Kaggle competition:

How many “useful” votes will a yelp review receive

YELP

[redacted]

Send
Follow This

7/16/2013 2 photos

The nicest staff and the best ice cream, that's all you need to know.

But other helpful info... the pistachio is amazing, even if you think you don't like pistachio.
The cotton candy may seem silly but it's actually very delicious - reminds me of Lucky Charms marshmallows!
The soft serve, while often sold out, can be worth the wait.

Try to go early in the day if you can, because the after-dinner crowd results in a crazy long line. But if you're in th
neighborhood, it's an absolute must.


Soft serve

Was this review ...? Useful Funny Cool

Bookmark Send to a Friend Link to This Review

Flag th

[redacted]

fari'ing. Be back July 30th. See you here: http://bit.ly/yelpheIpsBK
, NY

Send Message
Follow This Reviewer

10/16/2012 1 check-in here

Yes, the ice cream is incredible. The salted crack caramel deserves its name. Every flavor I've had has been the decadent treat that ice cream should be. And I can't wait to try all the other ridiculous fun and tasty-sounding flavors like the Bourbon Street, the Barclays Gridlock, the Gather 'Round the Campfire and more.

But the vibe of this place (illustrated in the names of the flavors as well) is what makes it a genuine pleasure to walk through the doors, spend my money, and get just a little bit fatter. Over and over again.

Everyone is so happy to be here. There are smiles on the faces of the scoopers, the parents, the kids (even the ugly ones!*), and the regular old adult customers as well. There's a sign on the wall that the owner can lift 10,000 lbs. (right up over his head!) and that he has perfect pitch, and can talk to animals. I believe it because he created this ice cream shop, and it is equally incredible.

On a warm and sunny day, there's nothing better than getting your ice cream and eating it on one of the benches situated on the fake grass outside. The only downside would be that I was nearly castrated by the spiked fence as I leapt over to chase the girl with the "Free Hugs" t-shirt who walked by.

Otherwise, Ample Hills is perfection.

---

*Just kidding. Ugly kids don't have the capacity for joy.

Listed in: yelp related., reasons why brooklyn is..., this is why i'm fat., delicious desserts in NYC., believe the hype.

Was this review ... Useful (8) Funny (10) Cool (5)

Bookmark Send to a Friend Link to This Review

Flag this review

Kaggle Competition

- Sponsored by Yelp.
- Ran from March 27th to June 30th.
- Yelp tracks 3 community-powered metrics of review quality:
  1. **Useful**
  2. **Funny**
  3. **Cool.**
- Other data includes **freshness of a review**.
- Problem statement: can we predict the number of useful votes a review will get ?

Data set

- Reviews: 230K
  - date, text, stars, cool, funny, useful
- Unique Users: 43K
  - # of reviews, # of useful reviews, avg stars, cool, funny, useful
- Business: 11K
  - #of reviews, # of stars
- Checkin: 8K
  - Time of day, type of biz

Methodology

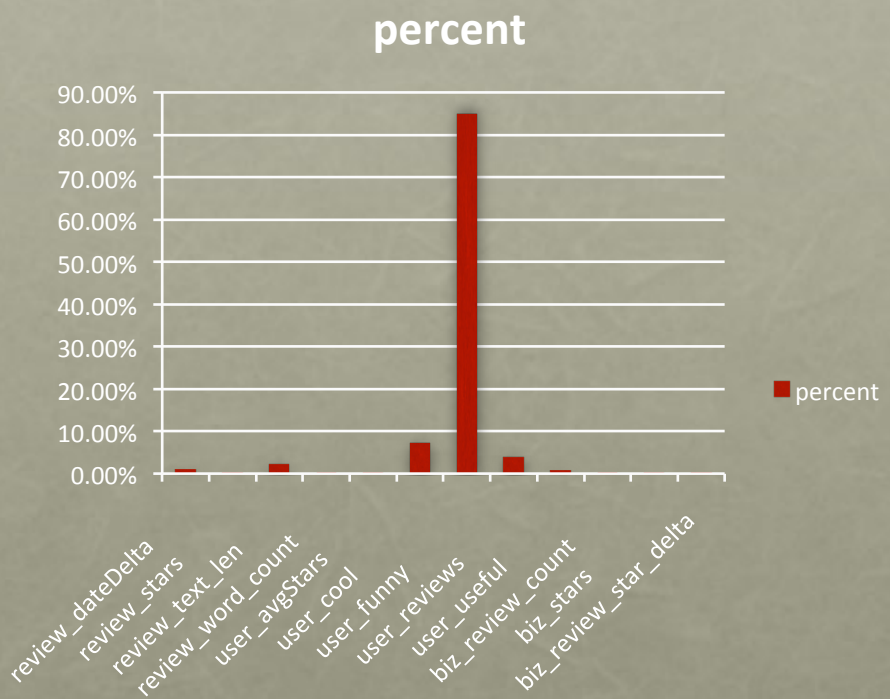
- Data:
  - Collection: Get JSON data into DB
  - Preparation: Normalize data across tables
  - Analysis: Choose descriptive features and calculate some features
- Algorithm: Random Forest
  - Training: ran it with 75% of the training data
  - Testing: ran it with the other 25% to get predictions

Descriptive Features

- Review:
  - Date
  - Stars
  - Text
- User:
  - Average Number of stars given
  - Cool votes
  - Funny Votes
  - Useful Votes
- Business:
  - Count of reviews
  - Stars
- Computed
  - Review star delta
  - Date Delta from 1/1/13
  - Text Length
  - Text Word Count

Observations

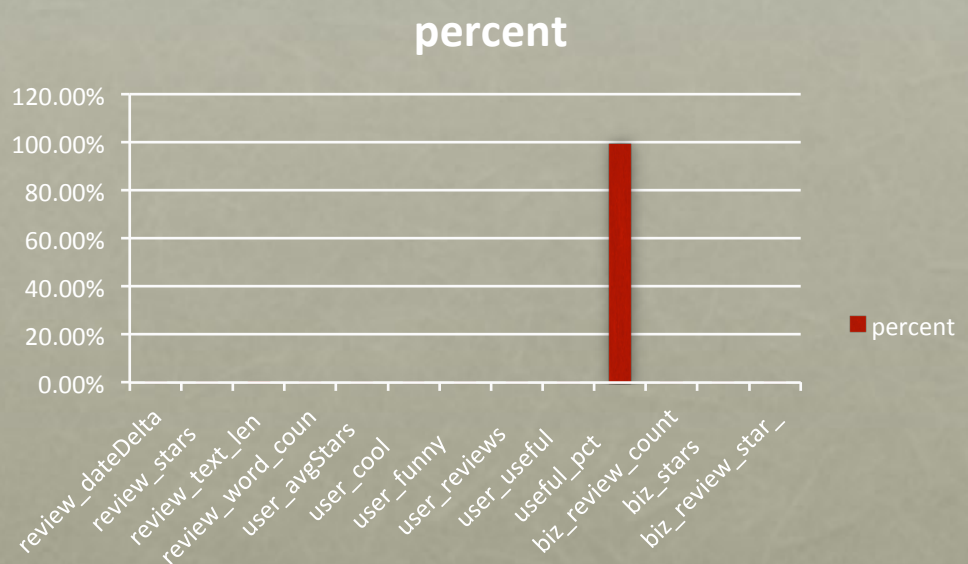
features	percent
review_dateDelta	0.90%
review_stars	0.23%
review_text_len	2.13%
review_word_count	0.17%
user_avgStars	0.17%
user_cool	7.17%
user_funny	84.73%
user_reviews	3.76%
user_useful	0.71%
biz_review_count	0.02%
biz_stars	0.00%
biz_review_star_delta	0.01%



RMSLE = 0.4797

Observations 2

Features	percent
review_dateDelta	0.23%
review_stars	0.06%
review_text_len	0.54%
review_word_coun	0.08%
user_avgStars	0.01%
user_cool	0.02%
user_funny	0.07%
user_reviews	0.01%
user_useful	0.01%
useful_pct	98.94%
biz_review_count	0.02%
biz_stars	0.00%
biz_review_star_	0.01%



RMSLE = 0.4791

Kaggle

- Winner: RMSLE => 0.4404
- Mine:
  - RMSLE => 0.4791
  - Place: 52 out of 352

Tools

- MongoDB
- Python
- Pandas
- Scikit learn: RandomForestRegressor
  - 150 trees
  - 6 mins
  -

Challenges

- Too many unknowns to start with (mongodb, python, ML algorithms)
- First version of the program took over 4 days to figure out. (2 weekends). Second version took me 45 mins and it was 1/6 the number of lines of code.

Take aways

- iPython: awesome interactive python shell. Really good for quickly testing stuff
- Pandas: great library! DataFrame has so much stuff like join, merge, PivotTable. Read_csv.
- Python: really quick to prototype
- Mongo: didn't really get to fully exploit it's value. Not sure of perf vs. straight python.