20182597-Final

Summary

In this project, I tried to upgrade the performance of the model using the technique called “dimension reduction” that was covered in my class.

It turns out that some selections show better performance compared to others. But there is no single selection that outdoes the rest of the selections.

Introduction

I used two approach to reduce the dimension of the features.

1. select features
2. primary component analysis (PCA)

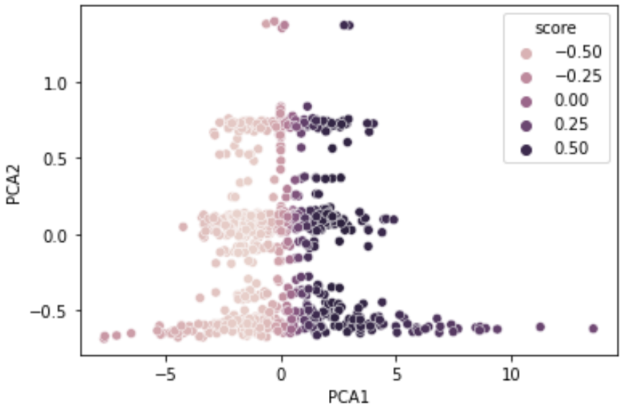
Luckily I had only 4 features I didn’t need to use greedy method. For example, forward selection or backward selection. I could use the brute force to calculate the all possible results. I only had to check 15 cases excluding the empty case.

And I used PCA method to reduce the dimension. I decided to use 2 PCAs for easy and clean plotting.

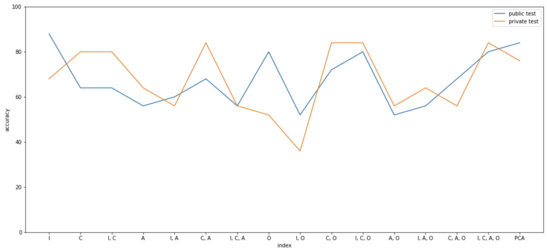
The maximum agreement rate is used as the performance level of the model. Each is calculated on the public test data and the private test data respectively.

Results

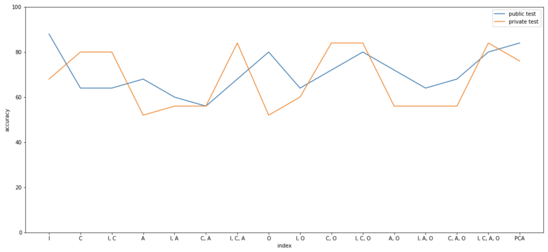
I separately performed PCA because selection is just the extension of the previous project but PCA is the whole new chapter. The first PCA is taking crucial role to decide the group.



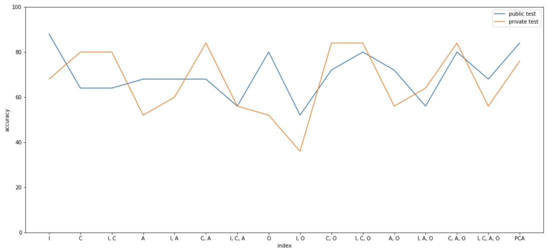
I performed the process many times. Because every time I run it each shows different performance level.



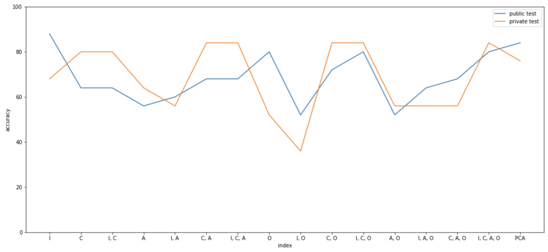
# first



second



# third



fourth

`I` is IsCorrect, `C` is Confidence, `A` is Age and `O` is Opacity. Some selections show good performance. However it doesn't show good performance in both public and private data, which indicates that the model is not stable. For example the case which only selects `IsCorrect` has about 85% agreement rate in the public test data but shows about 65% rate in the private test data.

`IsCorrect, Confidence, Opacity` Selection and `PCA` gave quite stable performance compared to other options.

Conclusion

I chose `IsCorrect, Confidence, Opacity` selection over `PCA` for the interpretability. This is the best result I could get. Correct rate, Confidence level of the students and the ambiguous level of the question are important. So when we design the question we need to consider these features.

Repository

https://github.com/HooKim/question_quality_analysis