Inferential and Predictive Models for Diabetes

(Project Abstract for MGT-256)

Monika Baloda, Zhuran Kang, Valeria Valenzuela, Yufei Zhu

October 17, 2022

Introduction, Research question and its importance—Changing lifestyle along with climate change is posing greater health challenges before us. Diabetes is one of such issue in our population these days. According to Global Disease Burden report [1], diabetes was responsible for 1.5 million deaths. As high as 48% of all deaths due to this disease occured before the age of 70. About 8.5% of adults (18+) reported to have diabetes. Therefore it becomes a question of vital importance to investigate what explains diabetes in a human-being. Further, model of predicting diabetes may be meaningful to allocate health resources to deal with this challenge. We, in this project, use regression techniques to explain the factors responsible for diabetes and then provide a framework for prediction.

Data– we use *Diabetes Health Indicators Dataset* available on Kaggle [2]. This dataset is originally based on Behavioral Risk Factor Surveillance System survey. This survey is done by Center of Disease Control (CDC), the United States of America. This is a telephone based survey where people are asked health related risk behaviours, chronic health conditions and use of preventive services. We have 22 total features of variables in this data along with more than 250000 observations. This richness of the dataset attracted us to choose this data. We will be able to apply the 'central limit theorem' with more confidence even in complicated regression models due to the large amount of observations. If required we can always randomly draw a sample of 1000 observations from this dataset.

Methodology– Our tentative project outlines is divided into four major tasks, one for each of the group member. The first task is to write introduction and perform exploratory data analysis (EDA). In this task we motivate the research question using standard existing references e.g. research articles from google scholar and WHO reports etc. EDA includes data visualization analysis– bar charts, scatter plots between two variables, and summary tables. We will offer some insights from these plots and graphs. Our second task is dedicated to inferential modeling i.e. to understand what set of factors are explains the probability of getting diabetes the most. We will employ the techniques learnt in the class such as multivariate linear regressions and logistic regressions. We will fit regression models, interpret the coefficients and discuss the limitations. The third task is to do a predictive modelling, in this section we use regression trees, we fit highly non-linear trees to predict diabetes. Our goal in this task is to get maximum predictive capabilities on test data. The final task is to aggregate all the results, summarize the outcomes, and tell a story based on the evidences found in the earlier tasks.

References

[1] Global Burden of Disease Collaborative Network, Global Burden of Disease Study (2019). *Institute for Health Metrics and Evaluation.*

[2] Diabetes Health Indicators Dataset (2016). *Behavioral Risk Factor Surveillance System*, Kaggle