How to create a highly engaged reddit post?

Using data from reddit hot post page

Carol Duan

Overview

- Problem Defining
- Data Gathering
- Data Exploring & Cleaning
- Modeling & Evaluation
- Suggestions
- Next Step

Problem Defining

Problem: How to create a highly engaged reddit post?

- **Target:** Engagement -> number of comment -> high / low (threshold: median)
- **Features**: Drivers of engagement -> highly related matrix (measure: feature importance)

Problem Type: Classification & **Inferences**

My Workflow:

- **Models:** Build classification models that are possible to interpret the importance of features (e.g. KNN can't be used) and choose one with a highest test score of accuracy.
- **Inferences:** Analyze the result of features (coefficient / feature importance) and find the most important features that can be used to drive more engagement for reddit post.
- **Suggestions:** Give specific suggestions to answer the how to question.

Data Gathering

Scraped thread info from reddit.com by sending GET requests

- Get all information listed in the homepage (87 columns total)
- Get 3 pieces of data from different time of the day - at 11pm, 7am and 3pm (5000 posts/rows with duplicates each time)
- Total 87 columns and 15,000 rows data as a starting point

Scraped comments from r/CringeAnarchy and r/funny subreddit pages using Python reddit API wrapper PRAW

- For each subreddit, go through top 1000 posts and top 10 comments (without any replies) for each post and pack the comments together for each post (1000 rows for each subreddit)
- Get number of comments / ups / title / name of post for further analysis

Data Exploring & Cleaning

Drop duplicates (by name)

Go through the data info and choose relative features (manually)

- required features: title, subreddit, duration
- add features: post_hint, ups, subreddit_subscribers, spoiler, stickied, over_18
- target: num_comments

Check data types / relationship of each features and do feature engineering

- Combine similar features together : use is_self to add missing text post category information in post_hint
- Create dummy variables: subreddit, spoiler, stickied, over_18
- Use CountVectorizer to create features (words) to deal with string variable: title
- Map num_comments to binary value
    - 0: low engagement, num_comments < median
    - 1: high engagement, num_comments >= median

Train / test dataset split (0.85 / 0.15)

Modeling & Evaluation

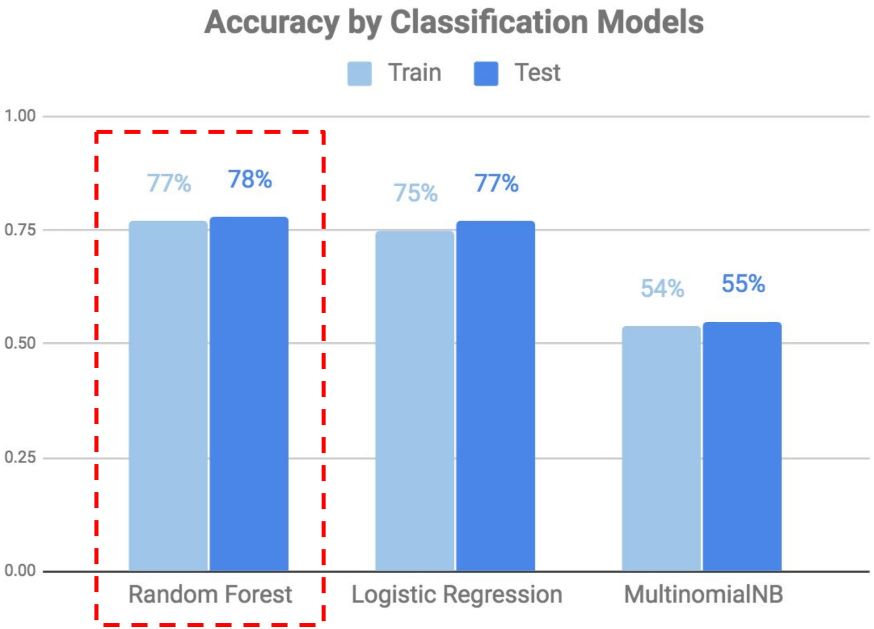
Model with data from reddit.com homepage

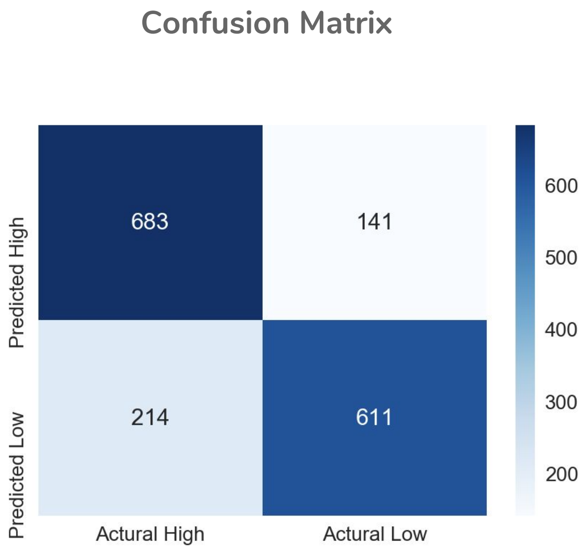
- Build models using pipeline and tune them using GridsearchCV
  - Random Forest
  - Logistic Regression
  - MultinomialNB
- Choose the one with the highest test accuracy score and analyze the feature importance

Explore comments on the top subreddits "CringeAnarchy" and "funny"

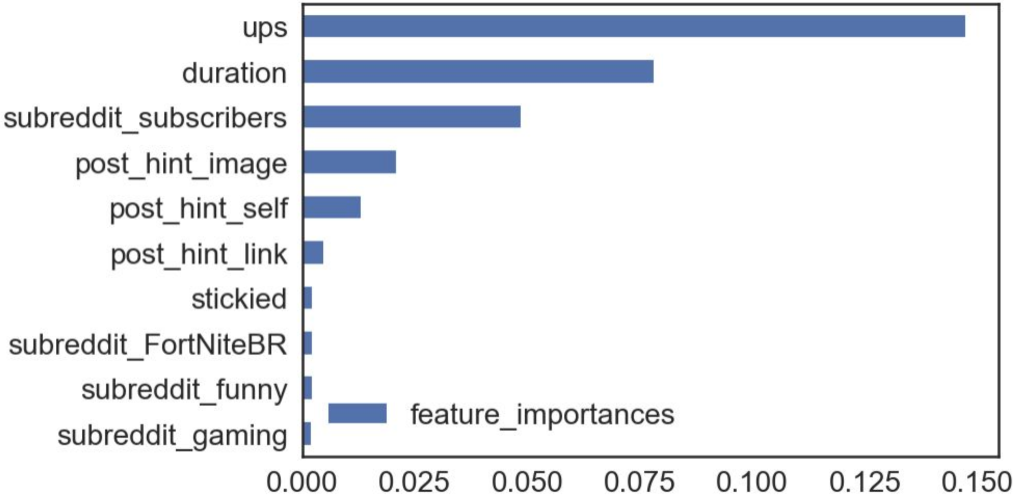
- Build a model (Random Forest) to analyze which words are important in the comment section of the top posts that drive more engagement
- Compare the keywords

Modeling & Evaluation





Suggestions



Influencers of post engagement

- Upvotes
- Duration of posting
- **Subreddit subscribers**
- **Post type**
- Stickied
- Subreddit name

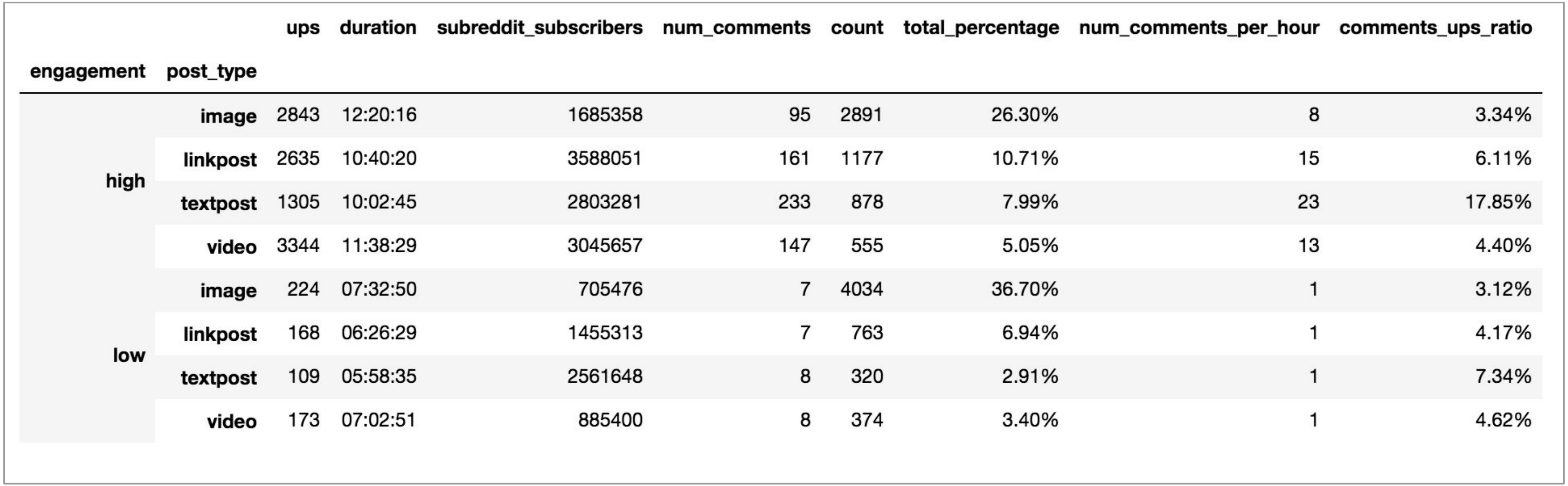
Suggestion #1 Post Position

Post in the subreddits that have higher number of subscribers

subreddit	feature_importances	subscribers (million)
FortNiteBR	1.08%	0.66
funny	0.95%	19.63
AskReddit	0.90%	19.30
gaming	0.78%	18.20
CringeAnarchy	0.73%	0.35
NintendoSwitch	0.58%	0.63
MMA	0.53%	0.53
aww	0.52%	17.22
todayilearned	0.52%	18.84
videos	0.50%	17.81

Suggestion #2 Post Type

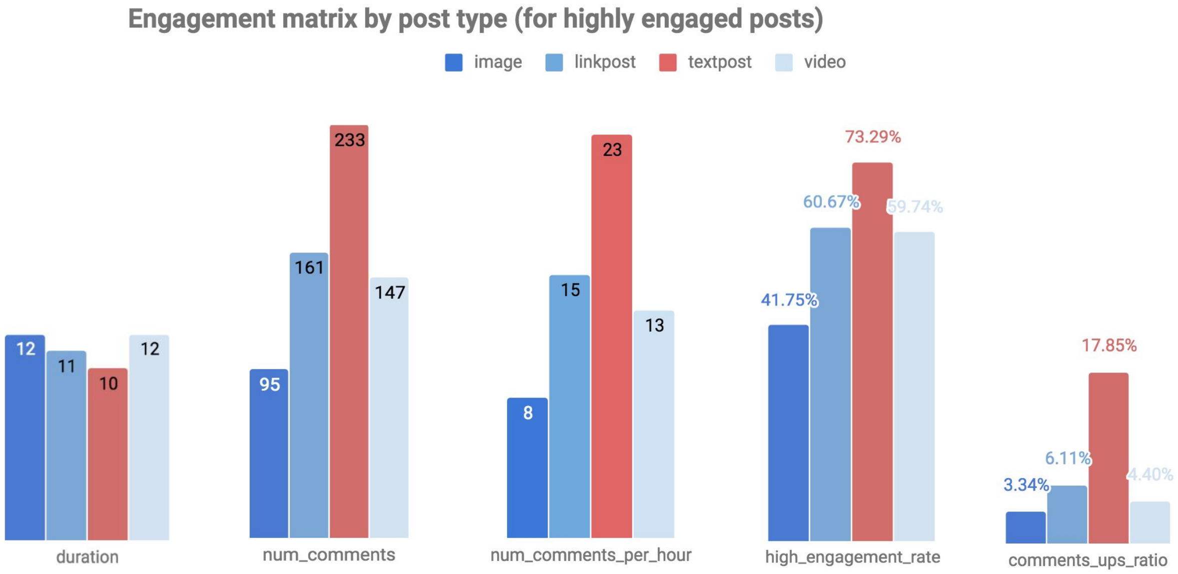
Text post (aka self post) is a better choice



		ups	duration	subreddit_subscribers	num_comments	count	total_percentage	num_comments_per_hour	comments_ups_ratio
engagement	post_type								
high	image	2843	12:20:16	1685358	95	2891	26.30%	8	3.34%
	linkpost	2635	10:40:20	3588051	161	1177	10.71%	15	6.11%
	textpost	1305	10:02:45	2803281	233	878	7.99%	23	17.85%
	video	3344	11:38:29	3045657	147	555	5.05%	13	4.40%
low	image	224	07:32:50	705476	7	4034	36.70%	1	3.12%
	linkpost	168	06:26:29	1455313	7	763	6.94%	1	4.17%
	textpost	109	05:58:35	2561648	8	320	2.91%	1	7.34%
	video	173	07:02:51	885400	8	374	3.40%	1	4.62%

Suggestion #2 Post Type

Text post (aka self post) is a better choice

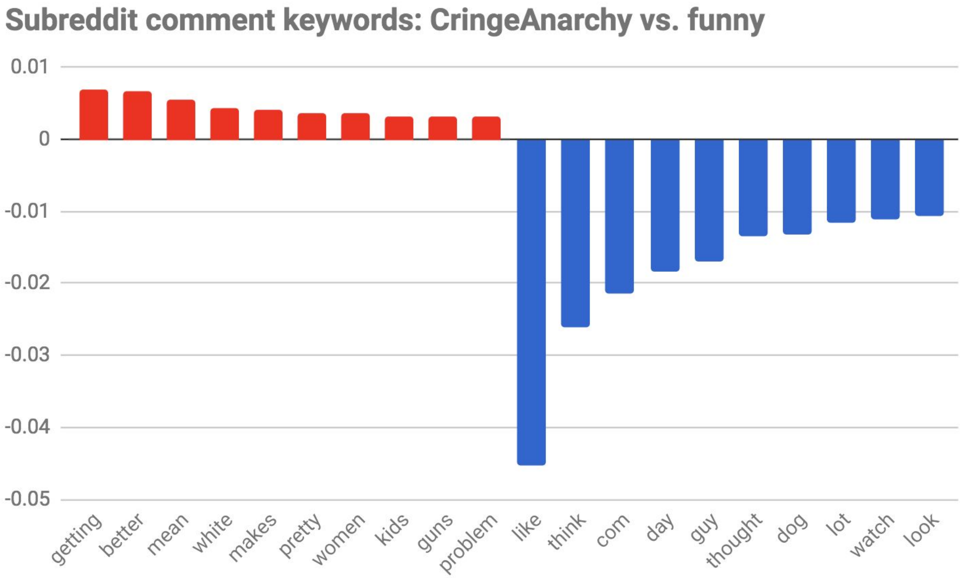


Text post get a higher performance on the engagement matrix

- Takes shorter time to become a highly engaged post
- Gets more comments per post on average
- Gets more comments per hour than other types of post
- Has a bigger chance to become a highly engaged post
- Gets more comments per upvote on average

Suggestion #3 Post Content

Focus on the specific keywords in the comments of subreddit



Keywords in comments shows the general topic in this subreddit

- r/CringeAnarchy:

  "The official subreddit room for all organized alt right trolls"

- r/funny:

  "reddit's largest humour depository"

Next Step

Continue explore the driver of **text post (self post)** engagement:

- Test post content: is there any specific words / phrases in the content driving engagement?
- Test post title: does the title matter for text post? will it be more important than other types of post?

Continually tune and improve my model to increase the accuracy score

Collect more data and consider about finding new features

Thoroughly Interpret model result and give more recommendations

Thank You for Watching!