Springboard Capstone Project Report: Coffee Shop

Starbucks is getting a run for its money. While they are still renowned for having a coffee shop on every corner, more coffee shop chains and local shops are opening. The competition is steep, and each chain or local shop needs something that sets them apart from the other competitors. My goal for this project was to identify the top coffee features a new coffee shop would need to consider when creating its signature blend to set it apart from Starbucks and other coffee shops.

As the project unfolded, I utilized the following python packages:

- numpy
- pandas
  - ProfileReport from pandas_profiling
- matplotlib.pyplot
- seaborn
- display from IPython.display
- os
- sklearn.preprocessing
  - StandardScaler
- sklearn.model_selection
  - train_test_split
  - GridSearchCV
- sklearn.linear_model
  - LinearRegression
  - ElasticNet
- sklearn.ensemble
  - RandomForestRegressor
- sklearn.svm
  - SVR
- sklearn.metrics
  - make_scorer
  - mean_squared_error
  - mean_absolute_error
- xgboost
- shap

Data Sourcing and Cleaning:

Data for this project was sourced from Kaggle. The data set comprises reviews collected on a website, rating coffees by features, and providing other descriptive information such as origin, roaster, and other notes. There were a total of 19 features and 7,041 entries. The original data set can be found here.

Viewing the first 5 rows of the dataset, I already noticed features with missing values. To better determine how to process missing values, I called a count of null values on each feature. The results are displayed to the left.

	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

The 'with_milk' feature was immediately dropped due to having nearly 86% null values. I considered dropping the 'agtron' feature since 257 of its entries were a simple '/' without any numeric features. With some research, I found that the agtron score is rating the coffee beans before and after grinding. Since this has a possible effect on the overall rating, I left this feature for now.

According to the display of the first 5 rows, 'rating' was based on a 1-100 scale, while 'acidity_structure', 'aftertaste', 'aroma', 'body', and 'flavor' were all based on a 1-10 scale. Whichever range, all these categories should be numeric, but were currently labeled as objects. I changed these categories to numeric categories and confirmed every category was either numeric for a rating or an object for a descriptive feature.

EDA:

To more extensively explore the data, I ran a ProfileReport. The report gave 28 alerts regarding features with high cardinality, features with high correlation, features with missing values, uniformly distributed features, and features with completely unique entries. The report also gives a summary of each feature, and visualizations, as well as notes features that are unique, common, rare, most frequent, and least frequent. I used this report to determine which features needed further cleaning.

The report showed several features that ended up needing to be dropped from the unusable data set. Several of these features were commentary rather than descriptive of the coffee itself or had high percentages of null or unusable data.

- Dropped features:
  - acidity_structure *(high correlation with the rating category, but about 62% null values)*
  - agtron *(high cardinality, empty rating frames, and alphabetic entries resulting in unusable data)*
  - est_price *($18.00/12 ounces was the most common entry, but only made up 2.2% of the data. This category may be useful for product pricing in a secondary project.)*
- Commentary features:
  - title
  - blind_assessment
  - bottom_line
  - notes
  - review_date
  - url

The following features had null values replaced. Numeric features with null values were replaced by the feature's rounded mean, and object features had null values and similarly formatted entries were replaced by the frequently used "Not disclosed". The features with replaced values are listed below:

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

Pre-Processing and Training:

Since the object features had such high cardinality, I created dummy variables for each object category. I also standardized the magnitude of the numeric features using StandardScaler. The fit and transformed scaled data and dummy variables were concatenated into the data frame, and the original features were dropped, then the data set was made into a pandas data frame. I split the data into X and y subsets with X being all features except for rating, and y being only the target feature, rating. The data subsets were entered into the train_test_split function at 80% training and 20% testing portions.

Modeling:

I ran five models on the data set: Linear Regression, Random Forest, Extreme Gradient Boosting (xgboost), Support Vector Regression (SVR), and Elastic Net. I ran these models a second time with GridSearchCV() to ensure I had optimized the hyperparameters for each model. The results are displayed below.

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

GridSearchCV() Model Matrics:

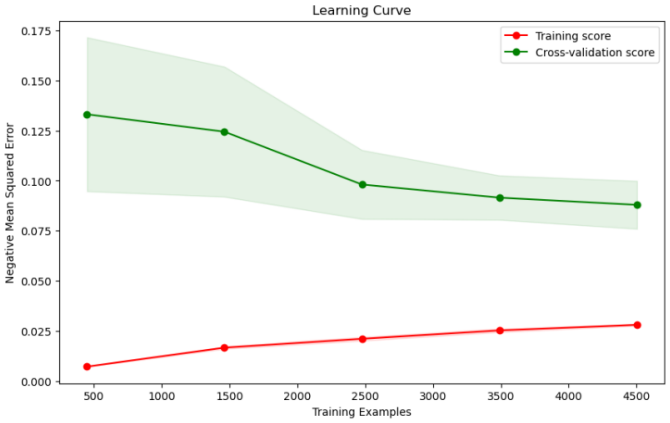
	Model with GridSearchCV	MAE	MSE	RMSE
0	Linear Regression	0.208727	0.095691	0.309340
1	Random Forest	0.184572	0.082398	0.287050
2	XGBoosting	0.184580	0.080327	0.283421
3	SVR	0.196410	0.084748	0.291114
4	Elastic Net	0.200381	0.090394	0.300656

I visualized a Learning Curve graph to visualize the accuracy of the train and test fit and used SHAP to visualize a force plot and summary plots. The SHAP visuals indicated that flavor, aroma, body, and aftertaste were the four primary feature contributors to the target rating. The other features had miniscule effect on the rating. I decided to see if I could gain better results by dropping all but the four primary features and the rating, then again running xgboost with GridSearchCV() to see if I could get better metrics on a more focused data set compared to the best original model. The comparison table of MAE, MSE, and RMSE metrics of the original and reduced data sets using xgboost modeling with GridSearchCV is displayed below. The Learning Curves and SHAP visualizations are on the following pages.

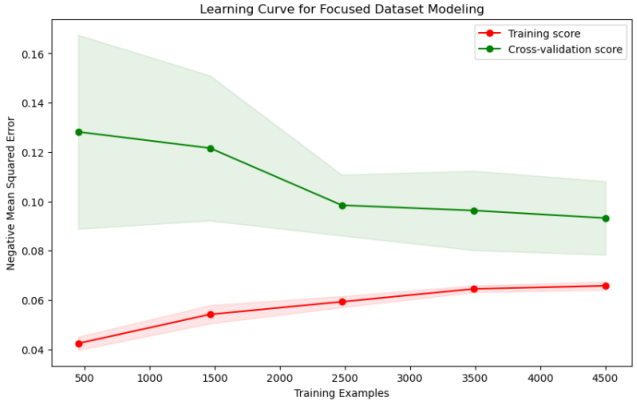
Original vs. Focused Data Sets Model Metrics:

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

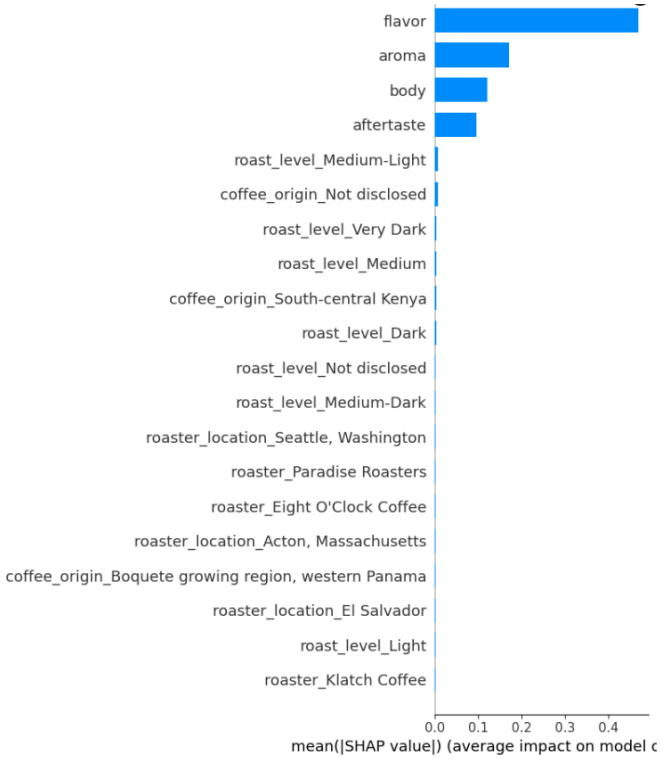
Orginal Data Set Learning Curve:



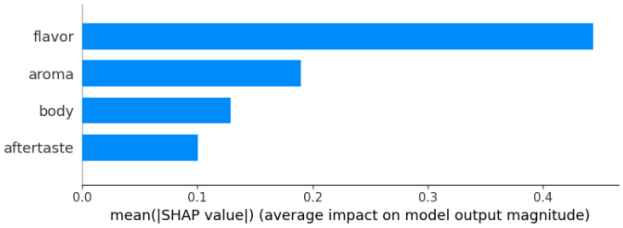
Focused Data Set Learning Curve:



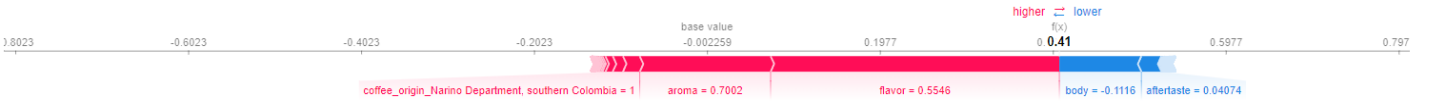
Original Data Set SHAP Summary Bar Graph:



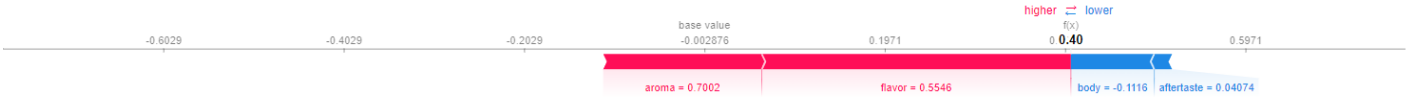
Focused Data Set SHAP Summary Bar Graph:



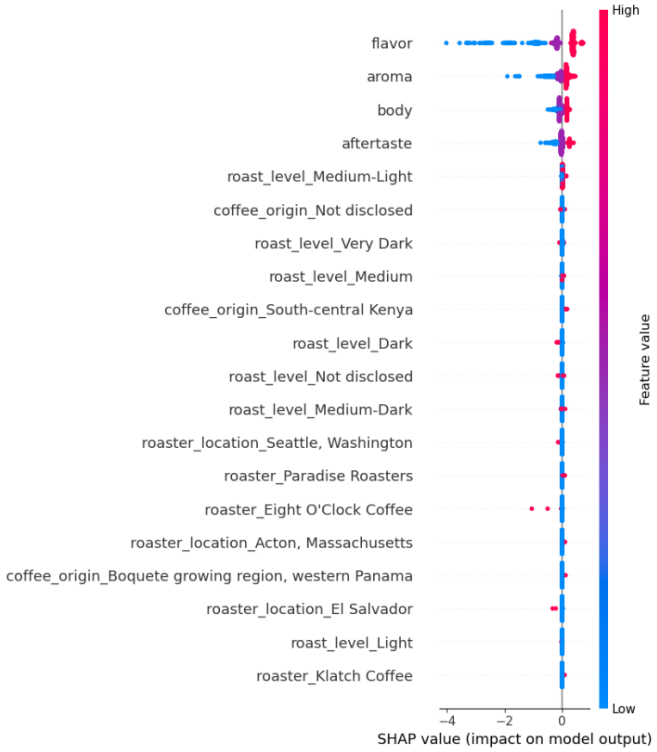
Original Data Set SHAP Force Plot:



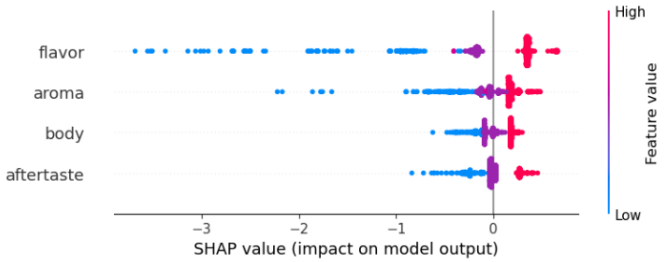
Focused Data Set SHAP Force Plot:



Original Data Set SHAP Summary Graph:



Original Data Set SHAP Summary Graph:



Next Steps:

The modeling revealed what the top four features affecting the rating are flavor, aroma, body, and aftertaste. However, in the original data set, these features are numeric categories, not objects that describe the flavor, aroma, body, or aftertaste. Now that the important features are known, the next project should determine specifically what kind of flavor, aroma, body, and aftertaste result in high ratings. The commentary section of this data set may be able to be parsed into the needed descriptive features, or a new data set may need to be found.

Similarly, a new project could be run to determine what quantity the coffee should be sold in, and what the price point should be. The est_price feature in this data set had extremely high cardinality and included worldwide pricing units. Standardizing the units and narrowing down the entries to the United States or the local market may be sufficient. Otherwise, more data will need to be identified.