Texas College Bridge Program Effectiveness

Darin Young

Applied Data Science – DSC 680

July 11, 2025

Business Problem

My project is focused on the effectiveness of a program recently introduced at Midwestern State University called "Texas College Bridge." This program is designed to help students be more prepared for college and is intended to enhance student outcomes (retention, course grades, graduation rates, etc.). Student quality has only decreased over the years, so if the Texas College Bridge (TCB) program is effective, it will have widespread effects on Midwestern State University as a whole. With all of that said, my hypothesis is:

The Texas College Bridge (TCB) program has little to no impact on student outcomes – specifically in barrier courses for first year students as well as retention rates to the following spring.

If my hypothesis is true, it would be wise to reconsider our participation in the program over the long term.

Background/History

As part of the completion of the capstone course of my Master's Program at Bellevue University, I was asked to complete three totally open-ended research projects. Considering my role as a Data Analyst at Midwestern State, I felt it was an excellent opportunity to answer some institutional research questions regarding our students. So, I sought out advice from executive leadership and other leaders on campus on questions to research. This project was originally brought to my attention by Dr. Kristen Garrison, Associate Vice President at MSU. Our university first implemented the program for the Fall 2024 cohort, and Dr. Garrison, among others, had questions regarding the program's efficacy. Generally, they felt that the program did not have relevant impact on student outcomes and naturally questioned MSU's involvement in the program based on that feeling.

Data Explanation

The data for this project is real student data from the Fall 2024 cohort. It has been anonymized for student privacy.

For context, each row in the dataset represents a student's enrollment in a given term and each column is an attribute for that enrollment. The columns included are described below:

- ADM_MATH
  - Contains various values denoting how a student was admitted for the math component.
  - For this analysis, the only relevant value is "BRIDGEO," denoting a student that participated in the TCB program for this component.
- ADM_WRIT
  - Same as above, but for writing.
- ADM_READ
  - Same as above but for reading.
- COURSELOAD
  - Numeric value of credit hours taken in the term.

- SECOND_SEM_ENROLL
  - 1 = Enrolled in next semester, 0 = Not.
- YEAR
  - Year value for the row (all will be 2024)
- TERM
  - Term for the row (all will be Fall)
- ETHNICITY
  - 1 = Hispanic/Latino, 0 = Not
- RACE
  - Coded values for the student's race.
  - 0 = Missing
  - 1 = Non-Resident Alien
  - 2 = Hispanics of Any Race
  - 3 = Black or African American
  - 4 = White
  - 5 = American Indian Alaskan Native
  - 6 = Asian
  - 7 = Native Hawaiian Pacific Islander
  - 8 = Two or More Races
  - 9 = Unknown
- LEGAL_SEX
  - M = Male, F = Female
- FGEN
  - 1 = First Generation, 0 = Not/Unknown
- PELL
  - 1 = Pell eligible, 0 = Not
- HSCODE
  - Code for high school attended
- ZIP
  - Zip code where the student lives
- HIST1133:
  - Letter grade for this course (will be either A, B, C, D, F representing their grade. Could also be W, WF or WS, representing that they withdrew from the course. Could also be NA if they didn't take this course in this term).
- POLS1333: Same as above.
- POLS1433: Same as above.
- BUAD1033: Same as above.
- ENGL1143: Same as above.
- PHIL1033: Same as above.
- PHIL2033: Same as above.
- MATH1203: Same as above.

- MATH1233: Same as above.

Data Preparation

This dataset was prepared mostly using excel. The data came from a few sources (classes data, enrollment data, admissions data, demographic data, etc.), so it required joining via student id, which was removed for privacy concerns. In addition, I needed to include a random sampling of students from the Fall 2024 cohort, so that required filtering down the data. So, once I joined the various datasets and filtered to only include relevant students, I was ready to start making transformations/changes.

The first change that I made was adding a "TCB" column which indicated whether a student was a TCB student or not. This was determined by whether or not they were marked with "BRIDGEO" in **any one** of the "ADM" columns. This allowed me to identify TCB vs Non-TCB students later in my analysis.

The next change I made was dropping the PHIL1033 and PHIL2033 courses from the dataset entirely, because **no students** attempted the PHIL2033 course and **only one student** attempted the PHIL2033 course. This sample was useless and I felt those could be excluded for now (but could be added on later if we have more years of TCB involvement).

Methods

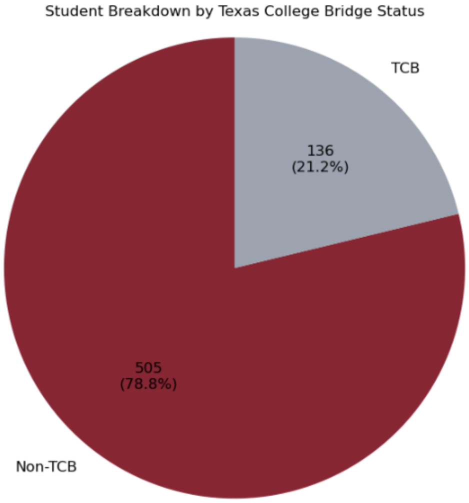
For my analysis, I used the following methods:

- Exploratory Data Analysis
    - I first looked at TCB vs Non-TCB students and their various relationships with student outcomes.
        - How many students passed the relevant courses for each TCB category?
        - How many students retained to the next semester for each TCB category?
- Chi-Square Testing
    - Then I looked at statistical significance for each of those populations to determine if the difference was worth exploring.
- Logistic Regression
    - Then, I created logistic regressions to predict student outcomes (retention and course grades) while controlling for other factors.
        - Specifically, I used Course Load, Pell Eligibility, First Generation Status and Legal Sex as **controls** to help bolster the regression models.
- Random Forest Modeling
    - Then, I created a random forest model – only for retention – to determine if the success/failure of the model was related to the modeling technique. I then compared the results to the regression analysis to evaluate performance overall.
- Propensity Score Matching
    - Then, I felt it would be valuable to match student's based on their "nearest neighbors" based on the same controls values mentioned above.

- ○ Once students were matched, I then compared outcomes to reinforce the causal inference of the hypothesis by assuming all other factors are equal aside from TCB status.

Analysis

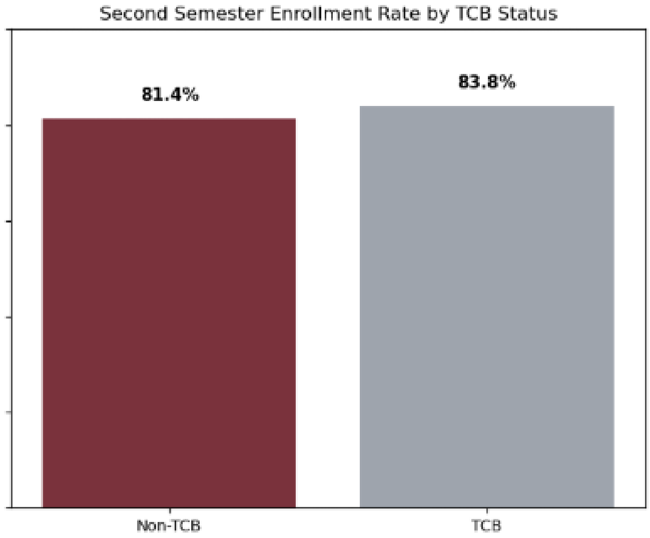
Exploratory Analysis



As mentioned above, the first step of my analysis was to explore the dataset related to the hypothesis and see what trends emerged. As shown, there were 641 total students in the dataset – 136 TCB students and 505 Non-TCB students. A further breakdown of each relevant demographic category can be found in Appendix A.

Once I visualized each of the demographics, I was interested to see trends for DFW rates for each course broken down by TCB status (this breakdown can be seen in Appendix B). I quickly found that, without taking into account any other factors, **students in the TCB program have a higher DFW rate than non-TCB students, for literally every course in the dataset.** In addition, I found that the difference was statistically significant for four courses – MATH1233, POLS1333, MATH1203, and HIST1133 – via chi-square testing.

Then, I was interested in whether there was any relationship between higher DFW rates for any singular course when comparing TCB vs non-TCB students. Essentially, are TCB/non-TCB students more/less likely to get a D, F or withdraw from one of the relevant courses? Similarly, I found that **overall DFW rates were higher for TCB students as compared to non-TCB students** (58.1% vs 35.4%, respectively). This was also reinforced by chi-square testing, suggesting that the difference was statistically significant.

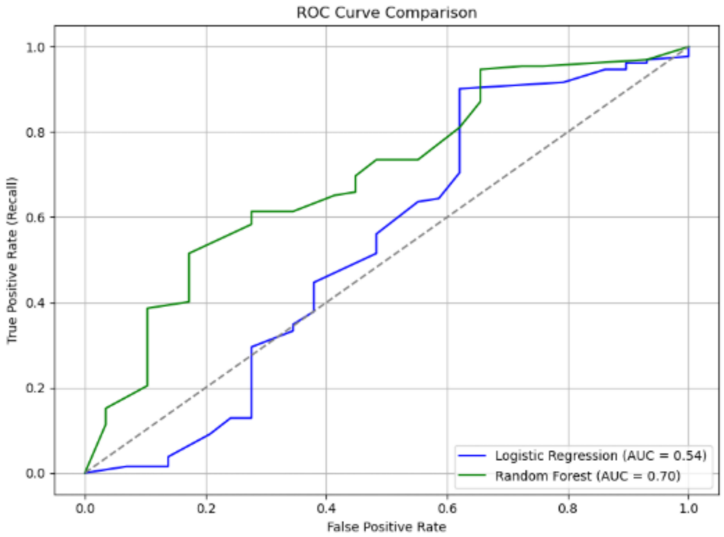


Finally, I was interested in comparing student retention for each TCB category and found that **TCB students have a slightly higher retention rate as compared to non-TCB students,** but upon chi-square testing, the difference between the two groups was not statistically significant.

Modeling

First, I built a regression model to predict student retention to the spring semester using the following predictors: TCB Status, Course Load, First Generation Status, Pell Eligibility and Legal Sex. The goal of this model was to identify whether TCB status was a relevant predictor of retention while controlling for other demographic characteristics. The model was fairly accurate, classifying students into the correct category 78% of the time. However, the model had many false negatives, indicating that many students were being predicted to retain when they actually did not retain. So generally, the model was defaulting to "retain" mainly because the number of students that retained was much higher overall for the entire dataset (around 80/20). **Regardless, the model suggested that students were only 2% more likely to retain if they participated in the TCB program, reinforcing the conclusion made earlier that the difference was statistically insignificant.**

To attempt to alleviate the false negatives, I then rebuilt the model using the "Random Forest" method. I felt that this modeling technique might be more effective considering the non-linear relationship between the variables. With that in mind, the accuracy of the model decreased overall (down to 70%), but it was more effective at identifying students at risk of not returning. **And still, the model suggested that TCB status is irrelevant when it comes to retention.** In fact, being a first-generation student turned out to be the biggest predictor of retention – first generation students are 1.5x more likely to retain to the next semester as compared to non-first-generation students. That's not relevant to our analysis, but I felt it was interesting discovery. **The ROC Curves for both models (that visualize their performance compared to random classification) are shown here.**



Once I completed the models for retention, I then moved to creating models for each course that had a statistically significant difference for TCB vs Non-TCB, identified earlier. I built a model for each course, with DFW rates as the outcome and the same predictors used previously. Some of the results were shocking, to say the least. **Specifically, for MATH1233, students that participate in the TCB program have 5.5x higher odds of DFW as compared to non-TCB students,** even while controlling for the other demographic variables. Similarly, TCB students in POLS1333 have 3x higher odds of DFW as compared to non-TCB students. The other two courses had high rates as well, but their p-values were just slightly over the 5% threshold, so it's possible that they are due to random change. The results of each of the regression analyses for these courses can be viewed in Appendix C.

Propensity Matching

Finally, I felt it would be valuable to match students based on their demographics and test if statistical significance still existed between TCB/non-TCB students. I matched students using the same characteristics as above (Course Load, Pell Status, First Generation Status, and Legal Sex) and evaluated DFW rates for each course after the matching occurred. In each case, the matching was fairly accurate, suggesting that similar students were grouped together. **Additionally, in every case, DFW rates were higher for TCB students as compared to non-TCB students** - the highest difference being a 28% increase in DFW rates for MATH1233 and the lowest being a 4% increase in DFW rates for HIST1133. The results of these tests can be viewed in Appendix D.

In any case, this matching reinforced the causal inference of my initial hypothesis by assuming all other factors are equal aside from TCB status.

Conclusion

Based on my analysis steps shown above, **it is quite clear that the TCB program has little or no impact on student outcomes.** If anything, it actually has a negative impact on student outcomes, as was demonstrated in multiple cases.

Limitations

The main limitations for this analysis are as follows:

- Small sample size
    - This dataset only includes students from the Fall 2024 cohort. In fact, it only includes a sample of students (641, to be exact) from that cohort.
    - With a larger n value (more students) and a larger span of time, the analysis would be much more meaningful and could be actioned.
- Student preparedness
    - Students that participate in the TCB program are likely less prepared for college overall, hence their participation in the program. They may have more issues socioeconomically, or when in comes to learning disabilities or potentially with their familiarity with college overall.
- DFW Simplification
    - When it comes to course grades, this analysis only considers whether a student passed or failed, but it does not consider actual grades in the courses. It’s possible that students who participate in the TCB program get better grades than their peers, when they pass the course.

Assumptions

The main assumptions for this analysis are as follows

- Students are similar across TCB categories
    - All analyses attempt to control for differences between students and only compare their participation in the program.
    - It’s possible that TCB students are simply more academically unprepared by definition – which explains their lesser outcomes.
- Student matching contains representative samples

  - I assume that when matching, each group has a representative sample of TCB vs non-TCB students that allows for meaningful comparison between groups.

Challenges

The main challenge I faced during this analysis was likely the same challenge that most data professionals face - cleaning and preparing the data. One of my professors during my undergrad at Wilmington University stated that data professionals spend about 80% of their time cleaning and preparing data and only 20% actually analyzing the data. I can attest to that. Cleaning and preparing is an iterative process that requires fine-tuning and adjusting. As you move further through your analysis, there may be new things to add or change to prepare your data appropriately. It can be extremely tedious but it's what makes this process so fun and rewarding.

Aside from that, the other challenge worth mentioning was inconsistency or unavailability of data. There are other variables that I believe may contribute a student's preparedness (and therefore my ability to match them). Unfortunately, high schools don't have a standardized GPA scale across the board (one school may use a 4-point scale while the next simply uses a scale from 0-100). Additionally, standardized testing scores (SAT/ACT) could be another valuable comparison tool when it comes to student preparedness – but many students choose to only take one or the other, and some number don't take either. These issues make it difficult to match students based on their academic backgrounds.

Future Uses/Additional Applications

I believe this analysis can and should be used again in the future to evaluate the efficacy of the TCB program at MSU. As mentioned previously, the dataset only includes one cohort of students – it would be extremely useful to evaluate these trends over 5 or 10 years to see if there is any value to continuing our involvement in the program.

Recommendations

Considering that this analysis represents such a small sample of students, I do not believe it would be wise for MSU to take action based on this report. My primary recommendation would be to **continue evaluating the TCB program with each new cohort and provide additional support to students that participate in TCB for the foreseeable future.** Additionally, I think it would be valuable to expand the analysis to include other success outcomes, like fall-to-fall retention and graduation rates, for example.

Implementation Plan

Due to the small sample, I do not believe any implementation should occur at this stage. Once more data is available and the conclusion hold true, a plan can be developed and implemented at that time.

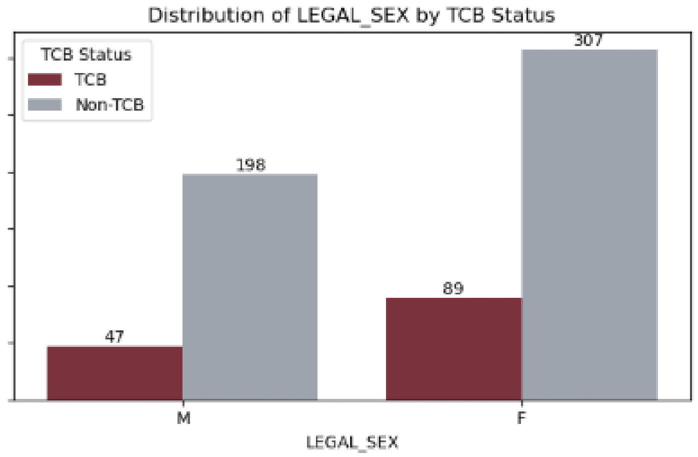
Ethical Assessment

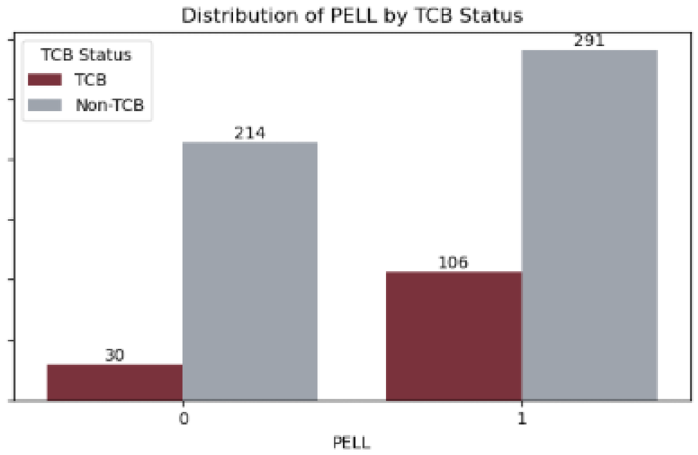
There are a few things to consider ethically when considering my conclusion:

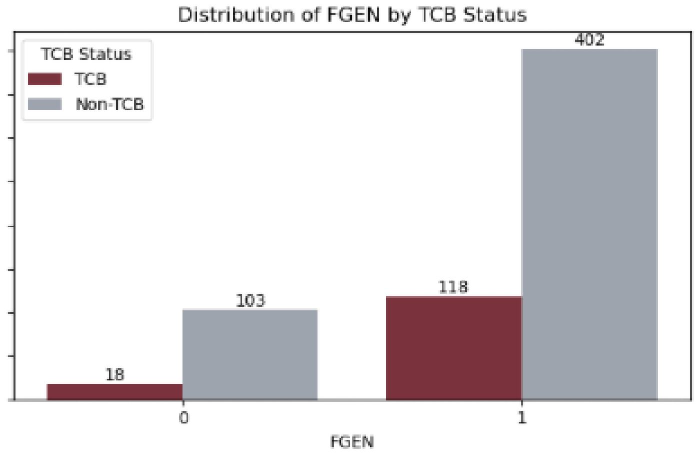
- Equity
  - My analysis suggests that TCB students perform worse in key courses, but it's possible that they perform *better than they would have* if they didn't participate in the TCB program.

- Evaluation of TCB as a whole
  - My analysis suggests that TCB is not effective, but it is only a relevant conclusion for MSU specifically, and does not imply the program does not work overall.

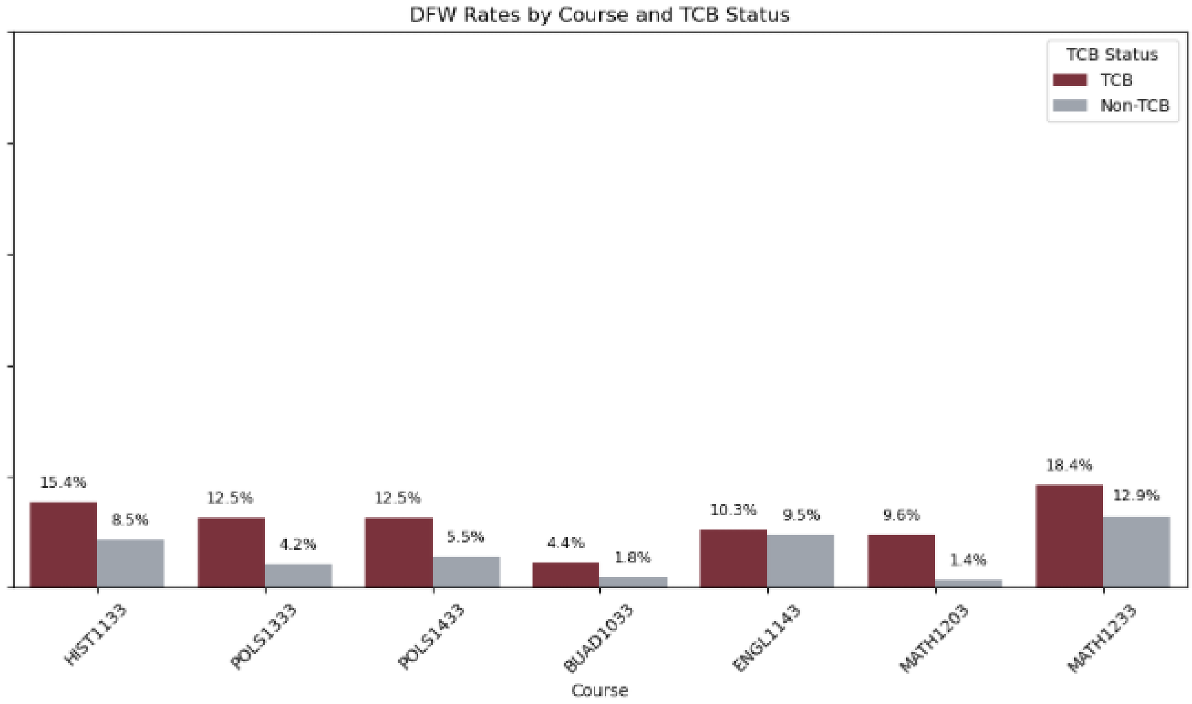
Appendix A – Demographic Distribution



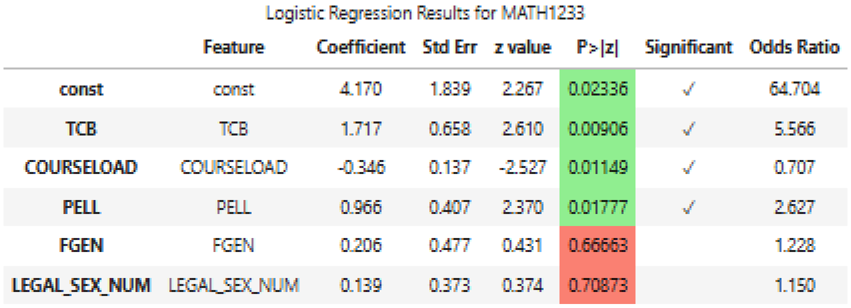




Appendix B – DFW Rates by Course

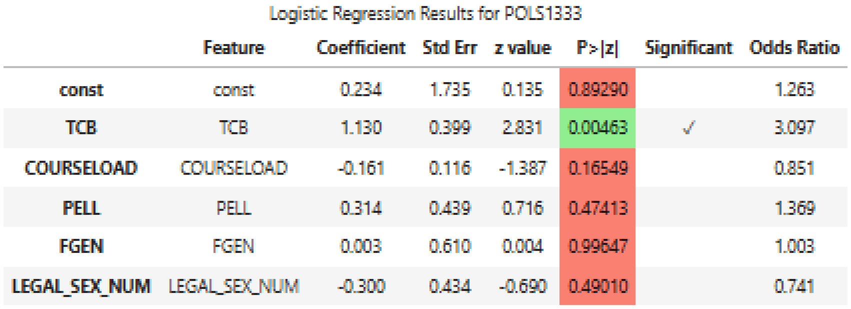


Appendix C – Regression Analyses for courses



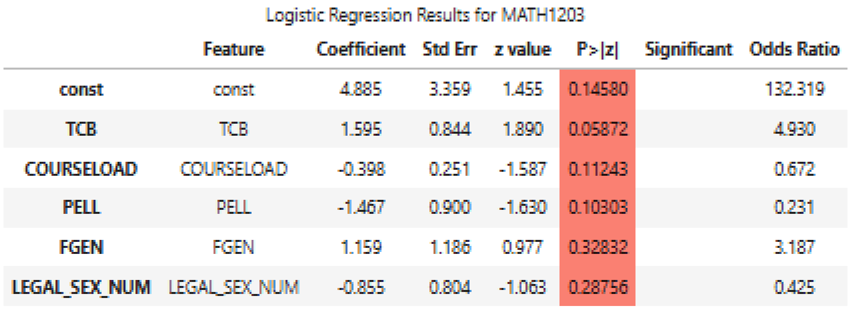
Logistic Regression Results for MATH1233

	Feature	Coefficient	Std Err	z value	P>|z|	Significant	Odds Ratio
const	const	4.170	1.839	2.267	0.02336	✓	64.704
TCB	TCB	1.717	0.658	2.610	0.00906	✓	5.566
COURSELOAD	COURSELOAD	-0.346	0.137	-2.527	0.01149	✓	0.707
PELL	PELL	0.966	0.407	2.370	0.01777	✓	2.627
FGEN	FGEN	0.206	0.477	0.431	0.66663		1.228
LEGAL_SEX_NUM	LEGAL_SEX_NUM	0.139	0.373	0.374	0.70873		1.150



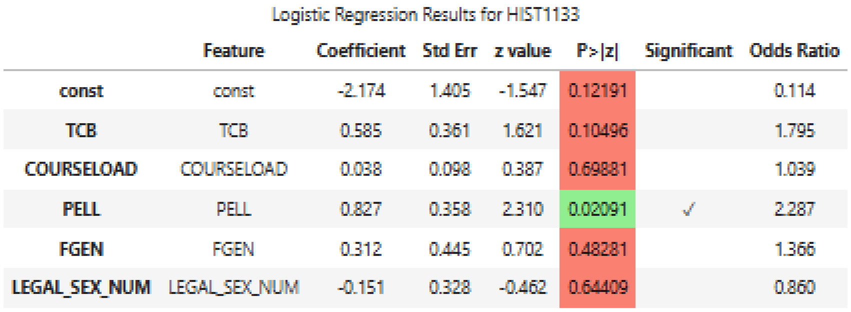
Logistic Regression Results for POLS1333

	Feature	Coefficient	Std Err	z value	P>|z|	Significant	Odds Ratio
const	const	0.234	1.735	0.135	0.89290		1.263
TCB	TCB	1.130	0.399	2.831	0.00463	✓	3.097
COURSELOAD	COURSELOAD	-0.161	0.116	-1.387	0.16549		0.851
PELL	PELL	0.314	0.439	0.716	0.47413		1.369
FGEN	FGEN	0.003	0.610	0.004	0.99647		1.003
LEGAL_SEX_NUM	LEGAL_SEX_NUM	-0.300	0.434	-0.690	0.49010		0.741



Logistic Regression Results for MATH1203

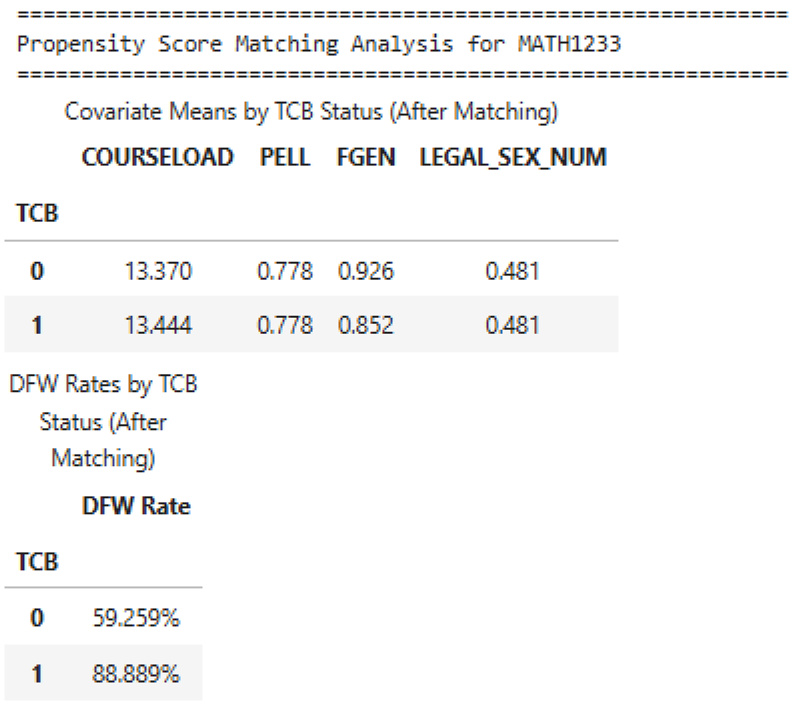
	Feature	Coefficient	Std Err	z value	P>|z|	Significant	Odds Ratio
const	const	4.885	3.359	1.455	0.14580		132.319
TCB	TCB	1.595	0.844	1.890	0.05872		4.930
COURSELOAD	COURSELOAD	-0.398	0.251	-1.587	0.11243		0.672
PELL	PELL	-1.467	0.900	-1.630	0.10303		0.231
FGEN	FGEN	1.159	1.186	0.977	0.32832		3.187
LEGAL_SEX_NUM	LEGAL_SEX_NUM	-0.855	0.804	-1.063	0.28756		0.425



Logistic Regression Results for HIST1133

	Feature	Coefficient	Std Err	z value	P>|z|	Significant	Odds Ratio
const	const	-2.174	1.405	-1.547	0.12191		0.114
TCB	TCB	0.585	0.361	1.621	0.10496		1.795
COURSELOAD	COURSELOAD	0.038	0.098	0.387	0.69881		1.039
PELL	PELL	0.827	0.358	2.310	0.02091	✓	2.287
FGEN	FGEN	0.312	0.445	0.702	0.48281		1.366
LEGAL_SEX_NUM	LEGAL_SEX_NUM	-0.151	0.328	-0.462	0.64409		0.860

Appendix D – Propensity Matching for courses



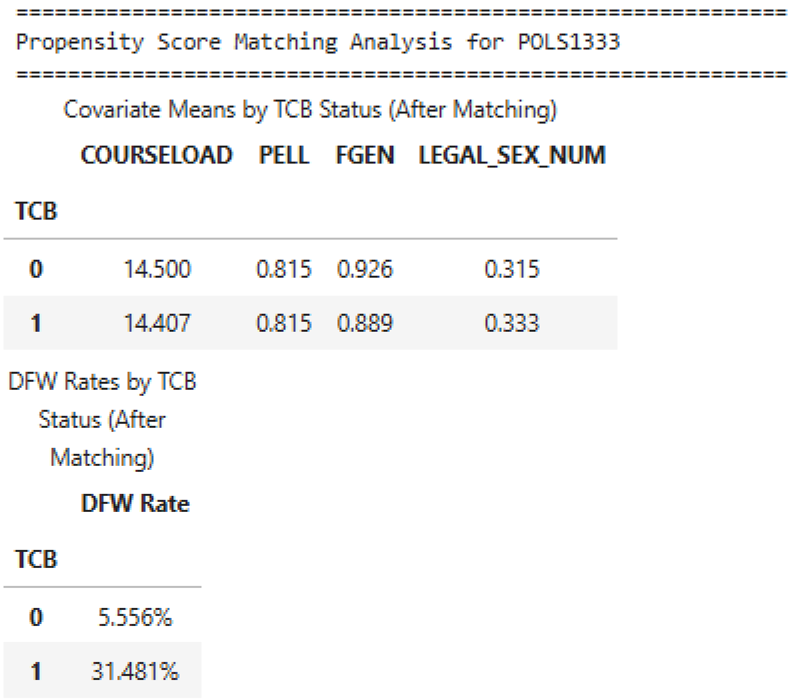
```
============================================================
Propensity Score Matching Analysis for MATH1233
============================================================
```

Covariate Means by TCB Status (After Matching)

	COURSELOAD	PELL	FGEN	LEGAL_SEX_NUM
TCB				
0	13.370	0.778	0.926	0.481
1	13.444	0.778	0.852	0.481

DFW Rates by TCB Status (After Matching)

	DFW Rate
TCB	
0	59.259%
1	88.889%



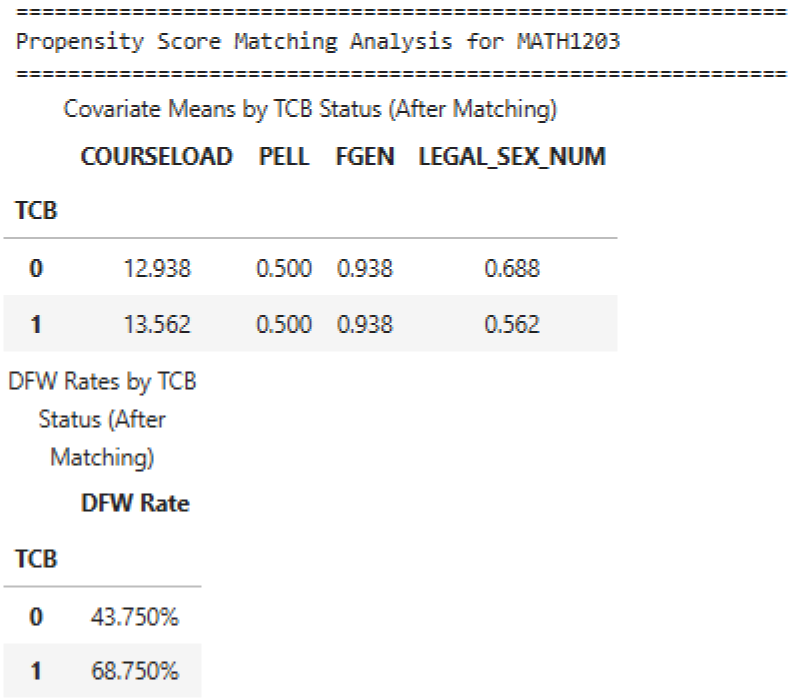
```
============================================================
Propensity Score Matching Analysis for POLS1333
============================================================
```

Covariate Means by TCB Status (After Matching)

TCB	COURSELOAD	PELL	FGEN	LEGAL_SEX_NUM
0	14.500	0.815	0.926	0.315
1	14.407	0.815	0.889	0.333

DFW Rates by TCB Status (After Matching)

TCB	DFW Rate
0	5.556%
1	31.481%



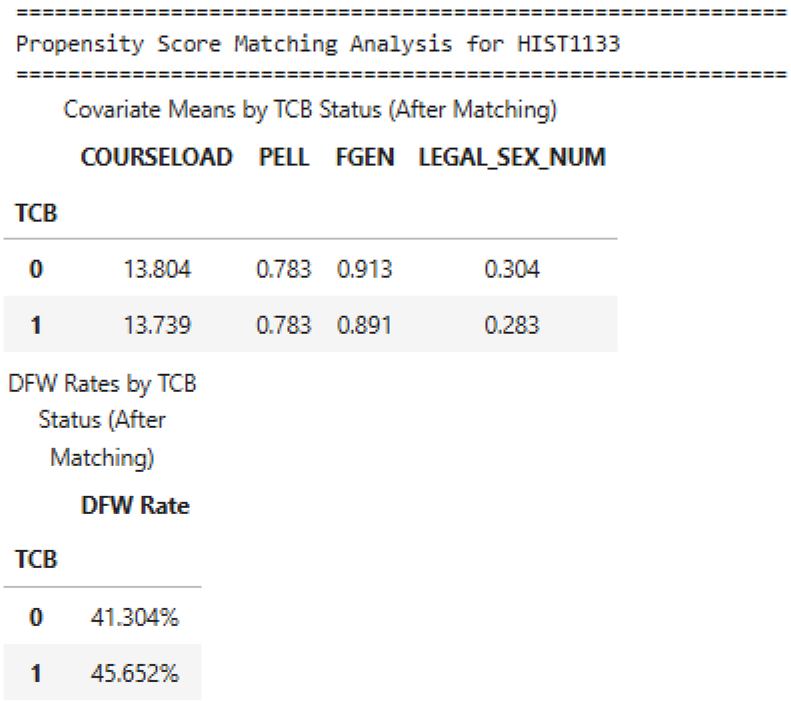
```
============================================================
Propensity Score Matching Analysis for MATH1203
============================================================
```

Covariate Means by TCB Status (After Matching)

TCB	COURSELOAD	PELL	FGEN	LEGAL_SEX_NUM
0	12.938	0.500	0.938	0.688
1	13.562	0.500	0.938	0.562

DFW Rates by TCB Status (After Matching)

TCB	DFW Rate
0	43.750%
1	68.750%



```
============================================================
Propensity Score Matching Analysis for HIST1133
============================================================
```

Covariate Means by TCB Status (After Matching)

TCB	COURSELOAD	PELL	FGEN	LEGAL_SEX_NUM
0	13.804	0.783	0.913	0.304
1	13.739	0.783	0.891	0.283

DFW Rates by TCB Status (After Matching)

TCB	DFW Rate
0	41.304%
1	45.652%