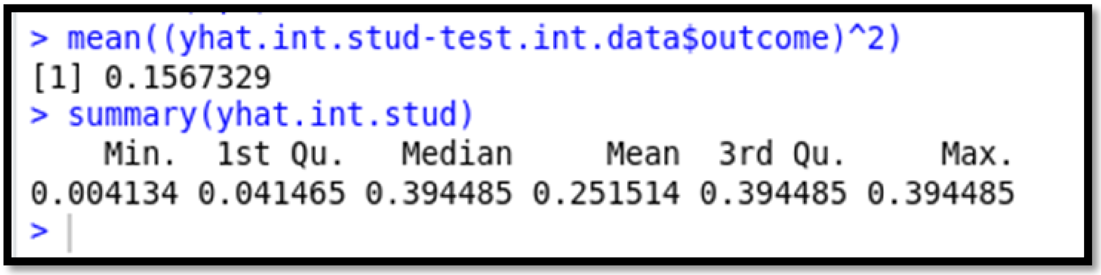
Figure 10-2 – Random Forest Confusion Matrix and R Output



```
> confusionMatrix(factor(yhat.int.stud.rf),factor(test.int.data$outcome))
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 20328   542
         1   476  6438

               Accuracy : 0.9634
                 95% CI : (0.9611, 0.9655)
    No Information Rate : 0.7488
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9023

 Mcnemar's Test P-Value : 0.04163

            Sensitivity : 0.9771
            Specificity : 0.9223
         Pos Pred Value : 0.9740
         Neg Pred Value : 0.9312
             Prevalence : 0.7488
         Detection Rate : 0.7316
   Detection Prevalence : 0.7512
      Balanced Accuracy : 0.9497

       'Positive' Class : 0
```

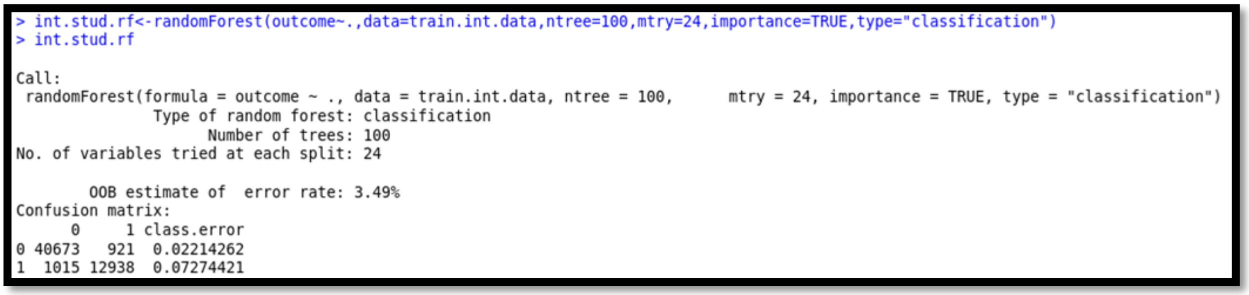
Additional model validations were completed, and summaries of the mean squared error results and the out-of-bag error estimates are provided below in Figure 11 and Figure 12 respectively. Although there is no perfect assessment of the mean squared error (MSE) metric, the 0.15 measure was considered a good result because it was relatively low. This is particularly important since MSE penalizes large individual errors that can make this metric grow quickly with numerous large prediction errors. the MSE is also assisted by the use of a classification model which does not allow for misestimations that are large enough to cause significant MSE penalties.

Figure 11 – Random Forest Mean Squared Error



```
> mean((yhat.int.stud-test.int.data$outcome)^2)
[1] 0.1567329
> summary(yhat.int.stud)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.004134 0.041465 0.394485 0.251514 0.394485 0.394485 
> 
```

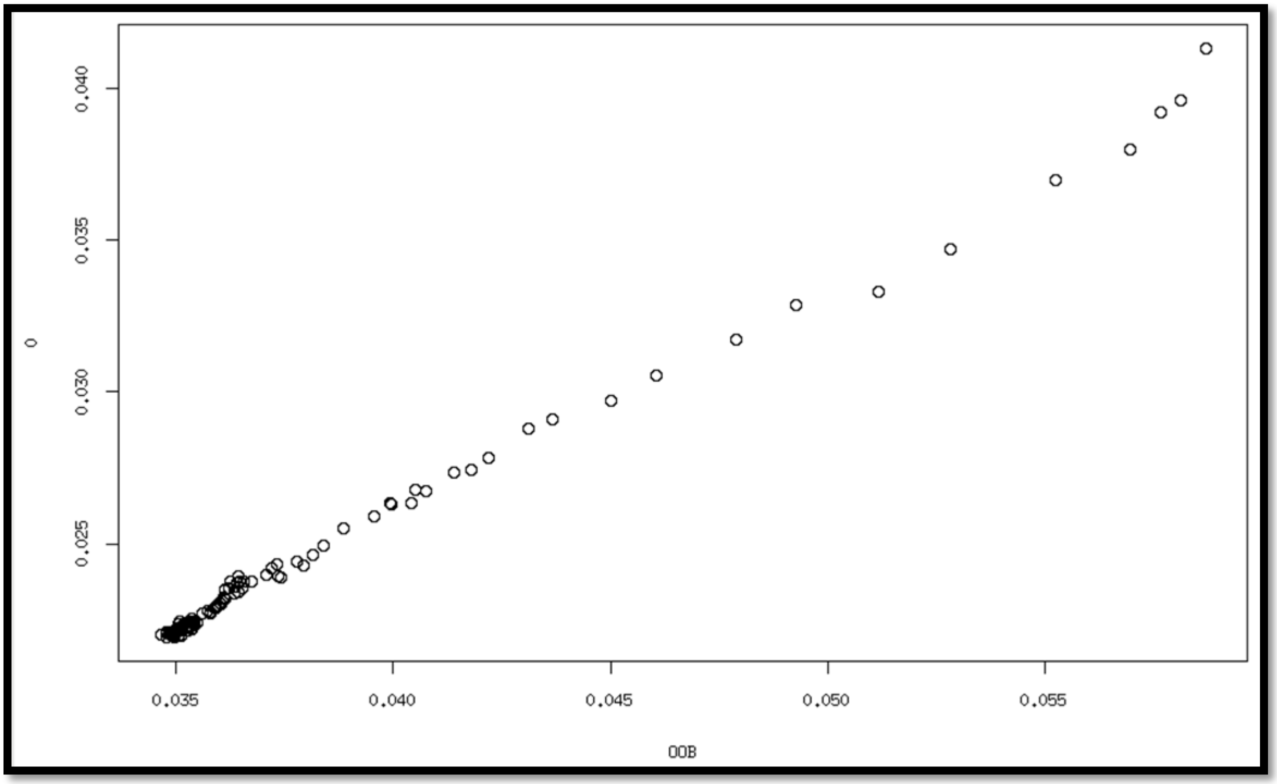
Figure 12 – Random Forest Out-of-Bag Error Rate



```
> int.stud.rf<-randomForest(outcome~.,data=train.int.data,ntree=100,mtry=24,importance=TRUE,type="classification")
> int.stud.rf

Call:
 randomForest(formula = outcome ~ ., data = train.int.data, ntree = 100,      mtry = 24, importance = TRUE, type = "classification") 
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 24

        OOB estimate of  error rate: 3.49%
Confusion matrix:
      0     1 class.error
0 40673   921  0.02214262
1  1015 12938  0.07274421
```

Out-of-bag (OOB) estimates are based on the training data set and showed low OOB error estimates with significant clustering around a 3.49% error estimate. Figure 13 provides a depiction of the OOB errors.

Figure 13 – Random Forest Out-of-Bag Error Plot



Conclusion

International student graduation rates had many of the same predictors that would be expected of all college students. For example, age, number of semester credit hours taken, and cost all played significant parts in the predictive power of the final random forest model. There were also several variables that strongly supported the predictive power of the model but did not have purely intuitive reasons for doing so. Variables such as selected major, sex, and school all demonstrated their importance to the model, but additional causal research would be useful in understanding the underlying reasons these variables matter to international student graduation rates.

Regardless of the potential for future research questions, the final random forest model was a reliable predictor of international student graduations when large testing data sets were available. For future development, incorporation of more variables that have causal and theoretical support for predicting international student college success would further improve this model's predictive power and add to its ability to be used in the assessment of appropriate interventions to support more international students persisting in college through to completion of a bachelor's degree.

References

Alsakran, R. I. 2018. "Differences in Persistence and Graduation Rates by the Institutional Status of International Students in Texas Community Colleges: A Multiyear, Statewide Study." Doctoral Dissertation.

Alsakran, R., and J. R. Slate. 2017. "A Descriptive Study of Persistence Rates for International Students in Texas Community Colleges." *AASCIT Journal of Education* 3 (3): 16–21.

Kwai, C. K. 2010. "Model of International Student Persistence: Factors Influencing Retention of International Undergraduate Students at Two Public Statewide Four-Year University Systems." Doctoral Dissertation, Dissertation Publishing, Ann Arbor.

Arellano, L. 2019. "Capitalizing Baccalaureate Degree Attainment: Identifying Student and Institution Level Characteristics that Ensure Success for Latinxs." *The Journal of Higher Education* 91. https://doi.org/10.1080/00221546.2019.1669119.

Arizmendi, C. J., M. L. Bernacki, M. Raković, R. D. Plumley, C. J. Urban, A. T. Panter, J. A. Greene, and K. M. Gates. 2023. "Predicting Student Outcomes Using Digital Logs of Learning Behaviors: Review, Current Standards, and Suggestions for Future Work." *Behavior Research Methods* 55 (6): 3026–54. https://doi.org/10.3758/s13428-022-01939-9.

Bird, K. A., B. L. Castleman, Z. Mabel, and Y. Song. 2021. "Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education." *AERA Open* 7: 23328584211037630. https://doi.org/10.1177/23328584211037630.

Brown, A. R., C. Morning, and C. Watkins. 2005. "Influence of African American Engineering Student Perceptions of Campus Climate on Graduation Rates." *Journal of Engineering Education* 94. https://doi.org/10.1002/j.2168-9830.2005.tb00847.x.

Flores, B., and C. K. Della Piana. 2000. "Undergraduate Student Retention Strategies for Urban Engineering Colleges." *Proceedings of the 2000 Frontiers in Education Conference* 1. https://doi.org/10.1109/fie.2000.897695.

Genschel, U., A. M. Gansemer-Topf, and J. Downey. 2018. "Overcoming Undermatching: Factors Associated with Degree Attainment for Academically Undermatched Students." *Journal of College Student Retention: Research, Theory & Practice* 22. https://doi.org/10.1177/1521025117753822.

Kovacic, Z. J. 2010. "Early Prediction of Student Success: Mining Students Enrolment Data."

Oseguera, L. 2005. "Four and Six-Year Baccalaureate Degree Completion by Institutional Characteristics and Racial/Ethnic Groups." *Journal of College Student Retention: Research, Theory & Practice* 7. https://doi.org/10.2190/e1tu-aw8j-5fya-glpw.

R Code for EPPS 6323 ERC Project 178

```
rm(list=ls())

library(haven)
# FAD21<-read_dta(DATA)

library(dplyr)
## head(FAD21)

# FAD21_Int<-FAD21 %>% filter(res_flag==2&classify==1)
# table(FAD21_Int$classify)
# table(FAD21_Int$proglvl)

# Rep9_21<-read_dta(DATA)
 # workingdf<-read_dta(DATA)
 # workingdf2<-read_dta(DATA)
 # workingdf3<-read_dta(DATA)
 # workingdf4<-read_dta(DATA)
# table(workingdf4$type)
# table(workingdf4$ethinter)

## Load data and subset strong variables in a subset
##################################################
rm(list=ls())
int_stud_data<-read_dta(DATA)

## 42 variable data
int_stud_data<-
subset(int_stud_data,select=c('outcome','age','sex','ethasian','ethblack','ethhispa','major','ethwhite',

'course','crs_point','crs_grd','begin_dt','rank','tenure','livarge','momed','daded','costatt','tfamcont',

'pell','atp_fhr','atp_sphr','tot_cumhr','nofund_c','nofund_d','nofund_i','nondeg','nondis','res','sch_c',
'sch_d',

'sch_dual','sch_grs','sch_on','sch_ug','schcode','school','semester','totchrs','tutstat','type','uglimit'))

## 25 varable data
int_stud_data<-
subset(int_stud_data,select=c('outcome','age','sex','ethasian','ethblack','ethhispa','major','ethwhite',
        'nofund_c','nofund_d','nofund_i','nondeg','nondis','res','sch_c','sch_d',
        'sch_dual','sch_grs','sch_on','sch_ug','schcode','school','totchrs','tutstat','uglimit'))
# add variables - nofund_c, nofund_d, nofund_i, nondeg, nondis, res, sch_c, sch_d, sch_dual,
sch_grs, sch_on,
#        sch_ug,schcode, school, semester, totchrs, tutstat, type, uglimit,
 # int_stud_data<-
subset(int_stud_data,select=c('outcome','age','sex','ethasian','ethblack','ethhispa','major',
#                'course','rank','tenure'))

library(ggplot2)
library(tidyverse)
library(tidyr)
library(gtsummary)

library(tree)
Int_tree<-tree(outcome~.,data=int_stud_data,control=tree.control(nobs=83331,mincut=8))
plot(Int_tree)
text(Int_tree,pretty=1)

 ## Load packages and run classification tree models
############################################################
library(rpart);library(rpart.plot);library(Cubist)

intstud_train<-int_stud_data[1:50000,]
intstud_test<-int_stud_data[50001:83331,]

set.seed(2)
train<-sample(1:nrow(int_stud_data),50000)
Int_train_tree<-tree(outcome~.,int_stud_data,subset=train)
Int_tree_predict<-predict(Int_train_tree,int_stud_data[-train,])
addmargins(table(Int_tree_predict,int_stud_data[-train,'outcome']))

help("cv.tree")
cv_int_tree<-cv.tree(Int_train_tree,K=10)

 par(mfrow=c(1,2))
plot(cv_int_tree$size,cv_int_tree$dev,type="b")
plot(cv_int_tree$k,cv_int_tree$dev,type="b")
par(mfrow=c(1,1))    # reset plot window

## Random forest model
############################################################
rm(list=ls())
library(haven)
int_stud_data<-read_dta(DATA)

## 42 variable data
int_stud_data42<-
subset(int_stud_data,select=c('outcome','age','sex','ethasian','ethblack','ethhispa','major','ethwhite',
'course','crs_point','crs_grd','begin_dt','rank','tenure','livarge','momed','daded','costatt','tfamcont',

'pell','atp_fhr','atp_sphr','tot_cumhr','nofund_c','nofund_d','nofund_i','nondeg','nondis','res','sch_c',
'sch_d',

'sch_dual','sch_grs','sch_on','sch_ug','schcode','school','semester','totchrs','tutstat','type','uglimit'))

## 25 varable data
int_stud_data25<-
subset(int_stud_data,select=c('outcome','age','sex','ethasian','ethblack','ethhispa','major','ethwhite',
       'nofund_c','nofund_d','nofund_i','nondeg','nondis','res','sch_c','sch_d',
       'sch_dual','sch_grs','sch_on','sch_ug','schcode','school','totchrs','tutstat','uglimit'))

library(tree)
library(ISLR2)
library(pROC)
library(caret)

set.seed(312)
split.int.data<-rsample::initial_split(int_stud_data25,prop=0.6666,strata="outcome")
train.int.data<-rsample::training(split.int.data)
test.int.data<-rsample::testing(split.int.data)

train.tree<-tree(outcome~.,data=train.int.data)
test.tree<-tree(outcome~.,data=test.int.data)

par(mfrow=c(1,1))    # reset plot window
plot(train.tree)
text(train.tree,pretty=1)

# cross validation of tree model
help(tree)
set.seed(7)
cv.tree.train<-cv.tree(train.tree,FUN=prune.tree,K=10)

par(mfrow=c(1,2))
plot(cv.tree.train$size,cv.tree.train$dev,type="b")
plot(cv.tree.train$k,cv.tree.train$dev,type="b")
par(mfrow=c(1,1))

# Evaluated pruned tree
pruned.train.tree<-prune.tree(train.tree,best=3)
plot(pruned.train.tree)
text(pruned.train.tree,pretty=1)
```

```
# Perform prediction
yhat.int.stud<-predict(train.tree,newdata=test.int.data)
plot(yhat.int.stud,test.int.data$outcome)
abline(0,1)
mean((yhat.int.stud-test.int.data$outcome)^2)  ## Mean squared error MSE
summary(yhat.int.stud)

# random forest specific model - # retry this model with the 42 variable data ###
library(randomForest)
library(ROCR)

 train.int.data$outcome<-as.factor(train.int.data$outcome)
# warnings()

set.seed(3)
int.stud.rf<-
randomForest(outcome~.,data=train.int.data,ntree=100,mtry=24,importance=TRUE,type="classification")
int.stud.rf

importance(int.stud.rf,type=1)
varImpPlot(int.stud.rf,type=1) # plot important factors

plot(int.stud.rf)

 yhat.int.stud.rf<-predict(int.stud.rf,newdata=test.int.data)
plot(yhat.int.stud.rf,test.int.data$outcome)

 library(ggplot2)
ggplot()
plot(int.stud.rf$predicted)
plot(int.stud.rf$err.rate)
plot(int.stud.rf$votes)
plot(int.stud.rf$y)

summary(int.stud.rf$y)
int_stud_data25$outcome<-as.factor(int_stud_data25$outcome)
summary(int_stud_data25$outcome)

confusionMatrix(factor(yhat.int.stud.rf),factor(test.int.data$outcome))
Matrix<-confusionMatrix(factor(yhat.int.stud.rf),factor(test.int.data$outcome))
plot(Matrix$table)
```