Predicting Dropout

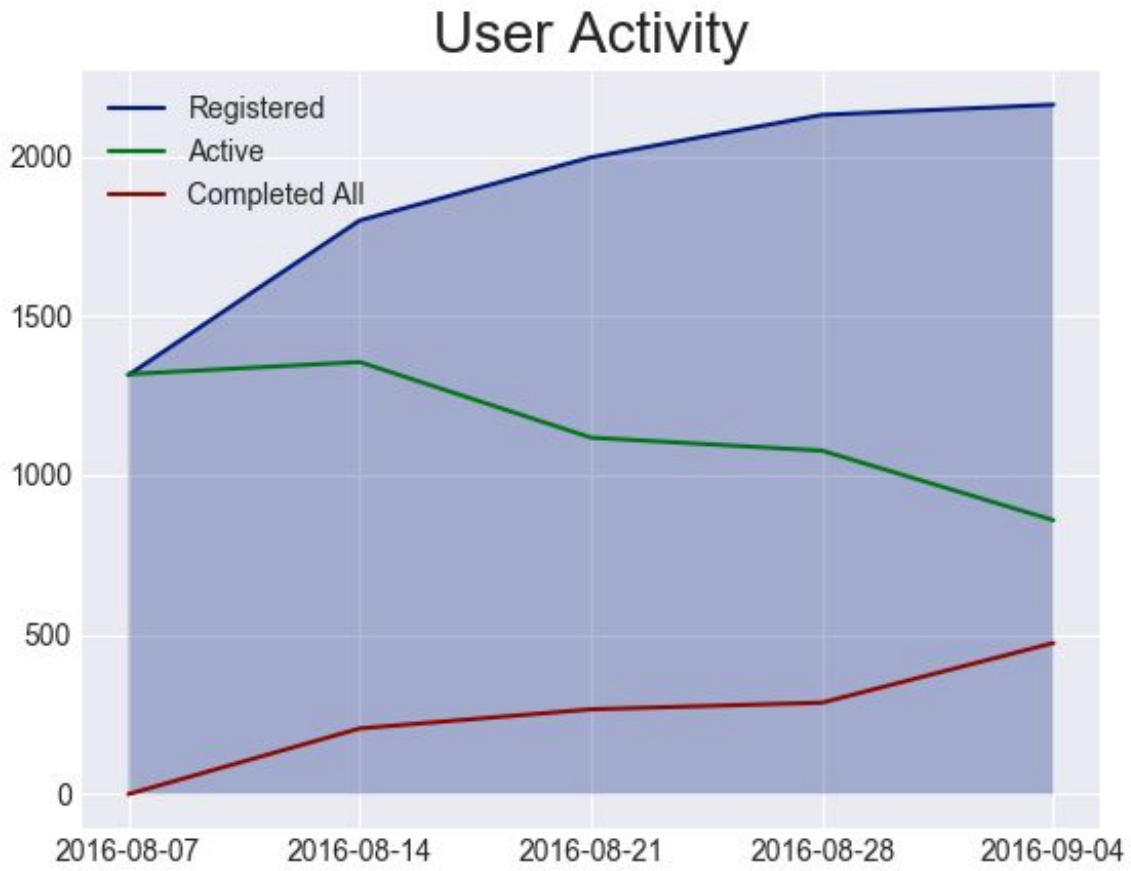
Predicting Student Dropout Risk in MOOC

by Dmytro Kovalchuk

github.com/KDmytro/moodle
linkedin.com/in/KDmytro

Business Problem

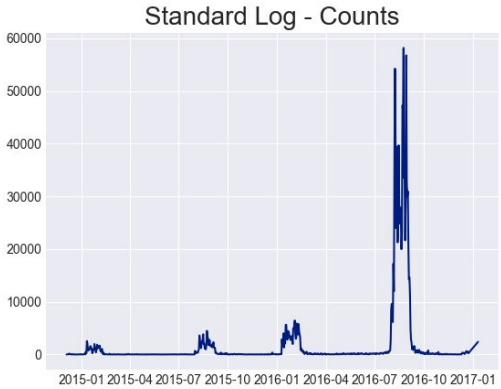
- Who is at risk of dropout?
- What are the predictive factors?
- Lack of available tools for instructors to automate intervention



Whitehill J, Williams J, Lopez G, Coleman C, Reich J. Beyond Prediction: First Steps Toward Automatic Intervention in MOOC Student Stopout, in Educational Data Mining. ; 2015.

LearnMoodle.com

- learnmoodle.com, August 2016 cohort, 4 weeks duration
- Anonymized dataset extract directly from moodle SQL DB
- Standard log file: 1.3 mil records



Feature Selection

Engagement

Total count of activities within a given timeframe

- Number of logins
- Page views
- Forum posts
- Messages sent
- Prior registrations
- Days since last login

Completions

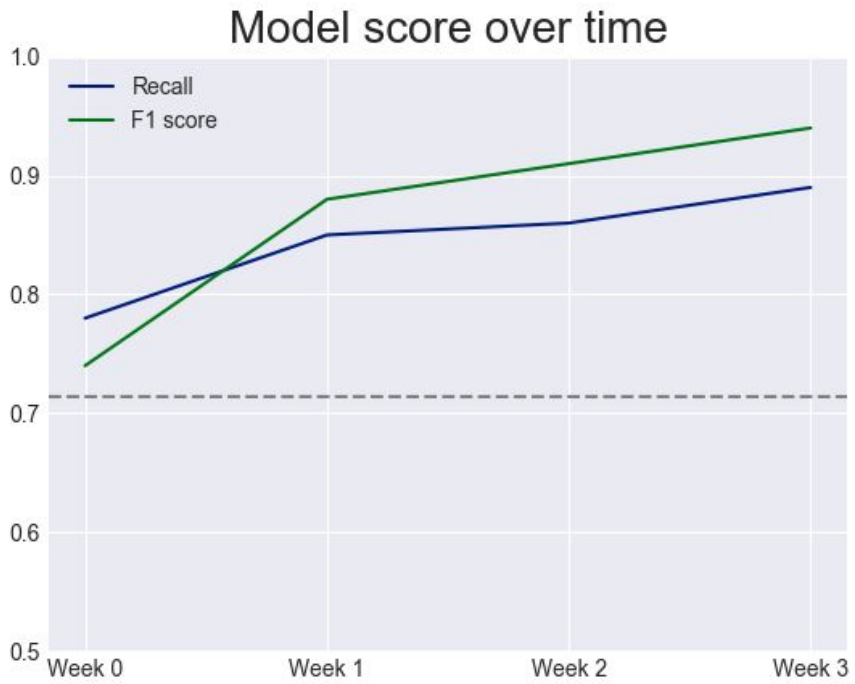
Did the user complete this week learning objects by the deadline?

- wk1_page
- wk1_forum
- wk1_quiz

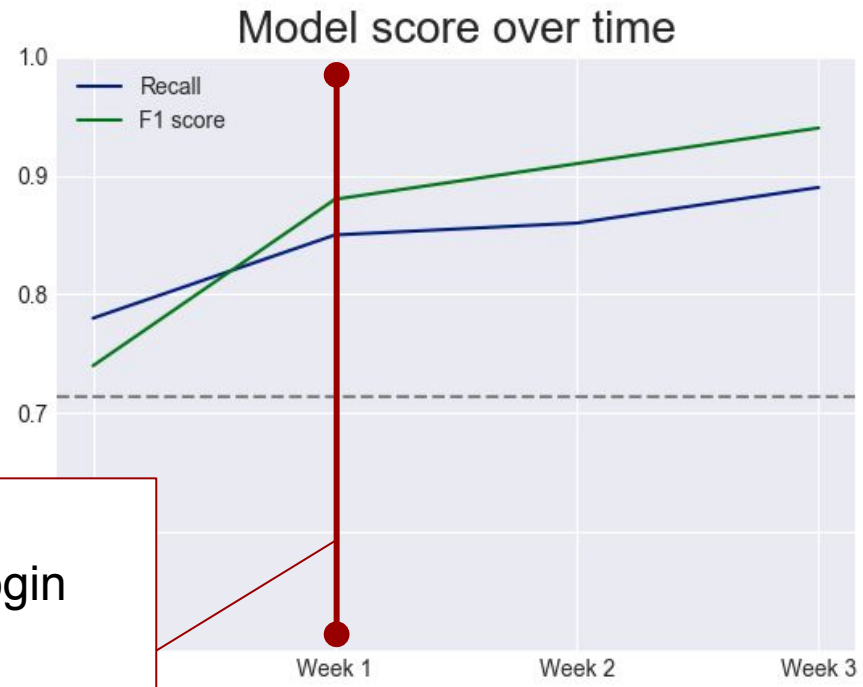
Dependent Variable

Did the user complete all required learning objects at the end of the course?

Model


Model score over time
Recall
F1 score
1.0
0.9
0.8
0.7
0.6
0.5
Week 0
Week 1
Week 2
Week 3

Model



- Page views
- Days since last login
- Wk1 - Forum
- Wk1 - Feedback

Model

- Page views
- Number of logins
- Posts created
- Wk2 - Quiz
- Wk2 - Glossary
- Wk2 - Wiki
- Wk3 - Quiz
- Wk3 - Forum

Data Pipeline

DataExplorer → FeatureFactory → ModelFactory

SQL Interface

Cleaning & Preprocessing

EDA functions

Generates Features

Dependent Variables

Trains, Stores and Loads Models

- WK1_model
- WK2_model
- WK3_model
- WK4_model

Next steps

- Collect more data from moodle:
  - To generalize model across different cohorts
  - To compare prediction accuracy across courses
- Develop and implement PHP plug-in for Moodle

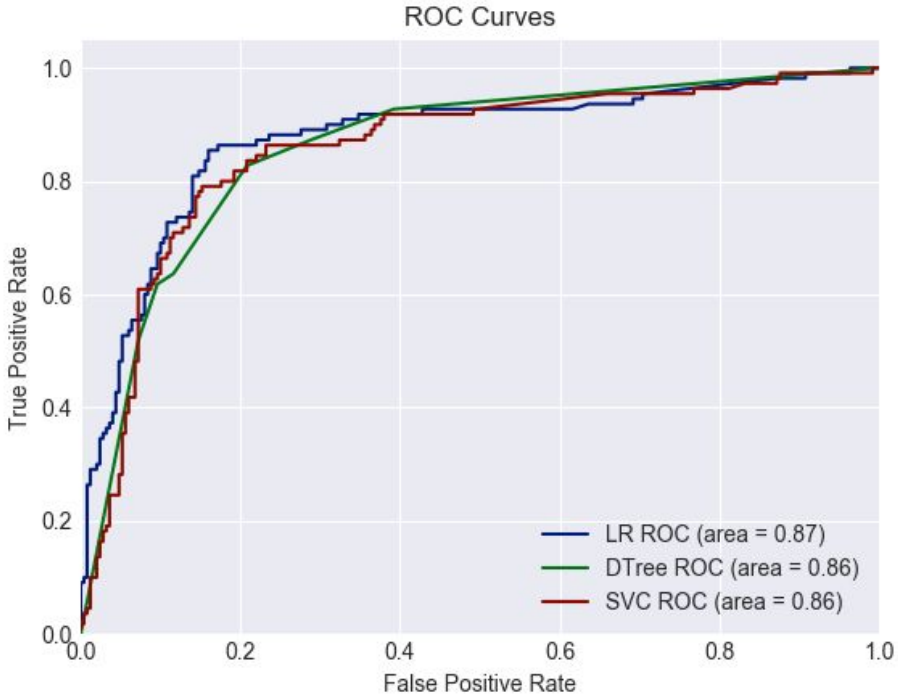
- Questions?

- Thank you

Dmytro Kovalchuk
github.com/KDmytro/moodle
linkedin.com/in/KDmytro

Appendix

Model


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LR ROC (area = 0.87)
DTree ROC (area = 0.86)
SVC ROC (area = 0.86)

Model

Week 1 Model

Logit Regression Results

Dep. Variable:	y	No. Observations:	1799
Model:	Logit	Df Residuals:	1781
Method:	MLE	Df Model:	17
Date:	Tue, 23 May 2017	Pseudo R-squ.:	0.1388
Time:	08:22:52	Log-Likelihood:	-953.76
converged:	True	LL-Null:	-1107.5
		LLR p-value:	3.110e-55

	coef	std err	z	P>|z|	[0.025	0.975]
n_logins	0.2099	0.118	1.780	0.075	-0.021	0.441
n_viewed	1.2509	0.288	4.338	0.000	0.686	1.816
n_posts_created	0.0517	0.234	0.221	0.825	-0.406	0.509
last_login	-0.2054	0.060	-3.411	0.001	-0.323	-0.087
n_msg_sent	0.1168	0.195	0.600	0.549	-0.265	0.499
n_prior_enrollment	0.0496	0.096	0.517	0.605	-0.138	0.237
0_forum	0.0471	0.084	0.563	0.574	-0.117	0.211
0_page_1	0.0184	0.113	0.164	0.870	-0.202	0.239
0_page_2	-0.1039	0.114	-0.908	0.364	-0.328	0.120
0_page_3	0.2077	0.100	2.077	0.038	0.012	0.404
1_forum	0.1802	0.080	2.259	0.024	0.024	0.337
1_book_1	0.2979	0.125	2.375	0.018	0.052	0.544
1_book_2	-0.1438	0.131	-1.094	0.274	-0.401	0.114
1_quiz	0.1897	0.098	1.935	0.053	-0.002	0.382
1_feedback	0.3316	0.095	3.490	0.000	0.145	0.518
1_choice	-0.1496	0.088	-1.698	0.089	-0.322	0.023
1_page_2	0.0212	0.107	0.199	0.842	-0.188	0.230
1_page_1	0.0067	0.077	0.087	0.931	-0.145	0.158

Model

Week 3 Model

Logit Regression Results

Dep. Variable:	y	No. Observations:	2130
Model:	Logit	Df Residuals:	2100
Method:	MLE	Df Model:	29
Date:	Tue, 23 May 2017	Pseudo R-squ.:	0.5303
Time:	08:23:12	Log-Likelihood:	-602.56
converged:	True	LL-Null:	-1282.9
		LLR p-value:	2.761e-268

	coef	std err	z	P>|z|	[0.025	0.975]
n_logins	0.6791	0.194	3.498	0.000	0.299	1.060
n_viewed	0.4724	0.206	2.291	0.022	0.068	0.877
n_posts_created	2.3156	0.721	3.210	0.001	0.902	3.729
last_login	-0.1240	0.114	-1.085	0.278	-0.348	0.100
n_msg_sent	-0.7332	0.510	-1.439	0.150	-1.732	0.265
n_prior_enrollment	0.0527	0.125	0.421	0.673	-0.192	0.298
0_forum	-0.0244	0.116	-0.210	0.834	-0.253	0.204
0_page_1	-0.0611	0.161	-0.379	0.705	-0.377	0.255
0_page_2	-0.0298	0.155	-0.193	0.847	-0.333	0.273
0_page_3	0.1626	0.118	1.381	0.167	-0.068	0.393
1_forum	-0.2203	0.115	-1.916	0.055	-0.446	0.005
1_book_1	-0.0895	0.152	-0.588	0.556	-0.388	0.209
1_book_2	-0.0495	0.174	-0.285	0.776	-0.390	0.291
1_quiz	0.0197	0.163	0.121	0.904	-0.300	0.340
1_feedback	0.2555	0.154	1.659	0.097	-0.046	0.557
1_choice	-0.0980	0.124	-0.792	0.428	-0.341	0.145
1_page_2	-0.0785	0.163	-0.483	0.629	-0.397	0.240
1_page_1	0.0022	0.101	0.022	0.983	-0.197	0.201
2_quiz	0.3487	0.146	2.388	0.017	0.063	0.635
2_book	-0.1665	0.159	-1.050	0.294	-0.477	0.144
2_glossary	0.5034	0.152	3.310	0.001	0.205	0.802
2_survey	0.0275	0.207	0.133	0.894	-0.378	0.433
2_wiki	0.4222	0.189	2.235	0.025	0.052	0.792
2_page_1	-0.1180	0.148	-0.800	0.424	-0.407	0.171
3_quiz	0.6583	0.156	4.221	0.000	0.353	0.964
3_book	-0.1175	0.160	-0.734	0.463	-0.431	0.196
3_forum	1.2086	0.227	5.313	0.000	0.763	1.654
3_data	-0.0188	0.126	-0.149	0.882	-0.266	0.228
3_page_1	0.1976	0.111	1.782	0.075	-0.020	0.415
3_page_2	0.2748	0.199	1.378	0.168	-0.116	0.666

Project Objectives

- Analyze clickstream data to identify factors impacting attrition rates
- Build predictive model to help instructors identify users at early risk of dropout
- Design data-pipeline that is suitable for future product integration (Moodle Plug-in)