Predict Student Performance From Game Play

Yueqi Xu

2023-05-23

Abstract

This report focuses on the analysis and prediction of student performance in a game-based learning environment. The study aims to identify the optimal combination of features and choice of model for predicting student performance and to explore the association between a student's performance on earlier questions and their performance on subsequent questions within the same gameplay session. Additionally, the report investigates how predictive models for student performance can be integrated into real-time feedback mechanisms to enhance the learning experience.

Logistic regression models and random forest models are employed. and the findings reveal that there is a significant association between a student's performance on earlier questions and their performance on subsequent questions within the same gameplay session. Moreover, the use of random forest models trained on balanced data with all available features as predictors yields the best prediction perofrmance.

1 Topic: Knowledge Tracing in Game-based Learning

Game-based learning is a highly engaging and effective teaching method that utilizes interactive games to facilitate student learning, and the effectiveness of game-based learning can be further enhanced by incorporating knowledge tracing techniques to track the knowledge and skills of individual students as they progress through the game. However, most game-based learning platforms do not sufficiently make use of knowledge tracing to support individual students. Therefore, the main objective of this project is to help advance research into knowledge-tracing methods for game-based learning, in order to to create more effective learning experiences for students.

To achieve this objective, the project will involve developing a predictive model to predict whether a player will answer in-session questions correctly in a game-based learning environment.

2 Dataset: Game Logs From Jo Wilder

The model will be built using the dataset from the Jo Wilder online educational game, available on Kaggle. The dataset records game log of 23,562 game play sessions, contains information about game setting (full screen or not, music on/off, etc.), gameplay events (click, hover, etc.), and player performance (whether the in-session questions were answered correctly).

The original dataset contains 26,296,946 observations, where each observation represents an event such as a mouse click or hover in a game play session. Each game play session comprises hundreds or thousands of such events. Below is a list of the variables in the original game log dataset:

- session_id: the ID of the session the event took place in

- index: the index of the event for the session
- elapsed_time: how much time has passed (in milliseconds) between the start of the session and when the event was recorded
- event_name: the name of the event type (e.g. object_click, object_hover, map_click, etc.)
- name: the event name (e.g. identifies whether a notebook_click is is opening or closing the notebook)
- level: what level of the game the event occurred in (0 to 22)
- page: the page number of the event (only for notebook-related events)
- room_coor_x: the coordinates of the click in reference to the in-game room (only for click events)
- room_coor_y: the coordinates of the click in reference to the in-game room (only for click events)
- screen_coor_x: the coordinates of the click in reference to the player's screen (only for click events)
- screen_coor_y: the coordinates of the click in reference to the player's screen (only for click events)
- hover_duration: how long (in milliseconds) the hover happened for (only for hover events)
- text: the text the player sees during this event
- fqid: the fully qualified ID of the event
- room_fqid: the fully qualified ID of the room the event took place in
- text_fqid: the fully qualified ID of the
- fullscreen: whether the player is in fullscreen mode
- hq: whether the game is in high-quality
- music: whether the game music is on or off
- level_group: which group of levels - and group of questions - this row belongs to (0-4, 5-12, 13-22)

Each game play session in the dataset consists of 18 questions, which are distributed across three checkpoints: 3 questions in the first checkpoint, 10 questions in the second checkpoint, and 5 questions in the third checkpoint. The student performance for each question is recorded in a separate dataset. The performance dataset contains two variables:

- session_id: the unique identifier for each game play session along with its corresponding question number
- correct: binary variable indicating whether the question was answered correctly (1 for correct and 0 for incorrect)

The original dataset is aggregated by game play sessions and checkpoints, and relevant information such as game settings, user activities, and player performance (correctness of answers to questions given at each checkpoint) are grouped and summarized for each checkpoint in each game play session.

3 Scientific Questions

This study will focus on addressing the following three scientific questions:

1. What is the optimal combination of features (e.g., time spent on task, number of mouse clicks, etc.) and choice of model for predicting student performance (correctness of answering questions) in a game play session?
2. Is there an association between a student's performance on earlier questions and their performance on subsequent questions within the same game play session. In other words, can we predict a student's performance in latter questions based on their performance on previous questions in the same game play session?
3. How can the predictive models for student performance (correctness of answers) during game play sessions be integrated into real-time feedback mechanisms in games to provide timely and personalized feedback to students, thereby enhancing their learning experience and performance in the game?

4 Descriptive Statistics

The following table presents the descriptive statistics of the related data for the first checkpoint:

	Mean	Std Dev	Min	25%	Mdn	75%	Max
fullscreen:	0.14	0.35	0.00	0.00	0.00	0.00	1.00
hq:	0.12	0.33	0.00	0.00	0.00	0.00	1.00
music:	0.93	0.26	0.00	1.00	1.00	1.00	1.00
hover_duration:	43722.46	1714429.29	0.00	8747.00	14100.50	23654.75	221783815.00
n_actions:	168.96	52.48	85.00	137.00	158.00	187.00	2628.00
elapsed_time:	1310528.62	23813846.31	846.00	199177.00	269920.00	367914.75	1986921747.00
n_event_name:	10.01	0.97	7.00	9.00	10.00	11.00	11.00
n_name:	3.65	0.62	3.00	3.00	4.00	4.00	6.00
n_fqid:	24.84	2.43	19.00	23.00	24.00	26.00	35.00
n_room_fqid:	6.41	0.49	5.00	6.00	6.00	7.00	7.00
n_text_fqid:	14.55	2.16	9.00	13.00	14.00	16.00	26.00
n_notebook_click:	3.47	4.41	0.00	0.00	2.00	6.00	86.00
n_object_hover:	4.55	2.53	0.00	3.00	4.00	6.00	26.00
n_map_hover:	1.92	1.50	0.00	1.00	2.00	3.00	22.00
n_cutscene_click:	33.43	7.62	24.00	29.00	32.00	36.00	293.00
n_person_click:	20.58	3.62	9.00	18.00	19.00	22.00	87.00
n_navigate_click:	76.73	38.83	25.00	53.00	68.00	89.00	1874.00
n_observation_click:	1.73	2.11	0.00	0.00	1.00	3.00	53.00
n_notification_click:	7.78	2.13	5.00	6.00	8.00	9.00	32.00
n_object_click:	15.49	12.35	6.00	9.00	12.00	17.00	451.00
n_map_click:	2.29	1.33	1.00	2.00	2.00	2.00	121.00
n_page1:	0.69	1.25	0.00	0.00	0.00	2.00	20.00
n_page0:	2.78	3.78	0.00	0.00	2.00	4.00	83.00
n_historicalsociety:	125.73	39.84	61.00	102.00	116.00	138.00	1673.00
n_kohlcenter:	39.87	17.89	16.00	30.00	36.00	45.00	839.00
n_capitol:	3.36	3.30	2.00	2.00	2.00	3.00	149.00
correctness:	0.88	0.19	0.00	0.67	1.00	1.00	1.00
q1:	0.73	0.45	0.00	0.00	1.00	1.00	1.00
q2:	0.98	0.14	0.00	1.00	1.00	1.00	1.00
q3:	0.93	0.25	0.00	1.00	1.00	1.00	1.00

Some findings from the table of descriptive statistics:

- There are no missing values in the dataset.
- Most players have music turned on during game play, but only a small portion have high quality and full screen settings enabled.
- Potential outliers: The maximum `elapsed_time` recorded is 1.9869217×10^9 milliseconds, which is equivalent to 23 days. This seems unlikely as each game play session typically takes only half to a few hours to complete, and thus the related observation may not be useful. Potential outliers are also observed for variables like `hover_duration`, `n_actions`, `n_observation_click`, and `n_kohlcenter`, and require further investigation.

The descriptive statistics for the related data of the second and third checkpoints also show similar findings.

After further investigating the outliers, observations with `elapsed_time` exceeding 1 day are removed. This is because a typical game play session for this game usually lasts half to an hour, with the content before the first checkpoint even able to be finished within 10 minutes. Therefore, it is highly unlikely for a player to spend more than a day on the first checkpoint. Such extreme `elapsed_time` are more likely to be the result of data collection errors or players who just started the game and did not complete it. Thus, removing these observations can help to reduce potential errors and ensure data accuracy. A table of descriptive statistics of the updated data is presented below.

	Mean	Std Dev	Min	25%	Mdn	75%	Max
fullscreen:	0.14	0.35	0.00	0.00	0.00	0.00	1.00
hq:	0.12	0.33	0.00	0.00	0.00	0.00	1.00
music:	0.93	0.26	0.00	1.00	1.00	1.00	1.00
hover_duration:	34329.23	924050.74	0.00	8747.00	14098.00	23638.50	73568746.00
n_actions:	168.86	52.34	85.00	137.00	158.00	186.00	2628.00
elapsed_time:	673075.61	4146213.44	846.00	199099.00	269652.00	367255.50	85970623.00
n_event_name:	10.00	0.97	7.00	9.00	10.00	11.00	11.00
n_name:	3.65	0.63	3.00	3.00	4.00	4.00	6.00
n_fqid:	24.84	2.43	19.00	23.00	24.00	26.00	35.00
n_room_fqid:	6.41	0.49	5.00	6.00	6.00	7.00	7.00
n_text_fqid:	14.55	2.16	9.00	13.00	14.00	16.00	26.00
n_notebook_click:	3.47	4.41	0.00	0.00	2.00	6.00	86.00
n_object_hover:	4.55	2.53	0.00	3.00	4.00	6.00	26.00
n_map_hover:	1.92	1.50	0.00	1.00	2.00	3.00	22.00
n_cutscene_click:	33.40	7.55	24.00	29.00	32.00	36.00	293.00
n_person_click:	20.57	3.59	9.00	18.00	19.00	22.00	87.00
n_navigate_click:	76.69	38.79	25.00	53.00	68.00	89.00	1874.00
n_observation_click:	1.73	2.11	0.00	0.00	1.00	3.00	53.00
n_notification_click:	7.77	2.12	5.00	6.00	8.00	9.00	32.00
n_object_click:	15.47	12.33	6.00	9.00	12.00	17.00	451.00
n_map_click:	2.29	1.33	1.00	2.00	2.00	2.00	121.00
n_page1:	0.69	1.25	0.00	0.00	0.00	2.00	20.00
n_page0:	2.78	3.78	0.00	0.00	2.00	4.00	83.00
n_historicalsociety:	125.65	39.73	61.00	102.00	116.00	138.00	1673.00
n_kohlcenter:	39.85	17.87	16.00	30.00	36.00	45.00	839.00
n_capitol:	3.36	3.30	2.00	2.00	2.00	3.00	149.00
correctness:	0.88	0.19	0.00	0.67	1.00	1.00	1.00
q1:	0.73	0.45	0.00	0.00	1.00	1.00	1.00
q2:	0.98	0.14	0.00	1.00	1.00	1.00	1.00
q3:	0.93	0.25	0.00	1.00	1.00	1.00	1.00

Similar changes are made to the datasets `checkpoint2` and `checkpoint3`.

A snippet of the table presents the descriptive statistics stratified by the correctness of question 3 is shown below:

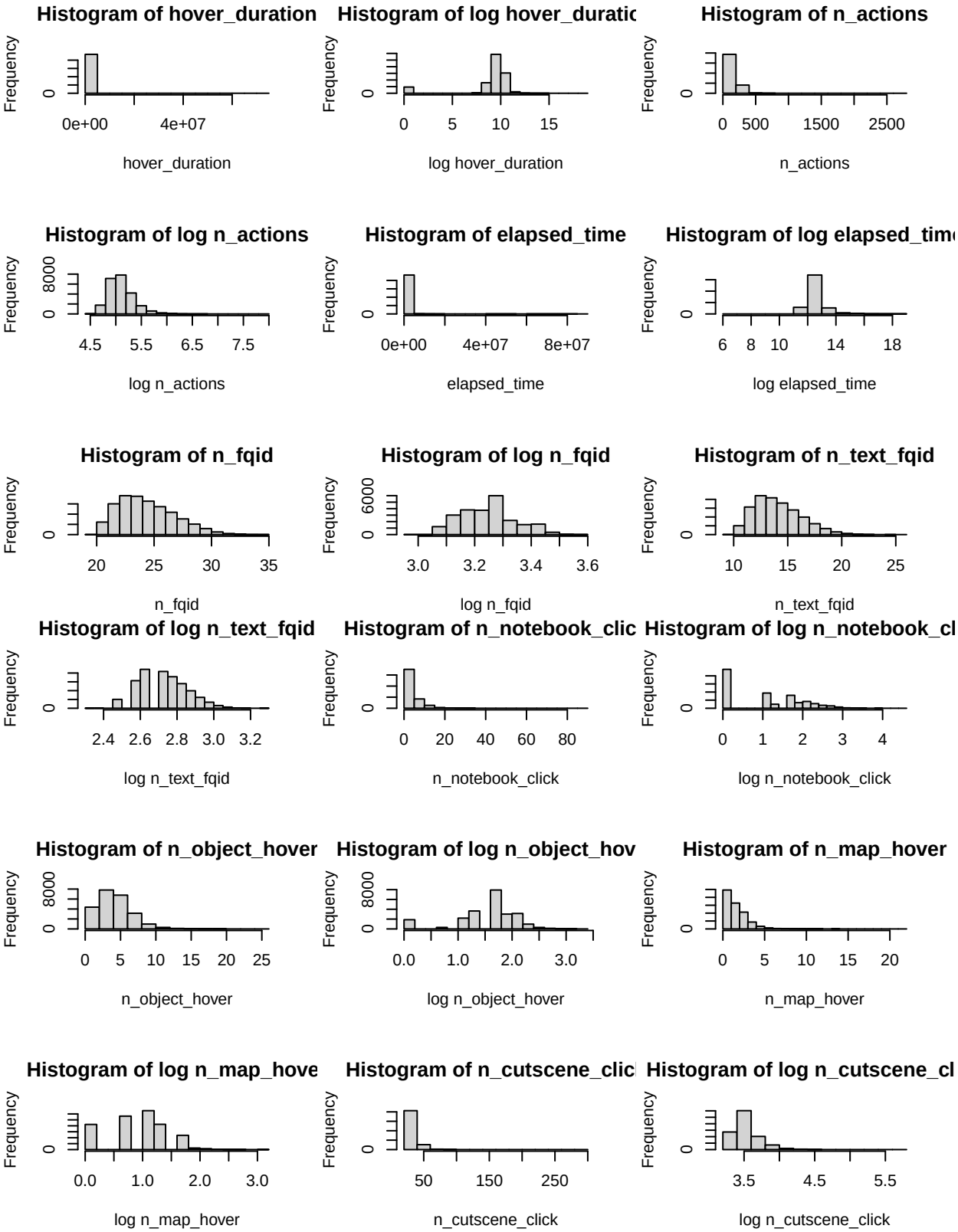
	0	1	Overall
	(N=1550)	(N=21969)	(N=23519)
fullscreen			
Mean (SD)	0.133 (0.340)	0.139 (0.346)	0.139 (0.346)
Median [Min, Max]	0 [0, 1.00]	0 [0, 1.00]	0 [0, 1.00]
hq			
Mean (SD)	0.115 (0.319)	0.121 (0.326)	0.120 (0.325)
Median [Min, Max]	0 [0, 1.00]	0 [0, 1.00]	0 [0, 1.00]
n_notification_click			
Mean (SD)	7.90 (2.31)	7.76 (2.11)	7.77 (2.12)
Median [Min, Max]	8.00 [5.00, 24.0]	8.00 [5.00, 32.0]	8.00 [5.00, 32.0]
n_object_click			
Mean (SD)	21.5 (20.1)	15.0 (11.5)	15.5 (12.3)
Median [Min, Max]	16.0 [6.00, 291]	12.0 [6.00, 451]	12.0 [6.00, 451]
n_page1			
Mean (SD)	0.954 (1.42)	0.673 (1.24)	0.691 (1.25)
Median [Min, Max]	0 [0, 11.0]	0 [0, 20.0]	0 [0, 20.0]
n_page0			
Mean (SD)	4.10 (4.86)	2.68 (3.68)	2.78 (3.78)
Median [Min, Max]	3.00 [0, 49.0]	2.00 [0, 83.0]	2.00 [0, 83.0]
q1			
Mean (SD)	0.484 (0.500)	0.745 (0.436)	0.728 (0.445)
Median [Min, Max]	0 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]
q2			
Mean (SD)	0.926 (0.261)	0.983 (0.131)	0.979 (0.144)
Median [Min, Max]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]

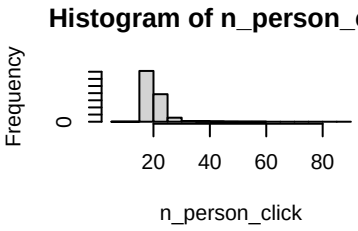
It is worth noting that the average values of certain variables, such as `hq` and `n_object_click`, vary between players who answered question 3 correctly and those who did not. This observation suggests that these variables may be potential predictors in a model for question 3. Furthermore, there are differences in the correctness of questions 1 (`q1`) and 2 (`q2`) between players with different behavior on question 3. This finding implies a possible association between the correctness of earlier questions and the correctness of later ones.

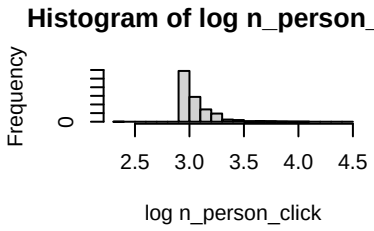
Based on the initial exploration of the dataset, it appears that the data aligns well with the scientific questions and objectives of the project. The information gathered from the game play sessions and checkpoints, including game settings, user activities, and player performance, is grouped and summarized in a way that will allow for analysis and modeling. The feasibility of the project is promising given the available data.

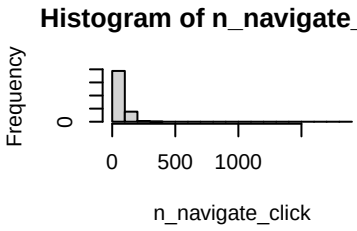
The histograms of the distributions of continuous variables are presented below, both on their original scale and on a log scale. It can be observed that some variables, such as `hover_duration` and `elapsed_time`, have a large range and long right tails, indicating that a log-transformation may be appropriate. On the other hand, variables such as `n_fqid` and `n_text_fqid` appear to be more normally distributed on their original scale and may not require log-transformation.

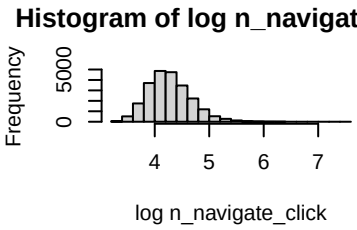
A bar plot of overall correctness by question is also presented. Note that the correctness of questions 3, 13, and 18 are approximately 0.93, 0.27, and 0.95, respectively, which are imbalanced. Therefore, techniques used to adjust the class distribution, such as oversampling or undersampling, may be considered.

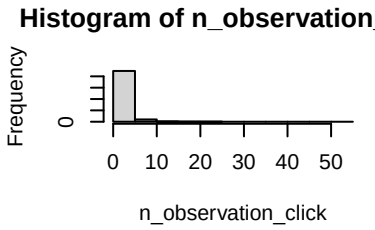

Histogram of hover_duration
Frequency
0
0e+00
4e+07
hover_duration
Histogram of log hover_duratic
Frequency
0
0
5
10
15
log hover_duration
Histogram of n_actions
Frequency
0
0
500
1500
2500
n_actions
Histogram of log n_actions
Frequency
0
8000
4.5
5.5
6.5
7.5
log n_actions
Histogram of elapsed_time
Frequency
0
0e+00
4e+07
8e+07
elapsed_time
Histogram of log elapsed_tim
Frequency
0
6
8
10
14
18
log elapsed_time
Histogram of n_fqid
Frequency
0
20
25
30
35
n_fqid
Histogram of log n_fqid
Frequency
0
6000
3.0
3.2
3.4
3.6
log n_fqid
Histogram of n_text_fqid
Frequency
0
10
15
20
25
n_text_fqid
Histogram of log n_text_fqid
Frequency
0
2.4
2.6
2.8
3.0
3.2
log n_text_fqid
Histogram of n_notebook_clic
Frequency
0
0
20
40
60
80
n_notebook_click
Histogram of log n_notebook_cl
Frequency
0
0
1
2
3
4
log n_notebook_click
Histogram of n_object_hover
Frequency
0
8000
0
5
10
15
20
25
n_object_hover
Histogram of log n_object_hov
Frequency
0
8000
0.0
1.0
2.0
3.0
log n_object_hover
Histogram of n_map_hover
Frequency
0
0
5
10
15
20
n_map_hover
Histogram of log n_map_hove
Frequency
0
0.0
1.0
2.0
3.0
log n_map_hover
Histogram of n_cutscene_clic
Frequency
0
50
150
250
n_cutscene_click
Histogram of log n_cutscene_cl
Frequency
0
3.5
4.5
5.5
log n_cutscene_click

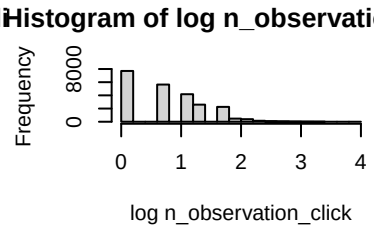

Histogram of n_person_click
Frequency
0
20
40
60
80
n_person_click

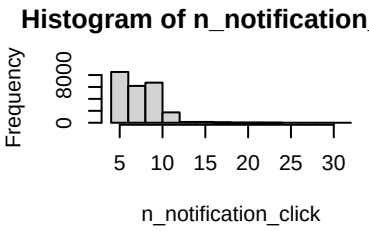

Histogram of log n_person_clic
Frequency
0
2.5
3.0
3.5
4.0
4.5
log n_person_click

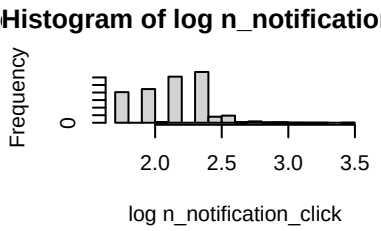

Histogram of n_navigate_clicl
Frequency
0
0
500
1000
n_navigate_click

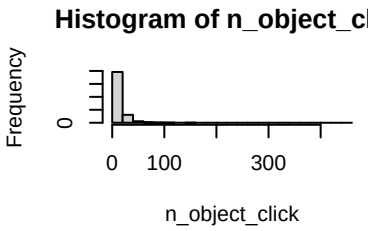

Histogram of log n_navigate_cli
Frequency
0
5000
4
5
6
7
log n_navigate_click

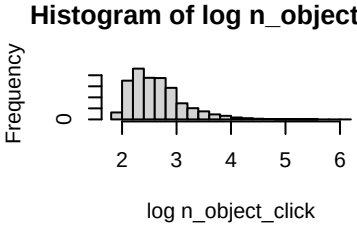

Histogram of n_observation_cli
Frequency
0
0
10
20
30
40
50
n_observation_click

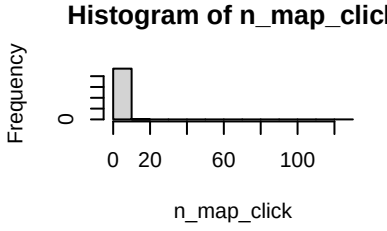

Histogram of log n_observation_
Frequency
0
8000
0
1
2
3
4
log n_observation_click

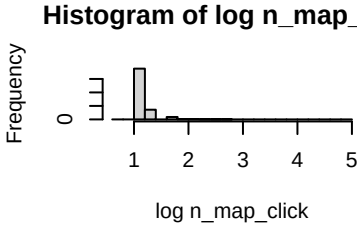

Histogram of n_notification_cli
Frequency
0
8000
5
10
15
20
25
30
n_notification_click

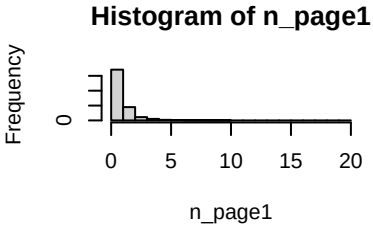

Histogram of log n_notification_
Frequency
0
2.0
2.5
3.0
3.5
log n_notification_click

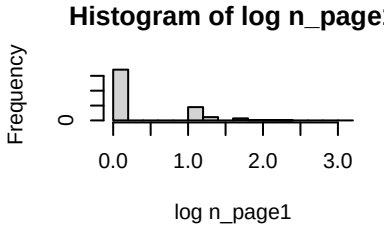

Histogram of n_object_click
Frequency
0
0
100
300
n_object_click

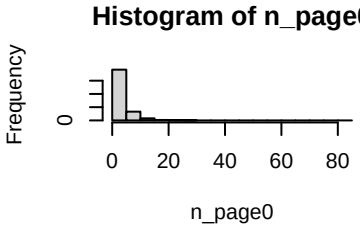

Histogram of log n_object_clic
Frequency
0
2
3
4
5
6
log n_object_click

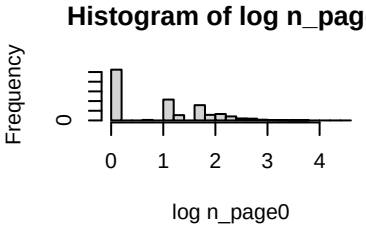

Histogram of n_map_click
Frequency
0
0
20
60
100
n_map_click

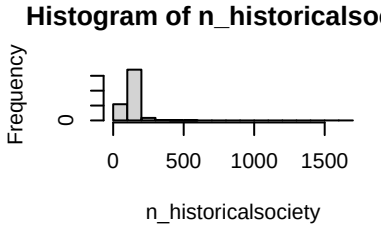

Histogram of log n_map_click
Frequency
0
1
2
3
4
5
log n_map_click

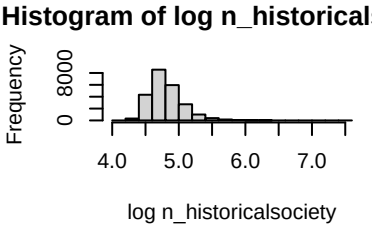

Histogram of n_page1
Frequency
0
0
5
10
15
20
n_page1

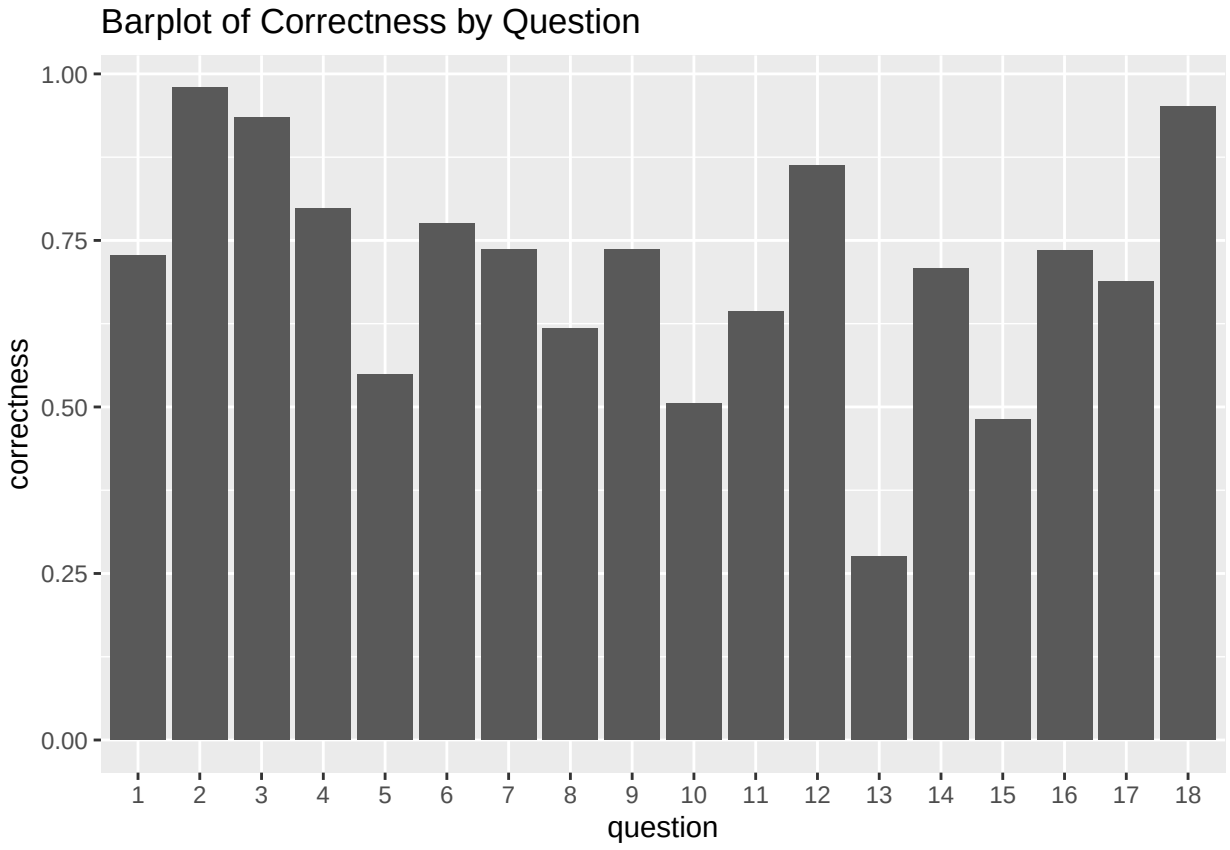

Histogram of log n_page1
Frequency
0
0.0
1.0
2.0
3.0
log n_page1


Histogram of n_page0
Frequency
0
0
20
40
60
80
n_page0


Histogram of log n_page0
Frequency
0
0
1
2
3
4
log n_page0


Histogram of n_historicalsocie
Frequency
0
0
500
1000
1500
n_historicalsociety


Histogram of log n_historicalsoc
Frequency
0
8000
4.0
5.0
6.0
7.0
log n_historicalsociety



5 Statistical Methods

Recall that there are 18 questions in each gameplay session, and each gameplay session consists of three sections. At the end of each section, there is a checkpoint, where a number of questions are given as an assessment (3 questions in the first checkpoint, 10 in the second, and 5 in the third). The dataset used for analysis records the game settings and user behavior in each section up until the checkpoints, and user actions in each checkpoint are not recorded. That is, all questions in the same checkpoint share the same set of data, but the information available at each checkpoint is different.

However, it's important to note that the sets of variables recorded in each section are different. The second section contains a few extra features than the first section, and the third section adds a few more new variables to the second section. Therefore, with the consideration of the difference in the amount of information and sets of features, I performed the analysis at the level of the checkpoint. And for simplicity, I only investigated questions 3, 13, and 18, which are the last question from each checkpoint.

Given the relatively small number of variables compared to the sample size, all available features are retained in the predictive model. Thus, the answer to the first scientific question "optimal combination of features for predicting student performance in a gameplay session" is "all features available". Besides, I am also interested in exploring whether different models yield different prediction performance. Hence, both logistic regression and random forest models are employed for each of the three questions.

Furthermore, we observed earlier that the correctness of questions 3, 13,and 18 are approximately 0.93, 0.27, and 0.95, respectively, which are highly imbalanced. To address this imbalance, models trained on oversampled data are also considered. As a result, for each of questions 3, 13, and 18, four models are fitted:

1. Model 1a/2a/3a: A logistic regression model of the correctness of question 3/13/18 vs. the overall correctness of all previous questions, adjusting for all other features that appeared in the first/second/third

checkpoint, using the original data.

2. Model 1b/2b/3b: A logistic regression model of the correctness of question 3/13/18 vs. the overall correctness of all previous questions, adjusting for all other features that appeared in the first/second/third checkpoint, using the oversampled data.

3. Model 4a/5a/6a: A random forest model predicting the correctness of question 3/13/18 given the overall correctness of all previous questions and all other features that appeared in the first/second/third checkpoint, using the original data.

4. Model 4b/5b/6b: A random forest model predicting the correctness of question 3/13/18 given the overall correctness of all previous questions and all other features that appeared in the first/second/third checkpoint, using the oversampled data.

The dataset is split into training and testing sets (80/20). All twelve models are fitted on the training set and evaluated on the testing set. The performance of models for the sample question is cross-compared in the result section.

6 Result

The resulting models for questions 3, 13 and 18 as well as their performance are summarized below. The model outputs are shown in the appendix. Overall, we have enough evidence from data that there is an association between a student's performance on earlier questions and their performance on subsequent questions within the same game play session.

Furthermore, when oversampling was applied to balance the data, the random forest model exhibited slightly higher prediction accuracy on the testing set in general. Therefore, for predictive purposes, I recommend utilizing a random forest model trained on balanced data, with all available features as predictors.

6.1 Question 3

According to model 1a in the appendix, for two populations with the same game setting and actions but differ in the overall correctness of previous problems by 1 unit, we estimate that the odds ratio of answering question 3 correctly between these two groups is 1.622 (95% CI based on non-robust standard errors: [1.405, 1.84]), with the group with higher overall correctness having higher probability of answering question 3 correctly.

At the 5% confidence interval, we are able to reject there is not association between the overall correctness of earlier questions in a gameplay session and the correctness of question 3 ($p = 1.73 \times 10^{-48}$).

Without oversampling the minority group, the logistic regression model achieved a testing accuracy of 0.934, while the random forest model achieved a testing accuracy of 0.934. After applying oversampling, the testing accuracy of the logistic regression model dropped to 0.712, while the testing accuracy of the random forest model remained at 0.934.

6.2 Question 13

According to model 2a in the appendix, for two populations with the same game setting and actions but differ in the overall correctness of previous problems by 1 unit, we estimate that the odds ratio of answering question 13 correctly between these two groups is 1.971 (95% CI based on non-robust standard errors: [1.755, 2.188]), with the group with higher overall correctness having higher probability of answering question 13 correctly.

At the 5% confidence interval, we are able to reject there is not association between the overall correctness of earlier questions in a gameplay session and the correctness of question 13 ($p = 4.63 \times 10^{-71}$).

Without oversampling the minority group, the logistic regression model achieved a testing accuracy of 0.732, while the random forest model achieved a testing accuracy of 0.736. After applying oversampling, the testing accuracy of the logistic regression model remained at 0.732, while the testing accuracy of the random forest model increased to 0.785.

6.3 Question 18

According to model 2a in the appendix, for two populations with the same game setting and actions but differ in the overall correctness of previous problems by 1 unit, we estimate that the odds ratio of answering question 18 correctly between these two groups is 4.666 (95% CI based on non-robust standard errors: [4.223, 5.112]), with the group with higher overall correctness having higher probability of answering question 18 correctly.

At the 5% confidence interval, we are able to reject there is not association between the overall correctness of earlier questions in a gameplay session and the correctness of question 18 ($p = 5.61 \times 10^{-94}$).

Without oversampling the minority group, the logistic regression model achieved a testing accuracy of 0.95, while the random forest model achieved a testing accuracy of 0.951. After applying oversampling, the testing accuracy of the logistic regression model stayed at 0.95, while the testing accuracy of the random forest model improved slightly to 0.953.

7 Discussion (Assumption & Limitations)

Based on previous analysis, it has been determined that at 5% confidence level, there is evidence from data supporting the association between a student's performance on earlier questions and their performance on subsequent questions within the same game play session. In terms of prediction, a random forest model with all available features as as predictor and balanced training data is recommended, as it has the best prediction accuracy on the testing set.

In order to provide timely and personalized feedback to students, one potential application is to leverage this predictive model to continuously monitor the students' actions and performance during the gameplay session. A system can be built based on this model to offer feedback and guidance whenever it detects a need for further attention or improvement. However, further investigation and development are necessary to establish a more comprehensive and refined feedback mechanism.

Other related assumptions and limitations of this analysis include:

1. It is assumed that the observations are independent, i.e. the game play sessions are not related to one another. Situations such as players play the game repeatly are not taken into consideration, and may affect the predicting result. However, given the large sample size utilized in the analysis, the impact is expected to be minimal.
2. The analysis is based solely on the data collected from Jo wilder online educational game, and may not generalize well to other educational games or populations. Care should be taken when applying these results to different contexts.
3. The original data exhibited significant class imbalance, resulting in biased model performance for models trained on original data. The use of oversampling techniques helped address this issue and produced more reasonable confusion matrices. However, it is important to recognize that oversampling may introduce its own biases, and caution should be exercised when interpreting the results.

Sources

Jo Wilder online educational game: https://pbswisconsineducation.org/jowilder/play-the-game/

Kaggle - Predict Student Performance from Game Play: https://www.kaggle.com/competitions/predict-student-performance-from-game-play/data

Appendix: Model outputs

Question 3

Model 1a: Logistic Regression Without Oversampling

```
##
## Call:
## glm(formula = q3 ~ ., family = "binomial", data = q3_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9030   0.2343   0.2901   0.3854   1.3544
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.016401   2.531319   0.006  0.99483
## fullscreen           -0.021642   0.095974  -0.226  0.82159
## hq                    0.013868   0.100681   0.138  0.89045
## music                 0.119380   0.115529   1.033  0.30145
## hover_duration        0.012963   0.026063   0.497  0.61892
## n_actions            -0.058500   1.336987  -0.044  0.96510
## elapsed_time         -0.105351   0.040631  -2.593  0.00952 **
## n_event_name          0.008209   0.081433   0.101  0.91970
## n_name                0.037032   0.087645   0.423  0.67265
## n_fqid                1.791509   0.765975   2.339  0.01934 *
## n_room_fqid          -0.106415   0.081091  -1.312  0.18942
## n_text_fqid           0.068514   0.527346   0.130  0.89663
## n_notebook_click     -0.123774   0.205449  -0.602  0.54687
## n_object_hover        0.241254   0.108342   2.227  0.02596 *
## n_map_hover          -0.164596   0.081779  -2.013  0.04415 *
## n_cutscene_click     -0.653436   0.282541  -2.313  0.02074 *
## n_person_click       -0.645175   0.276277  -2.335  0.01953 *
## n_navigate_click     -0.890250   0.480150  -1.854  0.06372 .
## n_observation_click  -0.405475   0.094032  -4.312 1.62e-05 ***
## n_notification_click  0.116175   0.158842   0.731  0.46454
## n_object_click       -0.820660   0.150958  -5.436 5.44e-08 ***
## n_map_click           0.160725   0.171465   0.937  0.34857
## n_page1              -0.153087   0.102573  -1.492  0.13558
## n_page0              -0.067127   0.167432  -0.401  0.68848
## n_historicalsociety   1.162133   1.199436   0.969  0.33260
## n_kohlcenter          0.559409   0.419472   1.334  0.18233
## n_capitol            -0.079761   0.092115  -0.866  0.38655
## correctness           1.622432   0.110876  14.633  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9141  on 18815  degrees of freedom
## Residual deviance: 8355  on 18788  degrees of freedom
## AIC: 8411
##
## Number of Fisher Scoring iterations: 6
```

Model 1b: Logistic Regression With Oversampling

```
##
## Call:
## glm(formula = q3 ~ ., family = "binomial", data = q3_balanced)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9399  -1.0531   0.6465   0.9732   2.6153
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.119883   1.039443  -3.001 0.002687 **
## fullscreen          -0.042791   0.036850  -1.161 0.245548
## hq                   0.010833   0.038907   0.278 0.780682
## music                0.138500   0.045219   3.063 0.002192 **
## hover_duration       0.015589   0.010139   1.538 0.124152
## n_actions           -0.065376   0.653697  -0.100 0.920336
## elapsed_time        -0.141588   0.017905  -7.908 2.62e-15 ***
## n_event_name         0.032951   0.032030   1.029 0.303599
## n_name               0.004426   0.035172   0.126 0.899866
## n_fqid               2.311585   0.298645   7.740 9.92e-15 ***
## n_room_fqid         -0.085890   0.031386  -2.737 0.006209 **
## n_text_fqid         -0.051843   0.214038  -0.242 0.808613
## n_notebook_click    -0.262074   0.085280  -3.073 0.002118 **
## n_object_hover       0.229527   0.042925   5.347 8.93e-08 ***
## n_map_hover         -0.177761   0.031880  -5.576 2.46e-08 ***
## n_cutscene_click    -0.722058   0.120014  -6.016 1.78e-09 ***
## n_person_click      -0.651109   0.120646  -5.397 6.78e-08 ***
## n_navigate_click    -0.829417   0.199427  -4.159 3.20e-05 ***
## n_observation_click -0.481521   0.038885 -12.383  < 2e-16 ***
## n_notification_click  0.129753   0.066124   1.962 0.049730 *
## n_object_click      -0.831797   0.060946 -13.648  < 2e-16 ***
## n_map_click          0.246401   0.068055   3.621 0.000294 ***
## n_page1             -0.064588   0.043330  -1.491 0.136065
## n_page0              0.048088   0.069668   0.690 0.490039
## n_historicalsociety  1.041265   0.576616   1.806 0.070946 .
## n_kohlcenter         0.536891   0.198423   2.706 0.006814 **
## n_capitol           -0.069095   0.040583  -1.703 0.088653 .
## correctness          1.703236   0.046183  36.880  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46826  on 33815  degrees of freedom
## Residual deviance: 41315  on 33788  degrees of freedom
## AIC: 41371
##
## Number of Fisher Scoring iterations: 4
```

Model 4a: Random Forest Without Oversampling

```
##
## Call:
##  randomForest(formula = q3 ~ ., data = q3_train, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 6.59%
## Confusion matrix:
##   0     1  class.error
## 0 6  1234 0.9951612903
## 1 6 17570 0.0003413746
```

Model 4b: Random Forest With Oversampling

```
##
## Call:
##  randomForest(formula = q3 ~ ., data = q3_balanced, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 0.07%
## Confusion matrix:
##       0     1 class.error
## 0 16240     0 0.000000000
## 1    22 17554 0.001251707
```

Question 13

Model 2a: Logistic Regression Without oversampling

```
##
## Call:
## glm(formula = q13 ~ ., family = "binomial", data = q13_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5407  -0.8430  -0.6131   1.1609   2.6220
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -4.49489    2.48553  -1.808 0.070541 .
## fullscreen            0.00770    0.05310   0.145 0.884695
## hq                    0.01608    0.05571   0.289 0.772899
## music                 0.12700    0.06657   1.908 0.056442 .
## hover_duration       -0.06149    0.02911  -2.113 0.034642 *
## n_actions             4.14932    0.86382   4.803 1.56e-06 ***
## elapsed_time         -0.01534    0.02840  -0.540 0.589102
## n_event_name          0.29396    0.11688   2.515 0.011899 *
```

```
## n_name                0.02541    0.04569   0.556 0.578105
## n_fqid               -1.59267    0.49737  -3.202 0.001364 **
## n_room_fqid          -0.15049    0.05303  -2.838 0.004543 **
## n_text_fqid           0.12228    0.25318   0.483 0.629113
## n_notebook_click     -0.27566    0.12483  -2.208 0.027220 *
## n_object_hover        0.06051    0.08325   0.727 0.467324
## n_map_hover          -0.12958    0.05006  -2.588 0.009646 **
## n_cutscene_click      0.62851    0.13692   4.590 4.43e-06 ***
## n_person_click       -0.33880    0.29059  -1.166 0.243658
## n_navigate_click     -0.11997    0.22303  -0.538 0.590630
## n_observation_click  -0.12134    0.06184  -1.962 0.049761 *
## n_notification_click  0.33005    0.16867   1.957 0.050367 .
## n_object_click       -0.92558    0.09425  -9.820  < 2e-16 ***
## n_map_click           0.13607    0.11023   1.234 0.217040
## n_page1               0.07305    0.06242   1.170 0.241850
## n_page2               0.05162    0.05118   1.009 0.313210
## n_page3               0.18090    0.05380   3.363 0.000772 ***
## n_page0               0.15462    0.08578   1.803 0.071449 .
## n_historicalsociety  -1.84766    0.43672  -4.231 2.33e-05 ***
## n_kohlcenter         -0.08387    0.03585  -2.340 0.019295 *
## n_capitol            -0.36054    0.06992  -5.156 2.52e-07 ***
## n_humanecology       -0.39734    0.14594  -2.723 0.006475 **
## n_drycleaner         -0.22422    0.14231  -1.576 0.115129
## n_library            -0.28480    0.18503  -1.539 0.123744
## correctness           1.97085    0.11058  17.824  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22101  on 18786  degrees of freedom
## Residual deviance: 20350  on 18754  degrees of freedom
## AIC: 20416
##
## Number of Fisher Scoring iterations: 4
```

Model 2b: Logistic Regression With Oversampling

```
##
## Call:
## glm(formula = q13 ~ ., family = "binomial", data = q13_balanced)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6033  -0.8404  -0.6147   1.1605   2.6535
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.4063381  2.1011529  -1.621 0.104980
## fullscreen           0.0003087  0.0444176   0.007 0.994455
## hq                   0.0451865  0.0467735   0.966 0.334009
## music                0.1347313  0.0560121   2.405 0.016155 *
## hover_duration      -0.0655294  0.0243898  -2.687 0.007215 **
```

```
## n_actions             2.7918631  0.7493526   3.726 0.000195 ***
## elapsed_time          0.0049406  0.0238675   0.207 0.836008
## n_event_name          0.3054553  0.0982879   3.108 0.001885 **
## n_name                0.0183587  0.0383696   0.478 0.632316
## n_fqid               -1.4962507  0.4146799  -3.608 0.000308 ***
## n_room_fqid          -0.1464956  0.0441459  -3.318 0.000905 ***
## n_text_fqid           0.1899618  0.2131350   0.891 0.372782
## n_notebook_click     -0.2742296  0.1047791  -2.617 0.008865 **
## n_object_hover        0.0965420  0.0700740   1.378 0.168291
## n_map_hover          -0.1452258  0.0423315  -3.431 0.000602 ***
## n_cutscene_click      0.6659036  0.1143247   5.825 5.72e-09 ***
## n_person_click       -0.4962760  0.2465548  -2.013 0.044131 *
## n_navigate_click      0.0092750  0.1901049   0.049 0.961088
## n_observation_click  -0.1274440  0.0523244  -2.436 0.014865 *
## n_notification_click  0.4934217  0.1377061   3.583 0.000339 ***
## n_object_click       -0.8847724  0.0802057 -11.031  < 2e-16 ***
## n_map_click           0.1947668  0.0920130   2.117 0.034283 *
## n_page1               0.0591927  0.0519860   1.139 0.254858
## n_page2               0.0601652  0.0430575   1.397 0.162316
## n_page3               0.2076493  0.0453937   4.574 4.78e-06 ***
## n_page0               0.1410989  0.0730758   1.931 0.053501 .
## n_historicalsociety  -1.3309003  0.3771975  -3.528 0.000418 ***
## n_kohlcenter         -0.0492646  0.0299948  -1.642 0.100499
## n_capitol            -0.2897848  0.0596371  -4.859 1.18e-06 ***
## n_humanecology       -0.2236850  0.1249466  -1.790 0.073415 .
## n_drycleaner          0.0362968  0.1211085   0.300 0.764402
## n_library            -0.1327690  0.1588398  -0.836 0.403230
## correctness           2.0277452  0.0931584  21.767  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31346  on 26786  degrees of freedom
## Residual deviance: 28924  on 26754  degrees of freedom
## AIC: 28990
##
## Number of Fisher Scoring iterations: 4
```

Model 5a: Random Forest Without Oversampling

```
##
## Call:
##  randomForest(formula = q13 ~ ., data = q13_train, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 26.72%
## Confusion matrix:
##       0   1 class.error
## 0 13066 554  0.04067548
## 1  4465 702  0.86413780
```

Model 5b: Random Forest With Oversampling

```
##
## Call:
##  randomForest(formula = q13 ~ ., data = q13_balanced, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 15.49%
## Confusion matrix:
##       0    1 class.error
## 0 19093  412  0.02112279
## 1  3736 3546  0.51304587
```

Question 18

Model 3a: Logistic Regression Without oversampling

```
##
## Call:
## glm(formula = q18 ~ ., family = "binomial", data = q18_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1907   0.1581   0.2178   0.3180   1.4532
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.406290   4.930863  -1.502  0.13309
## fullscreen           -0.279500   0.105923  -2.639  0.00832 **
## hq                    0.114461   0.124524   0.919  0.35800
## music                -0.194686   0.155984  -1.248  0.21199
## hover_duration        0.113892   0.060801   1.873  0.06104 .
## n_actions             1.405784   1.661445   0.846  0.39749
## elapsed_time         -0.076877   0.043899  -1.751  0.07991 .
## n_event_name          0.023725   0.221264   0.107  0.91461
## n_name                0.115252   0.066081   1.744  0.08114 .
## n_fqid                2.700186   0.820956   3.289  0.00101 **
## n_room_fqid           0.038310   0.064744   0.592  0.55404
## n_text_fqid          -0.653342   0.662241  -0.987  0.32386
## n_notebook_click     -0.033676   0.180195  -0.187  0.85175
## n_object_hover       -0.248681   0.155230  -1.602  0.10915
## n_map_hover          -0.231888   0.111659  -2.077  0.03783 *
## n_cutscene_click      0.280247   0.355373   0.789  0.43035
## n_person_click       -0.579084   0.469930  -1.232  0.21785
## n_navigate_click     -0.809490   0.558697  -1.449  0.14737
## n_observation_click   0.017268   0.024498   0.705  0.48090
## n_notification_click  0.335771   0.136958   2.452  0.01422 *
## n_object_click        0.081814   0.126771   0.645  0.51869
## n_map_click          -0.460379   0.172969  -2.662  0.00778 **
## n_page1              -0.089936   0.064004  -1.405  0.15997
## n_page2               0.023601   0.080252   0.294  0.76869
```

```
## n_page3              -0.123781   0.092647  -1.336  0.18153
## n_page4               0.043605   0.076127   0.573  0.56679
## n_page5              -0.048379   0.081127  -0.596  0.55096
## n_page6              -0.098619   0.081131  -1.216  0.22416
## n_page0              -0.014758   0.042546  -0.347  0.72868
## n_historicalsociety  -0.722248   0.768522  -0.940  0.34733
## n_kohlcenter          0.059867   0.048177   1.243  0.21399
## n_capitol             0.256546   0.097073   2.643  0.00822 **
## n_humanecology       -0.045839   0.050656  -0.905  0.36552
## n_drycleaner         -0.009458   0.055212  -0.171  0.86398
## n_library            -0.032854   0.225091  -0.146  0.88395
## n_wildlifecenter      0.239190   0.355216   0.673  0.50071
## n_flaghouse          -0.257179   0.170128  -1.512  0.13062
## correctness           4.665832   0.226879  20.565  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7338.2  on 18722  degrees of freedom
## Residual deviance: 6281.4  on 18685  degrees of freedom
## AIC: 6357.4
##
## Number of Fisher Scoring iterations: 6
```

Model 3b: Logistic Regression With Oversampling

```
##
## Call:
## glm(formula = q18 ~ ., family = "binomial", data = q18_balanced)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5526   0.1242   0.1866   0.2864   1.7888
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -14.877860   3.638089  -4.089 4.32e-05 ***
## fullscreen           -0.255241   0.082037  -3.111 0.001863 **
## hq                    0.263775   0.096430   2.735 0.006230 **
## music                -0.562537   0.135139  -4.163 3.15e-05 ***
## hover_duration       -0.125731   0.043239  -2.908 0.003640 **
## n_actions             0.305818   1.194209   0.256 0.797886
## elapsed_time         -0.168737   0.028518  -5.917 3.28e-09 ***
## n_event_name          0.470361   0.164708   2.856 0.004294 **
## n_name               -0.191039   0.043319  -4.410 1.03e-05 ***
## n_fqid                3.635791   0.587727   6.186 6.16e-10 ***
## n_room_fqid          -0.006038   0.048844  -0.124 0.901619
## n_text_fqid           1.603237   0.491243   3.264 0.001100 **
## n_notebook_click      0.013914   0.131215   0.106 0.915549
## n_object_hover        0.187158   0.110637   1.692 0.090714 .
## n_map_hover          -0.139197   0.080443  -1.730 0.083561 .
## n_cutscene_click      0.964871   0.258316   3.735 0.000188 ***
```

```
## n_person_click        -0.631021   0.338619  -1.864 0.062390 .
## n_navigate_click      -0.200899   0.423962  -0.474 0.635599
## n_observation_click    0.009454   0.017844   0.530 0.596259
## n_notification_click   0.218808   0.101312   2.160 0.030792 *
## n_object_click         0.151170   0.092568   1.633 0.102455
## n_map_click            0.254622   0.126878   2.007 0.044768 *
## n_page1               -0.115129   0.051309  -2.244 0.024842 *
## n_page2                0.112101   0.056066   1.999 0.045561 *
## n_page3                0.237429   0.068608   3.461 0.000539 ***
## n_page4               -0.138331   0.057245  -2.416 0.015672 *
## n_page5                0.008438   0.058568   0.144 0.885446
## n_page6               -0.192096   0.058145  -3.304 0.000954 ***
## n_page0               -0.030255   0.035456  -0.853 0.393490
## n_historicalsociety   -1.592708   0.564122  -2.823 0.004753 **
## n_kohlcenter           0.318942   0.035875   8.890  < 2e-16 ***
## n_capitol              0.162678   0.069211   2.350 0.018750 *
## n_humanecology        -0.196863   0.036577  -5.382 7.36e-08 ***
## n_drycleaner          -0.198509   0.040007  -4.962 6.98e-07 ***
## n_library              0.008440   0.162835   0.052 0.958662
## n_wildlifecenter      -0.060678   0.260297  -0.233 0.815677
## n_flaghouse           -0.027792   0.123077  -0.226 0.821352
## correctness            5.543215   0.169418  32.719  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14707  on 38722  degrees of freedom
## Residual deviance: 11665  on 38685  degrees of freedom
## AIC: 11741
##
## Number of Fisher Scoring iterations: 7
```

Model 6a: Random Forest Without Oversampling

```
##
## Call:
##  randomForest(formula = q18 ~ ., data = q18_train, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##         OOB estimate of  error rate: 4.92%
## Confusion matrix:
##   0     1  class.error
## 0 0   920 1.0000000000
## 1 2 17801 0.0001123406
```

Model 6b: Random Forest With Oversampling

```
##
## Call:
##  randomForest(formula = q18 ~ ., data = q18_balanced, ntree = 500)
```

```
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##         OOB estimate of  error rate: 2.26%
## Confusion matrix:
##     0     1 class.error
## 0 947   877   0.4808114
## 1   0 36899   0.0000000
```