Diabetes Prediction & Risk Factor Analysis

Group 2
Team Members: Taylor Erickson

Questions we sought to answer

What are the strongest predictors of prediabetes and diabetes?

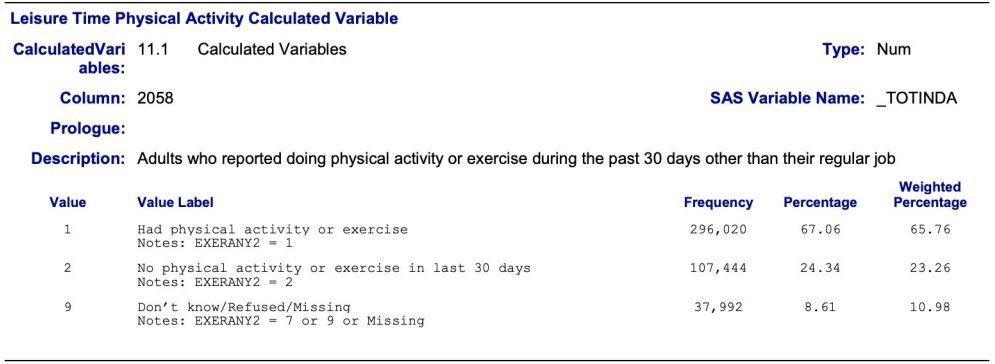
Are there meaningful differences between the predictors of prediabetes and diabetes, or are they similar?

Do these sets of predictors match existing screening standards?

Data Preparation Work

Datasets were extracted from Kaggle and transformed using the question format of the BRFSS code book.

- Records containing any null values were dropped
- Records with attribute values encoded as ‘missing’ were dropped
- Attributes were transformed to categorical numerical variables for modeling
- Records were aggregated into one large dataset



Leisure Time Physical Activity Calculated Variable

CalculatedVariables: 11.1 Calculated Variables **Type:** Num

Column: 2058 **SAS Variable Name:** _TOTINDA

Prologue:

Description: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Had physical activity or exercise Notes: EXERANY2 = 1	296,020	67.06	65.76
2	No physical activity or exercise in last 30 days Notes: EXERANY2 = 2	107,444	24.34	23.26
9	Don’t know/Refused/Missing Notes: EXERANY2 = 7 or 9 or Missing	37,992	8.61	10.98

Tools Used

Code prepared and written in Jupyter Notebook

Data Cleaning & Transformation:

- pandas
- os
- numpy
- random

Descriptive Statistics & Analysis:

- scipy
- Yellowbrick
- seaborn

Data Visualization:

- matplotlib
- collections

Modeling:

- statsmodels
- sklearn

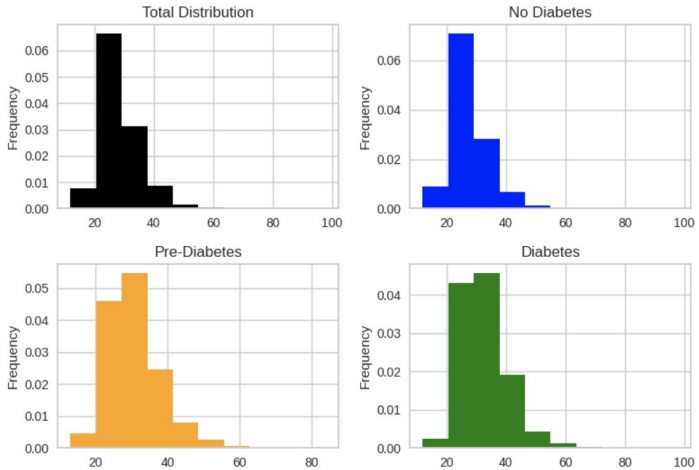
Training Data Strengthening (SMOTE):

- imblearn

Data Mining Methods: Feature Selection

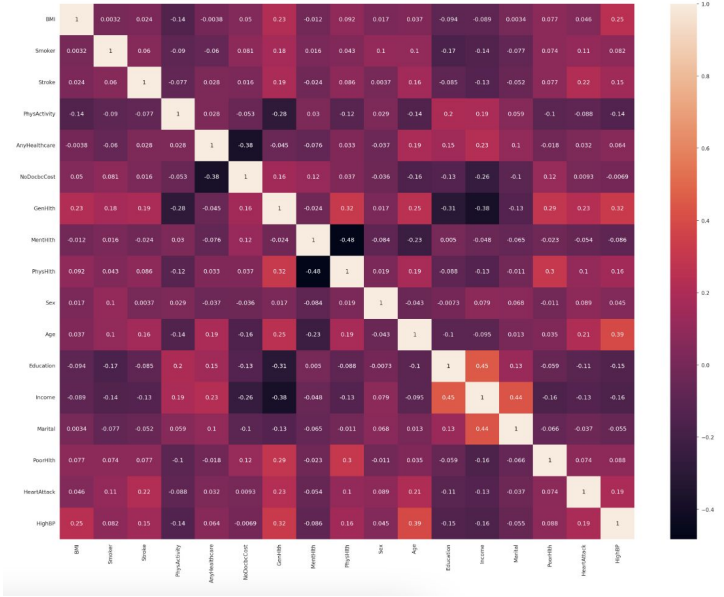
Feature Selection:

- T-test, F-Test
- ANOVA
- Chi-squared test
- Histograms, box plots
- Recursive Feature Elimination



Collinearity Analysis:

- Pearson correlation coefficient heatmap
- VIF score analysis



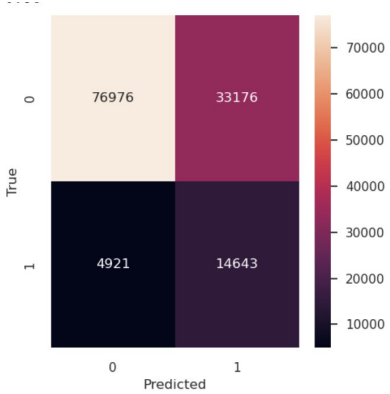
Data Mining Methods: Model Tuning, Performance Analysis

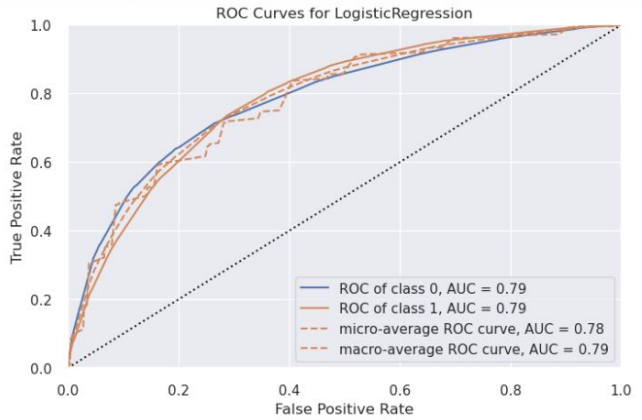
Modeling:

- Multinomial Logistic Regression
- Logistic Regression

Performance Analysis

- ROC/AUC curves
- Confusion matrix
- Accuracy, recall, and precision score
- Coefficient p-values
- Pseudo R-squared and log-likelihood





Knowledge Gained

There is strong similarity between predictors of prediabetes and diabetes.

All of these features were shown to increase the risk of prediabetes and diabetes.

BMI **Stroke** **Physical Health** **Heart Attack** **High BP**

However, there are **many more attributes that are predictors of diabetes.**

Smoker **No Doc (Cost)** **Mental Health** **Sex** **Marital Status** **Poor Health**

There is strong collinearity between many features thought to be associated with prediabetes and diabetes. Additionally, there is **high dimensionality** of the problem space, suggesting that there are **many factors that contribute to risk of prediabetes and diabetes.**

Knowledge Application

Support for existing screening standards

- There were not any features identified that conflict with existing screening standards
- However, not all features that are in existing screening standards were considered due to collinearity and lack of data availability

Identification of additional features that are usually classified as outcomes of diabetes

- Aligns with existing medical research on comorbidity of prediabetes,diabetes with other conditions
- For example - increased Stroke risk is usually considered an outcome of diabetes, rather than a risk factor leading to diabetes

Strong collinearity of features eliminated commonly identified attributes such as Age from analysis - suggests that it is likely that prediabetes and diabetes risk may not be able to be meaningfully modeled as a linear system.

Prediabetes and diabetes share the same strongest predictors, but there are others that are not consistent between the two conditions, including demographic and health outcome attributes