Gestational Diabetes Mellitus Prediction

Xuechun Wang

May 8, 2020

#Abstract

Due to the increase of gestational age, the prevalence of obesity, the change of dietary habits, and the lack of exercise, the prevalence rate has increased a lot in the past 15 years. Getting gestational diabetes mellitus is getting more and more common for pregnant women and threaten more and more pregnant women and babies' life. Therefore, based on an open access physiological data, we build up several gestational diabetes prediction models and among them, trees perform the best. With the model, we also find the most important factor in the gestational diabetes prediction is pregestational bmi and central armellini fat. Therefore, we suggest that for pregnant women and those who have a pregnancy plan, keeping an ideal weight is really important.

#Introduction

##What is gestational diabetes mellitus?

Diabetes mellitus, also called diabetes, is a term for several conditions involving how your body turns food into energy. When people eat a carbohydrate, the body turns it into a sugar called glucose and sends that to the bloodstream. The pancreas releases insulin, a hormone that helps move glucose from blood into cells, which use it for energy. For people who have diabetes and don't get treatment, their body doesn't use insulin like it should. Too much glucose stays in blood, a condition usually called high blood sugar. This can cause health problems that may be serious or even life-threatening. There's no cure for diabetes. But with treatment and lifestyle changes, they can still live a long, healthy life. Diabetes comes in different forms, depending on the cause. There're mainly 4 types of diabetes: Type 1 Diabetes, Type 2 Diabetes, Gestational Diabetes, Other Types of Diabetes(rare, about 1%-5%).

Gestational diabetes is a type of diabetes that is first seen in a pregnant woman who did not have diabetes before she was pregnant. Pregnant women with pre pregnancy diabetes are not classified as gestational diabetes. Because of the high risk, pregnant women with pre pregnancy diabetes are advised to terminate their pregnancy during screening. Some women have more than one pregnancy affected by gestational diabetes. Gestational diabetes usually shows up in the middle of pregnancy. Doctors most often test for it between 24 and 28 weeks of pregnancy. Often gestational diabetes can be controlled through eating healthy foods and regular exercise. Sometimes a woman with gestational diabetes must also take insulin. Unless other diabetes, most patients with gestational diabetes no longer need insulin after delivery, only a few still need insulin treatment.

##Why we need to talk about gestational diabetes mellitus?

However, no need for insulin treatment doesn't mean gestational diabetes can be despised. Getting gestational diabetes will threaten both moms' and babies' health during and after pregnancy.

During pregnancy, gestational diabetes increases the possibility of abortion, dystocia and postpartum hemorrhage, and may lead to diabetic ketoacidosis. For fetus, maternal gestational diabetes may lead to fetal overweight, poor growth, hyperinsulinemia, and increased risk of malformation.

Moreover, the effect of gestational diabetes does not stop at the end of pregnancy. For most of the pregnant women with GDM, their postpartum blood glucose will return to normal, but the risk of diabetes is significantly increased. According to the follow-up survey data published in diabetes care in 2002, the incidence of

postpartum type 2 diabetes in GDM patients is as high as 70%. Women with a history of gestational diabetes mellitus are also included in the guidelines for the prevention and treatment of type 2 diabetes. Another study noted that a meta-analysis of 17 studies showed that the risk of postpartum metabolic syndrome (MS) in GDM patients increased nearly fourfold. In addition, the risk of atherosclerosis and cardiovascular in pregnant women with GDM increased significantly. For newborns, offspring of GDM pregnant women had a significantly increased risk of obesity and impaired glucose metabolism. Data show that obese adolescents exposed to GDM hyperglycemia intrauterine environment during fetal period have a nearly six fold increased risk of impaired glucose tolerance or type 2 diabetes.

The prevalence of GDM is very high. The overall prevalence of GDM in the world is as high as 17.8%. The prevalence of GDM in different centers varies from 9.3% to 25.5%. The prevalence of GDM continues to increase due to the increase of gestational age, the prevalence of obesity, the change of dietary habits, and the lack of exercise. Here in U.S., according to a study comparing GDM prevalence rate in 2006 and 2016 using data from from the National Health Interview Survey, the prevalence of GDM increased by 3.6% from 2006 to 2016; and the rise was more marked among non-white, overweight, low income, age 45-64 years, and insufficient activity groups.

As the most common disease of pregnant women, gestational diabetes is bringing life threats to a large number of pregnant women and fetuses. Based on available physiological data, early prediction of the risk of pregnant women acquiring gestational diabetes will be really helpful. Based on the predicted results, helping them prepare early and adjust their lifestyle will help improve the health of pregnant women and fetuses.

##My motivation

I come from a family of doctors and most of my relatives work in the hospital. Especially, my grandma is a gynecologist and obstetrician. Since I grew up living with my grandmother, I am particularly sensitive to the health of pregnant women and fetal health. At the same time, as a woman, the structure of the body determines that for most women, pregnancy is a topic that cannot be bypassed. So we pay more attention to related fields than men. Therefore, as a female and also a descendant of doctors, I hope I can make some tentative explorations in the field of maternal health and fetal health based on what I have learned.

#Methods

##Data

###Source and Introduction

This data is an open access data from the "Visceral adipose tissue measurements during pregnancy" project. Researchers did a cohort study of pregnant women up to 20 weeks of pregnancy and followed until delivery. Study sample consisted of a cohort of 154 women approached from October 2016 to December 2017 at the Ultrasound Department of the Murialdo Teaching Health Center, a clinic that provides fetal medicine services to users of the Unified Health System in the city of Porto Alegre, Rio Grande do Sul, Brazil. Participants were followed until delivery at five Unified Health System hospitals in the city. Of the 154 women selected initially, 21 (13%) were lost to follow-up, resulting in a final sample of 133 women. The inclusion criteria were singleton pregnancy and gestational age 20 weeks. The exclusion criteria was pre-existing type 1 or 2 diabetes mellitus.

###Definition

There are 15 variables in the original data set:

- Number: Unique ID for the case.
- Age (years): Age in years.
- Ethnicity: Ethnicity (0 = white; 1 = not white).
- Diabetes mellitus: Previous diabetes mellitus (0 = no; 1 = yes).
- Mean diastolic BP: Mean diastolic blood pressure in mmHg.
- Mean systolic BP: Mean systolic blood pressure in mmHg.

- Central Armellini fat (mm): Maternal visceral adipose tissue measurement in mm.
- Current Gestational age: Age (weeks of pregnancy, days of pregnancy).
- Pregnancies (number): Number of pregnancies.
- First fasting glucose (mg/dl): First measured fasting glucose.
- BMI pregestational (kg/m): Pregestational body mass index.
- Gestational age at birth: Age (weeks of pregnancy, days of pregnancy).
- Type of delivery: 0 = vaginal birth; 1 = cesarian section.
- Child birth weight (g): Birthweight in grams.
- Gestational DM (current gestational diabetes): 0 = no; 1 = yes.

###Data Cleaning

There are 3 main problems with the data:

####Character Data

There are 2 columns are character data: "Current Gestational age" and "Gestational age at birth". The way researchers deal with these 2 data is that they put the data into a form (weeks of pregnancy, days of pregnancy) and put them in one column. So we can't build a model using these 2 columns directly. I separate the weeks of pregnancy and days of pregnancy and put them in another 2 columns labeled as "…(week)" and "…(day)". Then with the math funciton, I define a new variable as "…(days)" which equals to (7*"…(week)" + "…(day)"), which is the total days of pregnancy. This varibale can be used in the model building.

####N/A Data

There's several N/A datapoints in the data set which will affect our computation and model building. With Rstudio, we can easily omit all these N/A datapoints.

####Useless Varaiable

There are several variables we've known already they're useless in the model building, such as " number","diabetes.mellitus" and so on. To keep our data table clean and facilitate subsequent operations, I hope to define a new data set with all useless variables excluded. I build up a very generaol baseline model with "gestational dm", my dependent variable versus all other variables in the data set. And then I exclude all the variable shows N/A in the regression table.

###Data Summary

```
## # A tibble: 10 x 8
##    var                            min    q25 median    q75    max    mean      sd
##    <chr>                        <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>   <dbl>
##  1 age                           16     21     25     31     43   2.60e+1   6.51
##  2 bmi.pregestational            15.8   22.4   26.2   30.8   44.9 2.74e+1   6.46
##  3 child.birth.weight          1105   2944   3200   3630   4534   3.26e+3 515.
##  4 current.gestational.age.d~    44     93    112    133    222   1.10e+2  27.2
##  5 ethnicity                      0      0      0      1      1   4.36e-1   0.498
##  6 gestational.age.at.birth.~   190    269    277    283    287   2.73e+2  15.0
##  7 mean.diastolic.bp             52.5   65     70     74.5  100.  7.01e+1   8.18
##  8 mean.systolic.bp              90    107    113    124.   165   1.16e+2  13.2
##  9 pregnancies                    1      1      2      3      9   2.32e+0   1.68
## 10 type.of.delivery               0      0      0      0      1   2.48e-1   0.434
```

##Prediction Models

During this semester, we learned several ways to build up a predication model: linear model, linear probabilty model (for classification dependent variable), stepwise/forward/backward selection, lasso regression, trees.

As "gestational dm", my dependent varaiable is a 0-1 varaiable, so I choose linear probability model with stepwise selection to improve, lasso regression and trees. For model comparison, we will use RMSE. The model is better if the RMSE is smaller.

###Train and Test set split

First, we'll randomize the data set. Then accroding to the number of rows in the data set, we take 80% of rows as train set and 20% of rows as test set

###Linear Probability Model

For the linear probabilty model, we'll first build up a baseline model with gestational.dm versus all other variables in the train set. Then with stepwise function we automatically choose the best variables and interactions to be included. After that, we do the assessment part with test set. We predict the result using stepwised model. As the dependent variable is 0-1 variable, we can only get the probability rather than the direct result of whether the sample get the gestational diabetes or not. So we define probability > 0.5 as getting the gestational diabetes. As this is a linear probability model, I try confussion matrix to get the accuracy rate and RMSE at the same time.

###Lasso Regression

For the lasso regression I will use "glmnet" to build up the model with train set and then show out a plot of the regression. As it is not appropriate to assess the lasso using RMSE or confusion matrix, we'll try assess the model by looking at the plot.

###Trees

Trees handles categorical/numeric x and y nicely and don't have to think about the scale of x's. In this case, we have a lot of explanatory variables. In stead of considering the scale of x's and make right transformation for using regression variable selection model, regression trees will be a really good choice. Regression trees' step function is crude, does not give the best predictive performance.So we use Random Forest and boosted regression tree rather than the basic tree models.

Random Forest is great with high dimensional data, has quicker training speed, has low bias and can handle unbalanced data. So using Random Forest here also gives us a better model. We will build the model using train set and using test set to predict and check the RMSE. Also, I'll show a plot of the forest (performance as a function of iteration number) and also a variable importance plot shows how much SSE decreases from including each variable.

Boosting is a numerical optimization technique for minimizing the loss function by adding, at each step, a new tree that best reduces (steps down the gradient of) the loss function. Using boosted regression trees can help us get a better model. I'll build up the boosted regression trees with relatively large number of trees, and then agjust the n.trees to the right number by checking the error curve. Then we use test set to check the RMSE and measure the relative importance.

#Results and Analysis

##Linear Probability Model

First, see the result of confusion matrix. According to the confusion matrix, we can see that the accuracy rate is slightly above 80%, which is quite good. We can also compute the false positive rate(FPR), true positive rate(TPR) and false discvoery rate(FDR). See from the result that, FPR=0.15, which is relative low, also show that the model is acceptable. However, because the test set is tiny, so the result of TPR and FDR is not very satisfactory.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 17  3
##          1  1  0
```

```
##
##                Accuracy : 0.8095
##                  95% CI : (0.5809, 0.9455)
##     No Information Rate : 0.8571
##     P-Value [Acc > NIR] : 0.8291
##
##                   Kappa : -0.0769
##
##  Mcnemar's Test P-Value : 0.6171
##
##             Sensitivity : 0.9444
##             Specificity : 0.0000
##          Pos Pred Value : 0.8500
##          Neg Pred Value : 0.0000
##              Prevalence : 0.8571
##          Detection Rate : 0.8095
##    Detection Prevalence : 0.9524
##       Balanced Accuracy : 0.4722
##
##        'Positive' Class : 0
##
```

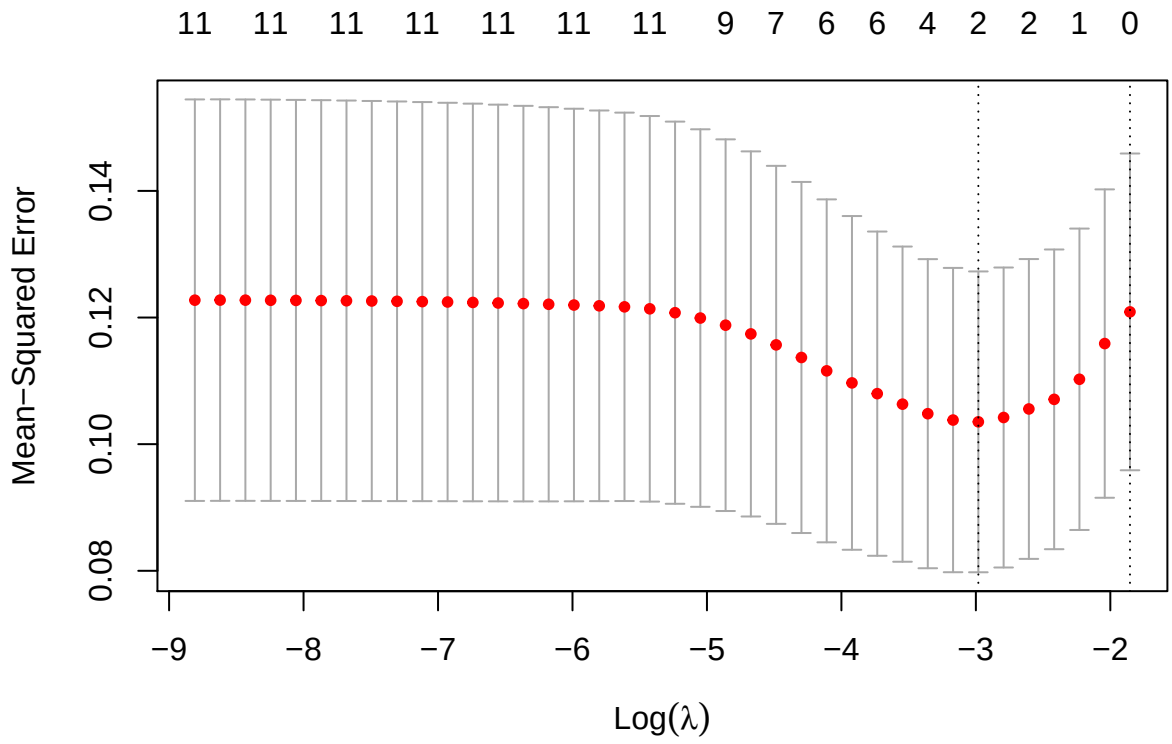
Then, we'll show a result of RMSE. As the result is slightly above 0.4, not very low but still acceptable.

```
## [1] 0.4364358
```

Finally, talk about the prediction result. According to the regression result, the increase of age, diastolic blood pressure, times of pregnancy and pregestational bmi will increase the probability of getting gestational diabetes mellitus.

##Lasso Regression

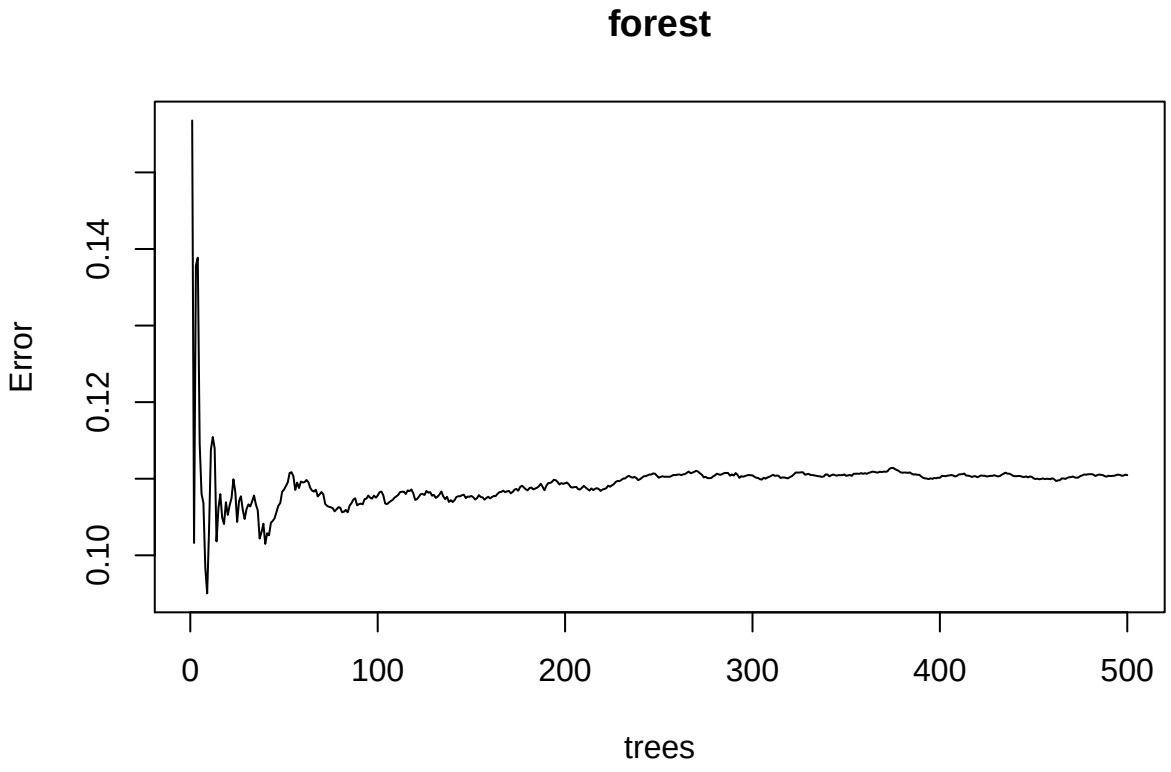
Due to the limitation of the data size, the lasso regression is not very well fitted. Therefore, we're not going to compare it with other models.



##Random Forest

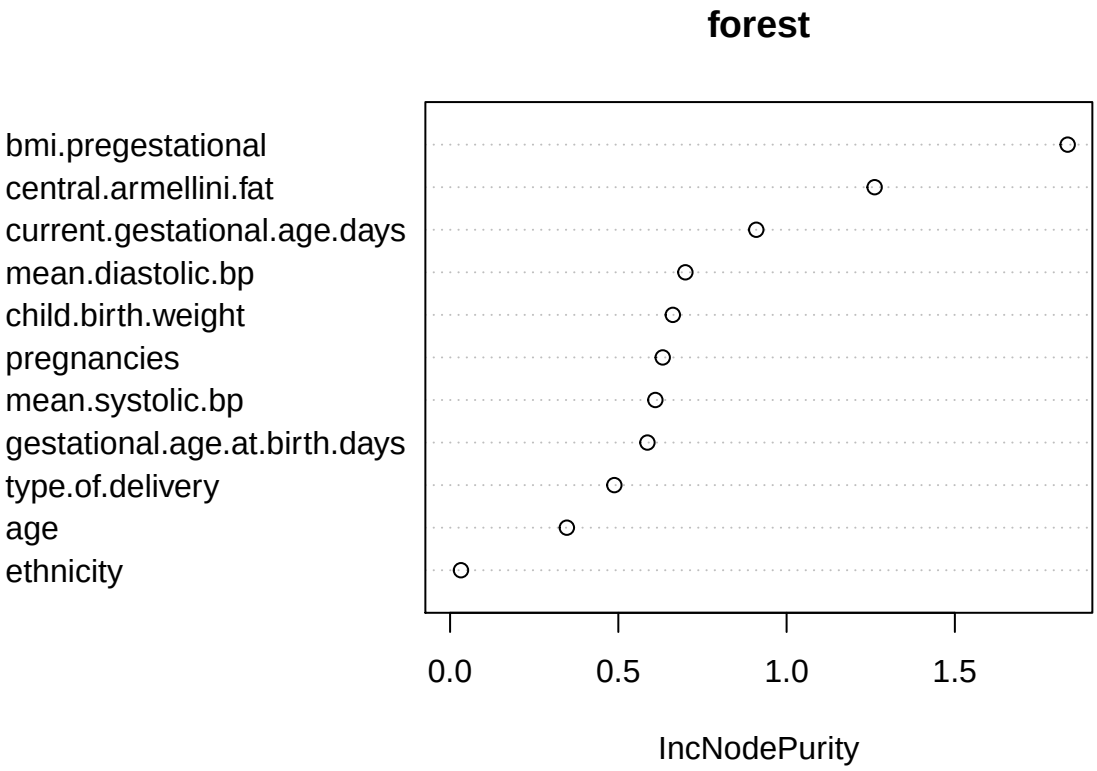
First, let's look at the out of sample RMSE results, which is slightly over 30%. It is an acceptable result and slightly lower than the RMSE of linear probability model.

```
## [1] 0.3366558
```



Then, let's talk about the prediction result. See the variance importance plot, pregestational bmi and central

armellini fat is the 2 most important factor that influnce the gestational diabetes mellitus.

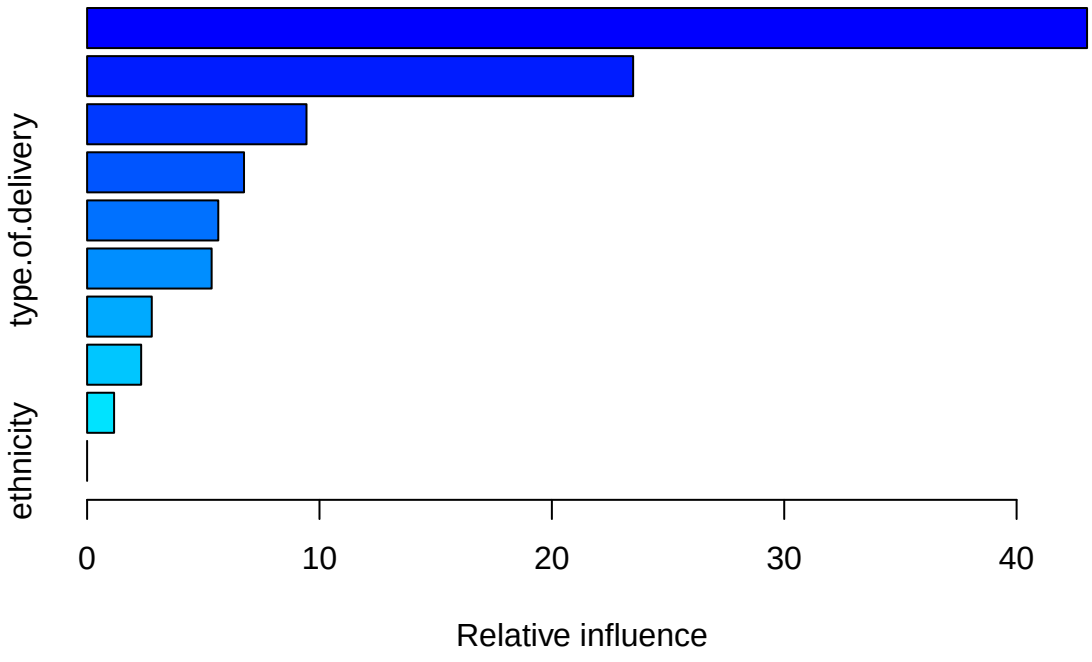


##Boosted Regression Tree

First, check the RMSE using test set. The RMSE is amazingly low, which made the boosted regression tree seemingly the best model. However, as the data set is really small, which may make the model not fit in the proper way or overfitted. So we keep reservations for the opinion the boosted regression tree fits best here.

```
## [1] 2.310236
```

Then, look at the relative importance plot. We get very similar results to the prediction result from random forest: pregestational bmi and central armellini fat are the most 2 important factors.




```
##                                                          var   rel.inf
## bmi.pregestational                         bmi.pregestational 43.031789
## central.armellini.fat                   central.armellini.fat 23.498860
## current.gestational.age.days     current.gestational.age.days  9.441653
## mean.systolic.bp                           mean.systolic.bp  6.752031
## type.of.delivery                           type.of.delivery  5.645632
## gestational.age.at.birth.days gestational.age.at.birth.days  5.359318
## child.birth.weight                       child.birth.weight  2.782802
## mean.diastolic.bp                         mean.diastolic.bp  2.326655
## pregnancies                                     pregnancies  1.161261
## ethnicity                                         ethnicity  0.000000
```

#Conclusion

Compare all the models, due to the limited data sets, lasso don't fits well. For three other models, boosted regression trees performs better than random forest and linear probability model.

From the prediction by 3 models, pregestational bmi is always the imporatnt or even the most important factor in the model. Central armellini fat is also quite important. Besides that, age, times of pregnancy, blood pressure and gestational age at birth will all increase the probability of getiting diabetes mellitus to some extent.

Therefore, for pregnant women, besides getting enough nutrition, it's also very important to exercise a lot to keep fit and stay away from overwighting. And for people with pregnancy plans, keeping an ideal weight and body shape is also important. The advice is to maintain an ideal weight before pregnancy. If you're already overweight, lose some weight and keep fit before starting the pregnancy plan. After all, maintain an appropriate weight can not only help you prevent gestational diabetes mellitus, it can also help you stay away from many other health problems. Also, for those who is older and has already been pregnant for several times, contraception may be the wiser approach.

#Deficiencies

- Lack of data. The limited data size limits the fitting of the model and makes the prediction result of the model more error prone
- Limited understanding of lasso regression

- The overall project is not problem orientation enough. Due to the limitation of technology and knowledge, some analysis that was originally intended to be realized could not be realized

#Refernces

Rocha, A. d. S., von Diemen, L., Kretzer, D., Matos, S., Rombaldi Bernardi, J., & Magalhães, J. A. (2020). Visceral adipose tissue measurements during pregnancy (version 1.0.0). PhysioNet. https://doi.org/10.13026/p729-7p53.

Kongtang. (2016). How is gestational diabetes different from diabetic pregnancy. Sohu. https://www.sohu.com/a/122108038_332478

What's the difference between type 1, 2 and gestational diabetes?. https://www.sohu.com/a/312381860_99977542

Yitang. (2019). The link between diabetes and hypertension. Sohu. https://www.medicalnewstoday.com/articles/317220#prevention

CDC. (2020). Gestational Diabetes and Pregnancy. https://www.cdc.gov/pregnancy/diabetes-gestational.html

Banma. (2017). R language: Logistic regression of binomial classification. Zhihu. https://zhuanlan.zhihu.com/p/28414024

Swinging BT. (2018). R language: ridge regression and lasso regression. https://amjiuzi.github.io/2018/11/11/lassoReg1/

Jason. (2017). Learn R | GBDT of Data Mining. Zhihu. https://zhuanlan.zhihu.com/p/25805870