Stat 141 Sugar Streak Final Report

Youhee Kil

Biying Deng

Yue Fei

Xinru Fang

Suyao Hu

This report is to predict the blood sugar

Statistics
University of California, Los Angeles
US
June 13, 2017

Contents

Abstract

Sugar streak is an application that records the daily blood sugars for its users. In this project, we are trying to use the most recent data of each individual patient to predict his expected blood sugar. Users are asked to provide the information of their blood sugar value, their gender, their birthday, their diabetes types, and what kind of meal period they just experienced. The application will record the date and time when they enter the information. Based on their information, we have treated them all as our potential predictors. In order to study the relations among all the variables, we start with exploratory analysis. Based on the plot of weekly change in blood sugar, we notice that the average blood sugar increases rapidly during the holiday season, and it has a significant decrease after the holiday season, so we we believe that the week of year is an important factor which can affect patients' blood sugar. Moreover, we find out that the average blood sugar for female tend to have more small values while the average blood sugar for male contain more large extreme values,therefore we considered gender as another important predictor when we create the model.

In addition to weekly change of blood sugar, we also explore the change of blood sugar before meal and after meal between male and female. The result is consisted with the fact that blood sugar will increase after the meal. In general, female patients tend to have a higher average blood sugar than male patients regardless of the time of the meal type. Besides, we believe that patients tend to have a higher blood sugar during the weekend since they may consume more food when they are enjoying their break. Therefore, we did further exploration on whether there will be significant change in blood sugar by each day of week for female and male patients. Surprisingly, the overall blood sugar maintains in a stable level for both female and male patients, which means that our patients maintain their blood sugar in a good condition even when they are enjoying their holiday. The sugar streak app help the patients maintain their sugar level by reminding them the risk of having high blood sugar when they are having too much sugar intake during the weekend. However, we also find that most of the patients doesn't keep track of their blood sugar all the time but with only once or twice posts per day. This may affect the accuracy of the prediction of blood sugar predicted by the app based on the insufficient data provided by the patients.

Considering the benefit of users, we think that Sugar Streak users prefer to see exact predicted blood sugar. Therefore we decide to predict blood sugar with one of prominent statistical model to predict data. At first, we use an algorithm of the 'tree' model which is called 'rpart', it is like creating a decision tree based on the app users input datas. But this decision tree has greater errors between predicted blood sugar and known user's blood sugar. Therefore, we decide to use 'randomForest', one of package of Random Forest statistical model packages, to predict the user's future blood sugar based on their input data. Random Forest predicts user's blood sugar better than predicting blood sugar level. But there is one weakness point of using Random Forest is that it cannot predict well when user's blood sugar changes suddenly so much.

Furthermore, we also try to use Random Forest to classify the blood sugar level. Refer to the medical knowledge, we define the blood sugar value less than 75 is in low level, the value between 75 to 180 is in normal level, while the value more than 180 is in high level. Similar to predicting the numeric blood sugar value, Random Forest gives us bunches of decision trees. From those trees, we can have an insight of the important factors that influencing the blood sugar level. In this analysis, we find that the most recent blood sugar condition and the week of year for each record play important roles on the blood sugar prediction. However, for those users experienced low sugar streak, our Random Forest does not work well on capturing the true blood sugar level.

In short, we believe that the week of year, the period of meals, the previous blood sugar condition, the frequency of monitoring the blood sugar are important elements for predicting the blood sugar in next few hours. For further study, we encourage users provide more medical condition and the detailed information about their meal, because some drugs and food may potentially have an impact on their blood sugar as well.

Introduction of Study

Extremely high or low blood sugar levels lead several high risk serious conditions such as high blood sugar levels can damage blood vessels and nerves which lead heart attack, stroke, coronary artery disease, loss of vision and kidney disease. On the other hand, persistent low blood sugar level which called hypoglycemia may lead lose to seizure and loss of consciousness. For those reasons, millions people with diabetes have to keep tracking on their blood sugar level. Sugar Streak was introduced to help users to monitor their blood sugar levels. Also, Sugar Streak provides the trend of blood sugar, for example, average, deviation, lowest, highest and predicted blood sugar. This report summarizes the trends of users' blood sugar and predict two types of blood sugar level - numerical type blood sugar level and categorical type blood sugar with three levels such as high sugar streak, low sugar streak and normal. - by using Random Forest Method with individuals' submission through the Sugar Streak App.

Research Question

After a rough data investigating, we had our preliminary focus on the questions below.

•What factors are important in predicting patients blood sugar?

•Can we create a prediction model from these 125 patients and generally use it for other new patients?

•Does the model perform well for different individuals, such as typical patients with low, normal, or high blood sugar?

Description of the sample and data collection

The data is collected from 125 users on the app Sugar Streak. Each user enters different numbers of posts daily. They input their value of blood sugar, their diabetes types, birthday, Sex, the period for their measurement. There are 62027 observations, each observation is for a specific period of the blood sugar for each user. The app records the date and time of user's' activities, such as creating the record, and updating the record. This data records those posts from June, 2015 to April, 2017.

Variables of the study

Variable Names	Description
id	unique number for each record
blood sugar	numeric value of blood sugar
blood sugar type	specific period for blood sugar(8 categories): bedtime, overnight, post breakfast, post dinner, post lunch, pre breakfast, pre dinner, pre lunch, and Unknown
created at	date and time of the record begin to enter
updated at	date and time of the record completely input
user id	unique number for different users
posts per day	the number of posts per day
ave blood sugar day	average blood sugar per day
sd blood sugar day	standard deviation of blood sugar per day
day of week	the day of week for each post
week of year	the week of year for each post
preMeal	True = pre meal, False = post meal, bedtime, or overnight
postMeal	True = post meal, False = pre meal, bedtime, or overnight
age	age = the day of today - birthday
Diabetes Type	the type of diabetes, 1, 2, or others
prev blood sugar	previous value of blood sugar, the most recent one
blood sugar cat2	levels of blood sugar: low = blood sugar < 75, normal = blood sugar [75,180), high = blood sugar >= 180
Sex	gender of users
overall sd	the overall standard deviation of patient's blood glucose level
posts per week	average number of posts the patient made per week
season	spring, summer, fall, winter
prev ave	average blood glucose level on previous day
prev sd	standard deviation of blood glucose level on previous day

Statistical methods

The main goal of this study is to predict sugar streaks in advance so that we can give diabetes patients warnings beforehand. This question is approached in two ways. Firstly, we can predict the numerical value of blood glucose level and then classify it into "normal" or "low/high sugar streak". In this case, our response variable will be numerical. Secondly, we can define a "low/high sugar streak" and a "normal" blood glucose level based on the standard medical definition, which lead to a categorical response.

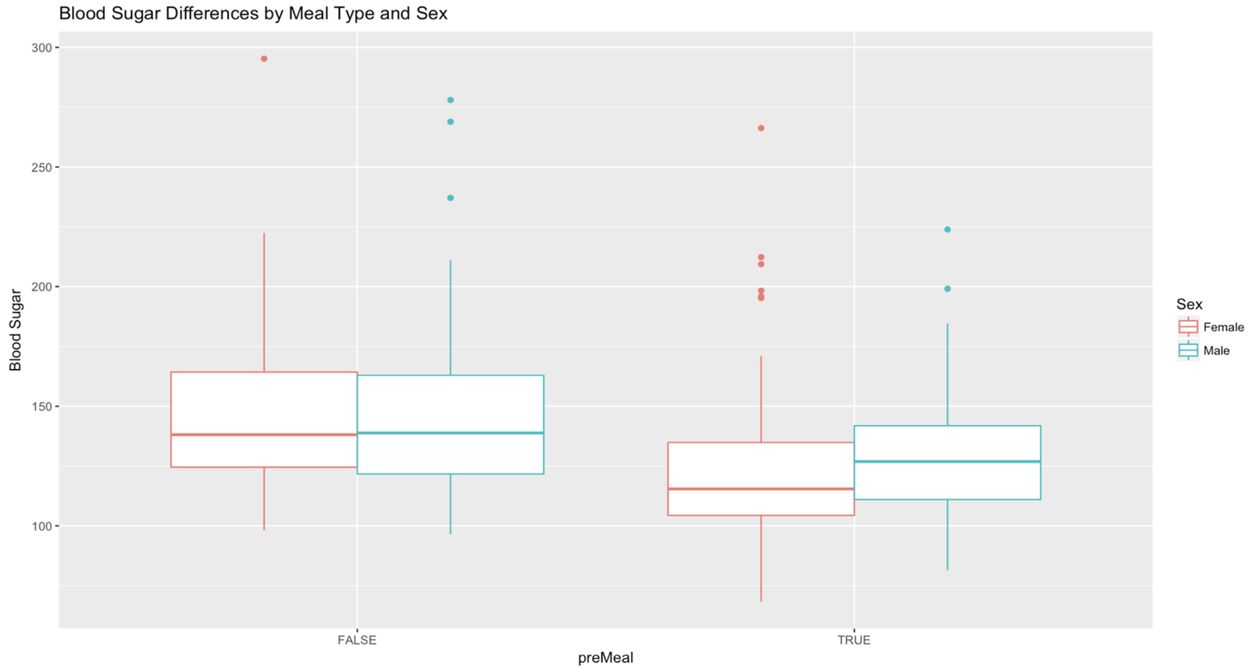
Random forest is used for both approaches since it is able to deal with both numerical and categorical responses with promising accuracy. It is a decision tree-based ensemble method in which the "random" part means to randomly create a subset of the data with replacement, and the "forest" part means to grow a tree for each of the random subset thus creating a "forest". In a decision tree, an input is entered at the top node and gets separated into different leaves based on a set of decision rules.

In the case of random forest regression (numerical response), the output is the mean prediction of individual trees. In the context of our study, it refers to the mean predicted blood glucose level of individual trees in our forest. While in the case of random forest classification (categorical response), the output is the mode of the categories for individual trees, which refers to the "majority vote" among "Normal", "Low", and "High" for all individual trees.

Summary of findings

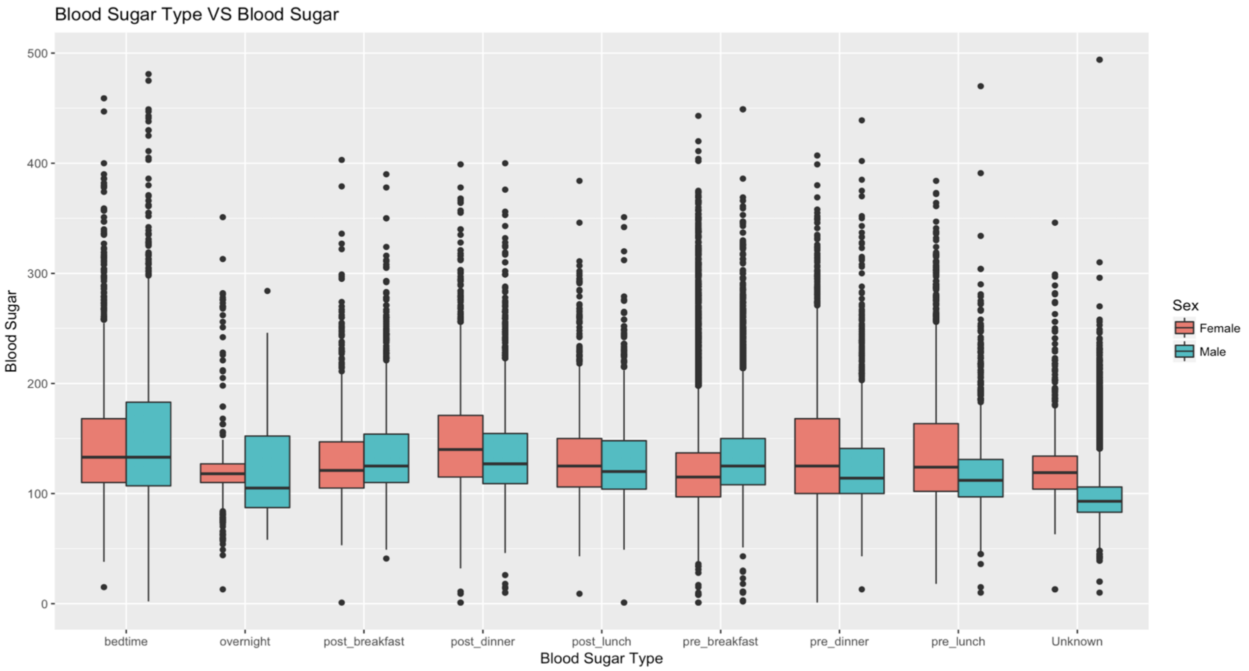
Exploratory data analysis

5.1 Relationship among Blood Sugar, Meal Type and Sex.



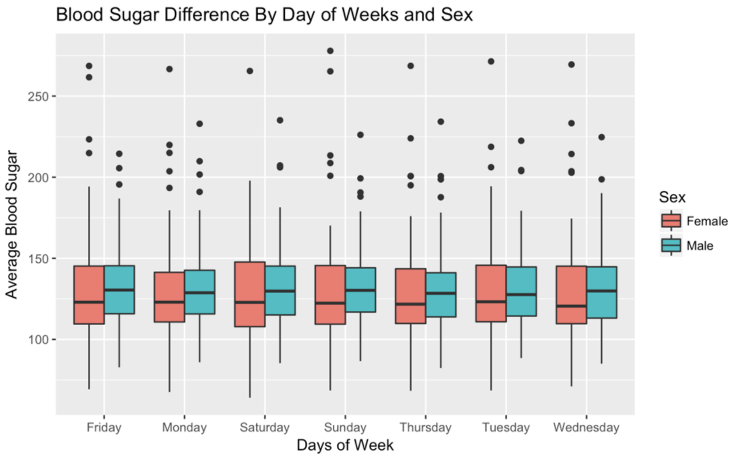
We are interested in exploring the change in blood sugar between different types of the meal (pre-meal and post meal) as well as gender (female and male). By implementing an Anova analysis, we discover that there is a statistically significant difference in blood sugar between each gender and meal types with p-value (<2.2e-16). From the plot, we can see that the female patients have similar average blood sugar about 147.1 mg/dl after the meal which very close to the average blood sugar 147.4 mg/dl for male patients. However, female patients have higher average blood sugar than male before meal while female has lower median than male as shown in the plot. Therefore, we find out that the blood sugar level after meal is generally higher than the blood sugar level before meal which is consistent with the fact that people tend to have their blood sugar increase after consuming food.

5.2 Relationship among Blood Sugar, blood sugar type and Sex.



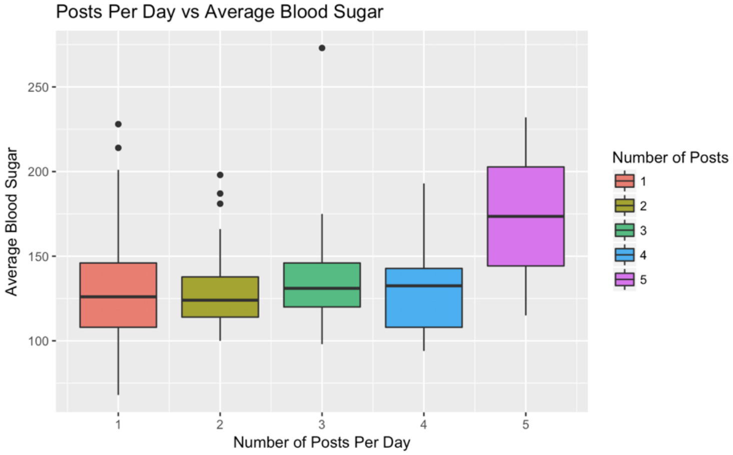
After exploring the relationship among the blood sugar, meal types and gender, we decide to look at the relationship among the blood sugar, sex for each specific blood sugar type (pre/post-breakfast/post dinner, pre/post lunch, overnight and unknown). By conducting an Anova test and a linear model, both of them indicate that there is a significant difference in blood sugar level between each sugar type and gender with p-value (<2.2e-16). From the plot, we find out that the blood sugar level for post-breakfast, post-dinner and post-lunch are all higher than the blood sugar level for pre-breakfast, pre-dinner and pre-lunch.

5.3 Blood sugar difference by day of weeks and sex.



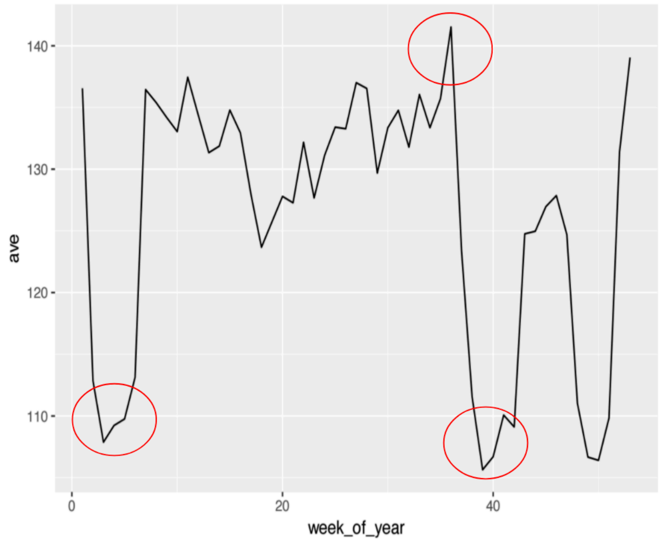
We are also interested in whether there is a difference between female and male on average blood sugar for each day of the week. At first, we believe that blood sugar will be significantly different between each day of the week since people tend to eat more during the weekend, which may result in a fluctuation in average blood sugar on Saturday and Sunday. However, we observe that the average blood sugar for the female is around 131.9 mg/dl for all days of the week while the average blood sugar for the male is also stable around 132.8 mg/dl. Besides, the median of the average blood sugar for the male is about 125.5 mg/dl while for the female is a little bit lower than the male that is about 120.5 mg/dl, which also indicates that there is no significant difference in blood sugar between male and female in each day of the week. To confirm this result, we again conduct an Anova analysis and observe that it has consisted with our conclusion that there is no significant difference in average blood sugar between each day of the week and the gender.

5.4 Relationship between Posts per day and Blood sugar



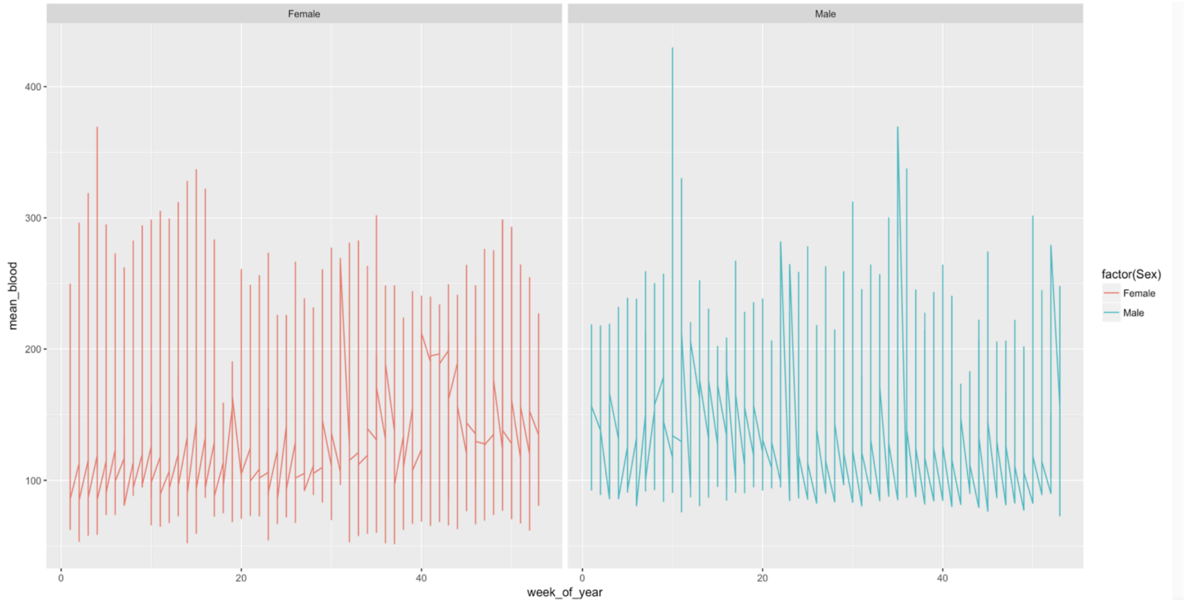
At first, We thought that there might be no relationship between the number of post per day and their blood sugar. However, considering that patients who have high blood sugar will be more worried about their change in blood sugar than those whose blood sugar is stable, and we believe that if the patient begins to worry about their blood sugar, they tend to measure and record their blood sugar more frequently. From the plot, we find that when there are 5 posts per day, it has the highest median and the highest mean 173.5 mg/dl while 1 post per day has average blood sugar about 126 mg/dl. However, we also observe a fact that there is only 2 patients post 5 times per day while 45 patients post 1 time per day. As the number of post per day increases, the number of patients decreases. By conducting the Anova analysis, we find out that post per day is not statistically significant in blood sugar change.

5.5 Relationship between week of year and average Blood sugar



In the exploratory analysis, we are interested in the change of the average blood sugar throughout the time, so we create a new variable called week . of . year which represents the week of the year, and we calculate the average blood sugar of all the patients by week . of . year. Based on the plot, we find that there are some significant changes. The first significant drop happens after the first week of the year, and it reaches the lowest blood sugar on the third week. The reason why we think this happens is that the users might celebrate the new year during the first week, and they do not control their blood sugar so strict as they used to do during the normal days. After the holidays, they start to control their blood sugar. After the first quarter, the patient's' blood sugar tends to be more stable. The average blood sugar reaches the peak on the week 36 and decreases dramatically right after it until it reaches the lowest blood sugar on week 39. We believe that the users control their blood sugars to prepare for the holiday season. As we expect, the average blood sugar increases during the November due to the Thanksgiving and decreases after that. Starting from week 50, the average blood sugar increases until the end of the year due to the Christmas and the New Year.

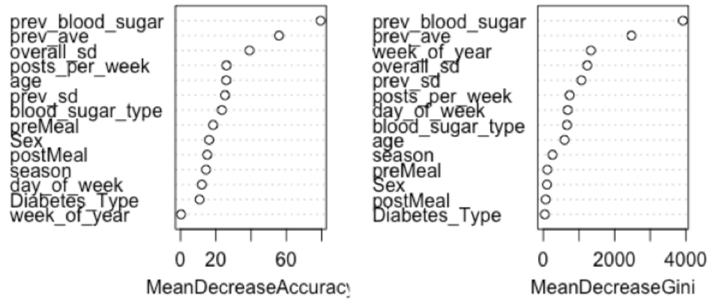
5.6 Relationship between week of year and average Blood sugar by gender



After this finding, we wonder if there is a significant difference on average blood sugar between female and male. We finally decide to calculate the average blood sugar based on the gender of each patient since we have 59 female patients and 61 male patients. Based on the plot of average blood sugar by gender, we can see that the average blood sugars for both genders are similar. Female patients have more low blood sugars than male while male patients have more extreme high blood sugars than female.

Random Forest Classification

With our random forest classification model, we are able to extract the important predictors for sugar streak episodes.



The variable importance plot on the left is based on Mean Decrease Accuracy, from which we can conclude that the most recent blood glucose entry, the average blood glucose level on previous day, the overall standard deviation of the blood glucose level for each individual patient and average number of posts the patient make per week are the top five predictors for sugar streak. The plot on the left is based on Gini index, in which the importance of predictors are slightly different from the one on the left, besides, it also indicates that week of year also influences blood glucose level.

Since the question of interest is whether we could build a model based on all 125 patients and generally use it on new patients, we use Leave One Out Cross Validation (LOOCV) to test the performance of our model. Generally speaking leave one out means to leave one observation out, in our case we define it as leaving one patient out.

Furthermore, we would like to know whether our classification model works on different types of patients, particularly for patients who often experience sugar streaks instead of normal patients. We manually selected three typical patients of interest: patient number 1508 who has rarely experienced sugar streak, patient 821 who has often experienced low sugar streaks, and patient 1258 who has often experienced high sugar streaks.

Patient 1508 (Normal)

Patient 1508			
Prediction	True Value		
	Low	**Normal**	**High**
Low	0	0	0
Normal	0	196	2
High	0	0	0

As we can see from the confusion table for patient 1508, we can successfully predict 196 normal blood glucose entries out of 198 total entries with accuracy 0.99. However, we fail to predict 2 high sugar streaks that the patient has experienced. Since our accuracy is relatively high, these mistakes do not necessarily suggest that our model fails. One possible interpretation could be that the model is robust enough not to give false warning to tension the patient.

Patient 821 (Low)

Patient 821			
Prediction	True Value		
	Low	**Normal**	**High**
Low	0	0	0
Normal	176	77	0
High	0	0	0

The confusion matrix above gives disappointing predictions for patients 821. Out of the 176 low sugar streaks that the patient has experienced, we predict all of them to be Normal. The only part we succeed is to predict 77 Normal entries, which leads to the accuracy level of 0.30. We have to admit that our model cannot be applied on this type of patients.

Patient 1258 (High)

Patient 1258			
Prediction	True Value		
	Low	**Normal**	**High**
Low	0	0	0
Normal	0	7	69
High	1	27	588

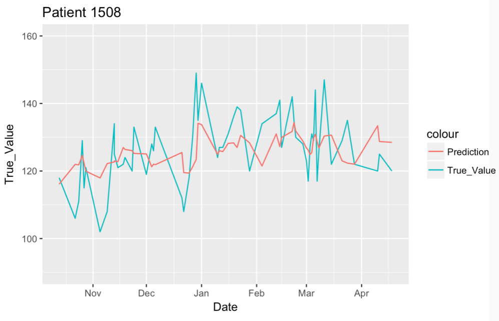
Given the overall accuracy of 0.86 on patient 1258, it is clear that our model provides moderate prediction power on patients who tend to experience high sugar streaks. Out of the 657(588+69) high sugar streak episodes, we can successfully predict 588 out of them. However, our model also predicts 27 Normal sugar entries and 1 Low sugar entry to be High, this pattern of mistakes may suggest that our model is hypersensitive to high blood glucose entries.

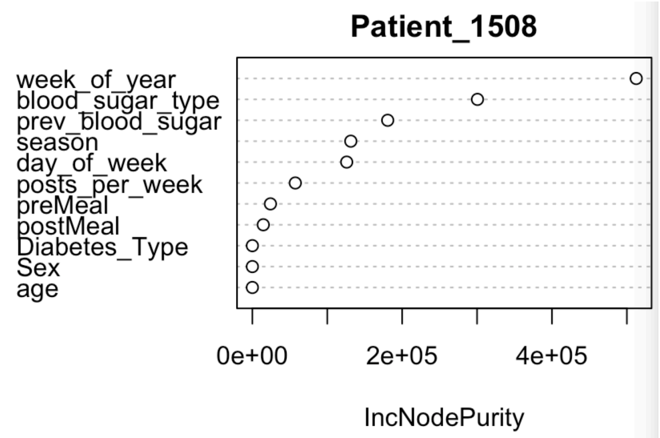
Random Forest Regression

By using predict.randomForest function to predict numerical blood sugar level. randomForest implements Breiman's random forest algorithm for classification and regression. At first I used rpart package to predict numeric blood sugar level, but it always gave bigger Root-Mean-Square-Error (RMSE). RMSE indicates how much error there is between predicted value and an observed or known value. Differences between the rpart package and the randomForest package is not so much distinctive, but it gives different results. the rpart package provides an algorithm of the 'tree' model, creating a decision tree. The results are easy to interpret, but the drawback of the rpart package is the predictive power is not great. On the other hand, the randomForest package produces a large number of tree by bootstrap, makes many decision trees on different subsets of the regressors and averaging them together. Therefore, it reduces the model variance but it is hard to interpret.

As we mentioned it above, we used three patients, 1508, 821 and 1258, who are experiencing high sugar streak, low sugar streak and normal sugar streak to test Random Forest model to predict the blood sugar. We selected important variables from randomForest classification model. When we predict blood sugar, important variables are different by patients. Random Forest predicted blood sugar better than Random Forest classification when we compared with observed data. But one of weakness point of using Random Forest to predict the blood sugar was that it could not captured sudden changes of blood sugar. However, most of statistical models would not predict well sudden changes of blood sugar.

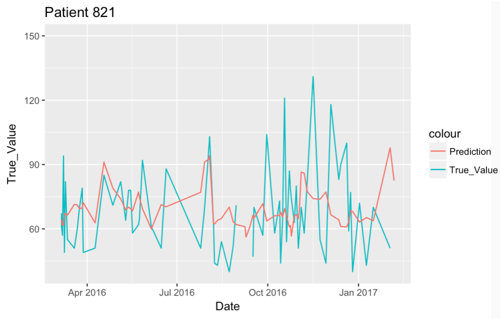
Patient 1508

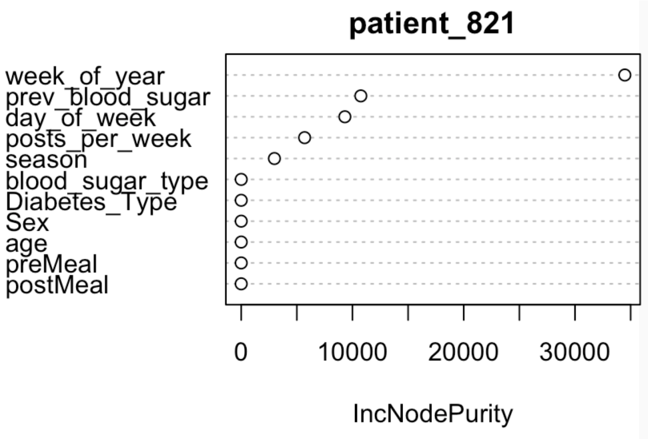




Predicting blood sugar through rpart package showed bigger error between predicted blood sugar and known blood sugar compared to randomForest package. (rmse from rpart package was 65.39716 and rmse from randomForest package was 63.72641). Important variables to predict blood sugar for patient 1508 are week.of.year, prev.blood . sugar, day . of . week, posts . per . week, season.

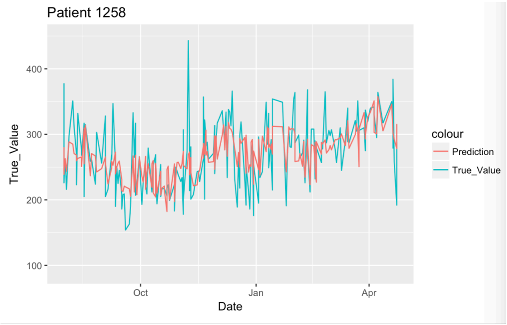
Patient 821

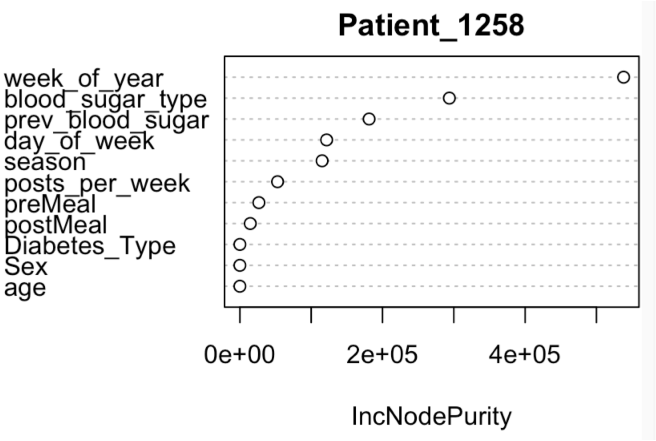




Predicting blood sugar through rpart package showed bigger error between predicted blood sugar and known blood sugar compared to randomForest package. (rmse from rpart package was 27.39695 and rmse from randomForest package was 20.85876). Important variables to predict blood sugar for patient 821 are week.of.year, prev.blood.sugar, day.of.week, posts.per.week, season.

Patient 1258





Predicting blood sugar through rpart package showed bigger error between predicted blood sugar and known blood sugar compared to randomForest package. (rmse from rpart package was 44.24823 and rmse from randomForest package was 40.38834). Important variables to predict blood sugar for patient 821 are week.of.year, blood.sugar.type, prev.blood.sugar, day.of.week, season, posts.per.week, preMeal, postMeal.

Conclusion

From exploration, we can see that there is significant relationship among blood sugar, the type of meal and gender. During the further analysis, we can conclude that the most recent blood sugar condition is an important factor to the trend of the blood sugar. The average blood sugar level on previous day, and the last entry of blood sugar value compose the definition of most recent blood sugar condition. What's more, the frequency of recording blood sugar in a week have a great impact on the accuracy of predicting the future blood sugar level. For classification, our model performs well on users who usually have high blood sugar level or normal blood sugar level, but it does not work quite well for those users with low blood sugar level. For predicting the numerical value of blood sugar, our overall predicted values are lower than the actual values. Hence, there's small root mean square error on predicting users with low blood sugar. Our regression model performs better on three types of users, but it's hard to capture those sudden changes.

Shortcoming of the study

In our study, we try to answer the question that if our model can be generalized to perform well on other new users. Due to the large size of the data, we fail to use the leave-one-out method, which is the method that leaving a specific user out of the train model and treating the information of that specific user as a test sample.We plan to take turn on testing these 125 users to estimate our model accuracy. However, due to this limitation, we have to split the train data into relatively small parts in order to perform it successfully. Therefore, we manually test on three typical users who frequently use the app and usually post their blood sugar in low, normal, or high level. At some extent, this may lose the generalization of using the model on other users.

Considering the diabetes types would be a potential factor influencing people's blood sugar, unfortunately, we find that there was only one user with diabetes type one. This situation blocks our analysis from comparing diabetes type one and diabetes type two on influencing the blood sugar.

Recommendation for future research

For further study, it's good to collect more medical information and habits of users, because certain drugs will have great impact on the level of blood sugar. Since post meal and pre meal will influence blood sugar differently, we recommend users provide more detailed information about their meals. Different food that users have taken will be a potential factor for blood sugar.

Reference

Winter 2107 Results, "Sugar Streak Final Report", by Aaron Garcia, Sophie Lellis-Petrie, Harsh Patel, Daezsha-Nay Williams.