

Project Proposal

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.

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.
- Earlier answers were included in the **training.csv** set, which is used for this project.
- All answers from users answering fewer than 6 questions are also included in the **training.csv** set.

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, duplicate values.
2. **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.
3. **Machine Learning** - Build, fit, and validate a method to model the data. Evaluate the performance of each model tested, including Logistic Regression, Random Forest, Linear Mixed Effects, Mixed Effects Random Forest, XGBoost.

DELIVERABLES

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