PREDICTING AND PREVENTING DIABETES USING DATA

Assignment: ISYE 7406 Project Presentation

Name: Regina Kang (rkang47, Group 2)

PROBLEM STATEMENT

- Diabetes is one of the most common chronic illnesses in the United States, affecting about 38.4 million people or around 11.6% of the population in 2021
- It was the 8th leading cause of death in the United States in 2021
- There is a significant amount of publicly available diabetes survey data collected by the CDC that can be used to:
  - Create models to predict whether a person has diabetes or prediabetes
  - Gain insights into the factors that are most likely to cause diabetes or prediabetes
- This project aims to create multiple models and compare their ability to predict diabetes or prediabetes using the given data, and provide other useful insights such as what features are most strongly associated with diabetes or prediabetes

DATASET DETAILS

Variable	Description
Diabetes_binary	Binary: 0 = no diabetes; 1 = diabetes
HighBP	Binary: 0 = no high BP; 1 = high BP
HighChol	Binary: 0 = no high chol; 1 = high chol
CholCheck	Binary: 0 = no chol check in 5 years; 1 = *
BMI	Int: body mass index
Smoker	Binary: 0 = smoked ≥ 100 cigs in life; 1 = *
Stroke	Binary: 0 = never had stroke; 1 = had stroke
HeartDiseaseorAttack	Binary: 0 = no CHD or MI; 1 = yes CHD or MI
PhysActivity	Binary: 0 = no phys act in past 30d; 1 = *
Fruits	Binary: 0 = eat fruit < once daily; 1 = *
Veggies	Binary: 0 = eat veg < once daily; 1 = *

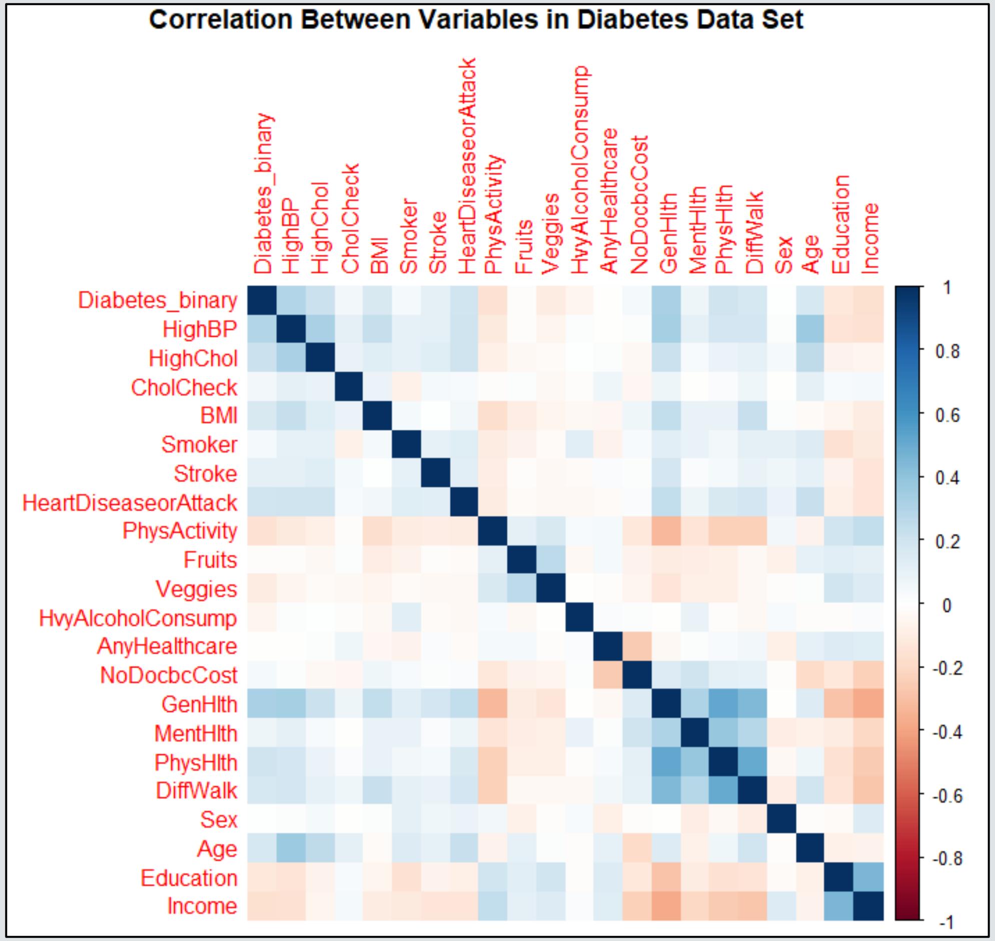
Variable	Description
HvyAlcoholConsump	Binary: 0 = not heavy drinker; 1 = *
AnyHealthcare	Binary: 0 = no health care coverage; 1 = *
NoDocbcCost	Binary: 0 = *; 1 = no doc bc cost w/i 12mo.
GenHlth	Int: rating of hlth 1 (excellent) – 5 (poor)
MentHlth	Int: past 30d - how many bad mental hlth
PhysHlth	Int: past 30d - how many bad phys hlth
DiffWalk	Binary: 0 = not hard walking/stairs; 1 = *
Sex	Binary: 0 = female; 1 = male
Age	Int: 13 age categories; 1 = 18-24; 13 = 80+
Education	Int: educat lvl; 1 (no school) – 6 (col. grad)
Income	Int: income scale; 1 (<10K) – 8 (≥75K)

*To save room, I used * to indicate that a binary value of 0 or 1 indicates the opposite of what is written for the other value

- The dataset used for this project is the *Diabetes Health Indicators Dataset* from Kaggle, and you can find the data by following this link: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
- The dataset contains 253,680 survey responses that the CDC collected in 2015; it has 21 predictor variables, and one response variable
- I used a stratified sample (1,269 observations) of the full dataset due to computational constraints
- The response variable, *Diabetes_binary*, has a value of 0 if the person does not have diabetes, and 1 if the person has diabetes or prediabetes
- Please see the table on the left for the descriptions of all the variables

EXPLORATORY DATA ANALYSIS

- To perform feature selection in the stratified sample dataset (which I will just refer to as "dataset" from now on), I performed stepwise variable selection using AIC in both directions
- I found that the model with the lowest AIC included *HighBP*, *HighChol*, *BMI*, *HeartDiseaseorAttack*, *GenHlth*, *MentHlth*, and *Age* as its predictor variables
- I also performed a VIF test to make sure I was not choosing any variables that had high multicollinearity
- None of the variables had VIF values that exceed 2.1, so I was able to conclude my feature selection to include all the features chosen by the stepwise variable selection above
- I also plotted a correlation heatmap shown on the right for an overall look at correlation between variables; as you can see, most of the colors are light, indicating low correlation, and confirming my VIF results



Correlation Between Variables in Diabetes Data Set

METHODOLOGY

- I chose to implement the following models: random forest, gradient boosting machine, logistic regression, linear discriminant analysis, Naïve Bayes, and a single tree
- I chose these methods because they are all known to be suitable for binary classification
- I performed parameter tuning on the ensemble methods (random forest, boosting), and used the features chosen from my previous feature selection for the non-ensemble methods (logistic regression, single tree, LDA, Naïve Bayes)
  - Feature selection is done implicitly within the random forest and gradient boosting machine models
- For both ensemble methods, I used k-fold cross-validation to find the best parameter combinations
- For random forest, I found that the best parameters were ntree = 500, mtry = 5, and nodesize = 7; the top four most important variables were *BMI*, *GenHlth*, *Age*, and *Income*
- For gradient boosting machine, I found that the best parameters were shrinkage = 0.05 and interaction.depth = 5; the optimal M number of trees was 70, and the feature variables with the highest relative influence were *BMI*, *Age*, *GenHlth*, and *PhysHlth*
- To make a robust decision on which model is best, I performed Monte-Carlo cross-validation with B = 50 iterations
  - I allocated 70% of the data to the training set and 30% to the test set using the sample() function, with different random observations chosen for each set in each iteration

METHODOLOGY CONT.

- Please see the table below for the pros and cons of each of the models I tested

Model	Pros	Cons
Random Forest	Implicit feature selection, robust to outliers, strong predictive accuracy	Black box, hard to interpret, computationally expensive
Gradient Boosting Machine	Good at handling complex data, strong predictive accuracy	Black box, computationally expensive, prone to overfitting
Logistic Regression	Easy to interpret, model coefficients can indicate feature importance	Does not perform well with complex relationships, linear decision boundaries
Single Tree	Simple, intuitive, easy to interpret, can handle non-linear problems	Prone to overfitting, sensitive to changes in data, computationally expensive
Linear Discriminant Analysis	Dimensionality reduction, simple, fast	Sensitive to outliers, prone to overfitting, assumes normal distribution of predictors
Naïve Bayes	Simple, fast, strong predictive performance when assumption of independence holds	Assumption of independence between predictors

ANALYSIS & RESULTS

Model	Mean Testing Error	Mean TE Variance	Mean Area Under ROC Curve
Random Forest	0.1442520	0.0003061934	0.5534754
Gradient Boosting Machine	0.1442520	0.0003104111	0.5714791
Logistic Regression	0.1472441	0.0003535132	0.5712868
Single Tree	0.1508136	0.0003076696	0.5247897
LDA	0.1510761	0.0003354924	0.5840230
Naïve Bayes	0.1790551	0.0002953202	0.6632690

- After performing Monte-Carlo cross-validation, I compared the performance of all my chosen models by averaging each model's testing errors and variances, as well as the AUC-ROC values across iterations
- Please see the table on the left for the results for each model ordered from lowest mean testing error to highest
- The two ensemble methods, random forest and gradient boosting machine, had the lowest mean testing errors
- This is expected because ensemble methods are known to have better predictive accuracy than standard models
- However, none of these models gave me a strong mean area under ROC curve

ANALYSIS & RESULTS CONT.

Model	Mean Testing Error	Mean TE Variance	Mean Area Under ROC Curve
XGBoost	0.0855118	0.0001524247	0.7521059
Random Forest	0.1442520	0.0003061934	0.5534754
Gradient Boosting Machine	0.1442520	0.0003104111	0.5714791
Logistic Regression	0.1472441	0.0003535132	0.5712868
Single Tree	0.1508136	0.0003076696	0.5247897
LDA	0.1510761	0.0003354924	0.5840230
Naïve Bayes	0.1790551	0.0002953202	0.6632690

- Therefore, I decided to try another ensemble method that is suitable for binary classification: XGBoost

- Following the same methodology I used for the other models, I tuned the hyperparameters of this model, making sure to test reasonable values that would not cause instability or overfitting

- I found the optimal hyperparameter values to be eta = 0.2, max_depth = 7, and subsample = 0.9; once again, I performed Monte-Carlo cross-validation with B = 50 iterations using these hyperparameters

- XGBoost performed significantly better than all the other models, with a mean testing error of 0.0855118 and a mean test error variance of 0.0001524247

- XGBoost also had a significantly better mean AUC-ROC value of 0.7521059

- While XGBoost and GBM are both boosting models, XGBoost offers improvements including regularization to prevent overfitting, parallel processing, speed, efficiency, and improved accuracy

- However, it does have its drawbacks including being computationally expensive and hard to interpret

CONCLUSIONS

- My results indicate that ensemble methods like random forest and gradient boosting machine are better at predicting diabetes status on this dataset than baseline methods like logistic regression, linear discriminant analysis, Naïve Bayes, and a single tree
  - However, low AUC-ROC values across all models led me to another model: XGBoost
  - XGBoost gave me the lowest mean testing error and variance, and the highest mean AUC-ROC value
  - I believe my XGBoost model is strong enough to conclude that the survey questions capture enough data to be able to accurately predict if a person has diabetes or prediabetes
- XGBoost is the best model to use if the goal is to predict whether someone has diabetes or prediabetes, but a baseline model like logistic regression might be more useful to understand what the most predictive features are due to its interpretability
  - While XGBoost also gives you feature importance, it is much harder to interpret because it is built on an ensemble of trees
- I have confidence in my results because I used k-fold cross validation to tune my hyperparameters for my ensemble methods, used feature selection when creating my baseline models, and performed Monte-Carlo cross-validation across 50 iterations and all models

REAL-WORLD APPLICATIONS

- If a person can answer all the questions in the CDC survey and wants to know right away if they are likely to have diabetes or prediabetes, this person should run their data through my XGBoost model
- Someone who wants to know which areas of their health they should focus on to avoid getting diabetes may want to look at my logistic regression model
    - My logistic regression model suggests that having high blood pressure, heart disease, poor general health and high cholesterol respectively have the strongest relationship with having diabetes or prediabetes
    - So to avoid getting diabetes, this person should look to improve these factors
- While I was unable to do so due to computational limitations, if my models were to be implemented for real-world use, I would recommend that the entire dataset be used
- If implemented for real-world use, my models have the potential to have a significant impact
    - They will give doctors and patients the ability to diagnose or self-diagnose diabetes by answering just a few questions, and
    - Give people data-driven insights into what factors to focus on to avoid diabetes

BIBLIOGRAPHY

- "Diabetes Statistics - NIDDK." *National Institute of Diabetes and Digestive and Kidney Diseases*, U.S. Department of Health and Human Services, www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics#:~:text=Estimated%20prevalence%20of%20diabetes%20in,8.9%25%20of%20the%20population). Accessed 25 Mar. 2024.
- "Statistics about Diabetes." *Statistics About Diabetes | ADA*, diabetes.org/about-diabetes/statistics/about-diabetes#:~:text=Deaths,a%20total%20of%20399%2C401%20certificates. Accessed 25 Mar. 2024.