Building Risk Prediction Models for Type 2 Diabetes

Martell N. Tardy

September 13, 2021

1 Domain Background

1.1 Overview

The World Health Organisation estimates that by 2030 there will be approximately 350 million people with type 2 diabetes. This is double the current number. Diabetes type 2 is associated with renal complications, heart disease, stroke and peripheral vascular disease, and individuals with diabetes have mortality rates nearly twice as high as those without. Of these projected diagnoses, only around one half will know they have the condition. This has been shown repeatedly in epidemiological surveys. Therefore, early identification of patients with undiagnosed type 2 diabetes or those at an increased risk of developing type 2 diabetes is an important challenge [1].

There are 3 main types of diabetes: type 1, type 2, and gestational. Of those 3, type 2 diabetes is the most prevalent and accounts for 90% to 95% of all cases. Type 2 diabetes is a predictable and preventable disease because it usually develops later in life (age >30) as a result of lifestyle (eg, low physical activity, obesity status) and other (eg, age, sex, race, family history) risk factors [2,3]. Type 2 diabetes is usually diagnosed using the glycated hemoglobin (A1C) test [6]. However, if a more complete identification of the risk factors that cause this disease were developed, providers could diagnose, treat and prevent the disease better.

2 Problem

This paper aims to explore the relation between the behavioral traits and lifestyle and the probability of being diabetic using machine learning techniques to build a predictive model for the identification of risk factors for type 2 diabetes. To achieve this objective, we first identify the possible risk factors contributing to diabetes from a set of selected features using a Light GBM (Gradient Boosting Machine) based predictive

model. Then we venture forward to analyse the impact of the individual selected features on the model outcome to prove the soundness of the feature statistically.

The target variable for this project is represented by a binary classification of Yes or No answers to the question "Have you ever been told you have diabetes?". In order to understand the dataset, each variable is visualized to initially assess the distribution and/or odds ratio of each variable regarding the target variable. It is important to note, there are prior studies of models built to predict the occurrence of type 2 diabetes. However, because of its causal complexity, the prediction performance (especially sensitivity) of models for type 2 diabetes based on survey data have room for improvement. In addition, although many risk factors, including obesity and age, are well established for type 2 diabetes, others remain to be identified. A comprehensive list of common risk factors compiled from literature on diabetes can be seen in Appendix A [1].

3 Datasets and Inputs

For this study the Behavioral Risk Factor Surveillance System (BRFSS) database is used. In 1984, the Centers for Disease Control and Prevention (CDC) initiated the state-based Behavioral Risk Factor Surveillance System (BRFSS) a cross-sectional telephone survey that state health departments conduct monthly over landline telephones and cellular telephones with a standardized questionnaire and technical and methodologic assistance from CDC. BRFSS is used to collect prevalence data among adult U.S. residents regarding their risk behaviors and preventive health practices that can affect their health status [4].

BRFSS survey data from 2015 was downloaded locally in ASCII format and converted into CSV format for simplicity. In total 330 features and 441,456 observations were initially observed in the dataset. Since the variables for this project are not predetermined, this study applies a hybrid (human + machine) approach to reduce dimensionality of the dataset.

3 .1 Feature Selection

First, the dimensionality of the data was reduced by manually filtering the variables which met the following criteria; a) completely irrelevant variables, which have no clinical or research implication (e.g., "Do you also have a landline telephone?"), b)

very highly correlated variables (e.g, “Have you ever been told by a doctor or other health professional that you have pre-diabetes or borderline diabetes?” is similar to the outcome variable), c) redundant variables (e.g., age categories, race categories). In addition, all missing values were removed, bringing the total number of responses down to 144 as recorded in the jupyter notebook Data Wrangling. This includes the target variable `TYPE2` identified during this process.

Secondly, the machine learning algorithm Light GBM was applied to reduce dimensionality. As a result of this algorithm, 25 features were selected as the final set of features for this project [Appendix A].

3.2 Feature Preparation

The 25 features selected were then modified using target encoding to prepare for the preprocessing and training process ahead. Note, the responses that were marked ‘Don’t Know/Not Sure’ and ‘Refused to Answer’ were not removed during this step and ordinal variables were not changed to a numeric scale since target encoding can handle these aspects of categorical variables. The replacement method was initially applied to the categorical features to turn them from numerical string values into character string values. During this process the original size and the target encoding size of the dataframes remained the same, however, five of the features experienced a reduction in the number of original categories after this process. For three of these features it was an issue of overfitting, a common issue with target encoding, that was addressed using a smoothing technique where weight was used as a hyperparameter. The remaining two affected features were simply experiencing target encoding combining the categories with only one observation into one new category with two recorded observations.

4 Experimental Setup and Methodology

For this study 10 research questions were proposed and explored for potential solutions using exploratory data analysis as seen in the jupyter notebook Research Questions.

Research question 1: In the diabetic sample population `TYPE2`, is there a difference between males and females and the incidence of diabetes type 2? The question looks to determine if gender has any relationship between increased prevalence of diabetes among either the male or female population.

Research question 2: What is the probability that participants with access to health care coverage will have been diagnosed with diabetes `DIABETE3`? and how many of these participants were diagnosed with type 2 diabetes `TYPE2`? The question looks to determine whether there is a relationship between participants with health care coverage and therefore, being able to afford the cost of diabetic screening and participants diagnosed with diabetes.

Research question 3: Is there an association between age `AGEG5YR` and a participant being told diabetes has affected their eyesight? The question looks to determine whether there is a relationship between blindness and age within the diabetic sample population. This is of interest because diabetic retinopathy is a leading cause of blindness among the elderly diabetic population.

Research question 4: Is there an association between being diagnosed with a heart attack and Body Mass Index `_BMI5CAT` among participants who are diabetic? This question looks to determine whether there is a relationship between Body Mass Index and heart attacks among diabetics.

Research question 5: Is there an association between participants who have ever smoked tobacco or consumed alcohol and the participants diagnosed as diabetic? This question looks to determine whether there is a relationship between the amount of tobacco and or alcohol consumed amongst those participants with diabetes.
Research question 6: Is there an association between a participant's intake of fruits and vegetables and a diagnosis of diabetes? This question looks to determine whether there is a relationship between a healthy diet and diabetes.

Research question 7: What is the probability that participants with regular exercise have diabetes? This question looks to determine whether there is a relationship between the various levels of physical activity a participant does and their likelihood of being diabetic.

Research question 8: Is there an association between where a participant lives and if she/he is diabetic? his question looks to determine whether there is a relationship between the physical state, demographic region (ie. rural, city), and diabetic status of a participant.

Research question 9: What is the probability of participants being diagnosed with some form of diabetes and having an advanced degree? This question looks to determine whether there is a relationship between pre-diabetes, gestational diabetes, or type 1 diabetes and the education level attained by these participants.

Research question 10: What is the probability of a diabetic veteran? This question looks to determine whether there is a relationship between the shared occupational experience of retiring from the military and being diagnosed diabetic.

5 Benchmark

The independent variables are then assigned to X and the target variable to y. Then the data was divided into two data sets which are training and testing in a 70/30 split approach. Next, a baseline model was built with the final set of selected variables using the default parameters of a LightGBM binary classification algorithm. The accuracy score for the baseline model is 90.8%. Lastly, hyperparameter tuning was performed to improve the performance of the model with a focus on parameters that improve model accuracy. These parameters were the number of trees(n_estimators), tree depth(max_depth), learning rate, and boosting type. The best parameters were n_estimators at '500' (default = 100), max_depth at '9' (default = -1), learning rate at '0.1' (default value), and boosting type at 'gbdt' (default value). A final prediction model was chosen based on classification accuracy using these hyperparameters.

6 Solution

After building our model using the best performing parameters we achieved an accuracy score of 91.5%, a 0.8% increase from our baseline model, with a precision accuracy of 92% for those who are not type 2 diabetic and 77% for those who are. We also confirmed the 5 leading characteristics of someone having type 2 diabetes are related to the parameters prediabetes, age, state, blood pressure medication and cholesterol checks. Specifically, we can conclude that if a participant has prediabetes or has had prediabetes before this is the leading indicator for a type 2 diabetes diagnosis. As the age of a participant increases so does the likelihood of that participant having a type 2 diabetes diagnosis. If a participant takes blood pressure medication they are more likely to have a type 2 diabetes diagnosis. Lastly, participants who got their cholesterol checked "within

the last year" and those who answered "don't know" are more likely to be diagnosed with type 2 diabetes [Appendix A].

7 Suggestions

Address the research questions with exploratory data analysis. Continue hypertuning the baseline model to see if there can be an improvement in model performance. Address imbalance within the training dataset. Examine the effects of various encoding techniques and how features with large numbers of categories are initially binned, especially for the feature `_STATE`. Explore the use of monotonic constraint on ordinal features within the final dataset after applying an encoding technique. Compare performance of LightGBM with other algorithms such as XGBoost and Catboost.

References

[1] Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011 https://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-9-103

[2] Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011 https://pubmed.ncbi.nlm.nih.gov/22123912/

[3] Sullivan PW, Morrato EH, Ghushchyan V, Wyatt HR, Hill JO. Obesity, inactivity, and the prevalence of diabetes and diabetes-related cardiovascular comorbidities in the U.S., 2000-2002. Diabetes Care 2005 https://pubmed.ncbi.nlm.nih.gov/15983307/

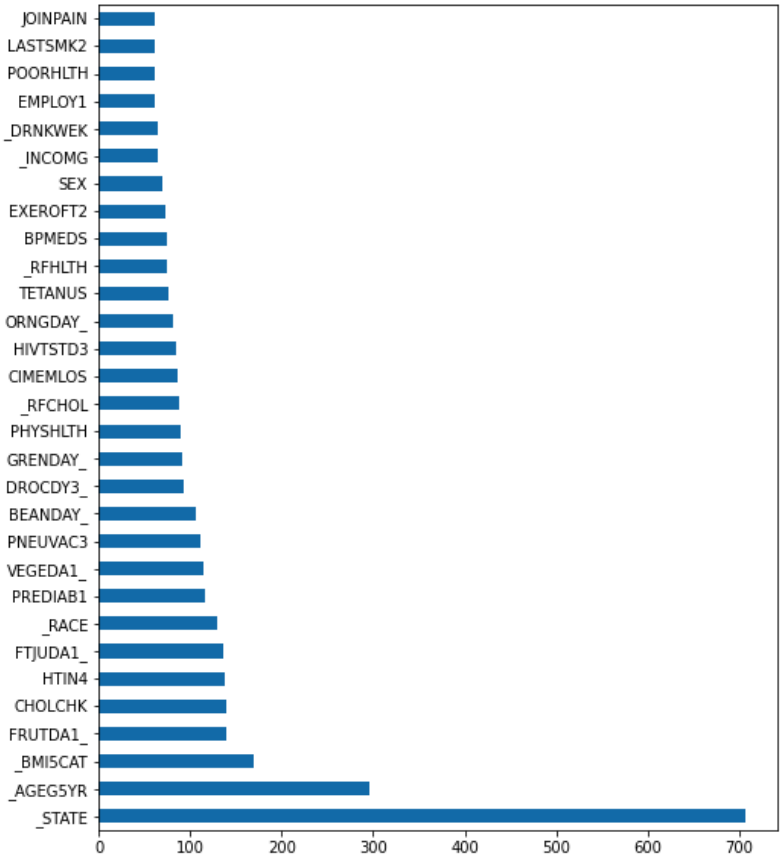
[4] BRFSS data https://www.cdc.gov/brfss/data_documentation/index.htm

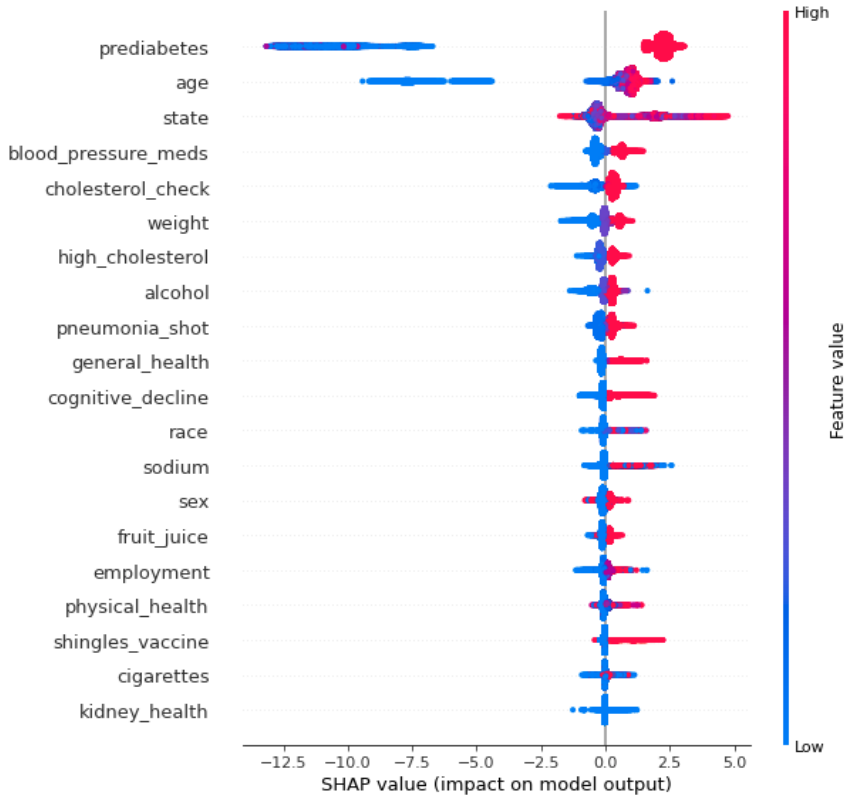
[5] Fayyad U, Piatetsky-Shapiro G, Smith P, Uthurusamy R. Advances in Knowledge Discovery and Data Mining. AAAI, MIT Press; MA: 1996. [Google Scholar]

[6] Type 2 Diabetes Diagnosis https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/diagnosis-treatment/drc-20351199

[7] Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer. 2008;8:37–49. [PMC free article]

Appendix A


JOINPAIN
LASTSMK2
POORHLTH
EMPLOY1
_DRNKWEK
_INCOMG
SEX
EXEROFT2
BPMEDS
_RFHLTH
TETANUS
ORNGDAY_
HIVTSTD3
CIMEMLOS
_RFCHOL
PHYSHLTH
GRENDAY_
DROCDY3_
BEANDAY_
PNEUVAC3
VEGEDA1_
PREDIAB1
_RACE
FTJUDA1_
HTIN4
CHOLCHK
FRUTDA1_
_BMI5CAT
_AGEG5YR
_STATE
0
100
200
300
400
500
600
700


prediabetes
age
state
blood_pressure_meds
cholesterol_check
weight
high_cholesterol
alcohol
pneumonia_shot
general_health
cognitive_decline
race
sodium
sex
fruit_juice
employment
physical_health
shingles_vaccine
cigarettes
kidney_health
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
SHAP value (impact on model output)
High
Feature value
Low