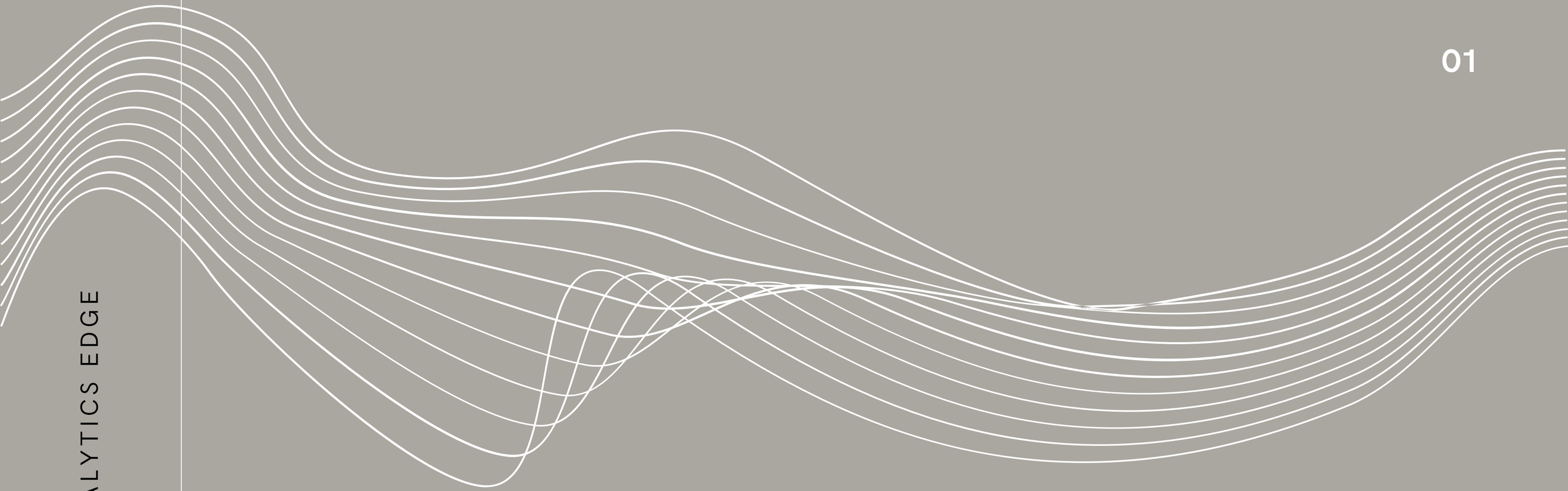


15.071 - THE ANALYTICS EDGE

Integrated Design & Management (IDM)

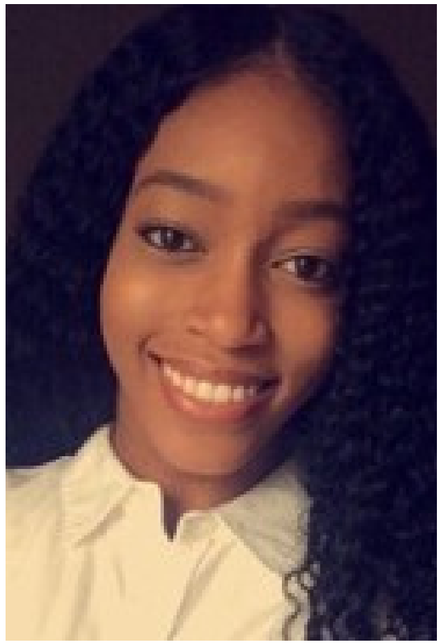
Course Recommendation Engine

The Team

ETHAN CARLSON
IDM '21

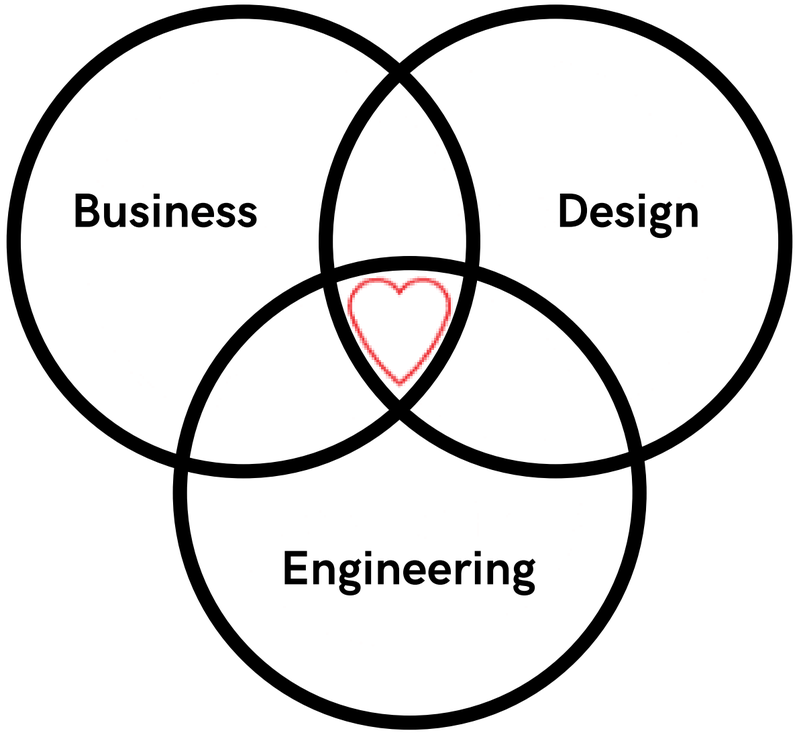
ANGELICA CHINCARO
IDM '22

JAKE DRUTCHAS
IDM '22

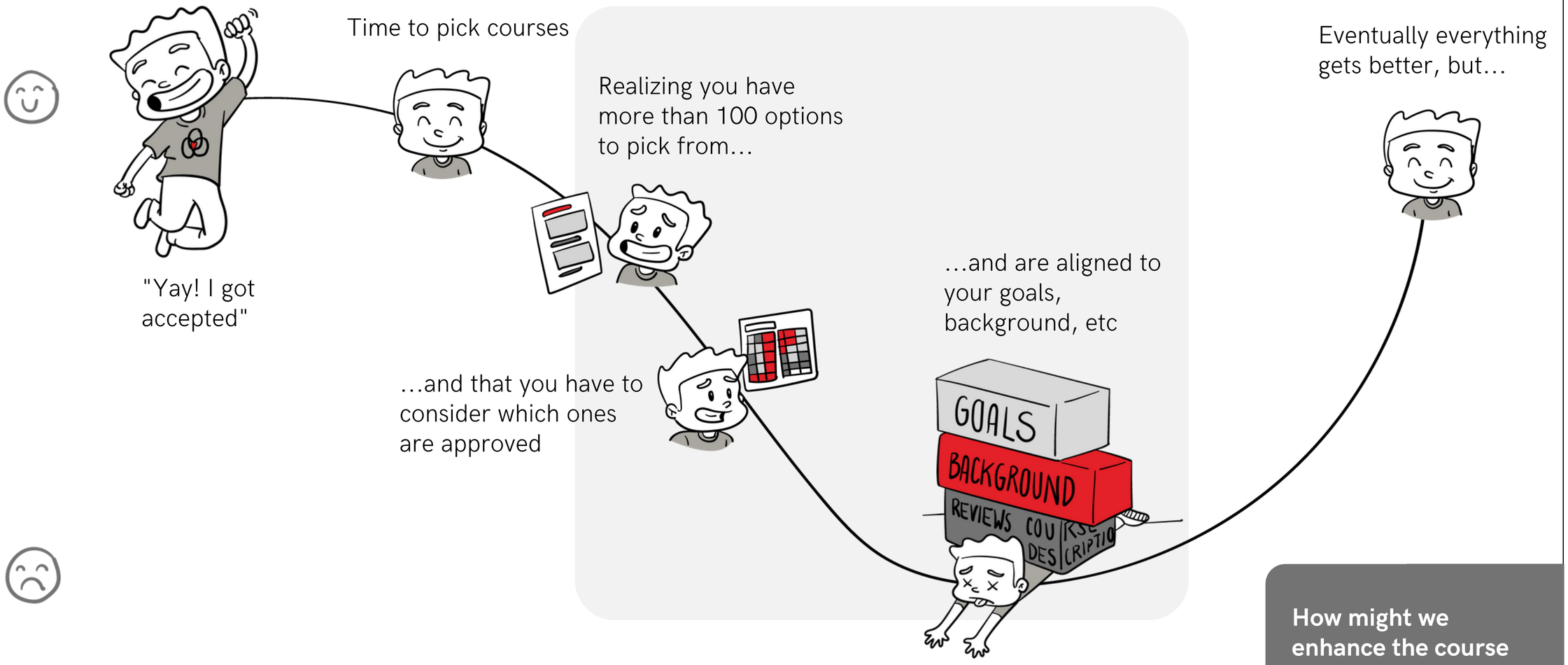


CHINELO ONUOHA
IDM '22


MITidm


Business
Design
Engineering

Picking a course @IDM



How might we enhance the course selection experience?

Project Brief

IDM gathers students with backgrounds in design, business, and engineering for a dynamic two year program. Because of the varying nature of interests and access to such a wide ranging course catalog the course selection process is often overwhelming and dissatisfying.

Our Goal: To build a recommendation engine to help provide insight into potential classes of interest.

Methods Used:

- XGBoost
- Logistic Regression
- Ensemble
- Hierarchical Clustering

Pursuing both clustering and predictive analysis to construct a tool for future students

The Data

IDM HISTORICAL MATRICULATION DATA

All courses current and former students have taken

MIT COURSE REVIEWS AND DESCRIPTIONS

Population level course reviews and course descriptions

SURVEY

Categorical data on individuals and class experiences

LINKEDIN

Career history

Cleaning & Collation

01 **INCOMPLETE DATA**
Manual search and entry

02 **CLEANING / COMBINING DISPARATE DATASETS**
Strategies for combining sparse datasets

03 **STANDARDIZATION / REGULARIZATION**
Maximizing effectiveness of data

Assumptions/Challenges

01

MODEL ASSUMPTION

Every time someone took a class they were happy they took it and every time they didn't take a class they were happy they didnt. The model is predicated on this idea of no regrets.

02

MISSING GROUND TRUTH LABELING

To recommend future classes we want to be able to predict beyond the historical binary, but without the course review data we were limited in our potential.

03

DATA COLLECTION

Collecting the data from disparate sources and data owners was extremly time intensive.

Analysis

Methods Used:

- XGBoost
- Logistic Modeling > Ridge, Lasso
- Ensemble > XGB into Lasso
- Clustering > Hierarchical

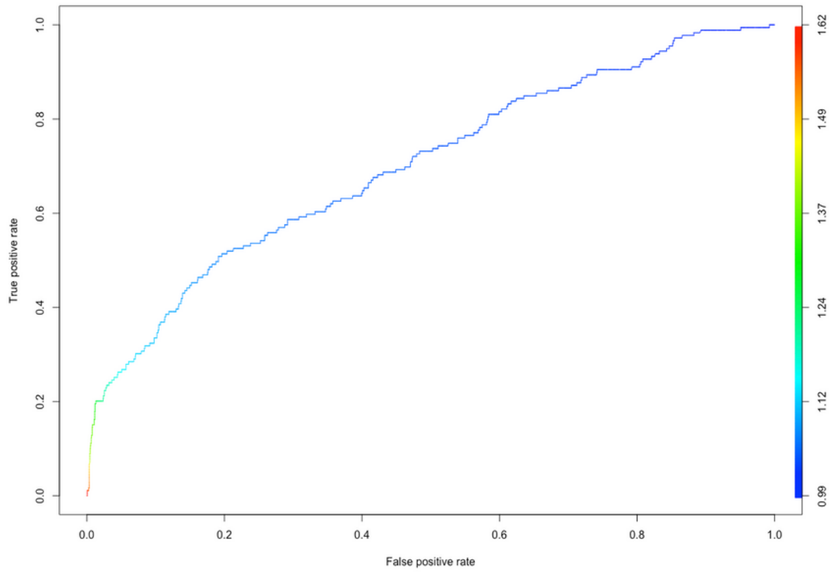
XGBoost (XGB)

BASE MODEL

AUC 0.710

- Parameter manipulation
  - max_delta_step
- We found that overall XGB prefers more information
  - 200 bag of words columns outperformed top 60
  - Counting instances of each word performed better than just using binary presence/absence
- The optimal tree depth was 3
  - We attempted to validate for a larger depth but at 13 layers our computers were unable to cope

Top variables:
- Matriculation year
- Years worked
- Eligible respondents



Logistic Model

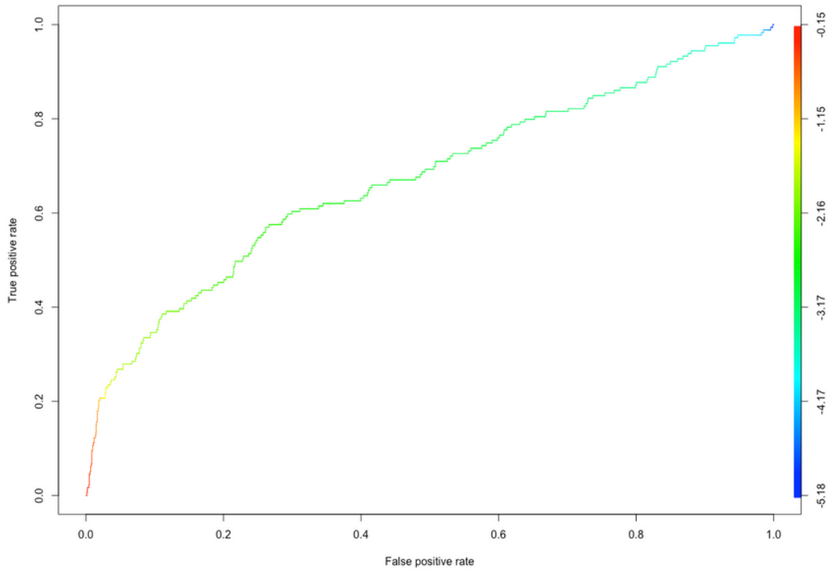
LASSO REGRESSION

AUC 0.673

- Tried both Lasso and Ridge regression
- This use of logistic regression also provided probabilities for each course that could be analyzed at various cutoffs to provide tranches of recommendations

Top 5 variables (of 88 total):

- AssignmentsContributedMean
- LearningObjectivesMetSD
- 'policy'
- 'primarily'
- 'implement'



Ensemble Model

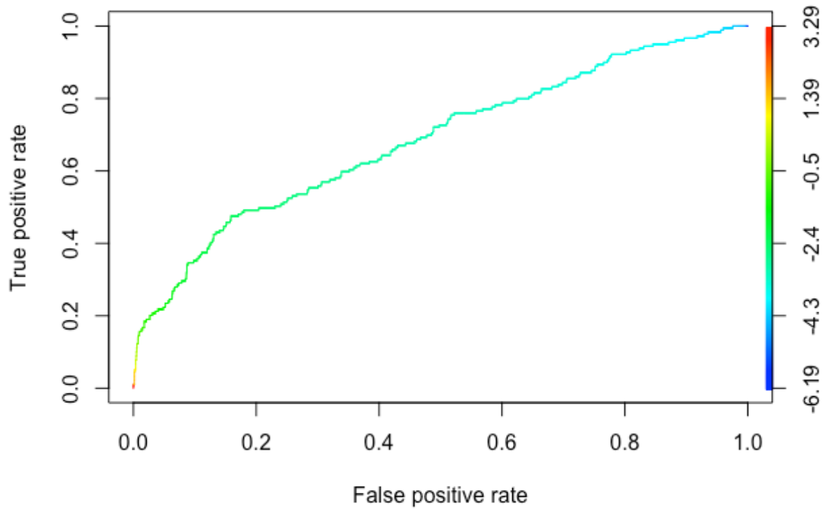
XGB > LASSO

AUC 0.725

- 2 Layers ensemble XGB feeding into LASSO

Top 5 variables (of 42 total)
  - XGBPreds
  - 'control'
  - 'simulation'
  - 'policy'
  - 'broad'

- This model in its current structure is well suited to improve over time with improved data.



Ensemble Findings

01

TOP 5 COURSES RECOMMENDED FOR NEW IDM STUDENTS

Rank	Course #	Course Name	Avg Score	Deviation
1	15.665	Power and Negotiation	-0.06	0.57081
2	16.453	Human Systems Engineering	-0.59	2.29394
3	15.871	Introduction to System Dynamics	-0.84	2.01354
4	15.846	Branding	-1.45	0.67186
5	2.888	Global Mfg & Entrepreneurship	-1.5	0.72279

02

TOP 3 COURSES WITH THE MOST GENDER DISPARITY

Rank	Course #	Course Name	MaleScore	FemaleScore	AbsValue
1	16.453	Human Sys Engineering	-1.336987338	-0.075215963	1.261771375
2	16.842	Fundamentals of Systems Engr	-1.625736703	-2.225417355	0.599680651
3	IDS.333	Risk and Decision Analysis	-2.003456451	-1.603555827	0.399900623

03

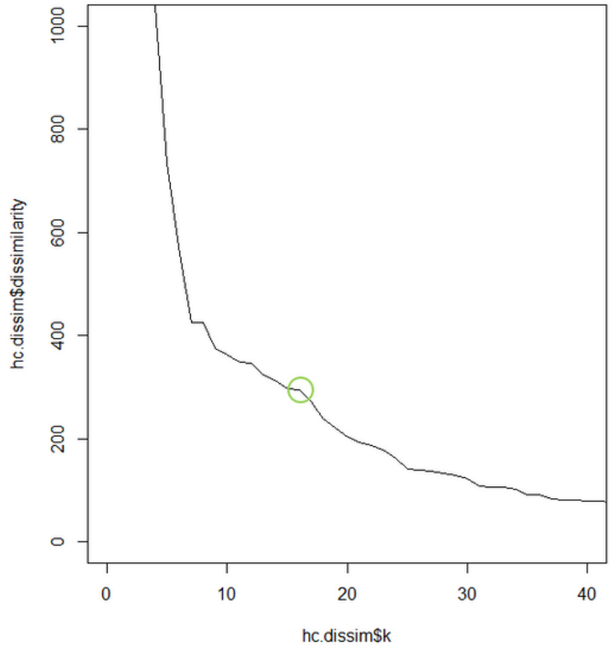
TOP 3 RECOMMENDED FOR STUDENTS WITH BUSINESS BACKGROUNDS

Rank	Course #	Course Name	Avg Score
1	15.871	Introduction to System Dynamics	0.2620758112
2	15.665	Power and Negotiation	-0.220395629
3	15.846	Branding	-1.513173445

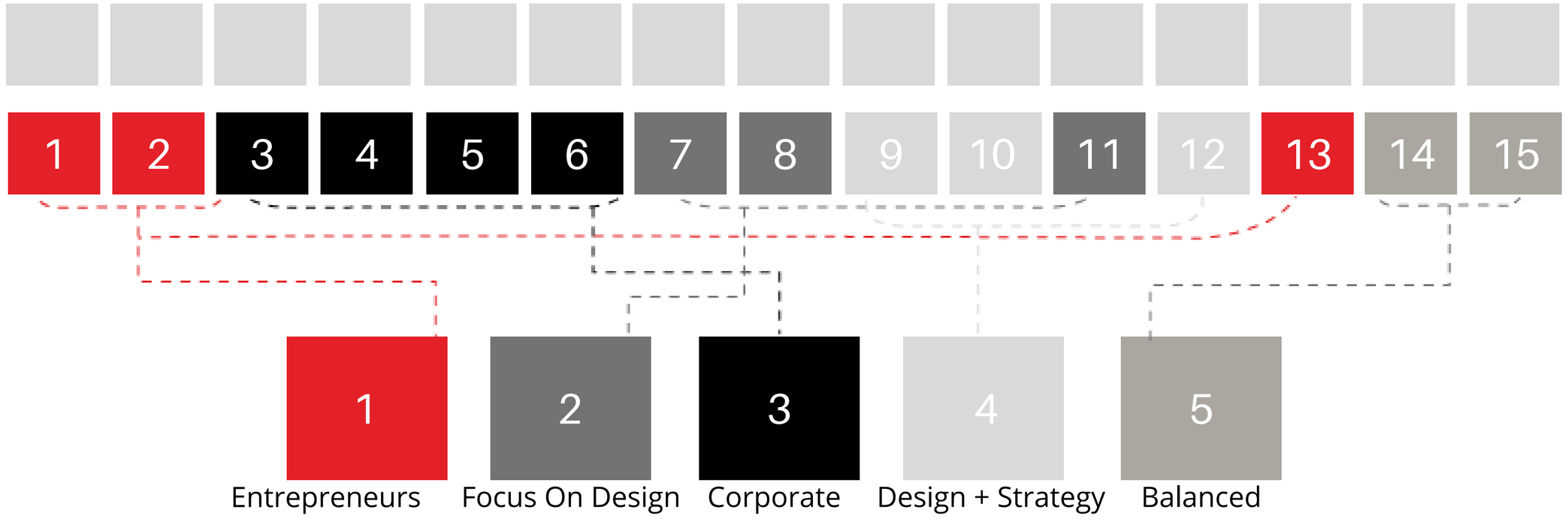
Clustering
Strategy

Using a hierarchical clustering approach, we analyzed the relationship between students and courses with the significant variables obtained after running Lasso + XGBoost.

As we can observe in the scree plot, **15 clusters** was a good number to start clustering the data.

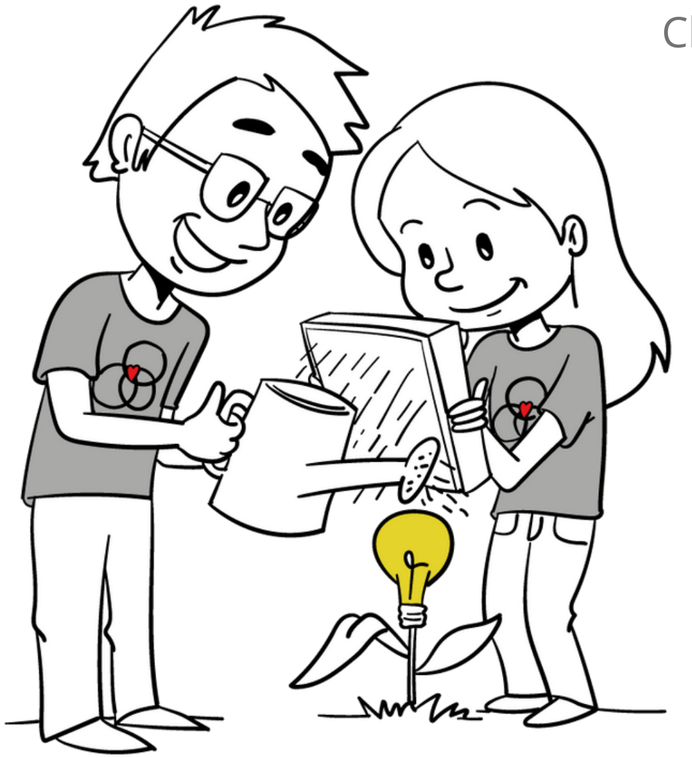


We observed the commonalities between each cluster and narrowed them to 5.



Clustering

The Entrepreneurs



Characterized for having the biggest percentage of entrepreneurs (founders, co-founders).

30%

vs 12% (average)

We found a sub-profile within this group, related to students with engineering background, focused on learning about the entrepreneurship world.

Background

B

D

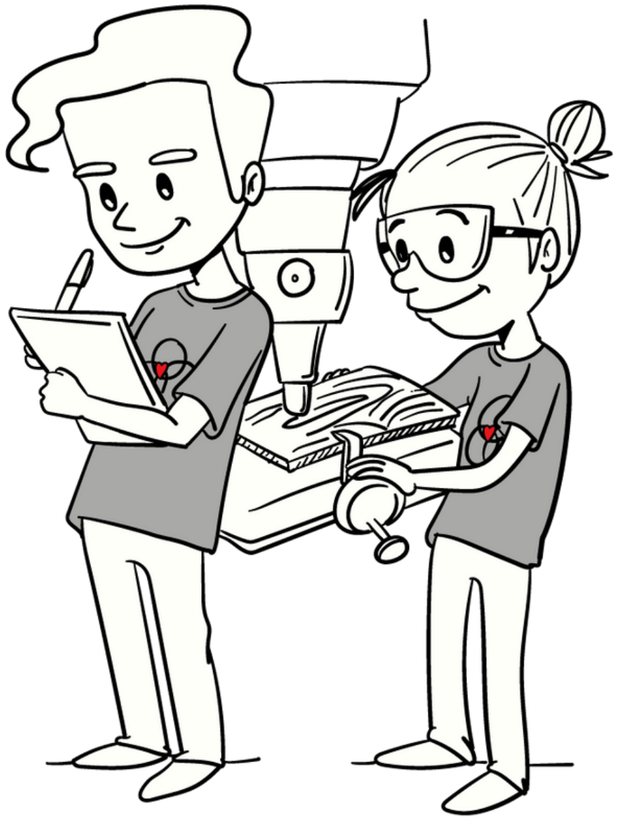
E

Most taken courses

- Power and Negotiation
- Human Factors Engineering
- Intro. to System Dynamics

Clustering

Focus On Design



Less presence of business background.

They love design and want to continue working in that type of role.

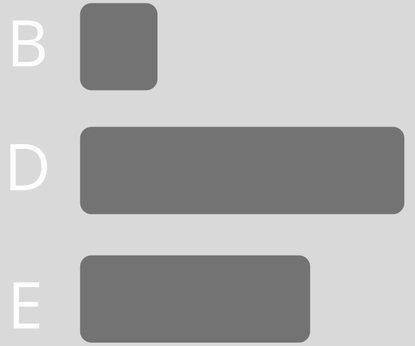
Previous and current roles as designers (Industrial, UI/UX)

40%

vs 21% (average)

Preference for non-business oriented courses.

Background

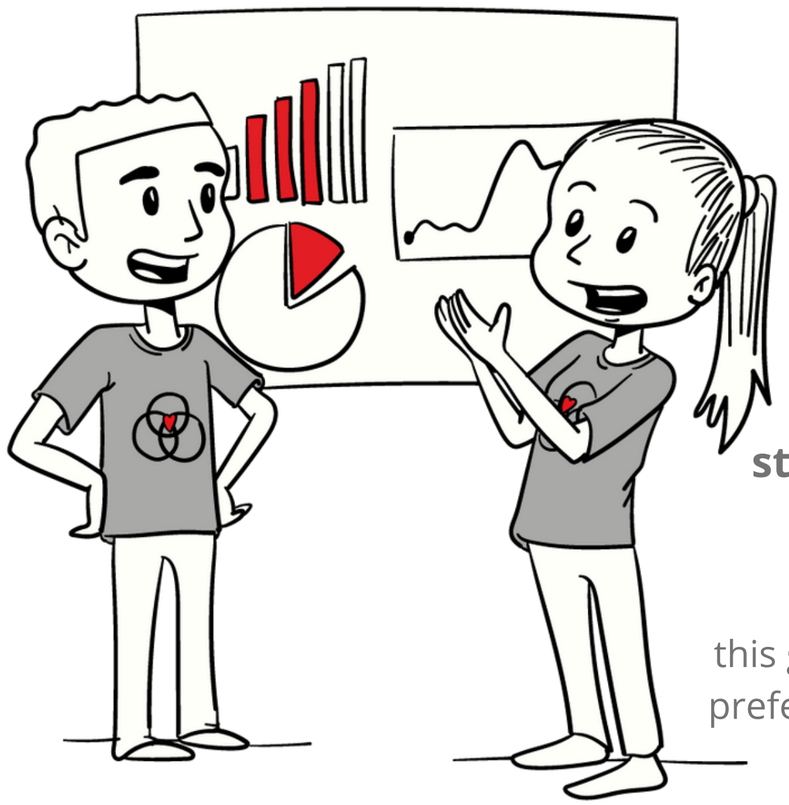


Most taken courses

- Computational Science and Engineering
- Data, Systems, and Society
- Ethics

Clustering

Corporate



Not interested in pursuing entrepreneurship (0% of current roles)

Preference for entering a corporation in **more strategic roles**, as a designer or engineer.

We found a subprofile within this group, of engineer women with high preference for courses that enhance their business toolbox.

Background

B

D

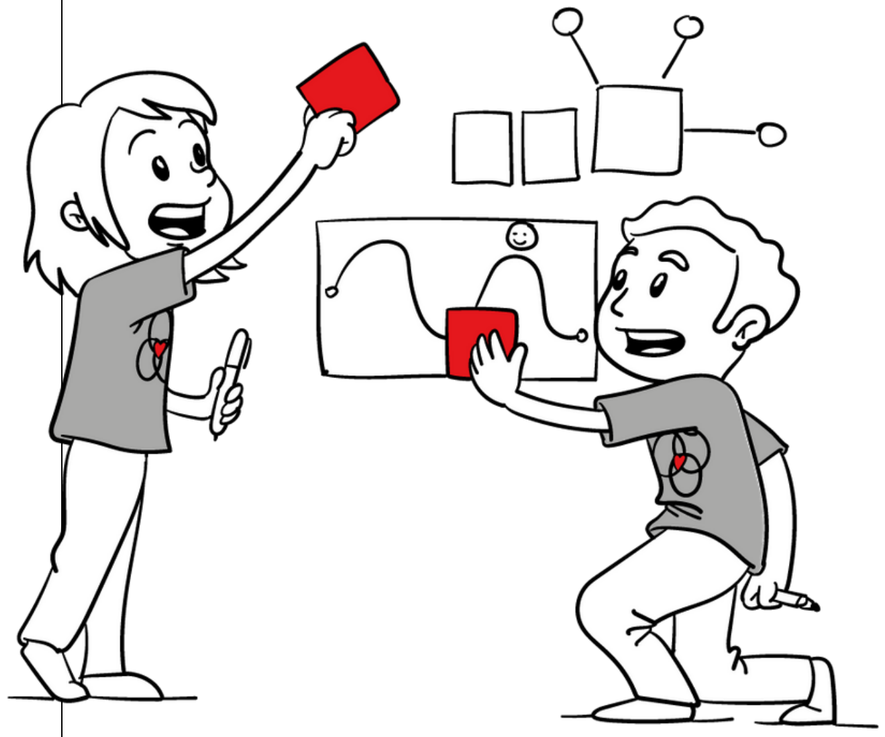
E

Most taken courses

- Supply Chain Management
- Power and Negotiation
- Economic Analysis for Business Decisions

Clustering

Design + Strategy



Interest in obtaining **design strategic roles.**

Less presence of engineering background.

Background

B

D

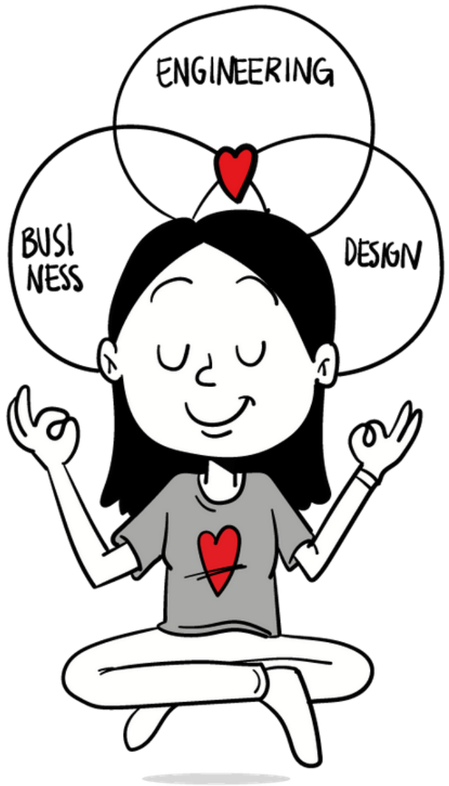
E

Most taken courses

- Engineering Systems Analysis for Design
- Manufacturing System and Supply Chain Design
- Business and Operations Analytics

Clustering

Balanced



Group with no clear preference for a role or background.

Interested in many roles with a **multidisciplinary approach.**

Background

B

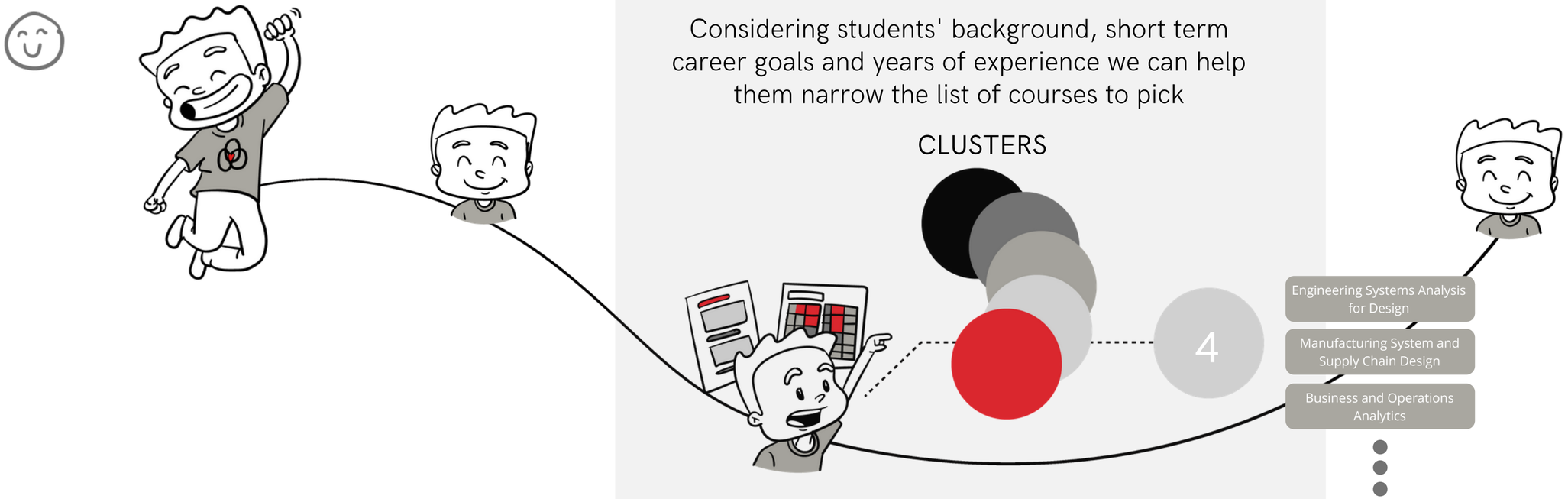
D

E

Most taken courses

- Improvisational Leadership
- Introduction to Operations Management
- Power and Negotiation

How can we make the journey less ovewhelming?



Considering students' background, short term career goals and years of experience we can help them narrow the list of courses to pick

CLUSTERS

4

Engineering Systems Analysis for Design

Manufacturing System and Supply Chain Design

Business and Operations Analytics

For example if a student with a designer background that wants to move into strategy is selecting courses, maybe looking at the courses of cluster n°4, will help him/her focus on specific courses. Making the selection less overwhelming and efficient.

Next Steps

APPENDIX

Takeaways

01 WHICH CLASSES ARE EASY TO CATEGORIZE VS. DIFFICULT

- Classes that are less popular are easier to categorize (the model always recommends not to take)
- Classes that have a favorable ratio of rating to rating difficulty are:
  - Power and Negotiation
  - Human Systems Engineering
  - Intro to System Dynamics
  - Branding
  - Seminar in Global Manufacturing

Operationalizing Outcomes

01

RICHER DATA COLLECTION
To improve outcomes we must improve data collection from future cohorts

02

EXTRAPOLATION
How do you recommend classes that have never been taken?

03

DATA INPUT
What data do we need to collect from students to improve recommendations and what is that mechanism?

04

INTUITIVE FRONT END
Tying it all together to maximize value

Using the tool!

In the current state the tool is best suited for students with no completed course history taking in model and cluster specific categorical data to provide recommendations.