

Projet Hubble

Success prediction in MOOCs *A case study*

Antoine Pigeau

DUKe, LS2N / PolytechNantes, Université de Nantes

Projet Hubble - 2015/2019

- LIP6, LIRIS, LIG, IFE, STEF, LIUM, labSTICC, LS2N
- Analysis of learner traces
- Open Classrooms (12 courses)

Success Prediction

- which indicators ?
- which model ?

Outline

Dataset

Features

Models

Experiments

Conclusion

Click stream dataset

- massive, open and online
- asynchronous
- no-stakes
- heterogeneous

Name	# users	# pass	# fail
Java	7761	34 (0%)	7727 (100%)
XML	855	10 (1%)	845 (99%)
Ionic	960	46 (1%)	914 (99%)
Rubys	149	5 (3%)	144 (97%)
Node JS	2227	81 (4%)	2146 (96%)
Arduino	2487	115 (5%)	2372 (95%)
Bootstrap	8402	727 (8%)	7675 (92%)
Audace Entr.	225	26 (12%)	199 (88%)
JavaScript	8105	1803 (22%)	6302 (78%)
Gestion Projet	1808	666 (36%)	1142 (64%)
Twitter	817	328 (40%)	489 (60%)
Web	7947	3502 (44%)	4445 (56%)

idUser	idCourse	idResource	start	end
00...43e54	2778161	17529	04-11 21:19:08	04-11 21:37:44
...

Groups

- passing group: average grade equal $\geq 70/100$
- failing group: average grade equal $\leq 70/100$ & dropout

Activity

- total duration
- duration per resource
- average session duration
- average # events in session
- ...

Regularity

- from Boroujeni et al. [10]

Inter-activity periods

- median of intertime
- duration before assessment
- # events before assessment
- # sessions before assessment
- ...

Assignment

- Marks

Baseline approaches

- Random Forest
- AdaBoost
- SVM
- logistic regression
- dense neural network

Temporal data mining

- LSTM neural network
- process mining
- sequence mining

Classification task

	25%			50%		
	A.	P.	F.	A.	P.	F.
Random forest	91%	28%	96%	93%	36%	96%
AdaBoost	91%	41%	94%	93%	48%	95%
SVM	88%	61%	90%	91%	64%	92%
Logistic Reg.	88%	66%	89%	91%	71%	91%
Dense NN	91%	37%	94%	93%	47%	95%
LSTM	90%	38%	93%	91%	47%	93%
Process mining	66%	47%	64%	72%	49%	71%
Seq. mining	24%	85%	13%	24%	88%	9%

To summarize

- best features depend on the model used for the prediction tasks
- failing and passing learners do not seem to present differences in the way they browse a course.
- the best models to detect failing and passing learners are respectively based on AdaBoost and logistic regression solutions.

Perspectives

- prediction on another courses
- model building from several courses