8.2 Logistics Regression

Form of regression analysis used for prediction of discrete variables using mix of continuous and discrete variables

→ Used when the research objective is focussed on whether or not an event occured rather than when it occurred i.e. time course information is not used

→ Instead of building a predictive model for "Y" ~~Reps~~ response directly, the approach models Log odds (Y); hence the name logistics or Logit

→ Used as an alternative to Multivariate Discriminant analysis when underlying assumptions for Discriminant analysis are violated.

Logistics Regression used when:

→ Dependent Variable : Categorical

→ Independent Variable: Continuous or Categorical

examples

- Customer default on Credit card Payment

- Customer reponse to direct mailer

- Customer will buy or not

Types of Logistics Regression

Binary Logit

→ Used when response variable is binary or dichotomous

→ It has only 2 outcomes

e.g. Good - Bad , Yes - No

Ordered Logit

→ used when response variable has more than 2 outcomes

→ and the outcomes can be ordered in meaningful way

e.g. High - Medium - Low

strongly Agree - Agree - Disagree - Strongly disagree

Multinomial Logit

→ used when response variable has more than 2 outcomes

→ and the outcomes cannot be ordered in any manner

e.g. Type choice of Bread

Travel Itineary

Odds Ratio

It is a standard statistical term that denotes probability of success to probability of failure

If probability of success is 0.75, then odds ratio $= \frac{0.75}{0.25} = 3$ $\left(\frac{P}{1-P}\right)$

i.e. there is 3:1 chance of success

Logistics Regression

$P(DV) = f(IV)$ (P: Probability)

But, Probability values can take values between 0 and 1

$\log 0 \Rightarrow -\infty$
$\log \infty \Rightarrow \infty$

One way solve this problem to take an odds ratio $(P/1-P)$ and then take log

$\frac{P}{1-P} \Rightarrow$ can take values of 0 to ∞

$f \log\left(\frac{P}{1-P}\right) \Rightarrow$ can take values of $-\infty$ to ∞

Prediction values can take values from $-\infty$ to ∞

So the Equation:

$$\log\left(\frac{P}{1-P}\right) = Y = f(IV)$$

Log Transformation has linear relationship with predictors

A unit change in x will lead to fixed % change in log y

In terms of y :=) A unit change in x will lead to multiplicative e^{β} change in y

% change in y approx = 100 x $(e^{\beta}-1)$

(for small values of coefficient)

Data Preparation for Logistic Regression includes:

1. Response Variable Coding

The reponse variable will need to be converted to 1/0

Code Repaid Loan as 1 and
Default Loan as 0

2. Missing value Treatment

- using logical rules

3. Outlier Detection

- to ensure we don't have highly skewed values.

4. Multicolinearity

2 independent variables do not provide similar info

5. Variable Transformation

• we have meaningful transformation of variables depending on the research and modelling scope.

6. Descriptive Statistics

• basic measures of central tendency need to be output to validate if correct data is being used for modelling

7. Predictors

• review if existing set of predictors are good enough or more meaningful predictors can be created.

Ideally : Proportion of 1's to 0's should not be less than 2%.

- If rare events proportion of 1's is < 2% oversample
- keep rare events and reduce the no of non events

e.g. fraudulent transactions which are rare

This artificial approach will not alter the model form. It has an impact on the constant term or intercept and has to be corrected once the model is finalized.

Logistics Regression model form is

$$\text{Log}\left(\frac{P_i}{1-P_i}\right) = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

then

bias correction in Intercept due to oversampling

$$\mathrm{Log}\left(\frac{P_i}{1-P_i}\right) = a + b_1 x_1 + b_2 x_2 + b_3 x_3 - \mathrm{LN}\left(\frac{\omega_0 u_1}{\omega_1 u_0}\right)$$

where

ω_0 – population proportion of 0

ω_1 – population proportion of 1

u_0 – derived sample proportion 0

u_1 – derived sample proportion of 1

LN – Natural Log

Estimation for Logistics

The coefficients for the Logistics equation are estimated using a technique known Maximum Likelihood Estimation (MLE)

MLE is a popular method of estimation

- It does not have any underlying assumptions of distribution

- when the underlying distribution of error terms is normal, MLE estimates are similar to OLS estimates

- OLS like many other distributions is a special case of MLE

Intution of MLE

- what values of the unknown parameters make the data we see least surprising ?

SAS Code

Split data into Training dataset and validation dataset

```
proc surveyselect data = inter1
method = SRS out = SAMP1 samprate = 0.5
outall ;
run;

data train;
set samp1 ;
where selected = 0;
run;
```

Logistics Regression model code

```
proc logistics data = xxx outest=modelx;
Model RepaymentStatus = CreditHist Loan
                        Amount Imcomelevel
                        run;
```

Outest - save the output in a temporary SAS dataset

Model - input dependent and independent variables

Logistics Regression steps

- Build a logistics model using all the variables given

- Choosing variables

    - Using the variables shortlisted from profiling & by using the significant variables, build multiple logistics regression models

- Best Model Summary will have

    - P values to choose most significant variables

    - Significant difference between null deviance and residual deviance

    - Least AIC value

glm ⇒ generalised linear model

Model Validation

- Plot the ROCR curve

- choose a point that has high TPR and low FPR. Using this choose the cut off

- Check the AUC value. Choose the model with highest AUC value

- Test the model on atleast 3 to 4 test samples to ensure model performance consistency

* TPR : True Positive Rate

* FPR : False Positive Rate

* AUC : Area Under Curve

R Code :

customers who spend more than average are good customers

family = "bianomial" ⇒ states that we want logistics regression output

step (mod, direction = "both")

stepwise regression

direction ⇒ forward as well as backward selection

It will give suggestion for attributes

type = "response" ⇒ will give predictions in terms of predictions

kappa matrix → if > 0.6 then it is good model

confusion matrix

positive = "1" ⇒ what we want to predict will ~~give~~ parse that as positiv

what is gains chart ?