Admission Strategy Memo

Zaiyang Guo, Zhirui Zhang

Part I

First of all, we want to check the probability of enrolling once admitted (yield) for different groups of students using conditional means, so as to get an idea of how different students' decision changes according to certain conditions and help us decide the target group(s) of our new policy. We grouped the admitted students by different variables and calculated the mean yield values within each group. For continuous variables, we divided the dataset into deciles and calculated mean yield for each; for binary variables, we divided the dataset by whether that particular variable is true/false, and calculated conditional means. Our potential policy changes concern the following variables: SAT scores, income, net price, and thus we will focus on those. The results are shown below.

For SAT scores, students with higher SAT scores have a higher yield rate. The same trend is observed with students coming from higher-income households, as they are more likely to enroll. However, the trend for net price is very different. When the net price is between $0 to $13053, the net yield rate increases from 0.512 to 0.921; the net yield rate then drops from 0.921 to 0.279 as the net price increases from $13053 to $22505; then, the net yield rate increases again from 0.279 to 0.860 as net price increases from $22505 to $45000 – this indicates different price sensitivity for students of different net price groups.

	income_decile	min_income	pct_yield
	<int>	<dbl>	<dbl>
1	1	8844.	0.205
2	2	41483.	0.265
3	3	55177.	0.447
4	4	68818.	0.591
5	5	82099.	0.651
6	6	99895.	0.809
7	7	119776.	0.884
8	8	143791.	0.967
9	9	178036.	1
10	10	241071.	1

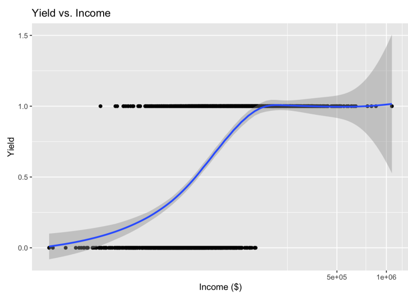
A tibble: 10 × 3

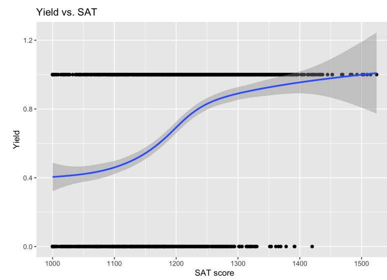
	sat_decile	min_sat	pct_yield
	<int>	<dbl>	<dbl>
1	1	1000	0.405
2	2	1072.	0.474
3	3	1114.	0.488
4	4	1145.	0.577
5	5	1173.	0.581
6	6	1200.	0.8
7	7	1227.	0.819
8	8	1252.	0.837
9	9	1285.	0.884
10	10	1329.	0.953

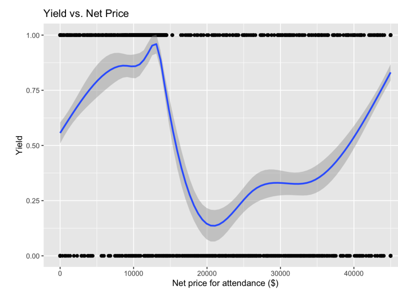
A tibble: 10 × 3

	net_price_decile	min_price	pct_yield
	<int>	<dbl>	<dbl>
1	1	0	0.512
2	2	0	0.786
3	3	9464.	0.884
4	4	11689.	0.921
5	5	13053.	0.926
6	6	14109.	0.414
7	7	22505.	0.279
8	8	31248.	0.419
9	9	43766.	0.819
10	10	45000	0.860

We also plotted yield against each of the three variables for better visualization. We can thus see clear patterns: as income increases, yield in general also increases. Thus, if the tuition revenue is below the 30,000,000 threshold, we may try to enroll more students from higher income households to make up for the difference. As for net price, high enrollment rate is observed when the net tuition is either low or high. The flat curve near the end indicates that there is a group of students whose willingness to pay does not vary a lot even with higher tuition, meaning their low price sensitivity. We can thus tap into that for making up tuition loss while we try to achieve the other goals of enrolling 200 more low-income students and not enrolling too few or too many students. The graphs are shown below.



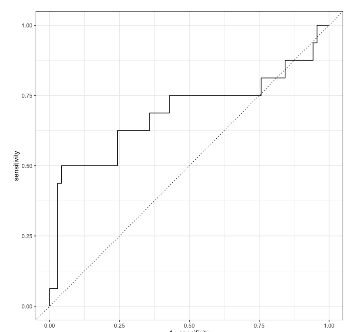




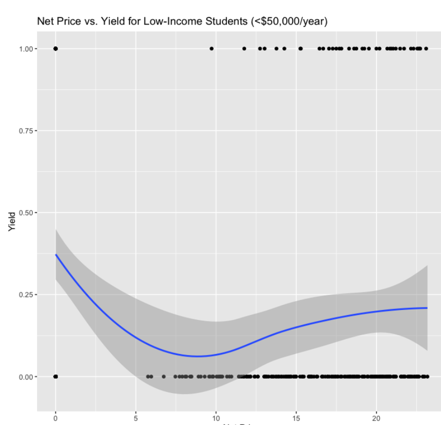
Part II

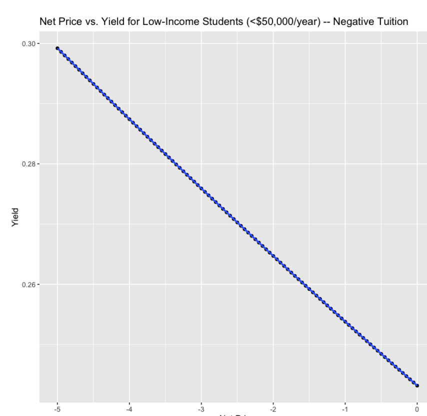
Next, we established a logistic regression model for students coming from households whose income is less than $50,000. The regression results are as shown below (coefficient and ROC curve). This model has an accuracy of 87.2%, a sensitivity of 50% and a specificity of 95.7%. It has a very poor capability to predict students who will enroll, but has a very good capability to predict those who will not enroll. We believe that the low sensitivity is a result of a small sample number as well as a low overall enrollment rate of 22.5% (as compared to the 76.8% enrollment rate of higher-income students). In terms of the coefficients, we noticed in particular that the coefficient of net_price is negative, which indicates that lower net tuition will increase their change of enrollment.

```
# A tibble: 10 × 5
   term        estimate std.error statistic   p.value
   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept) -11.3       4.04      -2.80  0.00515
 2 legacy        0.218     0.449      0.487 0.627
 3 visit         0.263     0.356      0.738 0.461
 4 registered    0.628     0.333      1.89  0.0593
 5 sent_scores   0.600     0.473      1.27  0.205
 6 sat           0.137     0.344      0.399 0.690
 7 income        0.0548    0.0206     2.66  0.00772
 8 gpa           1.16      0.848      1.37  0.171
 9 distance     -0.837     0.201     -4.17  0.0000307
10 net_price    -0.0568    0.0304    -1.87  0.0617
```



We plotted yield against net price for low-income students. We observed that if net price is lower than about $8,000, the student's probability of enrollment increases as net price goes down.

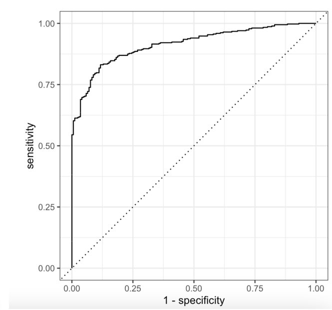




We thus want to test whether a negative net tuition (i.e. the college gives our merit-based scholarships or aid) would affect their probability of enrollment. We used hypothetical data points and found that as net price becomes more negative, the probability of enrollment increases slightly.

In addition to that, we set up a general linear regression model in order to reveal the relationship between the dependent variable "whether the student yielded or not" and each independent variable, including whether is legacy, visit, registered, sent scores, sat scores, household income, high school gpa, distance, and net price. In the model, we split the data into training and testing for checking its accuracy, and added seed to avoid variations. As a result, we found that legacy, visit, registered, sent_scores, income, and gpa all have a positive correlation with the probability of yielding, while SAT scores, distance, and net price have a negative correlation with yielding (see the left figure below). However, the correlation for SAT is not very statistically significant. The model has an accuracy of 83.6%, a sensitivity of 85.9%, a specificity of 78.5%, and an AUC of 91.3%, all well above the baseline of 68.2%, meaning the model is very accurate (see the right figure below). This is to get us a general view of how changes in different variables may affect the final outcome of yielding, thus helping us to decide the direction of policy changing later.

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-10.7	1.80	-5.95	2.66e- 9
2	legacy	0.554	0.177	3.13	1.73e- 3
3	visit	0.308	0.159	1.94	5.22e- 2
4	registered	0.502	0.153	3.29	1.00e- 3
5	sent_scores	0.727	0.208	3.50	4.66e- 4
6	sat	-0.0881	0.147	-0.599	5.49e- 1
7	income	0.0523	0.00363	14.4	4.18e-47
8	gpa	1.97	0.393	5.00	5.59e- 7
9	distance	-0.488	0.0806	-6.05	1.43e- 9
10	net_price	-0.0565	0.00879	-6.43	1.29e-10



To get an idea of the current student group's baseline characteristics, we applied a filter to summarize the target variables. Now, there are 1466 students who yielded, their average SAT score is 1226, among them 77 come from families with an annual income of less than $50,000, and the total revenue for the school is $30,674,149. Based on this information, our policy will need to 1) increase the average SAT score by at least 75, which can be achieved by either admitting more students with higher scores or deterring students with low scores; 2) increase the number of low-income students by 200, i.e. to 277+, which can be achieved by applying tuition deduction/financial aid/scholarship for the low-income; 3) there is still space for us to admit fewer students and lowering the price, but not a lot; 4) to balance the revenue budget, we can charge a higher price for students coming from high-income families as they are less price sensitive.

Part III

We first decided to look at SAT score as our first anchor. To begin with, we figured out the distribution of the student population based on the SAT threshold of 1300 and their yielding situation. This gives us the following matrix: There are only 13 students (0.8% of total yielded students) with higher-than-1300 SAT who did not yield, while students with lower scores (1104) are the majority (75%). Then we filtered out students who either yielded or have scores >=1300 and checked their average SAT, meaning that even if we admit all of these students, the SAT is increased to 1234 but is still below the expected level, not to mention the effect of other possible variations that may reduce yielding probability. Such a disproportional distribution of high and low SAT students makes the goal of achieving an average 1300 SAT score very difficult. Therefore, we decided to forgo the "average SAT1300" goal.

```
# A tibble: 4 × 3
# Groups:   sat >= 13, .pred_class [4]
  `sat >= 13` .pred_class     n
  <lgl>       <fct>       <int>
1 FALSE       No            693
2 FALSE       Yes          1114
3 TRUE        No             11
4 TRUE        Yes           332
```

```
> ad_pred_filtered <- ad_pred %>%
+   filter(sat >= 13 | .pred_class=="Yes") %>% #to have students who either yielded or have scores >=1300
+   summarize(mean(sat * 100))
> ad_pred_filtered
# A tibble: 1 × 1
  `mean(sat * 100)`
              <dbl>
1             1234.
```

With our aforementioned strategy direction in mind, we come up with the following financial policy changes: for students who have higher SAT scores, such that students with SAT score higher than 1350 could get a deduction of $30,000 in their tuition for every 100 points they exceeded the threshold, while students with lower scores will be added $20,000 for every 100 points below the threshold; here we do not allow negative net price so a negative number will simply be charged $0. To attract more low-income students whose family income is less than $50,000 per year, we will deduct \$28k; and we will allow a negative net price, which simply means the school is giving merit scholarships to those students with lower income. To balance the total revenue and the total number of students enrolled, we will charge students whose family income is higher than $180,000 per year an additional $50,000 as they are less price sensitive.

Our results suggest that the mean sat score is now increased to 1236 from 1226, and the net price of $31,660,936 is still above the required level, the number of students becomes 1456 which is within expected, and we increased the number of low-income students to 279 that exceeds the expectation; 202 more low-income students are predicted to enroll. Therefore, three of the four goals are achieved under our new pricing policy. Also, though the first goal was not achieved, we increased the average SAT score by more than 10 points, which is still a step closer to the initial goal.

There are a few ethical implications to our policy changes. On the positive side, we managed to enroll over 200 additional low-income students. This would certainly mean that these low-income students are given a more leveled playing field in the form of equitable education, and would give them opportunities that they otherwise might not have without attending college. However, we also managed to keep the overall tuition revenue relatively the same, which means that the money used to subsidize low-income students are mostly from students who come from high-income households or students who are willing to pay more to attend college. For those students, by choosing to enroll, they essentially give an implicit consent to financially aid students from low-income households. This implicit consent is often not obvious to many students and those who pay for their college education. Thus, the implicit consent to aid others is rather questionable.