Statistical Learning, Homework 1

Virginia Leombruni 247160

2024-03-30

Introduction

The data come from a study conducted in a UK hospital, investigating possible factors influencing pregnant women's decision to breastfeed their children. This homework aims to understand the impact that different variables have on women's choice to breastfeed, to help promote breastfeeding to women who are less likely to choose it. The study is based on 135 mothers-to-be. All the factors are two-level factors. The first listed level of each factor is used as the reference (and coded with 0).

The data included the following variables:

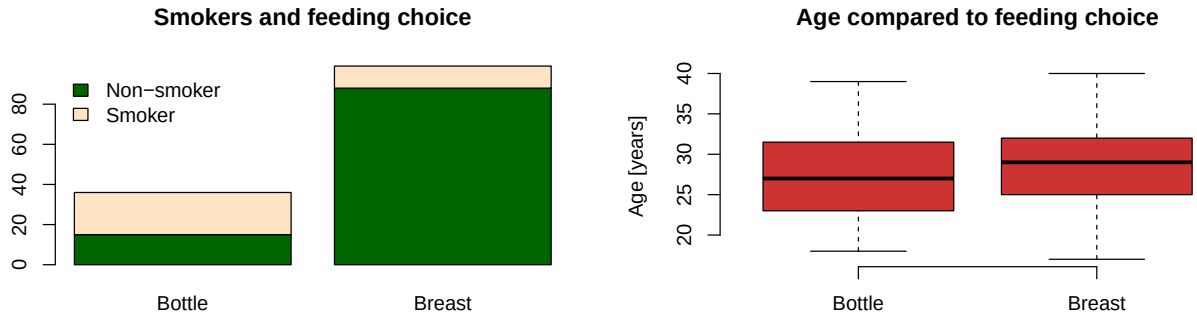
1. **breast** is the attribute to predict with the model. It is a nominal binary variable categorizing the will of bottle feeding only from other methods including partly or completely breast feeding;
2. **pregnancy**: concerns the course of pregnancy and it is coded as 'beginning' or 'end';
3. **howfed**: is the way mothers were fed as infants;
4. **partner**: is associated with the presence or absence of a partner;
5. **smokenow**: if they stopped smoking;
6. **smokebf**: if they have ever smoked;
7. **age**: the age of each woman;
8. **educat**: the age at which they left full-time education;
9. **ethnic**: the ethnic group they belong to.

Data exploration

The exploration of the data consists of visualising and evaluating them.

```
head(data)
typeof(data$breast)
summary(data$age)
table(data$breast)
sum(is.na(data))
data <- na.omit(data)
```

When exploring the data, no outliers are found, as 8 out of 10 variables are factors and the variables '*age*' (ranges between 17 and 40) and '*educat*' (ranges between 14 and 38) have plausible limits. The discriminatory power of each predictor can be visualised with a bar graph for the categorical columns and a boxplot for the numerical ones. All dichotomous binary variables were converted as factors of value 0 and 1. During the exploration of the data, a class imbalance was observed within the dataset of the response column, with 73% of women breastfeeding.




```
columns_to_convert <- c("breast", "pregnancy", "howfed", "howfedfr", "partner", "smokenow",
    "smokebf", "ethnic")
data <- data %>%
    mutate_at(vars(columns_to_convert), ~factor(., labels = c(0, 1)))
# View(data)
```

Data splitting

The model is evaluated by dividing the data into a training test and an evaluation test. 70% of the data were included in the training set due to the limited number of samples. During the division of the set, the proportion was considered, and the appropriate function was applied to maintain the proportion.

```
set.seed(22)
train_indexes <- caret::createDataPartition(data$breast, p = 0.7, list = FALSE)
train_data <- data[train_indexes, ]
test_data <- data[-train_indexes, ]
response_train <- train_data$breast
```

To clarify the choice of fit, two models are produced using the Generalised Linear Model (GML) and the K-nearest Neighbours (KNN) algorithms.

GLM

The GLM is a statistical generalised linear model used to discuss the results after the train data set has been established and preliminary analyses completed. In a GLM, a linear relationship is assumed between the expected value of the dependent variable and the independent variables. The objective of this statistical model is to estimate the coefficient that best describes the relationship between the dependent variable and the independent variables and usually minimises the sum of the squared residuals. The GLM

provides interpretable coefficients that quantify the direction and magnitude of the relationship between the response variable and the predictors.

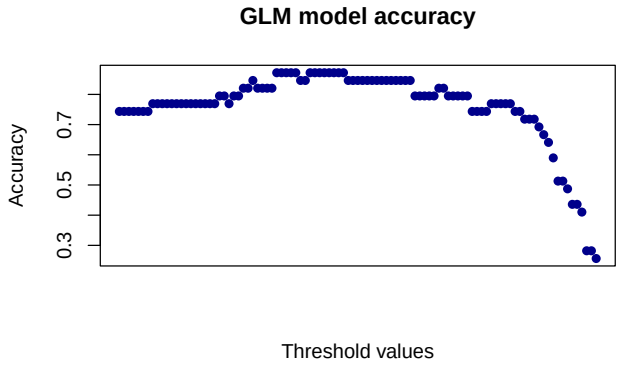
```
glm_fits <- glm(breast ~ ., data = train_data, family = "binomial")
glm_prob <- predict(glm_fits, newdata = test_data, type = "response")
```

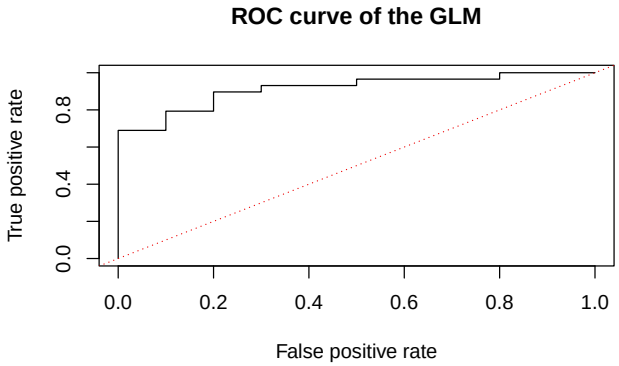
```
Call:
glm(formula = breast ~ ., family = "binomial", data = train_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.43006    2.88248  -0.843  0.39920
pregnancy1   1.28079    0.66763   1.918  0.05506 .
howfed1      0.40823    0.70921   0.576  0.56488
howfedfr1    1.07368    0.69506   1.545  0.12241
partner1    -1.07146    0.78689  -1.362  0.17332
smokenow1   -3.14973    1.08957  -2.891  0.00384 **
smokebf1     1.64068    1.06797   1.536  0.12447
age          0.03403    0.06425   0.530  0.59634
educat       0.15058    0.13290   1.133  0.25723
ethnic1     -1.93316    0.83339  -2.320  0.02036 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.144  on 95  degrees of freedom
Residual deviance:  69.926  on 86  degrees of freedom
AIC: 89.926

Number of Fisher Scoring iterations: 5
```

The function *summary()* provides a statistical or descriptive summary. The first column lists the coefficients of each predictor. Being at the end of pregnancy, having been breastfed as an infant and having a breastfeeding friend are positively correlated with a woman's choice to breastfeed. On the other hand, being single, smoking, or being of white ethnicity are factors that reduce the logarithm of odds of breastfeeding. These data suggest a higher reliability of current smoking habits and ethnicity than of having a partner. In contrast, age and education play only a discrete role in the decision to breastfeed.

Based on the probabilities of the model, a label is assigned, observing the thresholds to assess the best cutoff value. The graph representing the accuracy of different threshold values has a peculiar shape, which is due to the imbalance of the dependent variable and does not provide meaningful information on which is the best threshold to choose. The accuracy of the glm is 0.84, which means that 84% of the predictions made by the model are correct.

The ROC (Receiver Operating Characteristic) curve is a graphical representation, which provides an effective way to evaluate and compare the performance of binary classification models. In this dataset, the ROC curve returns an acceptable value of 0.91 for the area under the curve (auc).

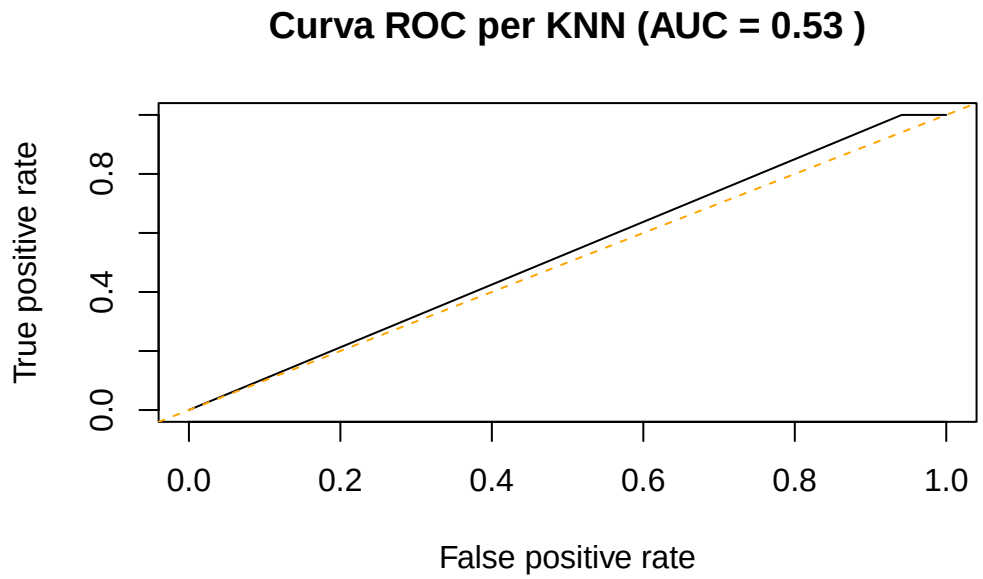




KNN

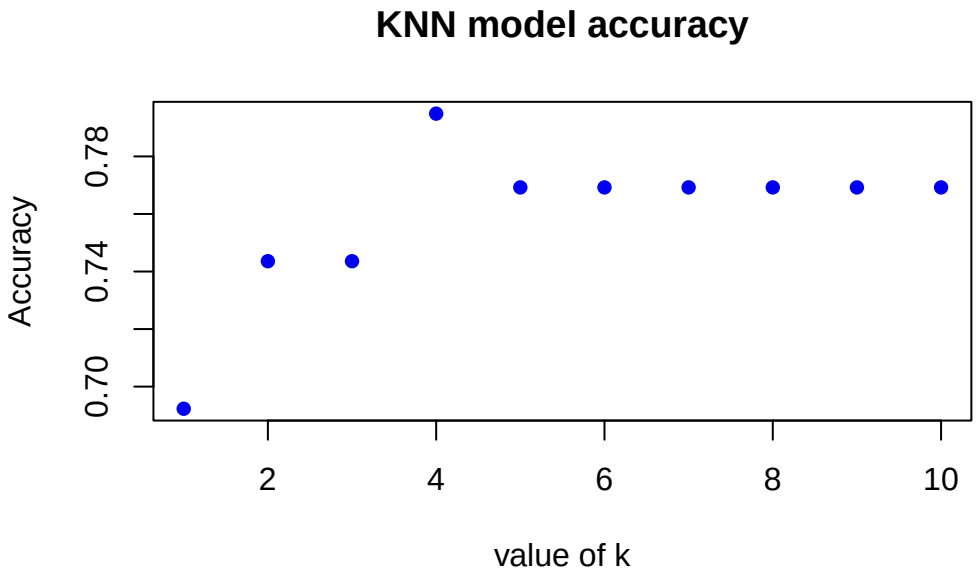
K-Nearest Neighbours is a non-parametric algorithm, which searches for the k closest data points to the point to be classified and assigns the most common class of these points to the point of interest, without assuming any functional form or probability distribution of the data. For this reason, it is useful for data sets where the decision boundary is complex or not well-defined. The KNN can capture complex relationships between the predictors and the response variable, and it is robust to noisy data; however, it is sensitive to the choice of the value of k.

```
KNN  0  1
  0  1  1
  1  9 28
```



Running a *for loop* allows the value of k that maximises the accuracy of the model to be assessed. The value of K that was assigned is 4, which was chosen based on the small size of the test dataset. In this way, the outputs are predicted. The confusion matrix evaluates the performance of a classification model

on a test data set and compares the model's predictions with the true class labels of the test data. In a confusion matrix, the rows represent the real classes of the test data, while the columns represent the predictions made by the model. The confusion matrix can be used for an initial evaluation of the model, calculating an accuracy of 0.74.



Performance of the two methods

An initial comparison between the two models is made through the representation of the accuracy levels, which are 0.84 for the GLM and 0.74 for the KNN algorithms. Furthermore, an examination of the confusion matrices reveals comparable values.

```
glm_l  0  1
    0  8  4
    1  2 25
```

```
knn_l  0  1
    0  1  1
    1  9 28
```

Conclusions

The variables '*breast*', '*pregnancy*' and '*ethnicity*' do not effectively represent the different composition of the population. In fact, including all non-white or non-breastfeeding women in a single label does not allow for an effective understanding of the distinctions between the various groups. The GLM should be chosen because it is an easy model to interpret and from this, it is inferred that the importance of the presence of a partner, friendship choices and ethnicity are parameters that play an important role in the choice of breastfeeding.