Technical Report of Quercus Activity and First-Year Retention

Amanda Ng

2025-02-13

1. Introduction

Student retention is a critical metric for universities, influencing both institutional success and student outcomes. At the University of Toronto (UofT), the Year 1 to Year 2 retention rate is estimated to be between 90-92%, reflecting the university's commitment to admitting high-achieving students. However, for the small percentage of students who do not complete their degrees, the long-term consequences can be significant—research suggests that individuals without a post-secondary degree face lower earnings and more precarious employment prospects in the labor market. Beyond individual outcomes, student retention has broader financial implications. Retaining just 1% more students for an additional year translates to approximately $3 million in tuition revenue, underscoring the institutional importance of understanding and addressing factors contributing to student attrition. In particular, UofT does not attract a significant number of upper-year transfer students, making it all the more essential to support first-year students in persisting to their second year.

This study aims to explore the relationship between first-year students' engagement patterns on Quercus, UofT's online learning management system, and their likelihood of continuing to Year 2. For the purposes of this study, we define dropping out as not returning for Year 2. A key hypothesis of this study is that student engagement with Quercus serves as a strong predictor of retention, while prolonged inactivity may be an early warning sign of attrition.

The study is guided by two primary questions:

1. Does the trajectory of Quercus access throughout the academic term influence a student's decision to drop out?
2. At what point (in terms of access frequency or specific dates/weeks), should we be particularly concerned about a student potentially dropping out?

By identifying behavioral patterns associated with retention or dropout, these analyses will inform interventions, such as assigning college registrars, who play a key role in student success initiatives, to conduct outreach to support students who are at risk of dropping out based on their interaction with the course platform. Through this research, UofT can develop data-driven strategies for early intervention, ensuring that students who show signs of disengagement receive timely support.

This report is structured to provide a comprehensive technical analysis of the association between first year students' Quercus engagement and drop-out risk. Section 2: Data describes the data collection process, alongside the cleaning and preparation of the data set. A summary of key variables and exploratory data visualizations is also presented. Section 3: Modelling explains the modeling approach, logistic regression model, used to analyze the relationship between the likelihood of dropping out and Quercus visit counts (and other confounders), along with justifications of appropriateness of the model design. The models' fitted results, predictors significance as well as prediction accuracy are also reported. Limitations of the study, including data quality and practical usability of the models, are also discussed. Finally, Section 4: Conclusion summarizes the key insights and outlines the next steps for future studies.

2. Data

2.1 Data Collection

Our data set was collected from 2022-2023 University of Toronto first year students between September 1, 2022 and December 7, 2022. The data set was extracted from two sources: Quercus Records Store and Institutional Retention.

Quercus is a U of T's learning management system (LMS) and all clicks are tracked in the transactional data set Quercus Records Store (QRS). QRS was collected with primary use to provide individual instructors insight into patterns of activity in their courses. It was updated daily and indicated what was clicked on and which items were viewed and when. However, in our extracted data set, we only have information about whether Quercus was accessed on a day. In other words, we do not have information about the items and the exact time of the student's interaction with Quercus. Note that students who never interacted in Quercus at all during the term are not recorded in the data.

Institutional Retention is a database that tracks the retention status (either active, inactive, or graduated) of students by cohort (defined as the year when they started at the University of Toronto). It gets updated once a year on November 1st. In our extracted data set, all students are from the same cohort (i.e. 2022) and we have information on the students' second-year retention status (either active or inactive), first-year academic performance, and some demographic information.

The data collected for this study aims to investigate the association between students' engagement patterns in Quercus and their likelihood of dropping out.

2.2 Data description

The raw data is in wide format, meaning that each row represents information collected from a student and no more than one row contains information from the same student. It consists of 3311 observations and 109 variables.

Below is a summary description of all 109 variables in the dataset.

Variable	Type	Categories/Range
Randomized student ID	Numerical	Unique numerical identifier
Incoming GPA before student began at UofT	Categorica	No admission average, <76%, 76-80%, 81-85%, 86-90%, 91-95%, >95%
Gender	Categorica	Male, Female, Unreported
Immigration Category	Categorica	Domestic, International
First Term Course Count (Nov 4, 2022)	Numerical	0 to 6
Student Registration Status (Nov 4, 2022)	Categorica	Registered, Financially cancelled, Deregistered, Cancelled, Invited
Year 1 CGPA	Numerical	0 to 4.0
Year 1 Final Academic Standing	Categorica	In good standing, On academic probation, No Registration, No assessment data, Suspended for one year
Year 2 Active (Fall 2023)	Binary	1 = Active
Year 2 Inactive (Fall 2023)	Binary	1 = Inactive
Quercus Access (1/9/2022 - 7/12/2022)	Binary	1 = Accessed Quercus on that day

Note that both the "End-of-year GPA" and "Year 1 Final Academic Standing" variables are based on information that becomes available only after the fall term.

2.3 Data Cleaning

For missing values, we found three students with no recorded "Incoming GPA before student began at UofT." These observations were assigned to the "No admission average" category. After cleaning, there are 347 observations with no admission average. Given the substantial size of this category, we retained these cases to preserve the representativeness of the dataset.

We also checked for inconsistencies in the "Year 2 Active" and "Year 2 Inactive" variables. No discrepancies were found since there were no cases where both variables were simultaneously marked as 1 (both active and inactive) or both marked as 0 (neither active nor inactive). This confirms that the variables were coded correctly and are mutually exclusive and consistently coded.

To ensure our analysis focuses on students who were properly registered in Fall 2022 and to accurately predict their probability of dropping out in the following year, we first excluded any student who was not officially registered that term and those whose Fall 2022 course count was 0. These cases were deemed irrelevant to our research question. This exclusion affected only a small number of observations (13 students), leaving a total of 3,298 students after cleaning.

Next, the "Yr 2 Active" variable was removed because it is redundant when paired with the "Yr 2 Inactive" variable. Since these variables are mutually exclusive (i.e., if one is 1, the other must be 0, and vice versa), retaining only "Yr 2 Inactive" provides all necessary information while reducing duplication and simplifying the dataset. The "Students registration status as of November 4, 2022" variable was also removed because it is irrelevant to this study after data cleaning. Furthermore, we verified that there were no duplicate observations in the dataset.

To facilitate our analysis, we created new variables representing the standardized cumulative Quercus visit counts over specific time periods. Instead of using daily visit counts, which can be highly granular and introduce noise, we aggregated the total Quercus visits for each student across different time intervals: Weeks 1-4, Weeks 1-8, Weeks 1-12, and Weeks 1-15 and divide the sum by the number of days in that period. These standardized cumulative Quercus visit counts essentially represent the proportion of days a student visited Quercus during that window. These variables help to capture broader patterns in student activity while minimizing the impact of short-term variations.

In the raw data's original wide format, each date is represented as a separate column, which can make it cumbersome to identify trends across all dates and perform trend comparisons between different student subgroups. For visualization purpose, we created a long format version of the data, since the data includes repeated measurements over time. By reorganizing the data so that all dates are listed in a single column, with their corresponding binary Quercus interaction values in another column, we created a data structure in which each row corresponds to an occurrence or missing Quercus interaction event on a specific date. Note that multiple rows in the long format data may refer to the same student if they accessed Quercus on multiple days. This format makes it easier to filter, group, and analyze data based on specific periods and simplifies the process of generating visualizations, such as the time-series plots in section 2.43.

2.4 Data summary and visualization

In total, there are 3,298 students in this study. Among them, 2952 students remained active in Fall 2023, while 346 students dropped out after their first year. This distribution highlights that the majority (nearly 90%) of first-year students in the data set remained enrolled at the University of Toronto in Year 2.

This study's observation window is *1/9/2022 - 7/12/2022*, totaling to 98 days or 15 weeks.

2.41 Incoming GPA

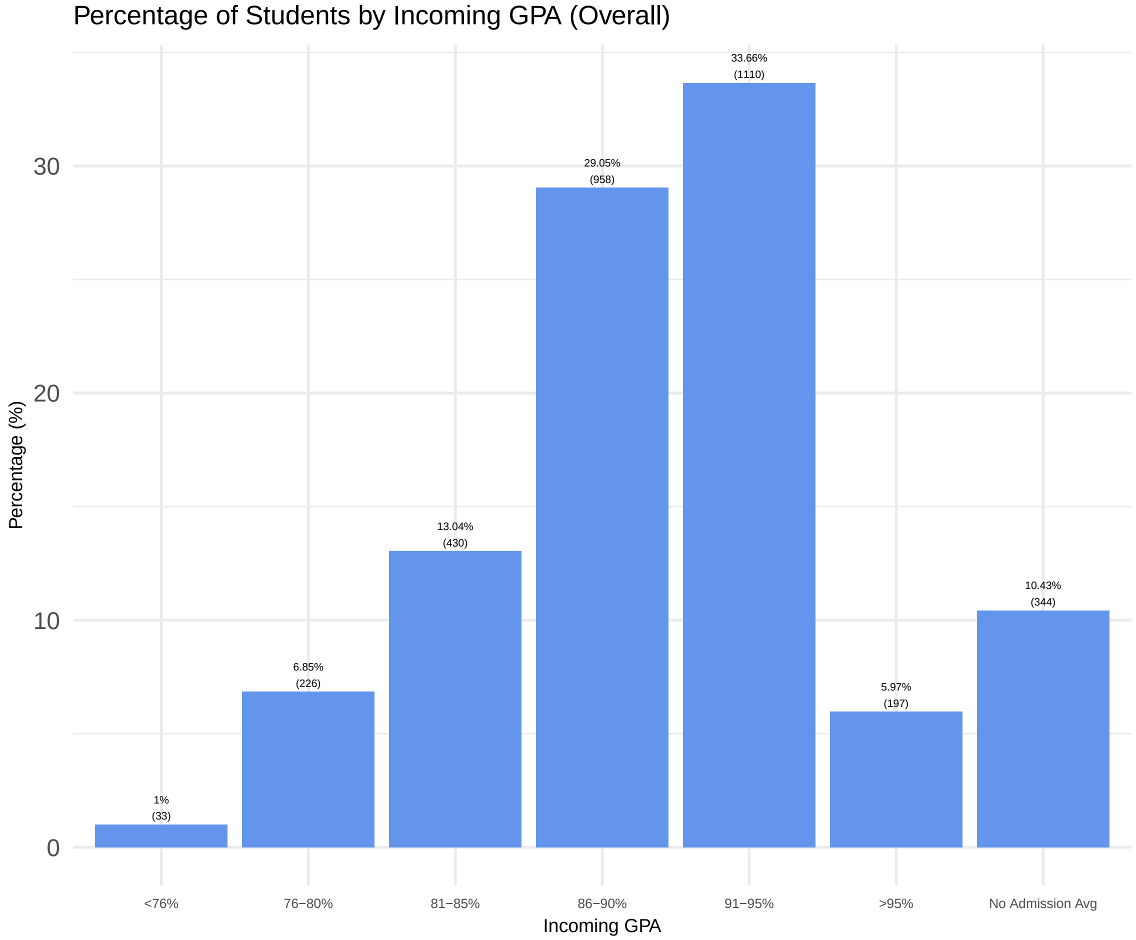


Figure 1: Percentage of Students by Incoming GPA (Overall). This bar plot illustrates the distribution of all students based on their incoming GPA, categorized into six groups: less than 76%, 76-80%, 81-85%, 86-90%, 91-95%, and above 95%. An additional category, "No Admission Avg," represents students without an admission GPA recorded. Each bar is labeled with both the corresponding percentage and the total counts in brackets within each category.

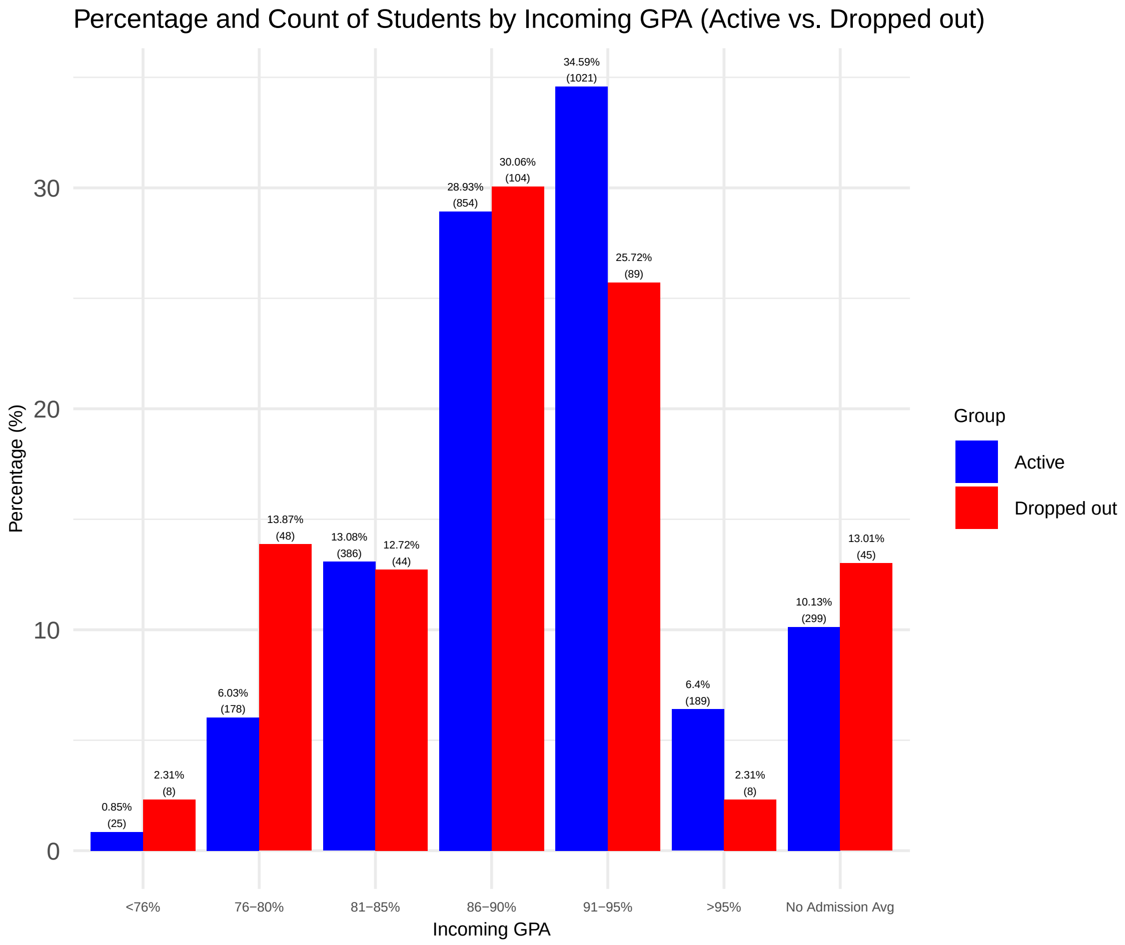


Figure 2: Percentage of Students by Incoming GPA (Active vs Dropped out). This bar plot illustrates the distributions of active and dropped-out students based on their incoming GPA, categorized into six groups: less than 76%, 76-80%, 81-85%, 86-90%, 91-95%, and above 95%. An additional category, "No Admission Avg," represents students without an admission GPA recorded. The blue bars correspond to active students, while the red bars represent dropped-out students. Each bar is labeled with both the corresponding percentage and the total counts in brackets within each category.

Figures 1-2 illustrate the distribution of students across different incoming GPA categories for all students, students who remained enrolled in their second year, and those who dropped out. For active students, the majority fall into the 91–95% and 86–90% GPA ranges, accounting for 34.59% and 28.93%, respectively. Dropouts show a similar trend, with 30.06% in the 86–90% range and 25.72% in the 91–95% range, though the distribution is more spread across lower GPA brackets compared to active students. Notably, we observe the percentage of students who had >95% incoming GPA among active students is 2.84 times higher than those who dropped out.

The greater proportion of students in lower GPA brackets among dropouts highlights that lower incoming academic performance may predispose students to disengagement and eventual withdrawal. Furthermore, the significantly higher proportion of active students with >95% incoming GPA suggests that high academic achievers may be more resilient to disengagement, which could influence the interpretation of engagement behaviors. This finding indicates that incoming GPA could be an important covariate in the model when analyzing the effect of Quercus engagement patterns on drop-out likelihood, ensuring that the model accounts for the baseline differences in academic preparedness.

2.42 First Term Course Count

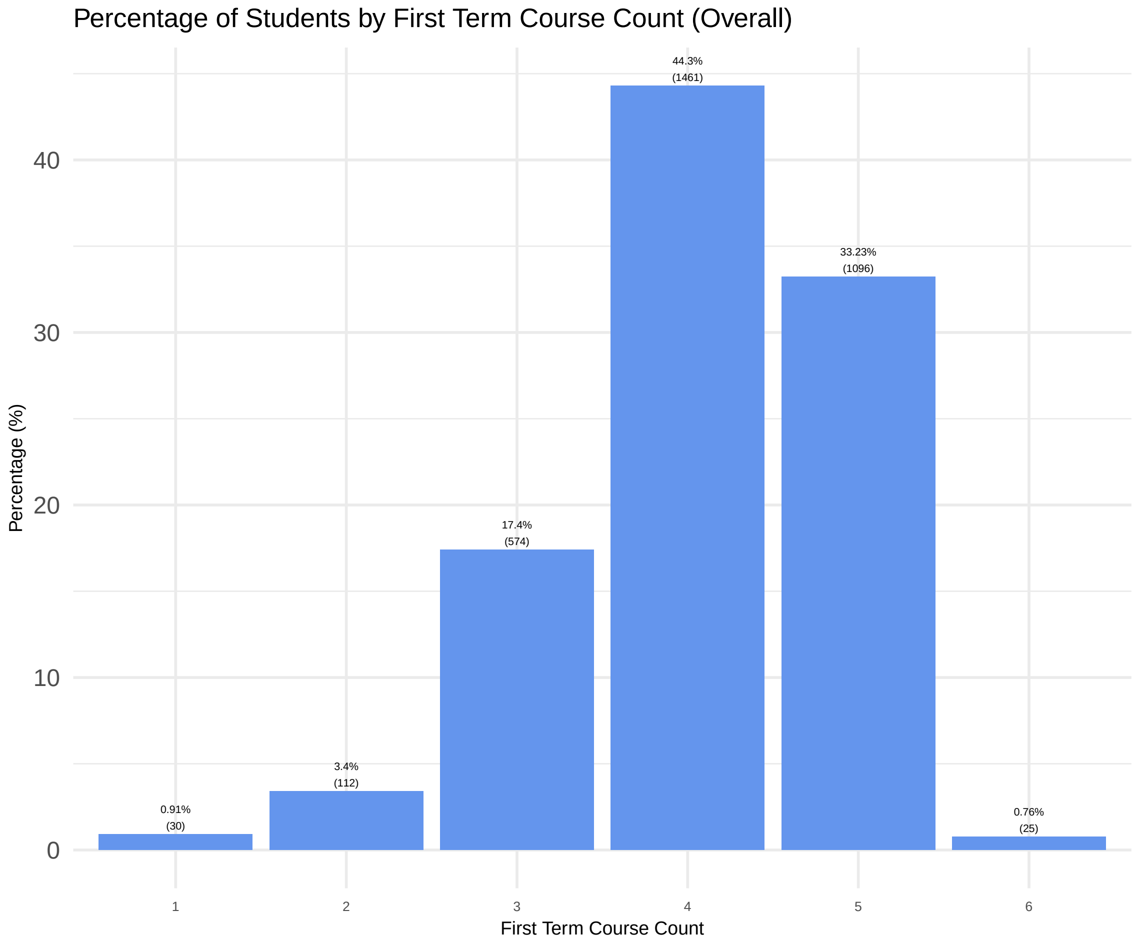


Figure 3: Percentage of Students by First Term Course Count (Overall). This bar plot illustrates the distribution of all students based on the number of courses they took during the 2022 Fall, ranging from 0 to 6. Each bar is labeled with both the corresponding percentage and the total counts in brackets within each category.

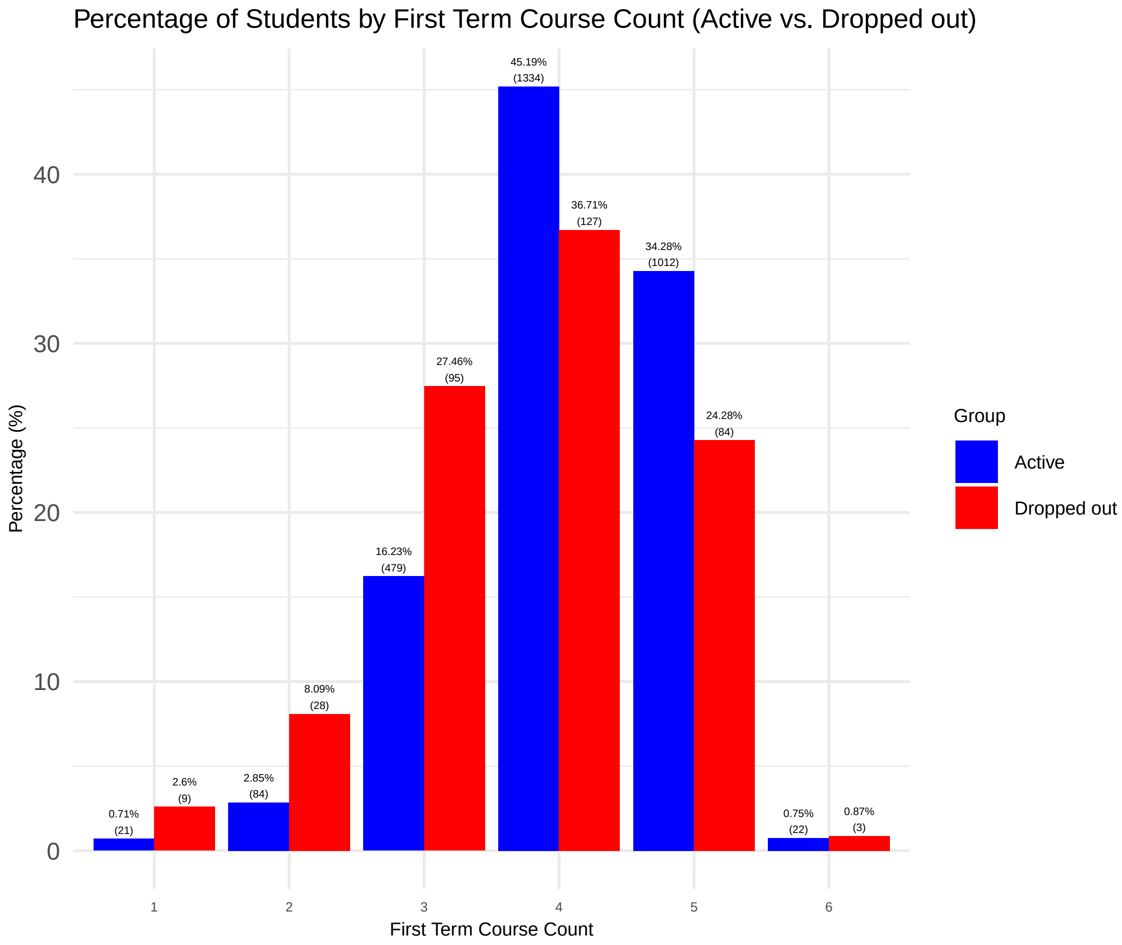


Figure 4: Percentage of Students by First Term Course Count (Active vs Dropped out). This bar plot illustrates the distributions of active and dropped-out students based on the number of courses they took during the 2022 Fall, ranging from 0 to 6. The blue bars correspond to active students, while the red bars represent dropped-out students. Each bar is labeled with both the corresponding percentage and the total counts in brackets within each category.

Figures 3-4 highlight the distribution of students by their first-term course count across overall, active, and dropped-out students. A consistent trend found across all groups is that the largest proportion of students concentrated in the middle range of course counts (e.g., around 3-5 courses), though the exact percentages vary. In the overall and active groups, the majority of students are clustered around course counts of 4 and 5, with noticeable tails at lower and higher counts. For students who dropped out, the distribution is somewhat more evenly spread and shows a higher percentage at lower course counts compared to active students. This suggests that a lower course load might be associated with higher dropout rates.

These results suggest that course load is a potentially important factor in understanding the likelihood of dropping out. The higher proportion of students with lower course counts among those who dropped out indicates that a reduced course load might be associated with disengagement or challenges that increase the likelihood of leaving. On the other hand, lower course loads could influence engagement patterns in Quercus, such as less frequent logins, since lower degree of participation is expected when students are enrolled to fewer courses. These findings highlight the need to include first-term course count as a covariate since it could serve as a proxy for workload and academic commitment affecting students' ability to manage their studies.

2.43 Quercus engagement patterns

To analyze the cumulative Quercus visits over time, we generated separate spaghetti plots for the active and inactive student groups.

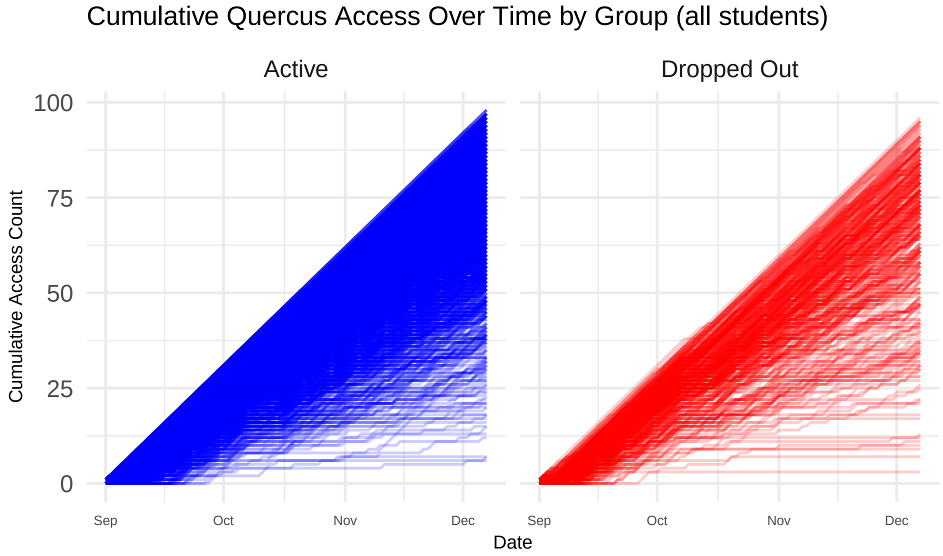


Figure 5: Cumulative Quercus access over time by group (Active/ Dropped out). The spaghetti plots compare the cumulative Quercus access counts over time for two groups: students who remained active (left, in blue; n = 2953) and students who dropped out (right, in red; n = 346) during 1/9/2022 - 7/12/2022. Each line represents the Quercus access visits trajectory of an individual student from a sample of each group. The diagonal lines in the plots represent students who visited Quercus daily, while horizontal lines indicate periods where students did not visit the platform.

Given the substantial group size difference between the active and dropout groups, we drew random samples of size 100 from each group to ensure a fair comparison in the visualization.

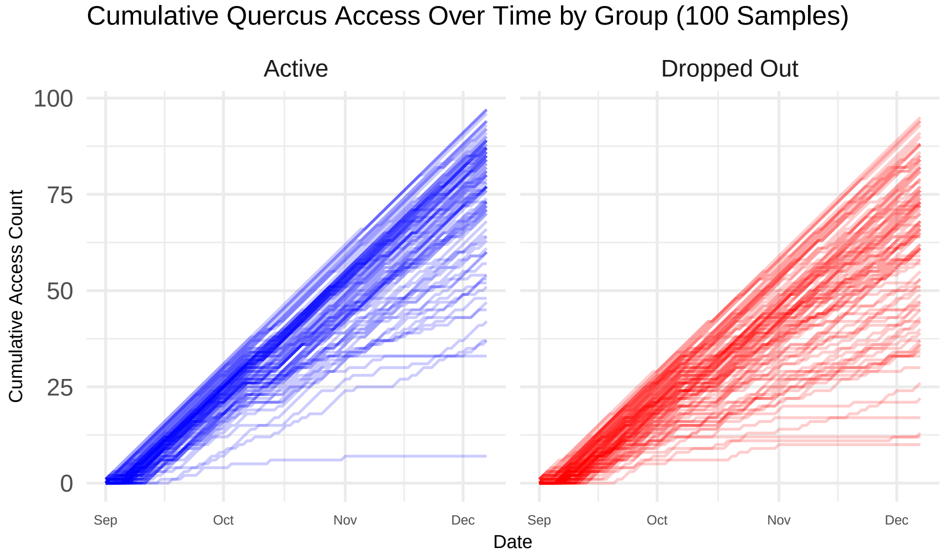


Figure 6: Cumulative Quercus access over time by group (Active/ Dropped out) of random samples n = 100. The spaghetti plots compare the cumulative Quercus access counts over time for two groups: students who remained active (left, in blue) and students who dropped out (right, in red) during 1/9/2022 - 7/12/2022.

Each line represents the Quercus access visits trajectory of an individual student from a sample of 100 in each group.

According to figures 5 and 6, both the active and dropped-out groups show an overall increase in cumulative access counts over time, but the trajectories differ in their spread and consistency. The active group exhibits tightly clustered lines, indicating more consistent and uniformly high engagement across students. Conversely, the dropped-out group shows more variability, with more students displaying slower growth in cumulative access counts or flatter trajectories, suggesting lower or irregular engagement. This difference implies that students with lower or irregular engagement are more likely to drop out, reinforcing the idea that cumulative engagement is a potential predictor of dropout risk.

When modeling the association between Quercus engagement and dropout risks, the findings above suggest that cumulative access counts can serve as a key predictor to capture Quercus engagement patterns. The variability in the engagement trajectories of the dropped-out group indicates that the timing and magnitude of access could play a significant role in identifying at-risk students.

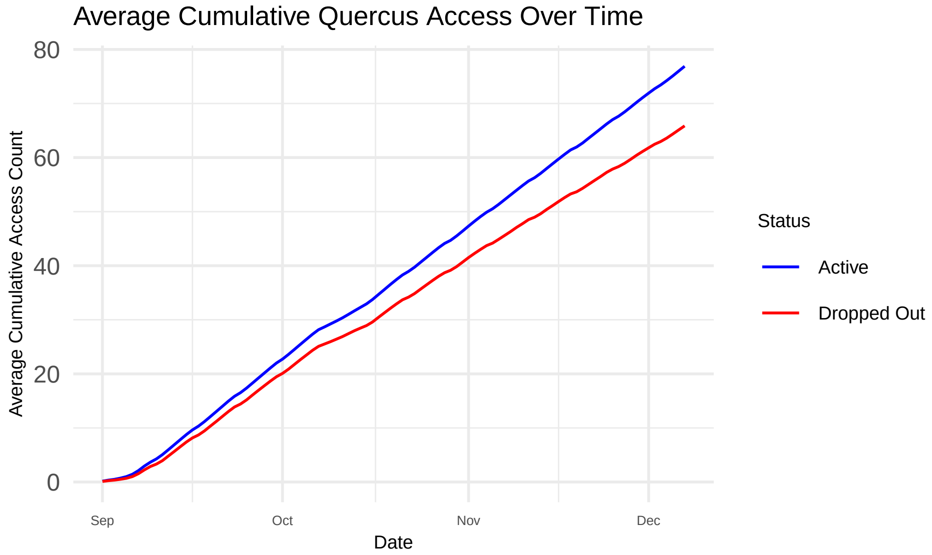


Figure 7: Average Cumulative Quercus access over time by group (Active/ Dropped out). The line plot displays the averaged cumulative Quercus visits of each group during 1/9/2022 - 7/12/2022. The blue represents the averaged cumulative Quercus access distribution among students who remained after the first year while the red line represents that among students who dropped out.

Additionally, we calculated and plotted the **average** cumulative Quercus visit line for each group. The steeper trajectory observed in the active group indicates sustained or increasing engagement over time, while the flatter trajectory in the dropout group suggests lower or declining engagement. This pattern implies that students who engage more frequently with Quercus are less likely to drop out, highlighting the potential association of Quercus engagement with dropout risks.

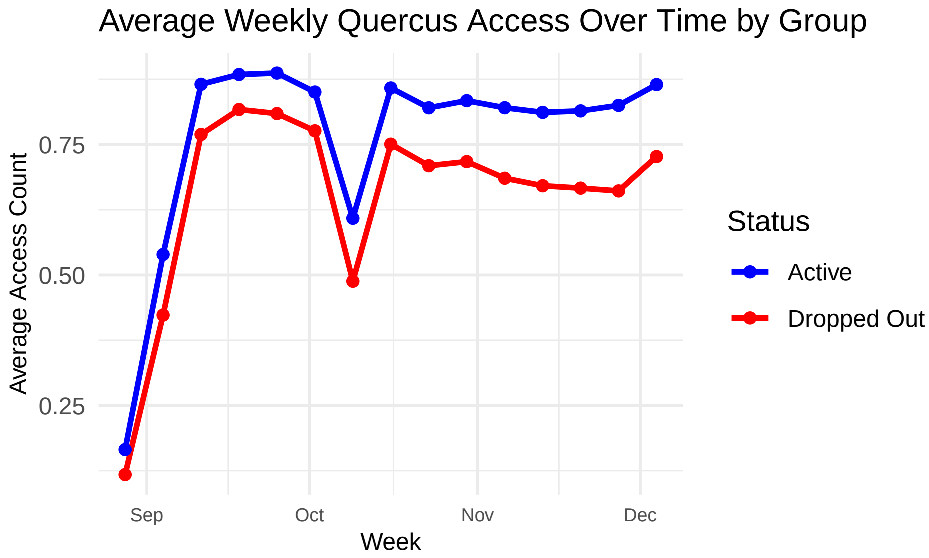


Figure 8: Average weekly Quercus access over time by group (Active/ Dropped out). The line plot displays the weekly averaged Quercus visits of each group during 1/9/2022 - 7/12/2022. Average weekly counts were calculated by first summing all visit counts within one group in a specific week, then divided by the number of students in each group and further divided by 7 (i.e. number of days in a week). The blue represents the averaged weekly Quercus access distribution among students who remained after the first year while the red line represents that among students who dropped out.

From figure 8, we observe an initial increase in visits from W1 to W3, followed by a noticeable dip in W7, which may be attributed to the midterm season. After this, the trend tends to stabilize in subsequent weeks. Similar to figure 7, students who dropped out generally had lower engagement, with their average weekly visit counts being consistently lower compared to those who remained enrolled. Potentially, a lack of an increase in visits during the first few weeks (W1 to W3) might be indicative of a student being at risk of dropping out. This could serve as an early warning signal for potential disengagement, emphasizing the importance of monitoring engagement patterns early in the term.

2.44 Correlation analysis of Quercus visit counts

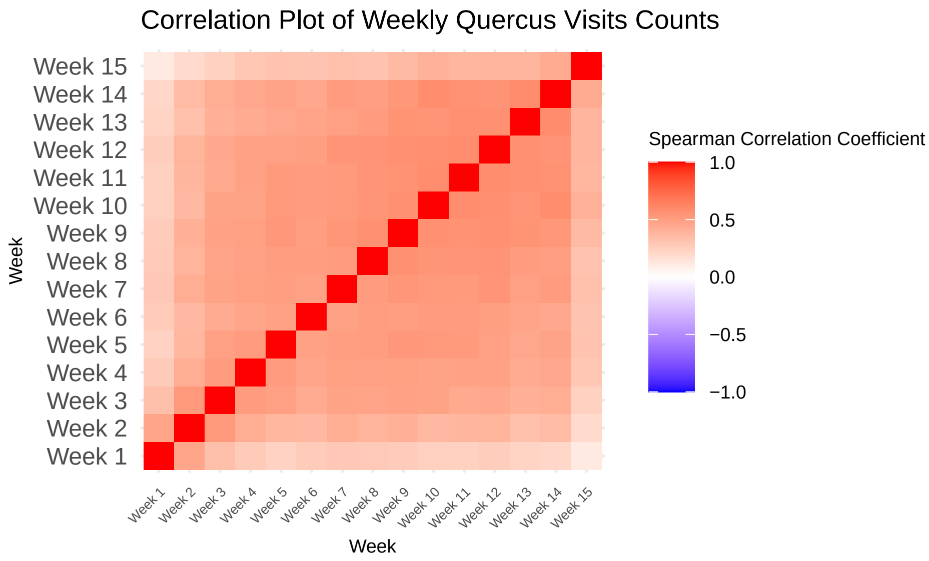


Figure 9: Correlation Plot of Weekly Quercus Visits Counts. This plot shows the correlation of the weekly Quercus visits counts between each pair of weeks. Darker red grids represent stronger positive correlations, while darker blue grids indicate stronger negative correlations.

Figure 9 reveals that all pairwise correlations weekly Quercus visit counts between weeks are positive, as indicated by the red tones across the grid. As expected, the strongest correlations are observed along the diagonal, where each week is perfectly correlated with itself (correlation coefficient = 1.0). As the time gap between weeks increases, the correlation weakens slightly, reflected by lighter shades of red further from the diagonal. These results suggest that weekly Quercus visits are consistently and positively correlated over time, with stronger relationships between temporally closer weeks and gradually weaken as weeks become farther apart.

This finding suggests that if our final model includes the sequence of weekly visit counts, there may be multicollinearity issue due to their high correlation. Instead, we may take a cumulative engagement count at a specific cutoff (e.g., total visits count up to Week 4, total visits count up to Week 8, etc) as the predictor in the model. This approach allows us to assess how engagement up to different time points relates to dropout probability while avoiding issues due to highly correlated predictors.

3. Modeling

We employed logistic regression models to predict the likelihood of a student dropping out after their first year. In our models, the outcome variable is the logit of the dropout probability, and the key predictor is standardized cumulative Quercus visits count within a certain period. Along side, we included a range of covariates: Incoming GPA, Gender, Immigration Category, First Term Course Count.

To examine how student engagement over different periods of the academic term influences dropout risk, we fitted four logistic regression models that differ only by how we defined the standardized cumulative Quercus visits count. Specifically, we consider standardized cumulative visits up to Week 4 (W1–W4), Week 8 (W1–W8), Week 12 (W1–W12), and Week 15 (W1–W15).

Below are the model equations:

Model 1 (Weeks 1-4)

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{Standardized Cumulative Quercus visit count within Weeks 1-4}$$

$$+\beta_2 \cdot I_{\text{Incoming GPA>95\%}} + \beta_3 \cdot I_{\text{Incoming GPA 91\%−95\%}} + \beta_4 \cdot I_{\text{Incoming GPA 86\%−90\%}}$$

$$+\beta_5 \cdot I_{\text{Incoming GPA 81\%−85\%}} + \beta_6 \cdot I_{\text{Incoming GPA 76\%−80\%}} + \beta_7 \cdot I_{\text{Incoming GPA No Admission Average}}$$

$$+\beta_8 \cdot I_{\text{Male}} + \beta_9 \cdot I_{\text{Unreported gender}} + \beta_{10} \cdot I_{\text{International student}}$$

$$+\beta_{11} \cdot I_{\text{First Term Course Count=2}} + \beta_{12} \cdot I_{\text{First Term Course Count=3}} + \beta_{13} \cdot I_{\text{First Term Course Count=4}}$$

$$+\beta_{14} \cdot I_{\text{First Term Course Count=5}} + \beta_{15} \cdot I_{\text{First Term Course Count=6}}$$

Model 2 (Weeks 1-8)

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{Standardized Cumulative Quercus visit count within Weeks 1-8}$$

$$+\beta_2 \cdot I_{\text{Incoming GPA>95\%}} + \beta_3 \cdot I_{\text{Incoming GPA 91\%−95\%}} + \beta_4 \cdot I_{\text{Incoming GPA 86\%−90\%}}$$

$$+\beta_5 \cdot I_{\text{Incoming GPA 81\%−85\%}} + \beta_6 \cdot I_{\text{Incoming GPA 76\%−80\%}} + \beta_7 \cdot I_{\text{Incoming GPA No Admission Average}}$$

$$+\beta_8 \cdot I_{\text{Male}} + \beta_9 \cdot I_{\text{Unreported gender}} + \beta_{10} \cdot I_{\text{International student}}$$

$$+\beta_{11} \cdot I_{\text{First Term Course Count=2}} + \beta_{12} \cdot I_{\text{First Term Course Count=3}} + \beta_{13} \cdot I_{\text{First Term Course Count=4}}$$

$$+\beta_{14} \cdot I_{\text{First Term Course Count=5}} + \beta_{15} \cdot I_{\text{First Term Course Count=6}}$$

Model 3 (Weeks 1-12)

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{Standardized Cumulative Quercus visit count within Weeks 1-12}$$
$$+\beta_2 \cdot I_{\text{Incoming GPA>95\%}} + \beta_3 \cdot I_{\text{Incoming GPA 91\%−95\%}} + \beta_4 \cdot I_{\text{Incoming GPA 86\%−90\%}}$$
$$+\beta_5 \cdot I_{\text{Incoming GPA 81\%−85\%}} + \beta_6 \cdot I_{\text{Incoming GPA 76\%−80\%}} + \beta_7 \cdot I_{\text{Incoming GPA No Admission Average}}$$
$$+\beta_8 \cdot I_{\text{Male}} + \beta_9 \cdot I_{\text{Unreported gender}} + \beta_{10} \cdot I_{\text{International student}}$$
$$+\beta_{11} \cdot I_{\text{First Term Course Count=2}} + \beta_{12} \cdot I_{\text{First Term Course Count=3}} + \beta_{13} \cdot I_{\text{First Term Course Count=4}}$$
$$+\beta_{14} \cdot I_{\text{First Term Course Count=5}} + \beta_{15} \cdot I_{\text{First Term Course Count=6}}$$

Model 4 (Weeks 1-15)

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{Standardized Cumulative Quercus visit count within Weeks 1-15}$$
$$+\beta_2 \cdot I_{\text{Incoming GPA>95\%}} + \beta_3 \cdot I_{\text{Incoming GPA 91\%−95\%}} + \beta_4 \cdot I_{\text{Incoming GPA 86\%−90\%}}$$
$$+\beta_5 \cdot I_{\text{Incoming GPA 81\%−85\%}} + \beta_6 \cdot I_{\text{Incoming GPA 76\%−80\%}} + \beta_7 \cdot I_{\text{Incoming GPA No Admission Average}}$$
$$+\beta_8 \cdot I_{\text{Male}} + \beta_9 \cdot I_{\text{Unreported gender}} + \beta_{10} \cdot I_{\text{International student}}$$
$$+\beta_{11} \cdot I_{\text{First Term Course Count=2}} + \beta_{12} \cdot I_{\text{First Term Course Count=3}} + \beta_{13} \cdot I_{\text{First Term Course Count=4}}$$
$$+\beta_{14} \cdot I_{\text{First Term Course Count=5}} + \beta_{15} \cdot I_{\text{First Term Course Count=6}}$$

where

- i represent the ith student
- p_i represents the probability of the ith student dropping out after first year
- Standardized Cumulative Quercus visits count within Weeks 1-K represents the total number of Quercus visit within a specified time period Weeks 1-K divided by number of days during the time period.

For all models, we set the reference level of:

- Incoming GPA as below 76%
- Gender as female
- Immigration category as domestic student
- First Term Course Count as 1

For model evaluation, we split the data into a training set and a testing set, using an 80/20 split. This means that 80% of the data was randomly selected to train the models, while the remaining 20% was reserved for testing. The random sampling ensures that the training and testing sets are representative of the overall data, which helps in assessing the models' predictive performance such as sensitivity and misclassification rate on unseen data.

3.1 Justification for considering the models

Initially, we considered using a longitudinal framework to build a time series model, but the nature of our response variable presented challenges. Specifically, a student's intention to drop out can occur at any point in time, yet the exact moment of this decision is unknown to us. A longitudinal model requires aligning predictors at time t with a response variable also observed at time t, which is not feasible given our data structure. Therby, we settled down with non-longitudinal logistic regression models instead.

Given that our response variable, the Year 2 inactive indicator, is binary, logistic regression is an appropriate modeling choice. This is because logistic regression models the outcome variable as the logit of the dropout probability, allowing us to predict the likelihood of dropout while accounting for the effects of various predictors. One of the key strengths of logistic regression is its interpretability. The beta coefficients of the predictors represent the log odds of dropping out for a one-unit change in each predictor, while holding all other variables constant. Odds represent the ratio of the probability that an event (dropping out) will occur to the probability that it will not occur and log odds is the natural logarithm scale of the odds of an event occurring. Positive log odds indicate the event is more likely to happen, while negative log odds suggest it is less likely. This provides a straightforward and quantifiable way to assess the influence of each factor on student drop out risk, which is especially valuable when reporting to our non-statistical collaborators.

We chose the specific time-point windows to capture key phases of the academic term and assess how student engagement at different stages relates to dropout risk.

- **Week 1-4**: This window focuses on early engagement before the mid-term period, which is critical for identifying students who might already be showing signs of disengagement early in the term. This is motivated by the findings from Figure 8 in section 2.43 which suggests that a lack of visits during the first few weeks might be indicative of a student being at risk of dropping out.
- **Week 1-8**: The mid-term season is a pivotal time when students are balancing exams and coursework. Engagement during this period could provide insight into whether students' involvement in the early weeks helps them navigate academic pressure during midterms.
- **Week 1-12**: After midterms, some students may experience a drop-off in engagement. This window allows us to investigate whether a decrease in activity post-midterms contributes to dropout risk.
- **Week 1-15**: This final window covers the entire fall term, providing a comprehensive view of engagement over the full term and allowing us to evaluate whether sustained or declining engagement throughout the semester predicts dropout risk.

In a real-world predictive setting, we wouldn't have access to a student's future Quercus engagement at the current time point. By comparing these models, we aim to determine how early engagement data can effectively predict dropout risk and whether shorter-term engagement measures (e.g., W1–W4) provide comparable predictive power to longer-term measures (e.g., W1–W15). This analysis helps us identify the optimal observation window for early intervention—allowing us to assess when declining engagement becomes a meaningful warning sign. If models with earlier cumulative visit counts (e.g., W1–W4 or W1–W8) perform well, we can intervene sooner rather than waiting for a full-term engagement profile.

We chose to use standardized cumulative visits counts rather than a sequence of weekly counts primarily due to multicollinearity concerns, as revealed from the correlation analysis of Quercus visit counts in section 2.44. Weekly visit counts tend to be highly correlated over time, which can introduce instability in the regression model, making it difficult to isolate the true effect of each weekly visit count predictor. Additionally, multicollinearity increases the variance of the estimated coefficients, resulting in larger standard errors. This reduces statistical significance, making it more difficult to detect meaningful relationships between the weekly visit count predictors and the outcome variable. Using cumulative Quercus visit counts helps avoid this issue by collapsing all weekly engagement data into a single variable that represents overall engagement over time rather than using multiple correlated weekly visit counts variables. We standardized the cumulative counts to account for the varying lengths of the different time periods as longer periods naturally accumulate higher visit counts, ensuring the models can be compared on a consistent scale. On top of that, using standardized cumulative visit counts simplifies the interpretation of the regression coefficients. If weekly counts were used, the beta coefficient for week K would need to be interpreted as: given that visit counts in all other weeks remain constant, an additional visit in week K is associated with a β_k change in the logit of dropping out. In practice, however, it is unrealistic to assume that engagement in other weeks stays the same. By contrast, the beta coefficient of standardized cumulative visits provides a more intuitive interpretation. For instance, $0.1 \times \beta$ represents the averaged effect on the likelihood of dropping out if the student visited Quercus 10% more days within the specified period, without requiring assumptions about holding the visit patterns constant in other weeks.

Incoming GPA and First Term Course Count are included in the model since lower incoming GPA and lower course load might be associated with increase in the likelihood of dropping out, as reflected from the barplots in section 2.41 and 2.42. Gender and immigration category are also included to ensure that any subtle baseline associations are accounted for, especially in capturing the underlying factors related to engagement patterns, such as differences in access to resources, cultural adjustment, or language barriers.

We intentionally excluded Year 1 CGPA and Year 1 Final Academic Standing variables because they are based on information that becomes available only after the fall term. These variables may, nonetheless, be useful for future analysis once full-year data is available, as they could provide a more comprehensive understanding of the factors influencing student outcomes throughout the entire academic year.

3.2 Model assumptions

The logistic regression model relies on 5 key assumptions:

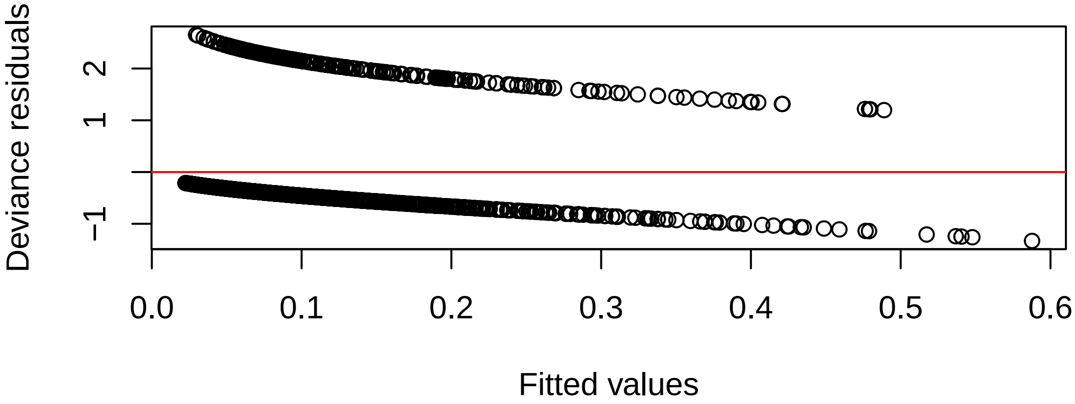
1. **Binary Dependent Variable**: The dependent variable in logistic regression must be binary, meaning it takes on only two possible outcomes. In our case, the dependent variable is the Year 2 inactive indicator, which indicates whether a student drops out (1) or not (0). This assumption is satisfied as the response variable is binary by design.

2. **Independence of Observations**: Logistic regression assumes that the observations are independent of one another. This assumption is satisfied since each observation in our dataset corresponds to an independent student.

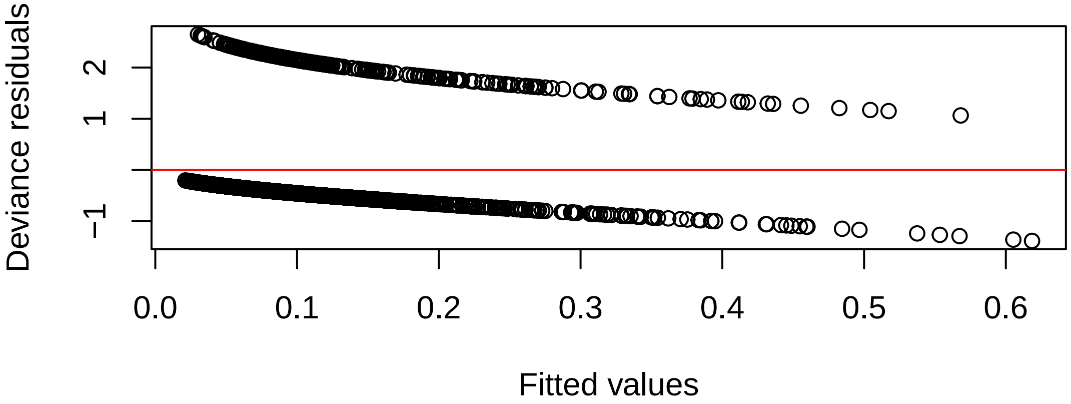
3. **No Multicollinearity**: Logistic regression assumes that there is little or no multicollinearity among the independent variables. This means that the predictors should not be highly correlated with one another, as high correlation can make it difficult to separate out the individual effects of each predictor on the dependent variable. To verify this assumption, we calculated VIF values for each predictor in each model. A VIF value below 5 indicates no significant multicollinearity in our model. Should any predictor exhibit a VIF value exceeding 5, we address multicollinearity by removing the predictor with the highest VIF and constructing a reduced model. Fortunately, multicollinearity check showed that all VIF values are below 5, indicating no severe issue is observed in any of the predictors in all 4 models.

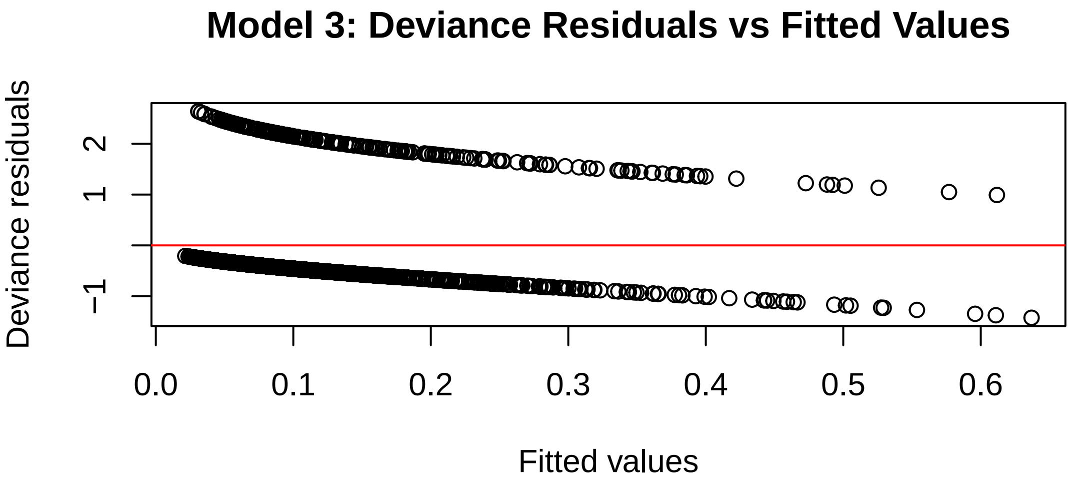
4. **Linearity of Independent Variables and Log Odds**: Logistic regression assumes a linear relationship between the independent variables and the log odds of the dependent variable. To check for this assumption, we identify if there are systematic non-linear patterns in the "deviance residual vs fitted" plot.

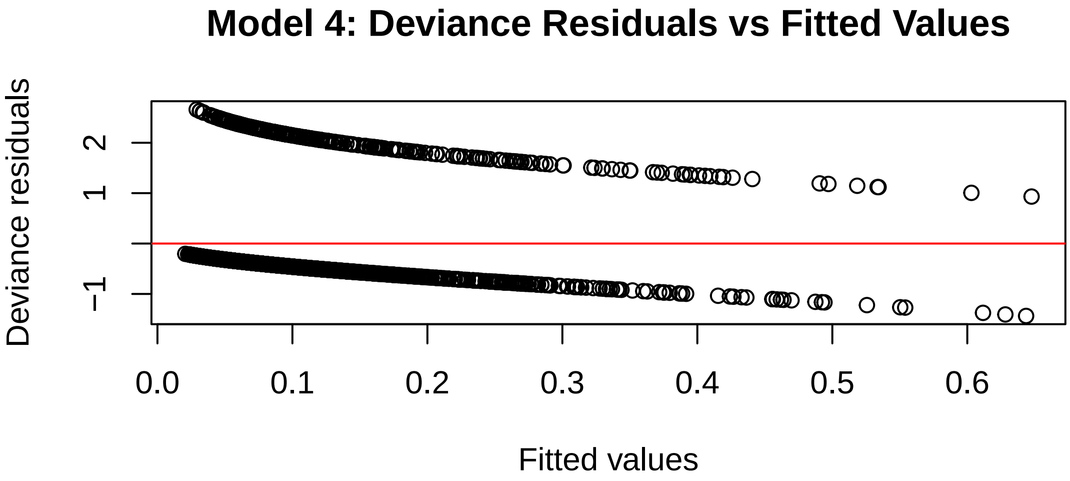
Model 1: Deviance Residuals vs Fitted Values



Model 2: Deviance Residuals vs Fitted Values







Figures 10-13: Deviance residuals vs. fitted values plot for Model 1/2/3/4. Each dot represents an individual observation from our dataset. The x-axis indicates fitted values, which is the predicted probability of the outcome from the logistic regression model. The y-axis indicates the deviance residual for that specific observation, which measures how much the observation deviates from the model's prediction. In all 4 plots, the residuals are spread in two distinct curves, which is expected in logistic regression due to the binary nature of the outcome variable. The lack of clear patterns suggests no major model misspecification. In general, the residuals decreases linearly as their fitted values increase, despite the slight curved shape for small fitted values. Nonetheless, there is no severe sign of linearity violation in any of the 4 models. The horizontal reference line at zero helps visualize deviations.

According to figures 10-13, despite the slight curvatures in the residuals with fitted values lower than 0.1, we observe no significant non-linear pattern observed in any of the model residual plots. Hence, we may assume that the linear requirement is satisfied in all 4 models.

5. **No Overdispersion**:

Overdispersion occurs when the variance of the outcome variable is greater than what the logistic regression model assumes. This implies that the model underestimates the true variance of the model parameter estimates, leading to incorrect standard errors and inflated Type 1 error. We check for overdispersion by examining the Pearson chi-squared statistic. A statistic which is significantly greater than 1 indicates overdispersion. If overdispersion is present, we might consider using alternative methods, such as quasi-binomial regression, which accounts for overdispersion. Fortunately, all Pearson chi-squared statistics are close to 1 (Model 1: 1.0025; Model 2: 1.0003; Model 3: 1.0006; Model 4: 1.0026), indicating no severe issue is observed in any of the 4 models.

3.3 Estimates and Hypothesis tests

3.31 Model 1: Weeks 1-4

Below is a summary table of all coefficient estimates of the fitted model 1:

Terms	Coefficient estimate	P-value	Standard error	95% confidence interval
Intercept	0.82155	0.2475	0.71038	(-0.57197, 2.21507)
Standardized Cumulative Quercus visit count within Weeks 1-4	-2.30865	3e^{-8}	0.41661	(-2.72526, -1.89204)
Incoming GPA > 95%	-1.12595	0.0755	0.63348	(-2.36765, 0.11575)
Incoming GPA 91%-95%	-0.57520	0.2615	0.51226	(-1.57921, 0.42881)
Incoming GPA 86-90%	-0.46652	0.3570	0.50648	(-1.45924, 0.5262)
Incoming GPA 81%-85%	-0.62984	0.2247	0.51873	(-1.64653, 0.38685)
Incoming GPA 76%-80%	-0.01466	0.9776	0.52205	(-1.03786, 1.00854)
Incoming GPA No Admission Average	-0.53620	0.3094	0.52754	(-1.57017, 0.49777)
Gender: Male	0.12471	0.3642	0.13743	(-0.14465, 0.39407)
Gender: Unreported	0.51719	0.1137	0.32694	(-0.12372, 1.1581)
International student	-0.14509	0.3790	0.16493	(-0.46835, 0.17817)
First Term Course Count = 2	-0.24360	0.6617	0.55678	(-1.33488, 0.84768)
First Term Course Count = 3	-0.53247	0.2998	0.51350	(-1.53893, 0.4740)
First Term Course Count = 4	-1.14244	0.0255	0.51145	(-2.14487, -0.14001)
First Term Course Count = 5	-1.18824	0.0225	0.52066	(-2.20872, -0.16776)
First Term Course Count = 6	-0.97576	0.2791	0.90144	(-2.74255, 0.79103)

Interpretation:

The baseline of the model is a female domestic student with incoming GPA less than 76%, enrolled to only 1 course during first term and has never visited Quercus within Weeks 1-4. Assuming fixed values for all other predictors, the exponentiated coefficients of Standardized Cumulative Quercus visit count within Weeks 1-4 is the odds ratio for dropping out associated with an "one-unit increase" in the proportion of days a student visited Quercus during the first four weeks of class. Since Quercus visit variable is defined as a proportion, it does not make sense to increase a proportion (which is between 0 and 1) by 1. So, we conduct interpretation in terms of a factor of the coefficient. For instance, the odds of dropping out if a student visited Quercus 10% days of the first four weeks of class (i.e. $0.1 \times 28 = 2.8$ days) is $exp(-2.30865 * 0.1) = 0.7938466$ times compared to those who have never visited Quercus, assuming fixed levels for all other predictors. Alternatively, we can translate the interpretation in terms of days: the odds of dropping out if a student visited Quercus for an additional day during the first four weeks of class is $exp(-2.30865 * 1/28) = 0.9208558$.

This suggests that more frequent engagement with the platform early on is linked to a lower likelihood of dropping out.

For other categorical variables, the exponentiated coefficients represents the odds ratio of dropping out between the indicator level as compared to the reference level of the predictor. When controlling for other variables, students with an incoming GPA above 95% experience the greatest reduction in dropout odds compared to those with a GPA below 76%. Students with incoming GPA in the 81%-85% and 91%-95% ranges also show a lower likelihood of dropping out. Additionally, male students generally have higher odds of dropping out than female students, whereas international students are less likely to drop out than domestic students. Finally, students enrolled in four or five courses in their first term are the least likely to drop out compared to those taking only one course. These findings suggest that both a higher course load and increased Quercus engagement may serve as indicators of greater academic involvement, which could help prevent students from falling behind and ultimately dropping out.

We conducted a Likelihood Ratio Test (LRT) to assess the overall significance of the model. The test compared the full model, which includes all covariates, to a null model containing only the intercept. The test result led us to reject the null hypothesis that none of the covariates are linearly related to the log-odds of dropping out ($p = 4.643e^{-16} < 0.05$). This suggests that at least one predictor in the model is significantly associated with dropout status .

Additionally, Wald tests were used to evaluate the significance of each individual predictors. It is revealed that "Standardized Cumulative Quercus visit counts in Weeks 1-4" ($p = 3e^{-08}$), "First Term Course Count = 4" indicator ($p = 0.0255$) and "First Term Course Count = 5" indicator ($p = 0.0255$) have significant p-values. This suggest that Quercus engagement and First Term Course Count have a significant linear relationship with the log-odds of dropping out.

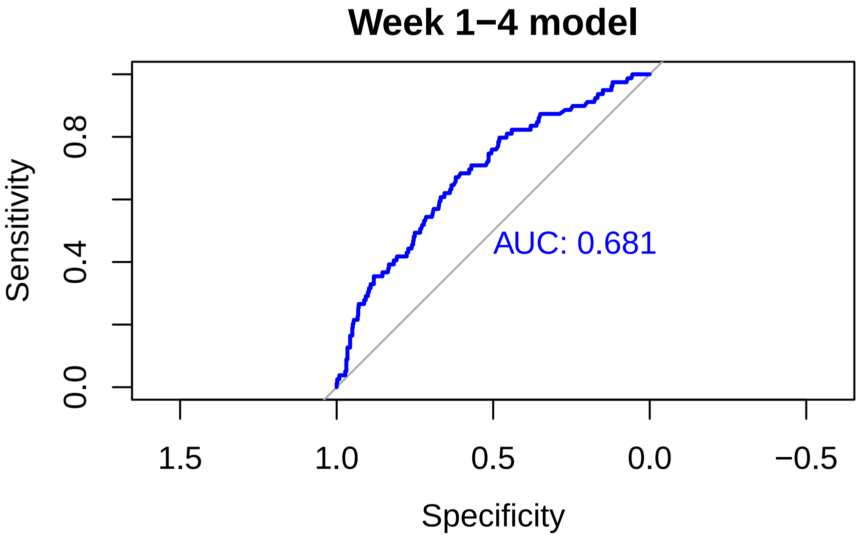


Figure 14: Receiver operating characteristic (ROC) Curve for the Week 1-4 Model Predicting Dropout. The ROC curve displays the diagnostic performance of the logistic regression model using Week 1-4 predictors, with sensitivity plotted against specificity. The area under the curve (AUC) is 0.681.

According to figure 14, the area under the ROC curve is 0.681, indicating moderate discriminative ability of the model to distinguish between students who dropped out and those who remained active after year 1.

3.32 Model 2: Weeks 1-8

Below is a summary table of all coefficient estimates of the fitted model 2:

Terms	Coefficient estimate	P-value	Standard error	95% confidence interval
Intercept	0.92722	0.1928	0.71195	(-0.46819, 2.32263)
Standardized Cumulative Quercus visit count within Weeks 1-8	-2.74207	7.37e^{-11}	0.42102	(-3.16309, -2.32105)
Incoming GPA > 95%	-1.02124	0.1085	0.63623	(-2.26824, 0.22576)
Incoming GPA 91%-95%	-0.49034	0.3416	0.51556	(-1.50083, 0.52016)
Incoming GPA 86-90%	-0.38389	0.4518	0.51022	(-1.3841, 0.61632)
Incoming GPA 81%-85%	-0.59099	0.2576	0.52207	(-1.61423, 0.43225)
Incoming GPA 76%-80%	0.00636	0.9903	0.52545	(-1.02351, 1.03623)
Incoming GPA No Admission Average	-0.46906	0.3772	0.53121	(-1.51023, 0.57211)
Male	0.11347	0.4104	0.13785	(-0.15672, 0.38366)
Unreported gender	0.49600	0.1315	0.32885	(-0.14853, 1.14053)
International student	-0.17722	0.2850	0.16575	(-0.50209, 0.14765)
First Term Course Count = 2	-0.091997	0.8693	0.55901	(-1.18767, 1.00368)
First Term Course Count = 3	-0.32569	0.5287	0.51699	(-1.33808, 0.6867)
First Term Course Count = 4	-0.91578	0.0756	0.51539	(-1.92594, 0.09438)
First Term Course Count = 5	-0.93116	0.0766	0.52582	(-1.96174, 0.09942)
First Term Course Count = 6	-0.74845	0.4083	0.90510	(-2.52241, 1.02551)

Interpretation:

The baseline of the model is a female domestic student with incoming GPA less than 76%, enrolled to only 1 course during first term and has never visited Quercus within Weeks 1-8. Assuming fixed values for all other predictors, the odds of dropping out if a student visited Quercus 10% days of the first eight weeks of class (i.e. $0.1 \times 56 = 5.6$ days) is $exp(-2.74207 * 0.1) = 0.7601747$ times compared to those who have never visited Quercus, assuming fixed levels for all other predictors. Alternatively, the odds of dropping out if a student visited Quercus for an additional day during the first eight weeks of class is $exp(-2.74207 * 1/56) = 0.9522139$. This suggests that more frequent engagement with the platform early on is linked to a lower likelihood of dropping out. Compared to what the Model 1 standardized cumulative visit predictor coefficient implies, the odds is smaller in size in Model 1, potentially indicating that more frequent visit within the first 4 weeks of class may be more effective in reducing dropping out likelihood than first 8 weeks of class (i.e. an odd ratio closer to 0).

Similar to Model 1, when controlling for other variables, students with an incoming GPA above 95% experience the greatest reduction in dropout odds compared to those with a GPA below 76%. Male students generally have higher odds of dropping out than female students, whereas international students are less likely to drop out than domestic students. Students enrolled in more courses in their first term are less likely to drop out compared to those taking only one course.

LRT test on the model rejected the null hypothesis of no statistically significant linear relationship between any of the predictors and log-odds of dropping out ($p < 2.2e^{-16}$). Wald test indicated significant p-values only for the “Standardized Cumulative Quercus visit counts in Weeks 1-4” predictor ($p = 7.37e^{-11}$). This implies that among all predictors, only Quercus engagement predictor has significant linear relationship with the log-odds of dropping out.

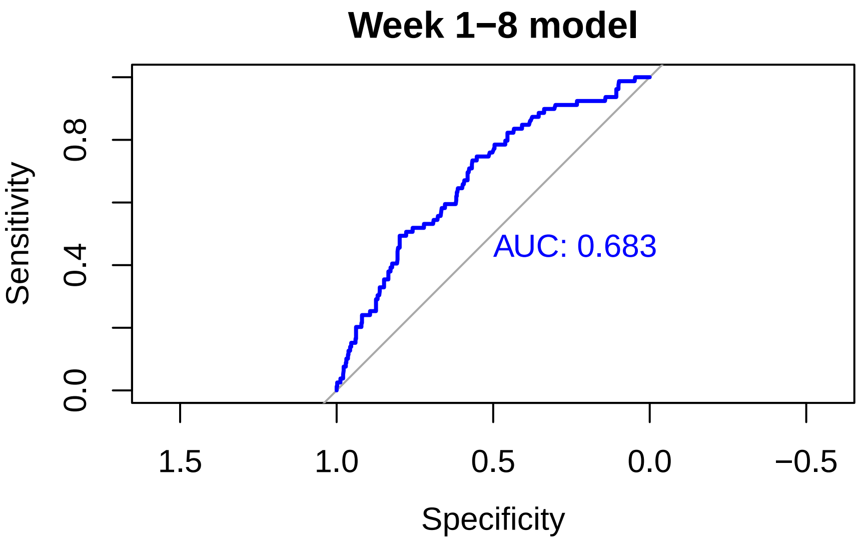


Figure 15: Receiver operating characteristic (ROC) Curve for the Week 1-8 Model Predicting Dropout. The ROC curve displays the diagnostic performance of the logistic regression model using Week 1-4 predictors, with sensitivity plotted against specificity. The area under the curve (AUC) is 0.683.

According to figure 15, the area under the ROC curve is 0.683, indicating moderate and improved (compared to Model 1) discriminative ability of the model to distinguish between students who dropped out and those who remained active after year 1.

3.33 Model 3: Weeks 1-12

Below is a summary table of all coefficient estimates of the fitted model 3:

Terms	Coefficient estimate	P-value	Standard error	95% confidence interval
Intercept	0.86669	0.224	0.71296	(-0.53072, 2.2641)
Standardized Cumulative Quercus visit count within Weeks 1-12	-2.865841	1.21e^{-12}	0.403370	(-3.269211, -2.462471)
Incoming GPA > 95%	-0.98491	0.123	0.63885	(-2.23706, 0.26724)
Incoming GPA 91%-95%	-0.46895	0.366	0.51875	(-1.4857, 0.5478)
Incoming GPA 86-90%	-0.35737	0.487	0.5137	(-1.3644, 0.64966)
Incoming GPA 81%-85%	-0.57967	0.27	0.52572	(-1.61006, 0.45072)
Incoming GPA 76%-80%	-0.00327	0.995	0.52913	(-1.0406, 1.03406)
Incoming GPA No Admission Average	-0.42246	0.43	0.53506	(-1.47117, 0.62625)
Male	0.10722	0.438	0.13828	(-0.1648, 0.37925)
Unreported gender	0.49185	0.136	0.32984	(-0.15455, 1.13825)
International student	-0.16923	0.307	0.16558	(-0.49376, 0.15531)
First Term Course Count = 2	0.0107	0.985	0.56305	(-1.09296, 1.11436)
First Term Course Count = 3	-0.17803	0.733	0.52228	(-1.20168, 0.84562)
First Term Course Count = 4	-0.73623	0.158	0.52158	(-1.75854, 0.28608)
First Term Course Count = 5	-0.73081	0.17	0.53283	(-1.77513, 0.31351)
First Term Course Count = 6	-0.57351	0.529	0.91105	(-2.35915, 1.21213)

Interpretation:

The baseline of the model is a female domestic student with incoming GPA less than 76%, enrolled to only 1 course during first term and has never visited Quercus within Weeks 1-12. Assuming fixed values for all other predictors, the odds of dropping out if a student visited Quercus 10% days of the first 12 weeks of class (i.e. $0.1 \times 84 = 8.4$ days) is $exp(-2.865841 * 0.1) = 0.7508239$ times compared to those who have never visited Quercus, assuming fixed levels for all other predictors. Alternatively, the odds of dropping out if a student visited Quercus for an additional day during the first 12 weeks of class is $exp(-2.865841 * 1/84) = 0.9664583$. This suggests that more frequent engagement with the platform early on is linked to a lower likelihood of dropping out. Compared to what the Models 1 and 2 standardized cumulative visit predictor coefficients imply, the odds is greater in size in Model 3, potentially indicating that more frequent visit within earlier period of fall term may be more effective in reducing dropping out likelihood (i.e. an odd ratio closer to 0).

Similar to Models 1 and 2, when controlling for other variables, students with an incoming GPA above 95% experience the greatest reduction in dropout odds compared to those with a GPA below 76%. Male students generally have higher odds of dropping out than female students, whereas international students are less likely to drop out than domestic students. Students enrolled in more courses in their first term are less likely to drop out compared to those taking only one course.

LRT test on Model 3 gives a significantly low p-value ($p < 2.2e^{-16}$), rejecting the null hypothesis of no statistically significant linear relationship between any of the predictors and log-odds of dropping out.

Similarly to Model 2, the Wald test indicated significantly low p-values only for the "Standardized Cumulative Quercus visit counts in Weeks 1-12" predictor ($p = 1.21e^{-12}$), implying that only the Quercus engagement predictor has significant linear relationship with the log-odds of dropping out.

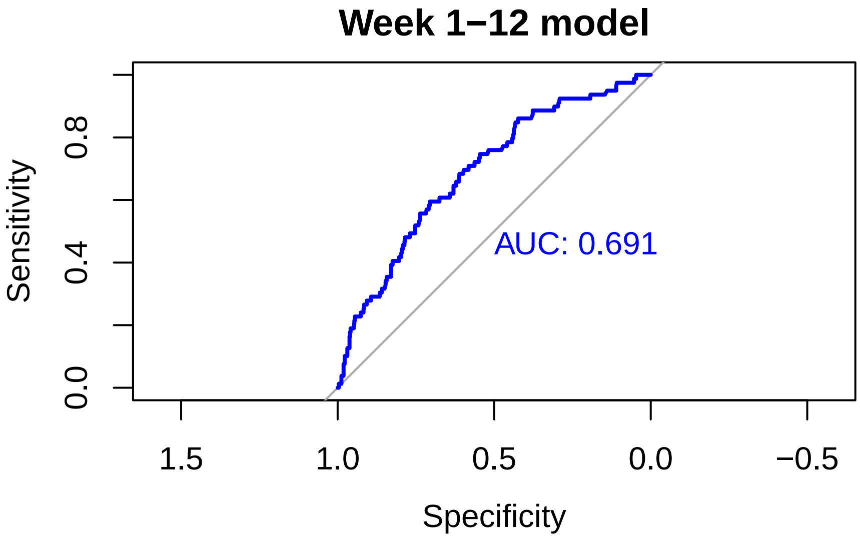


Figure 16: Receiver operating characteristic (ROC) Curve for the Week 1-12 Model Predicting Dropout. The ROC curve displays the diagnostic performance of the logistic regression model using Week 1-4 predictors, with sensitivity plotted against specificity. The area under the curve (AUC) is 0.691.

According to figure 16, the area under the ROC curve is 0.691, indicating moderate discriminative ability of the model to distinguish between students who dropped out and those who remained active after year 1. Note that the AUC is higher than both Model 1 and Model 2, indicating improved performance as we increase the cumulative counts to a longer period.

3.34 Model 4: Weeks 1-15

Below is a summary table of all coefficient estimates of the fitted model 4:

Terms	Coefficient estimate	P-value	Standard error	95% confidence interval
Intercept	0.87564	0.22	0.71464	(-0.52404, 2.27532)
Standardized Cumulative Quercus visit count within Weeks 1-15	-3.089662	1.5e^{-14}	0.401910	(-3.491572, -2.687752)
Incoming GPA > 95%	-0.94997	0.139	0.6415	(-2.2084, 0.30846)
Incoming GPA 91%-95%	-0.44484	0.394	0.52176	(-1.46749, 0.57781)
Incoming GPA 86-90%	-0.33604	0.516	0.51694	(-1.34931, 0.67723)
Incoming GPA 81%-85%	-0.56593	0.285	0.52917	(-1.60305, 0.47119)
Incoming GPA 76%-80%	0.00328	0.995	0.53257	(-1.04057, 1.04713)
Incoming GPA No Admission Average	-0.39153	0.467	0.53851	(-1.446, 0.66294)
Male	0.0991	0.475	0.13871	(-0.17276, 0.371)
Unreported gender	0.47705	0.149	0.33033	(-0.17039, 1.12449)
International student	-0.18109	0.276	0.16617	(-0.50686, 0.14468)
First Term Course Count = 2	0.10649	0.85	0.56493	(-1.00174, 1.21472)
First Term Course Count = 3	-0.05092	0.923	0.52459	(-1.07811, 0.97628)
First Term Course Count = 4	-0.57803	0.271	0.52468	(-1.60648, 0.45042)
First Term Course Count = 5	-0.55197	0.304	0.53679	(-1.60419, 0.50025)
First Term Course Count = 6	-0.43337	0.636	0.91497	(-2.22671, 1.35997)

Interpretation:

The baseline of the model is a female domestic student with incoming GPA less than 76%, enrolled to only 1 course during first term and has never visited Quercus within fall term.

Assuming fixed values for all other predictors, the odds of dropping out if a student visited Quercus 10% days of the fall term (i.e. 0.1 × 98 = 9.8 days) is $exp(-3.089662 * 0.1) = 0.7342056$ times compared to those who have never visited Quercus, assuming fixed levels for all other predictors. Alternatively, the odds of dropping out if a student visited Quercus for an additional day during fall term is $exp(-3.089662 * 1/98) = 0.9689646$. This suggests that more frequent engagement with the platform early on is linked to a lower likelihood of dropping out. Compared to what all previous models' cumulative visit predictor coefficients imply, the odds has the largest absolute value in Model 4, potentially indicating that more frequent visit within earlier period of fall term may be more effective in reducing dropping out likelihood (i.e. an odd ratio closer to 0).

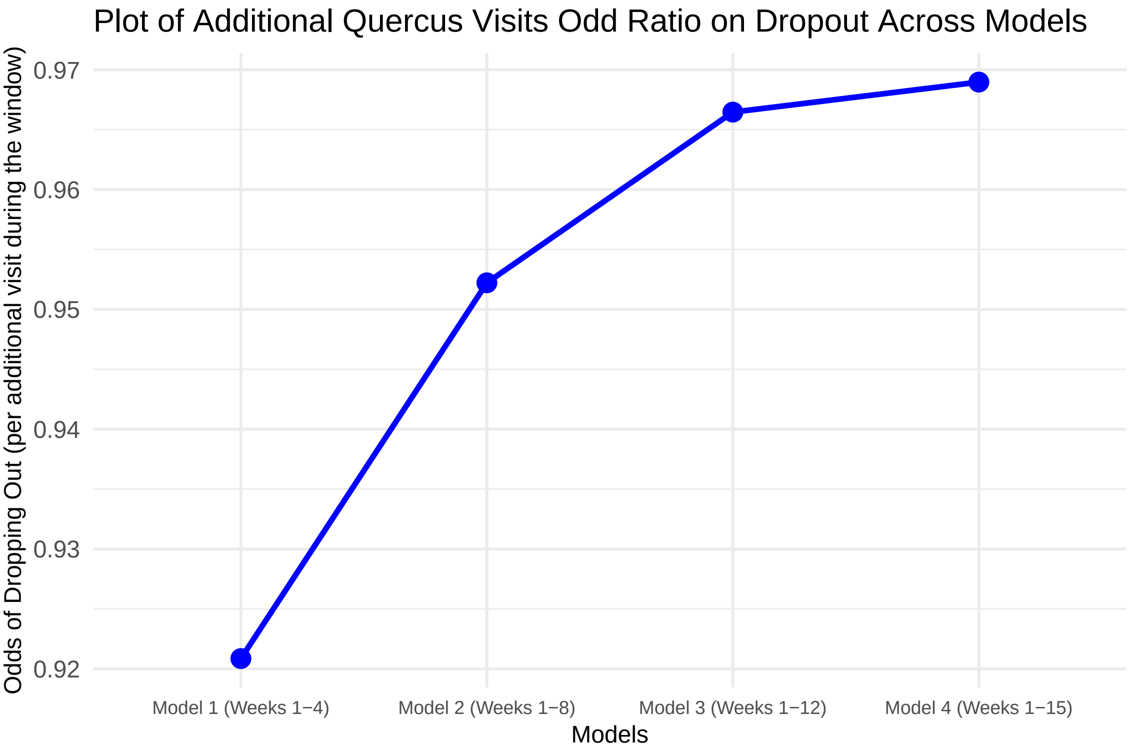


Figure 17: Plot of Additional Quercus Visits Odd Ratio on Dropout Across Models. The plot shows the relationship between the observation window length (x-axis) and the odds ratio of dropping out given an additional Quercus visit (y-axis). The x-axis represents four models with increasing observation windows: Weeks 1-4, Weeks 1-8, Weeks 1-12, and Weeks 1-15, while the y-axis represents the corresponding odds ratios. The trend indicates that as the observation window lengthens, the odds ratio increases toward 1 and there is a diminishing impact of additional Quercus Visits on dropout risk over longer time windows.

Figure 17 illustrates how the impact of an additional Quercus visit on student dropout risk diminishes as the observation window lengthens. The odds ratio, which represents the likelihood of dropping out given one additional day of Quercus usage, increases towards 1 as the time window extends from Weeks 1-4 to Weeks 1-15. This indicates that early engagement (shorter windows) has a stronger effect on decreasing dropout risk, whereas visits over the entire term (longer windows) show a weaker association. As the odds ratio approaches 1, the effect of an additional visit becomes less significant, suggesting that differences in dropout risk due to Quercus engagement levels are more pronounced earlier in the term.

Similar to previous models, when controlling for other variables, students with an incoming GPA above 95% experience the greatest reduction in dropout odds compared to those with a GPA below 76%. Male students generally have higher odds of dropping out than female students, whereas international students are less likely to drop out than domestic students. Students enrolled in more courses in their first term are less likely to drop out compared to those taking only one course.

LRT test on Model 4 gives a significantly low p-value ($p < 2.2e^{-16}$), rejecting the null hypothesis of no statistically significant linear relationship between any of the predictors and log-odds of dropping out.

Similarly to Models 2 and 3, the Wald test indicated significantly low p-values only for the "Standardized Cumulative Quercus visit counts in Weeks 1-12" predictor ($p = 1.5e^{-14}$), implying that only the Quercus engagement predictor has significant linear relationship with the log-odds of dropping out.

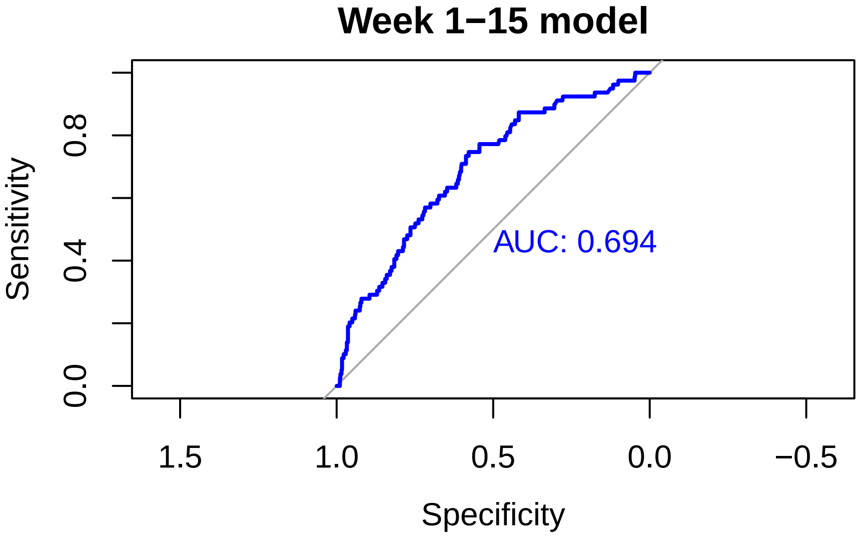


Figure 18: Receiver operating characteristic (ROC) Curve for the Week 1-15 Model Predicting Dropout. The ROC curve displays the diagnostic performance of the logistic regression model using Week 1-4 predictors, with sensitivity plotted against specificity. The area under the curve (AUC) is 0.694.

According to figure 18, the area under the ROC curve is 0.694, indicating moderate discriminative ability of the model to distinguish between students who dropped out and those who remained active after year 1. Note that the AUC is higher than any previous models, indicating improved performance as we increase the cumulative counts to a longer period.

3.4 Sensitivity, Specificity, and Misclassification rates

Sensitivity, specificity, and misclassification rate are key performance metrics used to evaluate the effectiveness of the classification models. Sensitivity, also known as the true positive rate, measures the model's ability to correctly identify students who actually drop out. A higher sensitivity indicates that the model is effective at detecting at-risk students. Specificity, on the other hand, is the true negative rate and represents the model's capacity to correctly identify students who remain active. High specificity means the model accurately classifies students who do not drop out, avoiding false alarms. The misclassification rate is the proportion of all predictions that are incorrect, combining both false positives (predicting dropout when a student remains active) and false negatives (failing to identify a dropout). A lower misclassification rate indicates better overall model accuracy, but it must be interpreted alongside sensitivity and specificity to understand the model's strengths and weaknesses in identifying dropouts versus active students.

Below is a table summarizing sensitivity, specificity, and misclassification rate on the unseen testing data for all 4 models using different threshold values (i.e. the cut-off probability to classify a student into the drop-out group):

Threshold	Model	Sensitivity	Specificity	Misclassification Rate
0.1	Model 1: Weeks 1-4 Model	0.5696	0.6799	0.3333
	Model 2: Weeks 1-8 Model	0.5443	0.6885	0.3288
	Model 3: Weeks 1-12 Model	0.5823	0.7074	0.3076
	Model 4: Weeks 1-15 Model	0.5696	0.7126	0.3045
0.15	Model 1: Weeks 1-4 Model	0.3671	0.8434	0.2136
	Model 2: Weeks 1-8 Model	0.3544	0.8451	0.2136
	Model 3: Weeks 1-12 Model	0.3165	0.8554	0.2091
	Model 4: Weeks 1-15 Model	0.3291	0.8451	0.2167

Threshold	Model	Sensitivity	Specificity	Misclassification Rate
0.2	Model 1: Weeks 1-4 Model	0.2532	0.9294	0.1515
	Model 2: Weeks 1-8 Model	0.2152	0.9208	0.1636
	Model 3: Weeks 1-12 Model	0.2278	0.9260	0.1576
	Model 4: Weeks 1-15 Model	0.2658	0.9225	0.1561
0.25	Model 1: Weeks 1-4 Model	0.1139	0.9656	0.1364
	Model 2: Weeks 1-8 Model	0.1266	0.9587	0.1409
	Model 3: Weeks 1-12 Model	0.1899	0.9587	0.1333
	Model 4: Weeks 1-15 Model	0.2025	0.9587	0.1318
0.3	Model 1: Weeks 1-4 Model	0.0380	0.9793	0.1333
	Model 2: Weeks 1-8 Model	0.0759	0.9742	0.1333
	Model 3: Weeks 1-12 Model	0.1013	0.9690	0.1348
	Model 4: Weeks 1-15 Model	0.1139	0.9690	0.1333

The performance evaluation of the four models highlights the fundamental trade-off between sensitivity and specificity when selecting a classification threshold for identifying students at risk of dropping out. Lower threshold values, such as 0.1, result in higher sensitivity (54.43%–58.23%), meaning the models capture more actual dropouts. However, this comes at the cost of lower specificity (67.99%–71.26%), meaning many students who are not at risk may be incorrectly flagged. While this ensures that more at-risk students are identified, it may also lead to unnecessary interventions. Conversely, higher threshold values, such as 0.3, drastically increase specificity (96.90%–97.93%), but this improvement in specificity comes with a sharp decline in sensitivity (3.80%–11.39%). While this minimizes false positives, it also means that many students who could benefit from intervention may go unnoticed.

Ultimately, the optimal threshold depends on the acceptable trade-off between false positives and false negatives. Since the goal is to provide early interventions, it may be preferable to flag more students who will ultimately remain after first year rather than miss those who are truly at risk. This suggests prioritizing thresholds that maintain reasonable sensitivity while keeping misclassification rates manageable.

When comparing models at the same threshold, they generally perform similarly in terms of specificity and overall misclassification rate. Sensitivity values are also comparable across models, except at a threshold of 0.3, where the Weeks 1-12 and Weeks 1-15 models perform notably better than the Weeks 1-4 and Weeks 1-8 models. However, given the extremely low sensitivity values at this threshold, such comparisons mainly highlight which models perform less poorly rather than which are truly effective. In practical applications, a threshold of 0.3 may not be viable due to its inability to capture at-risk students.

Overall, the results suggest that different time-window models yield similar performance. However, given that none of the models achieve sensitivity above 60%, they appear overly conservative, favoring predictions of student retention. This conservatism represents a critical weakness in identifying at-risk students, limiting the models' practical utility for proactive dropout intervention strategies.

3.5 Limitations

Firstly, since our dataset is not longitudinal, it does not capture how probabilities evolve over time. As a result, we are unable to assess changes in probability of dropping out at different time points, such as reporting the probability at Week 4 versus Week 5, or determining the rate of change in probabilities over time. This limits our ability to explore dynamic shifts in the dataset and the potential influence of temporal factors.

Secondly, our models exhibit very low sensitivity. In this study, sensitivity refers to the model's ability to correctly identify students who are likely to drop out. The primary goal of this study is to identify students who may drop out, so that interventions can be targeted to help them stay engaged. However, the low sensitivities from the models indicate that our models struggles to correctly predict students who are at risk of dropping out, hence hinders the opportunities to intervene and offer support to these students before they disengage. In summary, the models are less useful than expected in providing insights for preventing dropouts reducing their overall value in the context of our study's objectives.

Thirdly, data imbalance is an issue in our dataset. Specifically, we observe a disproportionate number of students who remained active after Year 1. This imbalance creates challenges in training and evaluating our model. In particular, when applying a naive "predict-all-negative" approach (i.e. predicting any student to remain active after year 1), the misclassification rate is 11.97% on the testing dataset. In fact, the final models demonstrate worse misclassification rates, this suggest that the models are not offering any improvement over the naive model and, as such, might not provide the expected level of predictive power.

4. Conclusion

4.1 Key findings

The study finds that early engagement with Quercus is a significant predictor of first-year student retention at the University of Toronto. Across all models, a higher standardized cumulative Quercus visit count was associated with lower dropout probability, indicating that reduced engagement correlates with increased attrition risk. Exploratory data analysis further reveals that students who remained active after Year 1 engaged with Quercus consistently throughout the term, whereas those who dropped out exhibited lower or irregular usage. This pattern suggests that students who disengage early or inconsistently may benefit from early intervention strategies.

Model comparisons suggest that the first four weeks of the term are particularly critical for identifying at-risk students. The model using standardized cumulative Quercus visits from Weeks 1–4 demonstrated comparable predictive accuracy to models incorporating a longer observation window, emphasizing that early engagement is a key factor in retention. Students who show minimal or no engagement with Quercus during this period face a heightened dropout risk, reinforcing the need for timely interventions within the first month.

Demographic and academic factors also played a role—male students, domestic students, those with an incoming GPA below 76%, and those enrolled in fewer first-term courses were more likely to drop out. However, while Quercus engagement was statistically significant, these individual characteristics had limited predictive power. Despite this, all models exhibited low sensitivity, highlighting the limited potential of leveraging early engagement data for timely interventions and underscoring the need to improve model sensitivity to enhance its practical utility in student retention efforts.

4.2 Next step

The Wald test results indicate that most non-Quercus engagement predictors are statistically insignificant. Building on this insight, the next step in this research would involve implementing feature selection techniques, such as LASSO or StepAIC, to systematically eliminate demographic and other non-contributory predictors that do not significantly enhance model performance. This process aims to refine the model by reducing overfitting, enhancing predictive accuracy, and improving overall interpretability. Furthermore, it is crucial to assess whether Quercus visit count predictors remain after selection. If these variables are excluded, it may suggest that Quercus engagement is not a reliable indicator of dropout risk, highlighting the need to explore alternative behavioral, academic, or engagement-related predictors.

Further validation is necessary to assess the model's generalizability. Currently, the dataset has been split into training and testing subsets within the same cohort. While this approach provides initial performance insights, testing the models on an independent dataset—such as data from a different student cohort—would offer a more robust evaluation of predictive reliability across different academic years.

Lastly, we may consider expanding this research by incorporating full-year Quercus engagement data along with end-of-year predictors, such as Year 1 CGPA and final academic standing, to improve dropout risk prediction. Leveraging full-year engagement patterns can provide a more comprehensive understanding of students' academic activity behaviors and trajectories, capturing both early warning signs and later shifts in performance. Additionally, integrating outcome-based variables like CGPA and academic standing at the end of the first year can enhance the predictive model's accuracy, as these measures often reflect cumulative challenges or successes that strongly correlate with retention. Note that relying on full-year data may limit the registrar's ability to conduct early outreach during the Fall term. Nonetheless, using full year data could still prove valuable for identifying students at risk of not returning for their second year, enabling targeted outreach efforts immediately after the first year and provides academic support during the summer to encourage students to re-enroll for the following Fall term.