Daytime vs Evening Attendance Effects on Undergraduate Student Dropout

Natalie Huante

Chapman University

CPSC 540-01 Statistical Machine Learning I

Dr. Chelsea Parlett

10/29/2024

Abstract

This report observes a dataset held by the University of California in Irvine related to students enrolled in various undergraduate programs and containing information from the time of their enrollment, academic performance from the first two semesters, and more. The paper aims to calculate the effect of whether student attendance in the daytime or evening has an effect on whether they remain in school or dropout. Included in the discussion are methods of how to control for any confounding variables, limitations of the analysis, etc.

Daytime vs Evening Attendance Effects on Undergraduate Student Dropout

This paper consists of three main sections. The first section focuses on describing the dataset and using exploratory data analysis to elaborate on what the data shows. The second section uses the information collected in the first section to propose an appropriate method of analysis to answer our research question. The third section reports the findings from running the proposed models and discusses the significance of the results as well as limitations, applications, and other reflective thoughts.

Introduction to the Dataset

The dataset used for our analysis can be found in the UC Irvine Machine Learning Repository. It was donated to UCI in December of 2021 and was supported by the program SATDAP - Capacitação da Administração Pública in Portugal. UCI took the data from this program and pre-processed it in order to handle missing values, unexplainable outliers, etc. Therefore, the dataset used in this paper's analysis is the product of this pre-processing and does not have any missing values itself.

The data collected is related to students who enrolled in various undergraduate degrees and includes information about the student's enrollment, academic performance, and target outcome. The target outcome has three possible values: dropout, enrolled, graduated. There are a total of 4424 students who were observed and 36 features per student noted. Each feature represents some piece of information about the student such as their nationality, parent's profession, age at time of enrollment, etc. The main two we focus on are the features of daytime vs evening attendance, which represents whether the student attended classes during the day or evening, and the target outcome, which represents the student's enrollment status. Before we tackle which other variables (or features) we should include in our discussion and our model, we must introduce our research question.

Given the data, the question this paper aims to answer is the following: what is the direct effect of daytime versus evening attendance on a student's enrollment status or dropout. Initially, it is apparent that there could be, and surely are, many different features that could impact whether a student will drop out of their undergraduate program or not. However, the goal of this paper is to find out what the direct effect of daytime versus evening attendance is after considering any other possible features that might contribute to the same outcome. In order to do so, we must identify and control for any confounding variables or mediators that might be present in the dataset. We will discuss how to control for these later, when we introduce our proposed models. Now that we have our research question to discuss, we will describe the data in more detail so that we can identify any relevant variables and interesting patterns.

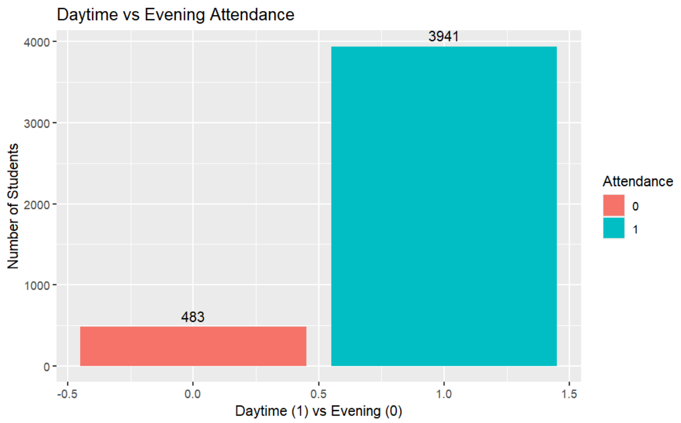
The table below includes all the features within the dataset, what values they may hold, and what they mean.

Variable Name	Type	Description
Marital Status	Integer	1-single, 2-married, 3-widower, 4-divorced, 5-facto union, 6-legally separated
Application Mode	Integer	1-1st Phase (general), 2-Ordinance No 612/935, 7-Holders of other higher courses, 10-Ordinance No 854-B/99, 15-International Student, 16-1st Phase (Madeira Island), 17-2nd Phase (general) 18-3rd Phase (general), 26-Ordinance No 533-A/99, 3-Over 23 years old, 42-Transfer, 43-Change of Course, 44-Tech Specialization Diploma Holder, 51-Change of Institution/Course, 53-Short Cycle Diploma Holder, 57-Change of Institution/Course (International)
Application Order	Integer	Between 0 (first choice) and 9 (last choice)
Course	Integer	33 - Biofuel Production Technologies 171 - Animation and Multimedia Design 8014 - Social Service (evening attendance) 9003 - Agronomy 9070 - Communication Design 9085 - Veterinary Nursing 9119 - Informatics Engineering 9130 - Equinculture 9147 - Management 9238 - Social Service 9254 - Tourism 9500 - Nursing 9556 - Oral Hygiene 9670 - Advertising and Marketing Management 9773 - Journalism and Communication 9853 - Basic Education 9991 - Management (evening attendance)
Daytime/Evening Attendance	Integer	1 - daytime, 0 - evening
Previous Qualification	Integer	1 - Secondary education 2 - Higher education - bachelor's degree 3 - Higher education - degree 4 -

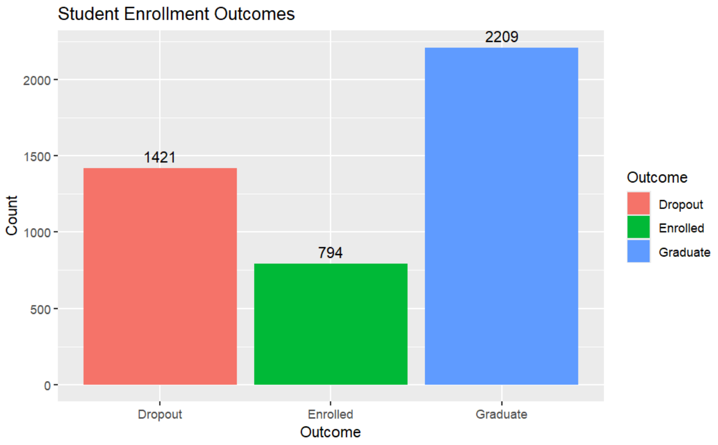
		Higher education - master's 5 - Higher education - doctorate 6 - Frequency of higher education 9 - 12th year of schooling - not completed 10 - 11th year of schooling - not completed 12 - Other - 11th year of schooling 14 - 10th year of schooling 15 - 10th year of schooling - not completed 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv. 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 42 - Professional higher technical course 43 - Higher education - master (2nd cycle)
Previous qualification (grade)	Continuous	Grade of previous qualification (between 0 and 200)
Nationality	Integer	1 - Portuguese; 2 - German; 6 - Spanish; 11 - Italian; 13 - Dutch; 14 - English; 17 - Lithuanian; 21 - Angolan; 22 - Cape Verdean; 24 - Guinean; 25 - Mozambican; 26 - Santomean; 32 - Turkish; 41 - Brazilian; 62 - Romanian; 100 - Moldova (Republic of); 101 - Mexican; 103 - Ukrainian; 105 - Russian; 108 - Cuban; 109 - Colombian
Mother’s Qualification	Integer	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) …
Father’s Qualification	Integer	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) …
Mother’s Occupation	Integer	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers …
Father’s Occupation	Integer	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers …
Admission Grade	Continuous	Admission grade (between 0 and 200)
Displaced	Integer	1 - yes, 0 - no
Educational Special Needs	Integer	1 - yes, 0 - no
Debtor	Integer	1 - yes, 0 - no
Tuition Fees Up to Date	Integer	1 - yes, 0 - no
Gender	Integer	1 - male, 0 - female
Scholarship Holder	Integer	1 - yes, 0 - no
Age at Enrollment	Integer	Age of student at time of enrollment
International	Integer	1 - yes, 0 - no
Curricular Units 1st sem (credited)	Integer	Number of curricular units credited in the 1st semester
Curricular Units 1st sem (enrolled)	Integer	Number of curricular units enrolled in the 1st semester

Curricular Units 1st sem (evaluations)	Integer	Number of evaluations to curricular units in the 1st semester
Curricular Units 1st sem (approved)	Integer	Number of curricular units approved in the 1st semester
Curricular Units 1st sem (grade)	Integer	Grade average in the 1st semester (between 0 and 20)
Curricular Units 1st sem (without evaluations)	Integer	Number of curricular units without evaluations in the 1st semester
Same data for 2nd sem	Integer	Same metrics as noted for 1st sem academic performance
Unemployment Rate	Continuous	Unemployment rate (%)
Inflation Rate	Continuous	Inflation Rate (%)
GDP	Continuous	GDP
Target	Categorical	Three category classifications (dropout, enrolled, graduate) at the end of the normal duration of the course

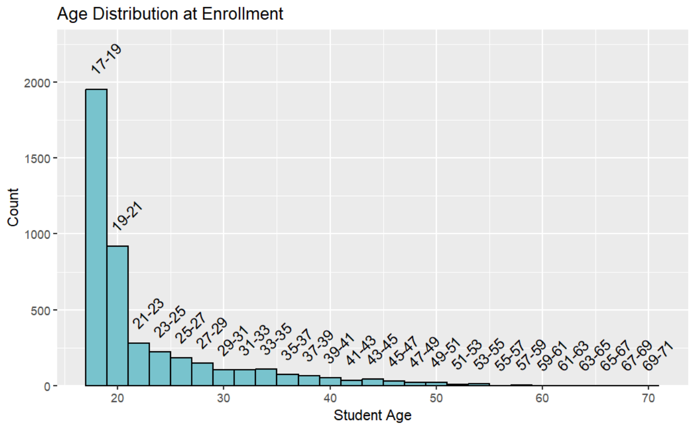
Given all of these features to consider, we graphed various relationships between them that we considered to be worth exploring and visualizing based on our team's knowledge of student enrollment and student life in general. Among these relationships are the following:



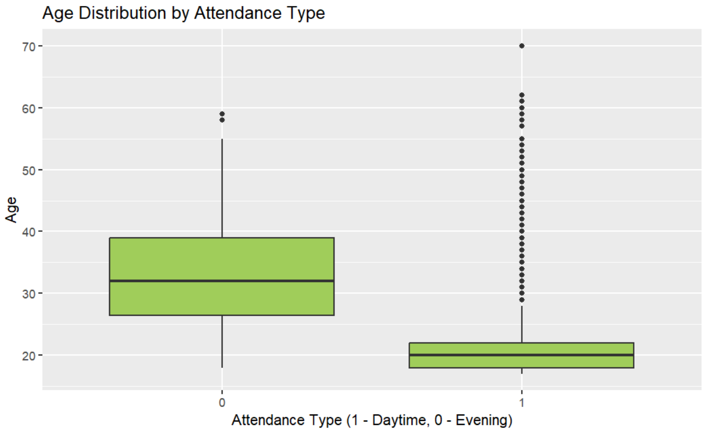
1. Number of Students Attending Daytime vs Evening Classes



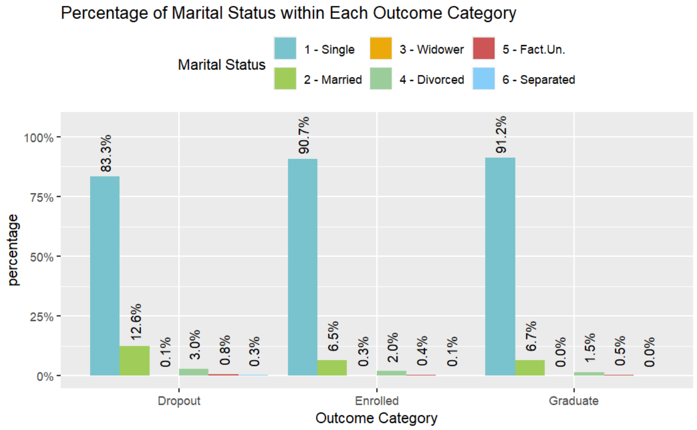
2. Number of Students by Outcome Category



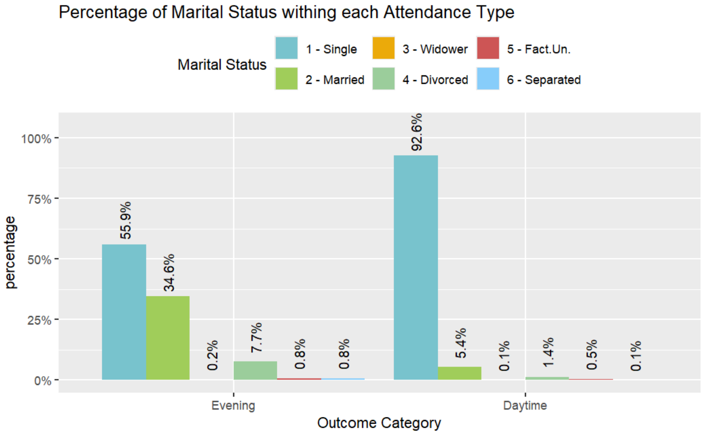
3. Student Ages at Time of Enrollment



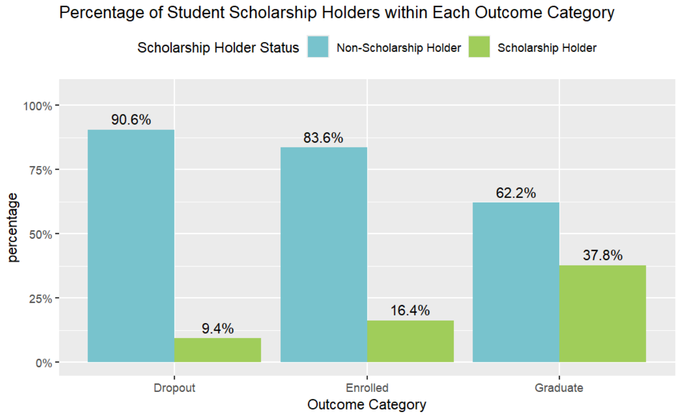
4. Student Ages at Time of Enrollment by Attendance Type



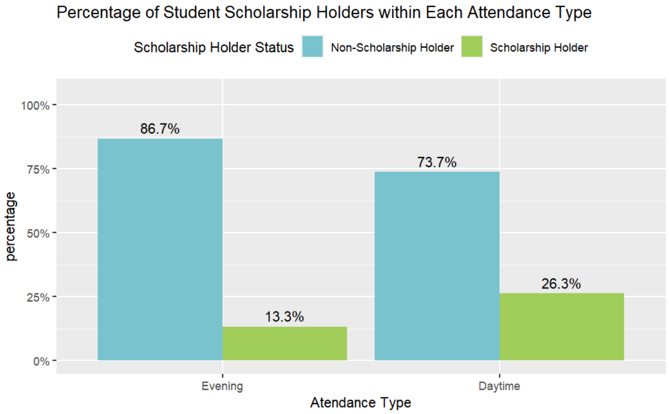
5. % of Students in each Marital Status Category by Outcome Category



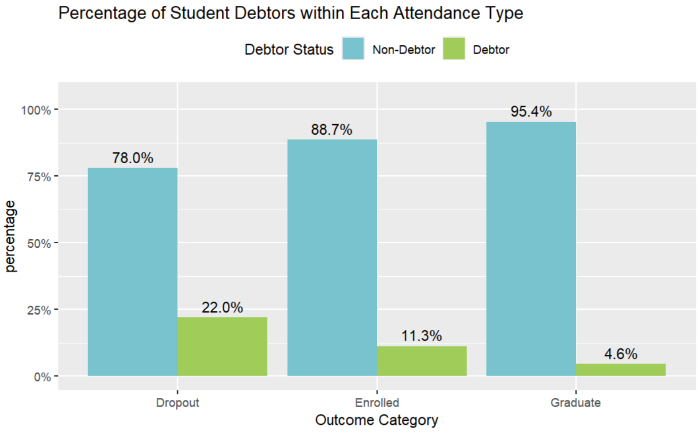
6. % of Students in each Marital Status Category by Attendance Type



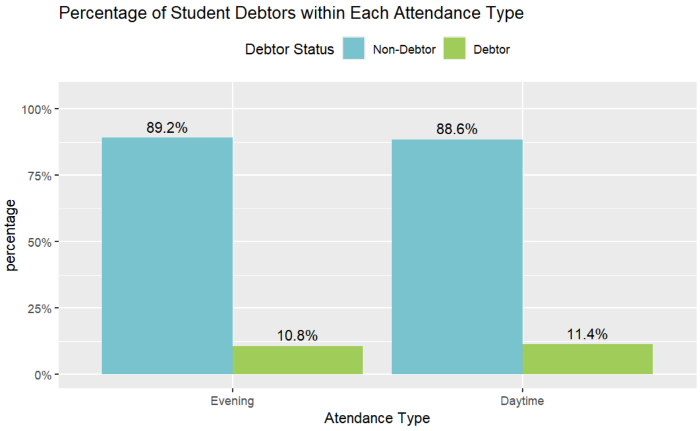
7. % of Student Scholarship Holders by Outcome Category



8. % of Student Scholarship Holders by Attendance Type



9. % of Student Debtors per Outcome Category



10. % of Student Debtors per Attendance Type

Having visualized a few graphs, we can make the following observations and along with the knowledge our team has on collegiate student life make the following considerations about what we might need to incorporate into our model when we begin to calculate causal estimates.

Graph 1 shows that the dataset we are using observed a much higher proportion of students who are attending daytime classes (3941 students) compared to evening classes (483 students). This is important because it means that whenever we are comparing the two groups, we have a lot more data to go off of for the daytime section and a lot fewer data points to use for the evening section. Therefore, any calculations or analyses we run on this sample of students could be more accurate for the daytime group than it is for the evening group, resulting in bias. This can make our results less reliable, so we should account for this in our model.

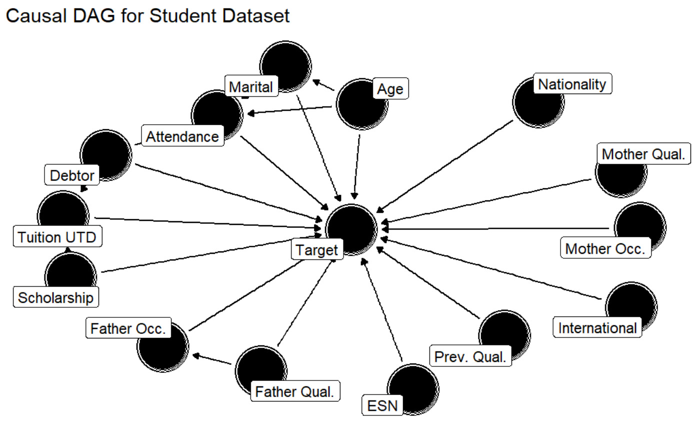
Graphs 3 and 4 are related to student ages. The latter shows the age distribution for each attendance type. We can see that those who attend evening classes are on average over 30 years old while those who attend daytime classes are on average about 20 years old. This is important because if age influences attendance type and other factors such as familial responsibilities and personal life commitments that in turn affect if a student drops out of school, then we should control for the student's age. Otherwise, age could be acting as a mediator or confounder between our variables.

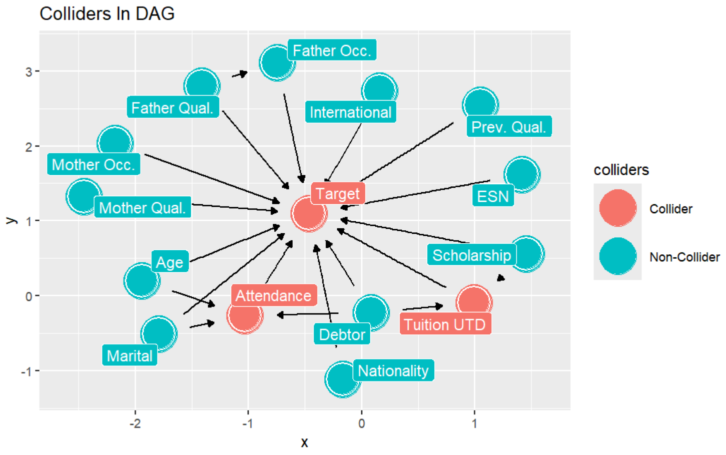
Somewhat similarly, Graphs 5 and 6 show us the relationship of student marital status by outcome category and attendance type, respectively. Students who are married make up 30% more of the evening group than the daytime group, again suggesting that marital status might indicate other responsibilities or factors that are true for those students, which could potentially impact them choosing evening classes in the first place and/or considering dropping out more than daytime students.

The last four graphs show similar comparisons by outcome category and attendance type. The first two are related to the percentage of scholarship holders in each group while the last two are related to student debtors in each group. We make similar observations with these visuals. Student debtors make up a higher percentage of the dropout group (22%) than they do in the enrolled (11.3%) and graduated (4.6%) groups combined. Students who hold scholarships make up a smaller percentage in the dropout group (9.4%) than they do in the enrolled (16.4%) or graduated group (37.8%). This could suggest that a student's ability to pay for schooling via a scholarship impacts their decision to stay in school and work towards graduation rather than prioritizing working a job over schooling. So, this means we might want to consider controlling for scholarships in our models as a confounding variable.

The visuals above provide only a couple examples of potential confounding variables in our dataset such as age, marital status, whether a student holds a scholarship, and whether a student has debt. These variables are examples of features that can influence a student's decision to take evening vs daytime classes as well as their decision to drop out of their undergraduate studies. If we do not incorporate these considerations into our model, our results may be biased and unreliable, therefore, not ideal for generalizing to a population.

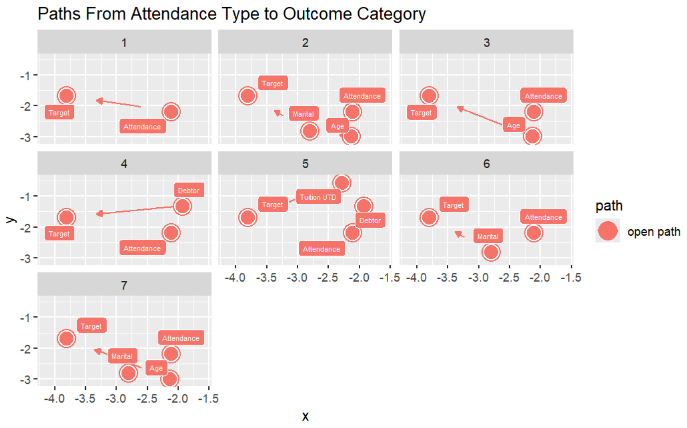
Given these examples, we discussed all 37 features in the dataset to identify potential causal relationships between them before we created our model. The figure below shows these relationships based on the graphs produced above and our team's background knowledge of collegiate student life and is in the form of a Directed Acyclic Graph (DAG). Each node represents a feature and an arrow represents a casual relationship between two nodes. So, the DAG represents any causal relationships between the features (if any one impacts another) as well as with the target outcome (if they impact whether the student drops out or continues).

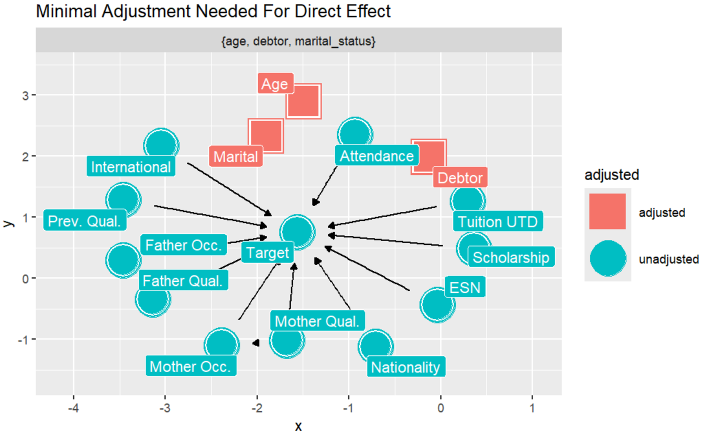




11. Causal DAG for Dataset Features

12. Colliders in Causal Dag





13. Possible Paths from Exposure to Outcome

14. Minimal Adjustments Needed for the Direct Effect from Exposure to Outcome

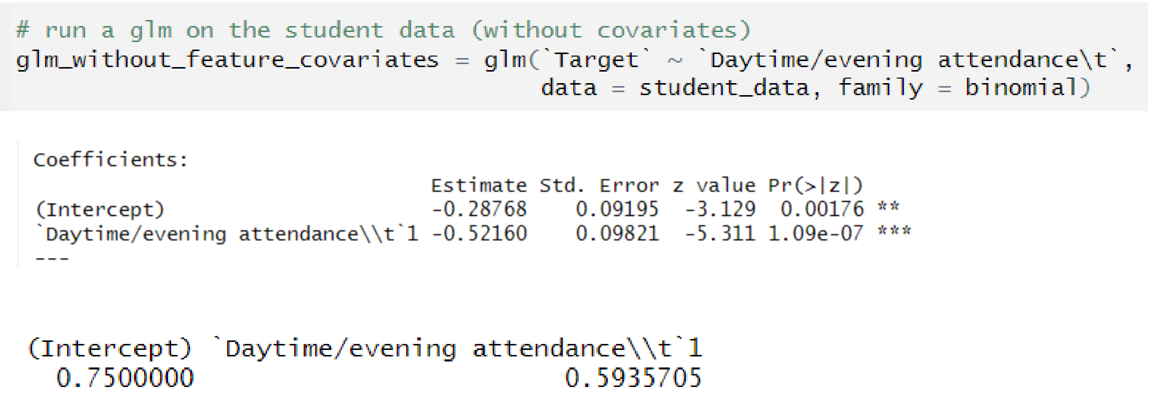
Proposed Analysis

Having discussed how our dataset is structured, relevant variables and relationships, and our research question, we can now shift into which analysis our team has proposed. A viable model should allow us to consider any variables, other than the exposure variable, that might have an impact on the treatment group and, therefore, might bias our results if not accounted for. Based on our previous discussions, some of these variables include age, debt status, marital status, and scholarship status. For example, if a student has debt, they might prefer evening classes so they can work during the day. So, if our model determines that evening attendance increases dropout likelihood, we must consider how much of that impact is actually due to a student's debt status rather than assuming it is purely the result from the attendance type. Our team created three levels of a model, all of which are proposed below.

Proposal 1: GLM with Features as Covariates

One way to consider the effects of any confounding variables is to ask the following: what is the effect of X (exposure) on Y (treatment) after subtracting the influence of Z (confounding) on Y. We can use a generalized linear model (GLM) and run a logistic regression to do so. This will allow us to predict a binary outcome, which in this case will be whether the student will drop out or not. We can begin by comparing the causal estimates calculated by this model before and after adding other features as covariates to the GLM. First, the target outcome needed to be transformed from its 3 categories to 2 where 1 represents the student did drop out and 0 represents the student is either enrolled or a graduate. The features used as covariates in the logistic regression were also transformed into factors so that the GLM would create dummy variables representative of membership in each possible category rather than treating their values numerically. Finally, we ran two models: one without covariates and one with 4 covariates.

Proposal 1 Results: GLM with Features as Covariates

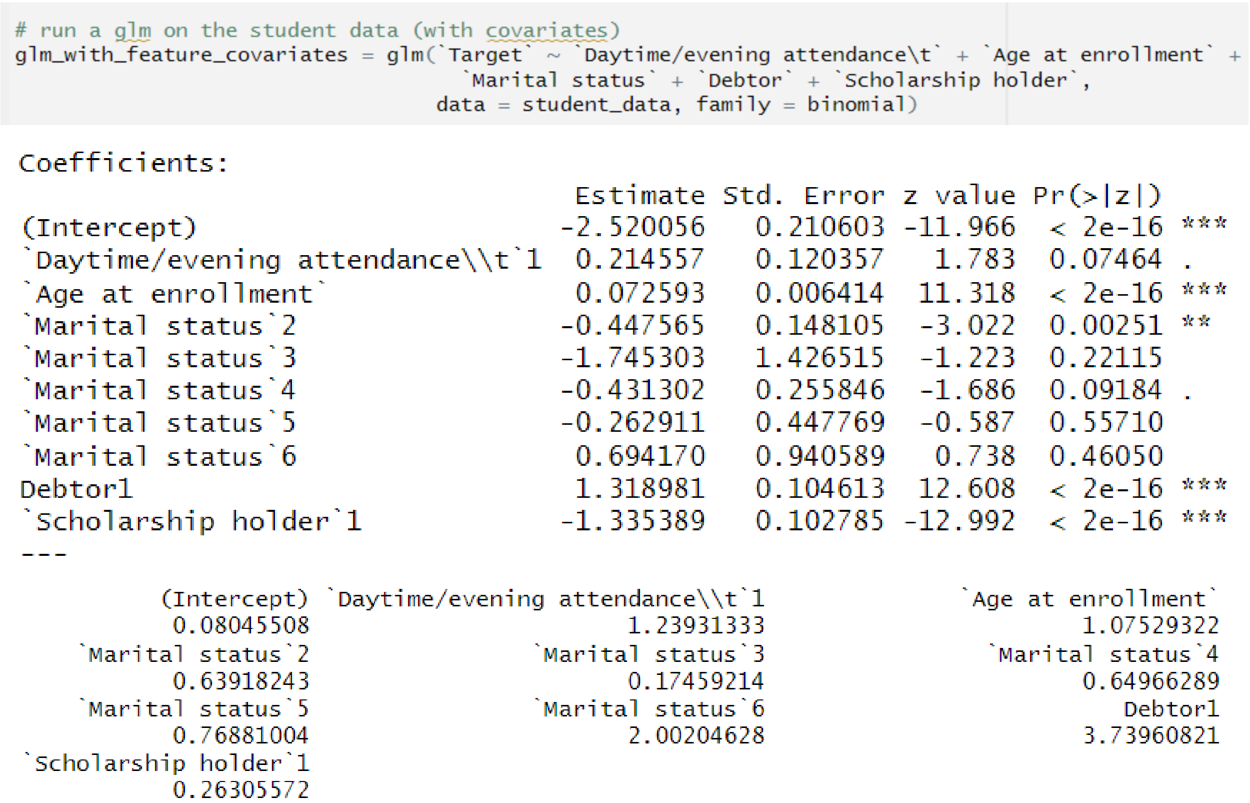


```
# run a glm on the student data (without covariates)
glm_without_feature_covariates = glm(`Target` ~ `Daytime/evening attendance\t`,
                                     data = student_data, family = binomial)
```

```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -0.28768    0.09195  -3.129  0.00176 **
`Daytime/evening attendance\\t`1 -0.52160    0.09821  -5.311 1.09e-07 ***
---
```

```
(Intercept) `Daytime/evening attendance\\t`1
  0.7500000                        0.5935705
```

15. GLM Output Without Covariates



```
# run a glm on the student data (with covariates)
glm_with_feature_covariates = glm(`Target` ~ `Daytime/evening attendance\t` + `Age at enrollment` +
                                  `Marital status` + `Debtor` + `Scholarship holder`,
                                data = student_data, family = binomial)
```

```
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -2.520056   0.210603 -11.966  < 2e-16 ***
`Daytime/evening attendance\\t`1  0.214557   0.120357   1.783  0.07464 .
`Age at enrollment`                 0.072593   0.006414  11.318  < 2e-16 ***
`Marital status`2                  -0.447565   0.148105  -3.022  0.00251 **
`Marital status`3                  -1.745303   1.426515  -1.223  0.22115
`Marital status`4                  -0.431302   0.255846  -1.686  0.09184 .
`Marital status`5                  -0.262911   0.447769  -0.587  0.55710
`Marital status`6                   0.694170   0.940589   0.738  0.46050
Debtor1                             1.318981   0.104613  12.608  < 2e-16 ***
`Scholarship holder`1              -1.335389   0.102785 -12.992  < 2e-16 ***
---
```

```
          (Intercept) `Daytime/evening attendance\\t`1              `Age at enrollment`
           0.08045508                       1.23931333                       1.07529322
     `Marital status`2                `Marital status`3                `Marital status`4
           0.63918243                       0.17459214                       0.64966289
     `Marital status`5                `Marital status`6                          Debtor1
           0.76881004                       2.00204628                       3.73960821
`Scholarship holder`1
           0.26305572
```

16. GLM Output With Covariates

Looking at figure 15, the intercept of the first GLM output means that when the attendance type is 0 (the student attends evening classes) the log odds of our target being 1 (the student dropping out) is -0.2768. If the attendance changes to 1 (the student attends daytime classes), then the log odds decrease by -0.52160. When we convert these numbers to odds, we can interpret them a little more intuitively. In the third picture of Figure 15, we can see that the intercept in odds means that students have a baseline 0.75 odds of dropping out when they attend evening classes. If they attend daytime classes, their odds of dropping out are 0.59 times the odds of those who attend evening classes. The p-value for this is categorized as statistically significant at p = 1.09e-07 < 0.001. So, we could conclude that students who attend daytime classes have a lower likelihood of dropping out compared to those who attend evening classes. However, the issue here is that we are not considering any of the possible confounding or mediator variables. Our second model does incorporate this.

Moving on to Figure 16, we can observe that the second GLM's intercept has decreased to -2.520056. This represents the log odds for a student with the baseline groups of all the predictors (evening student, single, non-debtor, non-scholarship holder, etc). Converted to odds, the intercept is 0.08, so a student with those baseline groups has 0.08 odds of dropping out, which is not very high. Looking at the coefficients for each variable we note the following:

- The odds of a student dropping out if we change their attendance type to daytime is 1.24 times greater relative to an evening student's. The p-value for this is not statistically significant at p = 0.07464 !< 0.001.
- The odds of a student dropping out increases by 1.075 times the previous for each increase of 1 year. So, older students have higher odds of dropping out than younger students.
- The odds of a married student dropping out is 0.639 times the odds of that for a single student. This is about a 36% reduction in dropout odds.

- The odds of a student with debt dropping out increases by 3.739 times the odds of that for a student without debt. So, a student's debt is very significant; students with debt have a higher likelihood of dropping out than otherwise.
- The odds of a student with a scholarship dropping out is 0.263 times the odds of that for a non-scholarship holding student; this is a 74% reduction in dropout odds. So, a student with a scholarship is less likely to drop out than otherwise.

Having considered these additional covariates in our model, we can make a different conclusion based on the coefficients. It appears that some of the covariates such as whether a student has debt or not are much more significant than whether the student attends evening or daytime classes in their decision to drop out of college. With this output, we would conclude that the p-value indicates the coefficient is not statistically significant and therefore there is not a significant impact of attendance type on the odds of a student dropping out. Instead, the change in odds could be explained by other factors such as debt or scholarship holding.

Proposal 2: GLM with Features as Covariates and Inverse Probability Weighting

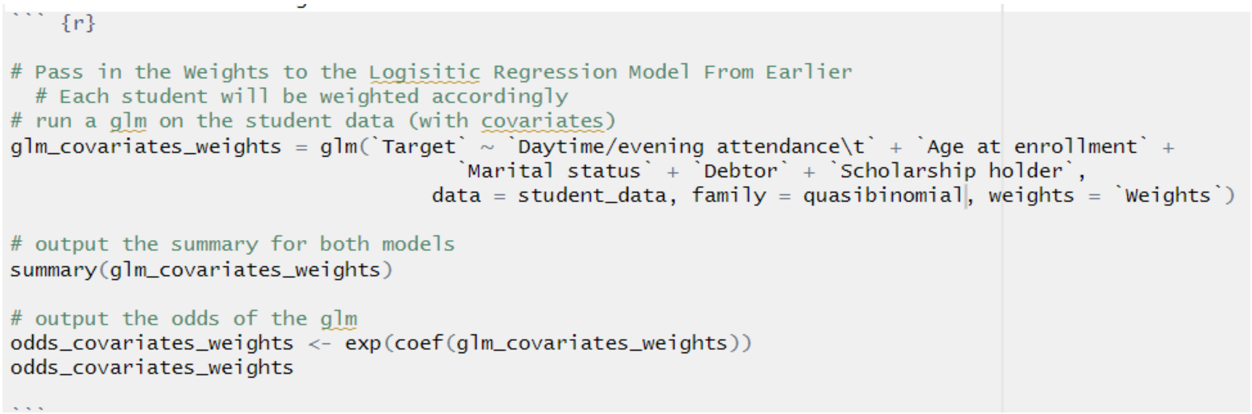
Due to the attendance type imbalance in our dataset discussion previously, we can improve the GLM model to account for bias in this respect by using propensity scores. There are a few ways this can be done, two of which are matching or applying weights to our samples based on their propensity scores. The first would imply matching pairs of students with similar scores and splitting them into two comparable groups for our control and treatment groups. The latter would require applying a weight to each sample relative to its propensity score.

The propensity score of a sample represents its likelihood of being in the treatment group, or of receiving treatment. In this case, the baseline for attendance type is evening classes, so each sample (or student's) propensity score represents the likelihood of them attending daytime classes (or of attendance type being equal to 1). The weight for each sample is then calculated by taking the inverse of the propensity score (1/prop_score). Therefore, it will give a higher weight to samples who are members of groups that are underrepresented and a lower weight to samples who are members of groups who are overrepresented in the data. This is the method we will use to help reduce the bias of the attendance type imbalance in our dataset, in an effort to get more accurate comparisons between the daytime and evening attendance groups.

To accomplish this in our model, we will first run a GLM to predict the attendance type of each student and calculate the propensity scores given the GLM we produce. Then, we will take the inverses of those scores and fit the weights calculated into the GLM that predicts the log odds of the student dropping out. So, we are essentially adding steps before which calculate the weights of each student and then add the weights to our model.

Another modification to note is the change from binomial to quasibinomial for the family parameter for the last GLM in our code. The reason for this is due to the weights not being a binomial categorical variable. In other words, the weights are not 0 or 1 but rather a probability ranging from 0 to 1. The quasibinomial family parameter makes this distinction.

Proposal 2 Results: GLM with Features as Covariates and Inverse Probability Weighting



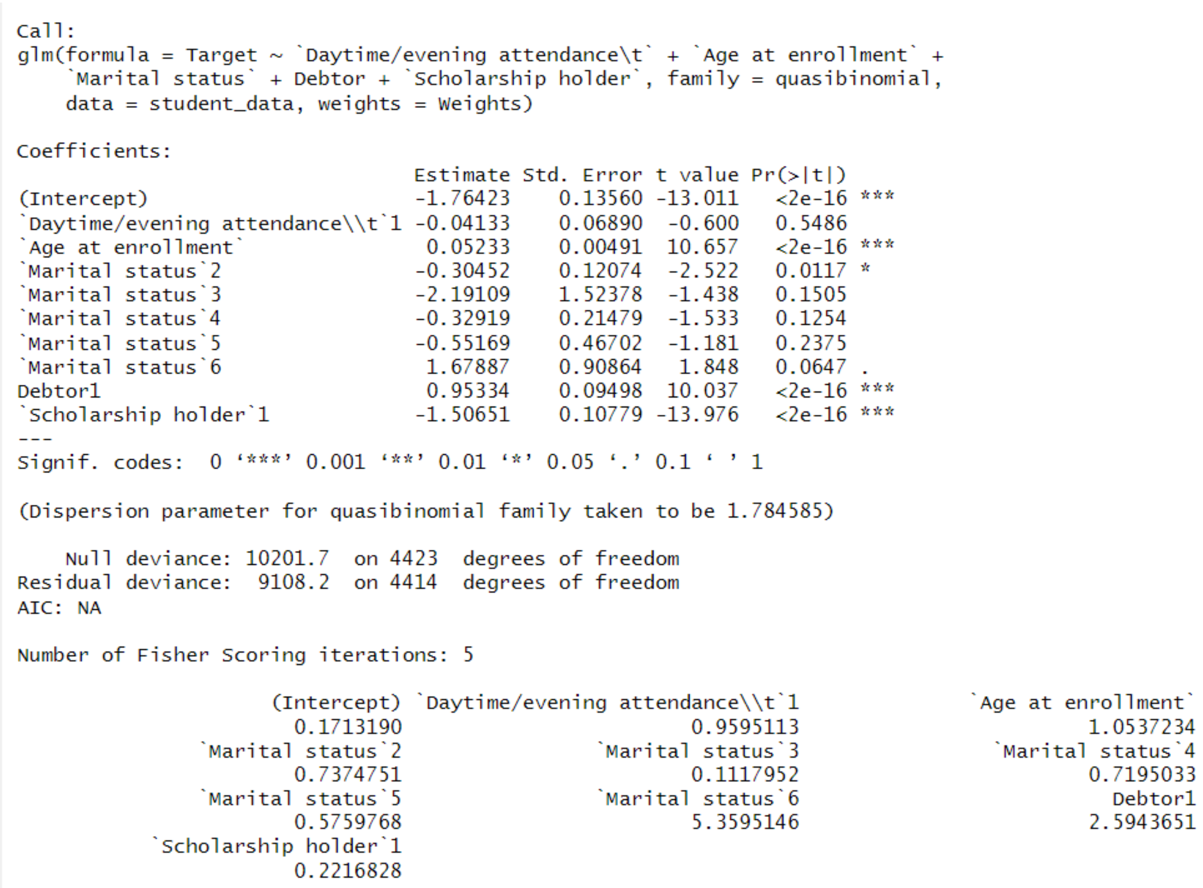
~~~
```{r}

# Pass in the Weights to the Logisitic Regression Model From Earlier
  # Each student will be weighted accordingly
# run a glm on the student data (with covariates)
glm_covariates_weights = glm(`Target` ~ `Daytime/evening attendance\t` + `Age at enrollment` +
                                    `Marital status` + `Debtor` + `Scholarship holder`,
                                  data = student_data, family = quasibinomial, weights = `Weights`)

# output the summary for both models
summary(glm_covariates_weights)

# output the odds of the glm
odds_covariates_weights <- exp(coef(glm_covariates_weights))
odds_covariates_weights

```
~~~



```
Call:
glm(formula = Target ~ `Daytime/evening attendance\t` + `Age at enrollment` +
    `Marital status` + Debtor + `Scholarship holder`, family = quasibinomial,
    data = student_data, weights = Weights)

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.76423    0.13560 -13.011   <2e-16 ***
`Daytime/evening attendance\\t`1 -0.04133    0.06890  -0.600   0.5486
`Age at enrollment`              0.05233    0.00491  10.657   <2e-16 ***
`Marital status`2               -0.30452    0.12074  -2.522   0.0117 *
`Marital status`3               -2.19109    1.52378  -1.438   0.1505
`Marital status`4               -0.32919    0.21479  -1.533   0.1254
`Marital status`5               -0.55169    0.46702  -1.181   0.2375
`Marital status`6                1.67887    0.90864   1.848   0.0647 .
Debtor1                          0.95334    0.09498  10.037   <2e-16 ***
`Scholarship holder`1           -1.50651    0.10779 -13.976   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.784585)

    Null deviance: 10201.7  on 4423  degrees of freedom
Residual deviance:  9108.2  on 4414  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

                     (Intercept) `Daytime/evening attendance\\t`1              `Age at enrollment`
                       0.1713190                        0.9595113                        1.0537234
                `Marital status`2                `Marital status`3                `Marital status`4
                       0.7374751                        0.1117952                        0.7195033
                `Marital status`5                `Marital status`6                          Debtor1
                       0.5759768                        5.3595146                        2.5943651
            `Scholarship holder`1
                       0.2216828
```

Having added the weights to each student in our data, we can compare the data to that of our model before we applied weights. To start, the coefficient for the independent effect of the attendance type on a student's dropout likelihood decreased from 0.214557 to -0.04133 with a p-value of 0.5486 !< 0.01, meaning it is not significant. Even though the unweighted model also led us to the same conclusion, the weighted model now indicates that the attendance type is less significant than we previously thought.

We can also make the following comparisons and observations about the other predictors:

- The coefficient for a student's age's effect decreases from 0.072593 to 0.05233 and remains strongly significant with a p-value of <2e-16 in both models.

- The coefficient for a student's marital status is not considered significant unless they have the marital status 2 (married). For this case, the coefficient of a status of 2 decreased from -0.447565 to 0.30452, but remains slightly significant with a p-value of 0.0117. So, the level of significance also decreased with the inclusion of IPW in the model.
- The coefficient for a student's debtor status being 1 decreased from 1.318981 to 0.95334 and remains strongly significant with a p-value of <2e-16 in both models.
- The coefficient for a student's scholarship status being 1 decreased from -1.335389 to -1.50651 and remains strongly significant with a p-value of <2e-16 in both models.

Discussion

By observing the output of our models presented above, we can conclude the following:

1. The attendance type does not significantly influence a student's likelihood of dropping out of their undergraduate program rather than remaining enrolled or graduating. In other words, the causal estimate of attendance type on student dropout likelihood is so small that we consider it to not have a significant effect at all. This means that whether a student attends classes during the daytime or the evening does not have a large enough effect on whether they are likely to drop out or not based on the data.
2. We did observe some other predictors that do significantly affect a student's likelihood of dropping out of their programs. These are the student's age, debt status, and scholarship holding. These three variables had p-values that indicated they have a significant impact and coefficients that describe what that impact is. The odds we calculate indicate the following:

- For each increase of 1 year to a student's age, the odds of them dropping out are 1.0759 times the odds for the previous age.
- If a student has debt, the odds of them dropping out are 3.7396 times the odds of if they did not have debt.
- If a student holds a scholarship, the odds of them dropping out are only 0.2630 times the odds of if they did not hold a scholarship.
- A student's marital status also had a significant effect, but only if the status was that they are married. In this case, their marital status reduced the odds of them dropping out.

Therefore, the results suggest that the attendance type of a student is not a leading or primary factor in their likelihood of dropping out. The independent effect that attendance type has on the outcome is simply not large enough nor significant enough to suggest we draw a different conclusion. However, the age, debt, scholarship holding, and marital status (conditionally) variables do have significant independent effects on the outcome. We could elaborate on why we believe this to be the case. Our team briefly discussed that two of these variables are related to financial needs and responsibilities of a student. So, we might suggest that variables that relate to finances typically affect the outcome. Similarly, the other two variables relate to a student's lifestyle outside of the classroom. For instance, a 22-year old student who is single most likely lives a different lifestyle and has different responsibilities than a 45-year old student who is married. So, we might also suggest that variables that are related to what a student's lifestyle might be outside of school also typically influence the outcome.

These findings can be helpful for undergraduate programs to know so they may apply or prioritize resources to programs that might help students stay in school. Knowing that, for instance, finances play a significant role in a student's decision to drop out or not, schools can use this information to prioritize programs that can help students manage their finances or become aware of other options that may help such as scholarships, grants, etc. Similarly, schools can take a similar approach for factors that affect a student's lifestyle. For example, if students who are older are more likely to drop out than otherwise, a viable option could be to first run another study and collect data to target why this is the case. If, for example, those results show that they feel disconnected and unwelcome on-campus, then the university could launch a program that focuses on helping older students become involved on-campus and prioritizes building the community on-campus. These are only two examples of how the results of this analysis can be applied to real universities or improved upon by other studies.

Limitations

As mentioned earlier, the dataset we use for this analysis has an imbalance of students observed who are members of the evening attendance group versus those who are in the daytime attendance group. This creates an imbalance in the data and therefore could introduce some bias since we have more data points to create more accurate assumptions about one group over the other. Although we incorporated IPW in our model to account for this, our team would be interested in how the analysis would change if the dataset were to be more balanced on this front.

Our analysis also operates on the assumption that we have identified all of the confounding variables to get an accurate estimate on each variable's effect on the target outcome. However, as we discussed causal estimates and potential confounding variables, our team came to the realization that there are many variables we would have liked to see included in the dataset. However, we completed the analysis with what was available. Therefore, there could still be other variables that should be included in the model that are simply not part of this data set and therefore could not be included for the purposes of this report. For instance, if a student grew up within a certain culture, this could influence how they approach the idea of dropping out of college. If their culture has instilled values of academic success and integrity and they value this highly, then that could potentially have an impact on whether or not they see dropping out as a reasonable option. This is a simple example of a variable we could argue we should incorporate, but is not included in the data.

The model only considers whether a student has dropped out or not due to the transformation we completed on the target outcome column in the data. Therefore, the model and its results can only be applied to questions related to the dropout likelihood. If one were to want to expand these findings to analyze the predictors' effects on a student's likelihood of dropping out versus remaining enrolled versus graduating, a multinomial regression would be the best improvement.

Another assumption our logistic regression model runs on is that the relationship between our predictors and our target outcome is linear. However, this might not always be the case. For instance, if the effect of a student's age is not linear as it increases, but rather begins to slope upwards and increase exponentially, then our model would meet one of its limitations. A similar issue would arise if this relationship varies when considering other variables. For instance, the influence of a student's age could differ for different debt statuses or scholarship holding statuses. For these cases, it would be beneficial to shift towards Generalized Additive Models (GAMs) in order to allow us to model non-linear relationships.

Personal Reflection

On a more personal note, this paper involved a good amount of research for me outside of class. Not having taken any data science or statistics classes in the past, the first phase of completing this assignment was to get comfortable with the programming language of R and the R Studio environment. It also made a huge difference to map out and plan out the report conceptually before figuring out the details of how to code the models. Once I got the models implemented, there was some debugging that needed to happen in order to fix the errors that R presented. For example, I had to transform some of the variables in the data set from integers to factors so that the model would interpret them correctly.

Other than that, the only other section that I had trouble with initially was interpreting the coefficients of the model's output. This semester, this has been one of the concepts that I have struggled with understanding. So, in order to interpret what the coefficients meant, I referenced a few online sites and examples related to not only coefficients but also log odds versus odds and likelihood. However, overall I enjoyed working on this paper and learned more about how to implement these models, considerations to make about the data set, and how to interpret a model's output. If I were to do a similar project again, I would like to explore other models or other improvements that could be made on the final model in this analysis. As mentioned in the limitations section, the models are operating on several assumptions as is, so it would be interesting to investigate that more.

References

Barrett, Malcolm. “An Introduction to Ggdag.” R-Project.org, 21 July 2024, cran.r-project.org/web/packages/ggdag/vignettes/intro-to-ggdag.html. Accessed 29 Oct. 2024.

“Causal Diagrams in R.” Lucymcgowan.com, 2020, user2020.lucymcgowan.com/02-dags.html#40. Accessed 29 Oct. 2024.

chelseaparlett. “CPSC540ParlettPelleriti/LectureSlides at Main · Chelseaparlett/CPSC540ParlettPelleriti.” GitHub, 2024, github.com/chelseaparlett/CPSC540ParlettPelleriti/tree/main/LectureSlides. Accessed 29 Oct. 2024.

“Colors in R.” R CHARTS | a Collection of Charts and Graphs Made with the R Programming Language, r-charts.com/colors/.

“FAQ: How Do I Interpret Odds Ratios in Logistic Regression?” UCLA: Statistical Consulting Group. stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/.

in. “What Is Quasi-Binomial Distribution (in the Context of GLM)?” Cross Validated, 28 Mar. 2014, stats.stackexchange.com/questions/91724/what-is-quasi-binomial-distribution-in-the-context-of-glm. Accessed 29 Oct. 2024.

Jankovic, Dina. “A Simple Interpretation of Logistic Regression Coefficients.” Medium, 7 Dec. 2021, towardsdatascience.com/a-simple-interpretation-of-logistic-regression-coefficients-e3a40a62e8cf.

Karim, M. “A Quick Introduction to Ggdag.” Netlify.app, 2021, ggdagcand3.netlify.app/?panelset1=using-dagitty.net2#38. Accessed 29 Oct. 2024.

OpenAI. “Discussion on Interpreting the Coefficients of a Model.” ChatGPT, Accessed 27 OCt 2024.

“Package “Ggraph” Reference Manual.” R-Universe.dev, 2024, thomasp85.r-universe.dev/ggraph/doc/manual.html#ggraph. Accessed 29 Oct. 2024.

“Quasibinomial Model in R Glm() | Random Effect.” Randomeffect.net, randomeffect.net/post/2020/10/12/quasi-binomial-in-r-glm/.

Realinho, Valentim, et al. "Predict Students' Dropout and Academic Success." UCI Machine Learning Repository, 2021, https://doi.org/10.24432/C5MC89.

“Reddit - Dive into Anything.” Reddit.com, 2024, www.reddit.com/r/rstats/comments/6zg2f2/struggling_with_including_weights_in_a_binomial/. Accessed 29 Oct. 2024.

“R: Family Objects for Models.” Stat.ethz.ch, stat.ethz.ch/R-manual/R-devel/library/stats/html/family.html.