Data Analytics Project

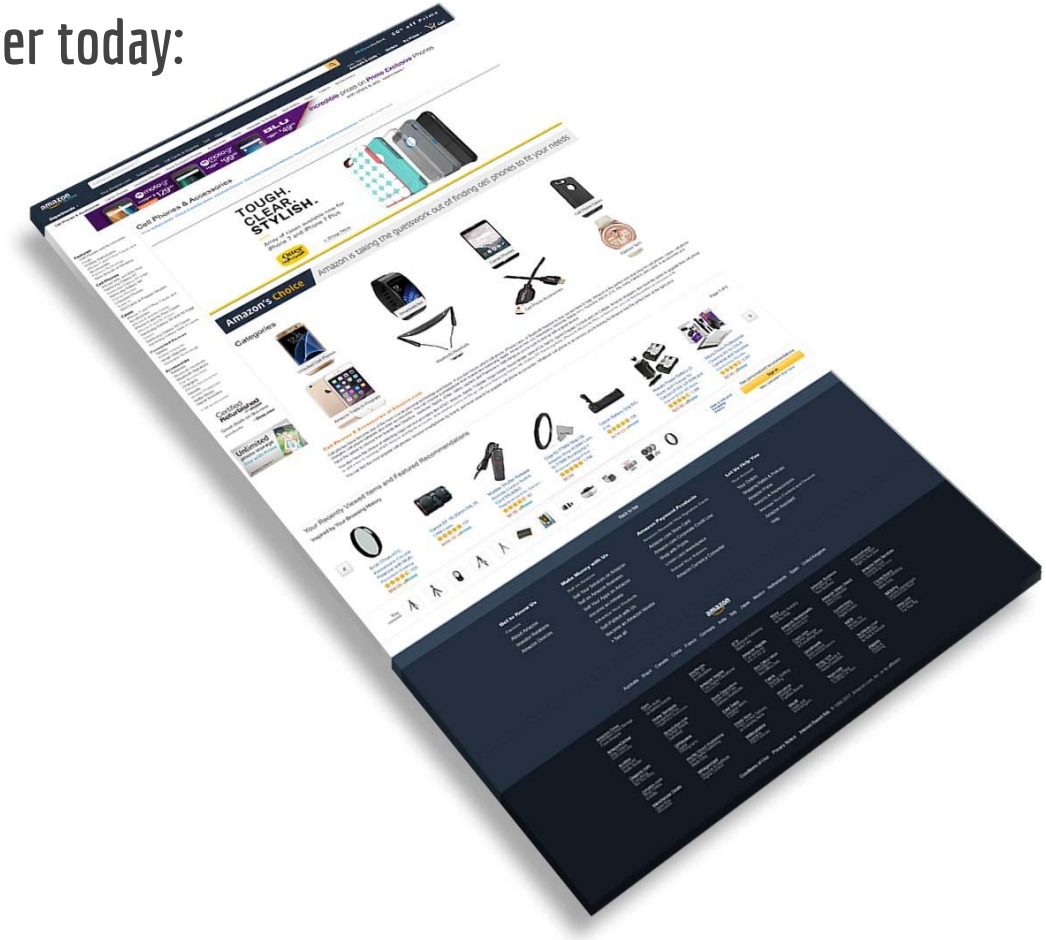
Insights from Amazon Product Reviews

Group 4

Anna Zhang, Anuj Shah, Egemen Can Gök,
Milisha Merchant, Sang Do, Ying Jin

Here are the topics that we will cover today:

- Introduction of Our Project
- Research Questions
- Data Mining Techniques
- Data Analytics
- Key Takeaways



Class Survey Time!

- How many of you purchased your cell phones through online retailers?
- How many of you read customer reviews before placing your order?
- How many bad reviews does it take to change your mind about purchasing a product online?

Let's take a look at some interesting stats!

105% purchase increases for site visitors who interact with both reviews and customer questions/answers (Bazaarvoice, Conversation Index, Q2 2011).

63% of customers are more likely to make a purchase from a site which has user reviews (iPerceptions, 2011).

61% of customers read online reviews before making a buying decision, and they are now essential for ecommerce sites.

Reviews produce an average **18%** uplift in sales (Reevoo).

What does a review look like on Amazon?

Rating

Summary

Reviewer

Helpfulness:
Was the review helpful to you? [Yes][No]

Uber Awesome, In Fact you can even order an Uber ride from it!!

By Michael S TOP 50 REVIEWER on July 24, 2016

Color: Black | **Verified Purchase**

LOVE OUR NEW ECHO! I have been watching the reviews online and checking with friends that have purchased the Echo to see how much they liked or disliked its features. Last person I talked to went on and on about all the things there were using it for and that persuaded me it was time and Amazon Prime Day was the perfect opportunity to go for it. Amazon did a fantastic job of creating this tubular info-taining command center! There are so many cool and awesome things its able to do that I'll hit the highlights that work for our household. First, we love that it follows your voice in the room (the circle lighting will show which direction it is 'listening'), the speaker is wonderfully balanced, so whether listening to music, the news or to Alexa speaking, I have nothing but high marks for its sound quality, given its size. Next, set up (after downloading the app to our iPhones) was quick, easy and very intuitive. The more you look over the app, the more you will realize a world of 'skills' (as Amazon refers to them - we've nicked named them "echolettes" LOL) that the unit is able to perform once they are turned on and you master the right sequence of keywords to initialize them. We've added things to shopping lists, while asking about the weather and our calendar of events and then asked Alexa to change the temp of our Nest thermostats in various parts of the house, simply by saying her name and then our commands, sometimes sitting in the living room or simply pass through - she is always there listening and ready. We've ordered some LED programmable lights and I can hardly wait for Alexa to help set the mood in the house, room to room, all from a simple voice request. Read more ›

68 Comments | 9,248 people found this helpful. Was this review helpful to you? Yes No Report abuse

"... Amazon makes the best of both the positive and negative reviews easy to find. And [the voting] feature, based on our calculations, is responsible for more than $2,700,000,000 of new revenue for Amazon every year. "

- Business Insider, 2009

What questions are we trying to answer?

- How do the **features of a product review** affect its level of **helpfulness**?
  - What ratings (1-star to 5-star) are the most helpful?
  - What is the correlation between each feature (i.e. word count) and the helpfulness of the review?

- **Customer Demographics**
  - Where do online customers of *Cell Phones and Accessories* live?
  - Which states are more likely to shop online?
    - Percentage of Online Buyers of the Product vs. Entire State Population

- What **words** are **important** in the reviews for *Cell Phones and Accessories*?
  - What are the most common words in these reviews?
  - How do the words used reflect consumer sentiment?

Where is our data coming from?

- **Source**: UCSD, Julian McAuley's Amazon Product Data
- **Data**: Millions of Product Reviews and Metadata Divided into Different Categories
  - Reviews:
    - Ratings
    - Text Information
    - Helpfulness Votes
  - Metadata:
    - Descriptions
    - Category Information
    - Price


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband...
                        ..Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


- **Analysis**: We chose to focus on the *Cell Phones and Accessories* datasets.

What does our data look like?

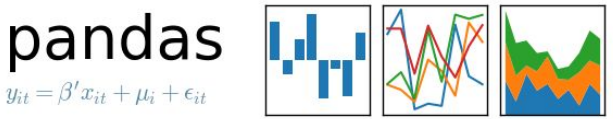
	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	120401325X	[0, 0]	4.0	They look good and stick good! I just don't li...	05 21, 2014	A30TL5EWN6DFXT	christina	Looks Good	1400630400
1	120401325X	[0, 0]	5.0	These stickers work like the review says they ...	01 14, 2014	ASY55RVNIL0UD	emily l.	Really great product.	1389657600
2	120401325X	[0, 0]	5.0	These are awesome and make my phone look so st...	06 26, 2014	A2TMXE2AFO7ONB	Erica	LOVE LOVE LOVE	1403740800
3	120401325X	[4, 4]	4.0	Item arrived in great time and was in perfect ...	10 21, 2013	AWJ0WZQYMYFQ4	JM	Cute!	1382313600

where

- `reviewerID` - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- `asin` - ID of the product, e.g. 0000013714
- `reviewerName` - name of the reviewer
- `helpful` - helpfulness rating of the review, e.g. 2/3
- `reviewText` - text of the review
- `overall` - rating of the product
- `summary` - summary of the review
- `unixReviewTime` - time of the review (unix time)
- `reviewTime` - time of the review (raw)

Techniques Used

- Python - Write Code
- Web Scraping - Collect Data
- JSON (JavaScript Object Notation) - Extract Datasets
  - Useful for Representing/Storing Semistructured Data
- Regular Expressions - Clean Data
- Pandas - Explore Data
- Matplotlib - Plot Results
  - Bar Graphs, Pie Charts, Scatter Plots
- NLTK (Natural Language ToolKit)
- WordCloud, Stopwords
- Machine Learning



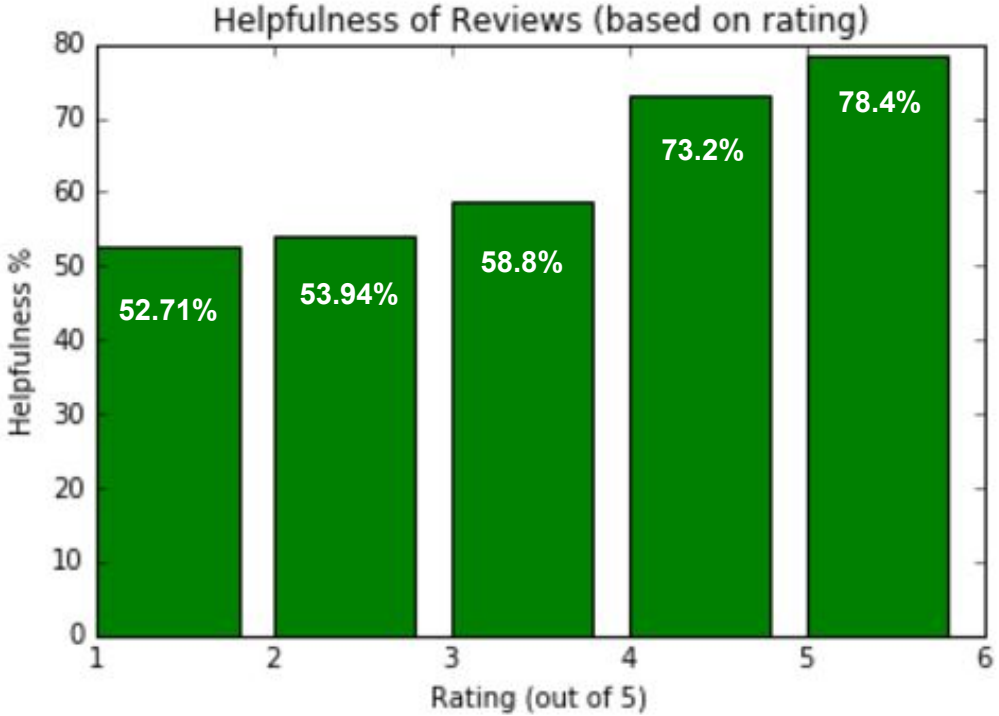
General Statistics

Cell Phones and Accessories

Number of Reviews: 194439
Average Rating (of those who left a review): 4.13

Number of Ratings: 3447249
Average Rating: 3.81
Date of First Rating: 11/16/1999
Date of Latest Rating: 07/22/2014
Number of Users (who only rated the products): 3252810

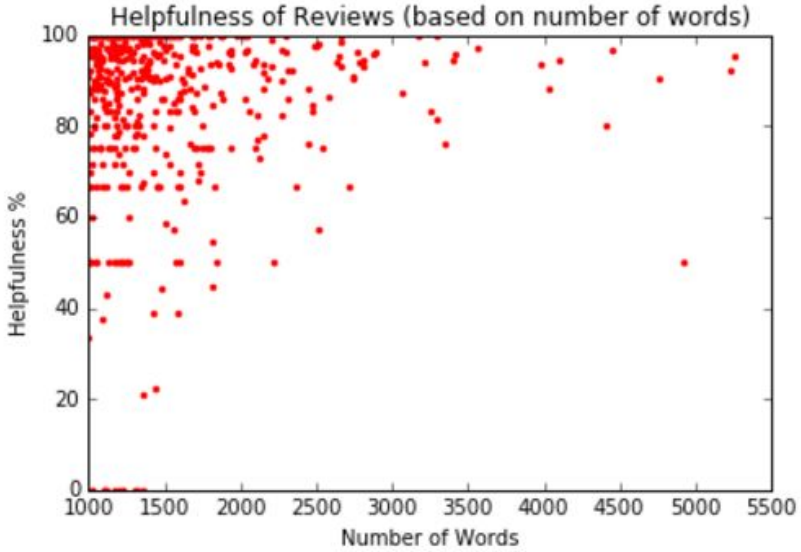
Rating vs. Helpfulness



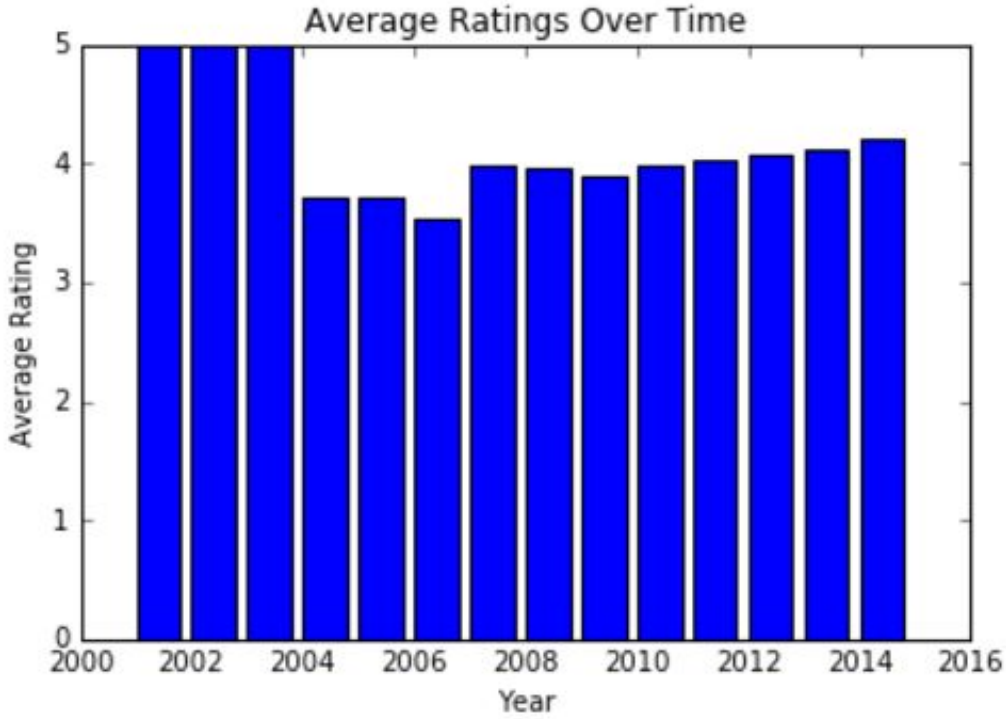
$$Helpfulness = \frac{Number of Upvotes}{Total Number of Votes} * 100$$

Length of Review vs. Helpfulness

```
def helpfulness_and_text(json_list):
    text_helpful = [count_words(x) for x in json_list if has_helpful(x) and count_words(x) >= 1000]
    helpful= [helpfulness(x)*100 for x in json_list if has_helpful(x) and count_words(x) >= 1000]
    plt.plot(text_helpful,helpful, 'r.')
    plt.xlabel("Number of Words")
    plt.ylabel("Helpfulness %")
    plt.title("Helpfulness of Reviews (based on number of words)")
    plt.show()
```



How have product ratings changed over time?



Web-scraping the Rater's Location

```
def user_locations(links):
    result = []
    num = 0
    for link in links:
        try:
            time.sleep(random.randrange(5,10))
            req = urllib2.Request(link, headers = {'User-Agent': 'Mozilla/5.0'})
            content = urllib2.urlopen(req).read()
            bs = bs4.BeautifulSoup(content, "lxml")
            if bs.find(class_ = "a-fixed-right-grid location-and-occupation-holder"):
                result.append(str(bs.find(class_ = "a-fixed-right-grid location-and-occupation-holder").span.string))
                num += 1
                print "{} location(s) found!".format(num)
        except urllib2.HTTPError:
            print "HTTPError: Unable to connect to webpage."
            return result
        except AttributeError:
            print "AttributeError: Bot was detected."
            return result
    return result

ul = user_locations(user_links)
print ul
```

Toocool

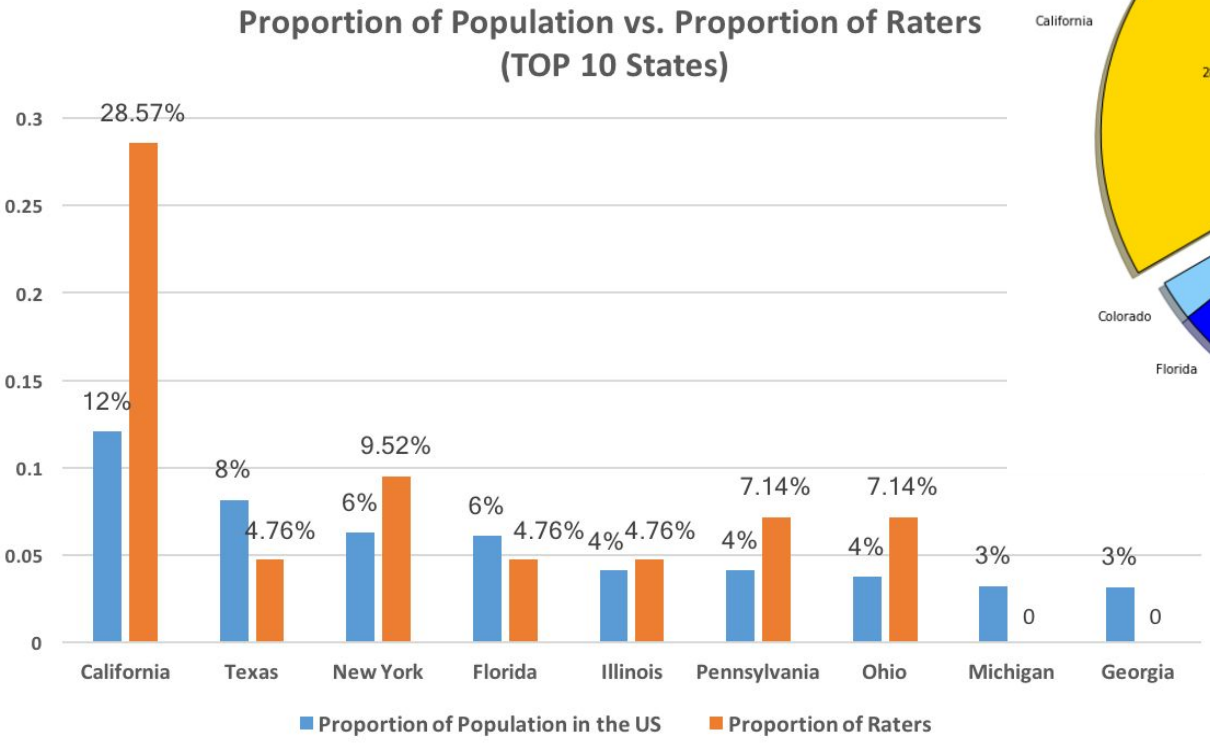
Professional Photographer | California

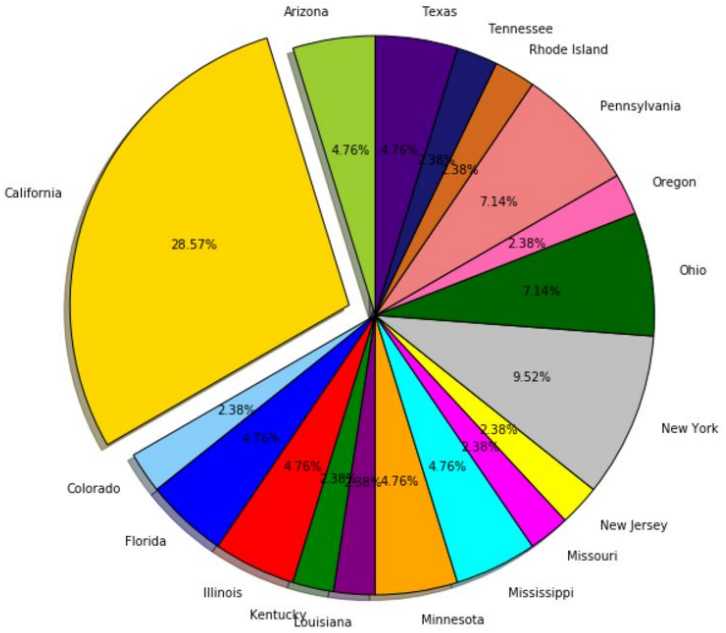
Nah.

See more

Example User Profile

Locations of Raters





Key Findings:

- California is indeed tech-savvy!
- Californians love to share their purchasing experiences! 😊

What do negative and positive reviews look like?

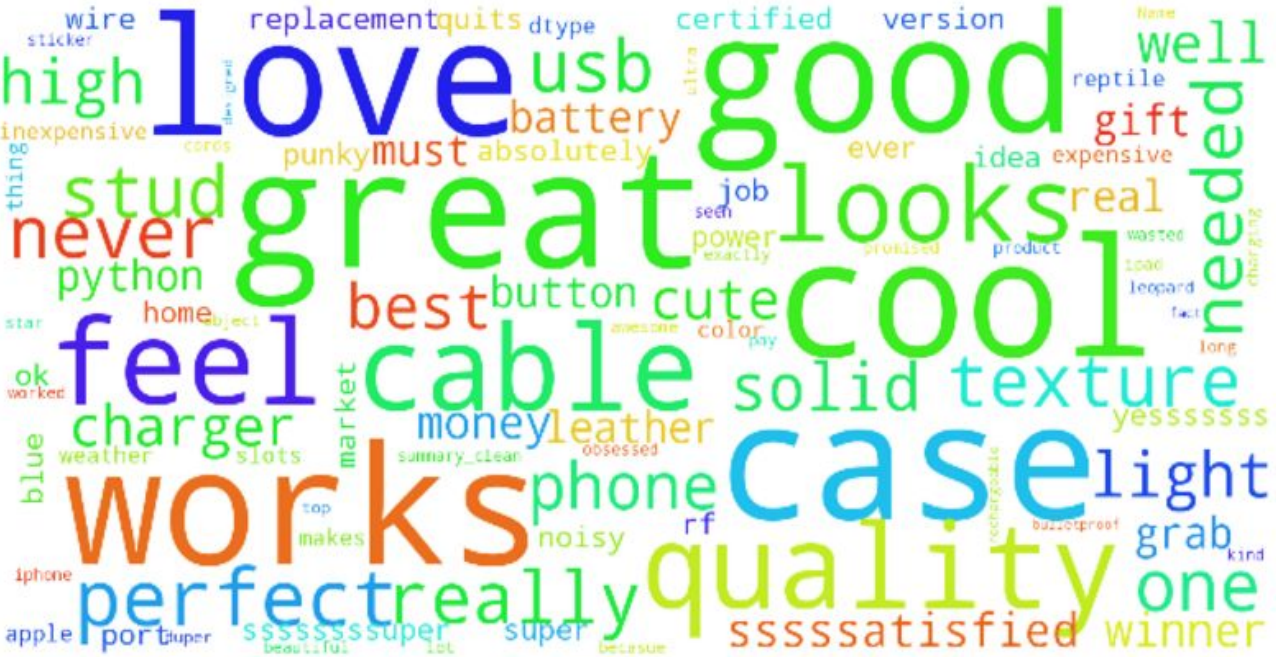
```
Summaries of Lowest Reviews:
1) not a good Idea
2) Horrible
3) don't waste your money, pay more and buy one at Walgreen s.
4) Be careful
5) bad experience


Summaries of Highest Reviews:
1) This works just perfect!
2) Great replacement cable. Apple certified
3) Real quality
4) I really like it becasue it works well with my Life Proof ...
5) I have wasted a lot of money on cords
```

Sentiment Analysis

	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime	sentiment
0	120401325X	[0, 0]	4.0	They look good and stick good! I just don't li...	05 21, 2014	A30TL5EWN6DFXT	christina	Looks Good	1400630400	positive
1	120401325X	[0, 0]	5.0	These stickers work like the review says they ...	01 14, 2014	ASY55RVNIL0UD	emily l.	Really great product.	1389657600	positive
2	120401325X	[0, 0]	5.0	These are awesome and make my phone look so st...	06 26, 2014	A2TMXE2AFO7ONB	Erica	LOVE LOVE LOVE	1403740800	positive

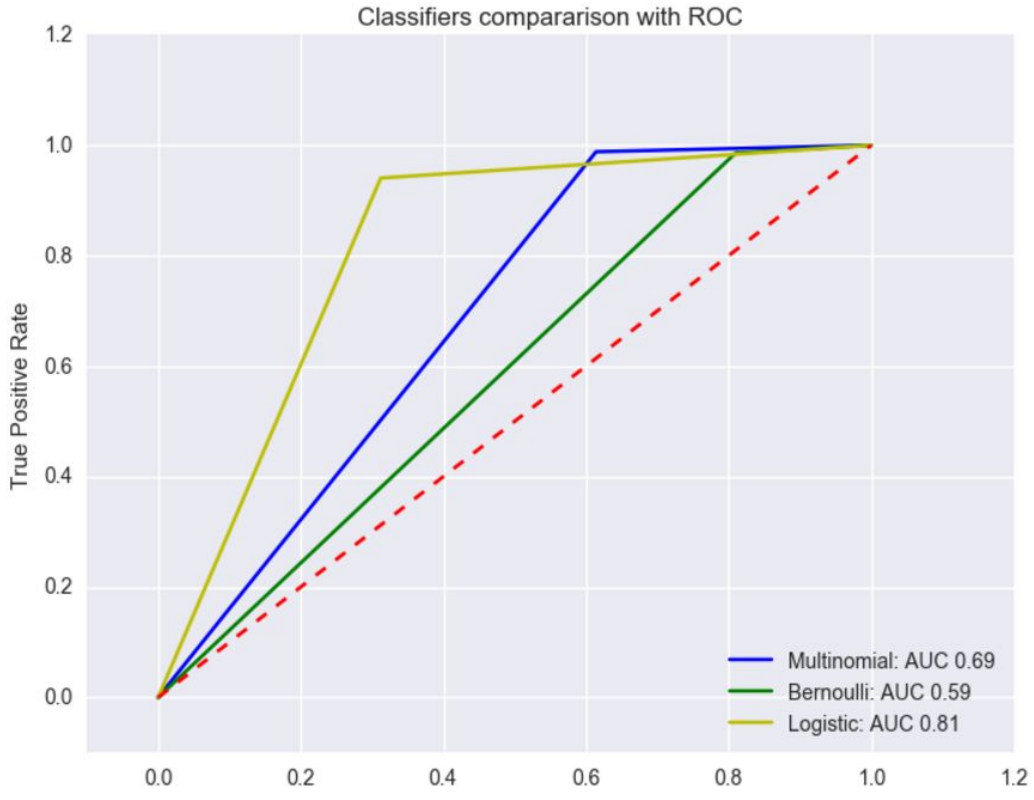
	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime	sentiment
5	120401325X	[1, 2]	3.0	These make using the home button easy. My daug...	10 12, 2013	APX47D16JOP7H	RLH	Cute	1381536000	negative
7	3998899561	[1, 2]	1.0	it worked for the first week then it only char...	11 21, 2013	A6FGO4TBZ3QFZ	Abdullah Albyati	not a good Idea	1384992000	negative
19	6073894996	[0, 0]	1.0	It worked great for the first couple of weeks ...	05 29, 2013	A2INSXDTE08WSJ	Barbie	Horrible	1369785600	negative

Word Cloud of Positive Summaries



Word Cloud of Negative Summaries

Building the Machine Learning Model



Building the Machine Learning Model

	feature	coef
279297	not	-24.652392
469819	worst	-22.908260
228689	junk	-19.259733
389513	terrible	-18.336070
198995	horrible	-18.154958
436949	useless	-18.074663
113841	doesn	-18.024830
320763	poor	-17.651222
132739	fake	-17.297670
111321	disappointing	-16.699974
110050	didnt	-16.095040
54082	broke	-15.739123

109533	didn	-15.543717
111166	disappointed	-15.512193
104951	defective	-15.222079
29013	at best	-15.191436
298196	one star	-13.916473
114783	doesnt	-13.851584
132392	failed	-13.811773
415475	three stars	-13.622162
470397	worthless	-13.414049
321297	poorly	-13.303238
53606	breaks	-13.291893
160576	garbage	-13.284023
469747	worse	-13.182011

Building the Machine Learning Model

281398	not for protection	9.844372
280099	not brick	9.941719
323977	powerful	10.016244
46216	better than	10.161186
371181	solid	10.180438
463385	wonderful	10.224061
284588	not too bulky	10.272810
278102	no problems	10.667337
199166	horrible review	10.851091
180563	great little	10.861661
440642	very good	10.875459
455309	why not	10.951802
424099	too cute	10.959413

282337	not knockoff	11.317576
283781	not so bad	11.377443
277637	no issues	11.389202
8111	amazing	11.659738
137476	finally	11.738526
259296	months broke	11.922342
277817	no more	12.509738
277023	no bubbles	12.957308
284831	not very bad	14.614576
133335	fantastic	14.744613
284177	not terrible	16.339913
42436	best	19.487968
279794	not bad	21.983740

Predicting the Positive and Negative Probability of Product Reviews

1 This is the first battery case I have had for my Galaxy S4. The S4 fits very well, is slim and doesn't add much weight to the Ga laxy S4. It doubles the battery life. You can charge either the battery, the phone or both. There is a handy on-off switch with leds to indicate the level of charge.The battery case came on time and was packaged well. Well worth the price.

Sample estimated as POSITIVE: negative probability 0.43%, positive probability 99.57%

2 Item arrived in great time and was in perfect condition. However, I ordered these buttons because they were a great deal and inc luded a FREE screen protector. I never received one. Though its not a big deal, it would've been nice to get it since they claim it comes with one.

Sample estimated as POSITIVE: negative probability 12.54%, positive probability 87.46%

3 Item arrived in great time and was in perfect condition. However, I ordered these buttons because they were a great deal and inc luded a FREE screen protector. I never received one.

Sample estimated as POSITIVE: negative probability 3.30%, positive probability 96.70%

4 I am disappointed that the 1A didn't work with my iPad. That's what I get for buying a cheap adapter.

Sample estimated as NEGATIVE: negative probability 97.60%, positive probability 2.40%

Predicting the Positive and Negative Probability of Product Reviews

This is the first battery case I have had for my Galaxy S4. The S4 fits very well, is slim and doesn't add much weight to the Ga
laxy S4. It doubles the battery life. You can charge either the battery, the phone or both. There is a handy on-off switch with
 leds to indicate the level of charge.The battery case came on time and was packaged well. Well worth the price.

Sample estimated as POSITIVE: negative probability 0.43%, positive probability 99.57%

	feature	coefficient
37568	battery life	-1.828719
36408	battery	-0.12111
34061	bad battery	-3.46806
85249	cheap battery	0.785204
166064	good battery	2.187514

Conclusion

- Reliability of Consumer Feedback
- Better CRM with Data Analysis: Cater to Location
- Other Features Worth Analyzing:
  - Evaluate the readability of each review (grammar, language, etc).
  - Evaluate each customer's expectation when purchasing.
- By predicting the helpfulness of each review, customers will be able to see actual helpful reviews before anyone has voted.
- What makes your product review more helpful?

Questions? Comments? Feedback?