Predicting Student Success with a Content-based Model

Executive Summary

Q3 2021

Problem identification - Context

- The number of students in the world has been steadily increasing.
- The ratio of teachers to students has been decreasing. (Source: UNESCO Institute for Statistics)
- Yet, even with these opposing shifts, the most common way to track student knowledge has been to pose question to a student and have a teacher evaluate the student's response

Problem identification - Context

- Should this trend continue, the most common approach used today will not scale to fit the education environment of the future.
- Machine learning models are well-adapted to this type of problem.
- What is the best approach to build a model that will predict the accuracy of students' responses to test questions?

Problem identification - Success Criteria

1. The best model will maximize ROC AUC.
2. Training time should scale linearly with the size of the data set.
3. Prediction time should scale linearly with the size of the data set.

Problem identification - Scope

- Binary classification problem
- Target: correctness of student response (true or false)
- Features: 200 features, derived from the content metadata and 8 non-derived features
- Prediction: there exists a model that is significantly better than the baseline (0.5>>model AUC >=1.00)
- This content-based approach will exclude time series constraints.

Key findings

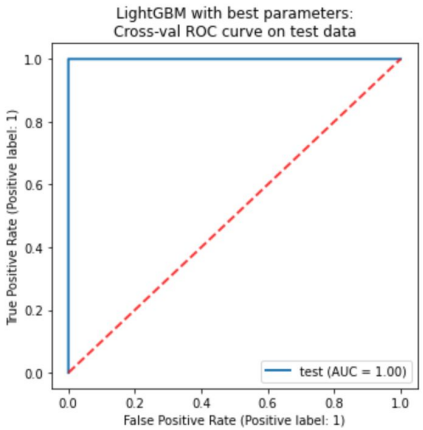
Among 6 models tested:

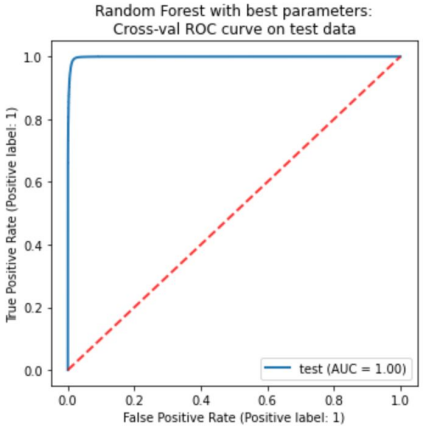
- LightGBM model is top-ranked, and
- Random Forest model is the second-ranked

in all three success criteria.

Modeling results and analysis - Performance

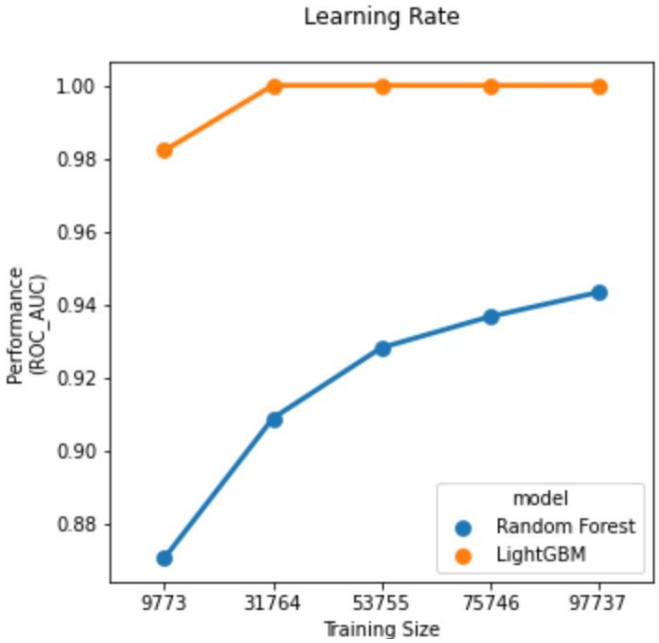
- With lots of training data, the two best-performing models--LightGBM and Random Forest--perform almost equally well.





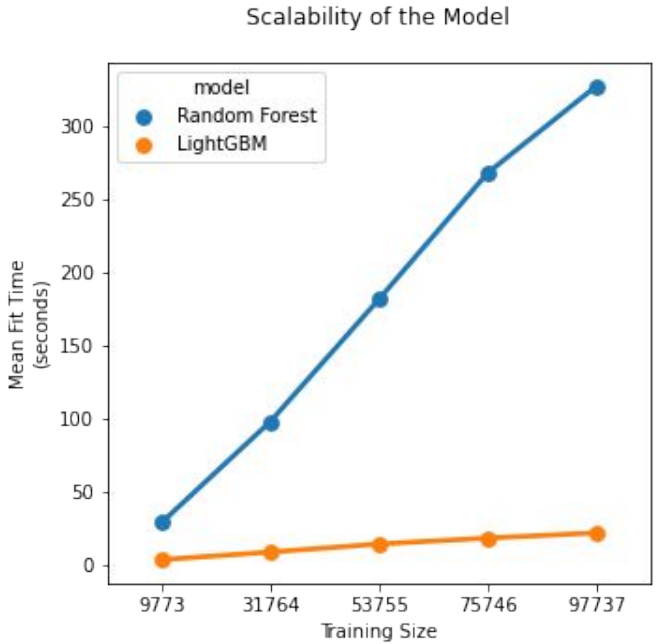
Modeling results and analysis - Performance

- With limited training data, LightGBM model outperforms the Random Forest model.



Modeling results and analysis - Scalability (training time)

- LightGBM model trains more than 10x faster than the Random Forest Model



Summary and conclusion

- Recommendation: use the LightGBM model for predicting student responses to test questions.

Further work

- Evaluate some deep-learning models as there is sufficient sample data to support these types of models.
- Evaluate the effect of ignoring time-series constraints in this analysis.