Identifying Influential Variables in the Prediction of Type 2 Diabetes Using Machine Learning Methods

IDENTIFYING INFLUENTIAL VARIABLES IN THE PREDICTION OF TYPE 2 DIABETES USING MACHINE LEARNING METHODS

by

AMANDA CHERNISHKIN

B.B.S., GEORGIA COLLEGE & STATE UNIVERSITY

A Thesis Submitted to the Graduate Faculty
Of Georgia State University in Partial Fulfillment
of the
Requirements for the Degree

MASTER OF PULIC HEALTH

ATLANTA, GEORGIA

30303

TABLE OF CONTENTS

List of Tables

List of Figures

APPROVAL PAGE

IDENTIFYING INFLUENTIAL VARIABLES IN THE PREDICTION OF TYPE 2 DIABETES USING MACHINE LEARNING METHODS

by

AMANDA CHERNISHKIN

Approved:

____
Committee Chair

Dora Il'yasova
____
Committee Member

____
Committee Member

07/31/2020
Date

Author's Statement Page

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Georgia State University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote form, to copy form, or to publish this thesis may be granted by the author or, in his/her absence, by the professor under whose direction it was written, or in his/her absence, by the Associate Dean, School of Public Health. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involved potential financial gain will not be allowed without written permission of the author.

Amanda Chernishkin_______________
Signature of Author

Abstract

This study investigates three alternative machine learning methods to explore influential predictors of type 2 diabetes. It compares ridge, lasso, and elastic net regression to linear regression, and focuses on 12 outcome variables that include age, sex, race, income, education level, body mass index, waist circumference, arm circumference, hip circumference, family history, smoking status, sleep duration, high blood pressure, and high-density lipoprotein. Ridge, lasso and elastic net regression do not outperform linear regression but do assist in choosing a simpler model which could be important for improving future modeling.

Chapter 1

INTRODUCTION

1.1 Background

According to the World Health Organization's (WHO) annual report, the number of people with diabetes has nearly quadrupled since 1980 and is one of the leading causes of death around the world. In 2012, there were 1.5 million deaths worldwide directly caused by diabetes, and an additional 2.2 million deaths attributed to high blood glucose levels, including cardiovascular disease, chronic kidney disease, and tuberculosis. Diabetes is also the number one cause of blindness, amputation, and kidney failure. In 2014, approximately 422 million people over the age of 18 had diabetes. The substantial increase in cases can be seen in countries of all income levels and reflects the gradual rise of obesity levels seen around the world (World Health Organization, n.d.). Symptoms of diabetes progress slowly over a long period of time, and thus are commonly overlooked. According to the American Diabetes Association (ADA), 34.2 million Americans had diabetes in 2018, and of those, 7.3 million were undiagnosed. Left unchecked, diabetes can cause irreversible damage and become a great financial burden at an individual and national level. Economic costs of diabetes have increased by 26% from 2012 to 2017 (American Diabetes Association, 2018). In 2018, the ADA estimated the total costs of diagnosed diabetes had risen to $327 billion in 2017 from $245 billion in 2012. This includes $237 billion in direct medical costs and $90 billion in reduced productivity. People with diagnosed diabetes incur an average of $16,750 a year in medical bills, which has shown to be 2.3 times higher than in those without diabetes.

There are three main types of diabetes: Type 1, Type 2, and gestational diabetes. Most cases of diabetes fall under the umbrella of either type 1 or type 2 diabetes. Type 1 diabetes is when the body stops making its own insulin; this type cannot be prevented. In type 2 diabetes, the body can no longer effectively use insulin to maintain normal blood sugar levels. While preventable through a healthy lifestyle, type 2 diabetes accounts for 90-95% of all diagnosed cases of diabetes. In addition, 88 million Americans are pre-diabetic, and of those, 84% don't know they have it (Centers for Disease Control and Prevention, n.d.). Therefore, intervention of diabetes should focus on the preventative stages, and studies should devote attention to determining predictors of this detrimental disease.

There have been many studies in the past that have explored predictors of type 2 diabetes. Turi et al. found that blood pressure, sleep duration, and family history of diabetes were all significant predictors of type 2 diabetes (2017). In addition, a study completed in 2014 found that body mass index (BMI), older age, family history, and hypertension were associated with higher risk of type 2 diabetes (Foley et al.). These results correspond with risk factors established by the CDC which include overweight/obesity, ages 45 and up, family history of diabetes, and physically inactive (Center for Disease Control, n.d.).

Unfortunately, establishing causation is an arduous task. In attempts to explore what variables are influential to an outcome, most analyses resort to linear or logistic regression, but choosing what variables to include in a statistical model can be complicated and distort results. More and more researchers are now beginning to experiment with machine learning to get over this obstacle of variable selection. Lasso, ridge, and elastic net regression are three types of machine learning tool that have risen in popularity to produce predictive models.

1.2 Purpose of Study

The purpose of this study is to model glycohemoglobin levels to determine what variables are most influential, and compare three machine learning methods: ridge, lasso, and elastic net against the highly utilized linear regression.

Chapter 2

REVIEW OF LITERATURE

2.1 Diabetes

After consuming food, your blood sugar levels rise, and your pancreas responds by releasing a hormone known as insulin. Insulin is responsible for lowering the blood sugar by transporting blood glucose into your cells, where it is stored and later used for energy. Over time, depending on various precursors, the pancreas can no longer keep up with the high demand of insulin, and blood glucose remains at an elevated state. Once muscle, liver, and fat cells can no longer use the insulin, the body suffers from a condition described as insulin resistance (National Institute of Diabetes and Digestive and Kidney Diseases, 2016). Risks that put individuals in danger for developing insulin resistance include being overweight, physically inactive, a family history of diabetes, older age, high blood pressure, and high cholesterol (Centers for Disease Control and Prevention, n.d.). There are various ways to diagnose diabetes. The American Diabetes Association (ADA) advocates for three types of tests: A1C test, fasting blood sugar test (FPG), and glucose tolerance test (OGTT). **Table 1** shows the various criteria for diagnosing diabetes based on the ADA's recommendations. Due to the inconvenience of measuring FPG and OGTT, the A1C test, which measures what percentage of hemoglobin is glycated, has risen in popularity. The use of A1C for glycemia control has been intensely investigated in the past (Nathan, Kuenen, and Borg 2008) and is also recognized by the Centers for Disease Control and Prevention as a suitable diagnostic measure for diabetes.

Table 1. Diagnostics criteria for diabetes			
	A1C Test	Fasting Blood Sugar Test	Glucose Tolerance Test
Diabetes	6.5% or above	126 mg/dL or above	200 mg/dL or above
Prediabetes	5.7-6.4%	100-125 mg/dL	140-199 mg/dL
Normal	Below 5.7%	99 mg/dL	140 mg/dL or below

2.3 Predictors

Diabetes often does not occur in isolation. A vast majority of patients suffering from diabetes also have multiple comorbidities. In 2016, a retrospective study was done using the Quintiles Electronic Medical Record database. They found that 97.5% of patients had at least one comorbid condition in addition to diabetes, and 88.5% had at least two. Comorbidity burden tended to increase with age and was higher in men. The most common conditions associated with type 2 diabetes were hypertension (82.1%), obesity (78.2%), hyperlipidemia (77.2%), kidney disease (24.1%), and cardiovascular disease (21.6%) (Iglay, Hakima, Patrick, et al. 2016). Results such as these show how multifaceted chronic diseases can be, and why understanding the cause and treatment for disease such as diabetes can be extremely complex. Between September 2011 and June 2013, a cross-sectional survey was done by Hilawe et al (2016) targeting adults 25-64 years old in Palau. A sample of 2,216 non-pregnant adults participated in the survey, and 301 were dropped due to missing values (N=1915). The following measurements were taken following the World Health Organization standards. Participants were asked to fast for eight hours before capillary whole blood samples were taken from the fingertip to test fasting blood glucose (FBG) and lipid profile. FBG was classified into three categories: normal (FBG ≤ 5.6 mmol/L), prediabetes (FBG 5.6-6.9 mmol/L) and diabetes (FBG ≥ 7 mmol/L). Body mass index (BMI) was stratified into three groups: underweight/normal

(<18.5 or 18.5-24.9), overweight (25.0-29.9) and obese (≥30.0). Abdominal obesity was defined as a dichotomous variable having waist circumference (WC) of ≥94 cm for men and ≥80 cm for women, and large waist-hip ratio (WHR) was defined as having ≥0.90 for men or ≥85 for women. Blood pressure (BP) was split into three groups based on previous studies: normal, pre-hypertension, and hypertension. Age was considered categorical (25-29, 30-39, 40-49, 50-59, and 60-64). Chi-squared, analysis of variance, or nonparametric median tests were used to compare characteristics across FBG status. Odds ratios (OR) were estimated by multinomial logistic regression using normal FBG as reference. The study identified older age, overall obesity (BMI), central obesity (large WC or WHR), hypertension and hypertriglyceridemia as significant predictors of prediabetes and/or diabetes. Studies such as these showcase the intricacies in determining predictors of type 2 diabetes, and how countless variables can be related to an outcome. Machine learning could be one possible solution for minimizing this complexity.

2.4 Modeling

A large part of modeling with machine learning has to do with the idea of bias-variance tradeoff. Overfitted models have high variance and will do poorly when generalized to new data. In contrast, underfitted models have high bias, meaning the model may be missing relationships between X and Y. Thus, a more complex model may have less variance but increased bias, and vice versa. The aim is to get the model to generalize and classify new input accurately. Ridge, lasso, and elastic net regression can assist in finding a variance-bias balance by introducing a penalty parameter called lambda, represented by the λ symbol. Finding the best value for λ can be accomplished through cross-validation and can range from any value starting from zero to positive infinity. As λ increases, the slope of the line approaches zero, meaning it introduces more and more bias. The main difference between lasso and ridge, is that lasso can eliminate insignificant coefficients by shrinking them completely to zero, while ridge regression can only shrink close to zero. In other words, lasso can eliminate the effect certain variables have on the model. Elastic net regression is a combination of the two. All assist in variable selection.

In 2019, Farbahari, Dehesh, and Gozashti did a cross-sectional study using machine learning techniques to explore influential variables that affect fasting blood sugar (FBS). The study consisted of 270 type 2 diabetic patients over 18 years of age from Iran and was based on a study done in 1999 by Haffner, Alexander and Cook. The metabolic variables assumed to be affecting FBS included glycated hemoglobin (HbA1c), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), thyroid-stimulating hormone (TSH), creatinine (Cr), and carbamide (Urea). Characteristic variables included age, gender, smoking status, drug use, and heredity, which were all self-reported. BP and BMI were also included. Those with other chronic disease or pregnant were excluded from the study, resulting in a total of 650 participants all together. Lasso, ridge, and linear regression were all utilized and compared by the mean squared error (MSE). Lasso regression had the lowest MSE of all three models. Hba1c, age, BMI, gender, smoking status, and urea were found to have a significant association with FBS. It should also be noted that all three models jointly introduced HbA1c as the most effective predictor of FBS, and the authors go on to state that HbA1c could be used instead of FBS in order to diagnose type 2 diabetes. The evidence from this study confirm the usage of lasso regression for clinical research.

Another study done by Oh, Yoo, and Park (2013), utilized ridge, lasso, and elastic net regression to identify the risk of blindness due to diabetic retinopathy (DR). The health records from the Korea National Health and Nutrition Examination Surveys (KNHANES) V-1 were used. Out of the 8,958 participants involved in the KNHANES V-1 study, 556 were selected based on their diabetic status in accordance to the A1C test criteria, and 66 were excluded because they did not receive an eye examination resulting in N = 490. After constructing the models, Bayesian information criterion (BIC) was used for model selection. The area under the curve (AUC), accuracy, sensitivity, and specificity of the models were calculated using Receiver Operator Characteristic curve (ROC) which is assists in evaluates the perforce of a classification model. Cut-off points were selected that maximized Youden's index. Those above the cut-off point were classified as being at high risk. Lasso predicted DR most efficiently and found the presence of DR was associated with 19 predictors. FPG, TG, low BMI, and insulin therapy were all strong predictors. This study concluded that lasso can contribute to our understanding of risk factors for DR and supports that lasso can be an effective prediction model in the analysis of high-dimensional health records. Studies like the ones mentioned, show how complex establishing a cause of diabetes can be, and how utilizing machine learning techniques could be advantageous in the explorations of predictors.

Chapter 3

METHODS

3.1 Data Source and Preparation

The National Health and Nutrition Examination Survey (NHANES) years 2017-2018 was used for this study. NHANES is a research program that began in the 1960's by the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC). NHANES is designed to evaluate the health of the United States population through physical examination and interviews. Health interviews are conducted in respondents' homes and consist of socio-economic, demographic, dietary, and health-related questions. Health measurements/ laboratory tests are performed by specialists in mobile centers, and consist of medical, dental, physical, and physiological measurements. All participants, excluding the very young, participate in having blood samples taken. To increase reliability of statistical estimates, NHANES over-samples persons 60 and older, African Americans, and Hispanics (*NHANES – About the National Health and Nutrition Examination Survey, 2018*).

NHANES 2017-2018 consisted of 9,254 participants. After the sample was limited to those who were 21 years and older and who have completed the glycohemoglobin blood test, the sample size dropped to 5,193. Variables of interest included demographic variables such as age, race, sex, education level, and income. Other variables included were body mass index (BMI), waist circumference, arm circumference, hip circumference, family history, smoking status, sleeping duration, high blood pressure, and high-density lipoprotein. The dependent variable, glycohemoglobin levels, was kept as continuous. Smoking status, family history of diabetes, sleep duration, and a report of high blood pressure were taken from the questionnaire branch of NHANES. Smoking status was defined as those who said yes to smoking at least 100 cigarettes in their life and said they still smoked cigarettes now. High-density lipoprotein and glycohemoglobin levels were taken from the laboratory data.

Of the 5,193 observations, 243 were missing. To handle missingness for the independent variables, multivariate imputation by chained equations (MICE) was used. **Table 2** shows the juxtaposition between before and after imputation. As shown, the distribution of variables remained nearly identical. The sample was roughly 50% female, with an average age of 52 years. The average glycohemoglobin level was 5.9 with a standard deviation of 1.1. Based on the American Diabetes Association (ADA) criteria in shown in **Table 1**, nearly 16% of the sample was diabetic, 29% prediabetic, and 51% falling within the healthy range.

To conduct the machine learning methods, the data was split into testing and training sets. One third of the data was placed in a testing set (n=1,683), and two thirds was placed in the training set (n=3,267). The training dataset was used to train the model, while the testing dataset was used to see how well the model performed. All data cleaning for this study was done in SAS and analyses in RStudio.

3.2 Analysis

As stated previously, ridge, lasso, and elastic net all incorporate a tuning parameter lambda (λ), which is chosen through cross-validation. The tuning parameter, λ , determines how much shrinkage will occur, and thus regulates the variable selection. Therefore, a very small tuning parameter, where little to no shrinkage is taking place (λ=0), results in a regular linear regression model. For this study, 10-fold cross-validation was used to choose the best lambda value for each method, excluding linear regression where lambda was set to zero. To complete cross-validation, data is split into ten subsets, also known as folds, and then a model is trained on all subsets excluding one. The subset left out is used to evaluate how well the model performed. This procedure is repeated k times, while a different subset is reserved for testing each time. This entire process can be achieved through the cv.glmnet function in R. **Figure 1** displays the formula for how this function is performed. When alpha (α) is set to 0, the lasso penalty equals 0, and the equation is reduced to ridge regression. When α is set to 1, the ridge penalty is set to 0, and the equation is reduced to only lasso regression. When α equals any number between 0 and 1, (aka. elastic net), both penalties are incorporated. **Table 3** shows the best lambda value chosen from each cross-validation procedure.

Figure 1. Detail on the function cv.glmnet

$$\lambda \text{ x } \overset{\text{Lasso penalty}}{\boxed{[\alpha \text{ x } (Coefficient_1 + \ldots + Coefficient_{\mathrm{x}})}} + \overset{\text{Ridge penalty}}{\boxed{(1- \alpha) \text{ x } (Coefficient_1^2 + \ldots + Coefficient_x^2)]}}$$

Table 2. Statistics before and after imputation

Variable	Before imputation	After imputation
Age, mean (SE)	51.6 (17.7)	51.6 (17.7)
Gender, frequency (%)		
Male	48.1%	48.1%
Female	51.9%	51.9%
Race, frequency (%)		
Mexican American	13.6%	13.6%
Other Hispanic	9.6%	9.6%
Non-Hispanic White	34.8%	34.8%
Non-Hispanic Black	22.9%	22.9%
Non-Hispanic Asian	14.1%	14.1%
Other	5.0%	5.0%
Education, frequency (%)		
Less than 9th grade	8.6%	8.6%
9th-11th grade & 12th with no diploma	11.4%	11.4%
High school graduate. GED	23.9%	23.9%
Some college	32.4%	32.4%
College graduate or above	24.7%	23.8%
Income, frequency (%)		
High	19.2%	19.3%
Middle	34.4%	34.1%
Low	46.3%	46.6%
Body mass index, frequency (%)		
Underweight	3.0%	3.0%
Normal	23.7%	23.7%
Overweight	31.8%	31.8%
Obese	41.5%	41.5%
Waist circumference, mean (SE)	100.8 (17.0)	100.6 (17.1)
Arm circumference, mean (SE)	33.4 (5.3)	33.3 (5.3)
Hip circumference, mean (SE)	107.1 (14.6)	106.9 (14.7)
Family history, frequency (%)		
Yes	23.0%	23.4%
No	77.0%	76.6%
Smoking status, frequency (%)		
Yes	18.0%	18.0%
No	82.0%	82.0%
Sleeping duration, mean (SD)	7.59 (1.6)	7.59 (1.6)
High blood pressure, frequency (%)		
Yes	38.8%	38.7%
No	61.2%	61.3%
High-density lipoprotein, mean (SD)	1.4 (0.4)	1.4 (0.4)
Glycohemoglobin level, mean (SD)	5.9 (1.1)	N/A

Chapter 4

RESULTS

4.1 Model Comparison

After splitting the data into training and test sets, choosing the best lambda values with cross validation, and fitting the model with the selected optimal tuning parameter, the mean squared prediction error (MSPE) was calculated to compare each model. Results are shown in **Table 3**. As discussed previously in this paper, ridge regression does not have the ability to shrink a coefficient to 0, and thus the model includes all 12 variables. Lasso only chose two variables: age and waist circumference. Elastic net regression chose 6 variables: age, non-Hispanic White, waist circumference, family history, high blood pression, and high-density lipoprotein. The selected variables align with previously done studies. Linear regression, also incapable of shrinking, includes all 12 variables. All four models produced a similar MSPE value ranging from the lowest at 0.95 for linear regression and the highest at 0.99 for elastic net regression. These results show that all four models performed similarly, and one could choose between either.

Table 3. Comparison of coefficients between ridge, lasso, and elastic net regression

Parameters	Ridge regression	Lasso regression	Elastic net regression	Linear regression
Intercept	5.27	4.69	5.01	5.38
Age	0.01	0.01	0.01	0.02
Gender				
Female	-0.01			0.16***
Race				
Other Hispanic	0.05			0.05
Non-Hispanic White	-0.11		-0.09	-0.22***
Non-Hispanic Black	0.03			0.09
Non-Hispanic Asian	0.06			0.12
Other	-0.02			-0.03
Education				
9-11th grade	0.01			-0.24**
High school graduate/GED	-0.03			-0.26**
Some college	-0.04			-.027***
College graduate or above	-0.06			-0.34***
Income				
Middle class	-0.02			-0.01
Upper Class	-0.02			0.04
BMI				
Normal	-0.05			-0.08
Overweight	-0.01			-0.17
Obese	0.04			-0.16
Waist circumference	0.01	0.01	0.01	0.02***
Arm circumference	0.00			0.00
Hip circumference	0.00			-0.01***
Family history				
Yes	0.11		0.03	0.20***
Smoking status				
Yes	-0.05			-0.05
Sleep duration	-0.00			-0.01
High blood pressure				
Yes	-0.20		0.11	0.15***
High-density lipoprotein	-0.20		-0.23	-0.40***
Lambda	1.25	0.10	0.18	0
Mean squared error	0.98	0.95	0.99	0.95

Chapter 5

DISCUSSION

5.1 Discussion of Results

Ridge, lasso, and elastic net regression are shrinking techniques that have the potential to limit model complexity. Especially in scenarios that include a myriad of variables, methods such as linear regression may result in an over-fitted and overly complex model. Thus, a major benefit from utilizing shrinkage techniques is the ability to choose a simpler model by selecting parameters based on prediction versus linear regression which just fits based on the given data. In this scenario, the machine learning methods did not outperform linear regression, but nevertheless revealed a model containing only 2 variables instead of all 12.

5.2 Limitations and Future Directions

One limitation of this study was that NHANES survey weights were not used. This may have introduced bias into the results of the analysis and interfere with the accuracy of the conclusion. Future studies should consider using the survey weights if they wish to have more precise outcomes that better describe the population.

5.3 Conclusion

This study's purpose was to compare three machine learning techniques to linear regression in determining possible predictors of high blood sugar levels. Results showed that ridge, lasso, and elastic net performed no better than linear regression, but demonstrated the parameter selection capability of said shrinkage methods. Most importantly, these results show a different approach for tackling the intricacies of predictive modeling.

REFERENCES

ADA. *Diabetes Overview* | American Diabetes Association. https://www.diabetes.org/diabetes

CDC. (2020, June 11). *What is Diabetes* | Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html

Farbahari A., Dehesh T., & Gozashti M. H. (2019). The Usage of Lasso, Ridge, and Linear Regression to Explore the most Influential Metabolic Variables that Affect Fasting Blood Sugar In Type 2 Diabetes Patients. *Romanian Journal of Diabetes Nutrition and Metabolic Disease*, 26(4), 371-379. https://doi.org/10.2478/rjdnmd-2019-0040

Foley D. L., Mackinnon A., Morgan V. A., Watts G. F., McGrath J. J., Castle D. J., Waterreus A., & Galletly C. A. (2014) Predictors of Type 2 Diabetes in a Nationally Representative Sample of Adults with Psychosis. *World Psychiatry*, 13(2), 176-183. https://doi.org/10.1002/wps.20130

Hiawe E. H., Chiang C., Yatsuya H., Chaochen W., Ikerdeu E., Honjo K., Mita T., Cui R., Hirakawa Y., Madraisau S., Ngirmang G., Iso H., & Aoyama A. (2016). Prevalence and Predictors of Prediabetes and Diabetes Among Adults in Palau: Population-based National STEPS Survey. *Nagoya Journal of Medical Science,* 78(4), 475-483. 10.18999/nagjms.78.4.475

Iglay K., Hannachi H., Howie P. J., Xu J., Li X., Engel S. S., Moore L. M., & Rajpathak S. (2016). Prevalence and Co-prevalence of Comorbidities Among Patients with Type 2 Diabetes Mellitus. *Current Medical Research and Opinion*, 32(7),1243-1252. https://doi.org/10.1185/03007995.2016.1168291

Nathan D. M., Kuenen J., Borg R., Zheng H., Schoenfeld D., & Heine R. J. (2008). Translating the A1C Assay into Estimated Average Glucose Values. *Diabetes Care*, 31(8), 1473-1478. 10.2337/dc08-0545

Oh E., Yoo T.K., Park E., (2013). Diabetic retinopathy risk prediction for fundus examination using sparse learning: a cross-sectional study*. BMC Medical Informatics and Decision Making* 13, 106. https://doi.org/10.1186/1472-6947-13-106

Turi K. N., Buchner D. M., & Grigsby-Toussaint D. S. (2017). Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors. *Preventing Chronic Diseases.* 14. http://dx.doi.org/10.5888/pcd14.160244external

WHO. Global Report on Diabetes | World Health Organization. file:///C:/Users/User/Downloads/9789241565257_eng%20(3).pdf