Sugar.IQ – A Cognitive Mobile Personal Assistant for Diabetic Patients

https://www.ibm.com/case-studies/medtronic

ISYE 6501 – Course Project

Fall 2019

Aadith Ramia

15 April 2019

Contents

Case Overview

This case study is regarding a product called Sugar.IQ developed by Medtronic. This is a sugar level monitor that assists diabetic patients in keeping their sugar levels under control by providing timely alerts and insights. The product equips the patient with a wearable device that constantly collects key metrics from the patients body and relays them to the cloud for analysis. The device also allows the patient to provide his own inputs such as details on his diet. Data collected in the cloud is used to predict a potential low-sugar or high-sugar situation the patient might run into based on all the collected data (diet, physical activity, sleep pattern etc) .The solution also throws light on aspects of the patient's life style that have the highest impact on the patients sugar level, thereby giving them an opportunity to take corrective actions.

Design considerations

Sugar levels of a person being time series data, the immediate approach that comes to mind is to use time series models like GARCH. However, using time series analysis here might not be the best approach in this case because:

1. Sugar level is a complex function of several factors. A time series model would only consider the past observed values of sugar levels. It wouldn't have any way to factor in any of the other data points collected by the device which would be crucial determinants of sugar level. Predictions which are just based on past values would have significant noise.

2. Time series models allow prediction for only the immediate next time interval. Here we need predictions for multiple time intervals

3. It would be of interest to understand the extent to which each lifestyle choice has an impact on sugar level. Time series models have no way to do this.

Regression model would be effective in addressing these requirements. We will use direct multi-step forecast strategy to predict sugar level of patients for several time intervals into the future. Basically, this involves building one regression model for each time interval we would want prediction for. We will also use the model to study the relation between the patients lifestyle choices and sugar level to find which choices have the largest impact on sugar levels.

Solution Overview

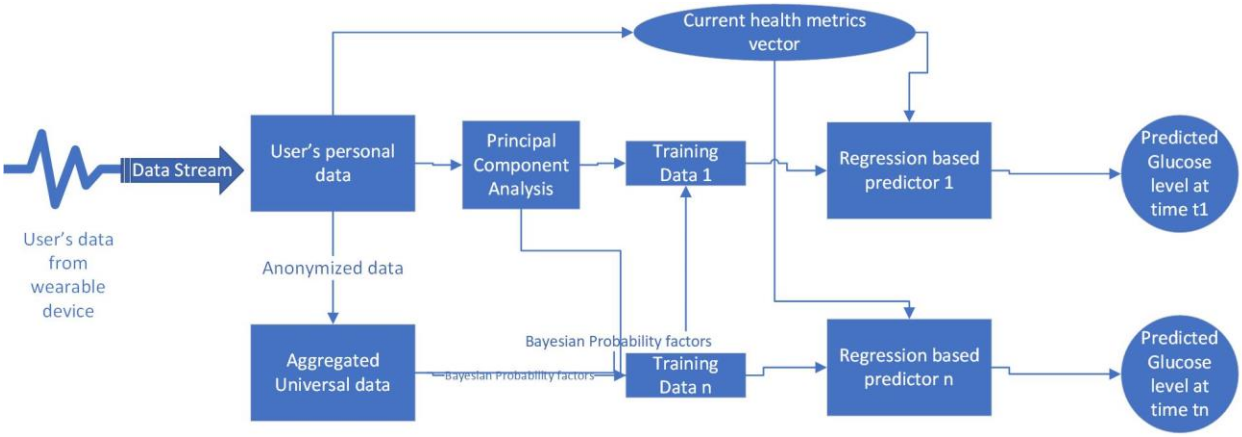
The solution would be primarily based on the following models:

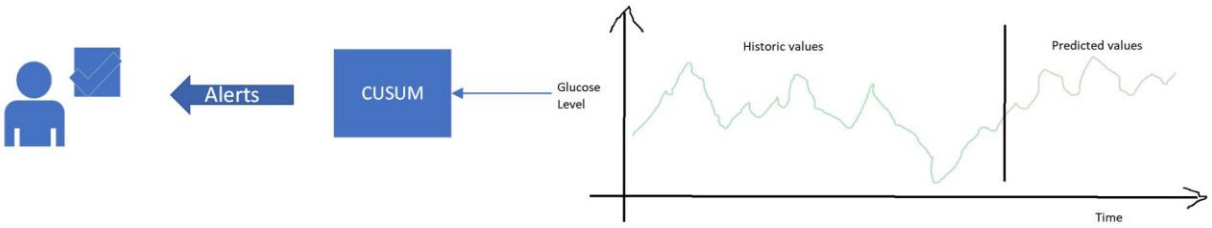
1. An array of **Linear Regression models**, each of which would predict sugar level for a time in future eg, model 1 would predict for 15^{th} min from present, model 2 would predict for 30 mins from present and so on.

2. An empirical **Bayesian inference model** would be used to find the probability of having a particular sugar level at particular point in time in future for each individual based on his current condition (represented by the feature vector). We will leverage data collected from the entire user base to compute the prior probabilities of various conditions occurring and to compute the likelihood of sugar level being at various levels given those conditions. The probability of having each sugar level will be used as factors in the regression model. The predictions made for individual users will have improved baseline accuracy even when the data for the given user has no precedent in current user's data.

3. The linear regression models will be based on a dataset that maps health metrics received for the past few time intervals to the current sugar level. We will use **Principal Component Analysis** to transform the raw dataset into a set of independent factors which account for most of the variability in original data.

4. Combining the recorded sugar levels for the past few intervals with the predicted sugar levels for the upcoming few intervals obtained from the regression models would give us time series data. We will use **CUSUM** to check whether there is excessive rise/drop expected, which can then be used to alert the patient so that he can take corrective action (eg, having a quick snack if the sugar level is dropping)





The Linear Regression Model

We know that current sugar level for a typical individual depends on several factors – both past and present. Our Regression model will have to include all these factors and fit a model that captures the relation between these factors and current sugar level.

Factors

We could collect several different kinds of metrics which can be used to derive insights on patterns in sugar levels of the patient over time. Following is a representative sample of factors that could be used.

Account level factors

Some of the details provided by the patient when he/she acquires the product and creates account as part of initial setup can be used as factors.

- Date of Birth (which will be used to compute age)
- Gender
- Ethnicity (which might be optional for legal reasons based on local laws)
- Medical history info – details on pre-existing conditions that might impact sugar levels

Health KPIs collected from the wearable device

Every 15 mins, the wearable device would be relaying health metrics that it would collect from the patient's body. These can be used as factors

- Heart Rate
- Blood Pressure
- Current Sugar level
- Calories burnt (measure of physical activity)

Event based metrics

These are metrics that might be getting collected upon occurrences of certain events, generally triggered by the patient's actions detected by the device.

Sleep metrics

- Start time of sleep
- End time of sleep
- Avg Sleep quality

Diet metrics

- Time of intake
- Amount of various nutrients

Derived factors/interaction parameters

It is evident that several of the parameters captured are not fully independent. Their interplay could also be a good determinant of sugar level. Here is an example:

Metabolism rate

We know that sugar level is strongly dependent on when and what a person ate. Suppose two people shad the same meal two hours back. Their current sugar levels would still be different depending on how fast the food gets digested. This is captured by metabolism rate, which would be a function of several factors such as age, gender, amount of physical activity etc. Such interaction parameters combined with the individual factors would help model the relation between the factors and sugar level better.

Using Probabilities from Empirical Bayesian inference model as factors

Since we would have data from several users collected over a significant length of time, we could leverage that to compute the probabilities of sugar level being at various value for a given patient, under his current conditions as indicated by his health metrics

Eg:

P(Sugar level = 80|age = 60, gender = male, heart rate = 85, calories burnt = 40,..)

= P(age = 60, gender = male, heart rate = 85, calories burnt = 40,..|Sugar level = 80) *P(Sugar level = 80)

/ P(age = 60, gender = male, heart rate = 85, calories burnt = 40,..)

The probabilities of sugar levels being at various levels would be highly correlated with the sugar levels, the target variable of our model. This will help improve the baseline accuracy of the model.

Training Data Preparation

Sugar level depends not only on the current conditions, but also on various factors from the immediate past, say past 48 hours. For example if a person did not have good sleep or has been involved in a lot of physical activities in the past 2 days, then his sugar level can be low because of that. To capture such relations, data points gathered from the past would have to be stitched together and mapped to the current sugar level.

Here is what our training data would look like, if we were to build a model to predict sugar levels 15 mins ahead.

Factors					**Target variable**
Account Level factors (age, sex, details on pre-existing conditions etc)	Metrics collected at time t – 15 mins	Metrics collected at time t – 30 mins	. . .	Metrics collected at time t – 48 hrs	sugar level at t^{th} interval

Model in Action - Predicting Sugar level for the next time interval

To get the prediction for the next time interval, just like we did to prepare training data, we would build up a vector of all factors from the past, using the same number of previous intervals as we did for training,starting with factors collected for current interval. We feed that vector to the model to get the prediction for the next time interval.

Input					**Prediction**
Account Level factors (age, sex, details on pre-existing conditions etc)	Metrics collected at time t	Metrics collected at time t – 15 mins	. . .	Metrics collected at time t – (48 hrs-15mins)	sugar level at $(t+1)^{th}$ interval

Challenges

While incorporating past data in this manner would give the model an opportunity to get a comprehensive view of what is going on in the person's life and relate that to the current sugar level, this brings up a couple of challenges:

1. Too many factors -- This will be especially problematic for new customers when there is not much data available. Having so many factors would in such a scenario would lead to overfitting.

2. What we have as factors is essentially time series data for each of the metrics. Values of the same metric across intervals (eg, heart rate across successive intervals) would be highly correlated with each other. There would also be correlations across metrics (eg, calories burnt and heart rate).

Dimensionality reduction with Principal Component Analysis (PCA)

The problems posed by having too many highly correlated factors can be addressed by applying principal component analysis on the data. This would give the following benefits:

- The process would give us a set of independent factors each of which would capture correlations between various original factors.
- The factors obtained from PCA would be ranked by importance ie, how much variance in data that each of them capture. We could select the first few of those factors that capture most of the variance and discard the insignificant components.

Before applying PCA, the factors would have to be scaled to ensure that original factors with vast range do not distort the transformed factors.

Generating Predictions for several future intervals -- Direct Multi-step Forecasting

The model outlined so far helps predict sugar level for one time interval. This by itself is of limited value. However, if we can generate predictions for several time intervals into the future, then we can stitch up past observed values with the predicted values and look for upward/downward trend.

We will achieve this using a technique called Direct multi-step forecasting. This involves building multiple linear regression models – one each for every time interval into the future that we would want prediction for. The basic structure of each of these models would be exactly identical, except for one minor difference, which is explained below. Note that the following discussion leaves out the account level factors because their treatment is the same across all the models.

Lets denote the health metrics for a patient at time t as H_t and the sugar level at time t as s_t

For the basic model we built earlier, the training data preparation involved mapping the health metrics for the past 48 hours starting from a given time to the sugar level observed at the end of subsequent interval. This model would predict the target value for one time interval ahead.

$$H_{t\text{-}48hrs}\ ,\ H_{t\text{-}30\ mins}\ ,\ H_{t\text{-}15\ mins}\ ,\ H_t \rightarrow S_{t+15\ mins}$$

Now, to build a model to predict another time interval ahead, we would have to map the same set of metrics to sugar level for the immediate next time interval and train the model against this data.

$$H_{t\text{-}48hrs}\ ,\ H_{t\text{-}30\ mins}\ ,\ H_{t\text{-}15\ mins}\ ,\ H_t \rightarrow S_{t+30\ mins}$$

Likewise, to predict another time interval into the future, we would train the model using data with the following mapping:

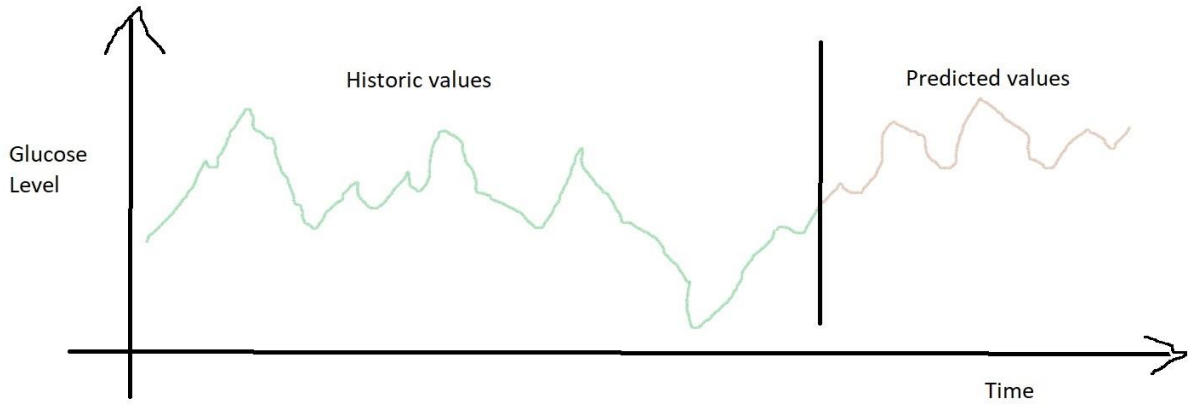
$$H_{t\text{-}48hrs}\ ,\ H_{t\text{-}30\ mins}\ ,\ H_{t\text{-}15\ mins}\ ,\ H_t \rightarrow S_{t+45\ mins}$$

And so on.

To get forecasts for several future intervals, we would build up the feature vector as before and use it against each of the models to get predictions for the respective intervals.

CUSUM - Alerting patient on updrward/downward trend in sugar levels

Now that we have the recorded sugar levels for the past several intervals and the predicted sugar levels for the next few time intervals, we can use a CUSUM model to examine this data to check whether the sugar level tends to continuously increase or decrease beyond critical levels.



We could then alert the user accordingly.

Detecting increasing trend:

For each time interval, we compute $I_t = \max\{0, I_{t-1} + (s_t - mu - C)\}$ for each time interval where s_t is the sugar level at time t, mu is the expected sugar level under normal conditions and C is a sensitivity parameter that be used to control how sensitive the change detection is to variations in sugar level. The higher the value of C, the lesser the number of false positive alerts.

Detecting decreasing trend:

Likewise, to detect decreasing trend, we compute $D_t = \max\{0, D_{t-1} + (mu - s_t - C)\}$ for each time interval.

In each case, we alert the user when I_t or D_t values exceed their respective thresholds.

User can have control over the sensitivity and threshold parameters so that he can fine tune the system to suit his preferences with respect to when he would want to be alerted.

Recommending personalized life style changes to keep sugar level under control

At periodic intervals, say once in a week or so, we could use the data collected for each user during the past week to fit a regression model on their data and study the sugar levels in relation to various aspects of life style such as sleep patterns, diet patterns, physical activity etc. As before, we would have to fit multiple models to study the relation between various activities and their effect on sugar level at various time gaps. Also, like before, we could use PCA to transform the data into independent components before fitting the model. If PCA is used, we would have to translate the factors in the model back to original factors using linear transformations to get the weights of the original factors.

We would have to average out the weights of coefficients from all the models built for the user. Coefficients of each aspect would then be an indicator of how much bearing that particular aspect has on sugar level. For example, if low sugar levels were consistently preceded by a period of low rest, the weights of sleep related factors would reflect that; that coefficient would be high. We could then report this to the patient. With the insight gained, the patient would then know that if he/she gets better sleep, he/she would have better control over sugar levels.

References

https://machinelearningmastery.com/multi-step-time-series-forecasting/