Capstone Two - Project Proposal

Riiid! Answer Correctness Prediction

Problem Statement

Create a model of all students' knowledge over time, and predict any student's response on future test questions.

Context

High-quality education at a low-cost is a nearly universal good at the levels of national public policy, local schools districts, single households, and even the individual. An computer-based educator that could teach as good as, or better than, a human educator--and at lower cost--would have broad applicability to improve the living conditions of humans all around the world. Riiid Labs, an e-learning startup, has shared their data from more than 1 million students who use their e-learning platform so that researchers can make new models to predict and guide student success.

Criteria for Success

When testing the model on unseen data, the area under the ROC curve between the predicted probability and the observed target must be greater than 0.5.

Scope of Solution Space

For this project, we will only focus on building a model that can accurately predict the student's response to a test question.

A high-performing model will have many broader applications (all of which are outside the scope of this project) such as: individualized curriculum design and planning, estimating the time to reach any milestone in the curriculum, and identify individual cognitive developmental differences or learning styles.

Constraints within Solution Space

- Learning history from the Riiid platform only - We do not have access to student learning data from outside of the platform, or other student non-learning activities data to may be relevant to
- Student metadata: only limited types were provided by Riiid - We dont not have access to student demographics data--such as age, gender, location, local time, or user device

(smartphone or desktop, etc)--which have proven useful in other personalized applications, like advertising.
- Question content metadata is obfuscated as numerical tags - We do not know the actual content of the questions, response choices, or lecture videos. We can not apply existing education domain expertise.

Stakeholders to provide key insight

As this is a public competition, there is no expectation of sourcing additional data from Riiid. The predictions on the test set are to be submitted directly via Kaggle Kernels for evaluation.

Key data sources

The sole source of data is the official repository associated with the Kaggle competition, Riiid! Answer Correctness Prediction | Kaggle | Data.

The most useful files are summarized here.

questions.csv:	file containing the test questions metadata,
lectures.csv:	file containing the lecture video metadata,
train.csv:	file containing over 100 million rows of students interactions with test questions and lecture videos