

EdNet: Predicting Student Performance

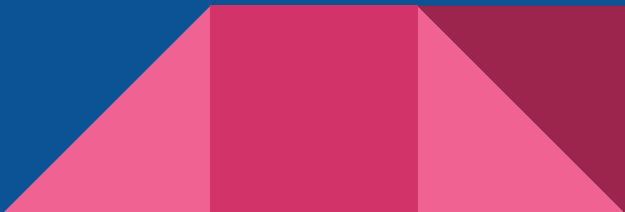
Springboard Data Science

A Report by Prem Ananda

10/13/2020

PROBLEM STATEMENT

*Online Education is expanding at about 28.55% each year 2017-2023.
By accurately predicting if a student will get an answer correct or incorrect, we can better recommend a learning path for the student and ultimately optimize their online learning.*

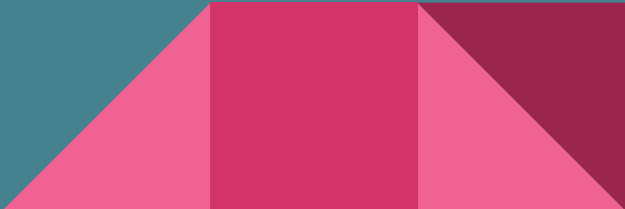


Objective

“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.”



DATASET

- 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
- 

DATA WRANGLING & DATA CLEANING

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.

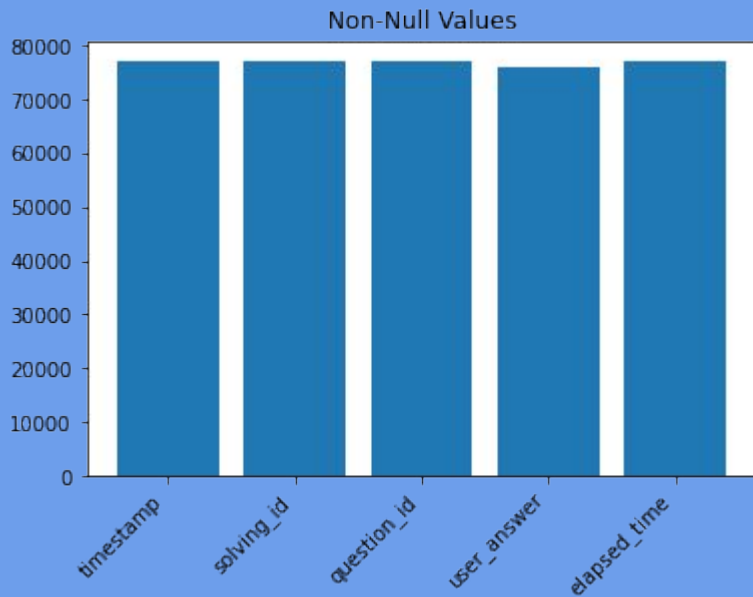
DATA WRANGLING & DATA CLEANING

	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

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.

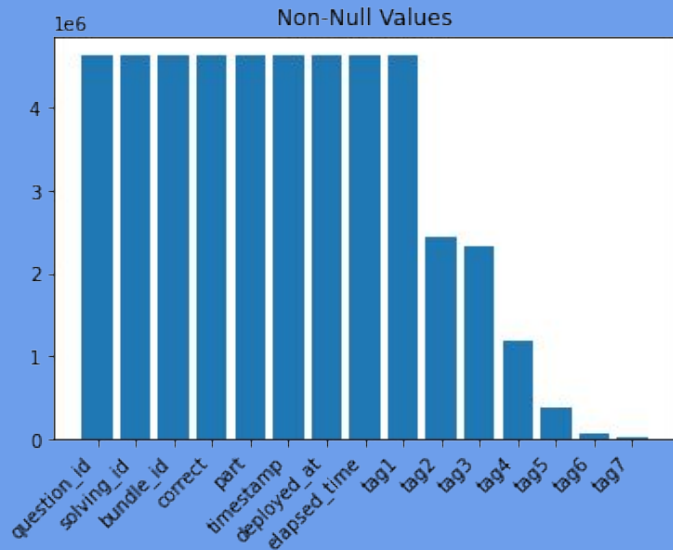
EXPLORATORY DATA ANALYSIS

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

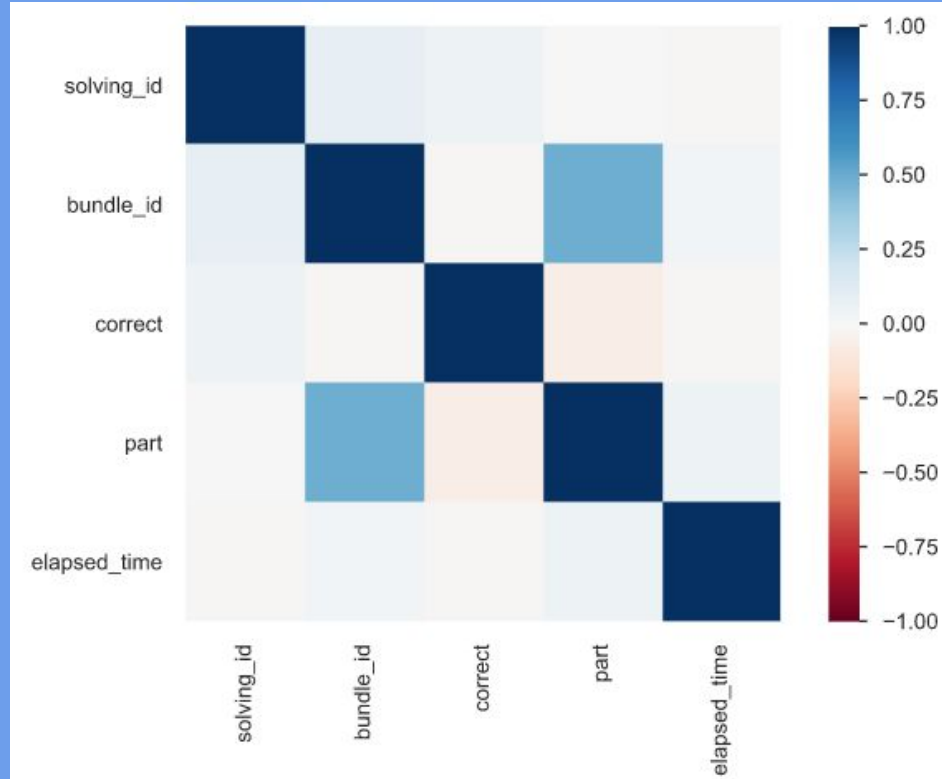


EXPLORATORY DATA ANALYSIS

The bar chart below 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.



EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

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

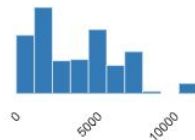
EXPLORATORY DATA ANALYSIS

bundle_id

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

Distinct count	8958
Unique (%)	0.2%
Missing	7845
Missing (%)	0.2%
Infinite	0
Infinite (%)	0.0%

Mean	4030.0536185917435
Minimum	1.0
Maximum	12206.0
Zeros	0
Zeros (%)	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

MODEL SELECTION

We compare various classification algorithms such as logistic regression, Gaussian Naive Bayes, Stochastic Gradient Descent Classifier, K Nearest Neighbors Classifier, and Random Forest Classifier.

We compare accuracy scores and ROC AUC (Area Under the Curve).



MODEL SELECTION

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

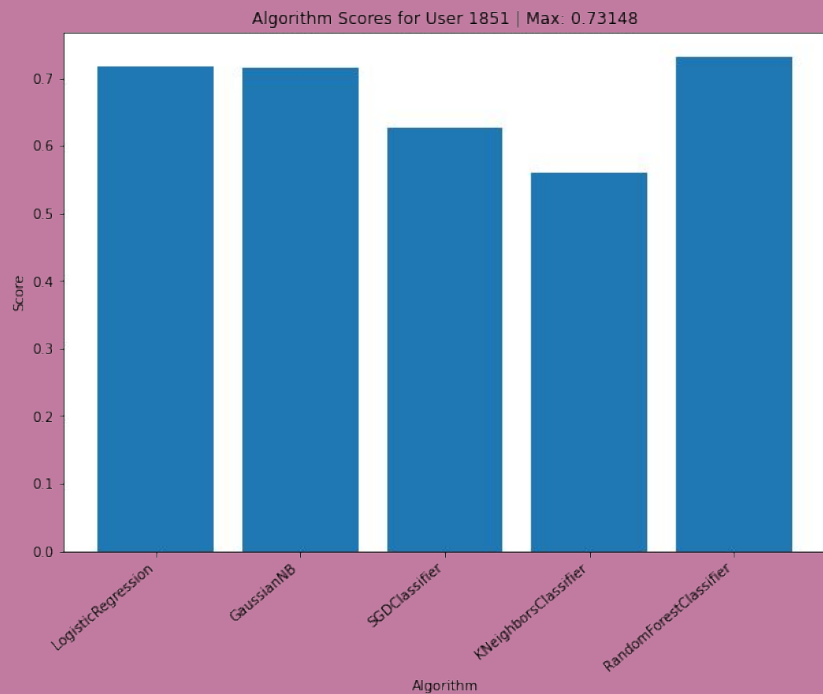
5. K-Nearest Neighbors

Mean Accuracy Classifier Score = 0.55950 ; AUC = 0.56222

*Best performing

MODEL SELECTION

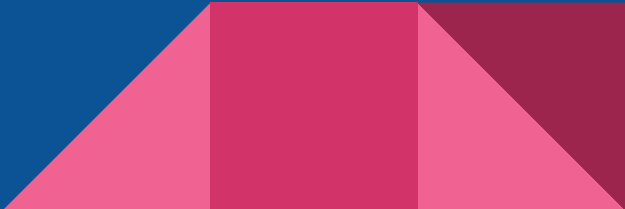
Ranked Classification Algorithm Results



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.

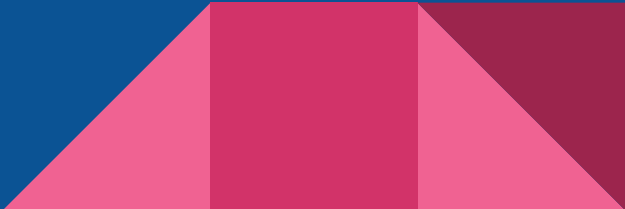


OPEN LOOPS & FURTHER RESEARCH

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.



EdNet: Predicting Student Performance

Springboard Data Science

A Report by Prem Ananda

10/13/2020