Predicting Mastery in Academic Assessments

Improve academic evaluations by predicting whether a student answers the next test question correctly.

Predicting Mastery in Academic Assessments

Students taking advantage of practice exercises through software applications seek an evaluation of their preparedness for formal academic assessments. The goal of this project is to help test developers understand user progress by predicting whether a student answers the next test question correctly.

Data are from students preparing for three test groups - ACT, GMAT and SAT. Question fields in the dataset indicate outcome, group, track (test subject), sub-track, ‘game’ structure, and times for when the question was started and answered or deactivated. Relationships that could predict observations under the field labeled **`correct`** are explored and modeled throughout the project.

Predicting Mastery in Academic Assessments

- Code and reports for this project live here:

  https://github.com/humburgc/predicting_assessment_mastery

- Detailed final report of steps summarized in this presentation:

  https://github.com/humburgc/predicting_assessment_mastery/blob/master/reports/final_report.pdf

- The dataset is available here:

  https://www.kaggle.com/c/WhatDoYouKnow/data

THE DATA

The dataset spans three years of students answering questions to prepare for the different tests. In order to predict a user's next test question, the data were split (by the test developers) in the following manner:

- If a user answered at least 6 questions, one of their answers was selected randomly from the 6th question to the last. This plus all later question attempts by this user were removed.
- Any earlier answers were included in the **`training.csv`** set, which is the dataset used for this project.
- All answers from users answering fewer than 6 questions are also included in the **`training.csv`** set.

Initial Inspection of the Dataset:

- 4,851,475 question attempts (rows)
- 17 variables (columns)
- 179,106 users
- 6,045 unique questions

DATA CLEANING

Organize columns, evaluate and resolve missing data, evaluate and resolve duplicate data, compute and explore time data.

Key Actions:

- Redundant columns were removed, and a new column was created so that one variable captures the end of a round (`round_ended_at`) rather than two (`answered_at` and `deactivated_at`).
- All observations missing all time-date values were eliminated. Another 21 rows were dropped for null `answered_at` and `deactivated_at` values since round duration cannot be calculated in those rows.
- The `date_of_test` column was dropped because 87% of users did not provide an expected test date.
- Missing data was expected for observations with the value of 6 in the `game_type` column, as this denotes a data collection failure, so they were dropped.
- Any observations that started and ended at the same time were dropped. As were duplicates of the combination of `user_id`, `round_started_at` and `round_ended_at`.

DATA WRANGLING

Before saving the clean dataset, a few variables were created to isolate outliers and help with exploratory data analysis:

- **`round_duration`** was created by subtracting **`round_started_at`** from **`round_ended_at`** values.
- Variables indicating **`round_duration`** as either fast (< 5 seconds) or slow (> 10 minutes).
- Variables indicating **`num_questions`** and whether a user attempted few (<6) or many (>4000).
- One variable indicating the number of **`prior_attempts`** of a particular question by the user, and one variable indicating when these become **`excessive_attempts`** (> 100).

	Total Observations	Total Variables	Number of Users	Number of Unique Questions
Pre Cleaning	4,851,475	17	179,106	6,045
Post Cleaning	**4,846,730**	**22**	**179,033**	**6,045**
Eliminated	4,745	(9 added) 4	73	0

EXPLORATORY DATA ANALYSIS

Create data visualizations to explain variables. Detect and possibly remove or mark outliers. Explore variable dependence and correlations. Consider a hypothesis to explore. Leverage statistical inference to test the hypothesis. More generally, begin to develop a preliminary likeness of the solution.

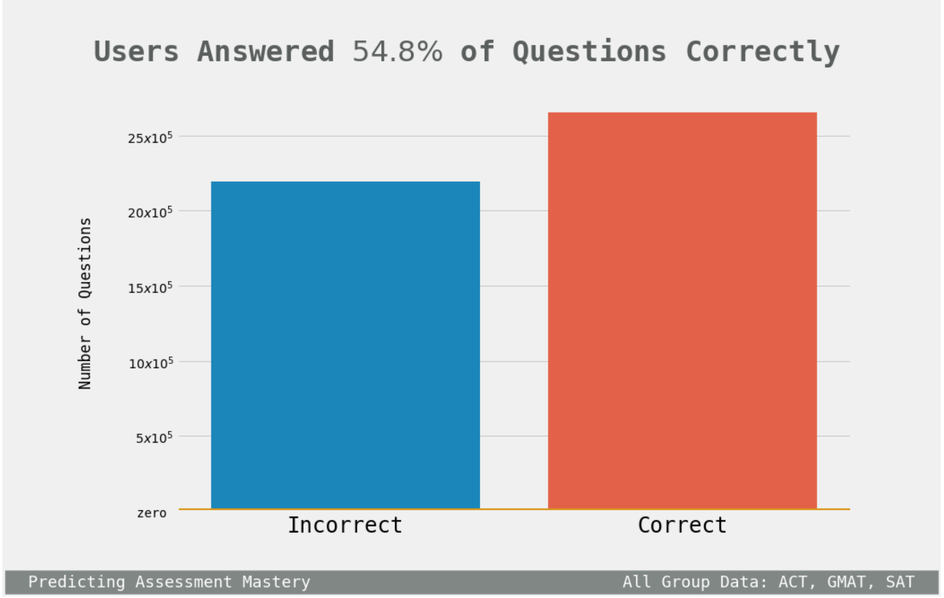
Key Features Explored:

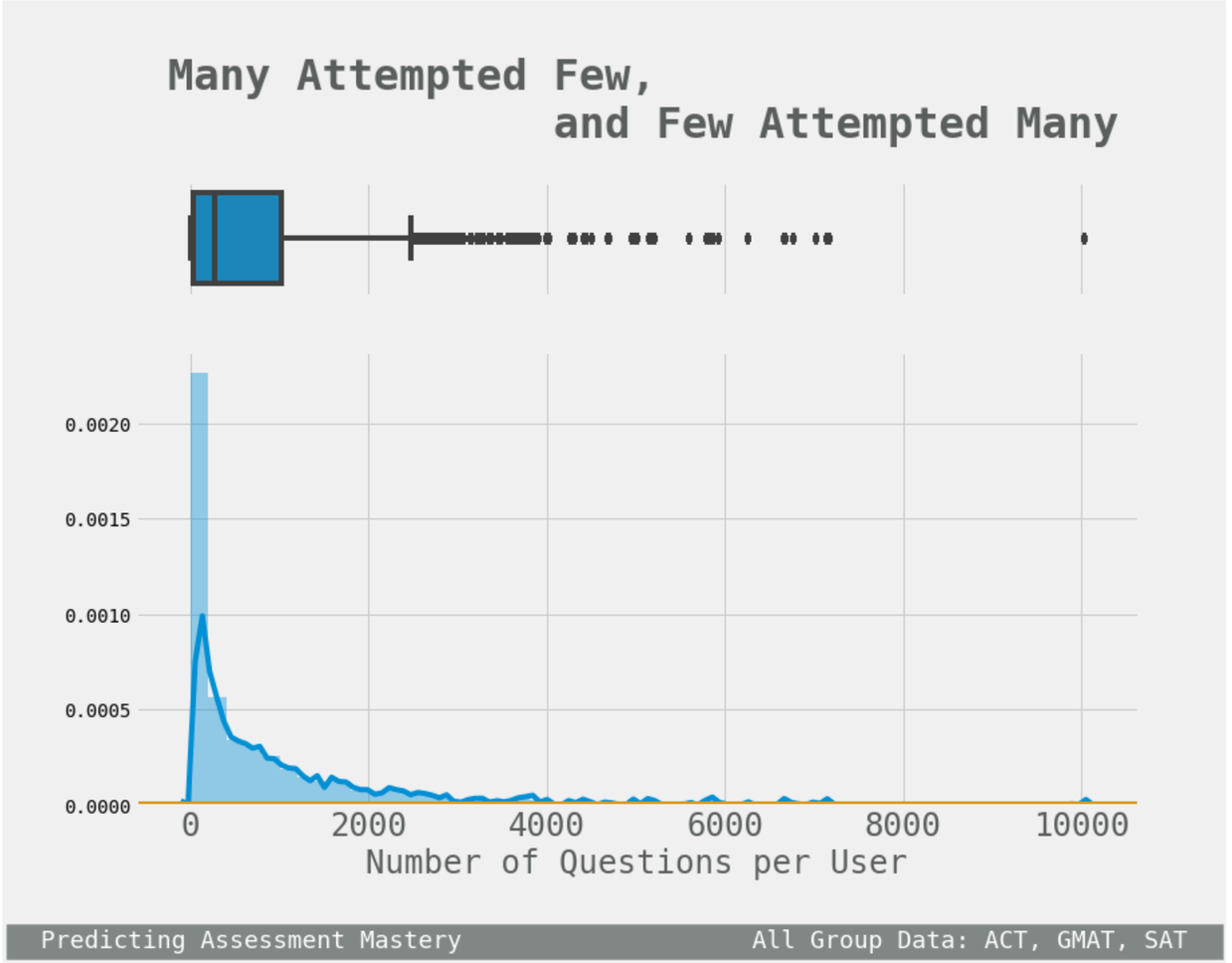
- Dataset Statistics
- User Data and Statistics
- Question Data and Statistics
- Group Specific Data
- Statistical Significance of Group Data

EXPLORATORY DATA ANALYSIS

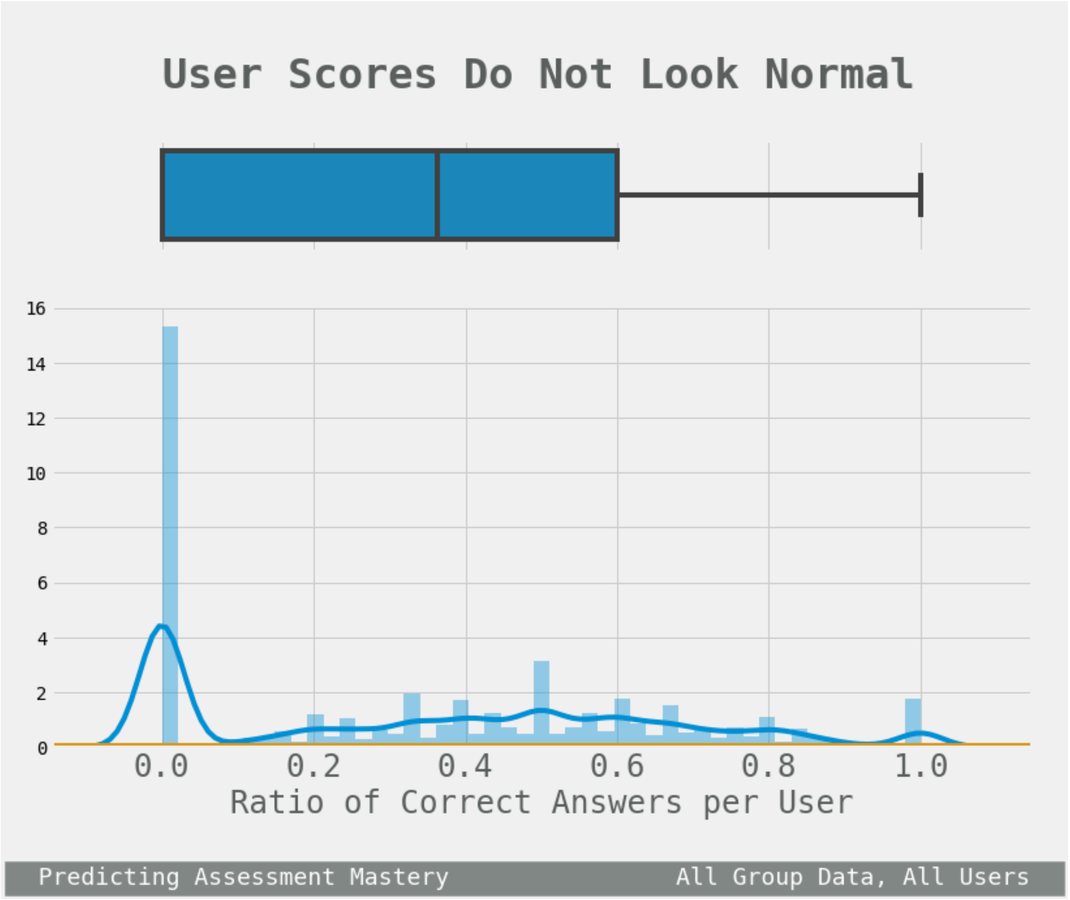
While more than half of all questions in the dataset were answered correctly, it's difficult to evaluate this alone from a performance point of view since so many users attempted so few questions.

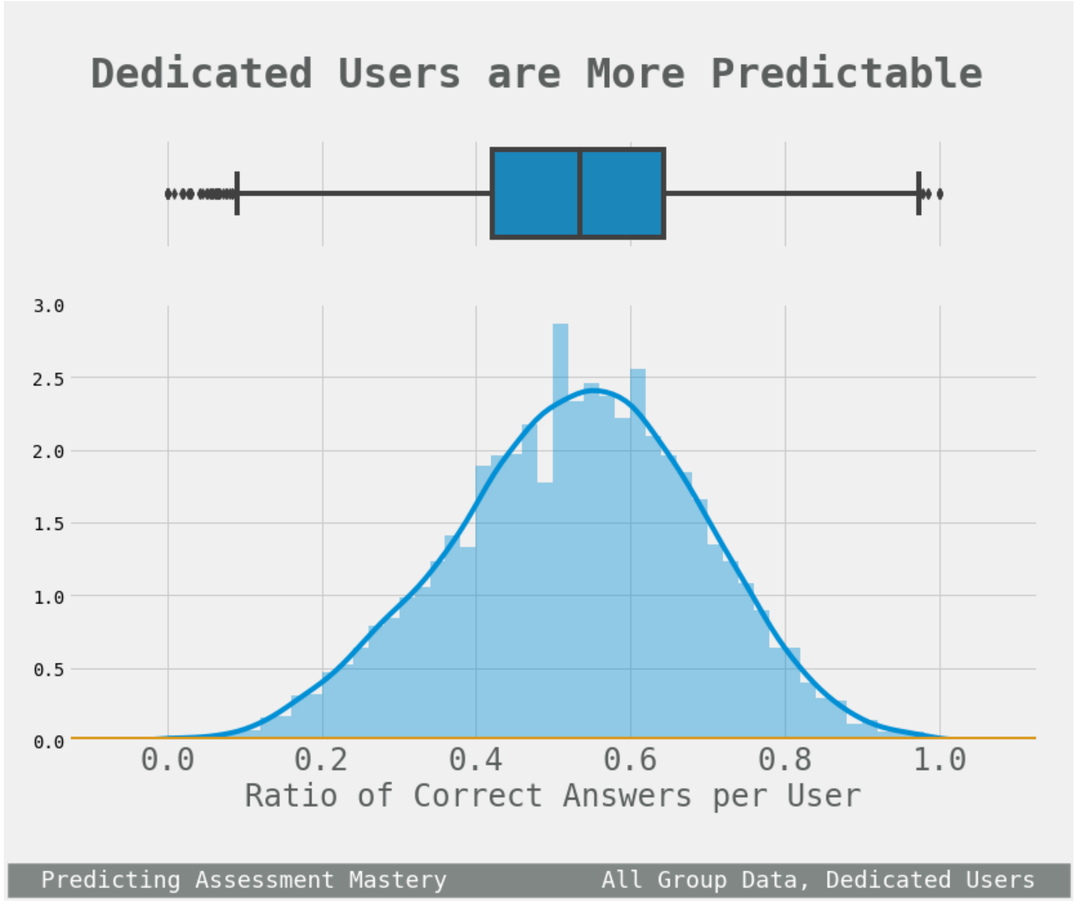
Users attempting more than 4,000 questions are scarce, but at least one attempted at least 10,000 questions. For all records in the clean dataset the round duration average is 1 minute 8 seconds.





EXPLORATORY DATA ANALYSIS

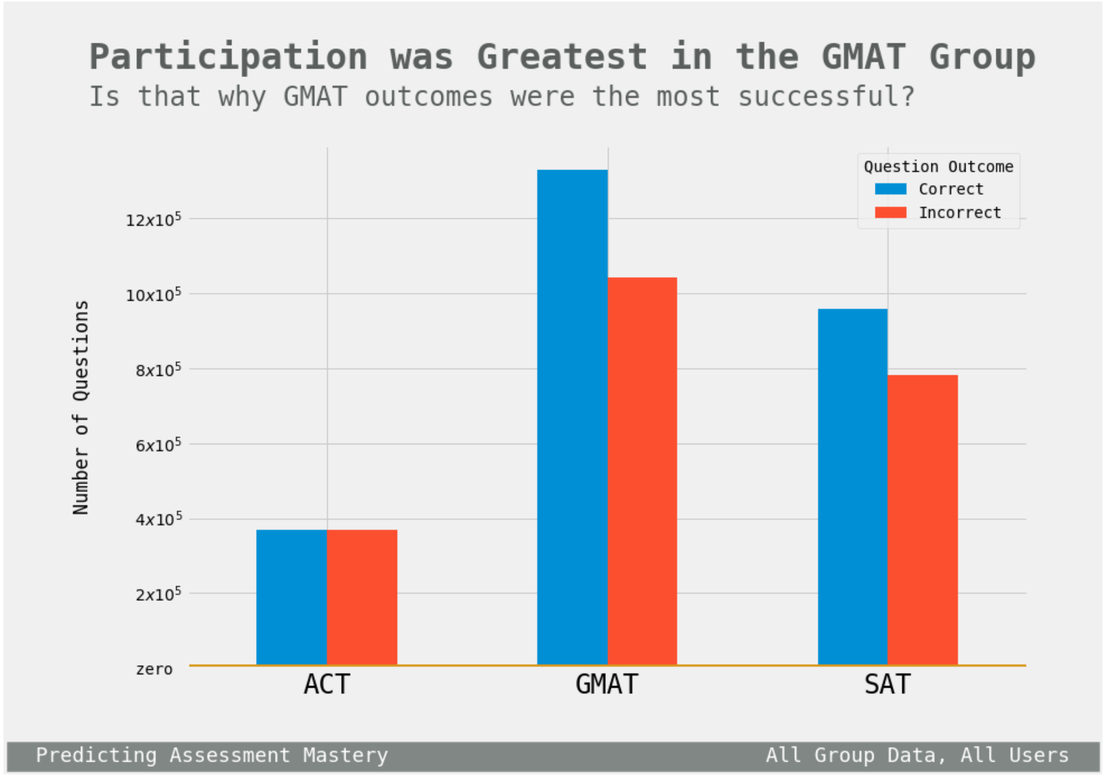




Dedicated User: One who has attempted at least 30 questions.

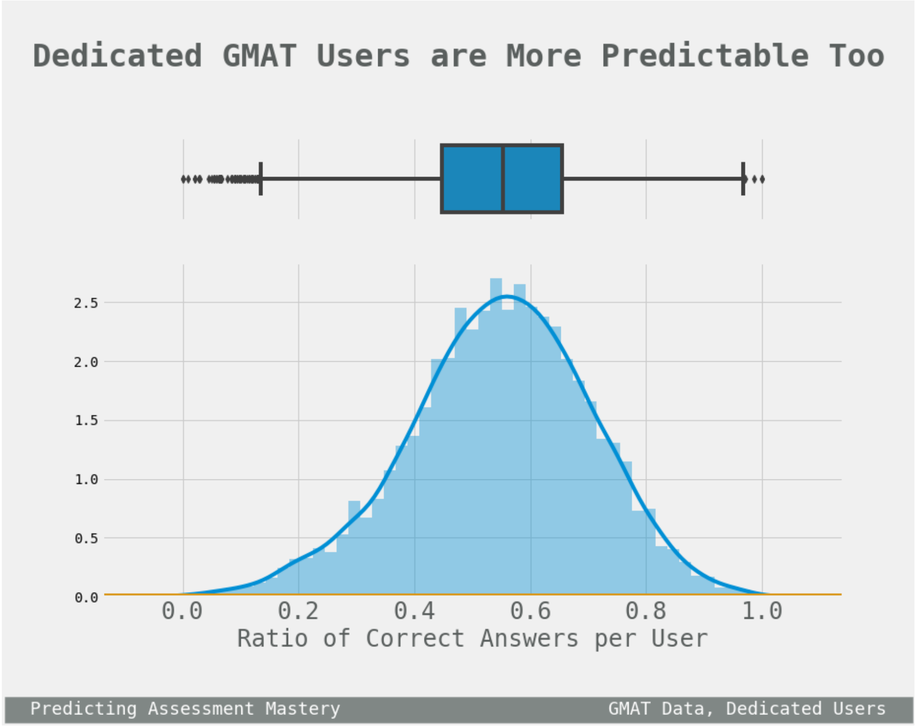
EXPLORATORY DATA ANALYSIS

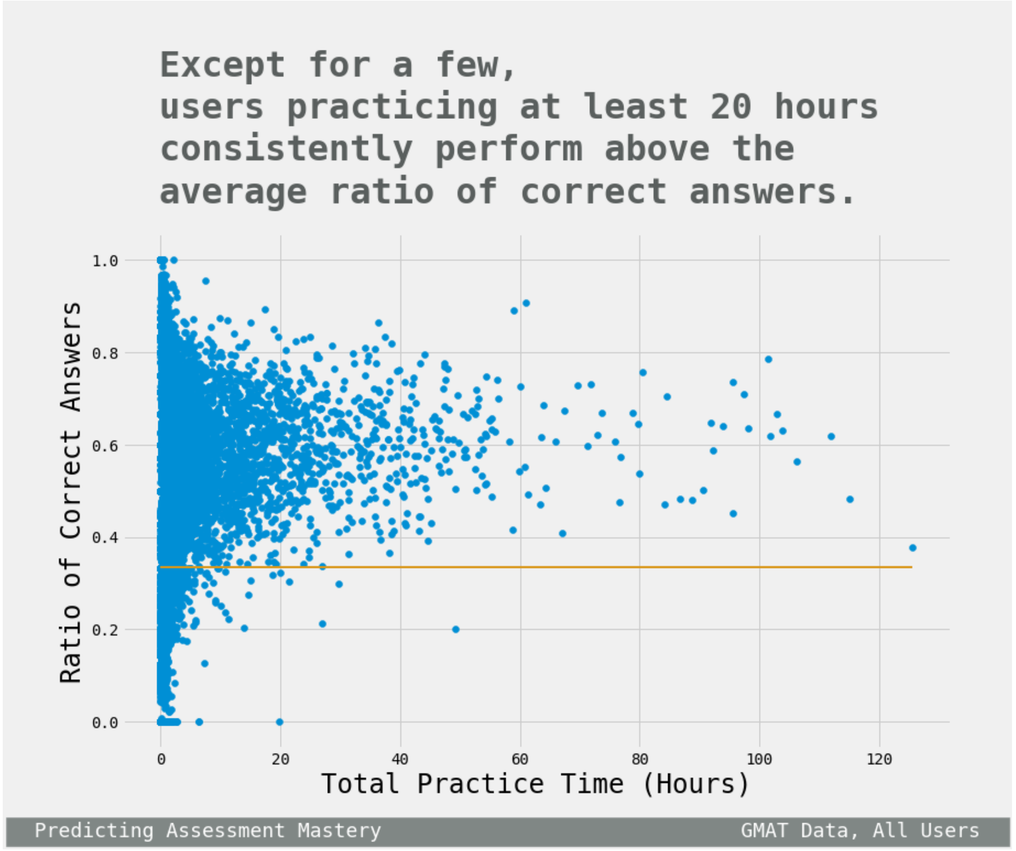
A cross-tabulation of **`group_name`**, **`track_name`**, and **`subtrack_name`** determined that there is no overlap of tracks or subtracks across groups. The groups seem to be independent, so the dataframe was split to explore each group individually. Helped by the graph below, GMAT was ultimately deemed best for modeling.



EXPLORATORY DATA ANALYSIS

However defined, either by number of questions attempted or by total practice time, user dedication seems to matter.





HYPOTHESIS TESTING

Null Hypothesis: All test groups are from the same distribution.

- For this, a one-way ANOVA test was performed on the average number of correct answers for each group.
- `alpha=0.05 > p-value=0.00`, and `F-value=4070.79 > F-critical=2.9957`
- Reject the null hypothesis - **At least one of the three test groups is from a different distribution.**

Null Hypothesis: The GMAT and SAT groups are from the same distribution.

- For this, a T-test for two independent samples was performed on the average number of correct answers for each group.
- `alpha=0.05 > p-value=0.00`, and `t-statistic=1.61`
- Reject the null hypothesis - **The GMAT and SAT groups are from different distributions.**

Null Hypothesis: GMAT users and dedicated GMAT users are from the same distribution.

- For this, a T-test for two related samples was performed on the average number of correct answers for each group.
- `alpha=0.05 > p-value=0.00`, and `t-statistic=-40.1945`
- Reject the null hypothesis - **GMAT users and dedicated GMAT users are from different distributions.**

DATA MODELING

Machine Learning Analysis: Build, fit, and validate a method to model the GMAT data. Make predictions for each model and evaluate the performance. Return the strongest predictors for the best model.

- **Logistic Regression** - A classification method for finding the probability of a given outcome, but assumes that observations are independent. Grouped by user, observations in this project are not independent. But this provides an easy and fast baseline for repeat simulations.
- **Random Forest** - This creates an ensemble of decision trees which are averaged to improve prediction accuracy, and it scores how important each feature was for classification.
- **XGBoost** - This implements gradient boosted decision trees, and scales very well to many samples. It works by iteratively learning a set of weak models on subsets of the data, weighing each weak prediction according to each weak model's performance, and combining the weighted predictions to obtain a single weighted prediction.
- **Linear Mixed Effects** - These models are useful for analyzing data that are dependent, hierarchical, or longitudinal. The tracks and subtracks suggest hierarchy and there is certainly a longitudinal nature of observations by user.
- **Mixed Effects Random Forest** - A method developed by a group of statisticians to account for mixed effects while using Random Forest as the estimator.

THE BEST MODEL

To compare models, two performance measures were computed for each: 1.) Accuracy of predictions on test data. Higher values are better. 2.) Capped Binomial Deviance, accounting for the probability of responses. Lower values are better.

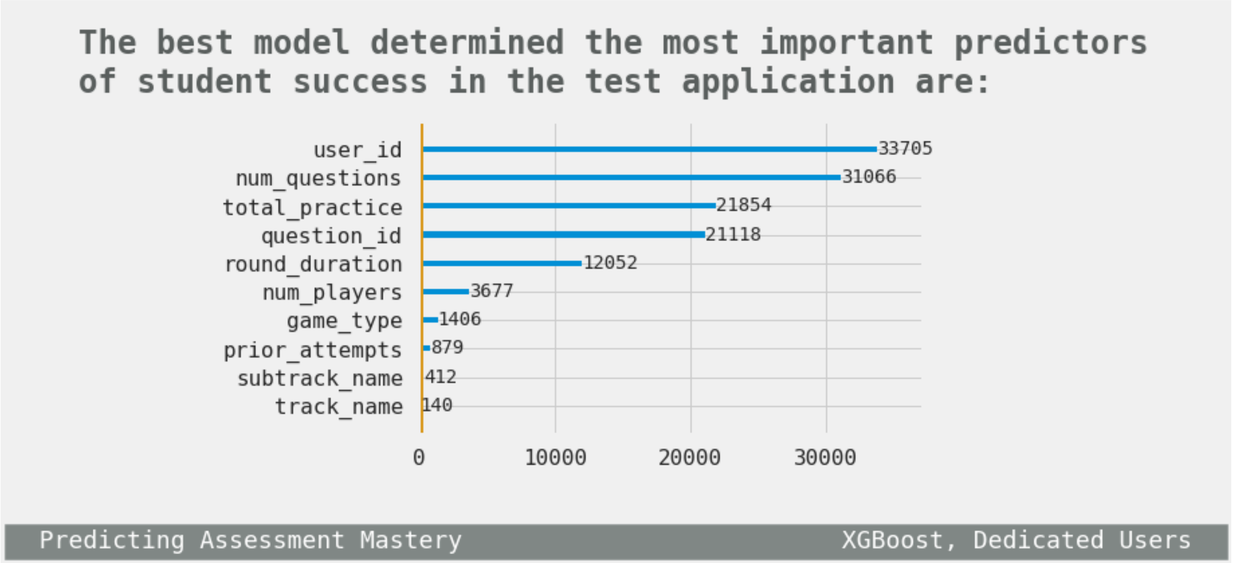
In order to make predictions for the outcome column labeled **`correct`**, all models looked for patterns in these variables: **`user_id, question_id, track_name, subtrack_name, game_type, num_players, round_duration, prior_attempts, num_questions, total_practice`**. Removing the least important amongst these did not improve performance.

Model Name	Accuracy	CBD
xgboost	0.6983	0.2454
random_forest	0.6687	0.2554
mixed_effects_randforest	0.6683	0.2583
logistic_regression	0.6164	0.2889
linear_mixed_effects	0.5692	0.3017

Since XGBoost was the best model by both measures, it was then used to fit data filtered for dedicated users. Both performance measures improved. Accuracy: **0.7109** CBD: **0.2411**

EVERYTHING POINTS TO PRACTICE, PRACTICE, PRACTICE

Exploratory data analysis and machine learning back up the intuition that the more a user practices the more successful they are in answering questions. More importantly, both show that dedicated users provide enough data for a more reliable analysis of their performance. Features related to subject matter and ‘game’ structure were less helpful as predictors.



Recommendation to Test Developers: To better understand user progress and improve test applications, developers are encouraged to incentivize users to attempt at least 30 questions.

PRACTICE!