Data Report 1

Course Feature Selection

Zachery McKinnon - 8/15/2017

SUMMARY

part1

The aim of this analysis is to assess the quality of courses and determine which features of courses are the best predictors of course quality through both parametric (OLS) and non-parametric (support vector machine) means. This information can be used by the education team to identify which dimensions of courses to consider, by the communications team to determine which courses to promote, and by the tech team to determine which features to improve upon and/or focus on. This analysis is an important first, not last, step to analyzing courses and their features. The methods presented herein can be improved upon with additional data (both new samples and new dimensions, particularly new input variables) or complemented with new models, so I welcome any input for improvement.

part2

After soliciting input from the Saylor team regarding the input variables and results from the first analysis, new variables were added. The inclusion of these variables in a regression introduces considerable collinearity issues; therefore, principal components regression and LASSO regression are used to reduce the dimensionality of our final models, which focus specifically on completions.

OUTPUT VARIABLES

part1

Course quality is a rather arbitrary concept and in fact probably cannot be captured by 1 single dimension. Therefore, I chose to define 3 output variables for 3 separate regressions:

completions	number of completions over the 6-month period of 1/1/2017 to 6/30/2017
pass_rate	number of students passing the final exam from 1/1/2017 to 6/30/2017 divided by the number of students attempting the final exam from 1/1/2017 to 6/30/2017
comp_rate	number of students completing the course from 1/1/2017 to 6/30/2017 divided by the number of students enrolling in the course from 10/1/2016 to 3/31/2017

The variable *completions* measures the overall popularity of the course, which of course is also largely influenced by the subject matter. The variable *pass_rate* measures the success of students learning the material and being prepared for the final. Lastly, the variable *comp_rate* measures students' retention and continued engagement in a course. Beyond these three independent variables, we also want to determine courses that succeed in multiple dimensions. To that end, I've created the following variable.

course_rating	3 if the course beats the averages for completions, pass_rate, and comp_rate
	2 if the course beats the averages for 2 of the three above variables,
	1 if the course beats the averages for 1 of the three above variables, and
	0 otherwise

It should first be noted that these variables are more indicators of student buy-in than they are indicators of course quality alone. That is, students may be predisposed to enroll in, engage in, and complete a course of a certain subject matter or type of subject (e.g., STEM), which really would not depend on how or what course content is presented. However, for the purposes of providing courses with which students will engage effectively, this difference is not necessarily important. That is, both improving course design and providing content that is inherently more popular are equally appealing low-hanging fruit for Saylor.

part2

Given the better performance of the models modelling completions and the greater importance of completions as an organizational goal, the second part focuses exclusively on that output variable.

INPUT VARIABLES

part 1

The following variables were selected as input features for the analyses:

course_length	sum of the approximated length (hours) of the units of the course
units	number of units in the course
study_guide	1 if the course has a study guide/review/summary and 0 otherwise
enrollment	enrollment over the 6-month period of 10/1/2016 to 3/31/2017
0_level	1 if the course is 0 level and 0 otherwise
200_level	1 if the course is 200 level and 0 otherwise
300_level	1 if the course is 300 level and 0 otherwise
400plus_level	1 if the course is 400+ level and 0 otherwise
STEM	1 if the course is STEM (Science, Technology, Engineering, and Mathematics) and 0 otherwise
for_credit	1 if the course has a for-credit option and 0 otherwise
1000degrees_2016	number (in 1000s) of degrees conferred for that major in 2016
computer	1 if the course focuses on a computer-based skill and 0 otherwise
iTunesU	1 if the course was featured on iTunesU and 0 otherwise
fb_campaign	1 if the course was the subject of a facebook promotion and 0 otherwise
youtube	1 if the Saylor.org youtube page has any information related to the course and 0 otherwise
modules	number of modules in the course
intercept	1 for all courses to serve as the intercept for the regression model

part2

The following input variables were used in addition to those in Part 1:

salary_report	average starting salary for the corresponding major
jetblue	1 if the course is part of the partnership with jetblue and 0 otherwise
TESC	1 if the course is part of the partnership with TESC and 0 otherwise
brandman	1 if the course is part of the partnership with brandman and 0 otherwise
memphis	1 if the course is part of the partnership with memphis and 0 otherwise
SIMCA	1 if the course is part of the partnership with SIMCA and 0 otherwise
adamjee	1 if the course is part of the partnership with adamjee and 0 otherwise
city_vision	1 if the course is part of the partnership with city vision and 0 otherwise
course_outcomes	number of course outcomes
unit_outcomes	number of unit outcomes
percent_open	percent of content that is OER
base_views	pageviews for course during 1/1/2016 to 3/31/2016, when there were few/no campaigns

SAMPLE SELECTION

part1

After consultation with Saylor's education team, the K12 and PRDV subjects were considered for exclusion from the sample as they differed notably from the other subjects. Given the small sample size to begin with, the analyses were run with and without these courses. The results with these courses excluded are shown in brackets. The full sample consists of 94 courses, and the subsample consists of 84 courses.

part2

Given the consistently better results with the subsample excluding K12 and PRDV courses, the analysis in part 2 considers just that subsample.

METHODS AND RESULTS

part1

The first step of the analysis is to get a sense of the strength of the (linear) relation between the input variables and output variables with Pearson's r. Below are the Pearson's r values between completions, pass_rate, and comp_rate and all of the input variables, as well as the VIFs, variance inflation factors, of the input variables..

Pearson's r

	completions	**pass_rate**	**comp_rate**	
completions	1.000 [1.000]	0.635 [0.636]	0.843 [0.851]	
pass_rate	0.635 [0.636]	1.000 [1.000]	0.727 [0.676]	
comp_rate	0.843 [0.851]	0.727 [0.676]	1.000 [1.000]	**VIF**
course_length	-0.294 [-0.291]	-0.345 [-0.191]	-0.348 [-0.206]	2.06 [1.32]
units	-0.112 [-0.062]	-0.132 [0.009]	-0.122 [0.025]	2.11 [2.05]
study_guide	0.247 [0.270]	0.009 [0.047]	0.119 [0.172]	1.44 [1.71]
enrollment	0.366 [0.369]	0.186 [0.209]	0.079 [0.082]	1.80 [2.11]
0_level	-0.033 [-0.043]	-0.053 [-0.124]	-0.040 [-0.135]	1.58 [1.55]
200_level	0.119 [0.150]	0.086 [0.159]	0.038 [0.101]	1.92 [1.99]
300_level	0.050 [0.063]	-0.072 [-0.050]	0.097 [0.143]	1.82 [1.92]
400plus_level	-0.069 [-0.057]	-0.037 [-0.004]	0.059 [0.107]	2.34 [2.50]
STEM	-0.163 [-0.129]	-0.163 [-0.075]	-0.227 [-0.155]	2.32 [2.77]
for_credit	0.291 [0.324]	0.108 [0.184]	0.047 [0.106]	1.96 [1.98]
1000degrees_2016	0.277 [0.315]	0.016 [0.083]	0.140 [0.223]	2.02 [2.02]
computer	-0.078 [-0.074]	-0.018 [-0.039]	-0.026 [-0.033]	2.78 [4.01]
iTunesU	0.229 [0.209]	0.193 [0.119]	0.201 [0.113]	1.58 [1.65]
fb_campaign	-0.035 [-0.040]	0.058 [0.033]	-0.086 [-0.117]	1.25 [1.38]
youtube	0.278 [0.290]	0.193 [0.147]	0.223 [0.199]	1.25 [1.25]
modules	-0.100 [-0.056]	-0.038 [0.111]	-0.177 [-0.065]	2.20 [2.25]

Notes: VIFs found using Python's statsmodels package.

For Pearson's r, a negative sign suggests an inverse relation, and a positive sign suggests a positive relationship. Furthermore, a good rule of thumb in the social sciences is that an absolute correlation of 0.9 to 1 is very high and in fact very unlikely, 0.7 to 0.9 is high, 0.4 to 0.7 is moderate, 0.2 to 0.4 is low, and 0.0 to 0.2 is negligible. All of the features' Pearson's r are in the range of negligible to low, which suggests that no singular feature will strongly linearly predict any dimension of course quality. In addition, though not pictured, we can also observe the correlation between input features from Pearson's r, which gives an idea of multicollinearity. None of the features are highly correlated, so this issue is preliminarily assuaged. The variance inflation factors confirm this observation, as all of the VIFs are less than 5 and in fact

relatively low. Because of the low multicollinearity, we can obtain reasonably stable coefficient estimates from an OLS regression including all variables, the results of which are shown below.

OLS Results - All Variables

	completions	**pass_rate**	**comp_rate**
course_length	-0.89*** [-0.73***]	-0.00*** [-0.00*]	-0.00*** [-0.00*]
units	0.15 [1.82]	-0.01 [-0.00]	0.00 [0.01]
study_guide	37.00* [54.24**]	0.01 [0.05]	0.05 [0.10*]
enrollment	0.13*** [0.17***]	0.00 [0.00*]	0.00 [0.00]
0_level	-9.10 [-38.11]	-0.09 [-0.12]	-0.04 [-0.09]
200_level	43.74* [54.10**]	-0.04 [0.08]	0.06 [0.11*]
300_level	58.27** [68.38***]	0.03 [0.06]	0.12** [0.16**]
400plus_level	65.21** [76.82***]	0.08 [0.11*]	0.16** [0.20***]
STEM	28.44 [32.25]	-0.01 [0.01]	-0.02 [0.02]
for_credit	34.60* [45.07**]	0.03 [0.05]	0.05 [0.07]
1000degrees_2016	-0.00 [-0.02]	-0.00 [-0.00]	-0.00 [-0.00]
computer	-52.90** [-62.74*]	-0.01 [-0.04]	-0.07 [-0.12]
iTunesU	-7.34 [-29.56]	0.03 [-0.02]	0.05 [-0.02]
fb_campaign	-19.00 [-37.29*]	0.01 [-0.02]	-0.04 [-0.07]
youtube	57.82* [56.50*]	0.10 [0.06]	0.10 [0.08]
modules	-0.08 [-0.04]	0.00 [0.00]	-0.00 [-0.00]
r-squared	**0.437 [0.474]**	**0.249 [0.209]**	**0.300 [0.298]**
Adj. r-squared	**0.317 [0.347]**	**0.089 [0.018]**	**0.150 [0.128]**

Notes: *, **, and *** indicate significance at 0.10, 0.05, and 0.01, respectively. Analyses conducted with Python's statsmodels package.

The small values of r-squared and the large difference between the r-squared and adjusted r-squared values indicate that the input variables do not strongly predict our measures of course quality in a parametric, linear fashion and that too many variables are being used to obtain a strong model overall, respectively. Therefore, I use a greedy algorithm combined with SVR (kernel = 'linear'), a non-parametric regression, to rank features in terms of importance. The results will then be used in an SVC, support vector classifier, to create a model predicting course quality.

Ranking of All Variables from Greedy Algorithm SVR

	completions	**pass_rate**	**comp_rate**	**course_rating**
course_length	10 [12]	13 [12]	13 [12]	13 [13]
units	13 [7]	11 [10]	11 [9]	12 [14]
study_guide	7 [5]	4 [8]	6 [7]	4 [2]
enrollment	14 [15]	16 [15]	16 [16]	15 [15]
0_level	5 [9]	6 [2]	7 [14]	3 [5]
200_level	6 [8]	12 [5]	9 [5]	8 [1]
300_level	11 [14]	7 [9]	12 [1]	10 [8]
400plus_level	9 [4]	10 [6]	3 [4]	11 [7]
STEM	4 [6]	3 [14]	1 [3]	5 [9]
for_credit	2 [1]	9 [3]	10 [6]	7[3]
1000degrees_2016	16 [16]	15 [16]	14 [15]	16 [16]
computer	15 [11]	2 [4]	5 [11]	6 [10]
iTunesU	1 [2]	5 [11]	4 [10]	2 [11]
fb_campaign	3 [3]	8 [7]	8 [8]	1 [6]
youtube	8 [10]	1 [1]	2 [2]	9 [4]
modules	12 [13]	14 [13]	15 [13]	14 [12]
intercept	17 [17]	17 [17]	17 [17]	17 [17]

There are not enough samples and generally too much variance for an ordinal regression to yield satisfactory results in a predictive sense. Therefore, as a last step, I selected the two output variables that are the most uncorrelated with each other (i.e., completions and pass_rate: 0.635) and created a binary classifier based on the following label:

course_classifier	1 if the course beats the median for completions and pass_rate and 0 otherwise

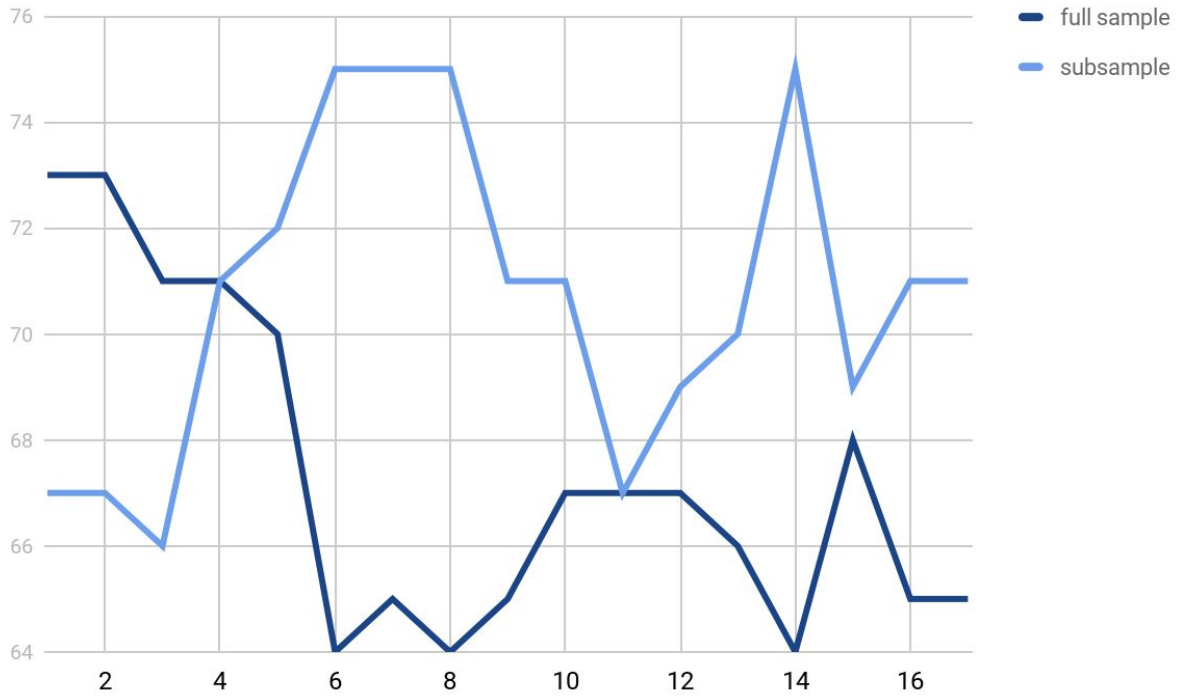
The median was chosen over the mean in this instance because it better balances the 'good' and 'bad' courses, which lessens the problem of majority class prediction. There were 35 good courses (see appendix) and 59 bad courses (a 37%-63% split).

Ranking of All Variables from Greedy Algorithm SVC

	full sample	**subsample**
course_length	10	15
units	8	12
study_guide	6	5
enrollment	16	14
0_level	7	2
200_level	14	8
300_level	2	3
400plus_level	9	7
STEM	5	10
for_credit	3	1
1000degrees_2016	15	16
computer	4	9
iTunesU	1	4
fb_campaign	13	11
youtube	11	6
modules	12	13
intercept	17	17

Following this list of variables ranked by importance, I retrained the classifier with progressively fewer input variables to observe the change in accuracy with variable inclusion. The results are presented below.

Change in Accuracy with Variable Inclusion



In general, the model based on the subsample is a better predictor of course quality, which confirms the education team's assertion that PRDV and K12 are outliers. For the subsample specifically, the top 8 features are the best predictors of course quality (in the name of parsimony). These 8 features are for_credit, 0_level, 300_level, iTunesU, study_guide, youtube, 400plus_level, and 200_level. For the full sample, the first two features appear to be more important than the rest. These two features are iTunesU and 300_level, which indicates that the use of iTunesU may be effective across all courses, including PRDV and K12.

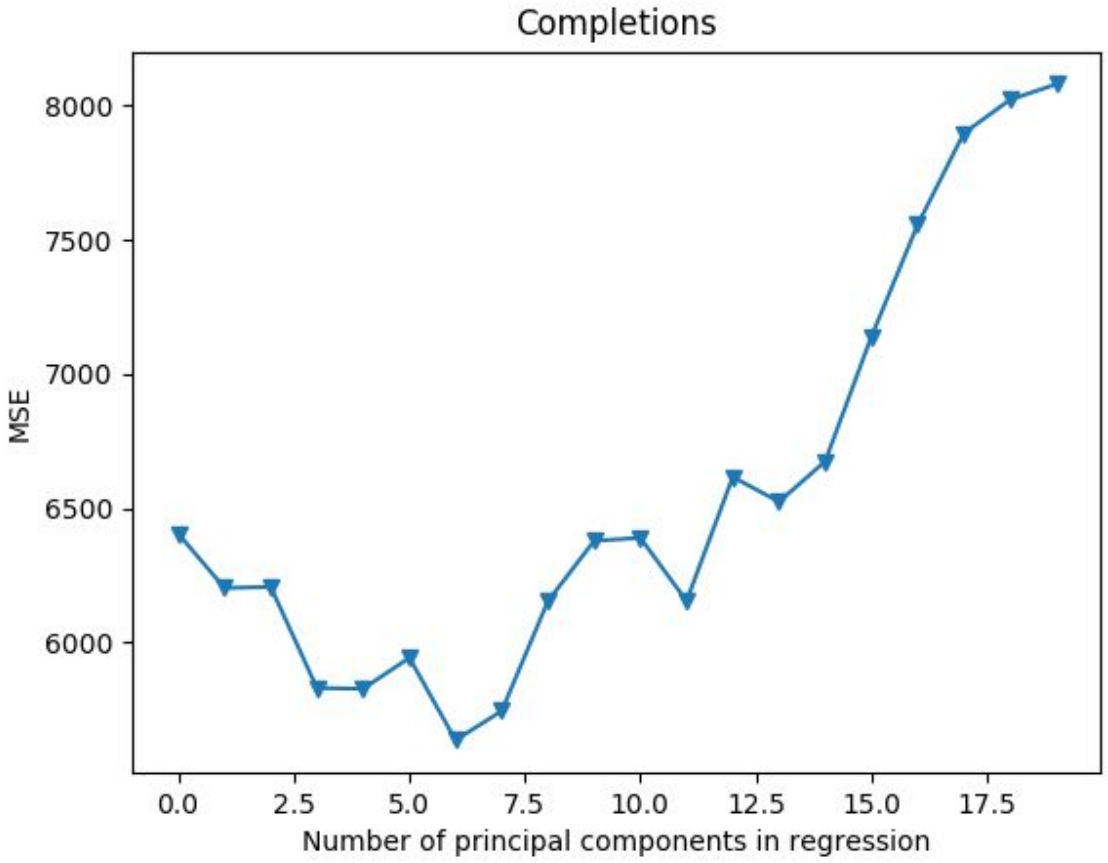
part2

Again, it is useful to look at VIFs to gain an understanding of collinearity issues associated with the input variables:

VIFs of Input Variables

Variable	**VIF**
course_length	2.33
units	2.99
study_guide	2.45
enrollment	12.23
0_level	4.19
200_level	2.49
300_level	4.71
400plus_level	5.84
STEM	4.53
for_credit	4.81
1000degrees_2016	15.77
computer	14.31
iTunesU	1.94
fb_campaign	2.05
youtube	2.16
modules	3.03
salary_report	12.24
jetblue	4.26
TESC	2.31
brandman	5.97
memphis	3.14
SIMCA	1.99
adamjee	2.5
city_vision	11.87
course_outcomes	1.65
unit_outcomes	2.19
percent_open	2.71
base_views	13.3

The large VIFs on a number of factors indicate that collinearity would be a serious issue using OLS. Therefore, I will adopt an approach to reducing dimensionality, i.e., principal components regression (PCR). PCR is a combination of principal components analysis (PCA) and regression analysis. PCA creates orthogonal (uncorrelated) components based on the variance in the data to establish what are essentially "super-variables," i.e., components containing loadings of (normalized) input features. The first task is to determine the optimal number of components. I did so by looking at when the average MSE based on a 5-fold cross validation is at a minimum with the inclusion of progressively more components using Python's sklearn and matplotlib.



*Notes: the average MSE decreases again to its global minimum as the number of components become 20+. However, at that amount, PCA is doing little to decrease dimensionality and would thus be unuseful.

The results indicate that 6 is the ideal number. From here, we can look at the factor loadings of the first 6 components. Below are the loadings above .25 for the first 6 components. The higher the absolute loading, the more that variable is part of that component. I provided an explanation at the bottom of the table of what each component means and its coefficient.

Factor Loadings of First 6 Principal Components

components	1	2	3	4	5	6
course_length						-0.30
units			-0.27	0.45		
study_guide			0.26			
enrollment		0.39			0.23	
0_level						
200_level					-0.36	
300_level						0.52
400plus_level					0.29	-0.40
STEM		0.34				
for_credit	0.25	0.28				
1000degrees_2016				0.41		
computer	-0.31	0.28				
iTunesU						
fb_campaign		0.25			0.34	
youtube						0.39
modules			-0.35	0.34		
salary_report	-0.30	0.30				
jetblue	0.31					
TESC	0.25					
brandman			0.35			
memphis		-0.29			0.45	
SIMCA						
adamjee						
city_vision			0.27	0.25		
course_outcomes					-0.30	
unit_outcomes			-0.36			
percent_open	0.28					
base_views		0.35			0.28	
description	for-credit, low-tech, OER classes	for-credit, high-enroll, tech classes	short courses with study guides	long classes with popular majors	unclear	short classes with youtube content
ideal partner	JetBlue, TESC		Brandman, CityVision	CityVision		
coefficient	7.21	3.57	14.32	4.46	-1.53	19.5

Another way to reduce dimensionality is with LASSO regression. LASSO achieves subset selection based on a value alpha. As alpha increases, more and more variables receive coefficients of 0. In other words, as alpha increases, the overall model becomes less predictive, but the coefficient estimates on the remaining variables become more reliable.

Coefficients of Input Variables

Variable	**Coefficient (alpha = 0)**	**Coefficient (alpha = 1)**	**Coefficient (alpha = 2)**
course_length	-0.21	-0.26	-0.31
units	3.44		-0.12
study_guide	43.97	21.58	14.11
enrollment	0.55	0.54	0.52
0_level	-60.36		
200_level	58.96	17.01	0.73
300_level	62.04	13.42	
400plus_level	47.93		-1.48
STEM	30.70	8.95	
for_credit	-15.47		
1000degrees_2016	-0.23	0.14	0.16
computer	-110.88	-30.59	
iTunesU	-7.53	-4.14	
fb_campaign	24.65	8.32	
youtube	62.73	8.21	
modules	-0.09	-0.01	0.01
salary_report	0.00		
jetblue	-33.94	-36.90	-31.56
TESC	-27.10	-16.55	-7.97
brandman	1.12		
memphis	-26.35		
SIMCA	7.14		
adamjee	-21.71	-0.27	
city_vision	83.17		
course_outcomes	-1.82	-1.43	-1.27
unit_outcomes	-0.28	-0.27	-0.18
percent_open	0.39	0.18	0.13
base_views	-0.05	-0.06	-0.06
r-squared	**0.77**	**0.71**	**0.66**

CONCLUSION

The following are the most salient findings from the analysis:

1) Study guides

Part 1: Study guides predict course completions and, to a lesser extent, the quality of the courses overall. This finding may indicate that people use Saylor as a way to refresh their knowledge on subjects with which they are already familiar and then seek out credentials/credit, rather than use Saylor courses to learn a topic from scratch.
Part 2: Study guides are the only aspect that reflected positively and significantly in all analyses. Therefore, they should be considered a high priority/necessity in courses going forward.

2) Credit Offering

Part 1: Providing credit for courses positively predicts all aspects of course quality (completions, pass_rate, and comp_rate).
Part 2: For credit is best applied to two types of courses: non-STEM courses with high amounts of OER (working with a partner) and, to a lesser extent, computer/STEM courses (no partner needed).

3) Alternative media and campaigns

Part 1: YouTube videos are positive predictors of course success. iTunesU appears to have a murkier, though mostly positive, relationship. More time and courses for iTunesU, as well as fb campaigns, are needed to assess their effectiveness.
Part 2: YouTube videos with short courses seem to be a recipe for great success (largest coefficient in PCR). I would recommend we work on creating YouTube videos for the short courses that we already have as well as create additional small YouTube-based courses, able to be completed in a much shorter time period than traditional courses. Since Saylor is concerned not just with course completions but also with skills gained, looking at these types of courses would also be a great way of quantifying our impact in a different way.

4) Enrollment

Part 1: Enrollment isn't generally important for course quality. I found this result rather surprising as enrollment is not just an indicator of face-value course appeal but also a proxy for the discussion community surrounding courses. The fact that it did not have high importance might suggest that the discussion forums are being underutilized or improperly utilized.
Part 2: The same conclusion as above applies. This to me indicates that focusing on the peer-to-peer support aspect of Saylor is much needed because there is no reason this should not be highly positive.

5) Popularity

Part 1: Surprisingly, course popularity, as proxied by the number of degrees conferred in 2016, is consistently an unimportant determinant of course quality.
Part 2: The relationship is more nuanced than as suggested in Part 1. Popular majors do in fact matter when considering longer courses designed to mimic traditional higher education, particularly when we have partnership programs already lined up. And in fact, users are most willing to tolerate longer courses when the corresponding major is a popular one.

6) Partnerships

Part 2: City Vision appears to be our most successful partnership in the first half of 2017, so we should continue to foster that relationship. We should also solicit more information from our partners to understand how they're recommending our content to students and whether we can change that for the better. Some partnerships have negative coefficients, and large ones at that, which would suggest that they're not using our most effective content.

7) Saylor's role

Part 1: Saylor appears to best serve as a complement to tech-based MOOCs, rather than as a substitute. This is evidenced by the negative and/or insignificant correlations and coefficients for the variable *computer*.

Part 2: Saylor's best role depends on the area being considered. Instead of thinking about our role overall, we should think about what areas we succeed at, as shown in the PCR table.

8) Course Level

Part 1: 200+ level courses positively predict course quality, whereas 0 level courses (particularly excluding PRDV) negatively predict course quality.

Part 2: The results are not conclusive enough to make concrete recommendations.

It's important to note that these results do not imply causality. There may be confounding variables that complicate the relationships. In addition, if an avenue is not important or significant, that does not necessarily mean it is not worth pursuing. It simply means large changes are likely needed for that avenue to become effective. For example, that the variable fb_campaign was not a selected as an important feature does not mean that campaigns are not entirely worth pursuing. After becoming familiar with the page views post-campaign, it's clear that there are some bottlenecks preventing the frictionless transformation of page clicks into enrollments and, ultimately, completions. In this case, further A/B tests of the landing pages, rather than changing course with the campaigns themselves, would be worthwhile.

Looking forward, there are a number of additional features that could improve this model. For example, I would like to add partnerships (e.g., courses heavily promoted by Memphis' Finish Line) to a model such as this. Therefore, I will end this report with a call to submit additional ideas for variables of interest.

APPENDIX

List of good courses in the SVC

- ARTH101: Art Appreciation and Techniques
- BIO101: Introduction to Molecular and Cellular Biology
- BUS203: Principles of Marketing
- BUS205: Business Law and Ethics
- BUS206: Management Information Systems
- BUS208: Principles of Management
- BUS210: Corporate Communication
- BUS303: Strategic Information Technology
- BUS401: Management Leadership
- CHEM101: General Chemistry
- COMM411: Public Relations
- CS101: Introduction to Computer Science I
- CS305: Web Development
- CS402: Computer Communications and Networks
- CS403: Introduction to Modern Database Systems
- CS405: Artificial Intelligence
- CS410: Advanced Databases
- ECON101: Principles of Microeconomics
- ECON102: Principles of Macroeconomics
- HIST103: World History in the Early Modern and Modern Eras
- MA001: College Algebra
- RWM101: Foundations of Real World Math
- RWM102: Algebra
- ENVS203: Environmental Ethics, Justice, and World Views
- POLSC101: Introduction to Political Science
- POLSC221: Introduction to Comparative Politics
- CUST105: Customer Service
- PRDV002: Professional Writing
- PRDV003: Word Processing Using Microsoft Word
- PRDV004: Spreadsheets
- PRDV005: Time and Stress Management
- PRDV102: Resume Writing
- PRDV103: Interviewing Skills
- PRDV104: Professional Etiquette
- SOC101: Introduction to Sociology