Individual Project: Beer Recommendations

CMPE 256 - Large Scale Analytics

Summer 2019

Mario Yepez <mario.yepez@sjsu.edu> - 012556050

Table of Contents

Abstract

One of the main motivators that drive data science in the industry is the ability to produce actionable insights that can then be utilized to provide a net benefit to a company. This can come in many forms such as saving money or boosting sales. Recommendation systems can help achieve this by predicting the preference that a user will give to a certain item. One of the industries that can highly benefit from doing so is the beer market. The beer market is estimated to be worth nearly $120 billion a year in the United States alone [1] and one of the questions that gets asked by companies is "How do I find more potential customers?". Likewise many people who consume beer often wonder "What's the next beer that I should try?". These are the questions I'll attempt to answer by applying the techniques learned in this class on a dataset of beer ratings from the RateBeer website and creating an algorithm that recommends beers to users and a beer to users.

Introduction

People all around the world consume alcoholic beverages and beer in particular is perhaps the oldest and most widely consumed of the bunch. Beer plays a huge role in various cultures and it is a staple of many social gatherings. Entire platforms have sprung up online where people are able to provide ratings for beers that they have had for the world to share such as RateBeer, Untappd, and BeerAdvocate to name a few. This has created a demand for beers to try, people want to taste new beers that they can then provide reviews for them. Likewise for brewers it is also important to find who these customers are so that they can market their beers to a certain set of users and maximize their profits. What we end up having is two groups that are trying to find each other. On one hand are users who are asking "What is the next beer that I should try?", and on the other hand we have brewers asking "How do I find more potential customers?". By applying the techniques learned in class for building a recommendation system we can attempt to provide an answer to these questions.

I will attempt to provide an answer for the questions above by building a model that allows one to specify the beer properties (i.e the Brewer, Style, and Alcohol by Volume) and then based on these properties find potential customers that exist in the dataset. With these inputs the model will be useful for not just beers that currently exist but also for future beers that a brewer might launch. Let's say that a brewer is planning on producing an IPA that is very strong with an ABV of 15%. The brewer can input the parameters to the model and receive a potential customers that might like said beer based on those properties.

Dataset

The dataset used for this project consisted of beer reviews from the website RateBeer, a popular site where users can give their input on the different types of beers that they have had. It contains data from April 2000 to November 2011 and contains nearly 3 million user reviews for 110,000 beers from 110,000 different users. The raw dataset size amounted to roughly about 1.1GB of data. The dataset was originally available for download on Stanford University's

SNAP project but apparently RateBeer requested that the dataset be pulled [2]. Fortunately someone at the UNC-Charlotte uploaded a copy of the dataset and made it publically available [3]. The following is what data was available for each rating in the dataset.

beer/name	Name of the beer
beer/beerId	Unique Id assigned to the beer
beer/ABV	Alcohol by Volume percentage of the beer
beer/style	The Style of beer in question. i.e India Pale Ale, Porter, Wheat Ale
review/appearance	User rating for the appearance i.e 5/10
review/aroma	User rating for the aroma i.e 6/10
review/palate	User rating for the palate i.e 3/5
review/taste	User rating for the taste i.e 7/10
review/overall	Overall rating for the beer i.e 15/20
review/time	Time that the rating was submitted
review/profileName	Profile name of the user
review/text	Review text from the rating the User gave

The following is a sample of the data from the dataset.

	name	beerId	brewerId	ABV	style	appearance	aroma	palate	taste	overall	time	profileName
0	John Harvards Simcoe IPA	63836	8481	5.4	India Pale Ale (IPA)	4	6	3	6	13	1157587200	hopdog
1	John Harvards Simcoe IPA	63836	8481	5.4	India Pale Ale (IPA)	4	6	4	7	13	1157241600	TomDecapolis
2	John Harvards Cristal Pilsner	71716	8481	5	Bohemian Pilsener	4	5	3	6	14	958694400	PhillyBeer2112
3	John Harvards Fancy Lawnmower Beer	64125	8481	5.4	K�lsch	2	4	2	4	8	1157587200	TomDecapolis
4	John Harvards Fancy Lawnmower Beer	64125	8481	5.4	K�lsch	2	4	2	4	8	1157587200	hopdog

Model Implementation

The model we're building will work as follows. The input is a list of beer properties (Brewer, Style, and Alcohol by Volume) and the output will be a list of potential customers with a score for the likelihood that the potential customer will like the beer, I'll dub this the "likeability score". The method to get the likeability score can be split into two parts. First we fetch the top 20 beers that a user likes the most given the current dataset. Once we have the top 20 beers we'll get the properties for each of them and attempt to find a pattern. The final likeability score will be determined based on how close the input properties and properties of the top 20 beers are.

Step 1: Data Pre-Processing

The first step for building the model was to preprocess the data and identifying what the most important features I'll need to use were. The dataset contains a massive amount of data and attempting to fit that into memory by using Pandas would be near impossible. What I'm predicting is the likelihood that someone will like a beer and for this I need three key pieces, the overall score, the beer properties, and the user. I stripped out the columns I didn't need from the raw dataset and produced a new dataset that only contained the following: beer/name, beer/beerid, beer/ABV, beer/style, review/overall, and review/profileName. By dropping the rest of the columns I ended up reducing the file size from 1.1GB to 192M, a more manageable size that can fit into memory.

The columns for profileName and beerId were then both encoded to numerical values so that we can run them through the algorithms that we'll be using. At first glance I thought the beerIds were all numerical to begin with but it turned out that some ids were given as strings. In the end this is what the final dataset looks like after filtering and encoding. For the purposes of this project I ended up not using all 3 million reviews but instead took a sample to get a prototype going for the final model. The following is a sample of what the final data frame looked like for the recommendation algorithm.

	name	beerId	brewerId	ABV	style	overall	profileName	beerId_cat	userId
0	John Harvards Simcoe IPA	63836	8481	5.4	India Pale Ale (IPA)	13	hopdog	105412	18486
1	John Harvards Simcoe IPA	63836	8481	5.4	India Pale Ale (IPA)	13	TomDecapolis	105412	10465
2	John Harvards Cristal Pilsner	71716	8481	5	Bohemian Pilsener	14	PhillyBeer2112	106618	8190
3	John Harvards Fancy Lawnmower Beer	64125	8481	5.4	K�lsch	8	TomDecapolis	105458	10465
4	John Harvards Fancy Lawnmower Beer	64125	8481	5.4	K�lsch	8	hopdog	105458	18486

Step 2: Finding the Top 20 Beers for Each User

Once our data is formatted nicely we can now proceed to the next step: finding the top 20 beers for each user. In class we went over various methods for predicting missing scores through collaborative filtering. In our case we have very sparse data and after researching online and doing homework assignments, it appears that for this problem Matrix Factorization was the way to go for a problem with this scale. For this project I used Alternating Least Squares to predict the missing ratings for each user and for the input I used a pivot table between the users and the beerIds where the value for each cell would be the overall score for the beer. The overall score is what we will be predicting for the users. The following is an example of the pivot table.

beerId_cat	13	21	33	38	49	57	77	82	99	100	...	109070	109194	109224	109246	109276	109820	109969	110051	110234	110429
userId																					
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
46	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
52	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The following are 5 of the top 20 predicted ratings for a particular user. Their username shows up as "mribm" in the dataset.

Beer Id: 85094
Beer Name: Cigar City Hunahpu�s Imperial Stout - Whiskey Barrel Aged
ABV: 11
BrewerID: 9990
Style: Imperial Stout
Predicted Score for Beer with Id (85094) : 20.918637532323952
Beer Id: 44870
Beer Name: Valley Brew Decadence 12 Cuvee Speciale
ABV: 13
BrewerID: 3490
Style: Abt/Quadrupel
Predicted Score for Beer with Id (44870) : 20.237246742111445
Beer Id: 20681
Beer Name: Valley Brew Uberhoppy Imperial IPA
ABV: 9.5
BrewerID: 3490
Style: Imperial/Double IPA
Predicted Score for Beer with Id (20681) : 20.120632737746007
Beer Id: 67716
Beer Name: N�rrebro Imperial Sk�rgaards Porter Cabernet Barrel
ABV: 9
BrewerID: 3992
Style: Imperial/Strong Porter
Predicted Score for Beer with Id (67716) : 19.924784703813692
Beer Id: 31410
Beer Name: Pizza Port Hop Suey Double IPA
ABV: 9
BrewerID: 1538

Style: Imperial/Double IPA

Predicted Score for Beer with Id (31410) : 19.570419198464386

Based on these alone we can start noticing some patterns. For example all of these beers have a very high ABV, suggesting that the user prefers drinking very strong beers. Some of the beers have similar styles giving the idea that the user probably prefers to drink Abt/Quadrupel and Imperial Stouts. In other cases the same brewer might show up more often and give the idea that the user might prefer particular brewer. These are the patterns that we'll be looking for when determining the likeability score.

Step 3: Determining the Likeability Score

Now that we have the top 20 beers for a user can calculate the likeability scores for each user. The likeability score is simply a numeric number that gives a rough idea of how likely it is that a person will like a beer based on the input properties. The input properties for calculating the score are ABV, BrewerID, and the Style of the beer. We'll go through each of the beers in the top 20 predicted beers and apply the following criteria:

- For the input ABV, if it is within a certain range from the ABV of the beer we add 1 point
- For the input BrewerID if it matches up we add 1 point
- For the input Style, if it matches up we add 1 point

We'll multiply the total above by the predicted score out of 20. If the user was predicted to give a 20/20 for a beer they'll retain the full score and likewise if they were predicted to give the beer a 10/20 they would retain half the score. This gives a higher score to the beers that the user really liked and a lower score to the beers that the user really hated.

For example, let's say we're introducing a beer that has the following properties.
- ABV 12%
- BrewerID of 3490

- Style that is Abt/Quadrupel

By using these parameters on "mribms" top 20 beers we might expect to get a high likeability score since the ABV, BrewerID and Style are present in the top 5 beers. We will run this against all the users in our dataset and then compile a list of the top users who could be potential customers. The following is the output from the running the likeability score for all users. Our old friend 'mribm' shows up once again and is the top potential customer for our new beer.

{'user': 'mribm', 'likeability_score': 10.213219018348461},
 {'user': 'Leighton', 'likeability_score': 9.718911921009681},
 {'user': 'hophead75', 'likeability_score': 9.693727438049795},
 {'user': 'lampeno420', 'likeability_score': 9.433628918755632},
 {'user': 'lusikka', 'likeability_score': 9.404424775458697},
 {'user': 'LilKem', 'likeability_score': 9.25259159169947},
 {'user': 'Hammy78', 'likeability_score': 9.202447052065768},
 {'user': 'OldGrowth', 'likeability_score': 9.198000006803122},
 {'user': 'wxman', 'likeability_score': 9.115318448508685},
 {'user': 'Defreni', 'likeability_score': 9.113605455405073},

So now that we have a set of likeability scores how do we know whether or not they are "accurate" or "reliable"?

Evaluation

For the likeability score the model was evaluated by testing beer properties for specific users, much like above with "mribm". I tested beer properties that were present in the top 5 beers with the expectation that "mribm" would show up at the top of the list for potential customers. By inspecting the top 5 ratings for "mribm" we can get the idea/make an assumption that they would like a beer with an ABV of 12, BrewerID of 3490 where the beer style is Abt/Quadrupel, and the final result was that the likeability score for "mribm" was relatively higher than most of the other potential customers.

Likewise I did the opposite and tested beer properties that were not present in the top 20 beers for "mribm" but instead the beers that he rated the lowest and expected to see them near the very bottom of the list of potential customers. More often than not the likeability score for those beers would be 0 since none of the criteria would match or fall within the range of the top 20 beers, that is just how the math ends up working.

One of the other things that I used to evaluate the model was to check the scores for newer beers that users rated that were not in the dataset. The dataset contained ratings only from April 2000 to November 2011 but "mribm" continued to provide ratings on the website. To evaluate the model I checked a beer that was rated not too long after the dataset was created. On Dec 25, 2011 "mribm" provided a high overall rating for a Rogue Double Chocolate Stout and gave the impression that he really liked the beer [4]. For this experiment we're pretending that we are the Rogue Ales brewer and we're introducing this new beer and want to find potential customers. We can use the model to predict the likeability scores for users on this beer and the assumption is that "mribm" would appear near the top of the list of potential customers.

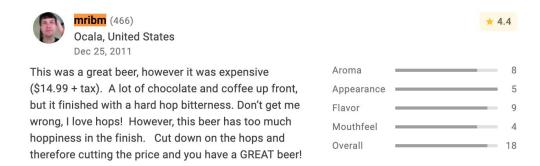The following are some screenshots for the beer info and the user rating that "mribm" gave.

Rogue Double Chocolate Stout
Newport, Oregon, United States

Style: Stout - Imperial   9.0%   72   270
Brewed by: Rogue Ales   ABV   IBU   Est. Cal

The input beer properties for the likeability score would be the following:

- ABV 9%
- BrewerID is 96
- Style that is Imperial Stout

The following are the top potential customers based on the likeability scores for the above input beer properties, and our old friend "mribm" appeared among them giving an 8.35 out of a max of 10.46.

{'user': 'Tejas', 'likeability_score': 10.464478132210425},

{'user': 'otakuden', 'likeability_score': 10.22642395488138},

{'user': 'BigBilly', 'likeability_score': 10.219930174512303},

{'user': 'WisconsinBeer', 'likeability_score': 10.158467249722007},

{'user': 'Elkas', 'likeability_score': 10.147180442662322},

{'user': 'ryan', 'likeability_score': 10.139934375460195},

{'user': 'illidurit', 'likeability_score': 9.99791288996612},

{'user': 'AgentSteve', 'likeability_score': 9.673481284011034},

... <truncated>

{'user': 'winkle', 'likeability_score': 8.364691490181027},

{'user': 'fromred2green', 'likeability_score': 8.364292140301396},

{'user': 'mribm', 'likeability_score': 8.356270105921467}

Conclusion

By calculating the likeability scores for each person we are able to get an idea of who the potential customers are for a certain new beer based on the beer properties. However, it is important to point out the caveats for such a model. For one the input beer properties might not be enough to accurately describe a beer. There are more features to a beer that could potentially improve the model. Things like the ingredients and the bitterness could help us describe the beer more and provide better results. One of the plans I had originally was scraping the IBU (International Bitterness Units) for the beers in the dataset but it never came into fruition. The likeability score too could also be computed in a more elegant fashion as opposed to adding an arbitrary amount of points if the beer properties lined up. There are definitely some refinements that could be done to provide better results.

With that being said, the likeability score model is able to give a ballpark figure of whether or not the user might like a beer, and can be used to recommend potential customers for beers and gets us closer to answering the questions posed in the introduction: "How do I find more potential customers?" and "What is the next beer that I should try?".

References

[1] https://www.nbwa.org/resources/industry-fast-facts

[2] https://snap.stanford.edu/data/web-RateBeer.html

[3] https://webpages.uncc.edu/~hli38/data/index.html

[4] https://www.ratebeer.com/beer/rogue-double-chocolate-stout/127436/278/