Modeling

Kelly Im, Brian Wang, Alvin Zou

Data Preprocessing

- Separated input and output variables
- Split training and test data (70%/30%)
- Fill numerical features with the feature's mean and filled categorical features with the feature's mode
- Removed features that implied that the participant already has diabetes
    - AB24 - Currently Taking Insulin
- Removed features that were the same except in different units
    - WGHTK_P - Weight in kilograms. Kept the alternative feature WGHTP_P - Weight in pounds.
- Used a Random Forest Classifier to get the important features for feature selection
    - Number of features: 763 → 52

Performance Metric Definitions

- Accuracy
  - The proportion of respondents who have or don’t have Type 2 diabetes and are correctly classified among the total number of cases examined.
- Precision
  - The proportion of people who have Type 2 diabetes and are correctly classified among all respondents who are predicted to have Type 2 diabetes.
- Recall
  - The proportion of respondents who have Type 2 diabetes and are correctly classified among those who actually have Type 2 diabetes.
- F1 Score
  - The harmonic mean of precision and recall. The value is between 0 and 1 - the closer the score is to 1, the more accurate the model.
- Confusion Matrix
  - A table used to describe the performance of a classification model on a set of test data for which the true values are known.

Performance Metric Equations

- Accuracy
  - (TP + TN) / (TP + FP + FN + TN)
- Precision
  - TP / (TP + FP)
- Recall
  - TP / (TP + FN)
- F1 Score
  - 2 * ((Precision * Recall) / (Precision + Recall))
- Confusion Matrix
  - Shown in the next slide.

Models

- Logistic Regression
  - Estimates the probability on whether a data point belongs to a certain class
- K-Nearest Neighbors
  - Takes the *k* nearest neighbors of a data point and assigns the data point the same label as the majority label within its neighbors
- Stochastic Gradient Descent
  - Iterative algorithm that starts from a random point on a function and randomly picks one data point at each iteration to travel down its slope until it reaches the lowest point of the function
- Decision Trees
  - Builds up a set of decision rules in the form of a tree which helps predict an outcome
- Random Forest
  - Consists of many decision trees and takes the prediction from each tree and selects the final prediction based on the majority votes of predictions

Confusion Matrix

Predicted

Actual

	Doesn’t Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn’t Have Type 2 Diabetes	True Negative (TN)	False Positive (FP)
Has Type 2 Diabetes	False Negative (FN)	True Positive (TP)

Methods Used To Deal With Unbalanced Classes

- Cost-Sensitive Learning
  - Takes the costs of prediction errors (and potentially other costs) into account when training a machine learning model.
  - Assigns different costs to the types of misclassification errors
- Synthetic Minority Oversampling Technique (SMOTE)
  - Oversampling method that selects a random sample in the minority class, finds its *k* nearest neighbors, randomly selects a neighbor, and then creates a synthetic example between the two examples

Logistic Regression

- Accuracy = 0.980118793732485
- Precision = 0.8885199240986718
- Recall = 0.9299900695134061
- F1 Score = 0.9087821445900048

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	49990	705
Has Type 2 Diabetes	423	5619

Logistic Regression Using Cost-Sensitive Learning

- Accuracy = 0.9853358478594215
- Precision = 0.8820768553828102
- Recall = 0.9953657729228732
- F1 Score = 0.9353032659409021

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	**49891**	**804**
Has Type 2 Diabetes	**28**	**6014**

Logistic Regression Using SMOTE

- Accuracy = 0.9850890952993637
- Precision = 0.882171226831421
- Recall = 0.9925521350546177
- F1 Score = 0.9341121495327102

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	**49894**	**801**
Has Type 2 Diabetes	**45**	**5997**

K Nearest Neighbors

- Accuracy = 0.9258861060683504
- Precision = 0.8048456687686691
- Recall = 0.40135716650115855
- F1 Score = 0.5356156819436775

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	50107	588
Has Type 2 Diabetes	3617	2425

K Nearest Neighbors Using SMOTE

- Accuracy = 0.8295292313657754
- Precision = 0.3473507148864592
- Recall = 0.6835484938762
- F1 Score = 0.46062904305152796

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	42935	7760
Has Type 2 Diabetes	1912	4130

Stochastic Gradient Descent

- Accuracy = 0.9386996140084953
- Precision = 0.9064679771718452
- Recall = 0.47318768619662366
- F1 Score = 0.6217920835145716

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	50400	295
Has Type 2 Diabetes	3183	2859

Stochastic Gradient Descent Using SMOTE

- Accuracy = 0.9711475756560974
- Precision = 0.8236590742101396
- Recall = 0.9276729559748428
- F1 Score = 0.8725772553903636

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	**47871**	**2824**
Has Type 2 Diabetes	**24**	**6018**

Decision Trees

- Accuracy = 0.974743112959797
- Precision = 0.8937297112591833
- Recall = 0.8657729228732208
- F1 Score = 0.8795292139554435

	Doesn’t Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn’t Have Type 2 Diabetes	**50073**	**622**
Has Type 2 Diabetes	**811**	**5231**

Decision Trees Using Cost-Sensitive Learning

- Accuracy = 0.9746197366797681
- Precision = 0.8961776859504132
- Recall = 0.8614697120158887
- F1 Score = 0.8784810126582278

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	**50092**	**603**
Has Type 2 Diabetes	**837**	**5205**

Decision Trees Using SMOTE

- Accuracy = 0.9761531275886988
- Precision = 0.8952959028831563
- Recall = 0.8788480635551142
- F1 Score = 0.8869957404159359

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	50074	621
Has Type 2 Diabetes	732	5310

Random Forest

- Accuracy = 0.9861994818196239
- Precision = 0.8881180811808118
- Recall = 0.9958622972525654
- F1 Score = 0.9389092611375517

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	49937	758
Has Type 2 Diabetes	25	6017

Random Forest Using Cost-Sensitive Learning

- Accuracy = 0.9860056048081499
- Precision = 0.8916417910447761
- Recall = 0.9887454485269779
- F1 Score = 0.937686391461309

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	**49969**	**726**
Has Type 2 Diabetes	**68**	**5974**

Random Forest Using SMOTE

- Accuracy = 0.9861818566367626
- Precision = 0.8869590815425376
- Recall = 0.9973518702416418
- F1 Score = 0.9389217824867561

	Doesn't Have Type 2 Diabetes	Has Type 2 Diabetes
Doesn't Have Type 2 Diabetes	49927	768
Has Type 2 Diabetes	16	6026

Best Model

- Random Forest Using SMOTE
- Highest F1 Score

Model	F1 Score
Logistic Regression	0.9087821445900048
Logistic Regression (Cost-Sensitive Learning)	0.9353032659409021
Logistic Regression (SMOTE)	0.9341121495327102
KNN	0.5356156819436775
KNN (SMOTE)	0.46062904305152796
Stochastic Gradient Descent	0.6217920835145716
Stochastic Gradient Descent (SMOTE)	0.8086535877452298
Decision Trees	0.8795292139554435
Decision Trees (Cost-Sensitive Learning)	0.8784810126582278
Decision Trees (SMOTE)	0.8869957404159359
Random Forest	0.9389092611375517
Random Forest (Cost-Sensitive Learning	0.937686391461309
Random Forest (SMOTE)	0.9389217824867561

Additional Models in Progress

- Neural Network Model
- Time Series Forecasting Model

PostgreSQL Databases

1. We have created multiple databases according to the different survey questionnaire sections. The database schema is attached to the email in a PDF.
   a. Screening Information (Ex: Household Size, Language of Interview, Proxy Interview, etc.)
   b. Demographic Information, Part I (Ex: Age, Gender, Race, Ethnicity, Marital Status, etc.)
   c. General Health Condition (Ex: Asthma, Diabetes, High Blood Pressure, Heart Disease, etc.)
   d. Health Behaviors (Ex: # Times Ate Fruit, # Times Ate Vegetables, # of Cigarettes per day, etc.)
   e. General Health, Disability & Sexual Health (Ex: Height, Weight, BMI, Type of Birth Control, etc.)
   f. Women's Health (Ex: Currently Pregnant)
   g. Mental Health (Ex: Feeling Nervous, Hopeless, Depressed, Restless, Chore Impairment, etc.)
   h. Demographic Information, Part II & Child Care (Ex: Citizen, English Proficiency, Education, etc.)
   i. Health Insurance (Ex: Covered by Medicare/Medi-cal, Covered for Prescription Drugs, etc.)
   j. Health Care Utilization/Access & Dental Health (Ex: Usual Source of Care, # Doctor Visits, etc.)
   k. Employment, Income, Poverty Status & Food Security (Ex: Work Status, # Hours Worked, etc.)
   l. Public Program Participation (Ex: Receiving TANF/CALWORKS, Food Stamps, SSI, Pension, etc.)
   m. Housing & Community Involvement (Ex: Currently Paying Off Mortgage, Active Member in Community, Own or Rent Home, Feeling Safe in Neighborhood, etc.)
   n. Demographic Information, Part III, Geographic Information (Ex: Rural or Urban)
   o. Voter Engagement (Ex: Frequency of Voting, Voter Engagement)