Using Machine Learning to Understand Class Data

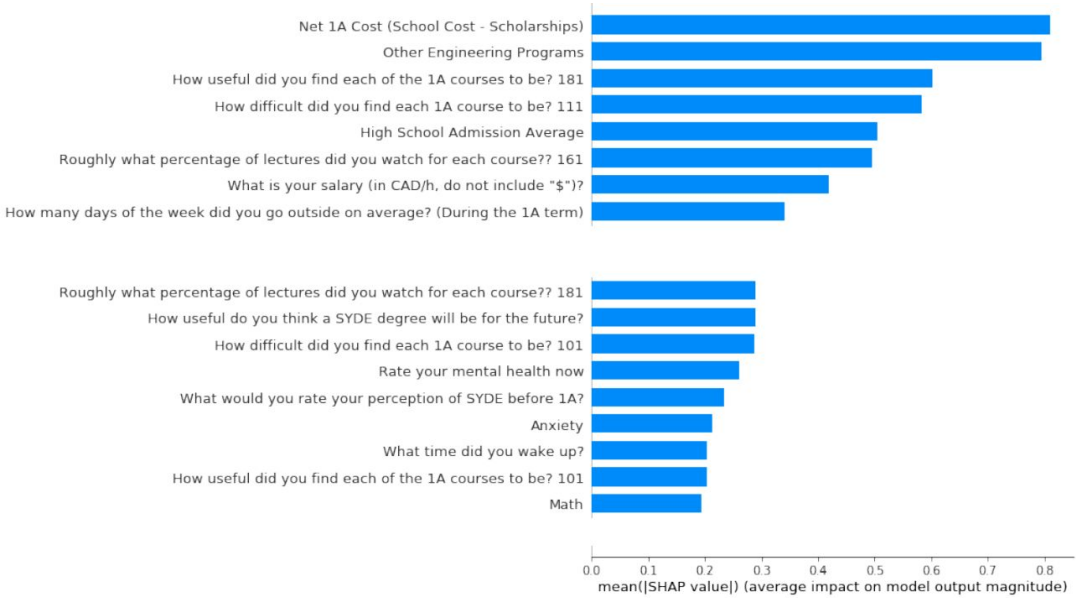
By: Nicolas Palmar

Factors That Predict Academic Averages

A machine learning model was trained to predict SYDE 1A student averages based on most of the survey questions. After training the model, these were the features that had the largest effect on predicting student averages.

The features near the top of this graph affected SYDE 1A student average predictions the most. These are only the top 20 features out of 261.

Note: 3 features have been excluded to prevent singling out any specific groups. Also, the results may vary depending on when the model was trained since this machine learning model includes random elements. For more information on what each feature means, refer to the appendix.



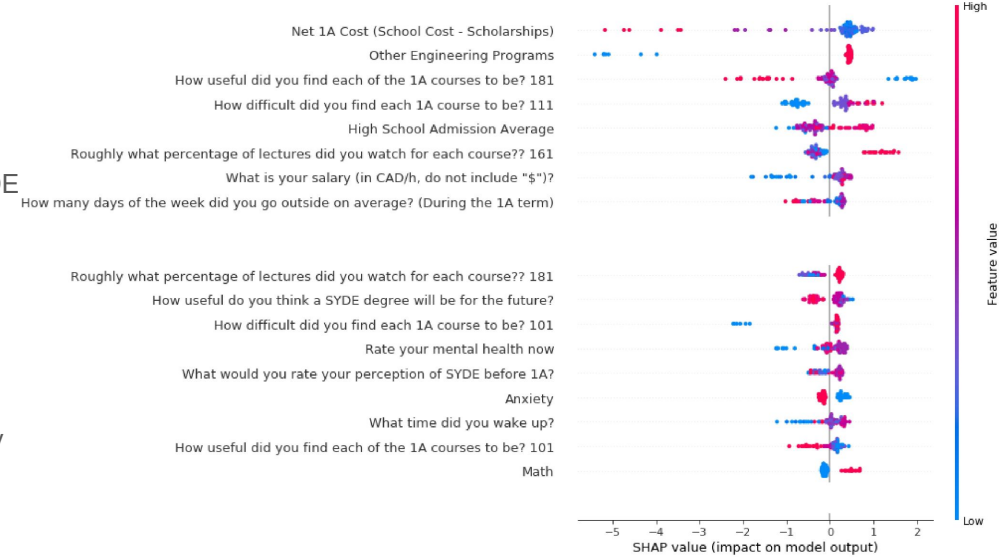
Factors That Predict Academic Averages (Continued)

This graph is a continuation of the past bar graph on feature importance and it shows **how** the past features affect SYDE 1A average predictions.

Here are interpretations for the top 6 features:

- High and medium-high *Net 1A Costs* predicted much lower averages; low and medium-low *Net 1A Costs* predicted slightly higher averages
- Not applying to any *Other Engineering Program* other than SYDE predicted much lower averages, applying to *Other Engineering Programs* predicted slightly higher averages
- Rating *SYDE 181* (Physics 1) as very useful predicted lower averages; rating *SYDE 181* as very useless predicted higher averages
- Rating *SYDE 111* (Calculus 1) as very difficult predicted slightly lower averages; rating *SYDE 111* as very easy predicted slightly higher averages
- In some cases, high *Admission Averages* predicted slightly higher 1A averages
- In general, watching a very high *Percentage of SYDE 161 (Introduction to Design) lectures* predicted higher averages

Please refer to the appendix for more information on how to interpret the rest of this graph.

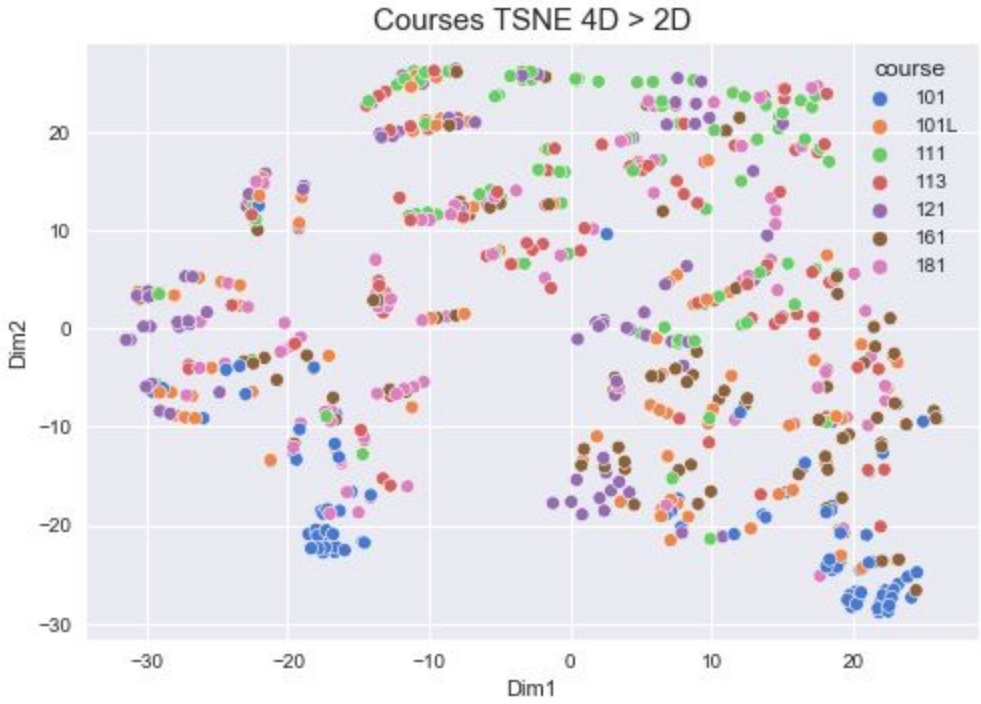


Finding Similar Courses

To find similarities between courses, 4 features/dimensions related to courses (average, percentage of lectures watched, difficulty, and usefulness) were reduced to 2 dimensions and then graphed.

Similar courses should be clustered together in this graph.

Note: The next two slides continue the analysis in greater depth.

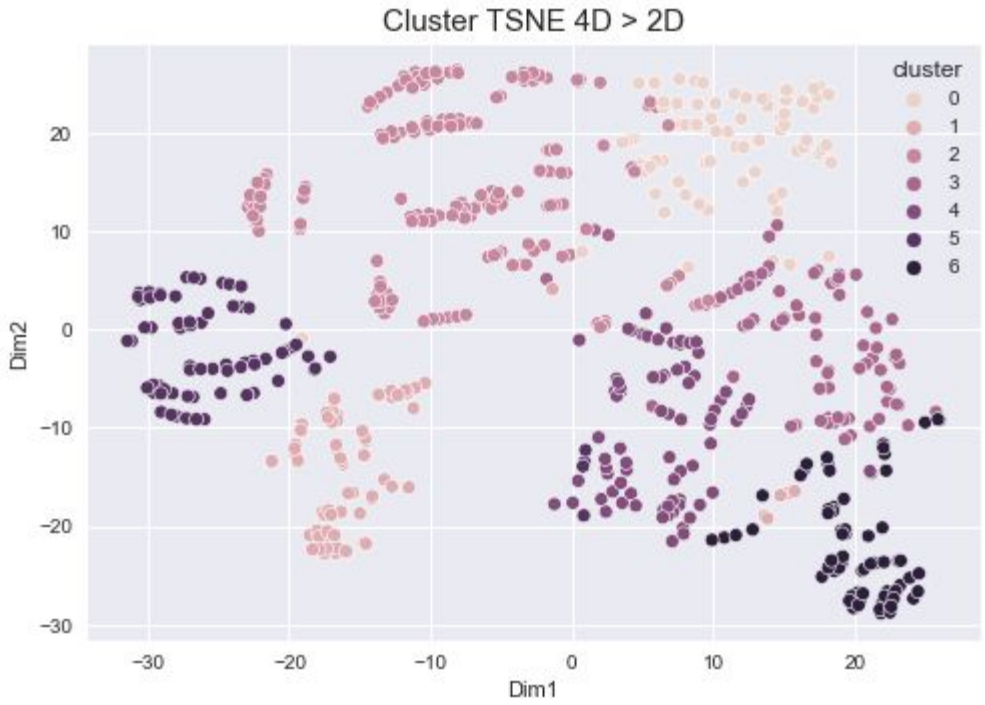


Finding Similar Courses (Continued)

The initial 4 dimensional data was clustered into 7 clusters using a machine learning algorithm. This way, each cluster could be inspected for similar courses.

This graphs shows the same data points as the previous graph, but it highlights the cluster rather than the course.

Finally, each one of these 7 clusters was graphed individually while highlighting the courses in each cluster to identify similar courses (next slide).



Finding Similar Courses (Continued)

Main Courses In Each Cluster (refer to SYDE opinion for each cluster below)

Cluster 0: SYDE 111, SYDE 113, SYDE 181

Cluster 1: SYDE 101, SYDE 181

Cluster 2: Mix of all courses except SYDE 101

Cluster 3: Mix of all courses except SYDE 101 and SYDE 121

Cluster 4: SYDE 101L, SYDE 121, SYDE 161

Cluster 5: SYDE 101L, SYDE 121

Cluster 6: Mainly SYDE 101, some SYDE 161

SYDE 101 - Oral/Written Comm.
SYDE 101L - Visual Comm.
SYDE 111 - Calculus 1
SYDE 113 - Linear Algebra 1
SYDE 121 - Digital Computation
SYDE 161 - Intro to Design
SYDE 181 - Physics 1

SYDE Opinion: Using Domain Knowledge for Analysis

Cluster 0: Mainly contains Math/Physics courses - all of these courses contain mathematical elements
Cluster 1: Both courses seem to have many extreme values for one specific bin; 101 was rated 'easy' by 69% of respondents and 62% of people said they watched '100% of lectures' for 181. Additionally, neither of the courses was rated highly in terms of usefulness
Cluster 2/6: SYDE 101 was generally rated the easiest, least useful, had the 2nd highest grade mean, and 2nd lowest average percent of lectures watched; it was an outlier (didn't appear in cluster 2 and dominated cluster 6)
Cluster 3: SYDE 101 and 121 had the highest course means and medians; best courses academically. Since cluster 3 does not contain SYDE 101 and 121, this cluster may highlight courses with lower course means
Cluster 4: Creative courses where the focus was on building projects; each one of these courses was project/lab based
Cluster 5: These were both of the courses that Professor Igor taught (the only two courses taught by the same professor)

Appendix

SHAP Plot Feature Meanings

NET 1A Cost: 1A School Cost - University Entrance Scholarships

Other Programs: 0 if only applied to Systems Design Engineering (SYDE) for engineering programs in high school, otherwise 1

How useful do you think a SYDE degree will be for the future: Rating from 1 to 10 on student’s opinions on SYDE degree usefulness

Rate your mental health now: Current mental health rating from 1-10

Anxiety: 1 if the student experienced anxiety throughout the 1A term, otherwise 0

Math: 1 if the student applied to math programs in high school, otherwise 0

Factors That Predict Grades: Extra Info

The further left a point was, the more it predicted lower averages. Conversely, the further right a point was, the more it predicted higher averages. Points near zero (near the y-axis) did not affect the predictions much.

The colour of a point specifies if the point is a higher value or lower value for its respective row/feature. For example, red points in the ‘Net 1A Cost’ feature represent high costs, purple point represent medium costs, and blue points represent low costs.

For a more in depth explanation, refer to:
https://www.kaggle.com/dansbecker/advanced-uses-of-shap-values