Diabetes Risk Prediction for Screening

Pavan Kalyan Reddy Kamasaani
Data Science Practicum-1
Regis University, Denver, CO, USA
pkamasaani@regis.edu

Abstract

This project is about finding the people who are suffering from diabetes before even doing the blood test or lab work. To find that, there is a need for data which answers the questions that relate to diabetes. That is how I got the BRFSS 2024 survey data, which answers the questions I needed. Now that I have the answers from the dataset, all I need is a plan to use the methods that are needed for that. In this case, I am using data science and machine learning methods to further understand the factors that are affecting and are related to diabetes. The workflow that includes these methods is in the order of data cleaning, EDA (exploratory data analysis), feature selection, feature engineering, model building, and model comparison. Here, different models are used and tested to find which model gives the best performance as per our requirement for prediction. Since the work is related to the healthcare industry, the main focus will not be just on accuracy, but also on other measures like recall, precision, calibration, fairness across age groups, and interpretability, which are very important to consider. The overall goal of this project is to create a screening model for diabetes that could simulate and work in the real world. For that reason, the data of almost half a million respondents, which represents the American population, has been used, so that it would be meaningful and relate to the real population in America.

I. Introduction

It begins with a simple question: why is finding diabetics so important? To answer that, let us start with what diabetes is. It is a common long-term health condition in the USA, as well as the world, and it creates long-term serious health concerns if it goes undiagnosed. The common problems faced due to diabetes are heart disease, kidney damage, nerve damage, and issues with eyesight. Because of these risks, early detection can be a lot of help in a person's life. So, with the help of data science and machine learning, I want to address these issues by analyzing health, lifestyle, and demographic factors to understand which people are having diabetes and are at higher risk.

The dataset I collected for this project is the 2024 BRFSS medical survey. This survey is conducted by the CDC every year from the citizens of the USA about their demographics, smoking status, health status, and many other variables that are related to diabetes. In our case, it will help me develop a prediction risk model for diabetes using this information. This topic will also help me understand how real data affects data science and machine learning models, which could in turn be useful in supporting the healthcare system and save a lot of people just by analyzing the respondents' answers for screening and decision-making, while also considering fairness, interpretation, and responsible model evaluation.

II. Problem Statement

The main problem in this practicum project is to build a model that could predict diabetes risk just by using the responses given by the respondents on things like their health behavior, demographics, and clinical indicators. Basically, these responses are the ones that can be answered without any medical checkups. In terms of data science, this really makes a lot of sense, but in practice there are several challenges that come with it. BRFSS is a large dataset with many coded variables and various columns that have different rules for values, missing responses, refused responses, and responses like "don't know." So, the data has to be cleaned very carefully, because mistakes here can result in wrong values. Secondly, the dataset has a lot of class imbalance. This dataset also contains both numerical and categorical data, so the preprocessing should be handled properly before building the model, and accuracy alone cannot

decide the quality of the model because the model may appear accurate, but too many diabetes cases can be missed.

The other challenges here are related to screening. The model should consider measures like recall, precision, threshold selection, calibration, fairness across subgroups, and interpretability as very important because, in the medical industry, missing a diabetic is much worse than flagging someone who does not have a problem. So, seeing these problems, the focus in this project is not just to train a machine learning model, but to create a usable prediction model that does not miss a person who is at risk.

III. Methodology/Approach End of the course

1. Data Collection

The first step for any research or project is data collection. This is the crucial step for anyone who is working in data science. In this case, the BRFSS 2024 dataset is collected from the CDC, available online. This dataset contains a large amount of information that connects to the health of near to half a million people. Since this is a large dataset, it will also be relevant to real-world screening.

2. Target Variable Definition

Once done with collecting the data, the next step comes to defining the target variable. In this case, whether a person has diabetes or not. Only valid responses are used for binary classification, whereas invalid are removed. Defining the target variable is a crucial part because if there is a mistake, the whole model will learn from wrong labels and give misleading results, which will not be of any use.

3. Data Cleaning

Here, the dataset is cleaned with proper caution because BRFSS uses a different coding system for valid values, missing responses, refused answers, and "don't know" responses. Every selected variable is checked and recoded properly, because every variable has its own coding rules, and using generic cleaning can remove valid values, and it is possible for it to happen vice versa for invalid ones.

4. Exploratory Data Analysis (EDA)

Once completed with the cleaning, EDA is performed to understand the data structure, class distribution, and the relationships between diabetes and variables such as age, BMI, general health, exercise, and income. These tables and plots help further to understand important patterns that can be useful in feature selection, feature engineering, and model decisions later.

5. Feature Selection

Feature selection is the place where variables are selected based on their relevance to the diabetes prediction, which includes demographic factors, lifestyle factors, and health-related factors. The variable selection must be done very carefully because unnecessary or weak variables can reduce the quality of the model.

6. Feature Engineering

New features are created from the already existing data. This is like a shortcut that makes it easy for the model to understand the pattern and improve the performance of the model. Features like BMI categories, missingness indicators, and combined risk are used here. Creating features can find the unseen connection in the data that the individual variable cannot see.

7. Train, Validation, and Test Split

Once we are ready with the dataset, this is where it is divided into train, validation, and test sets to train, tune, and evaluate the model properly. To keep the class distribution balanced, stratified splitting is used here. Avoiding data leakage is very important for this step because there is a possibility the model can look better than it originally is because of the leakage.

8. Data Preprocessing

Data preprocessing involves steps like handling missing values, encoding categorical variables, and scaling numerical variables. Doing this gives a correct format for the machine learning models. Messing up this step will directly affect the performance.

9. Model Building

With this prepared dataset, different models are trained, which include Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, LightGBM, and XGBoost. Each model takes its own approach to solve the same problem. In this case, the same prepared dataset is used, and since each model works in its own way, that captures different patterns, a model can perform better in one aspect than another, and vice versa for another model in another aspect.

10. Model Comparison and Evaluation

After training, these models are compared with different metrics like recall, precision, F1 score, F2 score, ROC-AUC, and PR-AUC. Since accuracy alone is not enough for healthcare-based design, because there is no use for accuracy if it misses too many diabetics.

11. Threshold Selection

This is one of the most important steps to get the desired results. Once the best-performing models are decided, the threshold must be set to an appropriate range because the default threshold is not suitable for these types of real datasets. The threshold is adjusted in such a way that it balances the recall and precision, with more focus on identifying true diabetics. Adjusting this threshold should always align with the goal of screening.

12. Calibration and Fairness Analysis

Calibration is a step used to check whether the obtained prediction probabilities by the model are accurate. Fairness analysis shows how the model is performing across different subgroups of age.

13. Model Interpretation

Interpretation methods like SHAP and others are used to tell how each feature contributes to the diabetes prediction. This will make it easy to understand how the model works and what factors or features affect the prediction, because understanding the models helps in understanding how the predictions are made.

14. Final Model Selection

Selecting the final model depends on the overall performance, screening usefulness, fairness, calibration, and interpretability. The final model selection does not depend on just one metric, it is done by considering its usefulness in practice because the goal is to create a model that can be used for screening

IV. Data Description End of the course

You need to share the basic information about the dataset during the initial proposal submission. Collection / Acquisition The data I used for this project is from the BRFSS 2024 survey dataset by the CDC. The reason for using this particular dataset is because the CDC conducts one of the largest public health surveys in the world that contains various information such as demographics, health-related, and behavioral data of a large number of people. Since the dataset contains enough information in large numbers, it is used for this real-world screening-based project.

shape of the data set: 457,670 rows, 301 columns

Preparation As discussed in the methodology section, a variable called DIABETE4 is used to create the binary target variable $y_diabetes.Toexplainindetail, respondentswhoreportedhavingdiabetesarecodedas1, andres$ $65,809$

Negative non-diabetes cases: 376,125

Excluded during target construction: 15,736

Weighted prevalence: 12.98

EDA The importance of EDA is that it helps in understanding the patterns and relationships between a particular variable and diabetes status. This information is useful in understanding how each selected feature will perform as a predictor. For example, using EDA revealed that there is a lot of class imbalance, where 14.89

The analysis also showed the difference between diabetics and non-diabetics across other variables like age, BMI, general health, and physical health. People who have diabetes usually have higher BMI and poor general health. Another observation is that diabetes increases as age increases, which means older age groups have more number of diabetics. This shows how age is an important factor for screening analysis. Add these values here:

Class distribution: 14.89
Mean BMI:
Diabetes: 31.465
No diabetes: 27.946
Mean general health:
Diabetes: 3.293
No diabetes: 2.518
Mean physical health bad days:
Diabetes: 7.995
No diabetes: 3.959

Visualization Visualization is just a part of EDA that makes the patterns in the data easier to understand. The plots below show the class imbalance, the diabetes rate across various age groups, and there are other plots that show different variables that affect diabetes in different ways.

Reporting From data preparation, EDA, and visualization, the process has led to the modeling stage. As it is understood that this is a healthcare screening project, the data analysis shows that there should be more focus on recall, precision, PR-AUC, calibration, fairness across age groups, and threshold selection, rather than only depending on accuracy. Overall, the data analysis concludes that variables such as age, BMI, general health, and physical health are strongly related to diabetes risk and are useful for prediction.

V. Expected Outcomes

This begins with obtaining the documentation for BRFSS 2024 and creating a data dictionary that clearly explains which variables I am going to use. Then, I will create a cleaned dataset for modeling and use it to produce EDA plots that describe the dataset and the patterns it contains, such as the number of positives and negatives, how the predicted risks are distributed, and the relationships between key factors.

used models such as logistic regression and random forest. Logistic regression will capture the basic patterns, which makes it easier to understand, while random forest will capture more complex patterns. After training these models, I will create performance tables and evaluation results that are needed for screening decisions.

Next, check whether the probabilities obtained are well calibrated and explain how calibration affects screening decisions. I will also analyze errors by examining the patterns of false negatives and false positives to understand why these errors occur. Then, I will explain how each factor affects the predictions. Finally, I will produce the final report that summarizes the full process, results, and conclusions.

VI. Timeline

The project will be completed over an eight-week practicum period with the following milestones:

- **W**eek 1: proposing the idea of project
- **W**eek 2: finalizing the project based on the feedback
- **W**eek 3:obtaining the data BRFSS 2024 published by CDC(Centers for Disease control and prevention) and identifying related variables
- **W**eek 4: cleaning and creating final target variable and perform EDA and create plots needed
- **W**eek 5: building baseline logistic regression model and choosing models needed for comparison and usage
- **W**eek 6: training and caomparing these models and selecting threshold.
- **W**eek 7:calibration, fairness checking, interpretation and selecting the final model
- **W**eek 8:presentations, complete the reports and submission.

VII. Conclusion End of the course

this practicum teaches on how to develop a diabetes risk prediction model and alos helps in understanding how real-world data affects it. It has provided with a deep understanding of the challenges that need to be faced when working with real data that has a lot of class imbalance. Trying different models helped in understanding the patterns and referring as much as possible to select one final model. The process has provided an understanding of how important machine learning is for solving public health problems. This concludes that these models are not only useful for prediction, but are also practical for healthcare screening.

Note: The final practicum report must cover the following components:

1) Project Title
2) Abstract
3) Introduction/Background of the Project
4) Problem Statement
5) Literature Review/Related Work Optional
6) Methodology/Proposed Approach End of the course
7) Data Analysis Collection/Acquisition, Preparation, EDA, Visualization, Reporting End of the course
8) Expected Outcomes
9) Timeline
10) Conclusion End of the course
11) Reference

References

[1] Centers for Disease Control and Prevention, "National diabetes statistics report," CDC website, n.d., accessed 2026-01-26. [Online]. Available: https://www.cdc.gov/diabetes/data/statistics-report/index.html

[2] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330, also published in PMLR, Volume 70. [Online]. Available: https://arxiv.org/abs/1706.04599

[4] Centers for Disease Control and Prevention, "BRFSS overview: 2024," 2025, accessed: 2026-01-30. [Online]. Available: https://www.cdc.gov/brfss/annual_data/2024/pdf/Overview_2024-508.pdf

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.