MasterChef – Gather insights from Recipe Data

Predict User Rating

Goals

We set out to solve two problems. First, we wanted to predict recipe review rating based on recipe information. Second, we wanted to figure out different types of recipes based on recipe information such as ingredients used.

Data

We scrapped 11,494 recipes presently in Epicurious.com as our dataset after filtering out any recipe with less than 10 reviews. For each recipe, we scrapped the following information:

- Title
- Image source
- Number of reviews
- Rating
- Make-it-again score
- List of ingredients
- List of tags
- Preparation

Feature engineering and testing association

We also engineered 5 additional features, that were also statistically significant, as features for modeling.

feature	test	statistic	P-value
Image uploaded	Chi-square test of independence	365.03	8.29e-79
Number of ingredients	1-way ANOVA	31.30	3.81e-20
Number of Tags	1-way ANOVA	11.70	1.17e-07
Length of the preparation text	1-way ANOVA	37.04	8.19e-24
Length of the title text	1-way ANOVA	14.96	1.027e-09

Class distribution and defining the baseline model(simplest possible prediction)

Rating class	Frequency
Exceptional **Rating: 4**	1,432
Good **Rating: 3.5**	6,452
Average **Rating: 3**	2,955
Negative **Rating: <3**	655

Since the problem is a classification, we have selected the class that has the most observations and use that class as the result for all predictions as our null model. Hence the baseline model accuracy is 56% (6,452/11,494).

Validation strategy

We randomly portioned the dataset into training and testing sets constituting 80% and 20% of the recipes respectively.

We approached a k-fold cross validation strategy to select a suitable model for prediction. We use the k-fold on the training dataset that we set up based on the split.

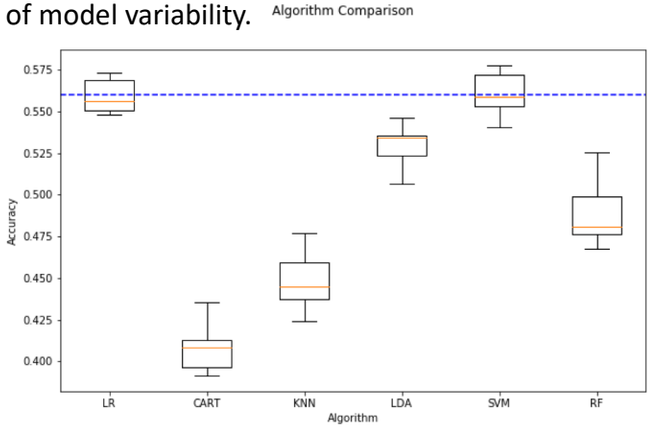
Feature extraction from text data

We apply text analytics to the preparation text to construct a sparse matrix of TF-IDF features. We ended up getting 7784 features. In order to reduce the complexity of fitting the model, we apply PCA to the sparse matrix to reduce the feature set to 500 (variance explained by PCA components reduced to 71.3%)

Selecting a machine learning model using k-fold cross validation

We evaluated cross-fold prediction accuracy on 6 models: Logistic regression(LR), Decision Tree(CART), K-nearest neighbor(KNN), Linear discriminant analysis(LDA), Support vector machine(SVM), and Random forest(RF).

Results: none of the algorithms were able to satisfactorily surpass the baseline model accuracy. LR fit was better than SVM in terms of model variability.



Prediction Accuracy on the test dataset(20%) using LR was 0.568

Further Research

- Use hyper parameter tuning
- Preprocess text data
- Fit the data using neural nets

MasterChef – Gather insights from Recipe Data

Clustering

Goals

We set out to solve two problems. First, we wanted to predict recipe review rating based on recipe information. Second, we wanted to figure out different types of recipe based on recipe information such as ingredients used.

Clustering recipes

We applied the unsupervised learning strategy of k-means clustering on ingredients(tf-tid vectorized), which we could then validate using recipe's tags.

With k=6, the algorithm gave a fair distribution of recipes per cluster. Below table depicts the recipes per cluster and the general theme found based on tags.

Cluster #	Recipes	Color	Theme
0	692	Blue	Dessert/Chocolate
1	2482	Green	Healthy/Salad
2	1800	Red	Healthy/Asian
3	1458	Cyan	Soup/Winter
4	1670	Yellow	Bakery/kid-friendly
5	3392	Magenta	Bakery /Sugar Conscious

Visualizing clusters

We applied PCA on ingredients to visualize the clusters.