Predicting Diabetes and Obesity

Team 6
Alex Huang, Brian Holligan, Christie Li

Goal

- Identify predictive factors of diabetes and obesity
- Help improve public health decision making
- Use machine learning classification methods

The Data Sets

1. CDC Health and Nutrition Survey
   from 2007 - 2014
2. Early Childhood Study for Kindergarteners
   from 2010 - 2011
3. Adolescent Health Study for 7th -12th graders
   from 1994

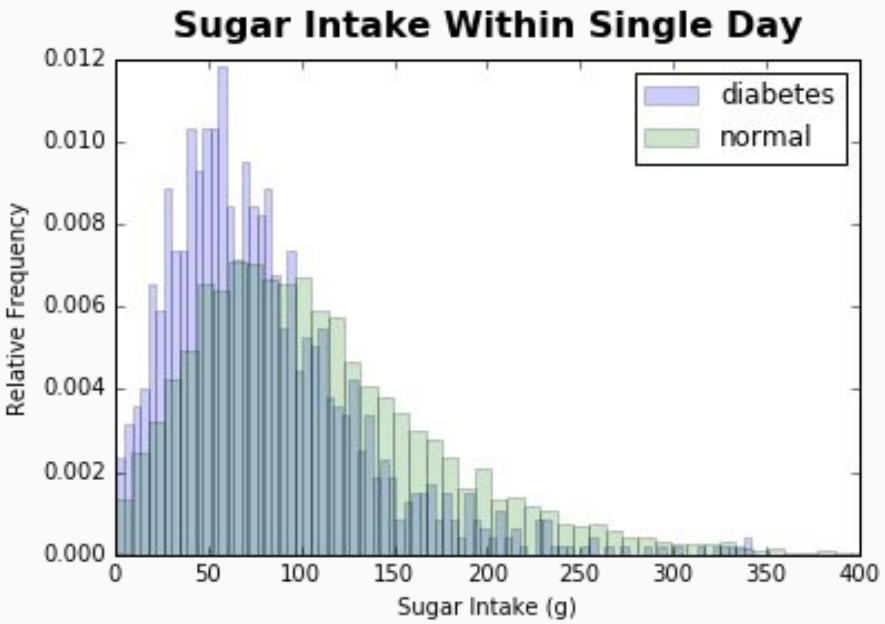
Our Classification Problem

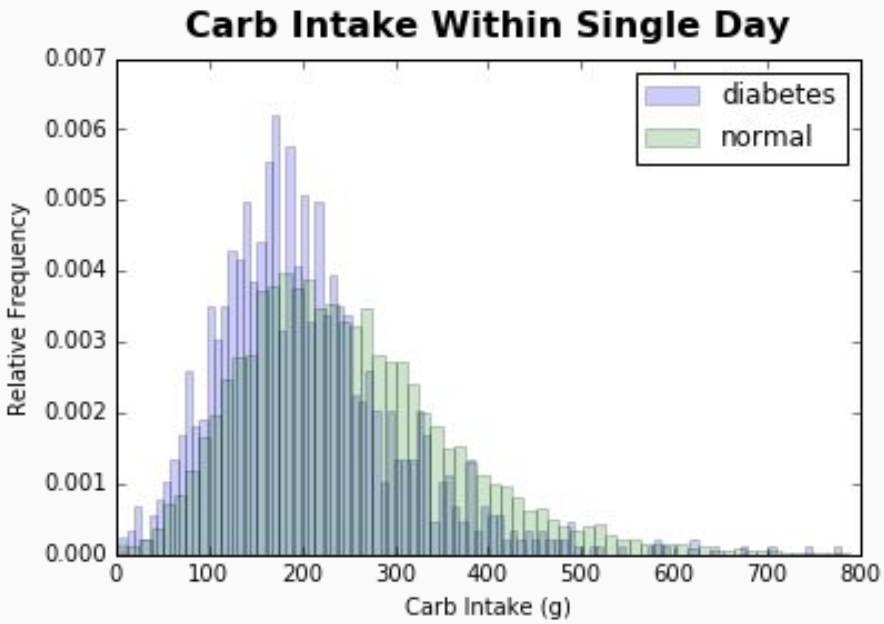
Treat diabetes and obesity as binary classifiers. Obesity defined as BMI over 95% of the national average for that age.

Models:

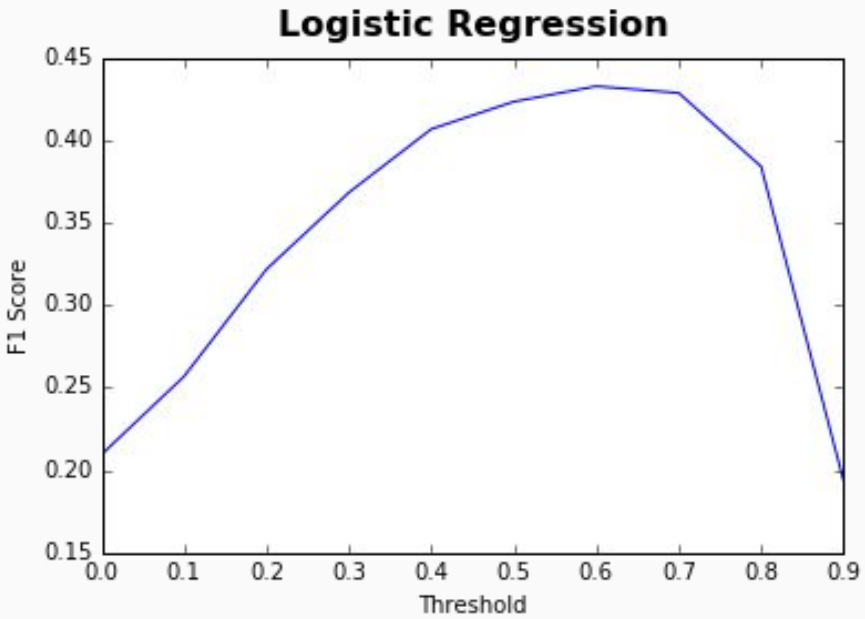
- Logistic Regression
- Random Forest
- Naive Bayes
- XGBoost

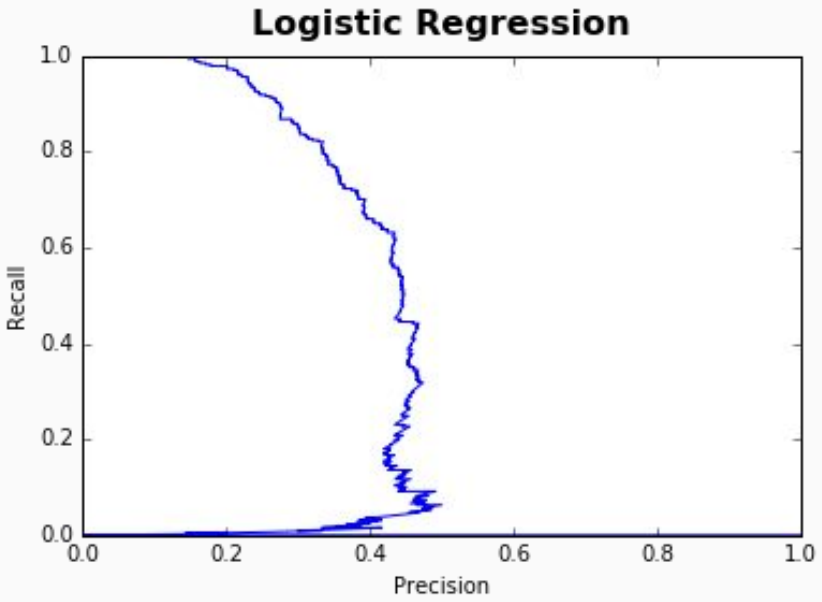
Feature Distributions - Diabetes Data





Logistic Regression - Diabetes Prediction





Some Logistic Regression Coefficients

Age: 7.56

Sugar: -5.46

Fat: 2.06

Cholesterol: 1.58

Protein: 1.58

Carbohydrates: -1.46

Polyunsaturated fat intake: 0.79

High blood pressure: 0.300

Hours watching TV: 0.22

Saturated fat intake: 0.18

Some Surprising Coefficients

Sugar: no effect

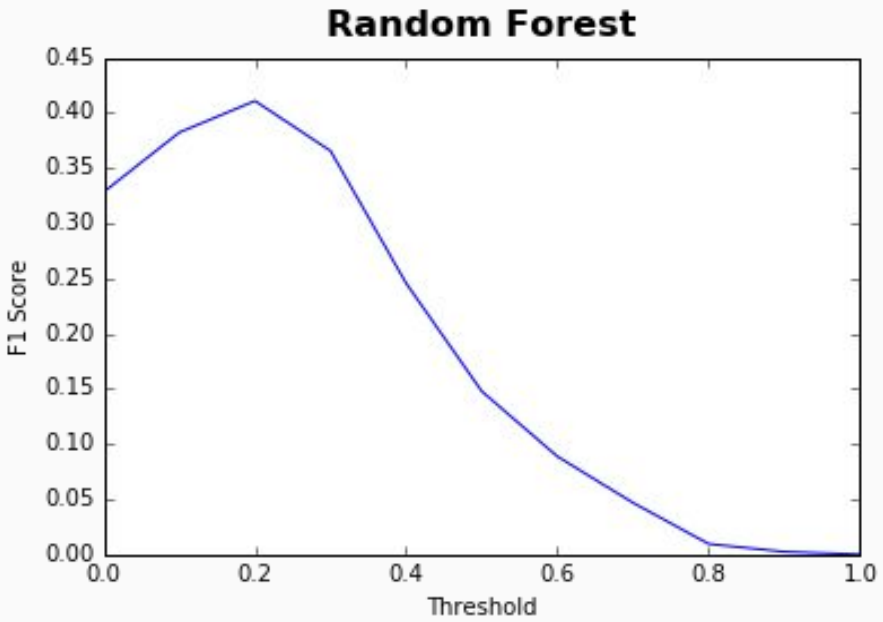
Moderate Activity: increase

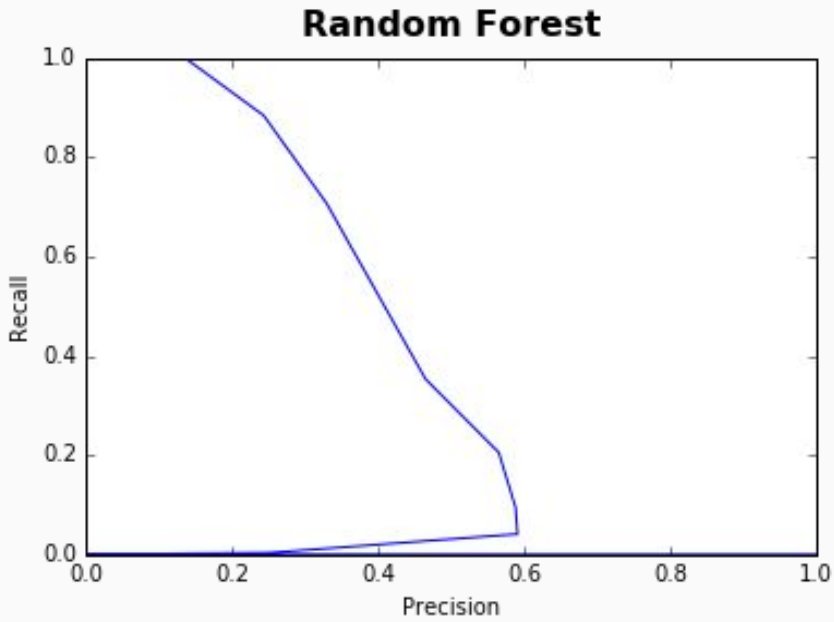
Walk/Bike: increase

Hypertension: decrease

High Blood Pressure: decrease

Random Forest - Diabetes Prediction





Random Forest Feature Importances

Main reason did not work last week:

- Taking care of house/family
- Going to school
- Retired
- Health Reasons
- Laid Off
- Disabled
- Other

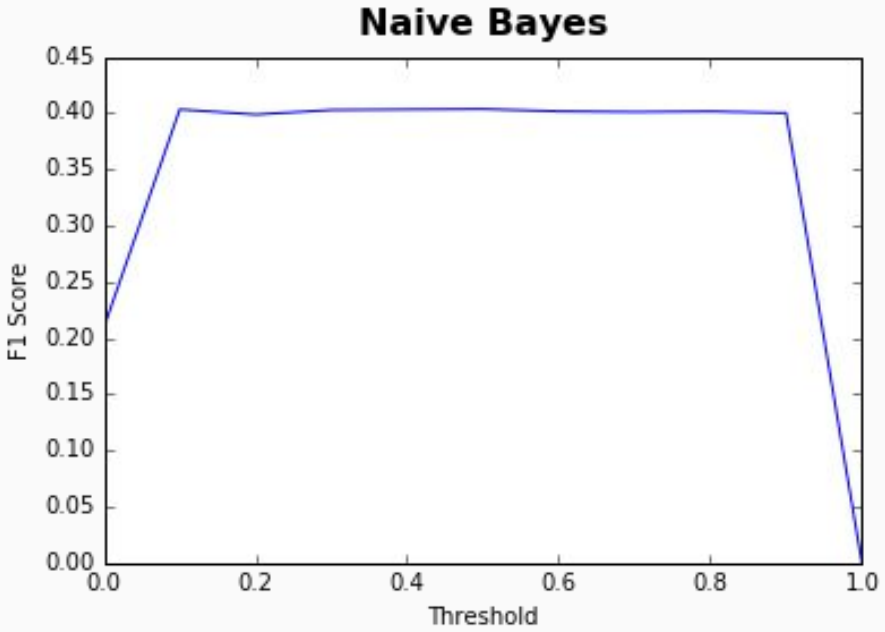
Age

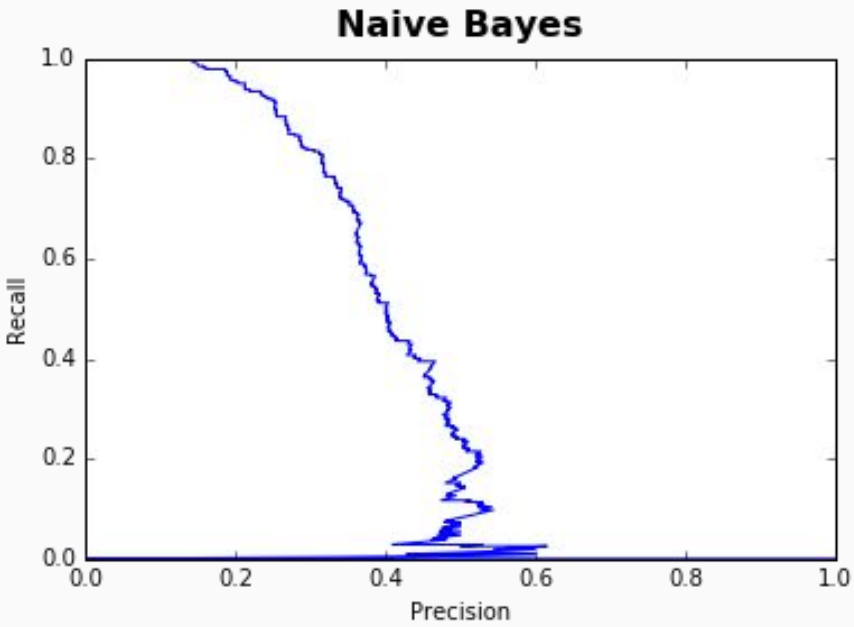
Sugar

Polyunsaturated Fat

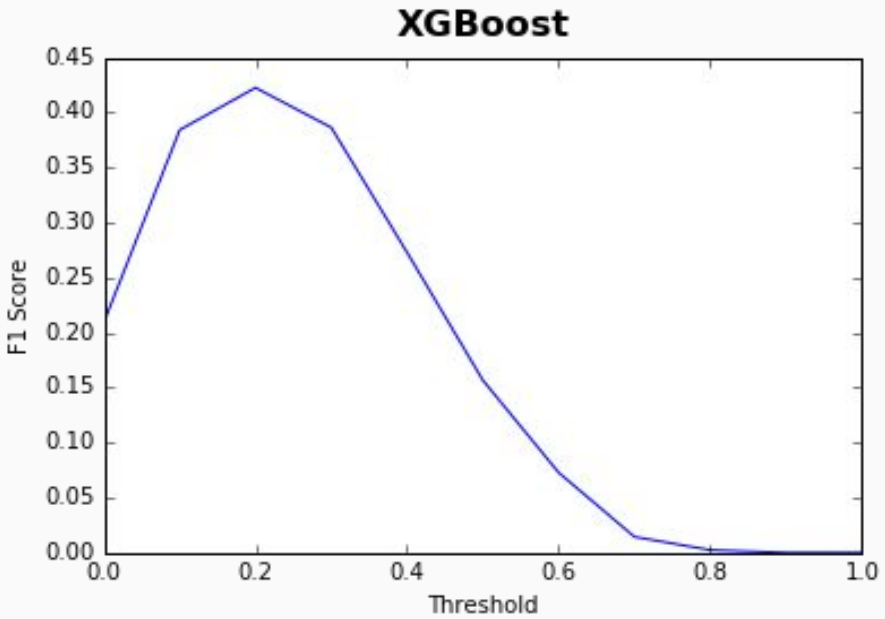
High Blood Pressure

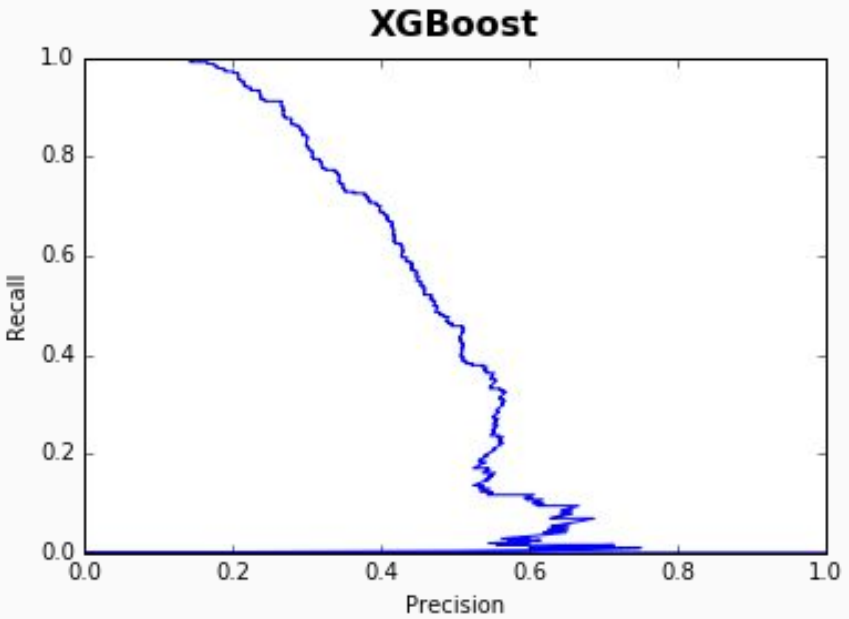
Naive Bayes - Diabetes Prediction





XGBoost - Diabetes Prediction





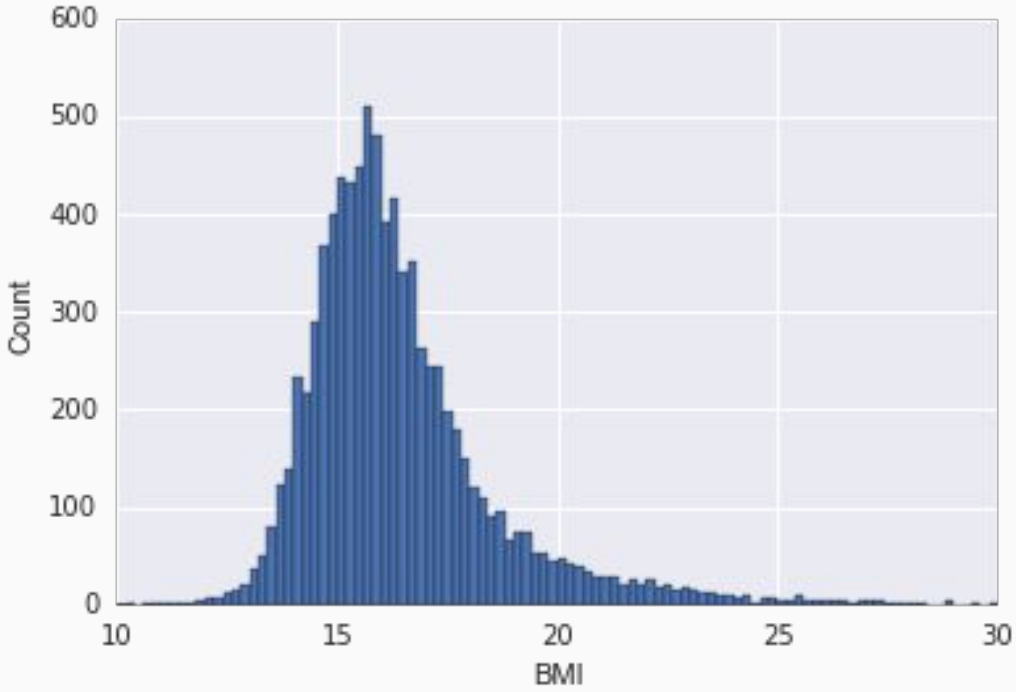
Final Model for Diabetes: XGBoost

Five Most Important Features:

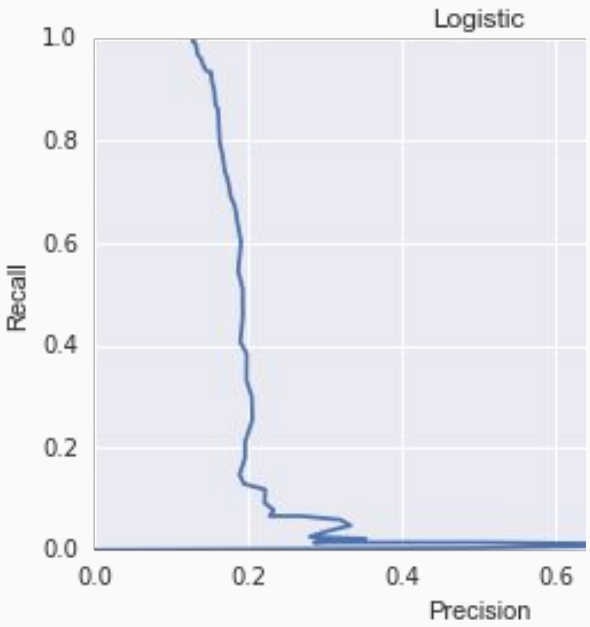
- Sugar Intake (per day)
- Age
- High Blood Cholesterol Level
- Polyunsaturated Fat Intake (per day)
- Alcohol Intake (per year)

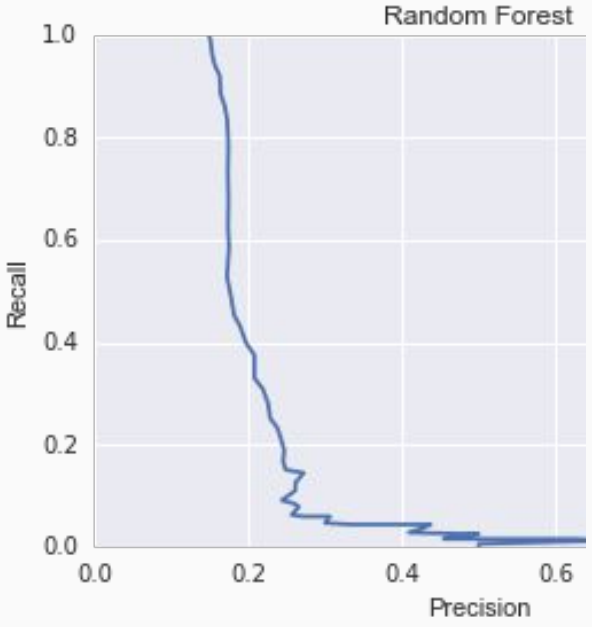
Kindergarten Data Set

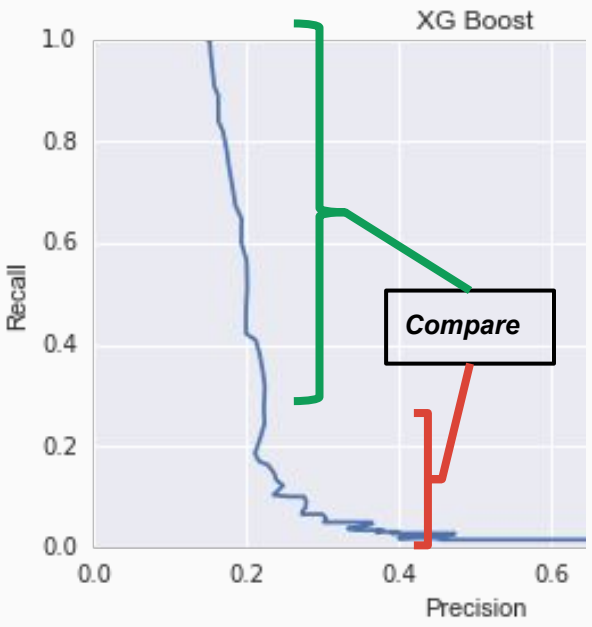
- Used features related to demographics and parental status.
- Geocode data was not made public.
- No behavioral data
- 8500 kids with no missing data
- Obesity cutoff is a BMI of ~18



Classification Results



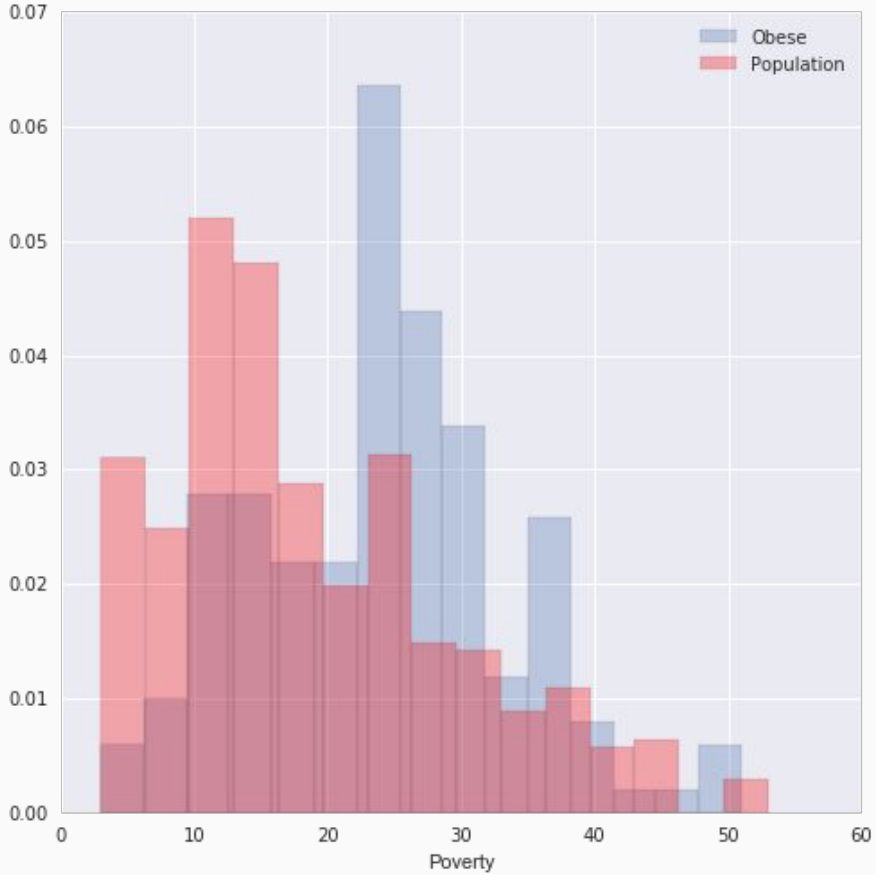




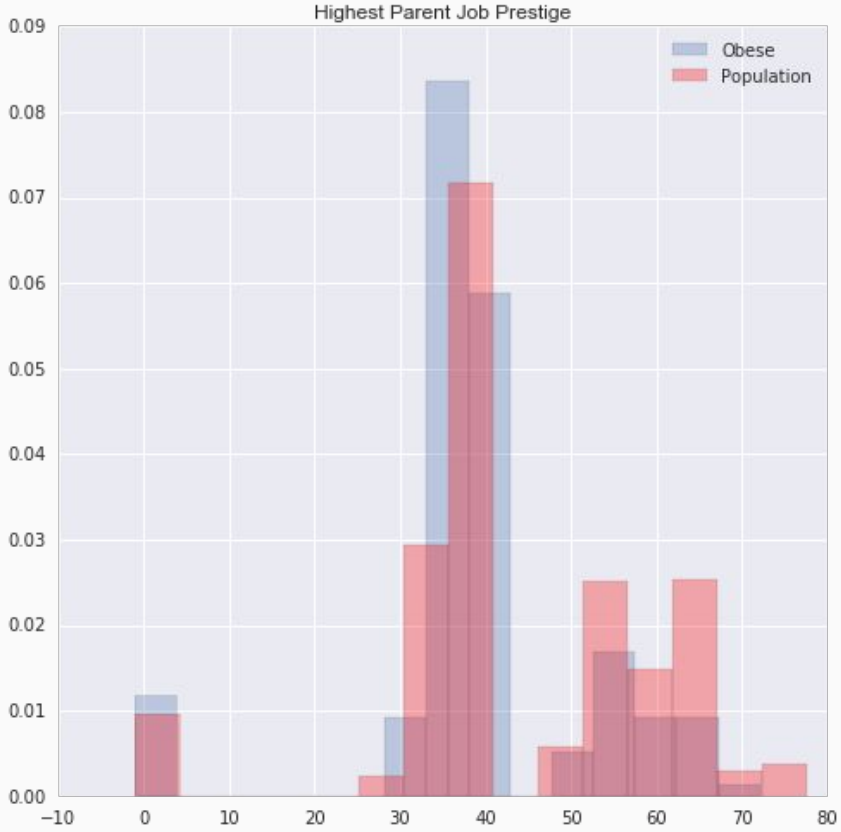
XG Boost

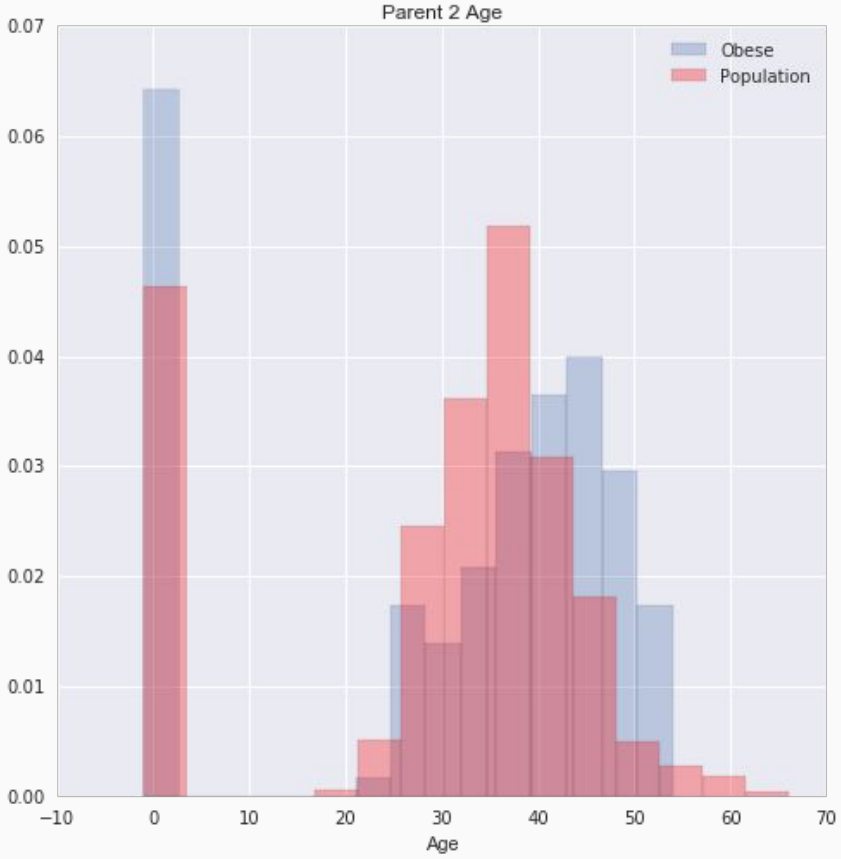
Most Important Features:

- Parents' Ages
- District Poverty
- Parents' Job Prestige
- Family Income Category



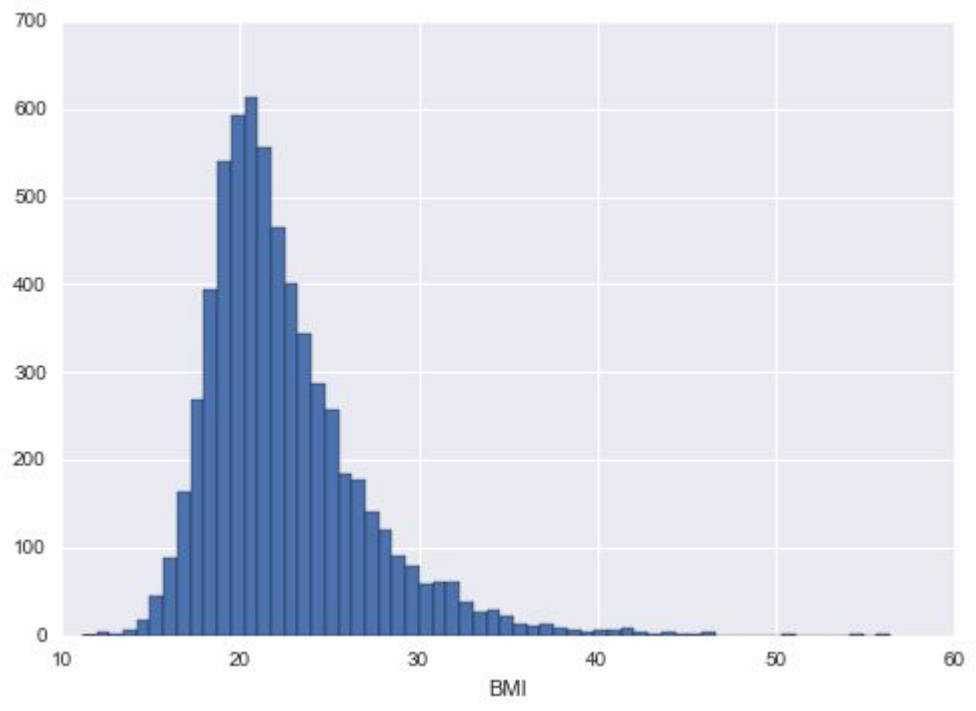
XG Boost


Highest Parent Job Prestige
Obese
Population
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−10
0
10
20
30
40
50
60
70
80


Parent 2 Age
Obese
Population
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−10
0
10
20
30
40
50
60
70
Age

High School Data Set

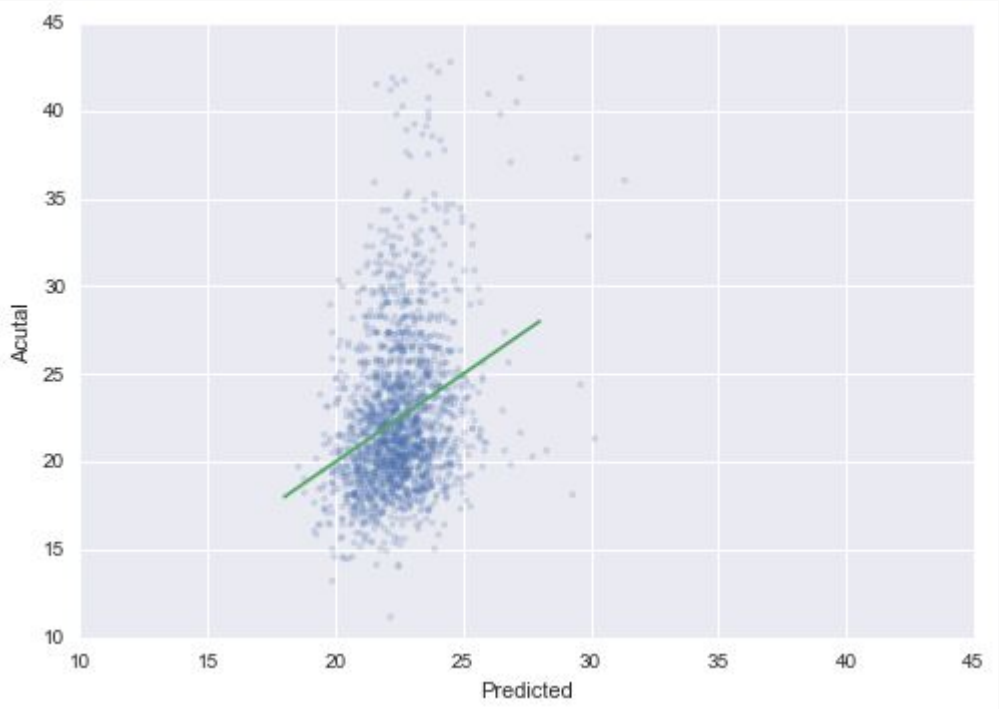
- Behavior features
  - Hobbies, exercise
  - Taught about diet/health in school?
  - Bullied at school?
  - Hours of sleep
- Race
- Census tract data:
  - Median household income
  - Urbanicity
  - Unemployment rate



GBM Regressor

- Classification of obesity more challenging when dealing with a range of children in their adolescence
- Model not predicting outliers well
- Median Household Income and hours of TV watching were the top features.

Predicting BMI



D3 Visualization

http://0.0.0.0:9000/