Methodological approach:

This section of the report describes the method for predicting graduation outcomes of Georgia Tech’s first-time undergraduate students. A quantitative study of student characteristics aims to produce generalizable knowledge of which factors contribute to graduation and predicting student outcomes.

Cohort 2011 – 2013 data from 8,394 students is a representative sample of current and future student populations. The predictive models use this data. These cohort years affectively provide data of students that have a time frame of six or more years to graduate.

In an attempt to classify which students are likely to graduate and which are not, several approaches were undertaken. Combining two classification approaches, logistic regression and random forest, produced a predictive model and insight into which factors have a significant impact on graduation.

Here is a list of the student factors studied:

* Cumulative GPA
* Transfer credit hours
* Ratio of Georgia Tech credit hours attempted to hours earned
* Number of changes in major of study
* College of study
* Residence
* Race/Ethnicity
* Sex
* Fraternity/Sorority participation
* Internship participation
* Cooperative participation
* Study abroad
* Federal Pell Grant
* HOPE Scholarship
* ZELL Miller Scholarship
* NCAA Athlete status

Together the classification methods provide more insight than a single model would for the Cohort 2011 –2013 data. The features selected for this study show less variation among the students who graduated compared with those who did not (see Table 4). This situation resulted in linearly separable data for which an infinite number of logistic regression models would technically differentiate students who are likely to graduate from those who do not (see Figure 6). Selecting anyone of these logistic models is likely to result in overestimation of the dominant graduating class.

There are large differences in the GPA’s between students who graduate and those who do not.

Table 4: Average Performance Metrics by Graduation Status

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Graduation Status** | **Cumulative GPA** | **Transfer Credit Hours** | **Ratio of GT Credits Attempted to GT Credits Earned** | **No. of Major Changes** |
| No | 2.23 | 4 | 78 | 1.4 |
| Yes | 3.33 | 7 | 128 | 1.0 |

The average GPA of graduating students is higher than students who did not graduate at 3.33 verses 2.23. This clear delineation between the classes of graduates and non-graduates poses a challenge for using logistic regression modeling for more than significant factor determination.

Graduates

Non-Graduates

Figure 6: Example of linearly separable classes

for graduating vs non-graduating students

In addition, before proceeding with an effective model determination the imbalanced proportion for the students that graduated compared with those who did not, required correction (see Figure 7). Fewer than 11% (918 out of 8,394) of students from the 2011 – 2013 cohorts did not graduate. Proportion difference larger than 70% require corrective action for use with many classification models.

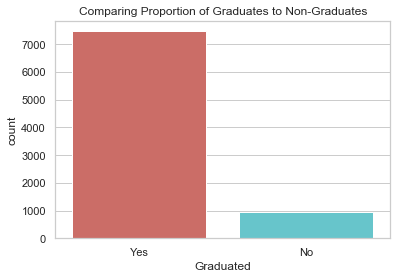


Figure 7: Comparing Proportion of Graduates to Non-Graduates

Correcting the imbalance required randomized oversampling of the students who did not graduate. This oversampling occurred on 70% of the original 8,394 records. Reserving the remaining 30% of the Cohort 2011 – 2013 data allowed for testing the random forest model’s predictive accuracy on the actual unbalanced data. The random forest classifier has the benefit of not being as sensitive to the imbalance and linear separable data as the logistic regression method. However, unlike linear regression the method’s model coefficients are not transparent.

Findings

Table 5: Logistic Regression Factor Significance Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Estimate** | **Odds Ratio** | **P Value** | **Notes** |
| (Intercept) | -8.759815 | 0.00 | P < 0.001 |  |
| Cumulative GPA | -0.990447 | 0.39 | P < 0.001 | Higher GPA is associated with increased chance for graduating |
| Transfer Credit Hours | -0.011927 | 0.98 | 0.07698 |  |
| GT Credits Attempted/Credits Earned | 9.92938 | 20910.86 | P < 0.001 | \*A higher ratio of credits attempted to credits earned indicates that the student is passing fewer courses. This suggests a one unit increase in the ratio (not practically plausible) will adversely affect the odds of graduating by 20910.86 |
| Number of Major Changes | -0.599449 | 0.61 | P < 0.001 | More likely to graduate than a student who does not change majors keeping other characteristics the same – \*Students who switch majors may be opting for programs that are easier or take less time to complete. |
| College Computing | -0.399063 | 0.46 | 0.108975 |  |
| College Design | 1.470716 | 1.08 | P < 0.001 | Less likely to graduate than a College of Engineering student keeping other characteristics the same |
| College Ivan Allen | 1.162309 | 1.47 | P < 0.001 | Less likely to graduate than a College of Engineering student keeping other characteristics the same |
| College Scheller | 0.629541 | 1.08 | 0.040371 | Less likely to graduate than a College of Engineering student keeping other characteristics the same |
| College Sciences | 1.008871 | 1.56 | P < 0.001 | Less likely to graduate than a College of Engineering student keeping other characteristics the same |
| Citizen Alien | -1.062697 | 0.55 | P < 0.01 | More likely to graduate than an American citizen keeping other characteristics the same |
| Citizen Resident Alien | 0.051302 | 1.54 | 0.87751 |  |
| Residence International | NA | NA | NA |  |
| Residence Out of State | -0.277311 | 0.89 | 0.297758 |  |
| Race/Ethnicity American Indian | -11.661949 | 13.65 | 0.983907 |  |
| Race/Ethnicity Asian | -0.525442 | 0.53 | P < 0.001 | More likely to graduate than a white student keeping other characteristics the same |
| Race/Ethnicity Black | -0.9423 | 0.71 | 0.066022 |  |
| Race/Ethnicity Hispanic | -2.118081 | 0.24 | P < 0.001 | More likely to graduate than a white student keeping other characteristics the same |
| Race/Ethnicity \_International | NA | NA | NA |  |
| Race/Ethnicity \_Pacific Islander | -12.58309 | 0.00 | 0.983781 |  |
| Race/Ethnicity Multi-racial | -0.61934 | 0.63 | 0.152948 |  |
| Race/Ethnicity Unknown | -0.321694 | 1.23 | 0.704854 |  |
| Under Represented Minority (Non-White or Non-Asian) | 0.843337 | 1.54 | 0.065558 |  |
| Sex Female | -0.239627 | 0.91 | 0.104443 |  |
| Fraternity/Sorority Engagement | -0.464265 | 0.48 | 0.049293 | More likely to graduate than a student not in a fraternity/sorority keeping other characteristics the same |
| Internship | -8.899287 | 0.00 | 0.991956 |  |
| Study Abroad | -1.339968 | 0.21 | P < 0.001 | More likely to graduate than student without study abroad keeping other characteristics the same |
| Federal PELL Grant | 0.252831 | 1.12 | 0.099204 |  |
| HOPE Scholarship | -0.933292 | 0.53 | P < 0.001 | More likely to graduate than student without a Hope Scholarship keeping other characteristics the same |
| Zell Scholarship | -0.627368 | 0.68 | P < 0.01 | More likely to graduate than student without a Zell Scholarship keeping other characteristics the same |
| NCAA Athlete | 0.100569 | 1.18 | 0.725748 |  |

P-values less than 0.05 indicate that the feature is a significant indicator of graduation. Variables with p-values greater than 0.05 are not significant. The table depicts factor variables broken out by their levels with the exception of the reference level. For example, the variable sex has the levels male and female. Since there are more male students at Georgia Tech the sex level – male is the reference group for this category. The other reference groups are College of Engineering, Georgia residents, and White students for race/ethnicity. For categorical variables, odds ratios greater than 1 indicate a higher chance of not graduating for the feature over the prescribed baseline, with the reverse also being true (see Table 5). A one-unit increase in the continuous variables indicate an increase in the odds of a student not graduating by the respective odds ratio. For example, a one unit increase in cumulative GPA increases the odds of not graduating by 0.39.

In summary, the logistic regression model indicates the following factors have the largest impact on graduation:

* Cumulative GPA
* GT Credits Attempted/Credits Earned
* Number of Major Changes
* College
* Citizenship
* Race
* Fraternity/Sorority Engagement
* Study Abroad
* HOPE Scholarship
* Zell Scholarship

Random Forest Model Results:

The subsequent diagnostics are based on the random forest model’s predictive performance on the remaining 30% of original cohort 2011 – 2013 data. Although this algorithm is less sensitive to imbalanced data, a balanced data set is still better for prediction. Again 70% of the cohort data was balanced leaving the remaining 30% for testing model accuracy. Testing the model accuracy on the remaining unbalanced data provides an indication of what the predictive results will be for new student data.

Table 6: Comparing Random Forest Model Predictions to Actual Graduation Outcomes

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Actual Graduation Status** | |  |
| **Predicted Graduation Status** | **Graduated** | **Did No Graduate** | **Row Totals** |
| **Graduated** | 2,189 | 108 | 2,297 |
| **Did Not Graduate** | 17 | 171 | 188 |
| **Column Totals** | 2,206 | 279 | 2,485 |

The numbers highlighted in gold indicate correct classification by the random forest model. Approximately 99% (2,189 out of 2,206) of the students who actually graduated were also predicted to graduate by the random forest model. While the prediction results of 61% (171 out of 279) accuracy in classifying the number of students who did not graduate is lower, this result is better than randomly guessing graduation status. Overall the model correctly classifies graduation status 95% of the time.

Table 7: Overview of Random Forest Model Predictive Viability

|  |  |  |
| --- | --- | --- |
| **Random Forest Model Diagnostics** | | |
| Accuracy (Acc) | 95.0% | The overall accuracy is higher than the no information rate indicating a good fit. |
| 95% Confidence Interval | (94.0%, 95.8%) | Use of the model should yield overall classification accuracy between 94.0% and 95.8% |
| No Information Rate (NIR) | 88.8% | Rate of correct classification from random guessing alone. |
| P-Value [Acc> NIR] | p < 0.001 | Indicator of model significance |
| Kappa | 0.7057 | The Kappa Coefficient is used to determine the agreement between the model and the reference data. |
| McNemar's Test P-Value | P < 0.001 | Indicator of model significance |
| Sensitivity | 61% | The model’s ability to correctly classify non-graduates alone. |
| Specificity | 99.2% | The model’s ability to correctly classify graduates alone. |

There are several indicators that the random forest model provides a relatively good fit to the data. These indicators are:

* Accuracy
* No Information Rate
* P-Values
* Kappa Statistic

The *No Information Rate* is the chance of obtaining the correct outcome (graduate or not graduate) for a student based on chance alone (somewhat like randomly selecting the student status with a weighted coin). Given that the accuracy of 95% (and subsequent confidence interval) is higher than 88.8% of the *No information Rate* the model is good fit for making predictions about graduation status overall. P-values of less than 0.05 also indicate good model predictive ability. The rubric for determining model sufficiency is represented by a scale for the Kappa Statistic[[1]](#footnote-1).

Table 7: Rubric for Interpreting the Model Efficiency with the Kappa Statistic

|  |  |
| --- | --- |
| **Interpretation of Kappa Statistic** | |
| **Kappa** | **Agreement** |
| < 0 | Less than chance agreement |
| 0.01-0.20 | Slight agreement |
| 0.21-0.40 | Fair agreement |
| 0.41-0.60 | Moderate agreement |
| 0.61-0.80 | Substantial agreement |
| 0.81-0.99 | Almost perfect agreement |

The kappa statistic for the predicting graduation status of Georgia Tech students is 0.71. The model provides substantial predictive ability. Sensitivity indicates the model will accurately predict when a student will not graduate 61% of the time. Specificity indicates the model will accurately predict when a student will graduate 99% of the time. Although the models ability to accurately predict when students will not graduate is 66% considering that 11% of student do not graduate, the model’s prediction is much better than randomly guessing which student will not complete.

1. Table sourced from: Viera, Anthony J. and Joanne M.Garrett, PhD. Understanding Interobserver Agreement: The Kappa Statistic. Page 362. Family Medicine: Research Series. May 2005 [↑](#footnote-ref-1)