

# Predicting Prediabetes and Diabetes with Machine Learning

...

National Health and Nutrition Examination Survey 2007 - 2016

Tracy Cardwell

# The Problem

- 10.5% of American population had diabetes as of 2018.
- One quarter of them didn't know it!
- 34.5% of American adults had prediabetes.
- Just over one in ten knows!

**WHO'S AT RISK**  
for prediabetes or type 2 diabetes?

You could have prediabetes or type 2 diabetes and not know it—there often aren't any symptoms. That's why it makes sense to know the risk factors:



The infographic displays six risk factors, each with a corresponding icon in a circle: 1. '45+ years old' with a person icon and the number 45+. 2. 'Physically active less than 3 times/week' with a person running and a crossed-out line. 3. 'Family history of type 2 diabetes' with a family silhouette. 4. 'High blood pressure' with a blood pressure cuff. 5. 'History of gestational diabetes\*' with a pregnant woman silhouette. 6. 'Overweight' with a scale icon.

\*Diabetes during pregnancy. Giving birth to a baby weighing 9+ pounds is also a risk factor.

**DID YOU KNOW...** African Americans, Hispanic/Latino Americans, American Indians/Alaska Natives, Pacific Islanders, and some Asian Americans are at higher risk.

If you have any of the risk factors, ask your doctor about getting your blood sugar tested.



CDC  
CENTERS FOR DISEASE  
CONTROL AND PREVENTION

# The Problem

## Diabetes:

- 7th leading killer of Americans
- Leading cause of end-stage kidney failure, blindness in working-age adults, amputation
- 2 to 4 fold risk of heart attack, stroke
- \$327 Billion cost

## Prediabetes:

- Increased risk of heart attack, stroke, kidney disease
- Increased risk of type 2 diabetes

# Can Machine Learning Help?

- Increasing use in health care to aid in diagnosis, create tailored treatment
- Leverage health care resources where spread thin
- Gain new insights from health data
- More people who know = earlier treatment and lower cost

# Predict Prediabetes and Diabetes with Machine Learning

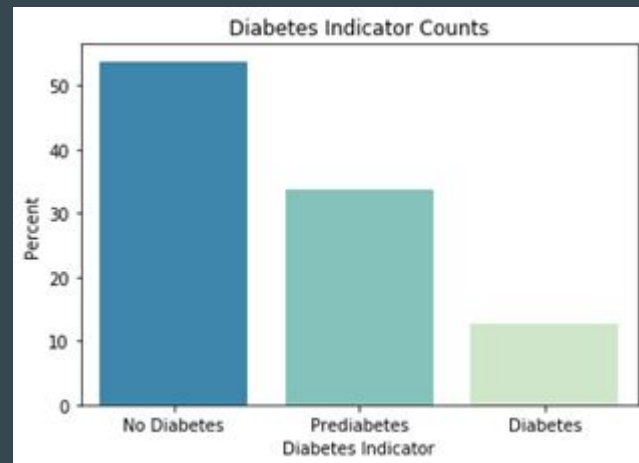
- National Health and Nutrition Examination Survey 2007-2016
- Python machine learning
- Predict prediabetes using American Diabetes Association diagnostic criteria with blood test results
- Predict diabetes using prescription medication records
- Look for risk factors for prediabetes and diabetes
- Help identify who should get tested

# NHANES

- Two year survey cycles
- Changes in survey content between cycles
- Demographics, socioeconomic, dietary intake, physical examination, laboratory testing, prescription medications, questionnaires on many health-related topics
- 5,000 participants per year
- Oversampling of certain groups to ensure subsample sizes adequate for analysis

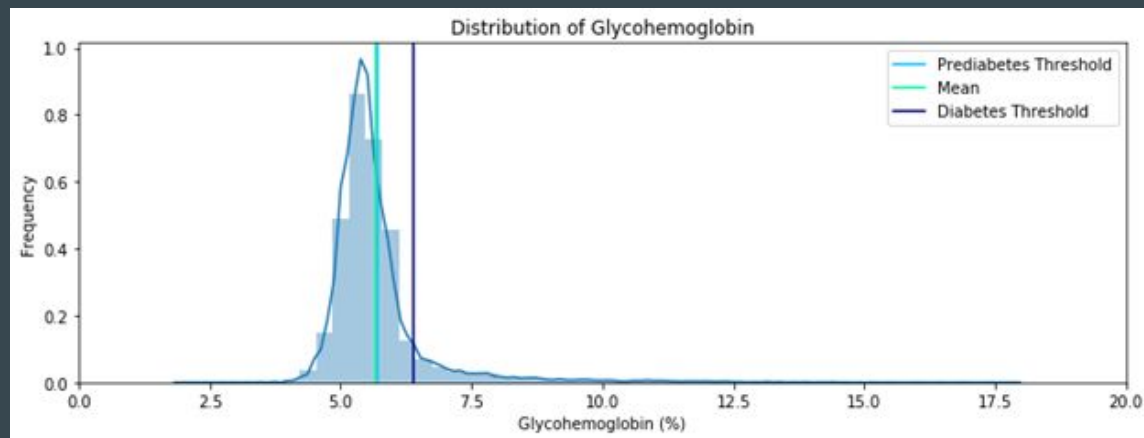
# What does the data say?

- Indicator based on laboratory results using American Diabetes Association diagnostic criteria
- Distribution similar to National Diabetes Statistics Report



# What does the data say?

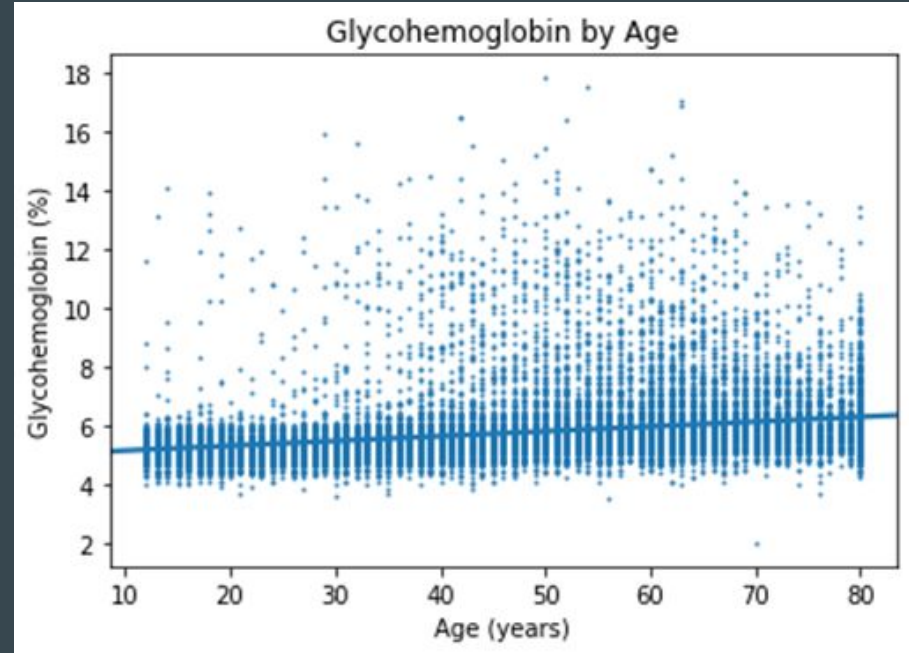
- Many features with right skews, long right tails
- Very high outliers
- Data verified, outliers are valid measurements
- Mean glycohemoglobin (long term blood sugar) at prediabetes diagnosis threshold!





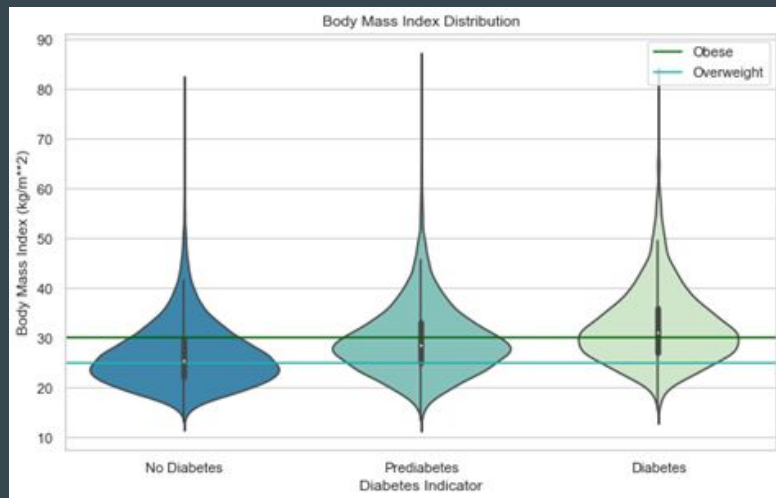
# What does the data say?

- Clear upward trend in glycohemoglobin with increasing age
- Proportion in diabetic range ( $\geq 6.5\%$ ) higher over 30 years old
- Many teens in diabetic range



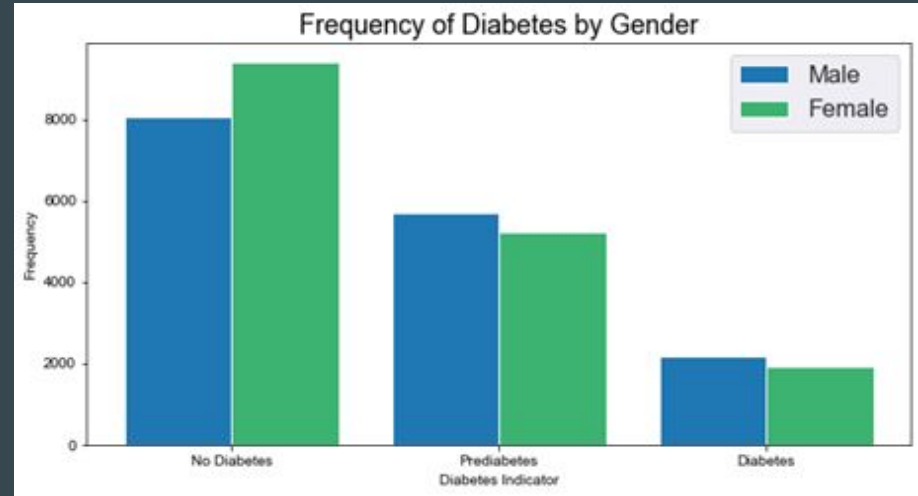
# What does the data say?

- Many features have a clear trend upward from healthy blood sugar through prediabetes to diabetes.



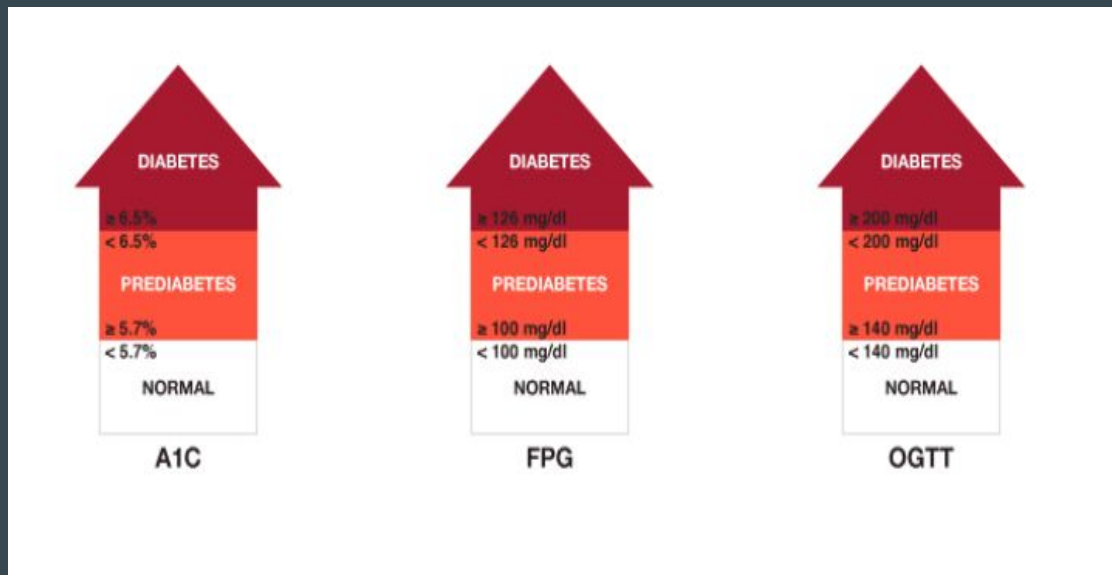
# What does the data say?

- Data analysis found relationship between diabetes and gender.
- Also high blood pressure, activity level, triglycerides, body mass index, serum iron, marital status, and aspartate aminotransferase (AST)



# Machine Learning Approach

- Two groups of models
- Group 1: multiclass classification with target determined by laboratory results and American Diabetes Association classifications.



# Machine Learning Approach

- Group 2: binary classification with target defined by participant medication use for diabetes
- Logistic Regression and tree based models due to very high dimensionality



<https://www.diabetes.org/diabetes/medication-management>

# Preprocessing

- Hundreds of features
- Duplicate features, different units
- Highly (nearly perfectly) correlated features
- Some features not collected entire ten years
- Some features moved during ten years
- Self-reported continuous features with code values above data range: “don’t know”, “refused to answer”
- IterativeImputer to fill nulls
- Dummy variables for categoricals
- StandardScaler for Logistic Regression continuous features

# Group 1 Results

- XGBoost performed best.
- No model was a strong predictor of prediabetes.
- They all had better precision with diabetes.

Summary of pre-tuning results:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Logistic Regression	.715	71.7 (.274)	.58	.63	.77	.60	.74
Decision Tree	.597	60.1 (.381)	.45	.46	.72	.46	.60
Random Forest	.674	67.3 (.540)	.55	.47	.80	.50	.74
Adaboost	.705	70.5 (.317)	.59	.53	.81	.56	.73
Gradient Boosting	.717	72.0 (.262)	.60	.56	.81	.58	.76
XGBoost	.712	71.7 (.457)	.60	.55	.81	.57	.77

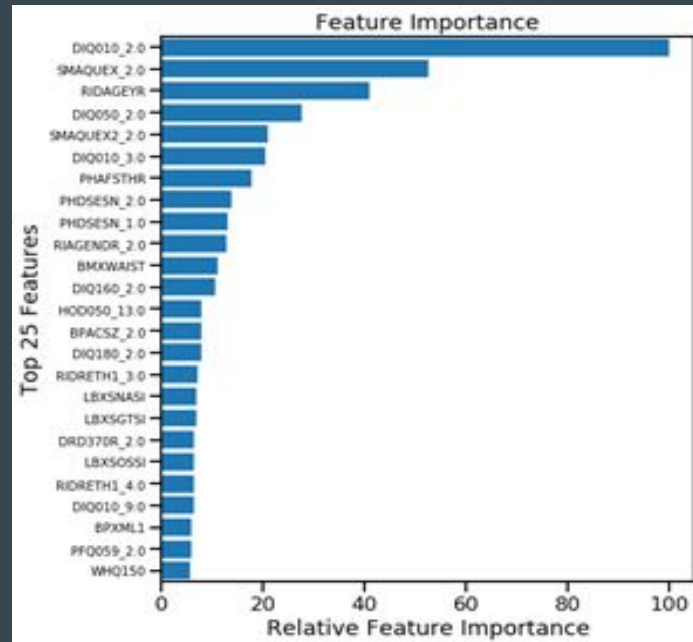
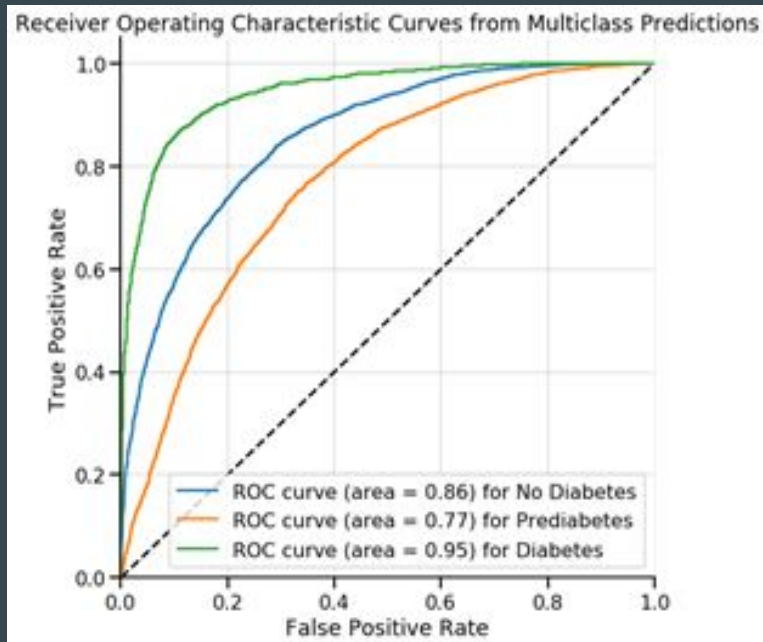
Post-tuning:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Random Forest	.702	70.7 (.584)	.57	.60	.78	.59	.76
XGBoost	.728	72.8 (.529)	.61	.59	.81	.60	.77

Note: Accuracy is for overall model. All other measures are for Prediabetes class.

As this is medical data, reported Sensitivity & Specificity in addition to Precision & Recall.

# XGBoost Model





# Group 2 Results

- XGBoost again outperformed other models.
- Better precision than Group 1 diabetes results

Summary of results pre-tuning:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Logistic Regression	.935	93.8 (.723)	.75	.88	.95	.81	.97
Decision Tree	.928	92.0 (.576)	.77	.78	.96	.77	.87
Random Forest	.917	91.7 (.703)	.85	.57	.98	.69	.97
Adaboost	.942	94.1 (.575)	.82	.81	.97	.81	.97
Gradient Boosting	.953	95.4 (.324)	.81	.92	.96	.86	.98
XGBoost	.952	95.5 (.537)	.82	.89	.96	.85	.98

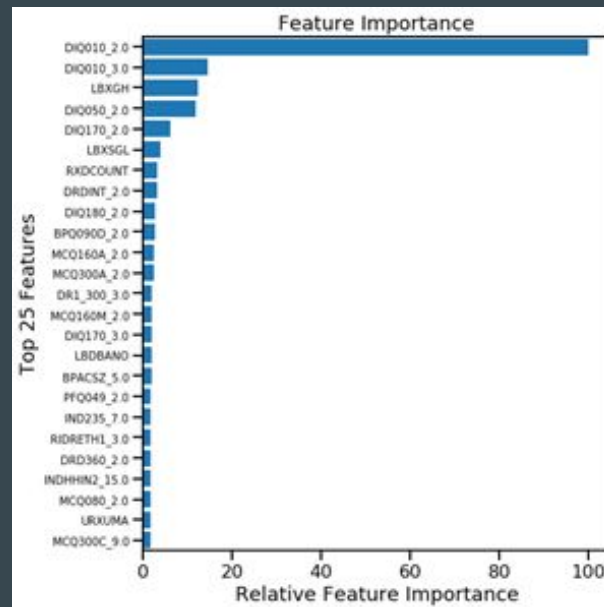
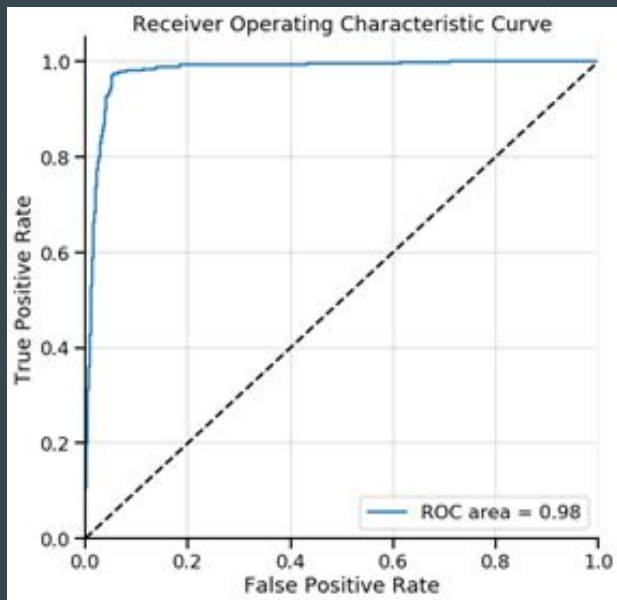
Post-tuning:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Random Forest	.948	95.4 (.448)	.78	.93	.95	.85	.98
XGBoost	.951	95.7 (.596)	.81	.89	.96	.85	.98

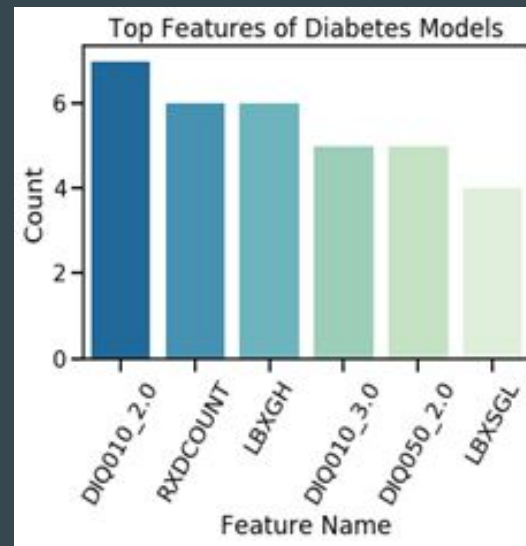
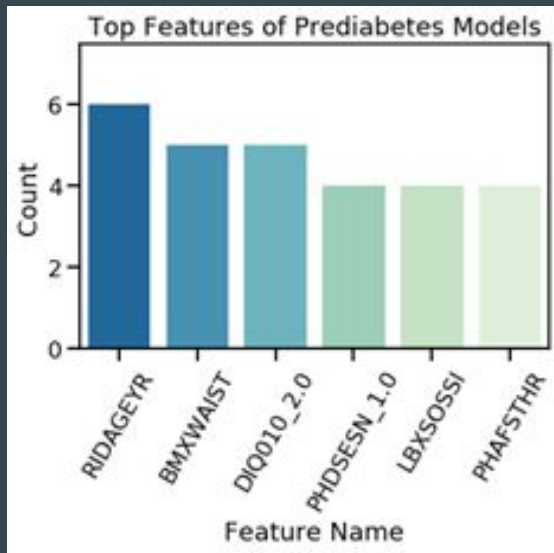
Note: Accuracy is for overall model. All other measures are for positive class: Diabetes.

As this is medical data, reported Sensitivity & Specificity in addition to Precision & Recall.

# XGBoost Model



# Most Important Features Overall



# Conclusion

- XGBoost performed best on data
- 60% precision predicting prediabetes in group 1
  - 33.7% of sample
- 80% precision predicting diabetes in group 2
  - 15.8% of sample
- Class imbalance
- More feature overlap between healthy and prediabetes than healthy and diabetes
- Group 2 included diagnostic labs as features
- Noise in features
- Novel important feature: Osmolality for prediabetes

# Future Work

- More feature engineering
  - Noisy features
  - Imputation for categoricals
  - Dummy definition for tree models
  - Class imbalance
  - Additional correlated features
  - Outliers
- Stacked models