Predicting new locations for Whole Foods

GA Data Science Project

Markus Huber

Data

- IRS data 2012:
  - 73 features
  - 27718 zip codes
- Whole Foods location:
  383 locations

Challenges

- Few number of positive labels and no negative labels
- Different feature scales
- “Prediction”

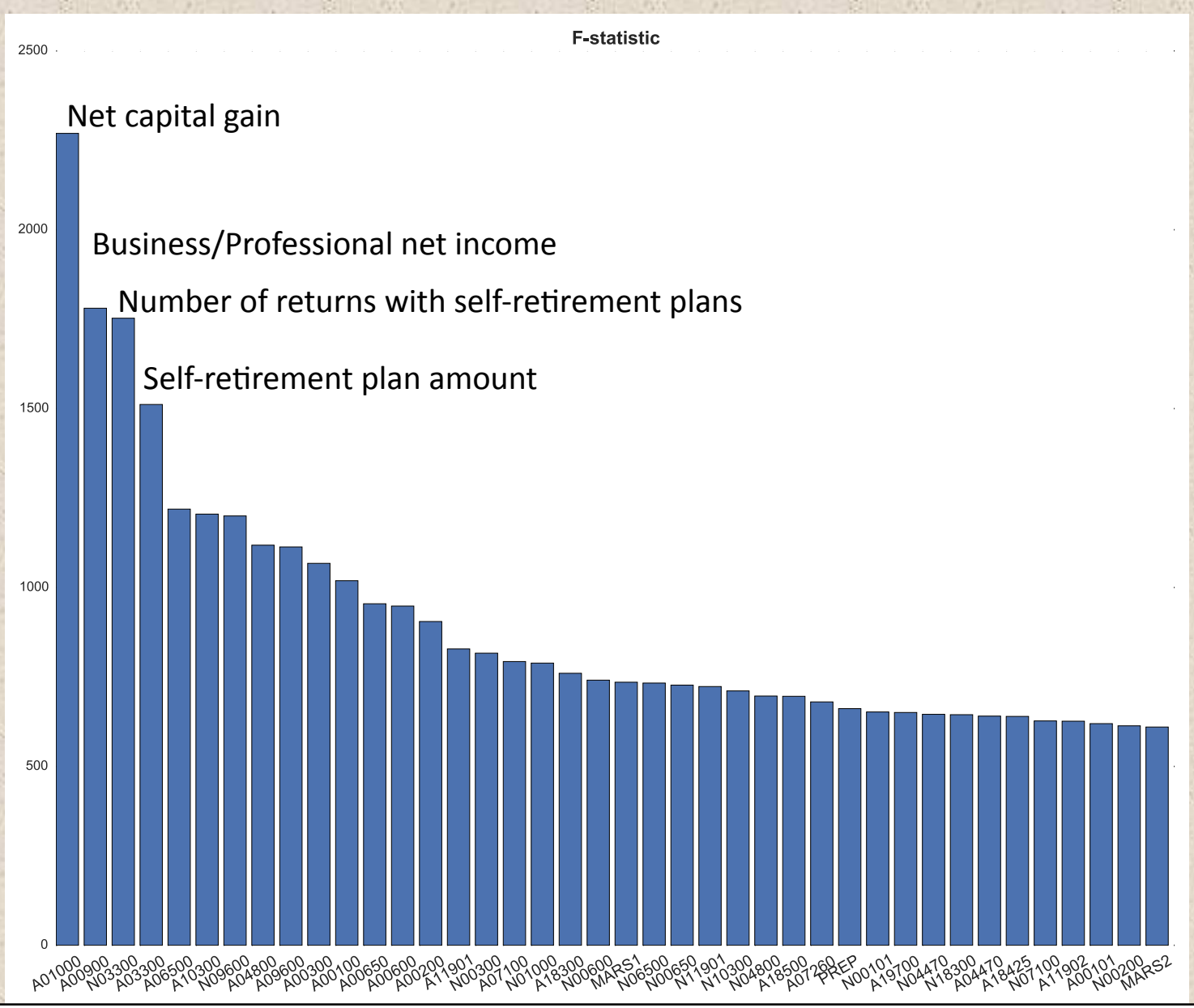
Approaches

- Dimensionality reduction
- Supervised learning

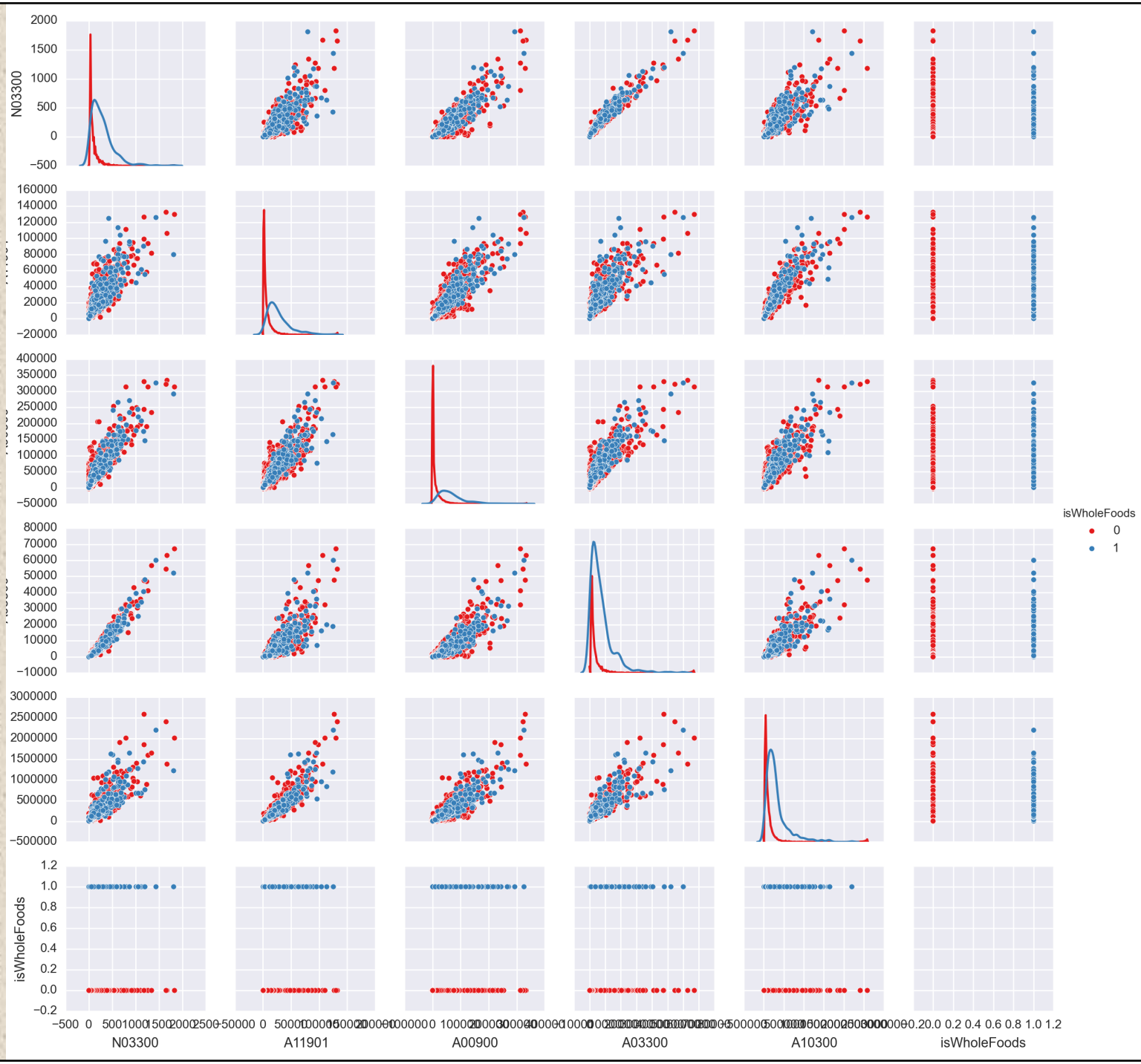
IRS data: 73 features

N1	Number of returns
MARS1	Number of single returns
MARS2	Number of joint returns
MARS4	Number of head of household returns
PREP	Number of returns with paid preparer's signature
N2	Number of exemptions
NUMDEP	Number of dependents
A00100	Adjust gross income (AGI) [2]
N00200	Number of returns with salaries and wages
A00200	Salaries and wages amount
N00300	Number of returns with taxable interest
A00300	Taxable interest amount
N00600	Number of returns with ordinary dividends
A00600	Ordinary dividends amount
N00650	Number of returns with qualified dividends
A00650	Qualified dividends amount [3]
N00900	Number of returns with business or professional net income (less loss)
A00900	Business or professional net income (less loss) amount
SCHF	Number of farm returns
N01000	Number of returns with net capital gain (less loss)
A01000	Net capital gain (less loss) amount
N01400	Number of returns with taxable individual retirement arrangements distributions
A01400	Taxable individual retirement arrangements distributions amount
N01700	Number of returns with taxable pensions and annuities
A01700	Taxable pensions and annuities amount
N02300	Number of returns with unemployment compensation
A02300	Unemployment compensation amount [4]
N02500	Number of returns with taxable Social Security benefits
A02500	Taxable Social Security benefits amount
N03300	Number of returns with self-employment retirement plans
A03300	Self-employment retirement plans amount
N04470/N00101	Number of returns with itemized deductions
N18500	Number of returns with real estate taxes
A18500	Real estate taxes amount
N18300	Number of returns with taxes paid
A18300	Taxes paid amount
N19300	Number of returns with mortgage interest paid
A19300	Mortgage interest paid amount
N19700	Number of returns with contributions
A19700	Contributions amount
N04800	Number of returns with taxable income
A04800	Taxable income amount
N09600	Number of returns with alternative minimum tax
A09600	Alternative minimum tax amount
N07100	Number of returns with total tax credits
A07100	Total tax credits amount
N07180	Number of returns with child and dependent care credit
A07180	Child and dependent care credit amount
N07220	Number of returns with child tax credit
A07220	Child tax credit amount
N07260	Number of returns with residential energy tax credit
A07260	Residential energy tax credit amount
N11070	Number of returns with additional child tax credit
A11070	Additional child tax credit amount
N59660	Number of returns with earned income credit
A59660	Earned income credit amount [5]
N59720	Number of returns with excess earned income credit
A59720	Excess earned income credit (refundable) amount [6]
N06500	Number of returns with income tax
A06500	Income tax amount [7]
N10300	Number of returns with tax liability
A10300	Total tax liability amount [8]
N11901	Number of returns with tax due at time of filing
A11901	Tax due at time of filing amount [9]
N11902	Number of returns with overpayments refunded

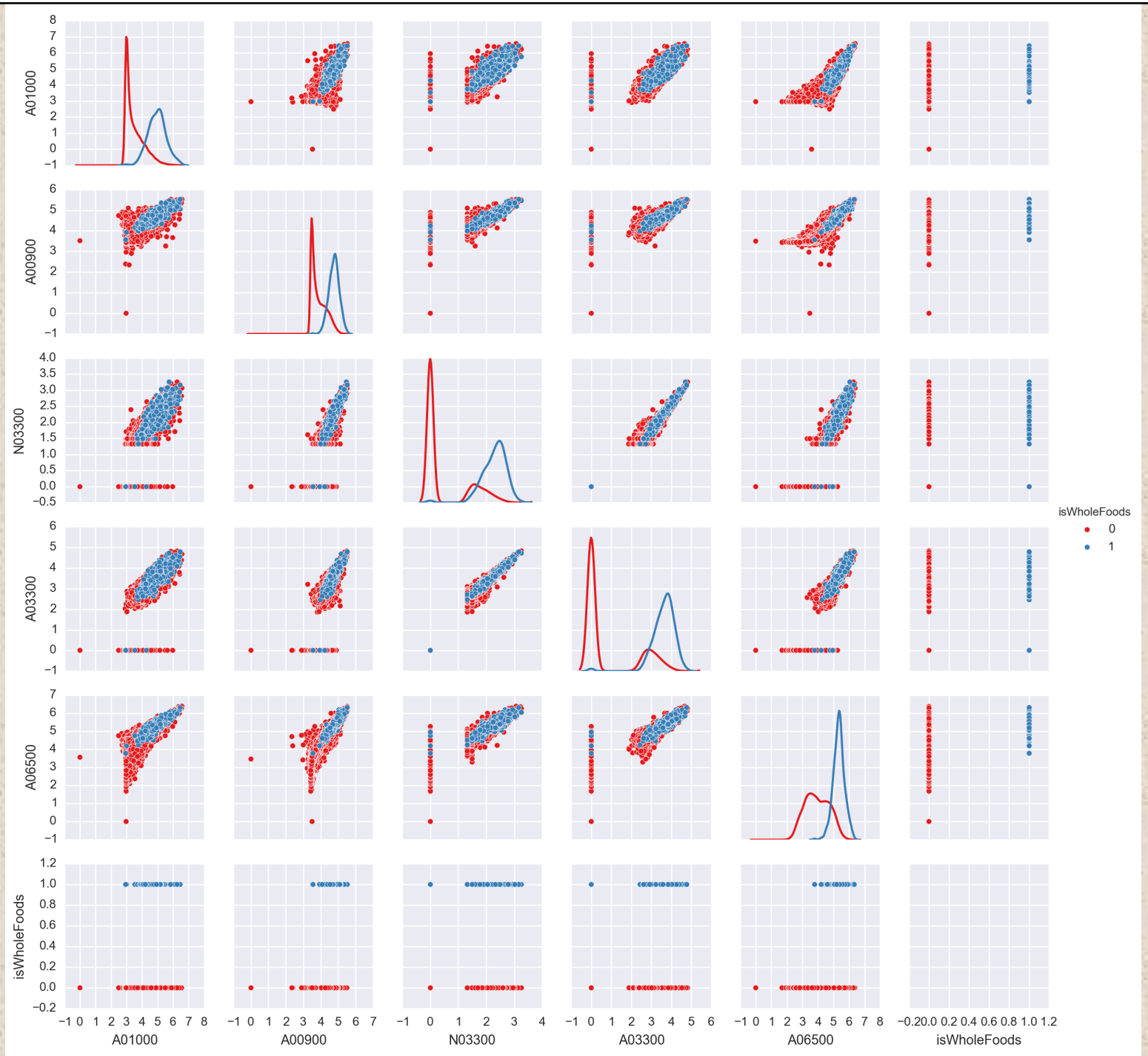
F-statistic


F-statistic
2500
2000
1500
1000
500
0
Net capital gain
Business/Professional net income
Number of returns with self-retirement plans
Self-retirement plan amount
A01000
A00900
N03300
A03300
A06500
A10300
N09600
A04800
A09600
A00300
A00100
A00650
A00600
A00200
A11901
N00300
A07100
N01000
A18300
N00600
MARS1
N06500
N00650
N11901
N10300
N04800
A18500
A07260
PREP
N00101
A19700
N04470
N18300
A04470
A18425
N07100
A11902
A00101
N00200
MARS2

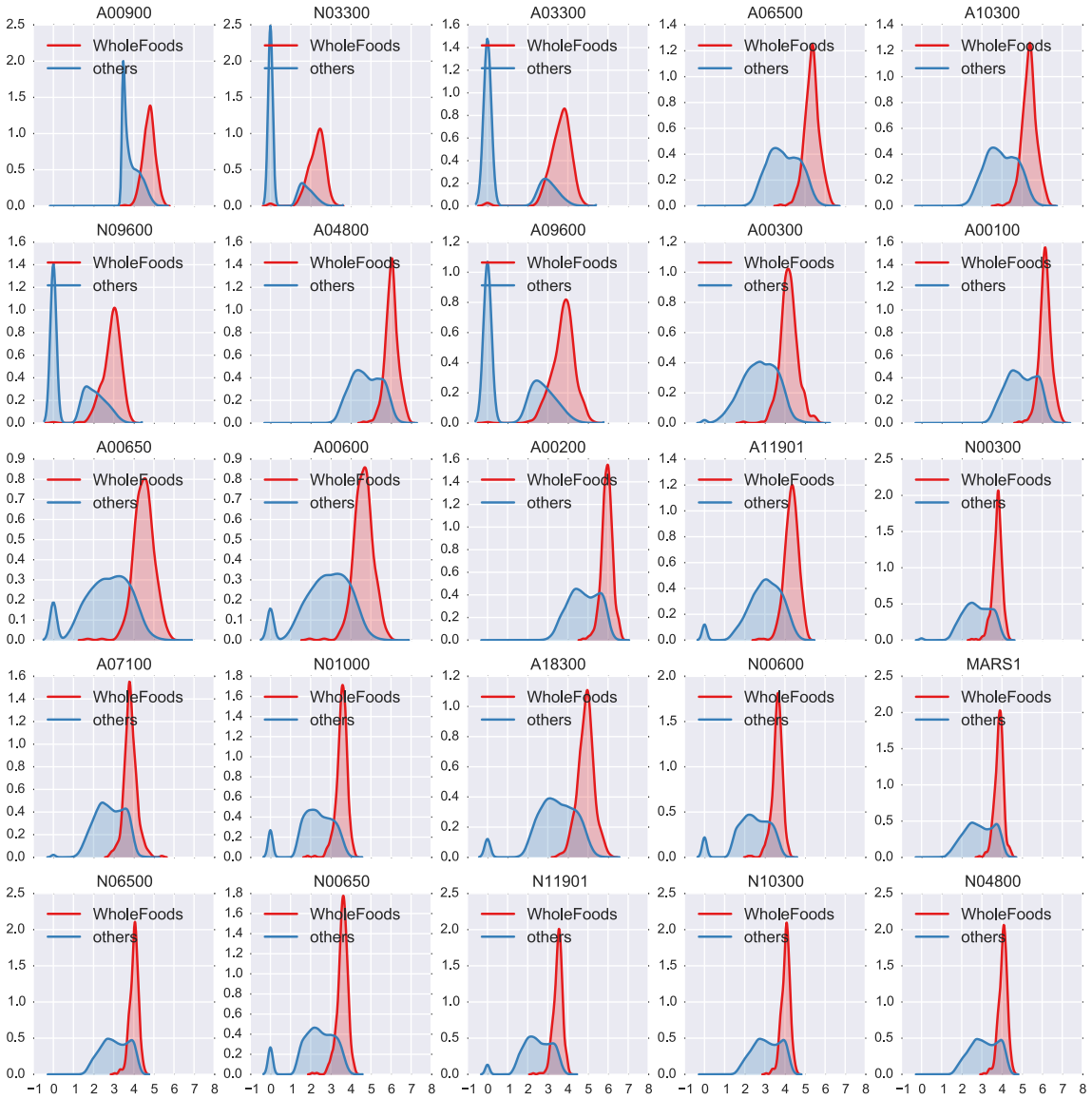
Original Units



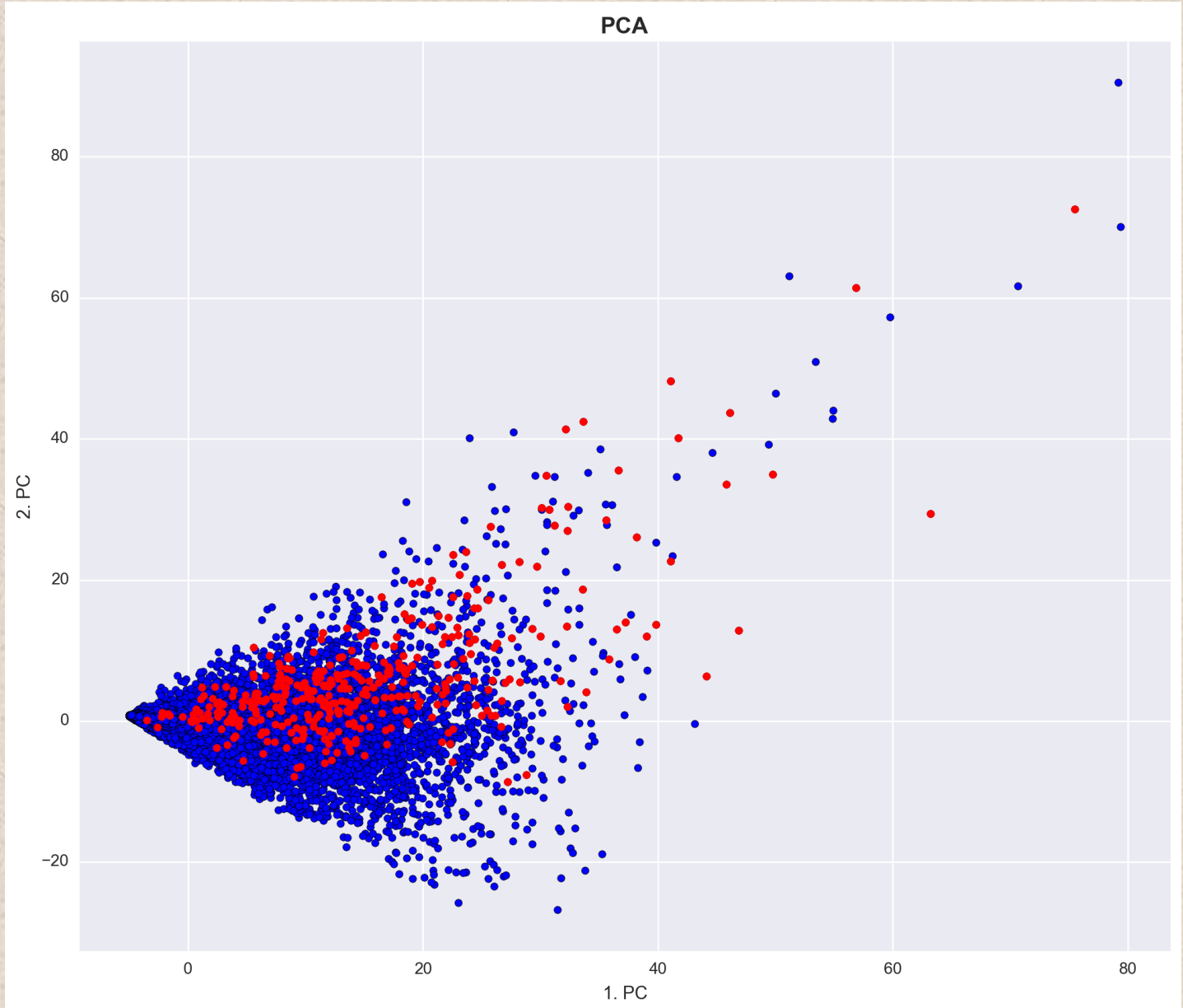
Log-transformed



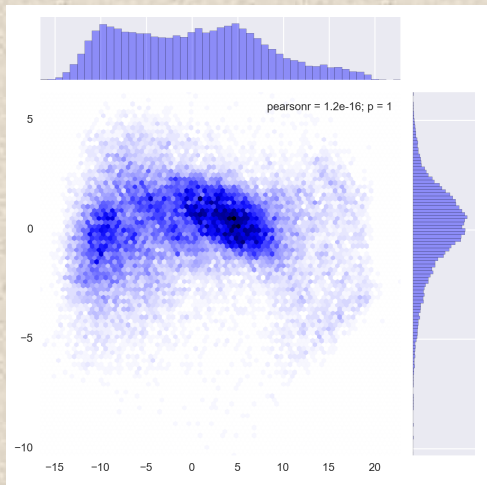
Histograms

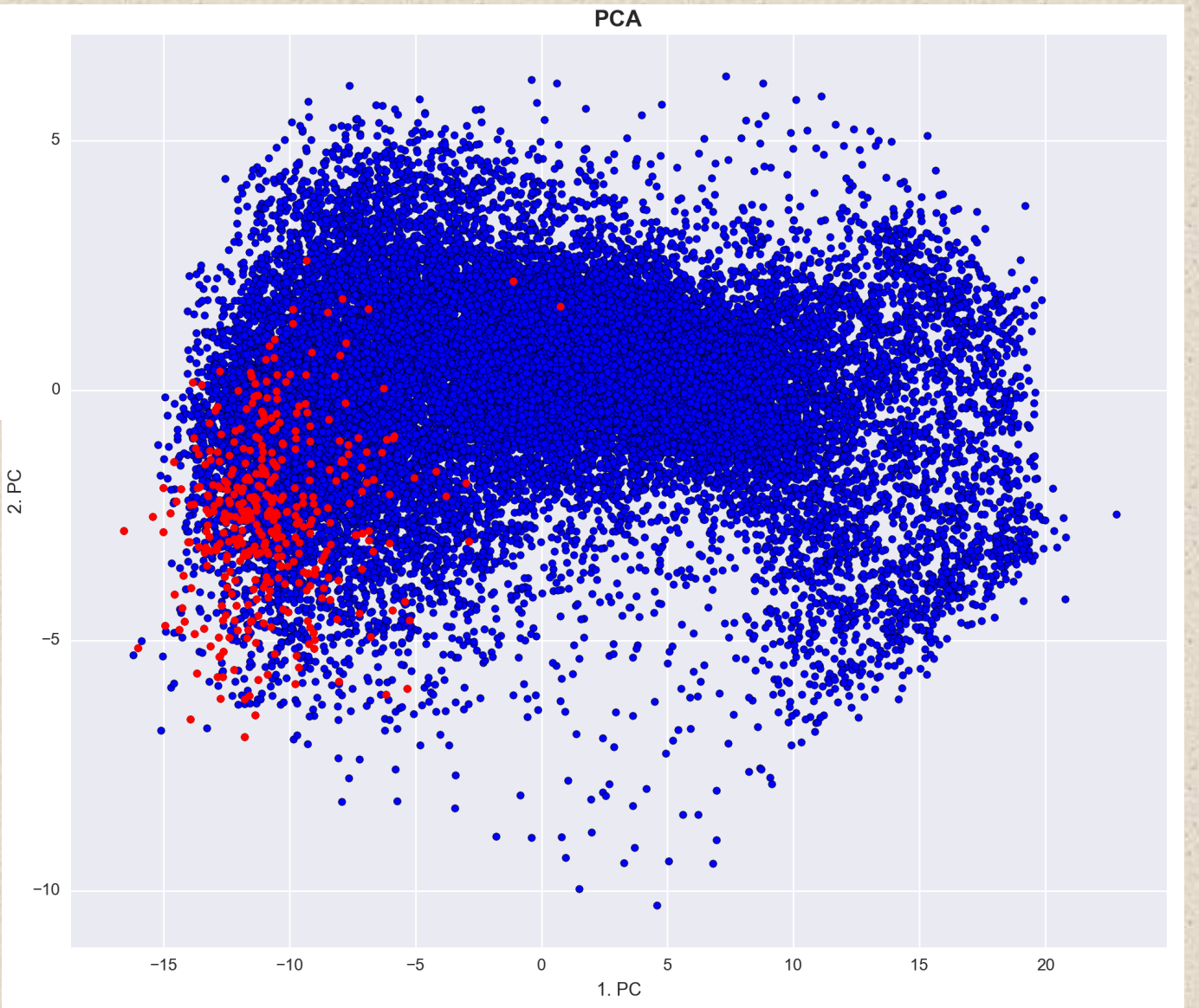

A00900
N03300
A03300
A06500
A10300
N09600
A04800
A09600
A00300
A00100
A00650
A00600
A00200
A11901
N00300
A07100
N01000
A18300
N00600
MARS1
N06500
N00650
N11901
N10300
N04800
WholeFoods
others

PCA:



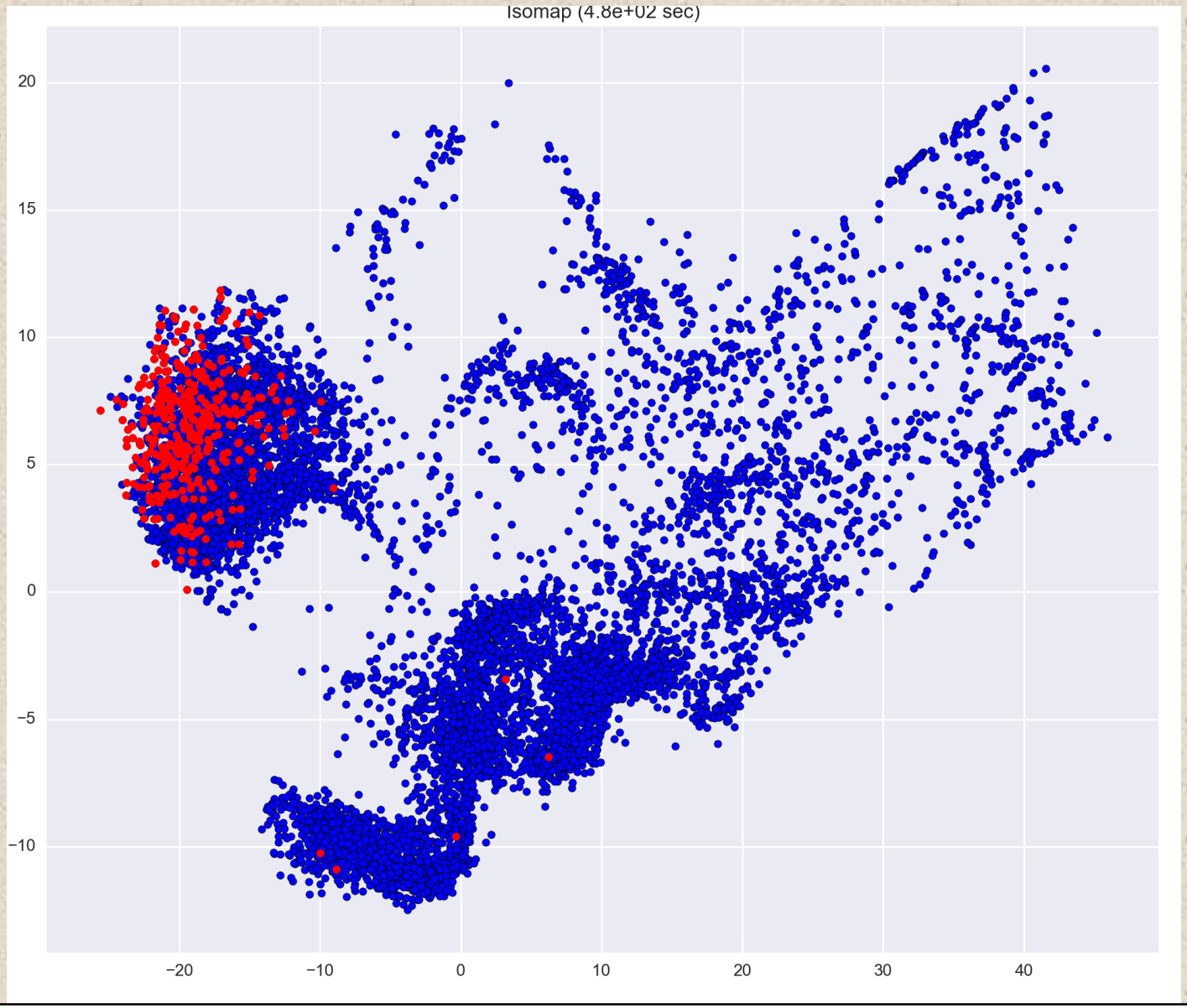
PCA:


pearsonr = 1.2e-16; p = 1
5
0
−5
−10
−15
−10
−5
0
5
10
15
20


PCA
2. PC
1. PC
5
0
−5
−10
−15
−10
−5
0
5
10
15
20

Non-linear dimensionality reduction: Isomap

intrinsic geometry of the data manifold (subsample)



Model: Logistic Regression

-Log-transformed features

-Cross-validation

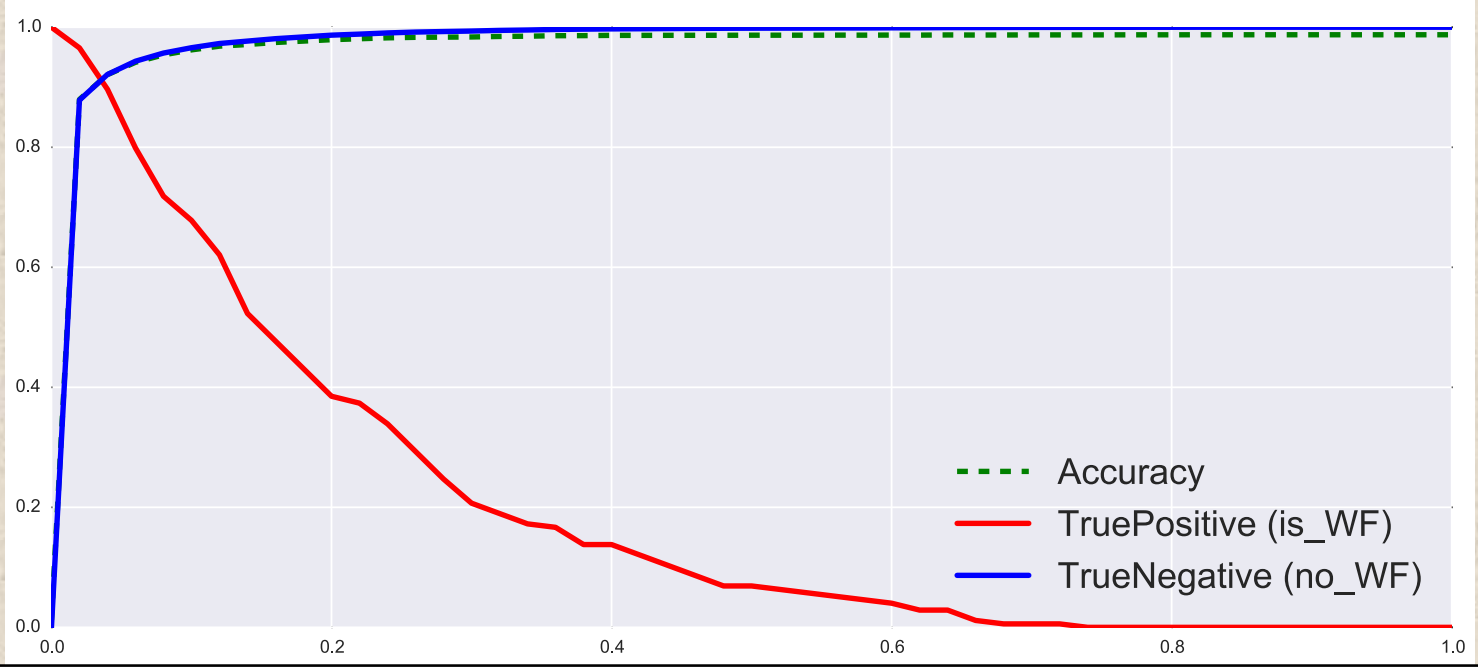
-Optimize for True Positives

-ROC curves, TruePositive curve, LR thresholds

-Consider False Positives as new locations for new locations for Whole Foods

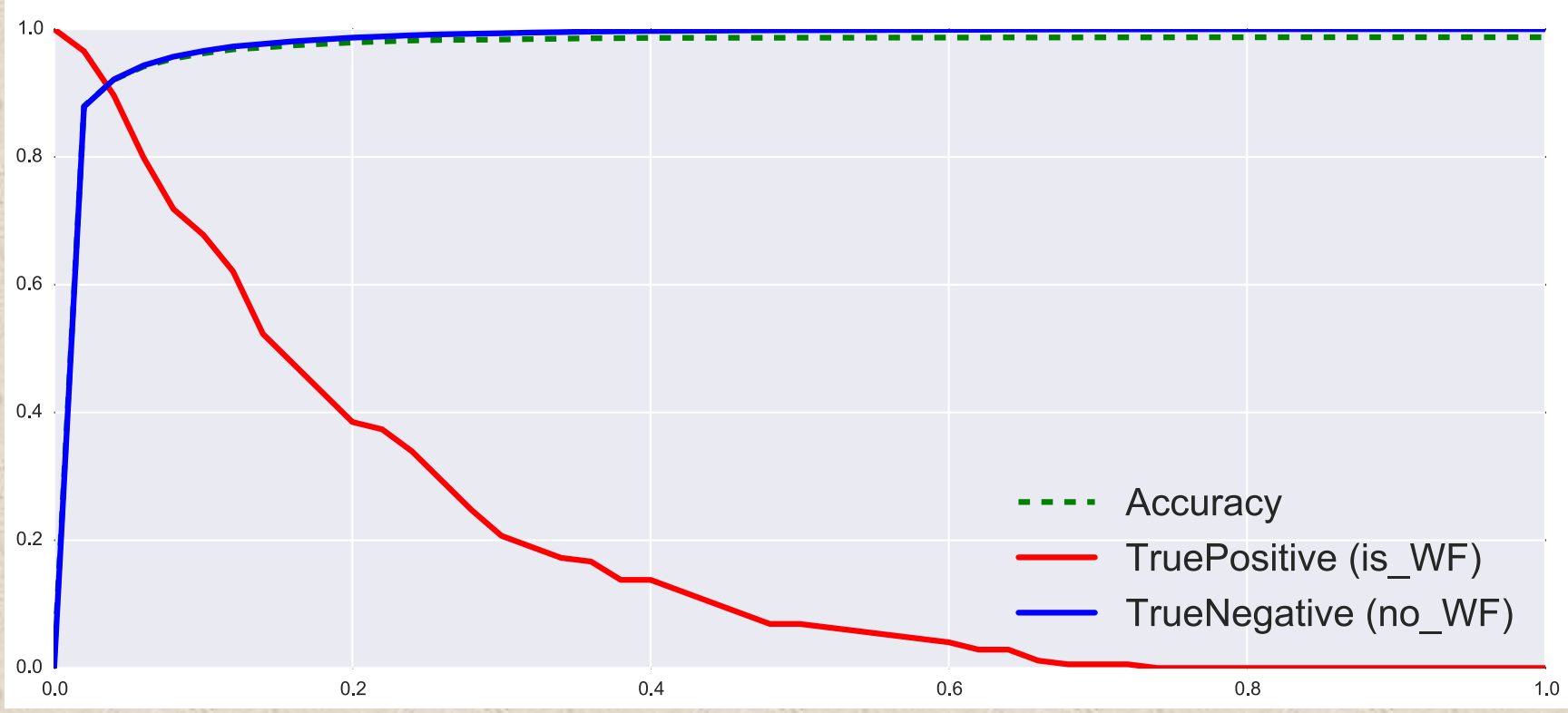
Logistic Regression:

		predicted	
		no_WF	is_WF
actual	no_WF	8094	124
	is_WF	83	15



Logistic Regression Thresholds:

TP: 97%
TN: 88%



Summary:

-Model can reduce potential zip codes by 88%.

-Prediction power of the data may be reached.

-More features may be necessary to exclude more zip codes.