Capstone Project
Coffee Shop

By Lindsey Caughlin
Special Thanks to Mentor: Kenneth Gil-Pasquel

Problem: What features most affect the ratings of coffee reviews?

Information benefactors:

- Widely known chains for purposes of creating new coffee blends
- Newly established coffee shops or regional chain coffee shops for creating signature blends
- Coffee roasters advertising their specialty coffee roasts
- Coffee bean sourcing companies determining what coffee bean to source

Data Sourcing and Cleaning

Data taken from a Kaggle dataset with 19 features and 7041 entries

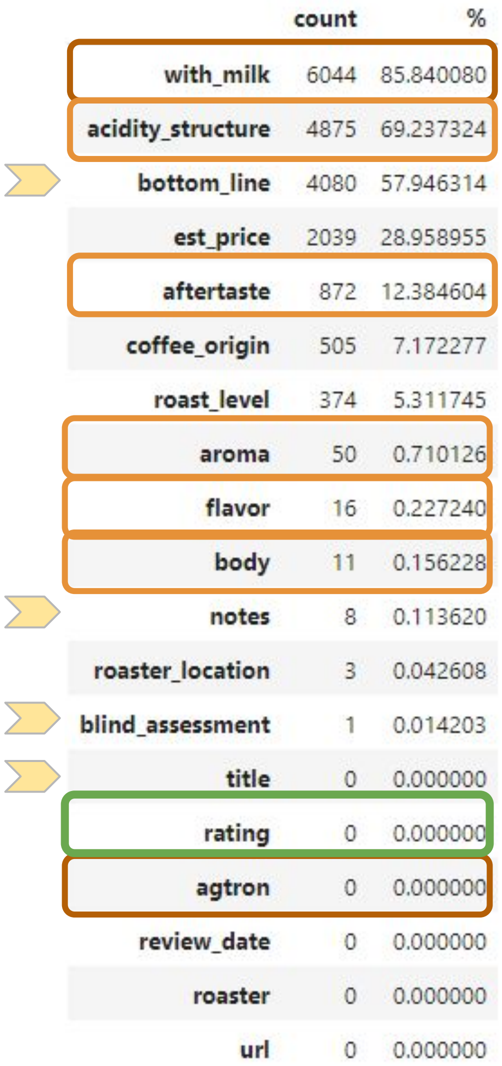
- Numeric, descriptive, and narrative features

Missing Values:

- 'with_milk' feature deleted
- 'agtron' was a descriptive feature, but had hidden unusable data

Numeric Features changed from object form:

- Target feature: 'rating' (1-100)
- 'acidity_structure' (1-10)
- 'aftertaste' (1-10)
- 'aroma' (1-10)
- 'body' (1-10)
- 'flavor' (1-10)



	count	%
with_milk	6044	85.840080
acidity_structure	4875	69.237324
bottom_line	4080	57.946314
est_price	2039	28.958955
aftertaste	872	12.384604
coffee_origin	505	7.172277
roast_level	374	5.311745
aroma	50	0.710126
flavor	16	0.227240
body	11	0.156228
notes	8	0.113620
roaster_location	3	0.042608
blind_assessment	1	0.014203
title	0	0.000000
rating	0	0.000000
agtron	0	0.000000
review_date	0	0.000000
roaster	0	0.000000
url	0	0.000000

Exploratory Data Analysis

Profile Report provided a summary, visualizations, and notes for each feature.

The report also raised 28 alerts:

high cardinality

features with high correlation

features with missing values

uniformly distributed features

Commentary features dropped:

- title
- blind_assessment
- bottom_line
- notes
- review_date
- url

Other dropped features:

- acidity_structure *(high correlation with the rating, 62% null)*
- agtron *(hidden null values)*
- est_price *(most common value only 2.2%)*

Features with null values replaced:

- aftertaste
- aroma
- body
- flavor
- coffee_origin
- roast_level
- roaster_location

Preprocessing:

Dummy variables for categorical features

StandardScaler to scale the magnitude of numeric features

80/20 train/test split

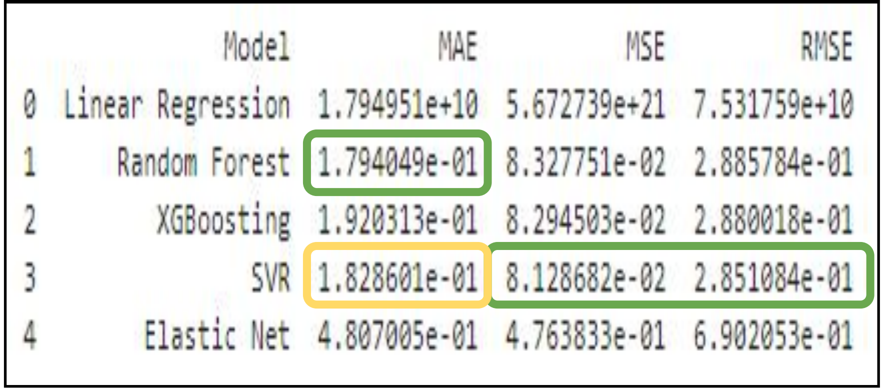
- Target feature: ‘rating’

Modeling:

Models run with and without GridSearchCV():

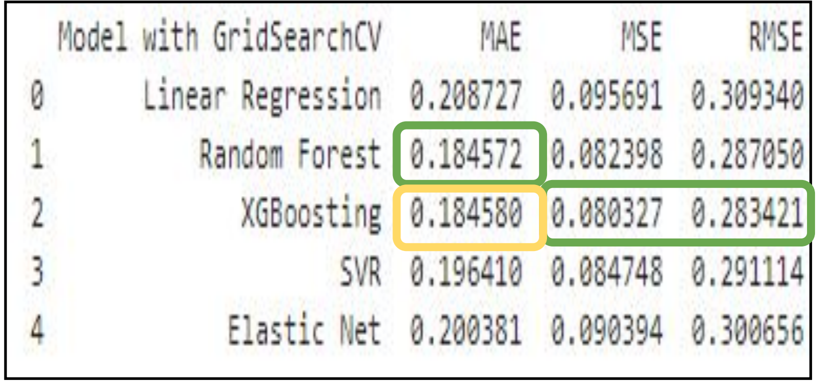
- Linear Regression
- Random Forest
- Extreme Gradient Boosting (xgboost)
- Support Vector Regression (SVR)
- Elastic Net

Original Model Metrics:



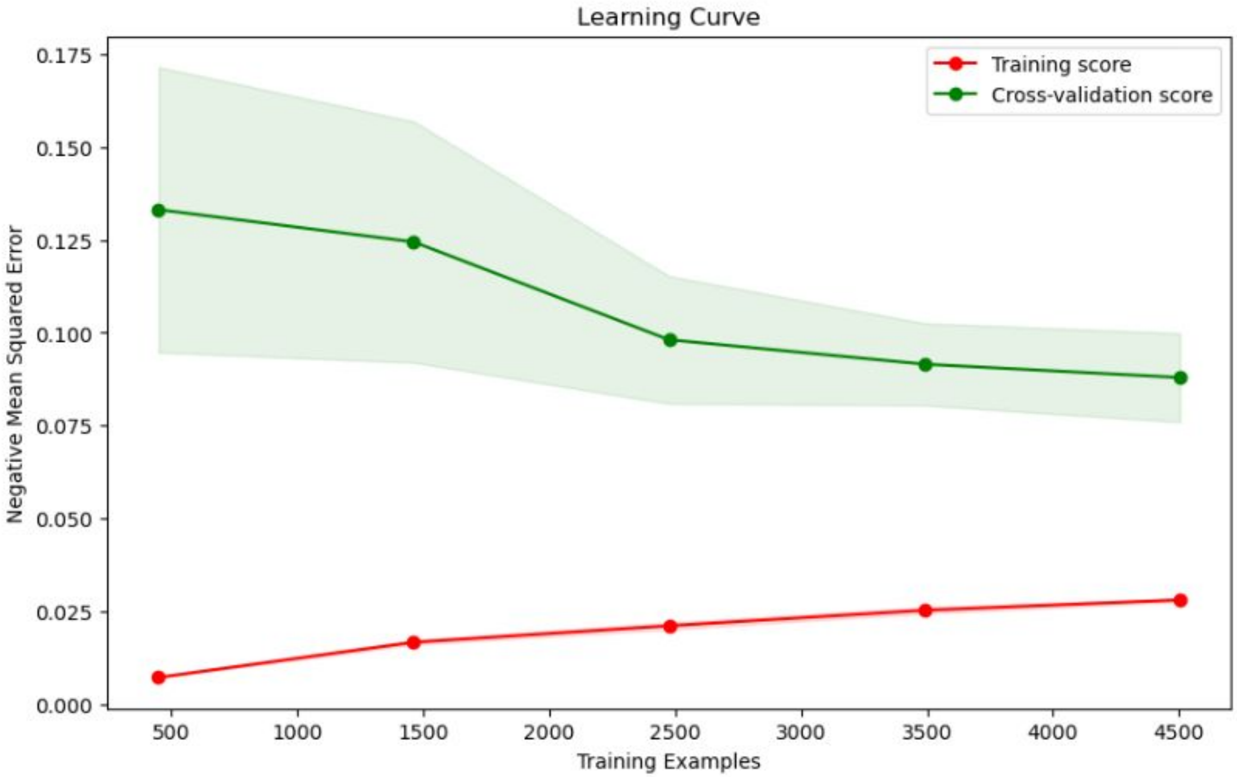
```
              Model           MAE           MSE          RMSE
0 Linear Regression  1.794951e+10  5.672739e+21  7.531759e+10
1     Random Forest  1.794049e-01  8.327751e-02  2.885784e-01
2        XGBoosting  1.920313e-01  8.294503e-02  2.880018e-01
3               SVR  1.828601e-01  8.128682e-02  2.851084e-01
4       Elastic Net  4.807005e-01  4.763833e-01  6.902053e-01
```

Model Metrics with GridSearchCV():



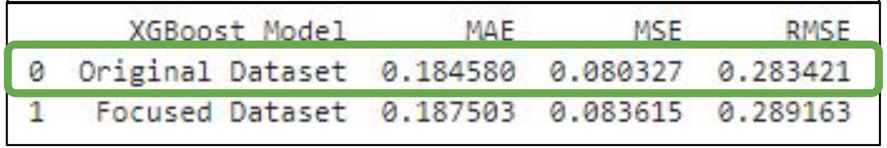
```
  Model with GridSearchCV       MAE       MSE      RMSE
0       Linear Regression  0.208727  0.095691  0.309340
1           Random Forest  0.184572  0.082398  0.287050
2              XGBoosting  0.184580  0.080327  0.283421
3                     SVR  0.196410  0.084748  0.291114
4             Elastic Net  0.200381  0.090394  0.300656
```

Visualizations: Learning Curve

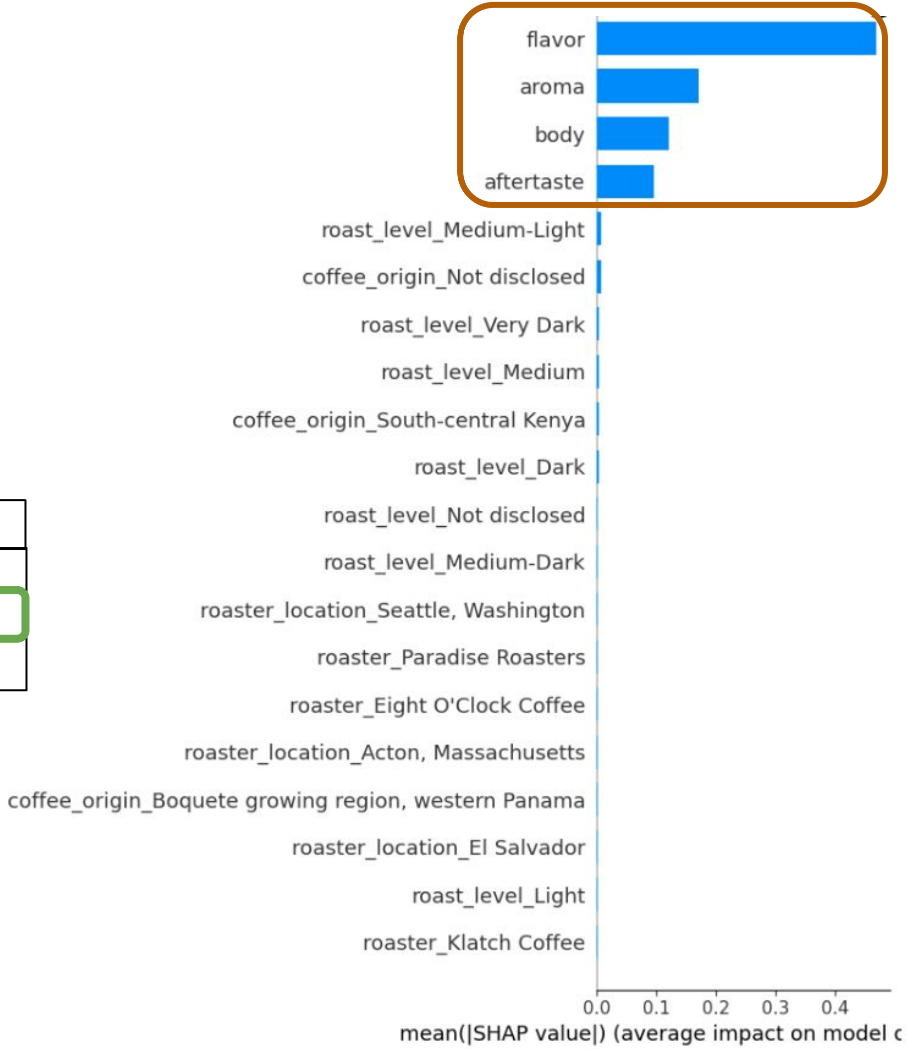


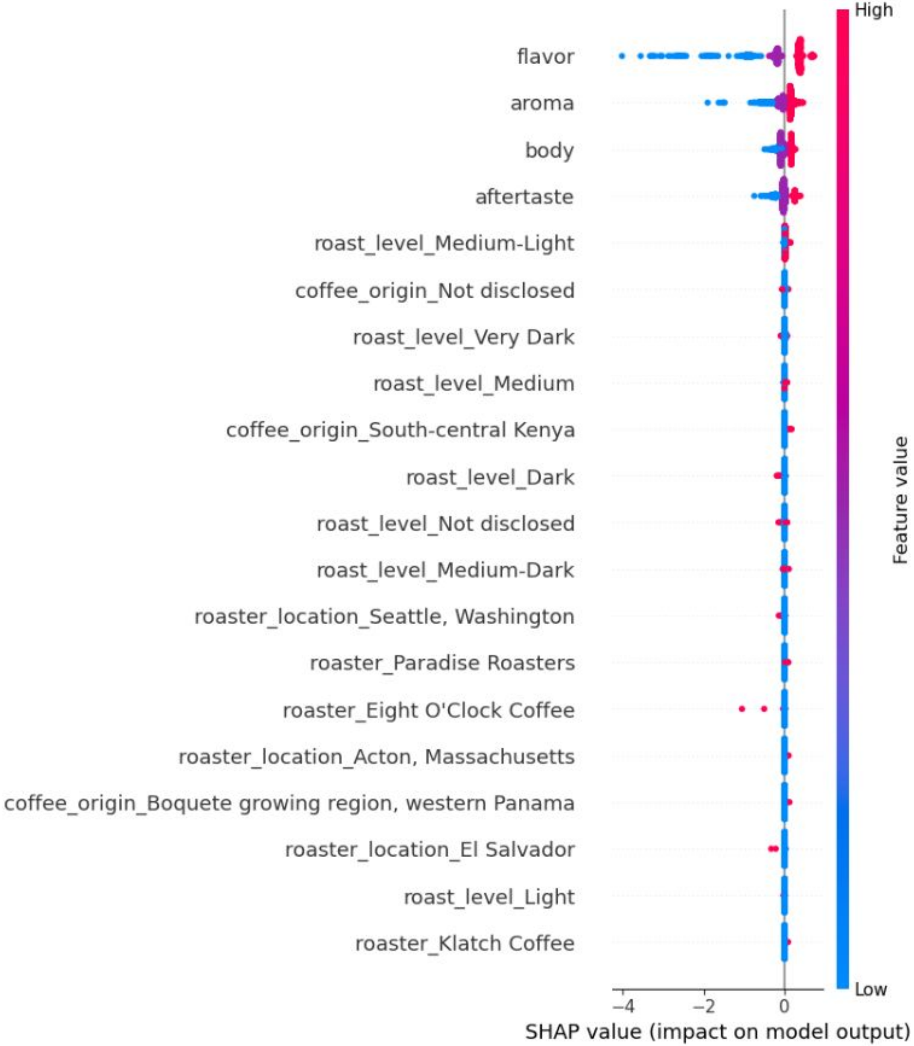
Visualization:
SHAP Summary Bar Plot

Original Data vs. Focused DataMetrics:



```
         XGBoost Model       MAE       MSE      RMSE
0  Original Dataset  0.184580  0.080327  0.283421
1   Focused Dataset  0.187503  0.083615  0.289163
```

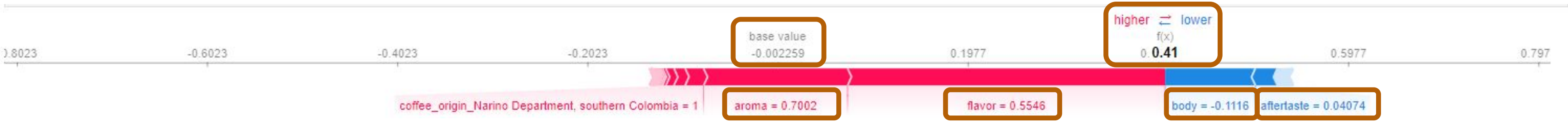




Visualizations: SHAP Summary Plot

Higher SHAP values have high positive impact on the rating for flavor, aroma, body, and aftertaste

Visualizations: SHAP Force Plot for Initial Review



- base value represents the predicted rating for all observations in the data set
- projected value represents the predicted rating for the initial review
- flavor and aroma are the two greatest components in the rating, both with a positive impact
- body and aftertaste are the next two greatest components in the rating, both with a negative impact

Next Steps:

Further Exploration of Key Features

Price Exploration

Top 4 features affecting rating: flavor, aroma, body, aftertaste

- descriptive info needed
- parse commentary features
- new data set recommended

Price Exploration

- est_price: $18.00/12 ounces (2.2% of feature)
- extensive data cleaning
- new data set recommended