Pima Indians Diabetes Prediction Using Logistic GAM

Yahui Peng

Department of Management Science and Statistics, College of Business, UTSA

Introduction

Diabetes is a major public health problem worldwide, and the Pima Indians of Arizona and Mexico are high-risk populations of that[1]. Their committed participation in diabetes research has contributed numerous data for scientists to study. In this project, the objective is to predict the onset of diabetes in Pima Indians with a model that is both flexible and interpretable. With a binary response variable, we naturally consider logistic regression. Additionally, since the real world is not always linear, we add more flexibility to it with the generative additive model (GAM) technique. Hence, logistic GAM is applied to the data.

Data Structure

The dataset is from the National Institute of Diabetes and Digestive and Kidney Disease[2]. Response variable *Outcome* is binary ($Y = 1$ for diabetes, $Y = 0$ for no diabetes). Eight biometrical predictors are included in the dataset.

Initial investigation is made into the predictors. By looking at the boxplots (**Fig. 1**), we find most observations have pregnancy history and are aged over 21, indicating the data is mainly from Pima Indian women aged over 21. Additionally, we observe unusual zero values in *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, and *BMI*, which are scientifically invalid. Even if these abnormal values are excluded, outliers are still observed for all predictors except *Glucose*.

Fig. 1 Boxplots of Predictors

In **Fig. 2**, the pairwise scatter plots (lower panel) are colored in red with *Outcome* $Y = 1$ and in dark grey otherwise. A likely trend for onset of diabetes is observed with a high *Glucose* level, a large *BMI*, or a large *Age*. Histograms (diagonal panel) suggest possible high skewness for *Pregrancies*, *SkinThickness*, *Insulin*, *DiabetesPedigreeFunction*, and *Age.* The correlation coefficients shown in the correlation matrix (upper panel) do not exceed 0.7 for all paired predictors, indicating no high correlation presents and no need to drop any predictor prior to further actions.

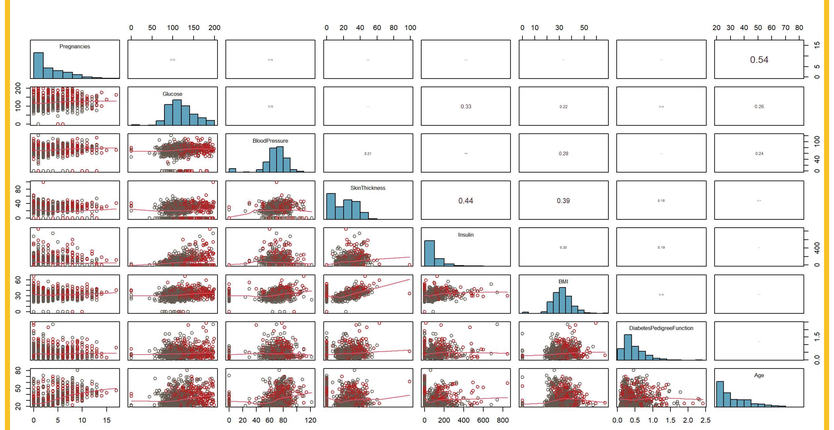


Fig. 2 Pairwise Scatterplots, Higtograms, and Correlation Matrix of Predictors

The data have issues in invalid values, outliers, skewness, and mentioned that logistic regression is sensitive to outliers. Therefore, data preprocessing is necessary before fitting a model to the training data.

Data Preprocessing

Missing Data Encoding and Imputation

The invalid zero values *in Glucose, BloodPressure, SkinThickess, Insulin, and BMI* are encoded as missing. Next, in an aggregation plot (**Fig. 3**), the left panel tells us nearly 50% (> 40%) of the samples are missing *Insulin*, and 30% are missing *SkinThickness*. Considering we only have eight predictors, we decide to keep *Insulin* in the data. The right panel helps us understand that almost 51% of the samples are complete, 25% are missing the *Insulin* and *SkinThickness* value, 18% are missing only *Insulin* value, and the remaining ones show other missing patterns. We conclude that the missing values are missing completely at random, and imputation is necessary.

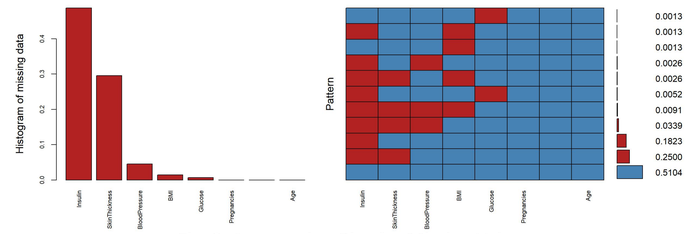


Fig. 3 Aggregation Plot for Missing Values

The *mice()* function is used to impute using predictive mean matching (pmm) method. After imputation, we draw a scatterplot (**Fig. 4** left) of *Glucose* against all the other imputed variables comparing the distributions of original (blue) and imputed (red) data. The shape of the imputed points matches the shape of the original. Also, the density of the imputed data for each imputed variable is shown in **Fig. 4** right panel. The distributions of imputed data (red) are similar to the original (blue), suggesting the imputed values are indeed plausible.

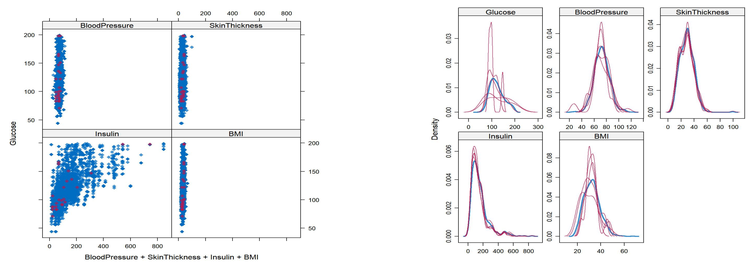


Fig. 4 Scatterplot (left) and Density Plot (right) of Imputed and Original Data

Data Splitting

Split the dataset by a ratio of 80/20 for training and testing data, accordingly.

Skewness and Outliers Resolving

Skewness check indicates that *Insulin*, DiabetesPedigreeFunction, and *Age* are highly skewed with absolute skewness values greater than 1. Skewness issue is resolved after BoxCox transformation (**Fig. 5** left panel) and outliers are gone after capping with 5% (for low outliers) or 95% (for high outliers) quantile value (**Fig. 5** right panel).

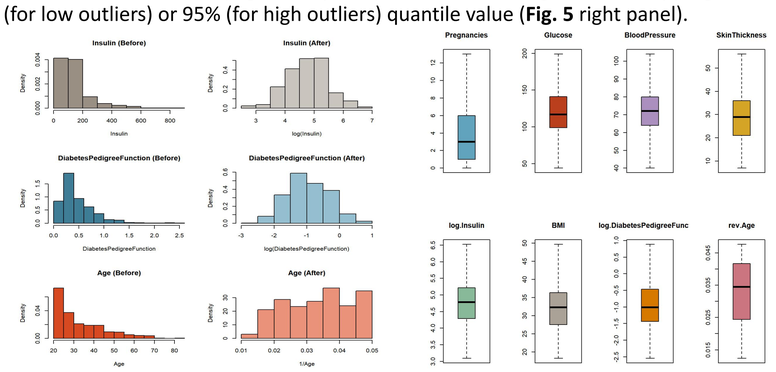


Fig. 5 Histograms of Skewed Predictors (left) and Boxplots After Capping Outliers (right)

Method

In logistic GAM, a logit link is used with a binomial error distribution, and GAM is added to capture the impact (can be nonlinear) of the predictive variables through smooth functions. We first probe a logit GAM model fitting spline functions for all predictors:

$$\text{logit}(P) = \beta_0 + \beta_1 s(Pregnancies) + \beta_2 s(Glucose) + \beta_3 s(BloodPressure) + \beta_4 s(SkinThickness) + \beta_5 s(\log(Insulin)) + \beta_6 s(BMI) + \beta_7 s(\log(DiabetesPedigreeFunction)) + \beta_8 s(Age^{-1}) + \varepsilon$$

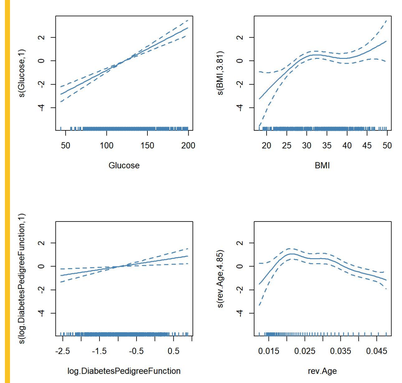
Then, a refined model is obtained by keeping only significant spline terms at 0.05 level as well as replacing the significant terms which have an estimated degrees of freedom (edf) of 1 to a linear term.

Results

The final model selected is

$$\text{logit}(P) = \beta_0 + \beta_1 Glucose + \beta_2 s(BMI) + \beta_3 \log(DiabetesPedigreeFunction) + \beta_4 s(Age^{-1}) + \varepsilon$$

The component smooth function plots (**Fig. 6**) provide us with visual aid in understanding the impact pattern of significant predictors. The logit of probability of onset of diabetes (logit(P)) increases linearly with increased *Glucose* levels or log(*DiabetesPedigreeFunction*) score, with other predictors held constant. logit(P) also shows an increasing trend for increased *BMI* except for a middle *BIM* range from about 32 to 40. *Age* has a more complicated impact on logit(P) with a more wiggly smooth function plot. Overall, logit(P) increases with increased *Age* from 21-50 and decreases dramatically with increased *Age* over 50. In basis dimension (k) checking, all p-value of spline functions greater than 0.05 indicates the basis dimension is statistically significant for each predictor.




```
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9869     0.1252  -7.884 3.17e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Approximate significance of smooth terms:
                                edf Ref.df Chi.sq  p-value
s(Glucose)                    1.000  1.001 78.490  < 2e-16
s(BMI)                        3.808  4.752 24.997 0.000142
s(log.DiabetesPedigreeFunction) 1.000  1.000  7.751 0.005373
s(rev.Age)                    4.848  5.877 41.782  8.5e-07

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                                k'  edf k-index p-value
s(Glucose)                    9.00 1.00    1.02    0.75
s(BMI)                        9.00 3.81    0.99    0.43
s(log.DiabetesPedigreeFunction) 9.00 1.00    0.96    0.17
s(rev.Age)                    9.00 4.85    0.98    0.32
```

Fig. 6 Plots of Component Smooth Functions in the Logistic GAM Model

When checking diagnostics, the interpretation of conventional residuals for GLM models can be problematic. Here we use the *simulateResiduals*() function to calculate scaled residuals[3]. Then, by using *plot*() function, we obtain two diagnostic plots (**Fig. 7**). We conclude from the plots that normality assumption holds and no significant deviation is detected. Therefore, the final logistic GAM model is valid.

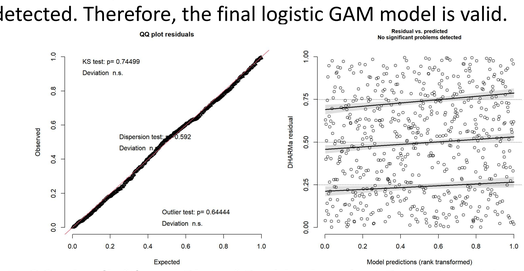
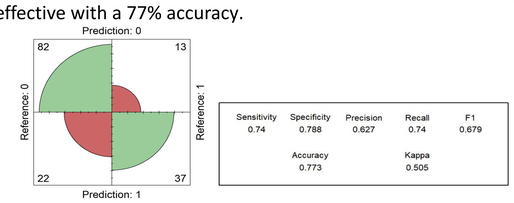


Fig. 7 QQ-plot (left) and Plot of Residual against Predicted Values (right)

We fit the model to test data, and the confusion matrix below shows the proposed logistic GAM model is effective with a 77% accuracy.



Conclusion

A 77% accurate logistic GAM model with good interpretability is built to help predict the onset of diabetes for Pima Indians. Overall, glucose level, age, BMI and diabetes pedigree function score have a significant impact on the outcome. Caution should be taken by Pima Indians who have high glucose levels, large body mass index, high diabetes pedigree function scores, or age near 50. Improvement can be made by refining methods for imputing missing values and handling outliers. Moreover, resampling methods should be used for more reliable model performance metrics.

References

1. High-Risk Populations: The Pimas of Arizona and Mexico
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4418458/
2. Pima Indians Diabetes Database
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
3. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models
https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html