Predict Diabetes with Demographics, Behaviors and Health Conditions

Yu Fu, Hao Li, Nina Nguyen

Background

- As of 2015, 30.3 million Americans
- 9.4 percent of the U.S. population have diabetes
- Another 84.1 million have prediabetes
- Diabetes was the seventh leading cause of death in the U.S.

------National Diabetes Statistics Report 2017

- Demographics and daily human behaviors have some correlations with diabetes

Research Goal

- Build predictive model using
  - Demographics
  - Daily behaviors
  - Current health conditions
- Identify important features in predicting diabetes
- Study correlations between diabetes and these features

Data Resources

- Behavioral Risk Factor Surveillance System Dataset (BRFSS)
  - Health-related telephone surveys that collect state data about U.S. residents
  - 50 states
  - >400,000 adults interviewees
  - Risk behaviors, chronic health conditions, and use of preventive services

Variables

- Target (y) variable: Binary indicator of diabetes status: 0 / 1
- Feature (X) variables:
  - Demographics
  - Health Status and Conditions
  - Healthcare Access, Check and Treatments
  - Behavior: Smoking, Alcohol Consumption, Sleep, Exercise, Drive and Sun Exposure

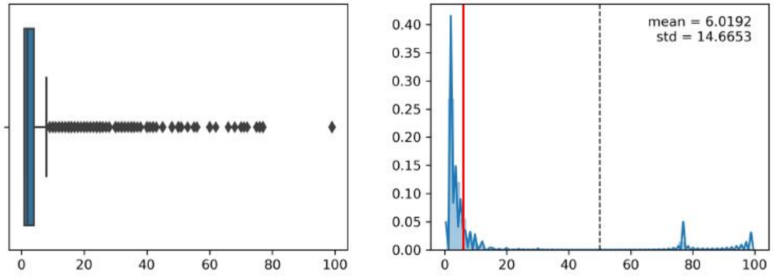
EDA and Data Cleaning

- Cleaning and transforming 176 variables
  - Continuous, Binary Categorical, Multi Categorical
    - For binary categorical variables, mapping values to 0 and 1
    - For multiple categorical variables, apply one-hot encoding
  - Most of cleaning are easy. Don't know / Refused / Missing -> NA values
  - Outlier detection and clipping by IQR. Bound by [Q1 - 1.5IQR, Q3 + 1.5IQR]
  - Unit Conversion (Example: ALCDAY5, number of drinking days)
    - 101 - 107: 1-7 days per week -> ((X - 100)/ 7)* 30
    - 201 - 230: 1-30 days per month -> X - 200
- Missing data in X variables
  - Number of variables with NA percentage 50% or higher: 112/176
- Plot data distribution before and after cleaning

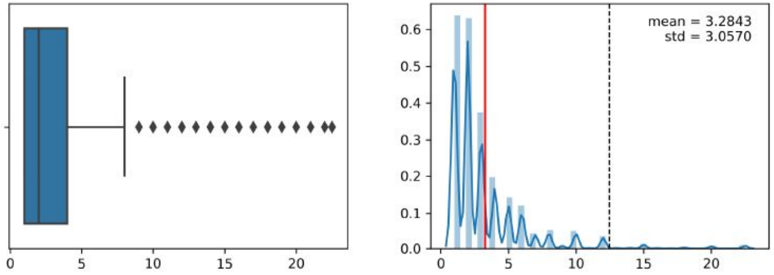
EDA and Data Cleaning

MAXDRNKS

Most drinks on single occasion past 30 days



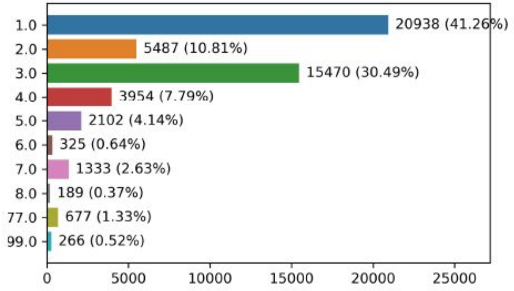
Cleaning Steps:
[1] 77 Dont know / Not sure -> NA
[2] 99 Refused -> NA
[3] Clip outliers out of 1.5 IQR



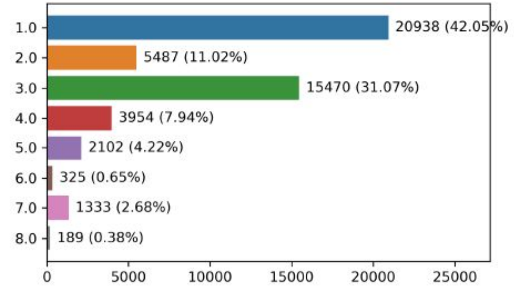
There are 231809 (52.99%) missing records.

HLTHCVR1

What is the primary source of your health care coverage?



Cleaning Steps:
[1] 77 Don't know / Not sure -> Missing
[2] 99 Refused -> Missing



There are 387638 (88.62%) missing records.

Solutions/Algorithm

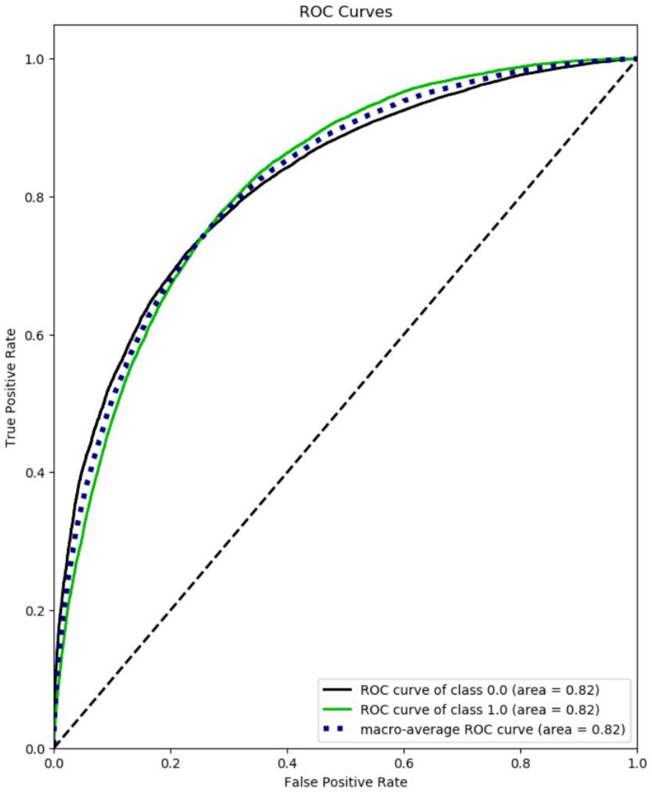
- Started with Random Forest
  - Doesn't handle missing values in predictors
- XGBoost
  - Directions for NA values of each feature is learned
  - Ensemble by gradient boosting - learn to cover mistakes (residual errors) of previous classifier
- Hyperparameter Tuning
  - Grid search with 5-fold cross validation
  - Use AUC ROC as metric (F1 is not good, need to consider more thresholds)
  - Hyperparameters to tune:
    - Number of trees in ensemble
    - Max depth of trees
    - Number of features used in each tree
    - Learning rate

Performance Checking

AUC(Area Under The Curve)

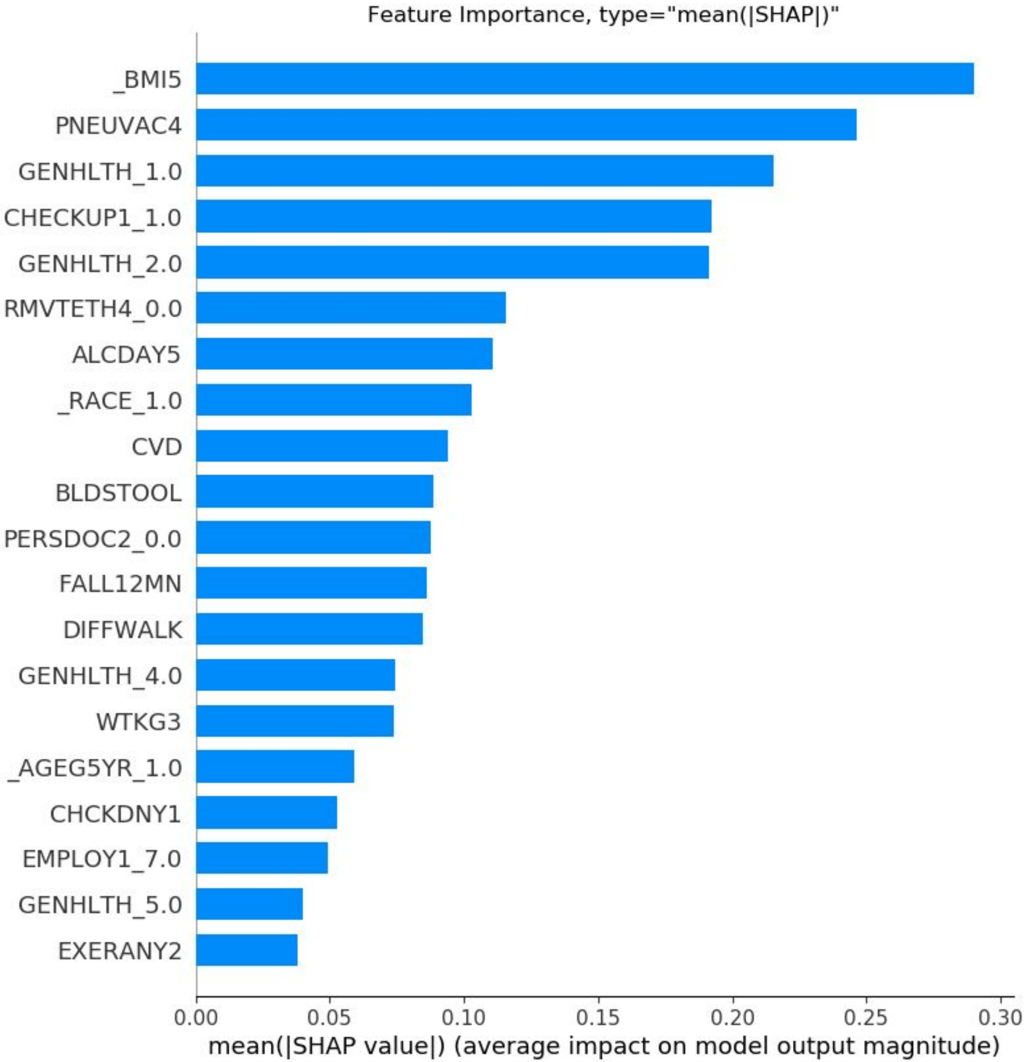
ROC (Receiver Operating Characteristics)

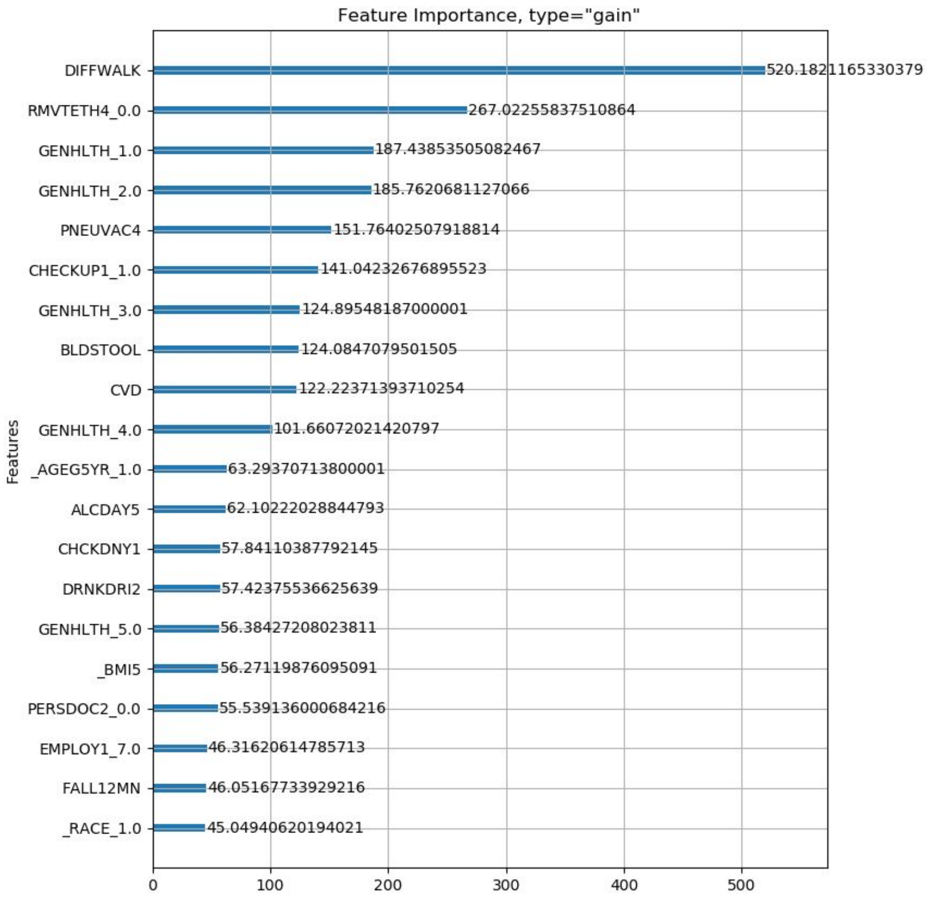
- Performance measurement for classification problem
- ROC is the curve of sensitivity and specificity on different thresholds
- Higher the AUC, better the model is at predicting
- Training model has AUC of 0.85 and Test model has AUC of 0.82

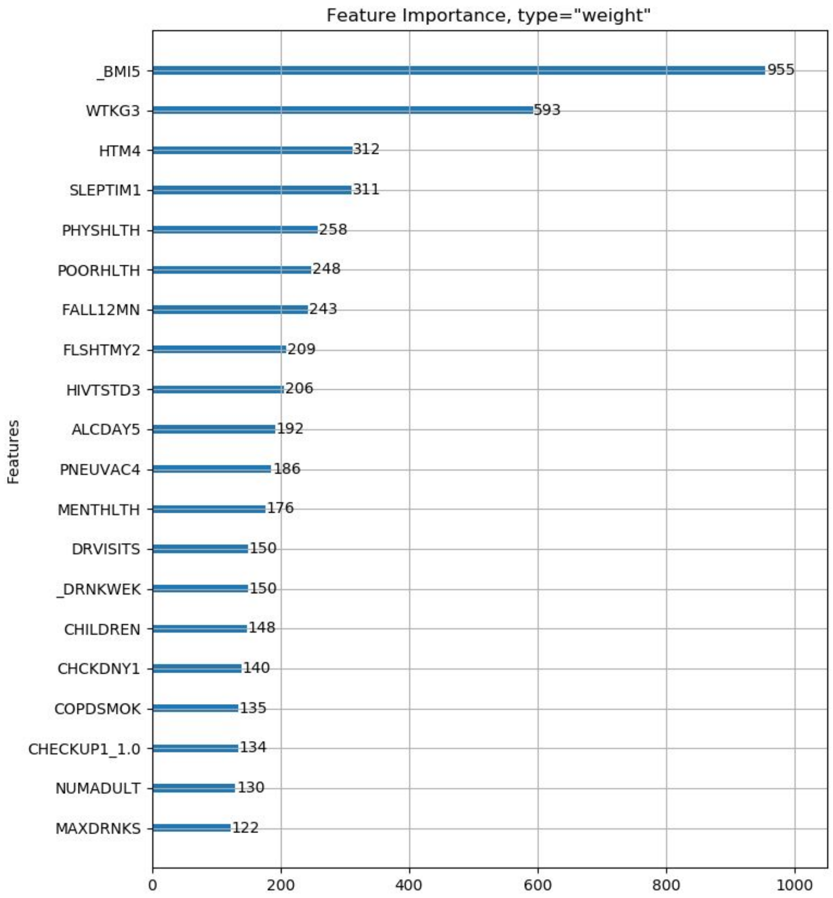


Feature Importance

- SHAP Values: A consistent and accuracy feature importance score that gives more explanation
- Weight: The number of times a feature is used to split the data across all trees
- Gain: The average training loss reduction gained when using a feature for splitting

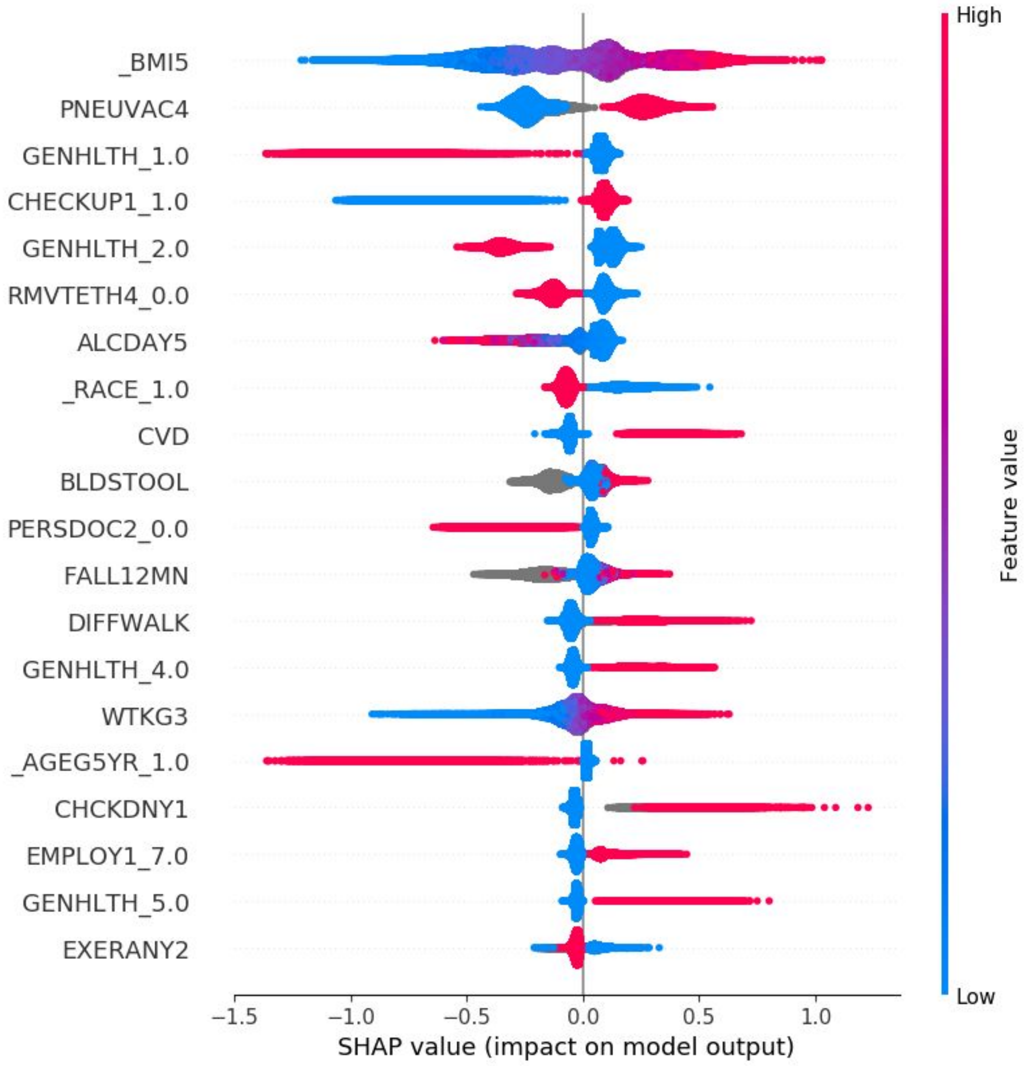



Feature Importance, type="gain"
Features
DIFFWALK 520.1821165330379
RMVTETH4_0.0 267.02255837510864
GENHLTH_1.0 187.43853505082467
GENHLTH_2.0 185.7620681127066
PNEUVAC4 151.76402507918814
CHECKUP1_1.0 141.04232676895523
GENHLTH_3.0 124.89548187000001
BLDSTOOL 124.0847079501505
CVD 122.22371393710254
GENHLTH_4.0 101.66072021420797
_AGEG5YR_1.0 63.29370713800001
ALCDAY5 62.10222028844793
CHCKDNY1 57.84110387792145
DRNKDRI2 57.42375536625639
GENHLTH_5.0 56.38427208023811
_BMI5 56.27119876095091
PERSDOC2_0.0 55.539136000684216
EMPLOY1_7.0 46.31620614785713
FALL12MN 46.05167733929216
_RACE_1.0 45.04940620194021
0
100
200
300
400
500


Feature Importance, type="weight"
Features
_BMI5 955
WTKG3 593
HTM4 312
SLEPTIM1 311
PHYSHLTH 258
POORHLTH 248
FALL12MN 243
FLSHTMY2 209
HIVTSTD3 206
ALCDAY5 192
PNEUVAC4 186
MENTHLTH 176
DRVISITS 150
_DRNKWEK 150
CHILDREN 148
CHCKDNY1 140
COPDSMOK 135
CHECKUP1_1.0 134
NUMADULT 130
MAXDRNKS 122
0
200
400
600
800
1000

Result Interpretation

- High BMI has a strong positive impact of diabetes, while low BMI has a strong negative impact.
- Pneumonia (lung infection) vaccination is recommended for diabetes patients. Our model catches the relationship.
- People identifying as being good health has a strong negative impact on diabetes. However, identifying as being poor health does not have a strong positive impact on diabetes.
- High alcohol consumption seems to have a negative impact on diabetes. Which direction is the causal relationship?



Future Work

- Model fine-tuning with more hyperparameters and larger search space
- The connection between multiple years of BRFSS survey data
- Population segmentation studies by demographics, by behaviors, or by unsupervised clustering
- Data collection or augmentation for minority groups.
- Experiments to test the causal hypothesis that can be derived from the current predictive model