# Final Report (Draft)

## Predicting Mastery in Academic Assessments

Students taking advantage of practice exercises through various 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 a test question correctly. The project examines data from students preparing for three tests: the GMAT, SAT and ACT. Question fields in the dataset indicate outcome, group (test for which the student is preparing), track (test subject), sub-track, tags, question format, and times for when the question was started and answered or deactivated (timed-out). Relationships that could predict observations under the field labeled 'correct' are explored and modeled.

## APPROACH

1. Initial Data Exploration - Import and inspect raw data. Isolate relevant variables, fill or calculate new variables, and organize the dataframe. Resolve missing, invalid, corrupted or duplicate values.

2. Exploratory Data Analysis - Create data visualizations to explain variables and analyze outliers. Possibly remove or mark outliers. Explore variable dependence and correlations.

3. Machine Learning - Build, fit, and validate a method to model the data. Evaluate the performance of each model tested. Methods to try: Logistic Regression, Random Forest, Linear Mixed Effects, Mixed Effects Random Forest, XGBoost.

4. Final Reports & Presentations

## INITIAL DATA EXPLORATION

**Organize Columns:** Redundant columns such as **`outcome, answer_id, question_set_id`** were removed. A new column was created, **`round_ended_at`**, filled with `answered_at` values if present. If no value was found in any `answered_at` field, then the value from `deactivated_at` was used to populate `round_ended_at`.

**Evaluate & Resolve Missing Data:** NULL values in the dataframe indicate that the value is missing, invalid or corrupted. These were expected for **`answered_at`** since this denotes that the question timed-out. Missing values for **`date_of_test`** were also no surprise since it was known that users may or may not enter an expected test date. NULL values for **`round_started_at`** and **`deactivated_at`** were less understood and, therefore, examined more closely. All observations with missing data for all timedate variables were eliminated. Another 21 rows were dropped for missing **`answered_at`** and **`deactivated_at`** values since round duration cannot be calculated in those rows. About 87% of users did not provide an expected test date, so the **`date_of_test`** column was dropped. NULL values also existed in the **`game_type`** column, as a value of 6

represents a data collection failure. Most, if not all, of the observations involve rounds with end times earlier than start times, so they were dropped.

**Evaluate & Resolve Duplicate Data:** Given the reasonable expectation that it is impossible to start and end a round at the same time, 4264 observations meeting such conditions were dropped. After weird time issues were resolved, duplicate observations for the combination of **`user_id`**, **`round_started_at`** and **`round_ended_at`** were dropped.

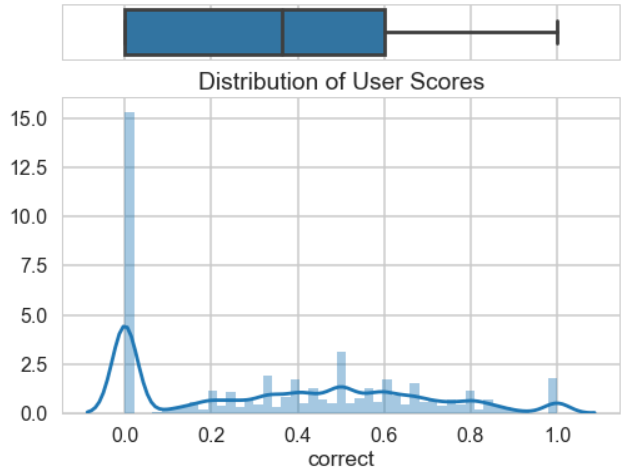
**Compute and Explore Time Data:** The **`round_duration`** column was created and filled by subtracting `round_started_at` values from `round_ended_at` values. A view of time data statistics revealed that some rounds were days long and many were over within only a few seconds. Rather than delete the 167,564 observations outside a threshold of 5 seconds to 10 minutes, new columns were created to categorize fast and slow extremes.

| | Total Observations | Total Variables | Number of Users | Number of Uniques 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

**Dataset Statistics:** For all records in the clean dataset the mean percentage of correct answers is 54.8% and the mean round duration is 1 minute, 8 seconds.

**User Data:** A small percent of users attempted a large number of questions and a large percent attempted a small number of questions. This certainly explains the distribution of the ratio of correct answers by user:




**Question Data:** The distribution of unique questions showed that a large number are underrepresented in the dataset. This is explained by the distribution of test groups (next section).

One user attempted one particular question 1268 times. This user was eventually dropped. There were at least 12 other questions associated with excessive attempts but the closest count amongst them was for 274 prior attempts.

**Group Name Data:** There is no overlap in the questions, tracks or subtracks across groups. Given the independence of test groups, the dataframe was divided by group for further exploration. Only a small number of users participated in more than one group.

**ACT Data:** About 50% of questions in the ACT group were answered correctly, and the average round duration was about 53 seconds.

- Average score for ACT users: 0.347707
- Number of ACT users with no correct answers: 12198
- Average score for dedicated (attempting at least 30 questions) ACT users: 0.507080
- Number of dedicated ACT users with no correct answers: 8

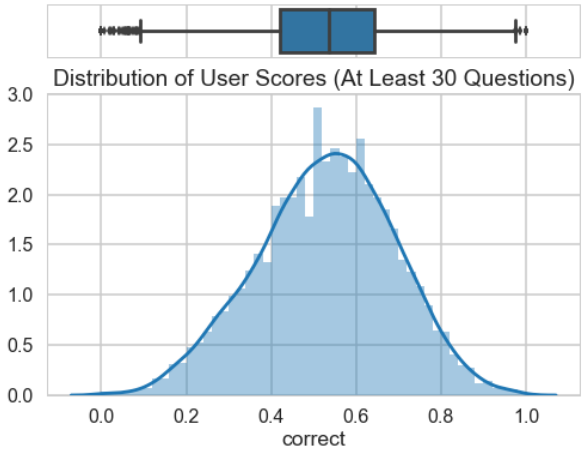
**GMAT Data:** About 56% of questions in the GMAT group were answered correctly, and the average round duration was about 1 minute 18 seconds.

- Average score for GMAT users: 0.333503
- Number of GMAT users with no correct answers: 28007
- Average score for dedicated GMAT users: 0.547034
- Number of dedicated GMAT users with no correct answers: 2

**SAT Data:** About 55% of questions in the SAT group were answered correctly, and the average round duration was about 1 minute.

- Average score for SAT users: 0.381569
- Number of SAT users with no correct answers: 19281
- Average score for dedicated SAT users: 0.524551
- Number of dedicated SAT users with no correct answers: 3

**Summary:** A small number of users practiced in more than one test group; however, no overlap in questions and subjects indicates the test groups are independent. For each test group, outcomes are better across all subjects for dedicated users (attempting at least 30 questions). In every scenario, user score distributions were more interpretable (normal) amongst the dedicated users, as evident in the graph below which includes data from all test groups:



## MACHINE LEARNING ANALYSIS

Next...

## CONCLUSIONS

User dedication seems to matter...

## DELIVERABLES

- All Code & Notebooks
- Project Reports
- Final Report
- Slide presentation