Beer Recommendation System

--Final report

Outline

1. Introduction

It's common to see someone staring at a wall of beers at local supermarket, contemplating for over 10 minutes before grabbing one. They are not alone. As a beer enthusiast, I'm always looking for something new to try but I also dread being disappointed so I often spend too much time looking up a particular beer over several websites to find some kind of reassurance that I'm making a good choice. So it's necessary to build a recommendation system to leverage users historical ratings to create predictions for un-tasted beers and create a personalized recommendation for which beer to sample next.

My target audience will be the beer enthusiast, believe that drinker will find the results interesting. There are more than 2,748 different breweries operating in the US as of June, 2018 and more than 5,000 unique brands in the U.S. which may cause newfound confusion for the average consumer when shopping for beer. Given the huge rise in the brewery scene in the United States, it can be overwhelming trying to decide which new brewery or beer to try. My recommendation system will help them decide which beer to try next given their unique tastes and historical ratings.

1.1 Data source information

This dataset contains around 1.5 million reviews of Beer from Beer Advocates. The dataset can be downloaded at https://www.kaggle.com/rdoume/beerreviews. The data (beer_reviews.csv) is in a .csv format consisting of 1,586,614 observations and 13 features: brewery_id, brewery_name, review_time, review_overall, review_aroma, review_appearance, review_profilename, beer_style, review_palate, review_taste, beer_name. The meaning of those features can be found in Table 1.

Table 1 Features information from beer reviews dataset

Feature name	Count	Missing	Type	Note
brewery_id	1586614	0	int64	An identifier for the brewery
brewery_name	1586599	15	object	The name of brewery
review_time	1586614	0	int64	A dict specifying when the review was submitted
review_overall	1586614	0	float64	Rating of the beer overall (1.0 to 5.0)
review_aroma	1586614	0	float64	Rating of the beer's aroma (1.0 to 5.0)
review_appearance	1586614	0	float64	Rating of the beer's appearance (1.0 to 5.0)
review_profilename	1586266	348	object	The profile name of the reviewers
beer_style	1586614	0	object	The style of beer
review_palate	1586614	0	float64	Rating of the beer's palate (1.0 to 5.0)
review_taste	1586614	0	float64	Rating of the beer's taste (1.0 to 5.0)
beer_name	1586614	0	object	Name of the beer
beer_abv	1518829	67785	float64	The alcohol by volume of the beer
beer_beerid	1586614	0	int64	A unique ID indicating the beer reviewed

1.2 Data Cleaning

The data cleaning aimed to deal with missing value, duplicates and outliers among all features. For review time, data were converted to YYYY-MM-DD format. As for the 5 review indexes, any values falling outside of 1 to 5 were dropped. Moreover, review recode with count less than 5 for any specific beer were discarded too. The main cleaning steps were listed in Table 2.

Table 2 Main cleaning steps for the dataset

Feature name	Cleaning steps
brewery_name	▪ Drop missing value. ▪ Keep name started with letter or number (a-zA-Z0-9)
review_time	▪ Convert to YYYY-MM-DD format
review_profilename	▪ Fill 'nan' with 'Unknown'
beer_abv	▪ Fill missing value with the average abv of the style
beer_beerid	▪ Count the number of id. ▪ Discard the id with review count <5
5 review indexes*	▪ Discard the review small than 1 and larger than 5

Note: 5 review indexes stand for review_overall, review_aroma, review_appearance review_palate, review_taste

2. Data Exploration

2.1 Describe statistics

Describe statistics analysis were conducted for the 5 review indexes and the alcohol content (Table 3): Describe statics showed that there are no big variations existing in features probably due to fixed rating range from 1 to 5. For instance, the overall review(review_overall) varies from 1.14 to 4.9 with a mean of 3.71. However, we observed large variation for the alcohol content (beer_abv), which varies from 0.05% to 41% with a mean of 6.58%.

Table 3 Describe statistics of the 5 review indexes and alcohol content for beers

	review _overall	**review _aroma**	**review _appearance**	**review _palate**	**review _taste**	**beer_abv (%)**
Count	19209	19209	19209	19209	19209	19209
Mean	3.71	3.62	3.74	3.63	3.66	6.58
Std	0.45	0.48	0.38	0.44	0.50	2.18
Min	1.14	1.17	1.50	1.23	1.15	0.05
25%	3.50	3.40	3.57	3.42	3.42	5.00
50%	3.79	3.69	3.81	3.70	3.75	6.00
75%	4.00	3.94	4.00	3.92	4.00	7.80
Max	4.94	5.00	4.93	4.94	4.94	41.00

The data encapsulated 5230 unique breweries; 4,900 unique beers; approximately 32908 unique user(reviewers) and 22638 of them have rated more than once; approximately 1.5million user-review pairings. It contains over 16 years of data leading up to 2012 (Fig. 1a). The most review number appeared in 2011.Based on the cumulative distribution of the review indexes (Fig. 1b), it is obvious that alcohol content is not normal distributed.

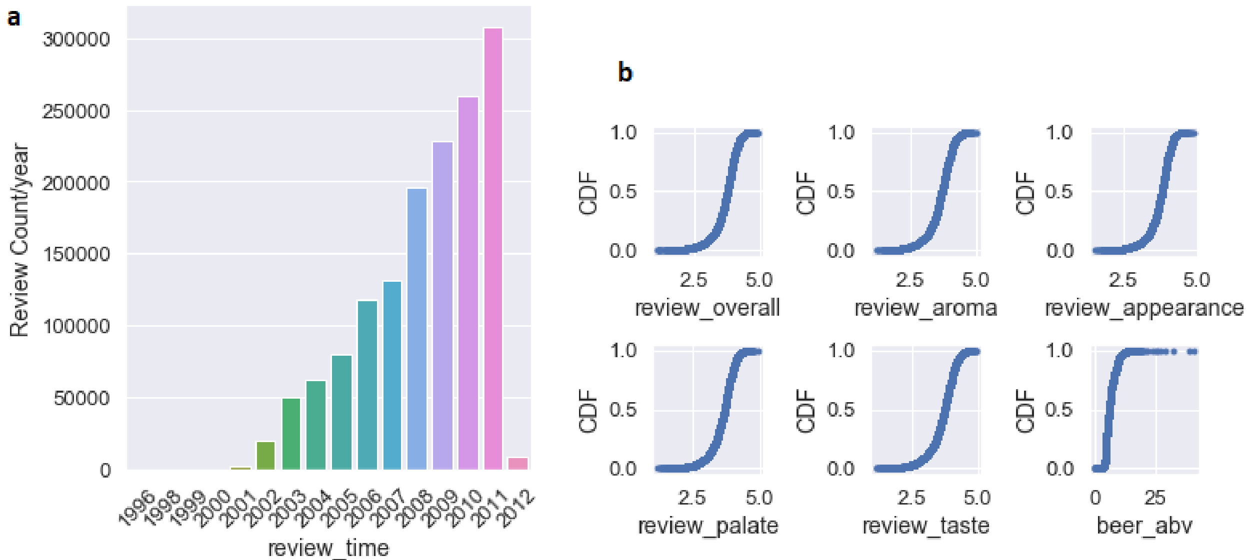

Figure 1 Review time (a) distribution and review indexes (b) cumulative distribution

Kolmogorov-Smirnov test was used to test normal distribution of the review indexes and alcohol content. Since p value of the all tests were close to 0, which suggested that their distribution sufficiently different from the normal distribution, that we can reject the hypothesis that the sample came from the normal distribution.

2.2 Ranking and sorting beer brand, style and brewery

In order to rank the different beer and brewery, mean value of the 5 review indexes were calculated for each beer brand and brewery. The ranking of beer brand, style and brewery were based on the mean value (Table 4,5,6).

Table 4 Top ten the most popular beer brand

Rank	**Beer Name**	**Review average**
1	Alesmith Speedway Stout - Vanilla And Coconut	4.887500
2	El Gordo	4.828571
3	M Belgian-Style Barleywine	4.735714
4	Capricho Oscuro - Batch 1	4.716667
5	Coffee Infused Imperial Stout Trooper	4.700000
6	Cantillon La Dernière Cuvée Du 89	4.683333
7	Barrel Aged Stout	4.680000
8	Armand'4 Oude Geuze Lente (Spring)	4.673846
9	Kopi Con Leche Stout	4.671429
10	Hitchhiker	4.666667

The number 1 of most popular beer is Alesmith Speedway Stout - Vanilla And Coconut. This brand of beer has chocolate and roasted malts dominate the flavor, supported by notes of dark fruit, toffee, and caramel. A healthy dose of locally-roasted coffee added to each batch brings out the beer’s dark chocolate flavors and enhances its drinkability.

The most popular beer style is Quadrupel (Quad). Inspired by the Trappist brewers of Belgium, the Quadrupel is a Belgian-style ale of great strength with even bolder flavor compared to its sister styles Dubbel and Tripel. Typically a dark creation that plays within the deep red and ruby brown end of the spectrum with garnet hues it is a full bodied beer with a rich, malty palate and spicy phenols than are usually kept to a moderate level. Sweet on the palate with a low bitterness yet well perceived alcohol, Quads are well suited to cellaring.

Table 5 Top ten the most popular beer styles

Rank	Beer style	Review AVG*
1	Quadrupel (Quad)	4.135505
2	American Double / Imperial Stout	4.131664
3	Russian Imperial Stout	4.113477
4	American Wild Ale	4.097805
5	Eisbock	4.088967
6	Gueuze	4.084845
7	American Double / Imperial IPA	4.061975
8	Lambic - Unblended	4.038159
9	Weizenbock	4.034948
10	Flanders Red Ale	4.026596

Note: Review AVG stand for the mean value of the 5 review indexes

The best brewery is Brouwerij Westvleteren based on the average reviewing of its beer product. This brewery founded in 1838 at the Trappist Abbey of Saint Sixtus in Vleteren, Belgium. The brewery's three beers have acquired an international reputation for taste and quality; Westvleteren is considered by some to be the best beer in the world. The beers are not brewed to normal commercial demands but are sold in small quantities weekly from the doors of the monastery itself to individual buyers on an advance-order basis.

Table 6 Top ten breweries based the rating of their beers

Rank	Brewery Name	Brewery ID	Review AVG
1	Brouwerij Westvleteren	313	4.51
2	Närke Kulturbryggeri AB	10902	4.48
3	The Alchemist	27039	4.47
4	Micro Cervejaria Falke Bier	10287	4.47
5	Old Chimneys Brewery	6560	4.43
6	Benny Brewing Company	24757	4.42
7	Kane Brewing Company	26676	4.41
8	Denison's Brewing Company & Restaurant	662	4.40
9	Breakwater Brewing	17988	4.38
10	Brauerei Zehendner GmbH	5983	4.38

2.3 Relationship among the features

Relationships among the features were analyzed using scatter plots and correlation analysis (Figure 2). The Person correlation coefficient ranged from 0.84 to 0.93. among the 5 review indexes. The results showed that, except for alcohol content, all 5 review indexes showed close relationships with each other.

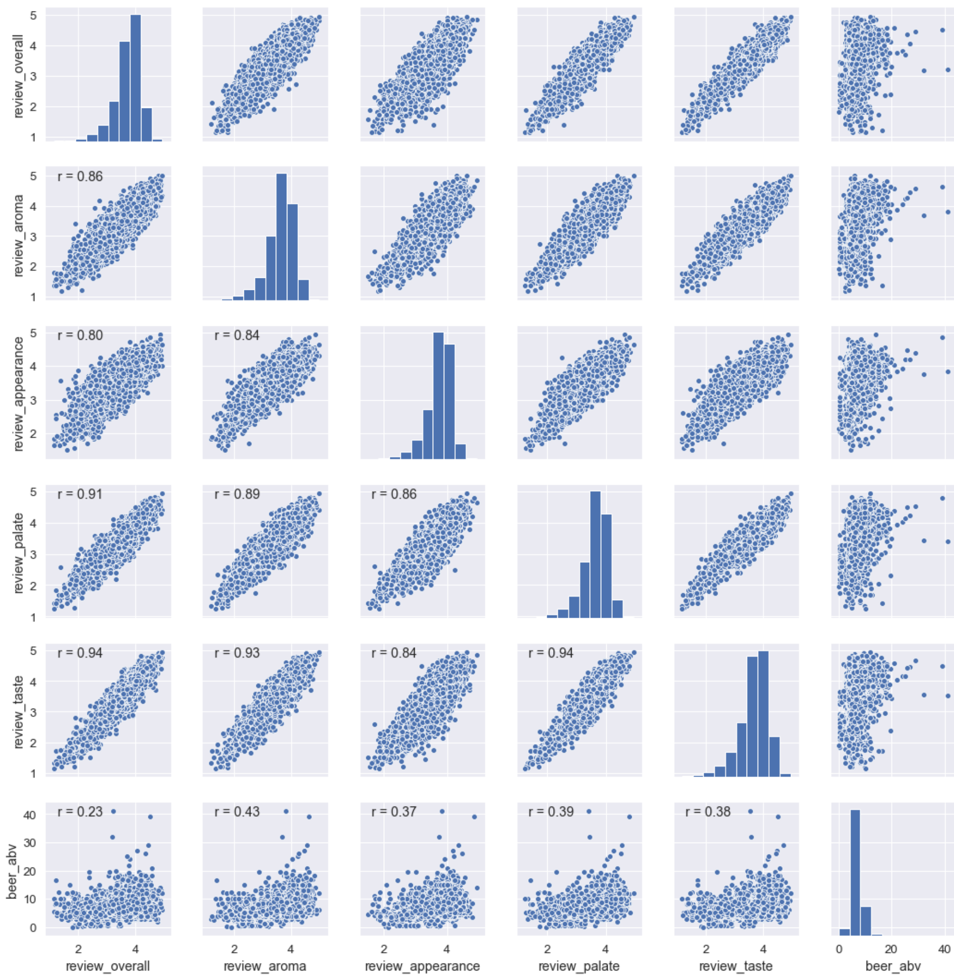

Figure 2 Histogram and scatter plot showing the relationships among the features.

However, the correlation coefficient between alcohol content and the 5 review indexes is relatively low with range of 0.23 to 0.43. The results suggest that alcohol content dose not play important role for beer rating. The histogram of the 5 review indexes (Fig. 2) showed that rating score slightly skewed to right side indicated that there are more people like the beer they rated than who did not.

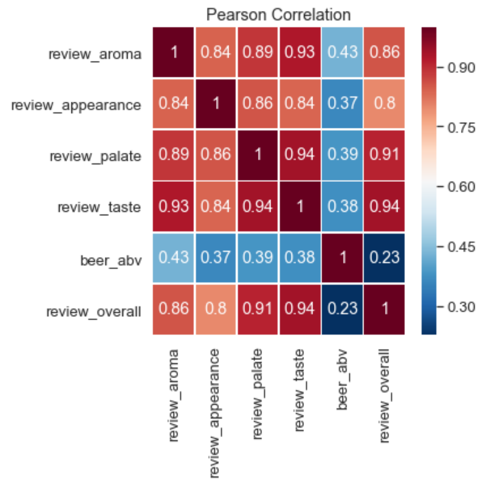

Figure 3 Correlation coefficient among 5 review indexes and alcohol content

The p-value for testing non-correlation were calculated for all pair-wise features (Table 7). The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. The results showed that the p-values are close to 0 or equal to 0, which indicate that all pair-wise features were significantly associated with each other.

Table 7 The p-value for testing non-correlation were calculated for all pair-wise features

	review_overall	**review_aroma**	**review_appearance**	**review_palate**	**review_taste**
review_aroma	0.0				
review_appearance	0.0	0.0			
review_palate	0.0	0.0	0.0		
review_taste	0.0	0.0	0.0	0.0	
beer_abv	9.28e-274	0.0	0.0	0.0	0.0

2.4 Relationship between the beer alcohol content and rating score

Based on the above analysis, I found that he alcohol by volume (ABV) is positively associated with the rating score. After classifying the beer into low (ABV<2%), medium (2%<ABV<8%) and high ABV (>8%) groups, it is obvious that beer with high ABV got relatively high rating score than those with low ABV. Analysis of variance (ANOVA) showed that there are significant differences in rating score among different ABV groups (Table 8).

Table 8 ANOVA analysis of rating difference among different ABV classes

	Sum_sq	df	F	PR(>F)
abv_class	128.006512	2	319.418558	3.43E-137
Residual	3848.387965	19206	NaN	NaN

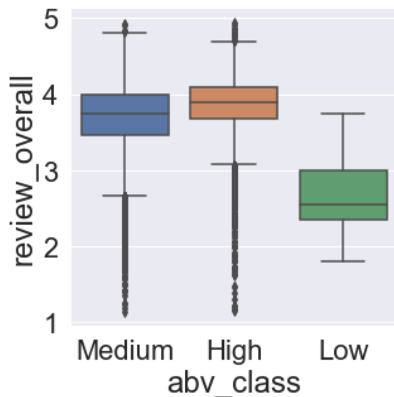

Figure 4 Box plot showing the difference in rating among three ABV classes

This tells us something important: while there seems to be a preference toward beers with higher alcohol. The exact causes may not be clear, alcohol content certainly has an influence on how we enjoy and rate the beers we drink, and each style comes with its own optimal balance or set of expectations.

2.5 Relationship between the beer style and rating

Cluster analysis was conducted to investigate whether specific beer style is more prefer than other. Based on rating scores, all 104 beer styles can be grouped into 11 clusters (Figure 4). I found that different clusters showed significant difference in rating (p=7.110717e-26, Figure 5). Cluster #1, #2 and #3 showed the highest rating score. Interestingly, more than half of the beer styles are Ale related. Except for Cluster #9, #10 and #11, there are no clear boundary among different clusters based on PCA (Figure 6).

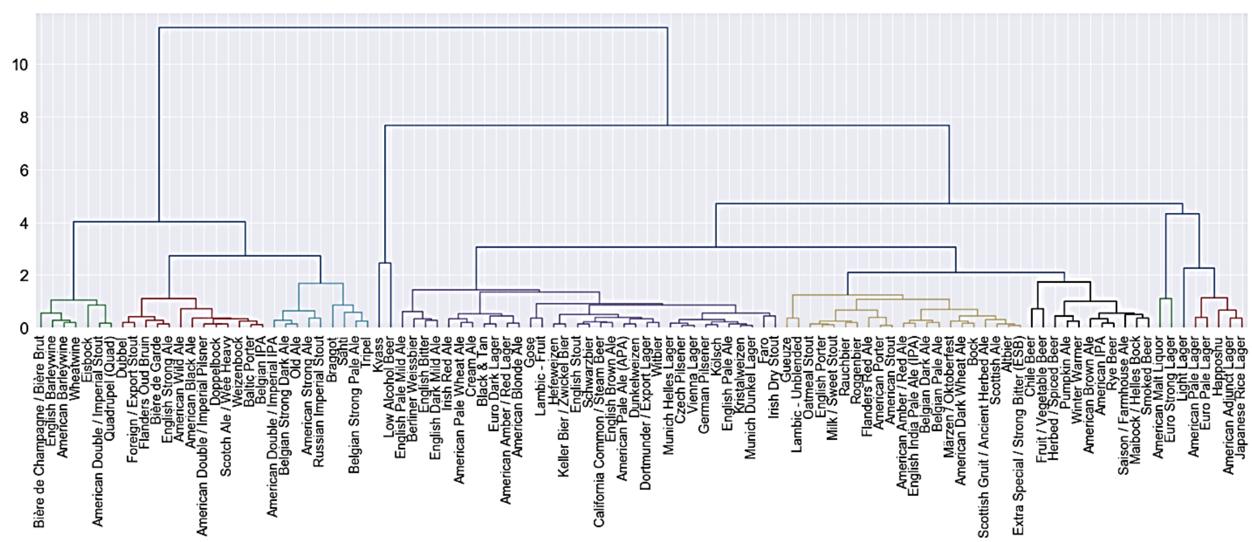

Figure 5 Cluster analysis for the beer style based on the rating data

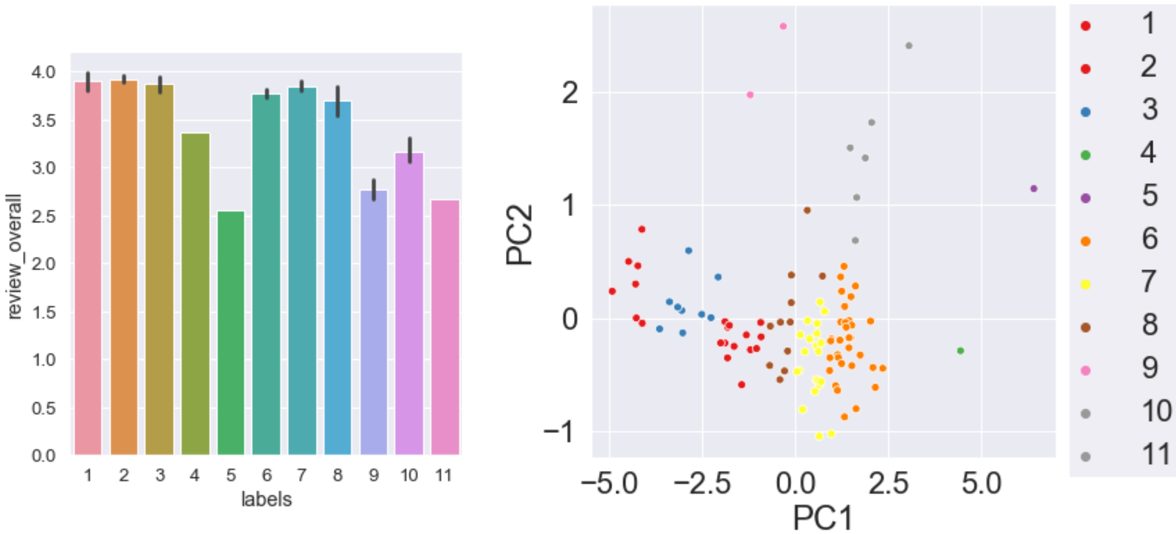

Figure 6 Bar plot showing the difference in rating score among 11 clusters and beer style distribution

3.Building and testing recommender systems with Surprise

3.1 Choose the best algorithms

In order to determine which algorithm perform best in building the recommender systems, nine algorithms were used to build the prediction models (Table9). Using RMSE (Root Mean Squared Error) as standard, SVDpp algorithms has best performance with the lowest RMSE (0.5975). However, this algorithm is also the most time-consuming one.

Table 9 Algorithms comparison for the beer dataset

Algorithm	test_rmse	fit_time	test_time
SVDpp	0.598	7313.033	337.3621
BaselineOnly	0.601	3.928861	6.072451
KNNBaseline	0.602	82.24511	232.1964
SVD	0.604	77.26623	6.828465
KNNBasic	0.620	76.56187	209.4991
NMF	0.625	79.14518	5.830669
KNNWithMeans	0.625	77.10334	219.8093
CoClustering	0.666	40.16992	6.11246
NormalPredictor	0.989	2.286215	6.465656

3.2 Tuning and evaluating the SVDpp model

It is a good practice to fit a model on the whole dataset after checking its performance and tuning parameters which affect the fit. The method SVDpp, will depend on a number of main tuning constants: 1) n_epochs, the number of iteration of the SGD procedure, 2) r_all ,the learning rate for all parameters, which affects the performance of the optimisation step, 3) the regularization term for all parameters, which affect the overfitting of the model. To tune the parameters of SVDpp and evaluate the performance of the method, GridSearchCV was used to test 8 combinations of different hyperparameters. The results showed that best hyperparameters combination is 10 for n_epochs 10, 0.01 for lr_all, and 0.1 for reg_all. The corresponding accuracy measured by RMSE score was 0.603.

3.3 Making recommendations with SVDpp model

After hyperparameters tuning, I used the train_test_split() to sample a trainset and a test-set with given sizes(3-fold), and use the accuracy metric of rmse. Then I used the fit() method which will train the algorithm on the trainset, and the test() method which will return the predictions made from the test-set. The following table showed the top 10 best predictions for the test dataset

Table 10 Top 10 the best predictions for the test dataset

uid	iid	rui	est	Details(was_impossible)	Iu	Ui	err
oteyj	44910	5	5	False	67	28	0
oteyj	51116	5	5	False	67	131	0
FARGO619	21690	5	5	False	36	472	0
FARGO619	7971	5	5	False	36	1907	0
billzav	7971	5	5	False	21	1907	0
paxtonthegreat	7971	5	5	False	10	1907	0
lemasney	7971	5	5	False	78	1907	0
oteyj	731	5	5	False	67	1472	0
kdawg105	6549	5	5	False	53	991	0
hippityhophead24	7971	5	5	False	23	1907	0

Note: uid — the user ID, for whom we carry out predictions, iid — item ID (here we treat movies as items), est — estimated rating for an item, as we expect the user to give, Ui : the set of all users that have rated item i. Iu : the set of all items rated by user u. err, the difference between predicted and true value.

Table 11 The best recommendation for user NaLoGra

Brewery name	Beer name	Beer id	Predicted rating
Southampton Publick House	Southampton Berliner Weisse	8626	4.87
Kern River Brewing Company	Citra DIPA	56082	4.77
Brouwerij Drie Fonteinen	Geuze Cuvée J&J Blauw (Blue)	23413	4.8
Brouwerij Drie Fonteinen	Armand'4 Oude Geuze Zomer (Summer)	70356	4.8
Brouwerij Drie Fonteinen	Armand'4 Oude Geuze Lente (Spring)	68548	4.89
Brouwerij Drie Fonteinen	Geuze Cuvée J&J	23414	4.76
The Lost Abbey	Veritas 004	45957	4.81
The Lost Abbey	Veritas 005	54147	4.79
FiftyFifty Brewing Co.	Imperial Eclipse Stout - Pappy Van Winkle	47668	4.77
Brouwerij Westvleteren	Trappist Westvleteren 12	1545	4.76

Moreover, I wrote a function (ck_score) to predict the score of each of the beer that a specific user didn't rate and find the best one. To do this, I created another dataset with the iids I want to predict in the sparse format as before of: uid, iid, rating. I just arbitrarily set all the ratings of this test set to 5, as they are not needed. In order to find the best, I converted the predictions object into an array of the predicted ratings. I then used this to find the iid (beer_id) with the best predicted rating. Taking user NaLoGra as an example, I apply the ck_score function on the user, the best predicted rating is 4.89 with beer_id 68548. So this beer is the best recommendation for NaLoGra(Table 11).

4. Discussion and Summary

I built a beer recommendation system with SVDpp algorithm, which is an extension of Probabilistic Matrix Factorization (PMF) algorithm. The PMF algorithm is good at dealing with scalability and sparsity, especially with users who have very few ratings. The algorithm was found to work best for the beer dataset in term of its RMSE as low as 0.59. However, it is computationally expensive. It took more than 24 h to perform hyperparameters tuning with only 8 combinations of different parameters. Note that it's necessary to try more diverse hyperparameters in the future.

There are other techniques for building recommender systems. Deep learning is one of the ways of doing so especially when you have massive datasets. Some of the algorithms that are being used to build advanced recommender systems include Autoencoders and Restricted Boltzmann machines. Haghighi et al. (2019) designed an explainable recommendation system using an Autoencoder model whose predictions can be explained using the neighborhood-based explanation style. Their work can be considered to be the first step towards an explainable deep learning architecture based on Autoencoders (https://arxiv.org/abs/2001.04344).

The age of the data is something that in future analysis can and should be considered. Since majority of the data were collected before 2012, a massive uptick in craft brewing and the corresponding new users have not been included in the current data. Other details that would be powerful is regionality. Knowing where beers are available geographically (i.e. internationally, nationally, regionally, brewery specific) would allow the recommendation engine to take that into account. It's obvious that users would not be recommended beers that they don't have access to. I believe that building a dynamic recommendation system will be a trend in the future.