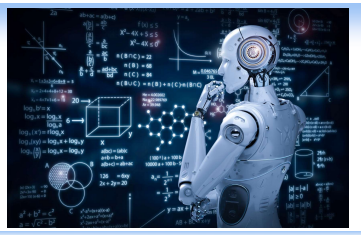
Riiid! Answer Correctness Prediction

Xiao Han

Department of Computer Science, Utah State University



Problem

This work mainly focuses on how to accurately predict the student's performance on the future interactions based on previous knowledge tracing.

Motivation and Background

The pandemic has changed how we work, learn and interact as social distancing guidelines have led to a more virtual existence, both personally and professionally.

With schools closed, lessons are being held remotely. A lot of sports, school activities, and events have been cancelled. Friendships and relationships have been transported to live chats and video calls.

Riiid Labs, an AI solutions provider launched an AI tutor based on deep-learning algorithms in 2017. The company released EdNet, the world's largest open database for AI education, and raise this challenge to help the online students get a better personalized learning experience in a post COVID-19 world.

Data

There are three files, train.csv, questions.csv, and lectures.csv, in the training dataset. The total size for the training dataset is 5.45 gigabytes.

1. train.csv

The original dataset has 10 features with 101 million records in the train.csv file.

In this project, I will use 8 columns listed below:

row_id, timestamp, user_id, content_type_id, task_container_id, user_answer, answered_correctly, prior_question_elapsed_time

2. questions.csv

The questions dataset has 5 features with 13,523 rows and the columns are listed below:

question_id, bundle_id, correct_answer, part, tags

3. lectures.csv

The lectures dataset has 4 features with 418 lecture ids in it and the columns are listed below:

lecture_id, part, tag, type_of

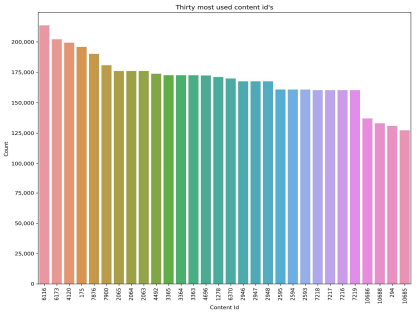
Limitation

The notebook needs to meet the following requirements to submit the result:

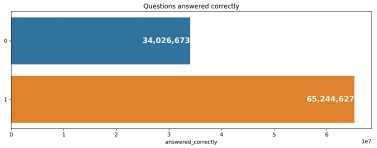
- CPU Notebook <= 9 hours run-time
- GPU Notebook <= 9 hours run-time

EDA

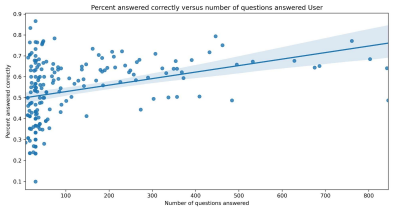
- There are 393,656 unique users in the training dataset.
- Content_id is a code for the user interaction. Basically, these are the questions if content_type is question (question_id: foreign key for the train/test content_id column, when the content type is question).



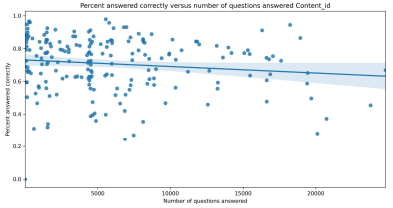
- Answered correctly is the target, and this project has to predict the probability for an answer to be correct. Without looking at the lecture interactions, we see about 1/3 of the questions was answered incorrectly.



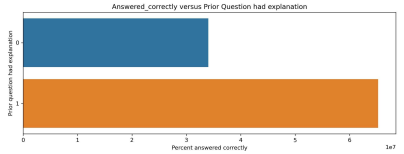
- After filtering out the outliers, the trends of the number of answers per user against the percentage of questions answered correctly is upward but there is a lot of variation among users that have answered few questions.



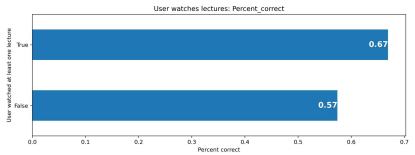
- The trends of the times of the questions have been answered against the percentage of questions answered correctly is downward.



- Does it help if the prior question had explanation? The percent answered correctly is about 17% higher when there was an explanation. N/A representative for the missing value.



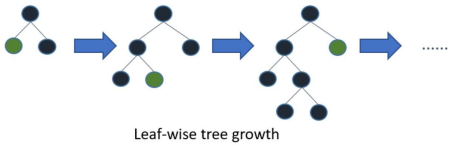
- Whether watching lectures will help the users to improve the accuracy of answering questions? Of course!



Approach

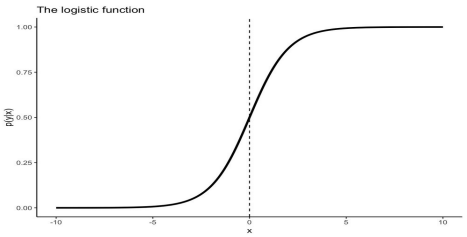
In this project, I'm trying four different models to solve the problem.

1. LGBM – Light gradient boosting framework that uses tree-based learning algorithms.

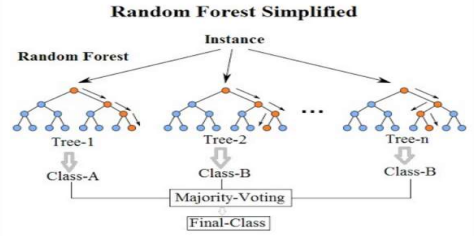


Leaf-wise tree growth

2. Logistic Regression

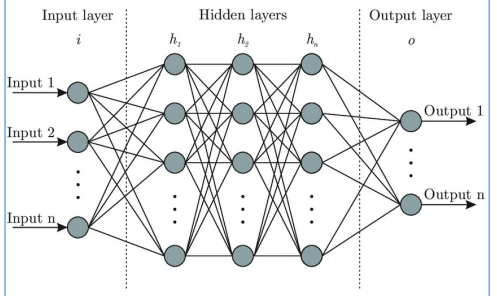


3. Random Forest



4. Neural Network Classifier

Dropout rate 0.1. two fully connected layers (1000, 50)



Results

Random Guess	LGBM	Logistic Regression	Random Forest	Neural Network Classifier
0.5	0.756	0.58	0.669	0.706

Analysis

The logistic regression could not achieve a good result with the basic data preprocessing. Random forest get the highest training accuracy and lower validation accuracy which means it will overfitting the training dataset. With the limitation of submission, I cannot get the best result with the random forest. LGBM is a tree-based learning algorithm which could achieve the baseline result for this competition. The neural network classifier could achieve with more preprocessing.

Conclusions

In this project, I use the latest 18 records for each users to predict the latest 5 records. With fine tune, feature engineering, and model design, the result could be better. I'll try the LSTM for time series in the next step.

Reference

Riiid: Comprehensive EDA + Baseline
https://www.kaggle.com/erikbruin/riiid-comprehensive-eda-baseline

Riiid! Answer Correctness Prediction EDA. Modeling
https://www.kaggle.com/isaienkov/riiid-answer-correctness-prediction-eda-modeling

Source Code

Please check the GitHub repository for more information.
https://github.com/hanxiao0607/CS5665