Riiid! Answer Correctness Prediction

Speaker：郭蕙綺、趙仰生

Progress

- Data preprocessing
- Data exploration

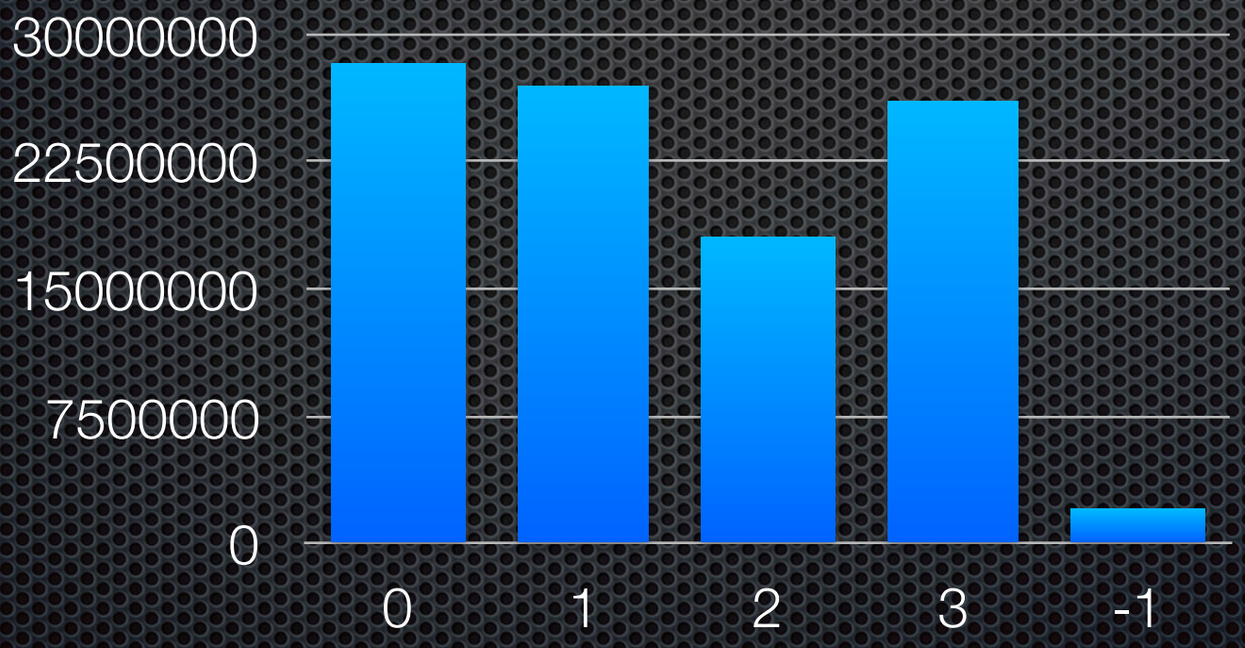
Data preprocessing

- The data is extremely huge, thus, we try to do some preprocessing on our data.

1. Feature selection -> It helps to reduce the dimensions.
2. Transfer data type -> especially for the feature “prior_question_had_explanation” (in train.csv) from object to boolean, which can reduce the memory usage from 3594972816 to 1012303332 bytes.

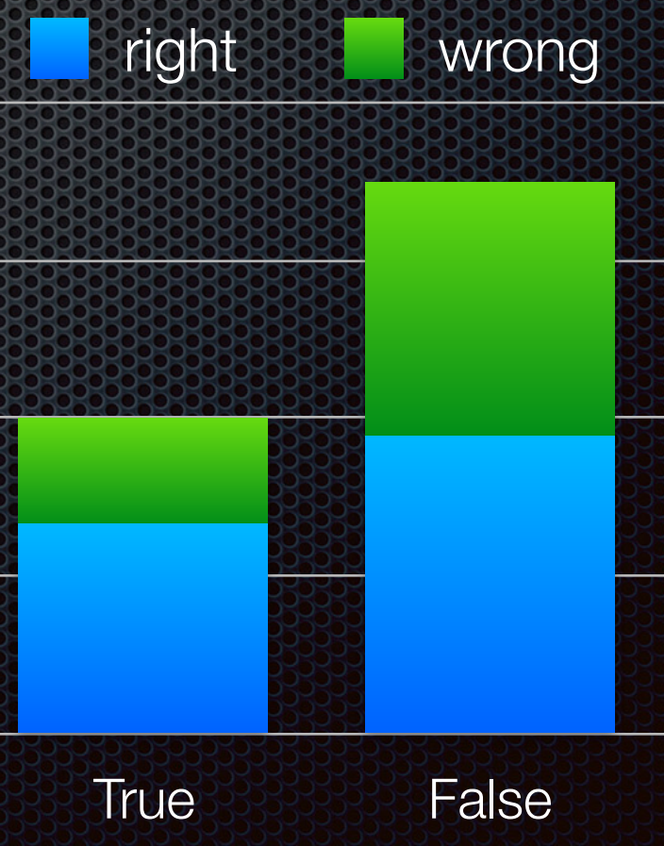
Data exploration

- We have 393656 unique users, 13523 unique questions, and 259 unique lectures.
- Answer distribution is quite interesting.



Data exploration

- Question which is explained has higher correct rate.
  (left pic. -> True:0.675, False:0.505)
- User who watches lectures has better performance.
  (right pic. -> True: 0.665, False:0.58)



Future Steps

- Baseline model : lightgbm(LGB), xbgoost(XBG).
- Features selections: “user_questions_amount”, “user_correct_rate”, “questions_answered_amount”, “questions_correct_rate”, “watch_lecture”
- Advanced model : AutoRegressive(AR) , RNN(LSTM).