Chocolate Bar Rating Analysis with Python

By: Daniel Abebe

January 4, 2021

Content

1) Project Overview
2) Understanding the Dataset
   - Data Wrangling
   - Exploratory Data Analysis
3) ML Modeling and Model Performance
4) ML Interpretation

Project overview

Project Background
Dataset Overview

Project Overview

Why should you care?

- Globally nine out of ten people love chocolate
- More than 2.8 billion pounds/year consumed in US

Stakeholders:

- Chocolate Manufacturing Companies
- Chocolate lovers

Goal:

- Classify chocolate bar based on rating

Objective:

- Help chocolate manufacturing companies to produce a quality chocolate

Problem Statement:

- What qualities make for a highly rated chocolate bar?

Dataset Overview

Dataset is compiled by Brady Brelinski (Manhattan Chocolate Society)

It contains **20** input features:

- 15 categorical & 5 numerical features
- No of records = 2224 rows

```
# head() function allows us to see the top N amount of records (in this case 5 records) of data frame
df.head()
```

	Unnamed: 0	ref	company	company_location	review_date	country_of_bean_origin	specific_bean_origin_or_bar_name	cocoa_percent	rating	counts_of_ingredients	beans	cocoa_butter
0	0	2454	5150	U.S.A	2019	Madagascar	Bejofo Estate, batch 1	76.0	3.75	3	have_bean	have_cocoa_butter
1	1	2458	5150	U.S.A	2019	Dominican republic	Zorzal, batch 1	76.0	3.50	3	have_bean	have_cocoa_butter
2	2	2454	5150	U.S.A	2019	Tanzania	Kokoa Kamili, batch 1	76.0	3.25	3	have_bean	have_cocoa_butter
3	3	797	A. Morin	France	2012	Peru	Peru	63.0	3.75	4	have_bean	have_cocoa_butter
4	4	797	A. Morin	France	2012	Bolivia	Bolivia	70.0	3.50	4	have_bean	have_cocoa_butter

vanilla	lecithin	salt	sugar	sweetener_without_sugar	first_taste	second_taste	third_taste	fourth_taste
have_not_vanila	have_not_lecithin	have_not_salt	have_sugar	have_not_sweetener_without_sugar	cocoa	blackberry	full body	NaN
have_not_vanila	have_not_lecithin	have_not_salt	have_sugar	have_not_sweetener_without_sugar	cocoa	vegetal	savory	NaN
have_not_vanila	have_not_lecithin	have_not_salt	have_sugar	have_not_sweetener_without_sugar	rich cocoa	fatty	bready	NaN
have_not_vanila	have_lecithin	have_not_salt	have_sugar	have_not_sweetener_without_sugar	fruity	melon	roasty	NaN
have_not_vanila	have_lecithin	have_not_salt	have_sugar	have_not_sweetener_without_sugar	vegetal	nutty	NaN	NaN

Understanding the Dataset

Data Wrangling
Exploratory Data Analysis

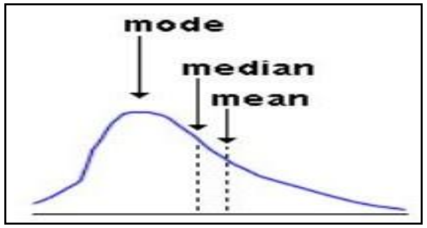
Data Wrangling

Raw Data Loading

↓

Missing Value Replacement

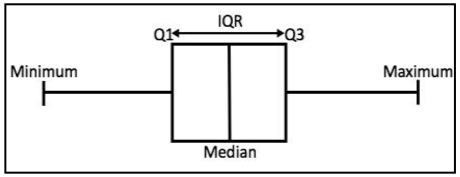
- Feature Drop
- Feature Imputation



↓

Handling Outliers

Interquartile Range:
LB = Q1 – 1.5*IQR
UB = Q3 + 1.5*IQR



↓

Resampling

- Up-sampling
- Down-sampling

Low-rating (50%)	High-rating (50%)

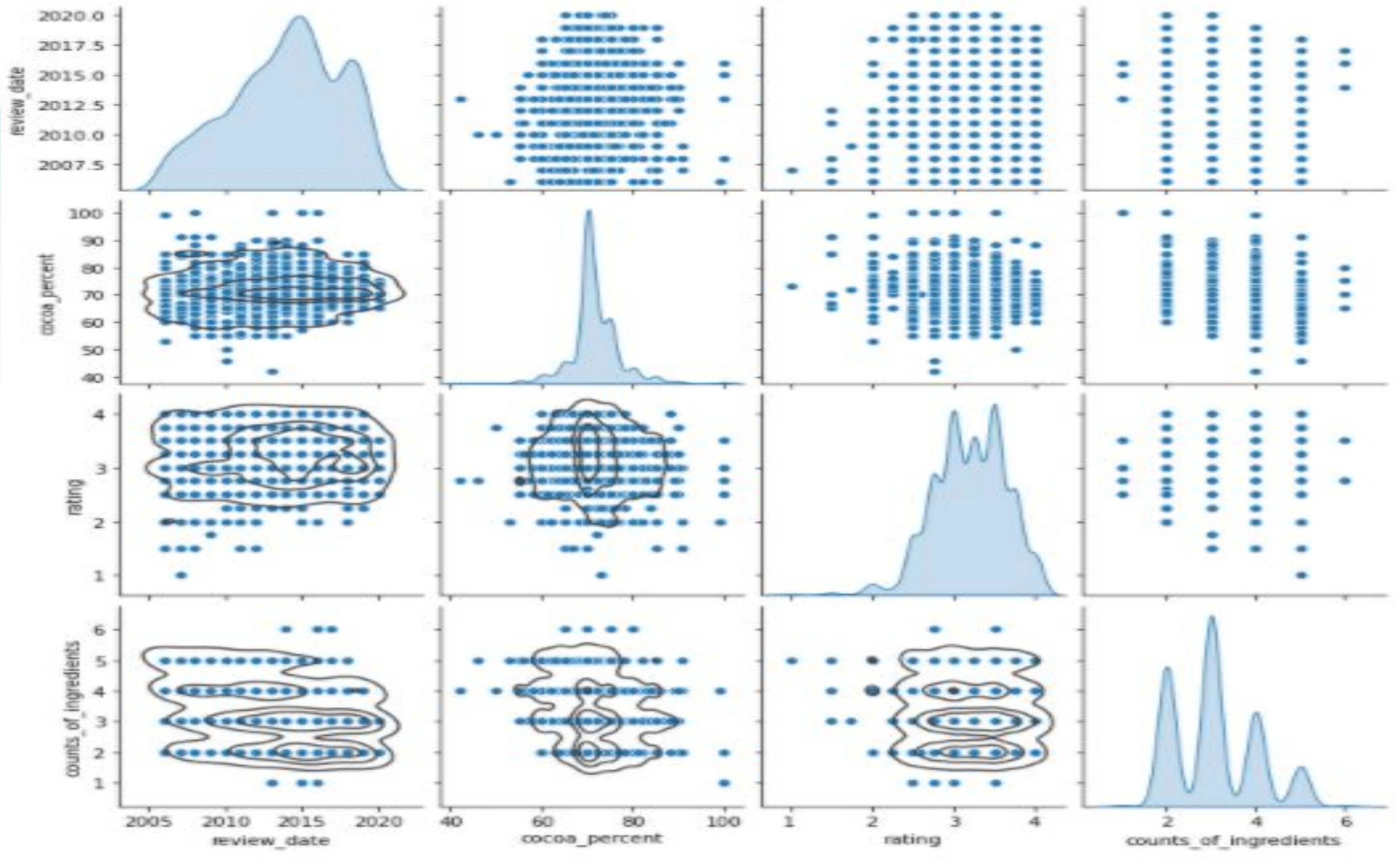
↓

Feature Transform

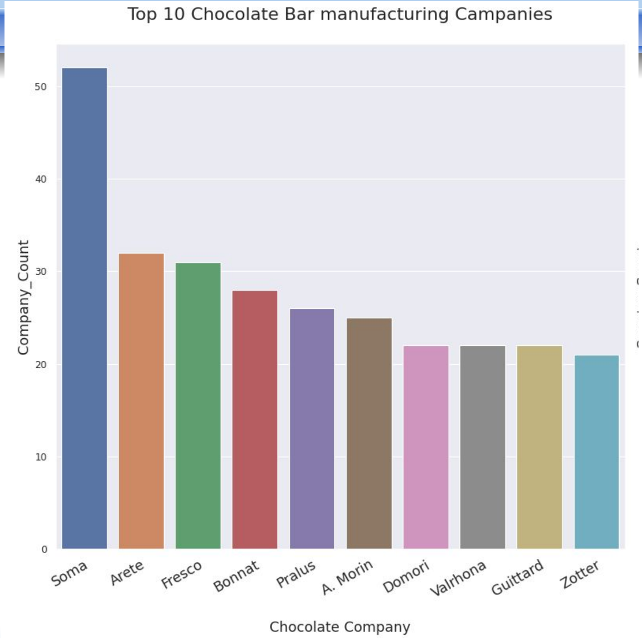
- Feature Encoding
- Feature Engineering
- Feature Scaling

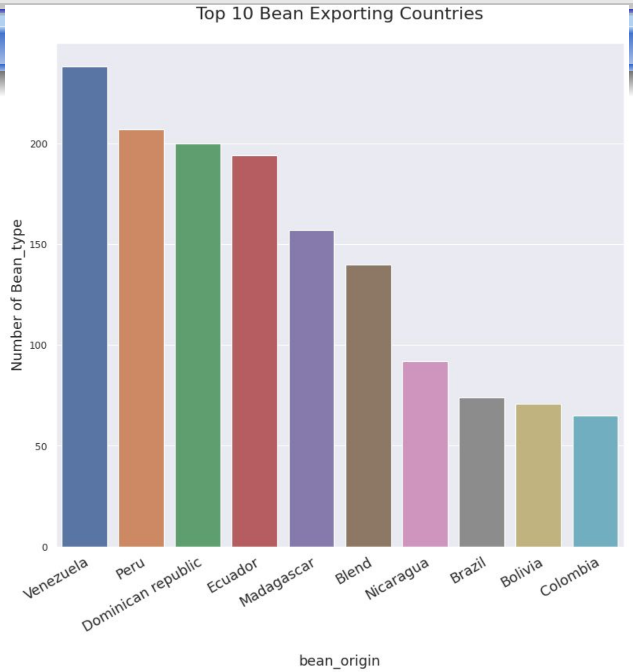
$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

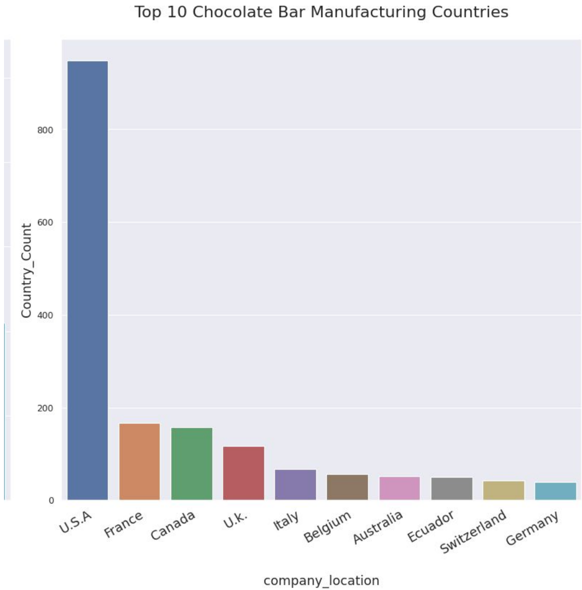
Exploratory Data Analysis



Exploratory Data Analysis



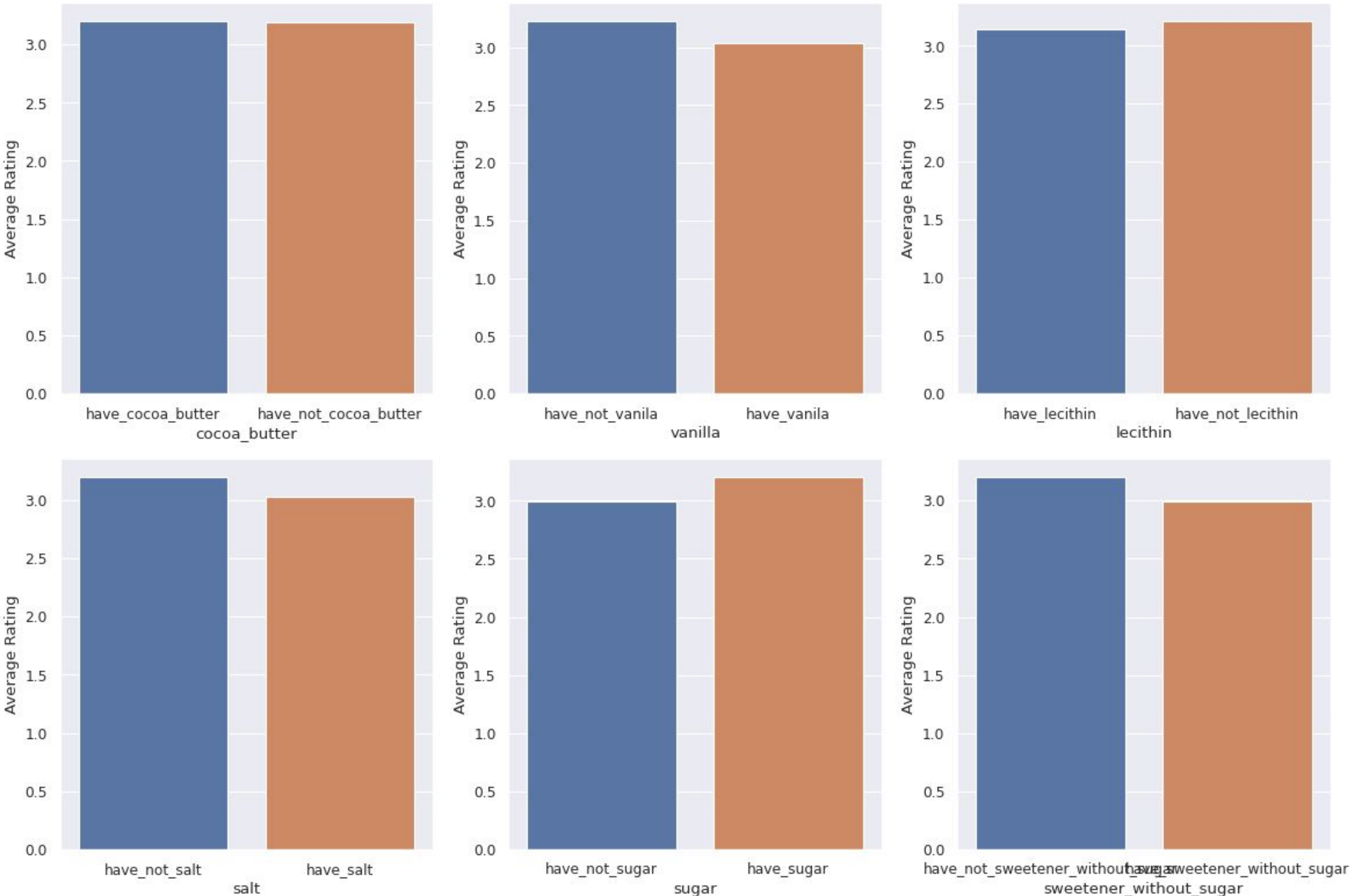




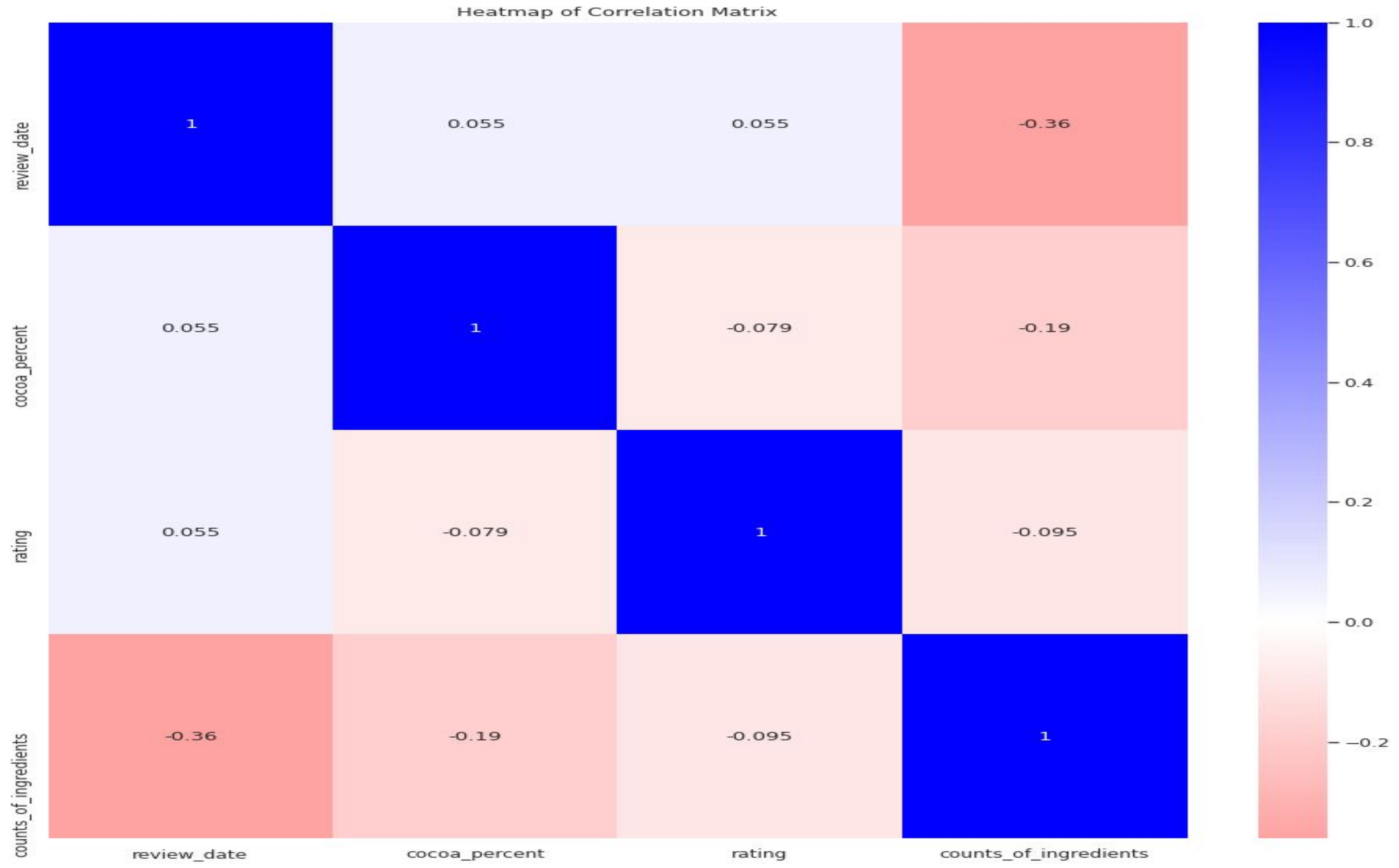
Rating of chocolate bar over time

Ave. cocoa % and Ave. rating vs over time

Effect of some categorical feature on rating



Correlation Matrix



Summary of EDA

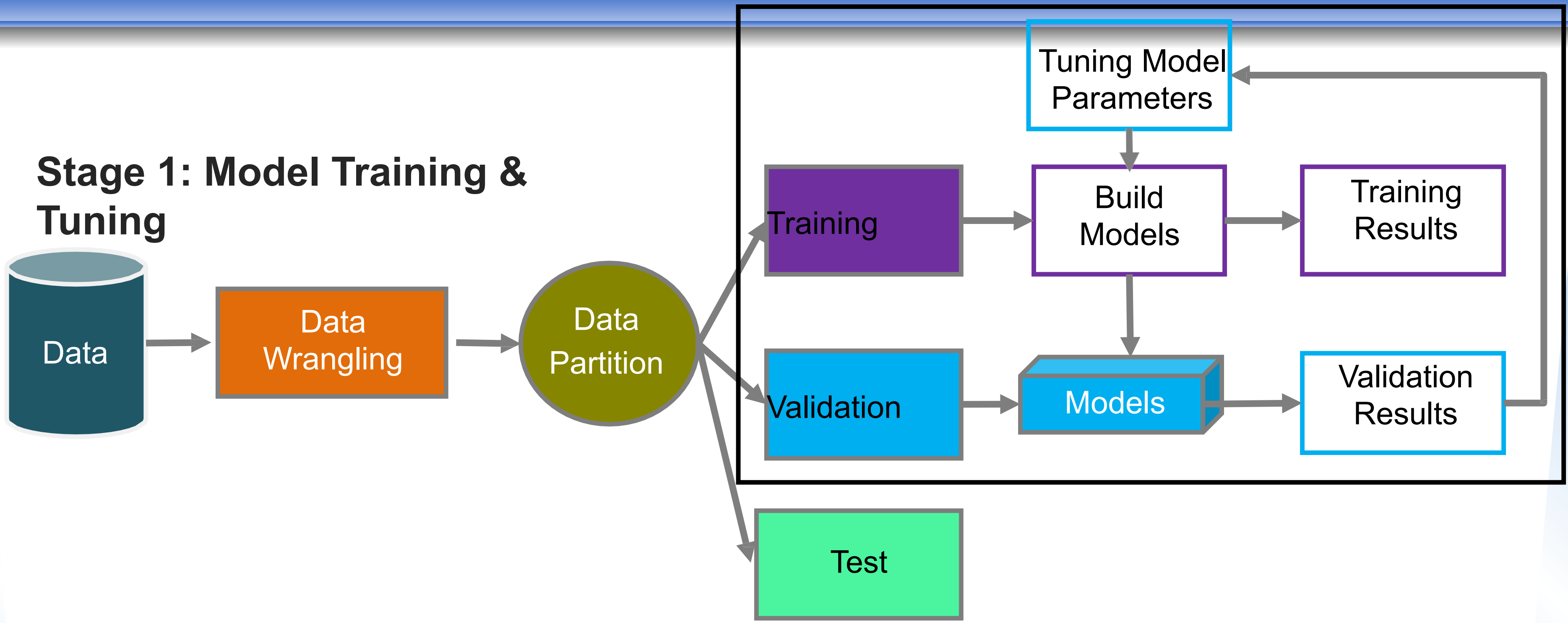
- The cocoa percentage of high rated chocolate bar is between 69%-72% and the best is for cocoa percentage of 71.5%.
- Bean from Venezuela and Peru are the best source of beans
- 2 and 3 types of ingredients are preferable amount ingredients
- The rating companies generally improved through time.
- Soma is the best chocolate bar manufacturing company and it is based in USA

ML Modeling and Model Performance

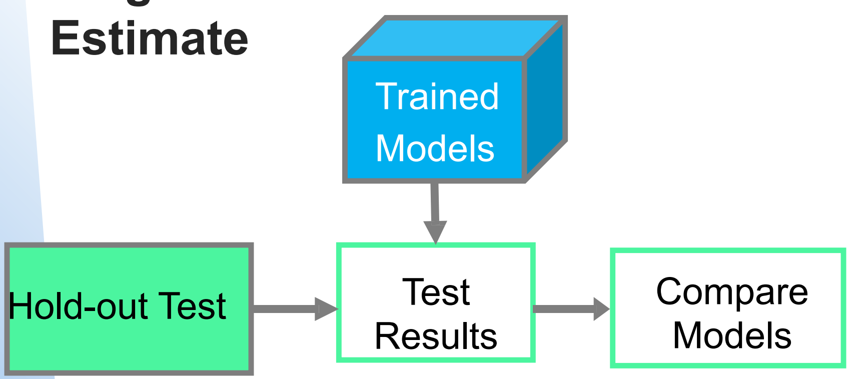
- SVM, LGBM
- SHAP, LIME

Model Workflow

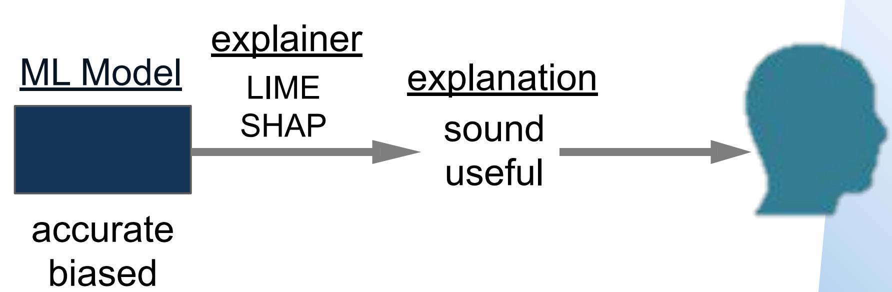
Stage 1: Model Training & Tuning



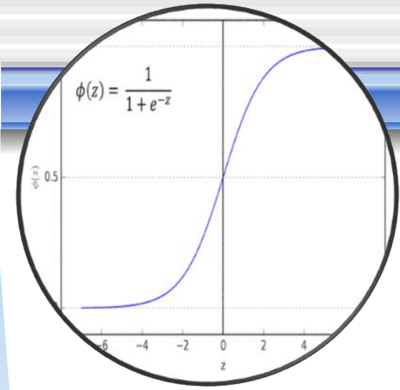
Stage 2: Model Performance Estimate



Stage 3: Model Interpretation



Model Selections

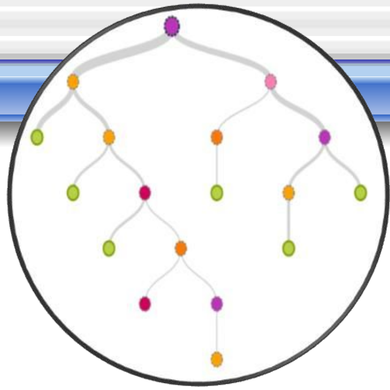


Logistic Regression

Sigmoid logit function:
log(p/(1-p))

Transforms:
Input values → estimate into prob. range (0, 1)

Works well on linearly separable classes.

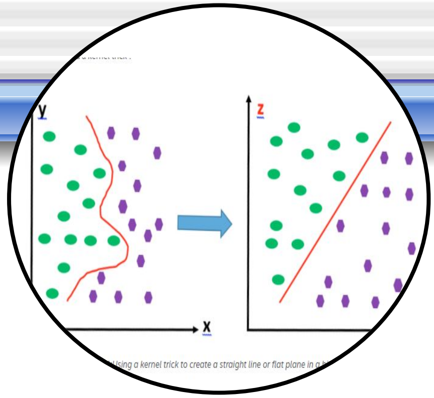


Decision Tree

Split data on features.

Repetitive splitting procedure.

Continue splitting until each node left with same class label.

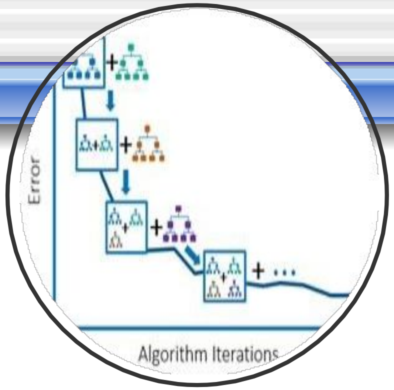


SVM

Creates hyperplane that separate classes.

raises the data to a higher dimensional space.

create a straight line or a flat plane in a higher dimension.



Gradient Boost

Sequential training.

Learn from residual errors.

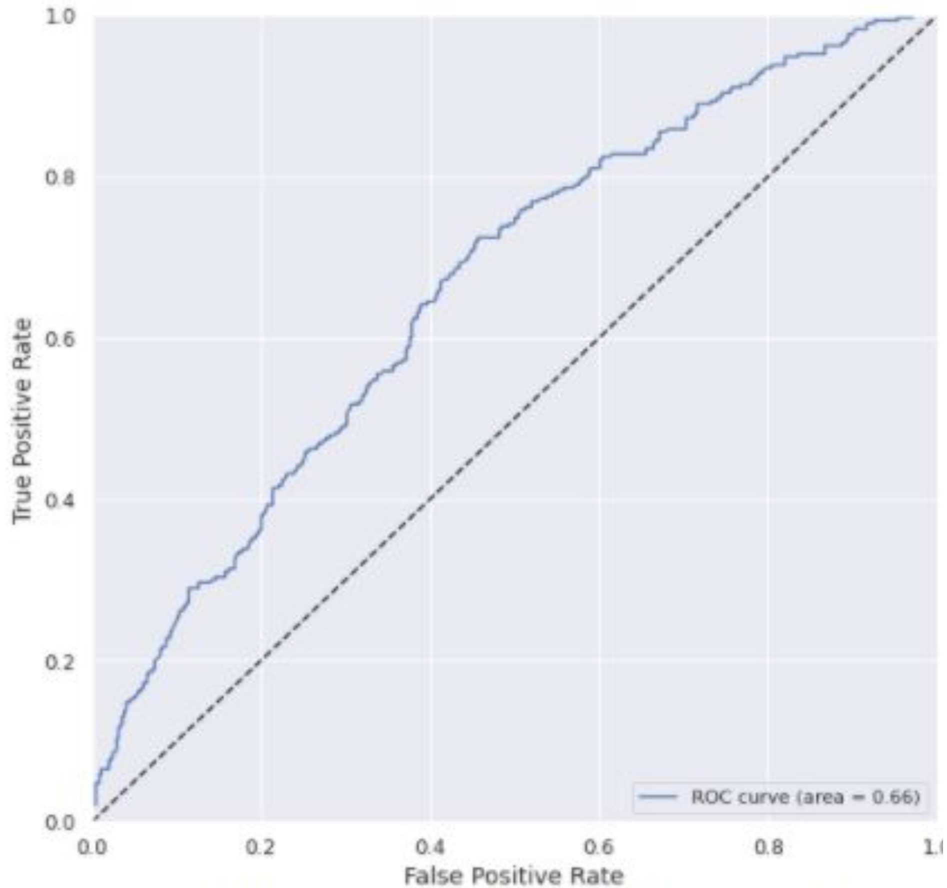
Step-wise forward

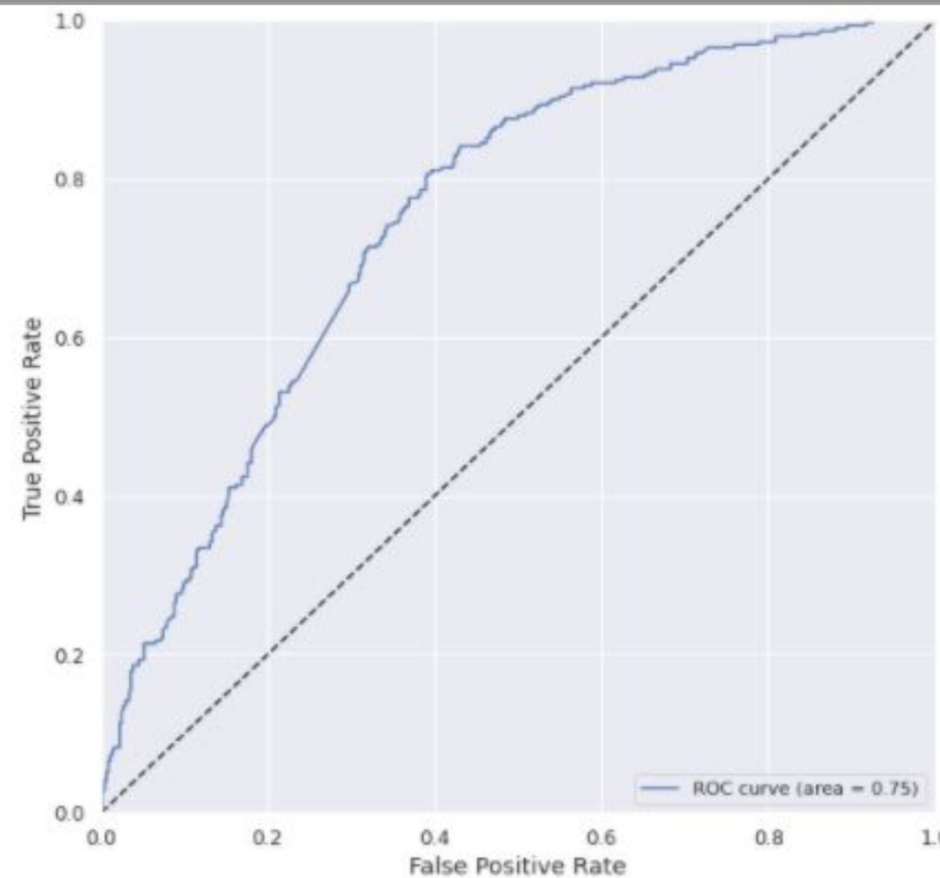
$$\boxed{\text{Label} = \text{mode}\ \{c_{lr}(x), c_{dt}(x), c_{svm}(x), c_{lgb}(x)\}}$$

Majority Vote
Meta-classifier
Combination of all models
Improves accuracy of model performances by majority vote

In this project we used SVM and Light Gradient Boost Model

ML Performance Evaluation (SVM vs LGBM Model)





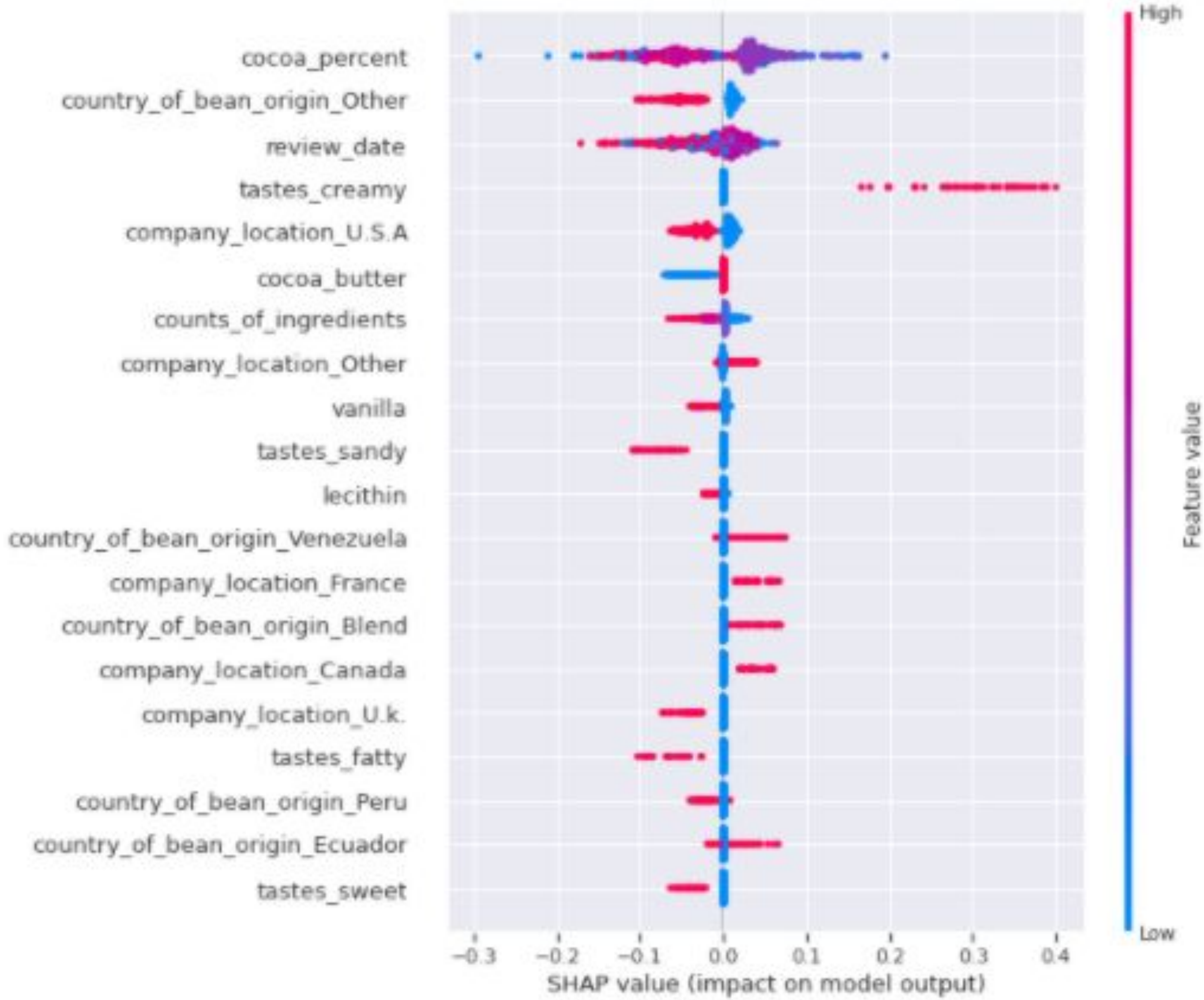
Accuracy train: 0.6537 Accuracy test: 0.6417
Precision test: 0.6216 Recall test: 0.2379
ROC-AUC_test: 0.6632 F1_test: 0.3441

Accuracy train: 0.7987 Accuracy test: 0.6771
Precision test: 0.6147 Recall test: 0.4897
ROC-AUC_test: 0.7506 F1_test: 0.5451

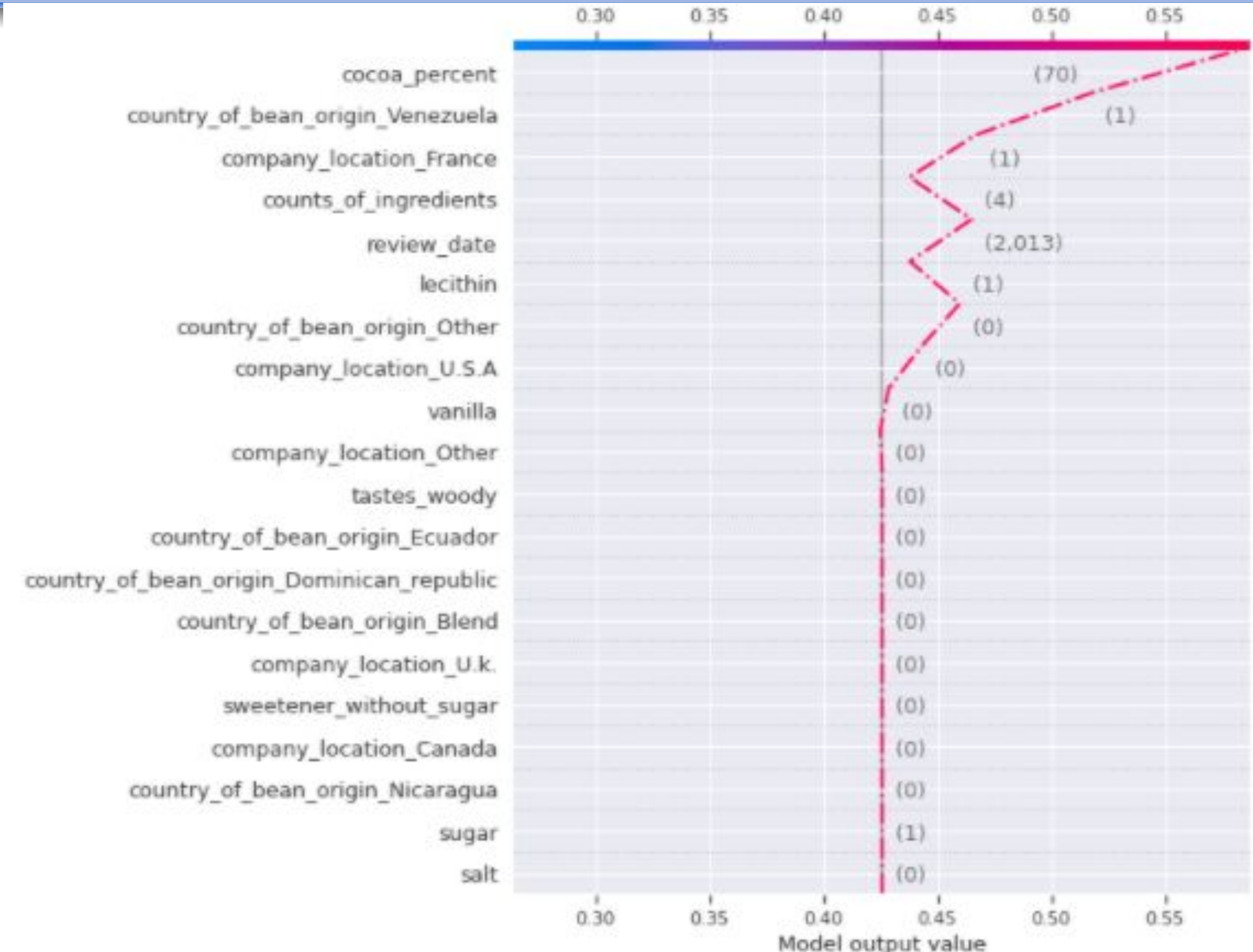
ML Model Interpretation

SHAP
LIME

Global Interpretation with SHAP

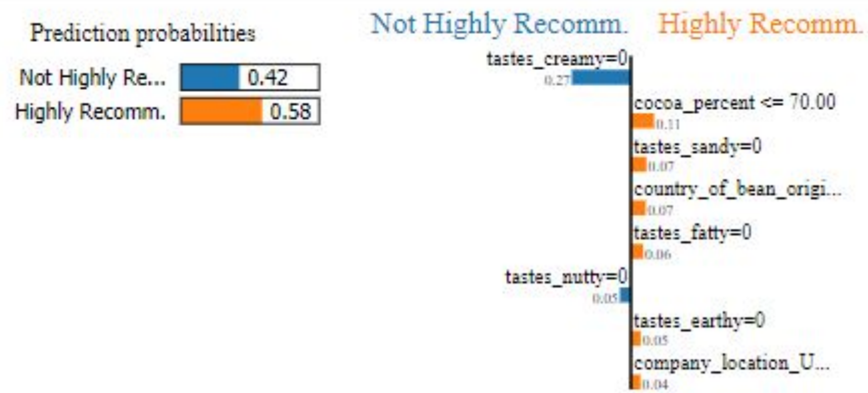


Isolate a Single Decision with a SHAP Decision Plot (Local interpretation)



Explain "Instances of Interest" with LIME Tabular Explainer

```
lime_svm_explainer.explain_instance(X_test[X_test.index==5].values[0],\
                                    fitted_svm_mdl.predict_proba,\
                                    num_features=8).\
                                show_in_notebook(predict_proba=True)
```



Feature	Value
tastes_creamy=0	True
cocoa_percent	70.00
tastes_sandy=0	True
country_of_bean_origin_Other=0	True
tastes_fatty=0	True
tastes_nutty=0	True
tastes_earthy=0	True
company_location_U.k.=0	True

```
#same as before but with all 5's replaced by 24
lime_svm_explainer.explain_instance(X_test[X_test.index==24].values[0],\
                                    fitted_svm_mdl.predict_proba,\
                                    num_features=8).\
                                show_in_notebook(predict_proba=True)
```

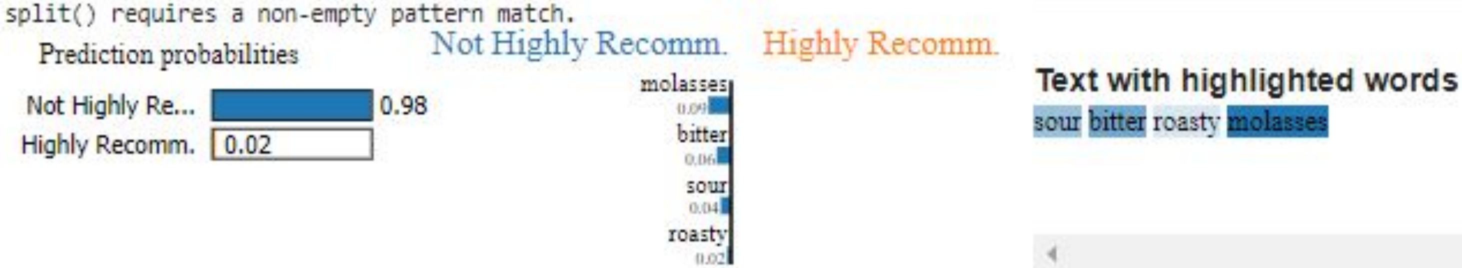
Feature	Value
tastes_creamy=0	True
cocoa_percent	70.00
tastes_fatty=0	True
tastes_rich=0	True
tastes_earthy=0	True
country_of_bean_origin_Other=1	True
tastes_sandy=0	True
tastes_spice=0	True

LIME NLP Explainer

```
lime_lgb_explainer.explain_instance('creamy rich complex fruity',\
                                    lgb_pipeline.predict_proba, num_features=4).\
                                    show_in_notebook(text=True)
lime_lgb_explainer.explain_instance('sour bitter roasty molasses',\
                                    lgb_pipeline.predict_proba, num_features=4).\
                                    show_in_notebook(text=True)
lime_lgb_explainer.explain_instance('nasty disgusting gross stuff',\
                                    lgb_pipeline.predict_proba, num_features=4).\
                                    show_in_notebook(text=True)
```

```
split() requires a non-empty pattern match.
```


```
split() requires a non-empty pattern match.
```




```
split() requires a non-empty pattern match.
```

Summary

Goal

Predict chocolate with bad and good rating

Results

- Model was able to predict whether the chocolate rating
- 66% accuracy on tabular data (SVM model) and 77% of accurate predictions NLP data (LGBM)

General Findings

- The amount of Cocoa and the flavor of chocolate has a significant effect on the rating (chocolate bars with 71.5% of cocoa has high rating)
- Bean origin Venezuela and Peru are good for chocolate bars
- 2 or 3 types number of ingredients in the chocolate are best
- Fruity, complex, creamy tastes are preferable where as molasses, bitter and sour tastes are not preferable

Next Steps

- Model improvement: algorithms, resampling and designs