

Sam Lieberman  
Recipe Rating Prediction Writeup  
12/3/24

## **What's in the Zip**

- 1) assignment2.html -> a jupyter notebook with my code for the project. I did my work in python
- 2) 2) writeup.txt -> a text file (that you're reading now) with my answers/explanations
- 3) Screenshots of the results of my model.
- 
- 
- The dataset is scraped from user-recipe interactions on Food.com:
- <https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions>
- 
- 

## **Recipe Model Feature Selection**

I built a recipe content rating predictor that used relevant features (i.e. ingredients, nutritional information, etc.) and to predict a recipe's mean rating and percentage of 5 star reviews. The envisioned use-case for this model is to be able to test a new recipe and predict its rating. (\*Note-> the model focuses solely on the recipe itself and not on temporal or interaction data).

In my model, the most prominent feature was a list of the 1000 most popular ingredients; implemented similarly to doing sentiment analysis on a bag of words in a text-recommendation system. However, instead of a bag of words, I replaced it with a one-hot encoded *bag of ingredients*. Fortunately, the Food.com dataset tokenized recipes with their relevant ingredients ids, accounting for variations in the different ingredients (e.g. 'iceberg' and 'romaine' lettuce were both tokenized as lettuce). I also parsed each recipe's nutritional data—calories, total fat, sugar, sodium, protein, saturated fat and carbohydrates—took the log of that number, and added them to the feature vector. Finally, I added the number of steps, the number of ingredients, and the (log of) number of minutes it took to prepare the recipe.

## **Recipe Model Set-Up**

I used ridge regression because the data was structured so that it was better to shrink their coefficients rather than eliminate them. Also, to make sure the model was not overfitting, I used a cross validation grid and tuned the regularization strength hyperparameter.

A major consideration for the model was setting a threshold for a recipe's minimum number of reviews. Intuitively, a recipe with more reviews will be closer to its 'actual' rating than a recipe with a single review. Therefore, I tried tinkering with different thresholds for a recipe's minimum number of reviews.

## Results

My model beat the baseline of predicting the mean by around 2-5% on the *whole test set*. This makes sense because most of the reviews are clustered around 4-5. However, the metric we ought to care about is the performance for the *top predictions*, since we care the most about the accuracy of our highest predicted items. Depending on the minimum number of recipe reviews, as well as the top number of predictions threshold (refer to the tables), the improvement in performance relative to the baseline is quite substantial. When the target variable is mean rating, the improvement in MSE ranges from 20-60%. When the target variable is the percentage of 5-star reviews, performance decreases slightly, but is still improved by 15-40%.

Recipe Content Rating Predictor MSE Based Performance for Predicting Mean Rating

Min reviews per recipe	Baseline MSE	Test MSE (whole dataset)	MSE Top 100 Predictions	MSE Top 500 Predictions	MSE Top 1000 Predictions
1	0.7175	0.7075	0.8450	0.6490	0.5541
3	0.3853	0.3707	0.1273	0.2457	0.2498
5	0.2683	0.2571	0.1742	0.1878	0.1717
8	0.1937	0.1851	0.0868	0.1079	0.1229
10	0.1915	0.1820	0.1173	0.1282	0.1609

Recipe  
Content  
Rating

Predictor MSE Based Performance for Predicting 5 Star Review%

Min reviews per recipe	Baseline MSE	Test MSE (whole dataset)	MSE Top 100 Predictions	MSE Top 500 Predictions	MSE Top 1000 Predictions
1	1384.8981	1358.9137	1311.2668	1193.3753	1258.3813
3	598.0704	567.6493	328.0806	488.5283	494.9241
5	402.1744	379.3309	280.0331	337.1758	322.2920
8	262.8341	244.6922	150.1732	189.5845	205.1515
10	249.8778	235.1857	241.6843	216.7618	219.1501

The most impactful feature on model performance was the bag of ingredients. In particular, the size of the bag of ingredients played an important role in prediction accuracy.

Smaller choices like 300 or 500 and larger choices like 2000 were not as accurate as the 1000 ingredient bag. 1000 ingredients seemed to be a balanced sweet spot. Furthermore, the other features (nutrition, steps, minutes) improved the bag of ingredients model's performance, but by themselves, were not as accurate. The reason why the bag of ingredients is a better predictor is because it relates better to the content of a recipe. Prep and nutrition are important factors into whether or not a recipe will be good, but the ingredients are what give it its taste.

Refer to the tables below for which ingredients had the largest positive and negative impacts on predicting a recipe's rating. The 5 ingredients with the highest coefficients (which predict a higher mean rating) when we set recipe reviews to a minimum of 3 are: *vanilla ice cream*, *Italian bread*, *poppy seed*, *ginger ale*, and *chuck roast*. On the flip side, the 5 ingredients with the lowest coefficients (which predict a lower mean rating) are: *artificial sweetener*, *firm tofu*, *graham crackers*, *cod fish filet*, and *bisquick*.

#### Top Ingredient Coefficients in Recipe Content Model

	Ingredient	Coefficient	raw_ingr	raw_words	processed	len_proc	replaced	count
0	7474	0.103120	low-fat vanilla ice cream	4	vanilla ice cream	17	vanilla ice cream	902
1	3923	0.097507	italian bread	2	italian bread	13	italian bread	361
2	5556	0.096044	poppy seed	2	poppy seed	10	poppy seed	695
3	3249	0.095903	sugar-free ginger ale	3	ginger ale	10	ginger ale	309
4	1485	0.093159	chuck roast	2	chuck roast	11	chuck roast	294
5	2901	0.092570	fresh strawberries	2	fresh strawberry	16	fresh strawberry	900
6	5627	0.089661	pork tenderloin	2	pork tenderloin	15	pork tenderloin	897
7	6846	0.088074	strawberries	1	strawberry	10	strawberry	1835
8	7117	0.086768	tea bag	2	tea bag	7	tea bag	231
9	4545	0.084906	mango chutney	2	mango chutney	13	mango chutney	195

#### Bottom Ingredient Coefficients in Recipe Content Model

	Ingredient	Coefficient	raw_ingr	raw_words	processed	len_proc	replaced	count
990	7559	-0.104652	vegetable shortening	2	vegetable shortening	20	vegetable shortening	386
991	2696	-0.107575	food coloring	2	food coloring	13	food coloring	179
992	5947	-0.108611	red food coloring	3	red food coloring	17	red food coloring	387
993	531	-0.111198	bicarbonate of soda	3	bicarbonate of soda	19	bicarbonate of soda	233
994	7396	-0.115144	unsweetened applesauce	2	unsweetened applesauce	22	unsweetened applesauce	676
995	537	-0.120868	bisquick	1	bisquick	8	bisquick	393
996	1622	-0.134458	cod fish fillet	3	cod fish fillet	15	cod fish fillet	213
997	3323	-0.138907	low-fat graham cracker	3	graham cracker	14	graham cracker	353
998	2643	-0.163377	low-fat extra-firm tofu	3	firm tofu	9	firm tofu	636
999	208	-0.166168	artificial sweetener	2	artificial sweetener	20	artificial sweetener	2256