STAT 196k Final Project

Kenta Miyahara

Abstract :

The objective of this analysis is to classify food category from ingredients. Let's say there are ingredients: baking powder, sugar, chocolate chips, and flower. It's expected to predict either cookies or pancakes. From the predicted category (pancake or cookies) it will give out all the branded cookie or pancake products. I have used the external dataset that contains some recipe ingredients to test out the classification. Here are some of the results : butter, cocoa, eggs, flour, sugar, white sugar. This has predicted Cookies & Biscuits. Overall, the classification does better job when there is less ingredients.

The first step I took before creating the classification was to learn some basic statistics about the columns that I would be using. In this case, I have used 'ingredients' and 'brandedFoodCategory'. Knowing what the highest counts for branded food category might give some insights about how the classification would perform. Figure is the plot for the occurrence of food categories. It shows that 'Candy' has the highest occurrence in the dataset. This might cause bias towards a food item that contains sugar to be classified as candy more frequently. If the ingredients include some poultry, salt, and sugar, it might predict that it's a candy more than it would predict that it's a meat product. I will check my assumption after the introduction of the classification.

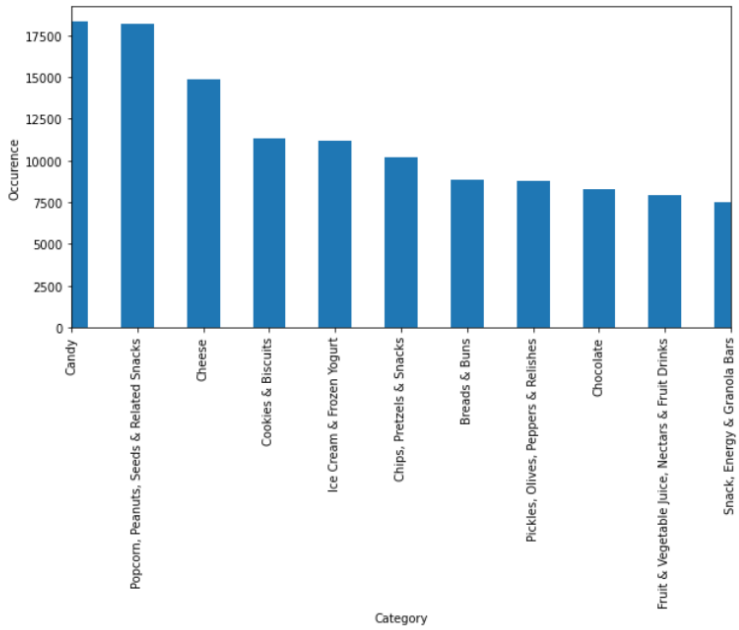

Figure 1

The classification algorithm that I will be using is scikit.learn naïve bayes multinomialNB. The reason why I chose this is because there are multiple categories instead of binary prediction. The training is done using the 'ingredients' and 'brandedFoodCategory' columns from USDA branded food dataset. Branded food category is the function and ingredients are the predictors. I have used count vectorizer to turn the ingredients into sparse matrix to make it suitable for classification. As of the input for classification, instead of using the dataset from USDA dataset I have imported external dataset 'simplified-recipies-1M Dataset' by Dominik Schmidt. This dataset includes all the ingredients from food recipes across multiple dataset. The data is in NumPy array format and each ingredients are separated by list so that I corresponds to each food. Figure 2 is the top 5 classification from recipe ingredients.

```
Input reci_ingr : bell pepper, black pepper, boneless skinless chicken, boneless skinless chicken breasts, buttermilk, celery,
celery seed, chicken, chicken breasts, chicken stock, corn kernels, dry mustard, flour, frozen corn, frozen corn kernels, garli
c, granulated garlic, honey, lemon, lemon juice, minced garlic, mustard, olive, olive oil, onion, onion powder, paprika, peppe
r, powder, red bell pepper, salt, skinless chicken breasts
predict category : Frozen Dinners & Entrees
----------------------------------------------------------
Input reci_ingr : baking soda, butter, buttermilk, cheese, chocolate chips, cinnamon, cocoa, cream cheese, eggs, flour, heavy w
hipping cream, margarine, powdered sugar, semi sweet chocolate, semi sweet chocolate chips, soda, softened butter, sugar, sweet
chocolate, vanilla, water, whipping cream
predict category : Cookies & Biscuits
----------------------------------------------------------
Input reci_ingr : dark rum, for color, grenadine, juice, lime juice, rum, simple syrup, syrup
predict category : Dips & Salsa
----------------------------------------------------------
Input reci_ingr : caster, caster sugar, corn flour, cream, demerara sugar, egg, egg yolks, flour, juice, milk, oranges, sugar,
whipping cream
predict category : Ice Cream & Frozen Yogurt
----------------------------------------------------------
Input reci_ingr : baby greens, black pepper, cheese, dates, dried dates, goat cheese, greens, ground, ground black pepper, mesc
lun, olive, olive oil, other, pepper, sauce, soy sauce, temperature, vinegar
predict category : Pickles, Olives, Peppers & Relishes
```

Figure 2

Figure 2 shows some valid classifications, such as cookies & Biscuits are made from flowers, eggs, and butter. However, there are some classifications that are questionable: dark rum, grenadine, juices, and syrup. These ingredients were classified as Dips & salsa.

In the classification outcome, the most frequent occurrence was first Frozen Dinners & Entrees, second pickles, olives, peppers & Relishes, and third cookies and biscuits. The parameter of figure 3 only contains 100 classifications since there were more than million recipes in the external dataset. The assumption made earlier about candy being biased category, in the outcome it was not effected at all. Candy only has been classified 4 times out of 100 recipes.

One of the things that I should have done is to analyze the performance and accuracy of the model. Also, there are some string inside the external dataset that are not processed perfectly such as in figure 2, prediction for dips & Salsa. In the ingredient there is 'for color ', this is not an ingredient name and it should be removed.

Overall, I was surprised that the classification worked pretty well. The things that I would do to make this better is to increase the accuracy of the classification by fine tuning the model or to have a cleaner dataset to train with.

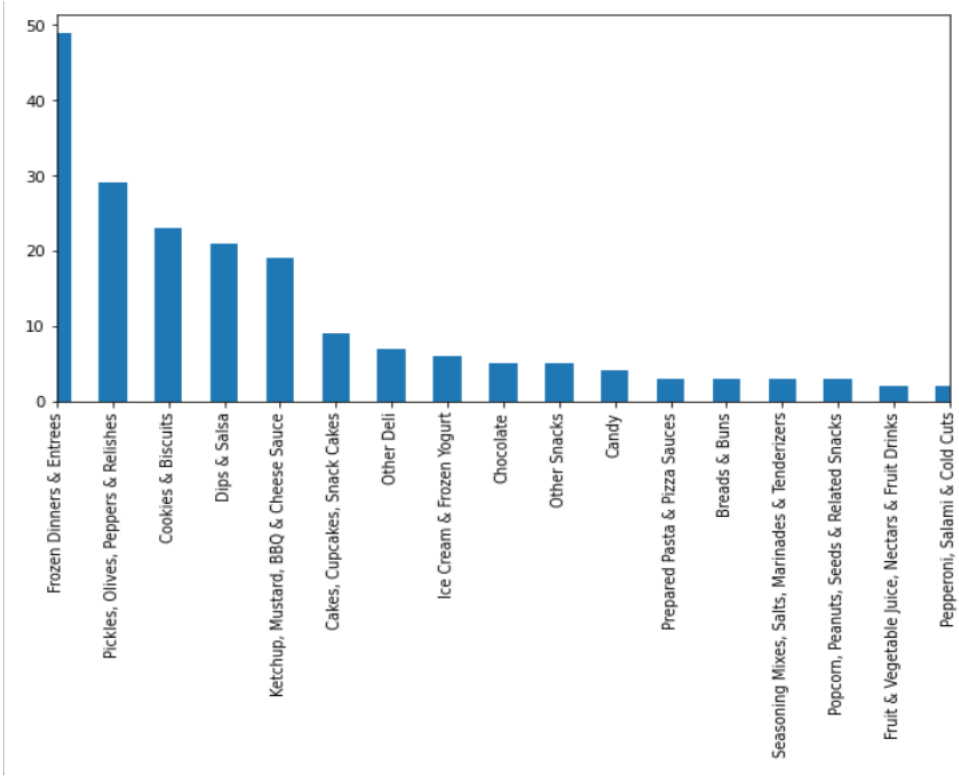
50
40
30
20
10
0
Frozen Dinners & Entrees
Pickles, Olives, Peppers & Relishes
Cookies & Biscuits
Dips & Salsa
Ketchup, Mustard, BBQ & Cheese Sauce
Cakes, Cupcakes, Snack Cakes
Other Deli
Ice Cream & Frozen Yogurt
Chocolate
Other Snacks
Candy
Prepared Pasta & Pizza Sauces
Breads & Buns
Seasoning Mixes, Salts, Marinades & Tenderizers
Popcorn, Peanuts, Seeds & Related Snacks
Fruit & Vegetable Juice, Nectars & Fruit Drinks
Pepperoni, Salami & Cold Cuts

Figure 3