MINDCRAFT

Illuminating Student Performance with Game Data Mining

TEAM: BIG DATA CRUSHERS

Hanish Singla
Sanka Naga Nitesh
Murad Salamov
Anchal Chaudhary
Sparsh Gupta
Sky McReynolds

TABLE OF CONTENTS

Description

The objective of this project is to predict student performance in real-time during game-based learning using a large dataset of game logs. The goal is to advance research into knowledge-tracing methods for game-based learning, ultimately helping developers create more effective educational games and providing educators with dashboards and analytic tools.
In game-based learning, educational games are used as a means of teaching and assessing students' knowledge and skills. However, there is a lack of open datasets available for analyzing and applying data science and learning analytics principles in this context. To address this, the project makes use of one of the largest open datasets of game logs provided by the Field Day Lab, a publicly-funded research lab that designs educational games for various subjects and age groups.

The dataset consists of several files, including the training set (train.csv), the test set (test.csv), a sample submission file (sample_submission.csv), and the correct values for all questions in the training set (train_labels.csv). The dataset is in CSV format and has a size of 4.74 GB.

https://www.kaggle.com/competitions/predict-student-performance-from-game-play/overview/description

The Product or Service

The product that we are ultimately looking to provide an enhancement and refining of game based learning platforms, empowering them to better serve current users, as well as be able to reach and create value for those with potentially underserved needs presently. That takes the form of developing new educational games incorporating our insights in order to better serve individual education goals, like increased information retention, engagement, or tailoring to specific needs. There will be games for both mobile and desktop.

It will also take place in a consistently occurring feedback loop with both students and educators, which in turn provides insights we can transform into new iterations. All of this is in pursuit of our primary goal: improving student performance.

The games will most likely employ a mixture of ad-supported and freemium pricing, though we do explore a subscription-based model as well. Both of these options are elaborated on in the revenue model section below. Depending on the data and feedback we receive, the pricing models may vary between mobile and desktop experiences.

As just mentioned, there will be games for both mobile and desktop, which creates two main ways that users will be interacting with our products. For mobile users, it is somewhat self explanatory, wherever they’re accessing the game on a compatible device. For desktop users, we envision a combination of tutoring/after school educational programs, in-school programs, and at-home usage.

Customer Mix, CLV and Customer Equity

Customer Segments:
Elementary to Middle School Students - 17.8M (US)
In-School Implementation
Tutoring Programs
High School Students - 12.4M (US)
Tutoring Programs
At-Home Use

When calculating CLV, the time frame shifts depending on age of end user, rate of churn by age and place (potentially acquisition method), all of which will have significant impacts on our final values. In turn, when looking at customer equity, it will be determined by what age demographics are most successful, and when in their educational journey it is. May also vary based on success rates on mobile vs. desktop. All of these demographic and user behavioral data will be more readily identifiable once it has become more widely available.

Revenue Model

1. **Advertisements**: We can incorporate ads along with the game play, where we can monetize off a cost-per-click model. For example: if the CPC(Cost-per-Click) is $1.5 and there are around 50000-100000 ad impressions with a click through rate (CTR) of 2-5%.
2. **Freemium Model**: The basic features could be free. Advanced features could play the most trending games and access the most fancy features in games, which will also let us measure elevated brain levels. This could be priced as a premium for $5-$10. We can assume that initially adoption would be low, so 5-10% of the users will go premium.
3. **Data Licensing**: Sell analytical gameplay data to researchers, educators and other interested partners and collaborators for something in between $1000-$2000.
4. **In-App Purchases**: Solutions to complex problems can be offered as in-app purchases for $0.99-$4.99.
5. **Subscription Model**: Alternatively, we can offer a subscription model once the product matures. The model could be basic($2.5), standard($5) and premium($10). These models can be differentiated between level of games and engagement.

Go to Market Strategy

1. **Positioning**: How measuring and immersing students in games leads to enhanced student learning, provides personalized insights and improves academic performance. Focus should be intuition and user friendliness to position itself differently from other game learning platforms out there.

2. **Place**: App Stores, In- Game advertisement (Mobile), Facebook Ads, forum groups, discord channels and community pages on social media (Desktop). We can also have launch events in schools. Also , partnering with game arcades and virtual reality equipment shops.
3. **Promotion**: Services like Kumon with which we can collaborate and promote our services. App store advertisements, word-of-mouth, school collaborations (for ex: state university collaborations to form legacy relationships) by offering trial versions, special discounts and beta versions of the game.. Partner with educational courses sites such as LinkedIn, Coursera Etc. Finally, newsletters and email campaigns are important along with Tv advertisements. Have educational bloggers and subject matter experts promote it as well.
4. **Product**: Educational gaming market, competitive analysis and market trends as we would gauge in customer equity. Gameplay mechanics and rewards scenarios in the game should be compelling to keep users motivated. The feedback mechanism for the product and services have also to be set up so that we can train our model better to improve our games.
5. **Segmentation**: High- School students(49M), College students(20M)

Business Objective

The main objective is to enhance the effectiveness of educational games and improve learning outcomes for students. The priority initiatives include:

1. **Personalized Learning:** Through this we customize the game experience to cater to each student's unique needs and abilities. Utilize adaptive learning algorithms to dynamically adapt game content and difficulty levels according to their performance and progress.
2. **Game Engagement and User Experience:** Through this we want to Craft an educational game that is captivating and immersive for students in the learning process. Designing compelling narratives, including intuitive gameplay mechanics, and building visually and audibly appealing elements will foster motivation and full engagement.
3. **Knowledge Retention and Transfer:** We want to develop an educational game that optimizes knowledge retention and transfer. In Order to achieve this we would want to implement techniques that encourage students to apply learned concepts in diverse contexts, fostering deeper understanding and long-term memory retention.

By prioritizing these initiatives, we aim to create impactful educational games that engage students, support personalized learning, and optimize learning outcomes.

Metrics of Success/KPI

1. **Engagement Rate:** By tracking the time spent per session, we can gain insights into how engaged students are with the game and its adaptive learning features. A higher average time spent per session generally indicates that students are actively involved in

the learning process, exploring various activities, and utilizing the adaptive elements provided by the game.

2. **Dropout Rate:** Measure the percentage of students who discontinue using the game or drop out before completing desired learning objectives. A lower dropout rate signifies higher user satisfaction and suggests that the adaptive learning features effectively maintain student interest.

3. **Learning Outcomes:** Measure the impact of the game's adaptive learning features on students' academic performance or learning outcomes. Compare the scores of students who have engaged with the adaptive learning system to those who have not, and evaluate the difference in achievement.

By monitoring these KPIs, we aim to gain insights into the effectiveness of the adaptive learning system and its impact on student engagement, retention, and learning outcomes.

Key Actionable Business Initiative

In order to address our business objective the key focus must be on addressing and incorporating the **"Adaptive Learning"** concept. To further deep dive into how we can address we can look at 3 pillers

1. **Predict New User Behavior:** With the existing user data, we can predict the competency levels of new users. This allows us to customize the game experience to their individual needs. By analyzing patterns and trends in user behavior, we can anticipate their learning preferences, strengths, and areas for improvement. This information enables us to create personalized learning paths, provide targeted guidance and support, and recommend specific game modules or activities that align with their skill levels and interests. Predicting new user behavior helps us optimize the onboarding process, improve user engagement, and maximize learning outcomes.

2. **Enhancement of Content and User Experience:** To improve educational game engagement, focus on enhancing narratives, gameplay mechanics, visuals, and audio. Develop captivating storylines and well-defined characters to immerse students in the game. Incorporate intuitive and interactive gameplay elements for active participation. Utilize high-quality graphics, animations, and sound effects to create an immersive environment. Address challenges by providing additional resources, hints, and alternative learning paths. By implementing these enhancements, educational games can offer an engaging and interactive learning experience that keeps students motivated and promotes effective knowledge acquisition and retention.

3. **Educator Support and Continuous Improvement:** By providing educators with access to F1 score predictions, they can identify common areas of difficulty and tailor classroom discussions and interventions accordingly. Establishing a feedback loop with students,

educators, and game developers enables continuous improvement of game design and content based on user feedback, research findings, and emerging best practices, ensuring ongoing alignment with educational goals.

Role of Analytics

Analytics will add value to the business initiative in several ways:

1. **Enabling the Business Initiative:** Analytics will enable us to extract valuable insights from the large dataset of game logs. By applying data analysis techniques, we can uncover patterns, trends, and correlations that are not immediately apparent. This information will guide decision-making and inform the design and implementation of targeted interventions.

2. **Refining the Business Initiative:** Analytics will help us refine the business initiative by providing a deeper understanding of the factors that contribute to student performance. Through exploratory data analysis and predictive modeling, we can identify the most influential variables and optimize game design elements, instructional strategies, and personalized learning pathways.

3. **Evaluating the Success of the Business Initiative:** Analytics will play a crucial role in evaluating the effectiveness of the implemented interventions and strategies. By measuring and analyzing key metrics, such as student engagement, knowledge retention, and learning outcomes, we can assess the impact of the initiatives and make data-driven adjustments to improve their effectiveness.

Analytics will serve as a powerful tool to inform decision-making, drive innovation, and continuously improve the game-based learning experience for students. It will provide actionable insights that can enhance student performance and contribute to the overall success of the business initiative.

Analytics Methodology

Data:

The dataset for this project consists of game logs collected from an online educational game named Jo Wilder. The game is an arcade game where users have to finish certain tasks, based on information provided, exploring elements in the game play and answering questions at certain check points. There are 3 check points in the game, where in the first check point, users answer 3 questions, 8 questions in the second check point and 7 questions in the last check point. This information has been collected as part of the Kaggle dataset which can be accessed here.

The data includes information about various game-related events, user behavior, and student performance. It is an existing observational dataset, meaning that we are relying on data that has already been generated rather than designing our own experiment.

The target/outcome variable for this analysis is student performance, specifically whether a student answers questions correctly at different checkpoints in the game. The explanatory variables/features include various aspects such as elapsed time, event type, game level, page number, coordinates of clicks, hover duration, text, fully qualified IDs, full screen mode, high-quality setting, and game music.

Exploring the variation in the data is an essential step in understanding its characteristics. This includes examining the distribution and range of the target/outcome variable (correct/incorrect answers) as well as exploring the distribution, summary statistics, and relationships among the explanatory variables/features.

Let's delve deeper into the various aspects of the dataset:

1. **session_id**: This field represents the ID of the session in which the game events took place. Each session is a unique gameplay session by a student.

2. **index**: It indicates the index of the event within a session, representing the order in which the events occurred.

3. **elapsed_time**: This field denotes the time in milliseconds that has passed between the start of the session and when the event was recorded. It provides information about the timing and duration of each event.

4. **event_name**: It represents the name of the event type, indicating the specific action or interaction that occurred during the gameplay.

5. **name**: This field provides additional information about the event, specifying the details or context of the event. For example, it could indicate whether a notebook was opened or closed, or provide other relevant information about the action performed.

6. **level**: It indicates the level of the game at which the event occurred. The game likely consists of multiple levels, and this field helps in understanding the progression of the gameplay.

7. **page**: This field represents the page number of the event, specifically for notebook-related events. It indicates the specific page or section within the game where the event occurred.

8. **room_coor_x, room_coor_y:** These fields provide the coordinates of the click event in reference to the in-game room. They indicate the location or position within the virtual environment where the click event took place.

9. **screen_coor_x, screen_coor_y**: These fields represent the coordinates of the click event in reference to the player's screen. They provide information about where on the screen the click occurred.

10. **hover_duration**: For hover events, this field indicates the duration in milliseconds for which the hover action took place. It measures the length of time the player hovered over a specific element or area.

11. **text**: This field contains the text that the player sees during the event. It could be instructional text, prompts, or any other relevant information displayed to the player.

12. **fqid**: The fully qualified ID of the event, which uniquely identifies each event in the dataset. It helps in tracking and linking events together.

13. **room_fqid**: The fully qualified ID of the room where the event took place. It provides information about the specific virtual room or environment within the game.

14. **text_fqid**: The fully qualified ID associated with the text displayed during the event. It helps in identifying and connecting events related to specific text elements.

15. **fullscreen**: This field indicates whether the player is in fullscreen mode during the event. It provides information about the player's viewing settings.

16. **hq**: It specifies whether the game is in high-quality mode. This field informs about the graphical settings of the game.

17. **music**: This field indicates whether the game music is turned on or off during the event. It provides insights into the audio settings of the game.

18. **level_group**: This field categorizes the events into groups based on the levels they belong to. The dataset mentions three level groups: 0-4, 5-12, and 13-22. It helps in organizing and analyzing events based on their corresponding levels and questions.

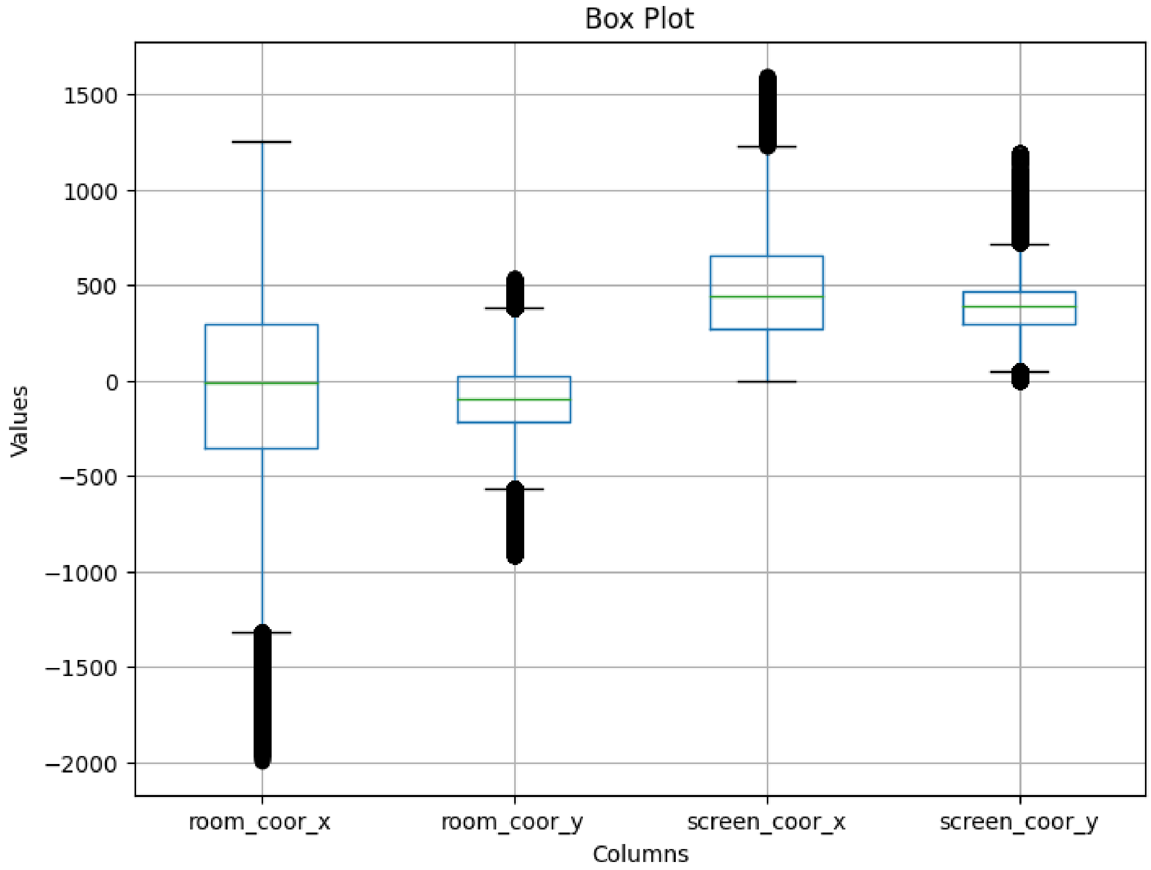
In addition to the mentioned fields, the dataset also includes information about three question checkpoints in the game which were discussed above.

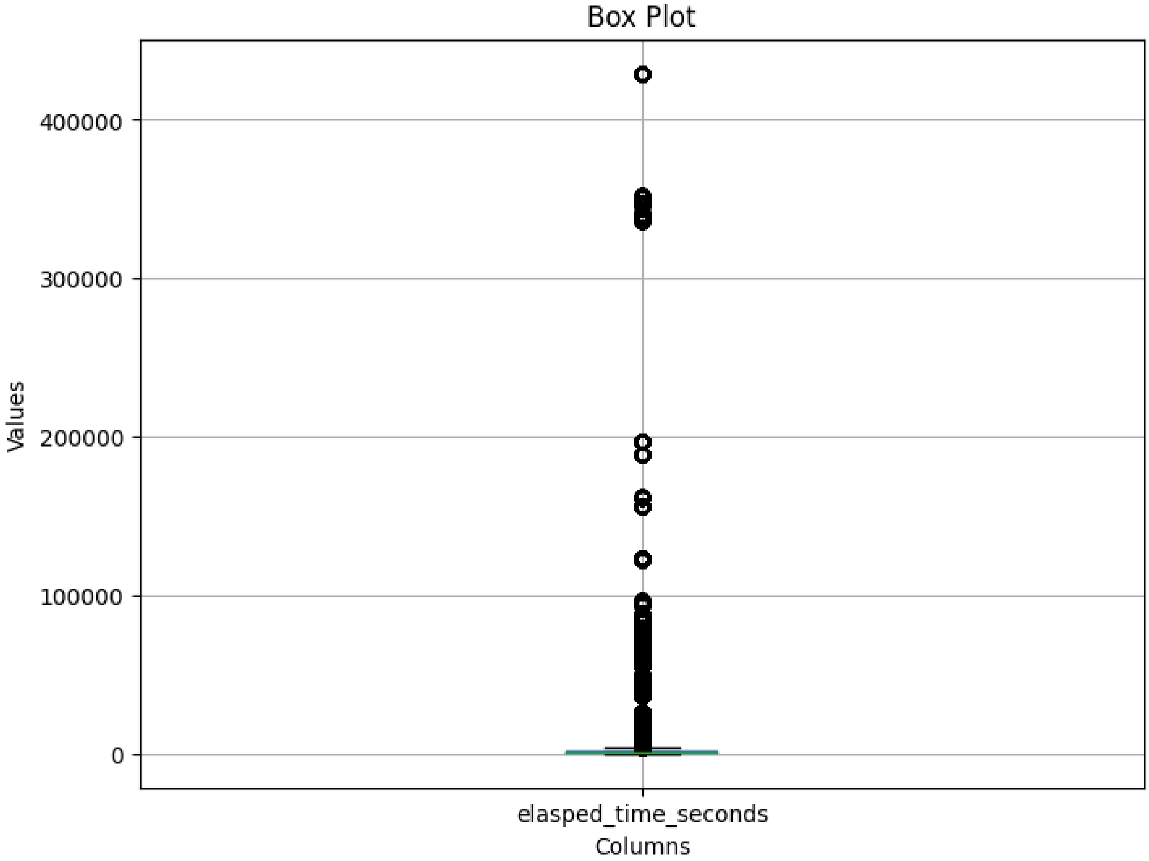
Understanding the patterns and behaviors of students during these checkpoints can provide valuable insights into their learning progress and help in predicting their performance in subsequent checkpoints.

Variation in data:

Session ID: We see a total of 25k different sessions (unique identifiers)

We also see sufficient variation in other data columns which is depicted below:





Elapsed time, also has enough variation (box is not available for the boxplot due to the scale of numbers) ranging between 0 to 400000 seconds

Type of Analytics and Methodology:

The analytics methodology for this project involves a combination of descriptive analytics, predictive analytics, and potentially prescriptive analytics.

1. **Predictive Analytics (Main Focus):** The primary focus of this project is predictive analytics. We aim to develop a predictive model that can accurately forecast whether a student will answer a question correctly based on the available features. So for 18 questions in the dataset, we will have 18 predictive models (binary classification) which will predict the probability of the student answering the questions correctly or not.

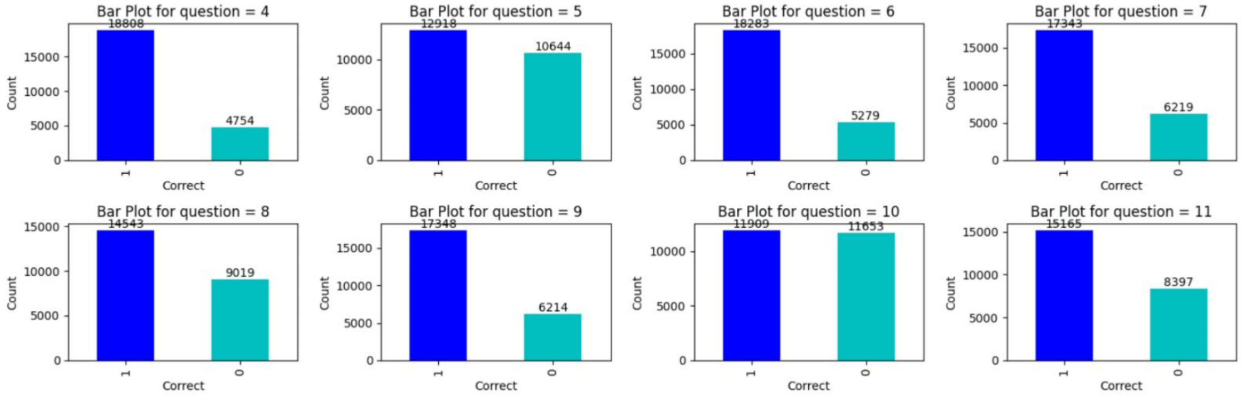
   One of the main aspect of this prediction is that it will take into account, the previous reponses from the user, as a predictor for the future questions. For example: User 1 has already passed check point #1 and answered 3 questions, based on the number of questions users answered correctly, we can input that in the prediction model for the question 4 onwards. Based on these probabilities, future paths for the students can be updated after each check point, and provide them with unique personalized experience.

2. **Descriptive Analytics:** Descriptive analytics will be used to explore and summarize the data. This includes data visualization techniques, summary statistics, and data profiling to gain a comprehensive understanding of the dataset's characteristics and identify initial patterns or trends.

Model Selection

The model selection will depend on the nature of the data and the specific objectives of the analysis. Potential models could include classification algorithms such as logistic regression, decision trees, random forests, or gradient boosting.

One quick note, before we further discuss on model selection - As part of exploratory data analysis, we saw that the training dataset has imbalance between correct and incorrect responses. Most of the responses captured are correct (depicted in the image below), thus considering accuracy as evaluation pattern will give incorrect results. Hence, we are using F1 score as a evaluation criteria for each model.



After careful consideration and analysis, we have chosen Gradient Boosted Trees as our model of choice. Its ability to handle complex relationships, robustness to outliers and missing data, feature importance analysis, and customization options make it a powerful algorithm for our predictive modeling task.

- Gradient Boosted Trees (GBT) is an ensemble machine learning algorithm that combines the power of decision trees and gradient boosting.
- It is a sequential learning algorithm where each subsequent tree in the ensemble corrects the mistakes made by the previous trees, leading to improved performance.

- GBT has shown impressive performance in a wide range of machine learning tasks, including classification and regression. The ensemble of decision trees learns to make accurate predictions by minimizing the loss function through gradient descent.
- By combining multiple weak learners, GBT can capture complex relationships and achieve high predictive accuracy. GBT is effective at capturing nonlinear relationships between features and the target variable.
- Each decision tree in the ensemble focuses on different subsets of features, allowing for the identification of intricate patterns and interactions.
- This capability is especially valuable when dealing with complex datasets where linear models may fall short.GBT is robust to outliers and missing data due to its ensemble nature.
- Outliers have a limited impact on the overall model as subsequent trees correct for the mistakes made by earlier trees. Missing data can be handled through appropriate imputation techniques, ensuring minimal loss of information during the modeling process.

It is important to note that the prescriptive analytics aspect will depend on the availability of actionable features and the ability to establish a causal relationship. If feasible, a suitable experimental or quasi-experimental design will be chosen to evaluate the impact of specific interventions on student performance.

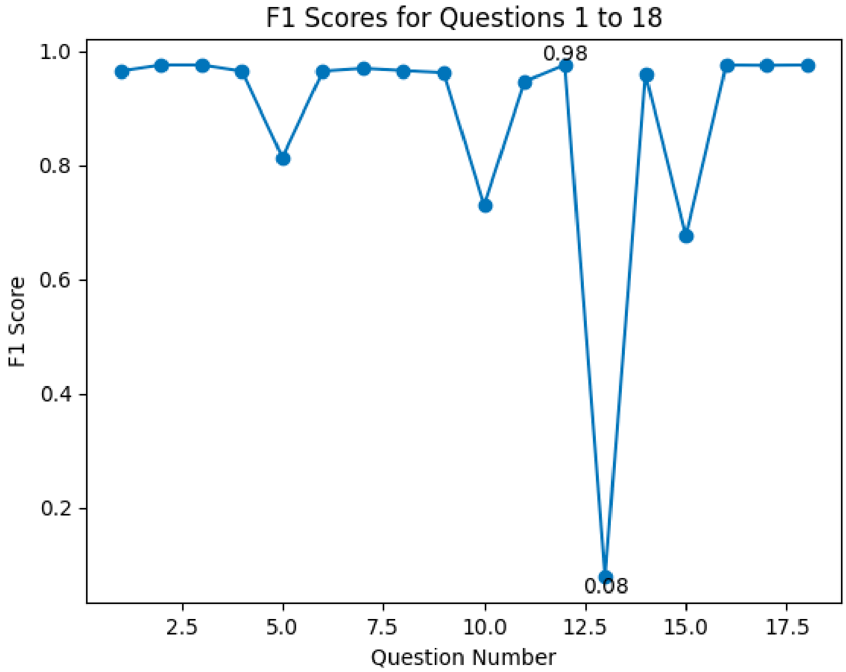
Main Analytics Results

1. F1 score/Accuracy Metrics (F1 score preferred due to imbalance)
2. Decision Trees

We performed question-specific predictions using different decision tree models. Each question is associated with a specific level group, and within that level group, different decision tree models are trained and used for prediction. The purpose of using different decision tree models for different questions is to capture the unique patterns and characteristics of each question. By tailoring the model to each question, it allows for more accurate and precise predictions based on the specific features and patterns associated with that question. By using question-specific decision tree models, we aim to improve the overall performance and prediction accuracy, as it takes into account the nuances and variations among different questions. This approach recognizes that different questions may have different levels of difficulty or distinct features that impact the prediction outcome.

Finding 1: We see that the F1 scores for the questions, have atleast a value of 70% which gives us enough confidence in the prediction algorithm and that the game will have accurate information to update and modify the users learning paths as we will have accurate estimations of their answering capability for each question

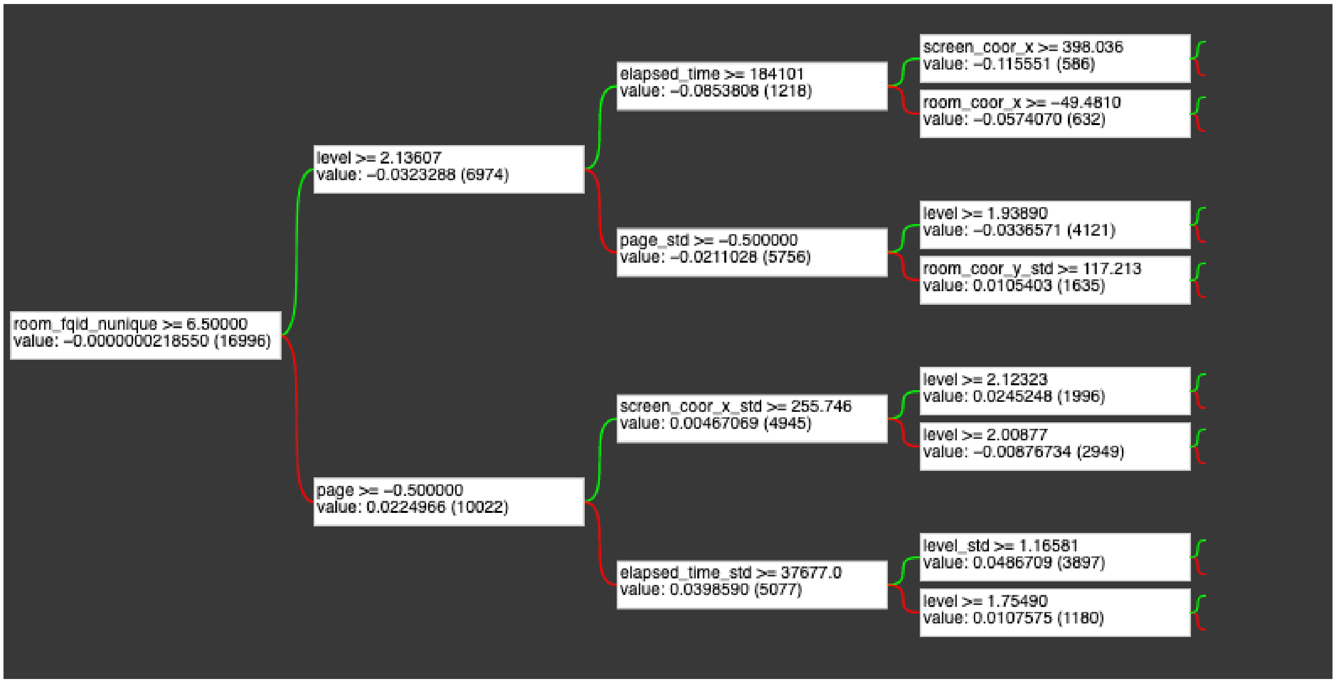
Finding 2: For some questions like Q13, F1 score is exceptionally lower than the other questions, so additional information may be required for such questions and specific steps should be taken in order to avoid incorrect decisions because of them



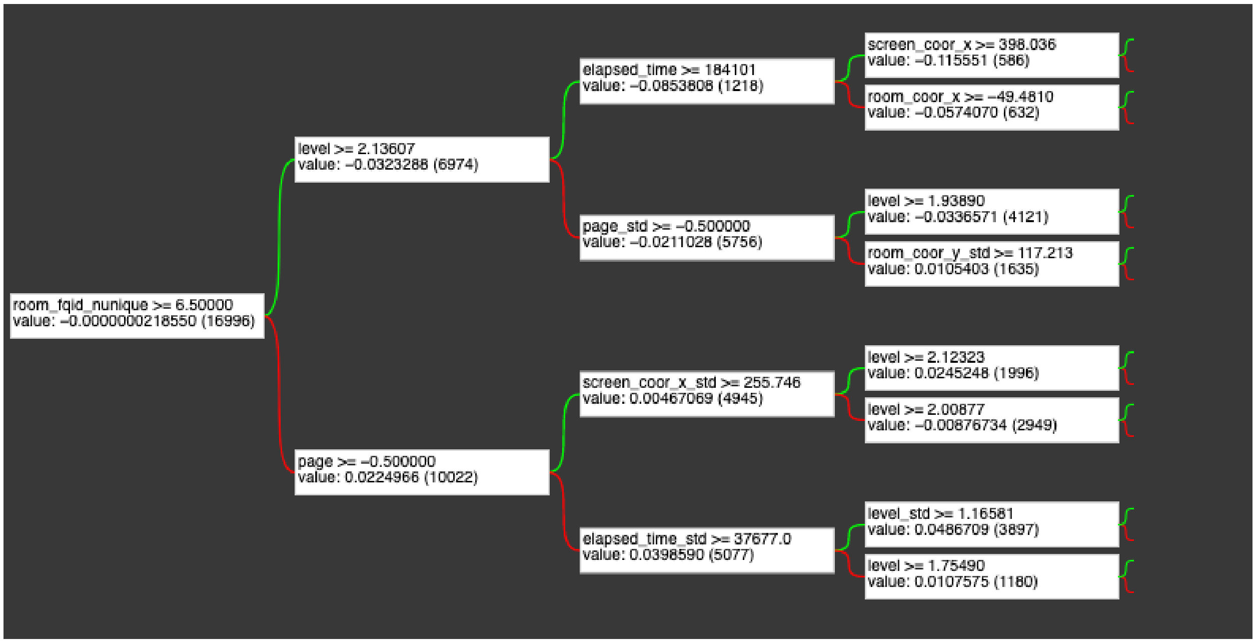
Decision Tree Findings: Through these decision trees, we are able to see the prediction algorithm followed and what factors were instrumental in decision making of successful answering probability.

We have three decision trees currently, where we have one question from each group and we see that for most of the questions the factors which were instrumental are:

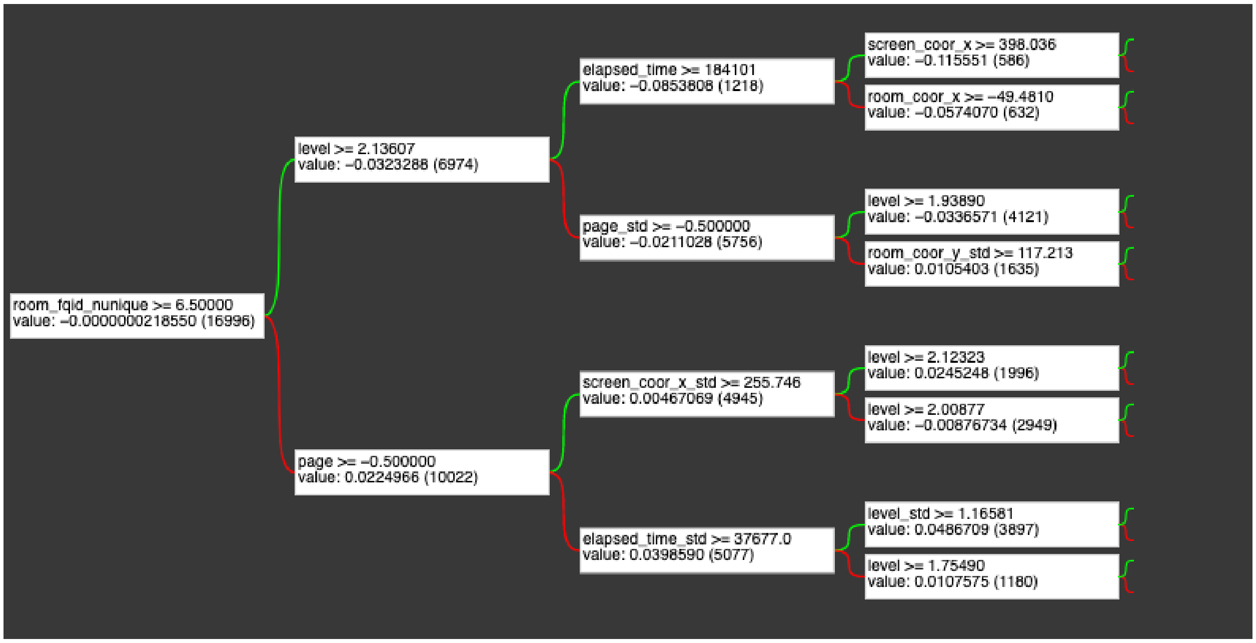
- Elapsed time on the ques - Higher time means more difficult ques for student, low probability for correct answer
- Screen co-ordinates where the user clicks - If the user clicks at incorrect place, low probability for correct answer
- Sum_Score: No of previous correct answers, if the user has more correct answers, high probability for correct answer



0- 4_1(for question 1 that belongs to group 0-4)



5-12_10(for question 10 that belongs to group 5-12)



13-22_18(for question 18 that belongs to group 13-22)

Executing the Analytics

We are recommending a whole funnel analysis for executing the analytics plan.

1. **Acquisition**: App downloads, user sign ups, Hover rates, Newsletter subscriptions.
2. **Activation**: Set up analytics tools and platforms such as Google analytics, Firebase and MixPanel to measure are users taking the first step by tracking user interactions, events and performance data.
3. **Retention**: The goal hers is to measure user behavior and measure retention. We have to track session durantions, frequency of gameplay and level upliftment rate pertaining to learning.
4. **Referral**: Word-of-mouth effects measured through CSLV (CLV+ CSV), signups through invites
5. **Revenue**: The monetization model we are using and the cost versus benefits analysis. Conversions if we are using a subscription model.

By executing our analytics plan as defined by the methodology and results could be executed in a funnel analysis to give us granular insights on which phase we need to focus and incorporate change.

Implementing the Analytics

1. **Defining User Demographics:** Defining the user segments based on educational levels. Data collection mechanism such as setting up data pipelines, data analytics models and visualization models to measure game level and user level data.
2. **Event Tracking:** Events in this case would be the most important user actions we want to track such as level completion, power up usage, quiz scores and time spent on specific activities.
3. **Integrating APIs:** Integrate multiple touchpoints such as for incorporating game data, academic performance measurements to measure against user game profiles, applying business analytics on this model to measure engagement, productivity and retention.
4. **Analyze Performance Improvement:** Comparing pre-games and post-games metrics to measure results on performance. Analyze the relationship between specific game activities or achievements and performance outcomes.
5. **Cohort and Revenue Analysis:** Evaluating demographics and psychographics of students over different time horizons to measure effect of intertemporal bias and heterogeneity on learning and game performance. Incorporate Revenue streams, from the model and also is monetization affecting performance or engagement.
6. **Refine and Reiterate:** Setting up a reporting system where weekly, quarter and annual reports would be generated to report analytics on user, games and an interaction of the two would be created. Review and use the data to refine and guide the feature enhancements, gameplay adjustments , feature updates and monetization updates. This would also have to be coupled with user feedback.

Scaling Up

1. **Marketing**: Targeted marketing campaigns aimed at students, referral programs and incentives

2. **User Experience**: Aim to better engagement by looping in feedback and incorporating their recommendations to enhance the user experience.

3. **Community Building:** Fostering support through social media forums, support channels and in-app collaboration features will encourage peer-to-peer interactions, collaboration and knowledge sharing.

4. **Scalable Infrastructure**: Have to focus on load balancing, server capacity and database management to handle increasing user demand. Maybe cloud object storage is something we could look into. Maintaining a proactive approach through stress tests to remove bottlenecks and improve efficiency would go a long way.

5. **Partnerships:** Focus on adding more tie-ups with educational platforms, bloggers and educational institutions to increase reach.

Sources and Citations

https://www.kumon.com/how-kumon-works

https://www.frontiersin.org/articles/10.3389/feduc.2023.1106679/full

https://www.axd.agency/post/analytics-and-adaptive-learning

https://www.youtube.com/watch?v=JkJltWH0Zrc

https://www.researchgate.net/figure/Adaptive-learning-mechanism_fig1_349351471

https://www.kaggle.com/competitions/predict-student-performance-from-game-play/overview/description