Chocolate Bar Ratings

Elisabeth Brockhaus

Karlsruher Institut für Technologie

Motivation

Outline

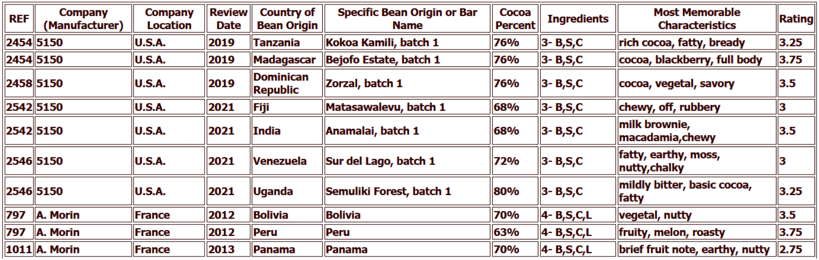
1. Motivation
2. Data
3. Progress
   - ▷ Baseline models
   - ▷ Preprocessing
   - ▷ Advanced ML
4. Conclusion
   - ▷ Future plans
   - ▷ Open questions

Motivation

- Many categorical features
  - How to handle categories with single observation?
  - Some categories not present in training data
- Approach: Regression vs. classification
- Cross-sectional data
  - Time series character negligible

Task

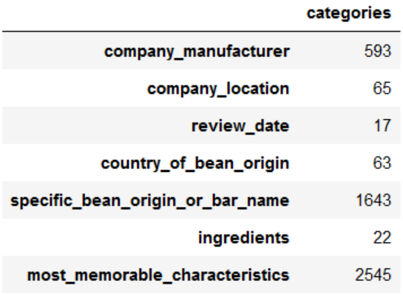
- Predict rating of chocolate bars
  - Discrete scale: 1.0, 1.25, ..., 4.0
  - 2588 observations



REF	Company (Manufacturer)	Company Location	Review Date	Country of Bean Origin	Specific Bean Origin or Bar Name	Cocoa Percent	Ingredients	Most Memorable Characteristics	Rating
2454	5150	U.S.A.	2019	Tanzania	Kokoa Kamili, batch 1	76%	3- B,S,C	rich cocoa, fatty, bready	3.25
2454	5150	U.S.A.	2019	Madagascar	Bejofo Estate, batch 1	76%	3- B,S,C	cocoa, blackberry, full body	3.75
2458	5150	U.S.A.	2019	Dominican Republic	Zorzal, batch 1	76%	3- B,S,C	cocoa, vegetal, savory	3.5
2542	5150	U.S.A.	2021	Fiji	Matasawalevu, batch 1	68%	3- B,S,C	chewy, off, rubbery	3
2542	5150	U.S.A.	2021	India	Anamalai, batch 1	68%	3- B,S,C	milk brownie, macadamia,chewy	3.5
2546	5150	U.S.A.	2021	Venezuela	Sur del Lago, batch 1	72%	3- B,S,C	fatty, earthy, moss, nutty,chalky	3
2546	5150	U.S.A.	2021	Uganda	Semuliki Forest, batch 1	80%	3- B,S,C	mildly bitter, basic cocoa, fatty	3.25
797	A. Morin	France	2012	Bolivia	Bolivia	70%	4- B,S,C,L	vegetal, nutty	3.5
797	A. Morin	France	2012	Peru	Peru	63%	4- B,S,C,L	fruity, melon, roasty	3.75
1011	A. Morin	France	2013	Panama	Panama	70%	4- B,S,C,L	brief fruit note, earthy, nutty	2.75

Features

- Numeric:
  - Cocoa percent
  - (Review date)
- Categorical:



	categories
company_manufacturer	593
company_location	65
review_date	17
country_of_bean_origin	63
specific_bean_origin_or_bar_name	1643
ingredients	22
most_memorable_characteristics	2545

Baseline Regression Model (BRM)

- Linear regression with “easy” features:
  - Cocoa percent
  - Review date (label-encoded)
  - Ingredients (one-hot-encoded)
- Performance:

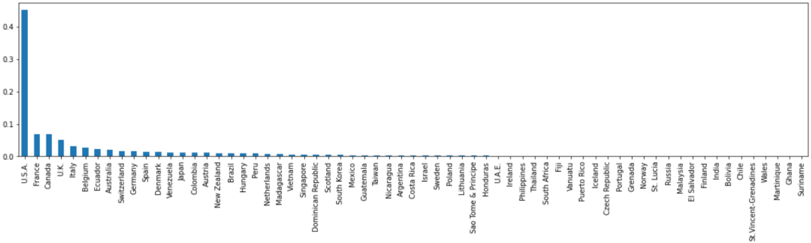
	R^2	Accuracy
Train	0.1030	0.2005
Test	0.0691	0.2490

Baseline Classification Model (BCM)

- Decision Tree Classifier with $max\_depth = 5$ and
  - Cocoa percent
  - Review date (label-encoded)
  - Ingredients (one-hot-encoded)
- Performance:

	R^2	Accuracy
Train	-	0.2966
Test	-	0.2201

Company Location

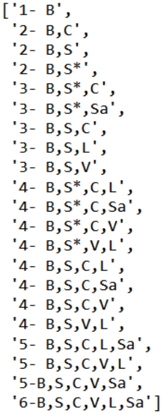


- Problem: Unbalanced categories
  - “New” categories in test data
- Solution: Dummies for countries with “enough” data points
- Same for Company Manufacturer and Country of Bean Origin

Impact on accuracy

	BRM	with new features	BCM	with new features
Train	0.2005	0.2155	0.2966	0.3048
Test	0.2490	0.2510	0.2201	0.2162

Ingredients



```
['1- B',
 '2- B,C',
 '2- B,S',
 '2- B,S*',
 '3- B,S*,C',
 '3- B,S*,Sa',
 '3- B,S,C',
 '3- B,S,L',
 '3- B,S,V',
 '4- B,S*,C,L',
 '4- B,S*,C,Sa',
 '4- B,S*,C,V',
 '4- B,S*,V,L',
 '4- B,S,C,L',
 '4- B,S,C,Sa',
 '4- B,S,C,V',
 '4- B,S,V,L',
 '5- B,S,C,L,Sa',
 '5- B,S,C,V,L',
 '5-B,S,C,V,Sa',
 '6-B,S,C,V,L,Sa']
```

- ▶ Split string
  - ▷ Number of ingredients
  - ▷ Dummy for each ingredient (except beans)
- ▶ NaN values in number of ingredients
  - ▷ Replace by 1 if cocoa percent = 100%
  - ▷ Else replace by mean from training data

Impact on accuracy

	BRM	with new features	BCM	with new features
Train	0.2005	0.2039	0.2966	0.2976
Test	0.2490	0.2375	0.2201	0.2181

Most Memorable Characteristics

- Problem: Descriptions almost completely unique
- Split into single words
  - Still words occur at most 5 times
- Autocorrect (autocorrect, spellchecker, textblob)
  - Unknown words are falsely corrected (macadamia → academia, buttery → battery)
- Word stemming (nltk)
  - Cuts off last letter(s) (oranges → orang, multiple → multipl)
  - Replaces ending y by i (earthy → earthi)
- Solution: Text embedding

Text embedding

- Trained embeddings available at TensorFlow Hub
- E.g.*
  - Trained on English Wikipedia corpus
  - Maps text into 250-dimensional numerical space
- Similar characteristics close to each other
- Additionally: PCA
  - 67 dimensions explain $> 95\%$ variation
  - 22 dimensions explain $> 80\%$ variation

*https://tfhub.dev/google/Wiki-words-250-with-normalization/2

Impact on accuracy

	BRM	with new features	BCM	with new features
Train	0.2005	0.2739	0.2966	0.3420
Test	0.2490	0.2857	0.2201	0.2548

More complicated models

- LASSO
    - With all processed features
    - Find alpha which maximizes R^2_{test}
- Random Forest Regressor
    - With all processed features
    - $min\_samples\_leaf = 50, max\_depth = 5$
- Performance:

	BRM	BCM	LASSO	RFR
Train	0.2005	0.2966	0.3425	0.2493
Test	0.2490	0.2201	0.3205	0.2606

Future plans

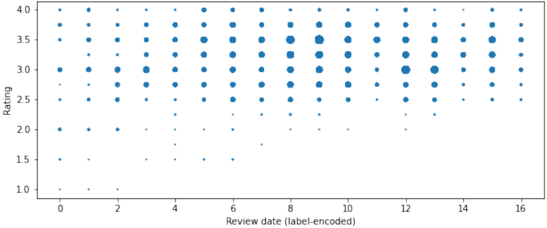
- More advanced/complicated models
  - More classifier models
  - Hyperparameter optimization in random forests
  - Neural networks
- Embedding for locations
  - geographically close ⇔ chocolate similar
- Text sentiment analysis
  - positive/negative label based on most memorable characteristics
- Try different encodings (frequency encoding for ingredients)

Open questions

- Mapping of regression results to discrete scale okay?
- What about reverse way: Treat predicted categories as floats and compute R^2?
- Is there anything you want to discuss or comment on?

Numerical features

- Transform in order to maximize correlation with the rating
- Candidates: x^2, x^3, x^4, $exp(x)$, $log(x)$
- (Cocoa percent)^4
- *exp*(Number of ingredients)
- Review date (label-encoded):



Impact on accuracy

	BRM	with new features	BCM	with new features
Train	0.2005	0.2019	0.2966	0.2966
Test	0.2490	0.2452	0.2201	0.2201