Machine Learning Approaches to Predict Learning Outcomes in Online Courses

Charles L Green[1], Yao Ma, Dr Jiangtao Huang and Dr Jiliang Tang[2]
[1]Electrical Engineering and Computer Science, Penn State University, State College, PA, 16801
[2]Computer Science and Engineering, Michigan State University, East Lansing, MI 48223

Introduction

MOOCs have been growing at a rapid rate as of recently.

As of 2018,

•101 million people who have signed up for an MOOC

•20 million new learners in 2018 alone

•11,000+ courses

•45+ degrees & many free courses

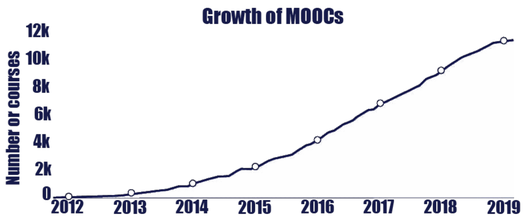

Figure 1: Growth of MOOCs chart

Downsides

- Compared to their in person counterparts
  - Lower pass rate
  - High withdraw rate

With Learning Analytics, data like demographics and interactions with the online course, in combination with Machine Learning, some possible reasons on why there are high drop rates can be figured out.

Objectives

•To identify which machine learning algorithms are best in predicting the final outcome in MOOCs

•To identify which influences/features impact the final outcomes to MOOCs the most

Definitions

- **Machine Learning (ML):** Scientific study of algorithms and statistical models used to perform specific task without explicit instructions
- **Learning Analytics:** The Collection of user data for the purpose of optimizing learning environments
- **K-fold Cross Validation:** divides the training data set k equal times, keeping same percentage of samples of each target class in each k set, and trains data with k-1 training and 1 testing.
- **Data Leakage:** when information outside the training dataset is used to create the model, can invalidate predictive models
- **Feature Importance:** found using an assortment randomized DT on various sub-samples of the dataset
- **Random-State:** used to randomize train/test data in ML algorithms in ski kit learn, can be set for consistent results
- **CSV :** Comma-Separated Values, stores tabular data in plain text

Methods

- Data
  - 32k+ users
  - 22 courses in STEM and Social Sciences
  - Demographics, Interactions and Assessments
- Organize CSVs into 1 CSV:
  - Organized so each user has an individual line of data
- Use the Jupyter Notebook opened through the Anaconda Navigator so all Libraries of code, such as Sci-Kit Learn, can be used

Model	Description	Architecture	Type	Algorithm
DT	Decision Tree	Recursive Partition Decision rules	Non Linear	CART algorithm
RF	Random Forest	Ensemble DT	Non Linear	Random Subset Features Bootstrap
SVM	Support Vector Machine	Hyperplane kernel trick	Non Linear	Quadratic Optimization
NB	Naïve Bayes	Bayesian Decision Rule	Linear	Maximum Likelihood Estimation
GBDT	Gradient Boosted Decision Trees	Ensemble DT	Non Linear	CART algorithm
KNN	K-Nearest Neighbors	Instance-based learning	Non Linear	KDTree
NN	Neural Networks	Multi-layer Perceptron	Non Linear	Backpropagation

Figure 2: Machine Learning Algorithms

- Follow directions found in Figure 3
  - Final Outcomes will be Pass, Pass with Distinction, Fail and Withdraw
  - Binary Outcomes are Pass and Not Pass

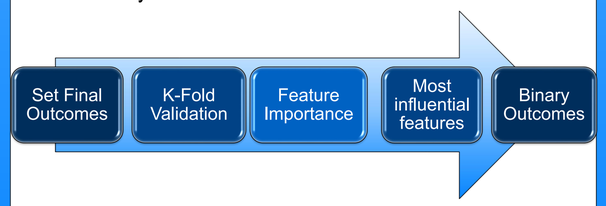

Figure 3: Machine Learning Testing Methods

Results

Alg = Algorithm
Acc = Accuracy(TP)/(TP+FN) ± 95% Confidence Interval
Eff = Efficiency in Run Time Seconds
Test 1 = All Features, Test 2 = Top 9 Features Subset

ML Alg.	Test 1 Acc.	Test 1 Eff.	Test 2 Acc.	Test 2 Eff.
GBDT	**0.753 ± 0.00455**	11.41	**0.739 ± 0.00329**	7.760
RF	0.744 ± 0.0038	**1.198**	0.732 ± 0.00278	**1.733**
DT	0.735 ± 0.00329	**0.281**	0.725 ± 0.00405	**0.223**
NN	0.71 ± 0.0053	5.877	0.703 ± 0.00683	9.955
KNN	0.716 ± 0.00354	2.413	0.713 ± 0.0038	1.225
NB	0.655 ± 0.00531	0.041	0.654 ± 0.00329	0.032
SVM	0.451 ± 0.00228	219.378	0.477 ± 0.00278	162.927

Figure 4: ML Test Results with 4 Final Outcomes

ML Alg.	Test 1 Acc.	Test 1 Eff.	Test 2 Acc.	Test 2 Eff.
GBDT	0.94 ± 0.0028	2.689	0.939 ± 0.00304	1.737
RF	0.939 ± 0.00278	**0.971**	0.937 ± 0.00304	**1.353**
DT	0.935 ± 0.00253	**0.209**	0.935 ± 0.00304	**0.168**
NN	0.932 ± 0.00278	3.249	0.929 ± 0.00304	7.787
KNN	0.932 ± 0.00304	2.316	0.934 ± 0.00304	1.17
NB	0.917 ± 0.00202	0.035	0.918 ± 0.00202	0.029
SVM	0.644 ± 0.00228	170.283	0.666 ± 0.00035	134.988

Figure 5: ML Binary Test Results with Pass/ Not Pass Final Outcomes

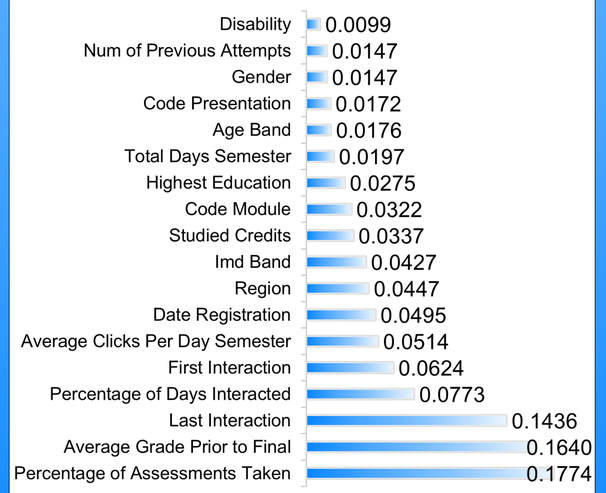

Figure 6: Feature Importance of Each Influence/Variable

Discussion/Conclusions

- GBDT had the highest accuracy but also the 2nd slowest
- RF is a close 2nd in accuracy but significantly for efficient than GBDT
- Test 2 is less accurate than test 1 while also more efficient except for the few variances to the norm
- Binary Results a significantly more accurate and is more efficient in correctly predicting the final outcome than testing with 4 outcomes
- Demographics and info known prior to starting the course hold less importance than the rest of the features

Future Studies

This information can help in future testing of MOOCs with Learning Analytics with ML by knowing which features are most important and which algorithms are best in testing these final outcomes. In future studies, people can:

- Study different datasets with more/different datasets
- Use the best algorithms with different parameters to find better combinations
- Focus on different possible influences/features to test importance
- Test for specific times a user may withdraw

References

- Al-Shabandar, Raghad, et al. "Machine Learning Approaches to Predict Learning Outcomes in Massive Open Online Courses." *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, doi:10.1109/ijcnn.2017.7965922.
- "By The Numbers: MOOCs in 2018 - Class Central." *Class Central's MOOC Report*, 21 Mar. 2019, www.classcentral.com/report/mooc-stats-2018/.
- Kuzilek, J., Hlosta, M. & Zdrahal, Z. figshare https://doi.org/10.6084/m9.figshare.5081998.v1 (2017).

Acknowledgments

I would like to thank the following people for their help and mentoring on this project. Yao Ma, Dr Huang, Dr Tang and Mr. Steven Thomas. I would also like to thank the Michigan State University Summer Research Opportunities Program (MSU SROP) and its faculty and facilitators along with the Millennium Scholar Program at the Pennsylvania State University.