In this project, I built a model to predict dropout 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. 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) * 30. Below is the learning curve with batch size=200.

**Feature Engineering**

**Feature Selection**

**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%.

Second Level of Model Ensemble

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

**Conclusion**

**Reference**