Food Recipe Recommender Systems

Faezeh Yazdi

April 2023

BrainStation Data Science Bootcamp Capstone Project

Eating the same food daily is dull, yet trying new dishes risks wasting ingredients if disliked. Hence, selecting what to cook is a dilemma for all cooks, and can be more time-consuming than cooking.

Background

One of the hot topics in data science is building recommendation systems which can be helpful in tackling the mentioned problem. Recommendation systems are among the areas in data science which have not become mature yet and much research is still going on. These kinds of models are being used in the most famous websites like Amazon, Spotify etc. to help users find the products they most probably like. In our case, we will try to use these models to find similar recipes and recommend ones to the user based on their recipes liked previously.

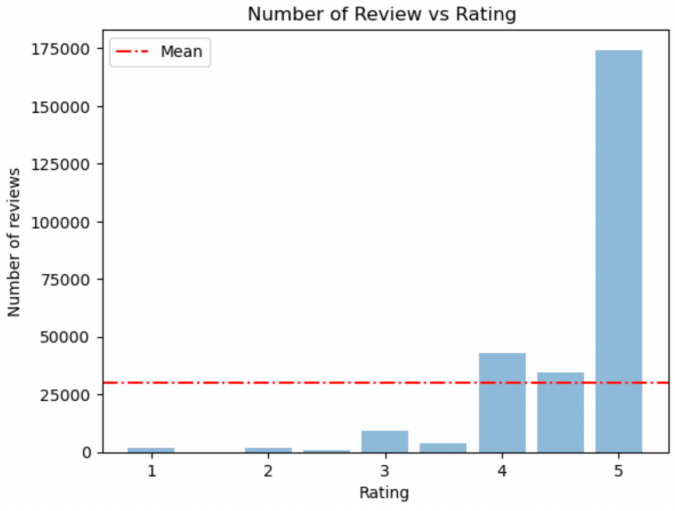
Dataset

The dataset used in this project is originally from the ‘food.com’ website, which is a platform for sharing recipes and reacting to others. It is retrieved from the Kaggle website including two main tables: recipes and reviews. Both are provided in CSV and Parquet format; however, the CSV file is not working properly with the date and keywords columns, and it is recommended to use Parquet for better parsing.

The recipe table is 0.5M and consists of information about recipes ranging from text columns like description, keywords, instruction etc. to numeric columns like nutritional factors including calories, fat, sugar etc. The review table has 1.5M rows about each review users have given to recipes.

Cleaning and Prepressing

About 50% of the recipe table has null values in the rating column which is our target and very important to us and cannot be filled, so they have been dropped and other null values are dealt with properly. The columns that were not in the proper format have changed to the right format. For example, the keywords were.
Nd array and the duration columns were in ISO format. The dataset had a lot of outliers in many columns like time and calories which forced us to implement more research to make sure they are correct data, not some typo. And as they were the right data, we kept them, but we used standard scaling normalization to reduce the effect of outliers and make different columns comparable to each other. Moreover, as shown below the left chart, we are facing a very unbalanced class in the recipe rating, which is our target column. Thus, a new feature is defined to bucket these classes into two main classes: zero for less than 5 and 1 for 5. However, two data are still not equally distributed in these two classes (below right chart). Regarding our huge dataset, an under-sampling approach has been taken to balance two classes.

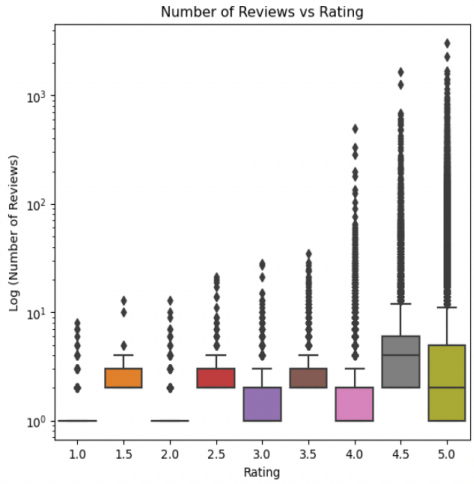


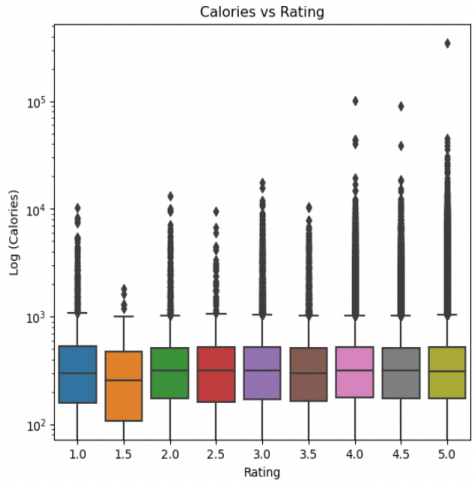
And finally, as the goodness of recommender systems depends on the historical information we have about our users, the users and recipes that had a few numbers of reviews are dropped.

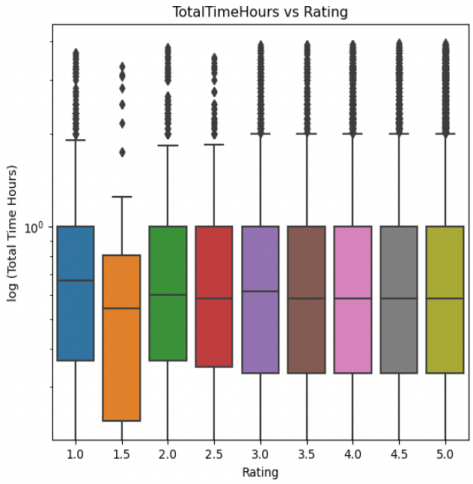
Insights, modeling, and results

Supervised Learning

Different classification models including logistic regression, Decision tree and K nearest neighbours have run to predict the goodness of the recipe's rating. Several technics like normalization, PCA, hyper-parameter tuning, and feature selection have been implemented. The result shows that the model can find out 80% of recipes which are going to be low-rated, while for the other class and in general it's not working well. This result could be predicted as we could not find any clear relationship between our features like calories and time and rating (below chart). However, as the below-left chart illustrates, the more a recipe receives reviews, the better the total score it might get. The number of reviews is not included in the model because it is not a part of the recipe characteristics and the author would not know how many reviews it will receive.







Recommender System

In terms of the recommendation system, we have arrived at good qualitative results. One example of the output of the content-based recommender for a user who liked peanut butter pie is lemon pie, Creamy Peanut Butter Fudge Pie, and Orange Creamsicle Pie.

Moreover, a more mathematical-based recommender system model, FunkSVD has been implemented in the dataset. The result of this model shows good quantitative model evaluations. The model mean absolute error is just 0.1 and the fraction of pairs whose relative ranking order is correct is more than 90%.

Findings and conclusions:

There is no clear relation between the characteristics of the recipe and the average rating it will be ended up receiving. Therefore, it is very hard to predict the final recipe rating. However, our model is able to predict 80% of low-rated recipes correctly.

The recommender systems do a great job of offering new recipes to try for users. The first model does this by finding the similarity between recipes based on their description, ingredients, nutritional factors, and other feature. And the second one is using historical ratings different users gave to different recipes as well as the previous ratings a certain user gave to predict the rating for all the other recipes and recommend the top ones.