# MOOC Dropout Prediction

Zixun Yang

## ABSTRACT

In this project, I built model to predict droupout in Massive Open Online Course(MOOC) platform, which is the topic in KDD cup 2015.

With my feature engineering result on this complicated three-dimensional dataset, I first explored different models and optimized parameters selection to reach best performance. With few models of good prediction on validation data, I ensembled them together using XGBoost Classifier to do a second level training based on the first level training prediction. In addition, I implemented Long Short Term Memory(LSTM) Recurrent Neural Networks with Keras on 30 days univariate Time series data and reached 0.857.

## CONTACT

Zixun Yang
Email: jasonyzx@stanford.edu
SUID# 06236719

## GOAL

**Predict whether or not a student will drop out MOOC.** Massive Open Online Course (MOOC) has been revolutionizing the way people getting education. However, it also raises up concern that MOOC has very high dropout rate relative to traditional classes. An accurate prediction of dropout becomes very important because it can help MOOC course developers to adaptively tune web designs to students with high dropout rate.

## Feature Engineering

72325 Observations

Train: 46288

Dev: 11572

Test: 14465

Activity Log

| ID | Activity | Time |
|---|---|---|
| 15962 | Wikipedia | 2013-12-26T08:47:02 |
| 15962 | Video | 2013-12-26T08:47:16 |
| 15962 | Navigate | 2013-12-26T08:57:24 |

Enrollment list

| ID | Student ID | Course ID |
|---|---|---|
| 15962 | hjsacfRSdoCE7ZGEhwZqKesr | djBWhu0JoDsrQ2a6Kzg6B4E2 |
| 15927 | 67Y9pdjnd1B5TSfVmd3RQxl | djBWhu0JoDsrQ2a6Kzg6B4E2 |

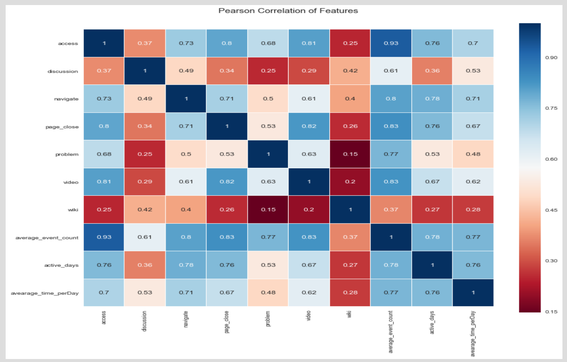
(7 + 1) x 30 | 30 | 39

Extracted features

| Day1 Activity Count | … | Day30 Activity Count | Day1 Online Time | … | Day30 Online Time | One Hot Encoded Course Type |
|---|---|---|---|---|---|---|
| 7 Overall Event Count | | Average Online Time | Unique Events Count | … | Active days | Average Event Count |

## Feature Selection

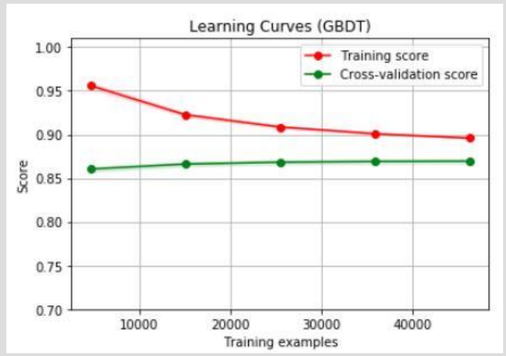
**Pearson Correlation of Features**

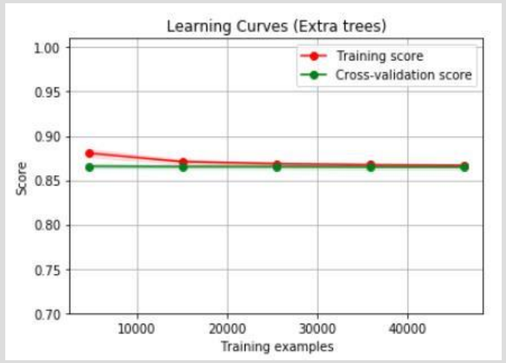



## Model Ensemble

**First Level of Model Ensemble**

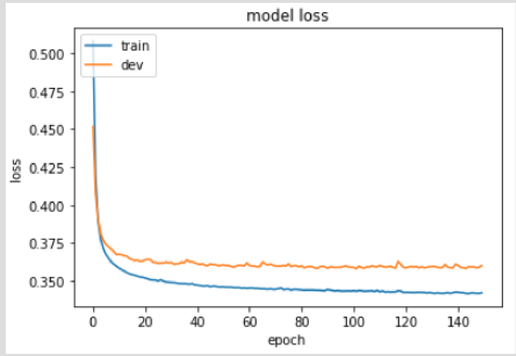
We fine-tuned SVC, Logistic Regression, Extra Trees and Gradient Boosting Decision Tree and Random Forest. Each of them have reached accuracy above 86%.

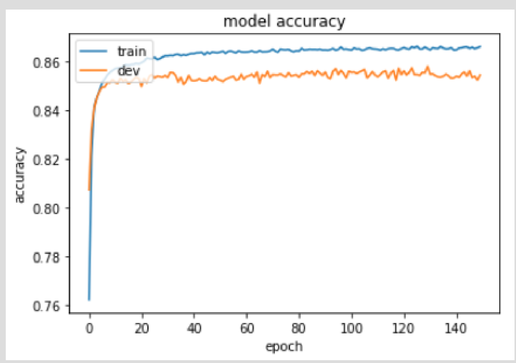



## LSTM

Since our data is time sequence data, and RNN is especially good at process sequence. We used all the day-based features, including seven types of event count, overall event count and online time, to train a stacked two layer LSTM model. Feature Size = 7+1+1 * 30. Below is the learning curve with batch size=200




## Conclusion

**Second Level of Model Ensemble**

After having these five classifiers, I ensembled/stacked the resuls of above five classifiers except LSTM and feed them to **XGBboost** classifier for the second level training, reached accuracy of **87.5%**

## Reference

- Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
- Liang, J., Li, C., & Zheng, L. (2016, August). Machine learning application in MOOCs: Dropout prediction. In Computer Science & Education (ICCSE), 2016 11th International Conference on (pp.52-57). IEEE.
- KDD Winner white paper: http://www.conversionlogic.com/wp-content/uploads/2016/06/Whitepaper- KDD2015-JYL.pdf