Characteristics of Chocolate: An associative and predictive analysis of the variables that contribute to satisfaction with chocolate

Kelsey Woods, Aayushi Verma, Manas Vani, Chandani Thumar

Seidenberg School of Computer Science and Information Systems, Pace University

Dataiku Hackathon

November 13th, 2022

Abstract

Chocolate is a well-known favorite in the world of dessert. There are many factors that contribute to the consumer’s satisfaction when enjoying the guilty pleasure. Through the use of tools such as Python and Weka, Team *Crunch “the data” Bar* analyzed the Dataiku-provided dataset, with the goal to discover not only the association between factors such as “nutty”, “earthy”, and “fatty” in terms of satisfaction, but also to predict the ratings of this satisfaction.

Research Problem

We are provided a dataset of over 2000 chocolates' defining characteristics such as cacao bean source, company, cocoa percentage, ingredient composition, most memorable characteristics, and ratings. We are interested in learning about the specific characteristics of chocolate that make it likable (or delicious) and in constructing models to predict ratings of chocolates.

Data Cleaning Methods

There are two data files provided, "data.csv", and "scoringData.csv". The first file is for training the data, and the latter for testing the data. For the training dataset, cleaning was performed using Python and packages like pandas and numpy. We removed missing variables, dropped "NA" or null values, encoded character strings to numeric variables, and converted objects to float or integer data types, coerced variables when necessary, and summed all possible null rows by variable. All these data cleaning steps resulted in us being able to use it for further analysis with no complications.

Exploratory Data Analysis (EDA)

The Python pandas package provided the tools to download, visualize, and transform the data. Visualization of the target variable (rating) accomplished by calculating the mean and other descriptive statistics by column. A box plot was used to visualize the descriptive statistics and determine outliers. All variables vs target variables (rating) were visualized and calculated for an overall view of the correlations. To further understand which features contributed the least and most to ratings. We found that features such as cocoa, creamy, vanilla, ingredient_beans, ingredient_sugar were highly correlated with rating, and had a high probability of affecting the target variable column the most.

Prediction and Findings

We used Weka to test various models on the data, in order to quickly ascertain which models might provide a good fit for the data. We found that the Random Forest Regression method provided the best results. We used Python's scikit-learn package to use a random forest regression model on the training data to train the model, and used the test data (scoringData.csv) to generate the predictions. We obtained a mean absolute error of 0.41 degrees, and a model accuracy of 86.67%. Fig. 1 shows the distribution of the predicted ratings we generated, and Fig. 2 shows a box plot distribution of the predicted ratings. We used the scikit-learn function *features_importance* to extract the weights of the features in the model. The key features that influenced rating in the model are Cocoa_percent, Country_of_bean_origin, Company_location, Number_of_ingredients.

Conclusions

We analyzed the dataset of chocolate characteristics to determine the features of chocolate that affect ratings. The data was prepared using methods in Python, primarily the pandas package. Using the cleaned data, we performed exploratory data analysis in order to examine the variables and their relationships with each other. We determined correlations of features which could be affecting the ratings. We then converted this clean dataset to an ARFF format and used WEKA to identify machine learning models, including the best fit; random forest regression. A set of predictions for chocolate ratings were created. A model was trained using the training data in Python with the help of scikit-learn. A high accuracy of 86.67% was obtained using the adjusted random forest regression, indicating that the predictions were highly reliable. Key features were identified. The features which affected the ratings of chocolate were Cocoa_percent, Country_of_bean_origin, Company_location, Number_of_ingredients. It can be concluded that the listed variables affected the ratings the most. Further analysis of this dataset should include a more in-depth comparison of machine learning models to identify all best models and features.

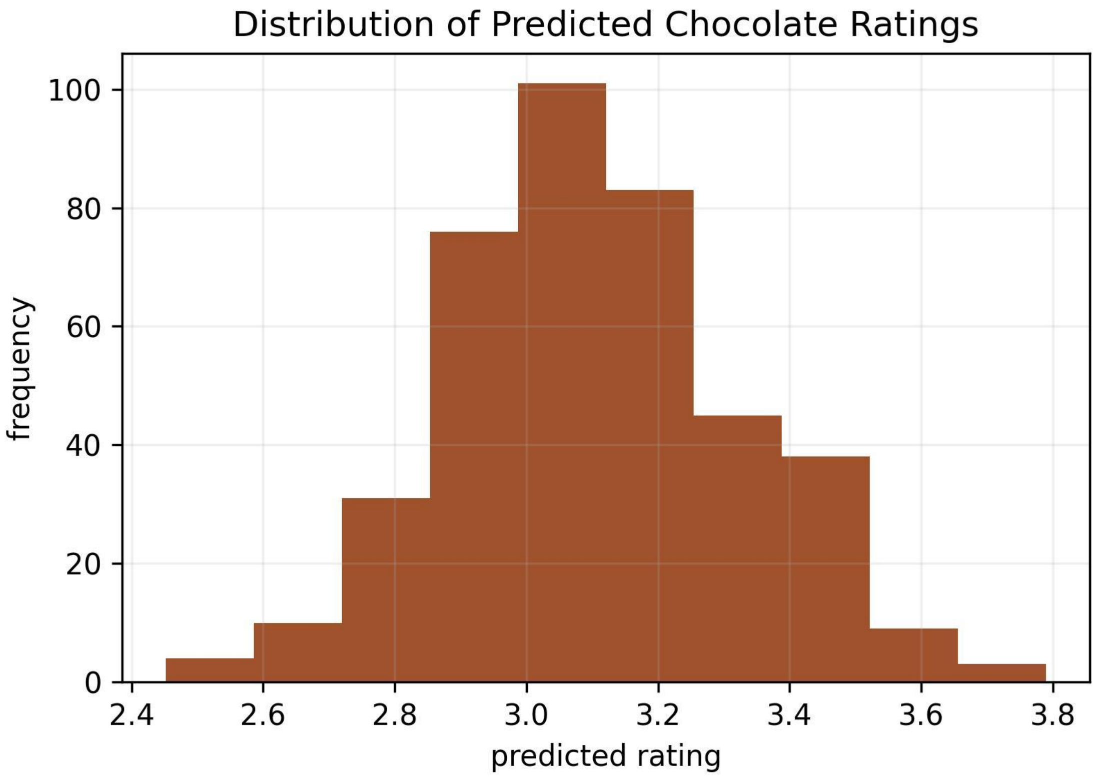

Fig. 1: Distribution of obtained predicted ratings of chocolate.

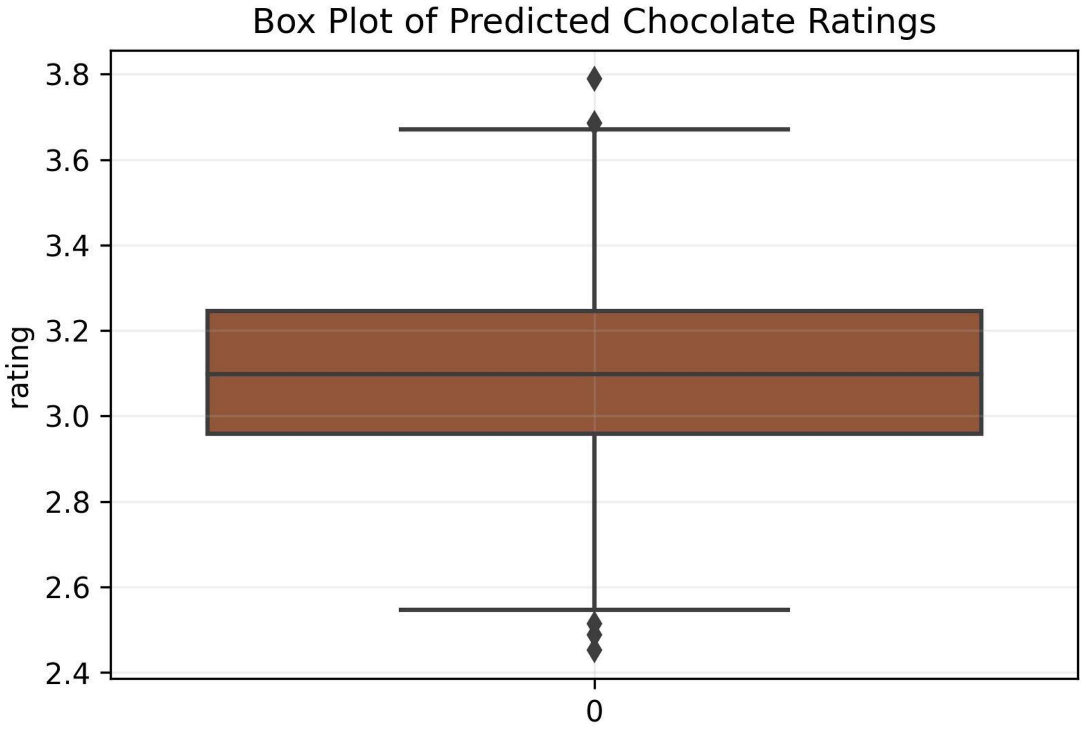

Fig. 2: Box plot distribution of obtained predicted ratings of chocolate.

Appendix

- Image source: https://www.thespruceeats.com/a-guide-to-chocolate-varieties-520311
- WEKA: https://www.cs.waikato.ac.nz/ml/weka/index.html
- Most Used Packages in Python: https://www.edureka.co/blog/python-libraries/