Chocolate!

Kimberly Cable

Belleview University

DSC 550: Data Mining

Dr. Brett Werner

August 13, 2022

Chocolate!

Chocolate. Who doesn't like it? It's been around since the 19th century BCE. But what makes chocolate highly desirable? In this study, I plan to see if certain factors determine whether certain chocolate in the United States is highly delectable.

Why predict the best chocolate?

Chocolatiers, home cooks, and lovers of chocolate are all looking for that one piece of chocolate that will impress. Knowing the correct combination of cocoa, cocoa butter, and/or vanilla or even relying on a reviewer's most memorable characteristic of the chocolate could make or break your company, cookie, or even relationship. Being able to find the best chocolate quickly and easily will save time and money in avoiding poor chocolates for your venture.

Data Selection and EDA

Data

The data I will use comes from the Manhattan Chocolate Society, Flavors of Cacao website, and their Chocolate Bar rating table. The data includes features such as cocoa content, where it was made, bean origin, the number of ingredients, and its most memorable characteristics from reviewers and their ratings. I will also use the USA Craft Makers table merging it with the Chocolate Bar ratings to look more closely at United States chocolates.

Analysis

In analyzing the data, I plan to see if I can predict the best chocolate in the United States based on its rating and see if any features play a part in making it good chocolate.

I will look at questions such as:

Which states have the best chocolate?

Which ingredients make the best chocolate?

How much cocoa makes for good chocolate?

What are the key characteristics of good chocolate?

Where do the best cocoa beans originate from?

Challenges

Some of the challenges I see are cleaning the dataset and separating out some of the columns. One major hurdle is knowing the best way to classify the ratings. Do I keep the ratings as they are? Do I group them into the 5 original rating scales or use some other grouping to best use to predict a good piece of chocolate?

Data Cleaning

The data was scraped and moved into CSV files to begin explanatory data analysis. Some cleaning steps that needed to be taken were to standardize the spelling of some states in the Chocolate Rating table to help merge with the United State Chocolate Makers table. Once the two tables were merged, I pulled out just the United States chocolate to analyze further.

Initially looking at the data, most of the chocolate in the United States is rated “Recommended” and “Highly Recommended,” (see Figure 1). Getting the average rate per state most States are rated as “Highly Recommended” (see Figure 2). This would indicate a good chance of getting good chocolate anywhere in the United States.

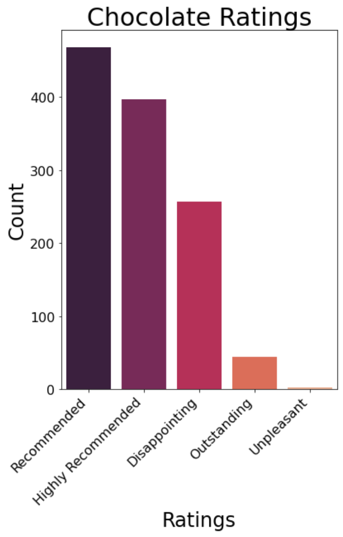


Figure 1: Chocolate Ratings for United States Chocolatiers

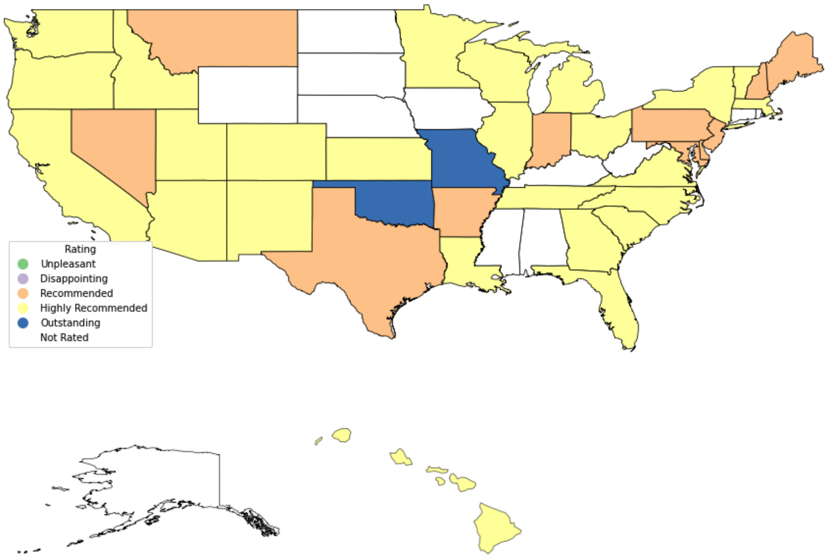


Figure 2: Average chocolate rating per State in the United States.

California has the most Chocolatiers in the United States (see Figure 3) and most of the cocoa beans originate from the Dominican Republic (see Figure 4).

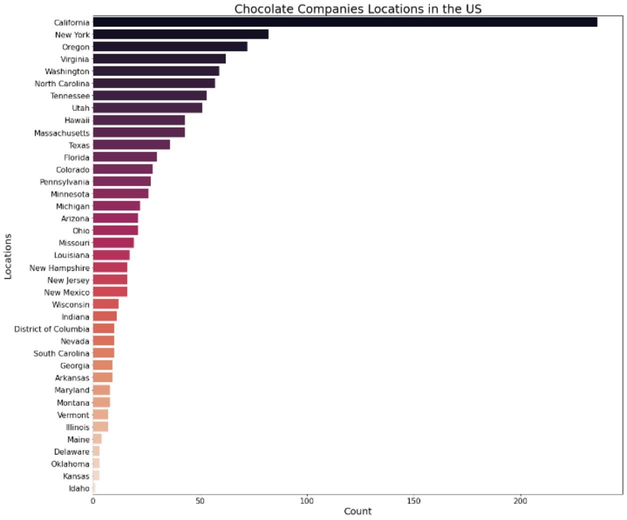


Figure 3: United States chocolate company locations.

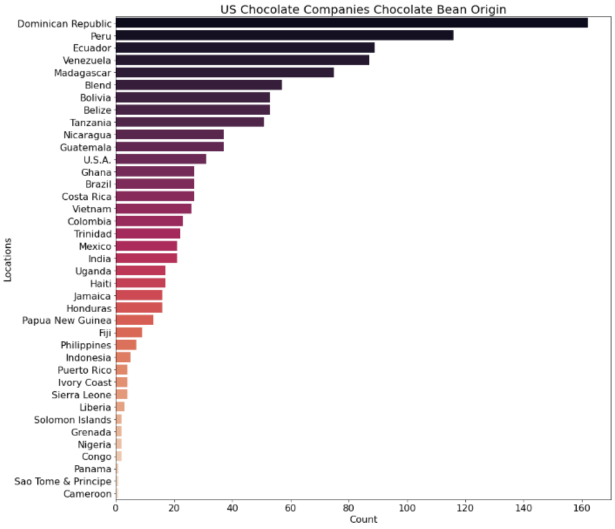


Figure 4United States cocoa bean origins.

Looking at ingredients, butter and sugar and butter, sugar and cocoa butter are the most popular ingredients (see figure 5), and cocoa percent is around 70% (see figure 6).

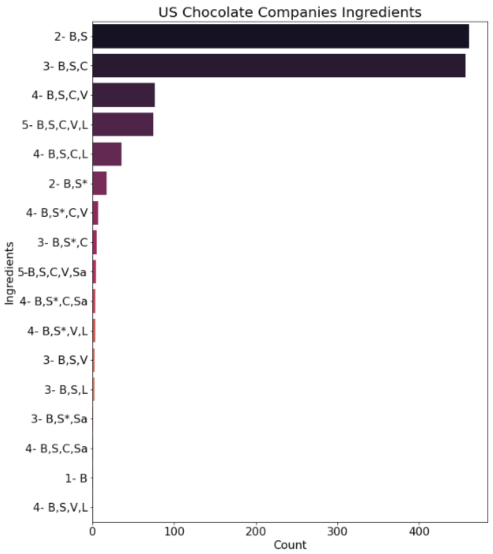


Figure 5: United States chocolate ingredient count

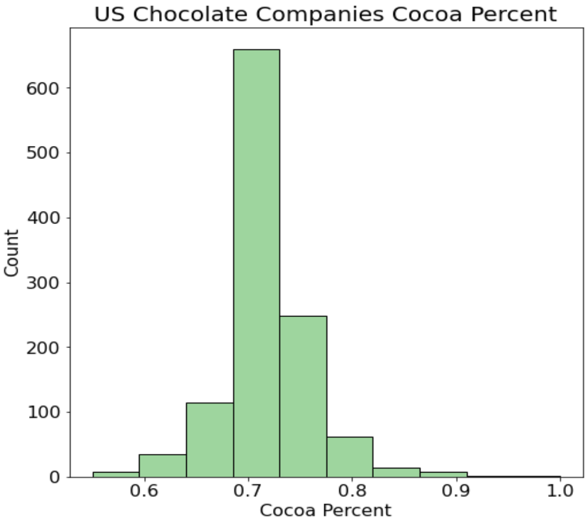


Figure 6: Cocoa percent in United State chocolate count

The top 20 most memorable characteristics in reviews used words such as “sweet”, “cocoa”, nutty”, and “roasty” (see figure 7)

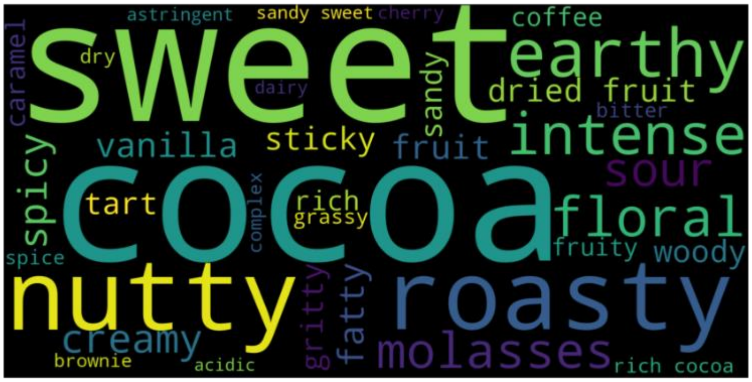


Figure 7: Most memorable characteristics from reviews about United States chocolates.

But do these features alone constitute a good piece of chocolate? Is the best chocolate made in Oklahoma, have 70% cocoa, be made with only butter and sugar, have their cocoa beans originate from the Dominican Republic and be called “sweet” and “nutty”?

Data Preparation

To prepare the data, I binned the ratings into 3 main classifications. Ratings above 4 were binned together for an “Above Average” rating, ratings between 3 and 4 were binned together for a “Recommended” rating, and ratings below 2 were binned together for a “Below Average” rating.

I reduced the number of memorable characteristics to the top 20 words used by reviewers. I also kept only the top 10 ingredients and bean origin countries. This helped eliminate very small samples in the lower-used features.

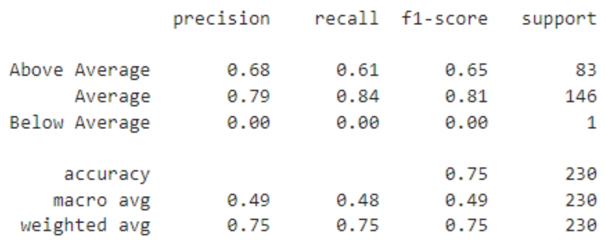
Dummy variables were created for all categorical features. This will create binary features for each class in each original categorical feature.

Using a 20% test and 80% train calculation, I split the data into training and testing datasets to properly train the best model and see how it will perform on new data.

Model Building and Evaluation

Wanting to predict good chocolate, I decided to look at three models: Random Forest Classifier, Logistic Regression, and KNN Classifier. All three were chosen due to the fact they were good at predicting a multi-class target. What we are looking for is a good prediction of ratings of “Above Average” and “Average”. This would indicate finding a good piece of chocolate would be easy in the United States.

The first step was to look at all the features I prepared and determine the best model that had a good accuracy score and correctly predicted average to above average scores. Using a grid search, I found that the Random Forest Classifier faired the best with an accuracy score of 75%. It correctly predicted a rating of “Average” 81% of the time, and “Above Average” 65% as indicated by the f1-score for each. It did not predict “Below Average” at all with an f1-score of 0%. This was probably due to the fact we did not have enough samples at the lower rating to successfully predict chocolates at that level (see Table 1).

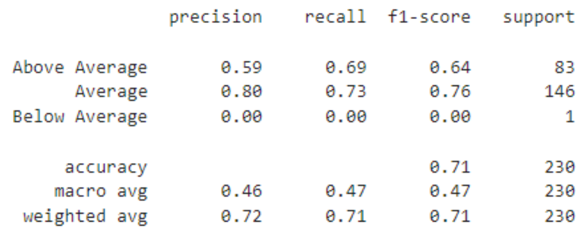


```
              precision    recall  f1-score   support

Above Average       0.68      0.61      0.65        83
      Average       0.79      0.84      0.81       146
Below Average       0.00      0.00      0.00         1

     accuracy                           0.75       230
    macro avg       0.49      0.48      0.49       230
 weighted avg       0.75      0.75      0.75       230
```

Table 1: Classification Report using all features.

The precision scores indicate a good ability to correctly predict each rating for “Average” and “Above Average” ratings, but it had trouble with correctly predicting positive “Below Average” ratings as indicated by the low recall score of 0%.

Next, I looked at reducing the number of features that do not help to predict the best chocolates using Principal Components Analysis. Using the same three models I used previously, using a grid search, I found that the Logistic Regression model predicted chocolates the best. With the reduction in features, the accuracy score was 71%. It correctly predicted a rating of “Average” 76% of the time, and “Above Average” 64%. It did not predict “Below Average” at all with a score of 0%. This indicated that we did not have enough samples at the lower rating to successfully predict chocolates at that level. (See Table 2).



```
              precision    recall  f1-score   support

Above Average       0.59      0.69      0.64        83
      Average       0.80      0.73      0.76       146
Below Average       0.00      0.00      0.00         1

     accuracy                           0.71       230
    macro avg       0.46      0.47      0.47       230
 weighted avg       0.72      0.71      0.71       230
```

Table 2: Classification Report with reduced number of features.

The precision scores indicate a good ability to correctly predict each rating for “Average” and “Above Average” ratings, but it also had trouble with correctly predicting positive “Below Average” ratings as indicated by the low recall score of 0%.

Conclusion

Rating	All Features	Using PCA
	Random Forest Classifier	Logistic Regression
	F1-Score	
Above Average	0.65	0.64
Average	0.81	0.76
Below Average	0.00	0.00
Accuracy	0.75	0.71
Weighted Average	0.75	0.71

Table 3: Summary of the Classification Reports of both models

In predicting chocolate in the United States, I believe personal taste is the true measure. Trying to predict good chocolate, and shown, can be tricky. Aside from that, looking strictly at accuracies, the Random Forest Classifier fared slightly better than using Logistic Regression, but both were above 70% (see Table 3). For me, the most important indicators were the F1 scores of each of the ratings. Using all features and the Random Forest Classifier, chocolates in the "Average" and "Above Average" ratings were correctly predicted 65 to 81% of the time. The weighted average was also 75% correctly predicting a rating. I believe using all features and the Random Forest Classifier would help initially get you good chocolate. Then, you can personally choose your tastes and preferences from the initial choices the model predicted.

The next step I would take to see if I could accurately predict good chocolate is to include all chocolates not just those made in the United States. This will give us more data to work with, especially in the bottom classification. Looking at other reviews besides the Manhattan Chocolate Society ratings will help broaden the data to different opinions.

References

Flavors of Cacao. (n.d.). Retrieved from Chocolate Bar Ratings: http://flavorsofcacao.com/