Shane Rodricks

CPSC 540 – Statistical Machine Learning I

Homework 1 Report

Background

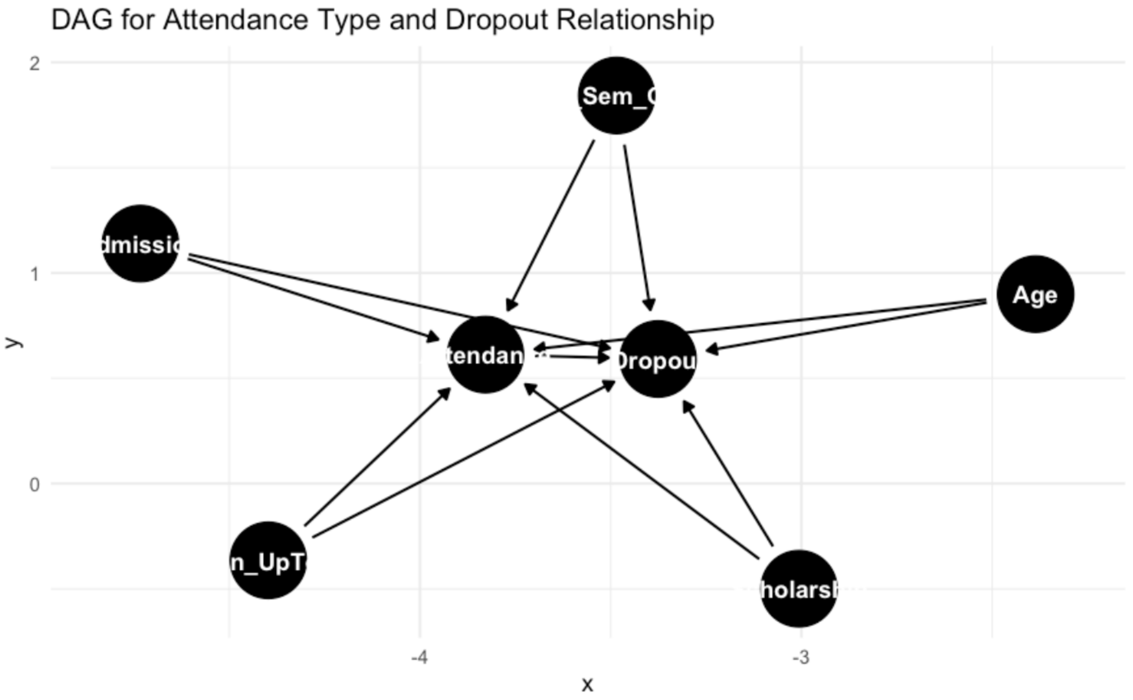
For this project, I used the ‘Student Dropout and Academic Success’ dataset from the UCI Machine Learning Repository to explore the causal relationship of attendance type (daytime vs. evening) on dropout rates among enrolled students. Dropout rates are an important statistic for any university as it directly impacts the student’s success, as well as the institution’s reputation. Using both a baseline logistic regression model and an advanced weighted model, I tried to answer this question: Is there a causal effect of daytime vs. evening attendance on students dropping out?

Dataset Overview:

The dataset contained 36 different features and 4424 instances of those features. The main variables I focused on were:

- Attendance Type (Daytime.evening.attendance) : Indicates whether a student attends daytime or evening classes.
- Target Variable (Target): Binary indicator of dropout status
- Age at Enrollment (Age.at.enrollment): The age of the student at the time of admission
- Admission Grade (Admission.grade): The student’s grade at the time of admission
- Scholarship Holder (Scholarship.holder): Indicates if the student holds a scholarship.
- Tuition Fees Up to Date (Tuition.fees.up.to.date): Shows if the student’s tuition fees are paid up to date.
- First Semester Grade (Cirricular.units.1st.sem.grade): Academic performance in the first semester.

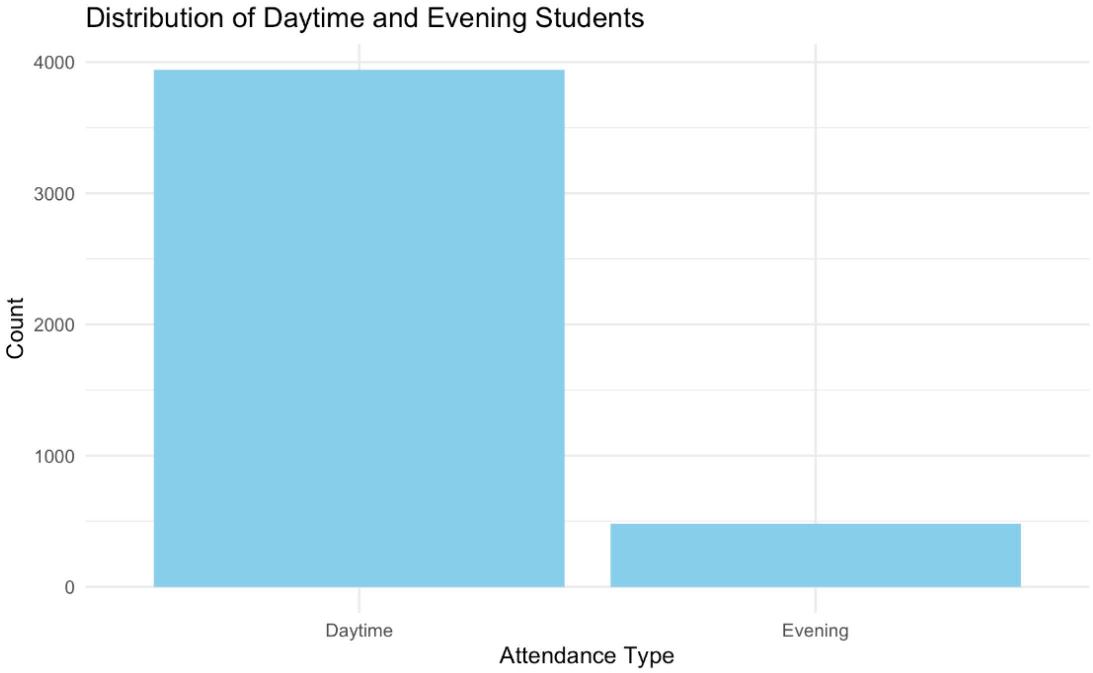
I created a DAG in R to model these relationships to the target variable:



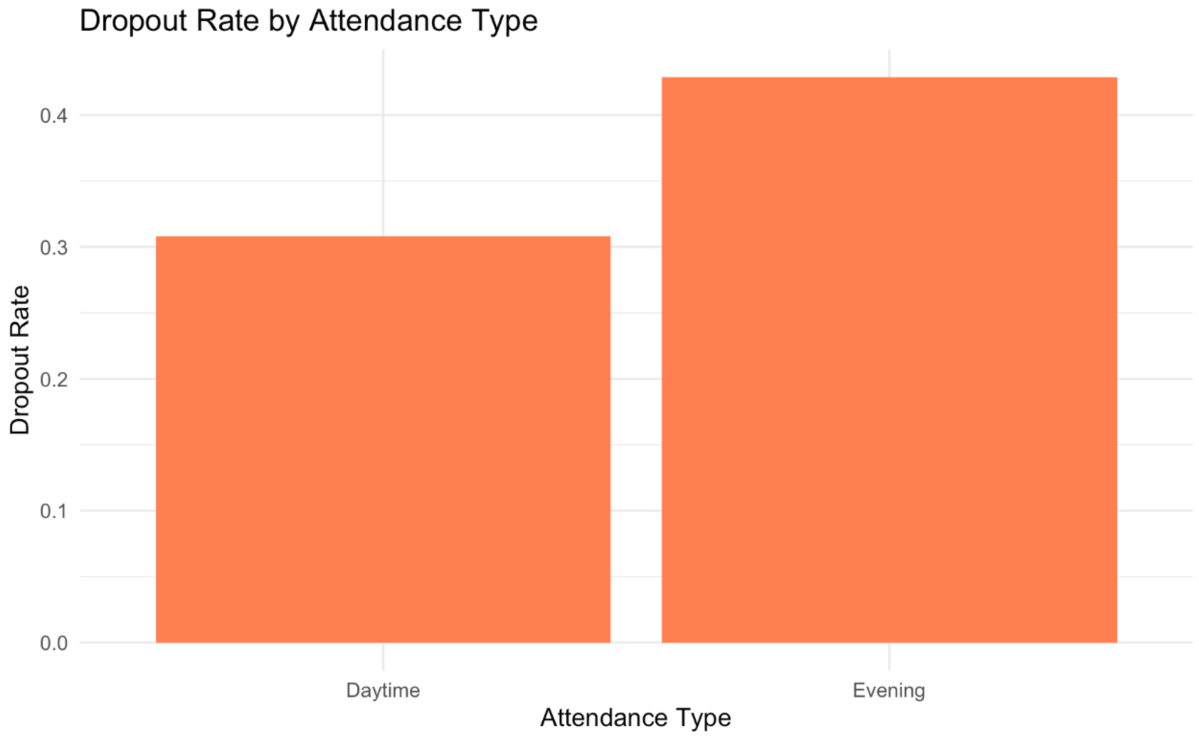
Exploratory Data Analysis (EDA)

To better understand the dataset and find some initial patterns, I conducted EDA steps, visualized distributions, looked for missing values, and analyzed relationships between variables of interest and the target variable. First, I checked to see if there were any missing variables and found none. When I looked at the description of the dataset on the UCI Machine Learning Repository, it said that the people who originally worked on the dataset had ‘performed rigorous data preprocessing to handle data from anomalies, unexplainable outliers and missing values’, so no work had to be done there. Next, I renamed a lot of columns of interest as they contained apostrophes (‘) and standardized the other column names by replacing the non-standard characters with underscores. I did so by using the colnames() function in R. Next, I used the as.factor() function to convert specific variables into factors, which will treat it’s values as categories instead of continuous numbers. I converted the following variables to use in my initial logistic regression model: Daytime.evening.attendance, Age.at.enrollment, Admission.grade, Scholarhsip.holder, Tuition.fees.up.to.date, Curricular.units.1st.sem and Target. For my most important predictor, Daytime.evening.attendance, I divided the variable into two categories (1 for ‘Daytime’ and 0 for ‘Evening’) and can now distinguish between the two class times.

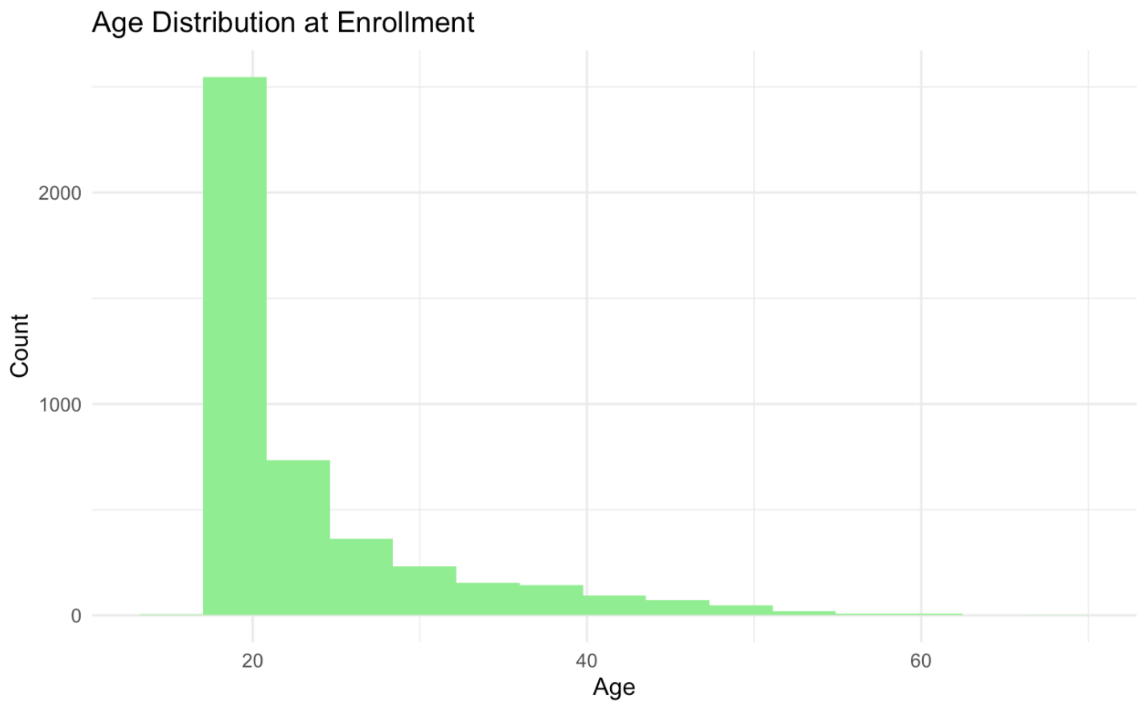
After the data preprocessing, I started EDA to help understand the distributions of key variables and relationships that may confound the effect of attendance type on dropout. Firstly, I looked at the distribution of attendance types by creating a bar plot. This helped me understand the balance between daytime and evening students in the dataset:



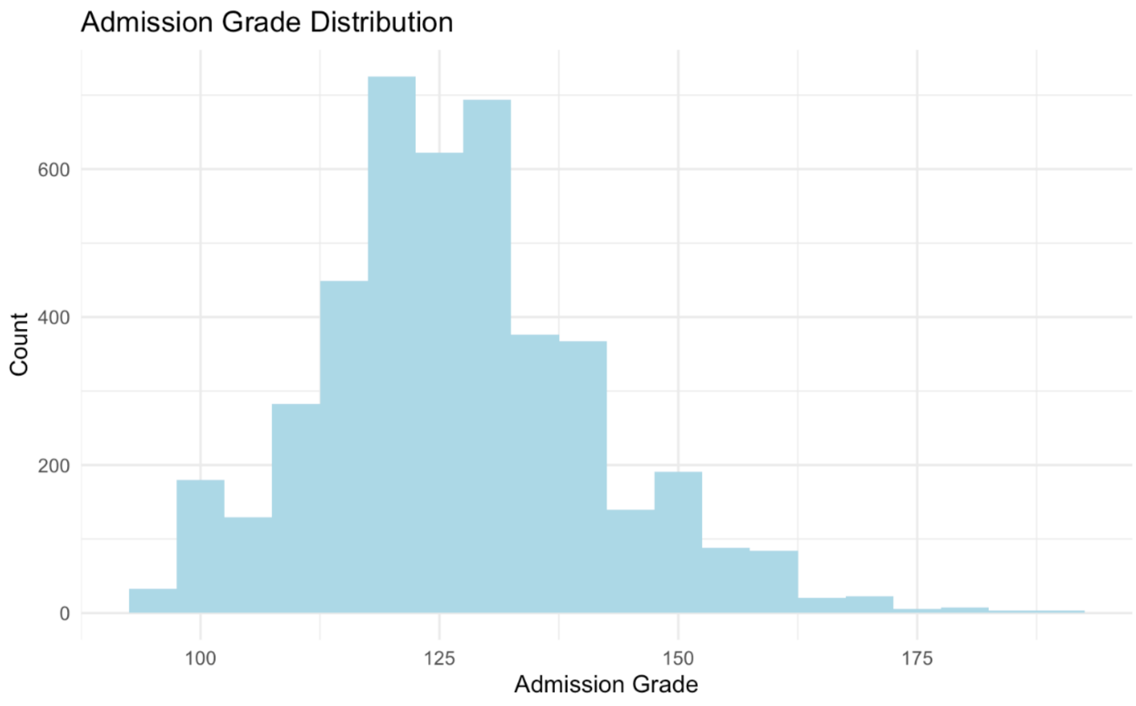
The bar plot indicates a significant imbalance between the two student groups, with a much larger proportion of students attending daytime classes (almost 4000). The large difference in attendance types may introduce a bias, as the smaller evening group may not represent the same variability as the daytime group, which I kept in mind when creating the model. Next, I looked at the distribution of dropout rates when filtered by attendance type to find out if there is an initial relationship between attendance type and dropout rate. First, I had to calculate the dropout rate. I started by recoding the Target variable where Dropout = 1 and others = 0. To calculate the dropout by attendance type, I created a new variable called 'dropout_rate', which grouped the data by attendance type and then calculated the mean of each category in the Target variable (dropout vs. others). This served as a dropout rate basis for each group.



The above plot indicates a higher proportion of evening students dropping out compared to the daytime students. This supports my initial hypothesis that the attendance type acts as a significant factor in determining if a student drops out or not. Lastly, I looked at the distribution of variables that I identified as key confounders. The two variables I looked at were ‘Age.at.enrollment’ and ‘Admission.grade’. I chose Age.at.enrollment because it could indicate socioeconomic status, as sometimes older students choose to work to save money to go to college later than typical high school graduates.



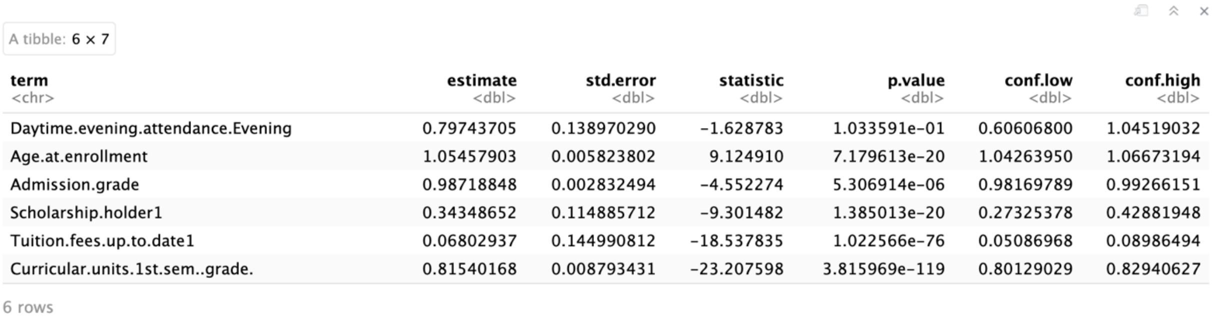
As shown by the plot above, majority of the students lie in the typical 20-23 range, but there are older students included in the dataset. Next, I looked at the distribution of Admission.grade:



This concluded the Exploratory Data Analysis I had conducted for this project.

Analysis

After conducting a thorough EDA, I began the analysis by fitting a logistic regression model with dropout status as the outcome variable and Daytime.evening.attendance as the main predictor, while controlling for all identified confounders. I chose a logistic regression model because it allows us to estimate the odds of dropout based on attendance type while adjusting for other variables. This baseline model provides an initial estimate of the relationship between the target variable and the main predictor. These were the results from the GLM:



A tibble: 6 × 7

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>	conf.low <dbl>	conf.high <dbl>
Daytime.evening.attendance.Evening	0.79743705	0.138970290	-1.628783	1.033591e-01	0.60606800	1.04519032
Age.at.enrollment	1.05457903	0.005823802	9.124910	7.179613e-20	1.04263950	1.06673194
Admission.grade	0.98718848	0.002832494	-4.552274	5.306914e-06	0.98169789	0.99266151
Scholarship.holder1	0.34348652	0.114885712	-9.301482	1.385013e-20	0.27325378	0.42881948
Tuition.fees.up.to.date1	0.06802937	0.144990812	-18.537835	1.022566e-76	0.05086968	0.08986494
Curricular.units.1st.sem..grade.	0.81540168	0.008793431	-23.207598	3.815969e-119	0.80129029	0.82940627

6 rows

From these results, I drew these causal interpretations:

- Daytime.evening.attendance.Evening: The odds ratio of 0.7974 for evening attendance suggests that when controlling for other factors, evening students have about 20% lower odds of dropping out than daytime students. But, with a p-value of 0.103, this effect is not statistically significant, and a strong causal estimate cannot be concluded.
- Age.at.enrollment: With an odds ratio of 1.055 it suggests that each additional year of age at enrollment increases the odds of dropping out by 5.5%. With a p-value < 0.001, this was a statistically significant effect.
- Admission.grade: The odds ratio of 0.987 suggest that higher admission grades slightly decrease the odds of dropping out. With a p-value < 0.001, this was a statistically significant effect.
- Tuition.fees.up.to.date: The odds ratio of 0.0680 suggests that students whose tuition fees are up-to-date have drastically lower odds of dropping out by 93.2% (p < 0.001)
- Cirricular.units.1st.sem.grade: An odds ratio of 0.815 suggests that a higher first-semester grade reduces the odds of dropping out (p < 0.001)

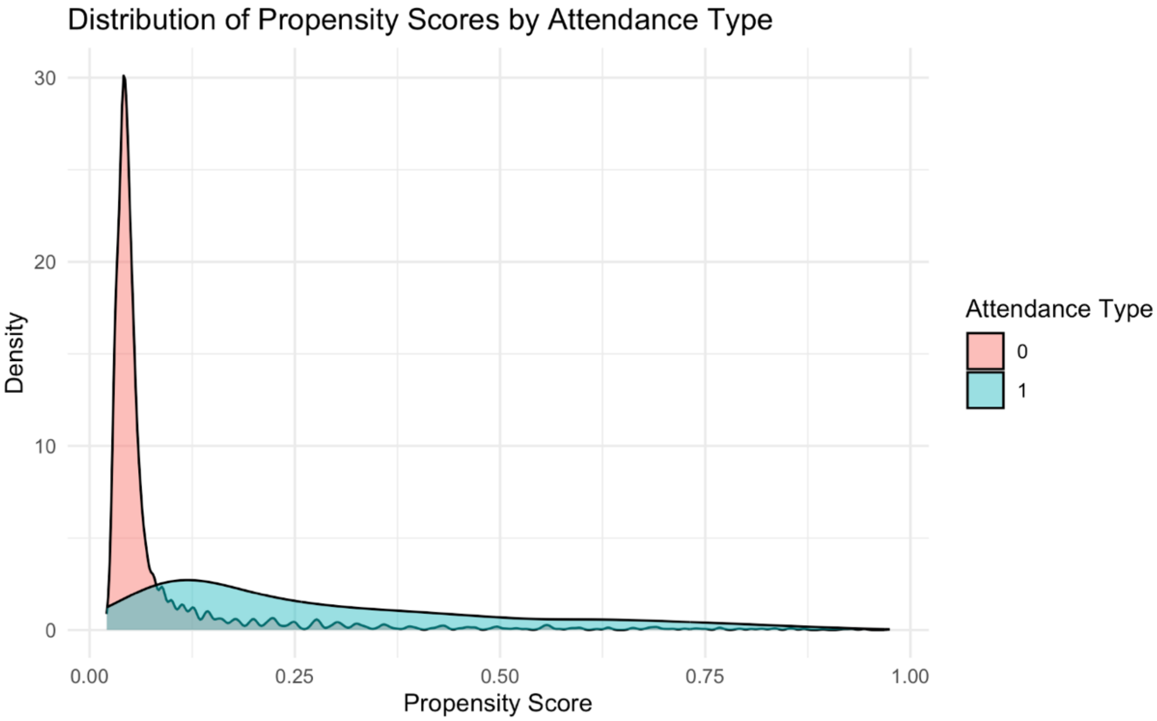
The result of my original model suggests that factors like scholarships, tuition up-to-date status and academic performance play substantial roles in chance of dropping out. However, the main predictor, attendance type, had a relatively minor effect and was not statistically significant. This indicated to me that adjustments needed to be made to my model to get a more accurate causal estimate.

Next, I attempted to improve the causal estimate by using inverse probability of treatment weighting (IPTW). As covered in class, IPTW is a method used to improve causal estimates by adjusting for confounding variables. This technique helps simulate a randomized trial by weighting observations based on their probability of being in the treatment group (evening attendance). I began by calculating propensity scores, which was done by modeling the probability of being in the evening attendance group based on confounders. Then, I applied weights to each observation by the inverse of its probability of being in the attendance group. This weighting created a balanced distribution between daytime and evening groups. Lastly, I fit the weighted logistic regression model on the Target variable. The IPTW model gave a much stronger causal estimate of my original predictor:

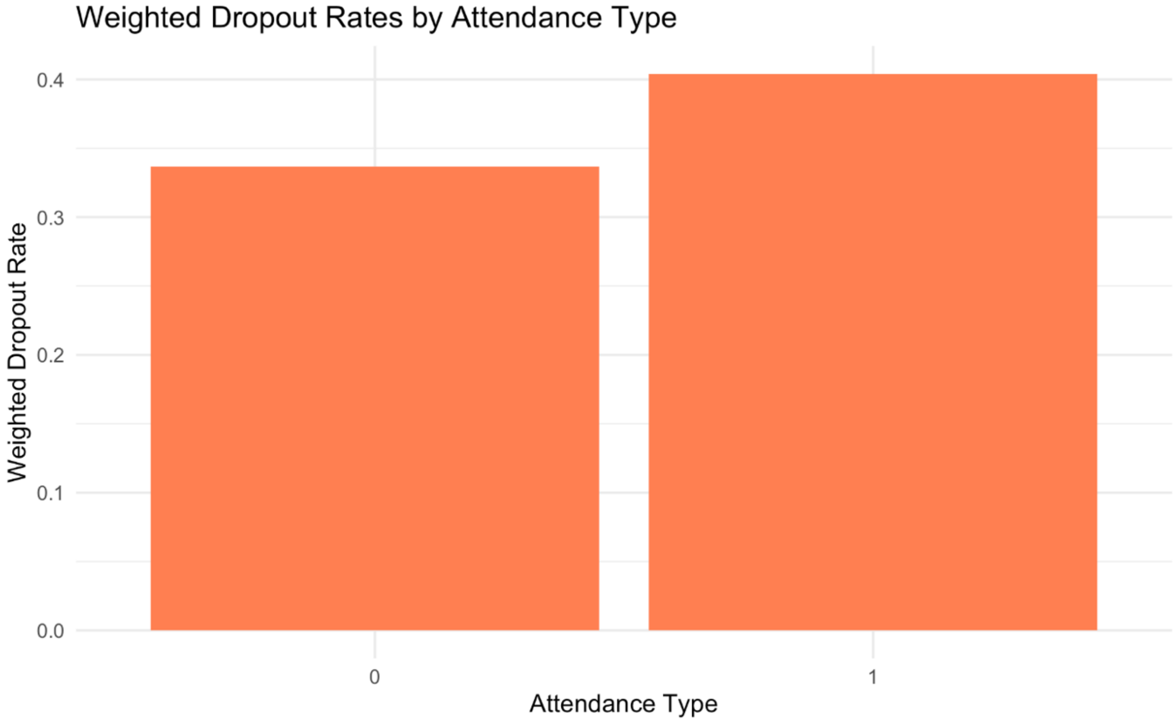
- Daytime.evening.attendance.Evening: The odds ratio of 1.335 suggests that when accounting for the weighted balance of covariates, evening students have about 33.5% higher odds of dropping out compared to daytime students ($p < 0.001$). Compared to the baseline GLM, the IPTW model yields a much stronger and statistically significant causal estimate, supporting my original hypothesis that evening students have a higher chance of dropping out compared to daytime students.

Results

To answer my original research question, if attendance type (daytime vs. evening) had a significant effect on dropout rates, the original logistic regression model suggested that evening students had 20% lower odds of dropping out compared to daytime students. However, this effect was not statistically significant, suggesting that attendance type along may not have a strong causal influence on dropout rates when controlling for other factors. To strengthen the causal estimate, I implemented an IPTW model, aiming to balance the covariates across attendance types by applying weights based on the propensity scores. From the results of the IPTW model, I concluded that students who attend evening classes have a higher chance of dropping out (33.5%) than students who attend daytime classes, after controlling for confounders. This effect was statistically significant and supported my original hypothesis that evening students had a higher chance of dropping out when controlling for confounders. To visualize my results, I plotted the distribution of propensity scores by attendance type, showing overlap between daytime and evening students after weighting.



Next, I plotted the weighted dropout rates by attendance type to see if there were any significant changes from my original plot:



The IPTW model indicated that attendance type does have a statistically significant impact on dropout rates. These results suggest that attendance timing does play a role in dropout risk, possibly due to additional student responsibilities or less engagement among evening students. The IPTW provided a stronger and more reliable causal estimate than the baseline logistic regression model.

Discussion

This project serves as a good insight for students and universities on student dropout rates and how it relates to class attendance type. The results suggest that students attending evening classes may be at a higher risk of dropping out compares to students attending daytime classes, even after adjusting for socioeconomic and academic factors. This can help universities in creating policies that support evening students, like additional academic resources or flexible scheduling options to lower the dropout risk. Since financial and academic support variables (scholarships, tuition up to date, etc.) were statistically significant in the analysis, universities should prioritize these areas for resource allocation. Initiatives like enhanced scholarship programs or support systems to help students keep up with tuition could lower the dropout rates for affected students. One of the limitations I encountered was the imbalance in attendance type. The dataset shows a significant imbalance between daytime and nighttime students, and although the IPTW model tried to address this, the smaller of the two groups (evening students) may not fully capture the range of characteristics seen in the larger group (daytime students). In addition, there were a lot of unobserved factors that could influence both attendance type and dropout rates. Factors like employment status, family responsibilities or motivation level were not included, and these could all have significant effects on dropout rates. If I redid this assignment in the future, I would collect and include those additional confounders to see if I could get a more accurate estimate on the effect of attendance type on dropout rates. In addition, I would try to implement Propensity Score Matching, which was another technique we learned in class to control for confounding variables, to see if I could get a more accurate result. In summary, this project provided me a lot of insight on the confounding variables that lead to student dropout rates. The analysis supported my original hypothesis that evening students drop out at a much higher rate than daytime students and provided me insight on what other variables attribute to that.