An Analysis of Machine Learning on Educational Data

Bhavana Jonnalagadda Erik Whitfield

Table of contents

Abstract

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

1.1 Background

Example citations are (haradai1262, 2020), and another (shshen-closer (2021)).

1.2 Research Questions

1.2.1 Objective

1.2.2 Rationale

2 Methods

2.1 Data

The data used format can be seen in Table 2.1. The description of the columns used are as follows:

- **QuestionId**: ID of the question answered.
- **UserId**: ID of the student who answered the question.
- **AnswerId**: Unique identifier for the (QuestionId, UserId) pair, used to join with associated answer metadata (see below).
- **IsCorrect**: Binary indicator for whether the student's answer was correct (1 is correct, 0 is incorrect).
- **SubjectId**: Each subject covers an area of mathematics, at varying degrees of gran- ularity. We provide IDs for each topic associated with a question in a list. Example topics could include "Algebra", "Data and Statistics", and "Geometry and Measure". These subjects are arranged in a tree structure, so that for instance "Factorising" is the parent subject of "Factorising into a Single Bracket". We provide details of this tree in an additional file subject metadata.csv which contains the subject name and tree level associated with each SubjectId, in addition to the SubjectId of its parent subject.
- **Category1**: Feature engineered. The first-level category of the question (given that there is hierarchical categories)
- **Gender**: The student's gender, when available. 0 is unspecified, 1 is female, 2 is male and 3 is other.
- **Age**: Feature engineered. The student's age, as calculated from `DateAnswered - DateOfBirth`.
- **PremiumPupil**: Whether the student is eligible for free school meals or pupil premium due to being financially disadvantaged.
- **DateAnswered**: Time and date that the question was answered, to the nearest minute.
- **Confidence**: Percentage confidence score given for the answer. 0 means a random guess, 100 means total confidence.
- **GroupId**: The class (group of students) in which the student was assigned the question.
- **QuizId**: The assigned quiz which contains the question the student answered.

Table 2.1: The source data used in this paper, tranformed by unions across several csvs, some columns dropped, and some created columns.

(a) First set of columns

	QuestionId	UserId	AnswerId	IsCorrect	DateAnswered	Confidence	GroupId
0	898	2111	280203	1	2019-12-08 17:47:00	nan	95
1	767	3062	55638	1	2019-10-27 20:54:00	25	115
2	165	1156	386475	1	2019-10-06 20:16:00	nan	101
3	490	1653	997498	1	2020-02-27 17:40:00	nan	46
4	298	3912	578636	1	2019-12-27 16:07:00	nan	314

(b) Second set of columns

	QuizId	Gender	PremiumPupil	SubjectId	Age	Category1
0	86	2	False	[3, 49, 62, 70]	12	Algebra
1	39	0	False	[3, 32, 144, 204]	nan	Number
2	39	0	False	[3, 32, 37, 220]	nan	Number
3	115	0	False	[3, 49, 81, 406]	nan	Algebra
4	78	2	False	[3, 71, 74, 180]	11	Geometry and Measure

2.2 Statistical Methods Analysis

2.2.1 Clustering

2.2.2 Supervised Learning Models

2.2.3 Principal Component Analysis (PCA)

2.2.3.1 Using PCA with Supervised Learning

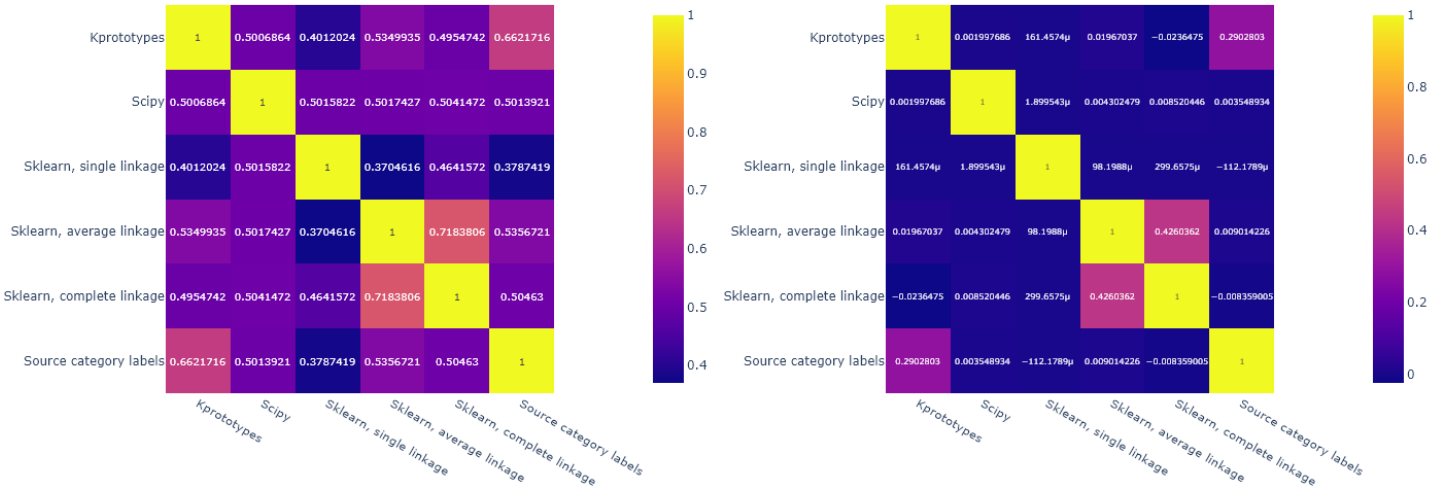
TODO:

- Compare results of clustering to the given question category labels
- Do PCA, do regression on both the PCs and original features, compare regression performance
- Do regression
  - Where label/output is the student's total score
  - Regress on student info, predict total student score across all answers
    - Also if it's predictive of score in the question category
- Random forest?

3 Results

3.1 Clustering

See Figure 3.1



(a) Rand Index Matrix

(b) Adjusted Rand Index

Figure 3.1: Matrices of the rand index as compared across all clustering methods, and compared to the original source category labels. The scipy and sklearn clustering methods were done using a precomputed Gower matrix.

3.2 Supervised Learning Models Performance

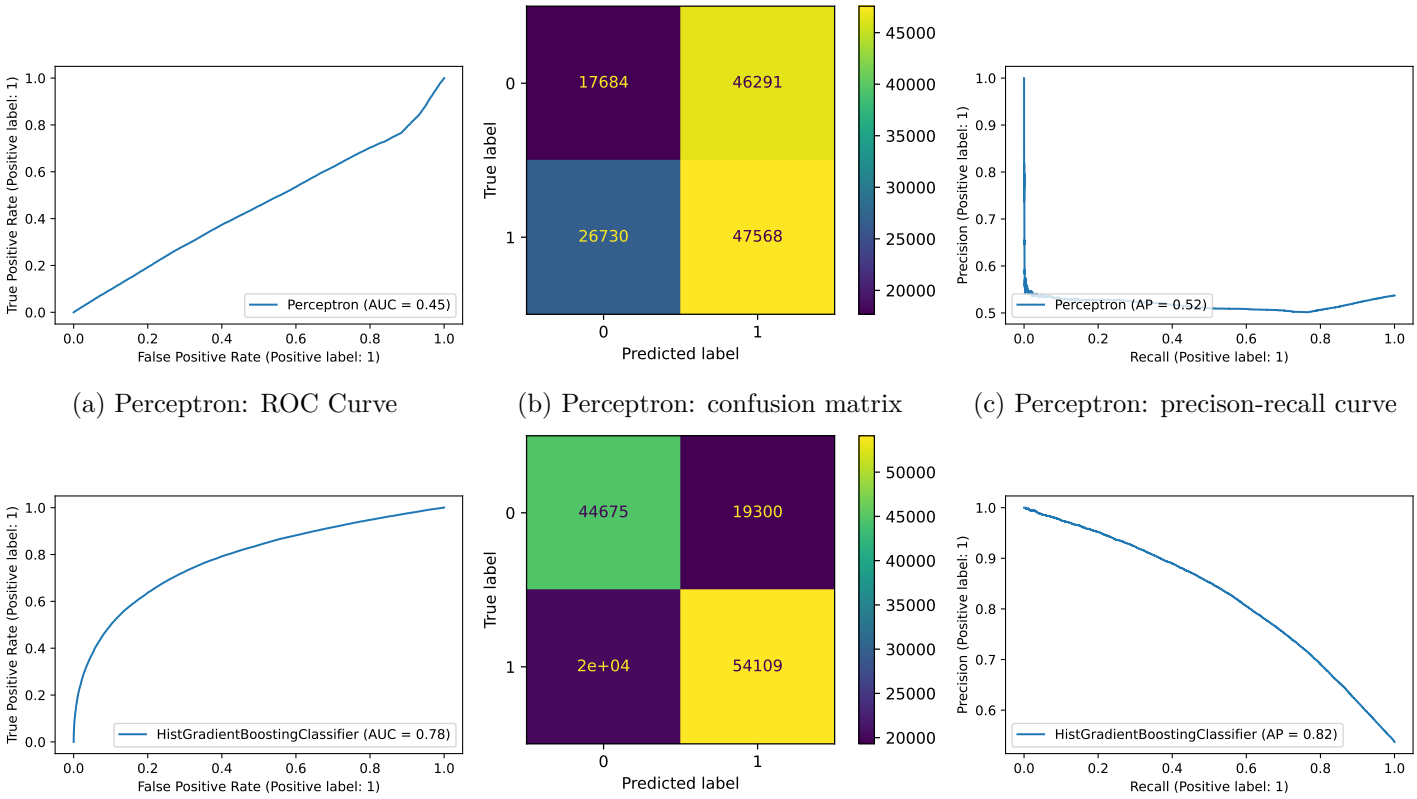
See Table 3.1

Table 3.1: The performance values and some settings for the supervized learning models ran, sorted by accuracy.

	Model	Accuracy	precision	recall	f1-score	support	TrainIters	LossFcn
7	HistGradientBoostingClassifier	0.7144	0.7147	0.7144	0.7145	138273.0	NaN	log_loss
6	HistGradientBoostingClassifier	0.6602	0.6600	0.6602	0.6601	138273.0	NaN	log_loss
5	RandomForestClassifier	0.6519	0.6519	0.6519	0.6519	138273.0	NaN	NaN
4	DecisionTreeClassifier	0.6108	0.6107	0.6108	0.6108	138273.0	NaN	NaN
2	LinearSVC	0.5703	0.5663	0.5703	0.5599	138273.0	12	squared_hinge
0	LogisticRegression	0.5703	0.5662	0.5703	0.5601	138273.0	[0]	NaN
3	SGDClassifier	0.5695	0.5655	0.5695	0.5550	138273.0	41	log_loss
1	Perceptron	0.4719	0.4565	0.4719	0.4550	138273.0	11	perceptron

3.2.1 Performance Plots

See Figure 3.2



(a) Perceptron: ROC Curve

(b) Perceptron: confusion matrix

(c) Perceptron: precison-recall curve

(d) HistGradientBoostingClassifier: ROC Curve

(e) HistGradientBoostingClassifier: confusion matrix

(f) HistGradientBoostingClassifier: precison-recall curve

Figure 3.2: The ROC, confusion matrix, and precision-recall curves for the best and worst performing model.

4 Conclusions

4.1 Findings

4.2 Limitations

4.3 Future Studies

References

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9365–9374.

Ghosh, A., & Lan, A. (2021). BOBCAT: Bilevel optimization-based computerized adaptive testing. In Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 2410–2417). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2021/332

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. http://www.jstor.org/stable/2528823

haradai1262. (2020). Solution of NeurIPS education challenge 2020. In *GitHub repository*. GitHub. https://github.com/haradai1262/NeurIPS-Education-Challenge-2020

shshen-closer. (2021). TOP1-for-task-2-in-the-NeurIPS-2020-education-challenge. In *GitHub repository*. GitHub. https://github.com/shshen-closer/TOP1-for-task-2-in-the-NeurIPS-2020-Education-Challenge

Vos, N. J. de. (2015--2021). *Kmodes categorical clustering library*. https://github.com/nicodv/kmodes.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernandez-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., & Zhang, C. (2021). *Results and insights from diagnostic questions: The NeurIPS 2020 education challenge*. https://arxiv.org/abs/2104.04034

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., & Zhang, C. (2020). Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv Preprint arXiv:2007.12061*.