PREDICTING STUDENT PERFORMANCE
780,000 STUDENTS
1 ALGORITHM

I. PROBLEM STATEMENT

What if I told you I had the algorithm that unlocks human learning potential?

I hope someday I will have that answer for you. For now, I'd like to focus on accurately predicting if an online student will answer correctly or incorrectly. This is a seemingly simple, yet critical, step in the process to enhance learning potential. Given a massive labeled dataset, what can we learn about the learning process?

BACKGROUND

Online education has been exploding exponentially in the past few years. "The Global Online Education Market is expected to expand at a 28.55% CAGR during the forecast period, 2017–2023" according to Market Research Future.[1] Growth has been catalyzed recently by the Covid-19 pandemic which has resulted in mandated distance learning in many places. Education Technology (EdTech) companies focus on how to better engage learners and how to optimize student performance using advanced machine-learning algorithms and even hyper-personalized 'AI-Tutors'. A human teacher can only reasonably track a certain number of students' specific strengths and weaknesses. Companies hope to overcome human teacher limitations by automatically recommending materials based on an individual's subject mastery level.

[1] Online Education Market by Type, Size, Growth and Forecast – 2023 | MRFR

Across the globe in Korea, English tests are a serious business. Each year, millions of South Koreans enroll in 'The Test of English for International Communication' (TOEIC) preparation courses. In 2019, 54% of South Korean TOEIC test takers indicated that their purpose was 'job application'.[2] A Korean person's TOEIC score is a "major factor in hiring people for most professional jobs" and "many universities in Korea still require a minimum score of 900".[3] Korean EdTech startup Riiid offers 'Santa TOEIC', "the AI-based web/mobile learning platform for TOEIC" with millions of enrolled users. The good news is, Riiid is into serious research[4] and they have shared user data anonymously in the aptly named "EdNet" dataset [5] of over 784,000 users & 131,441,538 interactions to encourage (gamify) crowdsourced research.[6]

GOAL

The primary purpose of this project is to predict a student getting a question correct or incorrect. Being able to predict a student answering correctly will inform the entire learning process. We can examine the strongest factors correlated with a student's likelihood to answer correctly and perhaps recommend enhancements to optimize the learning process.

II. DATASETS

The EdNet hierarchical dataset is large-scale consisting of:

- '131,441,538 interactions collected from 784,309 students of Santa since 2017'
- 441.20 interactions/student on average
- '13,169 problems and 1,021 lectures tagged with 293 types of skills'
- 'this is the largest dataset in education available to the public in terms of the total number of students, interactions, and interaction types'
- Student behavior logs: watching lectures, pause, purchase, stop, etc
- Multi-platform homogenization: iOS, Android, Web data all collected in a consistent manner
- Anonymized to protect user identities
- "As the level of the dataset increases [from KT1 - KT4], the number of actions and types of actions involved also increase."

[2] 2019 Report on Test Takers Worldwide : TOEIC Speaking & Writing Tests
[3] https://en.wikipedia.org/wiki/TOEIC#TOEIC_in_South_Korea
[4] https://arxiv.org/abs/1912.03072
[5] riiid/ednet
[6] http://ednet-leaderboard.s3-website-ap-northeast-1.amazonaws.com/ RETRIEVED ON 9/28/2020

KT1 = 784,309 csv files relating to each user (student) - 5.6GB

KT1 FEATURES = timestamp, question_id, bundle_id, user_answer, elapsed_time

KT2 = 297,444 csv files relating to each user (student) - 3.1GB

KT2 FEATURES = timestamp, action_type, item_id, source, user_answer, platform

KT3 = 297,915 csv files relating to each user (student) - 4.3GB

KT3 FEATURES = timestamp, action_type, item_id, source, user_answer, platform

KT4 = 297,915 csv files relating to each user (student) - 6.4GB

KT4 FEATURES = timestamp, action_type, item_id, cursor_time, source, user_answer, platform

CONTENTS = 4 csv files: questions, lectures, payment items, coupons

QUESTIONS FEATURES = question_id, bundle_id, explanation_id, correct_answer, part, tags, deployed_at

LECTURES FEATURES = lecture_id, part, tags, video_length, deployed_at

PAYMENT ITEM FEATURES = payment_item_id, payment_type, duration, number_of_bundles

COUPON FEATURES = coupon_id, coupon_type, duration

TOTAL = **1,677,587 csv files - 19.4GB, 29 features, 131,441,538 observations**

III. DATA WRANGLING & DATA CLEANING

The data organization was a major task in this project. The dataset is available publicly on the project's official Github as compressed (zip) files. Upon downloading and decompressing, there are 19.4 GB of data consisting of 1.6 million individual (csv) files. The files are divided hierarchically into the 4 tiers: KT1 - KT4. Initially, I merged the 4 levels per student. Upon inspection, I noticed that a great number of students had missing values for the higher tiers. Noticing the sparsity of the higher-tiers I decided to focus this study on KT1, since this tier represents all students in the dataset.

After managing the scope of the project, I proceeding to:

1. Cleaned up the data by converting the unix timestamps into human readable datetime

2. Dropped duplicate rows

3. Converted the elapsed time from milliseconds to seconds

4. Transformed data columns with unneeded leading alpha characters into an integer series.

5. Created a new master file (using glob and a python list comprehension),
6. Used a “pandas” concatenation

7. Exporting a cleaned and augmented (csv) file

8. Merged the user data with the ‘questions’ data on the rows with the same questions

9. Feature engineered a new binary ‘correct’ column filled with ‘1’s if correct and ‘0’s if not correct. At that point the data is ‘labeled’ and I had a clear dependent variable to create an algorithm around.

10. Cleaned up the order of the features and organized them into logical groups

The questions file had a ‘tag’ feature relating to expertly tagged categories of the questions. The tag feature column had up to 7 comma-separated integers.

11. Created a new dataframe by splitting the integers into independent columns and

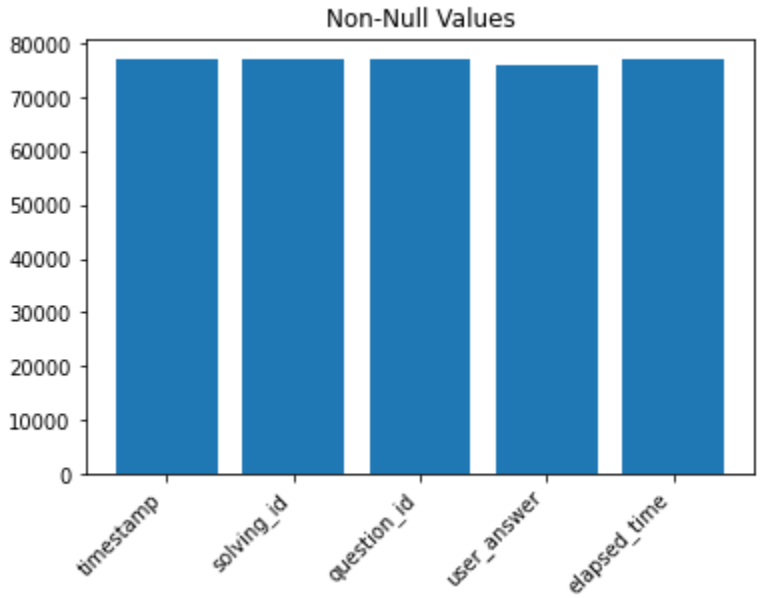
12. Merged the data frame back into the master data frame resulting in 15 features

The data cleaning process took a good deal of experimentation and massaging resulting in a “Tall” (csv) file with 23 million observations and 15 features.

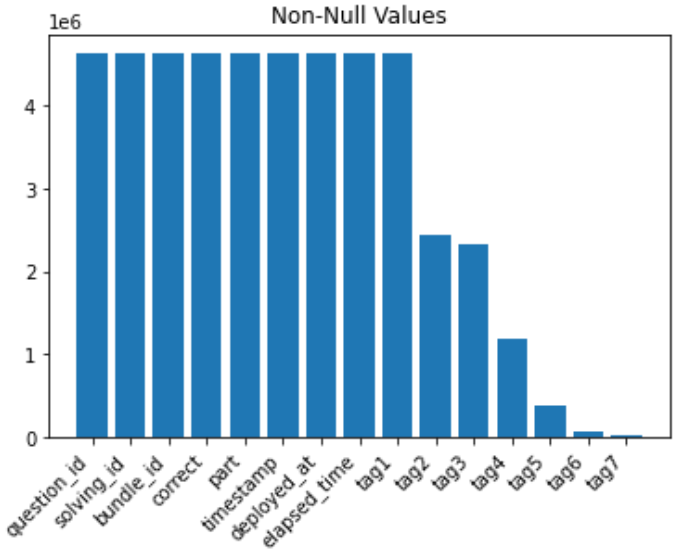
	question_id	solving_id	bundle_id	correct	part	deployed_at	elapsed_time	tag1	tag2	tag3	tag4	tag5	tag6	tag7
480004	9709	6550	7060	1	5.0	2018-05-18 09:42:11.702	20.0	74	0	0	0	0	0	0
480005	10205	6551	7556	0	5.0	2019-10-17 03:15:28.665	3.0	106	0	0	0	0	0	0
480006	7013	6552	5231	1	7.0	2019-09-16 12:01:51.089	67.2	152	147	163	179	178	149	178
480007	7014	6552	5231	1	7.0	2019-09-16 12:01:52.761	67.2	152	147	163	179	178	149	178
480008	7015	6552	5231	0	7.0	2019-09-16 12:01:54.417	67.2	152	147	163	179	178	151	178

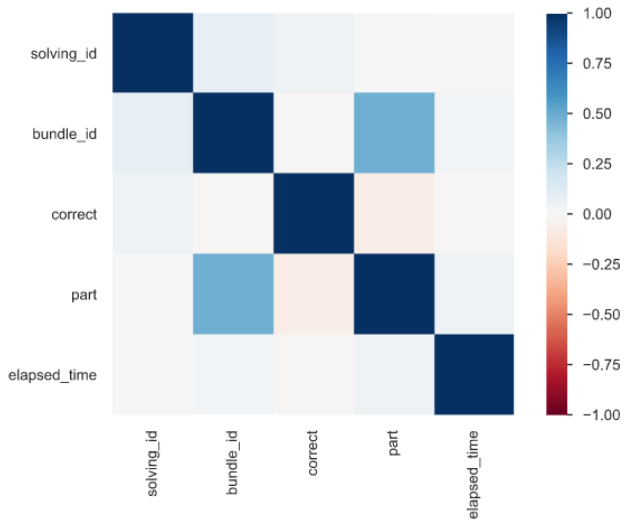
IV. EXPLORATORY DATA ANALYSIS

In order to ensure the veracity of the data, I graphed the 'non-null' values per column at various stages of progress.



The bar chart above represents almost 80,000 student interactions on the online platform as represented by non-null values in the data-set. The cleaned dataset used for the study included 23 million interactions.





question_id

Categorical

HIGH CARDINALITY

Distinct count	13239
Unique (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	37.4 MiB

1316	1423
854	1419
739	1411
952	1339
878	1302
Other values (13234)	4890132

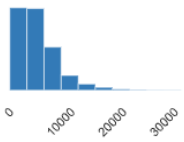
Toggle details

solving_id

Real number ($\mathbb{R}_{\geq 0}$)

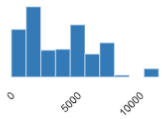
Distinct count	33132
Unique (%)	0.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	5381.372666185558
Minimum	1
Maximum	33132
Zeros	0
Zeros (%)	0.0%
Memory size	37.4 MiB



bundle_id
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	8958	Mean	4030.0536185917435
Unique (%)	0.2%	Minimum	1.0
Missing	7845	Maximum	12206.0
Missing (%)	0.2%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	37.4 MiB



Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Quantile statistics

Minimum	1
5-th percentile	406
Q1	1625
median	3751
Q3	5844
95-th percentile	8480
Maximum	12206
Range	12205
Interquartile range (IQR)	4219

Descriptive statistics

Standard deviation	2811.192382
Coefficient of variation (CV)	0.697557067
Kurtosis	-0.2873034075
Mean	4030.053619
Median Absolute Deviation (MAD)	2116
Skewness	0.6120642152
Sum	1.970366158e+10
Variance	7902802.61

The cleaned and augmented data set has 15 features. The 'tag' columns were generated from a split of the original 'tags' columns which are expert-coded classifications of the questions. Each question had at least 1 tag and some had as many as 7 tags.

V. MODEL SELECTION

Will the student answer correctly or incorrectly? The question is binary. The student is either correct or incorrect. This is a binary classification problem and there are numerous possible machine learning models that could provide predictions. My approach was simply to survey a range of classification algorithms and identify the top performer. The algorithms I compared are logistic regression, Gaussian Naive Bayes, Stochastic Gradient Descent Classifier, K Nearest Neighbors Classifier, and Random Forest Classifier. I compared accuracy scores and ROC AUC (Area Under the Curve).

Classification Algorithm Results

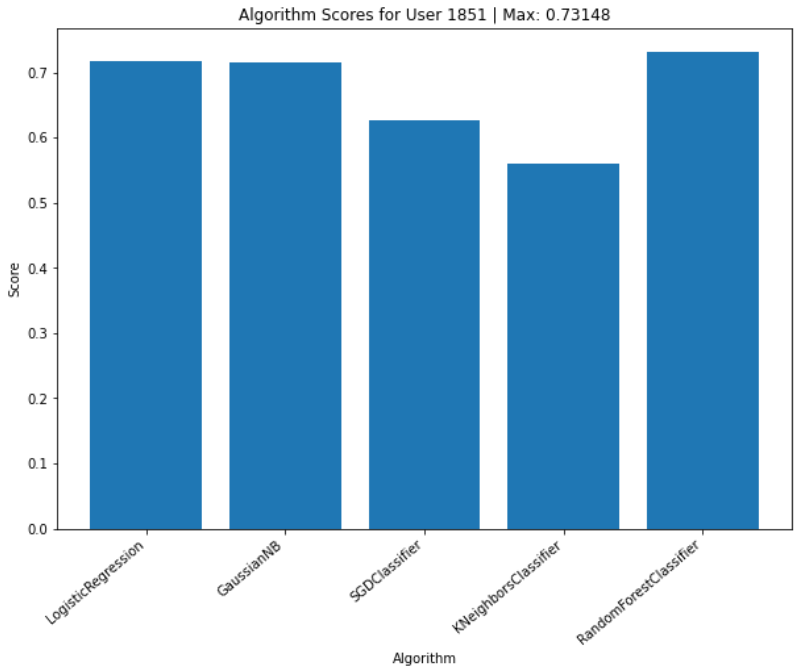
1. **Random Forest Classifier Score: 0.73148 ; AUC = 0.71084**

2. Logistic Regression Mean Accuracy Score = 0.71724 ; AUC = 0.56379

3. Gaussian Naive Bayes Mean Accuracy Score = 0.71660 ; AUC = 0.56808

4. Stochastic Gradient Descent Classifier Mean Accuracy Score = 0.62579 ; AUC = 0.47581

5. K-Nearest Neighbors Mean Accuracy Classifier Score = 0.55950 ; AUC = 0.56222



VI. SUMMARY

Upon reviewing the results of the machine learning classification algorithms, the Random Forest Classifier algorithm emerged as the winner. Random Forest best predicted if the students will choose correctly with a score with an ROC AUC of 0.71084. Given the input features, we were able to predict fairly well. The R-Squared score for the Random Forest Classifier is 0.73148, meaning that the regression fit the data fairly well.

VII. OPEN LOOPS & FURTHER RESEARCH

The depth and breadth of the dataset opens up many possibilities for further research. In fact, the research paper titled: 'EdNet: A Large-Scale Hierarchical Dataset in Education' (SOURCE: https://arxiv.org/pdf/1912.03072.pdf) describes knowledge tracing as a crucial step towards "learning path recommendation, score prediction and dropout prediction".

The research that interests me most is 'learning path recommendation' akin to an AI hyper-personalized tutor. This to me represents an important transition in the history of teaching. If the learning materials are proven effective and we can accurately trace a student's 'knowledge state' based on their previous answers, machine-learning systems (Reinforcement Learning) can recommend the learning path, just as a human teacher or tutor can assess a student and suggest for them strengthen their weaknesses.

1 Million Flashcards

An analogy I find useful is imagining the entire curriculum of a class as a set of flashcards with a question on the front and an answer on the back. Students attempt the flashcards and they put the flashcards they have mastered in pile #1 and the flashcards they have not mastered in pile #2. Obviously, the items in pile #2 should be reviewed more often until they are mastered. Now imagine being a teacher an attempting to keep track of 50 students' pile #2 cards. This approach can not scale and some students will have a disproportionately large pile #2. Since the memory of a computer is superior to a human's, once the learning path recommendations are well-calibrated (perhaps according to a human teacher's feedback), a single teacher could 'teach' thousands of students. This approach is still human-teacher-centered, but automation allows teachers to focus more on the teaching materials and less on the assessment and knowledge tracking aspect of teaching.

MOOCs vs Teachers

Recently, with the boom of massive open online courses (MOOC), EdTech, online learning, and even 100% online college degrees the emphasis is on the materials. This to me begs the question, "Can we just make 1 'ultimate' MOOC version of each class?" For example, if every student has access to the entire curriculum of High-School Algebra 1, and this is a standardized, highest quality course, is it really necessary for millions of teachers to

re-teach it each year? Education is trending in this direction, and machine learning systems are accelerating the effectiveness and efficiency of how people learn. Yes, there are many subjects that will still require an in-person, human teacher, but the trend towards EdTech solutions is only growing.

Chinese Cyborgs & Elon Musk

In China, some classrooms are experimenting with “brain-reading” headbands (LINK: https://qz.com/1742279/a-mind-reading-headband-is-facing-backlash-in-china/) to track attention and the data is aggregated in a dashboard on a teacher’s device and shared with the students' parents. This is an extreme example and very experimental, but it demonstrates one possible future. Elon Musk’s Neuralink company intends to implant internet connected devices into people’s skulls (already achieved with pigs). It’s easy to imagine a not-so-distant future where society will no longer need their fingers or voice to “Google” the world of knowledge on the Internet. This will change the nature of ‘memorization-based’ education, and potentially shift mankind’s educational focus on higher-level tasks such as synthesizing new insights. Cyborg speculation aside, it’s a very exciting time to get involved in the education revolution.