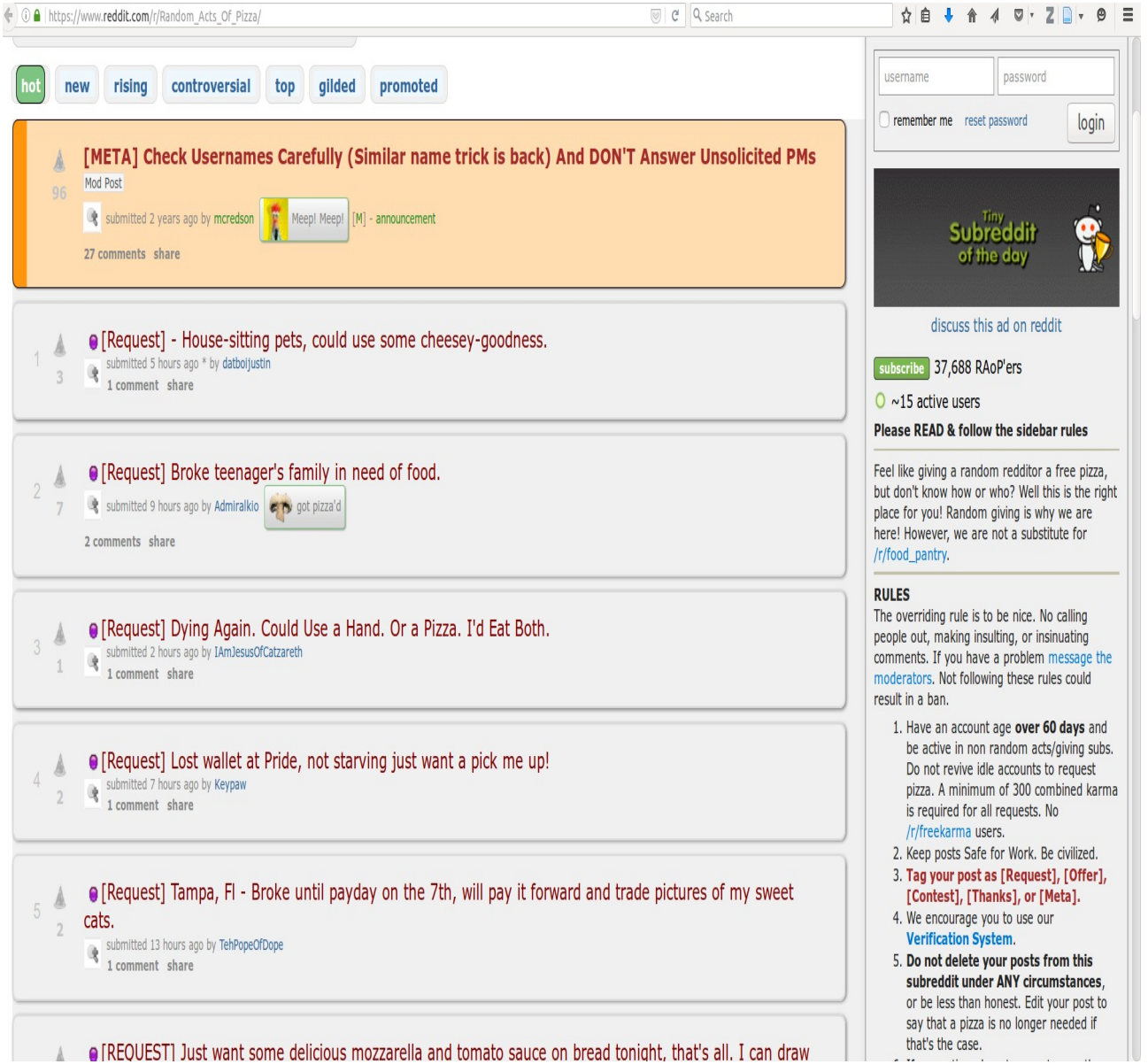
Hootsuite

(Random Acts of Pizza: My first Text Mining project)

Overview:

- My Goals
  - Give you a data-driven advice in how to write your post and increase the chance to receive a free pizza (How can I do it?)
  - Learn/Get more experience in text mining (**learn on demand**). Explore NLP and text mining packages (**try a new thing**) in R
  - Test my EDA Start Kit project (**GTD**)
- Overview
  - What is RaoP?
  - Data:
    - Data introduction;
    - Features and text engineering
    - Data prep and descriptive
  - First model development
    - Features selection
    - Model assessment
  - Recommendation and next steps

RAoP: Random Action of Pizza



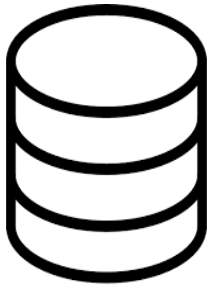
If I want a free pizza. I write a compelling post on the on r/Random_Acts_Of_Pizza/ asking for a pizza with a tag: [REQUEST]

If I want to offer a pizza. I write a post with a tag [Offer] or select a [REQUEST] post

So you can help people and restore Restoring Faith in Humanity, One Slice at a Time - ***Reddit***

https://www.reddit.com/r/Random_Acts_Of_Pizza/

Data



Data

- Data size: 5671 (rows) x 33 (vars)
- #users: 5671 (no user with 2+ post **!?**)
- Received pizza: 1397 (25%)
- Investigated **actionable** predictors (features)
  - Available (in the moment I write the post)
  - Control our partial control (I can change it)

N	Field	Description	Comm
1	post_was_edited	*Boolean indicating whether this post was edited (from Reddit).*	
2	request_id	*Identifier of the post on Reddit, e.g. "t3_w5491".*	
3	***request_text***	***Full text of the request.***	***Info***
4	request_title	*Title of the request.*	
5	***requester_account_age_in_days_at_request***	***Account age of requester in days at time of request.***	
6	requester_days_since_first_post_on_raop_at_request	*Number of days between requesters first post on RAOP and this request (zero if requester has never posted before on RAOP).*	
7	requester_number_of_comments_at_request	*Total number of comments on Reddit by requester at time of request.*	
8	requester_number_of_comments_in_raop_at_request	*Total number of comments in RAOP by requester at time of request.*	
9	requester_number_of_posts_at_request	*Total number of posts on Reddit by requester at time of request.*	
10	requester_number_of_posts_on_raop_at_request	*Total number of posts in RAOP by requester at time of request.*	
11	requester_number_of_posts_on_raop_at_retrieval	*Total number of posts in RAOP by requester at time of retrieval.*	
12	***requester_received_pizza***	***Boolean indicating the success of the request, i.e., whether the requester received pizza.***	***Y***
13	***requester_upvotes_minus_downvotes_at_request***	***Difference of total upvotes and total downvotes of requester at time of request.***	***karma***
14	requester_username	*Reddit username of requester.*	
15	unix_timestamp_of_request	*since most RAOP users are from the USA).*	Date

Kept my focus on:

- His/Her history was compelling?
- He/She was polite?
- He/She was able to prove his history?
- He/She provided what was required/asked to him?
- Is He/She a good/nice reddit user?
- How long is he/she a reddit user?

Features Engineering

- Post related features (total control)
  - #words in the post (**compelling**)
  - post sentiment score: (**polite and positive**)
  - post received 5 scores: money, job, student, family and desire narratives (kind of history) (**Some narratives strategy might be more effective than others**)
  - Has.link (**reciprocity: you gave something back**)
- User/requester related features (partial control)
  - status or karma: requester upvote minus downvote
  - Age account in days
- Community related features (No control)
  - Community age (**people are more excited in the begin**) (**WIP or TODO**)
- Temporal (moment) related features (control)
  - First half of the month (When was it post?)(I do not have money in the end of the month)

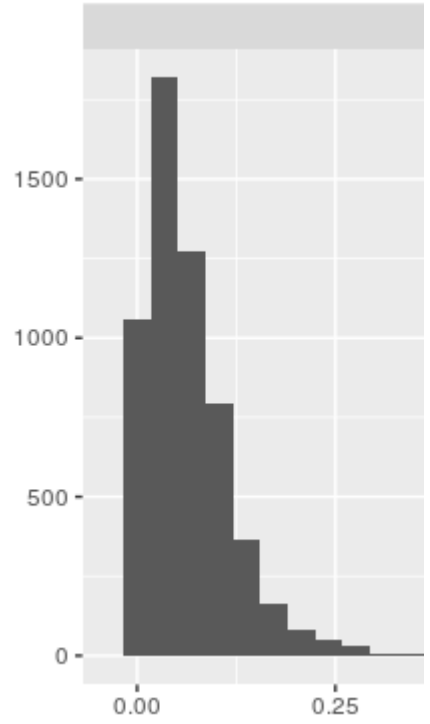
Text Engineering (80% of the time)
(Simple approach) – weekend :) --

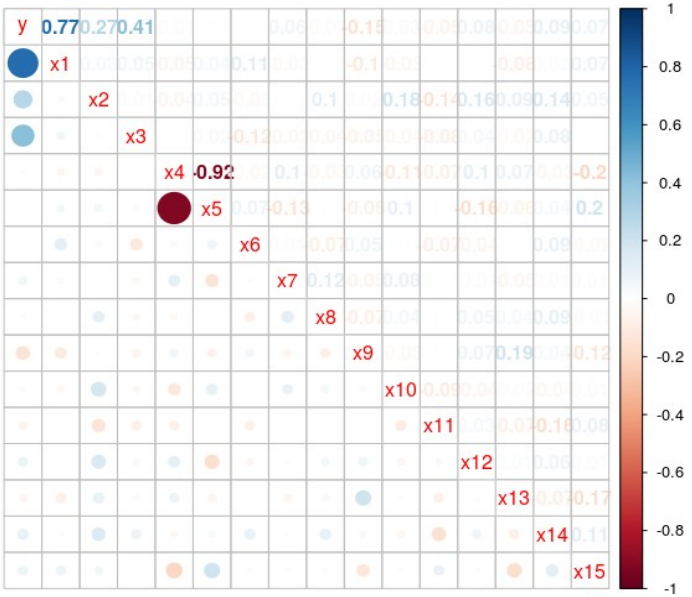
- Sentiment Score (count shared words)
  - 2 Dictionary: positive (+1) and negative (-1) words
  - {pos words: happy, friendly,...} => high positive score
  - {neg words: bad, broke,*#$@#,...} => high negative score
  - Score: **#pos shared words - #neg shared words**
- Narrative Scores (count shared words and divide by the number of word in the dictionary)
  - 3 dictionaries: money, job, family, desire, student
  - Scores: **count shared words/(#words in the dictionary)**
  - Shameless stolen (***Inspired***) in the article: **How to Ask for a Favor: A Case Study on the Success of Altruistic Requests (Stanford , Max Planck Institute)**

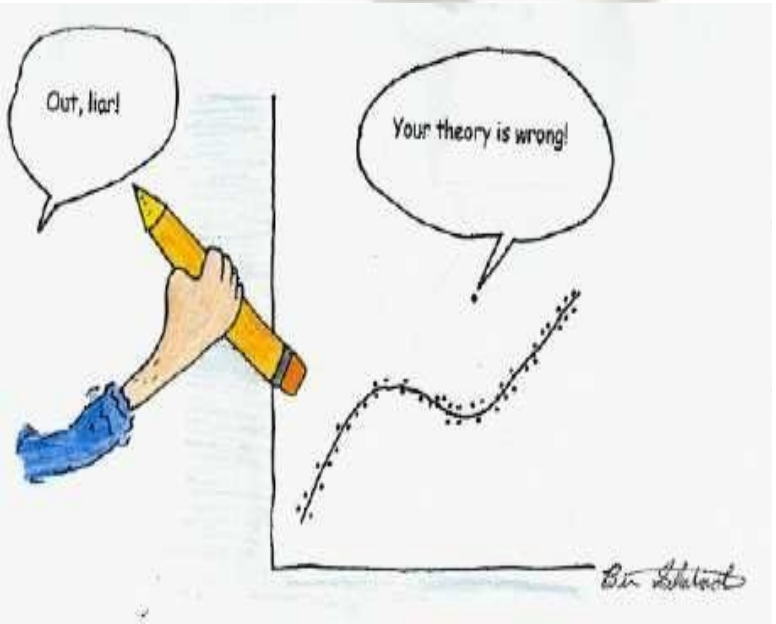
Data Preparation and issues

Issues:

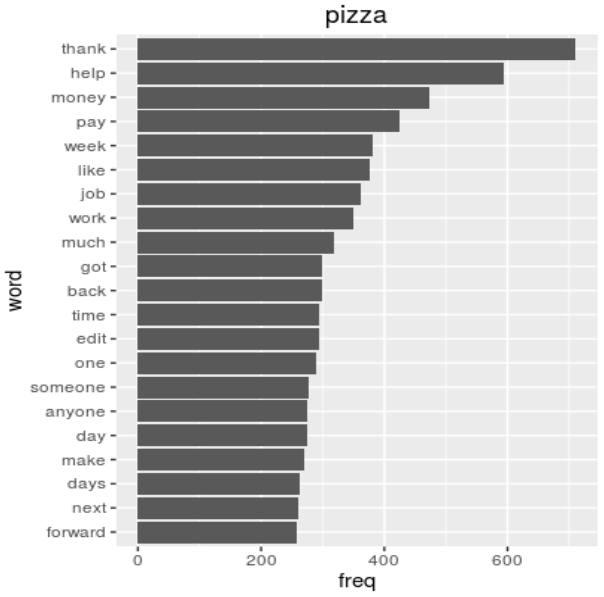
- Skewness and long tails (use median values) **:: Ignored** (work with median as possible)
- Concentration (You need variation to discover relation) **:: removed var**
- Outliers **:: removed points**
- Discrepant value ranges (X1: [0 – 10], X2: [0 – 10k]) **:: scale** [0 - 1]
- Multi-colinearity (karma and age account)**:: ignored** (predictive model)
- Classes were Unbalanced (pizza: 1397; no pizza: 4274) **:: re-balance**
- Text: stop words (a, by, the, in); punctuations, derived words working); synonyms :: **removed**

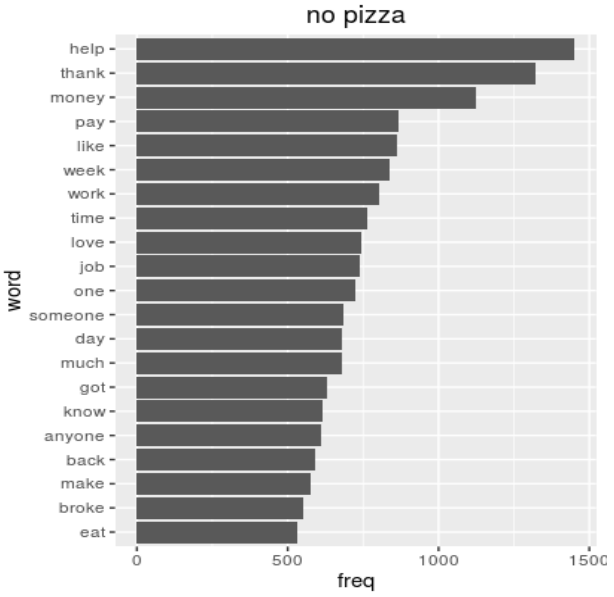




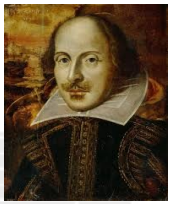


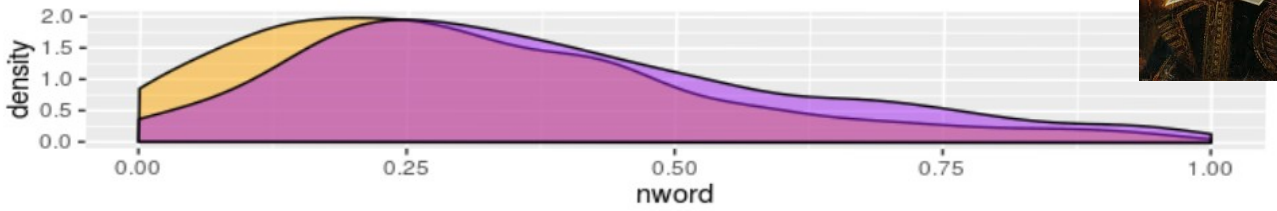
What are the most frequent words in each group?

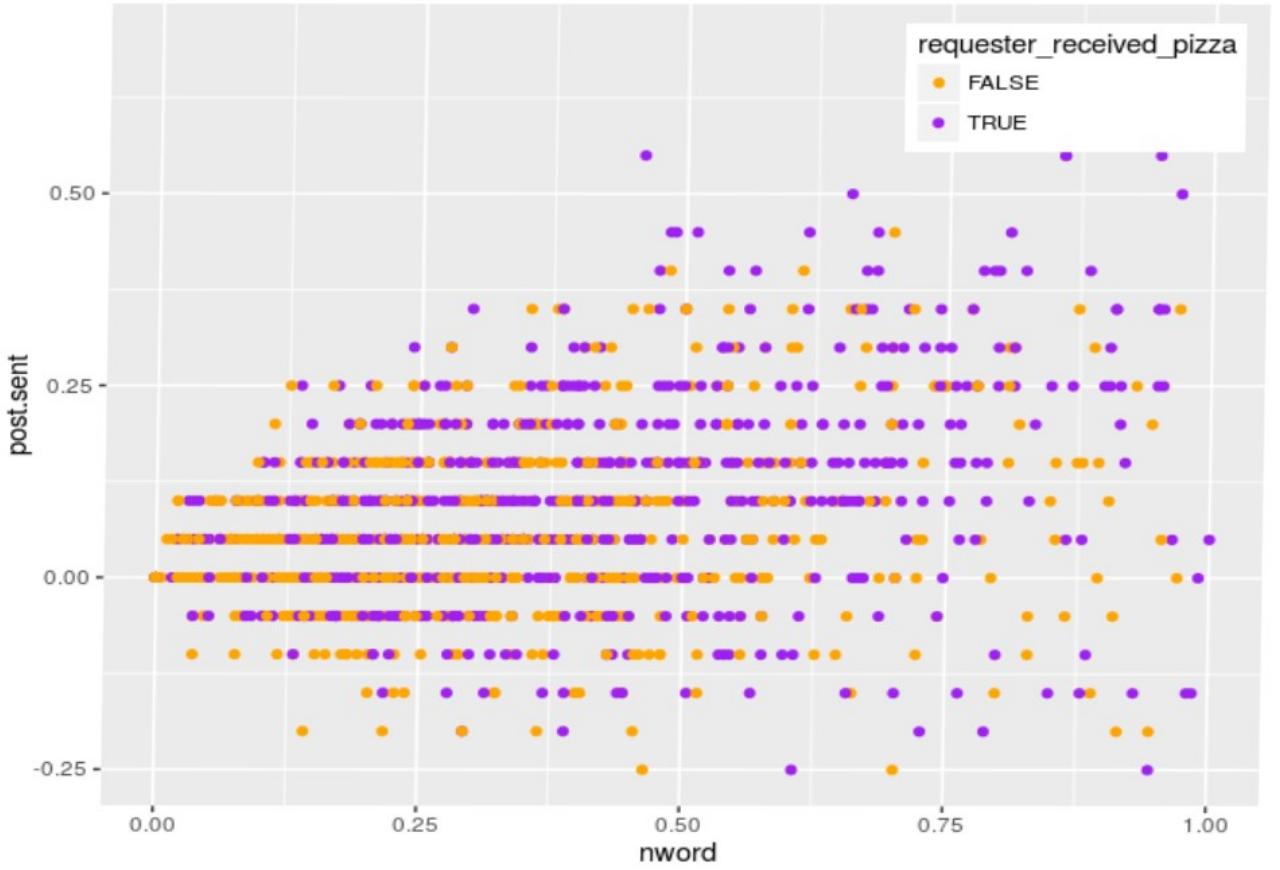


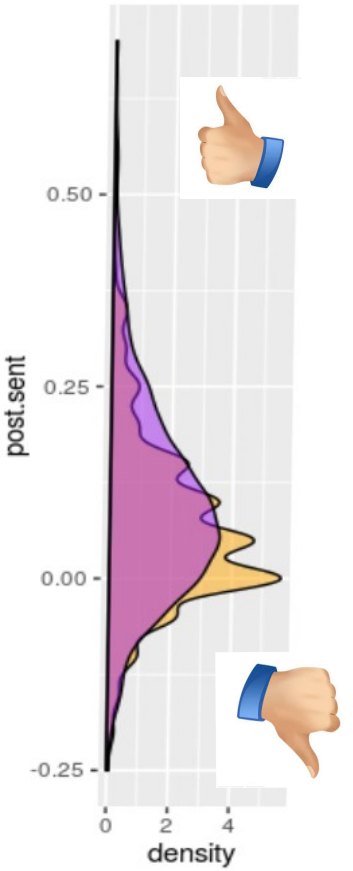


Get Pizza; #words and being positive





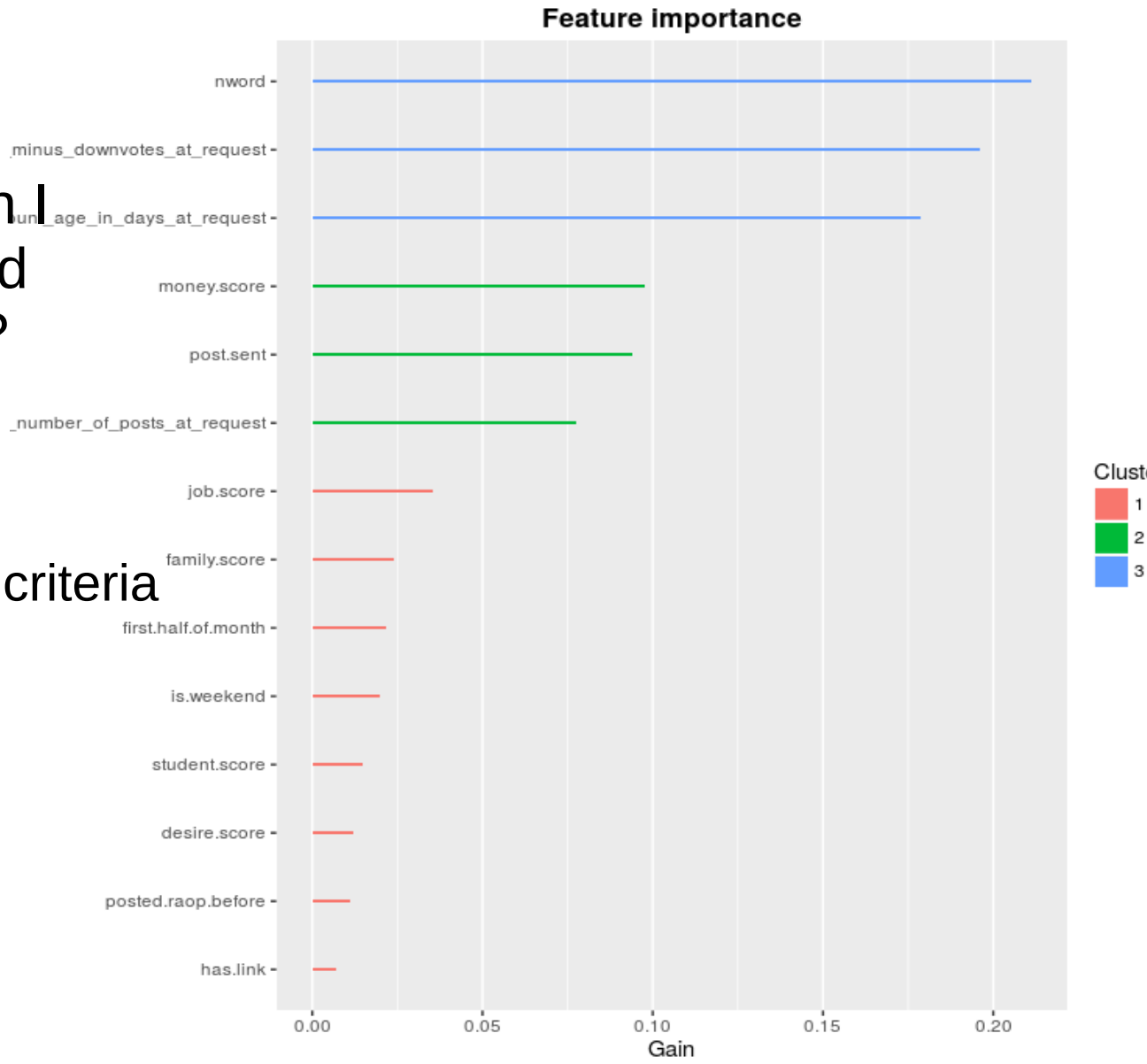




Model

Features Selection: Xgboost relative importance

- Built a ensemble of tree
- What happened in Y when I change one X variable and keep the rest unchanged? What is the impact in Y?



Features selected base this criteria

- nword (compelling)
- karma (nice user; status)
- requester account age
- money.score
- post sentiment score
- #of post at request

RAoP Model: Logistic Regression

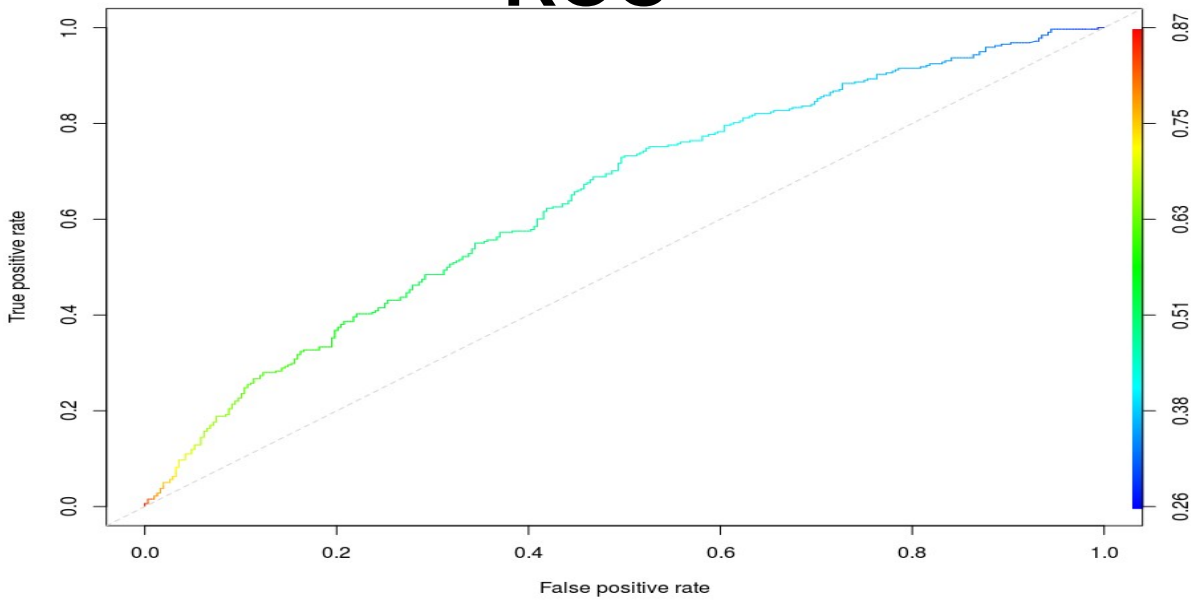
```
Call:
glm(formula = requester_received_pizza ~ requester_upvotes_minus_downvotes_at_request +
    nword + requester_account_age_in_days_at_request + money.score +
    post.sent + has.link + first.half.of.month + posted.raop.before,
    family = binomial(link = "logit"), data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0625  -1.0878  -0.8319   1.1491   1.5680

Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -0.8833     0.1298  -6.804 1.01e-11 ***
requester_upvotes_minus_downvotes_at_request  0.2255     0.2779   0.811 0.417089
nword                                       0.5858     0.3337   1.756 0.079156 .
requester_account_age_in_days_at_request    0.2270     0.2721   0.834 0.404005
money.score                                 5.5512     1.5327   3.622 0.000293 ***
post.sent                                   1.7254     0.4884   3.532 0.000412 ***
has.linkTRUE                                0.5138     0.2011   2.554 0.010635 *
first.half.of.monthTRUE                     0.1525     0.1084   1.407 0.159497
posted.raop.beforeTRUE                      0.9909     0.2807   3.530 0.000415 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **AUC: 0.64 (TEST)**
- **Accuracy: 0.59** (be right in both direction)
- **Sensitivity or Recall: 0.65** (hability to detect)

ROC



Recommendation based what I learned

- Try to be compelling (***my opinion and intuition, not analysis***). But write more than 100 words (pizza: median 74)
- Add a link in your post (image of a dog)
- If you are facing temporary problem with money, write about that but try to be polite and positive
- Be a nice user helps (Improve your karma)

Next Steps

- Improve variables transformations
  - segment continuous variables (quantiles,deciles)
  - create narrative categorical var: is.money; is.desire; ...
- Explore more variables
  - Add time (trend) in the model (I forgot to include it)
- Make the model “interpretability” (make easy to understand the results)

Post Examples: High pos scores, nwords < 100

“My wife and I are avid pizza fans, but this month has been exceptionally tight fiscally. Having recently put together a new monthly budget (with more reasonable expenses) we've acted more responsibly, but had considerably less fun. **If anyone would be so kind as to reward our new financial responsibility with dietary irresponsibility,** that would be grand. I'm sure the kids would love something more than just bland-yet-wholesome Spaghetti for the third night in a row. Thanks for taking the time to read, and have a great day no matter what."

Like the title says I'm in a tight spot and could really use a warm pizza to lift my spirits and give me a break from KD. **As it stands I won't be able to both pay bills and afford groceries when I get paid so this would be a huge boost. I'm a long time redditor and am happy to PM you from my main account to verify this as well as provide any other info you would like**. EDIT: Thanks to the wonderful DEStudent I will be having pizza for supper! Thank you so much!