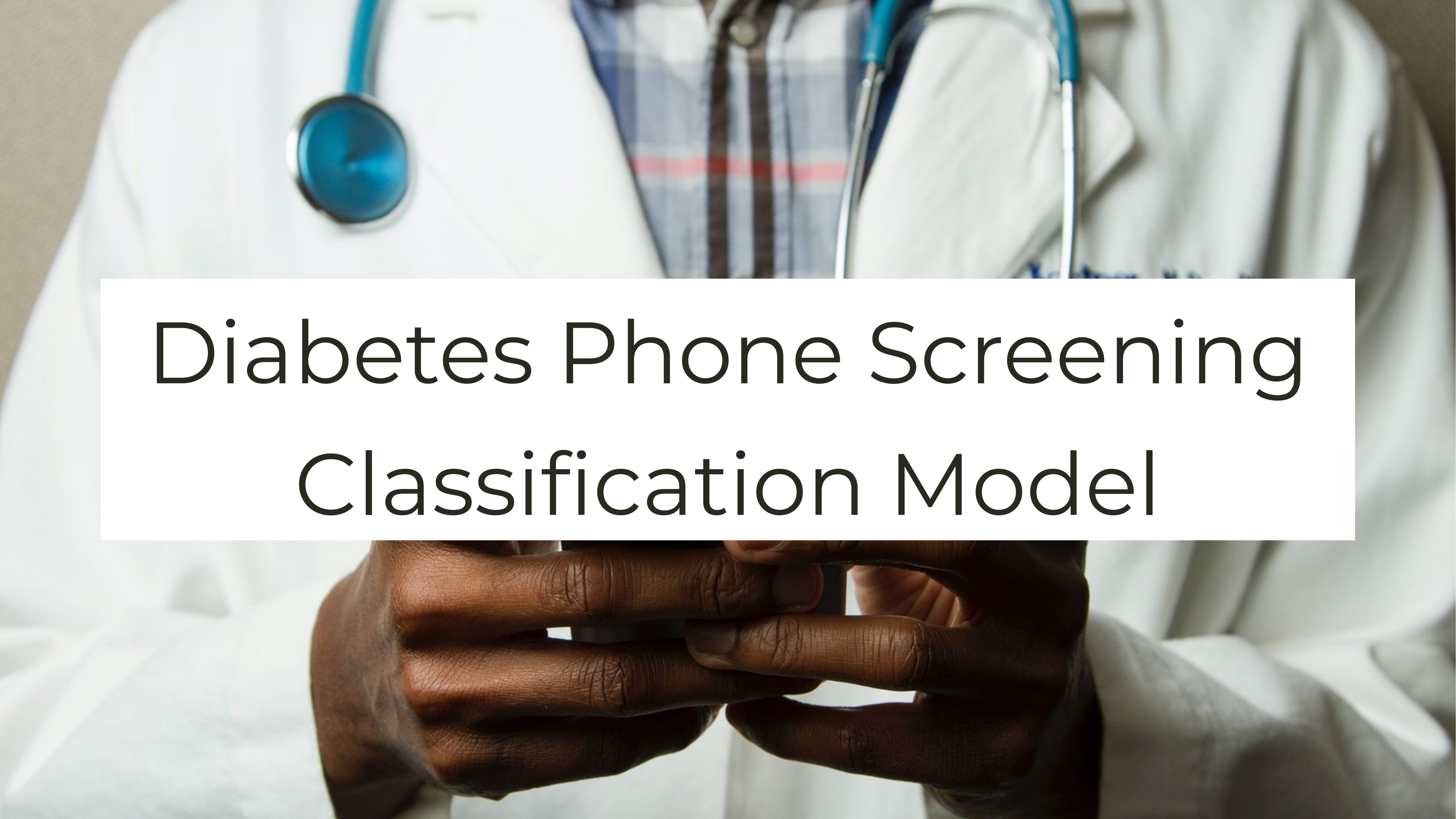
Diabetes Phone Screening Classification Model



Agenda

Why Are We Here?

1. The Business Problem

The Data

2. Data Understanind

3. Exploratory Data Analysis

Data Modeling

4. Model Results

Our Future

5. Recommendations + Next Steps

- Undiagnosed Diabetes Effects
  - The CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 pre-diabetics are unaware of their risk.
  - Undiagnosed diabetes and pre-diabetes approaching $400 billion dollars annually.

- Early diabetes monitoring can allow for quick action
  - Regular preventative monitoring can reduce diabetes risk
  - Closely tracking and analyzing indicators helps prompt necessary preventative measures.

- Key indicators:
  1. What are your blood sugar levels?
  2. What is your blood pressure?
  3. What are your cholesterol levels?
  4. What is your body weight and BMI?
  5. How much physical activity do you get?
  6. What is your family history?

Business Understanding

- Mt. Sinai's telemedicine targets pre-diabetics through preventive monitoring.
  - Regular check-ins with healthcare providers, including self-management classes, nutrition counseling, and prevention programs, can help reduce readmissions.
- Mt. Sinai has a limited staff and equipment to serve those at-risk.
  - Targeted preventative measures should be implemented for those individuals who are most likely to become afflicted with the disease.
- How can Mt. Sinai target at-risk patients?
  - Develop a classification model to identify at-risk diabetics using BMI, age and diet data from phone screening.

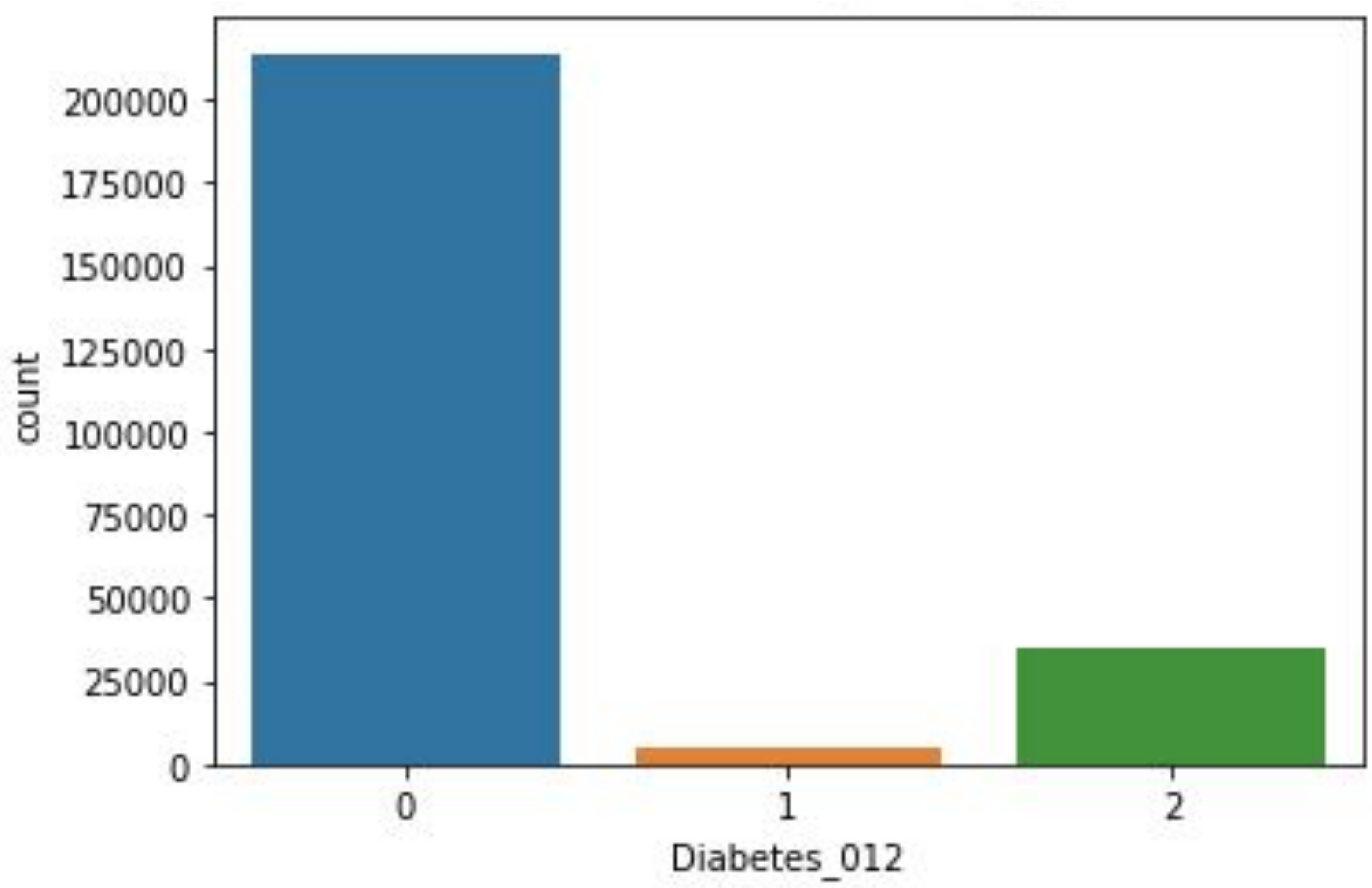
Dataset has 253,680 survey responses.. The target variable Diabetes_012 has 3 classes. 0 is for no diabetes, 1 is for prediabetes, and 2 is for diabetes. This dataset has 21 features of binned categories into discrete variables.

Data Understanding

BRFSS 2015 Diabetes Indicators (Phone Screen)

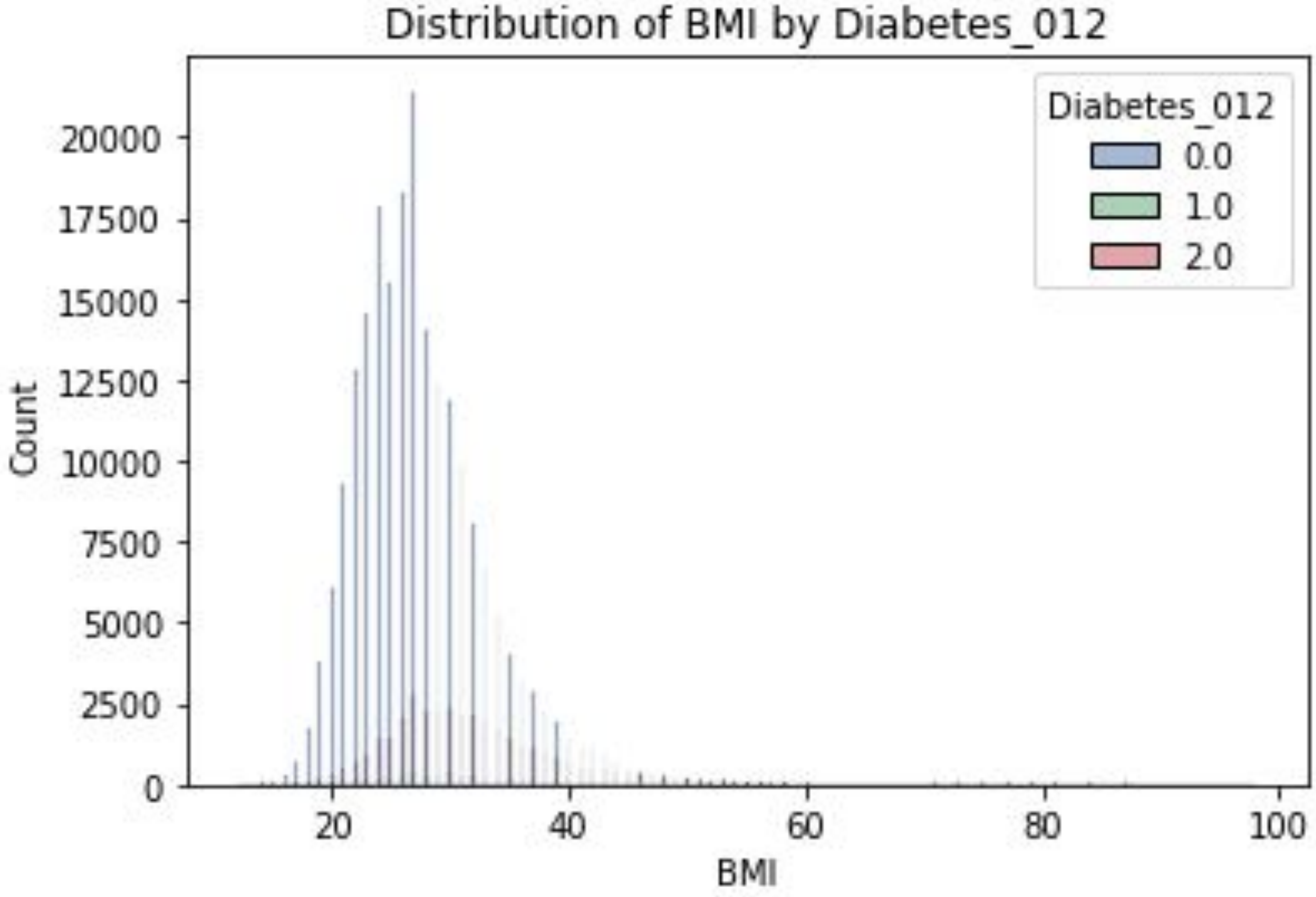
- Blood Pressure
- Stroke
- Healthcare Coverage
- Smoker
- Diet
- Education
- Physical Activity
- BMI
- Gender
- Cholesterol
- Heart Disease
- Income

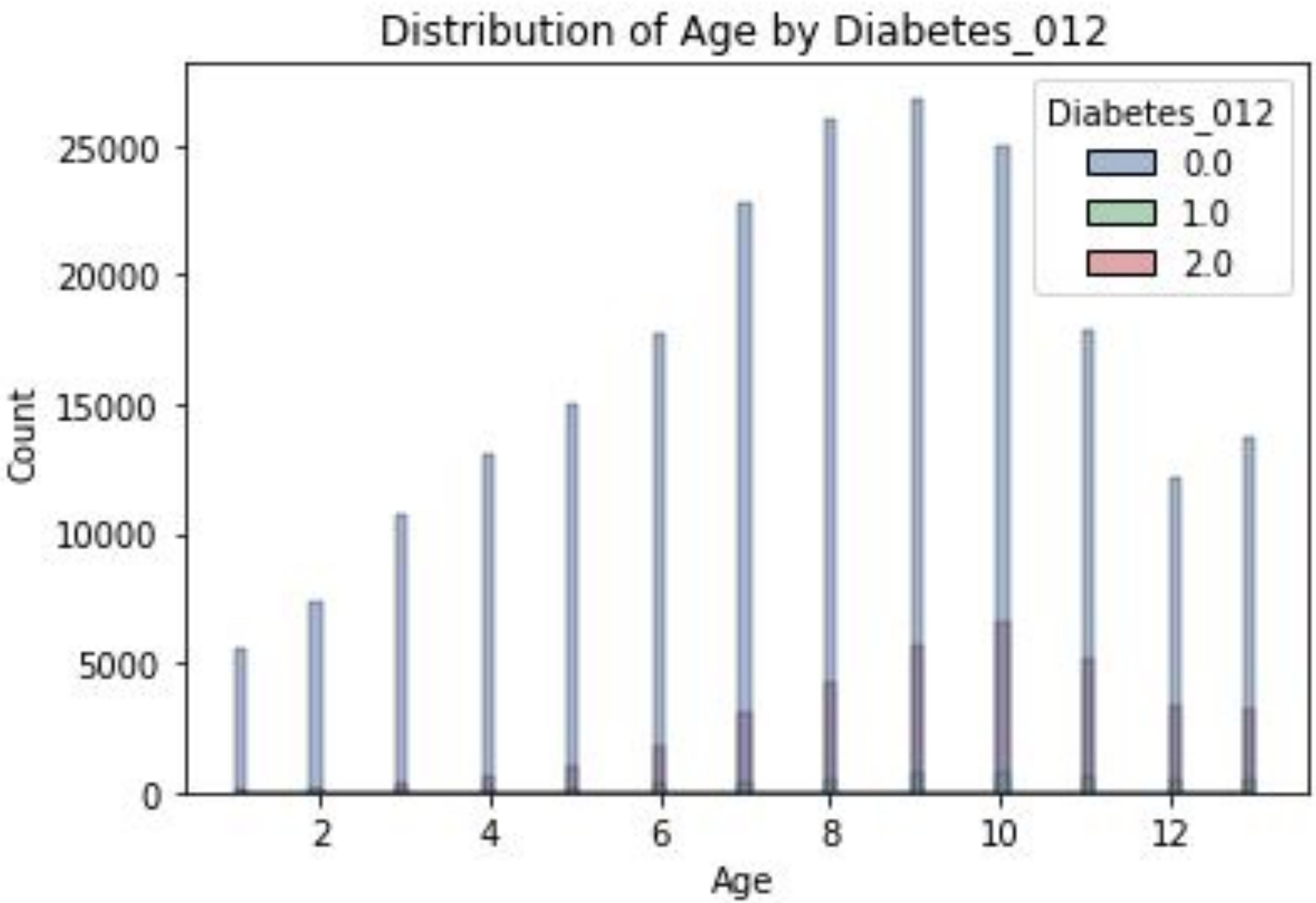
Diabetic Target Variable



Class Imbalance

- There are more individuals who do not have diabetes than those who do have diabetes.
- Under-sample majority class (Diabetes = 0) to create a balanced dataset


Distribution of BMI by Diabetes_012
Diabetes_012
0.0
1.0
2.0
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
20
40
60
80
100
BMI


Distribution of Age by Diabetes_012
Diabetes_012
0.0
1.0
2.0
Count
25000
20000
15000
10000
5000
0
2
4
6
8
10
12
Age

Data Modeling

1. Data processing

   Two sets of models were run for multi-class classification (predicting non-diabetic, pre-diabetic and diabetic) and binary classification (non-diabetic and either pre-diabetic or diabetic).

2. Iterative models

   8 different models per each multi-classification and binary classification were explored.

3. Final model

   The strongest model for each set was chosen as the final model.

Results - Multiclass

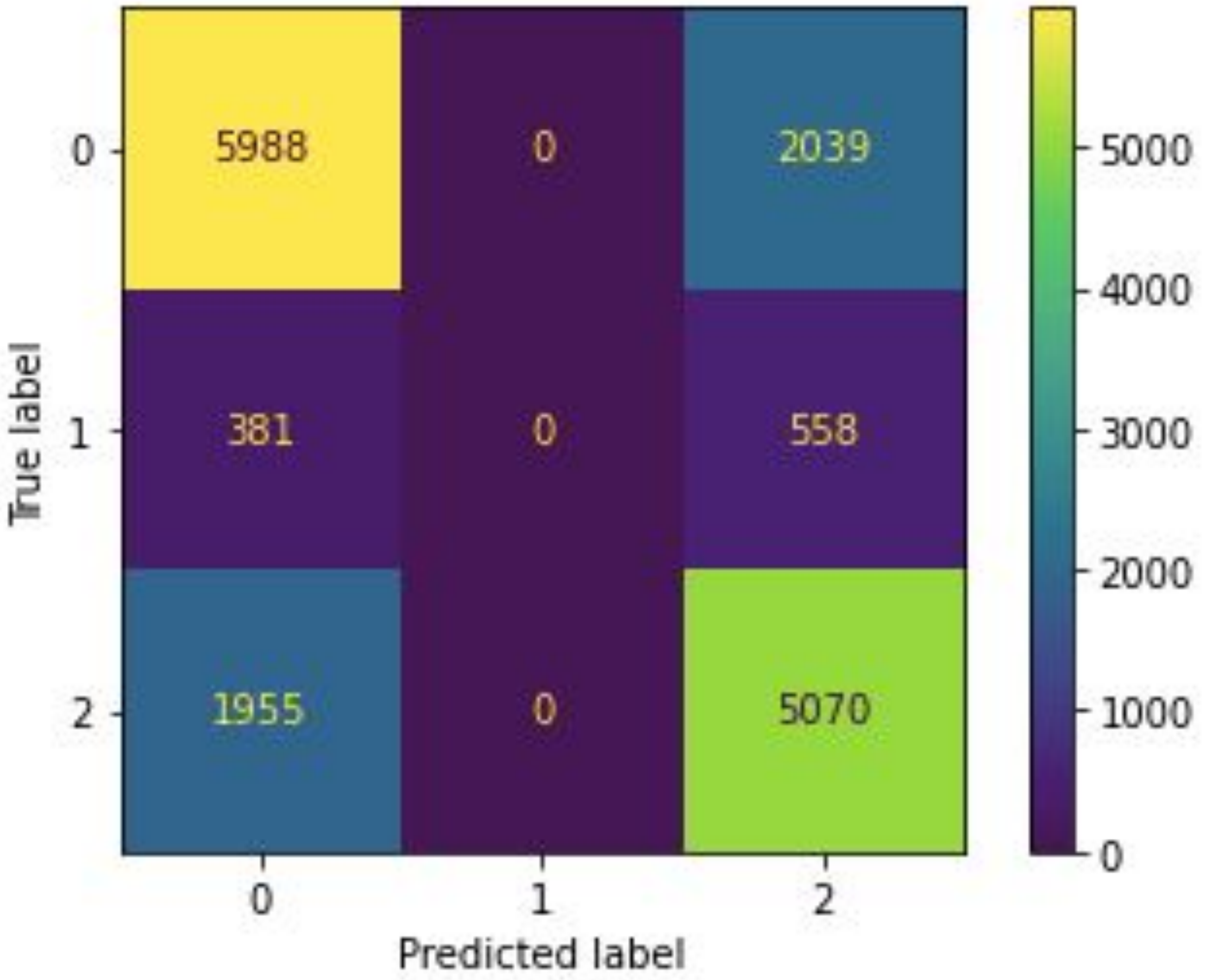
- Classification model: GradientBoost
- F1 score: 63%

  Balanced accuracy between accuracy target diabetic/pre-diabetic and non-diabetic survey respondents.

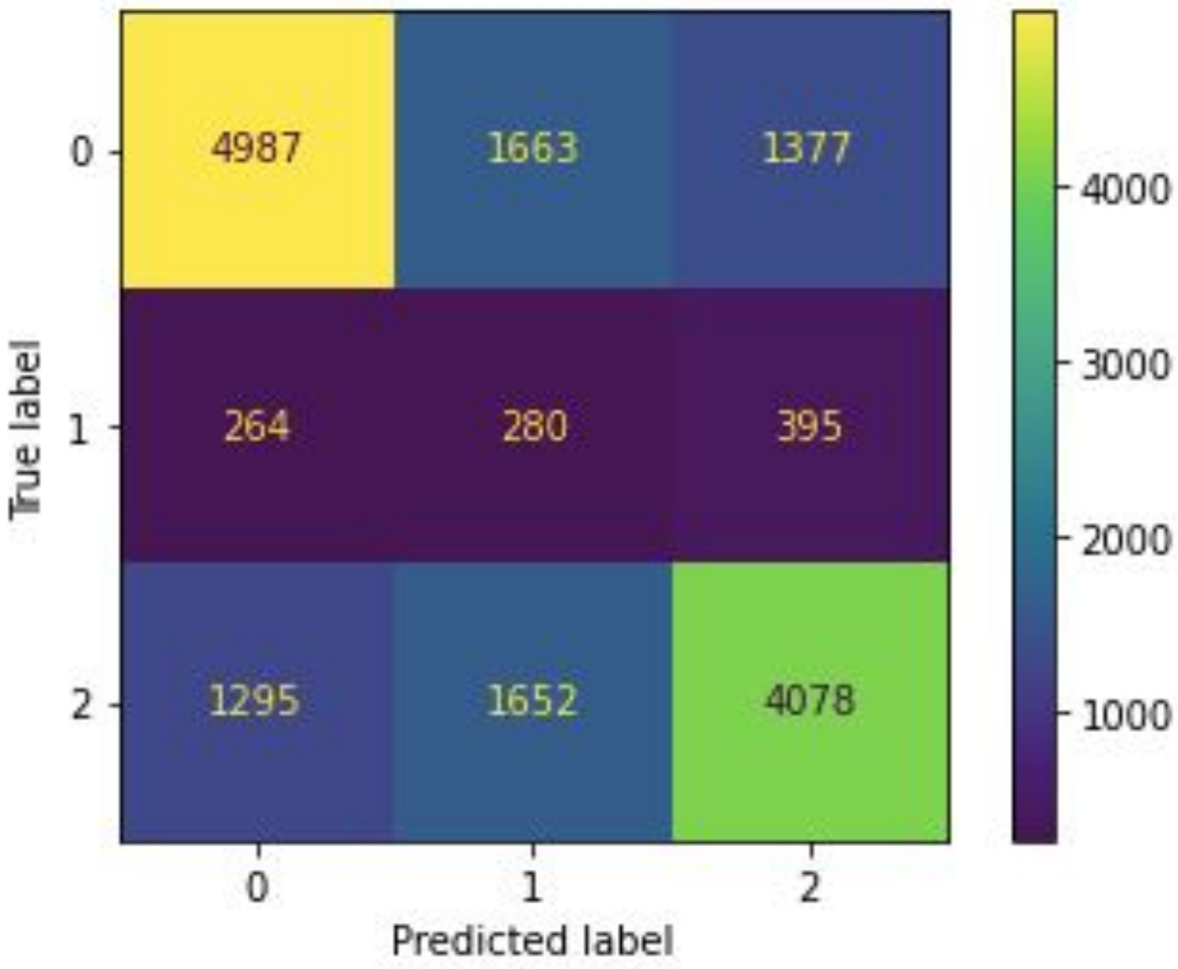
- Accuracy score: 60%
- Best feature classifiers

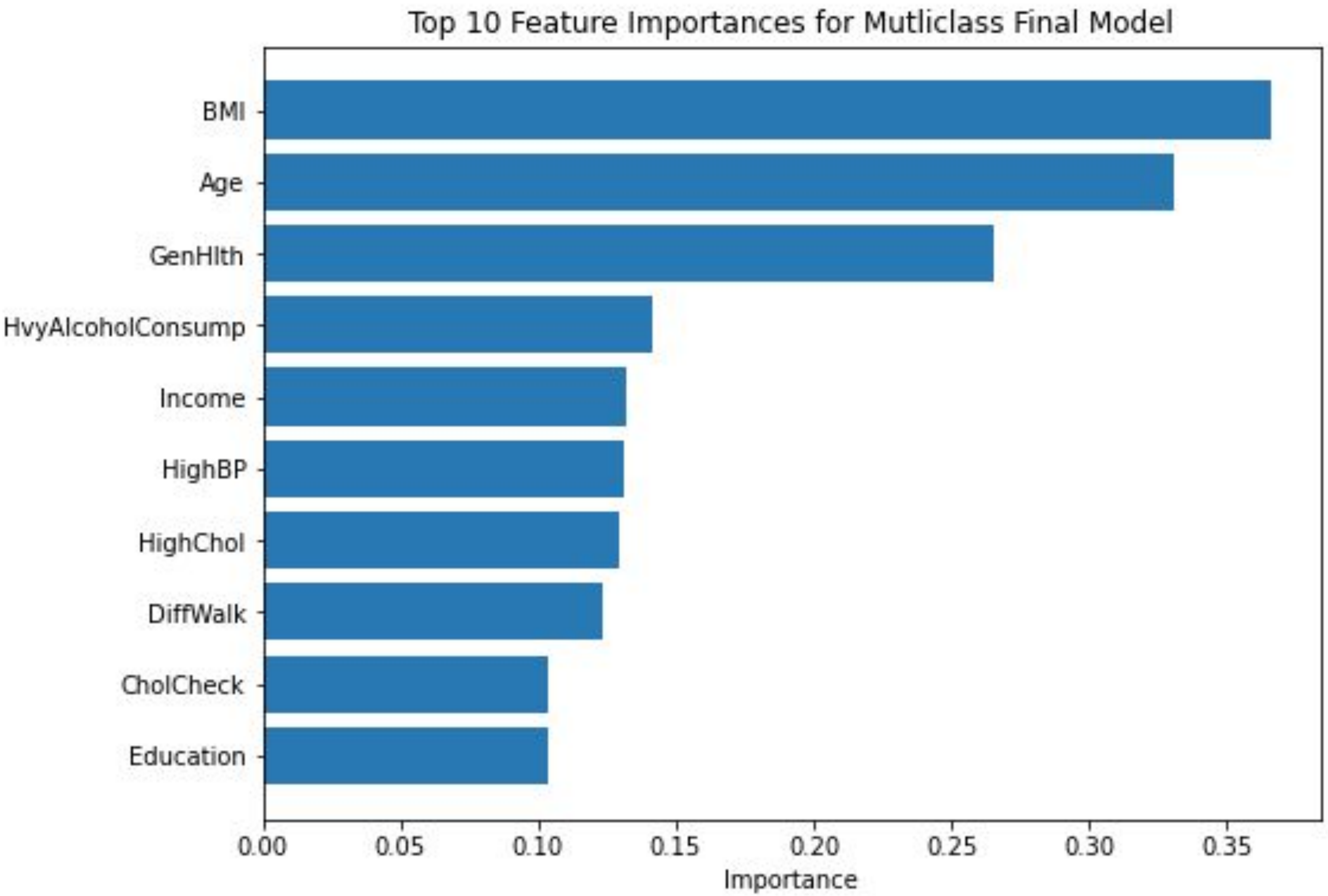
1. BMI
2. Age
3. GenHlth

Baseline Model - Multi-class (LR)



Final Model - Multi-class (GradientBoost)




Top 10 Feature Importances for Mutliclass Final Model
BMI
Age
GenHlth
HvyAlcoholConsump
Income
HighBP
HighChol
DiffWalk
CholCheck
Education
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance

Results - Binary Classification

- Classification model: AdaBoost

  Model where output of the other learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier.

- F1 score: 75%

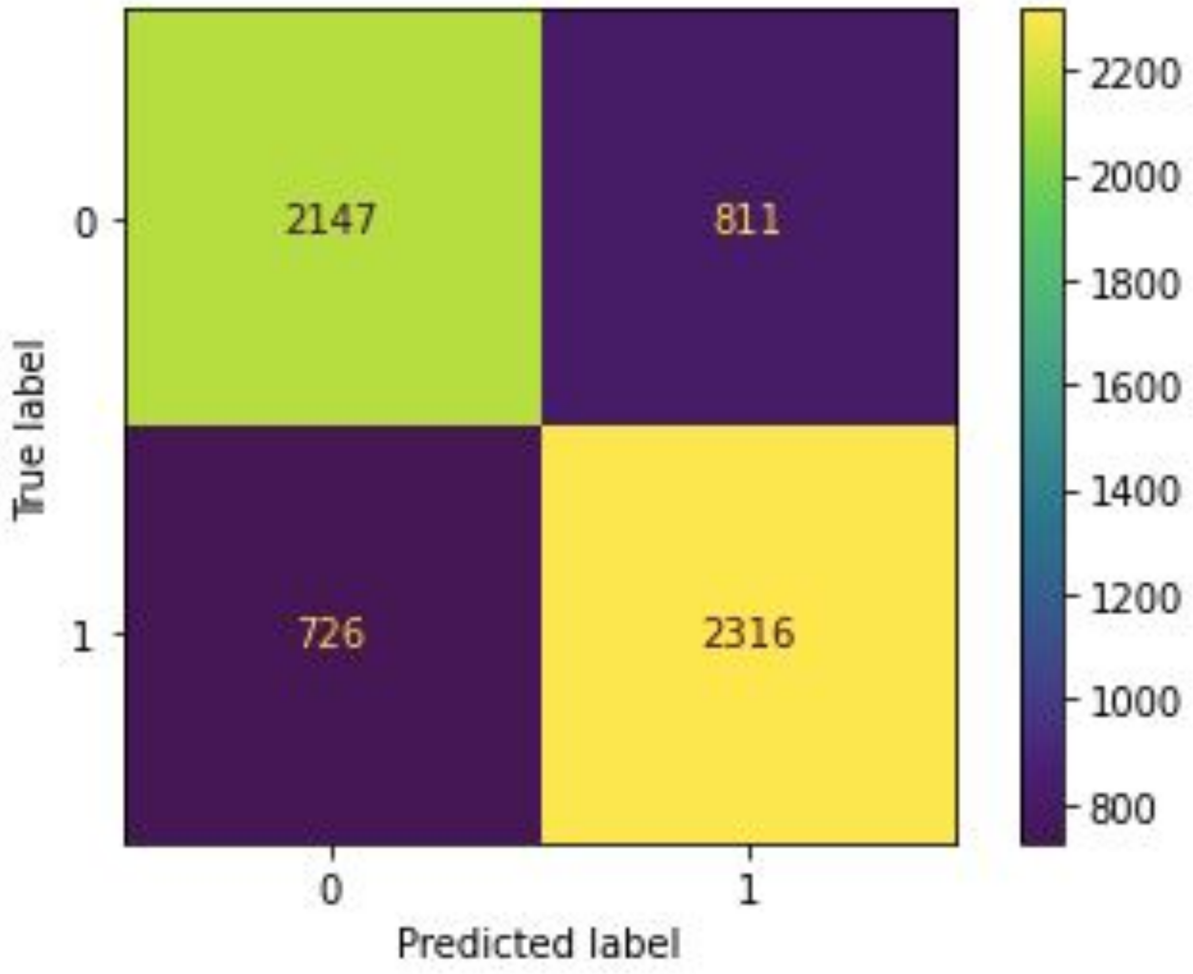
  Balanced accuracy between accuracy target diabetic/pre-diabetic and non-diabetic survey respondents.

- Accuracy score: 75%

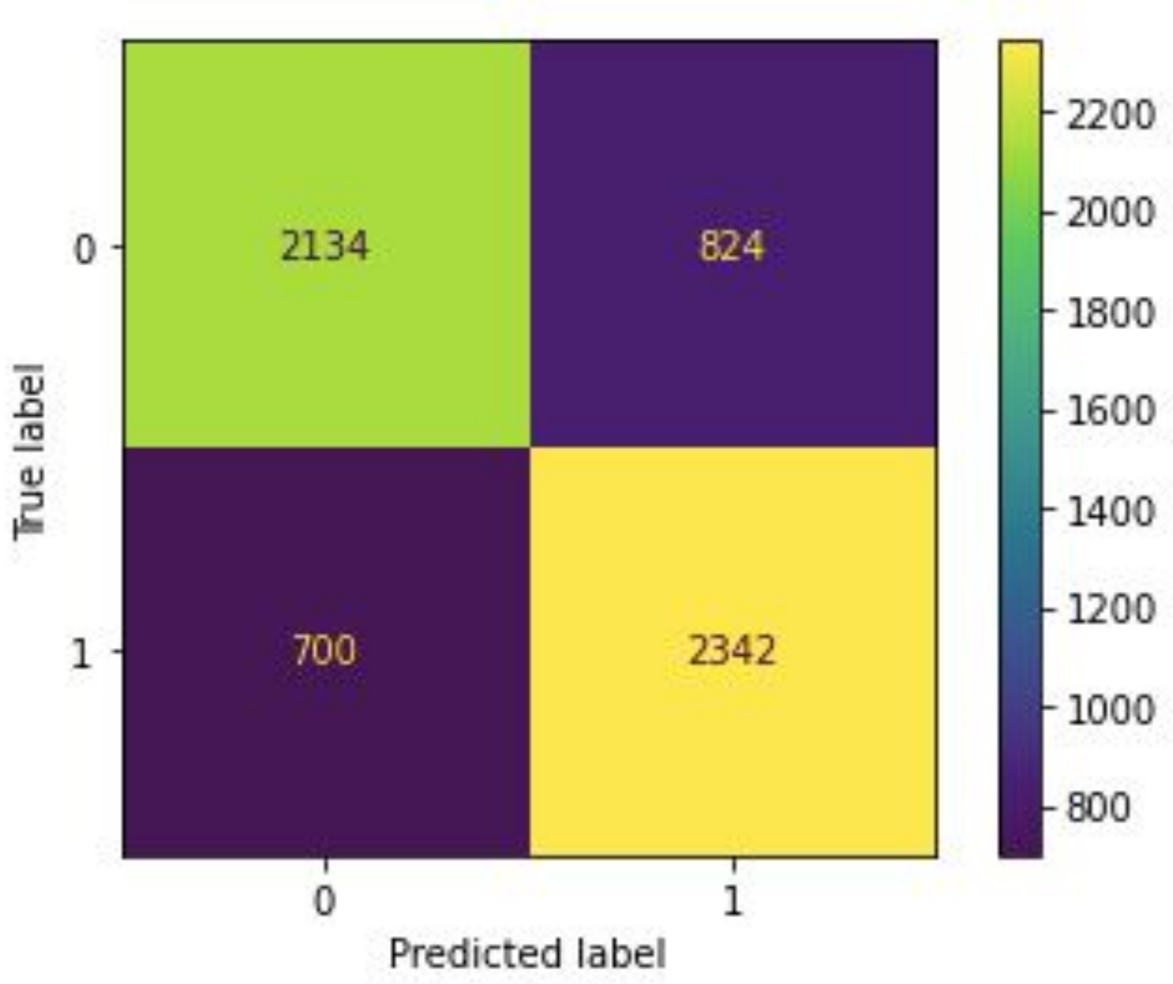
- Best feature classifiers

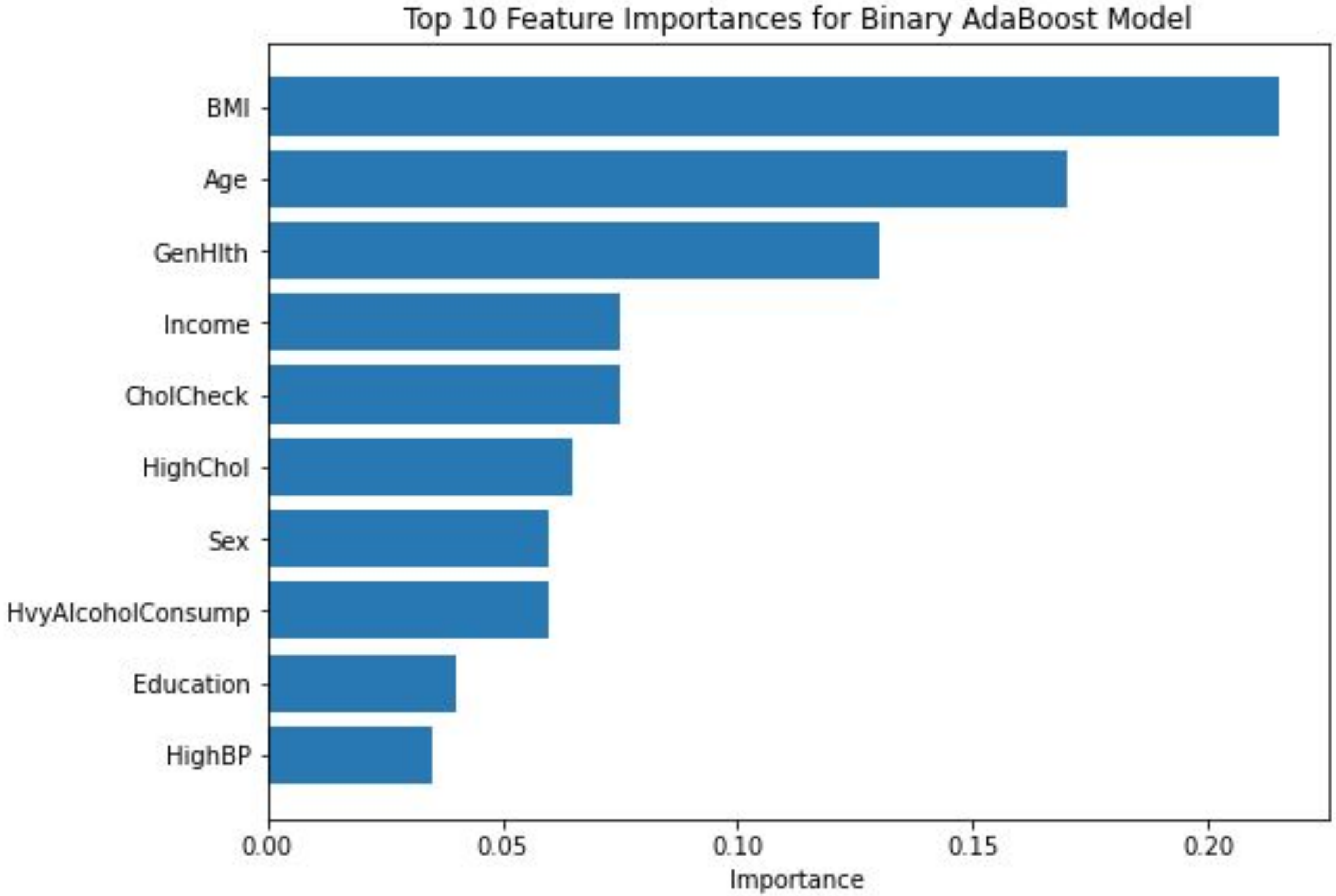
1. BMI
2. Age
3. GenHealth

Baseline Model - Binary (LR)



Final Model - Binary (AdaBoost)





Top 10 Feature Importances for Binary AdaBoost Model
BMI
Age
GenHlth
Income
CholCheck
HighChol
Sex
HvyAlcoholConsump
Education
HighBP
0.00
0.05
0.10
0.15
0.20
Importance

Recommendations + Next Steps

1. Predicting diabetes with phone screening is not very reliable
2. Binary model performs better than multi-class and still applies to our business case
3. Model can still be useful for a generalization of risk
4. Need to have final "test" on imbalanced set
5. Improve model by tuning hyper-parameters, try more models, polynomial feature engineering etc.
6. Look in the future to add additional biometric data to strengthen predictions
7. Prioritize feedback on BMI, age and general health on future surveys

Questions?

Geoff Vogt

Email: geoffrey.m.vogt.com

LinkedIn: https://www.linkedin.com/in/geoffvogt/

Github: https://github.com/gvogt2023