# Additive Logistic Regression to Predict Diabetes in a Homogeneous Population

Edbert Jao

June 2, 2024

**Contents**



## 1 Introduction and Research Question

**Research Question: For the dataset described in the following, can an additive logistic regression be fit on the data that predicts diabetes better than guessing randomly?**

Collected originally by the National Institute of Diabetes and Digestive and Kidney Diseases, the Pima Indians Diabetes dataset measures the presence of diabetes in 768 female Pima Indians of at least 21 years of age. Maintained by the UC Irvine Machine Learning repository[1], this dataset is widely used as a benchmark in diabetes classification[2] and in benchmarking predictive machine learning algorithms.

The response variable is *Outcome* $\in \{0, 1\}$, which is a binary variable denoting whether an individual has diabetes.

There are 8 covariates:

- *Pregnancies*: Number of Times Pregnant
- *Glucose*: Plasma glucose concentration at 2 Hours in an oral glucose tolerance test (mg / Dl)
- *Blood Pressure*: Diastolic Blood Pressure (mm Hg)
- *Skin Thickness*: Triceps Skin Fold Thickness (mm)
- *Insulin*: 2-Hour Serum insulin (µh/ml)

- *BMI*: Body mass index [weight in kg/(Height in $m^2$)]
- *Diabetes Pedigree Function*: Likelihood of developing diabetes based on age and family history
- *Age*: Age (years)

Given the binary nature of *Outcome*, a logistic regression would allow for predictions, classifying based on whether

$$\mathbb{P}\left[\,Outcome = 1\right] > 0.5$$

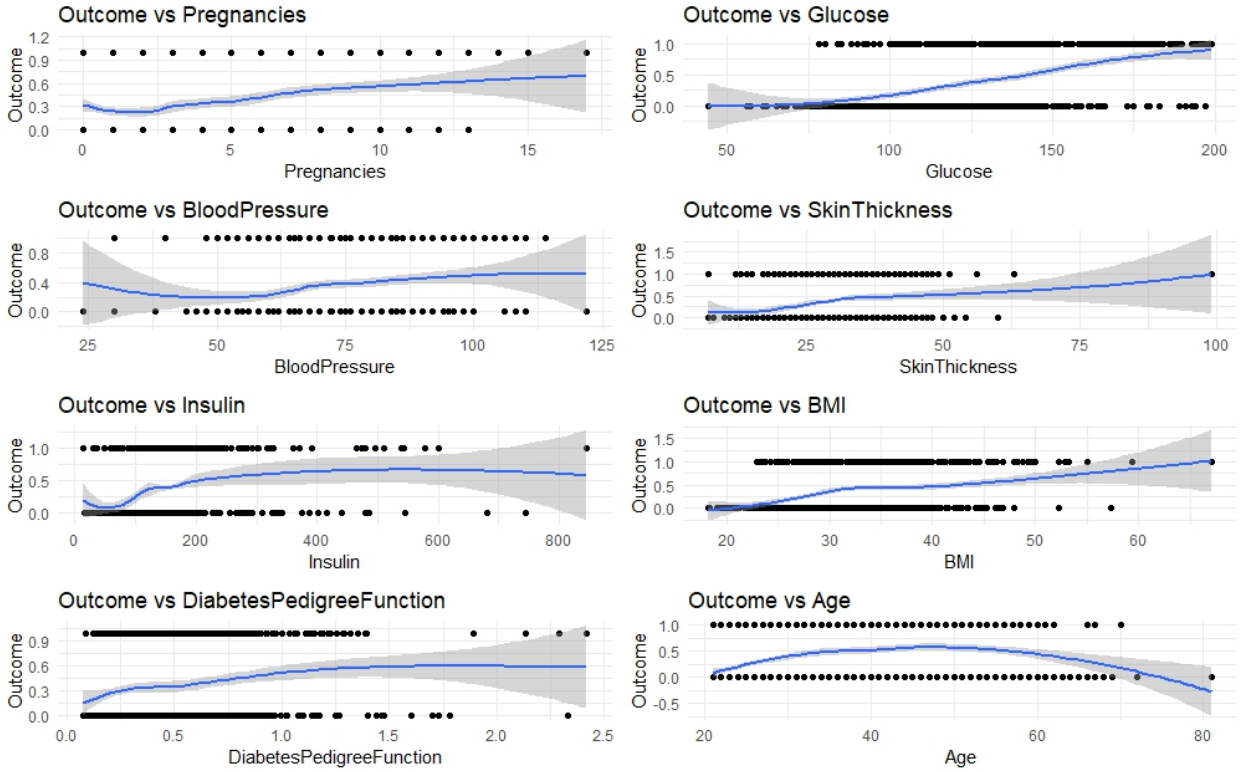
## 2 Data Cleaning and Exploratory Data Analysis

The dataset has 768 observations.

Some of the observations for the covariates *Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI* have a value of 0, which is biologically impossible. These observations are treated as N/A and are replaced with mean imputation.

```
[1] "Number of observations of Glucose with a value of 0: 5"
[1] "Number of observations of BloodPressure with a value of 0: 35"
[1] "Number of observations of SkinThickness with a value of 0: 227"
[1] "Number of observations of Insulin with a value of 0: 374"
[1] "Number of observations of BMI with a value of 0: 11"
```

Using R's built-in locally estimated scatterplot smoothing (i.e. local polynomial regression) the data can be inspected to see which covariates might have a non-linear relationship with *Outcome*.




*Blood Pressure, Insulin,* and *Age* appear to be non-linearly related with *Outcome*, although for *Age* this may be the result of the outlier with the greatest age. Because of this non-linearity which was non-rigorously

identified by visual inspection, there may be some merit to addittively modelling *Outcome* versus its covariates.

## 3 Model Specification

4 models are tested. Models 2, 3, and 4 are additive logistic regressions whereas Model 1 is a parametric logistic regression. The inclusion of Model 1 is meant to provide a parametric "benchmark" to compare the nonparametric Models 2, 3, and 4 against.
Below, "~" refers to logistic regression.

Model 1: *Outcome* is regressed on global order 3 polynomials of all covariates.

$$\begin{aligned} Outcome \sim & Pregnancies + Pregnancies^2 + Pregnancies^3 + Glucose + Glucose^2 + Glucose^3 + \\ & BloodPressure + BloodPressure^2 + BloodPressure^3 + \\ & SkinThickness + SkinThickness^2 + SkinThickness^3 + \\ & Insulin + Insulin^2 + Insulin^3 + BMI + BMI^2 + BMI^3 + \\ & DiabetesPedigreeFunction + DiabetesPedigreeFunctin^2 + DiabetesPedigreeFunction^3 \end{aligned}$$

Model 2: *Outcome* is regressed on local polynomials up to degree 2 of the covariates. Here, Local$(\cdot)$ refers to local polynomials of a variable.

$$\begin{aligned} Outcome \sim & \text{Local}(Pregnancies) + \text{Local}(Glucose) + \text{Local}(BloodPressure) + \text{Local}(SkinThickness) + \\ & \text{Local}(Insulin) + \text{Local}(BMI) + \text{Local}(DiabetesPedigreeFunction) \end{aligned}$$

Model 3: *Outcome* is regressed on basis splines up to degree 3 of the covariates. Here, BS$(\cdot)$ refers to basis splines of a variable.

$$\begin{aligned} Outcome \sim & \text{BS}(Pregnancies) + \text{BS}(Glucose) + \text{BS}(BloodPressure) + \text{BS}(SkinThickness) + \\ & \text{BS}(Insulin) + \text{BS}(BMI) + \text{BS}(DiabetesPedigreeFunction) \end{aligned}$$

Model 4: *Outcome* is regressed on natural cubic splines. Here, NS$(\cdot)$ refers to natural cubic splines of a variable.

$$\begin{aligned} Outcome \sim & \text{NS}(Pregnancies) + \text{NS}(Glucose) + \text{NS}(BloodPressure) + \text{NS}(SkinThickness) + \\ & \text{NS}(Insulin) + \text{NS}(BMI) + \text{NS}(DiabetesPedigreeFunction) \end{aligned}$$

The dataset is randomly divided into a training dataset and a testing dataset with a size ratio of 70:30. The training dataset has 537 observations and the testing dataset has 231 observations. Each of the 4 models above will be trained on the training dataset and it's predictive performance will be evaluated on the testing dataset with respect to log-loss.

The configuration of Local$(\cdot)$, BS$(\cdot)$, and NS$(\cdot)$ in Models 2, 3, and 4 align with the default settings of base R and `library(gam)`, which will be used for implementation.

`library(gam)` [3] employs the local scoring backfitting algorithm to fit component functions. This method is referred to in Elements of Statistical Learning Algorithm 9.2 [4]:

**Algorithm 9.2** Local Scoring Algorithm for the Additive Logistic Regression Model

1: Compute starting values: $\hat{\alpha} = \log\left(\frac{\bar{y}}{1-\bar{y}}\right)$, where $\bar{y} = \text{ave}(y_i)$, the sample proportion of ones, and set $\hat{f}_j \equiv 0 \ \forall j$.

2: Define $\hat{\eta}_i = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij})$ and $\hat{p}_i = \frac{1}{1+\exp(-\hat{\eta}_i)}$. Iterate:

(a) Construct the working target variable

$$z_i = \hat{\eta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}.$$

(b) Construct weights $w_i = \hat{p}_i(1-\hat{p}_i)$.

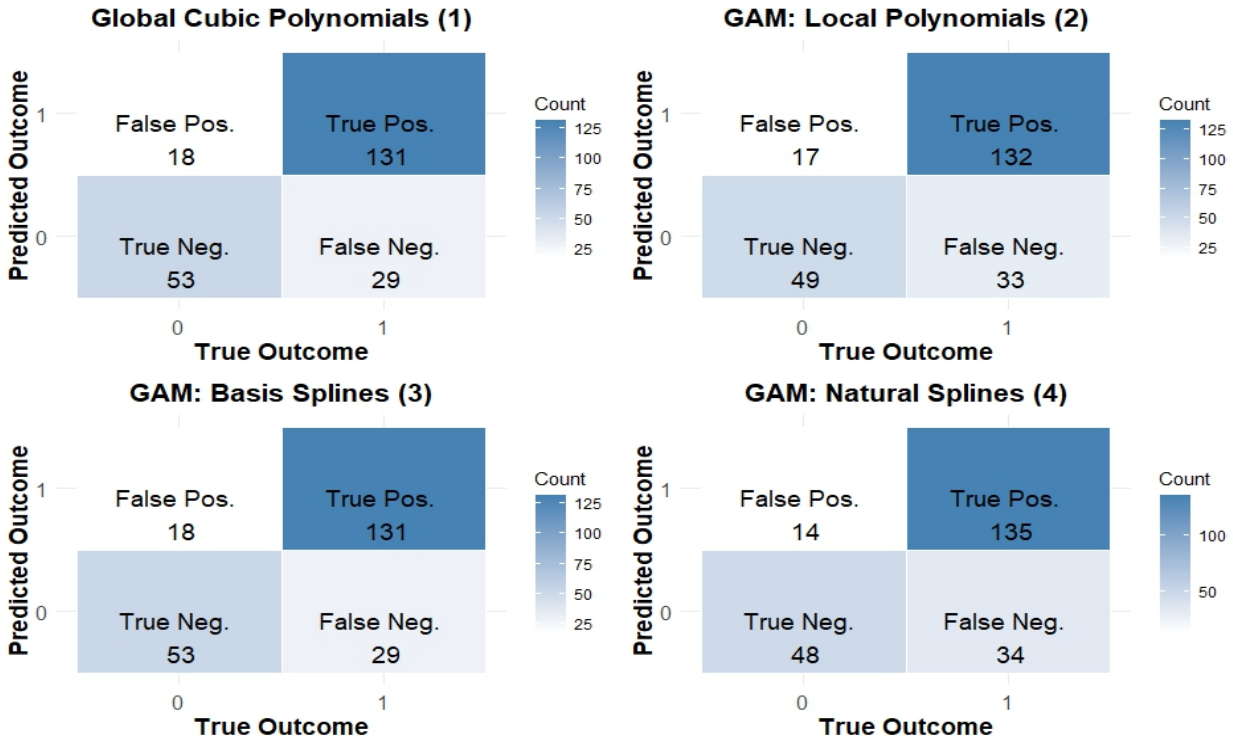
(c) Fit an additive model to the targets $z_i$ with weights $w_i$, using a weighted backfitting algorithm. This gives new estimates $\hat{\alpha}, \hat{f}_j, \ \forall j$.

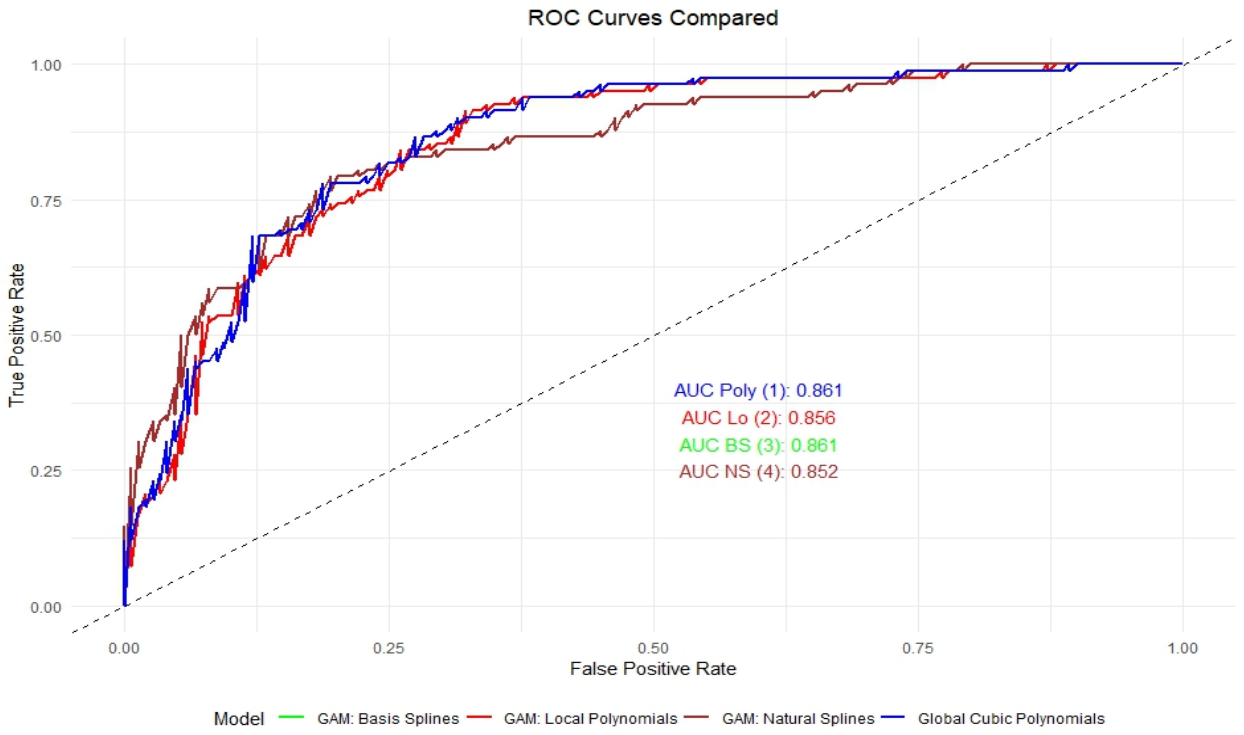
3: Continue step 2. until the change in the functions falls below a pre-specified threshold.

## 4 Results and Model Comparison

Here, the performances of the models are visualized with confusion matrices and receiver operating characteristic curves.

Fitting Model 1, Model 2, Model 3, and Model 4 in R with `library(gam)`, the models perform as follows:






Note: AUC refers to "Area under Curve"

When inspecting the predictions of the models, it is found that Model 1 and Model 3 performed identically. (This is why they perfectly overlap on the ROC plot.)

Model 1 and Model 3 have the lowest False Negative rate (Type II error rate), which would be the more costly to a patient than a False Positive (Type I error).

Furthermore, in overall prediction as well, Model 1 and Model 3 had the lowest mean log-loss.

| **Model** <chr> | **LogLoss** <dbl> |
|---|---|
| Global Cubic Polynomials (1) | 0.4481480 |
| GAM: Local Polynomials (2) | 0.4560115 |
| GAM: Basis Splines (3) | 0.4481480 |
| GAM: Natural Splines (4) | 0.4520909 |

Therefore, both from the perspectives of minimizing Type II errors and in overall predictive capability, Model 1 and Model 3 perform the best.

All things considered however, the models perform similarly and with moderate error. That is, the performance of the model is more informative than random guessing with a probability of $\frac{1}{2}$ (which would have a log-loss of $\approx 0.693$). In this sense, the answer to the research question is **"Yes"** but with room for improvement.

## 5 Closing Remarks

Given the identical performance of Model 1 and Model 3, the data in the Pima Indians Diabetes dataset appear to be either or both of

(a) sufficiently smooth, lacking extreme variance

(b) well-approximable by cubic models

Assuming that the dataset is representative of Pima Indians, it is possible to fit an additive logistic regression with for Pima Indians overall and perhaps other homogeneous populations. But it may be unnecessary to do so. The data appear "well-behaved" enough that parametric modelling is sufficient, and perhaps preferred if compute time is a serious concern.

## 6 References

### References


[1] UC Irvine Machine Learning Repository. Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[2] Chang, T., et al. (2022). Predictive Machine Learning Algorithms for Diabetes Classification. *Journal of Medical Research*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8943493/.

[3] Hastie, T. (2023). *Package 'gam'*: Generalized Additive Models. R package version 1.20. https://cran.r-project.org/web/packages/gam/index.html.

[4] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. https://hastie.su.domains/ElemStatLearn/.


Relevant code has been submitted separately.