To cocktail or not to

Joshua Banks Mailman

Predicting whether a food is used in a cocktail recipe

Shed light on

- drinkers' tastes and preferences

- drink-mixing creativity

- cocktails-to-be

Sources

Beautiful Soup

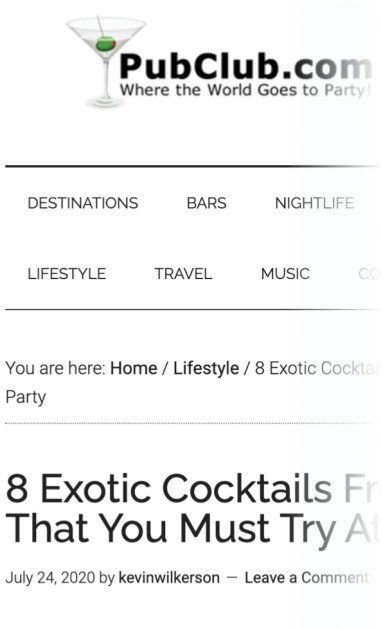
for *food* data

Foodb.ca

for *cocktail* recipes

CocktailDB.com (json)

PubClub.com



150+
cocktail ingredients

550+
cocktail recipes

900+
foods

5,000,000+
food compound concentration measurements

Features of foods

100+
categorical features
(food groups and subgroups)

food categories

id	name	food_group	food_subgroup	food_type
829	pita bread	Cereals and cereal products	Flat breads	Type 2
775	hot dog	Dishes	Sandwiches	Type 2
389	jackfruit	Fruits	Tropical fruits	Type 1
585	charr	Aquatic foods	Fishes	Type 1
744	popcorn	Snack foods	Snack foods	Type 2
30	common cabbage	Vegetables	Cabbages	Type 1
681	fruit preserve	Fruits	Fruit products	Type 2
310	wild boar	Animal foods	Swine	Type 1
137	anise	Herbs and Spices	Herbs	Type 1
660	pastry	Confectioneries	Desserts	Type 2
1019	white bread	Cereals and cereal products	Cereals	Type 1
423	nopal	Vegetables	Other vegetables	Type 1
438	pheasant	Animal foods	Poultry	Type 1
164	sorrel	Herbs and Spices	Herbs	Type 1
517	tea leaf willow	Herbs and Spices	Herbs	Type 1
823	tree fern	Vegetables	Other vegetables	Type 1
915	green bean	Pulses	Beans	Type 1
147	european plum	Fruits	Drupes	Type 1
416	moth bean	Pulses	Beans	Type 1
206	ginger	Herbs and Spices	Spices	Type 1

Processed food

Unprocessed food

Ingredients in cocktails

GIGO ?

Literal interpretation of words can be thorny

	name	value	in_cocktail	ingredient
0	Gin	101	True	gin
1	Vodka	94	True	vodka
2	Sugar	51	True	sugar
3	Orange juice	50	True	orange
4	Lemon	44	True	lemon
...	...	...	...	...
361	Coffee brandy	1	True	coffee brandy
362	Dubonnet Rouge	1	True	dubonnet rouge
363	Apricot Nectar	1	True	apricot
364	Fresh Lemon Juice	1	True	lemon
365	Apple cider	1	True	apple cider

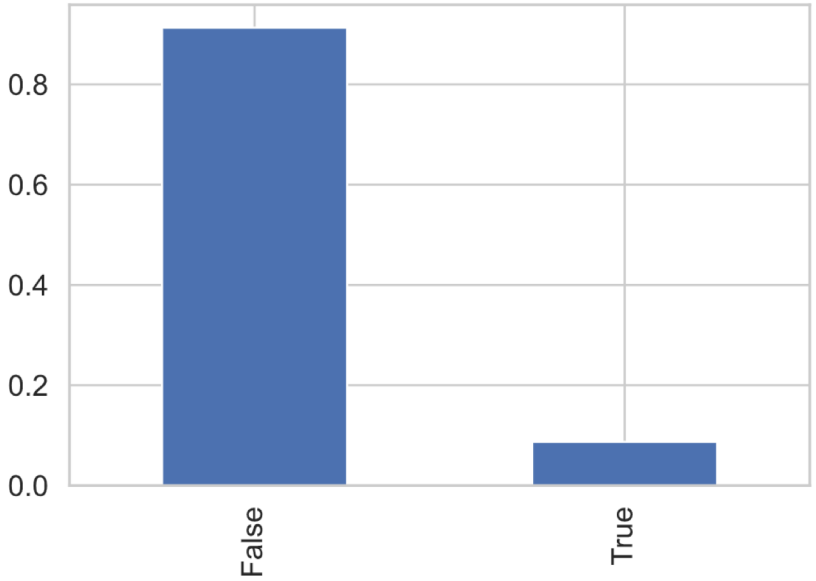
Linking datasets to provide target labels

A significant pre-processing step

The Python *NLTK Lemmetizer* helped a little

id	name	food_group	food_subgroup	food_type
829	pita bread	Cereals and cereal products	Flat breads	Type 2
775	hot dog	Dishes	Sandwiches	Type 2
389	jackfruit	Fruits	Tropical fruits	Type 1
585	charr	Aquatic foods	Fishes	Type 1
744	popcorn	Snack foods	Snack foods	Type 2
30	common cabbage	Vegetables	Cabbages	Type 1
681	fruit preserve	Fruits	Fruit products	Type 2
310	wild boar	Animal foods	Swine	Type 1
137	anise	Herbs and Spices	Herbs	Type 1
660	pastry	Confectioneries	Desserts	Type 2
1019	white bread	Cereals and cereal products	Cereals	Type 1
423	nopal	Vegetables	Other vegetables	Type 1
438	pheasant	Animal foods	Poultry	Type 1
164	sorrel	Herbs and Spices	Herbs	Type 1
517	tea leaf willow	Herbs and Spices	Herbs	Type 1
823	tree fern	Vegetables	Other vegetables	Type 1
915	green bean	Pulses	Beans	Type 1
147	european plum	Fruits	Drupes	Type 1
416	moth bean	Pulses	Beans	Type 1
206	ginger	Herbs and Spices	Spices	Type 1

Class imbalance



So ***accuracy*** will **not** be a good metric

	Not in a cocktail	In a cocktail
Proportions:	0.85624	0.14375

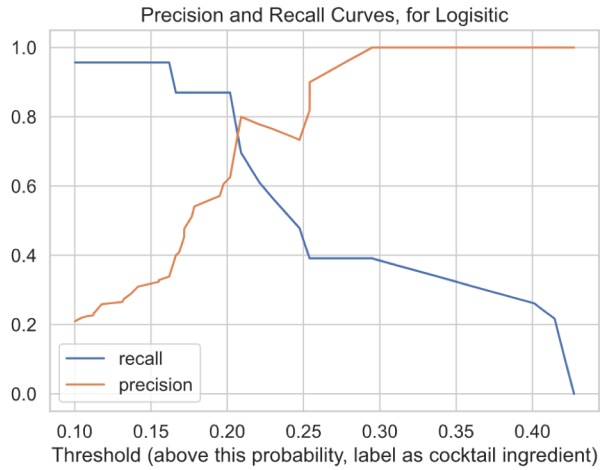
Logistic Regression based on just Categorical features

Logistic Regression based on just Categorical features

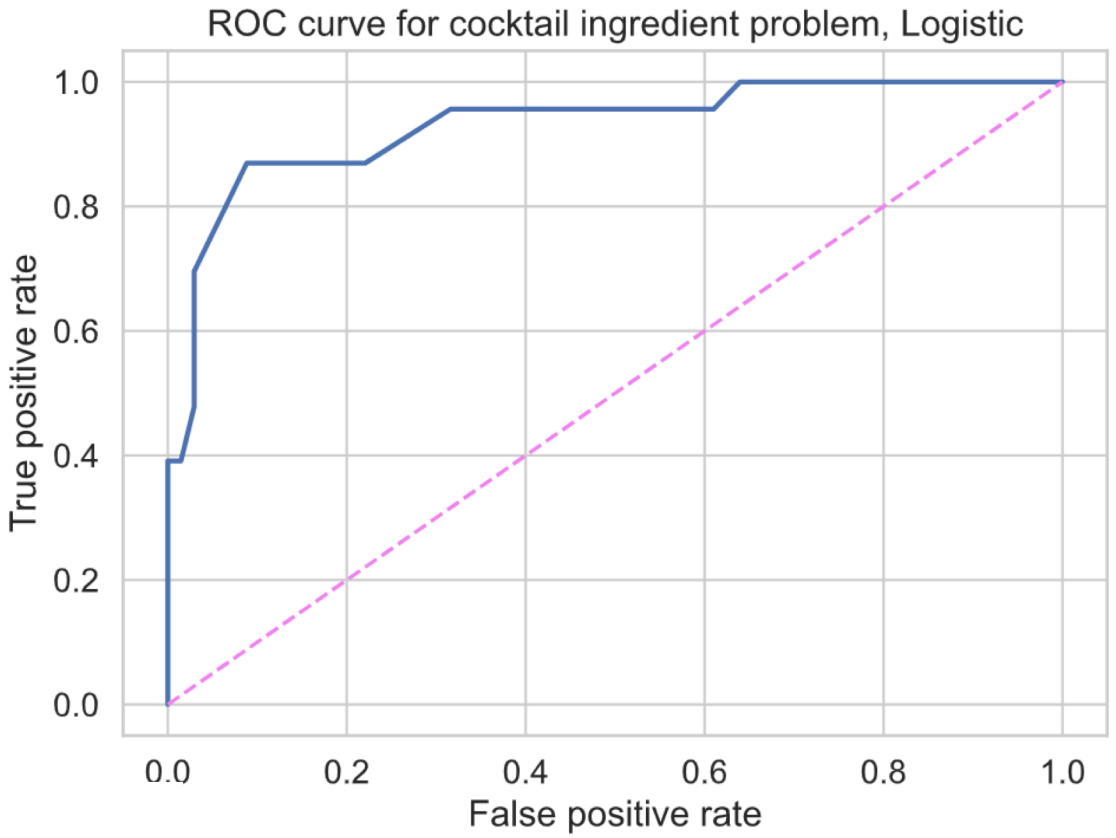
one trial

Log loss cross-entropy

train	val	difference to gauge overfitting
0.3340	0.3047	-0.0293




```
ROC AUC score, for Logistic =  0.9309462915601023
```

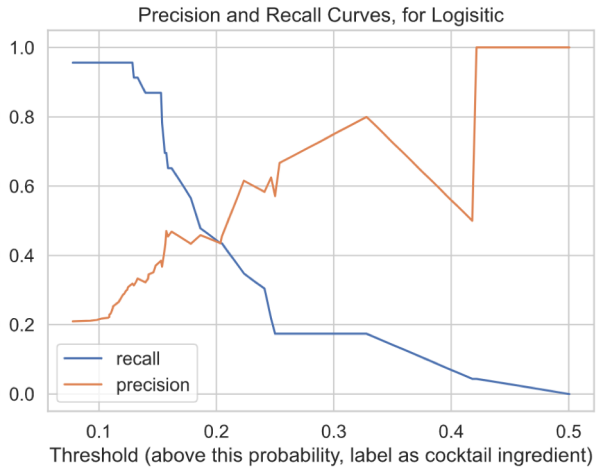


Logistic Regression based on just Categorical features

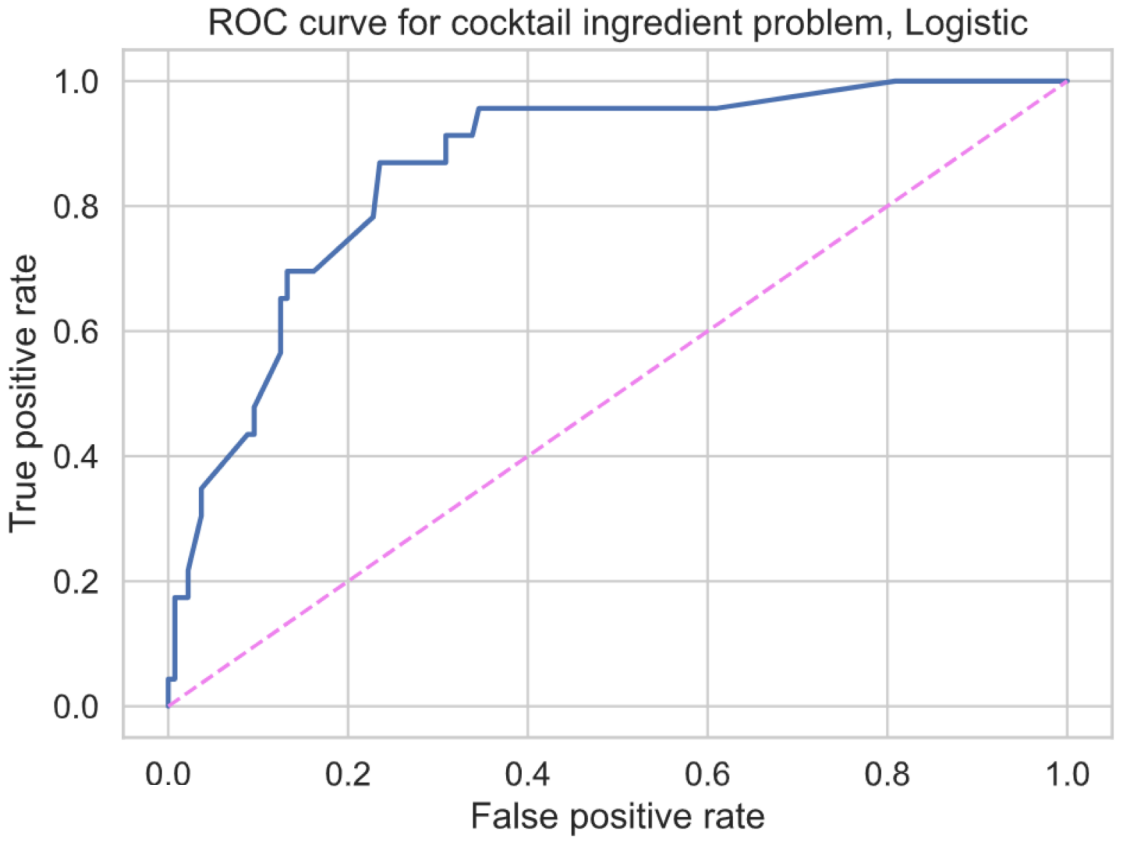
another trial

Log loss cross-entropy

train	val	difference to gauge overfitting
0.3210	0.3382	0.0172



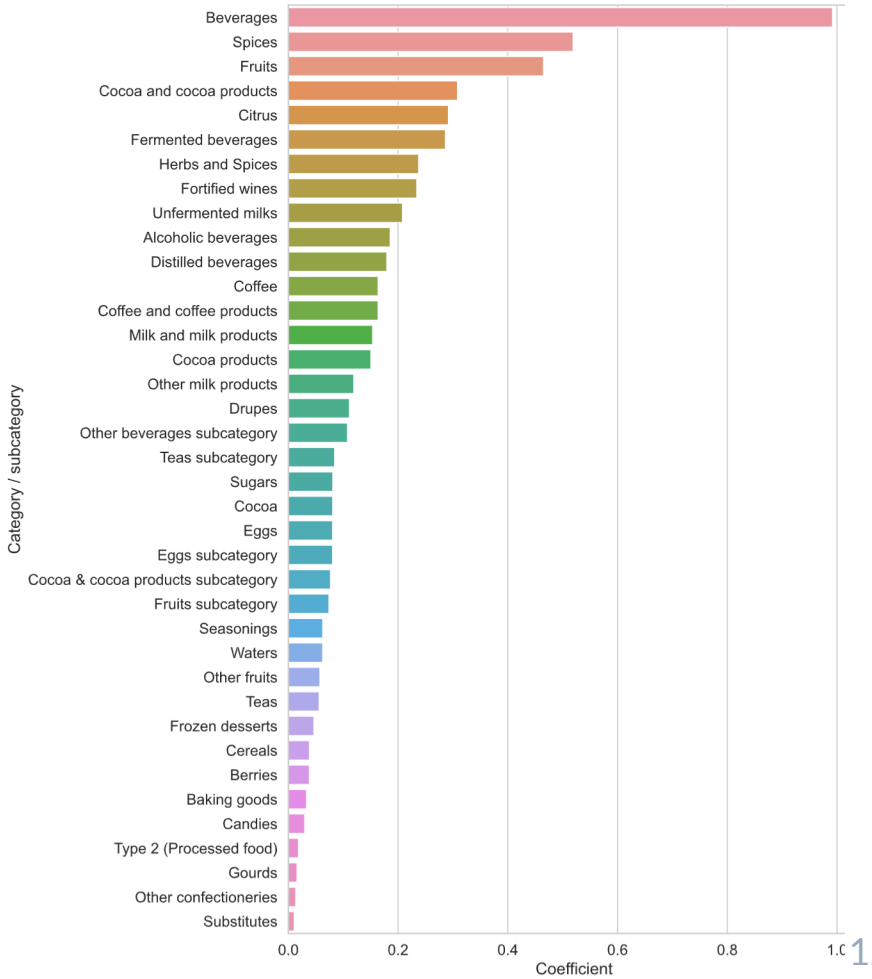

```
ROC AUC score, for Logistic =  0.8618925831202046
```



Logistic Regression (on continuous features) coefficients

contributors

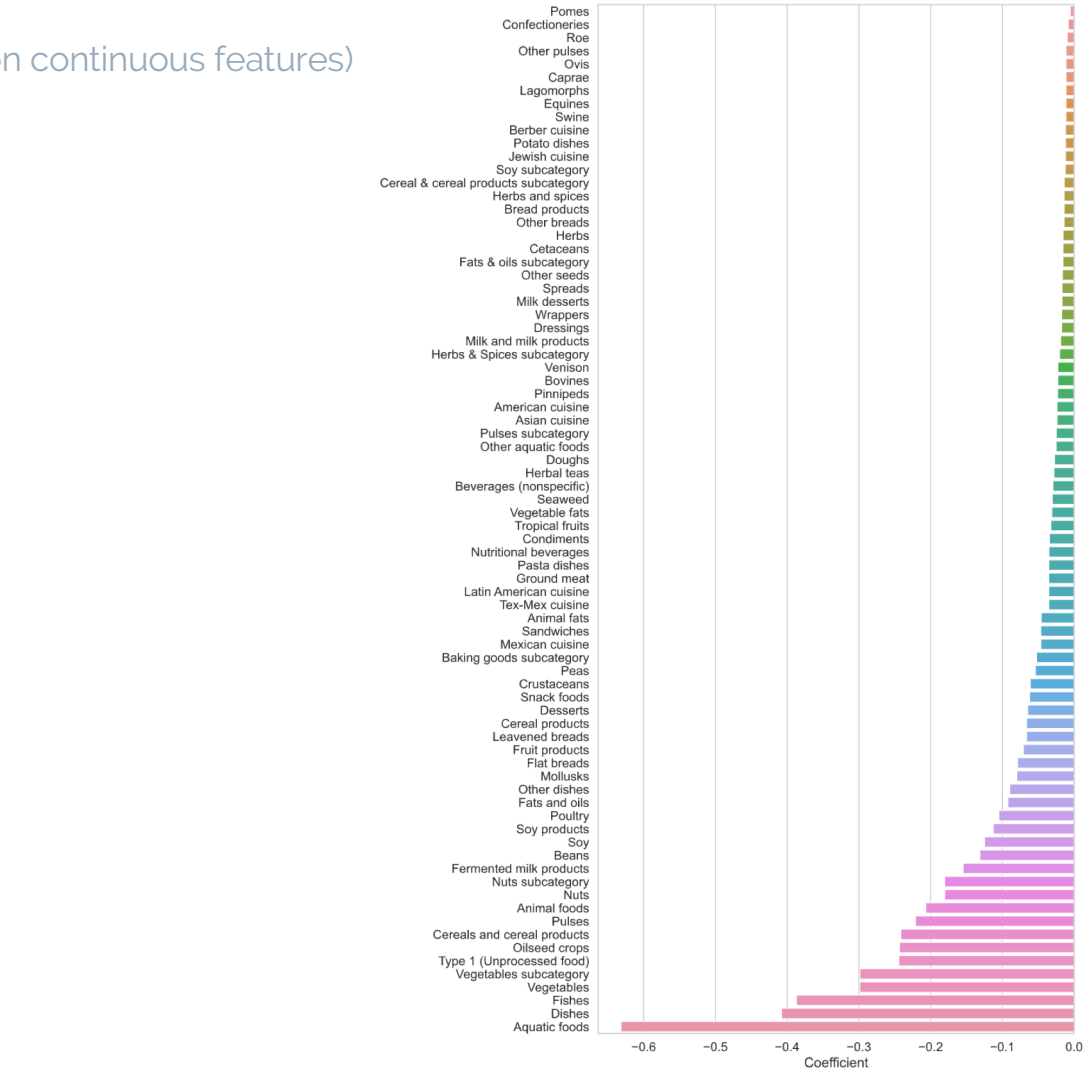
(positive coefficients)



Logistic Regression (on continuous features)

detractors

(negative coefficients)



Multiple scenarios

Categorical features

Insight into food feature importance

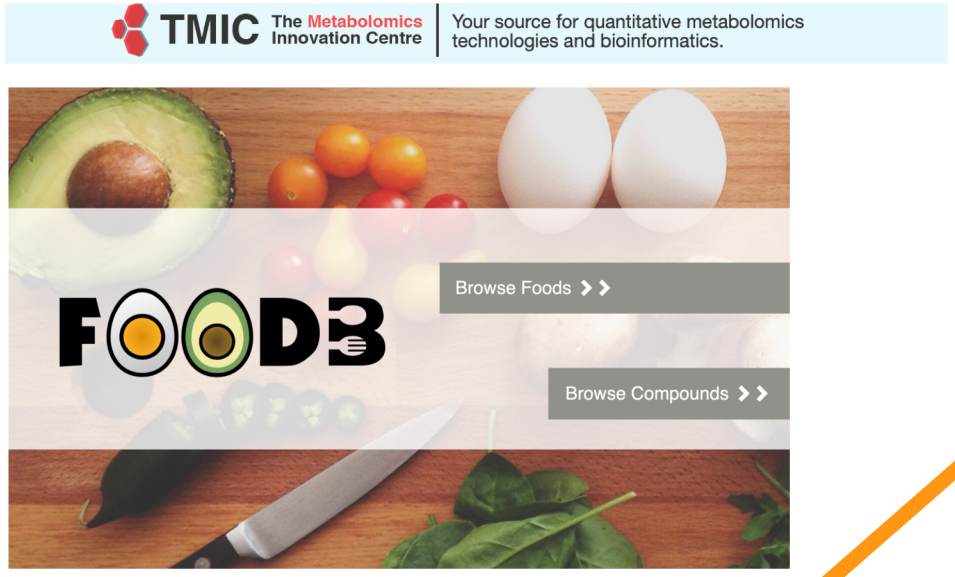
***Independent* of conventional food groupings**

Features of foods

Continuous features

Source for *food* data

Foodb.ca



- foodb_2020_04_07_csv
  - AccessionNumber.csv
  - Compound.csv
  - CompoundAlternateParent.csv
  - CompoundExternalDescriptor.csv
  - CompoundOntologyTerm.csv
  - CompoundsEnzyme.csv
  - CompoundsFlavor.csv
  - CompoundsHealthEffect.csv
  - CompoundsPathway.csv
  - CompoundSubstituent.csv
  - CompoundSynonym.csv
  - Content.csv
  - Enzyme.csv
  - EnzymeSynonym.csv
  - Flavor.csv
  - Food.csv
  - FoodTaxonomy.csv
  - HealthEffect.csv
  - MapItemsPathway.csv
  - NcbiTaxonomyMap.csv
  - Nutrient.csv
  - OntologySynonym.csv
  - OntologyTerm.csv
  - Pathway.csv
  - PdbIdentifier.csv
  - Pfam.csv
  - PfamMembership.csv
  - Reference.csv
  - Sequence.csv

900+ foods

5,000,000+

food compound concentration measurements

id	source_id	source_type	food_id	orig_food_id	orig_food_co	orig_food_sc	orig_food_pa	orig_source_i	orig_source_i	orig_content	orig_min	orig_max	orig_unit
1	1	Nutrient	4	29	Kiwi	Actinidia chir	Fruit	FAT	FAT	1955	70	3840	mg/100g
2	1	Nutrient	6	53	Onion	Allium cepa I	Bulb	FAT	FAT	1853.95	100	3607.9	mg/100g
3	1	Nutrient	6	53	Onion	Allium cepa I	Leaf	FAT	FAT	4150	600	7700	mg/100g
4	1	Nutrient	9	55	Chives	Allium schoe	Leaf	FAT	FAT	3900	300	7500	mg/100g
5	1	Nutrient	11	70	Cashew	Anacardium (	Fruit	FAT	FAT	2500	100	4900	mg/100g
6	1	Nutrient	11	70	Cashew	Anacardium (	Leaf	FAT	FAT	1300	600	2000	mg/100g
7	1	Nutrient	11	70	Cashew	Anacardium (	Seed	FAT	FAT	42600	37000	48200	mg/100g
8	1	Nutrient	12	74	Pineapple	Ananas como	Fruit	FAT	FAT	2188.6	100	4277.2	mg/100g
9	1	Nutrient	13	83	Dill	Anethum gra	Plant	FAT	FAT	4500	4300	4700	mg/100g
10	1	Nutrient	13	83	Dill	Anethum gra	Seed	FAT	FAT	12410	6800	18020	mg/100g
11	1	Nutrient	14	92	Custard Appl	Annona retic	Fruit	FAT	FAT	1150	200	2100	mg/100g
12	1	Nutrient	14	92	Custard Appl	Annona retic	Seed	FAT	FAT	40000	40000	40000	mg/100g
13	1	Nutrient	15	96	Celery	Apium grave	Fruit	FAT	FAT				
14	1	Nutrient	15	96	Celery	Apium grave	Leaf	FAT	FAT	2165	130	4200	mg/100g
15	1	Nutrient	15	96	Celery	Apium grave	Seed	FAT	FAT	26110.65	24273	27948.3	mg/100g
16	1	Nutrient	16	101	Groundnut	Arachis hypo	Leaf	FAT	FAT	1700	600	2800	mg/100g
17	1	Nutrient	16	101	Groundnut	Arachis hypo	Plant	FAT	FAT	5850	2200	9500	mg/100g
18	1	Nutrient	16	101	Groundnut	Arachis hypo	Seed	FAT	FAT	36081.7	19480	52683.4	mg/100g
19	1	Nutrient	17	104	Burdock	Arctium lapp	Fruit	FAT	FAT	16500	15000	18000	mg/100g
20	1	Nutrient	17	104	Burdock	Arctium lapp	Leaf	FAT	FAT	481.75	130	833.5	mg/100g
21	1	Nutrient	17	104	Burdock	Arctium lapp	Root	FAT	FAT	450	100	800	mg/100g
22	1	Nutrient	18	115	Horseradish	Armoracia ru	Root	FAT	FAT	700	200	1200	mg/100g
23	1	Nutrient	19	123	Tarragon	Artemisia dra	Seed	FAT	FAT	38100	38100	38100	mg/100g
24	1	Nutrient	19	123	Tarragon	Artemisia dra	Shoot	FAT	FAT	7500	7200	7800	mg/100g
25	1	Nutrient	20	127	Mugwort	Artemisia vul	Leaf	FAT	FAT	3550	800	6300	mg/100g
26	1	Nutrient	21	137	Asparagus	Asparagus off	Root	FAT	FAT				
27	1	Nutrient	21	137	Asparagus	Asparagus off	Shoot	FAT	FAT	2150	200	4100	mg/100g
28	1	Nutrient	22	144	Oats	Avena sativa	Plant	FAT	FAT	2900	1900	3900	mg/100g
29	1	Nutrient	22	144	Oats	Avena sativa	Seed	FAT	FAT	5400	1100	9700	mg/100g
30	1	Nutrient	23	1207	Carambola	Averrhoa car	Fruit	FAT	FAT	2640	80	5200	mg/100g
31	1	Nutrient	24	156	Brazilnut	Bertholletia e	Seed	FAT	FAT	67459.5	64919	70000	mg/100g
32	1	Nutrient	26	166	Beebread	Borago offici	Leaf	FAT	FAT	5350	700	10000	mg/100g
33	1	Nutrient	26	166	Beebread	Borago offici	Seed	FAT	FAT	38300	38300	38300	mg/100g
34	1	Nutrient	27	171	Mustard Gree	Brassica junc	Leaf	FAT	FAT	7535	1270	13800	mg/100g
35	1	Nutrient	27	171	Mustard Gree	Brassica junc	Seed	FAT	FAT	34000	30000	38000	mg/100g
36	1	Nutrient	32	175	Brussel-Sprou	Brassica olera	Leaf	FAT	FAT	1528	200	2856	mg/100g
37	1	Nutrient	33	180	Kohlrabi	Brassica olera	Stem	FAT	FAT	605.5	100	1111	mg/100g
38	1	Nutrient	34	2255	Asparagus Br	Brassica olera	Leaf	FAT	FAT				
39	1	Nutrient	37	193	Pigeonpea	Cajanus cajar	Fruit	FAT	FAT	1300	600	2000	mg/100g
40	1	Nutrient	37	193	Pigeonpea	Cajanus cajar	Leaf	FAT	FAT	6900	6900	6900	mg/100g
41	1	Nutrient	37	193	Pigeonpea	Cajanus cajar	Plant	FAT	FAT	6000	6000	6000	mg/100g

Concentration: mg/100g

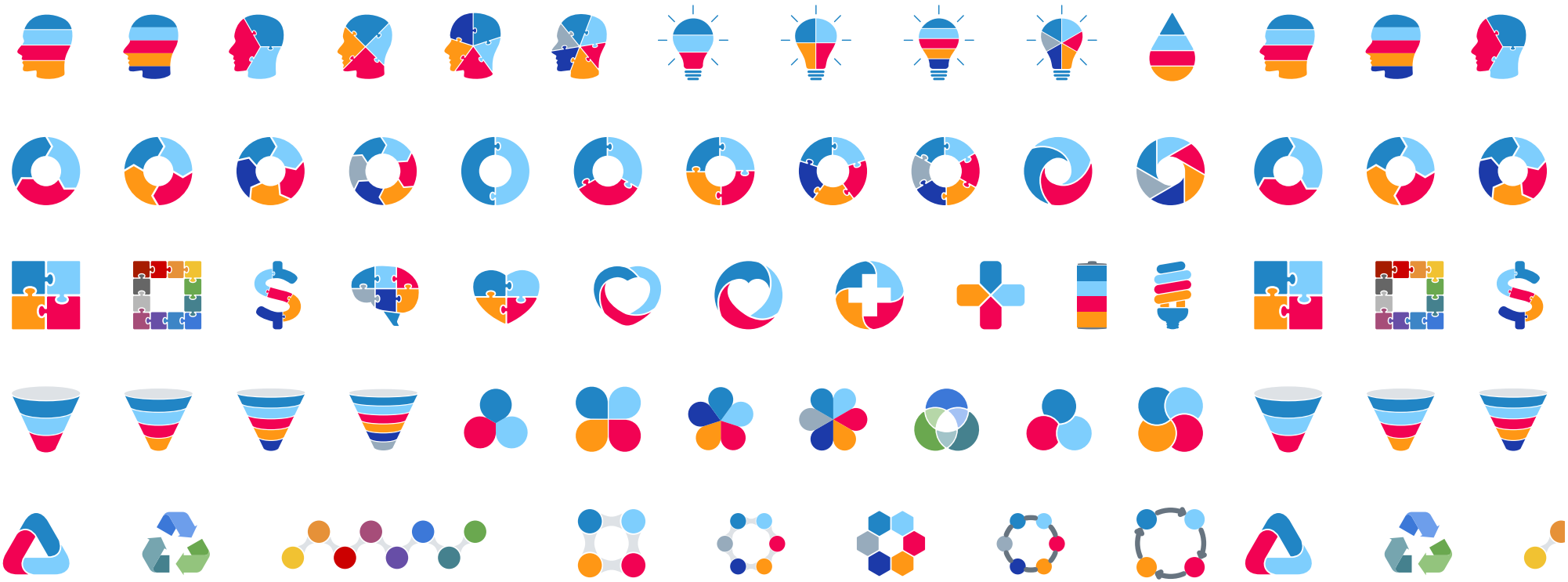
orig_food_id	orig_food_co	orig_food_sc	orig_food_pa	orig_source_i	orig_source_	orig_content	orig_min	orig_max	orig_unit
29	Kiwi	Actinidia chir	Fruit	FAT	FAT	1955	70	3840	mg/100g
53	Onion	Allium cepa I	Bulb	FAT	FAT	1853.95	100	3607.9	mg/100g
53	Onion	Allium cepa I	Leaf	FAT	FAT	4150	600	7700	mg/100g
55	Chives	Allium schoe	Leaf	FAT	FAT	3900	300	7500	mg/100g
70	Cashew	Anacardium (	Fruit	FAT	FAT	2500	100	4900	mg/100g
70	Cashew	Anacardium (	Leaf	FAT	FAT	1300	600	2000	mg/100g
70	Cashew	Anacardium (	Seed	FAT	FAT	42600	37000	48200	mg/100g
74	Pineapple	Ananas como	Fruit	FAT	FAT	2188.6	100	4277.2	mg/100g
83	Dill	Anethum gra	Plant	FAT	FAT	4500	4300	4700	mg/100g
83	Dill	Anethum gra	Seed	FAT	FAT	12410	6800	18020	mg/100g
92	Custard Appl	Annona retic	Fruit	FAT	FAT	1150	200	2100	mg/100g
92	Custard Appl	Annona retic	Seed	FAT	FAT	40000	40000	40000	mg/100g
96	Celery	Apium grave	Fruit	FAT	FAT				
96	Celery	Apium grave	Leaf	FAT	FAT	2165	130	4200	mg/100g
96	Celery	Apium grave	Seed	FAT	FAT	26110.65	24273	27948.3	mg/100g
101	Groundnut	Arachis hypo	Leaf	FAT	FAT	1700	600	2800	mg/100g
101	Groundnut	Arachis hypo	Plant	FAT	FAT	5850	2200	9500	mg/100g
101	Groundnut	Arachis hypo	Seed	FAT	FAT	36081.7	19480	52683.4	mg/100g
104	Burdock	Arctium lapp	Fruit	FAT	FAT	16500	15000	18000	mg/100g
104	Burdock	Arctium lapp	Leaf	FAT	FAT	481.75	130	833.5	mg/100g

Food part
Fruit
Bulb
Leaf
Seed
Plant
Root
Shoot
Stem
Flower
Latex Exudate
Bark
Silk Stigma Style
Rhizome
Stem Bark
Sprout Seedling
Pericarp
Hay
Aril
Twig
Herb
Pt
Tuber
Bud
Hull Husk
Pith

900+ foods

id	source_id	source_type	food_id	orig_food_id	orig_food_co	orig	ood_sc	orig_	od_pa	orig_	ource_i	orig_	ource_	orig_content	orig_min	orig_max	orig_unit
1	1	Nutrient	4	2	Kiwi	Acti	dia chir	Fruit		FAT		FAT		1955	70	3840	mg/100g
2	1	Nutrient	6	53	Onion	Alli	n cepa I	Bulb		FAT		FAT		1853.95	100	3607.9	mg/100g
3	1	Nutrient	6	53	Onion	Alli	n cepa I	Leaf		FAT		FAT		4150	600	7700	mg/100g
4	1	Nutrient	9	150	Olive	Alli	n schoe	Leaf		FAT		FAT		3900	300	7500	mg/100g
5	1	Nutrient	11	70	Cashew	Ana	ardium (	Fruit		FAT		FAT		2500	100	4900	mg/100g
6	1	Nutrient	11	70	Cashew	Ana	ardium (	Leaf		FAT		FAT		1300	600	2000	mg/100g
7	1	Nutrient	11	70	Cashew	Ana	ardium (	Seed		FAT		FAT		42600	37000	48200	mg/100g
8	1	Nutrient	12	74	Pineapple	Ana	as como	Fruit		FAT		FAT		2188.6	100	4277.2	mg/100g
9	1	Nutrient	13	83	Dill	Ane	um gra	Plant		FAT		FAT		4500	4300	4700	mg/100g
10	1	Nutrient	13	83	Dill	Ane	um gra	Seed		FAT		FAT		12410	6800	18020	mg/100g
11	1	Nutrient	14	92	Custard Appl	Ann	na retic	Fruit		FAT		FAT		1150	200	2100	mg/100g
12	1	Nutrient	14	92	Custard Appl	Ann	na retic	Seed		FAT		FAT		40000	40000	40000	mg/100g
13	1	Nutrient	15	96	Celery	Api	n grave	Fruit		FAT		FAT					
14	1	Nutrient	15	96	Celery	Api	n grave	Leaf		FAT		FAT		2165	130	4200	mg/100g
15	1	Nutrient	15	96	Celery	Api	n grave	Seed		FAT		FAT		26110.65	24273	27948.3	mg/100g
16	1	Nutrient	16	101	Groundnut	Ara	is hypo	Leaf		FAT		FAT		1700	600	2800	mg/100g
17	1	Nutrient	16	101	Groundnut	Ara	is hypo	Plant		FAT		FAT		5850	2200	9500	mg/100g
18	1	Nutrient	16	101	Groundnut	Ara	is hypo	Seed		FAT		FAT		36081.7	19480	52683.4	mg/100g
19	1	Nutrient	17	104	Burdock	Arc	um lapp	Fruit		FAT		FAT		16500	15000	18000	mg/100g
20	1	Nutrient	17	104	Burdock	Arc	um lapp	Leaf		FAT		FAT		481.75	130	833.5	mg/100g
21	1	Nutrient	17	104	Burdock	Arc	um lapp	Root		FAT		FAT		450	100	800	mg/100g
22	1	Nutrient	18	115	Horseradish	Arm	racia ru	Root		FAT		FAT		700	200	1200	mg/100g
23	1	Nutrient	19	123	Tarragon	Arte	isia dra	Seed		FAT		FAT		38100	38100	38100	mg/100g
24	1	Nutrient	19	123	Tarragon	Arte	isia dra	Shoot		FAT		FAT		7500	7200	7800	mg/100g
25	1	Nutrient	20	127	Mugwort	Arte	isia vul	Leaf		FAT		FAT		3550	800	6300	mg/100g
26	1	Nutrient	21	137	Asparagus	Asp	agus off	Root		FAT		FAT					
27	1	Nutrient	21	137	Asparagus	Asp	agus off	Shoot		FAT		FAT		2150	200	4100	mg/100g
28	1	Nutrient	22	144	Oats	Ave	a sativa	Plant		FAT		FAT		2900	1900	3900	mg/100g
29	1	Nutrient	22	144	Oats	Ave	a sativa	Seed		FAT		FAT		5400	1100	9700	mg/100g
30	1	Nutrient	23	1207	Carambola	Ave	hoa car	Fruit		FAT		FAT		2640	80	5200	mg/100g
31	1	Nutrient	24	156	Brazilnut	Ber	olletia e	Seed		FAT		FAT		67459.5	64919	70000	mg/100g
32	1	Nutrient	26	166	Beebread	Bor	o offici	Leaf		FAT		FAT		5350	700	10000	mg/100g
33	1	Nutrient	26	166	Beebread	Bor	o offici	Seed		FAT		FAT		38300	38300	38300	mg/100g
34	1	Nutrient	27	171	Mustard Gree	Bra	ca junc	Leaf		FAT		FAT		7535	1270	13800	mg/100g
35	1	Nutrient	27	171	Mustard Gree	Bra	ca junc	Seed		FAT		FAT		34000	30000	38000	mg/100g
36	1	Nutrient	32	175	Brussel-Sprou	Bra	ca olera	Leaf		FAT		FAT		1528	200	2856	mg/100g
37	1	Nutrient	33	180	Kohlrabi	Bra	ca olera	Stem		FAT		FAT		605.5	100	1111	mg/100g
38	1	Nutrient	34	2255	Asparagus Br	Bra	ca olera	Leaf		FAT		FAT					
39	1	Nutrient	37	193	Pigeonpea	Caj	us cajar	Fruit		FAT		FAT		1300	600	2000	mg/100g
40	1	Nutrient	37	193	Pigeonpea	Caj	us cajar	Leaf		FAT		FAT		6900	6900	6900	mg/100g
41	1	Nutrient	37	193	Pigeonpea	Caj	us cajar	Plant		FAT		FAT		6000	6000	6000	mg/100g

15,000 continuous features were engineered by aggregating the 5 million compound concentrations for each part of all 900 foods.



Multiple scenarios

Categorical features

Categorical & continuous features

Continuous features only

using continuous features only

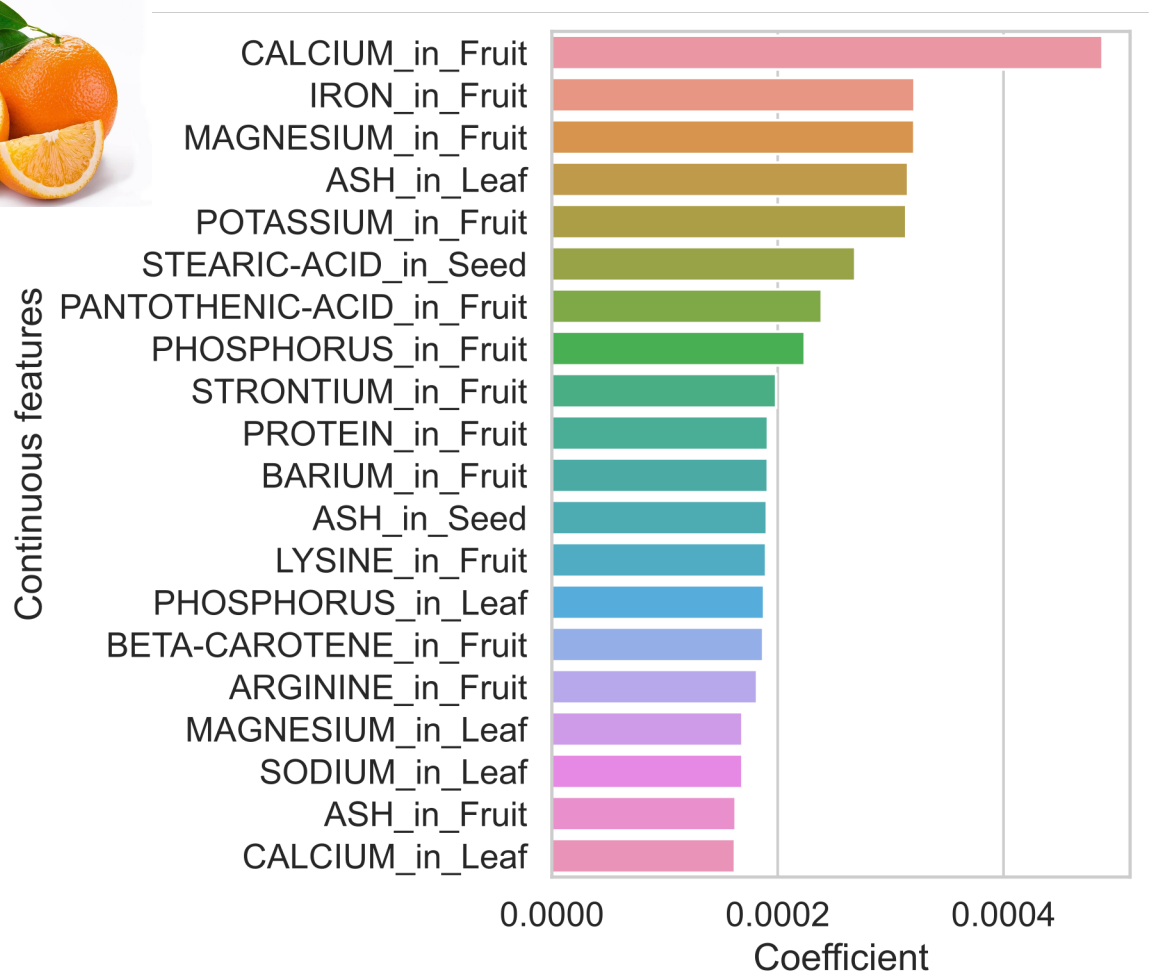
15000+ features, each a concentration of a compound in a food

Log-loss Cross-Entropy Train	Val	difference to gauge *overfitting*	
0.4113	0.4133	0.0020	**Logistic**
0.4097	0.6236	0.2139	**KNN**
27.8380	28.6737	0.8357	**Naïve Bayes**
0.3927	1.0500	0.6574	**Decision Tree**
0.3875	0.6070	0.2195	**Random Forest**
0.3931	0.4021	0.0091	**XGBoost**

Logistic Regression (on continuous features) coefficients

contributors

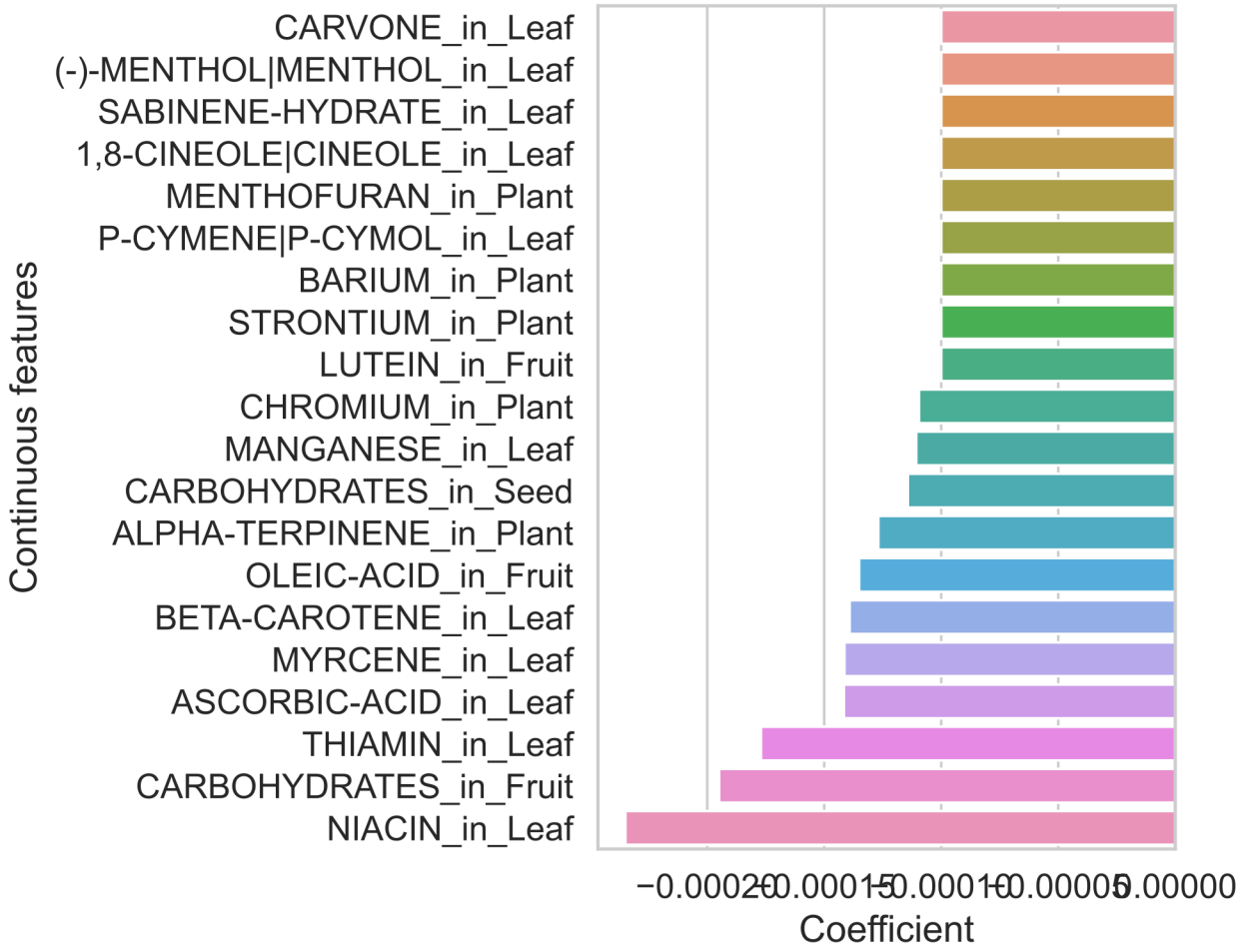
(positive coefficients)



Logistic Regression (on continuous features) coefficients

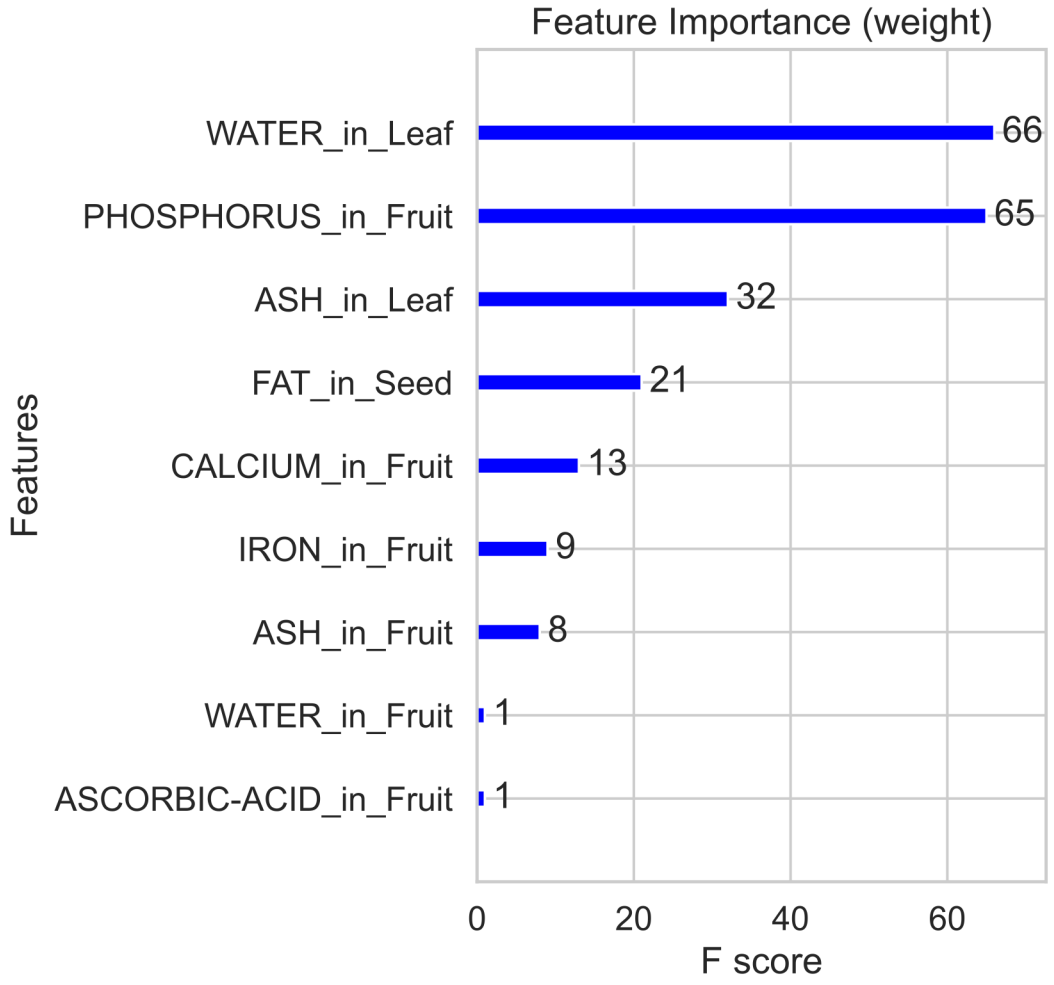
detractors

(negative coefficients)



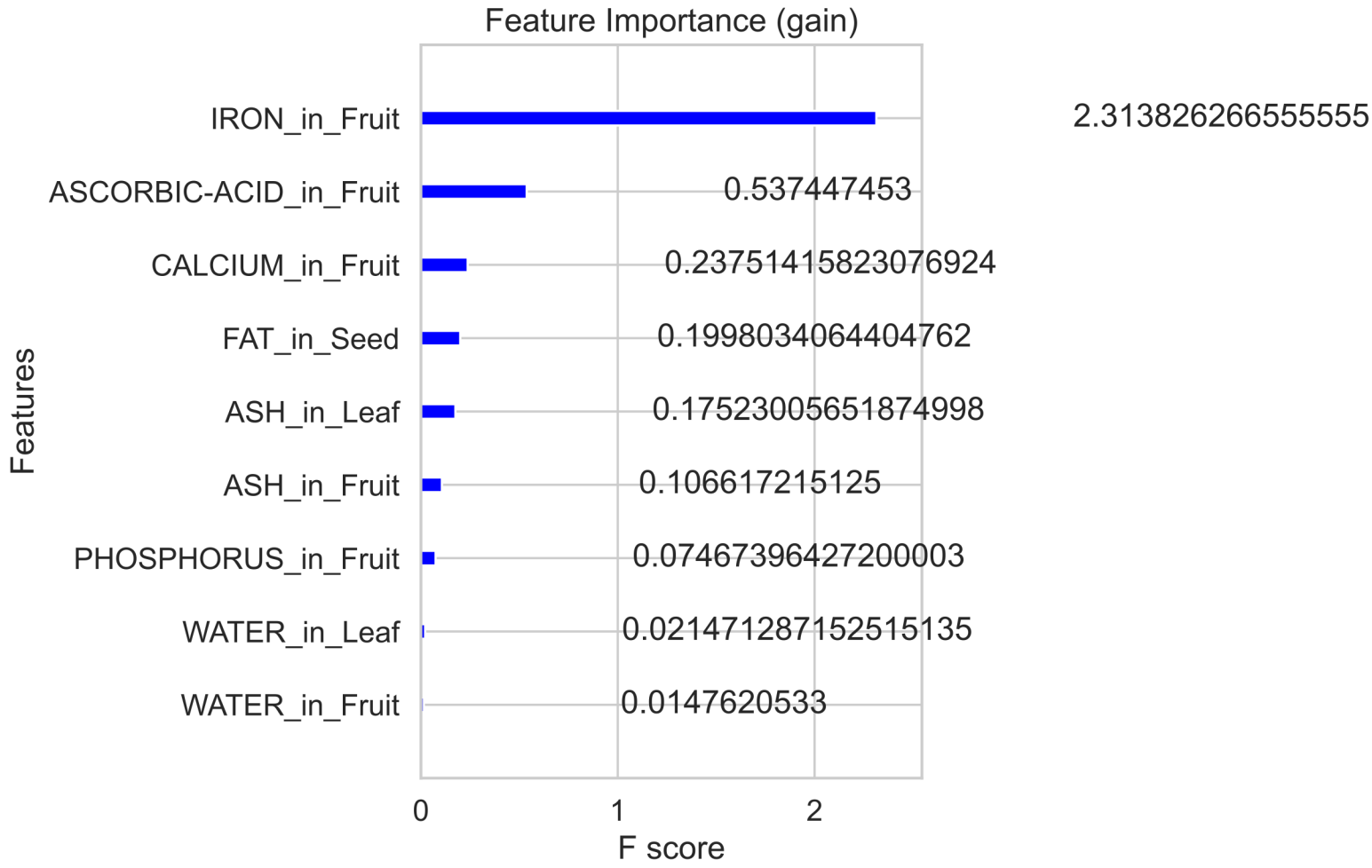
XGBoost 'Important Features'

'weight'



XGBoost 'Important Features'

'gain'



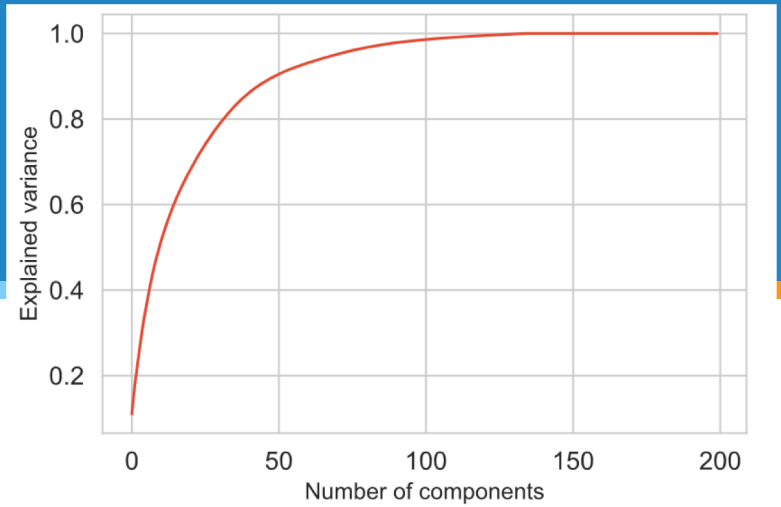
15,000 is a lot of features for only 900 rows. Danger of overfitting

15,000 is a lot of features for only 900 rows. Danger of overfitting

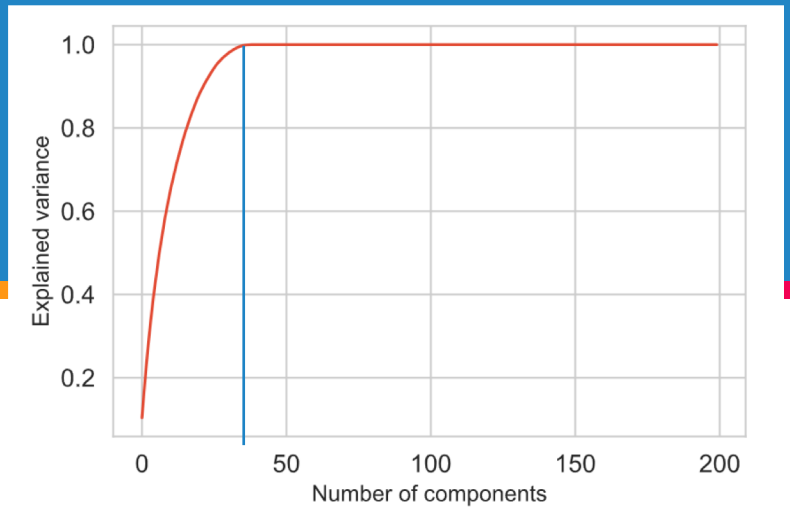
Principal Component Analysis

To reduce the given features down to a set of synthesized features that capture as much variance as possible

PCA on all features



PCA on continuous features



Multiple scenarios

Categorical features

Categorical & continuous features

Continuous features only

PCA on all features

Continuous features w/ PCA

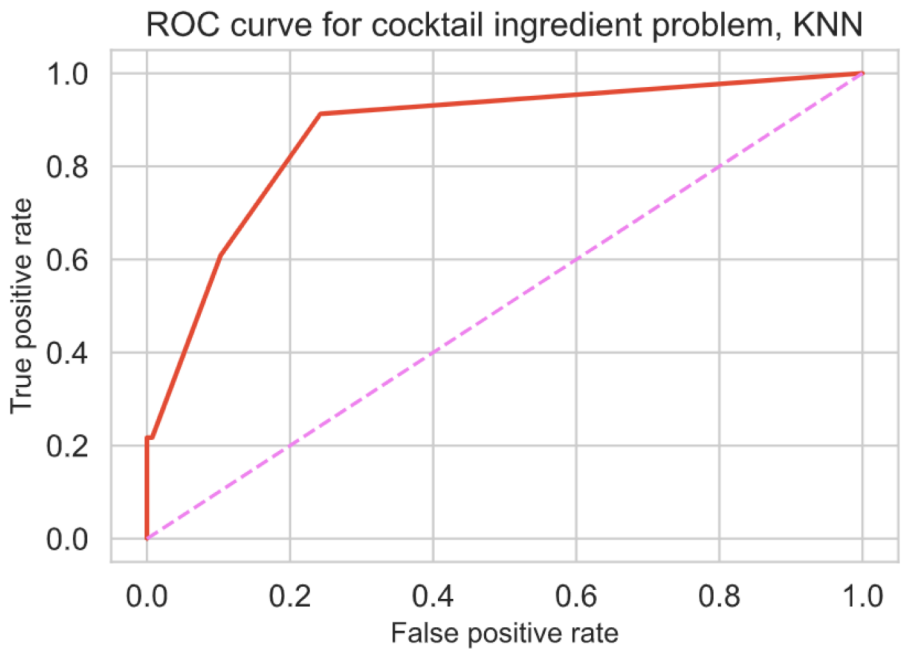
Categorical as-is with continuous features w/ PCA

Categorical features and 40 *principal components* based on the continuous variables

With oversampling

KNN

```
ROC AUC score, for KNN =  0.8718030690537084
```



Multiple scenarios

Because of *class imbalance*,...

Categorical features

Categorical & continuous features

Continuous features only

PCA on all features

Continuous features w/ PCA

Categorical as-is with continuous features w/ PCA

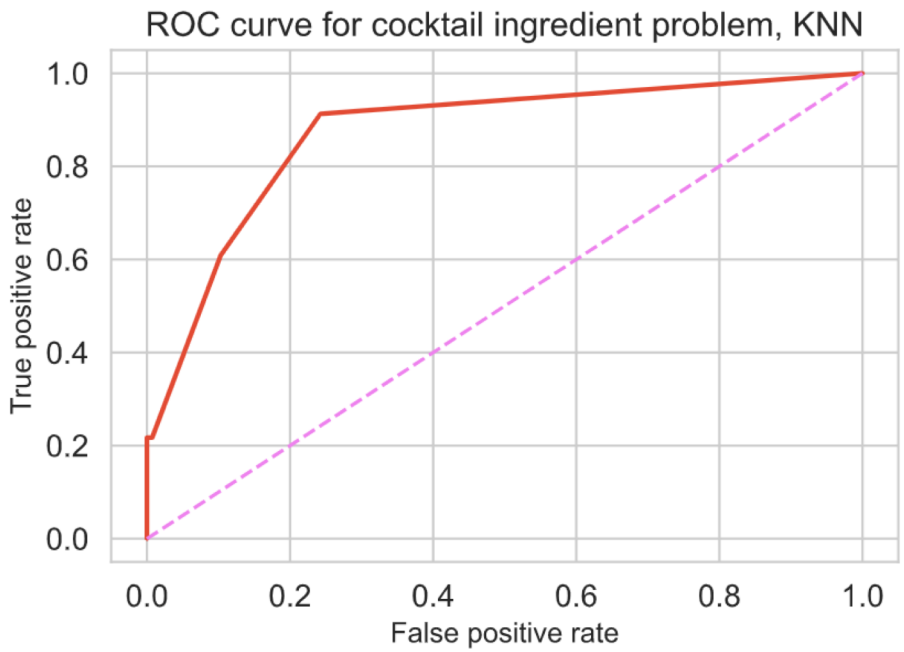
Categorical as-is with continuous features w/ PCA, *oversampled*

Categorical features and 40 *principal components* based on the continuous variables

With oversampling

KNN

```
ROC AUC score, for KNN =  0.8718030690537084
```

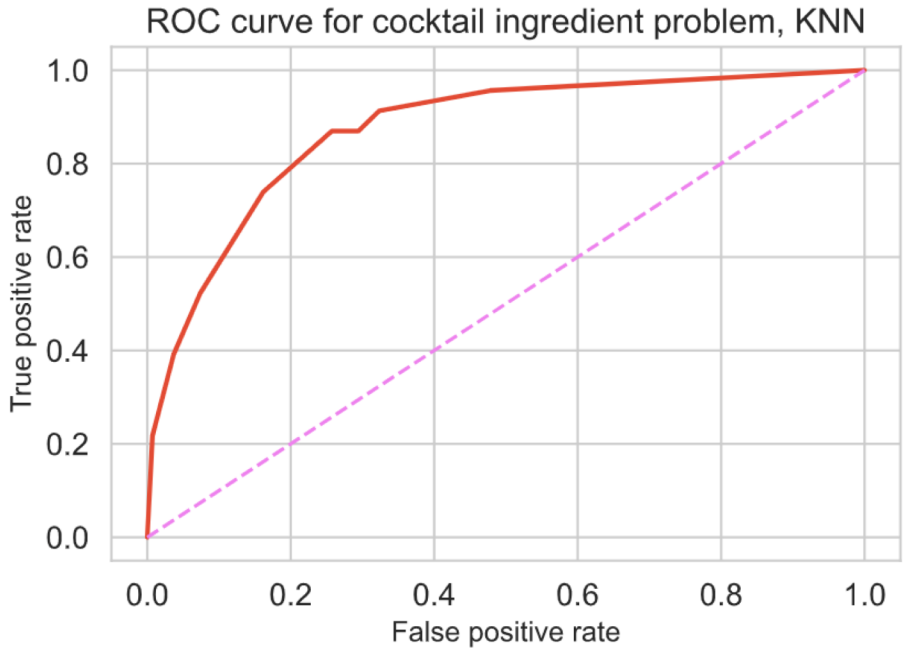


Categorical features and 40 *principal components* based on the continuous variables

With oversampling

KNN

```
ROC AUC score, for KNN =  0.8722826086956521
```

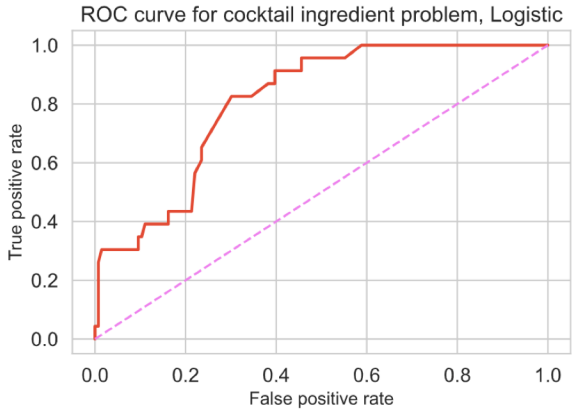


Categorical features and 40 *principal components* based on the continuous variables

With oversampling

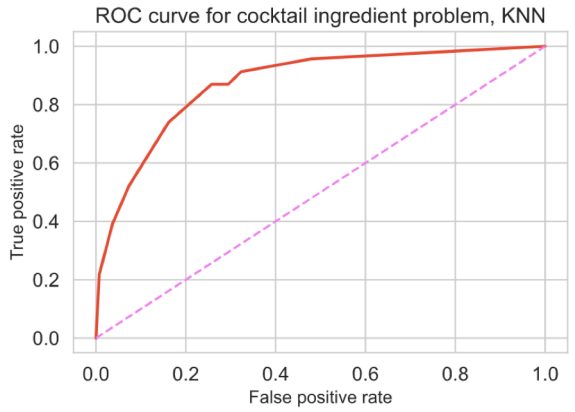
Logistic

```
ROC AUC score, for Logistic =  0.8093030690537084
```



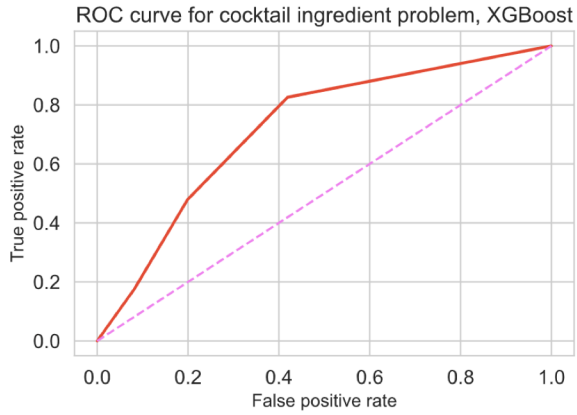
KNN

```
ROC AUC score, for KNN =  0.8722826086956521
```



XGBoost

```
ROC AUC score, for XGBoost =  0.7196291560102301
```



Categorical features and 40 *principal components* based on the continuous variables

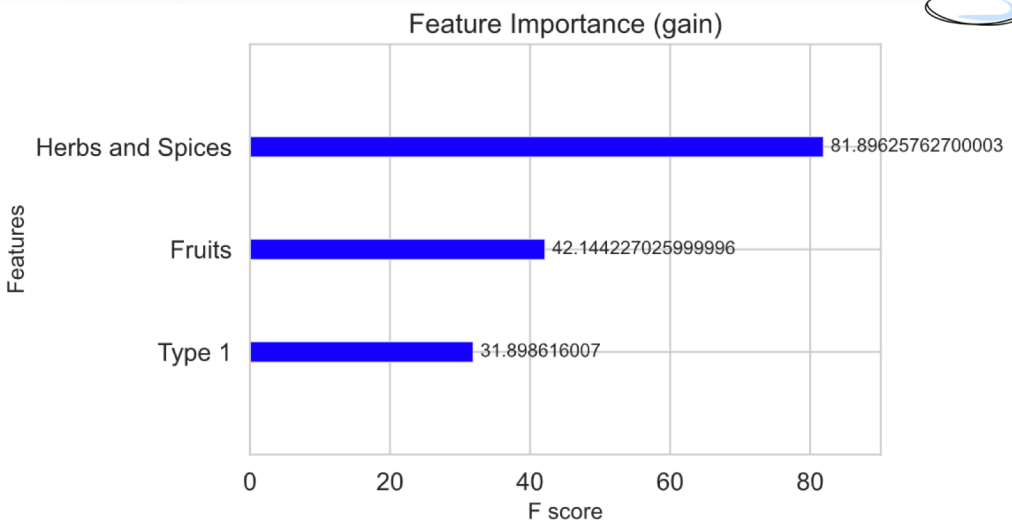
With oversampling

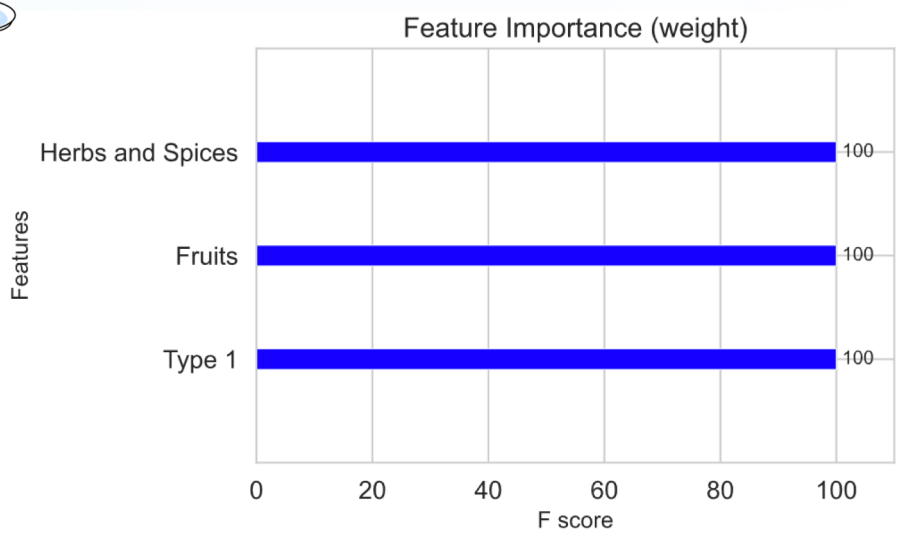
Log-loss Cross-Entropy Train	Val	difference to gauge *overfitting*	
0.6871	0.6861	-0.0010	**Logit**
0.4190	0.7592	0.3401	**KNN**
0.6790	0.6737	-0.0052	**XGBoost**

Although logistic doesn't overfit , the *log loss cross entropy* is suboptimal.

XGBoost revealing important categorical features

With oversampling





More speculative feature engineering

relative concentration
of each compound

Food DB provides

$$\textit{Concentration of compound c in food part P} = \frac{\textit{mg of c}}{\textit{100g of material in food part P}}$$

Another possible engineered feature:

$$\textit{Relative concentration of compound c in food part P} = \frac{\textit{Concentration of compound c in food part P}}{\sum_{\textit{counted compound d} \in \textit{Food part P}} \textit{concentration of compound d}}$$

It **EXAGGERATES**, ***amplifies***, any compound that dominates a food part in terms of compounds that are counted.

It privileges compounds in food parts whose weight is not fully counted in terms of compounds.

Thus
8,000 new continuous features

relative concentration
of each compound in each part of each food

Multiple scenarios

Categorical features

Categorical & continuous features

Continuous features only

PCA on all features

Continuous features w/ PCA

Categorical as-is with continuous features w/ PCA

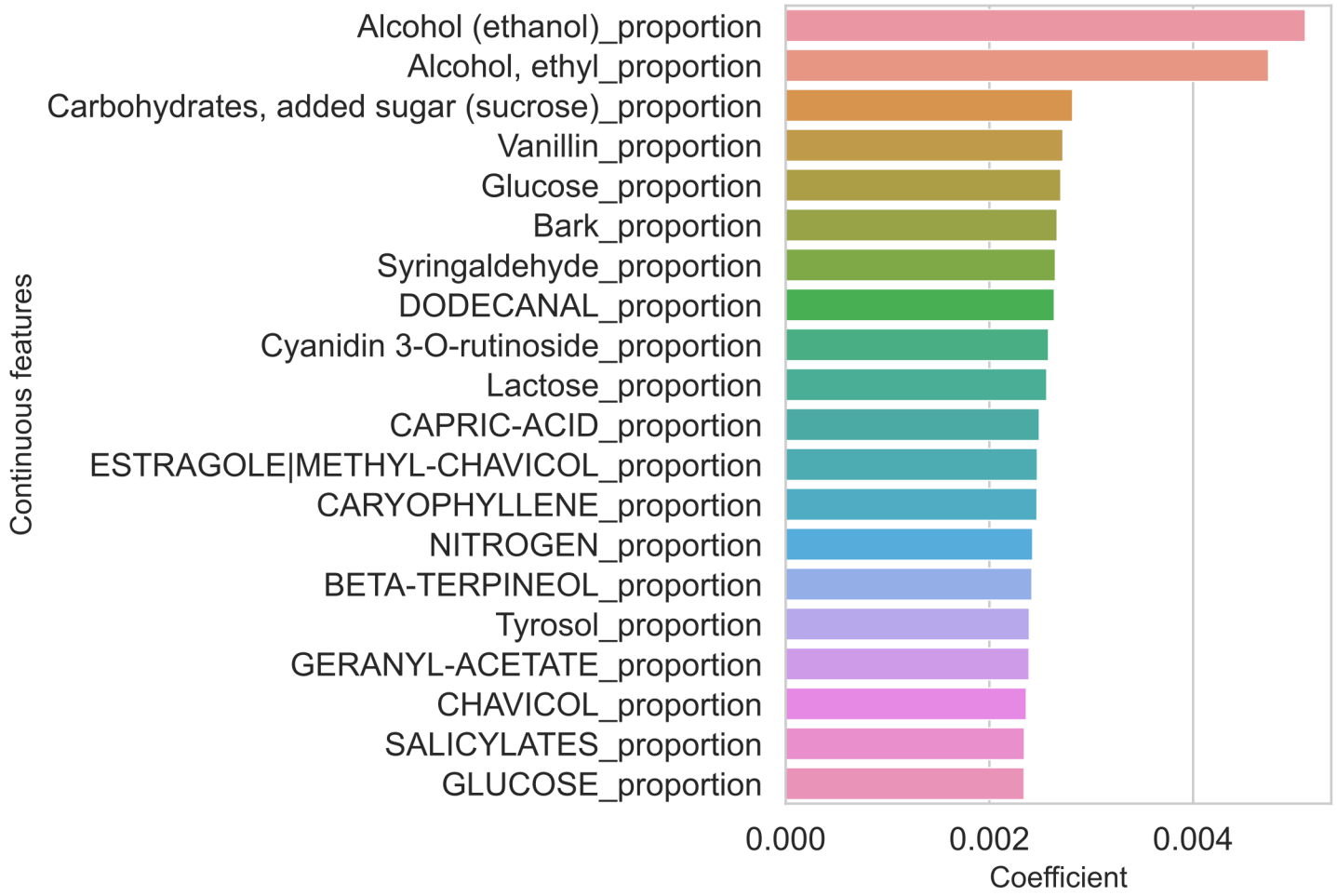
Categorical as-is with continuous features w/ PCA, *oversampled*

All the above with 8,000 engineered continuous features

Logistic Regression (on continuous features) coefficients

contributors

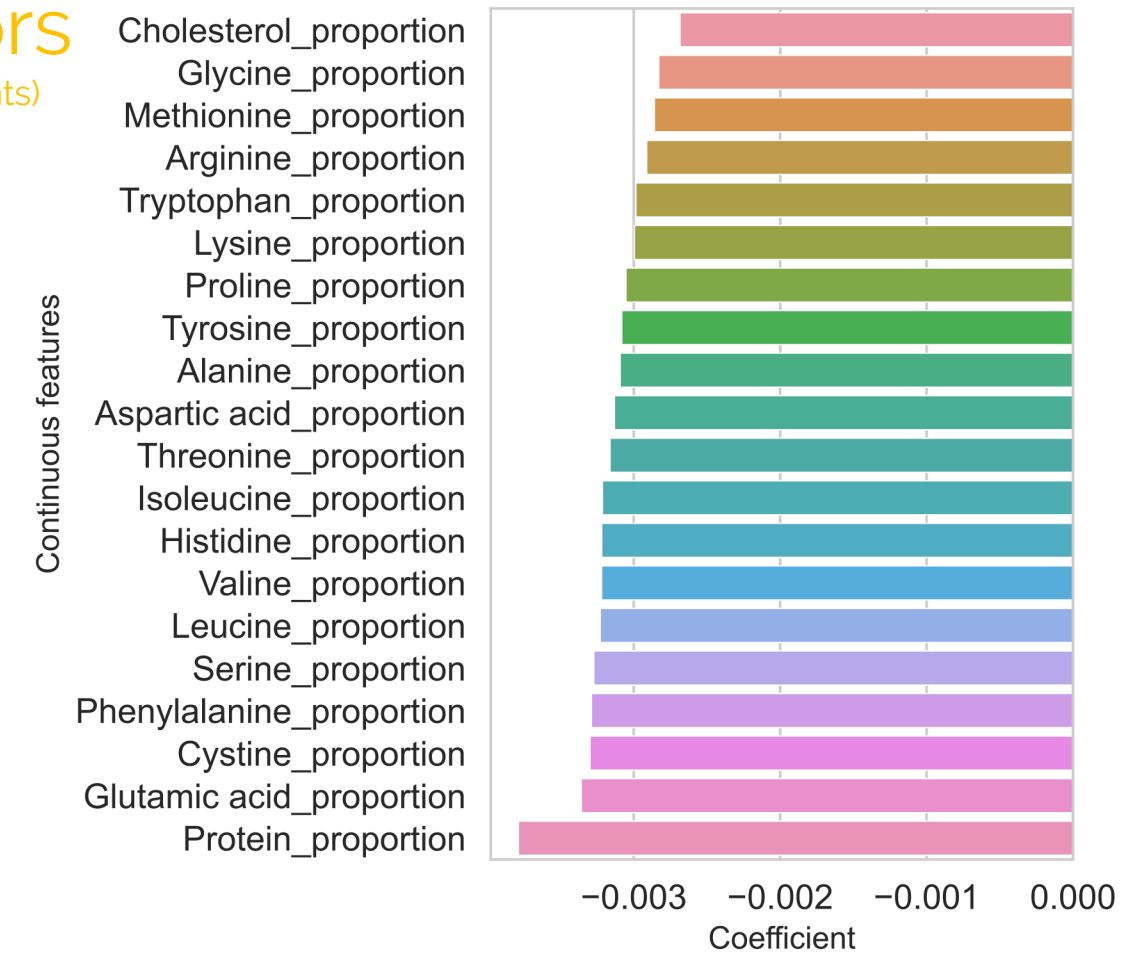
(positive coefficients)



Logistic Regression (on continuous features) coefficients

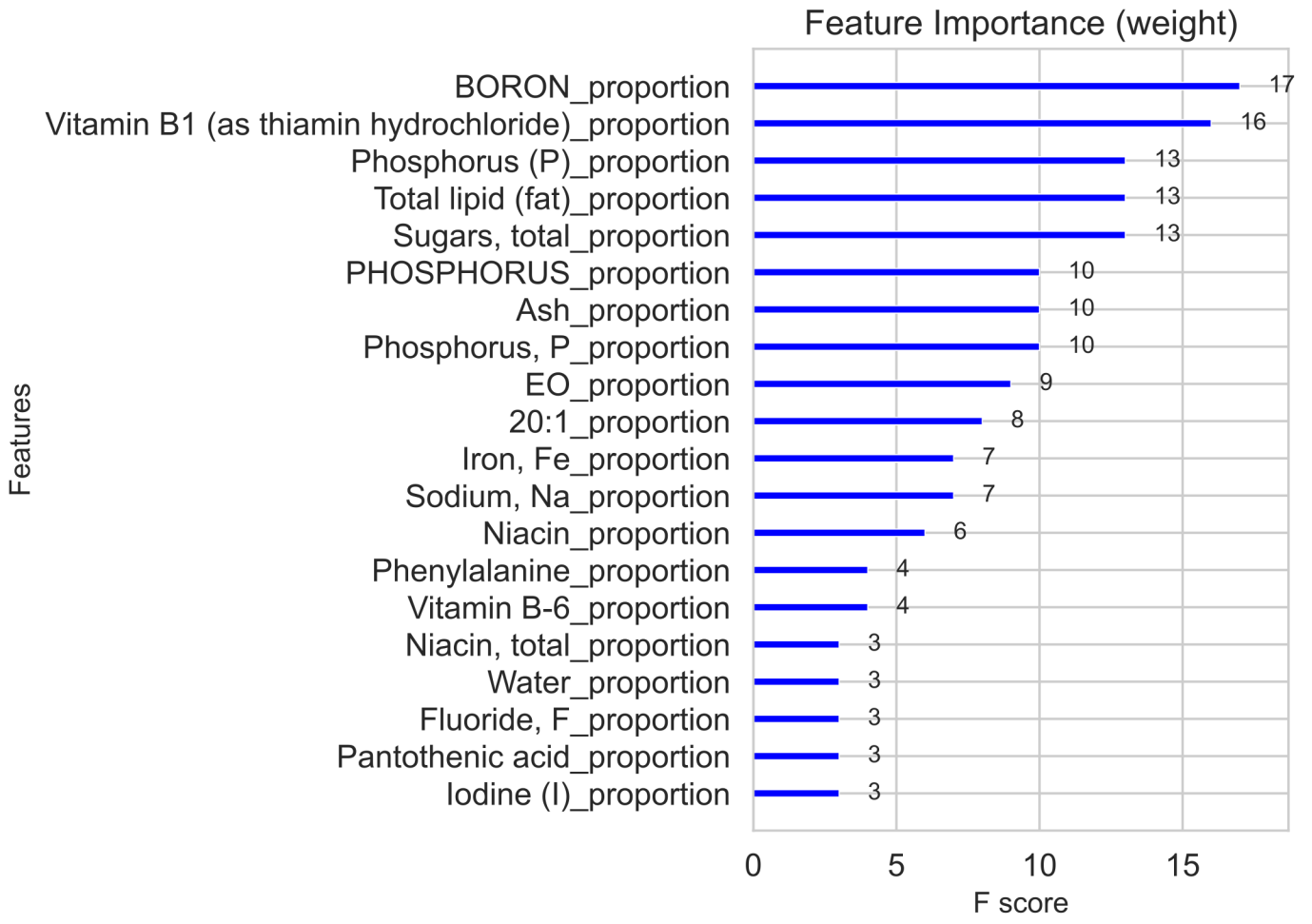
detractors

(negative coefficients)



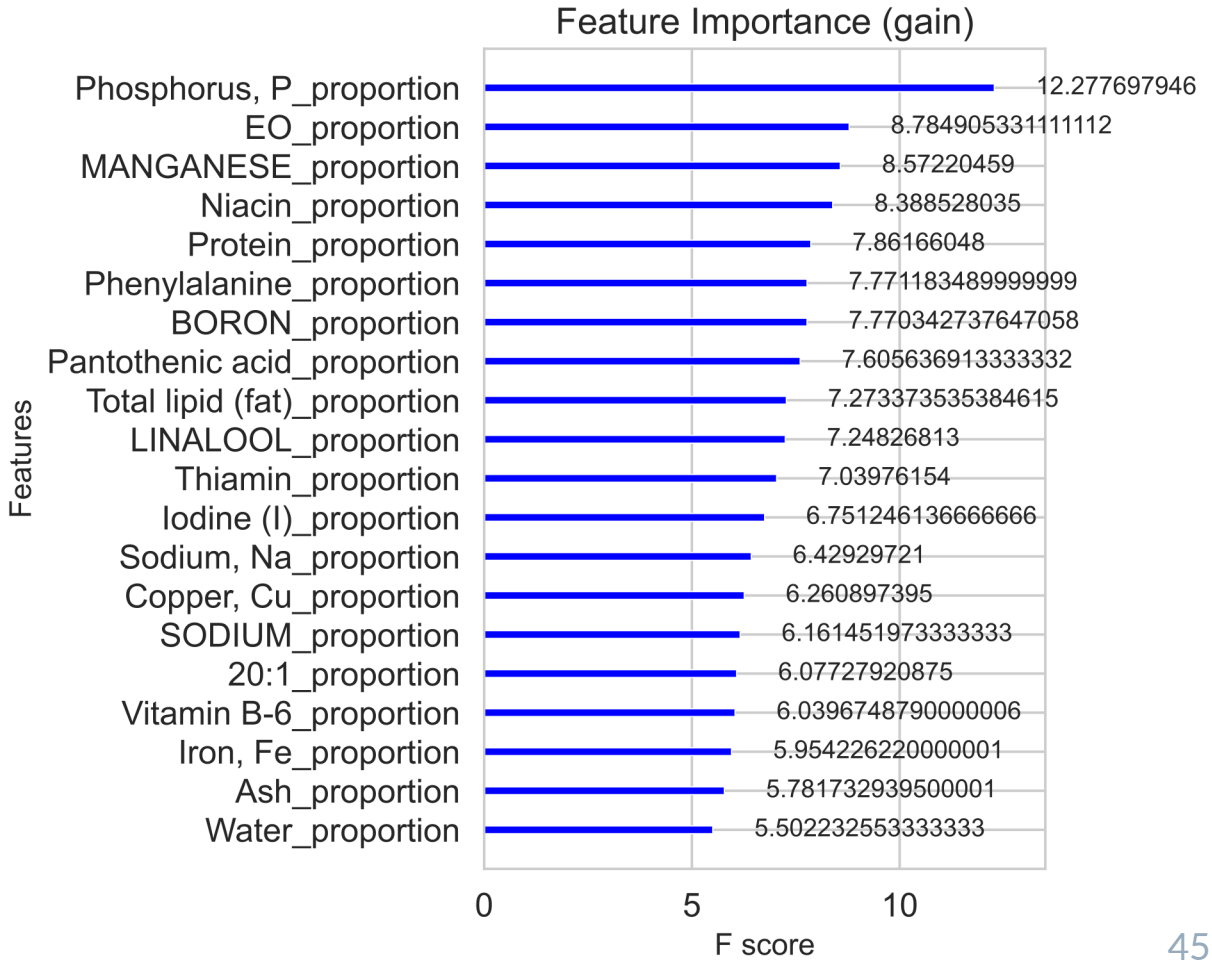
XGBoost 'Important Features'

'weight'



XGBoost 'Important Features'

'gain'



Exploring *False Positives*

to identify candidate ingredients for new cocktails

Satureja montana
(winter savory)
like thyme but stronger

Tinda, Indian squash, mild like a cucumber

Pouteria sapota, *mamay, a Mexican and Central American tree fruit*

ZiziphusJujube,
Chinese date

capers

mustard ?

Intriguing ***false positives***

white cabbage	Brassica oleracea L. var. capitata L. f. alba DC.	Cabbage (Brassica oleracea or variants)...
cheese		Cheese is a food derived from milk that is pro...
ketchup		Ketchup (sometimes catsup in American English ...
pear	Pyrus communis	The pear is any of several tree and shrub spec...
saffron	Crocus sativus	Saffron is a spice derived from the flower of ...
mexican oregano	Lippia graveolens	Lippia graveolens, a species of flowering plan...
winter savory	Satureja montana	Winter savory (Satureja montana) is a perennia...

Future steps

incorporate data on food shape and texture

Consider feature *interactions*

Interactive scatter plot with *predict_proba* against continuous PCA features, with tooltips

Apply to other food-drink combinations, such to coffee or ice cream

To cocktail or not to

Joshua Banks Mailman

Appendix

categorical only

Logit 0.3340 0.3047 -0.0293

train val difference
to gauge *overfitting*

Non-speculative continuous features

Log-loss Cross-Entropy Train	Val	difference	
0.4113	0.4133	0.0020	**Logistic**
0.4097	0.6236	0.2139	**KNN**
27.8380	28.6737	0.8357	**Naïve Bayes**
0.3927	1.0500	0.6574	**Decision Tree**
0.3875	0.6070	0.2195	**Random Forest**
0.3931	0.4021	0.0091	**XGBoost**

Non-speculative cat & pca oversampling

Log-loss Cross-Entropy Train	Val	difference	
0.6871	0.6861	-0.0010	**Logit**
0.4190	0.7592	0.3401	**KNN**
0.6790	0.6737	-0.0052	**XGBoost**

categorical only

to gauge *overfitting*

```
Log-loss:    0.3227  0.3333    diff =  0.0106  logit
Log-loss:    0.4233  1.4834    diff =  1.0601  knn
```

speculative continuous features

```
Log-loss:    0.3714  0.5773    diff =  0.2059  logit
Log-loss:    0.2352  1.5022    diff =  1.2670  knn
Log-loss:   13.6265 15.8574    diff =  2.2309  nb
Log-loss:    0.3344  0.3653    diff =  0.0309  d_tree
Log-loss:    0.6281  2.3566    diff =  1.7285  randomforest
Log-loss:    0.0891  0.3031    diff =  0.2140  xgb
```

speculative continuous features pca

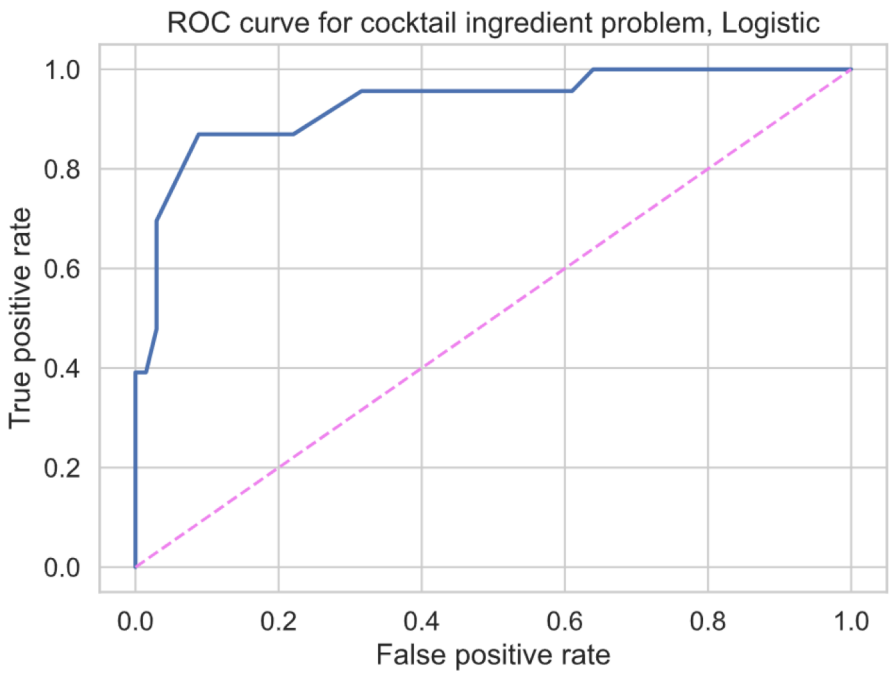
```
Log-loss:    0.3813  0.7984    diff =  0.4171  logit
Log-loss:    0.1974  1.0413    diff =  0.8439  knn
Log-loss:    0.3179  0.3713    diff =  0.0534  xgb
```

speculative continuous features pca oversampling

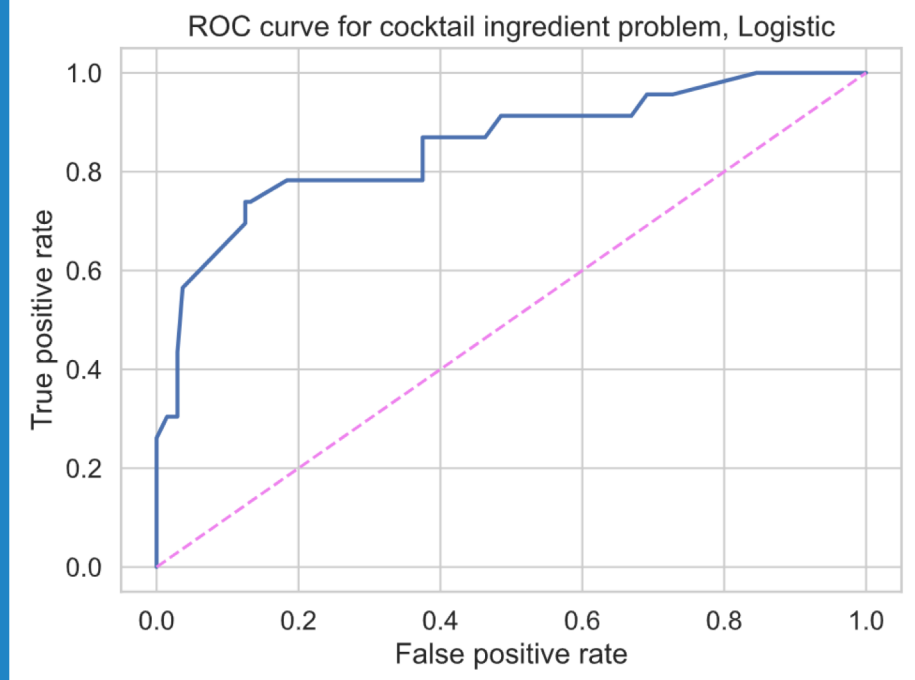
```
Log-loss:    0.5563  1.0635    diff =  0.5072  logit
Log-loss:    0.2893  1.4676    diff =  1.1784  knn
Log-loss:    0.6683  0.6764    diff =  0.0080  xgb
```

Two trials of Logistic Regression on categorical features

ROC AUC score, for Logistic = 0.9309462915601023



ROC AUC score, for Logistic = 0.8516624040920715



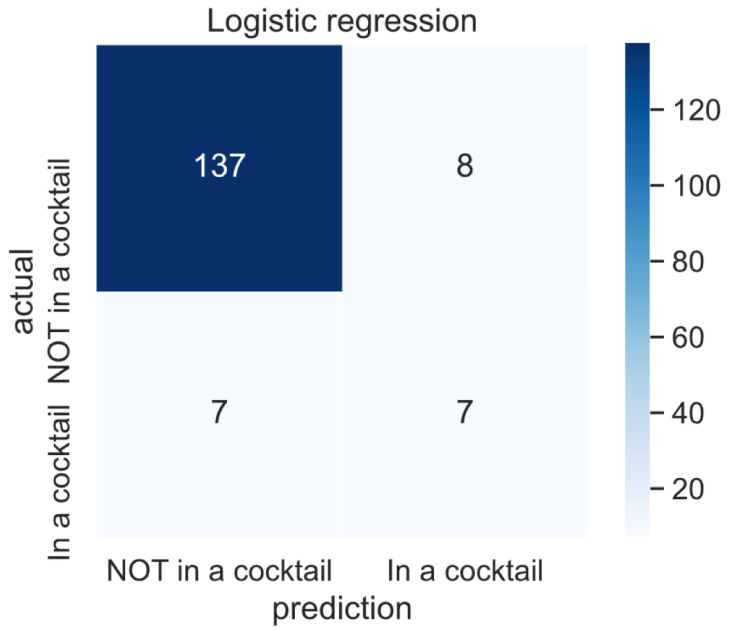
train	val	difference
0.3340	0.3047	-0.0293

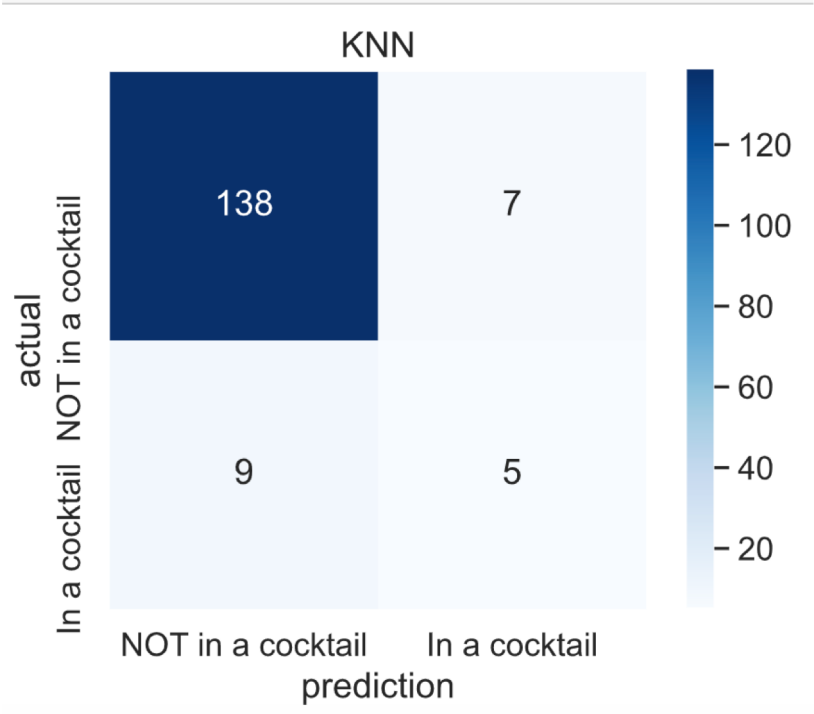
Log-loss: 0.3227 0.3333 diff = 0.0106 logit

Initial results

(early phases of analysis)

```
thresh = .20
make_confusion_matrix(lm, thresh, model_type="Logistic regression")
```





```
The score for logistic regression is
Training:  92.59%
Validation set:  93.71%
```

Logistic

```
label_predict = lm.predict(X_val)
print("Default threshold:")
print("Precision: {:6.4f},   Recall: {:6.4f}"
```

```
Default threshold:
Precision: 0.5000,   Recall: 0.3571
```

```
thresh = 0.2
label_predict = (lm.predict_proba(X_val)[:,1]
print("Threshold: ", thresh)
print("Precision: {:6.4f},   Recall: {:6.4f}"
```

```
Threshold:  0.2
Precision: 0.4667,   Recall: 0.5000
```

KNN

```
label_predict = knn.predict(X_val)
print("Default threshold:")
print("Precision: {:6.4f},   Recall: {:6.4f}".
```

```
Default threshold:
Precision: 0.5000,   Recall: 0.2143
```

```
thresh = 0.2
label_predict = (knn.predict_proba(X_val)[:,1]
print("Threshold: ", thresh)
print("Precision: {:6.4f},   Recall: {:6.4f}".
```

```
Threshold:  0.2
Precision: 0.4167,   Recall: 0.3571
```

```
false_neg
```

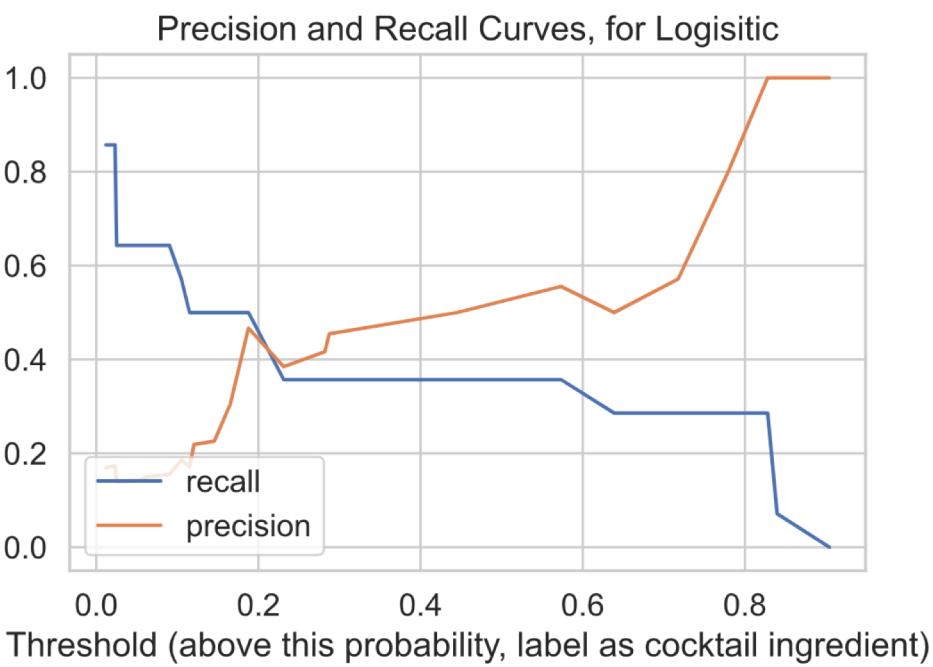
	id	name	food_group	food_subgroup	food_type	in_some_cocktail	predicted
562	577	agave	Vegetables	Vegetables	Type 1	True	False
205	206	ginger	Herbs and Spices	Spices	Type 1	True	False
952	985	sour cream	Milk and milk products	Other milk products	Type 2	True	False
649	669	cream	Milk and milk products	Other milk products	Type 2	True	False
161	162	red raspberry	Fruits	Berries	Type 1	True	False
39	40	pepper	Vegetables	Vegetables	Type 1	True	False
255	256	grapefruit	Fruits	Citrus	Type 1	True	False
647	667	butter	Milk and milk products	Other milk products	Type 2	True	False
56	57	sweet orange	Fruits	Citrus	Type 1	True	False

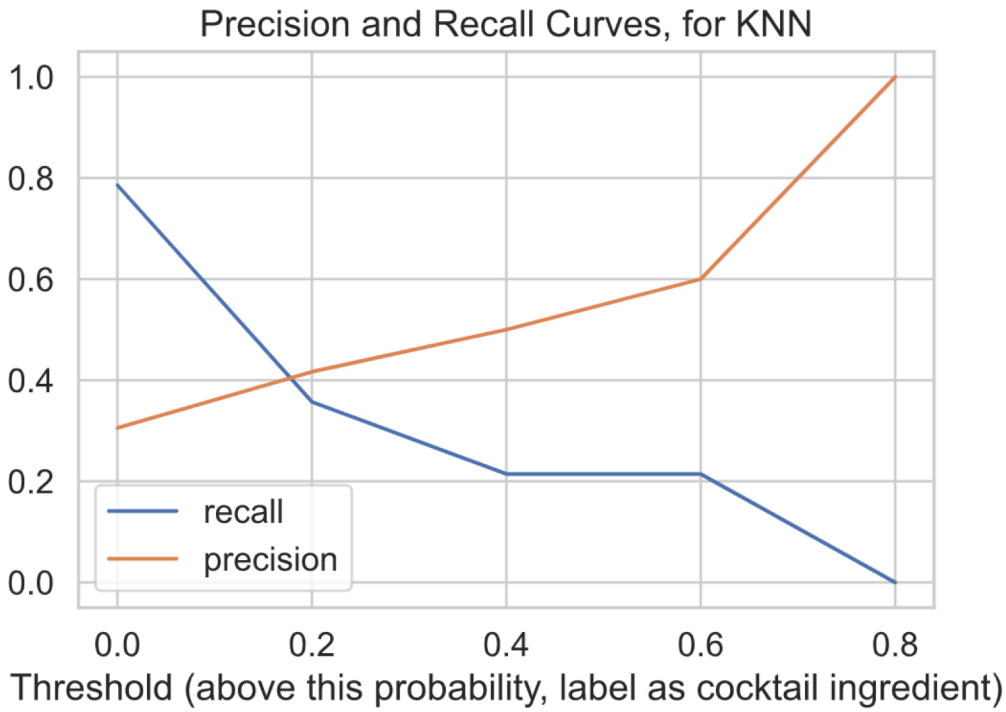
```
false_pos
```

```
<ipython-input-116-40c5ff02703a>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user
e/indexing.html#returning-a-view-versus-a-copy
  df_val['predicted'] = label_predict
```

	id	name	food_group	food_subgroup	food_type	in_some_cocktail	predicted
561	576	common chokecherry	Fruits	Drupes	Type 1	False	True
142	143	prunus (cherry, plum)	Fruits	Drupes	Type 1	False	True
685	708	cocoa liquor	Cocoa and cocoa products	Cocoa products	Type 2	False	True
757	783	adobo	Baking goods	Seasonings	Type 2	False	True
618	633	eggs	Eggs	Eggs	Unknown	False	True


Precision and Recall Curves, for Logisitic
1.0
0.8
0.6
0.4
0.2
0.0
recall
precision
0.0
0.2
0.4
0.6
0.8
Threshold (above this probability, label as cocktail ingredient)


Precision and Recall Curves, for KNN
1.0
0.8
0.6
0.4
0.2
0.0
recall
precision
0.0
0.2
0.4
0.6
0.8
Threshold (above this probability, label as cocktail ingredient)

```
The score for logistic regression is
Training:  92.59%
Validation set:  93.71%
```

F1 score

Logistic

```
label_predict = lm.predict(X_val)
f1_score(label_val, label_predict)
```

```
0.41666666666666663
```

```
label_predict = (lm.predict_proba(X_val)[:, 1] > 0.27)
f1_score(label_val, label_predict)
```

```
0.3703703703703704
```

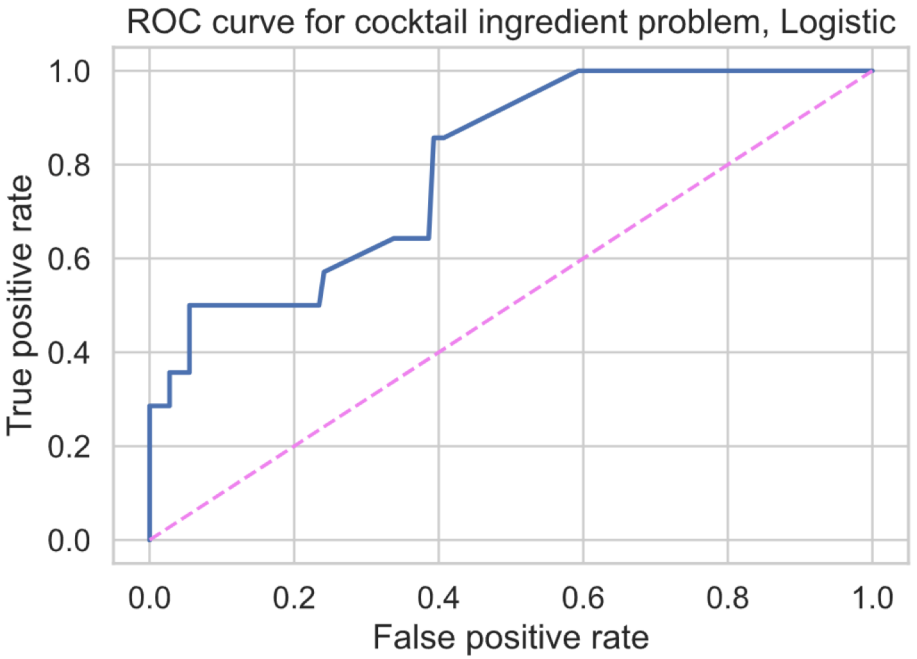
KNN

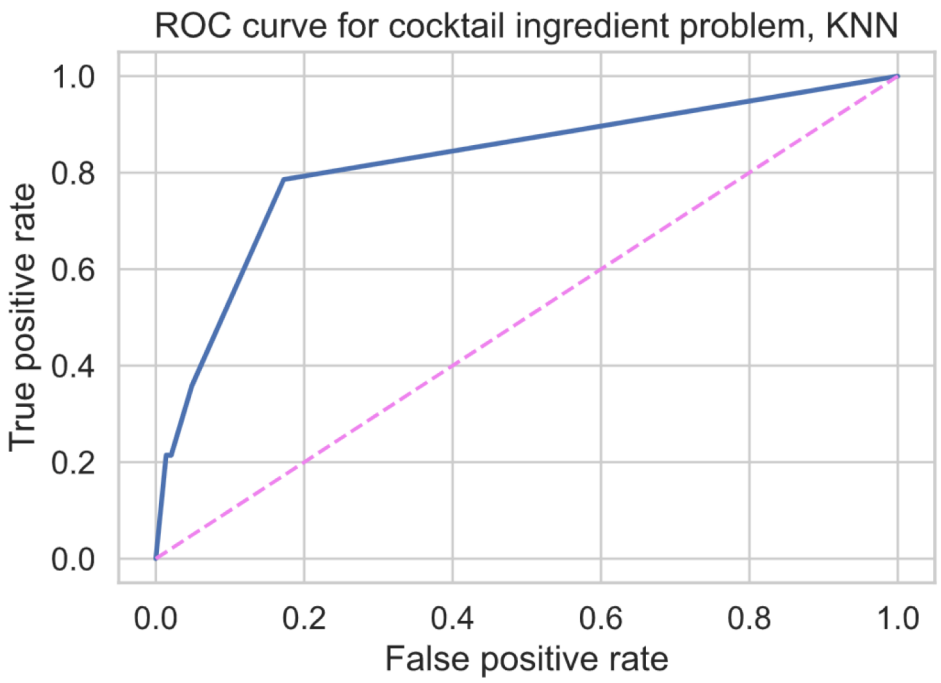
```
label_predict = knn.predict(X_val)
f1_score(label_val, label_predict)
```

```
0.23999999999999996
```

```
label_predict = (knn.predict_proba(X_val)[:, 1] > 0.2)
f1_score(label_val, label_predict)
```

```
0.2857142857142857
```


ROC curve for cocktail ingredient problem, Logistic
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve for cocktail ingredient problem, KNN
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

```
content_df['orig_food_part'].value_counts().head(50)
```

```
Fruit                    8167
Leaf                     8041
Seed                     6806
Plant                    4405
Root                     2252
Shoot                    1491
Essential Oil            1113
Flower                    724
Bulb                      588
Seed Oil                  499
Leaf Essent. Oil          487
Sprout Seedling           433
Rhizome                   422
Stem                      395
Seed Essent. Oil          394
Fruit Juice               350
Bark                      343
Tuber                     332
Pericarp                  324
Tissue Culture            288
Fruit Essent. Oil         229
Silk Stigma Style         203
Shoot Essent. Oil         181
Rhizome Essent. Oil       143
Pt                        142
Resin, Exudate, Sap       137
Pollen Or Spore           127
Petiole                    93
Herb                       91
Root Bark                  91
Endosperm                  88
Stem Bark                  86
Bud                        72
Pericarp Essent. Oil       71
Latex Exudate              68
```

```
content_df.groupby(by='orig_food_common_name').size().sort_value
```

```
orig_food_common_name
Lipid from Arabidopsis (PathBank)                          2043150
Endogenous compounds from human (HMDB)                     1230412
Tea                                                          11816
Milk, skimmed                                                 3002
Milk, 3.25% fat                                               2903
Milk, 2% fat                                                  2717
Other fruits                                                  2649
Milk, 1% fat                                                  2613
Herbs, condiments and spices                                  2569
Bell Pepper                                                   1840
Alliums                                                       1740
Fungi                                                         1722
Milk, raw                                                     1530
Tea, leaves                                                   1414
Tea, ready-to-drink                                           1351
Cereals and bakery products                                   1310
Coffee                                                        1229
Other green vegetables                                         937
Coffee, instant, decaffeinated, powder                         897
Edible fats and oils                                           861
Carrot                                                         794
Alcoholic beverages                                            745
Carrot, raw                                                    720
Legumes                                                        716
Milk                                                           710
Corn                                                           672
Celery                                                         644
Pork, loin with rind, raw                                      639
Coffee, instant, decaffeinated, prepared with water            633
Coffee, instant, prepared with water                           633
Coffee, beverage                                               633
Soybean                                                        631
Coffee, instant, powder                                        606
Squash, all varieties, raw                                     595
Coffee bean, roasted, ground                                   579
Ginger                                                         541
Spinach, raw                                                   534
Celery, raw                                                    529
```