Rating 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

- **Facts:** Coffee is one of the world's most popular beverages. Therefore, it is important to have a further study on coffee recommendation for the coffee lovers around the world.

- **Targeted Users:** An estimated 154 million adults (75%, age ≥ 20 y) of the US population reported drinking coffee. (Loftfield, Erikka, et al., 2016)

- **Our approaches:** We approach this question in two different trials.
  1. Classify the country of origin/ altitude/ bean processing method.
  2. Predicting the overall rating based on other features.

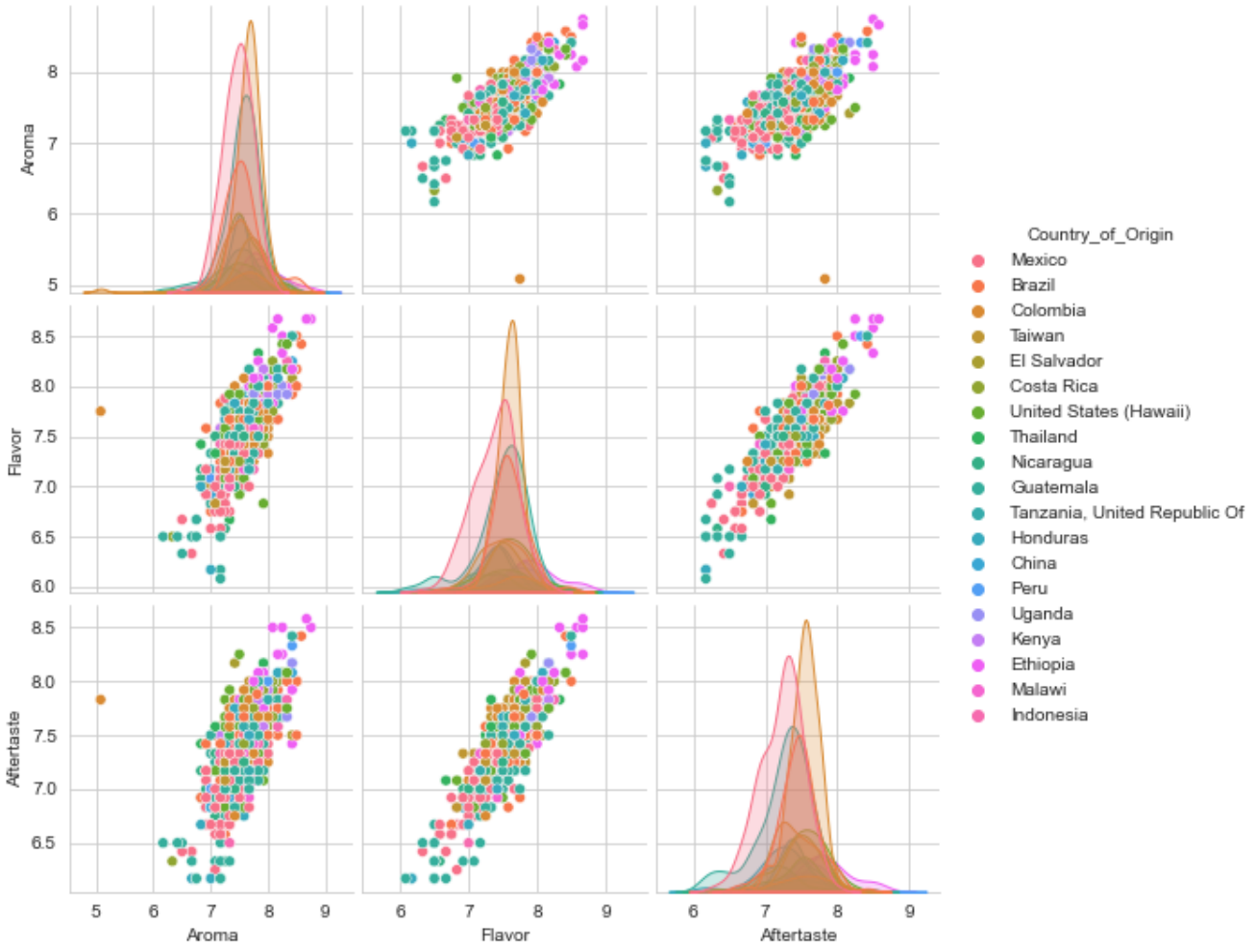
Data Source:

1. https://www.kaggle.com/datasets/ankurchavda/coffee-beans-reviews-by-coffee-quality-institute
2. https://www.kaggle.com/datasets/aryansakhala/coffee-recommendation

The First Trial

EDA

- *We removed irrelevant columns and any entry missing review scores or information on the country of origin.*
- *After cleaning, a portion of pairplot exploring the relationships between variables can be seen on the right.*
- *Although certain countries can be separated from others, many of them are highly overlapping each other.*



Mode Training

- *Applying 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*
- *We use accuracy as a base metric to compare the models. However, the a ccuracy is around 30 – 35% for all the model described above.*

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

- *Upon reflection, we think several factors cause the low accuracy of the models.*
- *The correlation between each predictors are pretty high.*
- *The classification models tend to place the samples into the categories with the largest number of samples. Since the data overlaps a lot. This leads to the low accuracy rate.*
- *We conclude that the Coffee Quality Institute ratings do not differ significantly between different countries, growth altitudes, or processing methods.*

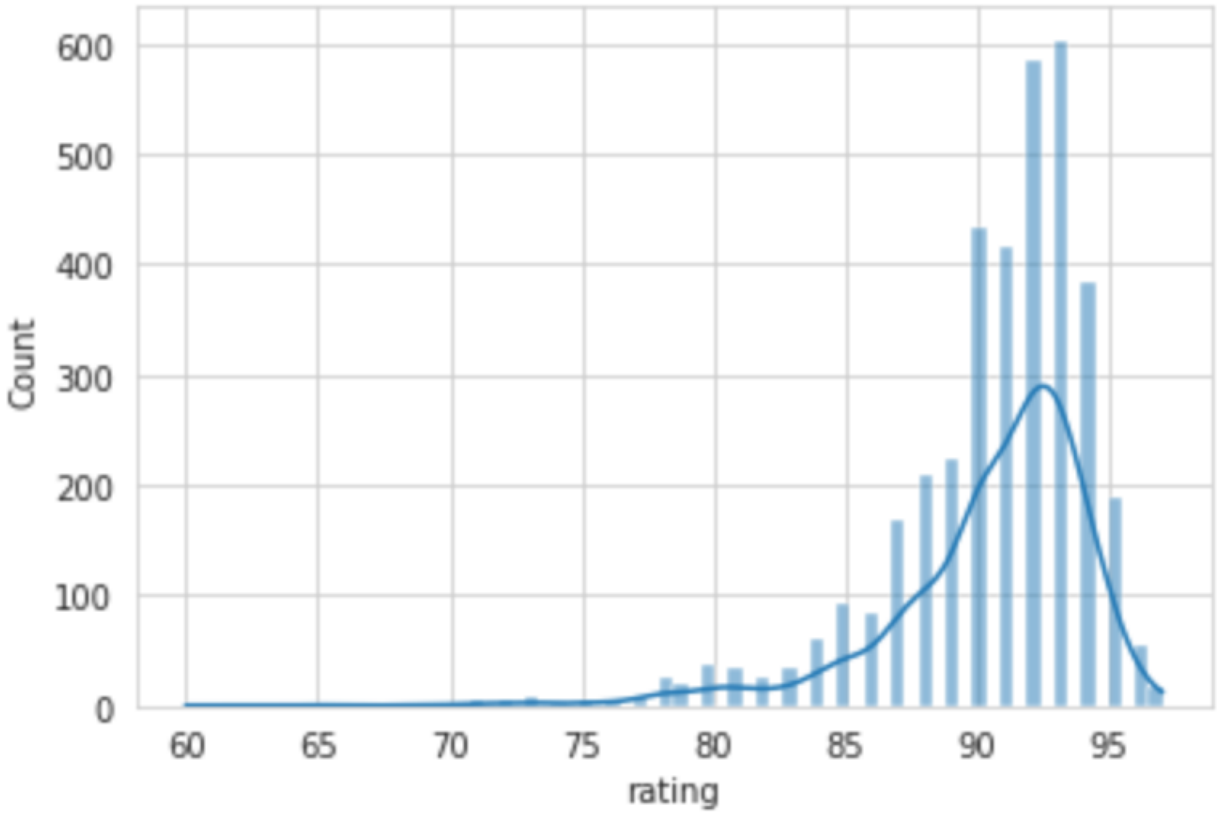
The Second Trial

Exploratory Data Analysis

- *After cleaning, this data contains 4,677 x 21 reviews and we separate the data into training (80%) and test set (20%).*

- *The distribution of rating is slightly left-skewed.*

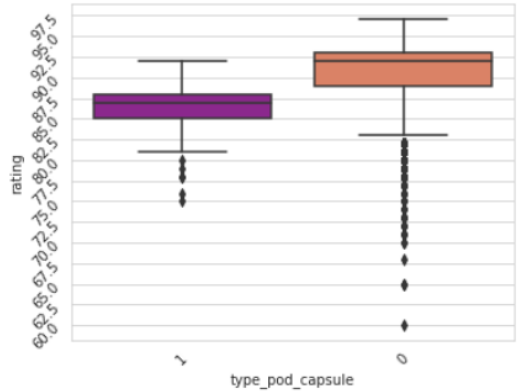
Histogram and Density Plot for Rating

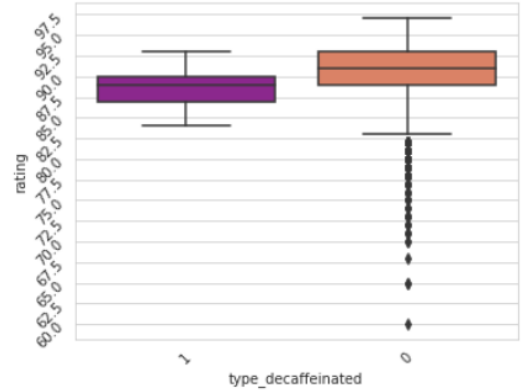


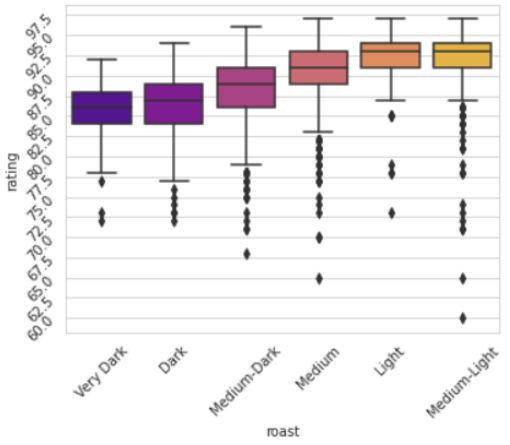
Exploratory Data Analysis

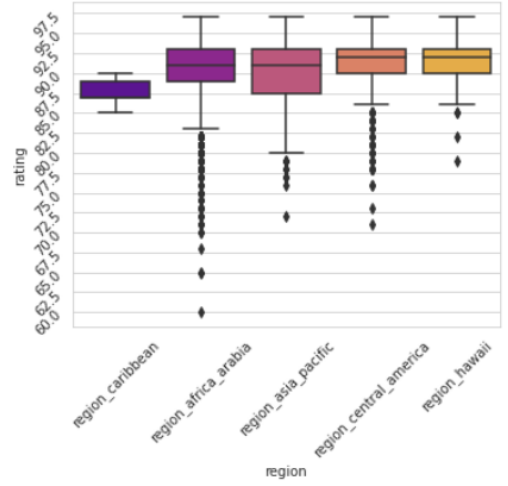
Correlation across Numerical Features

Boxplots for Categorical Features



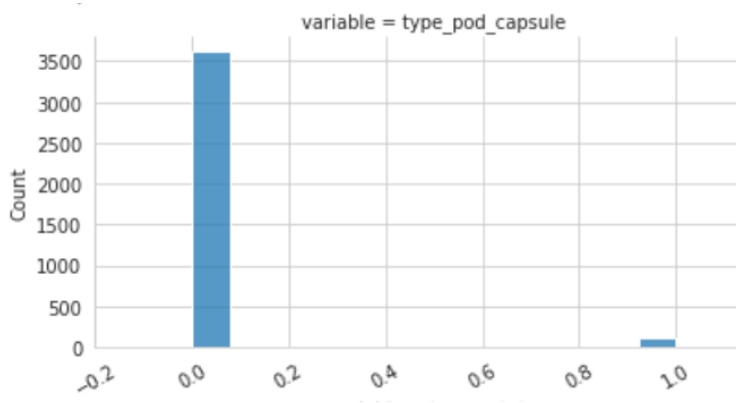


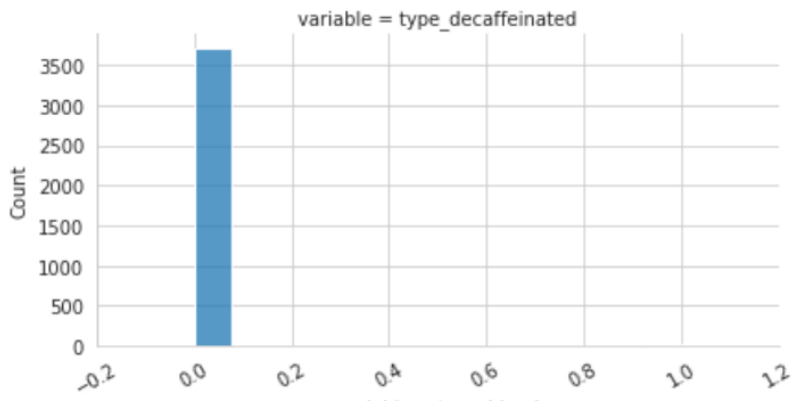


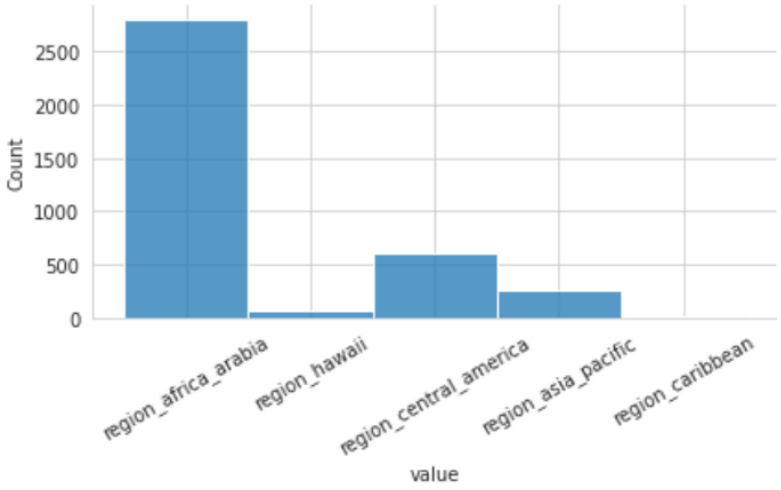


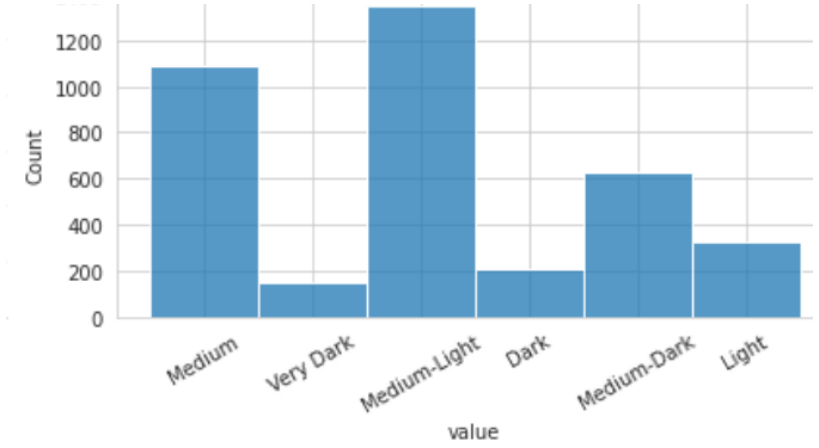
Exploratory Data Analysis

Imbalanced Predictors









Model Training

- *Since the numerical predictors are still highly correlated, we decided to use only the categorical predictors.*
- *We apply Multiple Linear, Losso, Ridge Regression to our training data.*
- *We create a baseline model that simply making prediction as the mean rating from the training set to compare the performance.*
- *We did not apply a polynomial regression because our predictors consist entirely of 0s and 1s.*
- *At the end, we also include interaction terms with predictors with stronger effect.*

Model Result

- *The following table summarizes the performance of each model we mentioned earlier.*
- *We applied these models to the test data set and got the mean square error (MSE) and mean absolute error (MAE).*

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

Key Takeaways

- *Lasso Regression also reveals which features are more important for a higher consumer rating.*
- *Lighter/ Midium-Light roast has more positive effect on the rating while darker roast tends to lower the rating.*
- *Africa/Arabia coffee tended to rate higher, while Caribbean coffees rated slightly lower.*
- *Features like Organic, fair trade, decaffeination, and blends does not affect the rating that much.*

	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

- *Tidy up the format of price data contained in our data and incorporate them into the predictors of our models.*

- *Add some sentiment analysis like natural language proces sing for the coffee review to extract the features of the comments from the customers.*

THANK YOU