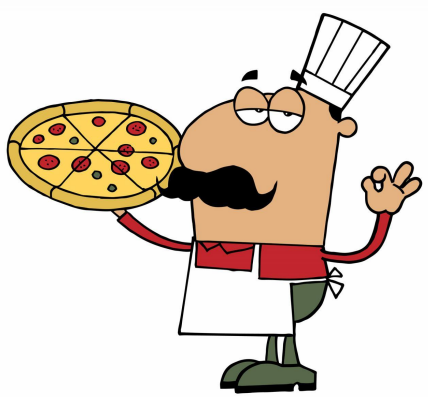
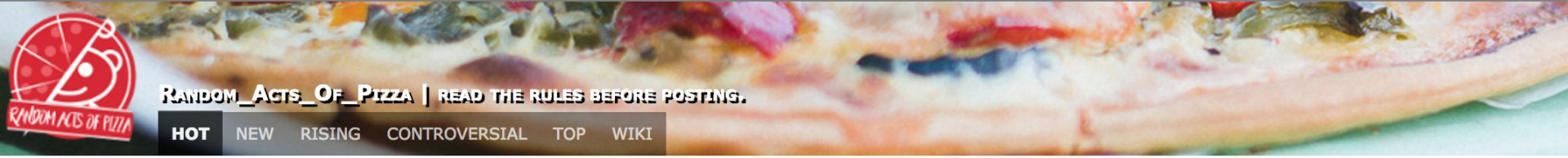
Random Acts of Pizza

An approach to the
Kaggle Random Acts of Pizza competition

8/28/2017

Omar Al Taher, Chris Sanchez, Ted Pham





Overview

- Reddit community
- Users post requests for a free pizza
- Requests can be up or down voted by community members

Problem

- Predict if a request will get funded
- Binary classification
- Pizza | No Pizza

Watch out for private message scams! Read sidebar and wiki for details.

83 [PSA] RAoP is back online! Sorry for the extended downtime!
submitted 5 months ago * by SantaHQ Pizza Tycoon (Gave 24) [MODERATOR (speaking officially)] - announcement
152 comments share report

1 21 [request] fallen on some hard times (not drug related) and have literally 72¢ to my name.
submitted 14 hours ago by ZaddyLongdicc Small Fish (15)
2 comments share report

2 • [Request] Tomorrow is my birthday!
submitted an hour ago by unliterate Small Fish (10)
2 comments share report

3 3 [Request] New Job New Place No Cash.[Indiana USA]
submitted 13 hours ago by BigBearMedic Small Fish (15)
2 comments share report

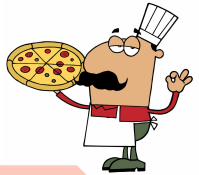
4 15 [Request] UK Master's student - just really could do with a pizza
submitted 23 hours ago by carolionel Small Fish (15)
2 comments share report

5 16 [REQUEST] I burnt my pancakes and could use a pizza?
submitted 1 day ago by Johninja321 Small Fish (15)
2 comments share report

6 6 [request] [UK] a tough past few weeks !
submitted 18 hours ago * by _N64 Small Fish (20)
2 comments share report

7 1 ~~[REQUEST] Birthday Pizza?~~ No Longer Needed
submitted 15 hours ago by AshieKyou Small Fish (15)
5 comments share report

Overview and Approach



Data Processing

- Json Format
- 80/20 train/dev
- 4040 observations
- 31 columns:
  - \+ 19 integers
  - \+ 4 floats
  - \+ 8 objects
  - \+ Boolean outcome
- At_retrieval excluded
- Text Vectorizer

Baseline Model

- Request Text feature only
- **Logistics Regression**

Also tried:
- Bernoulli Naive Bayes
- Support Vector Machine
- KNN

Feature Engineer

- **Numeric**
  - \+ 3
- **Binary**
  - \+ 10

Improved Model

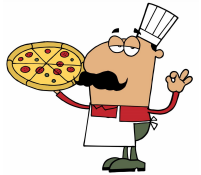
- **Logistic Regression**
  - \+ Eng features
  - \+ Combined TFIDF unigram vector with eng features
  - \+ Prediction on unigrams with eng features

Predict

- **Logistic Regression:**
  - \+ Combined TFIDF unigram vector with eng features
- Train on *full original data*
- Make prediction on test data
- Submit to Kaggle

ROC AUC as METRIC

Data Processing and Feature Engineering



Raw Data

giver_username_if_known
number_of_downvotes_of_request_at_retrieval
number_of_upvotes_of_request_at_retrieval
post_was_edited
request_id
request_number_of_comments_at_retrieval
request_text
request_text_edit_aware
request_title
requester_account_age_in_days_at_request
requester_account_age_in_days_at_retrieval
requester_days_since_first_post_on_raop_at_request
requester_days_since_first_post_on_raop_at_retrieval
requester_number_of_comments_at_request
requester_number_of_comments_at_retrieval
requester_number_of_comments_in_raop_at_request
requester_number_of_comments_in_raop_at_retrieval
requester_number_of_posts_at_request
requester_number_of_posts_at_retrieval
requester_number_of_posts_on_raop_at_request
requester_number_of_posts_on_raop_at_retrieval
requester_number_of_subreddits_at_request
requester_received_pizza
requester_subreddits_at_request
requester_upvotes_minus_downvotes_at_request
requester_upvotes_minus_downvotes_at_retrieval
requester_upvotes_plus_downvotes_at_request
requester_upvotes_plus_downvotes_at_retrieval
requester_user_flair
requester_username
unix_timestamp_of_request
unix_timestamp_of_request_utc

Data Available at Posting

request_text
request_text_edit_aware
request_title
requester_account_age_in_days_at_request
requester_days_since_first_post_on_raop_at_request
requester_number_of_comments_at_request
requester_number_of_comments_in_raop_at_request
requester_number_of_posts_at_request
requester_number_of_posts_on_raop_at_request
requester_number_of_subreddits_at_request
requester_received_pizza
requester_subreddits_at_request
requester_upvotes_minus_downvotes_at_request
requester_upvotes_plus_downvotes_at_request
unix_timestamp_of_request
unix_timestamp_of_request_utc

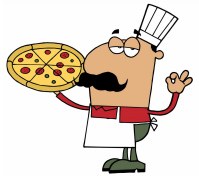
Data with all engineered features

request_text
request_text_edit_aware
request_title
requester_account_age_in_days_at_request
requester_days_since_first_post_on_raop_at_request
requester_number_of_comments_at_request
requester_number_of_comments_in_raop_at_request
requester_number_of_posts_at_request
requester_number_of_posts_on_raop_at_request
requester_number_of_subreddits_at_request
requester_received_pizza
requester_subreddits_at_request
requester_upvotes_minus_downvotes_at_request
requester_upvotes_plus_downvotes_at_request
unix_timestamp_of_request
unix_timestamp_of_request_utc
request_text_n_title
total_length
image_incl
karma
karma_low
timestamp
month
day
time
first_half
requester_grateful
requester_payback
narrative_money
narrative_job
narrative_family
narrative_student
narrative_craving

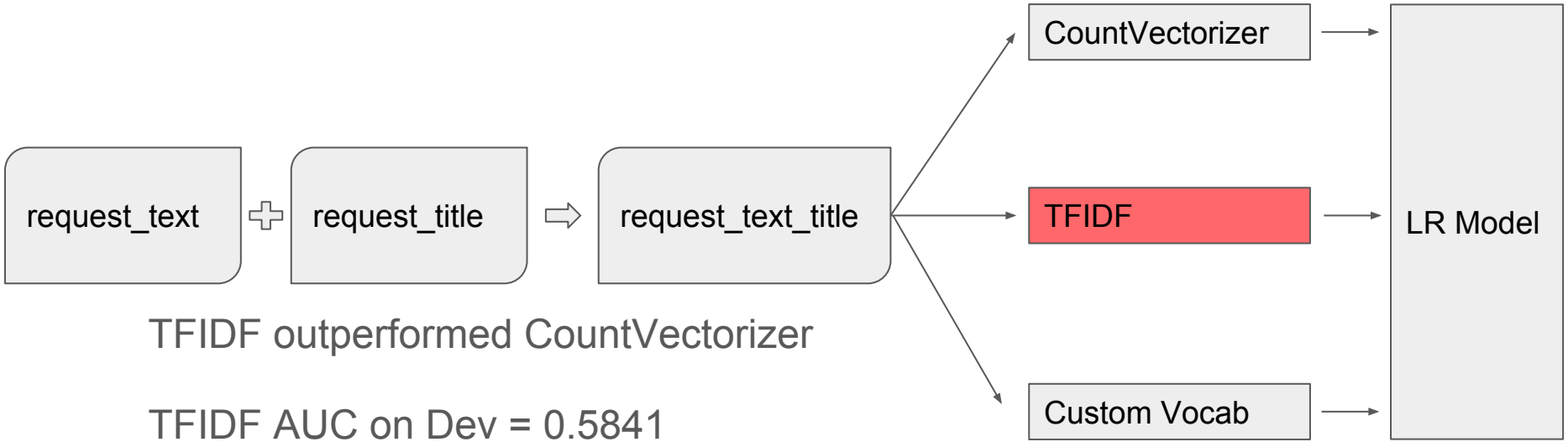
Pruned Data

image_incl
karma_low
time
first_half
requester_grateful
requester_payback
narrative_money
narrative_job
narrative_family
narrative_student
narrative_craving
total_length

Base Line with Unigrams



Based on text features only, fitted to a logistic regression model



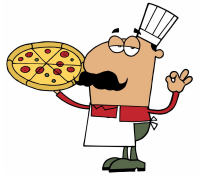
TFIDF outperformed CountVectorizer

TFIDF AUC on Dev = 0.5841

TFIDF Kaggle Score AUC = 0.5845

Custom Vocabulary worked well on the dev set but did not generalize

Feature Engineering from Text and Numeric



- Community Standing
- Request Length
- Time

Numeric — normalized — Binary

- Image_incl
- karma _low
- First_half
- Requester_payback
- Requester_grateful
- Narrative
  - Money
  - Job
  - Family
  - Student
  - Craving

Metadata Based

Text Based

Engineered Features

- TFIDF vectorizer

Unigrams

Unigram Probability (Logistic Regression)

Logistic Regression

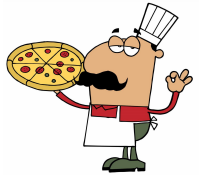
AUC 0.6121

AUC 0.5906

AUC 0.5876

Ref: Althoff, DN-Mizil, Jurafsky, "How to ask for a favor a case study on the success of altruistic request," arXiv 2014

Final Model Dev Scores



Model One: Engineered Features only

- ROC AUC score on dev data = 0.5876

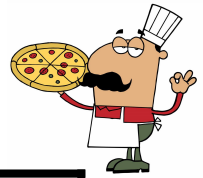
Model Two: Engineered Features + unigram text tokens as features

- ROC AUC score on dev data = 0.6121

Model Three: Unigrams predict_probability + engineered features

- ROC AUC score on dev data = 0.5906

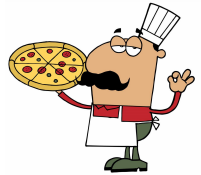
Kaggle Scores



Submission #	Model Features	Score (AUC)
1	Unigram text only	0.5845
2	Optimized Unigram Text (max features)	0.5939
3	Combined with engineered features	0.6028 Good enough for a..... Top 225!

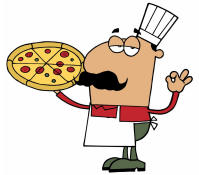
kaggle

Challenges and Lessons Learned



- Feature engineering is a lot of work, but it’s where it’s at
- Care must be taken not to build a model for the dev set
- AUC is a harsh critic, but better to not be fooled by false sense of success
- Did not find a magic algorithm that did particularly better than others

Thank You!



Fun Learning Experience!!!