MSDS 660

Week 7: Logistic and Polynomial Regression

Dr. Ksenia Polson

kpolson@regis.edu

Agenda

- Understand when to use polynomial regression and how to interpret it
- Understand when to use logistic regression and how to interpret it

Polynomial regression

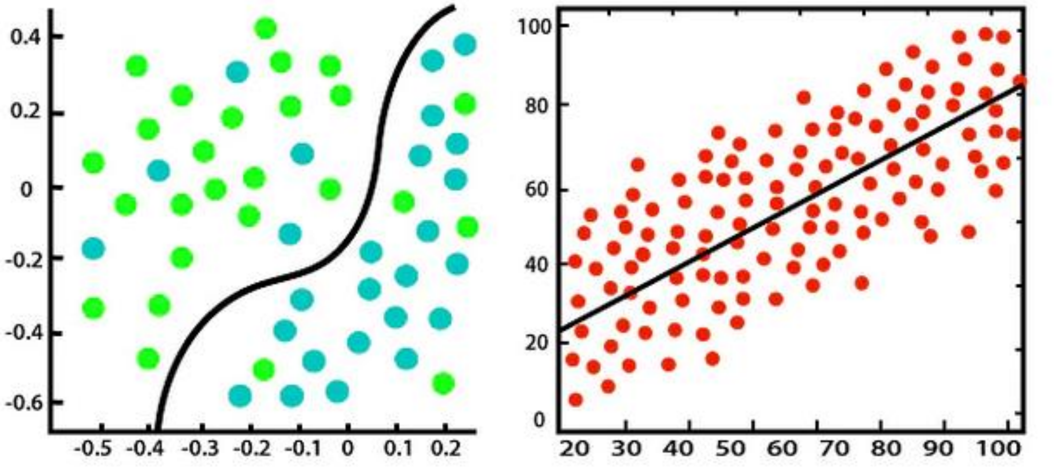
- Very similar to multiple linear regression model
  - Same assumptions, same ways of evaluating fit
  - If there is a curve in the fitted values vs. residuals plot, try polynomial regression
- Polynomial CO2 RegressionDemo.r

Logistic Regression Use Cases

Can be used when the outcome is binary or multinomial:

- Fraud detection
- Pass/fail rates
- Disease detection (e.g. cancer, heart disease, diabetes)
- Survival prediction
- Email spam detection

Classification vs. Regression



Classification   Regression

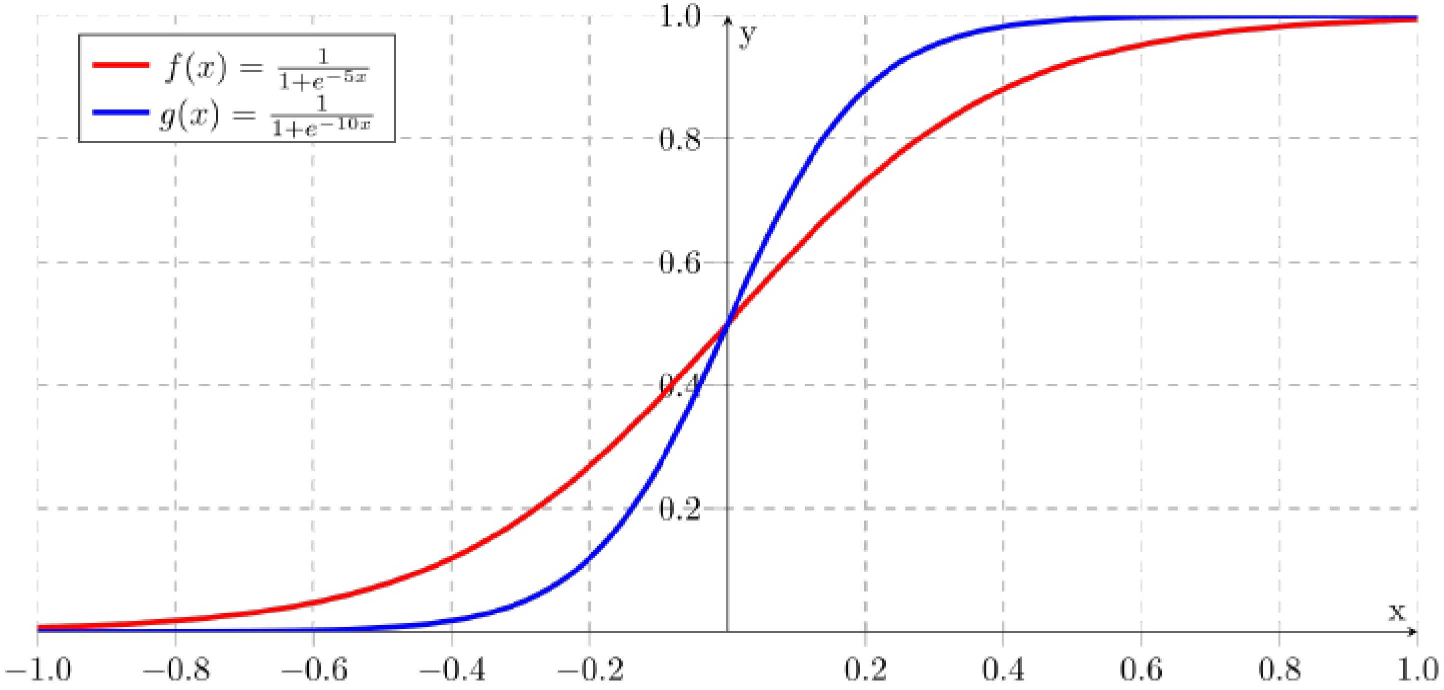
source

What is the difference?

- Classification: discrete labels (i.e. person, Index Fund, loan status)
- Regression: predicting a continuous variable (mpg of a car, loan amount in dollars, interest rate)
- In a classification, we may have two-class predictions:

In our demo for this week, loan term is 36 or 60 months. So, it could be coded “0”: for 36 and :1” for 60.

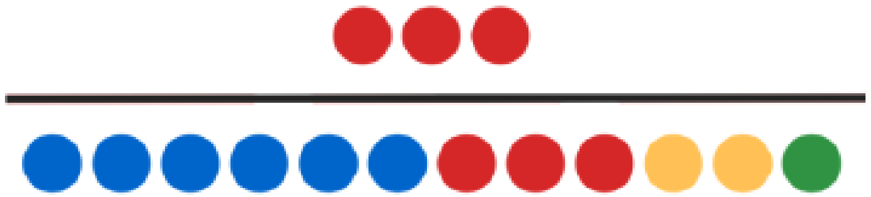
What is this function called?



Odds ≠ Probability

$$Probability = \frac{Chances\ for}{Total\ chances}$$

Probability of Red

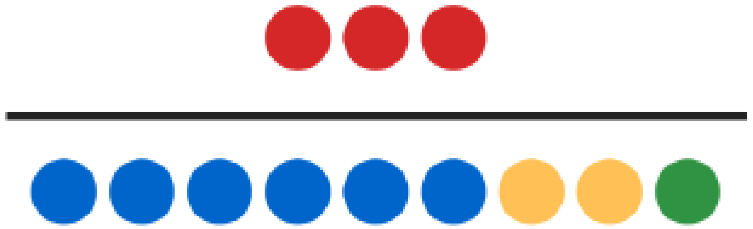


OR

$$P(RED) = \frac{3\ RED\ marbles}{12\ TOTAL\ marbles} = 25\%$$

$$Odds = \frac{Chances\ for}{Chances\ against}$$

Odds For Red



logarithms

The log, base b, of a number, x, is the exponent, y, that b is raised to in order to get x.

Default is usually base 10

- For example: if b=10, y=2, x would be =100 and the log of 10*100=2

$$b^y = x$$

$$\log_b x = y$$

$$b^{\log_b x} = x$$

The odds

The left side is the odds

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

Logit

Same equation as before, but we took the log of both sides.

log(e^x) = x

The left side is the log odds, or logit

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Logistic function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

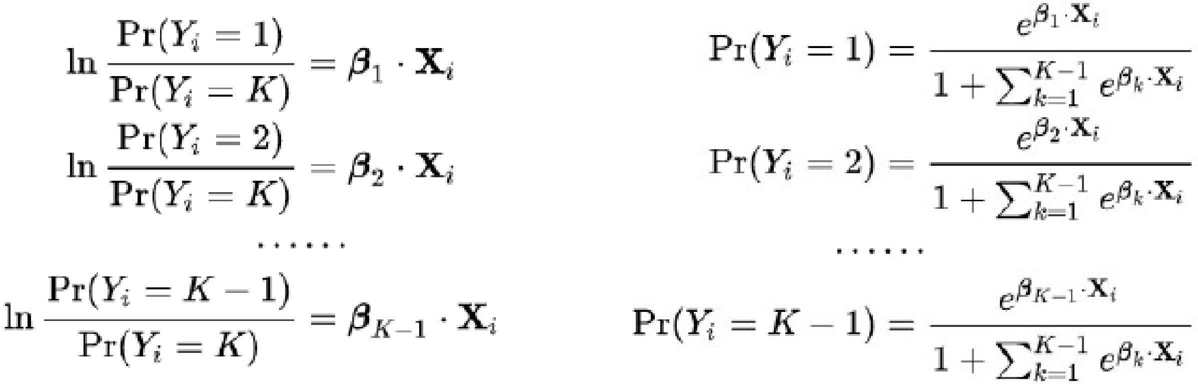
$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- The equations are all the same, they were rearranged
- We have p (*X*), or the probability of our outcome being a 1 for a given value of *X*

Logistic Regression Assumptions

- Independence of cases
- No multicollinearity between predictors
- No homogeneity of variance (HOV)
- No normality of errors

In multiple logistic regression, it’s as if we are running several logistic regressions at once- we keep adding variables



$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i$$

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i$$

$$\cdots\cdots$$

$$\ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i$$

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

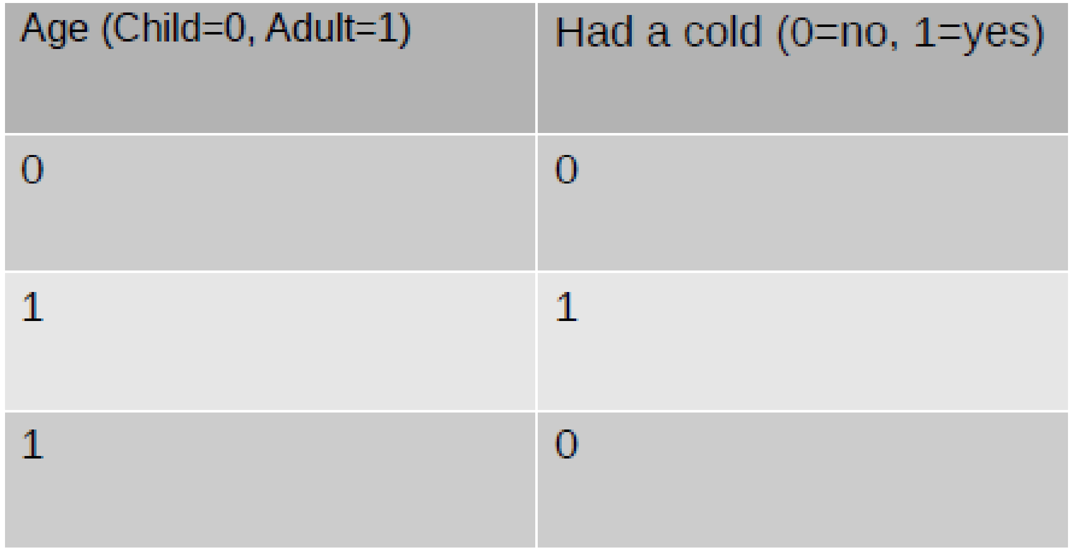
$$\Pr(Y_i = K-1) = \frac{e^{\boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

https://en.wikipedia.org/wiki/Multinomial_logistic_regression#As_a_set_of_independent_binary_regressions

Risk Ratio

Typically used for something ‘bad’, like getting a disease, defaulting on a loan, car accident, etc

Example:



Age (Child=0, Adult=1)	Had a cold (0=no, 1=yes)
0	0
1	1
1	0

Risk Ratio

Once we have fit the logistic model, we can calculate the risk ratio that an adult will get a cold vs a child:

$$P(x)=\frac{e^{B0+x*B1}}{1+e^{B0+x*B1}}$$

$$P(x=1)=\frac{e^{B0+B1}}{1+e^{B0+B1}} \qquad P(x=0)=\frac{e^{B0}}{1+e^{B0}}$$

$$\frac{P(x=1)}{P(x=0)}=\frac{\frac{e^{B0+B1}}{1+e^{B0+B1}}}{\frac{e^{B0}}{1+e^{B0}}}$$

Fitting the model

We use maximum likelihood

- This maximizes how close our predictions are to the true values in a non-linear way

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Interpreting coefficients in LR

Logistic Regression model:

- -1.2345+0.456x_1x1+0.02x_2

- The effect of the odds of one unit increase in x_1 is exp(0.456) =1.58.

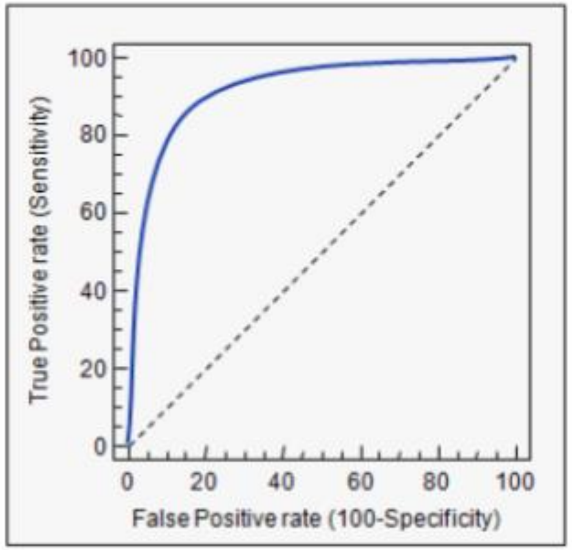
Evaluating Models

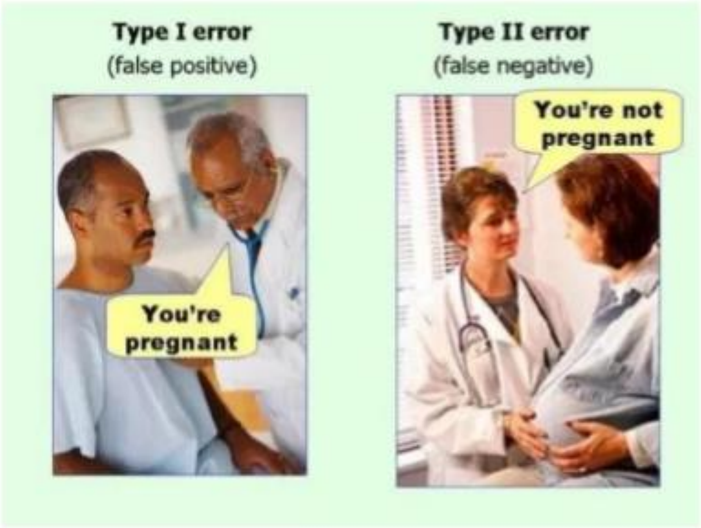
If we get probabilities, how do we know which cutoff to use?

We can set the threshold (up to 1) to anything we want, but it will change the performance of our classifier.

A Receiver Operating Characteristic (ROC) curve is used to visualize and measure performance overall.

- We evaluate model with the Area under the curve (AUC), we want the area to be close to 1.

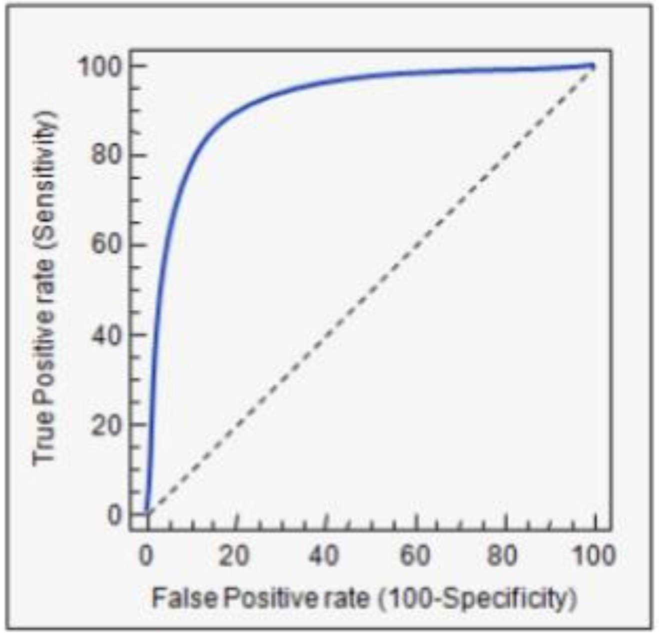


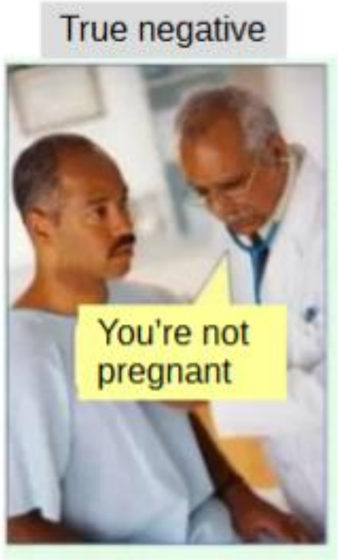


Evaluating Models

Sensitivity= True Positive rate

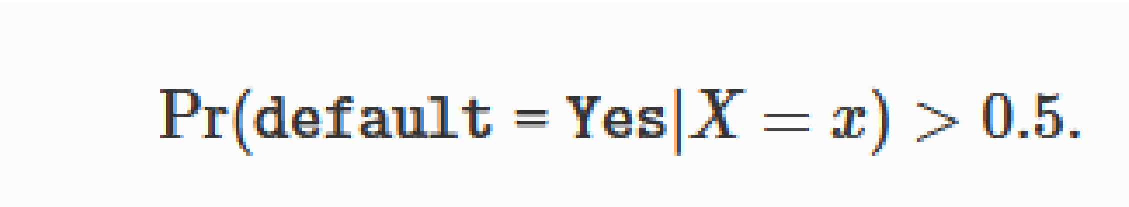
Specificity =True Negative rate





Evaluating Models

The Bayes classifier assigns an observation to the default class if:



$$\Pr(\texttt{default} = \texttt{Yes} | X = x) > 0.5.$$

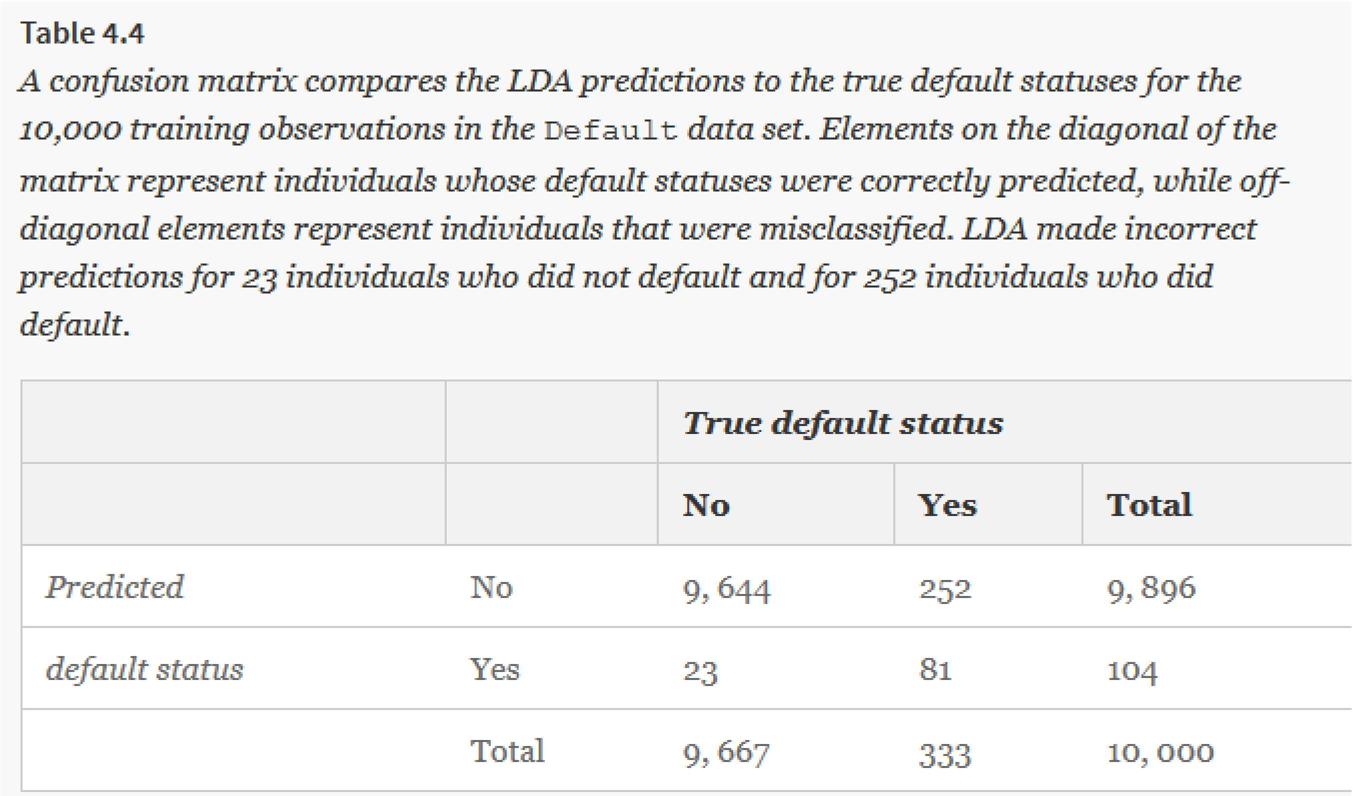


Table 4.4

A confusion matrix compares the LDA predictions to the true default statuses for the 10,000 training observations in the `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.*

		True default status		
		No	**Yes**	**Total**
Predicted	No	9, 644	252	9, 896
default status	Yes	23	81	104
	Total	9, 667	333	10, 000

An Introduction to Statistical Learning with Applications in R (ISLR) by Gareth James et. al.
chapter 4.3 (logistic regression)

Logistic Regression Demo

Download the loans_full_schema.csv and add the file path to the Lab_Week7.Rmd file.

Review the code in the file.

Discussion Activity

Perform logistic regression on the diabetes.pima dataset.

Fill in the code in the R file.

Post R code and comment on the model and its accuracy.

Assignment

Perform a logistic regression on the marketing dataset, marketing.csv on worldclass or on a data set of your choosing.

- Predict customer’s response on marketing campaign (i.e. 1 if customer accepted the offer in the last campaign, 0 otherwise)
- Interpret the model and comment on its accuracy
- Provide a summary of your findings. What are the implications of your analysis for company XYZ? (or if you choose a different data set include appropriate response)
- Post Rmd file and a knitted pdf file to the assignment dropbox.