Logistic Regression Primer

[1] Theoretical Framework of Logistic Regression

Logistic Regression applies a regression methodology with Logistic Function (Sigmoid Function) to conduct a binary classification analysis.

It may sound oxymoronic to suggest a regression model for classification analysis. Despite its paradoxical name, it has solid real-life applications. Here is how.

While the dependent variable of linear regression is unbounded, Logistic Function (Logistic Function (Sigmoid Function)) projects a bounded mapping of a continuous variable of an infinite range (-∞, ∞) into a bounded range within (0, 1): positive values between 0 and 1. Since the bounded range corresponds to the range of the probability, the regression result can be interpreted as the probability of an event. And using a probabilistic threshold, we can project the probability outputs into a Bernoulli variable, which represents two classes of the results of a binary classification. This is the reason why this method has a confusing name: a regression model for a binary classification application.

Now, we have a classification objective of the binary loan statuses: non-default and default. Now, we assign Bernoulli variable, class 0 and class 1, to non-default and default statuses respectively in the binary classification objective. Let Y denote the probability of being in class 1 (default case).

As a precaution, Y is not the ultimate goal that we want to predict, which is a binary classification. We need to set a probabilistic threshold to map Y to the Bernoulli variable, V consisting of class 0 or class 1. V is our ultimate objective. Y is an intermediary product of the Logistic Regression.

Now, the higher the value of Y, the more chance of default; the lower the value of Y, the lower chance of default.

Odds of being in class 1 as opposed to class 0 can be calculated as:

$$\frac{Y}{1-Y}$$

Logistic regression measures ***the logarithm of the odds*** and plugs in the dependent variable in the linear regression formula as below.

$$\ln\left[\frac{Y}{1-Y}\right] = w * X + b$$

We can transform this equation as follows.

$$Y = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}, where\ z = w * X + b$$

This is called Logistic Function (Sigmoid Function), $\sigma(z)$. Therefore, Y can be described by Logistic Function (Sigmoid Function) as follows.

$$Y = \sigma(w * X + b)$$

“In the case of the logistic regression model, it is not appropriate to apply the method of least squares that is used in linear regression. Using the fact that the variable Y has a known (binomial) distribution, the parameters are estimated with ***the maximum likelihood method***.

It is not possible to give an algebraic expression for the estimators of the parameters. The estimators are obtained by iteratively solving a system of equations, called maximum likelihood equations.” (Girimonte, n.d., p. 3)

That’s a quick summary of the theoretical profile of Logistic Function (Sigmoid Function).

1) ROC

ROC curve visually displays the ability of a binary classification model to correctly separate positives and negatives. An ROC curve is a graph of true positive rate on the Y-axis and false positive rate on the X-axis.

True positive rate is also called 'recall' and calculated based on the following formula:

$$True\ Positive\ Rate/Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

False positive rate is calculated based on the following formula:

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

"*The shape of an ROC curve suggests a binary classification model's ability to separate positive classes from negative classes.*" (Google Developers, n.d.)

https://developers.google.com/machine-learning/glossary#roc-receiver-operating-characteristic-curve

2) AUC: Area under the ROC curve

As a binary classification model improves its ability to correctly separate positives and negatives, the area under the ROC curve above approaches 1.0.

In this sense, the area under the ROC curve summarises the result of ROC curve and measures the ability of a binary classification model to correctly separate positives and negatives.

[2] Exercise

1) Dataset

The credit default historical dataset is obtained from the source in the link below.

- Source: https://r-data.pmagunia.com/dataset/r-dataset-package-islr-default#:~:text=The%20Default%20data%20set%20is,into%20a%20variable%20called%20Default.
- Dataset link: https://r-data.pmagunia.com/system/files/datasets/dataset-63314.csv

We try two models below and compare ROC and Confusion Matrix between them.

- Model 1: Probability = $\sigma(\mathrm{a} * \mathrm{balance} + \mathrm{c})$
- Model 3: Probability = $\sigma(\mathrm{a} * \mathrm{balance} + \mathrm{b} * \mathrm{student} + \mathrm{c})$

2) Model 1: Only one variable, 'balance'.

The trained model is as follows:

➢ Prob[default] = $\sigma(0.0054 * \mathrm{balance} - 10.4324)$ (Model 1 Formula)

where σ represents Sigmoid (logistic function), which is the inverse of the logit function.

```
print(Model_1_summary)
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 default   No. Observations:                 7000
Model:                             GLM   Df Residuals:                     6998
Model Family:                Binomial   Df Model:                            1
Link Function:                   logit   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -590.80
Date:                Tue, 13 Dec 2022   Deviance:                       1181.6
Time:                        10:37:32   Pearson chi2:                 5.36e+03
No. Iterations:                     9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -10.2946      0.407    -25.293      0.000     -11.092      -9.497
balance        0.0053      0.000     21.230      0.000       0.005       0.006
==============================================================================
```

The interpretation of the statistical results

a) Intercept:

- z-score of the intercept (-25.313) is negative and its absolute value is very large.
- p-value is negligible

b) Coefficient of Balance:

- z-score of `balance` (21.411) is positive and its absolute value is very large.
- p-value is negligible

The statistical significances of the results are sound for both the intercept and the independent variable. There is no reason to reject the regression result above

3) Model 3: Two variables, 'balance' and 'student'

The trained model is as follows:

$$\sigma(0.0056 * \text{balance} - 0.697 * \text{student} - 10.5308) \text{ (Model 3 Formula)}$$

where σ represents Sigmoid (logistic function), which is the inverse of the logit function.

```
: 1 print(Model_3_summary)
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 default   No. Observations:                 7000
Model:                             GLM   Df Residuals:                     6997
Model Family:                Binomial   Df Model:                            2
Link Function:                   logit   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -580.19
Date:                Tue, 13 Dec 2022   Deviance:                       1160.4
Time:                        10:37:32   Pearson chi2:                 5.19e+03
No. Iterations:                     9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -10.4081      0.418    -24.904      0.000     -11.227      -9.589
balance        0.0056      0.000     21.026      0.000       0.005       0.006
student       -0.7729      0.173     -4.471      0.000      -1.112      -0.434
==============================================================================
```

The interpretation of the statistical results

a) Intercept:
   - z-score of the intercept (-25.009) is negative and its absolute value is very large.
   - p-value is negligible

b) Coefficients:

   ➢ Balance:

   - z-score of `balance` (21.239) is positive and its absolute value is very large.
   - p-value is negligible

   ➢ Student:

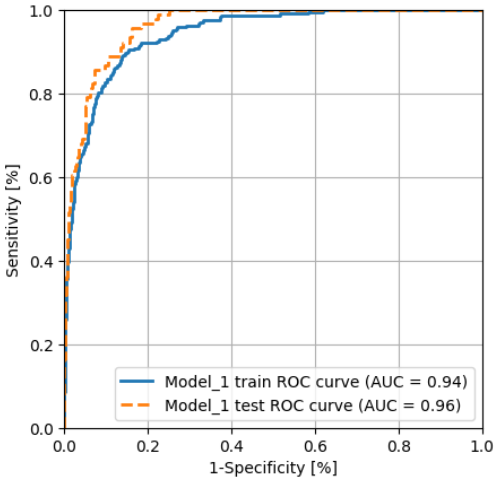
   - z-score of `student` (-4.104) is positive and its absolute value is sufficiently large.
   - p-value is negligible

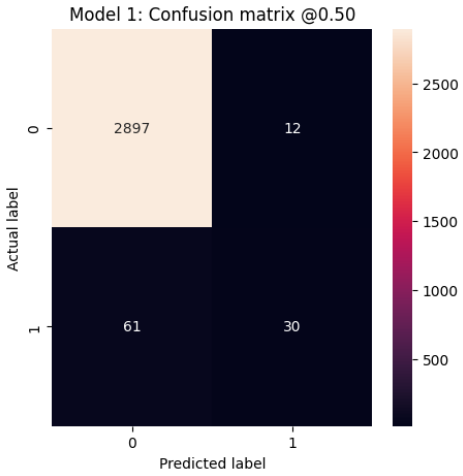
The statistical significances of the results are sound for the intercept and these two independent variables.

There is no reason to reject the regression result above (Model 3 Formula).

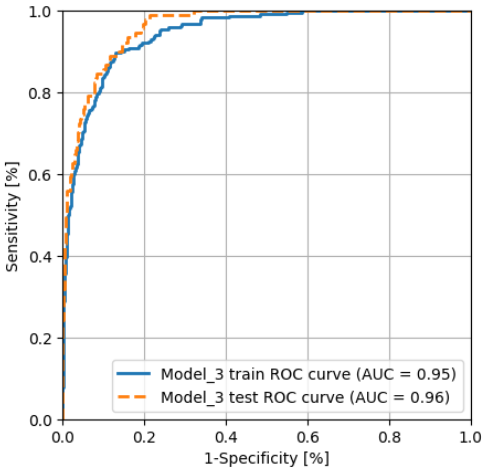
ROC and Confusion Matrix

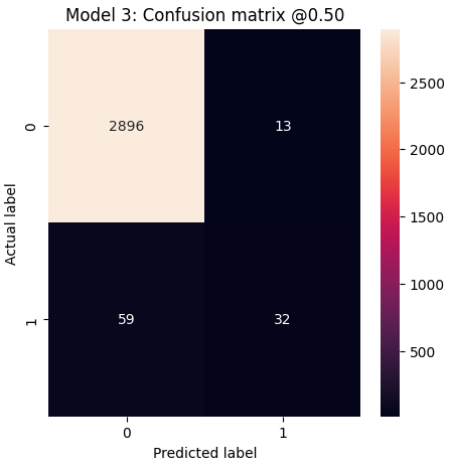
➢ Modelo 1





- Modelo 3





We cannot observe no material differences between these 2 models in ROC and Confusion Matrix.

References

- ArcGIS Pro 3.0. (n.d.). What is a z-score? What is a p-value? Retrieved from esri.com: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm
- Games, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). R Dataset / Package ISLR / Default. Retrieved from R-Data: https://r-data.pmagunia.com/dataset/r-dataset-package-islr-default#:~:text=The%20Default%20data%20set%20is,into%20a%20variable%20called%20Default.
- GeeksforGeeks. (2022, 11 26). Logistic Regression using Statsmodels. Retrieved from geeksforgeeks.org: https://www.geeksforgeeks.org/logistic-regression-using-statsmodels/
- Girimonte, P. (n.d.). REGRESIÓN LOGÍSTICA. Buenos Aires, Argentina: ENAP.
- Google Developers. (2022, 11 24). Logistic regression for binary classification with Core APIs. Retrieved from Tensorflow.org: https://www.tensorflow.org/guide/core/logistic_regression_core
- statsmodels.org. (n.d.). Generalized Linear Model. Retrieved from statsmodels.org: https://www.statsmodels.org/dev/glm.html