Feature Engineering and Selection… Chapter 5: Encoding Categorical Variables, pt 2

Agenda

Encoding Categorical Predictors

1) 5.6 Creating Features from Text Data

- **Simple text features**
- **Keywords**
- **Sentiment**

5.6 Creating Features from Text Data

Simple text features

OkC data – does open text predict STEM?

- Profile contains simple text features, e.g. hyperlinks; mentions; other specific items

Odds ratios:

- Odds : p / 1 - p
- Odds ratio: ratio of two quantities of odds
  - Odds hyperlink in STEM: 0.21 / 0.79 = 0.27
  - Odds hyperlink non-STEM: 0.142
  - Odds ratio {hyperlink STEM} : {hyperlink non-STEM}: 0.27 / 0.142 = 1.9

Should ask is it significant?

- 95% confidence interval gives a lower bound of 1.7
- because lower-bound does not overlap with 1 → is significant
  - (remember odds ratio of 1 equates to equal ratios of hyperlinks in STEM and non-STEM profiles)

Table 5.4: A cross-tabulation between the existence of at least one hyperlink in the OkCupid essay text and the profession.

	stem	other
Link	1063	620
No Link	3937	4380

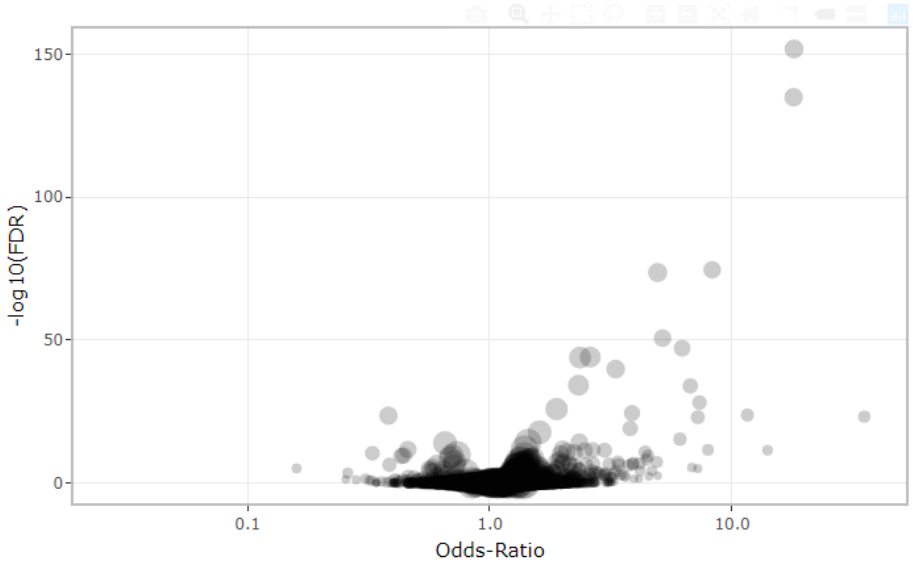
5.6 Creating Features from Text Data

Keyword analysis

- 63,440 words → cut-down to 4,918 (>50 occurrences)
- Calculate odds ratios on the occurrence of these terms in profiles

Limitations of p-values:

- P-value answers “is there a difference” does not tell magnitude, i.e. “how much”
- If you test many p-values, will get False Discovery Rate (will use correction to account for this)
- http://www.feat.engineering/text-data.html

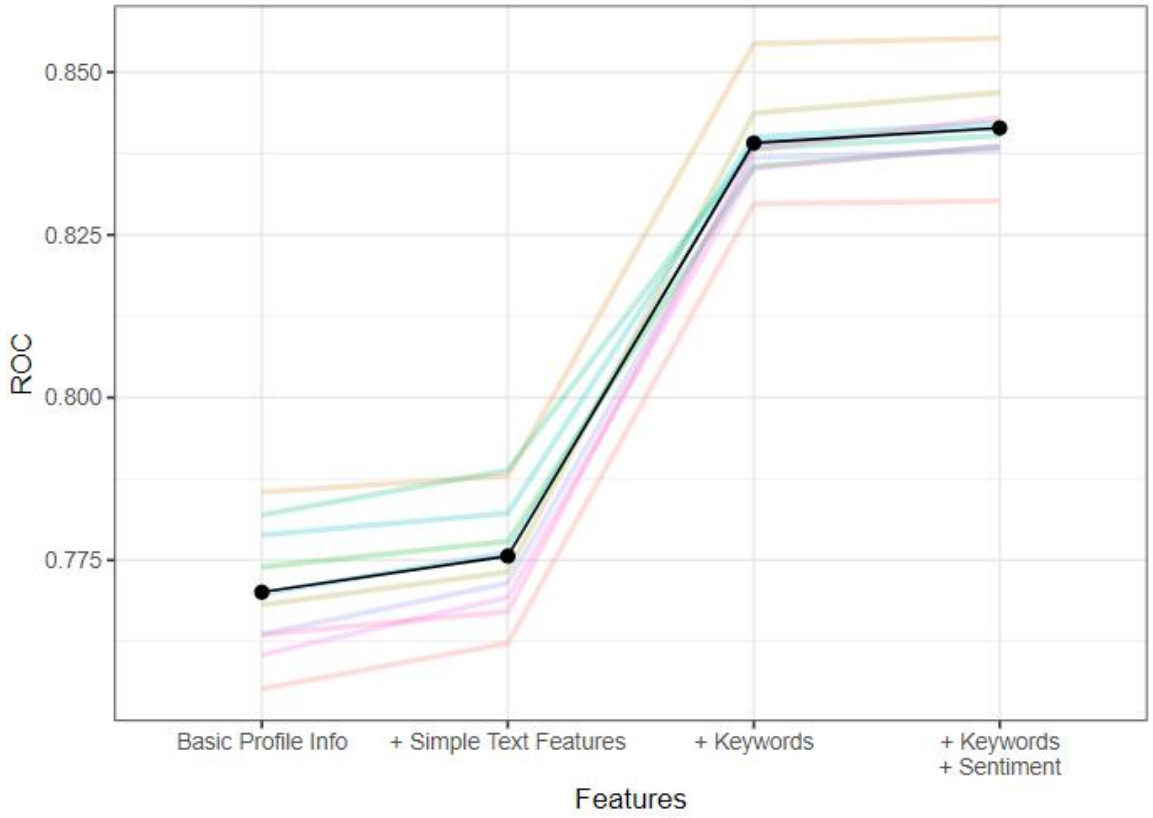


modeling. This results in 52 keywords:

alot	apps	biotech	code	coding
computer	computers	data	developer	electronic
electronics	engineer	engineering	firefly	fixing
futurama	geek	geeky	im	internet
lab	law	lawyer	lol	math
matrix	mechanical	mobile	neal	nerd
pratchett	problems	programmer	programming	science
scientist	silicon	software	solve	solving
startup	stephenson	student	systems	teacher
tech	technical	technology	valley	web
websites	wikipedia			

Performance with features included

- Basic set of profile characteristics (160 predictors) w/o open-text, **AUC: 0.77**
- Basic set + simple text features, e.g. hyperlinks (173 predictors), **AUC: 0.776**
- Basic set + keywords**, AUC: 0.839**
- Above + sentiment/language features, **AUC: 0.841**



5.6 Creating Features from Text Data

Common processing steps

- removing commonly used *stop words*, such as “is”, “the”, “and”, etc.
- *stemming* the words so that similar words, such as the singular and plural versions, are represented as a single entity.
  - For example, these 7 words are fairly similar: "teach", "teacher", "teachers", "teaches", "teachable", "teaching", "teachings". Stemming would reduce these to 3 unique values: "teach", "teacher", "teachabl“
  - Stemming would reduce unique word count in OkC data from 62,928 → 45,486

5.6 Creating Features from Text Data

Term frequency – inverse document frequency

- As an example, suppose the term frequency of the word “feynman” in a specific profile was 2. In the sample of profiles used to derive the features, this word occurs 55 times out of 10,000.
- Compare the frequency of a word in a doc against it’s frequency in the corpus of docs
  - In production – need to specify the corpus of docs for comparison ahead of time (likely using training set)

Other…

- N-grams…
- Sophisticated embedding techniques…
- Etc.