SDAL SOCIAL & DECISION ANALYTICS LABORATORY
DATA SCIENCE FOR THE PUBLIC GOOD

Virginia Tech Metrics | Degree Completion

Jay Warajuntano (VT), Doug Mattingly (VT) and Devika Nair (American University)
SDAL: Aaron Schroeder, Gizem Korkmaz and Sallie Keller
Sponsor: Ken Smith (Virginia Tech University, Vice Provost)

Background

Virginia Tech University is a land-grant university situated in Blacksburg attracting students from across the country from a diverse set of backgrounds. The Provost's office is tasked with oversight of all educational activities and student affairs. Ken Smith, the Vice Provost charged our team with exploring and identifying trends around the university's ability to attract students with particular emphasis on vulnerable populations.

As students progress through their undergraduate studies, their goal is ultimately to obtain a degree. Our goal is to understand how student characteristics correlate with student success, as defined by graduation within 6 years. This study made use of Virginia Tech's University Data Commons, the school's data lake that houses student and operational data.

Data

The project leveraged and combined several data sources to identify a number of meaningful variables that break out across the following categories:

- **Individual Background** Demographic and family characteristics
- **Academic Performance** Scholarly performance, either for applicants during their high school years or for students within their degree program
- **Financial Status** Financial obstacles to degree completion
- **Social Embeddedness** Student involvement with campus life and Blacksburg community

Methods

The study utilized logistic regression with LASSO-selected features to model enrollment decision and successful student progression behaviors. LASSO (least absolute shrinkage and selection operators) is a regression analysis method that seeks to reduce the dimensionality of the data. A second technique, binomial logistic regression, was applied to the reduced datasets' variables to better model the relationships between our responses and the independent variables and capture the accuracy of our model.

Results

LASSO highlighted the variables as having significant impact on progression. Of the original 128 variables, 21 were selected. A correlation plot was drawn against the variables to identify any potential variables with collinearity, and the resulting plot is shown in Figure 1. The plot indicates a correlation score of 0.6 between 6 year graduation and credit hours.

Logistic regression was run on these 21 variables to model graduation outcomes and the resulting model's coefficients are diagrammed below. The regression estimates indicate the number of times the probability shifts with that variable, while the bars show the standard deviation. The model's prediction showed an accuracy rate of 94%. The ROC curve below shows significant deviation from baseline.

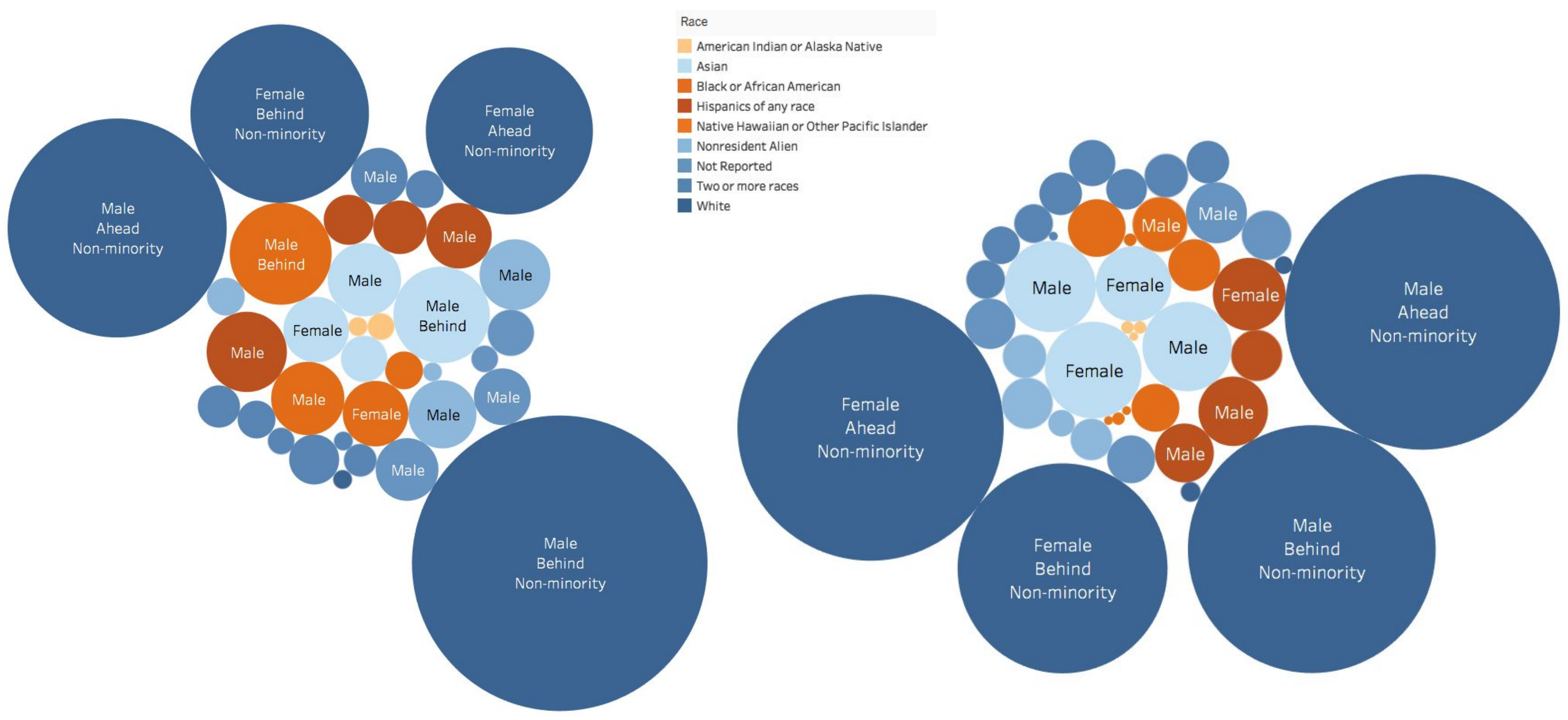

Figure 1. Student populations most (left-side) and least (right-side) likely to graduate. Size indicates number of students, color indicates race (orange hues indicate minority status), and labels indicate minority status and two year credit rate for the largest sub-populations.

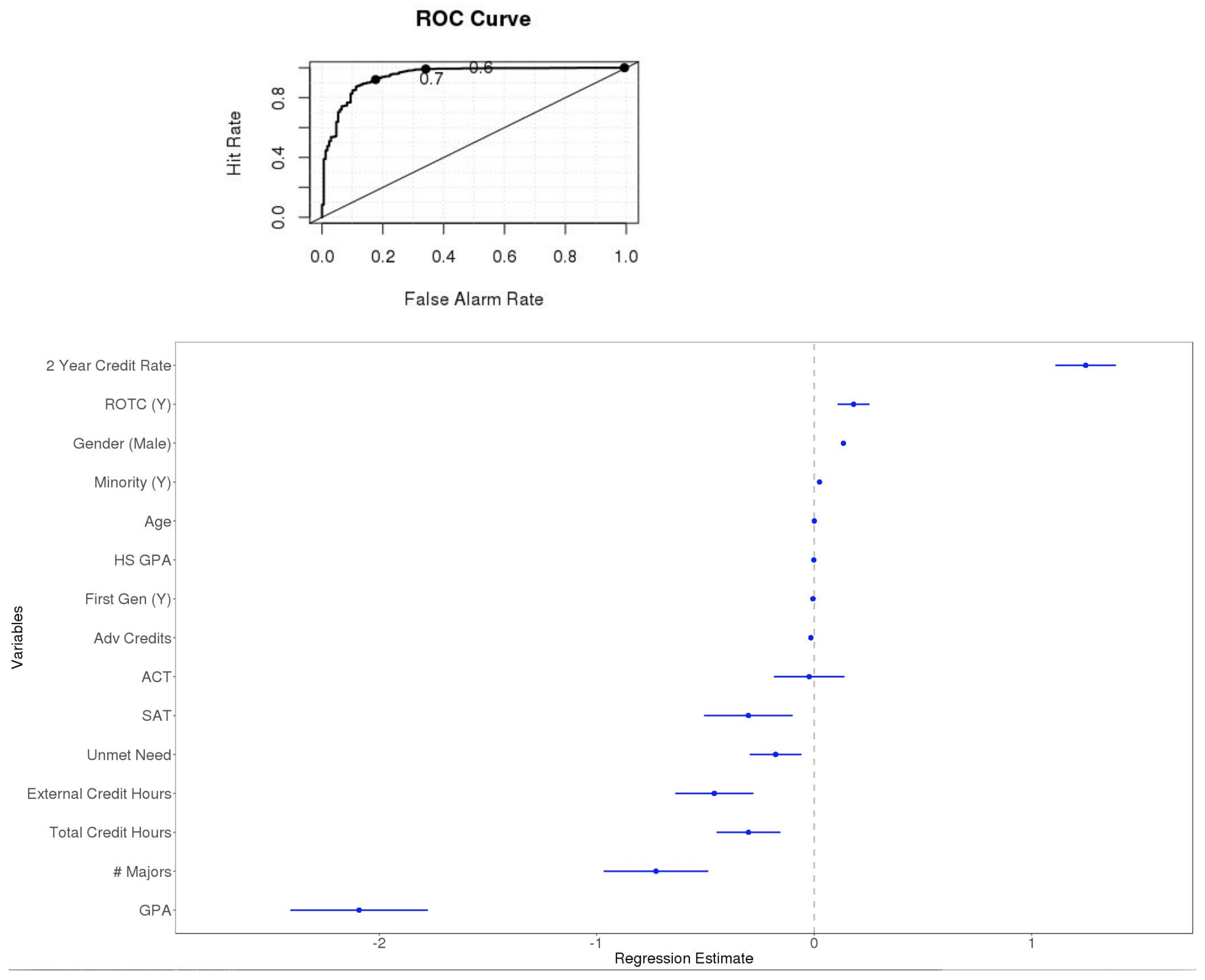

Figure 2. Coefficient plot for logistic regression model & corresponding ROC plot. Majority of variables showed significance.

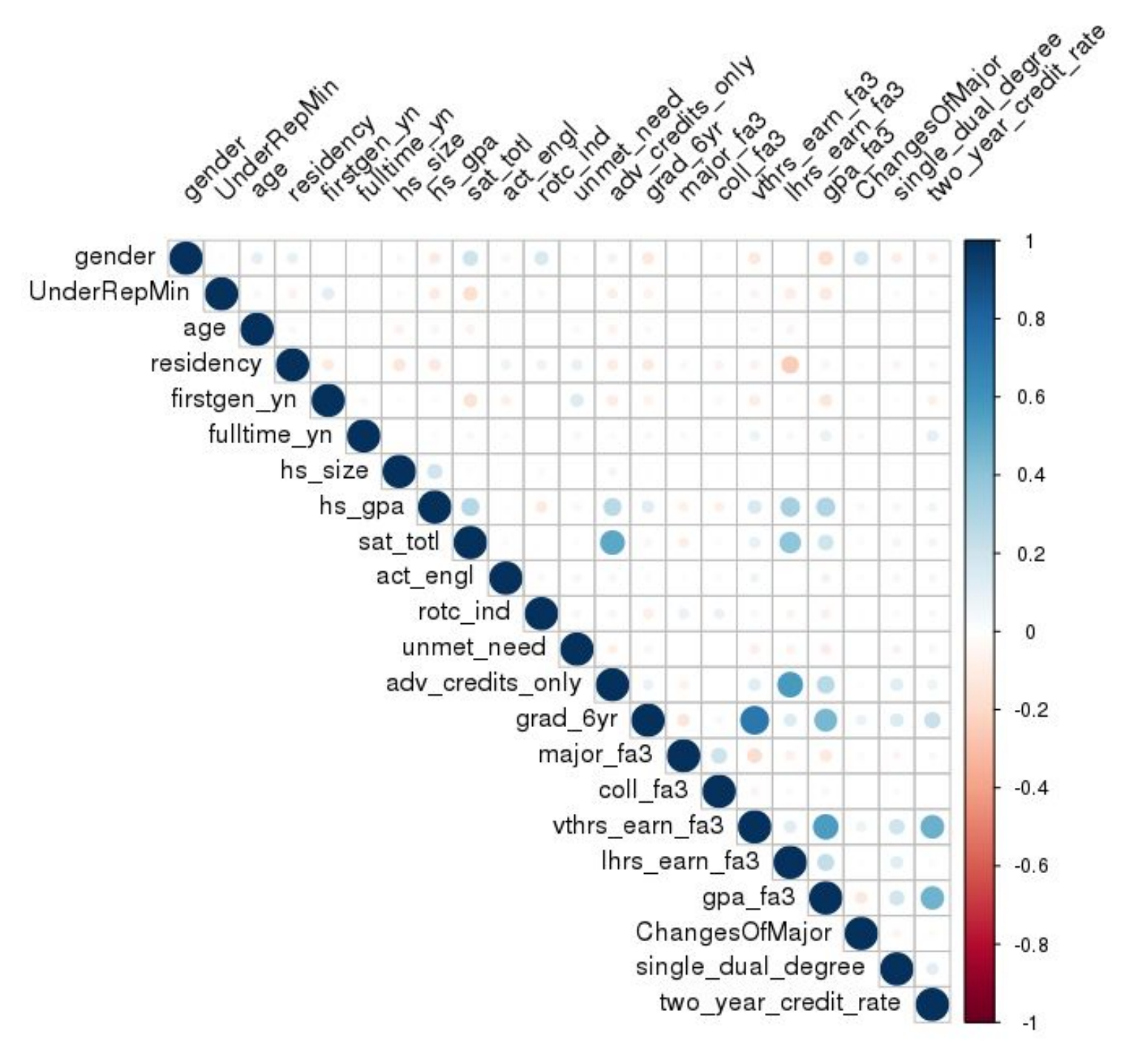

Figure 3. Correlation plot for all of the 18 variables to show the significance of the relationship between variables.

Conclusions

The model highlighted several variables of interest in relation to graduation likelihood. Positive impacts arise from factors such as 2-years credit hour, ROTC, and being male; meanwhile, test scores, unmet need, and GPA, tend to have more of a negative effect on likelihood of graduating.

This study has implications for how universities may approach degree completion across their undergraduate populations. Strategies may be developed to help provide targeted support to students with particular vulnerability to non-completion based on prior coursework, individual background, or social engagement. Furthermore, it may be beneficial to conduct further research on the social embeddedness of student populations and how these groups interact via support networks. Identifying positive social behaviors associated with increased graduation likelihood may help the university to develop strategies to encourage student success.

References

1. Murphy, Joel P. and Murphy, Shirley A. "Get Ready, Get In, Get Through: Factors that Influence Latio College Student Success." *Journal of Latinos and Education*, vol. 17, issue 1, 10 February 2017, p. 3-17. *https://www-tandfonline-com.ezproxy.lib.vt.edu/doi/full/10.1080/15348431.2016.1268139?scroll=top&needAccess=true*. Accessed 11 June 2018.
2. Wilson, Rick L. and Hardgrave, Bill C. "Predicting Graduate Student Success in an MBA Program: Regression Versus Classification" *Educational and Psychological Measurement*, vol. 55, issue 2, 1 April 1995, p. 186-195. https://doi.org/10.1177/0013164495055002003. Accessed 19 June 2018.
3. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodological) 58(1):267–288

BIOCOMPLEXITY INSTITUTE
bi.vt.edu/sdal
Stay Connected: facebook.com/biocomplexity
twitter.com/biocomplexityvt
youtube.com/c/biocomplexity
google.com/+biocomplexity

VIRGINIA TECH.