What are users eating? A study of the most popular food.com recipes

By Jessica Williams

Problem: Finding the link

- Food.com is looking to have more diverse cuisine types within the recipes that they offer on their site.
- Overwhelming majority of recipes fall within a few select cuisine types because these recipes have had the highest user interactions thus far.
- The goal for this research is to find a significant correlation between the recipe attributes and user interaction and ratings.

Data

To examine this question, we will use a dataset that consists of food.com recipes. The files are as follows:

- ingr_map (Ingredient categories)
- RAW_recipes (Detailed information about recipes)
- PP_recipes (coded/tokenized recipe data)
- Raw_interactions(Recipe rating and review data)
- PP_users (information about users)

Method

After examining the 5 different data sets, I determined that the raw recipes and raw interactions files contained the most useful data. I would like to explore the problem statement through the following steps:

- Descriptive statistics
- Reduce dimensionality
- Create new features
- Determine Predictive power

Data Wrangling

This was a fairly tidy dataset but there were a few structural issues I needed to fix:

- General Cleaning
- Creating text features
- Merging effectively

EDA

- Hypothesis: Ingredient features will have highest correlation.
- Correlations: No significant correlations with rating mean or review polarity.
- New features: Increased ngram range for text features.

Modeling:Testing Models

Since this project aimed to find features that had predictive power for the target variable, I tested a few different regression models on the data. I tested 3 different regression models and the results were as follows:

- Linear Regression: R2 =0.069 and 0.063. mae= 0.646 and 0.642.
- Random Forest Regression: R2 =0.093 and 0.088. mae= 0.647 and 0.642
- K-Neighbors Regression: R2 = 0.028. mae= 0.656

Best Model: Random Forest Regression

Modeling: Hyperparameters

Because I am dealing with a very large data set, I ran a RandomizedSearchCV for the hyperparameters individually. This decreased the time for the CV search dramatically. The following are the parameters were determined best after running each search:

- max_features: Best Param= 50
- n_estimators: Best Param=300
- max_depth: Best Param= 8

Modeling: Results

After applying the model to both data splits with our tuned hyperparameters the results I got were as follows:

- R2 for first target variable(rating mean) =0.134 and 0.105. mae= 0.632 and 0.634
- R2 for second target variable(review polarity average) =0.108 and 0.089. mae= 0.632 and 0.634

The results for the metrics were in the same ball park but since the results were higher for the rating mean as the target variable I decided to stick with those results.

Conclusions

After applying a few different models to the data, we were unable to find a model the showed any real predictive power from the features on our user interaction target variables. A few Ideas for the future:

- Re-working the text features to find significant correlations.
- Create another attribute: “Cuisine Type”

Overall it looks like there is more work that needs to be done to use if we want to make use of our recipe attributes for predicting user interactions. Although this first pass hasn’t shown a significant relationship between the features and user engagement, I think it is worth pursuing further.

The End!