Logistic Regression

Elsayed Hemayed

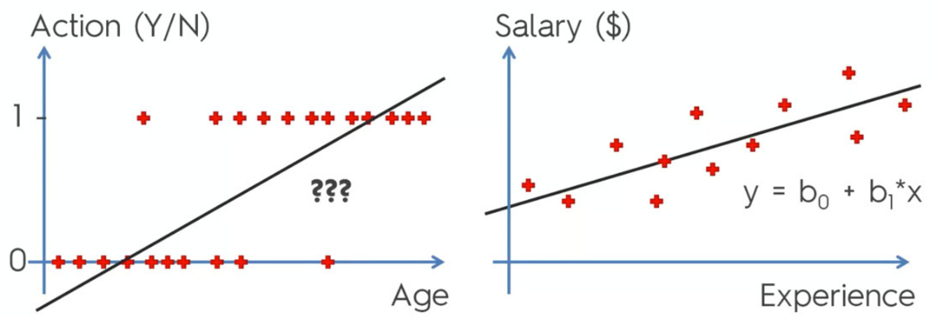
The original slides are from EMC Data Analytics Course and from Udmey course by SuperDataScience Team

Overview of Logistic Regression

– Why we need logistics regression
– Technical description of a logistics regression model
– Interpretation and scoring with the logistics regression model
– Diagnostics for validating the logisitics regression model
– The Reasons to Choose (+) and Cautions (-) of the logistics regression model

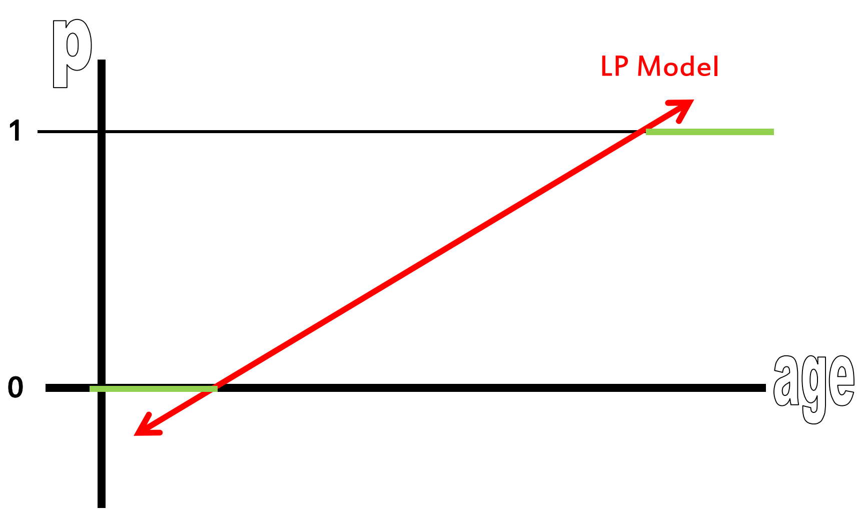
2

Why Logistic Regression?

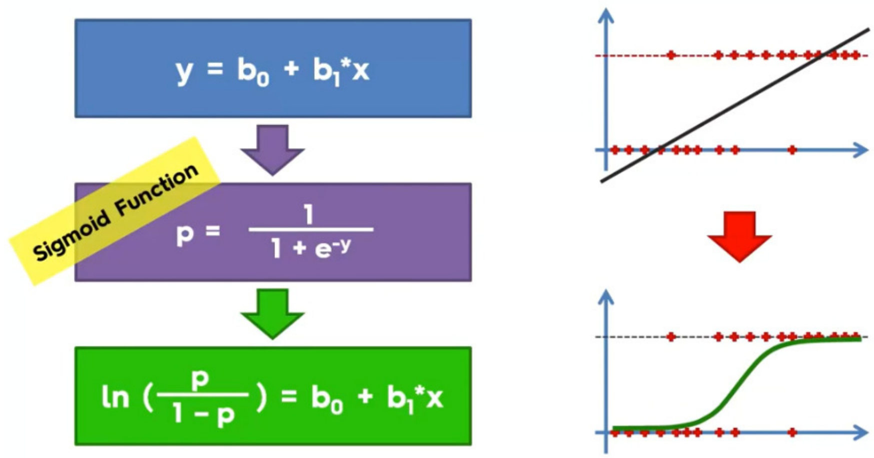


3

Linear Model Plot

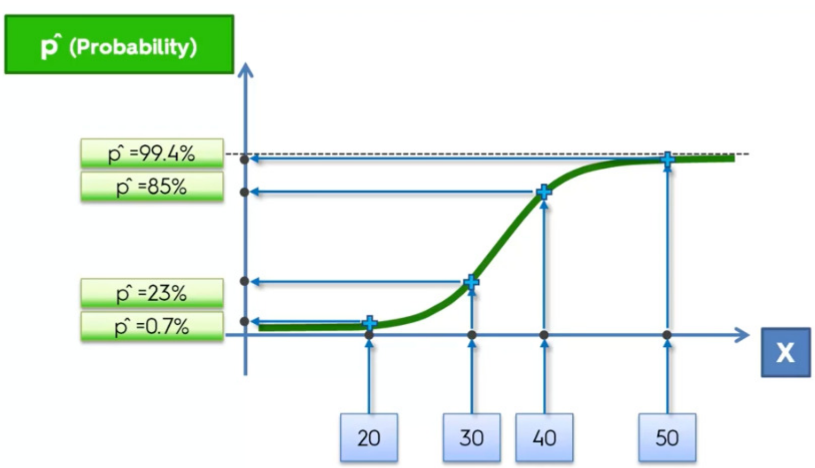


Logistic Regression



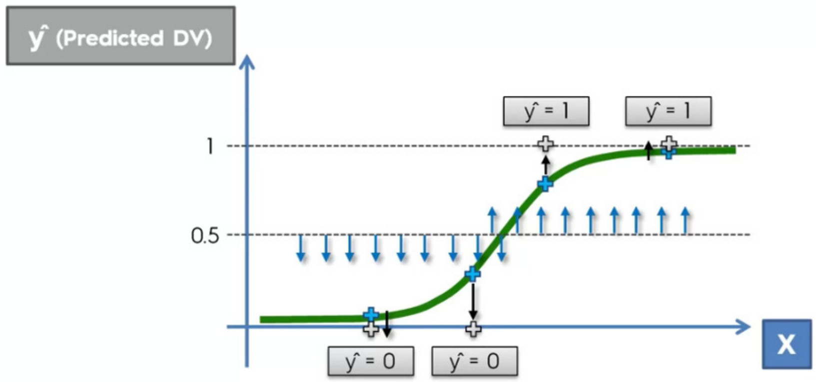
5

Predicted P



6

Predicted y



7

Examples of Binary Outcomes

- Should a bank give a person a loan or not?
- Is an individual transaction fraudulent or not?
- Which people are more likely to vote against a new law?
- Which customers are more likely to buy a new product?

8

Data for Example: Customers' Subscription

- We have data on 1,000 random customers from a given city. We want to know what determines their decision to subscribe to a magazine.
- Subscribe: indicates if a customer has subscribed to the magazine.
- Ages: we will start by examining how age influences the likelihood of subscription.
- Gender: May also influence likelihood of subscription.

9

A linear model

$$Subscribe = b_0 + b_1\, age + \varepsilon$$

	Coefficient
Const	-1.70073
age	0.0645

- *P(subscribe=1) = p = -1.700 + 0.064 age*
- *Every additional year of age increase the probability of subscription by 6.4%*

10

Problems with the Linear Approach

- Probabilities should be bounded 0 <= p <= 1
- The range of age is the data 20<= age <=55
- P(Subscribe =1 | age = 25) = -1.7 + 0.064 x 25 = -0.09 (<0)
- P(Subscribe =1 | age = 45) = -1.7 + 0.064 x 45 = 1.2 (>1)

11

Fixing the Linear Model

- p=f(age)
- What f to use
  - f(.) must be >=0
  - f(.) must be <=1
- (to be >=0)
  - $p=\exp(b_0 + b_1\, age) = e^{(b0 + b1\, age)}$
- (To be <=1)
  - $p= \exp(b_0 + b_1\, age) / (\exp(b_0 + b_1\, age) +1)$

12

Logistics Regression

$p = \exp(b_0 + b_1\, age) / (\exp(b_0 + b_1\, age) + 1)$

Can be rewritten as

$ln\ (p/(1-p)) = b_0 + b_1\, age$

The *ln* term is called the **logit**

And the ratio *(p/(1-p))* is called **the odd ratio**

13

A Logistics model

$ln\ (p/(1-p)) = b_0 + b_1\, age$

	Coefficient
Const	-26.524
age	0.78105

$ln\ (p/(1-p)) = b_0 + b_1\, age = -26.524 + 0.78\, age$

Or

$p = \exp(b_0 + b_1\, age) / (\exp(b_0 + b_1\, age) + 1)$

14

A Logistics model

$ln\ (p/(1-p)) = b_0 + b_1\ age$

	Coefficient
Const	-26.524
age	0.78105

$ln\ [p/(1-p)] = b_0 + b_1\ age = -26.524 + 0.78\ age$

So

For every unit increase in age, *ln [p/(1-p)]* increases by 0.78 units.

For age =35; y = ln [p/(1-p)] =0.813*

p = exp (y*)/[exp(y*)+1] = P (subscribe =1 | age =35) = 0.693

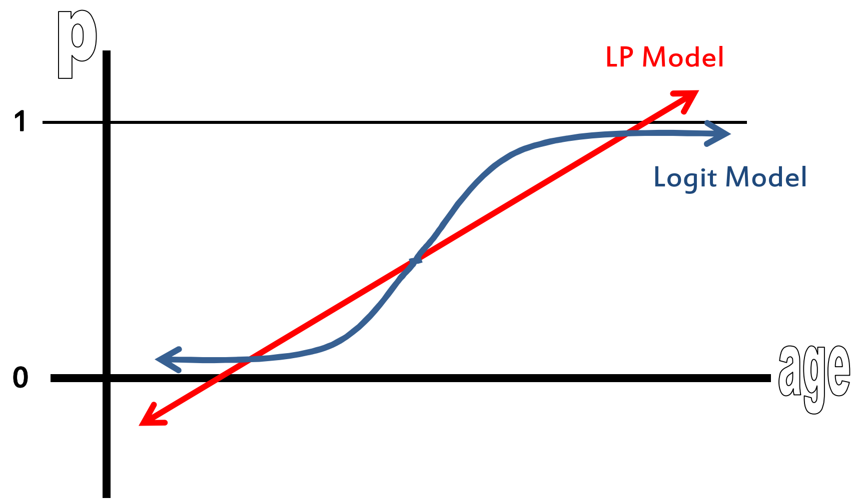
15

A logistics model results

- Change in p from age = 35 to 36 is 0.138
- Change in p from age= 25 to 26 is 0.001
- Change in p from age 45 to 46 is 0.0001

16

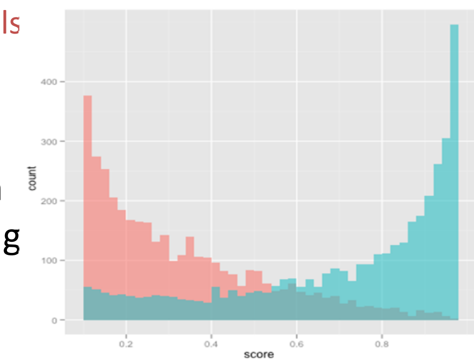
Comparing Linear and Logit Models



An Interesting Fact About Logistic Regression

"The probability mass equals the counts"

If 13% of our loan risk training set defaults then
The sum of all the training set scores will be 13% of the number of training examples



Logistic regression returns a score that estimates the probability that a borrower will default.
Blue = defaulters
Red = non defaulters

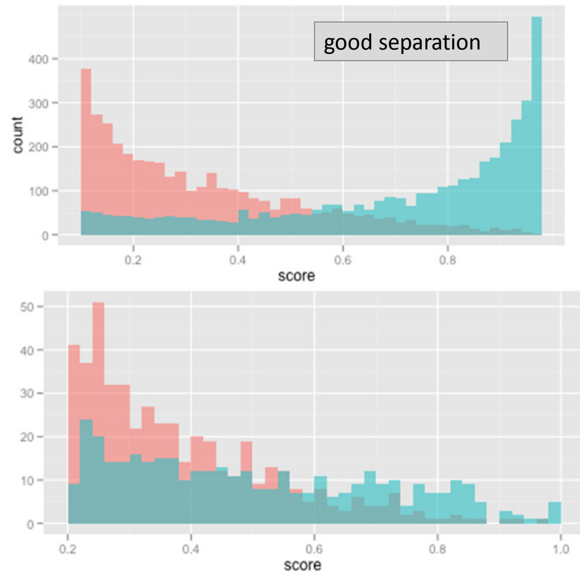
Diagnostics

- Hold-out data:
  - Does the model predict well on data it hasn't seen?
- N-fold cross-validation: Formal estimate of generalization error
- "Pseudo-R^2" : 1 – (deviance/null deviance)
  - Deviance, null deviance both reported by most standard packages
  - The fraction of "variance" that is explained by the model
  - Used the way R^2 is used

Diagnostics (Cont.)

- Sanity check the coefficients
  - Do the signs make sense? Are the coefficients excessively large?
    - Wrong sign is an indication of correlated inputs, but doesn't necessarily affect predictive power.
    - Excessively large coefficient magnitudes may indicate strongly correlated inputs; you may want to consider eliminating some variables**, or using regularized regression techniques.**
      - Unfortunately, regularized logistic regression is not standard.
    - Infinite magnitude coefficients could indicate a variable that strongly predicts a subset of the output (and doesn't predict well on the rest).
      - Try a Decision Tree on that variable, to see if you should segment the data before regressing.

Diagnostics: Plot the Histograms of Scores



Logistic Regression - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Explanatory value: Relative impact of each variable on the outcome in a more complicated way than linear regression	Does not handle missing values well
Robust with redundant variables, correlated variables Lose some explanatory value	Assumes that each variable affects the log-odds of the outcome linearly and additively Variable transformations and modeling variable interactions can alleviate this A good idea to take the log of monetary amounts or any variable with a wide dynamic range
Concise representation with the the coefficients	Cannot handle variables that affect the outcome in a discontinuous way. Step functions
Easy to score data	Doesn't work well with discrete drivers that have a lot of distinct values For example, ZIP code
Returns good probability estimates of an event	
Preserves the summary statistics of the training data "The probabilities equal the counts"	