One-Hot Coffee

Team Members: Cassidy Madison, Ethan Semrad, Ching-Lung Hsu

https://github.com/madisonc27/Team-Dragonfly

May 2022 The Erdos Institute
Data Science Bootcamp

Introduction

- **Rationale:** Coffee is one of the world's most popular beverages. An estimated 75% of the US adult population reported drinking coffee. (Loftfield, Erikka, et al., 2016)
- **Target Audience:** Coffee importers and distributors.
- **Main Question:** Can we find a correlation between coffee taste rating and other features?
- **Our approaches:** We approach this question in two different trials.
  1. Classify the country of origin/altitude/bean processing method.
  2. Predict the overall rating based on other features.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4997286/

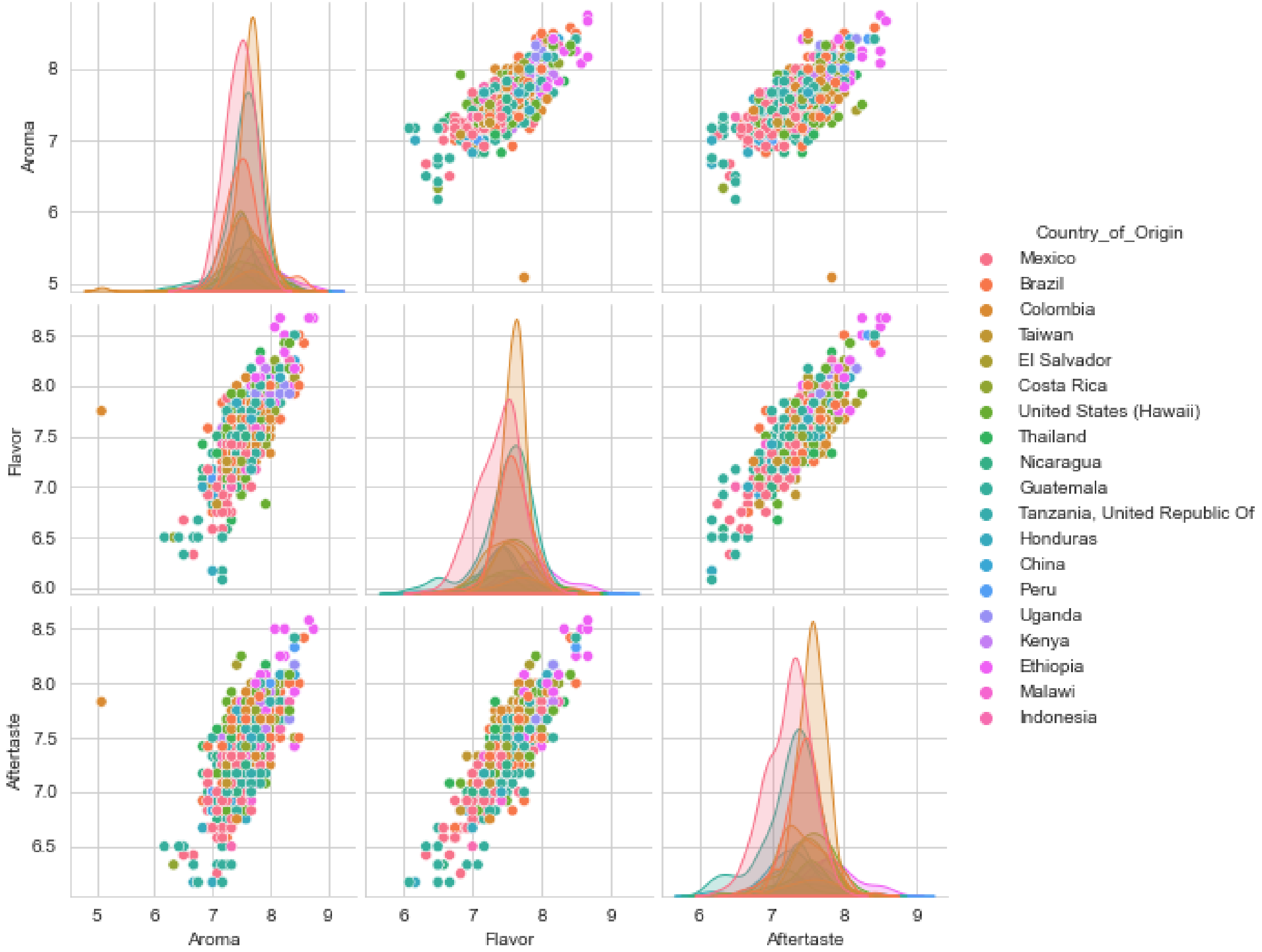
The First Trial

EDA

- *Selected predictors:*
  Aroma, Flavor, Aftertaste, Acidity, Body, Balance, and Uniformity

- *Selected feature of interest:*
  Country of origin

- *Although certain countries appear to have slight separation, most have a high degree of overlap*



Model Training

- *Applied several supervised learning models to the clean data including*
  1. *K-nearest neighbors*
  2. *Decision Tree*
  3. *Random Forest*
  4. *AdaBoost*
  5. *Support Vector Machine*
- *Used accuracy as a base metric to compare the models*
- *Accuracy was around 30 – 35% for each model*

Confusion Matrix for SVM

	Predicted Mexico	Predicted Colombia	Predicted Guatemala	Predicted Brazil	Predicted Taiwan	Predicted United States (Hawaii)	Predicted Honduras
Actual Mexico	28	4	2	1	0	2	0
Actual Colombia	4	20	3	1	0	1	0
Actual Guatemala	8	6	14	0	0	0	0
Actual Brazil	7	5	2	3	0	2	0
Actual Taiwan	5	2	1	1	1	0	0
Actual United States (Hawaii)	5	2	2	0	0	1	0
Actual Honduras	7	1	0	0	0	0	0

Conclusion

Several factors could contribute to the low accuracy:

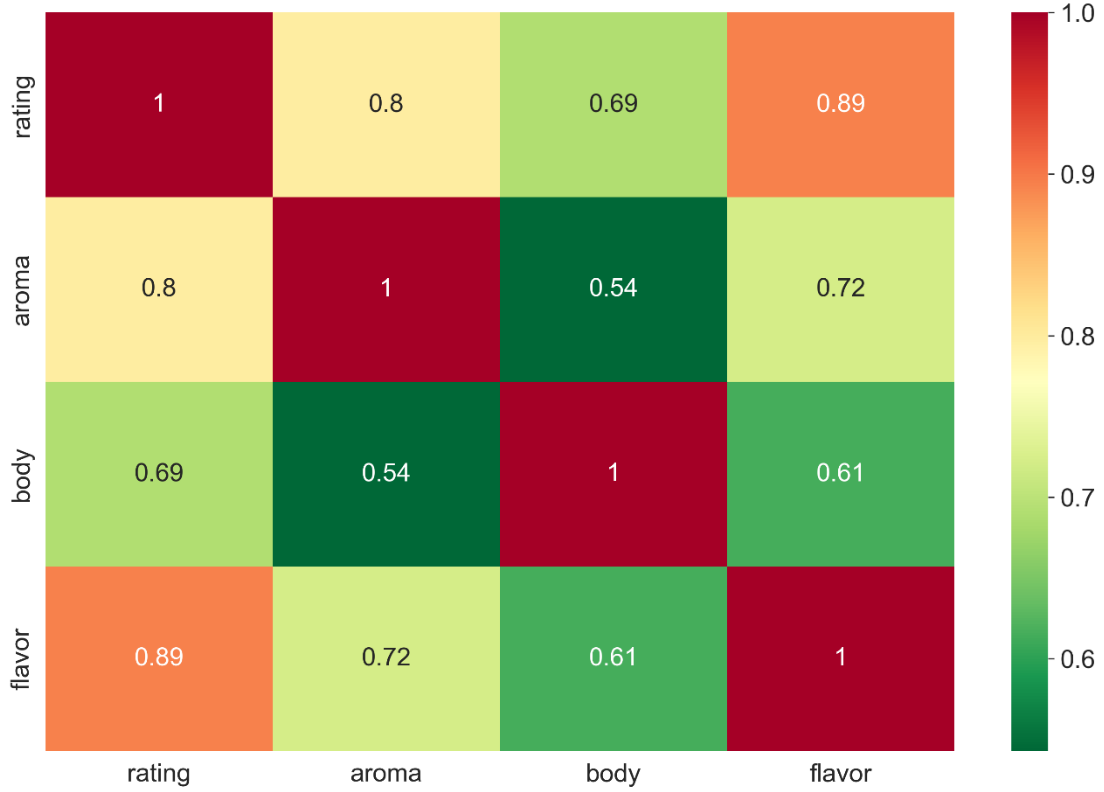
- *High correlation between the predictor variables*
- *High degree of overlap between countries*
- *Models tend to place predictions into categories with the largest number of samples*
- *Coffee Quality Institute ratings are not able to distinguish between different countries in these models*

The Second Trial

Exploratory Data Analysis

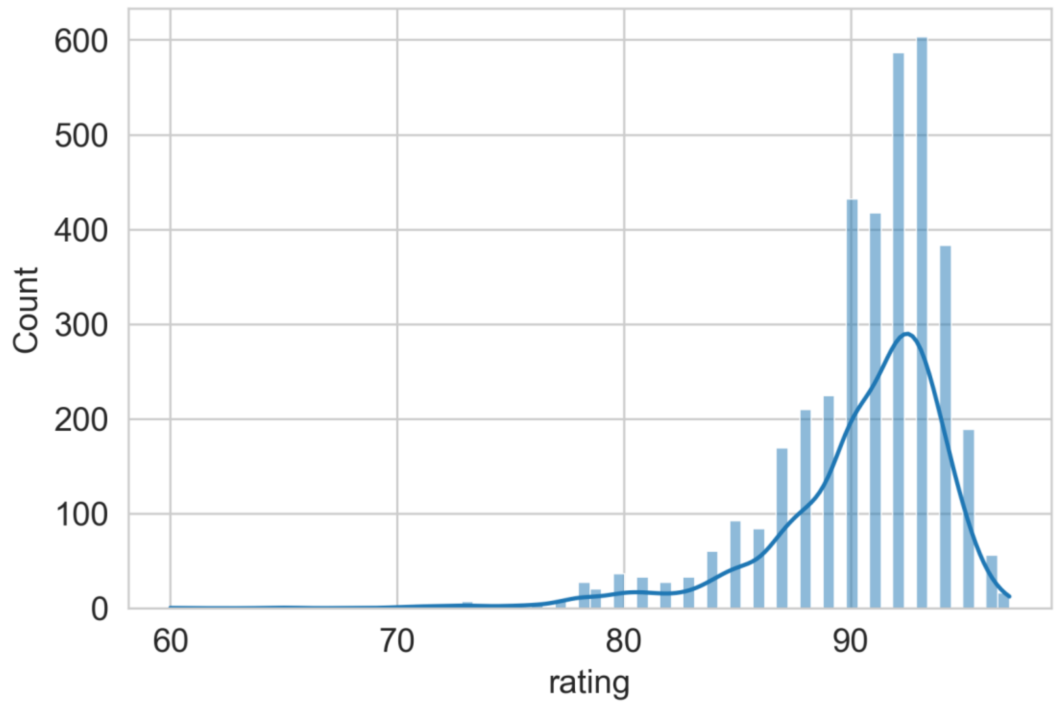
- *Numerical features are highly correlated*
- *Chose only categorical predictors*
- *Selected predictors:*
  Region, Roast, Espresso, Organic, Blend, Fair Trade, Decaffeinated, Pod/Capsule, Estate
- *Selected feature of interest:*
  Rating

Correlation across Numerical Features

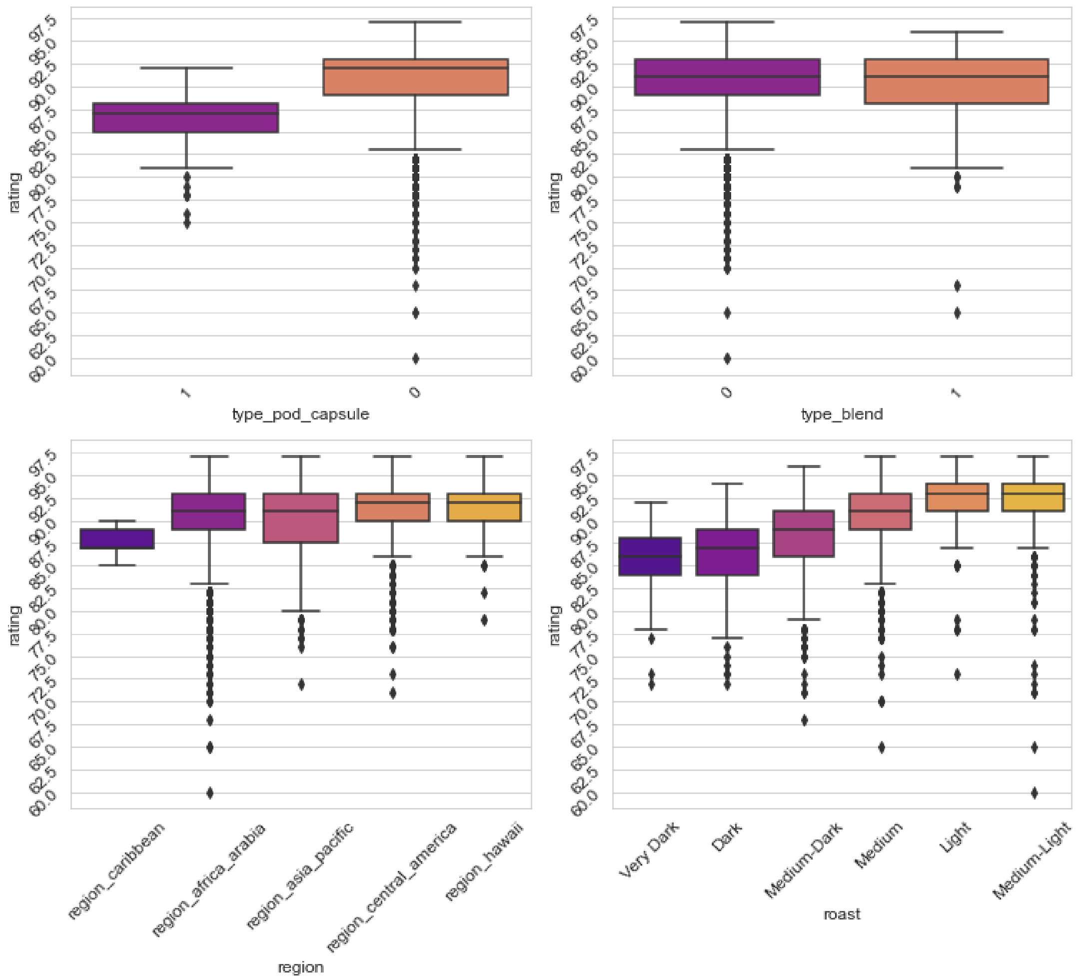


Exploratory Data Analysis

Histogram and Density Plot for Rating



Boxplots for Categorical Features



Model Evaluation

- *Applied multiple linear, lasso, and ridge regression*
- *Used mean rating as baseline prediction*
- *Added interaction terms to multiple linear regression*
- *Evaluated using mean squared error (MSE) and mean absolute error (MAE)*

Test	MSE	MAE	RMSE
Baseline	13.711274	2.753970	3.702874
MLR	8.944063	2.081188	2.990663
Ridge	8.943361	2.081775	2.990545
Lasso	9.044396	2.075086	3.007390
MLR_Interaction	8.815056	2.039720	2.969016

Conclusion

Key Takeaways

- *Strong positive correlation – features to seek out*
  Africa/Arabia, espresso, estate, light and medium-light roast

- *Negative correlation – features to avoid*
  pod/capsule, medium-dark, dark, and very dark roast

- *No correlation – features that have little impact*
  Regions: Asia/Pacific, South America
  Organic, fair trade, decaffeinated, blend, medium roast

	alpha=0.1
region_africa_arabia	0.660580
region_caribbean	-0.038320
region_central_america	0.044232
region_hawaii	0.015138
region_asia_pacific	0.000000
region_south_america	0.000000
type_espresso	0.458025
type_organic	0.000000
type_fair_trade	0.000000
type_decaffeinated	0.000000
type_pod_capsule	-0.267004
type_blend	0.000000
type_estate	0.155451
Light	0.345396
Medium-Light	0.625126
Medium	0.000000
Medium-Dark	-0.849087
Dark	-0.954651
Very Dark	-0.817816

Future Directions

- *Incorporate price data as a predictor*

- *Utilize natural language processing to extract key words from the professional flavor descriptors*

THANK YOU