

In Class Activity – More ML Models

In this activity we will continue to work with the data we cleansed in a prior activity.

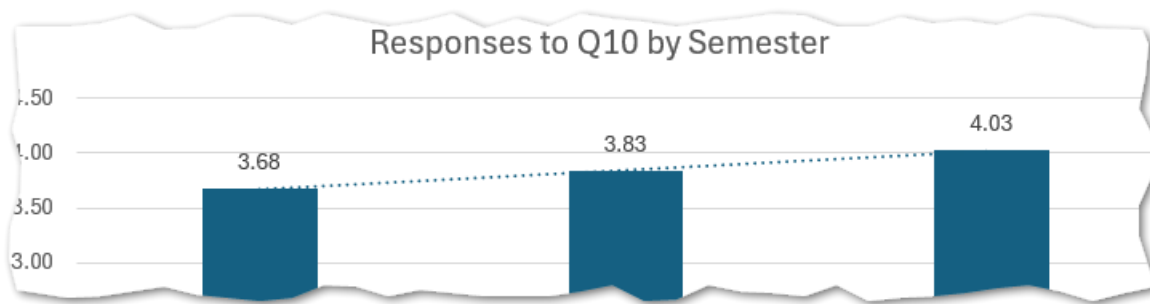
The survey data is real. It was collected over three semesters here at Temple. The survey was constructed to investigate student attitudes towards “flipped classroom” instruction of a STEM topic, post COVID.

We want to use appropriate models and techniques to determine answers to the following questions:

- A. Is student receptiveness to flipped classroom instruction increasing or decreasing?
- B. What characteristics best describe the students who do not want flipped classroom instruction?
- C. What groupings of student responses naturally exist in the data?

In the last activity we focused on question A. To answer this question, we used Excel to create a bar chart with a trend line. (The trendline was our rough approximation of a linear regression line. Sadly, our data does not lend itself to a proper regression analysis. Typically, we would use a regression model to predict a single *continuous* outcome variable.)

That got us a chart that looked like this:



We observed that, while the trend appeared to be upward, it was not dramatic. The only way we could improve the predictive power of our model would be with more data ... either more semesters, more student responses, or both. (Side note: the growth we see is *not* statistically significant. See your instructor if you want more details!)

This is what's known as a *scaling law*.

Our first scaling law is – the quality of the ML model improves with more training data.

Exploring Question B – Using a Decision Tree Model

Now we are going to use a decision tree model to predict a **categorical** outcome variable. We want to know which survey question(s) predict a student's approval of this instructional technique.

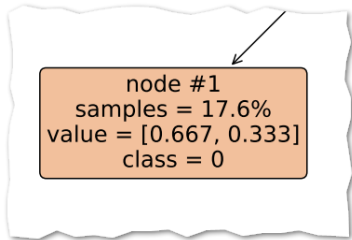
Instructions

1. Download ica-ml03.zip and unzip the ica-ml03 folder into your mis3536workspace.
2. Using Anaconda Navigator, open the folder and the Jupyter Notebook named decision_tree.ipynb.

3. Your instructor will review the contents of this script with the class.

DISCUSS: Why aren't we considering question 10 to predict approval?

4. Run this script to generate a decision tree model with a max depth of one.
 - a. What survey question is the most powerful predictor of approval?
 - b. Take note of the model's estimated accuracy (reported at the bottom of the script)

<p><i>What do all these numbers mean?</i></p> 	<p><i>Well...</i></p> <ul style="list-style-type: none">• This is node #1.• 17.6% of the data in my training set ended up in node 1.• 2/3 of these "node 1" people did not approve. They were coded 0.• 1/3 of these "node 1" people did approve. They were coded 1.• That is to say... this is a "class 0" bunch of people... most of them were coded 0 because they did not approve.
---	---

5. Now rerun the script with a max depth of 3, and again with a max depth of 5.

DISCUSS: What do we learn about the importance of various questions? What do we learn about the accuracy of the model as the depth increases?

Here we have a new ***scaling law***.

Our second scaling law is – the quality of the ML model ***tends to*** improve with increased model complexity. The more complex the model, the greater the accuracy.

Careful! This scaling law is not true all the time. At a certain point of complexity, all models can suffer from overfitting.

Overfitting happens when a model has high accuracy on training data but low accuracy on new, unseen data. Their predictive power improves with complexity up to a point and then starts to worsen as the model becomes overly complex.

Neural networks are unique in that they demonstrate an effect called the “double descent”. As the complexity of a neural network grows, its error rate decreases, then increases, then decreases again. This is the “double descent”.

The model we are looking at here today is a decision tree. Not a neural network.

6. To document your participation, your instructor will ask you to upload your decisiontree.pdf with a max depth of 1 to a corresponding activity on canvas.

As stated in the syllabus: Deliverables from in-class activities will be graded as success (100), some problems (80), unacceptable (50) or failure (0) .

The late penalty, also described in the syllabus, applies to this activity.

7. A second upload is expected. Keep reading.

CONTINUED

Exploring Question C – Looking for groups of students in the data

In this section we are going to look for groupings of responses that naturally exist in the data. Notice that, here, we are not trying to predict an outcome. We are not relying on the “semester” column or the “approve” column in our data set. Those are columns that we added to the data set – that is to say, we **labelled** the data, because we wanted to **predict** something.

If your model requires **labelled** data, to **predict a specific outcome**, then it is **supervised** model.

If your model explores the data as it is, to identify an unseen/underlying pattern or grouping, then it is an **unsupervised** model.

8. Open the Jupyter Notebook named correlation-matrix.ipynb.
9. Run it.
10. This correlation matrix shows which question responses are highly correlated, either positively or negatively. A correlation matrix is not really a machine learning model, but it is a powerful way to identify variables of interest.

DISCUSS: Looking at the correlation matrix, and the original question documented in the Excel spreadsheet, what are two questions that are interesting to you? (Other than i6, and i8 that we have already seen!)

FUN FACT: The correlation matrix is not really a model. It’s just a technique for exploring data.

11. Open the Jupyter Notebook named clustering.ipynb.
12. Your instructor will review the contents of this script with the class.
13. Run this script, generating three clusters.
14. DISCUSS: Notice the graph depicting the decrease in error as the number of clusters increases. Careful though, it’s not magic. This is not a predictive model, so you actually want a certain amount of generalization (that is, error) in your model. A cluster model that is too precise ceases to be useful!
15. Edit the script, taking the number of variables from 10 down to 4. (You do this by commenting / uncommenting the variable COLUMNS_FOR_ANALYSIS.)
16. Rerun the script, generating three clusters.

17. DISCUSS: How do we read the following table?

Average of variables in each cluster				
	i4	i6	i8	i9
Cluster 1	-0.0857	0.2702	0.283	-0.0213
Cluster 2	0.1663	-0.7543	-2.7852	0.2396
Cluster 3	0.8876	-2.5349	-0.3745	-0.0064

(Some students may get different results depending on operating system and/or versions of Python being run. That's OK. The important thing is that you know how to read the table.)

18. Edit the script, taking the number of variables from 4 down to 2. (You do this by commenting / uncommenting the variable COLUMNS_FOR_ANALYSIS.)
19. Rerun the script, generating three clusters.
20. DISCUSS: How do we read, **and interpret**, the following table? To **interpret** this output, you must relate it back to the survey questions. Use words that a person not familiar with the survey questions can understand.

Average of variables in each cluster		
	i3	i6
Cluster 1	0.4181	0.268
Cluster 2	-1.8702	-0.0523
Cluster 3	0.1842	-2.4617

21. Finally, on your own, edit the script specifying two columns for analysis. One should be i10, the other of your choosing. Consider prior results and choose an interesting column for analysis.
22. Run the script. Look at the “Average of variables in each cluster” table again.
23. Copy / Paste that table into a MS Word document. Then, write a paragraph that **interprets** what you see there. Also, write a second paragraph justifying your choice of columns (i10 and something else). What was your rationale for choosing those columns?
24. To document your participation, your instructor will ask you to upload your word document to a corresponding activity on canvas.

As stated in the syllabus: Deliverables from in-class activities will be graded as success (100), some problems (80), unacceptable (50) or failure (0) .

The late penalty, also described in the syllabus, applies to this activity.

Survey Questions

i1	<p>Choose the answer that BEST describes you.</p> <p>When do you watch the video lectures?</p> <p>1 - Within approximately 1 day of being made available.</p> <p>2 - Within approximately 1 week of being made available.</p> <p>3 - Roughly 24 hours before the relevant quiz or exam.</p> <p>4 - I don't watch the video lectures.</p>	1 thru 4	Outside of class
i2	<p>Choose the answer that BEST describes you.</p> <p>When I watch the video lectures ...</p> <p>1 - I am writing code / following along with the instructor. I might go back later and take some notes.</p> <p>2 - I am taking detailed notes.</p> <p>3 - I am multitasking. I am watching other media, doing other tasks.</p> <p>4 - I am skimming for the quiz. I don't watch the full video.</p> <p>5 - I don't watch the video lectures.</p>	1 thru 5	Outside of class
i3	<p>Indicate if you agree with the following statement:</p> <p>When the video lecture has a practical demonstration of a concept (for example, typing a command, or writing some code) I always stop the video and try it myself.</p>	1 thru 5	Outside of class
i4	<p>Indicate if you agree with the following statement:</p> <p>I feel like I spend more time on the video lectures than I should have to.</p>	1 thru 5	Outside of class
i5	<p>Indicate if you agree with the following statement:</p> <p>This "flipped classroom" approach has forced me to spend more time thinking about MIS3502 topics outside of class.</p>	1 thru 5	Outside of class
i6	<p>Indicate if you agree with the following statement:</p> <p>When I come to class I am prepared to work.</p>	1 thru 5	During class
i7	<p>Indicate if you agree with the following statement:</p> <p>When I leave class I feel as though I was able to follow along..</p>	1 thru 5	During class
i8	<p>Indicate if you agree with the following statement:</p> <p>I make a strong effort to follow along with the instructor during class time.</p>	1 thru 5	During class
i9	<p>Indicate if you agree with the following statement:</p> <p>I spend more time writing code (inside and outside of class) because of the "flipped classroom" approach.</p>	1 thru 5	Whole course
i10	<p>Indicate if you agree with the following statement:</p> <p>Future versions of this course should use the "flipped classroom" approach.</p>	1 thru 5	Whole course