Machine Learning Models for International Student Success in Texas Higher Education

EPPS 6323 – Knowledge Mining

Sonali Singh - Maitreyi Pillai - Greg Argueta

Overview and Motivation

Rapid Growth: The international student population in U.S. postsecondary institutions grew by over 84% from 2006 to 2017.

Challenges for International Students: Face higher challenges in persistence and graduation compared to domestic students due to factors like socioeconomic backgrounds, academic preparedness, and integration into campus life. Evidence for different drop-out rates for domestic and international students (Alsakran 2018)

Need for Advanced Data Analytics: There is a highlighted need for advanced predictive analytics to better understand and improve educational outcomes for this group. Previous scholarship attempts to measure why students did not continue with their education and some interesting work to measure international drop-out rates, but none using advanced quantitative approaches.

Existing Data and Models: While there is descriptive data and theoretical models accounting for factors such as country of origin, financial support, and academic metrics (e.g., GPA), these have not been sufficient. (Kwai 2010)

Gaps in Prediction and Support: There remains a significant gap in accurately predicting and improving graduation rates of international students.

Exploration of New Methods: Traditional models fail to capture the complexities affecting these students' success, prompting the exploration of machine learning techniques for more nuanced insights and predictions.

Research Question and Data

- **Research Question:** "How do student characteristics and institutional factors influence the degree attainment of undergraduate international students in Texas?"
- **Purpose**: To analyze and understand the complex interplay of personal and environmental factors affecting international students' academic outcomes, aiming to enhance accuracy in predicting their degree attainment.

Data:

- **Source**: Data collected by the **Texas Higher Education Coordinating Board (THECB)**, covering all postsecondary institutions in Texas.
- **Focus**: Specifically targets international students in undergraduate programs.

Components of the Data Set:

- **Demographic Information:** Includes basic student demographics.
- **Academic Records:** Details on degrees awarded, course schedules.
- **Faculty Information:** Data related to the faculty involved in student courses.
- **Scope of Data:** Analysis involves over 83,000 records focusing on the graduation outcomes for the 2021 school year, supplemented with data from the preceding six years to adhere to standard graduation timelines.
- **Methodology:** Utilizes descriptive analysis and machine learning models, such as random forests, to evaluate various predictors of graduation outcomes. Key variables analyzed include age, academic load, ethnicity, and socioeconomic status.

Data Utilization: Texas Schools Project – Education Research Center:

- **Data Integration**: Merged enrollment data with degree attainment and financial aid records to create a foundational data set.
- **Enhanced Data Set:** Further enriched by merging course and faculty information spanning the previous six years (12 semesters).
- **Final Data Set Details:** Resulted in 83,331 observations with 197 variables after filtering for international status and bachelor's program enrollment.
- **Outcome Variable:** A binary variable was developed to signify whether a degree was attained, serving as the primary outcome variable for the analysis.

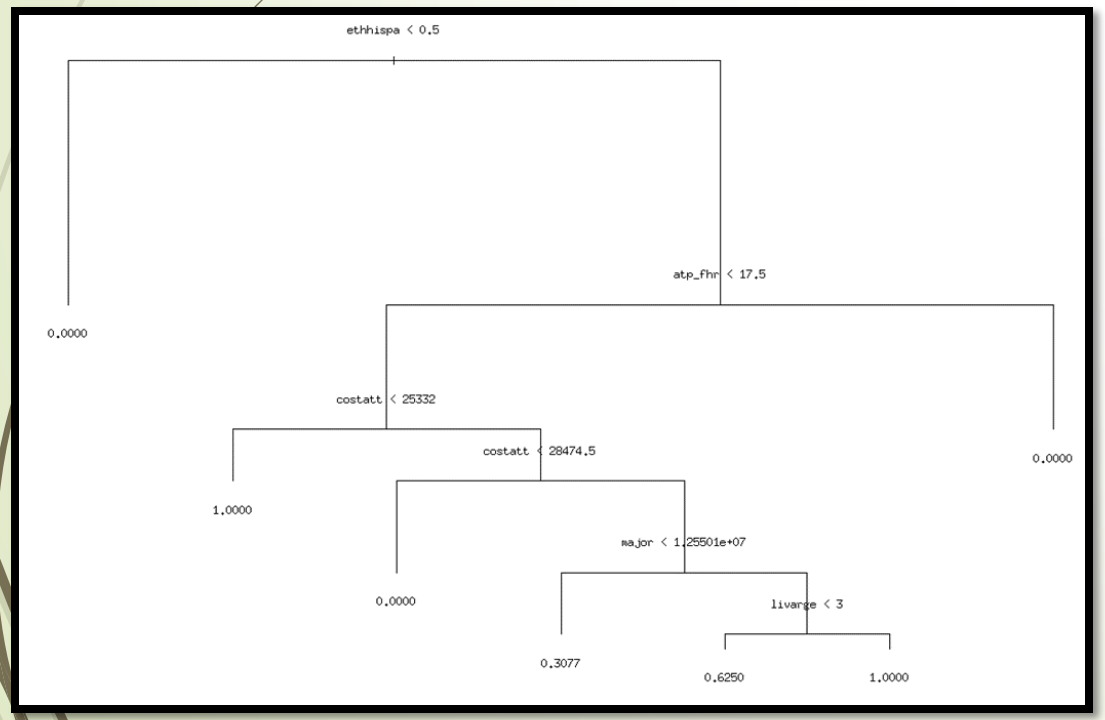
Methods and Results

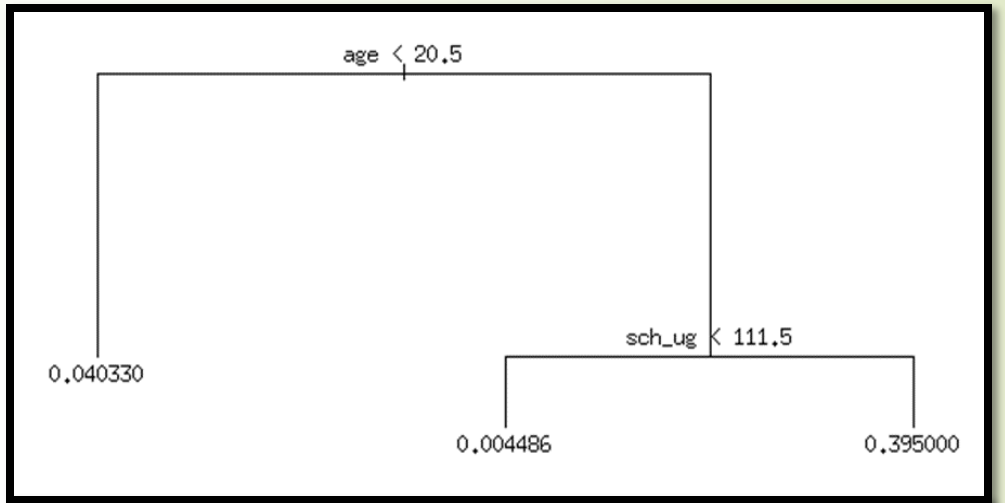
Variable	Obs	Mean	Std. dev.	Min	Max
outcome	83,331	.251203	.4337077	0	1
age	83,331	21.94393	3.435927	14	56
sex	83,331	.5763401	.4941408	0	1
ethasian	83,331	.1677527	.3736488	0	1
ethblack	83,331	.0414132	.199245	0	1
ethhispa	83,331	2.476605	1.093148	0	3
ethwhite	83,331	.0438012	.2046538	0	1
major	83,331	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

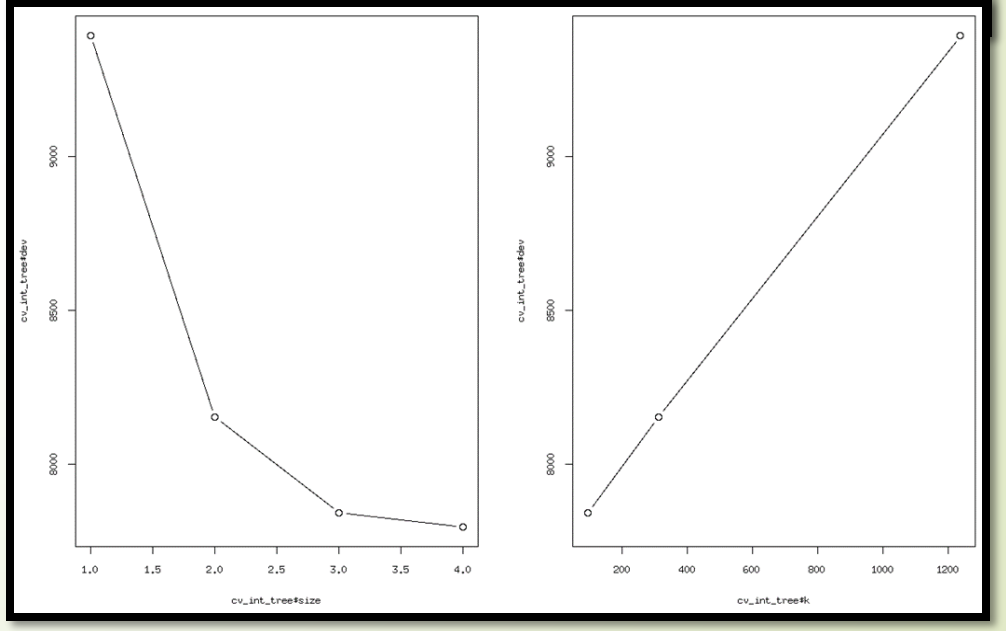
Variable	Obs	Mean	Std. dev.	Min	Max
atp_sphr	5,180	14.01873	3.481814	0	21
tot_cumhr	5,180	55.89486	33.73398	.6	256
nofund_c	83,331	.3437616	1.867849	0	23
nofund_d	83,331	.000612	.033759	0	3
nofund_i	83,331	.0107523	.3171631	0	16
nondeg	83,331	.0058802	.0764569	0	1
nondis	83,331	.1527163	.5311437	0	2
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

Methods and Results – Trees

- Focus on classification models
- Tree models helped identify major variables that split completers and non-completers – 42 variable model (left) were reduced to 25 (right) due to significant missingness

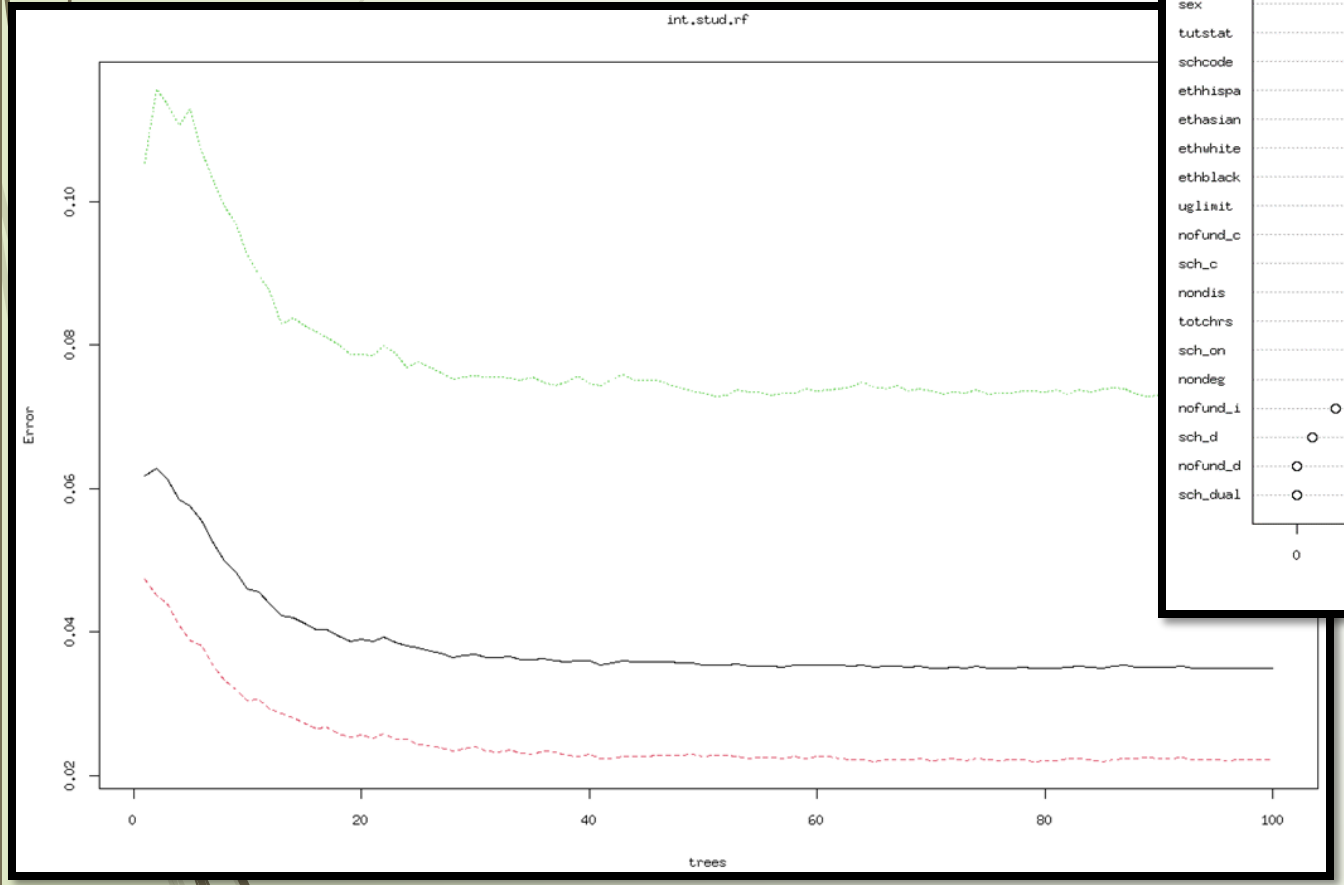


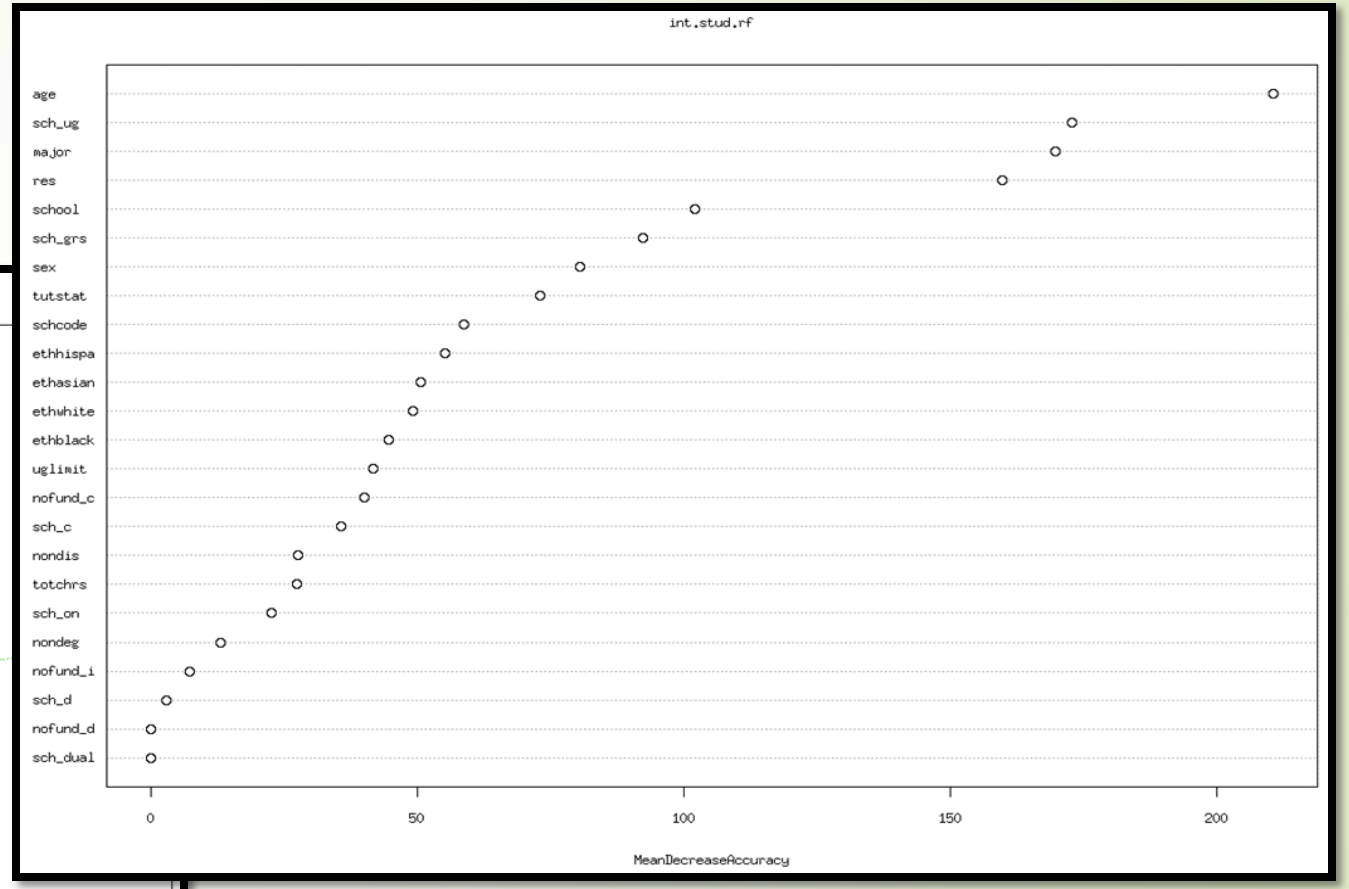




Methods and Results - Forests

- Graph depicts the optimum number of trees in the forest



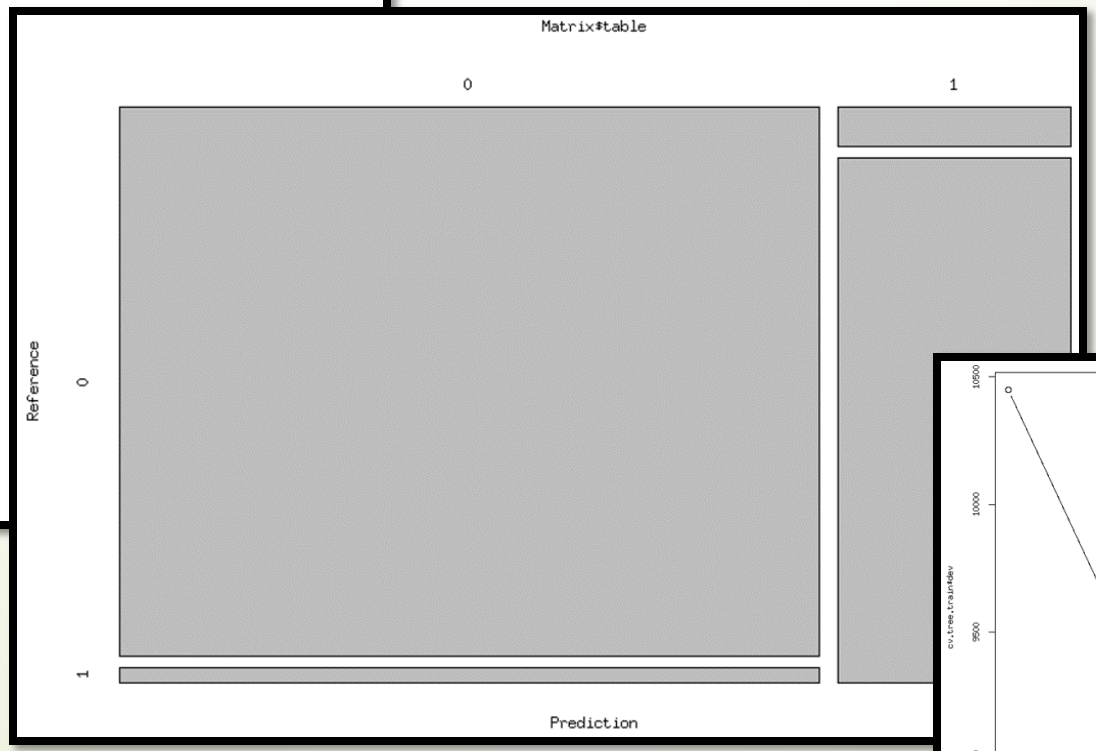


- Graph illustrates the importance of model features ordered by predictive power

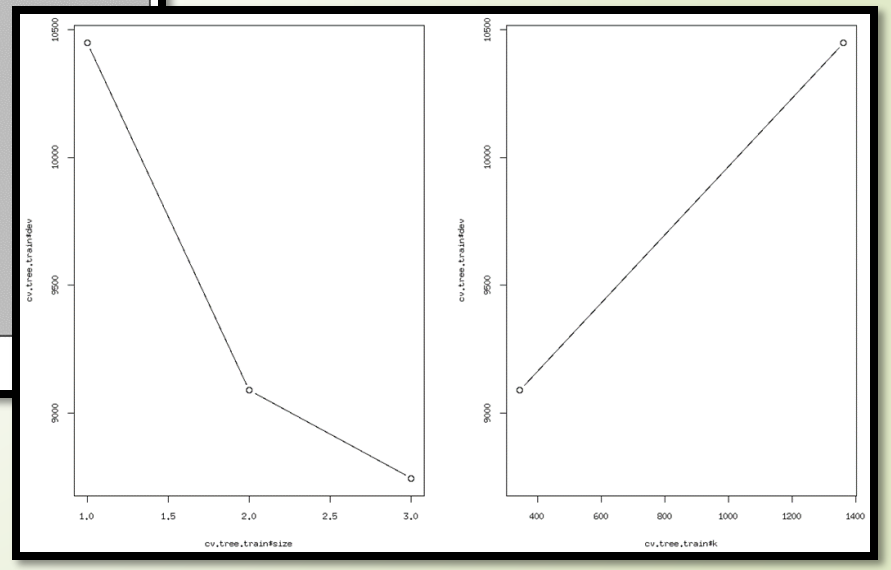
Methods and Results - Forests

```
> 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
```

- The model demonstrated strong accuracy at over 96%
  - Specificity (true negative) 92% and Sensitivity (true positive) 97%



- Predictive power against the testing data set is provided by the confusion matrix



- The cross-validation assessment indicates that the size of the trees within the forest minimize deviance at three splits and deviance increases as the test data set gets smaller

Conclusion

- Single tree models were useful when data was complete
- Random forest predictions were strong when data was complete, and the testing data set is larger
- Although the data was a good predictor, the variables selected would not be useful for causal applications
- Access to additional data files could have improved predictions on smaller testing data sets

POSTSECONDARY SUCCESS FOR INTERNATIONAL STUDENTS

EPPS 6323 – KNOWLEDGE MINING

GREG ARGUETA - MAITREYI PILLAI - SONALI SINGH

OVERVIEW

- Evidence for different drop-out rates for domestic and Int'l students (Alsakran 2018)
- Pre-existing student characteristics - finance, race, sex, abroad experiences (Kwai 2010)
- Previous scholarship attempts to measure why students did not continue with their education and some interesting work to measure international drop-out rates, but none using advanced quantitative approaches.
- Contribution - introduce novel variables that isolate the drop-out rates of international students compared to domestic students using advances predictive statistical models.
- EX. English language – pre-existing or language competency (TOEFL scores)

RESEARCH QUESTION

How do student characteristics and institutional factors influence the degree attainment of undergraduate international students in Texas?

COMMON THEMES IN LITERATURE

- Descriptive
- Narrow populations
- Cover momentum similar to non-international students
- Transferring between institutions

CONTRIBUTION TO THE FIELD

- Machine learning methods not found
- Large datasets studies not found
- The amount of data available in Texas is unique

METHODS

- Outcome variable – bachelor's degree completed
- Focus on classification models
- Tree models will help identify major variables that split completers and non-completers
- Forest models should help to improve outcome predictions
- May look at clustering models to look for student or institution similarities in characteristics

DATA AVAILABILITY

- Texas Schools Project – Education Research Center
- Utilize micro-data panel from a current project
  - K-12 administrative and data
  - College administrative and course-level data
  - Work and income-level data for assessment of labor market outcomes
- All data is from Texas
  - The large population, diverse economy, and number of colleges make the state a reasonable representation of the US

POLICY IMPLICATIONS

- A better understand of decision points that contribute to a student completing a bachelor's degree will improve college administrator's ability to develop interventions
- Better predictions of completion will allow student support professionals to take action to reduce the number drop-outs
- A deeper knowledge of the factors influencing enrollment and attendance may also contribute to re-enrolling students that previously left the institution

CONCLUSION

- International students face unique challenges
- Existing literature has been descriptive with focus on common educational variables
- Machine learning methods may illuminate previously omitted predictors
- The large-variable data set available in Texas will be a strong basis for the study
- Policymakers and college administrators should benefit from a better understanding of international student needs