STA 325 Final Project Report

Emily Mittleman & Julia Rosner

2022-12-05

Introduction

A few paragraphs which (i) motivate problem importance & relevance (supported by relevant literature, if any), (ii) describe project goals and how such goals address the problem, as well as (iii) a high-level roadmap of the proposed methodology, and (iv) other relevant information for the reader. See project rubric for details.

Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood. There are different types of diabetes, but Type II diabetes mellitus is the most common. If left untreated, Type II diabetes can cause major health complications, including heart attack, kidney failure, stroke, and eye damage. In fact, Type II diabetes was the 7th leading cause of death in 2019; and unfortunately, its prevalence is rapidly increasing worldwide (CDC). According to the CDC, more than 37 million people in the United States have diabetes, and 1 in 5 people are unaware that they have it. Furthermore, approximately 96 million Americans (1 in 3) have prediabetes, and a shocking 80% of those Americans are unaware of their risk (CDC). Type II diabetes and prediabetes often begin as silent conditions, and so they often go undiagnosed for years with no clear symptoms, until serious health complications develop.

Although diabetes is an irreversible disease, it is largely preventable. The risk of developing diabetes can be reduced significantly through early detection of prediabetes and lifestyle interventions. While type 2 diabetes and prediabetes can be easily diagnosed through glucose blood testing, many people fail to test regularly. Therefore, evaluating diabetes risk through metrics other than glucose levels can prove to be extraordinarily valuable.

The prevalence of type II diabetes varies by age, education, income, other social determinants of health, and a variety of health indicators/factors regularly screened for at routine check ups. We aim to build an ML model to (1) identify individuals with diabetes or prediabetes, who could otherwise go undiagnosed, and (2) indicate individuals who show early signs (or are at risk) for prediabetes. Clinicians can then implement this model as a part of annual physical exams. If a clinician sees a result that indicates diabetes risk, then they can proceed with a glucose level test to determine whether or not there is a diagnosis. If there is a diagnosis they can proceed with the medical protocols/advice established. However, what separates our model, even if there is no diagnosis (for diabetes or prediabetes aka no extreme blood sugar levels yet), the model still indicated the patient was at risk, and so the patient can then be proactive in lowering their risk for pre diabetes / type II diabetes, and implement preventative measures.

Data

Our data was obtained from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), which is a health-related telephone survey collected annually by the CDC and gathers responses from over 400,000 Americans on health-related risk behaviors and chronic health conditions. For this project, a CSV of the dataset available on Kaggle was used. This original dataset contains responses from 441,455 individuals and has 330 features. The dataset originally had 330 features, but based on diabetes disease research regarding factors influencing diabetes disease and other chronic health conditions, only 21 select features are included in this analysis. After removing observations with missing values, we were left with 253,680 observations in our dataset.

The response variable is a binary indicator of whether someone does not have diabetes (0), or they do have diabetes or prediabetes (1).

There are 21 predictors: most are binary indicators, and some are discrete data such as age, BMI, health over the past month, etc. All of these predictors are noninvasive measurements commonly taken in medical settings, and can easily be collected by doctors at physical checkups to be able to run our predictive algorithm in order to determine diabetes risk.

Table 1: Predictor Descriptions

Predictor	**Description**	**Data Type**
High BP	Has high blood pressure	Binary
High cholesterol	Ever had high cholesterol	Binary
Cholesterol check	Cholesterol check within past five years	Binary
BMI	Body Mass Index	Discrete (1-98)
Smoker	Smoked at least 100 cigarettes in entire life	Binary
Stroke	Ever had a stroke	Binary
Heart disease	Ever had coronary heart disease	Binary
Physical activity	Exercised within the past 30 days	Binary
Fruits	Consume fruit 1 or more times per day	Binary
Vegetables	Consume vegetables 1 or more times per day	Binary
Heavy alcohol use	Men: >14 drinks weekly, Women: >7 drinks weekly	Binary
Any healthcare	Has any kind of health care coverage	Binary
No doctor (cost)	Needed doctor in past year but couldn't go due to cost	Binary
General health	Scale of 1-5	Discrete (1-5)
Mental health	Days of poor mental health in past 30 days	Discrete (1-30)
Physical health	Physical illness or injury in past 30 days	Discrete (1-30)
Difficulty walking	Difficulty walking or climbing stairs	Binary
Sex	Male or Female	Binary
Age	Which age group (18-24, 24-30,...)	Discrete (1-13)
Education	Highest level of education (None, elementary,...)	Discrete (1-6)
Income	Annual income bracket (<$10k, $10k-$15k,...)	Discrete (1-8)

This dataset is sufficient in meeting our project goals since it has a significantly large number of observations (253,680), and a large number of predictors that can all easily be measured noninvasively in clinical settings.

TODO: TALK ABOUT EDA

Methodology

Discussion & justification of model choice and features, and how the proposed model(s) fully addresses project goals. Any "downstream" uses of the model (e.g., for prediction, optimization, ranking) should be discussed in detail here. See project rubric for details.

Inference

Prediction

Results

Statistical analyses of the fitted model(s), and a translation of these findings into meaningful & understandable conclusions for the target audience (e.g., engineers, business managers, policy-makers, etc). See project rubric for details.

Inference Results

Prediction Results

Conclusion

A summary of key findings and potential impacts of your project.