Predicting Diabetes Onset Using Logistic Regression in R: A Case Study on the Pima Indians Diabetes Dataset

Author: David Heller

May 2024

Abstract:

The accurate prediction of diabetes onset is crucial for effective medical management and patient care. Leveraging the Pima Indians Diabetes Database from Kaggle, this project aims to develop a logistic regression model using R to predict diabetes onset based on medical attributes such as glucose concentration, blood pressure, BMI, and age. Through data exploration, model building, optimization, and evaluation, the project seeks to identify key risk factors associated with diabetes and assess the model's predictive performance. By practicing logistic regression techniques on a real-world dataset, this project serves as an opportunity to enhance data analysis and machine learning skills while contributing to the development of effective screening tools for diabetes.

INTRODUCTION

The prediction of diabetes onset is a critical challenge in medical diagnostics, with significant implications for patient care and healthcare systems. Diabetes is a chronic disease that affects millions of individuals worldwide, leading to severe health complications if not managed effectively. Early detection of diabetes can facilitate timely intervention, potentially reducing the severity of the disease and improving patient outcomes.

This project leverages the Pima Indians Diabetes Database, sourced from Kaggle, to develop a predictive model for diabetes onset using R. The dataset contains medical information from female patients of Pima Indian heritage, including variables such as glucose concentration, blood pressure, body mass index (BMI), and age, among others. The primary objective is to build a logistic regression model that can accurately predict whether a patient has diabetes based on these medical attributes.

As an exercise to enhance data analysis and machine learning skills, this project provides an opportunity to practice applying logistic regression techniques to a real-world dataset. By analyzing the relationships between the predictor variables and the target variable (diabetes outcome), the aim is to identify key risk factors associated with diabetes and predict the likelihood of its onset.

The analysis involves several key steps:

1. **Data Loading and Exploration**: Reading and examining the dataset to understand its structure and contents.
2. **Model Building**: Fitting an initial logistic regression model using all available predictors.
3. **Model Optimization**: Refining the model using stepwise selection to retain the most significant predictors.
4. **Model Evaluation**: Assessing the model's performance using metrics such as accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC).
5. **Visualization**: Plotting the ROC curve to visualize the model's discriminative ability.

By completing this exercise, the goal is to demonstrate proficiency in logistic regression analysis and showcase the ability to apply machine learning techniques to solve real-world problems. The findings can provide valuable insights for healthcare professionals and contribute to the development of more effective screening tools for diabetes.

Dataset

The Pima Indians Diabetes Database, originating from the National Institute of Diabetes and Digestive and Kidney Diseases, embodies a curated selection of diagnostic measurements from a subset of individuals. A notable facet of this dataset is its focus on females of Pima Indian heritage, aged at least 21 years. This targeted approach ensures a homogeneous cohort, facilitating nuanced analysis while maintaining relevance to a specific demographic group.
Size and dimensionality: 768 observations and 9 attributes.

Within this dataset lie crucial diagnostic variables:

- **Pregnancies**: Reflecting the obstetric history of patients, indicating the number of times pregnant.
- **Glucose Level**: Indicative of blood sugar concentration, measured in plasma glucose concentration at 2 hours in an oral glucose tolerance test.
- **Blood Pressure**: Capturing cardiovascular health metrics, specifically diastolic blood pressure (mm Hg).
- **Skin Thickness**: Offering insights into adiposity and metabolic health, measured as triceps skin fold thickness (mm).
- **Insulin Levels**: A crucial biomarker for glucose metabolism, measured as 2-hour serum insulin (mu U/ml).
- **BMI (Body Mass Index)**: Gauging overall adiposity and metabolic health, calculated as weight in kg divided by height in meters squared.
- **Diabetes Pedigree Function**: Accounting for familial predisposition to diabetes, although the exact calculation method is not specified in the dataset.
- **Age**: A fundamental demographic variable, reflecting the life stage and potential health risks.
- **Outcome**: The target variable indicating the presence (1) or absence (0) of diabetes. Out of the 768 instances in the dataset, 268 are positive for diabetes.

METHODS AND EVALUATION

Logistic Regression

Logistic regression predicts the probability that an observation belongs to one of two classes (binary outcome) by fitting data to a logistic curve. It calculates the odds of the probability (p) of being in the default class (e.g., event happening) as a function of the independent variables. The logistic function ensures that the probability estimate is bounded between 0 and 1. Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$\log\left(\frac{P(y=1\ given\ X)}{p(y=0\ given\ X)}\right) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$$

Evaluation of a Logistic Regression Model

Null Deviance measures the deviance of a model with no predictor variables. It provides a baseline for comparison when assessing the improvement in model fit with the inclusion of predictors.

Residual Deviance measures the deviance after fitting a model with predictor variables. It represents the unexplained variability in the response variable after accounting for the predictors. The comparison between null deviance and residual deviance is commonly used for model assessment, with a significant reduction from null to residual deviance indicating that the predictors contribute significantly to explaining the variability in the response variable.

Information Criteria: Akaike Information Criteria (AIC) and BIC (Bayesian Information Criterion)

Both AIC and BIC provide a method for assessing the quality of a model through a comparison of related models. They are based on the Deviance but penalized for making the model more complicated. If there is more than one similar candidate model (where all the variables of the simpler model occur in the more complex models), then select the model that has the smallest AIC or BIC.

Confusion matrix is the most crucial metric commonly used to evaluate classification models. A confusion matrix is formed from the four outcomes produced from a binary classification. It provides a summary of the predictions made by a model compared to the actual outcomes. A binary classifier predicts all data instances of a test dataset as either positive or negative. This classification (or prediction) produces four outcomes – true positive, true negative, false positive and false negative. A confusion matrix of binary classification is a two-by-two table formed by counting the number of the four outcomes of a binary classifier. We usually denote them as TP, FP, TN, and FN.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Image source: https://towardsdatascience.com/demystifying-confusion-matrix-29f3037b0cfa

• Overall error rate is calculated as the number of all incorrect predictions divided by the total number of the dataset

$$Error\ Rate = \frac{FN + FP}{N}$$

• Overall Accuracy rate is calculated as the number of all correct predictions divided by the total number of the dataset.

$$Accuracy\ Rate = \frac{TP + TN}{N}$$

• Sensitivity is calculated as the number of correct positive predictions divided by the total number of positives. It is also called true positive rate (TPR).

$$Sensitivity = \frac{TP}{TP + FN}$$

• Specificity is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).

$$Specificity = \frac{TN}{FP + TN}$$

A Receiver Operating Characteristic (ROC) curve is a graphical representation used to assess the performance of a binary classification model, such as logistic regression.
The ROC curve illustrates the trade-off between sensitivity (true positive rate) and specificity (true negative rate) at various decision thresholds. If the predicted probability is above the threshold, the observation is classified as the positive class; otherwise, it's classified as the negative class. The ROC curve visualizes the model's performance across different threshold values.

The area under the ROC curve (AUC) summarizes the overall performance of the model across all possible thresholds. AUC ranges from 0 to 1, where higher values indicate better performance.
An AUC of 0.5 corresponds to a model that performs no better than random chance, while an AUC of 1 indicates perfect performance. A steeper ROC curve, closer to the top-left corner of the plot, suggests better overall performance.
The diagonal line (45-degree line) represents random guessing, and a model with good discrimination ability should have a curve above this line. An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier. We expect a classifier that performs no better than chance to have an AUC of 0.5.
ROC curves are useful for comparing different classifiers since they consider all possible thresholds.

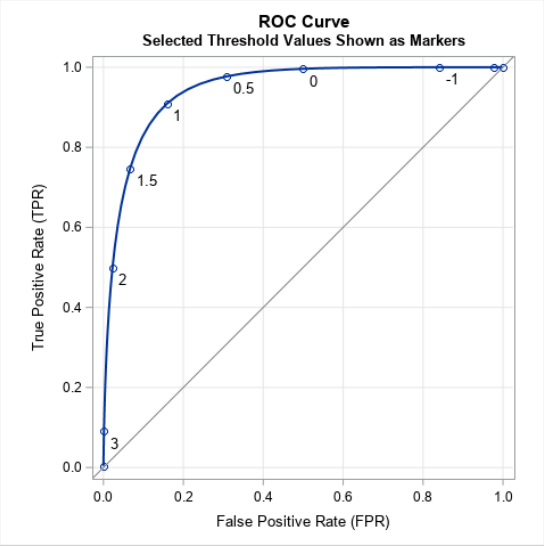


Image source: https://blogs.sas.com/content/iml/2020/02/26/binormal-model-roc-curve.html

ANALYSIS & RESULTS

Full Model

log(p/1-p) = -8.4047 + 0.1232 * Pregnancies + 0.0352 * Glucose - 0.0133 * BloodPressure + 0.0006 * SkinThickness - 0.0012 * Insulin + 0.0897 * BMI + 0.9452 * DiabetesPedigreeFunction + 0.0149 * Age

Where p = probability of having diabetes.

Summary:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.4046964  0.7166359 -11.728  < 2e-16 ***
Pregnancies              0.1231823  0.0320776   3.840 0.000123 ***
Glucose                  0.0351637  0.0037087   9.481  < 2e-16 ***
BloodPressure           -0.0132955  0.0052336  -2.540 0.011072 *
SkinThickness            0.0006190  0.0068994   0.090 0.928515
Insulin                 -0.0011917  0.0009012  -1.322 0.186065
BMI                      0.0897010  0.0150876   5.945 2.76e-09 ***
DiabetesPedigreeFunction 0.9451797  0.2991475   3.160 0.001580 **
Age                      0.0148690  0.0093348   1.593 0.111192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

Coefficients Interpretation:

The coefficients estimated by the logistic regression model provide insights into the effect of each predictor variable on the log-odds of diabetes onset, with all other predictors held constant.

- **Pregnancies**: For each additional pregnancy, the log odds of diabetes onset increase by 0.1232 units. This suggests that an increase in the number of pregnancies is associated with higher odds of diabetes.
- **Glucose**: A one-unit increase in plasma glucose concentration at 2 hours leads to an increase in the log odds of diabetes onset by 0.0352 units. Higher glucose levels are

associated with an elevated risk of diabetes.

- **Blood Pressure**: An increase in diastolic blood pressure by one unit results in a decrease in the log odds of diabetes onset by 0.0133 units. Elevated blood pressure levels may be inversely correlated with the risk of diabetes.
- **Skin Thickness**: The coefficient for skin thickness is not statistically significant (p = 0.9285), indicating that skin thickness does not have a significant impact on the log odds of diabetes onset.
- **Insulin**: Similarly, the coefficient for insulin is not statistically significant (p = 0.1861), suggesting that insulin levels may not have a significant effect on the log odds of diabetes onset.
- **BMI (Body Mass Index)**: An increase in BMI by one unit results in an increase in the log odds of diabetes onset by 0.0897 units. Higher BMI values are associated with an elevated risk of diabetes.
- **Diabetes Pedigree Function**: For each unit increase in the diabetes pedigree function, the log odds of diabetes onset increase by 0.9452 units. A higher diabetes pedigree function indicates a stronger familial predisposition to diabetes.
- **Age**: With each additional year of age, the log odds of diabetes onset increase by 0.0149 units. Advancing age is associated with a higher risk of diabetes.

Interpretation in Terms of Probability

To interpret the coefficients in terms of probability, we can exponentiate each coefficient. For example, exponentiating the coefficient for pregnancies (0.1232) yields approximately 1.131. This means that for each additional pregnancy, the odds of diabetes onset increase by approximately 13.1%.

Significance of Coefficients

In the logistic regression model for predicting diabetes onset, the significance of each coefficient provides valuable insights into the influence of predictor variables on the probability of diabetes. If the significance level is 0.05, then:

Significant Coefficients:

- Pregnancies
- Glucose
- BMI (Body Mass Index)
- Diabetes Pedigree Function
- Age

Non-significant Coefficients:

- Blood Pressure
- Skin Thickness
- Insulin

Understanding the significance of coefficients is pivotal for model interpretation and refinement. Significant coefficients provide valuable insights into influential predictors of diabetes onset, guiding clinical decision-making and risk assessment. Conversely, non-significant coefficients may suggest variables that contribute less to the predictive power of

the model and could potentially be excluded in future iterations to streamline model complexity and improve interpretability.

Optimal Model

To find the optimal model, I used the *step* function in R.
The stepwise model selection process aims to refine the logistic regression model by iteratively evaluating the inclusion or exclusion of predictor variables based on their impact on the model's performance, as measured by the Akaike Information Criterion (AIC). It aims to simplify the model by including only those variables that significantly contribute to the prediction of the dependent variable.

log(p/1-p) = -8.4051+ 0.1232 * Pregnancies + 0.0351 * Glucose - 0.0132 * BloodPressure - 0.0012 * Insulin + 0.0901 * BMI + 0.9476 * DiabetesPedigreeFunction + 0.0148 * Age

Optimal Model Summary:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.4051362  0.7167033 -11.727  < 2e-16 ***
Pregnancies              0.1231724  0.0320688   3.841 0.000123 ***
Glucose                  0.0351123  0.0036625   9.587  < 2e-16 ***
BloodPressure           -0.0132136  0.0051537  -2.564 0.010350 *
Insulin                 -0.0011570  0.0008142  -1.421 0.155275
BMI                      0.0900886  0.0144619   6.229 4.68e-10 ***
DiabetesPedigreeFunction 0.9475954  0.2980063   3.180 0.001474 **
Age                      0.0147888  0.0092897   1.592 0.111393
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 760  degrees of freedom
AIC: 739.45

Number of Fisher Scoring iterations: 5
```

Full Model vs Optimal Model Comparison

Anova

```
> anova(model, optimal_model)
Analysis of Deviance Table

Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
    Insulin + BMI + DiabetesPedigreeFunction + Age
Model 2: Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +
    DiabetesPedigreeFunction + Age
  Resid. Df Resid. Dev Df   Deviance
1       759     723.45
2       760     723.45 -1 -0.0080518
```

-Degrees of Freedom (Resid. Df):

- Model 1 (Full Model) has 759 degrees of freedom.
- Model 2 (Optimal Model) has 760 degrees of freedom.

The increase in degrees of freedom from Model 1 to Model 2 indicates that Model 2 has one fewer parameter (simpler model).

The difference in degrees of freedom between the two models is -1, indicating that one parameter was removed in Model 2.

-Residual Deviance:

Model 1 and Model 2 both show a residual deviance of 723.45. This indicates that both models have similar goodness of fit to the data, with no significant difference in residual deviance between them.

-Change in Deviance (Deviance):

The deviance change is -0.0080518, reflecting a minimal increase in residual deviance despite the reduced complexity. This suggests that the omitted parameter in Model 2 (SkinThickness) did not significantly contribute to improving the model's ability to fit the data, making the Optimal Model a more efficient choice without losing significant predictive power.

BIC

```
> BIC(model, optimal_model)
               df      BIC
model           9 783.2395
optimal_model   8 776.6037
```

The lower BIC for optimal_model suggests that it is the preferable model when considering both the fit and complexity. The reduction in BIC from [full] model to optimal_model (a difference of about 6.6358 points) implies a significant improvement in terms of a balance between model simplicity and the ability to explain the dataset.

AIC

```
> AIC(model, optimal_model)
               df      AIC
model           9 741.4454
optimal_model   8 739.4534
```

The lower AIC for the optimal_model suggests that it is the preferable model when considering model selection criteria. The reduction in AIC from the full model to the optimal model (a difference of about 1.992 points) indicates an improvement in model fit and parsimony, reinforcing the effectiveness of the optimal model in explaining the dataset.

Therefore, the optimal model will be used for making predictions.

Evaluating Optimal Model

Confusion Matrix:

```
> conf_matrix
          Actual
Predicted   0   1
        0 445 111
        1  55 157
```

-True Positives (TP): 157 instances

-True Negatives (TN): 445 instances

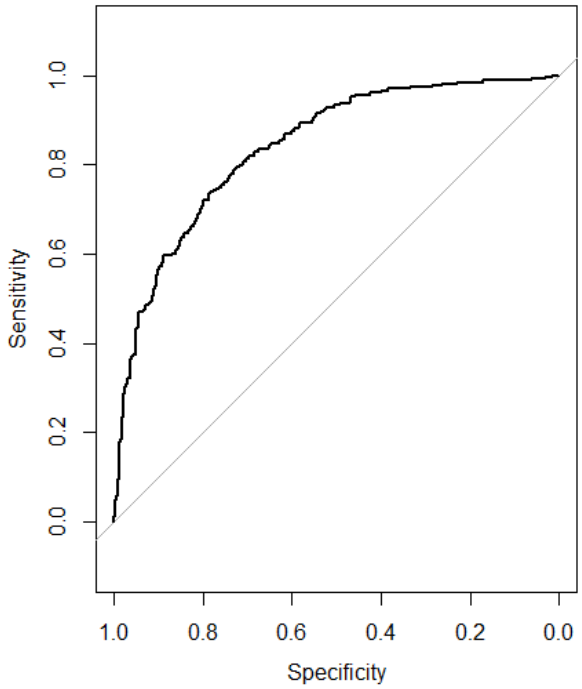
- False Positives (FP): 55 instances

- False Negatives (FN): 111 INSTANCES

Model's Performance

```
> print(precision)
[1] 0.740566
> print(recall)
[1] 0.5858209
> print(accuracy)
[1] 0.7838542
> print(f1_score)
[1] 0.6541667
```

ROC Curve:



AUC:

```
> auc_value
Area under the curve: 0.8396
```

CONCLUSIONS

Strengths of the Model:

- The model demonstrates a strong ability to distinguish between individuals who will develop diabetes and those who will not, as indicated by the high Area Under the Curve (AUC) value of 0.8396.
- It achieves a high overall accuracy of 78.39%, correctly classifying a significant majority of instances, which is crucial for reliable predictions.
- With a precision of 74.06%, the model minimizes false alarms and unnecessary interventions by correctly identifying positive outcomes.

Weaknesses of the Model:

- However, the model's performance is hindered by its relatively low recall of 58.58%, suggesting that it misses identifying about 41.42% of actual positive instances. This could lead to missed opportunities for early intervention.
- The moderate F1 score of 65.42% indicates room for improvement in balancing the trade-off between precision and recall.
- Due to the imbalanced nature of the dataset, there is a potential bias towards predicting the majority class (no diabetes onset), which may lead to suboptimal performance in identifying individuals at risk.

In conclusion, while the model demonstrates proficiency in accurately predicting negative outcomes (no diabetes onset) and exhibits high precision, its effectiveness in identifying positive instances (diabetes onset) is limited by its lower recall and F1 score. Improving the model's ability to detect true positive instances without significantly increasing false positives is essential for enhancing its utility in practical healthcare scenarios.

Suggestions for Model Improvement:

1. **Balancing the Dataset**: Employ techniques such as undersampling the majority class or oversampling the minority class to address class imbalance and improve model performance.
2. **Adjusting Class Weights**: Utilize class-weight parameters to give more importance to the minority class during model training, ensuring better representation of both classes.
3. **Feature Engineering**: Explore the inclusion of interaction terms or additional relevant features to capture complex relationships within the data and improve predictive accuracy.
4. **Model Selection**: Consider exploring alternative algorithms or ensemble methods that may better handle imbalanced data and capture subtle patterns more effectively.

These improvements have the potential to enhance the model's performance and reliability in predicting diabetes onset, thereby facilitating early intervention and better patient outcomes.

Acknowledgment of Data Handling:

It's important to note that in this analysis, the dataset was not split into separate training and test sets for model evaluation. This approach may introduce some limitations to the model's assessment and generalizability. Specifically, without a dedicated test set, we may inadvertently overestimate the model's performance on unseen data, as it has not been rigorously evaluated on independent samples. Therefore, the reported performance metrics should be interpreted with caution, and future iterations of this analysis could benefit from implementing a proper train-test split methodology to ensure more robust model evaluation and validation.

This project served as a personal endeavor to enhance my skills in logistic regression using R. Rather than solely focusing on achieving the highest predictive accuracy, my primary goal was to delve deep into the intricacies of the modeling process. Moving forward, I remain committed to refining my analytical skills and embracing best practices for model evaluation and validation.