Capstone 1 Project Report

by Anna Kantur

date: 03/21/2020

Contents

i. Problem Statement

This project was inspired by a personal interest in food and cooking and trying to answer the question of what successful recipes have in common. Is it nutritional value, number of steps required, time required, a specific ingredient, or something else?

One way to answer this question is to analyze the recipes that are popular online. Recipe websites collect large amounts of data over time with recipes details, as well the recipe rating by site visitors.

There are multiple groups of stakeholders who would be interested in the findings in this report. The primary customers for our findings would be amateur cooks and online bloggers who are trying to create the best recipes. Recipe websites and online cooking classes are currently on the rise with the recent developments of Coronavirus pandemic. Lots of people are trying to find culinary inspiration online and try something new. The secondary customers for our findings would be grocery store owners who need to make a decision on what products to put on display for advertisement and how to organize shelves in a supermarket. Grocery owners usually don’t have much information about what ingredients/ recipes are trending, so we think this information would be valuable for them as well.

ii. Data Collection and Wrangling Summary

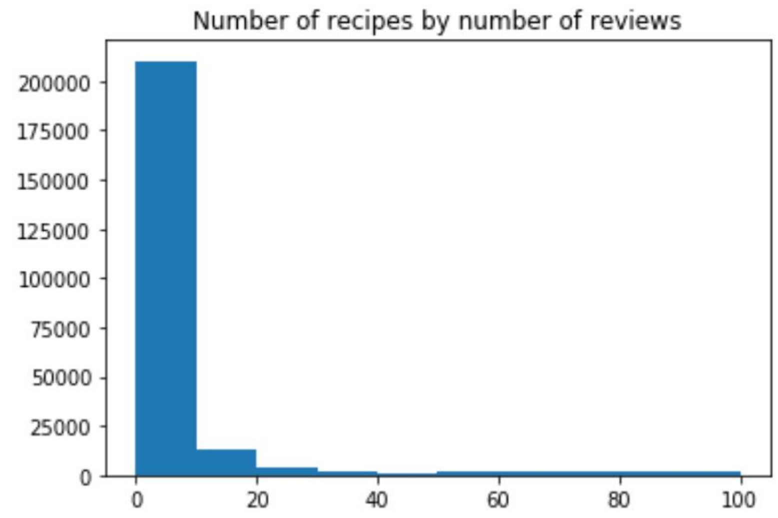
Data Collection

We collected the data from Kaggle:
https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions

This dataset consists of 180K+ recipes and 700K+ recipe reviews covering 18 years of user interactions and uploads on Food.com (formerly GeniusKitchen). We use the 'RAW_recipes.csv' and 'RAW_interactions.csv' files and upload them into Pandas data frames.

Irrelevant Data and Missing Values

For the purposes of this Capstone 1 project we decide to remove the recipes with less than 10 reviews due to them not having enough data to make meaningful predictions.The original reviews dataframe has 210,244 recipes, but we keep only 18,762 recipes with more than 10 reviews.



Upon investigation of the recipes data frame we also discover that the description column has 406 empty values. 0 star ratings are present in 58% of unique recipes and represent 6% of all the reviews in the reviews dataframe. We decide to keep these values because food.com allows to have an empty recipe descriptions and 0 star ratings (even if logged by error).

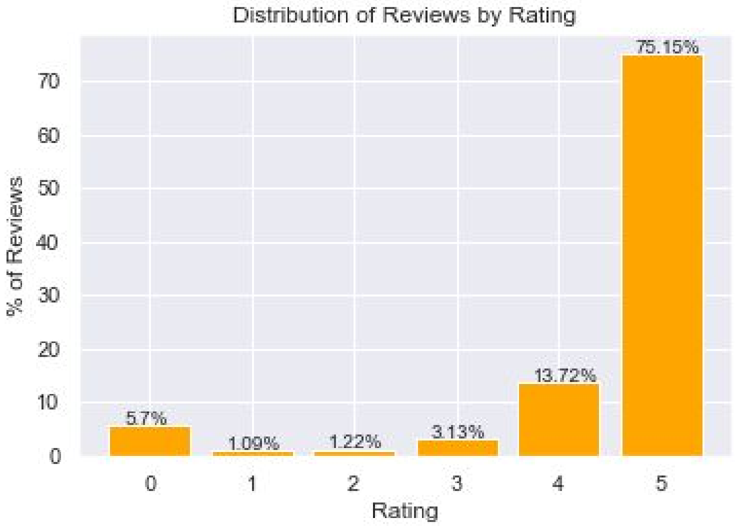
Recipe Tags and Ingredients

We created ingredients_matrix with 149 unique ingredients as columns, recipe ids as index, and 0/1 as values for an ingredient being not present/present in the recipe. Similarly, we created a tags_matrix with 495 unique tags.

iii. Exploratory Data Analysis

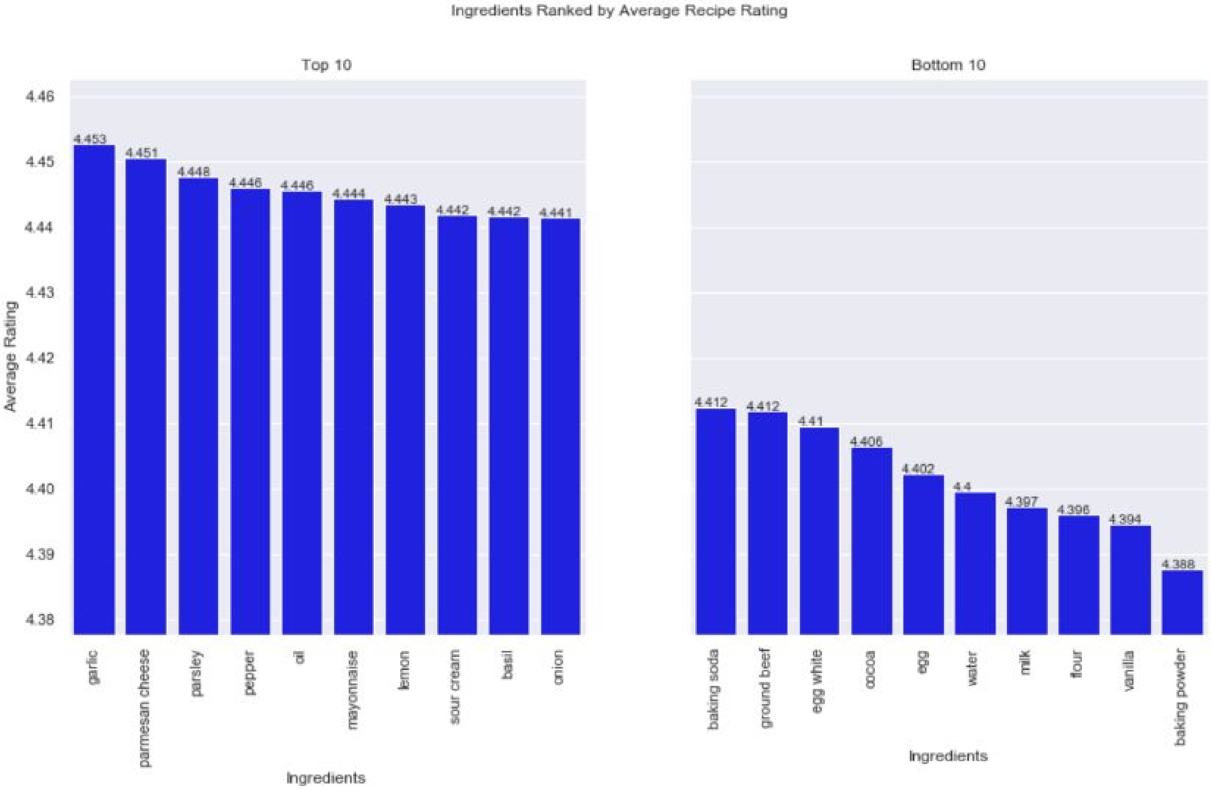
Average recipe rating across all reviews

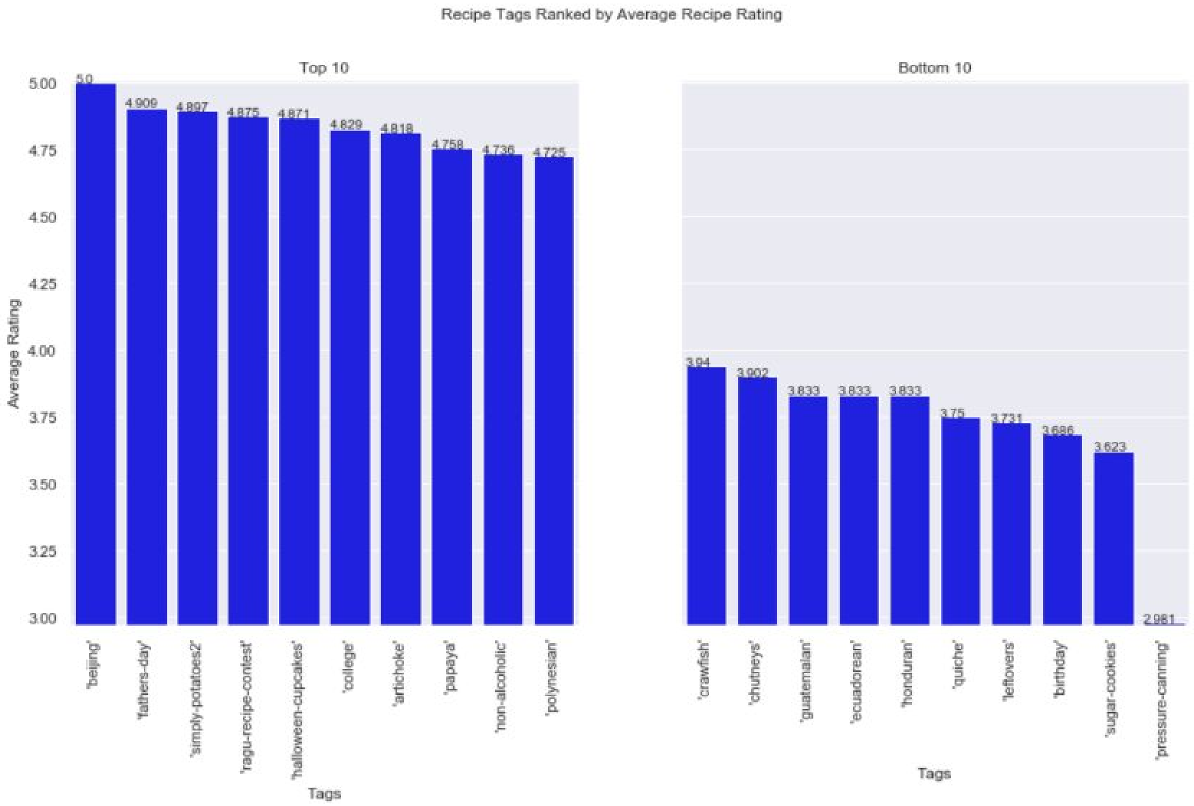
Average recipe rating across all reviews is 4.435.Also, over 75% of all reviews are5 star reviews:



Top 10 Ingredients and Tags by Average Recipe Rating

We identify the best and the worst ingredients and tags by recipe rating.

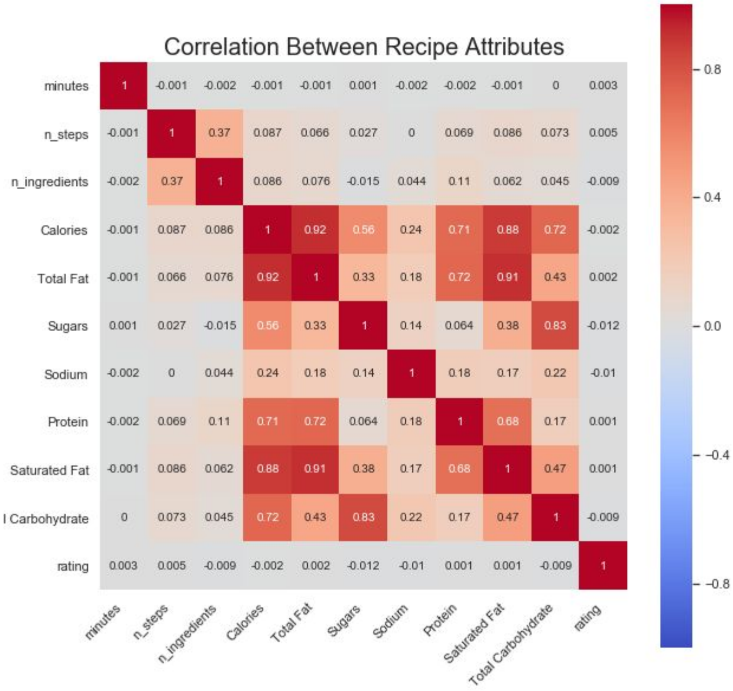




The best ingredient is garlic. It has the average recipe ranking of 4.453. The worst ingredient is baking powder. It has the average recipe ranking of 4.388. The best tag is 'beijing'. It has the average recipe ranking of 5.0. The worst tag is 'pressure-canning'. It has the average recipe ranking of 2.981. (As a reminder,t we are looking at recipes with over 10 reviews, and we know from the previous analysis that the majority of such recipes have an average rating of 3.5 or higher. Also, we are only analyzing the ingredients present in 100 or more recipes.)

Correlation Matrix

We also show the nutritional values correlation matrix.



The correlation matrix shows that there is a strong correlation between calories content and total fats (saturated fats specifically), protein and carbohydrates. However, there is no strong correlation between the recipe rating and any of the mentioned recipe attributes.

The Best and Worst Ingredients and Tags

We further investigate the data in EDA and discover the following findings:

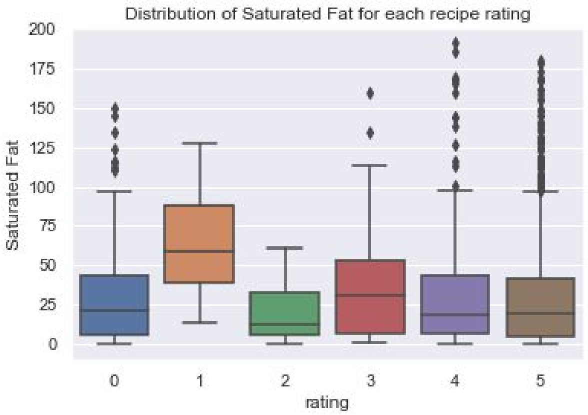
1. The best ingredient is garlic. It has the average recipe rating of 4.453
2. The worst ingredient is baking powder. It has the average recipe rating of 4.388
3. The best tag is 'beijing'. It has the average recipe rating of 5.000
4. The worst tag is 'pressure-canning'. It has the average recipe rating of 2.981
5. In terms of nutritional value, successful recipes have more total fats, sugars and carbohydrates, but less sodium and saturated fats, and slightly less protein.

T-Tests

We perform a two sample t-tests to confirm our EDA findings. The two independent samples are the recipes with a specific attribute and the recipes without it. The null hypothesis $H0$ is that the mean recipe rating for the two samples is identical. The alternative hypothesis Ha was that the rating means are different (e.g. the mean recipe rating is affected by a specific recipe attribute). If the t-test results were statistically significant (e.g. p-value > a, a = 0.05), then we rejected the $H0$ and accepted the Ha.

t-test results:

1. **Garlic** The test had p-value < 0.001 which is very small, so we can reject the $H0$ and confirm that garlic in recipes contribute to a higher recipe rating
2. **Baking Powder** The test had p-value < 0.001 which is very small, so we can reject the $H0$ and confirm that baking powder in recipes contribute to a lower recipe rating
3. **‘Beijing’ recipe tag** The test had p-value of 0.13 > a of 0.05, so we can reject the Ha. This means that 'beijing' tag positive effect on the recipe rating is statistically not significant.
4. **'pressure-canning' recipe tag** The test had p-value < 0.001 which is very small, so we can reject the Ho. This means that 'pressure-canning' tag negative effect on the recipe rating is statistically significant
5. **Saturated Fats Content** was the only other attribute that affects the recipe rating in statistically significant way (the test p-value was 0.03). All the other recipe attributes did not significantly affect the recipe rating.



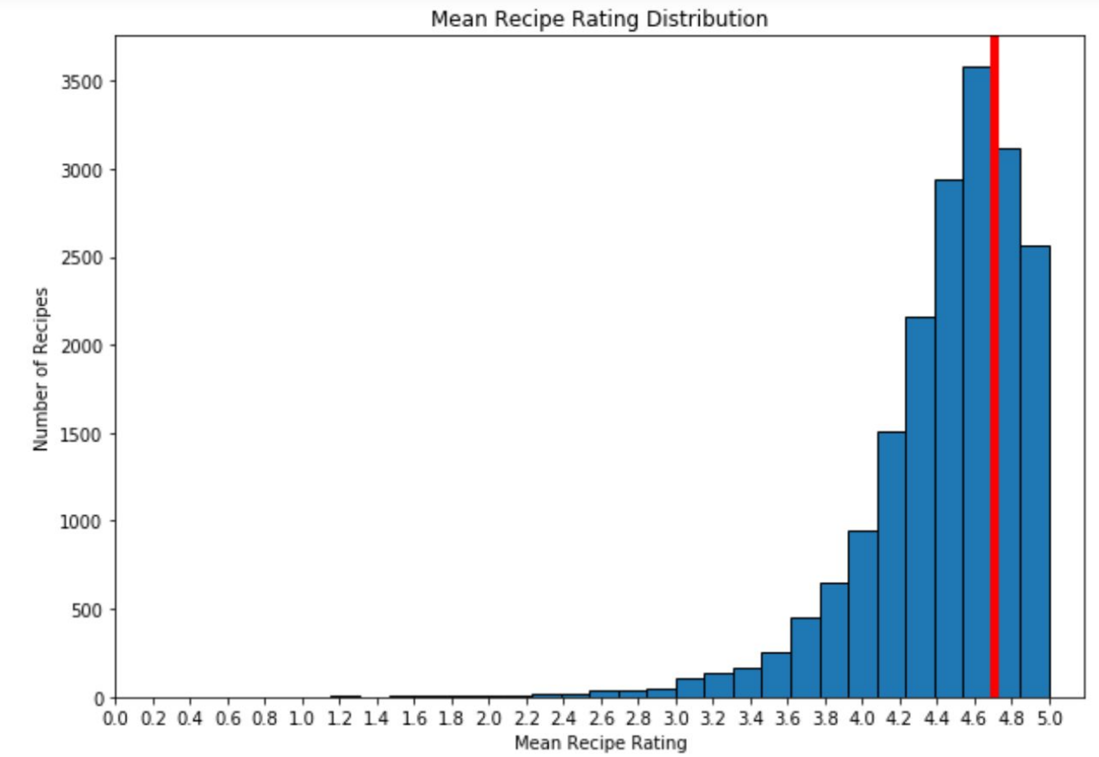
As part of EDA, we also attempt clustering the recipes into 3 clusters (see details in Appendix 1).

iv. In-depth Analysis Using Machine Learning

Predictive Modelling

We will now dive deeper and build a model to classify recipes as “good” or “bad” by mean recipe rating.

We choose 4.7 average recipe rating as a divide for “good” vs “bad” recipes:



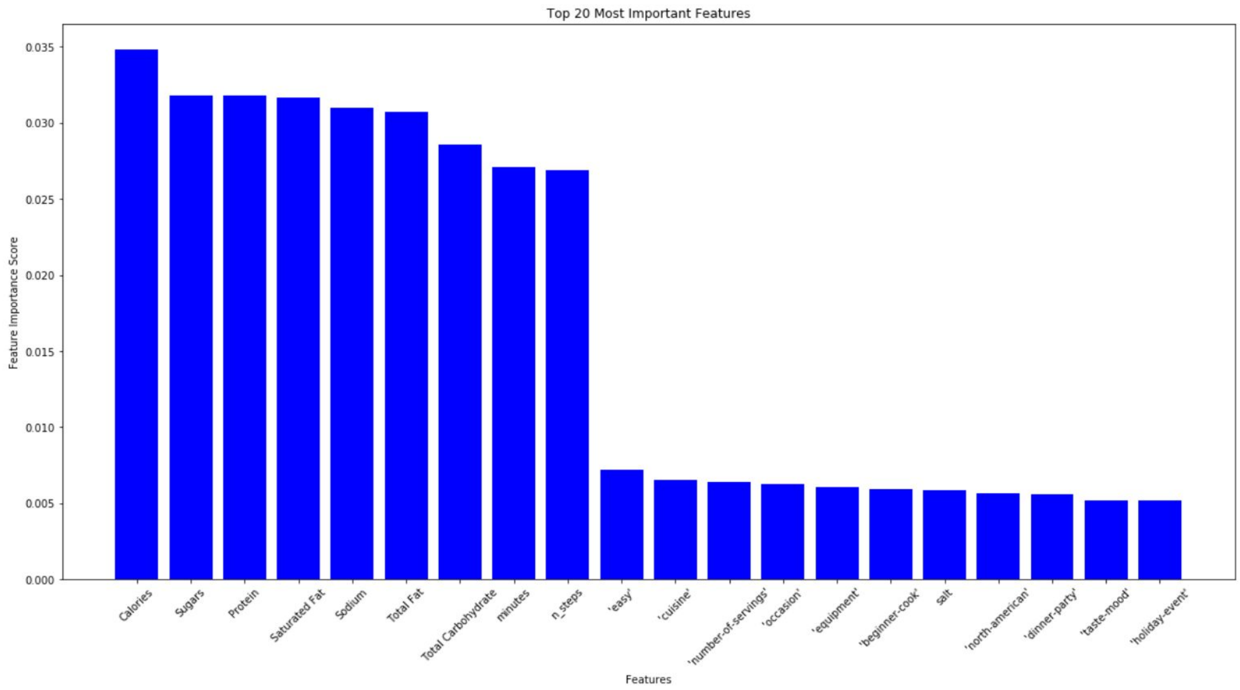
There are 5,519 “good” recipes and 13,302 “bad” recipes in our dataset.

Choosing the Best Model

We will try Random Forest, Logistic Regression and KNN classifier models. We use the same approach for each of the three models:

1. Build the model using sklearn package
2. Tune in hyperparameters
3. Estimate the ROC_AUC score

In addition, for Random Forest we find top 20 best features by using feature_importances method on the classifier. Here is the list of top 20 most important features by the importance score:



All the nutritional values and time to cook the recipe ("minutes", "n_steps", "easy", "60-minutes-or-less") are in the top 20 most important features to predict the goodness of the recipe average rating. Feature importance score drops sharply after 9 features, which include all nutritional values. We then rerun the Random Forest model with the best features to see if we can get a score improvement.

The summary of all the models that we try:

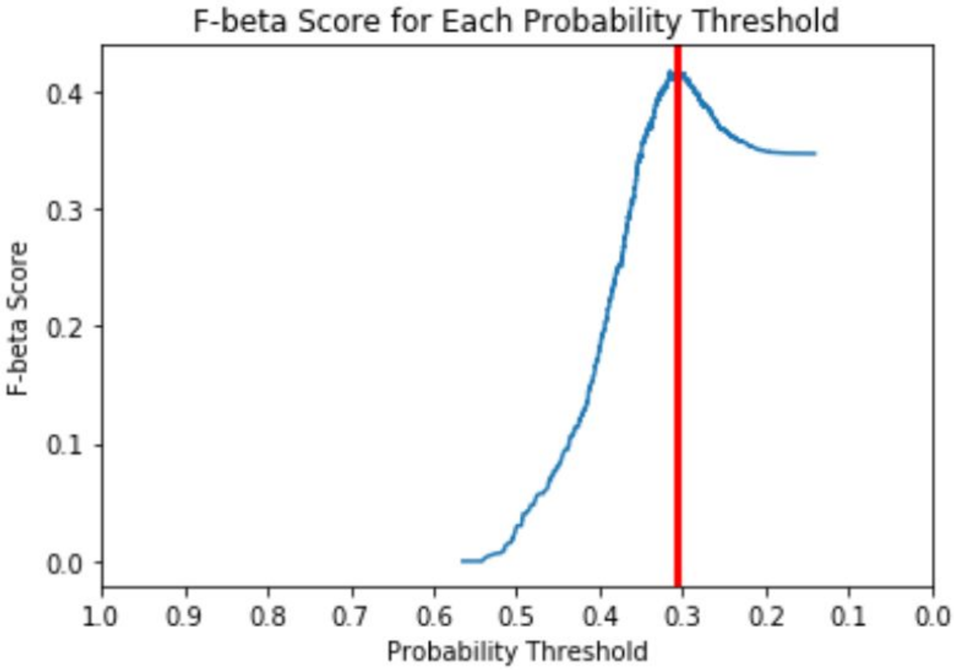
Model	**Best Score**	**Best Parameters**
Random Forest	62%	{'max_depth': 11, 'max_features': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0, 'n_estimators': 100}
Random Forest with 20 Best Features	57%	{'max_depth': 7, 'max_features': 5, 'min_samples_leaf': 3, 'min_samples_split': 5, 'min_weight_fraction_leaf': 0, 'n_estimators': 100}
Logistic Regression	54%	{'C': 8.483428982440725e-05, 'class_weight': None}
KNN	54%	{'n_neighbors': 20, 'weights': 'distance'}

The best ROC_AUC score is for the Random Forest model (62%), so this is the best model to predict the average recipe rating being good or bad.

Thresholding Probabilities by the Best F-Beta Score

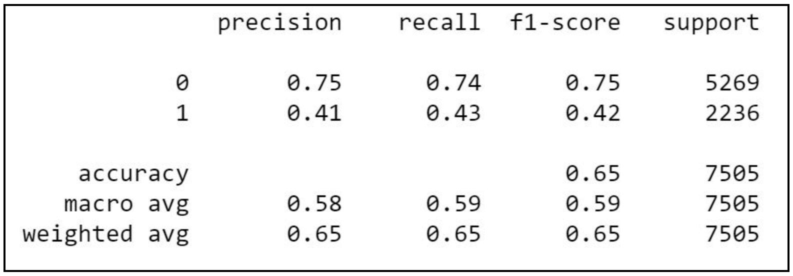
The default probability threshold is 0.5, but we decide to challenge that.

1. We loop over the probability thresholds from 0 to 1
2. We estimate the probabilities of predicting class 1 (the "good" recipes)
3. We transform the probabilities from the previous step into 1s and 0s, based on whether the estimated probability is higher or lower than the threshold probability
4. We use our ultimate accuracy metric to pick the best probability threshold. The key metrics that constitute any accuracy metric are precision and recall. Recall for us is a portion of "good" recipes correctly labelled by the model out of all "good" recipes. Precision is a portion of "good" recipes labelled by the model out of all the recipes labelled as "good". It is more important for us that our model correctly labels only "good" recipes. We choose F-beta score as our ultimate metric so that the accuracy score is weighted more towards precision (e.g. use parameter beta <1) rather than weighting precision and recall evenly with the harmonic mean (e.g. beta = 1) in F1 score. We use beta = 0.05 in F-beta score.



The chart above shows probability thresholds on X axis and F-beta score on Y axis. The vertical red line represents the optimal threshold of 0.308, at which F-beta score is 42%.

Classification report for the Random Forest model with the best probability threshold shows high precision and recall scores for both classes:



```
              precision    recall  f1-score   support

           0       0.75      0.74      0.75      5269
           1       0.41      0.43      0.42      2236

    accuracy                           0.65      7505
   macro avg       0.58      0.59      0.59      7505
weighted avg       0.65      0.65      0.65      7505
```

Recommendations

There are several recommendations that result from our analysis.

Everybody who is trying to make a good recipe should try recipes with garlic, as people love this ingredient and would likely rate the recipe higher. Avoid recipes requiring pressure-canning or high amounts of saturated fats, as these recipes would likely get a lower rating. To get an idea of how much people would like your recipe, think about nutritional values and how long it takes to make the recipe.

For grocery store owners it might be wise to arrange products on the shelves around the three recipe clusters that we found: Desserts/Baking Cluster, Mid Day Meal Cluster, and Dinner Cluster. Also we would advise to question carrying products with high saturated fat content and products that add to a high saturated fat content in food,

Ideas for Further Research

The findings in this project can be extended through further research.

One interesting finding in our project is that while the correlation matrix shows that nutritional values do not have strong correlation with the average recipe rating, they are the most important features in the Random Forest model. It would be interesting to take a closer look at the recipe nutritional values to understand a true relationship to the recipe rating.

Another research idea is to use descriptive data that was not used in our project. For example, the recipe description or the detailed descriptions of each recipe step can be used to provide more insights and predictions.

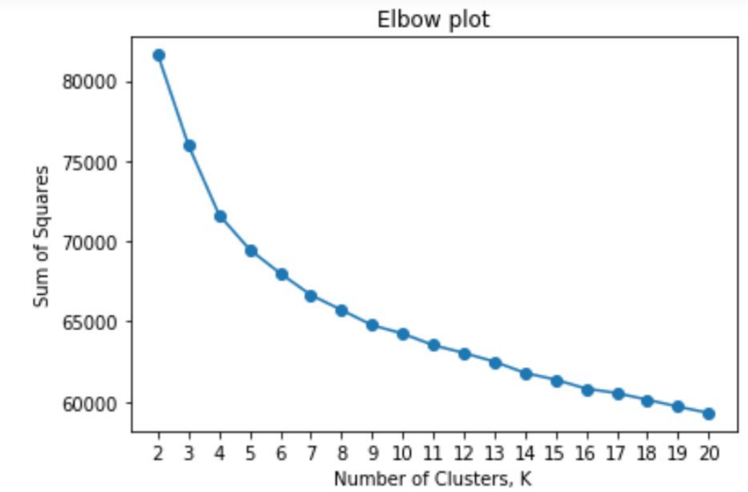
Since we know that 75% of the reviews were 5 stars, we can also recommend gathering more data, especially for recipes with lower ratings. A curious food-loving researcher can further our analysis by segmenting out specific groups of recipes.

Appendix 1 - Clustering

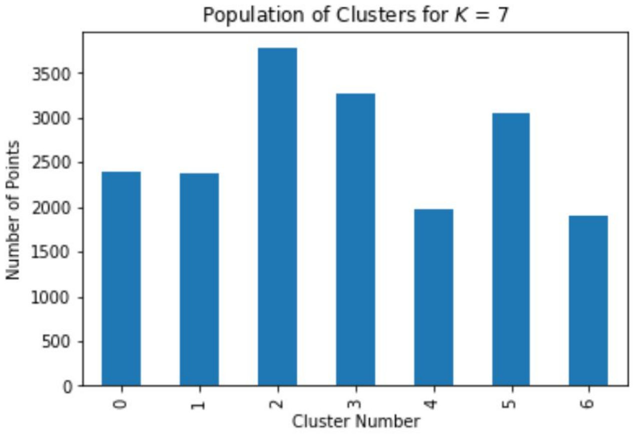
We use ingredients data to cluster recipes into groups and see if there are any patterns. We build a K-Means Clustering model and choose the number of clusters K by using the Elbow Method and Silhouette method.

The Elbow Method

We construct an Elbow plot showing the sum-of-squares for each K between 2 and 20:



From the chart above we infer that the best K is somewhere between 3 and 7. We try K=7 and make a bar chart showing the number of points in each cluster for K-means under the best K:



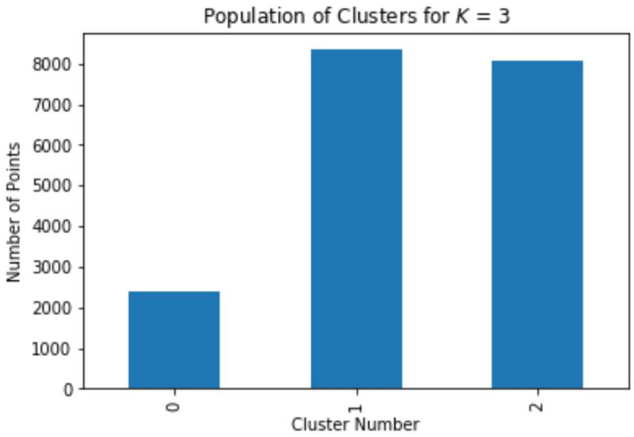
The Silhouette Method

We decide to use a second method for selecting K-value: the Silhouette method. This method measures how well each datapoint "fits" its assigned cluster and also how poorly it fits into other clusters. This is a different way of looking at the same objective. We need to choose K with the silhouette score closest to 1.

For K between 4 and 7 we get the following results:

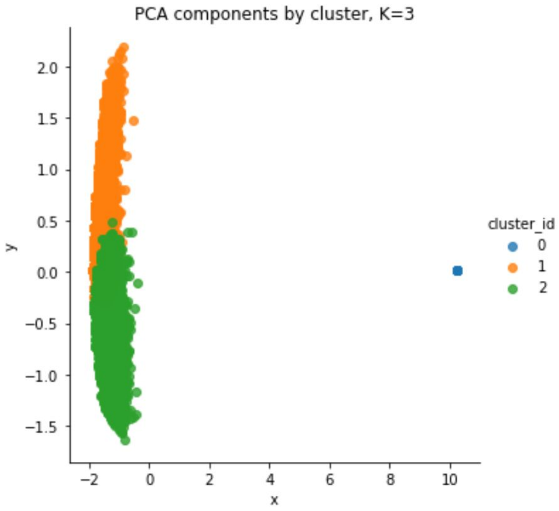
- For n_clusters = 3 The average silhouette_score = 0.19172585856463387
- For n_clusters = 4 The average silhouette_score = 0.18957449729577003
- For n_clusters = 5 The average silhouette_score = 0.18111405590654364
- For n_clusters = 6 The average silhouette_score = 0.16267324366196434
- For n_clusters = 7 The average silhouette_score = 0.1632525460602517

The silhouette score is best for K=3. And even then the score is below 0.25, which indicates that no substantial structure has been found. We decide to use K=3 as the best K for K-means clustering.



PCA Dimension Reduction

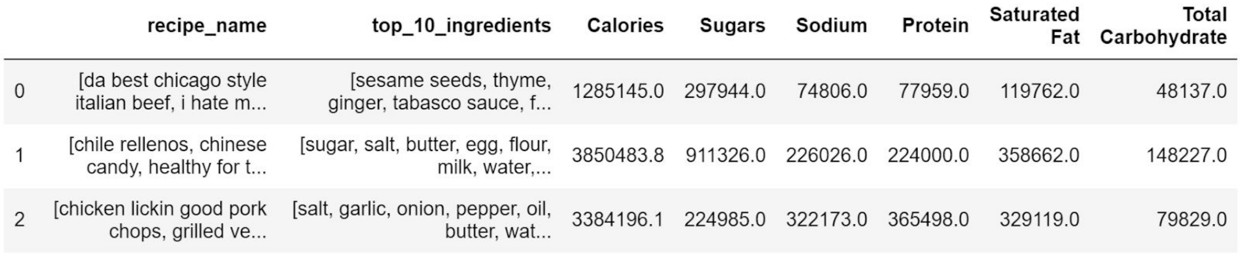
To visualize clusters, we use PCA to reduce the dimensionality of our data from 149 ingredients (dimensions) to 2 dimensions:



To summarize what we did so far, we took the columns of 0/1 for ingredients and transformed them into a 2-D dataset. We took one column and arbitrarily called it x and then called the other y. We showed the x and y points on a scatterplot. We color coded each point based on it's cluster so it's easier to see them. As seen from the scatterplot, there was a clean break between the clusters, and the clusters did not overlap a lot.

Resulting Clusters

We observe the data points in 3 clusters:



	recipe_name	top_10_ingredients	Calories	Sugars	Sodium	Protein	Saturated Fat	Total Carbohydrate
0	[da best chicago style italian beef, i hate m...	[sesame seeds, thyme, ginger, tabasco sauce, f...	1285145.0	297944.0	74806.0	77959.0	119762.0	48137.0
1	[chile rellenos, chinese candy, healthy for t...	[sugar, salt, butter, egg, flour, milk, water,...	3850483.8	911326.0	226026.0	224000.0	358662.0	148227.0
2	[chicken lickin good pork chops, grilled ve...	[salt, garlic, onion, pepper, oil, butter, wat...	3384196.1	224985.0	322173.0	365498.0	329119.0	79829.0

Cluster 1 is the most obvious. The recipes have the highest sugar and carbohydrates content. By observing the recipe names it seems like there are a lot of desserts and sweet things in this cluster.

Cluster 0 seems to have a combination of desserts, salads and meat dishes. It is the lowest cluster by all nutritional values, so we conclude that this is a "mid-day meal" cluster.

Cluster 2 has a lot of recipes with meat and savory recipes, and has the highest protein and sodium content. We conclude that this is the "Dinner" cluster.

To summarize, while no substantial cluster structure was found, we can classify the 3 clusters as:

1. Desserts/Baking Cluster
2. Mid Day Meal Cluster
3. Dinner Cluster