Author’s response to reviews

Title: Predictive Models for Diabetes Mellitus Using Machine Learning Techniques

Authors:

Hang Lai (hang68@mathstat.yorku.ca)
Huaxiong Huang (hhuang@mathstat.yorku.ca)
Karim Keshavjee (karim.keshavjee@ryerson.ca)
Aziz Guergachi (a2guerga@ryerson.ca)
Xin Gao (xingao@mathstat.yorku.ca)

Version: 1 Date: 15 May 2019

Author’s response to reviews:

Authors’ Response to Reviewers

Title: Machine Learning Techniques Outperform Current Diabetes Mellitus Prediction Rules

Authors’ Response to Reviewers:

Response to Reviewers’ comments
We would like to thank the Reviewers for their constructive comments and suggestions to improve the manuscript. Please see below our response to the reviewers’ comments.

Reviewer #1:
1. Four different machine learning methods were compared on Diabetes Mellitus detection data consisting of records from Canadian patients. Regarding AROC criterion computed using hold-out error estimate, they found that Gradient Boosting and Logistic Regression outperformed random forest and Rpart model. Authors found that their results were comparable to Wilson et. al., Mashayekhi et al. and some scoring systems without using a parental history as predictor. Such a results does not sound convincingly. I do not want to force the authors to obtain 90% AROC, but would like to ask them to point out, what is their added value and why such a result is success and should be published.

RESPONSE:
We used both current machine learning techniques namely GBM, Random Forest, Decision Tree, and the classical statistical Logistic Regression model to predict Diabetes Mellitus and compared the four models based on the results we obtained. This could give physicians different alternative methods to make an informed decision on the probability that a patient has Diabetes Mellitus.
Parental history is shown to be a significant factor in detecting Diabetes Mellitus in Wilson et.al. Without using this important factor, our AROC is comparable to Wilson et. al.’s result. Our results can be used to answer the question on what is the ability to detect Diabetes Mellitus when parental history is not available in the sample data.

Also, instead of using the scoring systems, our algorithm in Logistic Regression model can give direct calculation of the probability of having Diabetes Mellitus. This can tell a patient directly the risk he or she has Diabetes Mellitus.

We also used misclassification cost method to improve the sensitivity of the models. When a DM patient is correctly detected, this will help reduce greatly the cost of medical care.

This research is an additional scientific research conducted on Canadian sample data to detect DM.

Therefore, these points above are our added values.

2. It is not clear if the random forest and decision tree were also pre-tuned using cross-validation or if some default parameters were used. For GBM, at least four parameters were tuned (n.trees, interaction.depth, n.minobsinnode, shrinkage). It is nor described in the article, if the parameters of the other methods were also tuned. I expect that no. This can lead to invalid conclusion that GBM is better method. Authors must clearly describe the tuning process for all models or they have to perform tuning for the other methods as well. Both the decision trees and random forests are highly sensitive on the parameters settings.

RESPONSE:
We thank reviewer #1 for this comment. We have added the description of the tuning process to the Additional File 3.

Here is our tuning process for each model.

The training dataset as described in the manuscript was used in the tuning process.

1. GBM model:
The parameter search space: We create a search space by defining a parameter space for hyperparameters as follows: the number of trees (n.trees) is an integer ranging from 200 to 600; the depth of tree (interaction.depth) is from 2 to 6; the minimum number of observations in the terminal nodes (n.minobsinnode) is from 30 to 80; and learning rate (shrinkage) is from 0.01 to 0.3.
Tuning method: We perform random search on the parameter space with the number of iterations of 100.
Evaluation method: We use 10-fold cross validation and use AROC as the performance measure. That is, for each iteration, a 10-fold cross validation is conducted and the average AROC is reported. Then from the 100 iterations, the model with the highest average AROC is chosen.

The tuning results are as follows: the number of iterations (n.trees) is 257; the interaction depth (interaction.depth) is 2; the minimum number of observations in the terminal nodes (n.minobsinnode) is 75; the shrinkage rate (shrinkage) is 0.126.
2. Logistic Regression model:
There are no hyperparameters for the Logistic Regression models so we do not use the tuning process for Logistic Regression models.

3. Random Forest model:
Parameter search space: the number of trees to grow (ntree) is from 80 to 500; the number of variables should be selected at a node split (mtry) is an integer ranging from 3 to 6; the number of observations at terminal nodes (nodesize) is from 20 to 50.
Tuning method: We perform random search on the parameter space with the number of iterations of 100.
Evaluation method: We use 10-fold cross validation and use AROC as the performance measure. Then from the 100 iterations, the model with the highest average AROC is chosen.

The tuning results are as follows: the number of trees to grow (ntree) is 407; the number of variables should be selected at a node split (mtry) is 3; the number of observations at terminal nodes (nodesize) is 22.

4. Decision Tree model:
Parameter search space: the smallest number of observations in the parent node that could be split further (minsplit) is from 30 to 50; the smallest number of observations that are allowed in a terminal node (minbucket) is an integer ranging from 10 to 50; depth of tree (maxdepth) can be 8, 12, or 16; the complexity parameter (cp) is from 0.001 to 0.2.
Tuning method: We create a grid for the parameter space with resolution of 10.
Evaluation method: We use 10-fold cross validation and use AROC as the performance measure. Then from the 3000 iterations, the model with the highest average AROC is chosen.

The tuning results are as follows: The smallest number of observations in the parent node that could be split further (minsplit) is 41; the smallest number of observations that are allowed in a terminal node (minbucket) is 19; the depth of tree (maxdepth) is 8; the complexity parameter (cp) is from 0.001.

Since GBM, Random Forest, and Decision Tree each has its own set of hyperparameters, we also tuned the hyperparameters of Random Forest and Decision Tree. The hyperparameters of these models were tuned on the same training dataset, and the AROC values for the four models were computed based on the testing dataset which is not included in the model training process. Therefore, our results are valid.

We agree with reviewer #1 that the decision trees and random forests are highly sensitive on the parameters settings. Therefore, we validate the performance of the tuned models using cross validation method.

5. Quality of figures is very poor, authors should use EPS format for images.
RESPONSE: We thank reviewer #1 for this suggestion and have created the pictures using TIFF format.
Reviewer #2:
1. This is a well written paper describing development of a diabetes mellitus prediction model. Mostly the methods are appropriate but there seem to be some significant oversights that render their conclusions unsupported.

The authors claim that the GBM model is superior to the logistic regression model and the random forest model based on a nominally larger AUC, but they do not account for uncertainty in the estimates and do not provide any confidence intervals. I am not convinced that the GBM model is truly superior and not acknowledging the uncertainty of their estimates is a major oversight. They also claim machine learning methods outperform current prediction rules, but did not compare their models to the current prediction rules they cite (because they do not have sufficient data).

RESPONSE:
We thank reviewer 2 for this comment. We have done additional cross validations and added to the manuscript (line 223 to line 264, the Results section).

The process is as follows:

We do 10-fold cross validation on the whole dataset with the following steps:

1. Divide dataset into 10 parts.
2. Use 9 parts as training dataset; apply all 4 models.
3. Measure AROC for each model based on the last part (the testing dataset).
4. Repeat for all 10 folds

Shuffle the whole dataset and repeat the above procedure 2 more times.

Based on 30 values of AROC obtained for each model (age is treated as a continuous variable), we are able to estimate the consistency of their predictive power for each model by finding the mean, the standard deviation, and the 95% confidence interval of their AROC values.

Mean and Standard Deviation of AROC for the four models from the cross-validation results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM</td>
<td>83.9%</td>
<td>1.27%</td>
<td>(83.4, 84.4)</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>83.5%</td>
<td>1.22%</td>
<td>(83.1, 83.9)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>83.0%</td>
<td>1.22%</td>
<td>(82.6, 83.4)</td>
</tr>
<tr>
<td>Rpart</td>
<td>77.1%</td>
<td>3.21%</td>
<td>(75.9, 78.3)</td>
</tr>
</tbody>
</table>
The standard deviations are small and the confidence intervals are not wide. This indicates that the values of AROC of the four models are consistent.

We created a box plot to compare the AROC values of the four models. The new plot is added to our revised version.

Figure 3. Box plot: Comparing the AROC of the four models in the cross-validation results. The box plot shows that the medians of AROC values for GBM, Logistic Regression and Random Forest are quite close to each other and they are all greater than that of the Rpart model. We conducted a paired t-test for every two models. The hypothesis we wanted to test is whether or not the mean AROC values are the same for every two models. We used the paired t-tests with Bonferroni adjusted Type I error rate and found that all the tests are very statistically significant. We also conducted one-sided t-tests to test whether or not one model performs better than the other and we obtained the following results.

Paired one-sided t-tests to compare the AROC means of the four models.

<table>
<thead>
<tr>
<th>Test Name</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM vs. Logistic Regression</td>
<td>6.289</td>
<td>3.617e-07</td>
</tr>
<tr>
<td>GBM vs. Random Forest</td>
<td>8.734</td>
<td>6.482e-10</td>
</tr>
<tr>
<td>GBM vs. Rpart</td>
<td>12.450</td>
<td>1.842e-13</td>
</tr>
<tr>
<td>Logistic Regression vs. Random Forest</td>
<td>4.452</td>
<td>5.799e-05</td>
</tr>
<tr>
<td>Logistic Regression vs. Rpart</td>
<td>11.663</td>
<td>9.02e-13</td>
</tr>
<tr>
<td>Random Forest vs. Rpart</td>
<td>10.219</td>
<td>2.013e-11</td>
</tr>
</tbody>
</table>

These results show that the mean AROC of GBM model is statistically significantly greater than that of Logistic Regression, Random Forest, and Rpart models. The Logistic Regression model also has a mean AROC greater than that of Random Forest and of Rpart. The mean AROC of Random Forest model is statistically significantly greater than that of Rpart model, as well. We also noted that although all the tests are very statistically significant, the mean differences of AROC among GBM, Logistic Regression, and Random Forest are not very large and are not practically significant.

2. Additional issues:

a) Excluding patients who are on insulin seems unjustified as many patients with type 2 diabetes receive insulin treatment. The authors should explore other methods for identifying individuals with type 1 diabetes if they need to remove them.

RESPONSE:
Identification of patients with Type 1 diabetes is an on-going area of research in this database. At the time of this study, it was not possible to definitively separate those with Type 1 from those with Type 2. Given that Type 1 diabetes is not predictable, and given the large sample of patients not on insulin, we felt it was a reasonable step to remove those with insulin from the sample --especially since insulin might have been used to make the diagnosis of diabetes in the database (please see the reference below).
b) The authors do not describe how the diagnosis of diabetes was made

RESPONSE:
Case-finding for diabetes is described in:


We also cited this reference under the Reference section.

c) The authors do not describe the time period that proceeded the predictor data used and the diagnosis of diabetes

RESPONSE:
The last data point included in the dataset was 1 year prior to diagnosis of diabetes.

3. The authors mention hba1c in the formulation of their dataset but then do not use it as a predictor, which seems unusual given that it is a strong predictor of diabetes. If the reason is that many individuals did not have this measured, why include it in the dataset in the first place?

RESPONSE: We thank reviewer 2 for this suggestion. Since there is 68.4% of missing values for this variable, we did not use this variable as a predictor in our models. We would consider this factor in our future research if the data are available.

4. Why did the authors treat age as a categorical variable only for the logistic regression model but not for the GBM model?

RESPONSE:
We tried both Age as continuous variable and as a categorical variable and obtained very similar AROC results as the followings.

<table>
<thead>
<tr>
<th>Model</th>
<th>AROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM</td>
<td></td>
</tr>
<tr>
<td>LOGISTIC REGRESSION</td>
<td>84.7% 84.1% 83.4% 78.2%</td>
</tr>
<tr>
<td>RANDOM FOREST</td>
<td></td>
</tr>
<tr>
<td>RPART</td>
<td></td>
</tr>
<tr>
<td>Age Continuous</td>
<td>84.3% 84.0% 83.3% 78.2%</td>
</tr>
<tr>
<td>Age Category</td>
<td>84.3% 84.0% 83.3% 78.2%</td>
</tr>
</tbody>
</table>
However, when we checked the Logistic Regression model at first we found that the coefficient of Age is negative although this term is very significant. We refitted the model using the quadratic term for Age which is \((I(age ^2))\) and we found that this term was statistically significant and the coefficient of Age becomes positive. It is more difficult for model interpretation when the quadratic term is in the model. On the other hand, when Age is treated as a categorical variable, it helps with model interpretation as demonstrated in our Results section. For this reason, we treated age as a categorical variable only for the logistic regression model but not for the GBM model.

Reviewer #3:

1. This study is conducted on DM dataset curated from restricted source and conclusions drawn may be valid for used dataset only. Only on the basis of reported results it is very difficult to generalize the findings on other DM datasets.

However, many ML models performed better than the proposed work using same evaluation measures on different DM datasets. Therefore, direct comparison will be more convincing with other datasets (publicly available) with existing contributions using same ML models.

RESPONSE: We thank reviewer #3 for these suggestions. We have used our methods to work on the PIMA Indians dataset which is a publicly available dataset. We also added the analysis of this dataset to the manuscript (line 266 to line 303, the Results section).

Here is the added work:

To see how our methods work on a different dataset, we used Pima Indians Dataset which is a publicly available dataset. All patients in this dataset are females at least 21 years old of Pima Indian heritage. There are 768 observations with 9 variables as followings:
- Pregnant, number of times pregnant;
- Glucose, plasma glucose concentration (glucose tolerance test);
- BP, diastolic blood pressure (mm/Hg);
- Thickness (triceps skin fold thickness (mm));
- Insulin (2-Hour serum insulin (mu U/ml));
- BMI (body mass index (weight in kg/(height in m) squared));
- Pedigree (diabetes pedigree function);
- Age (Age of the patients in years);
- Diabetes (binary variable with 1 for Diabetes and 0 for No Diabetes).


We applied our methods on this dataset to predict the probability a patient has diabetes. When working on this dataset, we noticed that there are many rows with missing data and the missing values in Glucose, BP, Thickness, and BMI are labeled as 0. For example, about 48.7% of Insulin values are missing.

For purpose of validating our methods, we chose not to impute the data but excluded all rows with missing values. There are 392 observations left in the working dataset in which 130 patients with diabetes and 262 without diabetes. We applied our methods on this dataset to predict whether or not a patient has diabetes. We also divided the PIMA dataset into the training dataset (80% of the observations) and the testing dataset (20% of the observations). We trained the four models on the training dataset and validate the models on the testing dataset. On the testing dataset, we obtained the
AROC of 84.7% for GBM model, 88.0% for Logistic Regression Model, 87.1% for Random Forest Model, and 77.0% for Rpart model. We also conducted 10-fold cross-validation and repeated the procedure for two more times.

Here are our results based on the 30 AROC values of the cross-validation results conducted on the PIMA Indian dataset.

Table 6. Comparing the AROC values of the four models using PIMA Indian dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM</td>
<td>85.1% (82.7, 87.5)</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>84.6% (82.2, 87.0)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>85.5% (83.1, 87.9)</td>
</tr>
<tr>
<td>Rpart</td>
<td>80.5% (74.9, 84.9)</td>
</tr>
</tbody>
</table>

The results we obtained for this dataset are quite consistent with what we observed in our main dataset. Based on these results, GBM, Logistic Regression, and Random Forest are comparable and they all give higher mean AROC than that of the Rpart model on the testing dataset. We also created a box plot to compare the sampling distributions of the AROC values for the four models. This new plot is added to our revised manuscript as Figure 4. The box plot shows that GBM, Logistic Regression, and Random Forest are comparable and they all give higher mean AROC than that of the Rpart model.

2. Most of the cited reference are old. Therefore, I recommend to include some more references and cite where necessary specially introduction section related work. Such as


(Optional) There are possibilities to include some critical review of recent development in this field.

RESPONSE: We thank reviewer #3 for providing these new published studies in this field. We have read and included some review of recent development of this topic (line 55, Background section; Reference section).
3. (Optional)-Perform some statistical significance test to validate the results.
RESPONSE:
We thank reviewer #3 for this insightful suggestion. We have done additional cross validations as presented in our response to reviewer #2. Based on the results obtained, we are able to conduct statistical significance tests to validate the results. We have added this part in the manuscript (line 243 to line 258, the Results section).

4. It is already mentioned in the background that Logistic Regression has achieved an AROC of 85% by Wilson et.al. while the author has achieved an AROC 84% using LR and GBM which is lower than the previous experiment.
RESPONSE:
As mentioned in the discussion section, parental history was used in Wilson et al. models. Mashayekhi et. al. evaluated Wilson’s clinical model on Canadian sample data without parental history as a predictor and obtained an AROC of 78.6%. We also noted that the characteristics of the sample data used in this study are more diversified than the one used by Wilson et al. For example, the age of the patients in our dataset ranges from 18 to 90, while the patients studied by Wilson et. al. ranges from 45 to 64.

5. The model diagram may help better in understanding the experimental procedure.
RESPONSE: We thank reviewer #3 for this suggestion. We have included a diagram for our experimental procedure to the Additional File 3.