Capstone Project Proposal

ACADEMIC MASTERY STUDY

Improve academic evaluations by predicting whether a student will answer the next test question correctly. This 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, topics, question format, and times for when the question was started and answered or deactivated (timed-out). The study investigates relationships that could predict observations under the field labeled 'correct'.

The Problem

Students taking advantage of practice exercises through various software applications seek an evaluation of their preparedness for formal academic assessments. Students tend to quit an application if it is too boring or give up if the questions are too challenging. They would like an an accurate evaluation of any gaps in knowledge and skills without sitting through an assessment that repeatedly tests established proficiencies and difficulties.

The Client

The Test Prep Company wants to decide how they can better organize their online practice tests to precisely evaluate competencies so that fewer demonstrations of proficiency can accurately determine mastery.

The Data

The data used for this study comes from a a Kaggle competition for Grockit. This is the information provided regarding the datasets:

> " The data… is a sample of Grockit students… The test/training split is derived by finding users who answered at least 6 questions, taking one of their answers (uniformly random, from their 6th question to their last), and inserting it into the test set. Any later answers by this user are removed, and any earlier answers are included in the training set. All answers from users not in the test set are also used for the training set (as they may be useful in estimating question parameters or baseline ability distributions).

The test data distribution is thus different from training data in ways that may be significant. First, it does not include 'timeout' or 'skipped' outcomes: all test results are from the student actually answering the question. Second, it is biased towards users with more questions in the training set and biased towards their later answers. Third, it is one entry per user, so the distribution of various aspects of the data (such as correct/incorrect) is over users, not over all answered questions.

We have attempted to provide a reasonable validation split on the training data by taking the previous correct/incorrect answer for each of the students in the test set, for those users who had at least one previous answer. The results are in the additional files valid_training.csv and valid_test.csv. ”

The Approach

1. **Data Collection & Initial Exploration**
   Download the grockit_all_data.zip data file from the Grockit competition on Kaggle, and begin creating data visualizations using Python technologies.

2. **Data Wrangling and Cleaning**
   Employ Python to isolate important information; rectify missing, invalid or corrupted values; analyze outliers and decide how to handle them.

3. **Exploratory Data Analysis & Inferential Statistics**
   Continue with data visualization, using plots to gain insights, look for correlations, consider a hypothesis to explore, then leverage statistical inference to test the hypothesis.

4. **Machine Learning**
   Apply (to-be-determined) machine learning algorithms, evaluating the performance of each, and decide on the right one to use for this study.

5. **Final Reports & Presentations**

The Deliverables

- All code/notebooks
- Final Report
- Slide presentation
- Video presentation