Predicting ACU Enrollment and Retention Rates

Exceptional. Innovative. Artificial.

Rachael Shudde and Preston Werner

Research question

Question: Using data on high school students who are admitted to ACU, can we reliably predict which students will attend ACU?

Division of work

- Code
  - Preston 60%
  - Rachael 40%
- Power points and data meeting
  - Rachael
- Graphs and visualization
  - Preston

Final product example

Input:

Name	*Age*	*SAT*	*Family Income*	*City*	*State*	*Received*	*Legacy*
John Smith	17	1200	$35,000	Fresno	CA	3/4/18	Yes

Output:

John Smith has a **56%** chance of coming to ACU. His **low family income** and **far distance from ACU** are the two factors that impact the low chance of attending.

Grade:

A+

Final project report

- We ended being able to
  - Give a prediction of if an admitted student would come to ACU
  - Tell which of four factors might change their probability of coming to ACU
  - We ignored retention rate (hard to create testing data)

Original data and simplification

- Admissions data from ACU office with 128 columns, over 550,000 rows
- Information on all types of prospective students
- Information ranging from SAT scores to family income
- We ended up keeping 19 columns and 12,485 rows

Student data types

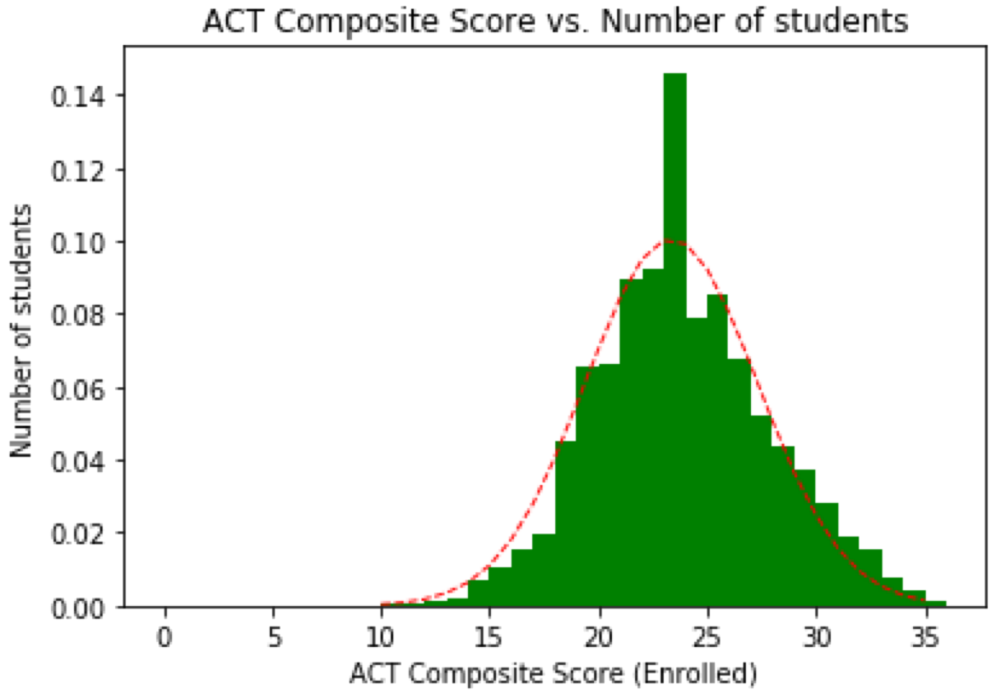
- **Suspect** - ACU has has bought student name
- **Prospect** - ACU has bought student name, student shown no interest, might possibly be a good fit for ACU
- **Inquiry** – student shown interest in some way, from contacting ACU to attending a camp
- **Application** – student has submitted application
- **Admitted** – student has been admitted to ACU
- **Confirm** – student confirms attendance by paying the deposit
- **Enroll** – student is registered for classes
- **Deny** – student is admitted but decides not to attend

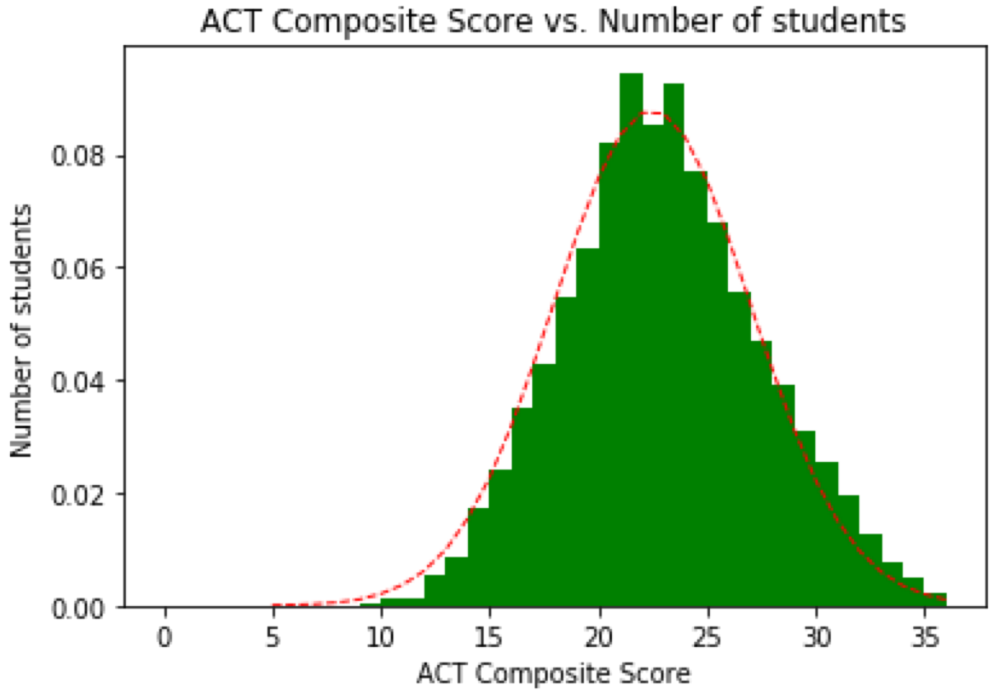
Rows kept

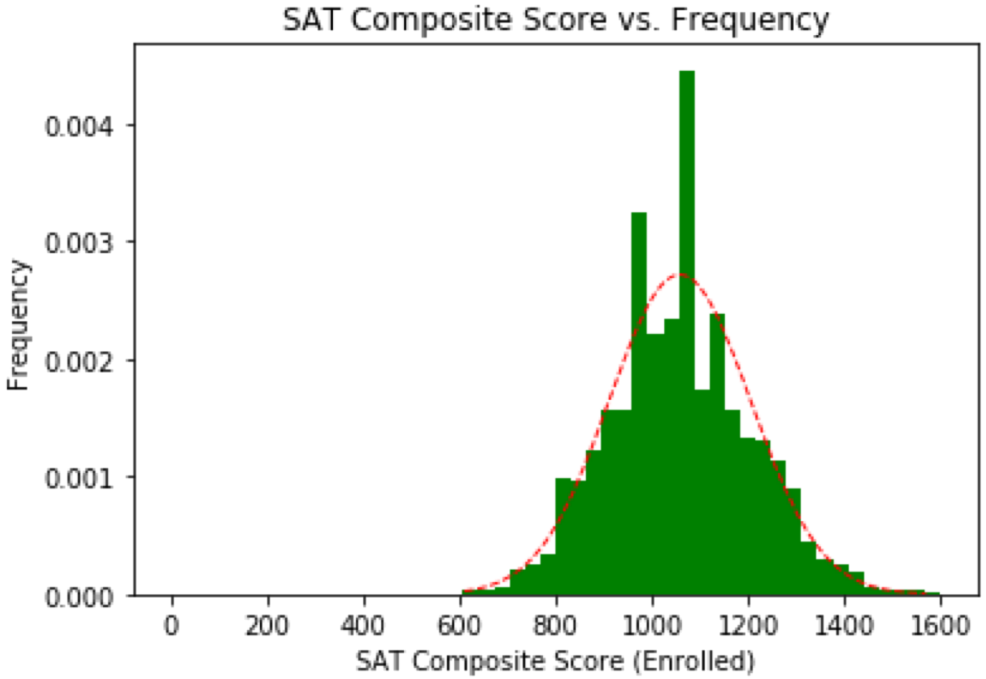
1. Gender
2. Core_Ind
3. Parent_Same_Last_Name
4. ACU_Distance
5. Household Income
6. QUAL_Familiarity
7. QUAL_rank
8. HS_Size
9. HS_Rank Percent
10. HS GPA
11. HS_GPA
12. NUM_advanced_classes
13. ACT
14. Visited
15. Taylor County
16. US
17. Abilene
18. Church of Christ
19. Ethnicity

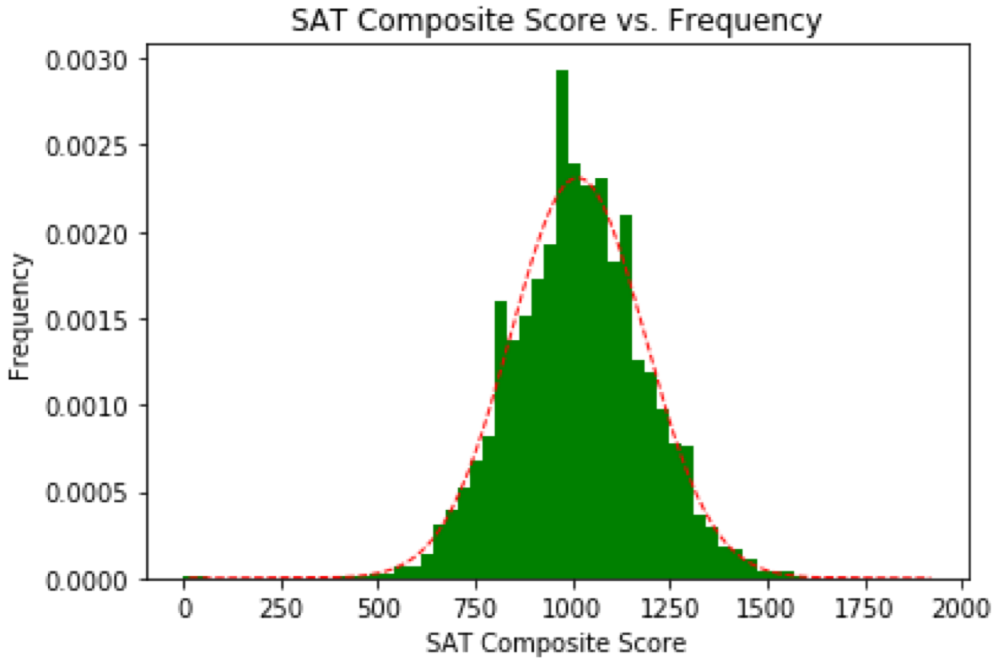
Summary statistics

- Female: 60%, Male: 40%
- Mean income: $70,590
- Median income: $63,587
- Mean ACT: 22.5
- Enrolled mean ACT: 23.3
- Mean SAT: 1013
- Enrolled mean SAT: 1058
- Legacy Enrolled: 98%
- Not Legacy Enrolled: 32%


ACT Composite Score vs. Number of students
Number of students
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
35
ACT Composite Score (Enrolled)


ACT Composite Score vs. Number of students
Number of students
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
35
ACT Composite Score


SAT Composite Score vs. Frequency
Frequency
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
1000
1200
1400
1600
SAT Composite Score (Enrolled)

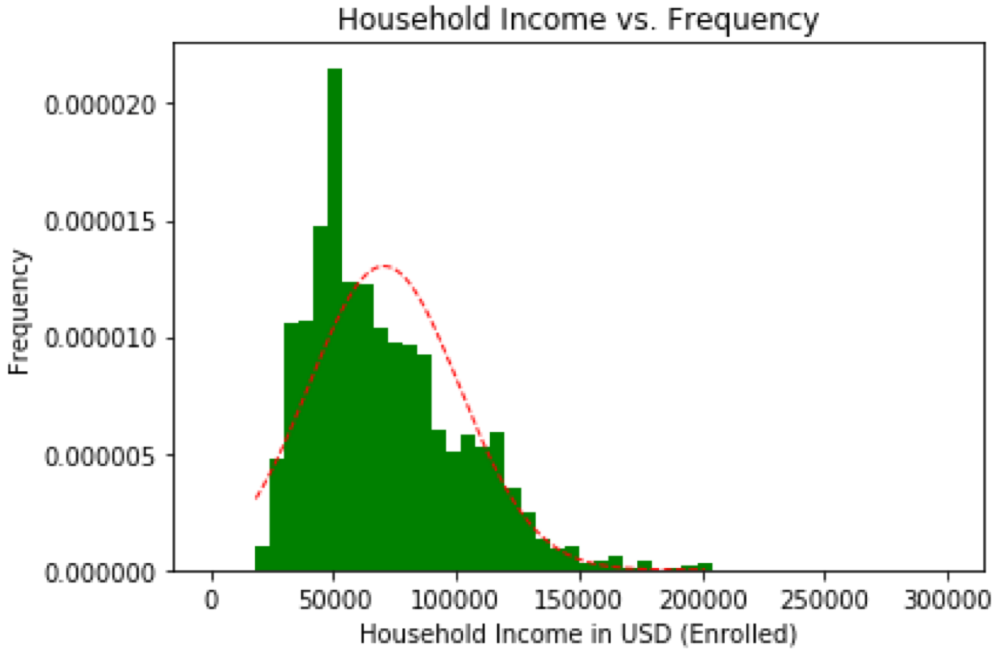

SAT Composite Score vs. Frequency
Frequency
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0
250
500
750
1000
1250
1500
1750
2000
SAT Composite Score

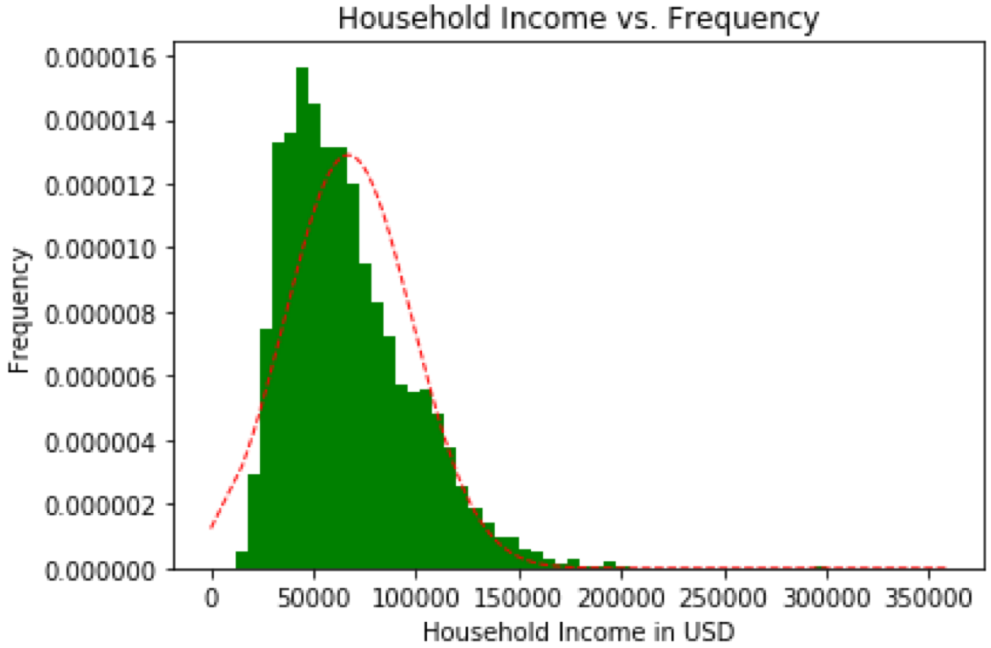
Location summary

State	Number Interested
TX	147386
CA	3760
OK	2481
CO	2464
TN	2166

County	Number Interested
Tarrant	15400
Harris	14375
Dallas	12798
Bexar	9253
Collin	6550

Country	Number Interested
USA	176903
Mexico	948
Honduras	674
Hong Kong	354
Nigeria	319


Household Income vs. Frequency
Frequency
0.000020
0.000015
0.000010
0.000005
0.000000
0
50000
100000
150000
200000
250000
300000
Household Income in USD (Enrolled)


Household Income vs. Frequency
Frequency
0.000016
0.000014
0.000012
0.000010
0.000008
0.000006
0.000004
0.000002
0.000000
0
50000
100000
150000
200000
250000
300000
350000
Household Income in USD

Yes, but keep in mind that all of this data was pulled from a production database as of 10/2017, meaning that for all but the most recent term (201810 - or Fall 2017) the scores in this field were likely generated at the end of the cycle when most of the data for determining enrollment were fully known. Imagine you're a part of the Fall 2017 class, so it's October of your Senior year in high school. You're a late starter, so maybe you haven't even applied to ACU yet, but if you have you probably haven't sent in all your transcripts and been admitted, you haven't registered for New Student Orientation or attended an Admitted Student Reception. Relatively speaking, we know very little about you. Now imaging you're a part of the Fall 2016 class or earlier. If you enrolled at ACU you did so about two months ago and we already know that. Even if we don't have an enrollment flag in the data we're looking at, we have a record of you attending Passport, an ASR, or even just know that you've sent 15 emails to your admissions counselor working through scholarships and making your decision. That data bleeds into just about everything in the file. If you include any of it in your model it's going to look very viable. You'll probably get validation statistics like AUC values or R2 in the .99+ range. Yet, if you apply that model to Fall 2017 data and score it, the scores will basically be meaningless, because most of the data you built the model on doesn't exist yet. Does all of that make sense?

The same will be true of these model scores if you're comparing them to your model, especially for the Fall 2017 class. It will appear that your model is able to predict much better, but in reality it's just telling us what we already know - that a certain subset of students already enrolled at ACU. You have to be very careful to pick a relative date for your model and not to include data in the model that would not have been known at that point in the recruitment cycle. This can be very challenging to do.

Sorry if I'm preaching to the choir here. I'm not sure what you've covered in your coursework and would rather explain all of this now than when you come back with what looks like a great model only to find out it's based on future data and operationally infeasible.

Good luck. Let me know if you run into any more questions.

Nick

It's going to depend on your objectives. Are you trying to build a model to predict who will enroll as of a given point in time in the year? For example, the file was pulled around 10/12/2017, I believe. So it would make sense to me to set that as your relative date. The term you would be trying to predict would be the entry term Fall 2018 (entry term code 201910 or entry term offset 0). That way you know all of the data for that term is point-in-time, and that's almost a year before the student will actually enroll. You would then need to filter everything else you throw into the model down to only include data that would have been known as of that date in the relative term. We usually do this using the entry term offset field.

If I want to know who had applied as of 10/12 for the Fall 2017 term that relative date would be 10/12/2017, for Fall 2016 it would be 10/12/2016, etc. The offset is set up to make this formula work for pulling relative dates:

AppDate <= getdate()-365*EntryTermOffset

So in SQL, if I wanted to pull students who had applied as of 10/12 I would say:

```
Select
CASE
WHEN [AppDate]<=cast('10/12/2017' as date)-365*[EntryTermOffset]
THEN 1
ELSE 0
END as AppliedFlag
from [TABLE]
```

I'm sure you could do the same thing with a formula in excel, though I've never tried it. It's much easier to manipulate data in SQL.

If your objective is to do a study after the fact and determine causal effects on enrollment your setup would be completely different. Does that answer your question?

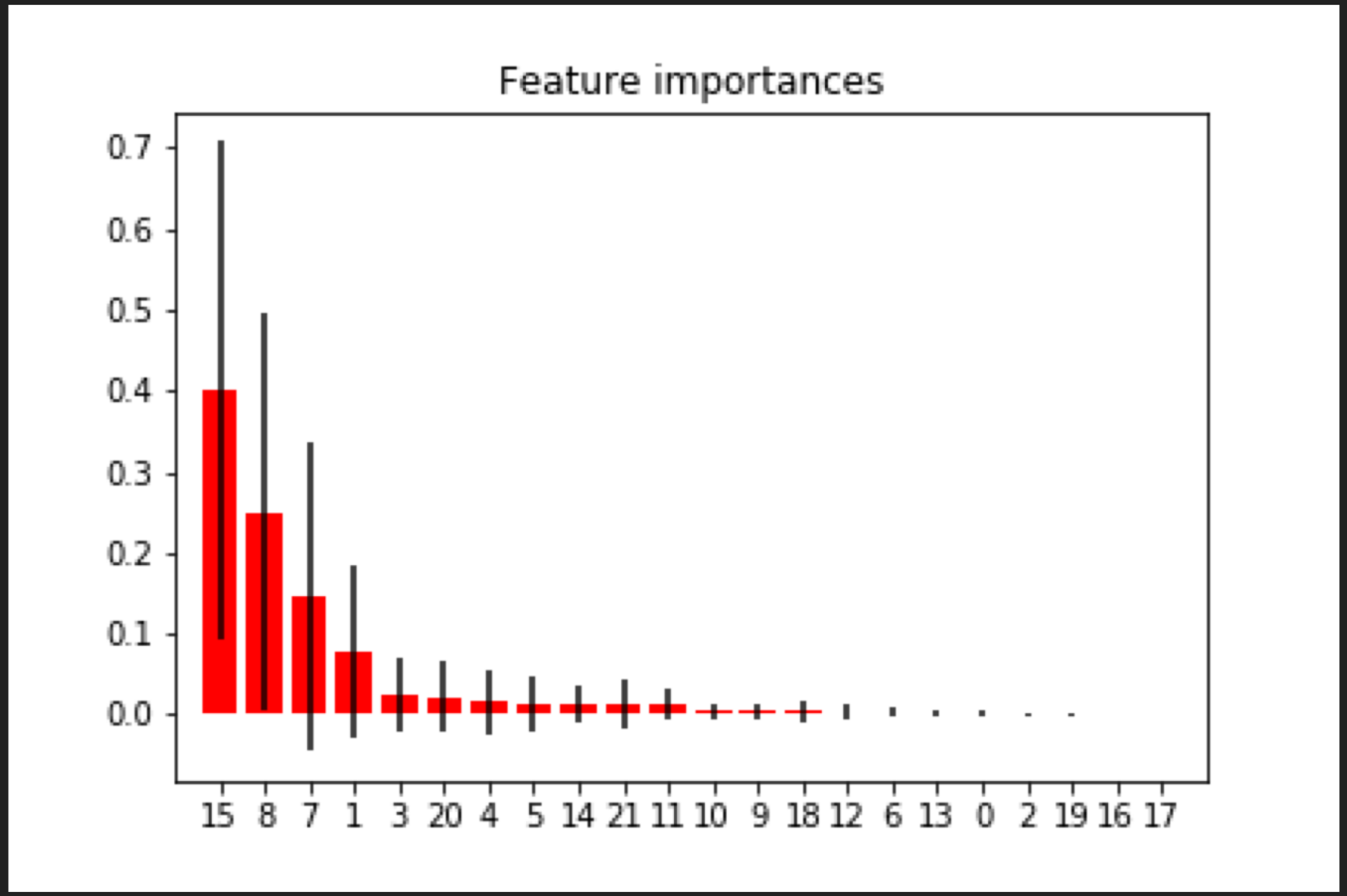
Nick Peterson

Decision tree

- Decision tree score: 88%
- We had the following parameters from a grid search:
  - Max depth: 4
  - Max leaf nodes: 13
  - Minimum samples in each leaf: 10
  - Minimum samples to split: 11

Feature importance ranking

Feature	Percentage Contribution
Visited	40%
Qual_Rank	25%
Qual_Familiarity	15%
Core_Ind	7%
ACU_Distance	2%
CoC	2%

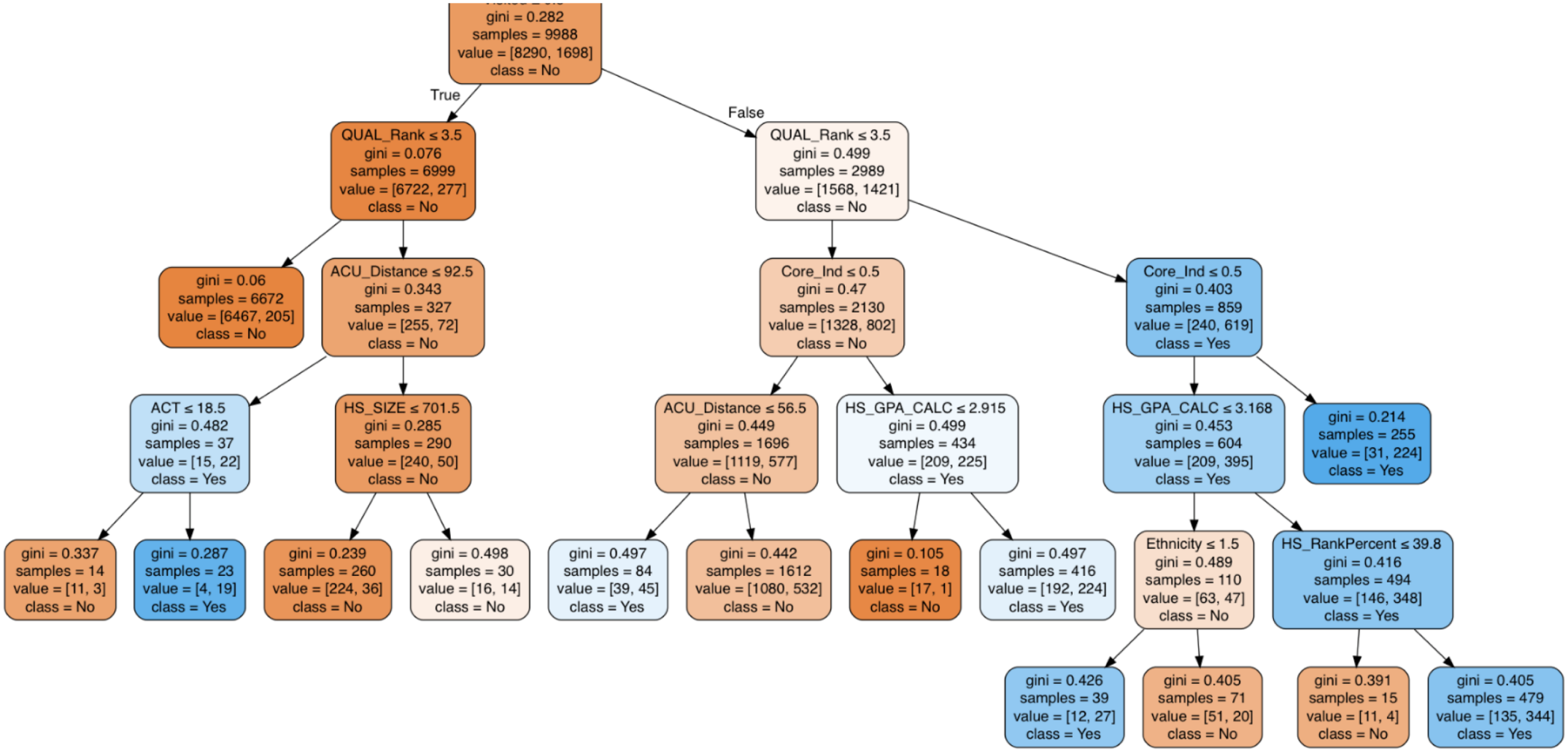

Feature importances
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
15 8 7 1 3 20 4 5 14 21 11 10 9 18 12 6 13 0 2 19 16 17

Confusion matrix

Actual	Predicted		
	191 True Positive	202 False Negative	393
	115 False Positive	1989 True Negative	2104
	306	2191	2497 Total

Gender	Core_Ind	Parent_Same_	ACU_Distanc	ACU_Distance_Band	Parent_ACU_	PRIZM_Hou	QUAL_Fami	QUAL_Rank	HS_SIZE	HS_RankPer	HS_GPA_CA	ACU_AcadRa	NUM_Advar	ACT	Visited	Taylor_Coun	US	Abilene	Texas	CoC	Ethnicity
1	0	2	1500	1	4	75000	3	2	500	70	3.8	5	5	25	1	0	1	0	0	1	1
1	1	1	100	8	1	90000	1	1	500	70	3.8	5	5	25	1	0	1	0	1	1	1
2	1	2	100	8	1	150000	4	2	1000	90	4	5	5	20	0	0	1	0	1	0	2
1	1	2	561	2	2	98951	4	3	435	47.36	3.25	4	3	23	1	0	1	0	0	0	1
2	0	2	500	2	2	86000	2	1	600	50	3	3	2	24	1	0	1	0	0	1	1
2	0	1	168	7	7	600000	2	1	400	14	3.4	5	4	32	1	0	1	0	1	0	1
1	1	2	30	10	10	150000	4	2	4	100	4	4	7	29	1	1	1	1	1	0	1

```
The student's probability of coming to ACU is:  0.459


The four most important factors in our model are:
1. If the student has visited or not
2. How the student ranked ACU on their list of colleges
3. How familiar the student said they are with ACU
4. If the student has a previous connnection to ACU


The student's visited status is: visited
Changing their visit staus to:  0  decreases their probability of coming to:  0.105


The student ranked ACU, on a scale of 1-5, on their preferences as:  3
Changing their ranking of ACU to:  2  does not change their original probability of coming to ACU


The student, on a scale of 1-4, said they were:  4  familiar with ACU
That's as high as they can rank ACU


The student's Core index status is: yes
```


gini = 0.282
samples = 9988
value = [8290, 1698]
class = No
True
False
QUAL_Rank ≤ 3.5
gini = 0.076
samples = 6999
value = [6722, 277]
class = No
QUAL_Rank ≤ 3.5
gini = 0.499
samples = 2989
value = [1568, 1421]
class = No
gini = 0.06
samples = 6672
value = [6467, 205]
class = No
ACU_Distance ≤ 92.5
gini = 0.343
samples = 327
value = [255, 72]
class = No
Core_Ind ≤ 0.5
gini = 0.47
samples = 2130
value = [1328, 802]
class = No
Core_Ind ≤ 0.5
gini = 0.403
samples = 859
value = [240, 619]
class = Yes
ACT ≤ 18.5
gini = 0.482
samples = 37
value = [15, 22]
class = Yes
HS_SIZE ≤ 701.5
gini = 0.285
samples = 290
value = [240, 50]
class = No
ACU_Distance ≤ 56.5
gini = 0.449
samples = 1696
value = [1119, 577]
class = No
HS_GPA_CALC ≤ 2.915
gini = 0.499
samples = 434
value = [209, 225]
class = Yes
HS_GPA_CALC ≤ 3.168
gini = 0.453
samples = 604
value = [209, 395]
class = Yes
gini = 0.214
samples = 255
value = [31, 224]
class = Yes
gini = 0.337
samples = 14
value = [11, 3]
class = No
gini = 0.287
samples = 23
value = [4, 19]
class = Yes
gini = 0.239
samples = 260
value = [224, 36]
class = No
gini = 0.498
samples = 30
value = [16, 14]
class = No
gini = 0.497
samples = 84
value = [39, 45]
class = Yes
gini = 0.442
samples = 1612
value = [1080, 532]
class = No
gini = 0.105
samples = 18
value = [17, 1]
class = No
gini = 0.497
samples = 416
value = [192, 224]
class = Yes
Ethnicity ≤ 1.5
gini = 0.489
samples = 110
value = [63, 47]
class = No
HS_RankPercent ≤ 39.8
gini = 0.416
samples = 494
value = [146, 348]
class = Yes
gini = 0.426
samples = 39
value = [12, 27]
class = Yes
gini = 0.405
samples = 71
value = [51, 20]
class = No
gini = 0.391
samples = 15
value = [11, 4]
class = No
gini = 0.405
samples = 479
value = [135, 344]
class = Yes

Things we might have done if there was more time

- Figured out just what ACU's models mean
- Looked at retention rates for the first three semesters
- Found data on how major impacts retention rates
- Trained a neural network
- Implemented a GUI

Lessons learned

- Data is often messy and a lot of our decisions were arbitrary
- Our model does not match what we initially thought, but it makes sense
- ACU should mainly focus on encouraging campus visits
- ACU should focus on getting students more informed about ACU
- ACU should foster more relationships between alumni and potential students
- We should get an A