Progress Update 1-15-19

1. What is the class imbalance in numbers?

Counts:

> lactose 291
> non-lactose 87

Proportion:

> lactose 0.77
> non-lactose 0.23

2. Please create two baselines as controls.

The first will just assume that the lactose content in any sample is the mean (or median if you have wild outliers). The second first classifies to none or some and then take the mean as the predicted lactose level. How much better the model is with respect to those baselines?

Results:

> Baseline Mean testing R^2: -0.02
> Baseline Median testing R^2: -0.29
> Baseline Perfect Classification, Mean Regression testing R^2: 0.12

The combined model R^2 value of .53 is an approximate 340% increase over the best R^2 value of the above baseline models.

3. R^2 of 0.38 is quite low. What is your SRC and Mutual Information (MI) here?

Update on model performance:

I realized the grid search cross validation utility I was using was not shuffling the data by default. Since the data is ordered by lactose vs. non-lactose, this was affecting CV results significantly. I turned on shuffling and average testing R^2 for the combined model (Lasso plus logistic regression) has **improved to ~.53**

Average correlations across all features:

> SRC (Spearman Rank Coef.) 0.20
> NMI (Normalized Mutual Information) 0.73
> MI (Mutual Information) 3.19

Individual feature correlations: correlations.xlsx

4. Any pre-processing steps (outliers, normalization, etc.)?

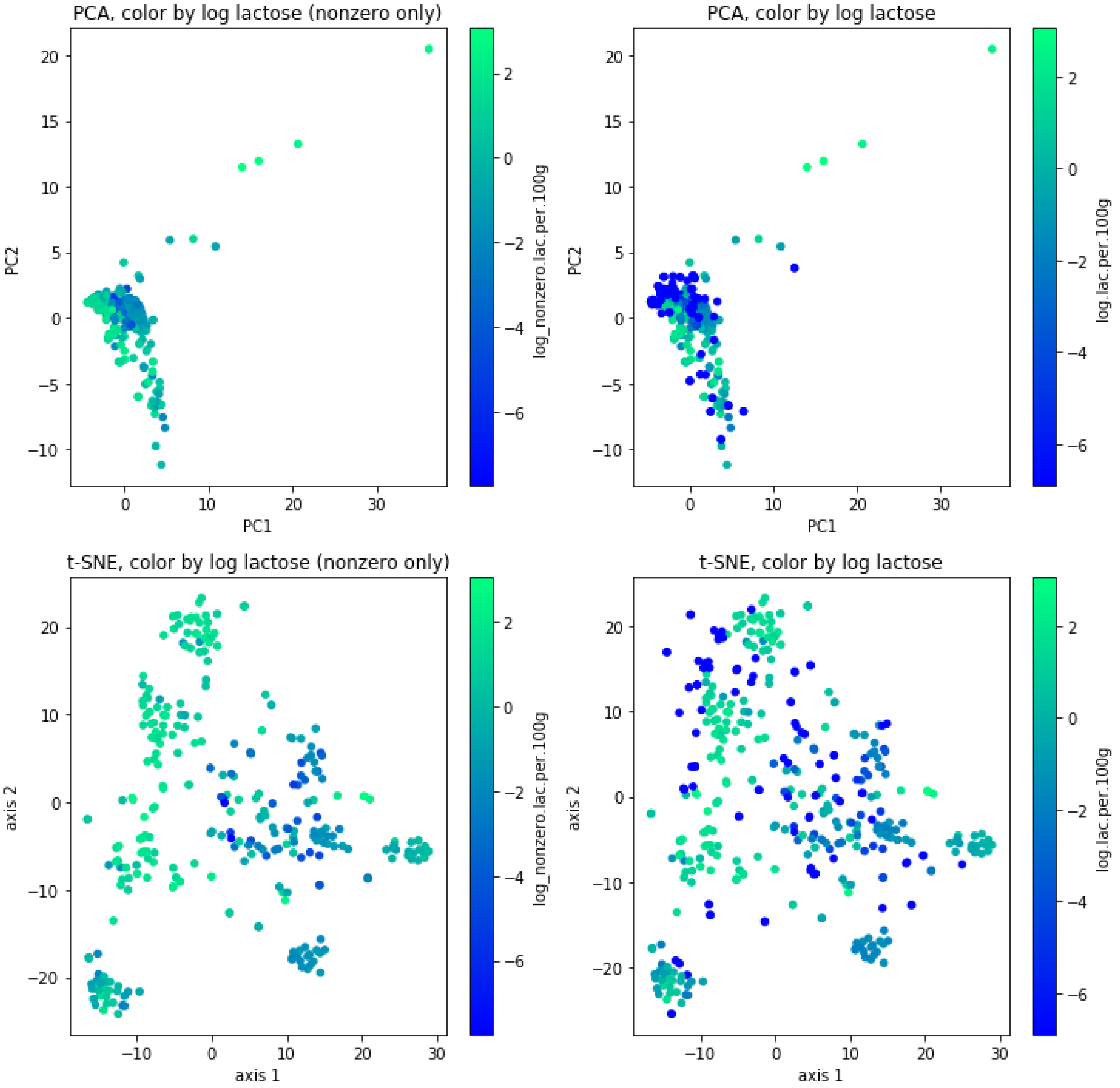
I thought outlier assessment would be a good idea, but was unsure what technique would be appropriate.

The features were normalized so that each column has 0 mean and unit variance.

5. Visualize the samples with PCA/t-SNE

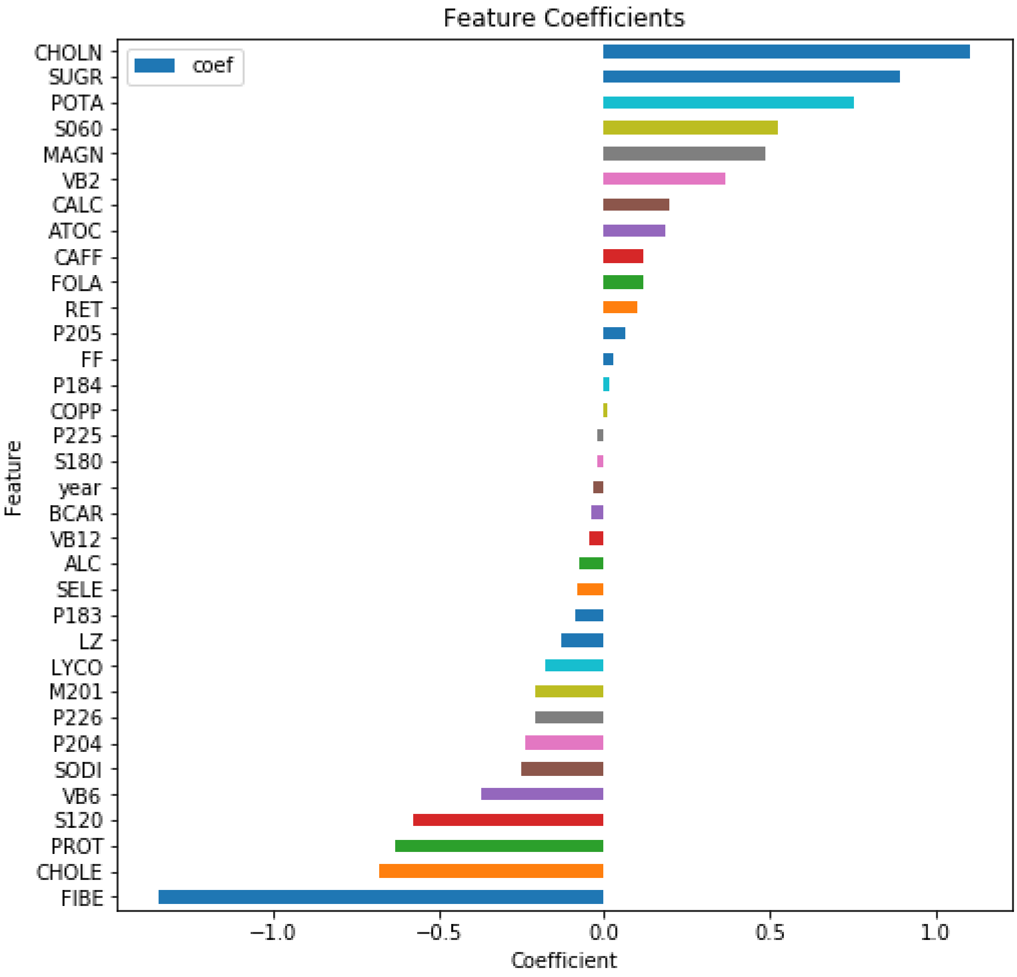
Do the samples cluster?

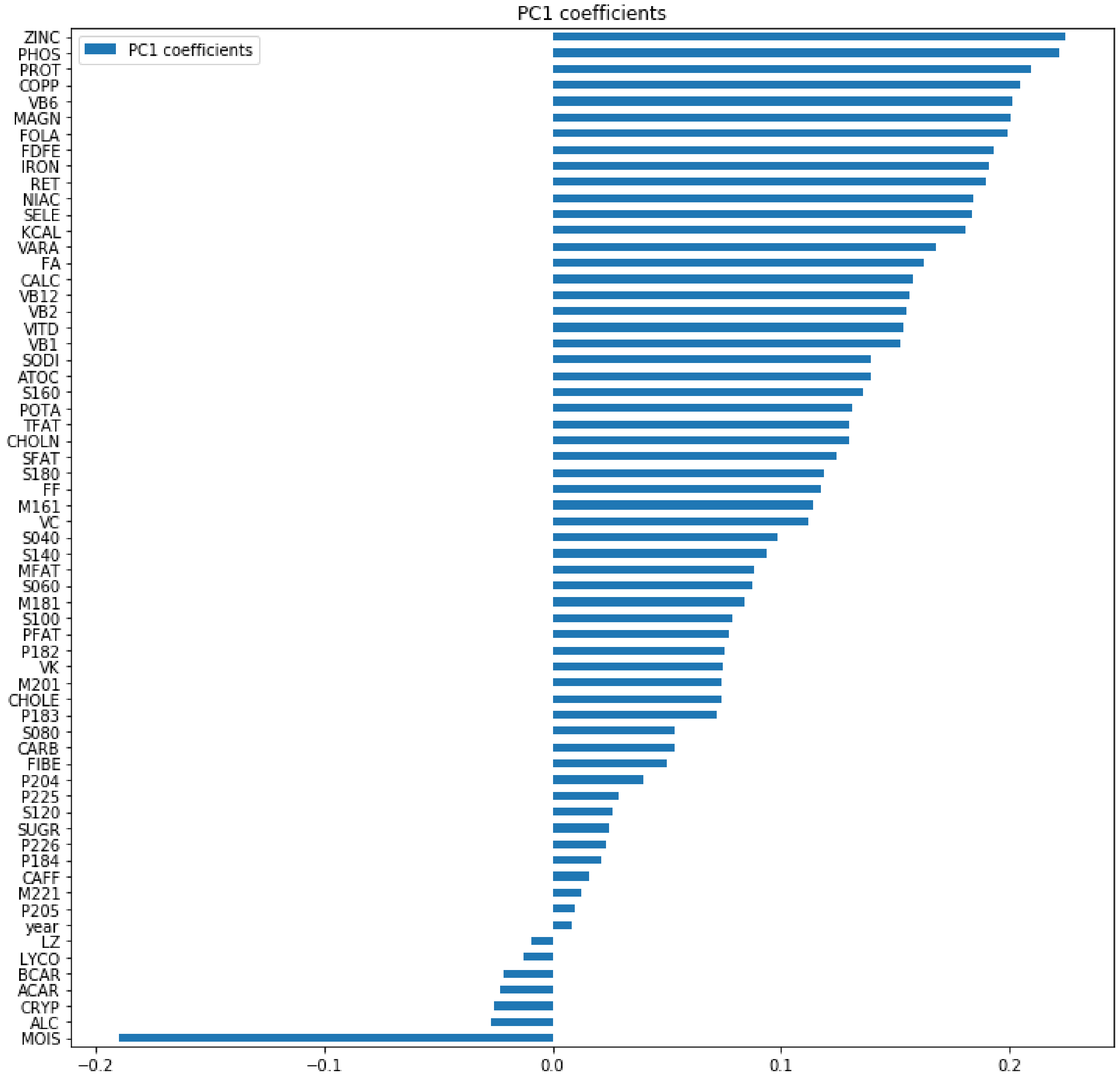
Both nonzero plots (left side) show some organization of the samples by log lactose value. PCA shows one dense area that looks somewhat stratified by lactose value along PC1. PCA also shows some possible outliers going towards the top right corner, but this area might be filled in if we had more samples. T-SNE shows a few distinct clusters and a less-dense area in the middle. For both visualization methods, comparing nonzero (left) and zero-included (right) plots shows that the zero-lactose foods (dark blue on right side plots) appear to be dispersed across the feature space pretty randomly.



Do you get the same weights with PCA on the features as with LASSO?

By inspection of the below, not really. SRC between the PCA weights and bounded Lasso weights is ~0.1


Feature Coefficients
coef
CHOLN
SUGR
POTA
S060
MAGN
VB2
CALC
ATOC
CAFF
FOLA
RET
P205
FF
P184
COPP
P225
S180
year
BCAR
VB12
ALC
SELE
P183
LZ
LYCO
M201
P226
P204
SODI
VB6
S120
PROT
CHOLE
FIBE
Feature
−1.0
−0.5
0.0
0.5
1.0
Coefficient


PC1 coefficients
PC1 coefficients
ZINC
PHOS
PROT
COPP
VB6
MAGN
FOLA
FDFE
IRON
RET
NIAC
SELE
KCAL
VARA
FA
CALC
VB12
VB2
VITD
VB1
SODI
ATOC
S160
POTA
TFAT
CHOLN
SFAT
S180
FF
M161
VC
S040
S140
MFAT
S060
M181
S100
PFAT
P182
VK
M201
CHOLE
P183
S080
CARB
FIBE
P204
P225
S120
SUGR
P226
P184
CAFF
M221
P205
year
LZ
LYCO
BCAR
ACAR
CRYP
ALC
MOIS
−0.2
−0.1
0.0
0.1
0.2