IMS Health & Quintiles are now IQVIA™

Day -3

Logistic regression

January 10, 2018

Agenda

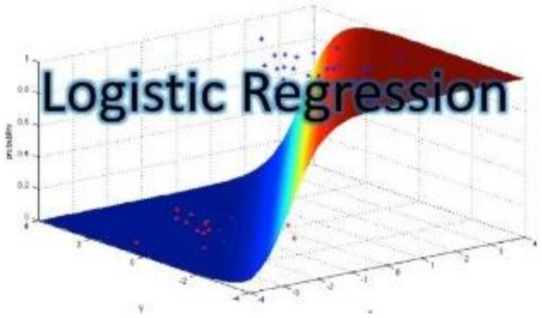
- An overview of logistic regression model
- Types of logistic regression model
- Logistic Regression Case Study
- Performance evaluation vs model validation
- Performance evaluation technique (In detail)
- Model validation technique (In detail)
- Reference
- Practical session (Implementation in R)

What will I learn?

- Concept behind logistic regression models (Non-Linear)
- Difference between linear and logistic regression
- How to build a robust logistic regression model and validate the accuracy using R

An overview of logistic regression model

- It is a form of regression analysis used for prediction of discrete variable using a mix of continuous and discrete predictors



Examples:

a. Does a customer default on credit card payment or not

b. To understand whether the HCPs will respond to the sales force campaign (i.e. Response = Yes) or not (i.e. Response = No)

c. Will a student get admission into business school (i.e. Response = Yes) or not (i.e. Response = No)

All of these are examples for categorical outcome variable

Types of logistic regression model

- There are three types of logistic regression model

1).Binary Logit :

Examples : Default vs Non-Default, Fraudulent vs Non – Fraudulent
(Note : All the examples that we discussed in previous slide fall under this category)

Binary

2).Multinomial Logit :

Examples : High / Medium / Low , Strongly Agree / Agree / Disagree / Strongly Disagree

Multinomial

3).Ordered Logit :

Examples : Choice of Bread (White , Wheat , Multigrain etc.), Mode of transportation (Road, Rail , Air etc.)

Ordered

Logistic Regression Case Study (1/2)

- In order to understand logistic regression let us start with an example
- Consider a sample customers who were granted loans by bank, and over time they have repaid the loan or have defaulted
- The bank wants to identify factors that can predict the likelihood of future customers defaulting

The bank will have access to :

Loan Related Data	Demographic Data
Loan amount	Age
Interest Rate	Gender
Tenure	Location
EMI	Marital status
Purpose of the loan	Employment status
Repayment Status	Income Level

Warning Engine through Logistic regression

Dataset: Copyright © by Jigsaw academy

Logistic Regression Case Study (2/2)

- For simplicity let us assume that bank has access to data on the following variables only

  - **Employment**
  - **The loan amount**
  - **The credit score of the customer**
  - **Repayment status on the loan taken by the customer**

Note : We are restricting ourselves to only three variables to understand how logistic regression model works

For future customer, we have to predict the repayment status based on loan amount, credit score and income category

ILLUSTRATIVE

Loan amount	Credit score	Employment	Repayment status
2,415	Delayed	Unemployed	No
1,813	Delayed	Unskilled employee	No
6,836	All credits paid at time	Skilled employee	Yes
7,356	All credits paid at time	management	Yes
8,567	All credits paid at time	self-employed	Yes
Factors / Independent variables			**Target Variable**

Variables	Types
Loan amount	Continuous
Credit score	Categorical
Employment	Categorical
Repayment status	Categorical

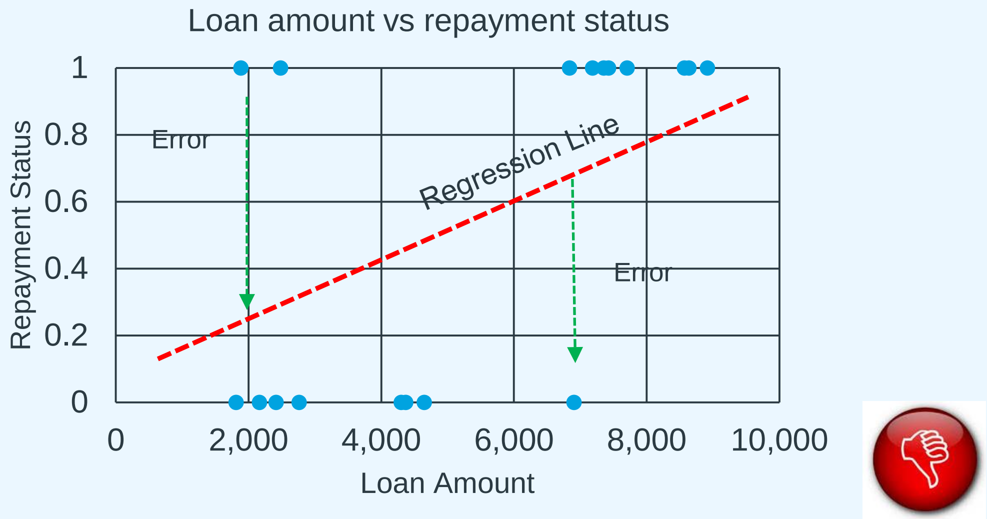
Logistic Regression Case Study (3/2)

ILLUSTRATIVE

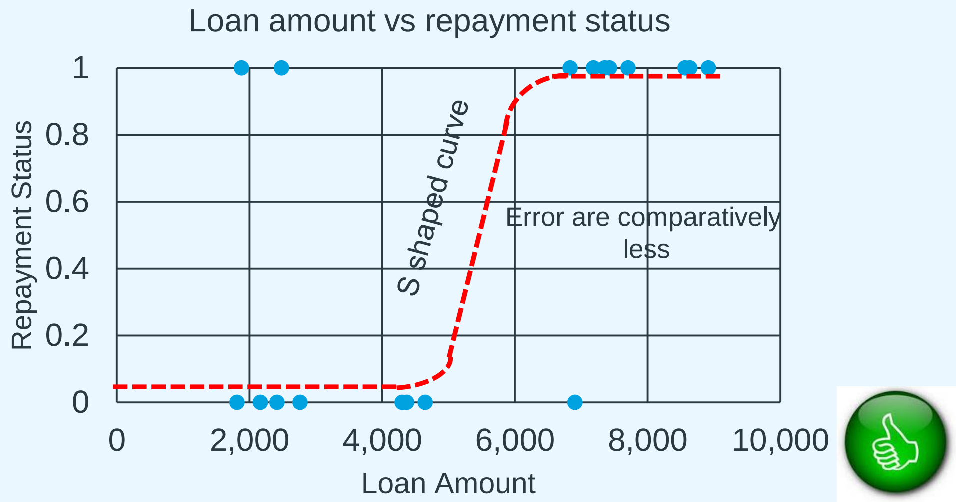
Repayment = f (loan amount) + Error

$$P(Repayment) = \frac{e^{(\beta_0 \; + \; \beta_1 \; (loan\ amount))}}{1 \; + \; e^{(\beta_0 \; + \; \beta_1 \; (loan\ amount))}}$$

Liner regression model



Logistic regression model (Non – Linear)



Performance evaluation vs model validation

Model performance evaluation : It is an assessment of how accurate the model is, and how well it answers the business question framed

Statistical evaluation

- How well is the model “predicting”/”explaining” ?
- **Metric :** Classification table / Confusion matrix

Business evaluation

- Are the relationship captured by the model intuitive and explainable?
- **Metric :** Look for business explanation

Model Validation : It is assessment of how valid and applicable the model is, beyond the sample on which it was generated

Training dataset

- Typically models should be **build on the training data set**

Test dataset

- Developed model should be used on the test data set **to ensure the general applicability of the model**

Performance evaluation technique

Performance of logistic regression model can be accessed through classification table / confusion matrix

<u>Confusion matrix looks as shown below</u> :

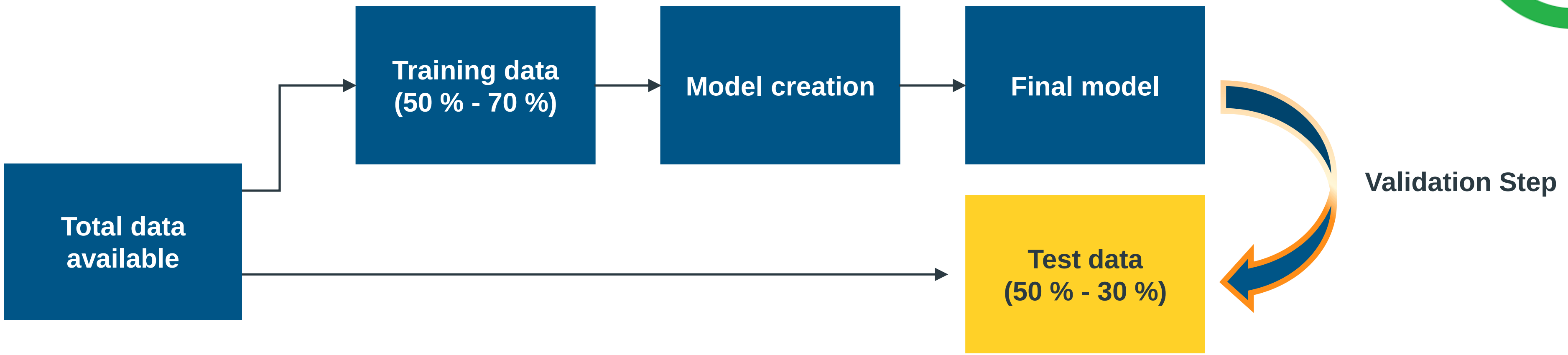
		Predicted Outcome	
		Replayed	Not Replayed
Actual Outcome	Replayed	True Positive (TP)	False Negative (FN)
	Not Replayed	False Positive (TP)	True Negative (TN)

$$\boldsymbol{Model\ Accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TP} + \boldsymbol{FN} + \boldsymbol{FP} + \boldsymbol{TN})}$$

Model validation technique

It is assessment of how valid and applicable the model is, beyond the sample on which it was generated

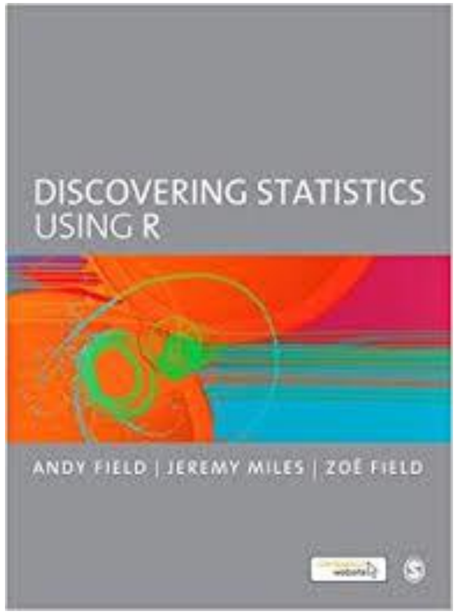
Steps involve in validation process :



Reference :

Discovering Statistics Using R

- Andy Field


DISCOVERING STATISTICS USING R
ANDY FIELD | JEREMY MILES | ZOË FIELD

Practical Session – Implementation in R