








COLLEGE EDUCATION & CAREER BUILDING

Using Natural Language Processing to understand the perceptions of Reddit users.

*Viviana V. Roseth
Jan 31, 2020*



Agenda

-  1. Research question
-  2. Description of Reddit data
-  3. Preprocessing of Reddit data
-  4. Model results
-  5. Conclusion





1

Research Question

General research question

Is college education a must-have to build a career and grow professionally?

This question could be tackled using a wide variety of methods:

- Simple comparisons of average income
- Tracer studies
- Analyses of returns to education
- Labor market demand analyses
- Skills mismatch analyses

OR



Two considerations:

1. Information in Reddit is more than abundant.

We need to select a couple of subreddits to analyze

Career Guidance

Higher Education

2. Information is produced by users and minimally curated

- It is subjective/anecdotal and rarely based on evidence we can corroborate
- Cannot be used to infer whether a college education is needed to develop a career

More concrete research question:

Do the concerns of users in both subreddits overlap to the extent that they cannot be distinguished?

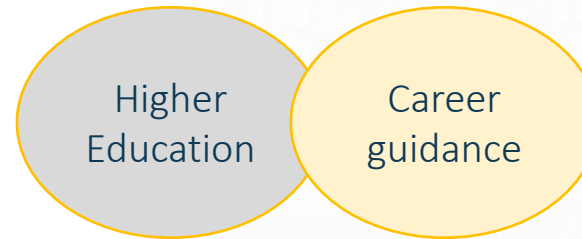
YES



Low score in classification models

People who are concerned about career guidance consistently use the same language as those who are interested in higher education

NO



High score in classification models

Concerns are mostly different in the two subreddits. Those who post in career guidance are not necessarily thinking about higher education



2

Description of Reddit Data

Description of Reddit Data

Career Guidance

- 126,000 members
- 10,000 posts between November of 2019 and last week
- Fewer serial posters
- Fewer empty post explanations (305)
- More comments on posts
- Final size: 9887 posts

Higher Education

- 26,300 members
- 10,000 posts between January of 2016 and last week
- More serial posters (advertisers, newscasters)
- More empty post explanations (+7600)
- Fewer comments on posts
- Final size: 9141 posts





3

Preprocessing of Reddit Data

Preprocessing in 4 steps



Data gathering

- Used Pushshift's API
- Gathered data from two subreddits
- Same number of posts per subreddit
- Different time periods



Data exploring

- Preselected 16 variables, including author, video, media, comments, cross posts, and age.
- Identified main characteristics of subreddits
- Dropped 8 variables



Data cleaning

- Eliminated duplicates
- Eliminated videos
- Changed datatypes
- Imputed missing values
- Eliminated serial irrelevant advertisings and content
- Concatenated the title and self-text



Data packing

- Packed data into 19,028 x 8 data frame
- Performed a train/test split
- Set 'X' as text and 'y' as subreddit
- Vectorized and fit/transformed X data



4

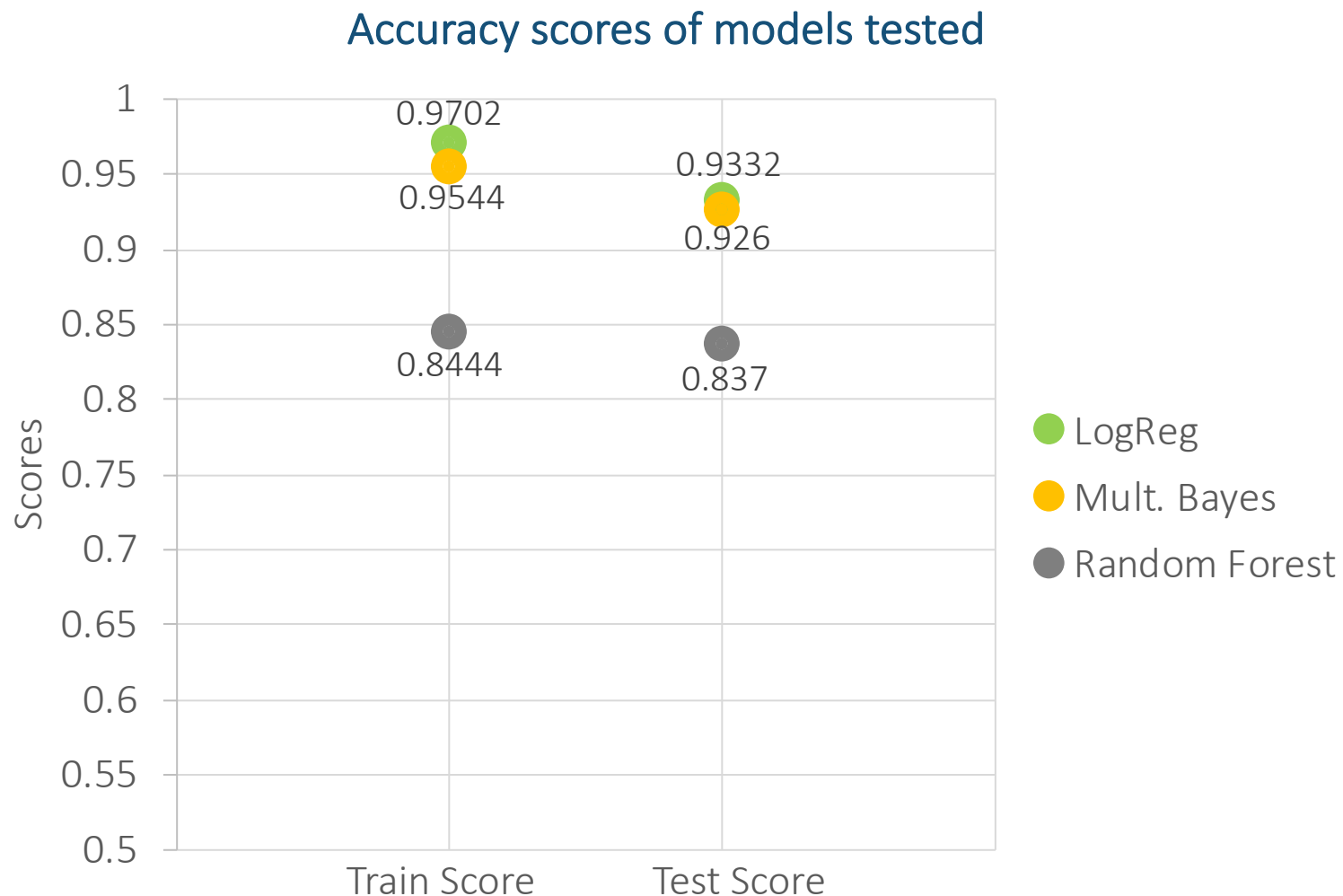
Model Results



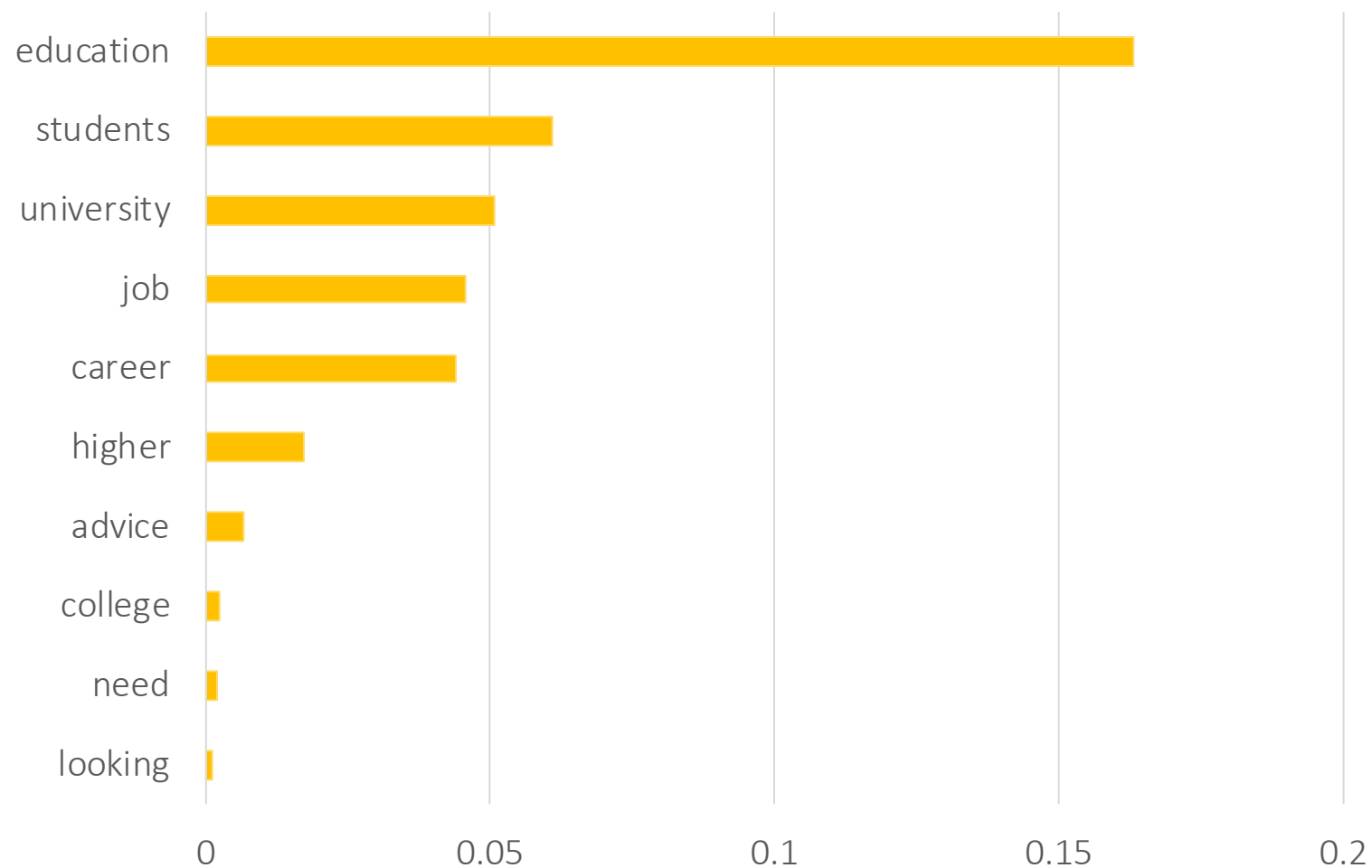
Baseline accuracy



Three classification models were tested



The 10 most important words in our models were:



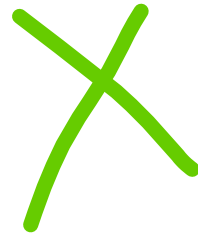


5

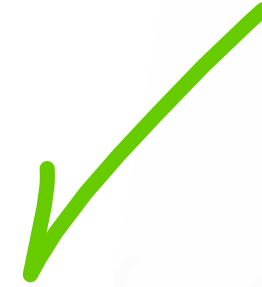
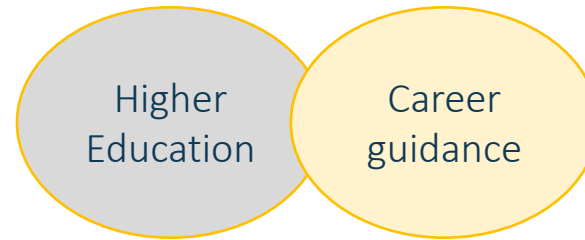
Conclusion

Do the concerns of users in both subreddits overlap to the extent that they cannot be distinguished?

YES



NO



Low score in classification models



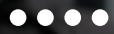
People who are concerned about career guidance consistently use the same language as those who are interested in higher education

High score in classification models



Concerns are mostly different in the two subreddits. Those who post in career guidance are not necessarily thinking about higher education

THANK YOU



Viviana V. Roseth



vvroseth@gmail.com