Cyber Barista

- prediction of arabica beans quality based on machine learning model

Jiayu Zheng
DSI student at Brown University
2022-10-18

link2repo: https://github.com/BubbleJoe-BrownU/hands-on-machine-learning-project

Workflow

01 Intro & Background

02 Exploratory Data Analysis

03 Splitting & Preprocessing

01

Intro & Background

motivation & background

How do you decide whether your choice of coffee beans is gonna give you a splendid experience?

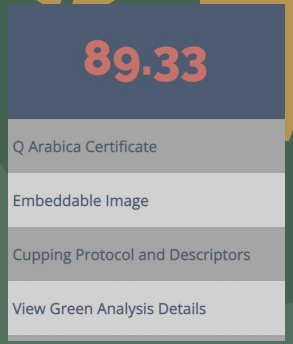
Processing method?
Country of origin?
Color of beans?
Scent of beans?

No?

What about let a seasoned barista help you ?

Or, a more fanciful way, a cyber barista!

COFFEE QUALITY INSTITUTE®



(1311, 43)

	Species	Owner	Country.of.Origin	Farm.Name	Lot.Number	Mill	ICO.Number	Company	Altitude	Region	...	Color	Category.Two.Defects
0	Arabica	metad plc	Ethiopia	metad plc	NaN	metad plc	2014/2015	metad agricultural developmet plc	1950-2200	guji-hambela	...	Green	0
1	Arabica	metad plc	Ethiopia	metad plc	NaN	metad plc	2014/2015	metad agricultural developmet plc	1950-2200	guji-hambela	...	Green	1
2	Arabica	grounds for health admin	Guatemala	san marcos barrancas "san cristobal cuch	NaN	NaN	NaN	NaN	1600 - 1800 m	NaN	...	NaN	0
3	Arabica	yidnekachew dabessa	Ethiopia	yidnekachew dabessa coffee plantation	NaN	wolensu	NaN	yidnekachew debessa coffee plantation	1800-2200	oromia	...	Green	2
4	Arabica	metad plc	Ethiopia	metad plc	NaN	metad plc	2014/2015	metad agricultural developmet plc	1950-2200	guji-hambela	...	Green	2

https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi

dataset description

Features from sample information:

- Owner
- Country of Origin
- Farm Name
- Lot Number
- Mill
- ICO Number
- Company
- Altitude
- unit_of_measurement
- altitude_low_meters
- altitude_high_meters
- altitude_mean_meters
- Region
- Producer
- Number of Bags
- Bag Weight
- In Country Partner
- Harvest Year
- Grading Date
- Owner_1
- Variety
- Processing Method

Features from cupping scores:

- Aroma
- Flavor
- Aftertaste
- Acidity
- Body
- Balance
- Uniformity
- Cup Cleanliness
- Sweetness
- Cupper Points
- Total Cup Points

Q CERTIFIED ARABICA™

Q Arabica Certificate | Sample #261473

Sample form Received a Q Arabica Certificate.

SAMPLE INFORMATION

Country of Origin	Colombia	Number of Bags	1
Farm Name	Finca El Paraiso	Bag Weight	35 kg
Lot Number	CQU2022015	In-Country Partner	Japan Coffee Exchange
Mill	Finca El Paraiso	Harvest Year	2021 / 2022
ICO Num			
Comp			
Altit			
Reg			
Produ			

CUPPING SCORES

Aroma	8.58	Uniformity	10.00
Flavor	8.50	Clean Cup	10.00
Aftertaste	8.42	Sweetness	10.00
Acidity	8.58	Overall	8.58
Body	8.25	Defects	0.00
Balance	8.42	Total Cup Points	89.33

89.33

Q Arabica Certificate

Embeddable Image

Cupping Protocol and Descriptors

View Green Analysis Details

dataset description

Features from Green Analysis:

- Moisture
- Category One Defects
- Quakers
- Color
- Category Two Defects

Features from Certification Information:

- Expiration
- Certification Body
- Certification Address
- Certification Contact

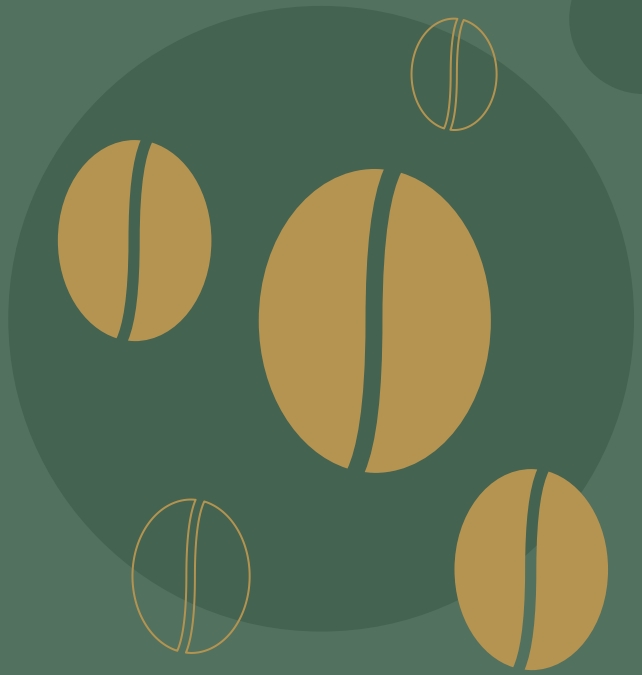
GREEN ANALYSIS

Moisture	11.8 %	Color	Green
Category One Defects	0 full defects	Category Two Defects	3 full defects
Quakers	0		

CERTIFICATION INFORMATION

Expiration	September 21st, 2023
Certification Body	Japan Coffee Exchange
Certification Address	〒413-0002 静岡県熱海市伊豆山１１７３－５８ ...
Certification Contact	松澤　宏樹　Koju Matsuzawa - +81(0)9085642901

Purpose of the project



data

preprocessing

data for training

training

model

regression

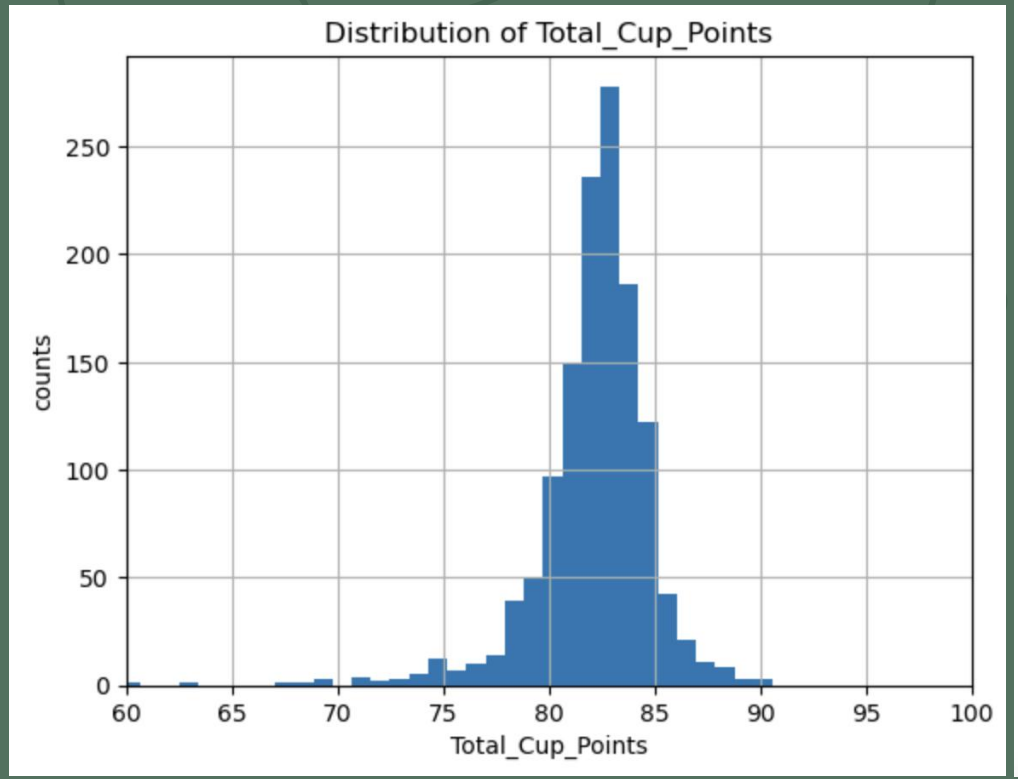
bean info

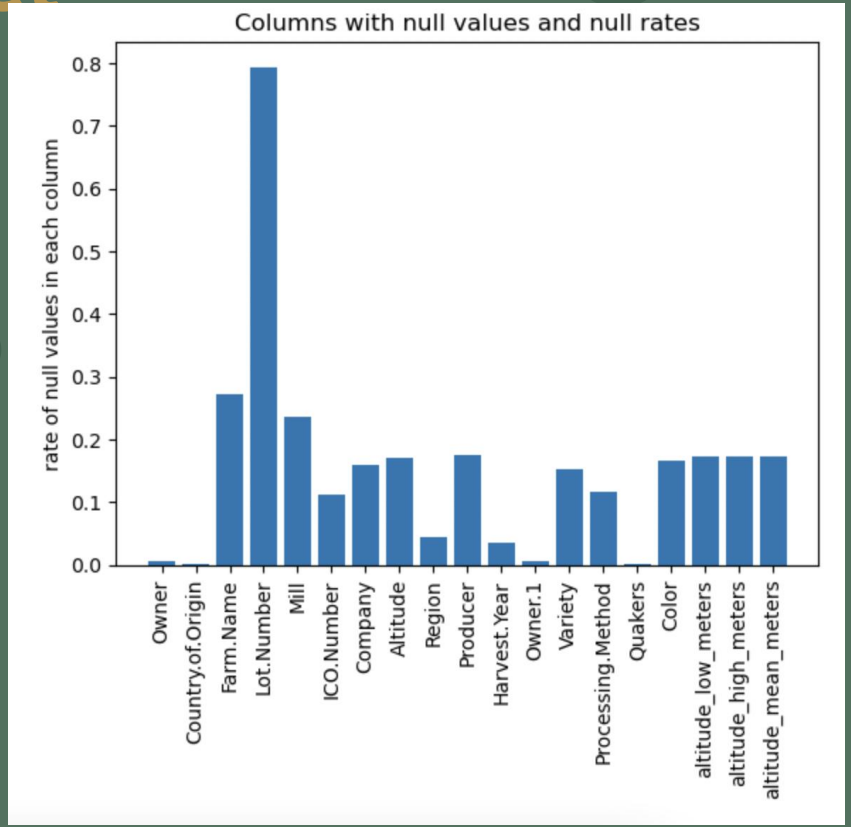
scores

02

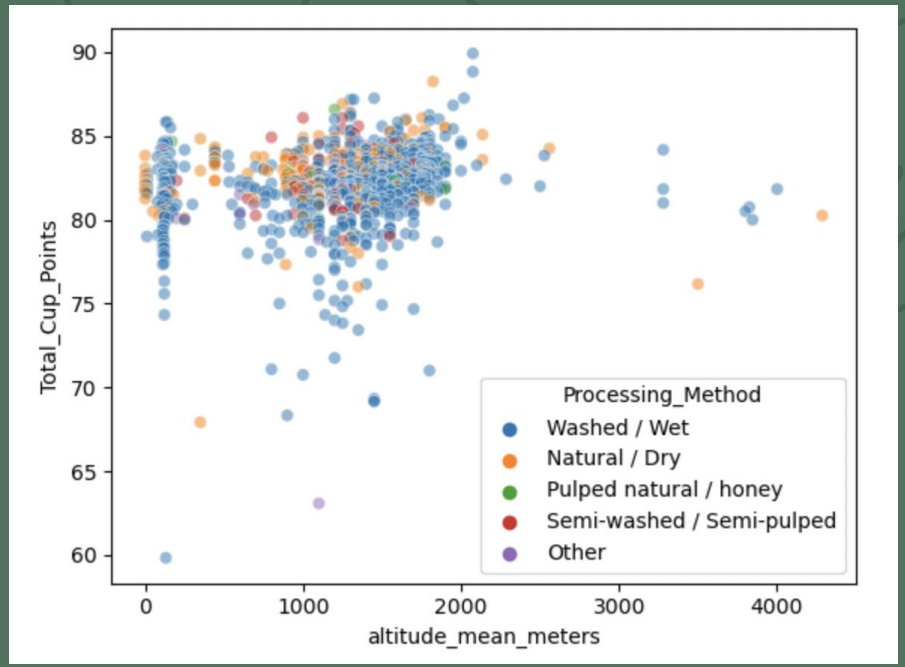
Exploratory Data Analysis

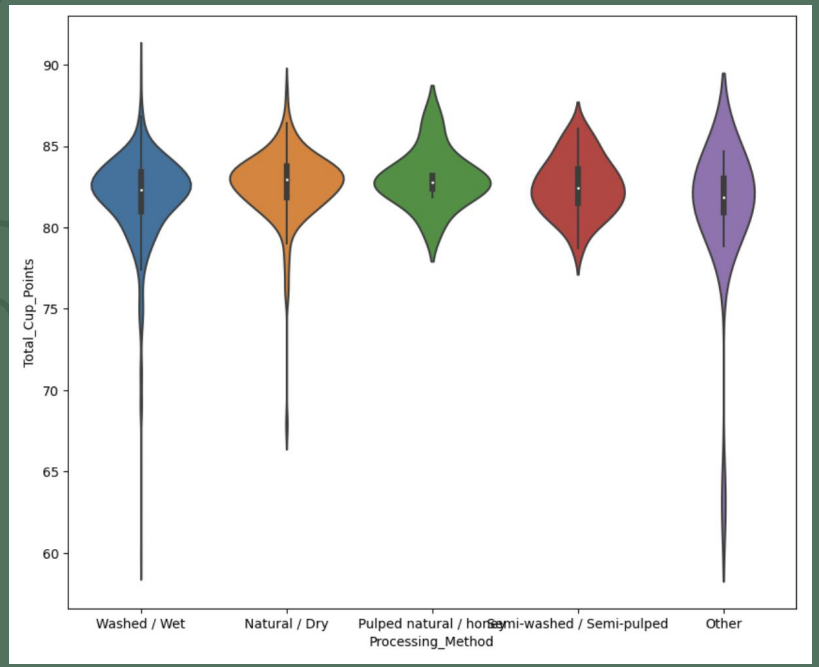
distribution of target





distribution of target





03

Splitting and Preprocessing

Preprocessing: Feature Selection

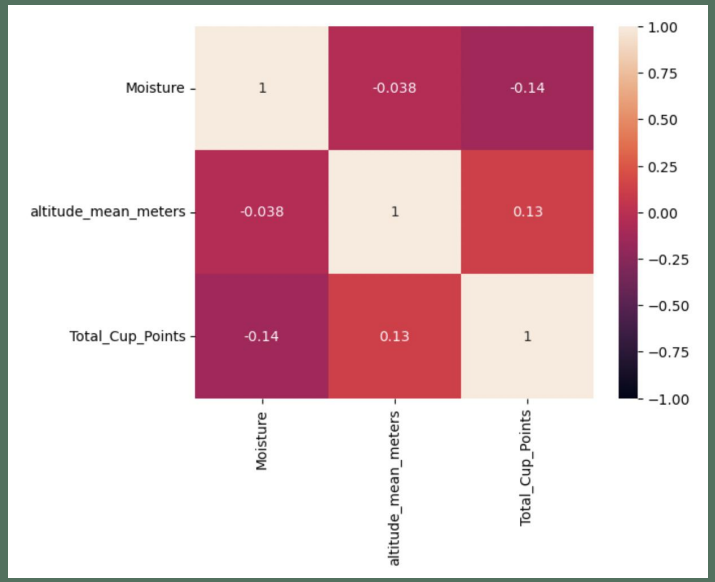
Customer-Oriented

Where beans come from
- Country of origin
- mean altitude
- unit of measurement

What type of beans
- Variety
- Processing method
- Moisture
- Color

Target Value:
- Total Cup Points

Columns before: 43
Columns after: 8



Preprocessing: unit convert

Before

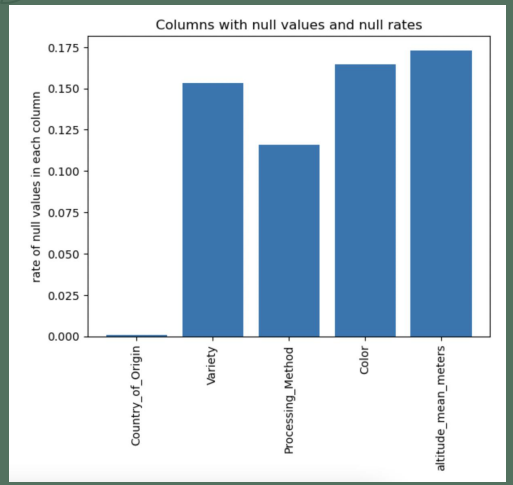
Processing_Method	Moisture	unit_of_measurement	Color	altitude_mean_meters	Total_Cup_Points
Washed / Wet	0.12	m	Green	2075.00	90.58
Washed / Wet	0.12	m	Green	2075.00	89.92
NaN	0.00	m	NaN	1700.00	89.75
Natural / Dry	0.11	m	Green	2000.00	89.00
Washed / Wet	0.12	m	Green	2075.00	88.83
...	...	...	...	...	...
Washed / Wet	0.11	m	None	900.00	68.33
Natural / Dry	0.14	m	Blue-Green	350.00	67.92
Other	0.13	m	Green	1100.00	63.08
Washed / Wet	0.10	ft	Green	1417.32	59.83
NaN	0.12	m	Green	1400.00	0.00

After

Variety	Processing_Method	Moisture	unit_of_measurement	Color
Other	Washed / Wet	0.12	m	Green
Other	Washed / Wet	0.12	m	Green
Other	Natural / Dry	0.10	m	Green
Catimor	Washed / Wet	0.10	m	Green
Other	Washed / Wet	0.00	m	None
...	...	...	...	...
Catuai	Washed / Wet	0.10	m	Green
Bourbon	Washed / Wet	0.11	m	None
Typica	Natural / Dry	0.14	m	Blue-Green
Caturra	Other	0.13	m	Green
Catuai	Washed / Wet	0.10	m	Green

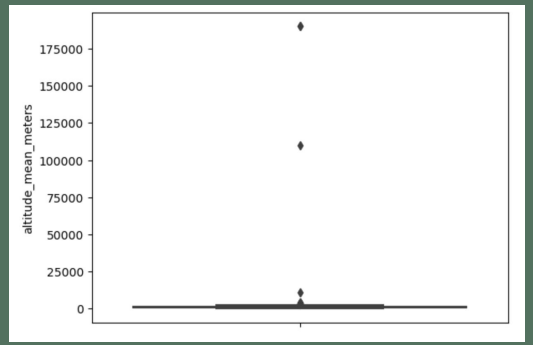
Preprocessing: null_value & outliers

Drop null

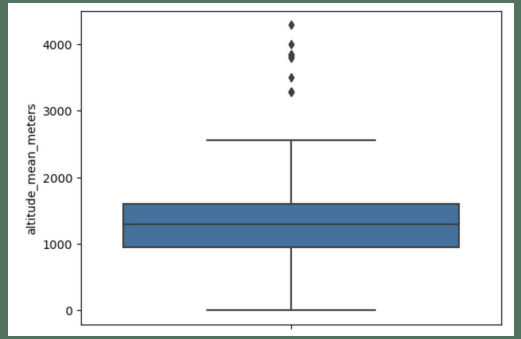


drop outliers

Before



After



Rows before: 1311
Rows after: 901

Preprocessing: Encoders

One-Hot Encoding:

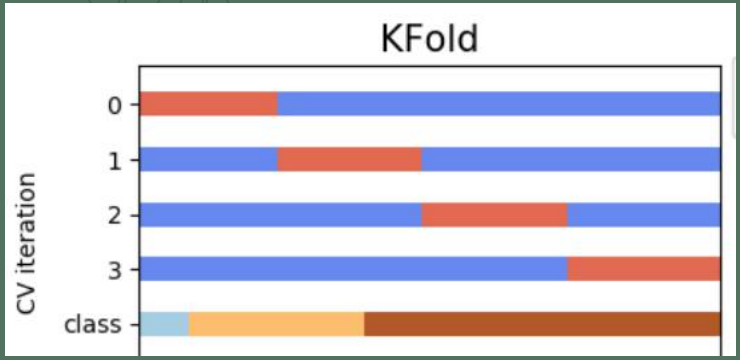
Processing Method,
Variety,
Color,
Country of Origin

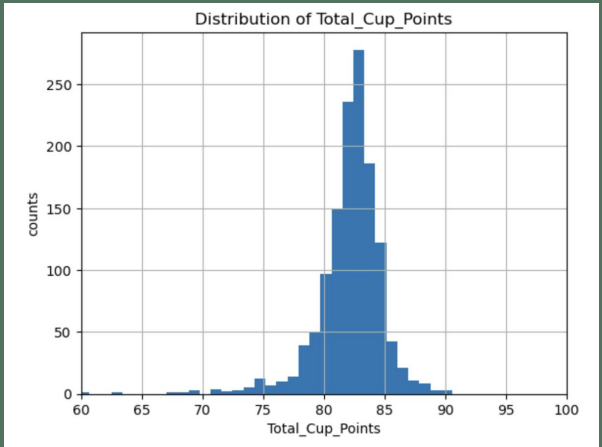
MinMax Encoding:

altitude_mean_meters
Moisture

Split: KFold

split_size:0.6, 0.2, 0.2






```
the 1/4 split:
   training set:  (540, 7) (540,)
   validation set:  (180, 7) (180,)
the 2/4 split:
   training set:  (540, 7) (540,)
   validation set:  (180, 7) (180,)
the 3/4 split:
   training set:  (540, 7) (540,)
   validation set:  (180, 7) (180,)
the 4/4 split:
   training set:  (540, 7) (540,)
   validation set:  (180, 7) (180,)
```

References

- https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi
- https://www.coffeeinstitute.org/
- https://github.com/jldbc/coffee-quality-database/tree/master/scraper

Thank you