

# Multi-Stage Ensemble and Feature Engineering for MOOC Dropout Prediction

**Jeong-Yoon Lee**  
Conversion Logic

12300 Wilshire Blvd. Los Angeles,  
CA 90025, USA

jeong@conversionlogic.com

**Kohei Ozaki**  
AIG Japan Holdings KK

Kamiyacho MT Bldg 18F, 4-3-20  
Toranomon, Minato-ku, Tokyo  
105-0001, Japan

ozaki.kohei@aig.co.jp

**Song Chen**  
American International Group,  
Inc. (AIG)

175 Water Street, New York, NY  
10038, USA

song.chen@aig.com

**Andreas Toescher**  
Opera Solutions

Hauptplatz 12, 8580 Koeflach,  
Austria

andreas.toescher@commendo.at

**Mert Bay**  
Conversion Logic

12300 Wilshire Blvd. Los Angeles,  
CA 90025, USA

mert@conversionlogic.com

**Tam T. Nguyen**  
Institute for Infocomm  
Research, A\*STAR

1 Fusionopolis Way, #21-01  
Connexis (South Tower), Singapore  
138632

nguyentt@i2r.a-star.edu.sg

**Michael Jahrer**  
Opera Solutions

Hauptplatz 12, 8580 Koeflach,  
Austria

michael.jahrer@commendo.at

**Peng Yan**  
NetEase Youdao

2nd Floor, Chuangye Building,  
Tsinghua Science Park, Beijing,  
100084, China

yanpeng@rd.netease.com

**Xiaocong Zhou**  
Tsinghua University  
Haidian District, Beijing, 100084,  
China

infinitezxc@gmail.com

## ABSTRACT

This paper describes the winning solution of KDD Cup 2015. The competition aims to predict dropouts in Massive Open Online Courses (MOOCs). Our approach begins with feature engineering to extract predictive features from activity logs of students and meta data. Then, we train sixty three individual classifiers with different subsets of features and seven algorithms. Lastly, we blend predictions of individual classifiers with the three stage ensemble framework. Our solution results in an AUC score of 0.9074 on the private leaderboard.

## Categories and Subject Descriptors

I.5 [Pattern Recognition]: Application

## General Terms

Application

## Keywords

KDD Cup, Feature Engineering, Ensemble Learning

## 1. INTRODUCTION

Our final solution is a joint work from 9 data scientists, distributed around the world. The pipeline from raw data to final solution is as follows:

- Hand crafted feature design (most of hard work)

- Automatic feature design (autoencoder)
- Individual models (gbm, nn, factor model,...)
- Stage2 ensemble (blends individuals)
- Stage3 ensemble (blends Stage2's)

This approach was published by a winning team of Otto kaggle competition [?],[?].

## 2. DATASET

We got history from 200k enrollments, from 120k we know the labels. bla bla  
bla bla

## 3. FEATURE DESIGN

bla bla  
bla bla

## 4. SINGLE MODELS

something like this:

- -Model 1: RandomForest(R). Dataset: X
- -Model 2: Logistic Regression(scikit). Dataset: Log(X+1)
- -Model 3: Extra Trees Classifier(scikit). Dataset: Log(X+1) (but could be raw)
- -Model 4: KNeighborsClassifier(scikit). Dataset: Scale(Log(X+1) )
- -Model 5: libfm. Dataset: Sparse(X). Each feature value is a unique level.

## **5. STAGE2 ENSEMBLE**

We used xgboost [?], neural nets and linear regression for stage2 ensembling.

## **6. STAGE3 ENSEMBLE**

We tried xgboost and linear stage3 ensembling. Finally we ended up with linear ensembling with 39 courseID correction factors. These 39 factors improved the score from 0.9091 to 0.90918.

## **7. CONCLUSIONS**

Our final AUC=0.90918 score results from a complex pipeline from raw data to final score. Every part of that pipe needs to be (sub-)optimal implemented by our team to get the best score at the end. The first part “feature design” is the most important one and needs expertise, experience and of course a bit luck to capture all signals in the data.

## **8. ACKNOWLEDGEMENTS**

Thanks to dropbox, github and skype to enable easy communication around the globe.