Lisa Everest, Grace Shin, Christine Vonder Haar, Kathleen Zhu
15.0711 The Analytics Edge Section A
Professor Robert Freund
Spring 2019

Investigating Business Analytics Degree and Advanced Algorithms Enrollment

Introduction

Our project explores the MIT undergraduate degree in two parts.

1. We first analyze the 15-2 Business Analytics undergraduate degree and course load. In this part, we optimize course load by maximizing the total utility a student receives from their degree. We then perform a sensitivity analysis on the objective function and the constraints and present our results and conclusions from our models.
2. We then focus on predicting enrollment for the course Design and Analysis of Algorithms (6.046), an undergraduate class required for an EECS degree. In this part, we utilize different models to predict 6.046 future enrollment based on data from the past 15 semesters. We compare and contrast the results from these various prediction models.

Part 1: Optimizing the 15-2 Course Road

Motivation:
Since course 15 has diverged into three different majors, many undergraduate students have declared 15-2 (Business Analytics). 15-2 offers a diverse and flexible range of classes to complete the degree. Though this range of classes gives students the benefit of choice to curate a more personalized major, it can also make it difficult for underclassmen to determine which electives to take and when to take them. Thus, we were motivated to create a model that optimizes the course load for underclassmen to enhance the undergraduate experience.

The results of this optimization model would allow undergraduates to better organize their course load. by finishing their course requirements by taking a specific number of classes per semester and only classes with the highest ratings, for example.

Data:
Our dataset contains all the classes that are pertinent to this model (core classes, pre-requisites, GIRs, HASS[1], electives), the units corresponding to the classes, when the classes are offered, the pre-requisites corresponding to the classes, whether the class is a GIR, whether the class is a core requirement for the major, and the type of elective of the class[2]. We were able to compile centralized data regarding number of hours, ratings from firehose.

Model:
Our model maximizes the utility of a course load, which we define as the sum of the ratings[3] of all courses taken. We then had to take into account the following constraints, based on the general MIT graduation requirements and requirements for a 15-2 degree:[4]

1. All possible elective classes
2. Core classes required by 15-2 major

---

[1] We have simplified the HASS GIR requirement and assumed all HASS classes are 12 units and require 12 hours per week. To allow for taking multiple HASS classes per semester, we've created 8 fake HASS classes with these specifications. Examples include 21.100, 21.200,...,etc.
[2] Please refer to appendix figure A for a more detailed specification of our variables in the dataset.
[3] Ratings for each course are calculated as an average of the past semesters of the course, at the time of this report.
[4] Graduation requirements taken from http://catalog.mit.edu/mit/undergraduate-education/general-institute-requirements and http://catalog.mit.edu/degree-charts/business-analytics-course-15-2

3. Credit limit respective to the year (for example: freshmen have a 54 credit limit)
4. Availability of classes (for example: fall, spring, both)
5. Pre-requisites

Please refer to Appendix Figure B for the mathematical formula of our model, which includes decision variables, notation, and constraints.

Baseline Model Results and Discussion:

In our baseline model, we restricted students to taking at most 4 classes/semester, 60 credits/semester, and 64 hours of weekly work/semester. This optimization model then produces the course load shown in Table 1, which gives students a utility of 170.72 and allows a student to graduate successfully within 8 semesters.

Semester	Courses
Semester 1	21.3, 8.01, 18.01, 7.012
Semester 2	5.111, 8.02, 18.02, 15.276
Semester 3	21.4, 21.5, 21.6, 15.874
Semester 4	21.1, 21.7
Semester 5	6.0001, 15.312, 15.0791
Semester 6	6.01, 15.075, 15.7611, 6.05
Semester 7	21.8, 15.780, 15.772, 6.034
Semester 8	21.2, 6.036, 15.053

Table 1: Utility-Maximizing Baseline Model Output

Based on personal experience as undergraduates, we find that Table 1's optimal course load seems reasonable because some semesters allow for less than 4 classes and almost all semesters contain a diverse mixture of classes. It is also interesting to note that, though the 15-2 degree requires at least 25 courses to be taken, our baseline model for maximizing total utility outputs a course load of 28 classes. This is because we maximize *total* utility, so taking more classes improves this objective; we could have maximized average utility per class and may have received different results. Furthermore, since our baseline results include 28 course and since the maximum utility possible from 28 courses is 7 x 28 = 196[5], our baseline model has therefore achieved 87% of the maximum utility possible.

Sensitivity Analysis: Optimizing Utility

We first performed a sensitivity analysis on our model's constraints, and we ran our utility-maximizing model with 8 different combinations of restrictions on the classes/semester, credits/semester, and hours of weekly work/semester. We represent these restrictions as a triplet (maxClasses, maxCredits, maxWorkload), respectively, and present the results in Table 2, with full results in Appendix B.

Most interestingly, perturbing these three values didn't affect the total utility received from the model's optimal course load, and all 8 models produced a utility of 170.2, the same as the baseline. Furthermore, we found that relaxing the constraints on maxClasses, seemed to allow a student to graduate in 7 semesters, rather than the 8

[5] The maximum possible rating of a course is 7.

semester in the baseline. Thus, we could infer from this that perturbations in the three values discussed here most likely do not affect *which* classes are taken, but rather *when* classes are taken. Finally, we note that if a student takes at most 3 classes/semester, that student will not graduate within 8 semesters, since the model was infeasible.

Model #	(maxClasses, maxCredits, maxWorkload)	Ratings-Maximizing Model		Utility-Maximizing Model	
		Total Utility	Semesters Needed	Total Utility	Semesters Needed
0 (Baseline)	(4, 60, 64)	170.2	8	145.3	7
1	(5, 60, 64)	170.2	8	145.7	5
2	(6, 60, 64)	170.2	8	145.95	5
3	(3, 60, 64)	N/A	N/A	N/A	N/A
4	(5, 72, 64)	170.2	7	145.7	5
5	(5, 64, 72)	170.2	7	145.7	5
6	(5, 72, 72)	170.2	7	145.7	5
7	(6, 72, 72)	170.2	7	145.0	5
8	(7, 72, 72)	170.2	8	156.2	4

Table 2: Sensitivity Analysis on Utility-Maximizing

Sensitivity Analysis: Optimizing Semesters

Our second sensitivity analysis was performed on our model's objective function; instead of maximizing the student utility, we instead minimized the number of semesters a student needs to finish the 15-2 degree. The mathematical formulation of this semester-minimizing model may be found in Appendix Figure B. We find that, by constraining a student to 4 classes/semester, 60 credits/semester, and 64 hours of weekly work/semester, the student has a decreased utility of 145.3 as compared to the utility of 170.2 in the rating-maximizing baseline model (with the same constraints). As expected, however, the semester-minimizing model allows a student to graduate in 7 semesters, which is less than the 8 semesters produced in the baseline.

With this same objective function that minimizes the number of semesters necessary for degree completion, we then perturbed the constraints with the same 8 combinations of class, credits, weekly work, as in Sensitivity Analysis 1. Results are included in Table 2.

Most notably, we find that relaxing the constraint on a student's maximum number of classes/semester significantly reduces the number of semesters a student needs to graduate. However, relaxing just the maximum credits/semester without relaxing the weekly hours of work/semester didn't affect the results; an increase in *both* the maximum credits/semester and weekly hours of work/semester were necessary in order to allow a student to reduce the number of semesters to graduate. This is likely due to the fact that credits and weekly hours of work/semester are highly related, since the credits of a class are supposed to be an expected value of the weekly hours of work for that class.

Finally, we observe that perturbations in the maximum number of classes/semester, credits/semester and weekly work/semester didn't have a large impact on the total utility of the course load produced by the model; in fact, the utility seemed to stabilize around 145. This is likely because students need at least 25 courses for a 15-2 degree, and there is probably not much flexibility in which courses are taken to count towards these 25 courses. In other words, if a student takes the minimum number of courses, then there is a rigid set of courses to take.

Conclusion:
Our model was successful in outputting a realistic and favorable course load for the 15-2 major. Our model does, however, have a few limitations that could be improved upon. For example, some constraints that we acknowledge from our baseline model is that a typical MIT student may not want to take a semester with only HASS classes and no technicals (as represented by Term 4 in Table 1). However, this limitation could be resolved by incorporating an additional constraint to our model that would constrain at least 1 class to be a technical class for each term. Another limitation of our model is that it generalizes all HASS classes. For future iterations of this model, we plan to allow students to pick a HASS concentration and allow for an optimal course load for their concentration specific HASS classes using the similar methodology that we used to optimize course 15-2 major.

Part 1: Predicting 6.046 Enrollment at MIT

Motivation:
Each year, professors must in some way predict how large their classes will be. Being able to accurately do so is crucial, as class size motivates many decisions that instructors need to make, including how many TAs, LAs, and graders to hire, how many sets of office hours to hold, and how many classrooms (along with the sizes of the classrooms) to book. For example, if more than 566 people (the capacity of the largest lecture hall at MIT) enroll in a class, an instructor will likely have to book an overflow room, or video-record the lectures so that all students have an opportunity to attend lectures.

Some ways that professors might try to intuitively predict enrollment is to look at historical data, pre-registration numbers, or trends in major declaration among different classes at MIT. We are interested in finding a more explicit and optimal way to predict enrollment for classes. This report will focus on Design and Analysis of Algorithms (6.046), a popular course at MIT that is a requirement for an EECS degree but is not a General Institute Requirement

Data:
Our task is to predict the number of students enrolled in 6.046 during any given semester at MIT. The data we used was collected from the MIT Registrar's office. There are 15 rows and 6 columns, with each row representing a semester at MIT (from Fall 2011 to Fall 2018), and each column representing a feature of that semester. The particular features that we used were time, semester, 6.046 enrollment, 6.046 rating, 6.006 (Introduction to Algorithms) enrollment, and 6.006 rating. We chose to include information for 6.006 because it is a prerequisite to 6.046, and might be helpful in predicting 6.046 enrollment.

Approach:
We used various models to predict 6.046 enrollment: namely, an autoregressive model with lag-1, an autoregressive model with lag-2, and other autoregressive models with significant features like pre-registration numbers or class ratings.

<u>Model 1: Autoregressive Model, Lag =1</u>
The idea behind an autoregressive model is that we can predict the enrollment of a given semester using historical data that we have available to us. For example, if an instructor notices that the enrollment of her class has been climbing steadily over the last few years, she might expect the enrollment to continue increasing in the future. Or, if she observes that enrollment tends to occur in cycles, she might use that information to motivate her choices as well.

Our first autoregressive model is the most simple. The idea is to predict this semester's enrollment by looking at last semester's enrollment. To do this, we created an independent variable representing last semester's enrollment, and then ran a linear regression with one independent variable.

Model 2: Autoregressive Model, Lag = 2
Our second model took into account the observation that enrollment in a given fall semester might be more correlated to enrollment in the previous fall semesters, and likewise in the spring. Thus, we created a lag variable representing the enrollment of the class two semesters ago.

Model 3: Autoregressive Model Lag = 1, with Spring or Fall
Our third model combines the utility of the first two models to better account for the general trend of the data, as well as the seasonality.

Model 4: Autoregressive Model Lag = 1, with Spring or Fall, with other features
Our fourth and last autoregressive model is the same as our third model, but includes other potentially relevant features like the enrollment last semester for 6.006.

Results and Conclusion:

Full results are in Appendix C. Most notably, however, we found that of all of the models we tried, Model 4 had the most predictive power, with an out-of-sample R^2 of 0.6679. This makes sense intuitively, as the model used the most information in its prediction. It is important to note, however, that our model (like many autoregressive models) has limitations. Because of the very limited data, the model is likely overfitted to the specific enrollment instances. Additionally, the upward trend of the data is misleading because it might predict a steadily increasing enrollment in the future. It is unlikely that this will be the case, because MIT only admits 1000 undergraduates per year, and tries to select for a diverse applicant pool. It is likely that the surge in 6.046 enrollment has reflected the rise in computer science interest over the last few years, and will not last.

Reflection and Further Work:

Our solutions to both problems we discussed in this paper can be further worked on to create an even better solution. For optimizing course load, we simplified some of the requirements (such as generalizing HASS classes as discussed above) which we could flesh out in more detail. Extending our optimization model to all other majors at MIT would also be an opportunity to extend our solution to other parts of MIT. For predicting enrollment numbers, we can analyze which professors are teaching the class each semester and if some professors are more popular or unpopular than others, and how that affects enrollment. We could also look at how many freshmen declare Course 6 or Course 18 in a given year and if this variable changes enrollment at all.

Though predicting class enrollment numbers and optimizing student course load address very different problems here at MIT, our solutions to these problems are broadly applicable across all undergraduate degrees and classes. Successfully addressing these two types of problems for more classes and majors would vastly improve the MIT undergraduate experience for every student.

Appendix A

Elective Type	Definition
0	Course is not an elective
1	Must pick one course out of the options
2	Must pick between 3 and 5 courses out of the options
3	Can pick up to 2 courses out of the options

Term	Definition
1	Offered only in fall
2	Offered only in spring
3	Offered both in fall and spring

GIR	Definition
0	Not a GIR
1	Non- HASS, non-bio GIR
2	HASS, non-bio GIR
3	Bio GIR

Core	Definition
0	Course is not a core requirement of major
1	Course is a core requirement of major

Appendix B
Optimization Models

Utility-maximizing model

$$\max \quad \sum_{i \in C} \sum_{j=1}^{8} r_i, x_{i,j} \tag{1}$$

s.t.

$$\sum_{i \in C} x_{i,j} \leq m_{classes} \qquad \forall j = 1, .., 8 \tag{2}$$

$$\sum_{i \in C} c_i x_{i,j} \leq m_{credits}[j] \qquad \forall j = 1, ..., 8 \tag{3}$$

$$\sum_{x \in C} c_i x_{i,j} \leq m_{workload} \qquad \forall j = 1, .., 8 \tag{4}$$

$$\sum_{j=1}^{2} x_{i,j} = 1, \sum_{j=3}^{8} x_{i,j} = 0 \qquad \forall i \in \text{GIR} \tag{5}$$

$$\sum_{j=3}^{8} x_{i,j} = 0 \qquad \forall i \in \text{BIO} \tag{6}$$

$$\sum_{j=1}^{8} x_{i,j} = 1 \qquad \forall i \in \text{CORE} \tag{7}$$

$$\sum_{i \in S} \sum_{j=1}^{8} x_{i,j} = 1 \tag{8}$$

$$3 \leq \sum_{i \in F_1} \sum_{j=1}^{8} x_{i,j} \leq 5 \tag{9}$$

$$\sum_{i \in F_2} \sum_{j=1}^{8} x_{i,j} \leq 2 \tag{10}$$

$$\sum_{i \in F_1 \cup F_2} \sum_{j=1}^{8} x_{i,j} = 5 \tag{11}$$

$$\sum_{j=1}^{8} x_{i,j} \leq 1 \qquad \forall i \in C \tag{12}$$

$$|P[i]| x_{i,j} \leq \sum_{k \in P[i]} \sum_{l=1}^{j=1} x_{k,l} \qquad \forall i \in C, \forall j = 1, .., 8 \tag{13}$$

$$x_{i,j} = 0 \qquad \forall i \in \text{SPRING}, \forall j = 1, 3, 5, 7 \tag{14}$$

$$x_{i,j} = 0 \qquad \forall i \in \text{FALL}, \forall j = 2, 4, 6, 8 \tag{15}$$

$$x_{i,j} \in \{0, 1\} \qquad \forall i \in C, \forall j = 1, ..., 8 \tag{16}$$

Semester-minimizing model

$$\min \quad \sum_{j=1}^{8} y_j \tag{1}$$

s.t.

$$\sum_{i \in C} x_{i,j} \leq m_{classes} \qquad \forall j = 1, .., 8 \tag{2}$$

$$\sum_{i \in C} c_i x_{i,j} \leq m_{credits}[j] \qquad \forall j = 1, ..., 8 \tag{3}$$

$$\sum_{x \in C} c_i x_{i,j} \leq m_{workload} \qquad \forall j = 1, .., 8 \tag{4}$$

$$\sum_{j=1}^{2} x_{i,j} = 1, \sum_{j=3}^{8} x_{i,j} = 0 \qquad \forall i \in \text{GIR} \tag{5}$$

$$\sum_{j=3}^{8} x_{i,j} = 0 \qquad \forall i \in \text{BIO} \tag{6}$$

$$\sum_{j=1}^{8} x_{i,j} = 1 \qquad \forall i \in \text{CORE} \tag{7}$$

$$\sum_{i \in S} \sum_{j=1}^{8} x_{i,j} = 1 \tag{8}$$

$$3 \leq \sum_{i \in F_1} \sum_{j=1}^{8} x_{i,j} \leq 5 \tag{9}$$

$$\sum_{i \in F_2} \sum_{j=1}^{8} x_{i,j} \leq 2 \tag{10}$$

$$\sum_{i \in F_1 \cup F_2} \sum_{j=1}^{8} x_{i,j} = 5 \tag{11}$$

$$\sum_{j=1}^{8} x_{i,j} \leq 1 \qquad \forall i \in C \tag{12}$$

$$|P[i]| x_{i,j} \leq \sum_{k \in P[i]} \sum_{l=1}^{j=1} x_{k,l} \qquad \forall i \in C, \forall j = 1, .., 8 \tag{13}$$

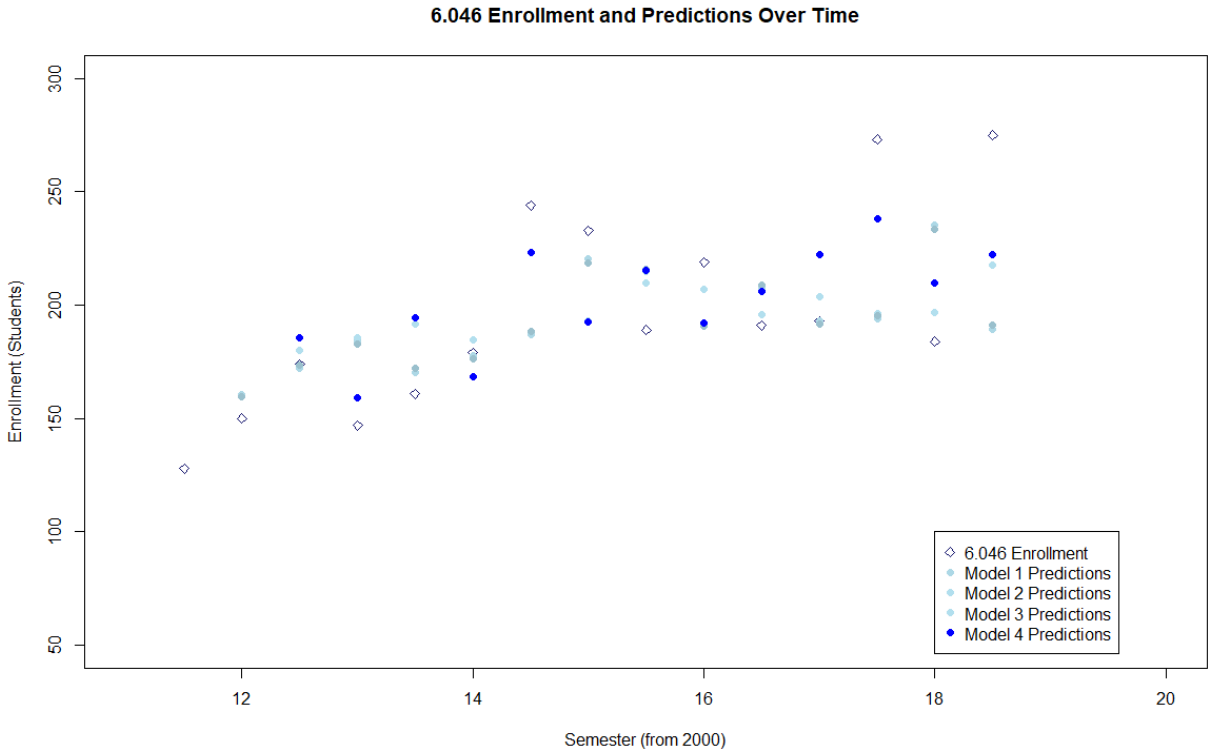
$$x_{i,j} = 0 \qquad \forall i \in \text{SPRING}, \forall j = 1, 3, 5, 7 \tag{14}$$

$$x_{i,j} = 0 \qquad \forall i \in \text{FALL}, \forall j = 2, 4, 6, 8 \tag{15}$$

$$8 y_j \geq \sum_{i \in C} x_{i,j} \qquad \forall j = 1, ..., 8 \tag{16}$$

$$x_{i,j}, y_j \in \{0, 1\} \qquad \forall i \in C, \forall j = 1, ..., 8 \tag{17}$$

Appendix C



Model Type	**Sample R^2**	**Sample $RMSE$**	**Out of Sample R^2**
Model 1 (Lag = 1)	0.3488	25.8463	0.0266
Model 2 (Lag = 2)	0.0938	29.4162	0.4294
Model 3 (Lag = 1 + Semester Information)	0.3509	25.8051	0.0657
Model 4 (Lag = 2 + 6.006 Enrollment + Course Rating)	0.3918	24.0995	0.6679