Binary Prediction

Arnab Kumar Laha

1 Introduction

Binary prediction is one of the most widely used analytical techniques having far-reaching applications in multiple domains. In the business context, it is used to predict which loans are likely to default, which policyholders are likely to discontinue an insurance policy, which customers are likely to change their service provider, which customers are likely to buy a newly released book, which transactions are likely to be fraud, etc. Apart from business applications, the binary prediction problem arises routinely in medicine, e.g., to determine whether a person has a certain disease or not (Shilaskar and Ghatol 2013), chemistry (Banerjee and Preissner 2018) and many other fields. Because of the huge importance of the binary prediction problem, a number of methods have been developed over the years. The more well-known and widely used methods are linear discriminant analysis, logistic regression, random forest, support vector machines and k-nearest neighbors (see James et al. 2013 for an introduction to these methods).

In this article, we concentrate on the binary prediction task. We discuss the well-known logistic regression predictor and compare its performance with a relatively less widely used predictor—the maximum score predictor using two real-life datasets. The two datasets considered in this paper are both unbalanced with one class having significantly larger number of observations than the other class. The maximum score predictor discussed in this article is based on a modification of the maximum score estimator introduced in Manski (1975). It is observed that the maximum score predictor performs better than the logistic regression predictor for these two real-life datasets.

The article is structured as follows: In Sect. 2, we briefly discuss the logistic regression from a prediction perspective; in Sect. 3, we discuss the use of the logistic

A. K. Laha (✉)
Indian Institute of Management Ahmedabad, Ahmedabad, India
e-mail: arnab@iima.ac.in

A. K. Laha (ed.), *Applied Advanced Analytics*, Springer Proceedings in Business
and Economics, https://doi.org/10.1007/978-981-33-6656-5_2

regression for binary prediction; in Sect. 4, we introduce the maximum-score predictor; in Sect. 5, we compare the performance of the logistic regression predictor and the maximum-score predictor using two real-life datasets, and in Sect. 6 we make some concluding remarks.

2 Logistic Regression

In binary classification problems, the response variable (Y) is dichotomous (i.e., takes only two values which are coded as 0 and 1). The predictor variables are typically either numeric or categorical though other types of variables have also been considered in the academic literature. We assume that we have k predictor variables $X_1, \ldots, X_k$ which are all numeric. In the logistic regression model, we try to fit a model

$$P(Y = 1) = \frac{\mathrm{e}^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}{1 + \mathrm{e}^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}$$

where $\beta_0, \ldots, \beta_k$ are unknown constants that have to be estimated from the given data.

Let $(y_i, x_{1i}, \ldots, x_{ki})$, $i = 1, \ldots, n$ be a random sample of size n from the target population. Then,

$$P(Y_i = y_i) = \left(\frac{\mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}{1 + \mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}\right)^{y_i} \left(1 - \frac{\mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}{1 + \mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}\right)^{1-y_i}$$

where $y_i = 0$ or 1.

The parameters of the logistic regression model are estimated using the maximum likelihood estimation (MLE) method. The likelihood is

$$L(\beta_0, \ldots, \beta_k) = \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}{1 + \mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}\right)^{y_i} \left(1 - \frac{\mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}{1 + \mathrm{e}^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}\right)^{1-y_i}$$

The values of $\beta_0, \ldots, \beta_k$ for which $L(\beta_0, \ldots, \beta_k)$ is maximized are the MLEs, and these are denoted as $\hat{\beta}_0, \ldots, \hat{\beta}_k$. Given a new observation for which the values of the predictor variables are known, say $(x_1^*, \ldots, x_k^*)$, but the value of the response variable Y^* is unknown we can estimate

$$P(Y^* = 1) = \hat{p} = \frac{\mathrm{e}^{\hat{\beta}_0+\hat{\beta}_1 x_1^*+\ldots+\hat{\beta}_k x_k^*}}{1 + \mathrm{e}^{\hat{\beta}_0+\hat{\beta}_1 x_1^*+\ldots+\hat{\beta}_k x_k^*}}$$

The delta method (Small 2010) can be used to obtain the approximate standard error of the estimated probability when the sample size n is large. Let $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k)$ and $\mathbf{x}^{*\prime} = (1, x_1^*, \ldots, x_k^*)$, then in matrix notation we have $\hat{p} = \frac{\mathrm{e}^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}^{*\prime}\boldsymbol{\beta}}}$. An application of delta method yields the estimated asymptotic standard error of $\hat{p}$ as $se(\hat{p}) = \hat{p}(1-\hat{p})\mathbf{x}^{*\prime}\hat{\mathbf{V}}(\hat{\beta})\mathbf{x}^{*}$ where $\hat{\mathbf{V}}(\hat{\beta})$ is the estimated variance–covariance matrix of the estimated coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$. An approximate 95% confidence interval for $\hat{p}$ can then be obtained as $(\hat{p} - 2\, se(\hat{p}), \hat{p} + 2\, se(\hat{p}))$.

3 Application to Binary Prediction

The logistic regression model provides an estimate of the probability $P(Y^* = 1)$. When a prediction of Y^* is desired, this information is converted to an estimate of Y^* by use of a threshold c on the magnitude of $\hat{p}$, i.e., $\hat{Y}^* = 1$ if $\hat{p} > c$ and is $= 0$ otherwise. In other words,

$$\hat{Y}^* = 1_{\hat{p}>c}$$

where 1_A is the indicator of the event A.

The value of $c = 0.5$ is sometimes used as in this case the estimated $P(Y^* = 1)$ is greater than $P(Y^* = 0)$. However, in many real-life business applications it is observed that the estimated value of $P(Y^* = 1)$ is much smaller than 0.5 for most (or sometimes all) values of the predictors within their observed ranges derived based on the training data. Using the threshold $c = 0.5$ would lead to prediction $\hat{Y}^* = 0$ in most/all cases. This often defeats the purpose of building the prediction model. Hence, a good choice of c is important for real-life applications.

The threshold value $\hat{c}$ is obtained by studying the variation of the performance of the logistic regression predictor for different values of c and choosing that value for which the predictor meets the business objective to the greatest extent.

Some of the popular measures that are useful for understanding the performance of a binary predictor are accuracy, sensitivity and specificity. Let the logistic regression predictor be used on a dataset of m observations for which the information about the predictors as well as the response is known. For each observation, the logistic regression model is used to estimate the probability that the response value is 1. Then, a threshold c is used to convert the estimated probabilities into predicted responses (i.e., 0 or 1) as discussed in the previous section. Now let m_{00} be the number of observations for which both the actual and predicted values of response are 0, m_{01} be the number of observations for which the actual response is 0 and predicted value of the response is 1, m_{10} be the number of observations for which the actual response is 1 and predicted value of the response is 0, and m_{11} be the number of observations for which both the actual and predicted values of response are 1. Note that $m_{00} + m_{01} + m_{10} + m_{11} = m$. The measure accuracy is defined as

$$\text{Accuracy} = \frac{m_{00} + m_{11}}{m}$$

and is often expressed as a percentage. While being a good measure in situations where the response is a balanced mix of 0s and 1s, accuracy can be a misleading measure if one of the classes is dominant. As an example, suppose it happens that there are only 5% observations in a dataset whose response is 1. In this case, a predictor that predicts all observations to be 0s would be 95% accurate but would not be able to correctly predict a single observation whose response is 1. To avoid such problems, it is important to look at performance measures that take a more granular view. The sensitivity of a binary predictor is its accuracy in predicting response 1, i.e.,

$$\text{Sensitivity} = \frac{m_{11}}{m_{10} + m_{11}}$$

and the specificity is the accuracy of the predictor in predicting response 0, i.e.,

$$\text{Specificity} = \frac{m_{00}}{m_{00} + m_{01}}$$

These are often expressed as percentages. An effective binary predictor should have both high sensitivity and specificity desirably close to 100%. However, it is generally not possible to have both specificity and sensitivity close to 100% when dealing with real-life datasets and therefore based on the application context a trade-off between sensitivity and specificity is carried out while choosing the threshold value c. Note that all the performance measures discussed until now are all dependent on the choice of the threshold value c. As the threshold value c is varied in the range $0 \leq c \leq 1$, we obtain a set of points (Specificity (c), Sensitivity (c)). The receiver operating characteristic (ROC) curve is a plot of $\big(1 - \text{Specificity}(c), \text{Sensitivity}(c)\big)$, $0 \leq c \leq 1$. The area under the ROC curve (AUC) is often used as a summary measure of binary predictor performance with its ideal value being 1.

In practical applications, it is advisable to determine c using a "validation" dataset that is separate from the training dataset to reduce the chance of overfitting. For this purpose at the initial stage itself, the given data is divided randomly into three parts, training, validation and test datasets containing $100\alpha\%$, $100\beta\%$ and $100(1-\alpha-\beta)\%$ of the data where $0 < \alpha, \beta < 1$ and $0 < \alpha + \beta < 1$. A popular choice for (α, β) is (0.7, 0.2). The test data is used to get an idea about the performance of the binary predictor with new data.

4 Maximum Score Predictor

It is easy to check that $\hat{p} > \hat{c} \iff \mathbf{x}'\hat{\beta} > ln(\frac{\hat{c}}{1-\hat{c}})$. Writing $\hat{\hat{\beta}}_0 = \hat{\beta}_0 - ln(\frac{\hat{c}}{1-\hat{c}})$ and $\hat{\hat{\beta}} = (\hat{\hat{\beta}}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$, we can rewrite $\mathbf{x}'\hat{\boldsymbol{\beta}} > ln(\frac{\hat{c}}{1-\hat{c}})$ as $\mathbf{x}'\hat{\hat{\beta}} > 0$. This suggests an alternative approach to the binary prediction problem, wherein we consider binary predictors of the form $\hat{Y}^* = 1_{\mathbf{x}'\beta > 0}$ and estimate the unknown parameters β by maximizing a "score function." The score function can be accuracy or can be a function of specificity and sensitivity as discussed below. Manski (1985) suggests maximizing the accuracy on the training data for estimating the parameter $\boldsymbol{\beta}$. Since

$\mathbf{x}'\boldsymbol{\beta} > 0 \iff k\mathbf{x}'\boldsymbol{\beta} > 0$ for any constant $k > 0$, to ensure the identifiability of $\boldsymbol{\beta}$ it is restricted to have unit Euclidean norm, i.e., $||\boldsymbol{\beta}|| = 1$. Other "score functions" that may be considered are Youden's index which is *Sensitivity − (1 − Specificity)* and the G-mean which is the geometric mean of the specificity and sensitivity. Note that a good binary predictor would have Youden's index and G-mean as high as possible. For both of these measures, the maximum possible value is 1.

5 Examples

In this section, we provide two examples based on real-life publicly available datasets. For Example 1, we analyze the Amazon books dataset from DASL (https://dasl.datadescription.com/datafile/amazon-books/). We aim to predict whether a book is paperback (P) or hardcover (H) based on the information about their list price, height, width and thickness given in the dataset. In Example 2, we analyze a telecom customer churn dataset (https://www.kaggle.com/mnassrib/telecom-churn-datasets) provided by Orange. Here, we aim to predict churn using the predictors: total day minutes, total evening minutes, total night minutes, total international minutes and number of customer service calls.

5.1 Example 1

The numbers of "*P*" and "*H*" in the given dataset are not balanced, and the ratio $P : H$ is roughly[1] 3:1. We split the given dataset into three parts: training (70%), validation(20%) and test(10%).[2] For the purpose of comparison, we use the same training dataset and test dataset for the logistic regression predictor and maximum score predictor. The validation dataset is not used when working with the maximum score predictor. For the logistic regression predictor, the validation data is used to determine the threshold in two different ways: (i) minimizing the misclassification error (where Misclassification error $= m_{01} + m_{10}$) and (ii) maximizing the G-mean. The sensitivity, specificity and the median G-mean value on the test data are noted for all the three methods, i.e., logistic regression with threshold chosen by minimizing the misclassification error (LR-Misclass), logistic regression with threshold chosen by maximizing the G-mean (LR-G-mean) and maximum-score method with G-mean score (Max-Score). This whole exercise is repeated 100 times, and then the median sensitivity, median specificity and the G-mean of the three methods on the test data are noted (see Table 1).

[1] Of the 318 observations in the dataset, 84 books were H and the rest were P.

[2] The number of observations in training, validation and test datasets was 223, 64 and 31, respectively.

Table 1 Comparison of performances of the LR-Misclass, LR-G-mean and Max-Score predictors on the Amazon books dataset

	Median sensitivity	Median specificity	Median G-mean
LR-Misclass	0.5	0.87	0.66
LR-G-mean	0.75	0.74	0.70
Max-Score	0.75	0.78	0.73

Table 2 Comparison of performances of the LR-Misclass, LR-G-mean and Max-Score predictors on the telecom churn dataset

	Median sensitivity	Median specificity	Median G-mean
LR-Misclass	0	1	0
LR-G-mean	0.49	0.48	0.49
Max-Score	0.51	0.50	0.50

It can be seen that the Max-Score predictor performs better than both the logistic regression-based predictors, LR-Misclass and LR-G-mean, in terms of the median G-mean.

5.2 *Example 2*

The telecom customer churn data consisted of 3333 observations of which 14.5% were churners and the rest were not churners. Of the several variables available in the dataset, we chose only the five variables mentioned earlier for this example. The same steps as those followed in Example 1 above were followed, and Table 2 gives the results.

It may be noted that the accuracy of the LR-Misclass predictor is the highest but is of no use as it always predicts every observation in the test dataset as belonging to the non-churner class. This is a typical problem when dealing with datasets in which one class has many more observations compared to the other class. In these situations, the trivial predictor which assigns every new observation to the majority class has high accuracy but no business relevance. The use of G-mean alleviates this problem to some extent. We find that the Max-Score predictor performs slightly better than the LR-G-mean predictor in terms of the median G-mean.

6 Conclusion

In this article, we discuss the predictive performance of the Max-Score predictor vis-a-vis the LR-misclass and LR-G-mean predictors when dealing with unbalanced datasets by analyzing two real-life datasets. The results suggest that the Max-Score predictor with the G-mean as the score has better predictive performance than the logistic regression-based predictors. This indicates the need for further studies, with both real-life and simulated datasets, to examine the efficacy of the Max-Score predictor and also its limitations.

References

Banerjee, P., & Preissner, R. (2018). Bittersweetforest: A random forest based binary classifier to predict bitterness and sweetness of chemical compounds. *Frontiers in Chemistry*, *6*, 93.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in r. Springer.

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, *3*(3), 205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, *27*(3), 313–333.

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis : Evaluation for cardiovascular diseases. *Expert Systems with Applications*, *40*(10), 4146–4153.

Small, C. G. (2010). *Expansions and asymptotics for statistics*. Chapman: Hall/CRC.