Chocolate Rating Project

Marcel Colvin

Intro

- Can we find the attributes that make up the best chocolate bar?
- A new chocolatier may find our analysis interesting to use ingredients and other attributes that may make a better chocolate bar.

Methodology

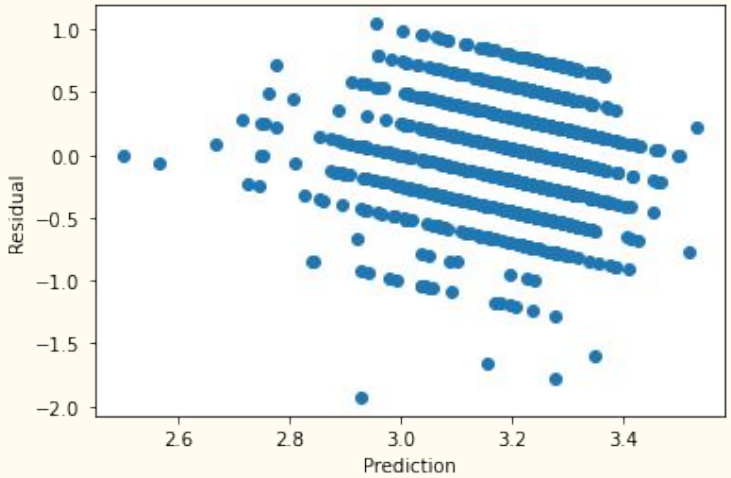
- Data: The data set is 2500 chocolate bars rated by a chocolate reviewer over the last 12 years. This includes some ingredient data, cocoa bean origin location, and other categorical data.
- Tools: Google Sheets, Pandas, Sci-Kit Learn, and Tableau

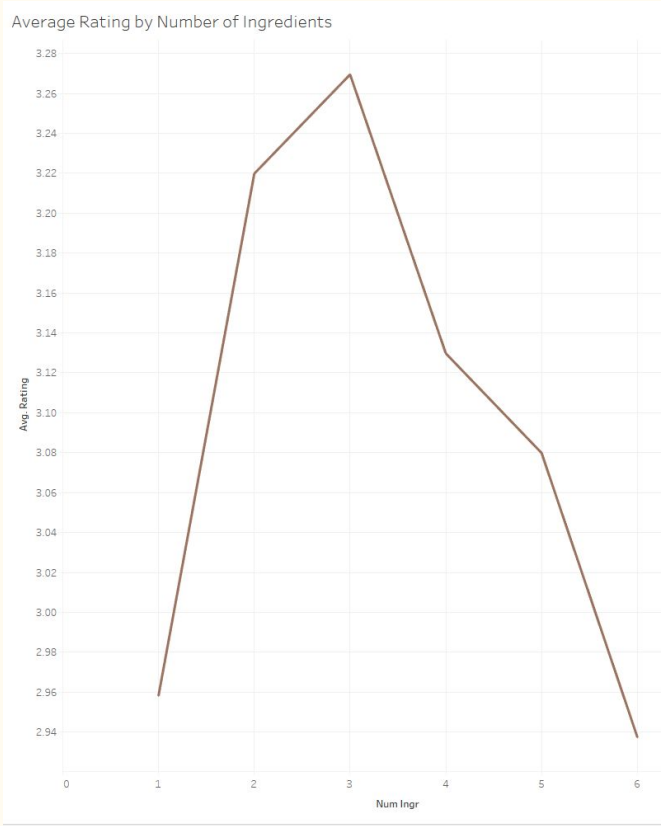
Methodology

- Metrics:
  - Manufacturer, Bean Origin, Main ingredients
  - Cocoa Percent, Latitude/Longitude, Number of Main ingredients
  - Predicting rating
- Models:
  - Linear Regression
  - Random Forest Regression
  - Extra Trees Regression

Results (Linear Regression)

- Linear Regression was poor due to the lack of continuous numerical data and poor linear correlation between the rating and any variables.
- R^2 ~0.08





Results (Extra Trees Regression and Random Forest Regression)

- Extra Trees performed worse than Random Forest after Grid Search Hyperparameter optimization.
- Random Forest is slower, but has better optimization.
- Both of these regressors worked better than the linear regression.
- Extra Trees Train R^2 ~0.2 Test R^2 ~0.10
- Random Forest Train R^2 ~0.4 Test R^2 ~0.10

Conclusions

- Using the Random Forest Model, I found that the most valuable attributes in predicting highest rating were:
  - Cocoa Percent, Vanilla, Latitude and Longitude, Manufacturers (with high numbers of reviews), and the number of major ingredients used.
- The model only predicts one person's subjective taste in chocolate, and does not perform particularly well.

Future Work

- Find Commodity Pricing Data and create an optimization model
- Perform NLP sentiment analysis on the “Most Memorable Features” column

Thank you