Best Predicitive Model for Diabetes*

Avijit Mallik (UT EID - AM99484)† Arindam Chatterjee (UT EID - AC83995)‡
Takehiro Hasimoto (UT EID - TH33985)§

Abstract

Diabetes is a prevalent chronic disease in the US, affecting millions of people and placing a significant financial burden on the economy. In this paper, we executed four analyses, (1) building the best predictive model for diabetes, (2) constructing the risk score for diabetes, (3) identifying the key factors that affect diabetes, and (4) finding other important factors that we can address in future. From our analysis, we found that the linear probability model with 92 variables (21 variables and 71 cross term variables) was the best predictive model of diabetes among the models we had considered. From the second analysis, we developed a risk score model, which can be used to identify the relative risk of developing diabetes. From the third analysis, we found that BMI, Age, General Health condition, High BP, Physical & Mental Health, presence of heart disease are the most critical factors that contribute to this disease. Physical activities, diets rich in fruits & vegetables etc. on the other hand help people reduce the risk of diabetes. Interestingly we found that poor & less educated people are more likely to develop diabetes. We finally found that the key words that are related to diabetes in the medical journal papers are fasting, metabolic, blood glucose level, cholesterol level, stress, inflammation and pregnancy. For future research, we can collect these data to construct a better predictive diabetes model.

1. Introduction

1-1. Background

Diabetes is a prevalent chronic disease in the US. It is characterized by the inability to regulate glucose levels in the blood due to insufficient insulin production or ineffective use of insulin. High blood sugar levels can lead to complications such as heart disease, vision loss, lower-limb amputation, and kidney disease. While there is no cure for diabetes, lifestyle changes and medical treatments can help mitigate its harms. Early diagnosis is important, and predictive models for diabetes risk can aid public health officials. Type II diabetes is the most common form, and its prevalence varies by social determinants of health such as age, education, income, race, and location. Diabetes also has a disproportionate impact on those of lower socioeconomic status. The economic burden of diabetes is significant, with costs exceeding $400 billion annually.

Here are some statistics on diabetes in the USA:

- As of 2021, approximately 34.2 million Americans, or 10.5% of the population, have diabetes.
- About 90-95% of cases are type II diabetes.
- Another 88 million American adults, or 34.5% of the population, have prediabetes.
- 1 in 5 people with diabetes, and 8 in 10 people with prediabetes, are unaware of their condition.
- Diabetes is the seventh leading cause of death in the United States.

*We would like to express our gratitude to Professor James Scott and Rui Zuo, University Texas Austin. Their expertise, guidance, and support were invaluable in shaping the research and ensuring its successful completion. Any errors or omissions are the sole responsibility of ours.

†avijit@iba-du.edu
‡arindamchatterjee@utexas.edu
§take.hashimoto0527@utexas.edu

- In 2017, the total cost of diagnosed diabetes in the United States was estimated to be $327 billion.

So given these statistics, we know that it is important to identify the key factors that lead to diabetes. If we can develop a predictive model that gives an idea how we can predict the chance of developing diabetes, we can take necessary actions to prevent diabetes. The finding will be quite useful for patients, healthcare providers or physicians as this disease is not curable. So people can take necessary actions accordingly to prevent this chronic disease.

1-2. Our Objectives and findings

The objectives of this paper are as follow:

- Can we develop a predictive model of diabetes?
- Can we assign a risk score for an individual given we know his/her individual biological & demographic characteristics?
- What risk factors are most predictive of diabetes risk?

Our findings will be useful for the following stakeholders:

- Patients: Diabetes models can help patients understand their risk of developing the disease, make informed lifestyle choices, and take steps to prevent or manage the disease.
- Healthcare providers: Predictive models can help healthcare providers identify patients who are at high risk of developing diabetes, allowing for earlier interventions and better disease management.
- Public health officials: Modeling diabetes can help public health officials understand the patterns and trends of the disease, identify populations at high risk, and develop targeted prevention and treatment strategies.
- Researchers: By analyzing data from diabetes models, researchers can gain insights into the underlying causes of the disease, identify new risk factors, and develop more effective treatments.

2. Methods: Data and Model

2-1. Data

Nature of Data: Dataset includes 70,692 observations of US individuals. 50% of them had diabetes & the rest did not. These data were collected by BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM in assistance with Chronic Disease Center. There are total 21 variables. There are a total 21 variables. The variables are defined as follows:

Diabetes_binary: A binary variable indicating presence of diabetes by 1 & absence by 0.

HighBP,HighChol,Smoker(Have you smoked at least 100 cigarettes in your entire life? 0 = no 1 = yes), **Stroke, HeartDisease PhysActiv, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk** (Any difficulty on walking, 1 indicate presence of difficulty) - all are binary variables where 1 indicate presence of the factor and 0 indicates absence. Cholcheck is also a binary variable where 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years

BMI: Body Mass index is a measure of body fat based on height and weight that applies to adult men and women. Here, BMI had a minimum value of 12 & maximum value of 98. BMI more than 25 is considered obese.

General Health: A measure of general health situation in which 1=Excellent, 2=Very Good, 3=Good, 4=Fair, 5=Poor.

Mental Health: During the past 30 days, for about how many days did poor physical or mental health keep an individual from doing your usual activities, such as self-care, work, or recreation? A value from 0 to 30 where 0 indicates best mental health & 30 indicates worst.

Physical Health: for how many days during the past 30 days was the individual's physical health not good? Responses range from 0 to 30 where higher the value worse the physical health.

Demographic Variables:

Sex: patient's gender (1: male; 0: female).

Age: 13-level age category ,where 1 = 18-24, 2=25-29, 3= 30-34, 4=35-39, 5=40-44, 6=45-49, 7=50-54, 8=55-59, 9=60-64, 10=65-69, 11=70-74, 12 = 75-79 & 13 = 80 or older etc.

Education: A value that ranges from 1 to 6. Higher value indicates higher level of education. 1= Never attended school or only kindergarten, 2 = Grades 1 through 8 (Elementary) 3 =Grades 9 through 11 (Some high school), 4= Grade 12 or GED (High school graduate), 5= College 1 year to 3 years (Some college or technical school), 6= College 4 years or more (College graduate)

Income: Indicates level of annual income for the households where 1= income less than $10000, 2= $ 10,000 to less than $15,000, 3= $15,000 to less than $20,000, 4= $20,000 to less than $25,000, 5= $25,000 to less than $35,000, 6= $35,000 to less than $50,000, 7= $50,000 to less than $75,000, 8= $75,000 or more

please see the Appendix 5-1 if you want to see other specific characteristics of the data.

2-2. Model

2-2-1. Best Predictive Model

Using different predictive models such as linear regression, KNN regression & logistic regression, we will develop a model which is the best for predicting diabetes.

2-2-2. Scoring

Based on the best predictive model, we will develop a risk score & based on the risk score we can give an early signal about the degree of risk exposure a person may have.

We have followed the following steps:

1. Use the estimated coefficients

we used the estimated coefficients the detail (see the Appendix 5-2) to estimate the probability of diabetes for individuals with a linear probability model and the logit model.

$$\hat{y} = \beta_0 + \beta_1(each\ variables) + \beta_2(cross\ temrs(only\ LPM))$$

2. 100 scaled (for only the Linear Probability model)

To show $\hat{y}$ as a score in the linear probability model, we scaled it into the format 0-100 points.

$$score = 100\hat{y}$$
$$where\ if\ \hat{y} > 100,\ then\ score = 100$$
$$if \hat{y} < 0,\ then\ score = 0$$

3. Risk score:

We defined and classified the risk score for individuals: 0-25="Low Risk", 26-50="Moderate Risk", 51-75= "High Risk", 76-100="Very High Risk"

2-2-3. Identifying critical factors responsible for diabetes

We have used correlation matrix, PCA, partial dependence plot etc to identify the factors responsible for diabetes.

2-2-4. Factors related to "diabetes"

We used the Natural Language Processing to identify factors that are much related to "diabetes". This is because if we can get additionally data of them, we can estimate our model more preciously.

To do that, at first, we collected data of the abstract of the recent academic paper with the keyword "diabetes" from the Pubmed. Specifically, we used the application of Publish or Perish 8, put "diabetes" into the "Keywords" and get information of the 1000 medical papers with "diabetes" keyword from 2022 to 2023.

Next, we excluded blank data and stopwords of "stopwords" library, and tokenized their abstracts using "tokenizers" library. So there were 950 papers.

After that, we applied "apriori" function for data that we got with support =.02, confidence=.001, and extract some data from them with lift>4. Then we showed the node graph that told us words that related to the word "diabetes" with Gephi.

Finally, from the node graph, we got factors that are related to "diabetes", and identified the additional factors while comparing them to our current factors of data.

Note: We did not use "Google Scholar" to collect in Publish or Perish as it did not give us all the abstract data of each paper. Also, Publish or Perish can only get up to 1000 data.

3. Results

3-1. Best Prediction Model for Diabetes

3-1-1. Linear model regression and KNN regression

First, to find the best combination of variables for the model, we used stepwise selection. We got the best model based on AIC criteria, which is so complex that we have shown it in the Appendix 5-2.

We also compared three models with the cross-Validation namely - the stepwise linear model, the basic model (which simply includes all variables), and the KNN regression model. Rmse of three models are as given below:

Table 1: Comparison of LPM and KNN model

RMSE_stndard	RMSE_step	RMSE_KNN
0.415	0.411	0.5

where the optimal k is 2. From the above result, it is safe to say the stepwise model is the best predictive model so far.

3-1-2. Logit model comparison

To find better predictive model, we compared the LPM & Logit models:

Table 2: Comparison of LPM and Logit

	Linear	Logit
accuracy	0.753	0.748
TPR	0.790	0.768
FPR	0.282	0.272
FDR	0.263	0.253

From these tables, the accuracy rate of the linear model is 0.753(=75.3%) and that of the logit model is 0.748(=74.8%). Also, each true positive rate (TPR) is 0.79 and 0.768, each false positive rate (FPR) is 0.282 (=Specificity: 70.8%) and 0.272 (=Specificity: 72.8%), and each false discovery rate (FDR) is 0.263 (=Precision: 73.7%) and 0.253 (=Precision: 74.7%).

Besides, The ROC curves shows that the linear model is slightly better than that the logit model (see the Appendix 5-3).

From the result, we see that the linear model is better than the logit model because the linear model has the lowest RMSE and the greater ability to predict diabetes. However, we need to be careful because sometimes the logit model might be better.

3-2. Scoring

3-2-1. What the crucial variables are to directly affect on the risk score?

From the linear probability model, we retrieve the coefficients as the weights of the risk score. Based on the coefficients, the direct marginal effects on the risk score are as follow:

Table 3: Coefficients(Weights) of LPM and Logit model

	linear	logit
HighBP	7.14e-02	0.365000
HighChol	5.28e-02	0.293000
CholCheck	3.46e-02	0.211000
BMI	8.71e-02	0.538000
Smoker	-6.22e+00	-0.000839
Stroke	1.67e-02	0.039100
HeartDiseaseorAttack	3.81e-02	0.089700
PhysActivity	-6.46e+00	-0.015100
Fruits	-2.30e-03	-0.016800
Veggies	-4.15e-03	-0.024900
HvyAlcoholConsump	-4.91e+00	-0.152000
AnyHealthcare	-3.62e-05	0.012600
NoDocbcCost	-8.20e-04	0.005540
GenHlth	1.14e-01	0.651000
MentHlth	-6.83e-03	-0.035600
PhysHlth	-1.31e-03	-0.083700
DiffWalk	2.77e-02	0.050000
Sex	2.14e-02	0.133000
Age	4.12e+00	0.434000
Education	-7.62e-03	-0.037900
Income	-2.52e-02	-0.128000

Here, these coefficients are the result of the estimation with scaled data (if you want to see all coefficients, see the Appendix 5-2). Among the binary variables, the highest weight is assigned to "HighBP"(high blood pressure) and the lowest one is assigned to "HvyAlcoholConsump" (heavy alcohol consumption). Outside the dummy variables, BMI seems to largely affect the risk score.

In the logit model, among the binary variables, "GenHlth"(general health) has the highest weight/coefficient and "HvyAlcoholConsump" (heavy alcohol consumption) has the lowest weight/coefficient. In addition to that, BMI seems to affect the risk score largely in logit model as well. We also note that the coefficient of HighBP is still positive and large. In both models, Physical activities, diets rich in fruits & vegetables have negative weights as these variables reduce the risk of diabetes.

From the above result, the crucial variables to increase the risk score of diabetes from the viewpoint of the linear model and logit model is **High blood pressure" and "General health"**. And, the variable that does not relate to diabetes or make risk score decrease is **"Heavy alcohol consumption"**. Besides, the risk score increases as someone's **"BMI"** increases.

Note that we shows the ratio of the number of people by each scoring thresholds in the Appendix 5-4.

3-2-2. Which risk socre model is better between LPM and Logit?

Next we will discuss, between LPM and Logit , which risk score model is better . Here, we have shown the distribution of the predicted risk scores from the actual data :

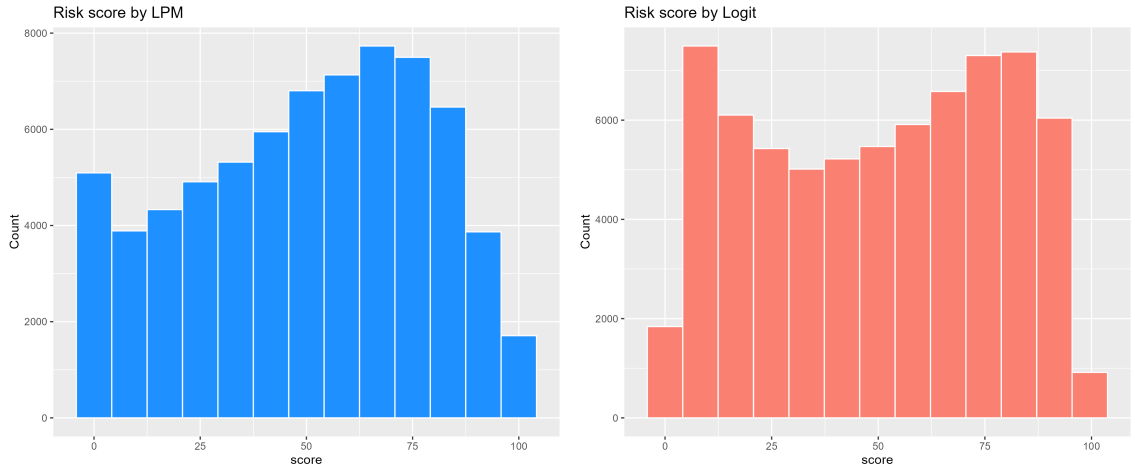


Figure 1: Distribution of scores by the linear and the logit

The correlation between predicted output in the linear model and the logit model is 0.975. From the graphs, we observe that the linear model roughly follows a normal distribution while top scores are slightly tilted to the right. Also, the logit model is a distribution with a dent in the middle. The correlation of the scores of these two predictive models is 0.975.

From these results we observe that the predicted scores under the two models are almost the same. So, we can focus on the shape of the distribution for scores. We think that a scoring distribution will be desirable to be normal because if there is the true score for diabetes, the distributed independent samples will be closer to from the Central Limit Theorem as the sample will be larger. Also, if we assume this distribution as a normal, we can easily use this score for other purposes.

We have already found that the linear probability model is a better scoring model than logit.

3-3. Identifying factors affecting Diabetes:

To identify the critical factors causing Diabetes, we used a correlation matrix, PCA analysis & variable importance plot (using a random forest model). The correlation matrix shows that there is a high positive correlation between **Diabetes & High BP, High cholesterol, age, Heart disease/attack, & stroke**.

We have also identified 3 latent groups of variables:

The most important factors according to the PCA analysis are General Health, HighBP, HeartDiseaseorAttack, HighChol , BMI, Mental Health & Age.

Since the sample size is large, we ran 2 random subsample to evaluate the random forest model & identify the most important variables using the 'varImp plot' (see the Appendix 5-5. The 2 different random forests gave us roughly similar important variables affecting diabetes. The variables based on importance are **BMI,**

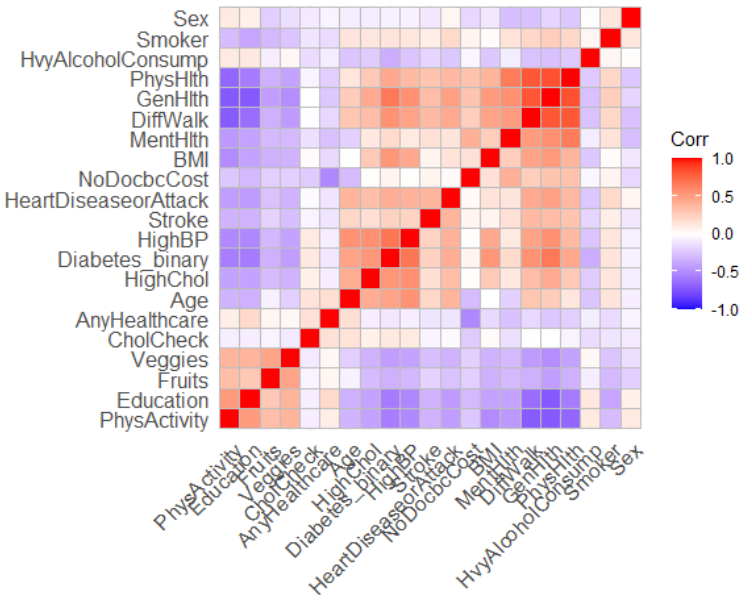


Figure 2: Correlation matrix in data

High BP, Age, General Health, Income, Education, Physical Activity, Mental Health, and High Cholesterol.

So, Using the correlation matrix, PCA analysis & the variable importance plot, we find that the most common factors affecting diabetes are: **BMI, Age, General Health, High BP, Income level, Education level,High cholesterol & Heart Disease.**

In order to identify how these variables might affect the probability of developing diabetes, we ran a gradient boost model & found some interesting insights. For instance, the **partial dependence plot** (see the Appendix 5-5) shows that as the **education level** goes up, the probability of diabetes goes down. Probably educated people live a more disciplined life & can take better care of their diet and health. In fact, diabetes is prevalent among less educated people. The probability of diabetes goes up after **age** 55. The plot also suggests that A person with **high cholesterol** will have a 20% greater chance of developing diabetes. The higher the BMI, the higher the probability of catching diabetes. The prevalence of diabetes also goes down as the **level of income** goes up. As people earn a higher level of income, they can spend more on healthcare, can eat a healthy diet & live a better lifestyle. It is also believed that diabetes is a disease closely related to one's lifestyle. On the other hand, the probability of occurrence of diabetes goes up as a person rates his/her **general health** fair or poor.

3-4. What kind of variables is more helpful to refine our model?

At first, as mentioned in the Model part, we used apriori function in R with support 0.02 and confidence 0.01 (see the Appendix 5-6) to look at the support-confidence. And with lift >4, we created data to show the node graph. the following graph is enlarged it around the word "diabetes":

From this graph, we can get some factors which are not included in our data are related to diabetes like these:

- fasting
- metabolic
- blood glucose level
- cholesterol level

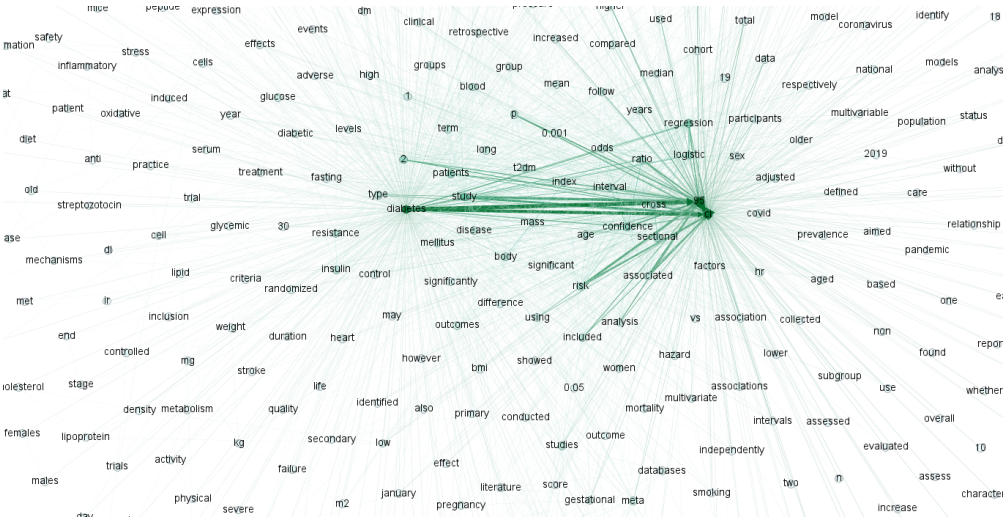


Figure 3: Nodes and words around the diabetes word

- stress
- inflammatory
- pregnancy

Therefore, if we want to refine our model in future, it is better to include these additional factors.

4. Conclusion

In this paper, we wanted to build a predictive model of diabetes. We used stepwise selection to find the best linear probability model. From the result, we got **the best model with 92 variables (21 variables and 71 cross-term variables)**, excluding the intercept terms, Also, we compared it to its KNN version's model with Cross Validation. Then, we found that this linear model is better than its KNN model from the view of RMSE.

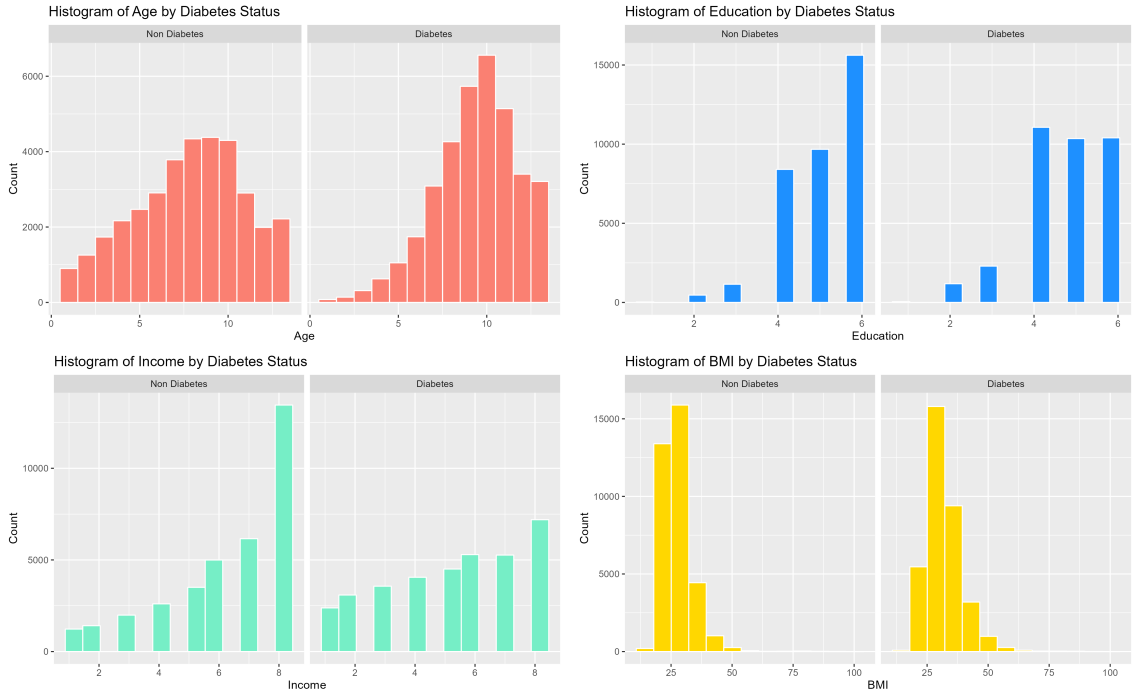
Then we used the linear probability model and the logit model with all variables data to construct the risk score for diabetes. The weight of the risk score is the coefficient of the linear probability model. Based on the risk score, we could identify whether someone has a low, moderate, high, or very high risk of developing diabetes. In this paper, we found that **BMI, Age, General Health condition, High BP, Physical & Mental Health, presence of heart disease are the most critical factors that affect diabetes**. Diabetes risk increases with age, high Blood Pressure & presence of heart disease. **People with poor mental, physical & general health are more likely to develop the risk of diabetes**. Finally, **Poor & less educated people have a greater risk of suffering from diabetes**.

Finally, we used the abstracts of 1000 medical academic papers from Pubmed to find other factors to sophisticate our model. In this analysis, with Natural Language Processing, we applied apriori function in R and created the node graph. The words related to "diabetes", which are most cited in academic papers, are 'fasting', 'metabolic', 'blood glucose level', 'cholesterol level', 'stress', 'inflammatory' and 'pregnancy'. We believe that in the future, we can collect these data to construct a better predictive diabetes model.

5. Appendix

5-1. Charactersitcs of Data

In the following graphs, we show the histograms of population distribution with & without diabetes based on age, income, education & BMI.



5-2. Stepwise selection model

The model that we got from the stepwise selection in 3-1 is as follow:

$$
\begin{aligned}
Diabetes_binary = \beta_0 + \beta[&HighBP + HighChol + CholCheck \\
&+ BMI + Smoker + Stroke + HeartDiseaseorAttack + PhysActivity + Fruits + Veggies \\
&+ HvyAlcoholConsump + AnyHealthcare + NoDocbcCost + GenHlth \\
&+ MentHlth + PhysHlth + DiffWalk + Sex + Age + Education \\
&+ Income + GenHlth \cdot DiffWalk + BMI \cdot Age + Sex \cdot Age + HighChol \cdot Age \\
&+ GenHlth \cdot Income + BMI \cdot DiffWalk + HighBP \cdot HeartDiseaseorAttack \\
&+ DiffWalk \cdot Age + GenHlth \cdot PhysHlth + HvyAlcoholConsump \cdot Age \\
&+ CholCheck \cdot GenHlth + GenHlth \cdot Sex + HighChol \cdot HeartDiseaseorAttack \\
&+ HighChol \cdot GenHlth + HeartDiseaseorAttack \cdot Age + Smoker \cdot GenHlth \\
&+ HeartDiseaseorAttack \cdot GenHlth + Fruits \cdot Education \\
&+ HeartDiseaseorAttack \cdot DiffWalk + HighBP \cdot HvyAlcoholConsump \\
&+ HighChol \cdot Stroke + AnyHealthcare \cdot Education + HighBP \cdot CholCheck \\
&+ Smoker \cdot Education + BMI \cdot HeartDiseaseorAttack + BMI \cdot Smoker \\
&+ MentHlth \cdot DiffWalk + CholCheck \cdot BMI + HighChol \cdot BMI \\
&+ HighBP \cdot HighChol + Fruits \cdot Age + PhysActivity \cdot Fruits \\
&+ Stroke \cdot GenHlth + HighBP \cdot Sex + Stroke \cdot Age \\
&+ HeartDiseaseorAttack \cdot NoDocbcCost + CholCheck \cdot Age \\
&+ CholCheck \cdot HeartDiseaseorAttack + Fruits \cdot Sex \\
&+ HvyAlcoholConsump \cdot GenHlth + HighChol \cdot MentHlth \\
&+ HighBP \cdot AnyHealthcare + HighBP \cdot Education + Smoker \cdot Stroke \\
&+ PhysActivity \cdot Age + PhysActivity \cdot Education + CholCheck \cdot AnyHealthcare \\
&+ Veggies \cdot NoDocbcCost + BMI \cdot Sex + MentHlth \cdot Income \\
&+ PhysHlth \cdot Income + NoDocbcCost \cdot DiffWalk + BMI \cdot MentHlth \\
&+ HighChol \cdot Sex + Sex \cdot Education + Smoker \cdot Sex + PhysHlth \cdot Age \\
&+ MentHlth \cdot Age \\
&+ HeartDiseaseorAttack \cdot PhysHlth \\
&+ PhysActivity \cdot PhysHlth + Stroke \cdot MentHlth + BMI \cdot NoDocbcCost \\
&+ AnyHealthcare \cdot Age + BMI \cdot PhysActivity + Smoker \cdot DiffWalk + \\
&+ Smoker \cdot Age + Smoker \cdot HvyAlcoholConsump \\
&+ AnyHealthcare \cdot DiffWalk + \\
&+ AnyHealthcare \cdot PhysHlth + Stroke \cdot HeartDiseaseorAttack \\
&+ PhysActivity \cdot Income] + \varepsilon
\end{aligned}
$$

The summary of the regression of the linear probability model is as follow:

Residuals:	Min	1Q	Median	3Q	Max
	-1.75341	-0.32024	0.05917	0.31446	1.22353
Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.40E-01	2.27E-03	237.646	< 2e-16	***
HighBP	7.14E-02	1.83E-03	39.041	< 2e-16	***
HighChol	5.28E-02	1.70E-03	31.086	< 2e-16	***
CholCheck	3.46E-02	2.16E-03	15.985	< 2e-16	***
BMI	8.71E-02	1.83E-03	47.567	< 2e-16	***
Smoker	-3.22E-03	1.62E-03	-1.984	0.047313	*
Stroke	1.69E-02	2.41E-03	6.982	2.93E-12	***
HeartDiseaseorAttack	3.81E-02	2.48E-03	15.355	< 2e-16	***
PhysActivity	-3.46E-03	1.74E-03	-1.989	0.046719	*
Fruits	-2.31E-03	1.63E-03	-1.416	0.156666	
Veggies	-4.15E-03	1.63E-03	-2.547	0.010868	*
HvyAlcoholConsump	-2.91E-02	1.82E-03	-16.003	< 2e-16	***
AnyHealthcare	-3.62E-05	2.06E-03	-0.018	0.985981	
NoDocbcCost	-8.19E-04	1.75E-03	-0.468	0.639457	
GenHlth	1.14E-01	2.15E-03	52.744	< 2e-16	***
MentHlth	-6.83E-03	2.01E-03	-3.391	0.000698	***
PhysHlth	-1.31E-03	2.60E-03	-0.503	0.614918	
DiffWalk	2.77E-02	2.43E-03	11.397	< 2e-16	***
Sex	2.14E-02	1.63E-03	13.156	< 2e-16	***
Age	6.12E-02	1.95E-03	31.438	< 2e-16	***
Education	-7.62E-03	1.81E-03	-4.207	2.59E-05	***
Income	-2.52E-02	1.97E-03	-12.795	< 2e-16	***
GenHlth:DiffWalk	-1.72E-02	2.25E-03	-7.627	2.44E-14	***
BMI:Age	2.98E-02	1.82E-03	16.352	< 2e-16	***
Sex:Age	2.02E-02	1.74E-03	11.617	< 2e-16	***
HighChol:Age	-1.46E-02	1.87E-03	-7.827	5.07E-15	***
GenHlth:Income	1.42E-02	2.04E-03	6.962	3.38E-12	***
BMI:DiffWalk	-1.51E-02	1.69E-03	-8.945	< 2e-16	***
HighBP:HeartDiseaseorAttack	-1.17E-02	2.04E-03	-5.731	1.00E-08	***
DiffWalk:Age	-1.46E-02	2.38E-03	-6.111	9.93E-10	***
GenHlth:PhysHlth	-1.58E-02	2.12E-03	-7.423	1.16E-13	***
HvyAlcoholConsump:Age	-8.42E-03	1.62E-03	-5.19	2.11E-07	***
CholCheck:GenHlth	7.51E-03	1.81E-03	4.147	3.37E-05	***
GenHlth:Sex	-5.58E-03	1.80E-03	-3.104	0.001912	**
HighChol:HeartDiseaseorAttack	-1.06E-02	1.90E-03	-5.574	2.50E-08	***
HighChol:GenHlth	1.25E-02	1.88E-03	6.636	3.25E-11	***
HeartDiseaseorAttack:Age	-1.42E-02	2.33E-03	-6.092	1.12E-09	***
Smoker:GenHlth	-3.72E-03	1.91E-03	-1.953	0.050827	.
HeartDiseaseorAttack:GenHlth	-1.17E-02	2.18E-03	-5.342	9.23E-08	***
Fruits:Education	-4.45E-03	1.61E-03	-2.769	0.00562	**
HeartDiseaseorAttack:DiffWalk	7.96E-03	1.70E-03	4.67	3.01E-06	***
HighBP:HvyAlcoholConsump	-6.73E-03	1.72E-03	-3.903	9.50E-05	***
HighChol:Stroke	-6.48E-03	1.79E-03	-3.618	0.000297	***
AnyHealthcare:Education	-4.97E-03	1.46E-03	-3.401	0.000673	***
HighBP:CholCheck	5.26E-03	2.01E-03	2.621	0.008768	**
Smoker:Education	5.78E-03	1.67E-03	3.468	0.000524	***
BMI:HeartDiseaseorAttack	-6.99E-03	1.75E-03	-3.984	6.79E-05	***
BMI:Smoker	7.85E-03	1.69E-03	4.648	3.35E-06	***
MentHlth:DiffWalk	2.88E-03	1.66E-03	1.734	0.082975	.
CholCheck:BMI	6.31E-03	1.75E-03	3.605	0.000313	***
HighChol:BMI	-7.41E-03	1.74E-03	-4.256	2.09E-05	***
HighBP:HighChol	8.00E-03	1.80E-03	4.437	9.15E-06	***
Fruits:Age	-4.88E-03	1.61E-03	-3.024	0.002496	**
PhysActivity:Fruits	-4.09E-03	1.58E-03	-2.587	0.009676	**
Stroke:GenHlth	-6.69E-03	1.84E-03	-3.647	0.000266	***
HighBP:Sex	-4.46E-03	1.80E-03	-2.477	0.013244	*
Stroke:Age	-7.11E-03	2.18E-03	-3.258	0.001124	**
HeartDiseaseorAttack:NoDocbcCost	-3.67E-03	1.54E-03	-2.385	0.017067	*
CholCheck:Age	5.79E-03	1.63E-03	3.545	0.000393	***
CholCheck:HeartDiseaseorAttack	-8.30E-03	2.65E-03	-3.134	0.001726	**
Fruits:Sex	-4.26E-03	1.58E-03	-2.695	0.00705	**
HvyAlcoholConsump:GenHlth	-5.62E-03	1.75E-03	-3.218	0.00129	**
HighChol:MentHlth	3.46E-03	1.78E-03	1.947	0.051543	.
HighBP:AnyHealthcare	-3.15E-03	1.73E-03	-1.825	0.067959	.
HighBP:Education	3.16E-03	1.64E-03	1.923	0.054425	.
Smoker:Stroke	-4.20E-03	1.63E-03	-2.571	0.010131	*
PhysActivity:Age	-2.57E-03	1.75E-03	-1.467	0.142433	
PhysActivity:Education	-2.38E-03	1.73E-03	-1.375	0.169065	
CholCheck:AnyHealthcare	1.50E-03	8.80E-04	1.703	0.088663	.
Veggies:NoDocbcCost	-2.04E-03	1.46E-03	-1.396	0.16277	
BMI:Sex	-1.50E-03	1.74E-03	-0.864	0.387737	
MentHlth:Income	3.45E-03	1.68E-03	2.058	0.039598	*
PhysHlth:Income	-6.63E-03	1.92E-03	-3.445	0.000571	***
NoDocbcCost:DiffWalk	-4.85E-03	1.55E-03	-3.128	0.001761	**
BMI:MentHlth	4.67E-03	1.54E-03	3.035	0.002409	**
HighChol:Sex	-4.35E-03	1.69E-03	-2.572	0.0101	*
Sex:Education	4.26E-03	1.68E-03	2.54	0.011081	*
Smoker:Sex	2.60E-03	1.62E-03	1.603	0.108878	
PhysHlth:Age	-4.36E-03	2.19E-03	-1.988	0.046767	*
MentHlth:Age	4.63E-03	1.89E-03	2.451	0.014241	*
HeartDiseaseorAttack:PhysHlth	2.28E-03	1.75E-03	1.307	0.191326	
PhysActivity:PhysHlth	-3.58E-03	1.55E-03	-2.312	0.020757	*
Stroke:MentHlth	-7.58E-04	1.37E-03	-0.555	0.579203	
BMI:NoDocbcCost	2.57E-03	1.48E-03	1.743	0.081257	.
AnyHealthcare:Age	-3.11E-03	1.80E-03	-1.728	0.083994	.
BMI:PhysActivity	2.87E-03	1.59E-03	1.799	0.072039	.
Smoker:DiffWalk	-3.61E-03	1.85E-03	-1.951	0.051119	.
Smoker:Age	2.84E-03	1.69E-03	1.683	0.092318	.
Smoker:HvyAlcoholConsump	2.19E-03	1.68E-03	1.306	0.19154	
AnyHealthcare:DiffWalk	2.89E-03	1.90E-03	1.518	0.128947	
AnyHealthcare:PhysHlth	-2.91E-03	1.80E-03	-1.622	0.104798	
Stroke:HeartDiseaseorAttack	2.49E-03	1.21E-03	2.057	0.039692	*
PhysActivity:Income	-2.21E-03	1.80E-03	-1.225	0.220572	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4109 on 70599 degrees of freedom
Multiple R-squared: 0.3256, Adjusted R-squared: 0.3247
F-statistic: 370.4 on 92 and 70599 DF, p-value: < 2.2e-16

Also, the summary of the logit model is as follow:

glm(formula = Diabetes_binary ~ ., family = binomial, data = data_cv)

Deviance Residuals:	Min	1Q	Median	3Q	Max
	-3.5606	-0.805	-0.0186	0.8388	2.9678
Coefficients:	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-0.039644	0.0093	-4.263	2.02E-05	***
HighBP	0.364689	0.009789	37.255	< 2e-16	***
HighChol	0.293265	0.009418	31.139	< 2e-16	***
CholCheck	0.211387	0.012632	16.735	< 2e-16	***
BMI	0.537938	0.011194	48.057	< 2e-16	***
Smoker	-0.000839	0.009426	-0.089	0.929075	
Stroke	0.039125	0.00988	3.96	7.50E-05	***
HeartDiseaseorAttack	0.089678	0.010092	8.886	< 2e-16	***
PhysActivity	-0.015138	0.00973	-1.556	0.119726	
Fruits	-0.016817	0.009548	-1.761	0.078185	.
Veggies	-0.02492	0.009524	-2.616	0.008884	**
HvyAlcoholConsump	-0.151613	0.009861	-15.375	< 2e-16	***
AnyHealthcare	0.012612	0.009782	1.289	0.197264	
NoDocbcCost	0.005539	0.009942	0.557	0.577452	
GenHlth	0.651008	0.012747	51.071	< 2e-16	***
MentHlth	-0.035561	0.010477	-3.394	0.000688	***
PhysHlth	-0.083731	0.011995	-6.981	2.94E-12	***
DiffWalk	0.049951	0.011239	4.445	8.81E-06	***
Sex	0.133136	0.00954	13.956	< 2e-16	***
Age	0.434217	0.011148	38.95	< 2e-16	***
Education	-0.037886	0.010518	-3.602	0.000316	***
Income	-0.127918	0.011295	-11.325	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 98000 on 70691 degrees of freedom
Residual deviance: 72388 on 70670 degrees of freedom
AIC: 72432

Number of Fisher Scoring iterations: 5

5-3. ROC curves

Besides, the ROC curves of the linear probability model and logit model is in the following.

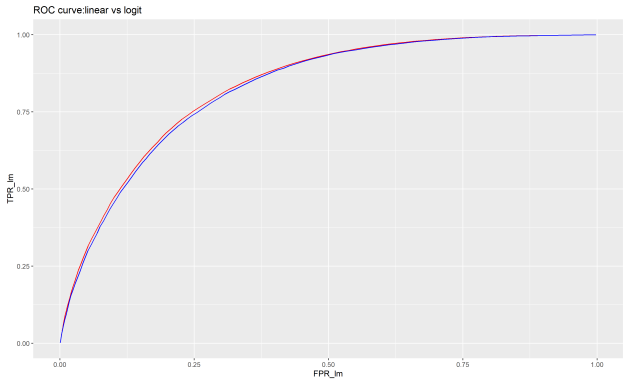


Figure 4: ROC Curves between the logit and the probit

The red line of this graph represents the best linear model and the blue line represents the logit model. These gaps are so little but the linear model is slightly better than that the logit model.

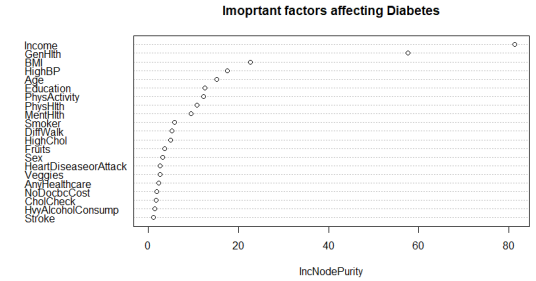
5-4. Ratio of people who got risk scores by each thresholds

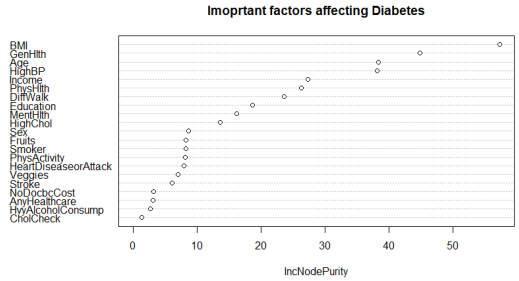
Table 4: The ratio of people who got risk scores by each thresholds
(%: all observations n= 70692)

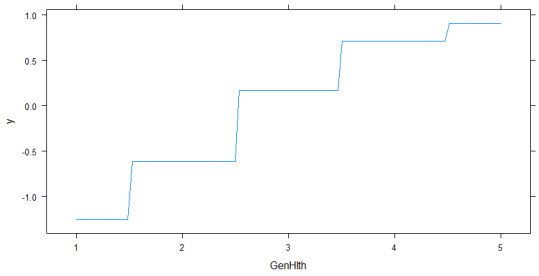
	linear	logit
low	22.3	25.6
moderate	23.4	21.3
high	30.2	25.9
veryhigh	20.1	20.9

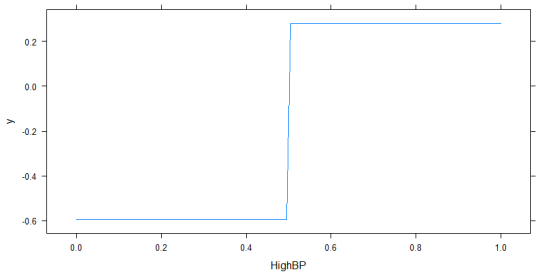
We saw the distribution of risk scores by thresholds which means the low score is below 25 points, moderate score is over 26 and below 50 points, high score is over 51 and below 75 points, and very high score is over 76 points. The above table shows the ratio of the number of people who get each score overall observations. The ratio of people who got high scores is 30.2% in the linear model and 25.9% in the logit model, which looks like a large difference. Similarly, The ratio of people who got low scores is 22.3% in the linear model and 25.6% in the logit model. However, the overall results seem us the same.

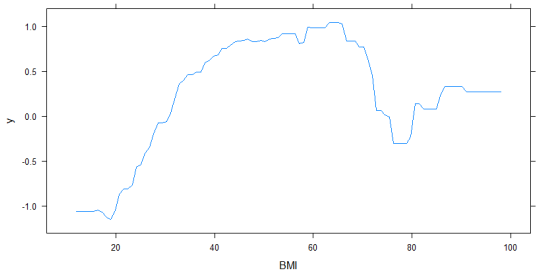
5-5. Variable Importance Plot based on Random Forest Model & Partial Dependence Plot based on Boost Model

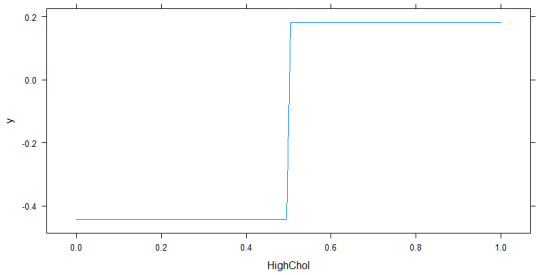


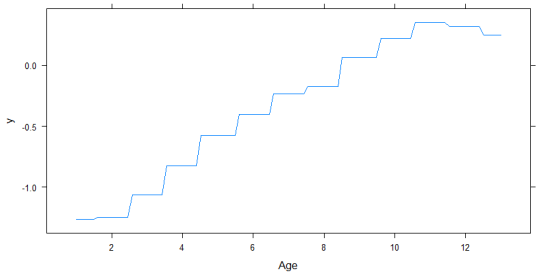


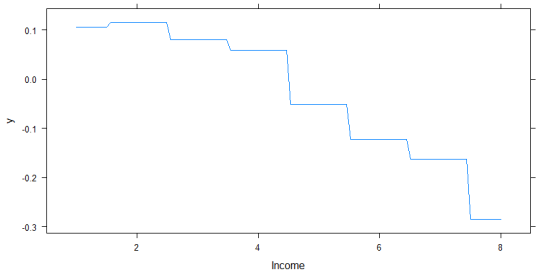


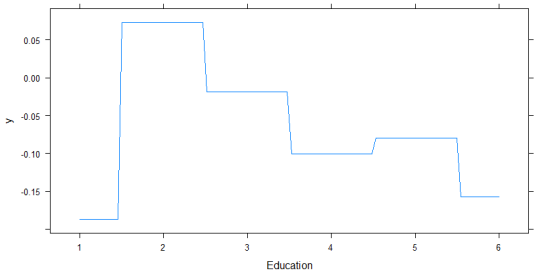






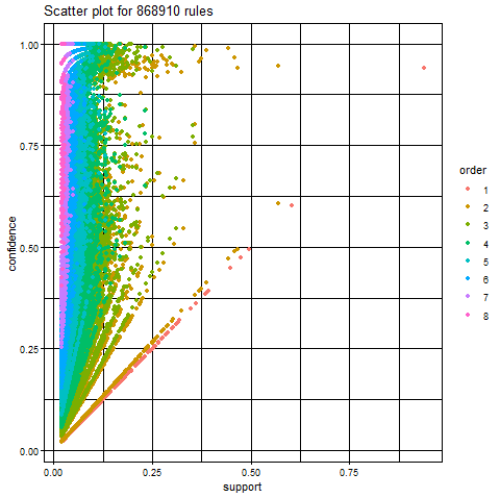






5-6. Support-Confidence plot

The support-confidence plot at the section 3-4 is as follow:



This looks many data is tend to the side of the confidence. Note that although we tried to do low confidence level, its result does not change mostly.