All About Chocolate

2022 May

Team CatDog: Yulin Guo, Moyi Tian, Cindy Zhang, Songhao Zhu

Introduction: Motivation and Goals

REF	Company (Manufacturer)	Company Location	Review Date	Country of Bean Origin	Specific Bean Origin or Bar Name	Cocoa Percent	Ingredients	Most Memorable Characteristics	Rating
2454	5150	U.S.A.	2019	Tanzania	Kokoa Kamili, batch 1	76%	3- B,S,C	rich cocoa, fatty, bready	3.25
2458	5150	U.S.A.	2019	Dominican Republic	Zorzal, batch 1	76%	3- B,S,C	cocoa, vegetal, savory	3.5
2454	5150	U.S.A.	2019	Madagascar	Bejofo Estate, batch 1	76%	3- B,S,C	cocoa, blackberry, full body	3.75

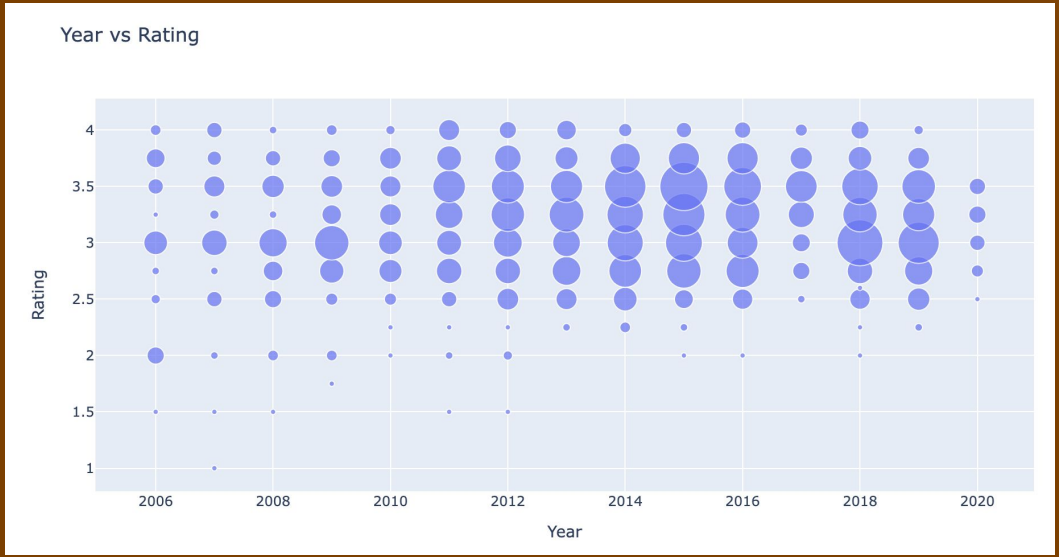
Which factors are the most important to the manufacturers for new product development?

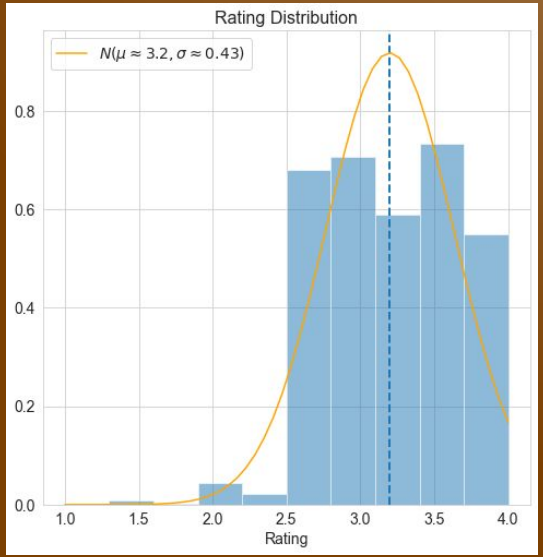
Given the information of a chocolate bar, what is the prediction of its rating?

Data

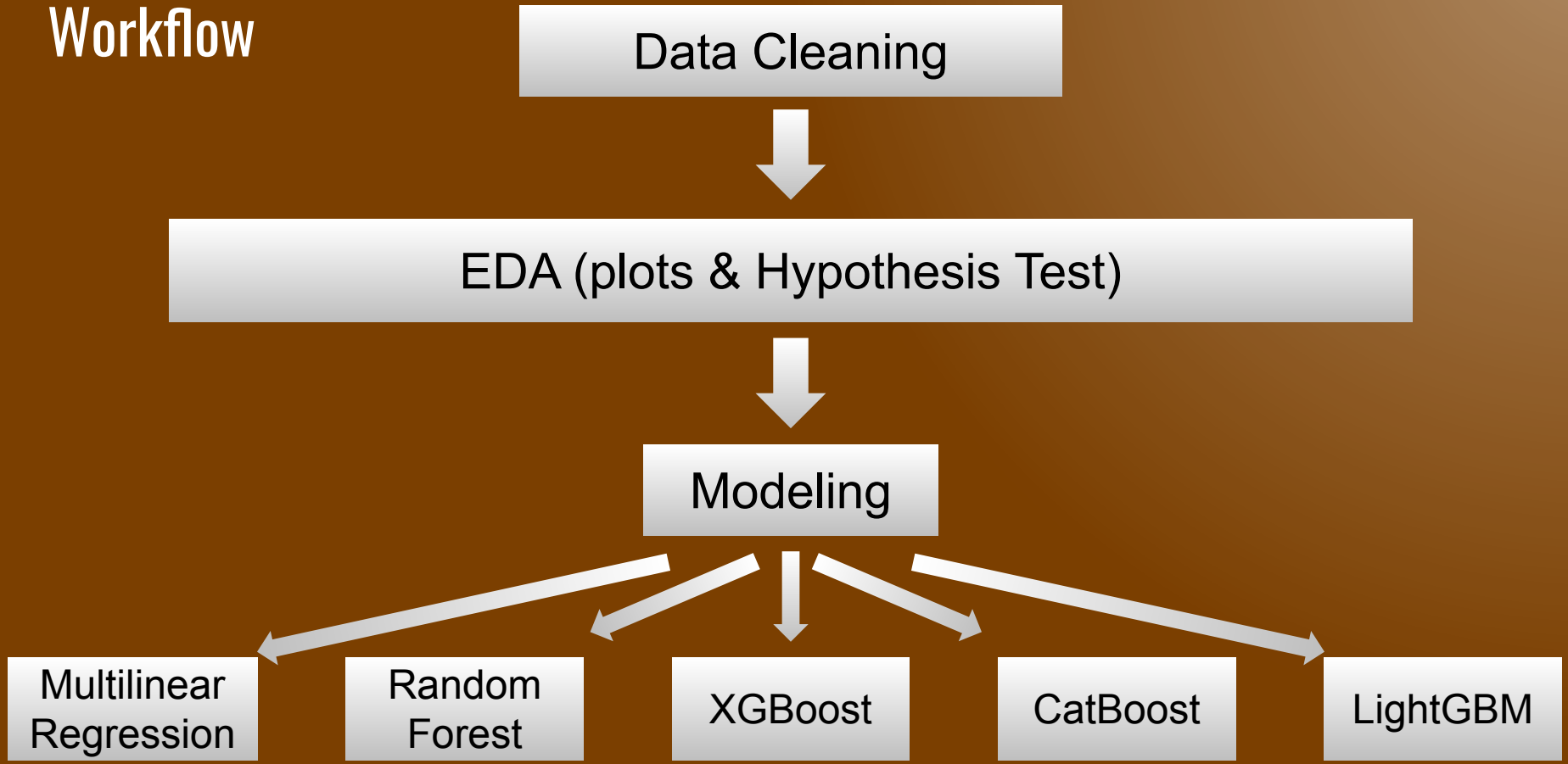
Features:
Ingredients (cocoa percent, sugar, etc), bean origin, taste profile...
Target: ratings (1 - 4)

Training data size: 2223
Testing data size: 223





Workflow


Data Cleaning
EDA (plots & Hypothesis Test)
Modeling
Multilinear Regression
Random Forest
XGBoost
CatBoost
LightGBM

Results

Mean absolute error is used to evaluate the trained models using the test data set.

	Baseline (mean)	Multilinear Regression	Random Forest	XGBoost	LightGBM	CatBoost
Test MAE	0.2935	0.2729	0.2765	0.2533	0.2545	0.3034

Takeaway

- Taste profile is the most important feature
  - Fruit flavors, vanilla, creamy, etc. have the greatest impact on ratings
- Cocoa percent is positively correlated to its rating
- Chocolate made from cocoa beans and companies from Africa and Asia have better ratings
- XGBoost has the best performance in predicting ratings in terms of MAE

Limitations

Our models:

- Do not handle time data very well
- Taste profiles are manually organized and encoded
  - Requires a lot of effort. It may not be able to handle unseen tastes
- Do not distinguish outliers very well

What could be done?

- Use a pre-trained word2vec model to get the embedding of each taste profile, then use a clustering model to group the tastes into a few categories based on cosine similarity
- Treat as classification problem and compare the result with regression models
- Treat as unsupervised problem to group chocolates and check if it is aligned with the original rating

Reference

- GitHub:
  - https://github.com/2022CatDog/Chocolate

- Data Set:
  - https://www.kaggle.com/datasets/soroushghaderi/chocolate-bar-2020
  - http://flavorsofcacao.com/chocolate_database.html