Springboard Data Science Career Track
Capstone Project 1 Proposal:
Predicting Student Test Outcome on an Online Tutorial
Cesur Dagli
December 2018

What is the problem you want to solve?

Background: Indiana University offers an online tutorial that aims to teach how to recognize and avoid plagiarism. This online tutorial is free and has been used by hundreds of thousands of users worldwide. The majority of the users of this tutorial are undergraduate and graduate students, who are required by their instructors to complete the tutorial and take and also pass a test to earn a certification. In order to earn a certification, students take a 10-question test, which is randomly selected from a very large question pool. Only those who correctly answer 9 out of 10 questions pass the exam and earn a certification. The problem I would like to solve is to predict whether not a student will pass a certification test based on the data available.

From a technical point of view, I will be solving a supervised learning classification problem since I will be using a dataset that consists of input and output variables to predict the class of output (pass or fail) based on the data. Various algorithms will be used to address this binary classification problem.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

Indiana University is the institution that designs and runs this online tutorial. When students fail certification tests at their first attempt, they become frustrated since they have to pass a certification test, which is required. The university receives a significant number of emails with complaints about the tutorial and the school from students who fail at their first attempt. The designers of this tutorial would like to increase the effectiveness and efficiency of their tutorial. In other words, they would like students to pass a certification test in their first attempt and therefore also to improve the pass rate. Using machine learning models one can identify the variables that influence a passing classification versus a failing, and this information can be used by the test designers to improve the test.

What data are you using? How will you acquire the data?

Anyone can access the tutorial. However, those who would like to earn a certification, have to complete the registration process, which collects *demographic data*. Also, students are asked to complete *an optional 30-item survey*, which seeks a student's perception of the quality and satisfaction of activities on the tutorial. And finally, students take a *10-item certification test,* whose questions are randomly selected from a test item pool. Since students can take the certification test multiple times, the dataset may include multiple data points per person. Also, the tutorial provides two types of certification tests: one for advanced high school and undergraduate students, and the other one for graduate students, which is harder. Overall, the dataset consists of

- **Demographic data (required):** Test ID, Age, Education Level, Current Level (of understanding of what plagiarism is), Reason for taking test, Duration,
- **30-item survey (optional):** Satisfaction, quality, data related to activities, etc.
- **Certification test (required for those who would like to earn a certification):** Test Types, Test Outcome (pass or fail)

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

1. Data acquisition and wrangling
2. Storytelling
3. Application of inferential statistics
4. Milestone report
5. Baseline modeling and analysis of preliminary results
6. Application of additional models and analysis of final results

Preparation and submission of deliverables: final report, Jupyter notebooks, presentation slide deck

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

My deliverables will be, as required, a final report, all notebooks I will develop, and a presentation slide deck.