DDA management with predictive modeling

go.ncsu.edu/ddamodel

John Vickery
NCSU Libraries
john_vickery@ncsu.edu
go.ncsu.edu/ddamodel
Will a DDA title be triggered for purchase within 12 months of being added to the collection?
About the data

- About 50,000 DDA records
- Added to the collection between Dec. 2011 and June 2015
- 11% had been purchased within 12 months
- 95% published between 2011 and 2016
- About 3000 publishers (Routledge and Palgrave Macmillan most common)
- Structured and unstructured data
Software and sources

- **SAS**
  - Extract DDA records from SIRSI
  - Data manipulation and clean-up

- **Python**
  - Summary text from Syndetic API
  - Publisher data from MARC record API
  - LC classification from OCLC API

- **SAS Enterprise Miner**
  - Text mining
  - Build and test models
  - Score new data
Model Inputs:

- Publisher
- Publication year
- LC class
- Summary
The Topic Modeling Process

Input:
Syndetic summaries
(Iterative and time consuming)

Adjust term weights

Do the topics make sense?

Generate document matrices

Model parameters
Reduce matrix dimensions

Get data
Parse text
Synonyms/refinement
Apply model
Interpret results

The Topic Modeling Process

Input:
Syndetic summaries
(Iterative and time consuming)

Adjust term weights

Do the topics make sense?

Generate document matrices

Model parameters
Reduce matrix dimensions

Get data
Parse text
Synonyms/refinement
Apply model
Interpret results
Regression Model vs. Decision Tree
Prob($Y = 1$)

Purchased within 12 months

X
Decision tree performed best

1. engineering, +engineer, +design, +application
2. clinical, +treatment, medical, +therapy, +professional
3. mathematics, +application, mathematical, +problem, +solution
4. +disease, molecular, +biology, +cell, +protein
5. software, +design, +basic, +guide, +learn
6. +woman, +man, male, +identity, sexual
7. +introduction, accessible, concise, comprehensive, +philosophy
8. +material, +engineer, +polymer, +technology, +material science
9. food, nutrition, +product, food
10. +law, legal, +court, +right, legal
11. data, big data, +process, +application, +technique
"all models are wrong, but some are useful"

George Box
Is this model useful?

- About 8% misclassification rate
- Tested against 9,769 DDA titles from 2015 and 2016
- 556 had been purchased within 12 months
- Correctly predicted 45 of the purchased titles
  - 511 false negatives
- Correctly predicted 8,972 of the non-purchased titles
- 241 false positives
- 48 of the false positives have been purchased
48 have been triggered
Thank You!

john_vickery@ncsu.edu

github.com/jnvickery/DDAmodel