Text Mining and Neural Networks

Group 17:
Yuxiao Liu
Jiashu Miao
Jiaye Pan
Breanna Qin
Sidong Wang
Chelsea Zhang

Part 1: Text Mining

- Research Question
- Data Manipulation
- Modeling and Classification
- Interpretation and Conclusion

Research Questions

1. We are using text mining on the STEM survey open-ended comments to extract meaningful words while conducting frequency analysis to determine students' view on help and inhibit factors
2. We are trying to see if the words can predict a students' transfer/non-transfer status, major and gender and provide insight based on the prediction.

Variables

- HelpFactor & InhibitFactor
- Approaches:
  - Word lists by Frequency
  - Word Count Table + background factor (e.g. stem, transfer, gender)

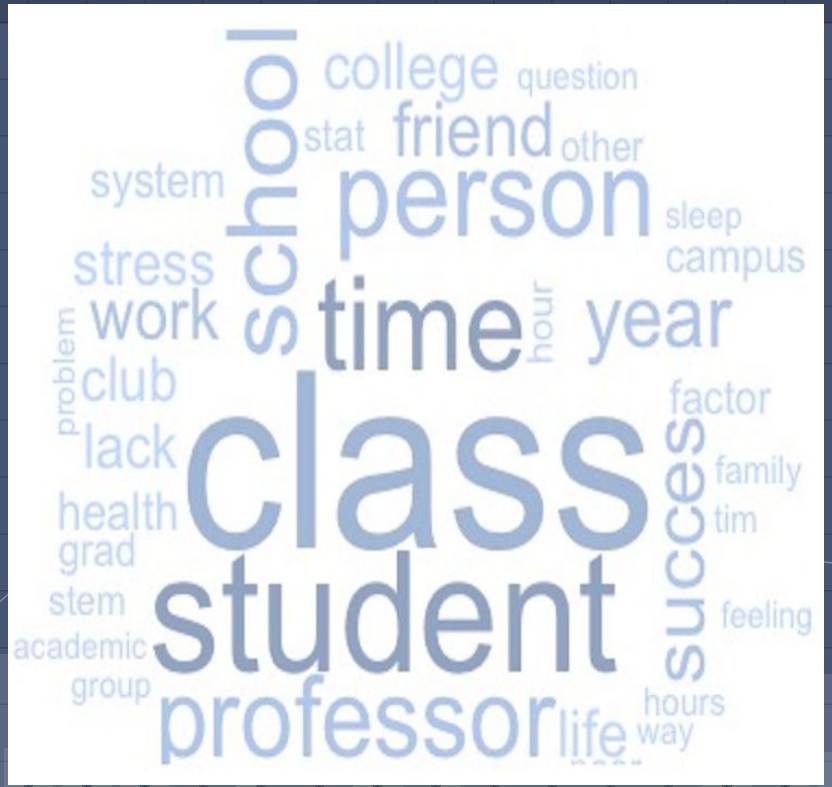
HelpFactor	InhibitFactor
Support Services and programs on campus like the Transfer ...	The transfer process was very difficult for the first year-havi...
The faultiest are nice and wish to help, my friends help me a...	I guess myself inhibited my success at UCLA, since I am not ...
The main factor that helped me was through building com...	Myself. I am my biggest enemy, weather it is being too scar...
I have to say that my pre-college education prepared me a l...	The most significant difficulty that comes to me is to figure ...
Definitely parental support. Without them it'd be possible.	Definitely parental support. Without them it'd be possible.
Having a support group in my friends has helped me imme...	Sometimes when I go to office hours, I'm scared to explain ...
I think my primary education played a huge role in my succ...	Although transitioning from high school to college was a br...
To be not procrastinate is really important in finishing tasks ...	procrastination sometimes inhibits my success and ruins my...
Love and support from family, friendships, guidance and me...	1. Many UCLA Clubs and Greek Life (especially professional ...
Finding the right friend group that I vibed with changed my ...	Professors who have a difficult time communicating/giving l...

Data Manipulation

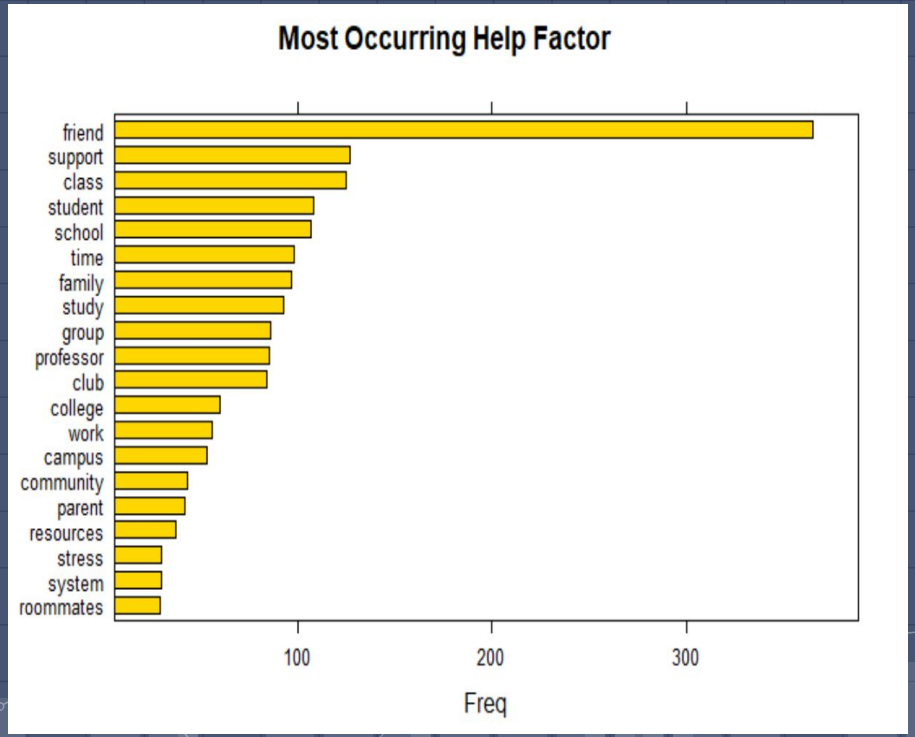
- Upper Case to Lower Case
- Punctuation
- “Non-breaking space” & “\n”
- Stopping words
  - "I", "me", "they", "what", "have"
- Words with similar format
  - “Service”, “Services” and some misspelling words
- Manually Selection
  - “success”, “etc” , “better”, “day” and vague words

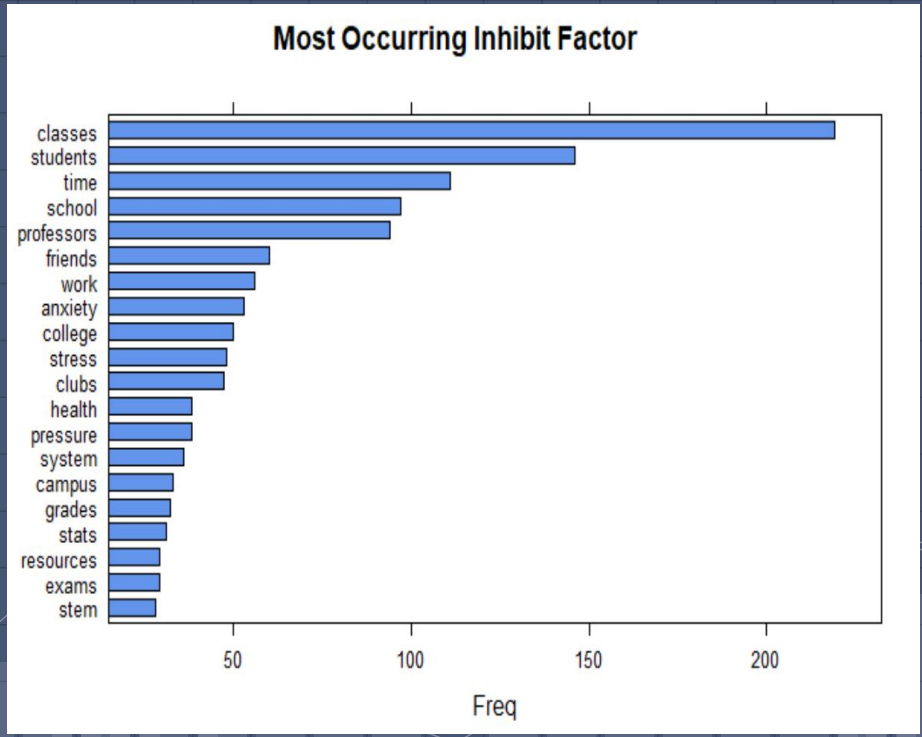
Word Cloud


roommates
experiences
advice
resources
college
group
community
factor
event
sleep
campus
timeschool
way
family
person
club
system
friend
activity
year
peer
support
class
life
health
opportunity
faculty
parent
student
part
succes
program
stem
stress
study
work
experience
professor
organization


college
question
school
stat
friend
other
system
person
sleep
campus
stress
work
time
hour
year
problem
club
factor
class
lack
family
succes
health
tim
grad
stem
feeling
student
academic
group
professor
hours
life
way

Word lists by Frequency Analysis





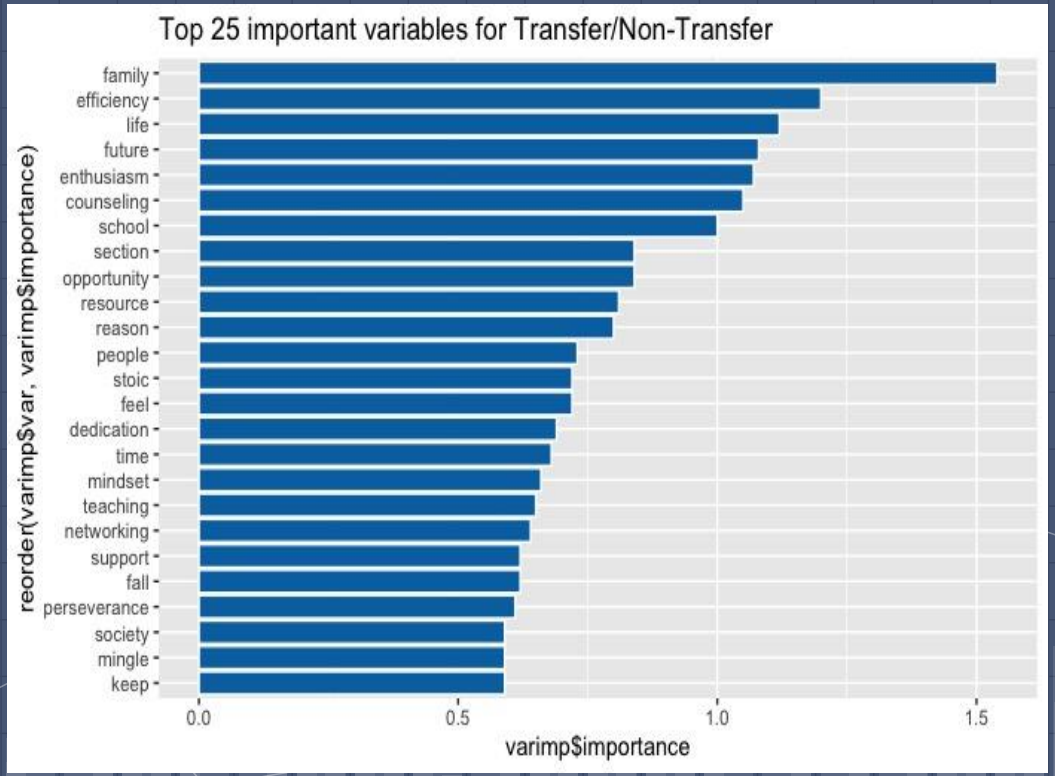
Word Count Table

- Prepare to determine the relationship

obs	anxiety	success	culture	language	professor	character	mindset	exam	homework	information	level	major
1	1	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	1	1	0	0	0	0	0	0	0
4	0	0	0	0	0	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	1	1	0	0	0
6	0	1	0	0	0	0	0	0	0	1	1	1
7	1	0	0	0	0	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	1	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	4
13	0	0	0	0	0	0	0	0	0	0	0	0

Transfer
No
No
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
No
No

Prediction -- Transfer/Non-Transfer

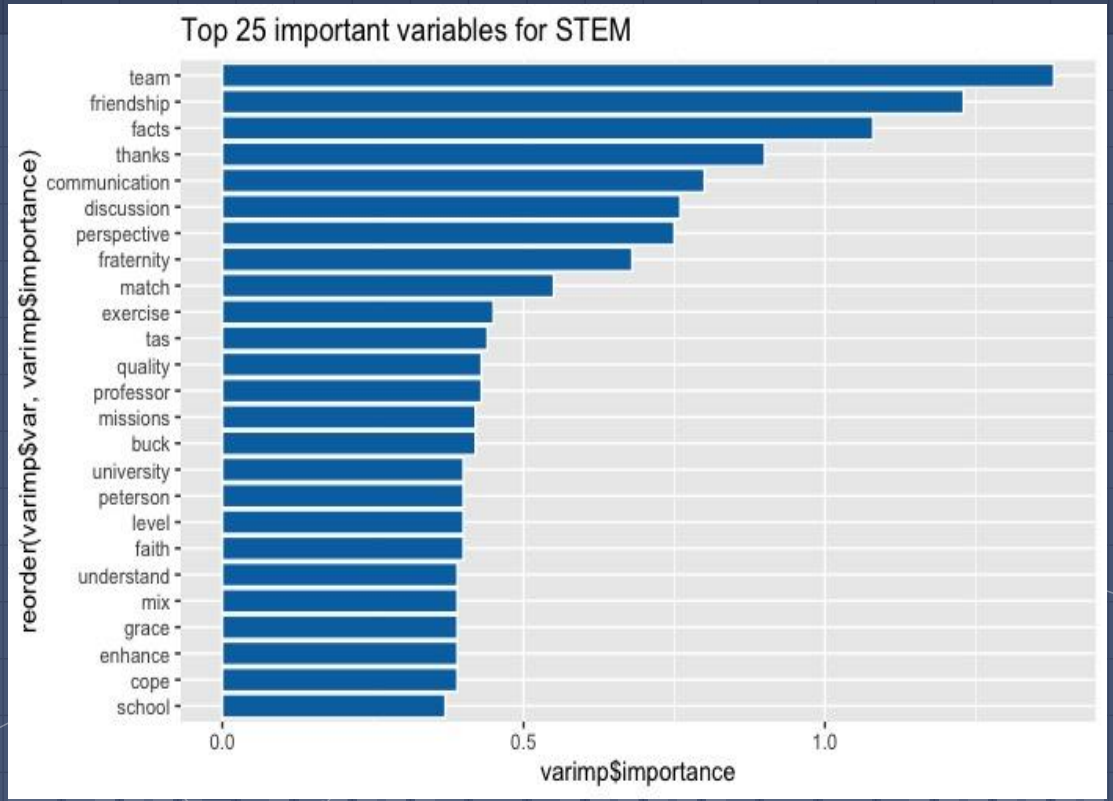



```
          Reference
Prediction   0   1
         0 135  33
         1   2   1

               Accuracy : 0.7953
                 95% CI : (0.727, 0.8531)
    No Information Rate : 0.8012
    P-Value [Acc > NIR] : 0.6198
```

From successfully classifying students into based on their comments on help factors, we can see that transfer and non transfer students are affected differently by these distinguishing factors. Non-transfer students are more likely to mention family, opportunity and resources, suggesting that we might need to provide more resources and support for transfer students.

Prediction -- Stem/Non-Stem

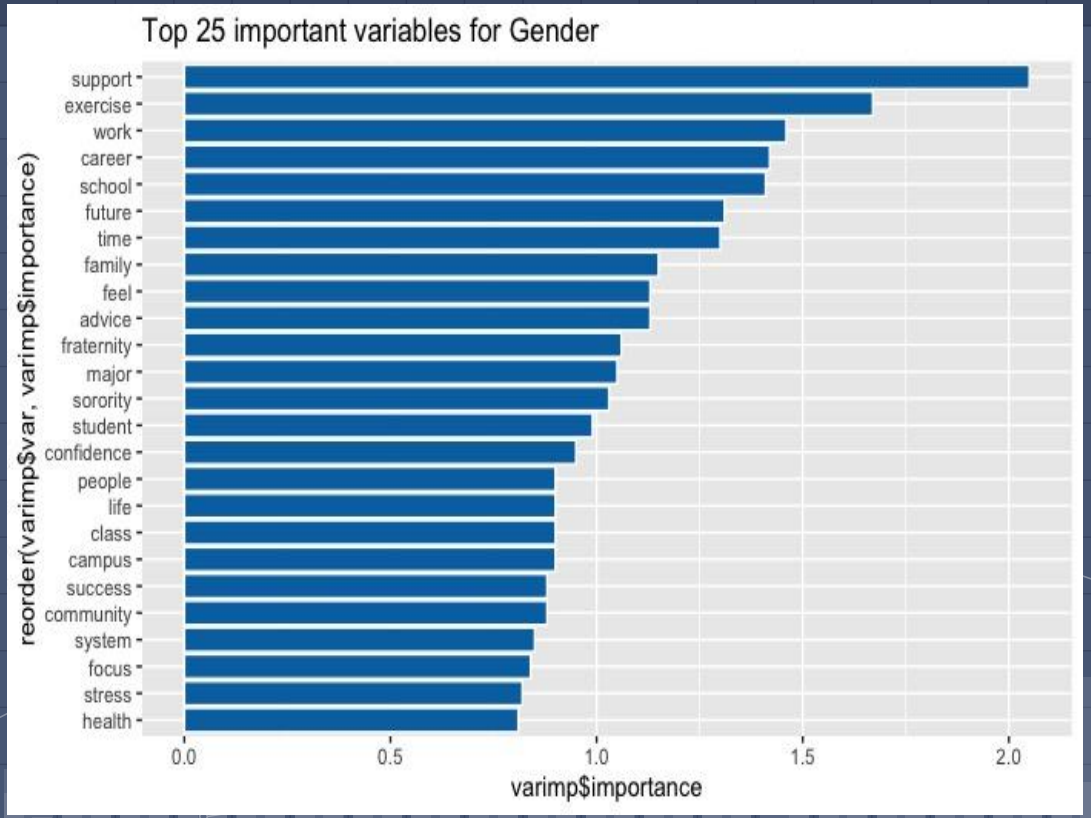



```
          Reference
Prediction   0   1
         0   0   0
         1  14 157

               Accuracy : 0.9181
                 95% CI : (0.8664, 0.9545)
    No Information Rate : 0.9181
    P-Value [Acc > NIR] : 0.570501
```

From successfully classifying students into based on their comments on help factors, we can see that stem and non stem major students are affected differently by these distinguishing factors. Non-stem students are less likely to mention these top words.

Prediction -- Gender




```
          Reference
Prediction  0  1
         0 15 17
         1 61 78

               Accuracy : 0.5439
                 95% CI : (0.4661, 0.6201)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 0.6506
```

For gender, we don't have a very high accuracy rate to correctly classify students by gender. Top factors include exercise, work, fraternity and sorority.This result may suggest that male and female students tend to mention the similar help factors in these comments.

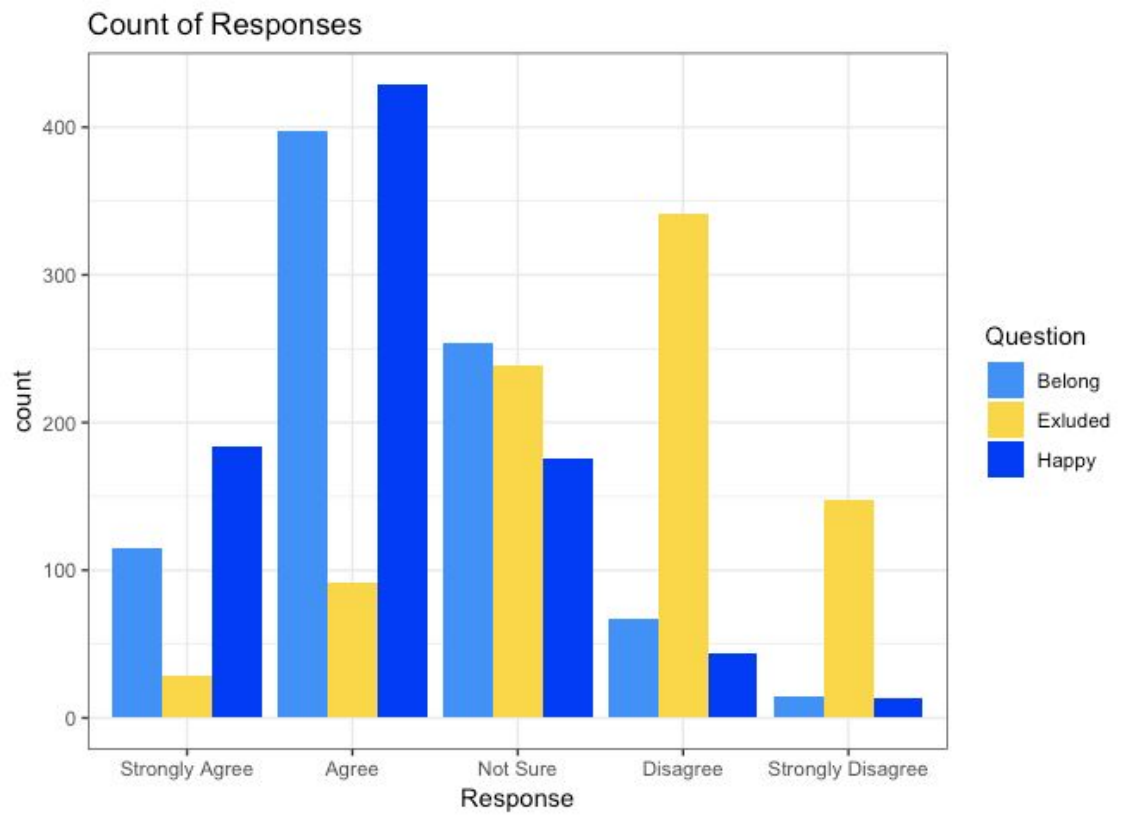
Part 2: Neural Networks

- Research Question
- Exploratory Analysis
- Variables
- Modeling, prediction, and visualization
- Conclusion and recommendations

2

What factors contribute to student sense of belonging? How can we predict this?

Exploratory Analysis



There is an inverse relationship between student sense of belonging and exclusion, as expected.

There also seems to be a positive relationship between sense of belonging and happiness at UCLA.

Recode Variables

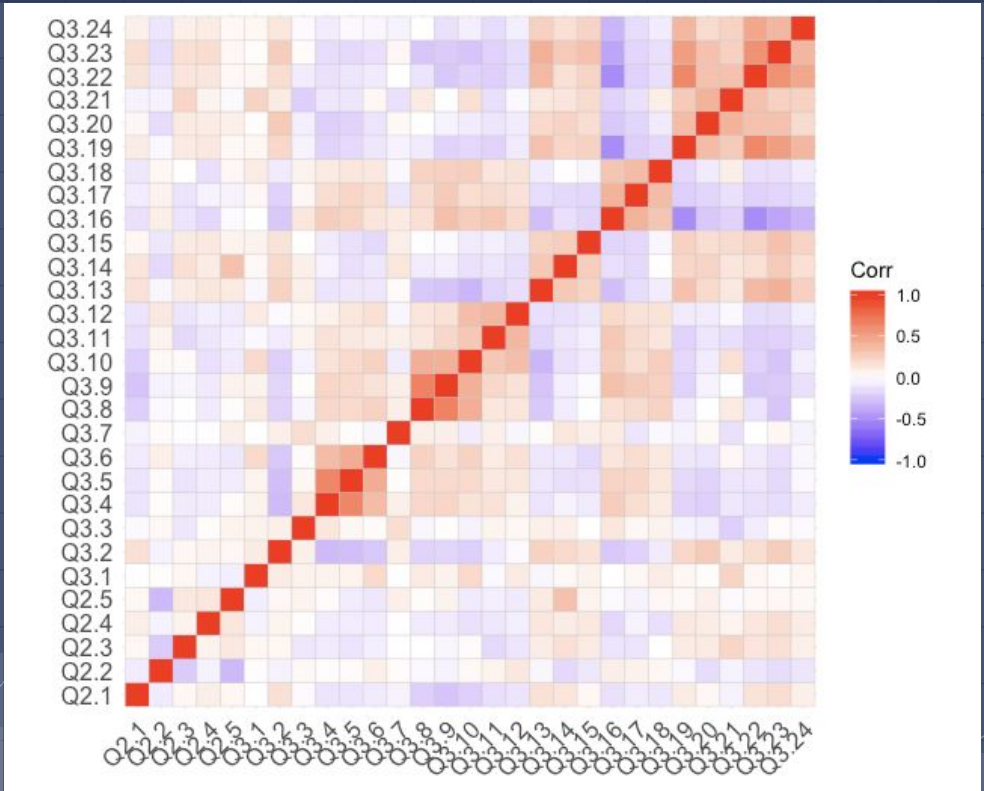
Question 2	Question 3
Never → 1	Strongly Disagree → 1
Rarely → 2	Disagree → 2
Sometimes → 3	Not Sure → 3
Often → 4	Agree → 4
Always → 5	Strongly Agree → 5

Categorical → Numeric
Neural networks handle numeric values

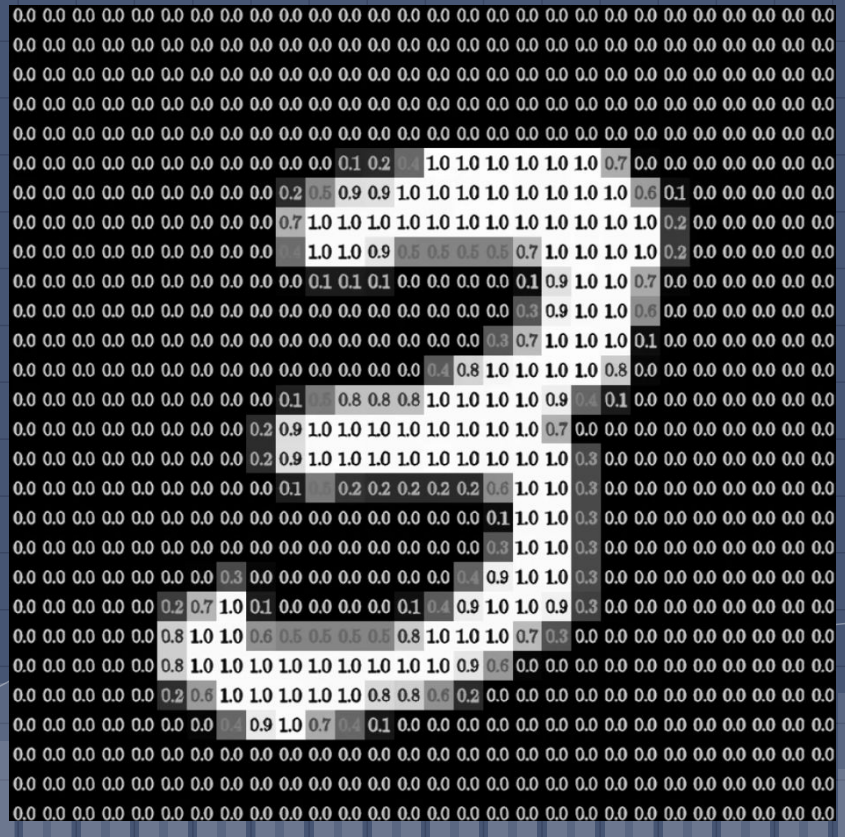
- Gender, International, Transfer, FirstGen, STEM: Yes, No → 0,1
- Year: Freshman, Sophomore, Junior, Senior → 1, 2, 3, 4

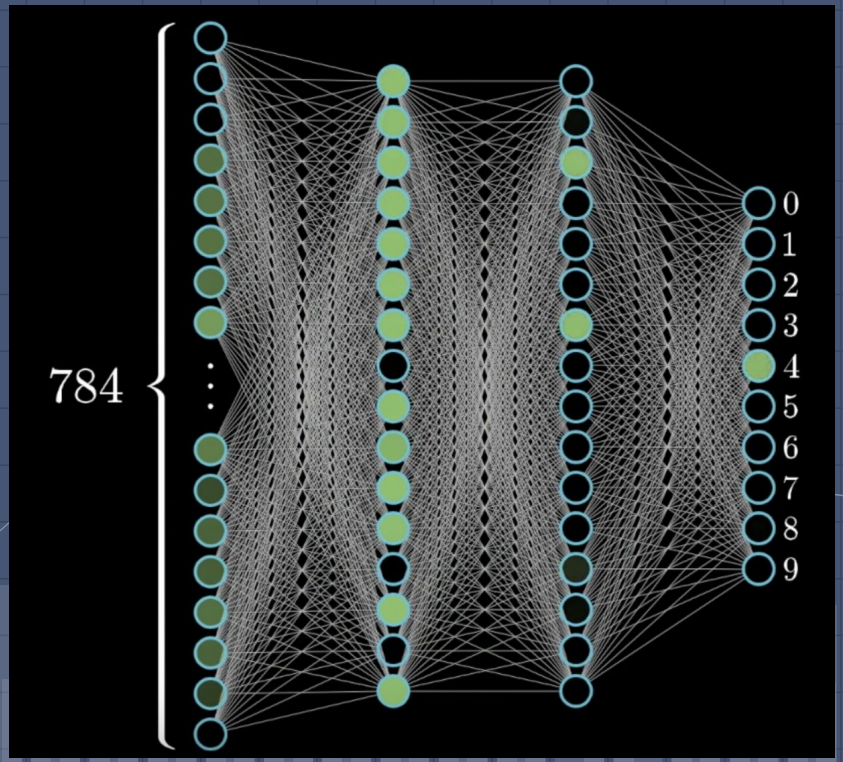
Outcome: Q3.19 (I feel I belong at UCLA)
Predictors: All other variables

Correlation

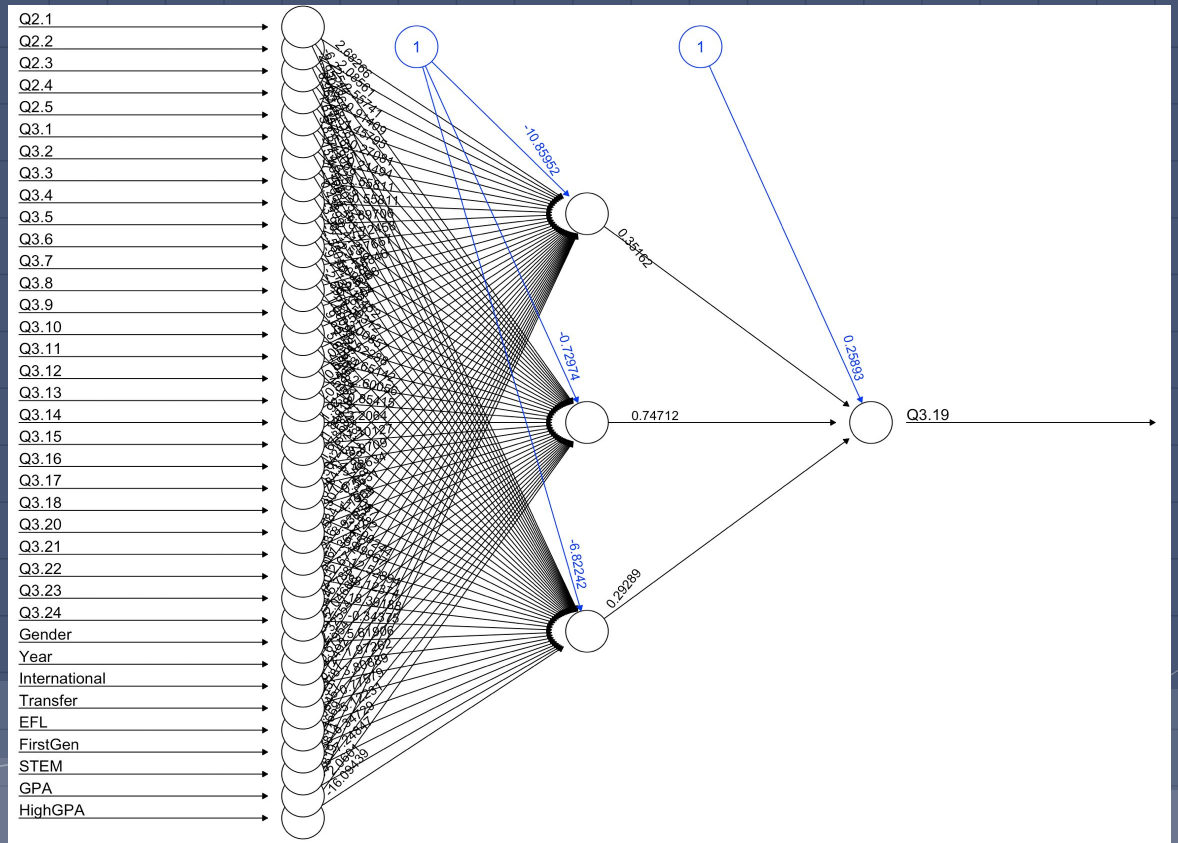


Neural Networks: Motivation

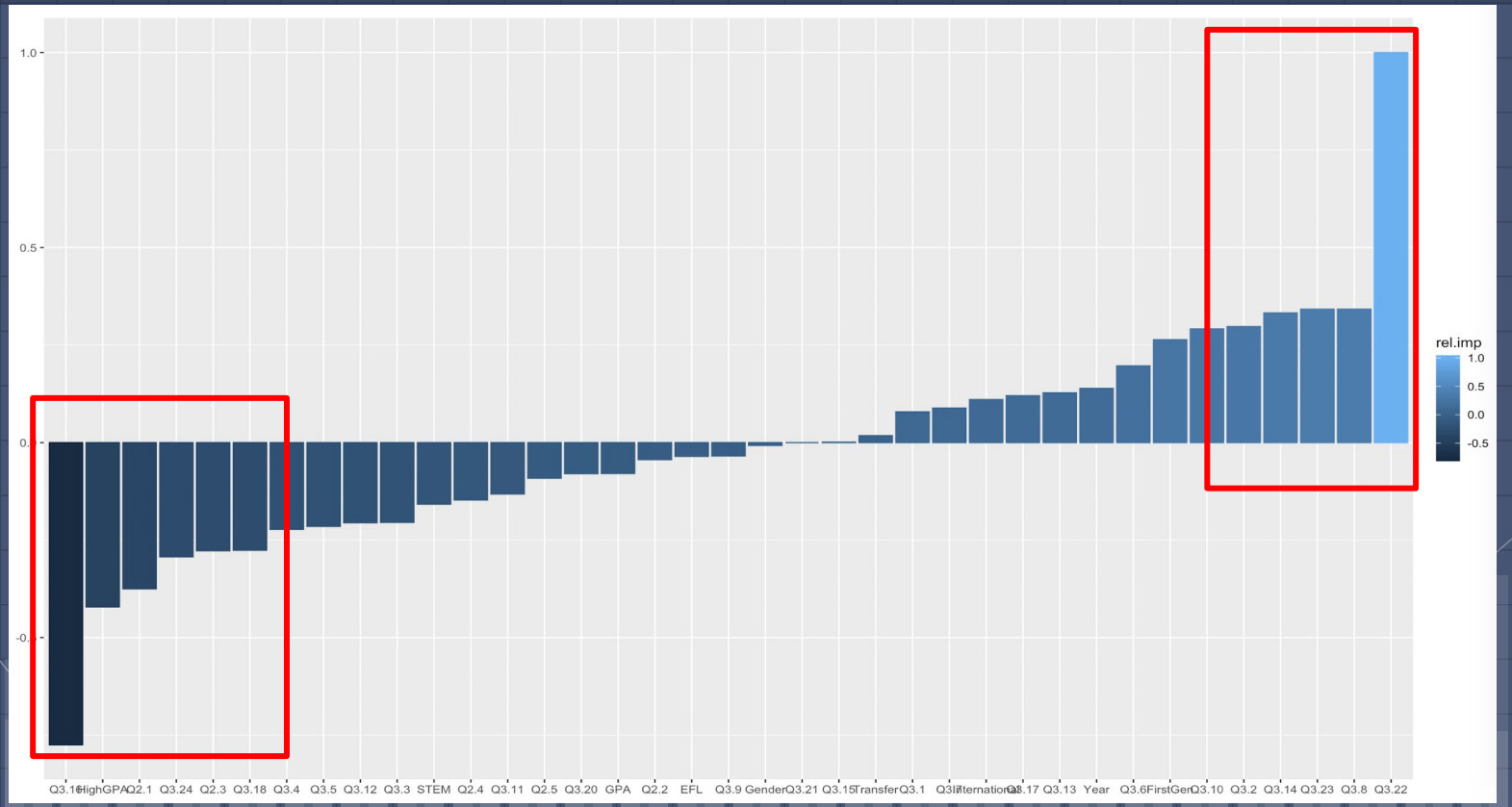




Plotting the Network - Inside the Black Box



What Factors Are Most Significant?



What Factors Are Most Significant?

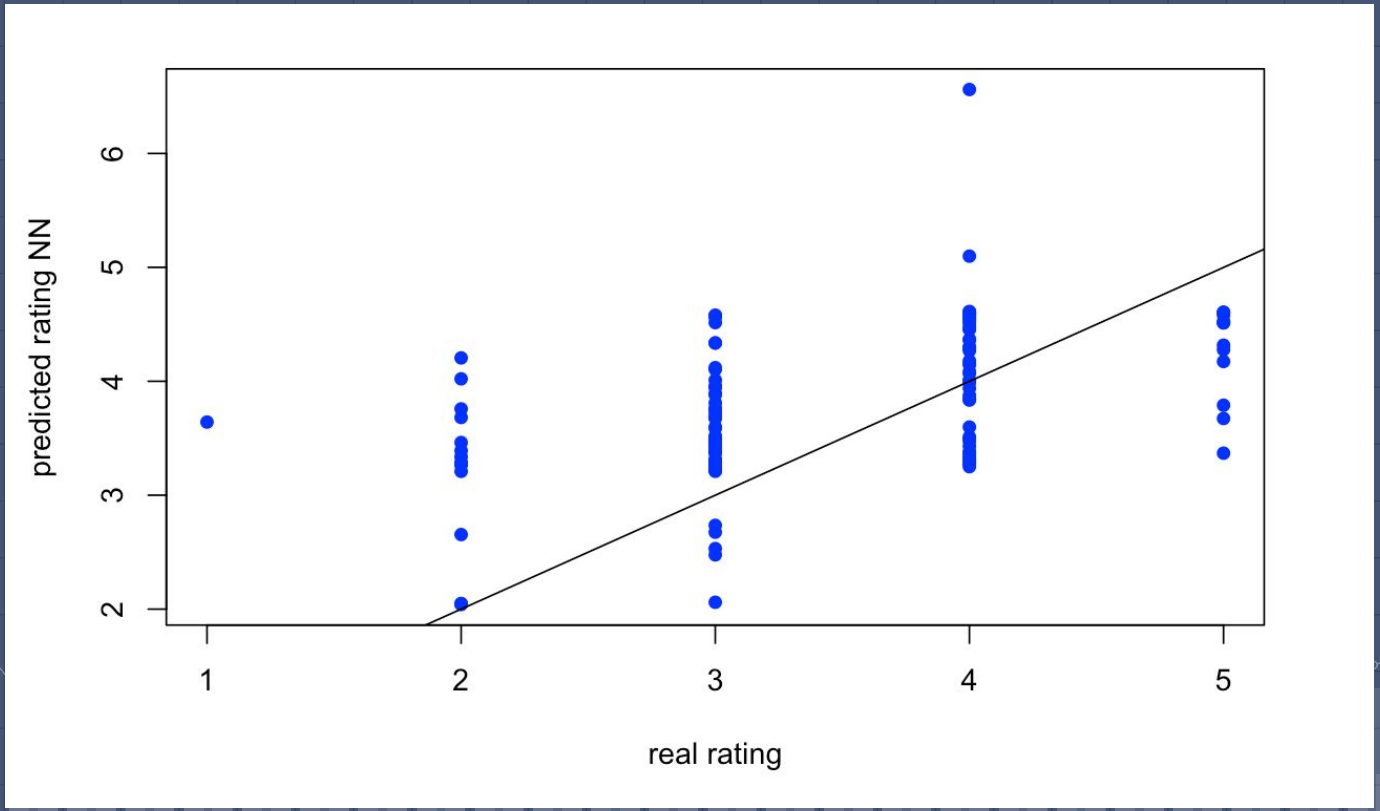
Positive Factors:

- Q3.22: I am happy at UCLA
- Q3.7: Most of my friends are from my own major
- Q2.2: I miss classes
- Q2.5: I go to office hours
- Q3.21: My parents understand the value of a college education
- First Gen Students

Negative Factors:

- Q3.16: I feel excluded at UCLA
- Q3.17: There is not enough students like me in my major
- Q3.4: My family struggled financially when I was growing up
- Q3.18: I wish there were more faculty I could identify with
- High School GPA

Prediction



RMSE: 0.84

Variation of RMSE

Conclusion and Limitations

- A sense of belonging is closely related to:
  - Peers and family support
  - How active you engage in college life
  - Your own perception of happiness
- Limitations:
  - Neural networks algorithm is largely a black box
  - Not many hyperparameters for tuning other than number of hidden layers
  - Neural networks mostly used in video and audio processing, although newer applications are booming