Logistic Regression

EDVANCER
EDUVENTURES

Class Case : Revenue Grids

A new product

- A particular portugese bank has come up with with share trading services

- They are offered with a fixed percent commission for each transaction to their existing customers

A pilot

- They offered the services to a small pool of their existing account holders

- They found that not all customers make significant profit for the bank

To discount or not to discount

- They manually divided the pool in two revenue categories .
- Ones which don’t really make any real money for the make
- And others who should be offered good discounts to retain because they turn in profits

Leveraging the pilot

- Now bank wants to roll this service to rest of its customer base

- Which customer should it entice with discounts on commission without compromising on profits?

Non Continuos Response

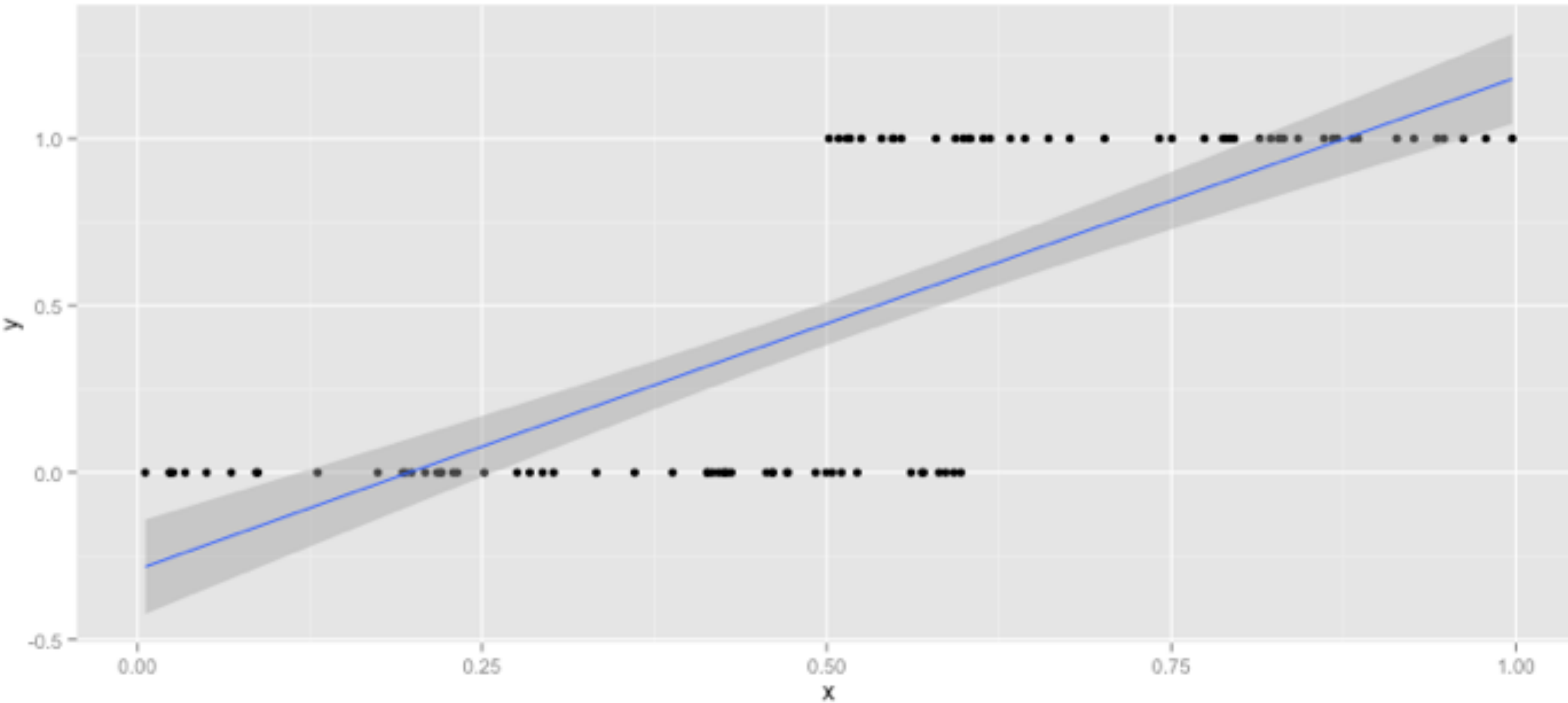
In many of the business problems, response is not as convenient as a continuous numeric variable which can be modelled through simple of multiple linear regression

Binary Response

- Response is binary. Default or non-default, success or failure, sold or not sold, good or bad etc.

- This can not be modelled efficiently with multiple linear regression. WHY?

Linear Regression for binary Response



EDVANCER EDUVENTURES

Proportions

- linear function can theoretically take values between –inf to +inf.

- We'll have to find some kind of transformation for our DV/Response here. First being instead of 0/1; proportions.

- Say we have a dataset with height and gender and for height 62 inches we have 4 M and 8 F then for H=62 inches the proportion or probability of some person being male is 4/12=33%

Odds & Log(Odds)

- Lets call this probability P, problem is that this takes values between 0 to 1 only.

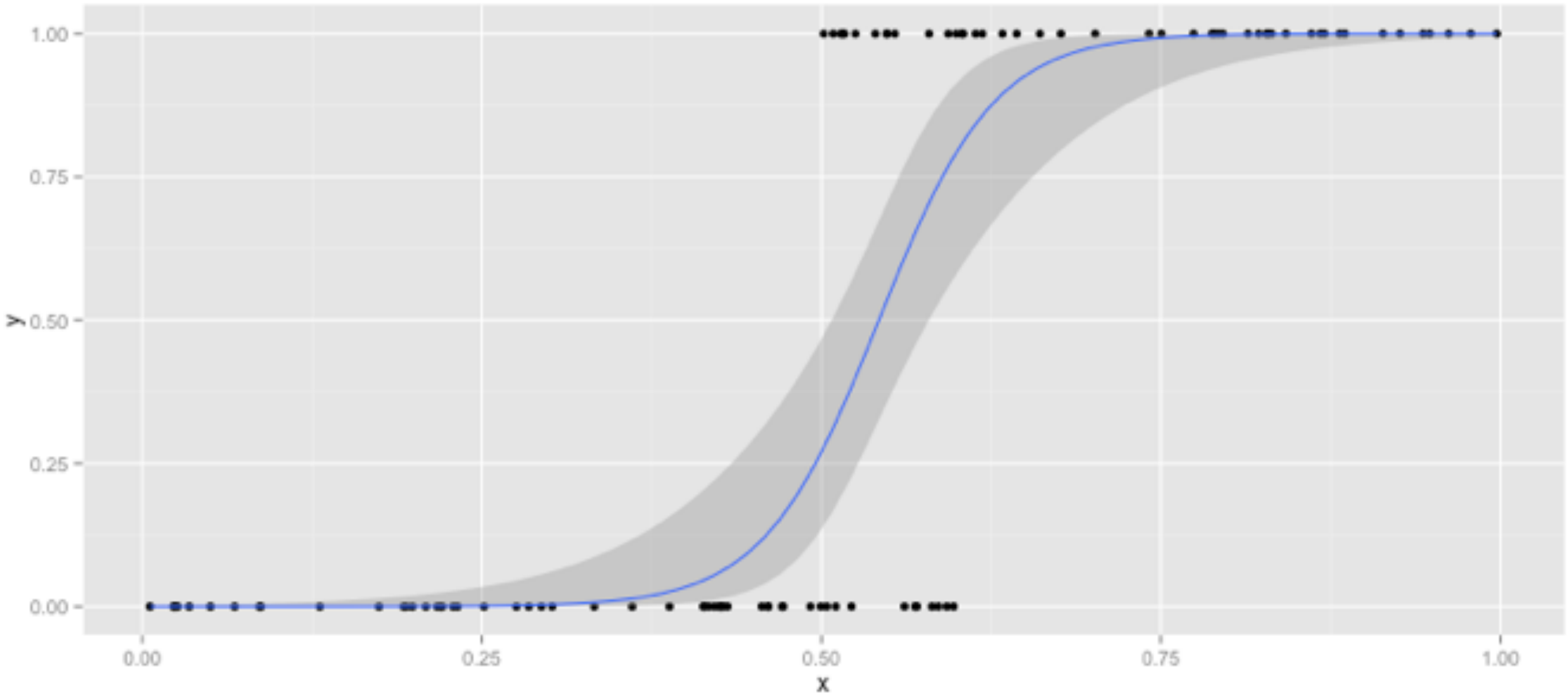
- Consider odds=P/(1-P) = (1/(1-P))-1, since P is in [0,1], odds will take values in [0,inf]. Still the range [-inf,inf] is not covered.

Contd..

- Finally we look at log(odds) or log(p/1-p), since odds are in [0,inf];log odds would take values in [-inf,inf]. Problem solved

- Now we look at the modelling equation which is very similar to MLR.

Modeling Log (Odds)



$$log(\frac{p}{1-p}) = b_0 + b_1 * x_1 + b_2 * x_2 ....$$

Maximum Likelihood Estimation

- One difference from linear regression here is that we are NOT trying to accurately predict log(p/1-p) here. So don't let the linear look of that equation tempt you to assume that we'd use same method of ordinary least squares to get estimates of parameters

- What we want here is the maximum separation between 1s and 0s.

Likelihood

$$P(y_i = 1) = M_i :: P(y_i = 0) = (1 - M_i)$$

$$L_i = M_i^{y_i} * (1 - M_i)^{(1-y_i)}$$

Mi

1-Mi

Maximum Likelihood Estimation

$$L = \prod_{i=1}^{n} M_i^{y_i} * (1 - M_i)^{(1-y_i)}$$

Explanation Contd

- Now this minimization is carried out with the help of numerical methods [mathematics for which would be too complex for scope of this course]

- These numerical methods need to converge for efficient estimation of the parameters.

Interpretation : Odds Ratio

$$\frac{p}{(1-p)} = e^{b_0} * e^{b_1 * x_1} * e^{b_2 * x_2} \ldots\ldots$$

What happens to linear regression assumptions?

- Independence of errors [from the values of log(odds) and predicted values of log(odds)]

- Multi-co-linearity : Detection of multi-co-linearity within IDVs through VIF is independent of how your DV is. Use lm and vif, ignore the parameter estimates.

What to do after we have our parameters?

- Goal is to move back from these estimates of log odds to your original DV 0/1

- Using mathematical operations we can arrive at P, lets call estimated log odds "L", then associated probability:
  - $P= \exp(L)/\{1+\exp(L)\}$

Cutoff

We need to come up with a cut-off for these Probabilities so that they can be converted into a binary decision outcome. Any suggestions?

Sensitivity, Specificity

- Accuracy= TP+TN/P+N
- Each cut-off would have different values for this terms
- What does high specificity and high sensitivity imply? Give business problem example where

Summarizing

		Condition positive	Condition negative	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
		Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Accuracy

KS/Lift or Gain

- Bin your response on score /P
- Plot cumulative 0/1 for each bin
- Bin score/cut-off with maximum lift/gain is chosen as cutoff
- These cut-offs can be different if we associate different costs with 0/1, response/non-response, goods/bads

ROC

- Plot between sensitivity and specificity [false positive rate Vs true positive rate{bin/decile wise}, each point on the ROC curve belongs to a different cutoff for the model.

- Point on the ROC curve which is nearest to top left corner [Sensitivity=1 ,Specificity=1] is considered be optimal cutoff, given your problem seeks to maximize both sensitivity and specifitivity with 1:1 cost equvialence for both the classes.

Validation

- Similar to MLR
- Also performance of the model on validation sample is checked