The Impact of Undergraduate Research on Doctoral Attainment of Spelman College Graduates

Jennifer Stanford Johnson
Associate Director, Student Success
Spelman College

7/22/2019

Table of Contents

List of Figures

List of Tables

Background

The hallmark of a Spelman educational experience is a challenging liberal arts curriculum that has prepared over six generations of African-American women to reach the highest levels of academic, community, and professional achievement.

Spelman's pride points include:

- Recognized among the Top 100 national liberal arts colleges (tied for No. 51)
- Among nation's most innovative liberal arts colleges
- Recognized as an institution noted for programs in study abroad and the first-year experience
- Among the Top 35 producers of Fulbright Fellows
- Leading producer of Black women who earn doctorates in STEM fields

These significant achievements deserve a deeper dive into which student experiences have contributed to these intrinsic successes, in particular the number of Spelman graduates with terminal degrees in STEM fields. An analytical exploration of this inquiry yielding scientific results would enable the College to determine which factors, if any, might be most predictive of the pursuit and completion of terminal degrees, broadly, by Spelman alumnae. Gaining a firm understanding of the most impactful practices strengthens academic excellence, and when these insights are operationalized within the curriculum and co-curricular activities, the reaped benefits can be scaled to the entire student body and future students.

Like most institutions of higher education, Spelman has a wealth of data ranging back several decades in various usable and semi-usable states. Therefore, a modest model allows for a baseline approach which can be enhanced as deeper insights are gained and as data are converted to more usable forms. Applying Occam's Razor, a reasonable starting place for any model which explains or predicts the pursuit of the highest levels scholarship and creative works must consider research and the role of the intellectual community. Examination of the long-standing, multidisciplinary practices of Spelman College reveals strong cohort communities, student/faculty mentoring relationships, and undergraduate research. The culmination of these three practices is showcased annually at Spelman's Research Day, a signature event that promotes undergraduate research among all majors. It is for this reason that Research Day is central to understanding the impact of undergraduate research on the doctoral attainment of Spelman College graduates.

Using ten years of data from the College's annual Research Day showcase, the National Student Clearinghouse database, affinity program records, graduation data, and other pertinent campus records, this analysis seeks to identify the ingredients of the "secret sauce" that have led to Spelman's various accolades and sustained success in developing the academic excellence of women of color in the liberal arts and the sciences.

Data

There is no shortage of academic, co-curricular, nor program data at Spelman. The challenge, though, is in gathering multi-year data that is already stored in a usable format meaning it contains a unique identifier, in this case the student identification number, and little to no variation in the data fields year over year. The bulk of the work for this and similar tasks is in identifying if and where the desired data exists, accessing and contextualizing the data, then cleaning or standardizing the data. The upside of such an effort is the incremental progress toward organizing the College's data.

The dataset most desirable for building the doctoral analytical model is a combination of raw data from multiple sources: Research Day database, student affinity programs, Banner, Navigate, and the National Student Clearinghouse. A detailed description of each is summarized below.

Research Day Database

Ten years of data was extracted from the Research Day database which is housed in IBM Lotus Notes. In total, there were 1787 Research Day presentations (observations) over the time period from 2008 to 2017.

Extracted Research Day data includes the following fields:

- Student ID (*when available)*
- Student name
- Presentation ID
- Research title
- Research abstract
- Primary research advisor
- Student's residential status (*when available)*

In order to glean insight and different representations of this data, additional data fields were derived from the Research Day data fields above. Data derived from the Research Day data are:

- Pres Type
- Advisor1-5
- Same advisor?
- Same topic?
- highest adv tier
- num_RD

From year to year, there were variations in the available data due to changes in focus and direction of the Research Day committee and the increase of student tracking efforts, among other factors. Listed below are challenges with the Research Day data:

- Prior to 2013, the **Student ID was not being collected** and stored in the Research Day database. As a result, a significant amount of time was devoted to manually cross referencing student name and enrollment records in other databases (e.g.,

Banner and Navigate) in order to find the unique identifier. Microsoft Excel was used in small part, but this task was mainly a manual process.
- For years with limited data stored in the database, **data had to be wrangled** from the Microsoft Word format of the Research Day program into Microsoft Excel
- **Student name reconciliation was complicated** due to the practice of updating maiden names with married names in Banner.
- There were **faculty advisor naming variations** from year to year (e.g., Joel Sokol and J. Sokol). In order to standardize faculty names, OpenRefine (a powerful tool for working with messy data) was used to cluster and edit the data.
- Lack of faculty advisor data (e.g., department and division) resulted in having to seek out and merge with data from the Office of the Provost. For external faculty advisors, this meant relying on Google searches to capture this information.
- **Inconsistent data collection** of residential status and research grant funder over the time period resulting in its omission from the final dataset.
- Due to **non-standard data collection practices**, there was a variation in how student research groups were handled. In cases where a group of student conducted and presented their research, only the lead/primary presenter was credited with presentation in the database. In earlier years, only a group name was provided, and therefore, those few groups were omitted entirely from the dataset.
- **Cross registered students** were removed from the dataset.
- **Disparate Research Data databases** did not allow for access to other desired data such as Research Day award winners

Student Affinity Programs

Spelman maintains several dozen grant and institutionally funded student affinity programs annually, some of which have a research focus. While this data does exist on campus, limited, usable data was readily available for the given time period. However, student participation in two hallmark Spelman programs were included in the analysis:

- **Research Initiative for Scientific Enhancement (RISE)** – The RISE (Research Initiative for Scientific Enhancement) program is a structured biomedical research-training program for underrepresented minorities and women who desire to pursue a career in biomedical research. The RISE program at Spelman College was established in 2000 and is structured into two segments:
  - The Research Development Program (academic year and summer)
  - The START Program (Summer Training About Research Techniques)

- **Ethel Waddell Githii Honors Program** – Founded in 1980, the Spelman College Honors Program, named for scholar-teacher Ethel Waddell Githii, is interdisciplinary in design recognizing the diversity of Spelman's faculty expertise and student creative scholarship. The Honors Program creates original programming and targeted supports for member students, and collaborates with academic departments and programs to provide a rich array of scholarly and creative venues.

Banner & Navigate

Banner is the higher education enterprise resource planning system which houses Spelman's academic, financial aid, institutional advancement, business and financial affairs, and enrollment management data.

Extracted Banner data includes the following fields:

- Student ID
- Student name
- Cohort
- GPA
- Hours attempted
- Hours earned
- Classification

Data derived from the Banner data are:

- Graduate?
- num_yrs_enrolled

Navigate is the comprehensive advising platform which receives nightly data updates from Banner. For this project, Navigate was helpful in identifying students with missing data.

National Student Clearinghouse

The National Student Clearinghouse is the source of subsequent enrollment and degree completion of the Spelman graduates who participated in Research Day during the given ten-year time period.

Illustrated below are the steps used to create the final dataset used for this analysis.

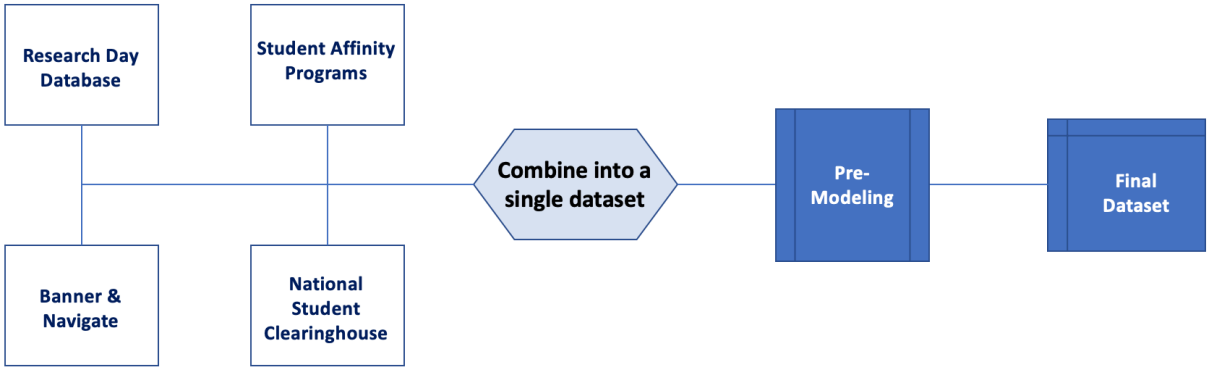

Figure 1. Dataset Preparation

After combining the data from the sources discussed above, initial decision tree and logistic regression models were run. Based on the results, some data fields were removed (see Preliminary Model section) resulting in the final dataset presented in the next section.

Final Dataset

Merging of the four data sources mentioned above and initial data analysis resulted in the following dataset of key data fields. The fields shown in blue represent derived data fields.

Table 1. Final Research Day Key Data Fields

Data Field	Field Description	Variable	Variable Type	Data Source
RD Year	Year	independent	numerical	Research Day database
Pres Type	Research Day presentation type (e.g., oral or poster presentation)	independent	categorical	Derived from Research Day database
Title	Title of research submission	independent	text	Research Day database
Abstract	Abstract of research submission	independent	text	Research Day database
Advisor1-5	Advisor categories based on the number of student mentees	independent	categorical	Derived from Research Day database
Same advisor?	Student had the same advisor each time she presented	independent	categorical	Derived from Research Day database
Same topic?	Student had the same research topic at least twice	independent	categorical	Derived using text analytics
Highest Adv Tier	Highest tier of the advisors over the time period	independent	categorical	Derived from Advisor1-5
Honors?	Honors program participant	independent	categorical	Honors Program
RISE?	RISE program participant	independent	categorical	RISE Program
Cohort	Entering class	independent	categorical	Banner
Class	Graduating class	independent	categorical	Banner
Graduate?	Graduation status	independent	categorical	Derived from Class
Num_RD	Number of times participating in Annual Research Day	independent	numerical	Derived from Research Day database
Num_yrs_enrolled	Number of years enrolled at Spelman	independent	numerical	Derived from Cohort and Class
Doctorate?	Earned doctorate	dependent	categorical	National Student Clearinghouse database

Exploratory Analysis

Before attempting to build the analytical model, exploratory analysis provides clues about behavior and relationships in the combined dataset (pre-final dataset). Descriptive statistics and network graphs are presented in the sections below.

Descriptive Statistics

There are 1787 observations and 30 variables in the combined dataset. Below is a descriptive overview of the dataset.

By Annual Presentations

Summarized below is the annual Research Day presentation count from 2008 to 2017.

Table 2. 2008 – 2017 Research Day Presentations

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Grand Total
Oral	46	60	45	45	92	82	97	93	119	94	**773**
Poster	76	71	76	90	124	114	109	100	127	127	**1014**
Grand Total	122	131	121	135	216	196	206	193	246	221	**1787**
Total Spring enrollment	**2231**	**2157**	**2113**	**2096**	**2058**	**2060**	**2042**	**2019**	**1695**	**2013**	**20484**
Percentage of enrollment (%)	5.47	6.07	5.73	6.44	10.5	9.51	10.1	9.56	14.5	11	**8.72**

By Affinity Group

On average 5 to 11% of the total enrollment showcased their research during the annual Research Day celebration. As a part of their research submission, students can chose either an oral or poster presentation or both. As such, the average overall rate of participation is 1.36 presentations per student with increased rates for students who are designated as RISE and Honors Program participants. Summarized below is the comparison of the affinity groups.

Table 3. Comparison of Affinity Group Participation in Research Day

	# presentations	# students	Rate of participation
RISE only	110	55	2.00
Honors only	352	239	1.47
RISE + Honors	60	25	2.40
Overall	**1787**	**1310**	**1.36**

Further examination of these statistics reveal:

- 13 of the RISE students who participated in Research Day earned a doctorate and showcased 26 presentations in alignment with the rate of participation above
- 22 of the Honors students who participated in Research Day earned a doctorate and showcased 26 presentation (rate of participation = 1.18)

- 4 students who were both RISE and Honors students and participated in Research Day earned a doctorate and showcased 9 presentations (rate of participation = 2.25). Given that there are students in both affinity programs, it might be useful to consider an interaction term in the logistic regression model.

By Classification

Provided below is a summary of Research Day by classification. Seniors are the largest classification to participate. The #N/As represent students who graduated a semester early, and therefore were not technically enrolled during the Spring semester when Research Day occurs.

Table 4. Participant Classification (All Years)

Row Labels	Count of Classification
Senior	1263
Junior	298
#N/A	126
Sophomore	92
First Year	8
Grand Total	**1787**

Illustrated below is the count of Research Day presentations by classification filtered to show those 136 participants who subsequently earned doctorates.

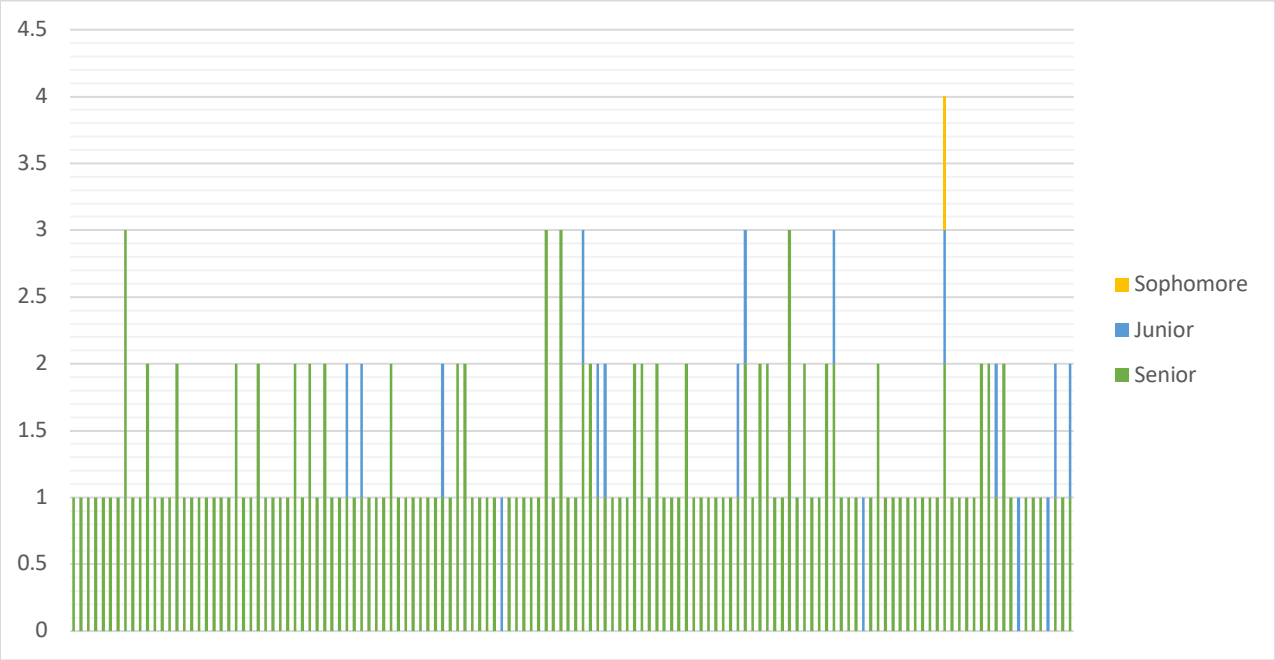

Figure 2. Presentation Count of Participants with Doctorates (with Classification shown)

By Terminal Degree

Comparing the count of doctorates between Research Day participants and non-participants there is a clear difference between the two groups. Of the 1310 students who participated in Research Day, 136 (10.4%) earned a doctorate. In contrast to the 19,174 students who did not participate, 301 (1.56%) earned a doctorate.

It is interesting to note that of the 136 Research Day participants who earned doctorates, 32.3% completed the PhD in contrast to 11.6% of non-participants. Nearly three times as many Doctorates of Philosophy for those students showcasing their research at Research Day. This indicates some relationship between participation Research Day and/or academic year research and the PhD.

Researchers studying the impact of summer undergraduate research experiences would agree with this preliminary finding. Undergraduate research adds to the student's credentials for being admitted to graduate school, so the experience has instrumental value in continuing the student's career trajectory [2]. Similar research found that participation in Research Experiences for Undergraduates (REU) Site programs was also effective at boosting research productivity. REU participants produced 2.14 times and 1.58 times as many scientific presentations and publications [5]. Recall Table 3 which summarizes similar output of students in research-focused programs such as RISE.

Table 5. Doctorates of Participants

Degrees	Count of Terminal Degree
DOCTOR OF PHILOSOPHY	**44**
JURIS DOCTOR	42
DOCTOR OF MEDICINE	27
DOCTOR OF DENTAL SURGERY	6
DOCTOR OF PHARMACY	6
DOCTOR OF MUSICAL ARTS	2
DOCTOR OF PSYCHOLOGY	2
DOCTOR OF CHIROPRACTIC	3
DOCTOR OF PODIATRIC MEDICINE	1
DOCTOR OF EDUCATION	1
DOCTOR OF DENTAL MEDICINE	1
DOCTOR OF PHYSICAL THERAPY	1
Grand Total	**136**

Table 6. Doctorates of Non-Participants

Degrees	Count of Terminal Degree
JURIS DOCTOR	186
DOCTOR OF MEDICINE	36
DOCTOR OF PHILOSOPHY	**35**
DOCTOR OF DENTAL SURGERY	12
DOCTOR OF PHARMACY	8
DOCTOR OF DENTAL MEDICINE	5
DOCTOR OF EDUCATION	3
DOCTOR OF CHIROPRACTIC	3
DOCTOR OF PHYSICAL THERAPY	2
DOCTOR OF NATUROPATHIC MEDICINE	2
DOCTOR OF PSYCHOLOGY	2
DOCTOR OF OSTEOPATHIC MEDICINE	2
DOCTOR OF OPTOMETRY	1
DOCTOR OF PODIATRIC MEDICINE	1
DOCTOR OF NURSING PRACTICE	1
DOCTOR OF SOCIAL WORK	1
DOCTOR OF VETERINARY MEDICINE	1
Grand Total	**301**

By Department of Research Advisors

Students who showcased their research during Research Day were primarily advised by Spelman faculty. By the numbers, 1648 presentations had Spelman research advisors and the remaining were advised by external advisors. Below is a count of the departments of the advising Spelman faculty.

Table 7. Top 10 Departments/Programs of Research Advisors

Department/Program	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Grand Total
Biology	17	15	29	50	82	69	58	41	48	57	**466**
Chemistry and Biochemistry	21	17	17	12	24	17	30	21	22	15	**196**
Psychology	10	17	13	10	23	15	21	14	18	26	**167**
Sociology and Anthropology	3	3	9	11	15	28	13	17	25	15	**139**
Computer & Information Sciences	13	7	6	8	8	9	8	10	12	6	**87**
Environmental Science	6	11	5	4	10	9	6	10	9	7	**77**
Political Science	3	11	8	8	10	4	7	2	9	11	**73**
Mathematics	7	7	8	4	7	8	9	8	5	6	**69**
English	5	2	3	3	6	2	11	7	13	5	**57**
History		3	1	5	7	6	3	7	8	1	**41**

Table 8. Top 10 Departments/Programs by Faculty Lines

Department/Program	Count
Psychology	24
World Languages and Literature	19
English	19
Chemistry and Biochemistry	18
Biology	16
Music	16
Mathematics	12
Art and Visual Culture	12
Education	11
Economics	10
Computer and Information Sciences	10

Department/program size is not necessarily correlated with the number of faculty serving as research advisors. Larger departments such as Biology, Chemistry and Biochemistry, and Psychology also lead the Research Day count. Sociology and Anthropology and Computer & Information Sciences have fewer faculty, but fall within the top 5 in presentation counts.

The role of research advisors as engaged mentors is crucial. Students at primarily undergraduate institutions are most likely to have faculty mentors. A great deal of responsibility is placed on the shoulders of the research mentor. Mentors are described as teachers, coaches, career advisers, and gatekeepers to the community of scholars [1].

Network Graphs

The relationship of the research advisor to the student is easily visualized with the use of networks graphs via the Gephi visualization platform. Two graphs were generated in order to lift up the important research mentoring bonds – one for all of the Research Day participants and their research advisors and another for Research Day participants with doctorates and their research advisors. Illustrated below are both relationships.

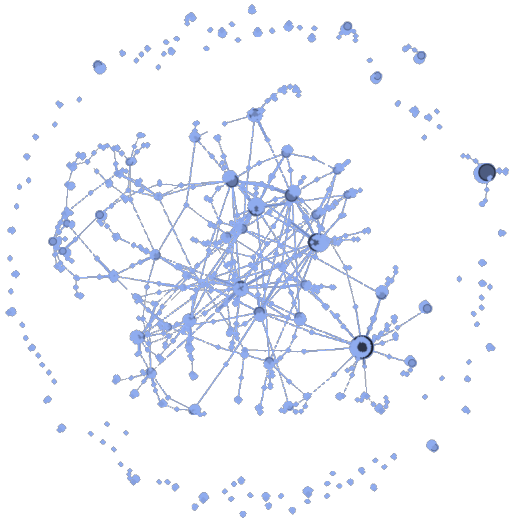

Figure 3. Overall Network Graph of Research Day Participant and Research Advisor

With 1310 students and 270 Spelman and non-Spelman research advisors, the graph is very cluttered. Using a Force Atlas layout for network spatialization and node degree settings, the larger, more central circles represent research advisors with the highest count of student participants over the ten-year period.

In contrast, the figure below illustrates the relationship between Research Day participants with doctorates and their research advisors.

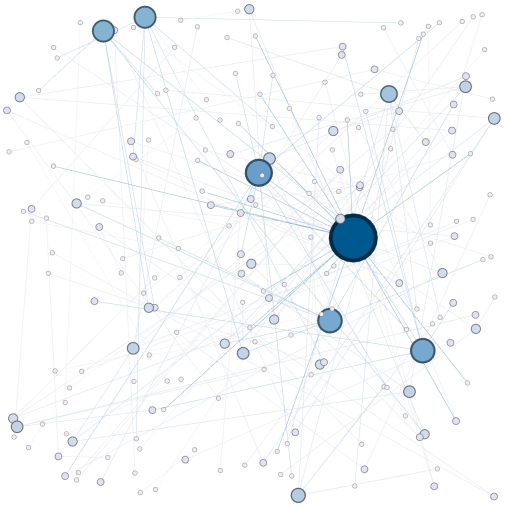

Figure 4. Network Graph of Research Day Participants with Doctorate and Research Advisor

This simple network graph illustrates 136 students and their 73 Spelman and non-Spelman research advisors. In this graph, though, the research advisor who advised the most students in Figure 3 is not the same advisor who advised the most students with doctorates. This indicates that the advisor-mentor relationship is an important consideration not necessarily favorable to faculty with large advising loads. Faculty advisors, therefore, should be included in the model.

Table 9. Top Research Advisors vs. Terminal Degrees

Research Advisor	JD	PhD	MD	Allied Health Field	Total
Advisor A	18				**18**
Advisor B	1	1	3	5	**10**
Advisor C			6	3	**9**
Advisor D		8	1		**9**
Advisor E		4	2	2	**8**
Advisor F	3	3		2	**8**

Preliminary Findings

This exploratory analysis was useful in providing the following insights and considerations which can be incorporated into the analytical models:

- Students in the RISE program have a higher presentation to student ratio.
- Students in both affinity programs also have a higher presentation to student ratio, and therefore, it might be useful to consider an interaction term in the logistic regression model.
- Recruiting younger students does not necessarily translate into higher rates of the doctorate. Evidence for the success of this strategy has not yet accumulated, and there are grounds for skepticism. Young students who have not committed to a career track may feel a strong desire to keep their options open and sample among a variety of valuable experiences, such as travel or internship experiences [2].
- 10.4% of RD participants earned a doctorate, in contrast to 1.56% of non-participants; Several studies have similarly suggested that undergraduate research participants were more likely to pursue graduate education than non-researchers [1].
- There is a relationship between participation Research Day and/or academic year research and the PhD.
- Student researcher-research advisor relationships matter.

Model

Preliminary Model

A preliminary decision tree model was constructed in R using the extracted data fields from the Research Day database, student affinity groups, Banner, and the National Student Clearinghouse. The exploratory analysis provided clues about which factors may prove to be most important in the models.

Conversely, a pre-model helped eliminate trivial factors from consideration. For example, academic data such as GPA, hours earned, and hours attempted tended to dominate the model and were more predictive of graduation from Spelman than completion of the doctorate. College graduation is certainly a prerequisite for completion of any graduate degree. Therefore, these three variables were removed from the model.

More Derived Features

A literature review on the impact of undergraduate research on underrepresented minorities agrees that self-efficacy of undergraduate researchers is built through successive, incremental, and iterative experiences proposing, conducting, and disseminating research to an increasingly broad and professional audience, an affordance of a multi-year program. Further, the personal mentor–student relationship is not forged as strongly over the course of several weeks as it is over several years. Mentors are an important source of confidence for students; having a respected role model who believes in your potential to pursue graduate school is critical [1].

As a result, two very important factors were derived from already extracted fields.

'Same advisor?' signals whether the student maintained the same research advisor for all of her Research Day presentations. This derived field is determined by comparing the number of Research Day participations ('num_RD') with the number of repeated advisors/mentors.

'Same topic?' gauges, through text analytics of the abstract, whether the student had the same research focus for all of her Research Day presentations. Using Python and the Natural Language Toolkit (NLTK), the abstract was tokenized and stop words were removed. For each Research Day participant, keywords were extracted and intersecting words across each year of participation were counted. As a starting baseline, one or more intersecting word is interpreted as the same topic; zero means not the same topic.

Correlations

Using the final dataset described in Table 1, the correlation matrix summarizes the correlation coefficients for the numerical values. The only meaningful correlation is between num_RD and cnt_adv which is an intermediate field used in determining 'Same advisor?'. The positive correlation between 'num_RD' and 'cnt_adv' supports the literature's assertion that the value of undergraduate research occurs over time and with the support of a research mentor.

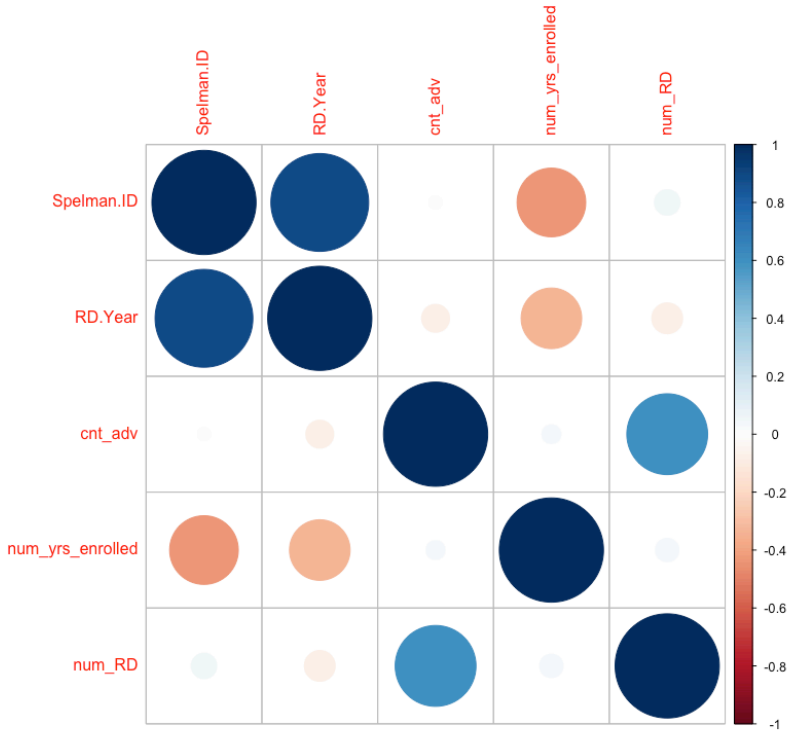

Figure 5. Correlation Matrix

Train & Test Datasets

The full dataset has 1310 observations which represents the number of unique students participating in Research Day over the ten-year period. The dataset is split using the ratio 75% used for training purpose and the remaining 25% used for test purposes. This translates into 982 observations in the train dataset and 328 observations in the test dataset.

Predictive Models

Two classification models were selected for this analysis: logistic regression and decision tree. Each was developed using the glm and rpart libraries in R, respectively.

Logistic Regression

Logistic regression was chosen because of the binary output of the dependent/response variable 'Doctorate?'. Stepwise regression was used to fit the regression model through a forward and backward automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of independent variables with the goal of achieving the lowest AIC.

The final models were:

Backward Model: Doctorate. ~ highest.adv.tier + Honors. + RISE. + Class
Forward Model: Doctorate. ~ Class + Honors. + highest.adv.tier + RISE.

Decision Tree

Decision tree is a simple way to visualize model decisions and their possible consequences. The variables used to grow the tree are: 'Class', 'Cohort', 'highest_adv_tier', 'Honors?', 'num_RD', and 'RISE?'. Figure 6 below illustrates the decision tree.

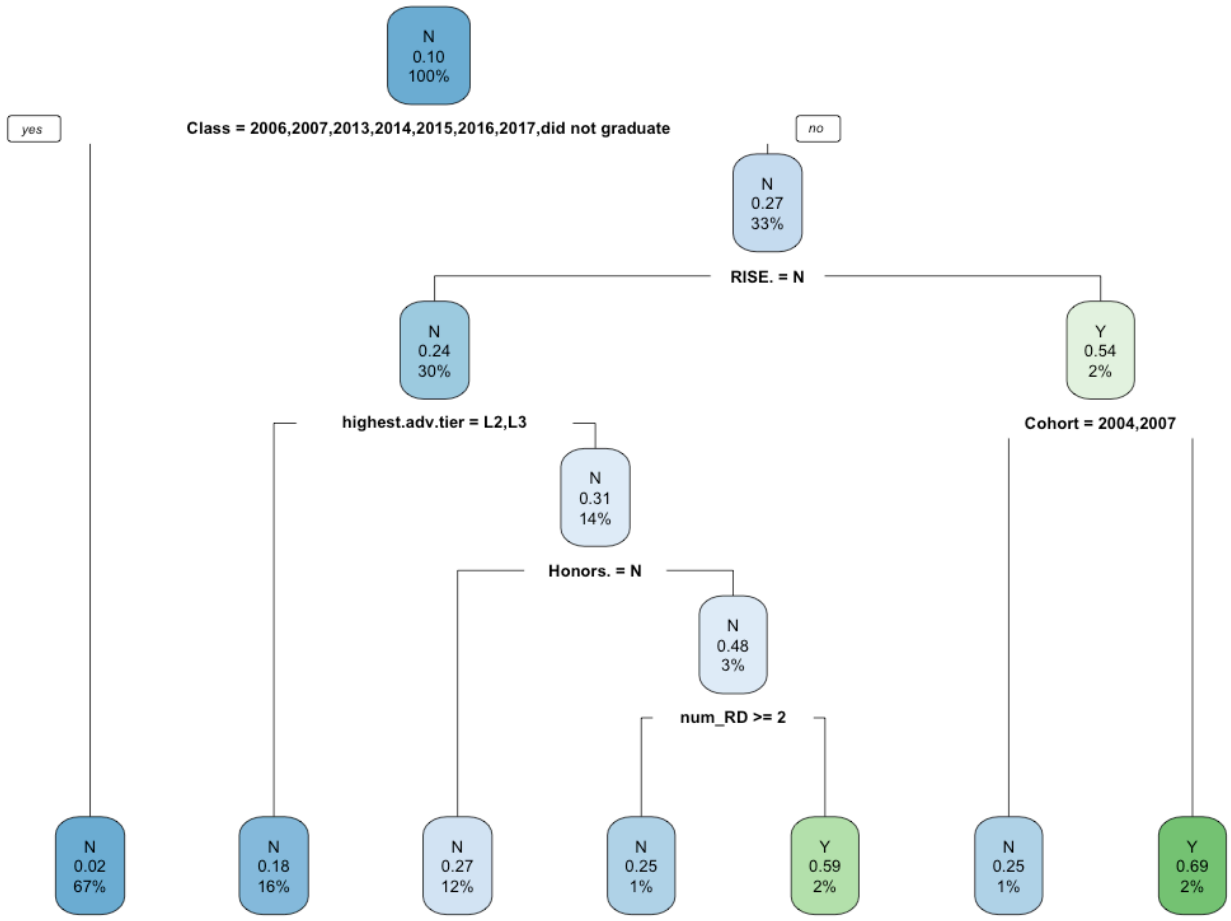

Figure 6. Decision Tree

Results

Model Performance

The performance of each of the models was assessed using at least one of the following statistical measures:

- Z-Tests
- Pseudo R^2
- Variable importance
- K-fold validation
- Confusion matrix

Logistic Regression

z-Tests

The z-tests show significance with a p-value less than 0.01 for the categorical variables 'Honors?' = Y and 'Class' = 2009 or 2010. 'RISE' and several of the other 'Class' variable have higher p-values ($p < 0.5$). Besides the intercept, none of the coefficients are statistically significant ($p < 0.001$) which means the independent variables of the model do not affect the dependent variable 'Doctorate?'. To test this theory, further performance measures are tested.

Pseudo R^2

Unlike linear regression, there is no R^2 statistic; instead use McFadden's R^2. The measure ranges from 0 to just under 1, with values closer to zero indicating that the model has no predictive power. Using the pscl package, McFadden's R^2 is 0.311.

Table 10. Pseudo R^2 Statistic

llh	llhNull	G2	McFadden	r2ML	r2CU
-221.068	-320.986	199.834	0.311	0.184	0.384

Variable Importance

This parameter is used to assess the relative importance of individual predictors in the model using the caret package. Values typically range from 0 to 100; for this logistic regression model, the values ranged from 0.92012182 to 3.04204329.

K-fold Validation & Confusion Matrix

K-fold cross-validation is used to assess how well the model performs in predicting the target variable on different subsets of the data. In this case, 10-fold validation was used and below is the confusion matrix.

n = 328	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 286	FP = 36	322
Actual: YES	FN = 5	TP = 1	6
	291	37	

Accuracy : 0.875
95% CI : (0.8343, 0.9088)
No Information Rate : 0.8872
P-Value [Acc > NIR] : 0.7865
Kappa : 0.0155
Mcnemar's Test P-Value : 2.797e-06
Sensitivity : 0.98282
Specificity : 0.02703
Pos Pred Value : 0.88820
Neg Pred Value : 0.16667
Prevalence : 0.88720
Detection Rate : 0.87195
Detection Prevalence : 0.98171
Balanced Accuracy : 0.50492
'Positive' Class : N

Figure 7. Logistic Regression Model Confusion Matrix and Computed Rates

Decision Tree

Model performance for the decision tree model is assessed with the confusion matrix based on test data.

Confusion Matrix

n = 328	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 286	FP = 35	321
Actual: YES	FN = 5	TP = 2	7
	291	37	

Accuracy : 0.878
95% CI : (0.8377, 0.9114)
No Information Rate : 0.8872
P-Value [Acc > NIR] : 0.734
Kappa : 0.0571
Mcnemar's Test P-Value : 4.533e-06
Sensitivity : 0.98282
Specificity : 0.05405
Pos Pred Value : 0.89097
Neg Pred Value : 0.28571
Prevalence : 0.88720
Detection Rate : 0.87195
Detection Prevalence : 0.97866
Balanced Accuracy : 0.51844
'Positive' Class : N

Figure 8. Decision Tree Model Confusion Matrix and Computed Rates

Conclusion

The logistic and decision tree models generated and the results of the exploratory analysis were of practical importance, but not statistical significance. Reasons for failing to establish statistical significance could be:

- **Dataset too small** – The current dataset has 1787 observations and fewer (1310) unique observations once duplicate student participation (i.e., multi-year Research Day participants) were removed. The next phase of this project will be to clean the remaining 20 years of data which will result in approximately 3000 more observations (assuming 150 presenters over the earlier 20 years).

  Another approach to increasing the size of the dataset is to incorporate more student affinity program data. This will require coordination with program leaders to gather their participant data dating back to program inception. Some of this data can also be acquired from the Office of Sponsored Programs based on annual grant reporting requirements to the funder. Lastly, the Office of Financial Aid may also be helpful since scholarships are awarded and stipends are paid from this office. The added benefit is that Financial Aid disbursement records will likely also contain the student ID.

- **Non-undergraduate research predictors** not captured in the model – In addition to expanding the size of the dataset, it would be worthwhile to consider career aspirations and parental influences of these participants as incoming first-year students and as graduating seniors. An alumnae survey could help the College understand their attitudes toward and motivations for pursing a doctorate before enrolling and Spelman and after graduating.

- **Academic year-over-year data** – Researchers believe that research experiences help students improve their academic performance. The presence of undergraduate researchers in a science course after they have had research experience may enhance their course experience [2]. Focusing on term improvements rather than absolute measures (e.g., GPA, hours earned, and hours attempted) will aid in understanding the broad impact of undergraduate research.

- **Faculty involvement as mentors beyond Research Day** – The role of faculty was narrowly concentrated on mentorship relative to this single annual event. Classroom and co-curricular mentorship might be equally as impactful as research mentorship.

There is an undeniable phenomenon occurring at Spelman College that dates back to its founding. In order to meet the changing needs of its students and achieve the College's ambitious strategic planning goals related to retention and graduation, institutional best practices must be deeply understood and scaled to continue to prepare alumnae for graduate school opportunities and careers in industry as well as support the most at-risk student populations in graduating with a competitive edge. Data analytics will certainly play a role in understanding ways in which to deliver on the Spelman promise.

References

1. Carpi, A., Ronan, D. M., Falconer, H. M., & Lents, N. H. (2016, August 05). Cultivating minority scientists: Undergraduate research increases self-efficacy and career ambitions for underrepresented students in STEM - Carpi - 2017 - Journal of Research in Science Teaching - Wiley Online Library. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/tea.21341

2. Research Day. (n.d.). Retrieved from https://www.spelman.edu/academics/research-programs/research-day

3. Review, P. (2014, December 29). Undergraduate Research as a High-Impact Student Experience. Retrieved from https://www.aacu.org/publications-research/periodicals/undergraduate-research-high-impact-student-experience

4. Thoman, D. B., Muragishi, G. A., & Smith, J. L. (n.d.). Research Microcultures as Socialization Contexts for Underrepresented Science Students - Dustin B. Thoman, Gregg A. Muragishi, Jessi L. Smith, 2017. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0956797617694865?rfr_dat=cr_pub=pubmed&url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&journalCode=pssa
5. Wilson, E, A., Pollock, L, J., Billick, Ian, . . . Adam. (2018, June 13). Assessing Science Training Programs: Structured Undergraduate Research Programs Make a Difference. Retrieved from https://academic.oup.com/bioscience/article/68/7/529/5034095