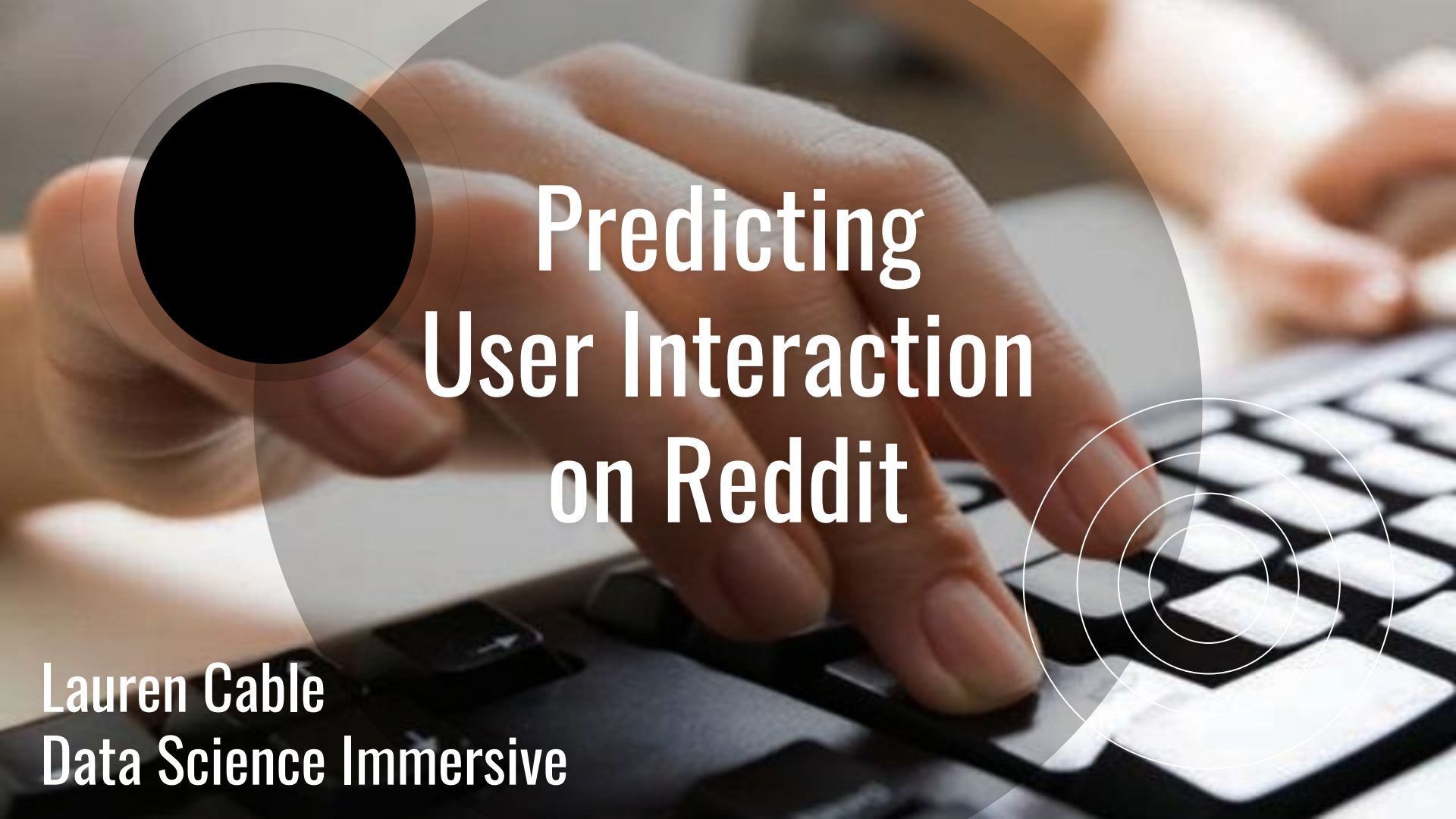


Predicting User Interaction on Reddit

Lauren Cable
Data Science Immersive

Why Are We Here?

- To Identify what characteristics of a post on Reddit are most predictive of overall user interaction, as measured by number of comments.

Reddit.com

...

ABOUT

- Social news and discussion website
- 542 visitors monthly

CONTENT

- Submitted as links, texts, and images
- Organized by subject into boards called “Subreddits”

USER INTERACTION

- Position on Subreddit page dependent on up/down votes a post receives
- Posts receiving adequate up-votes will appear as “hot thread” on site’s homepage

reddit

hot | new | rising | controversial | top | gilded | wiki

popular in: United States ▾ select state: Georgia ▾

5

AIRFORCE.COM

Ride along in 360 as an Air Force Special Ops team trains to infiltrate an enemy hideout. (http://www.airforce.com)

promoted by USAirForceOfficial

promoted save report

trending subreddits r/GlitchInTheMatrix r/catsareliquid r/puns r/SupermodelCats r/BoneAppleTea 4 comments

1 68.4k

Tammy Waddell taught school for 25 years. Her obituary asked that in lieu of flowers, mourners should bring backpacks filled with school supplies, to honor her commitment to students in need. (i.redd.it)

submitted 6 hours ago by PatFlynnEire to r/pics

574 comments share save hide report

2 11.3k

What would you do? (i.redd.it)

submitted 4 hours ago by ReallyBadNuggets to r/BlackPeopleTwitter

239 comments share save hide report

3 23.4k

He woke (i.redd.it)

submitted 5 hours ago by crackerjackbundy to r/funny

206 comments share save hide report

4 39.3k

Neat ball trick (i.imgur.com)

submitted 4 hours ago by langtuvn to r/gifs

611 comments share save hide report

5 17.5k

This massive sign was designed to trick people into getting off the highway 97.9 miles before Niagara Falls. (i.redd.it)

submitted 5 hours ago by NobleStumble to r/assholedesign

441 comments share save hide report

6 8289

minty is just cold spicy (i.redd.it)

submitted 3 hours ago by jcb9009 to r/funny

142 comments share save hide report

7 19.6k

Shoutout to KharkivForge for crafting me the Leviathan Axe! (i.redd.it)

submitted 5 hours ago by HistoryGuardian to r/gaming

245 comments share save hide report

The Data

...

WEBSCRAPER

- Constructed with Requests and BeautifulSoup libraries
- Scraped “hot threads”
- Aggregated 2500+ unique results

FOR EACH THREAD

- The following data was acquired :
  - Title
  - Subreddit
  - Length of time since posting
  - Number of comments

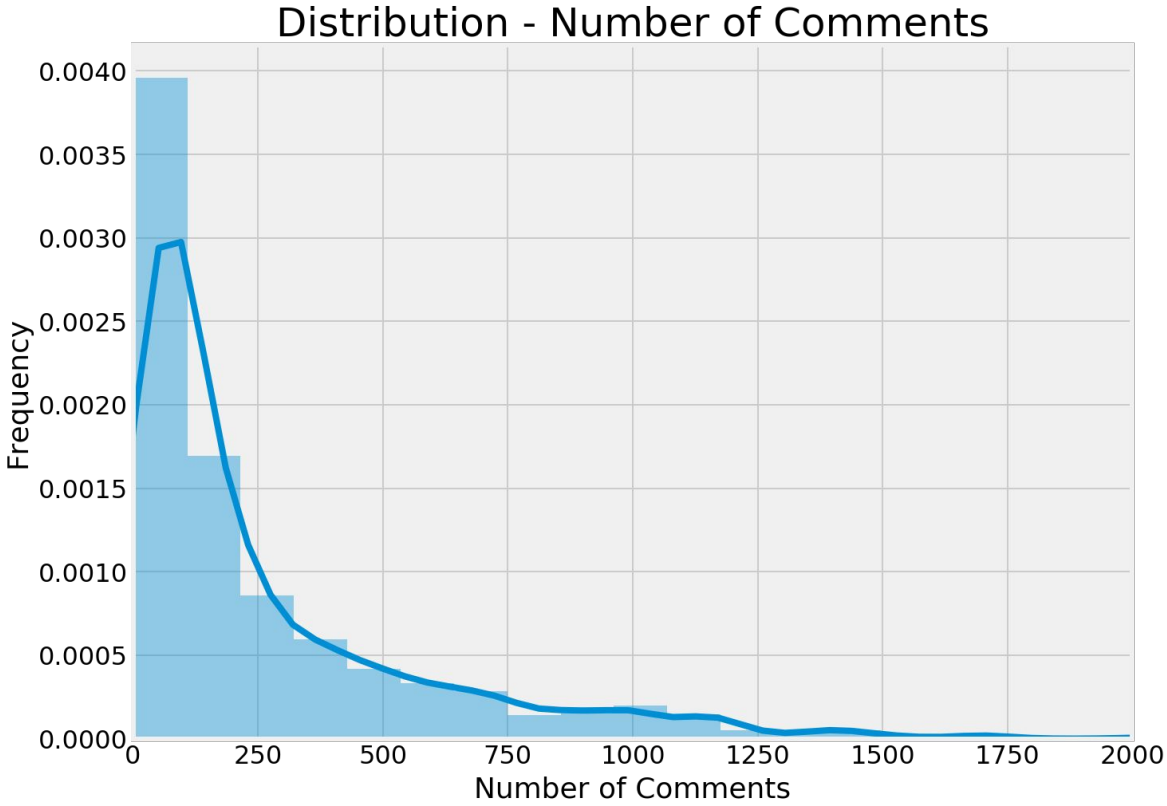
Comments	Subreddit	Time	Title
715 comments	r/funny	3 hours ago	Well...can anyone tell me the answer? (i.redd.it)
642 comments	r/mildlyinteresting	3 hours ago	We have a Party City coming soon. Next to Part...
1118 comments	r/oddlysatisfying	4 hours ago	Fun at work with 1377 pallets (i.imgur.com)
1359 comments	r/Showerthoughts	4 hours ago	Your future self is talking shit about you (se...
129 comments	r/BlackPeopleTwitter	5 hours ago	Until his daughter gets kidnapped again. (i.im...
257 comments	r/woahdude	5 hours ago	The Skeleton Flower's petals become transparen...
63 comments	r/funny	3 hours ago	Get her told Frank (i.redd.it)
556 comments	r/TwoXChromosomes	3 hours ago	I called the police. Just need to vent./r/all ...
693 comments	r/pics	5 hours ago	Flying into Bora Bora (i.redd.it)
145 comments	r/todayilearned	3 hours ago	TIL all "warranty void if removed" stickers ar...

Exploratory Data Analysis

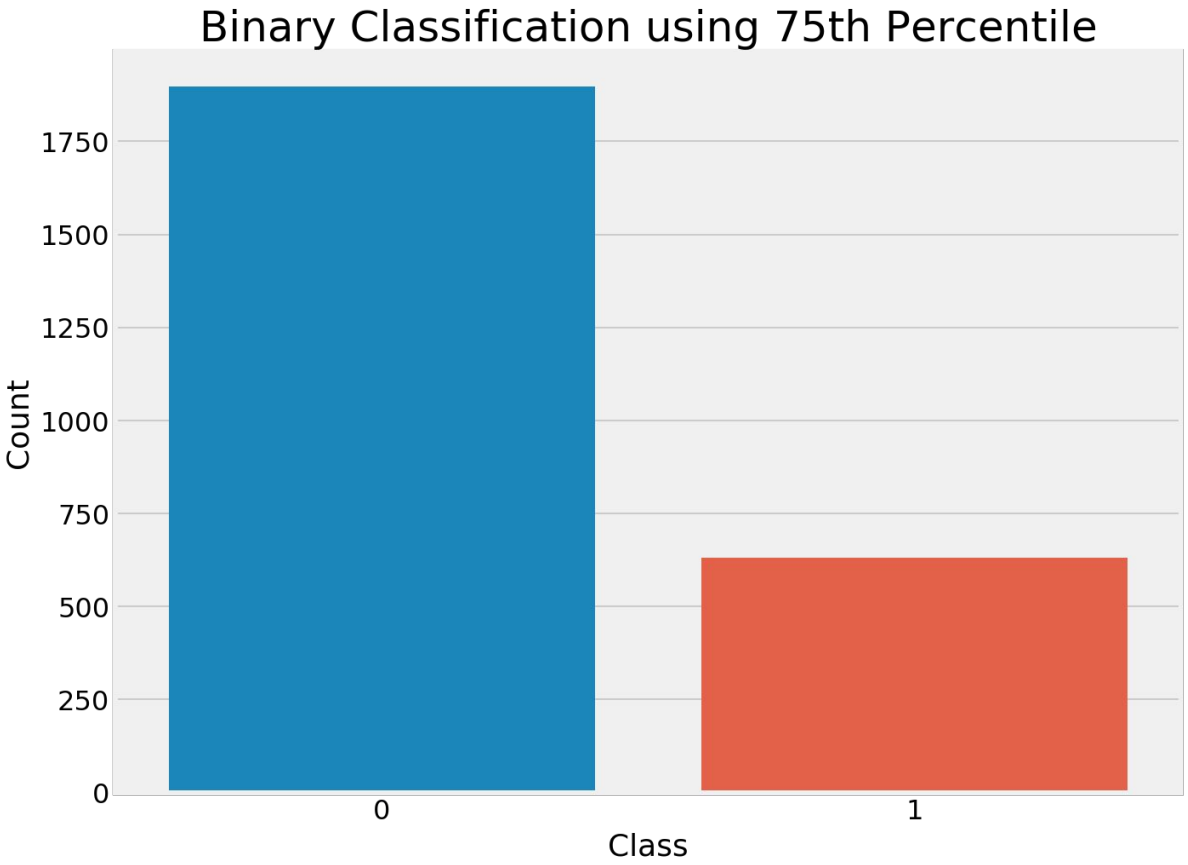
...

Creating Binary Classes

- In order for models to predict user interaction, convert target data (Number of Comments) into binary variable :
    - High Comments -> 1
    - Low Comments -> 0
- To determine cut-off for classes, calculate 75th percentile :
    - 418 Comments



- Histogram reveals heavy positive skew
- Majority of threads receiving low comments
- Simply using median as cut-off for classes is unsatisfactory




```
==========
Value Counts:
0    1897
1     632
Name: Comment Dummy,
==========
```

Determining Baseline Accuracy

- Simple summary statistic used as point of reference to consider whether value is being added to model
- For classification problems, use statistic for the class model seeks to predict

```
==========================================================
Baseline Accuracy:  0.7500988533017003
This is the accuracy our modeling techniques seek to improve on
==========================================================
```

The Models

...

Predictor Variable : Title

- Count-Vectorizer :
  - Transforms title words into sparse predictor matrix
  - N-Grams tuned to identify different patterns of words
- Random Forest Model :
  - Best Cross-Validated Score = 84.57%
  - N-Gram Range = 3

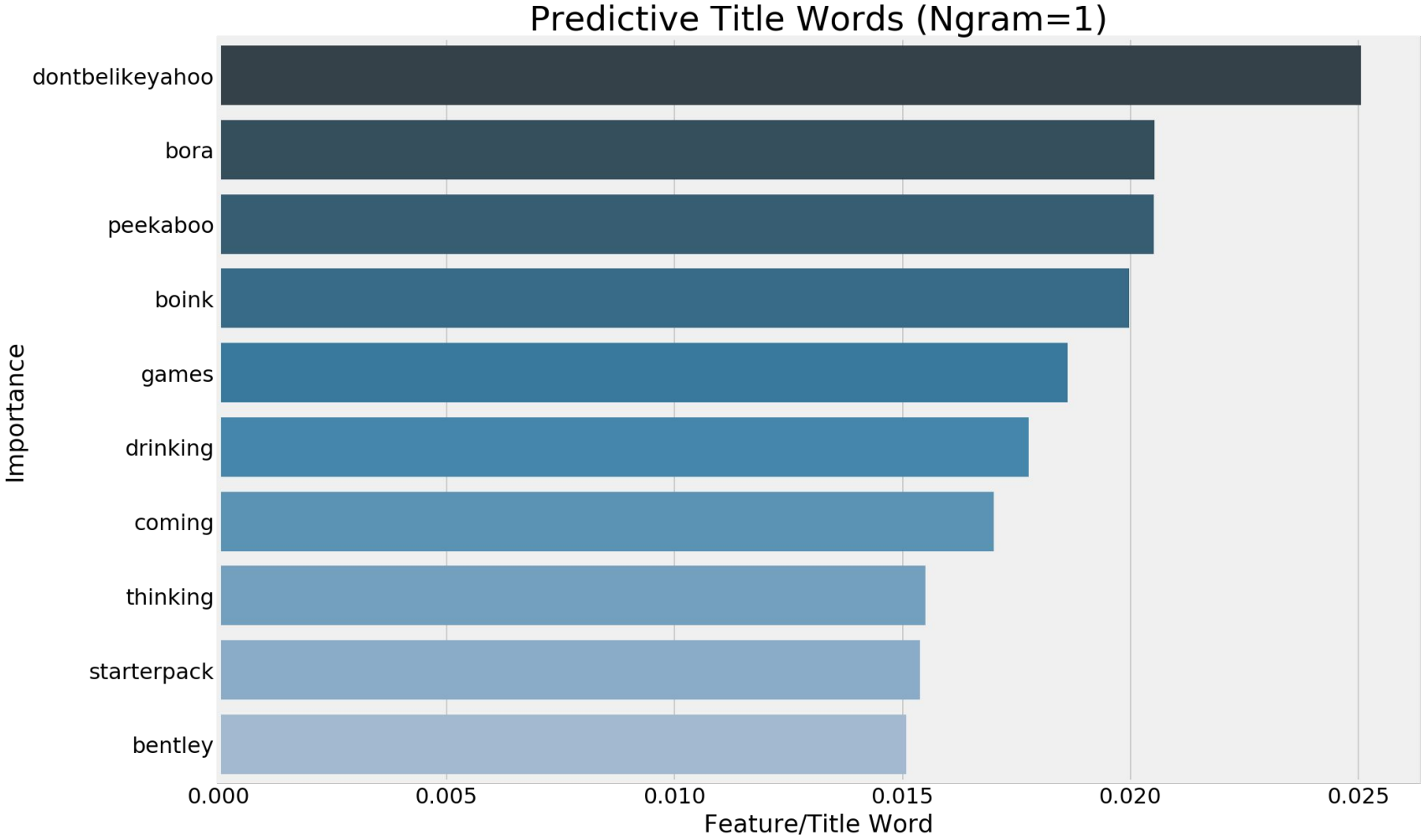
Predictor Variable : Subreddit

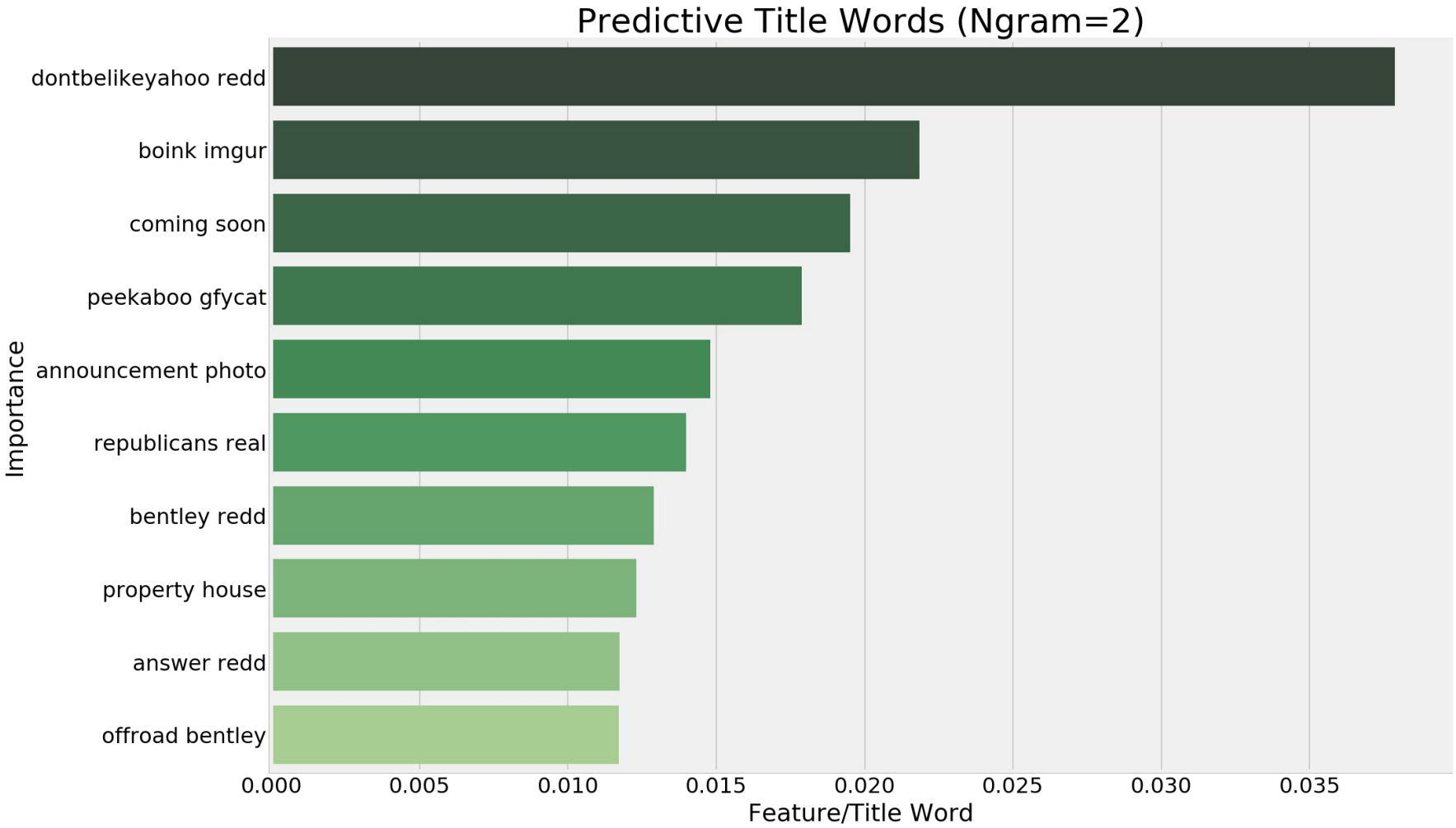
- Dummy Variable :
  - Quantifies and transforms categorical variables w/no relationship into sparse binary predictor matrix
- Random Forest Model :
  - Cross-Validated Score = 83.38%
- Logistic Regression Model :
  - Cross-Validated Score = 86.64%

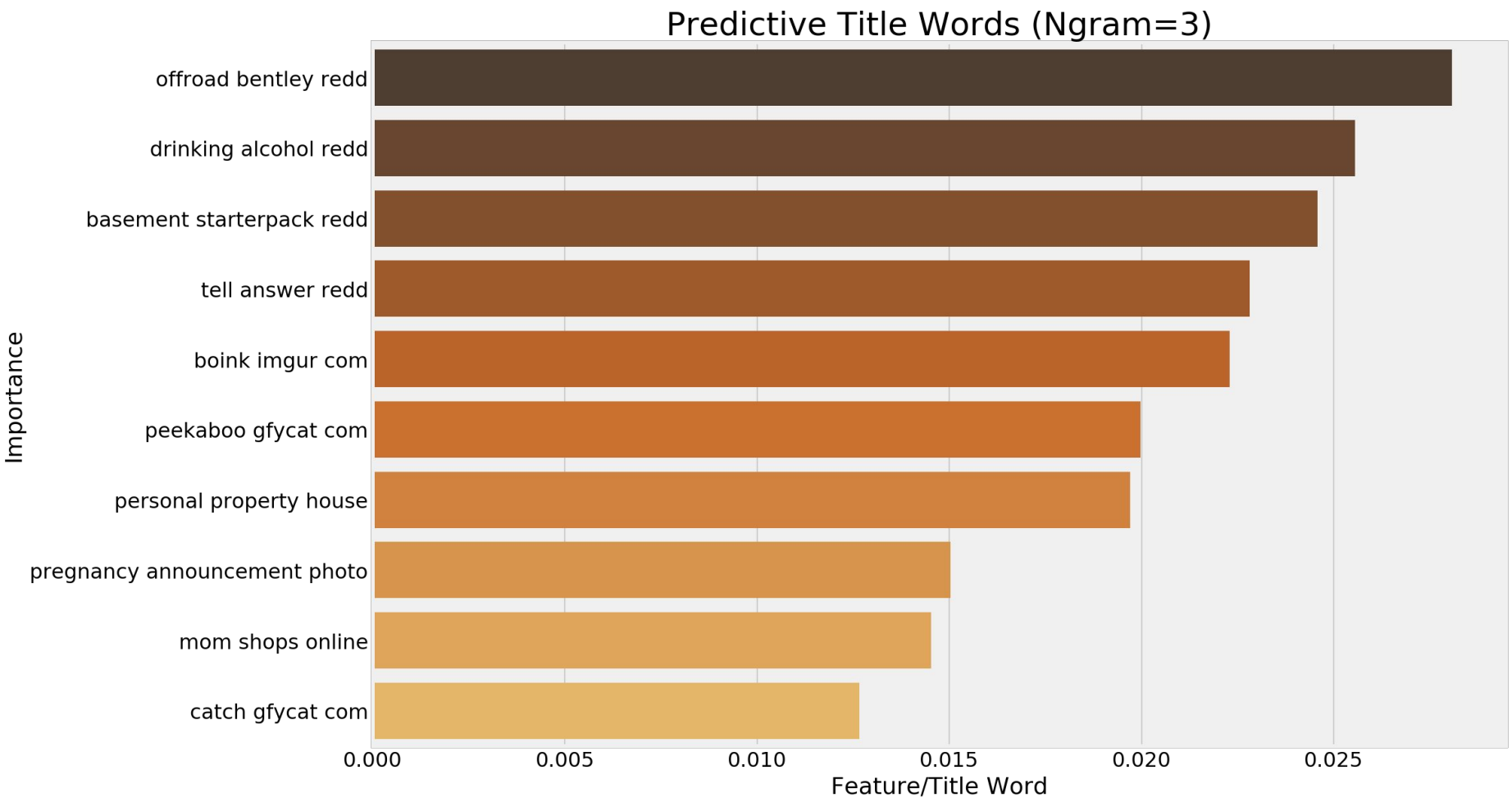
Optimization

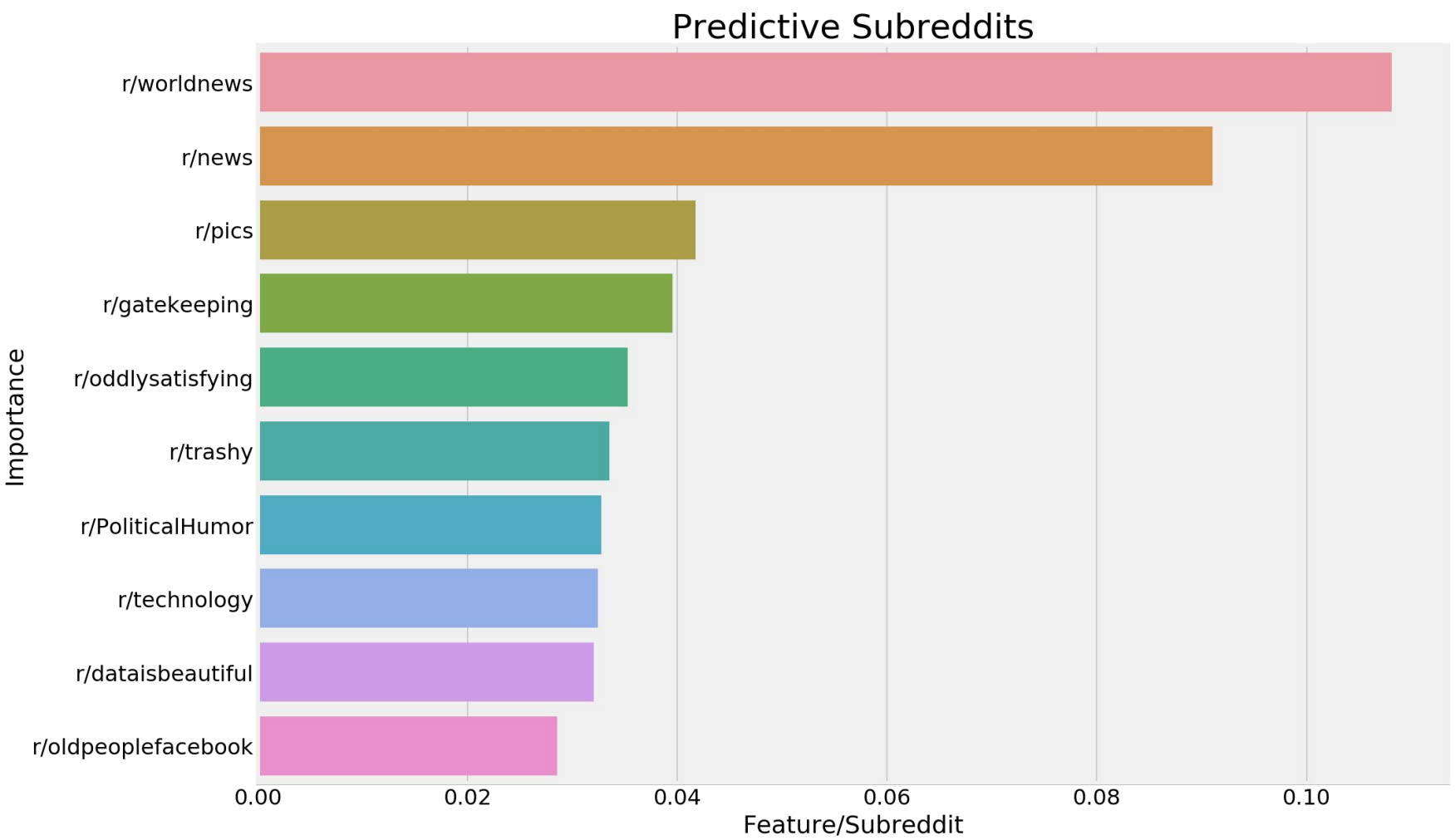
- Grid Search Cross Validation :
  - Builds and evaluates a model for each combination of algorithm parameters specified in a grid
- Logistic Regression Model :
  - Cross-Validated Score = 88.62%

The Conclusions


Predictive Title Words (Ngram=1)
dontbelikeyahoo
bora
peekaboo
boink
games
drinking
coming
thinking
starterpack
bentley
Importance
0.000
0.005
0.010
0.015
0.020
0.025
Feature/Title Word


Predictive Title Words (Ngram=2)
dontbelikeyahoo redd
boink imgur
coming soon
peekaboo gfycat
announcement photo
republicans real
bentley redd
property house
answer redd
offroad bentley
Importance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Feature/Title Word

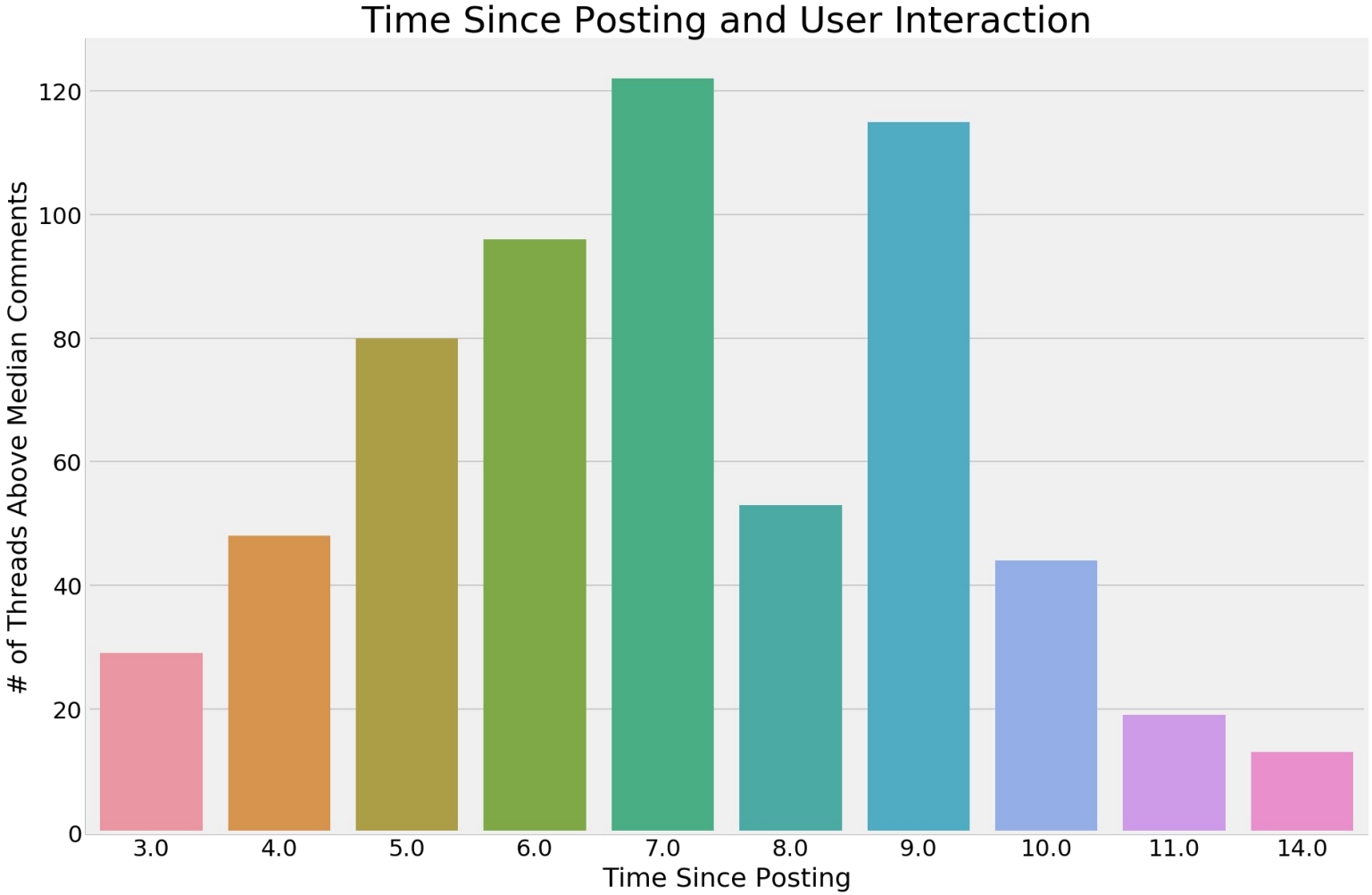

Predictive Title Words (Ngram=3)
offroad bentley redd
drinking alcohol redd
basement starterpack redd
tell answer redd
boink imgur com
peekaboo gfycat com
personal property house
pregnancy announcement photo
mom shops online
catch gfycat com
Importance
0.000
0.005
0.010
0.015
0.020
0.025
Feature/Title Word

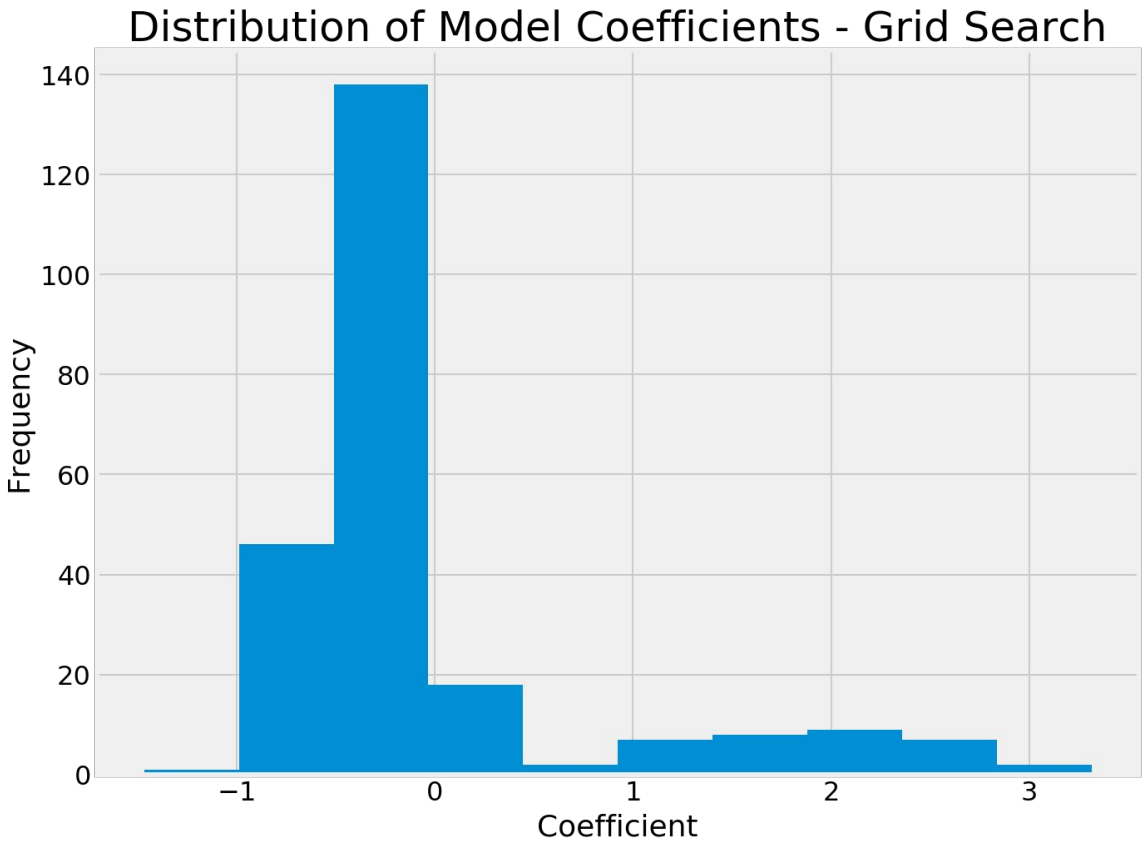

Predictive Subreddits
r/worldnews
r/news
r/pics
r/gatekeeping
r/oddlysatisfying
r/trashy
r/PoliticalHumor
r/technology
r/dataisbeautiful
r/oldpeoplefacebook
Importance
0.00
0.02
0.04
0.06
0.08
0.10
Feature/Subreddit

Most Predictive Subreddits

- Comparing time since posting with number of threads receiving high comments

- Prime-time for user interaction appears to be within the 6th-7th hour
  - r/pics :
    - 6.41 hours - 135 threads
  - r/worldnews :
    - 7.72 hours - 50 threads
  - r/news :
    - 9.23 hours - 47 threads


Time Since Posting and User Interaction
# of Threads Above Median Comments
0
20
40
60
80
100
120
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
14.0
Time Since Posting



- Majority of coefficients are negative
- Model better equipped to predict low comments rather than high comments


Thank you!
Q & A