BRFSS Data Analysis : Early Diabetes Prediction

Marcus Chery
Dept. of Computer Science
University of Central Florida
Oviedo, US
marcus.chery@knights.ucf.edu

Sanjay Shanbhag
Dept. of Computer Science
University of Central Florida
Oviedo, US
sanjayshan@knights.ucf.edu

Priyanka Gopi
Dept. of Computer Science
University of Central Florida
Oviedo, US
priyankagopi@knights.ucf.edu

Abstract* — The study analyzed pooled data from the Behavioral Risk Factor Surveillance System (BRFSS) (2016-17), a nationally representative cross-sectional survey collected by the CDC. An active research area where the experts from the medical field are trying to envisage the problem with more accuracy is diabetes prediction. Diabetes generally remains in dormant mode, and it boosts the other diseases if patients are diagnosed with some other disease such as damage to the kidney vessels, problems in retina of the eye, and cardiac problem; if unidentified, it can create metabolic disorders and too many complications in the body. The main objective of our study is to draw a comparative study of different classifiers and feature selection methods to predict the diabetes with greater accuracy. In this paper, we have studied Logistic Regression, Random Forest, and Histogram Boosting Gradient Classifier and feature selection was applied on the best classifier to detect the diabetes at an early stage.

***Keywords*—BRFSS, Data analysis, health, US**

I. Objective

To explore the dataset and develop a data model to predict diabetes and perform feature selection to detect diabetes at an early stage. The aim is also to perform imputation and exploratory analysis to address the raised question.

II. Related Work

The main motive behind the project was to identify how the CDC (Centre for Disease Control and Prevention) works on identifying and controlling the health issues across the U.S. from the survey data. It was interesting to learn the various aspects the data investigates, in order to obtain the relevant information and utilize them for analysis purposes. Diseases like Heart attack, diabetes, COVID19, pulmonary lung disease and many more are being analyzed by the CDC t learn and understand the causes and consequently, obtain important features and basically, encourage healthy lifestyle. Our study mainly focuses on Diabetes prediction and how the factors are responsible for the onset of Diabetes.

III. Introduction

The Behavioural Risk Factor Surveillance System (BRFSS) is a national survey that collects health-related data by telephone about U.S. residents (adults +18 years old) regarding their health-related risk behaviours, chronic health conditions, and use of preventive services. It was established in 1984 with 15 states, but not BRFSS collects data in all 50 states as well as the District of Columbia and three U.S. territories. The survey completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

BRFSS are often obtained through complex design, which involves stratification, clustering, multistage sampling, and unequal probability of selection of participants and responding rates. In order to make valid inference for the interested population where samples originated, appropriate statistical methods are required to analyse the complex survey data.

BRFSS is a continuous surveillance system created to assess behavioral risk factors in the non-institutionalized adult population of the United States (18 years of age and older). The goal of the BRFSS is to gather standard, state-specific data on preventive health measures and risk factors for chronic illnesses, accidents, and infectious diseases that affect adult population. Tobacco use, HIV/AIDS knowledge and prevention, exercise, immunization, health status, healthy days — health-related quality of life, access to health care, insufficient sleep, hypertension awareness, cholesterol awareness, chronic health conditions, alcohol consumption, and fruit consumption are among the factors evaluated by the BRFSS in 2013.

Since the data were gathered taking into account the population of all 50 U.S. states as well as the three U.S. territories, this study is population-based. Furthermore, telephone interviews were used to collect data in a random manner. It entails that participants are chosen at random, and the study only includes those who pick up the phone and consent to take the survey.

BRFSS has been conducting surveys using both landline and mobile phones since 2011. Interviewers obtain information from a randomly chosen adult in a home when conducting the BRFSS landline telephone survey. Interviewers gather information from an adult who participates in the BRFSS questionnaire using a cellular phone and lives in a private residence or college housing during the cellular telephone version of the study. The non-institutionalized adult population in the US who is 18 years of age or older and has health characteristics that are estimated from the BRFSS. Additional question sets were included as optional modules in 2013 to give a measure for several markers of children's health and wellness.

IV. Proposed Methodology

Our model begins with importing the SAS dataset from the official BRFSS website for the sample year, 2021, which is also the latest available data, currently. Data preprocessing steps are carried out to clean the data and make it suitable for EDA and modeling. As part of the pre-processing steps, feature engineering, dropping of irrelevant and duplicate columns were performed as they would, otherwise affect our model building and results, ultimately. Additionally, we handled missing values which was the major procedure in our analysis, by using the Imputation methods from Scikit class like the Iterative Imputer using Linear Regression.

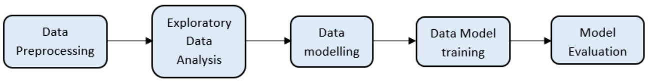


Fig 1 – Representing the model methodology pipeline

Modeling was done using the Logistic Regression, which is our baseline model, Histogram Boosting Gradient Classifier algorithm and Random Forest showed improved results. Important model evaluation metrics being used are Accuracy score, Confusion matrix, precision-recall and F1 score which are some of the best measures to evaluate a classification model. Following this, we performed feature selection which showed the most correlated features using the HG Boost algorithm.

V. Technical Details of Proposed Methodology

A. *Data Cleaning*

(i) Feature Engineering

One of the most crucial jobs when working on a typical classification problem is feature engineering, which involves extracting new features from the data. New columns were created for analysis, as the dataset contains variables with values specifically to represent missing data or in other terms, those left unanswered by individuals who had taken the survey. So separate variables were obtained to distinguish the data values from one another. This helps in achieving a data visualization more efficient and readable.

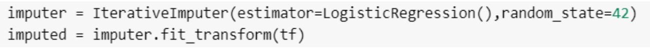
Features for important variables like diabetes, sex, smoking, alcohol consumption, BMI, heart attack etc. were created by creating a function and applying to the newly built variables.

(ii) Handling Null Values

The BRFSS is primarily a cross-sectional, self-report survey; as such, it is susceptible to recall bias and social desirability bias, which may affect which events respondents recollect or describe during the interview. For it being a survey data, we came across many null values in the data for most of the columns which was the most challenging aspect of the BRFSS data, and our motive was to fix the Nan values before proceeding with other preprocessing steps and further, modeling and analysis. There could be two reasons for missing values – Missing at Random (MAR) and missing at not Random (MANR). MAR is when several survey participants of the same gender leave a particular question in a survey blank. We can deal with this circumstance by doing a grouped mean/median replacement with the help of other features; the data that is absent can still be recovered. MNAR is the moment when we see a distinct pattern in the missing values. An instance of this is when respondents to a survey skip a particular question category since the question itself may be sensitive to the respondent the missing data will be difficult to retrieve unless additional research is conducted. MNAR is nonignorable, in contrast to the other two kinds of missing data.

To overcome the missing values due to the above reasons, we proceeded to dropping rows containing null values as it is not suitable for modeling. The pandas.dropna() was used to clear the rows containing NULL values in the dataset. Apparently, we ended up with 0 out of the total, 438693, which shows there are null values in every row of the data and realized that this approach is not appropriate.

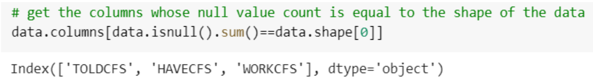
The second step we took was to use the Iterative imputer, an Ensemble based Imputation technique which has two types, i.e., Univariate and Multivariate. In our case, Multivariate Imputation technique was adopted as this can be used in cases where the data are MCAR, MAR, and even when the data are MNAR. Multiple imputation methods are known as multivariate imputation. Multivariate imputation basically aims to forecast the missing values in the current feature by using other features in the dataset. Iterative Imputer is a multivariate imputing technique that employs an estimate for imputation by modeling a column of missing values, like the target variable as a function of other features, like the predictor variables. As the dataset contained a combination of both categorical and continuous variables, we used the regressor for continuous and classifier for categorical columns separately and clubbed the columns back to one data frame. This method was able to impute all the columns effectively.



```
imputer = IterativeImputer(estimator=LogisticRegression(),random_state=42)
imputed = imputer.fit_transform(tf)
```

Fig 2 – Execution of the Iterative Imputer

We achieved good results from the imputation method and achieved zero null values in our dataset, as depicted below.

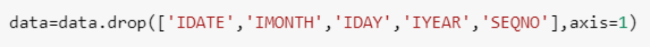


```
# get the columns whose null value count is equal to the shape of the data
data.columns[data.isnull().sum()==data.shape[0]]

Index(['TOLDCFS', 'HAVECFS', 'WORKCFS'], dtype='object')
```

Fig 3 – Checking for Null values post running the Iterative Imputer

(iii) Drop Irrelevant and Redundant Columns

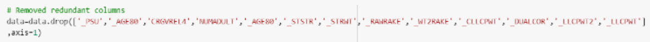
We also want to remove any irrelevant observations from the dataset. This is where our data doesn't fit into the specific problem we're trying to analyze. This will help us make our analysis more efficient and sensible.



```
data=data.drop(['IDATE','IMONTH','IDAY','IYEAR','SEQNO'],axis=1)
```

Fig 3 – Dropping unnecessary columns like the survey time, date and month

Removed redundant columns, by further looking up in the codebook to avoid duplication, which makes



```
# Removed redundant columns
data=data.drop(['_PSU','_AGE80','CRGVREL4','NUMADULT','_AGE80','_STSTR','_STRWT','_RAWRAKE','_WT2RAKE','_CLLCPWT','_DUALCOR','_LLCPWT2','_LLCPWT']
,axis=1)
```

Fig 4 – Dropping redundant columns like the methods of survey conduction

B. *Exploratory Data Analysis*

After extracting the features, we proceed to Exploratory Data Analysis – Univariate, Bivariate and Multivariate analysis. We were able to obtain some amazing insights.

About the response variable, 'diabetic', which is the diagnosis of diabetes, the dataset obtained after preprocessing them contains about 14% of diabetic class and 86% of non-diabetic class. This is a graph on overall 438693 samples and 303 variables.

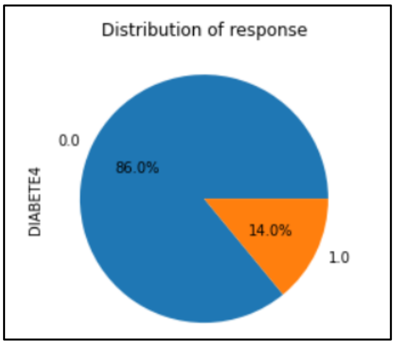


Fig 5 – Distribution of response

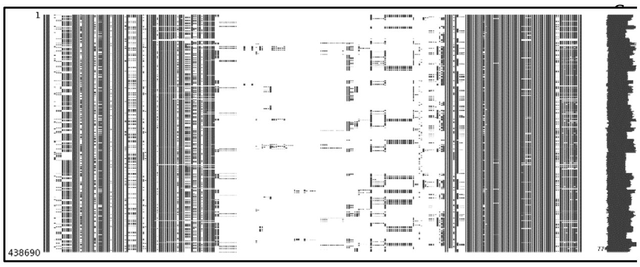
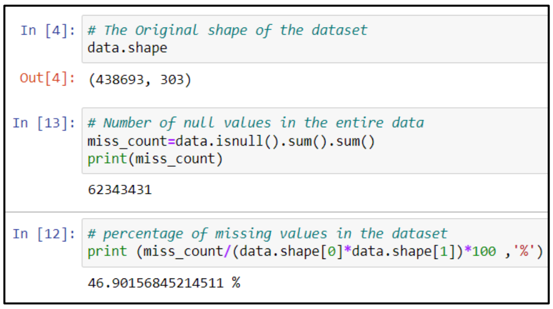


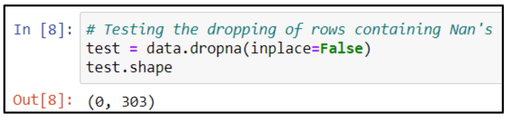
Fig 6 – NaN values in the dataset using missingno library

a. The above graph explains the number of missing values in the dataset. This matrix plot was plotted using the *missingno* python library which is extensively used to perform the analysis of Nan's in the datasets. The dark lines in the matrix plot represent the presence of the data whereas the white represent the absence or Nan's. It is also pretty much evident in the Fig 5 that the BRFSS data consists of almost **50%** Nan values, which makes it impossible to perform data modelling and get good classification accuracy.



```
In [4]:  # The Original shape of the dataset
         data.shape

Out[4]:  (438693, 303)

In [13]: # Number of null values in the entire data
         miss_count=data.isnull().sum().sum()
         print(miss_count)

         62343431

In [12]: # percentage of missing values in the dataset
         print (miss_count/(data.shape[0]*data.shape[1])*100 ,'%')

         46.90156845214511 %
```

Fig 7 - Numerical Figures and percentage of missing values in the dataset



```
In [8]:  # Testing the dropping of rows containing Nan's
         test = data.dropna(inplace=False)
         test.shape

Out[8]:  (0, 303)
```

Fig 8 – Elimination of sample data after dropping null values

b. From the Fig 8 below, on comparing the diabetic classes among the male and female population, we notice an increase in the non-diabetic cases for females and comparatively, lesser for males. The same applies to diabetic classes

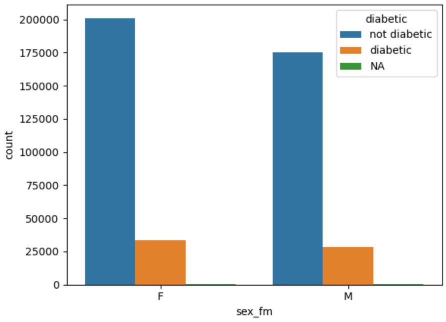


Fig 9 – Bivariate analysis of Sex with response, diabetic variables.

The above chart is supported by the pie chart below, which shows that our dataset consists of females count slightly on the higher side compared to the Male population.

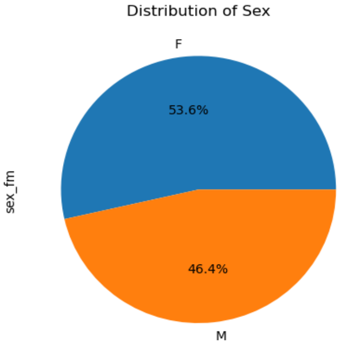
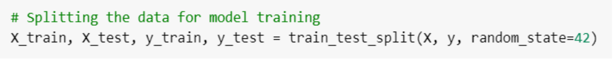


Fig 10 – Distribution of Male and Female count in dataset. 'M' and 'F' represent the male and female population, respectively

VI. Data Modeling

Over three machine learning models were trained and tested to note their performances on the test data. We have the predictors and target, and in our case, there is a good combination of categorical and continuous features. The data was split into the train and test data as below.



```
# Splitting the data for model training
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)
```

Fig 11 – Train and Test data splitting code

A. *Logistic Regression*

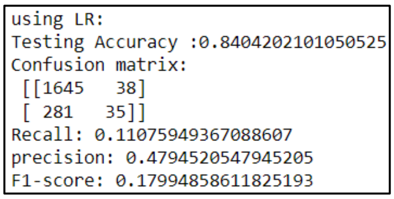
Logistic regression models are used to predict dichotomous outcomes (e.g., success/failure), or in our case diabetic and non-diabetic individuals with respect to predictors like individual's sex, race, occupation, number of medical visits, sore feet checkups, pharmacy visits and many more. These models can be fit with numerous approaches. One of the drawbacks faced during the modeling was each variables had to contain categorical data. In our dataset, as we reduced the data count by sampling, we missed a major portion of the variables data due to which they were not binary.

The logistic is of the form,

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

(1) – Amount of Nan values in the dataset

where intercept, β_0 = -μ/s of line y = β_0+ β_1x and rate parameter, β_1=1/s these are the *y*-intercept and slope of the log-odds as a function of *x*. Using the Logistic regression model, we obtained the following results.



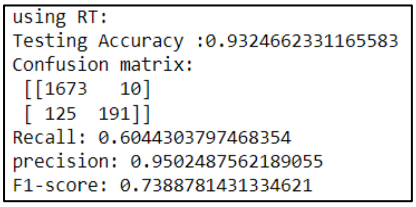
```
using LR:
Testing Accuracy :0.8404202101050525
Confusion matrix:
 [[1645   38]
 [ 281   35]]
Recall: 0.11075949367088607
precision: 0.4794520547945205
F1-score: 0.17994858611825193
```

Fig 12 – Logistic Regression model Results

From the above statistical results, we notice that although the accuracy is about 84%, the precision and recall scores are very low. This means that our model is not able to recognize and categorize the classes in its place accurately.

B. *Random Forest*

Next, we used the Random Forest classifier model, which showed results improved from the logistic classifier results. Here, many decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks. The class that the majority of the trees chose is the output of the random forest for classification problems. The mean or average prediction of each individual tree is returned for regression tasks.



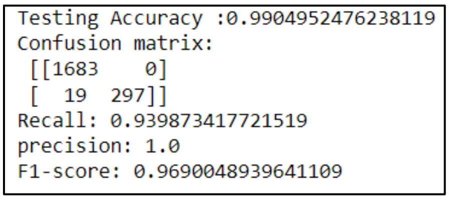
```
using RT:
Testing Accuracy :0.9324662331165583
Confusion matrix:
 [[1673   10]
 [ 125  191]]
Recall: 0.6044303797468354
precision: 0.9502487562189055
F1-score: 0.7388781431334621
```

Fig 13 – Random Forest model results

Evidently, the RF-model predicted the test response variables with an accuracy of 93.2% with precision-recall score equal to 95%-73% which is still an improvement from the Logistic regression model but not yet, the best model to conclude. The F1 score is still low and this only means that the model is able to recognize the classes but in not correctly predicting the test rows in its appropriate class.

C. Histogram Boosting Gradient Classifier

A group of decision tree techniques make up gradient boosting. Given that it performs so well in practice across a wide range of datasets, it may be among the most well-liked strategies for structured classification and regression predictive modeling issues. Gradient boosting has a significant issue with how slowly the model is trained.



```
Testing Accuracy :0.9904952476238119
Confusion matrix:
 [[1683    0]
 [  19  297]]
Recall: 0.939873417721519
precision: 1.0
F1-score: 0.9690048939641109
```

Fig 14 – HG Boost model results

The HG Boost model performs the imputation operation inbuilt without us having to do it before training. This is one of the reasons why this model is giving out the best results in terms of accuracy, Precision-recall and F1 scores. The testing accuracy is up to 99%, which is unimaginably great as the model can predict the classes accurately with a precision of score 1. We used the same model and performed feature selection.

VII. Feature Selection

Each property in the dataset has its significance estimated explicitly, enabling attributes to be ranked and contrasted with one another. Predicted target values against several criteria will be used to assess the feature relevance of the data.

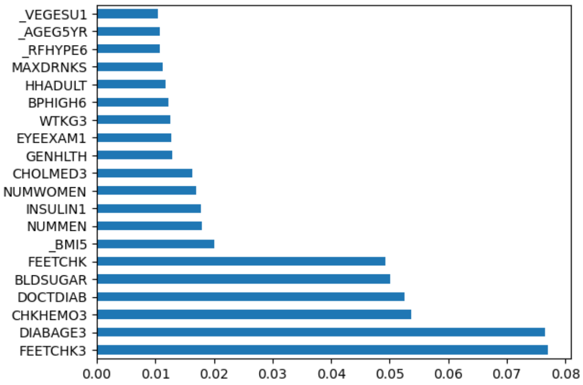


Fig 15 – Feature selection results

Features like count of medical visits for feet sore (FEEKCHK3), age of diabetes diagnosis (DIABAGE3), cholesterol medicine pharmacy visits (CHOLMED3), drinking frequency (MAXDRNKS), blood sugar (BLDSUGAR), insulin intake (INSULIN1), Body Mass Index (BMI5), individual's weight (WTKG3), eye exam results (EYEEXAM1), and high blood pressure (BPHIGH6) proved to be the most relevant features, which also makes sense. A person with increase in these features tends to have diabetes in real life too.

VIII. Model Evaluation Metrics

(i) Accuracy

Accuracy is the percentage of data points that are accurately anticipated. It is among the most basic types of evaluation measures. The accuracy score is calculated as the number of correct points divided by the total number of points.

(ii) Recall

Recall is a binary classifier metric that assesses the percentage of positive labels that are correctly predicted out of all positive labels. Formally, it is the ratio of the number of actual positive answers to the total of false negative responses that were incorrectly categorized as positive (false negative).

(iii) Precision

Precision is a binary classifier statistic that assesses the consistency of all positive labels. A binary classifier only produces two values (i.e., positive, and negative). To solve an issue with multiple values (in our case, three), we must first convert it into a binary classification problem.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

(2) – Precision-Recall calculation

(iv) F1-Score

The harmonic mean of a classifier's precision and recall is used to create the F1-score, which integrates both metrics into a single number. To compare the effectiveness of two classifiers, this method is usually employed.

$$\boxed{F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}}$$

(3) – F1 score equation

IX. Conclusion

In this paper, we developed a classification model to explicitly categorize diabetic classes from non-diabetic classes. We visualized various aspects of the data, understood the correlation between all the features, performed feature extraction, and performed predictive modelling using three algorithms. From the experiments the result that has been concluded is that Random Forest Classifier has an accuracy of 93% while Logistic Regression shows accuracy of 84%, but the best results are obtained by HG Boost that received a precise accuracy of 99.6%. The results obtained thus conclude that Histogram Boost gradient Classifier shows the most precise and high accuracy of 96.90%. Contrary to Logistic regression and Random Forest, which lack native support for data imputation and necessitate a data imputation phase, Histogram Boost gradient performs all data imputation on its own and yields outstanding results.

X. Future Prospects

Our model can be improvised much ahead by using the survey design using the stratification, clustering, multistage sampling in R and performing descriptive statistics by using the survey statistics like the svymean and svyby for bivariate analysis of independent and dependent variables. As BRFSS is a survey dataset, this is the best course of action to obtain good statistics as the survey design represents the entire dataset. Another element of the data is that each variable in the BRFSS treats "No response received," "Missing" or "Refused to answer" as categories, making it crucial to exclude them by further cleaning the data for all columns.

Acknowledgment

We thank Dr. Rui Xie (University of Central Florida) for giving us the opportunity to build this real-time project as part of the course Data Mining Methodology I (STA 5703). Thanks to CDC for releasing the BRFSS dataset for every year. Receiving guidance and help throughout the class and research assignment was crucial.

Appendix

The code for the analysis can be accessed from the link – BRFSS Data Analysis: Early Diabetes Prediction

References

[1] Handling Missing data in Machine learning (Towards Data Science) https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

[2] Handling Missing Data (Analytics Vidhya) https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

[3] Histogram gradient Boosting Classifiers (medium) https://towardsdatascience.com/meet-histgradientboostingclassifier-54a9df60d066

[4] Guide to deal with Missing values (Analytics Vidhya) https://www.analyticsvidhya.com/blog/2021/10/guide-to-deal-with-missing-values/

[5] Data Imputation using Amazon Science Data Wig (medium) https://towardsdatascience.com/imputation-of-missing-data-in-tables-with-datawig-2d7ab327ece2