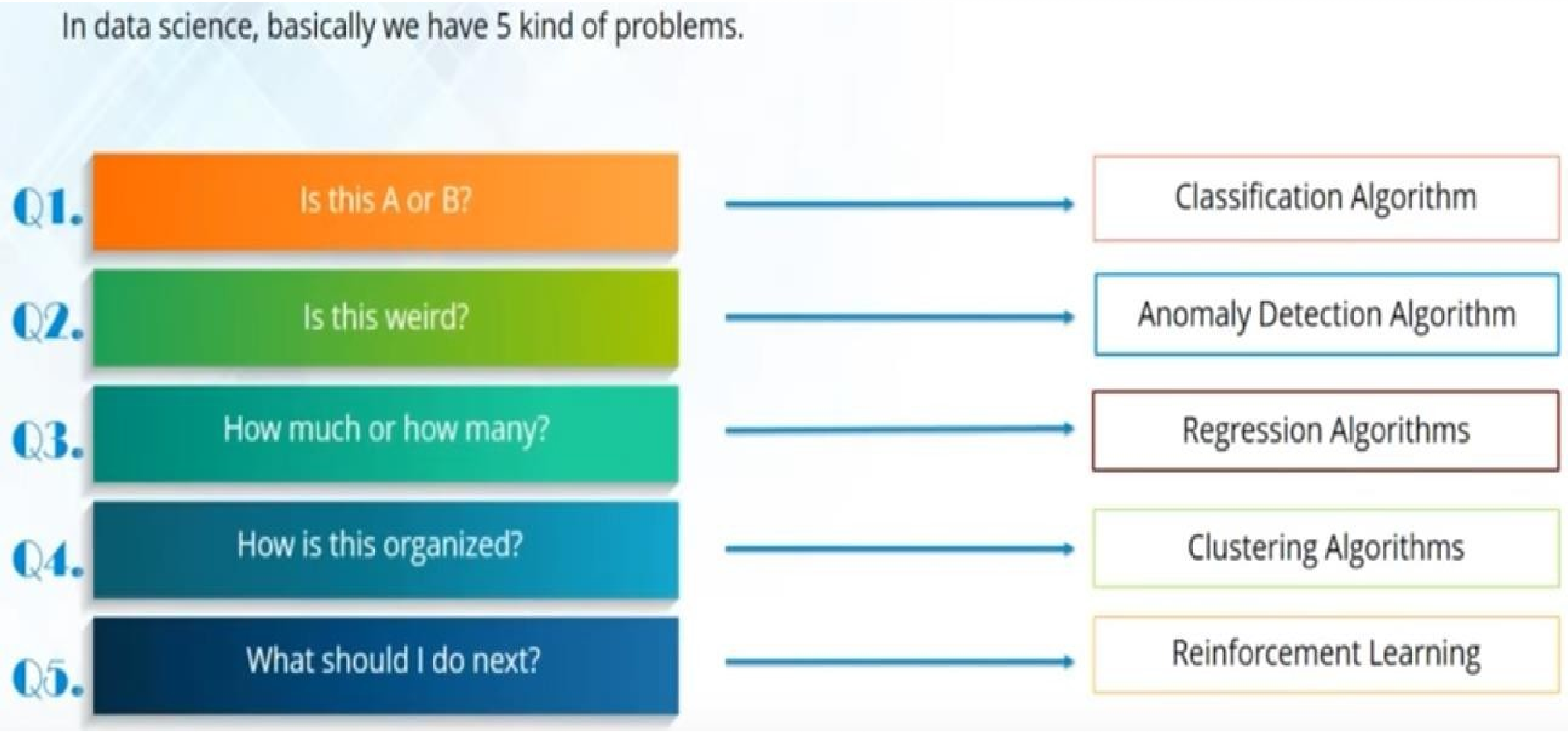
LOGISTIC REGRESSION

UTKARSH KULSHRESTH
Data Scientist – TCS
LearnBay

Algorithms to solve Business Problems



In data science, basically we have 5 kind of problems.

Q1. Is this A or B? → Classification Algorithm

Q2. Is this weird? → Anomaly Detection Algorithm

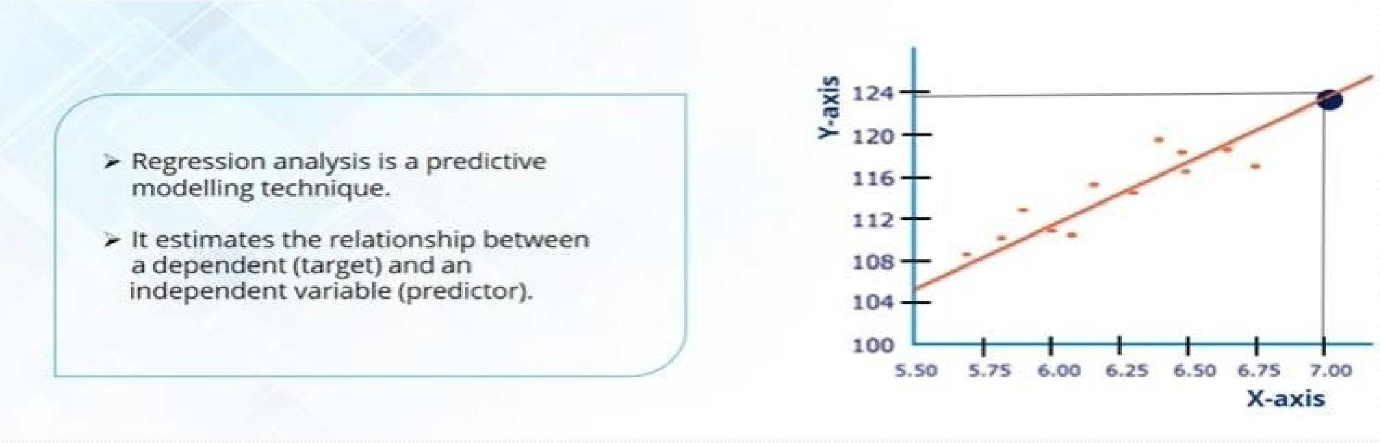
Q3. How much or how many? → Regression Algorithms

Q4. How is this organized? → Clustering Algorithms

Q5. What should I do next? → Reinforcement Learning

What is Regression?

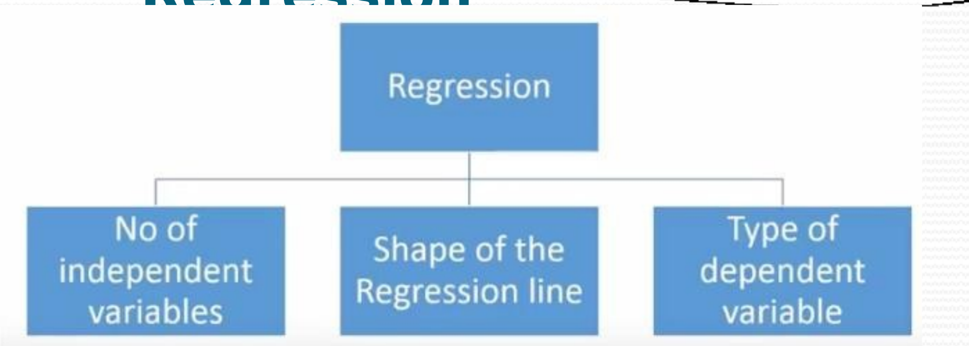
- In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').



- Regression analysis is a predictive modelling technique.
- It estimates the relationship between a dependent (target) and an independent variable (predictor).

- Correlation and Regression are not the same. Correlation quantifies the degree to which two variables are related. Correlation does not fit a line through the data points. Simply You are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does.

Type of Regression


Regression
No of independent variables
Shape of the Regression line
Type of dependent variable

1. **Linear Regression :** Dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

2. **Logistic Regression:** Dependent variable is binary (0/ 1, True/ False, Yes/ No)

3. **Polynomial Regression:** A regression equation is a polynomial regression equation if the power of independent variable is more than 1, the best fit line is not a straight line. It is rather a curve that fits into the data points.

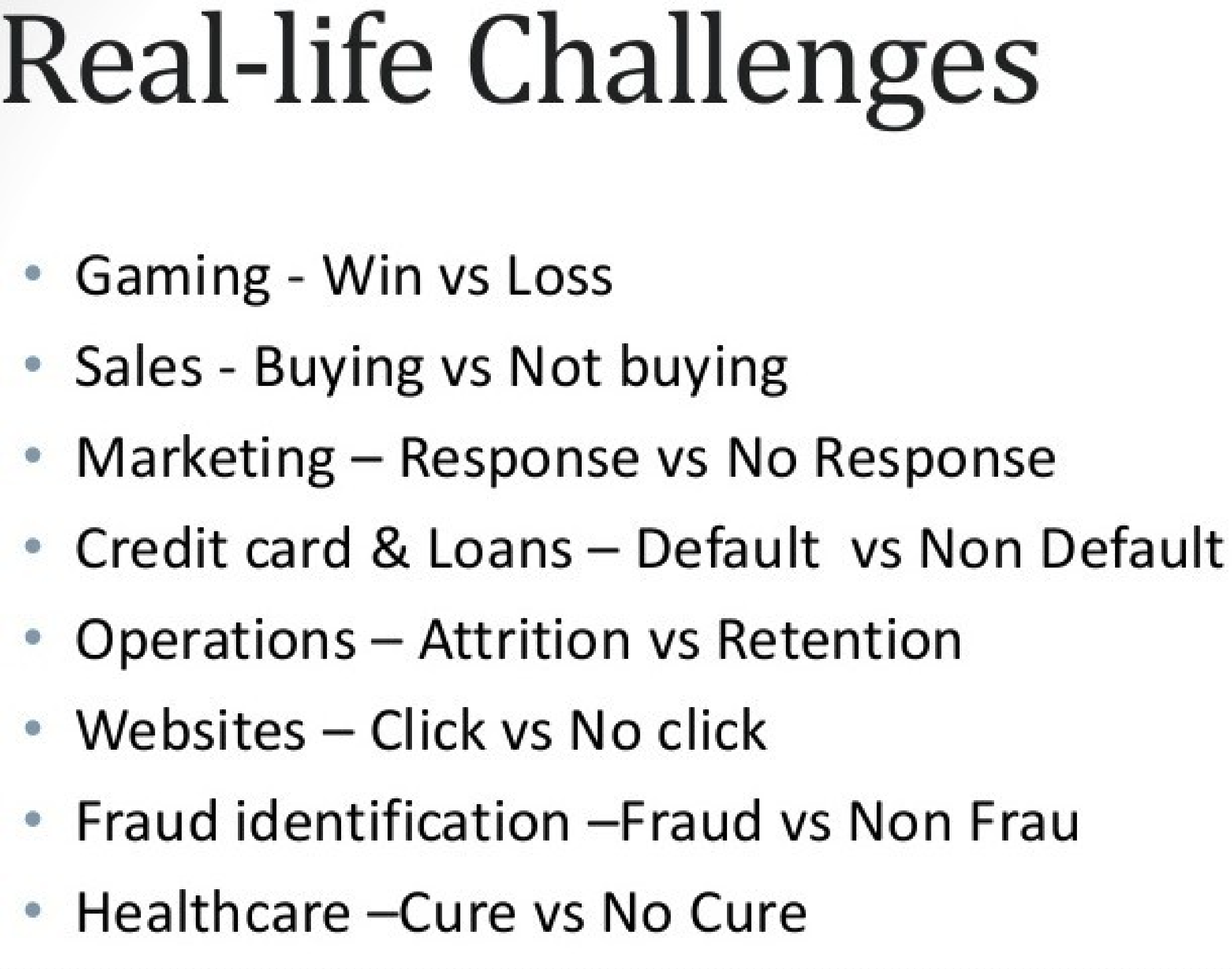
4. **Stepwise Regression:** when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention.

5. **Ridge Regression:** Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value.

6. **Lasso Regression:** Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models.

7. **ElasticNet Regression:** ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated.

Logit Use cases

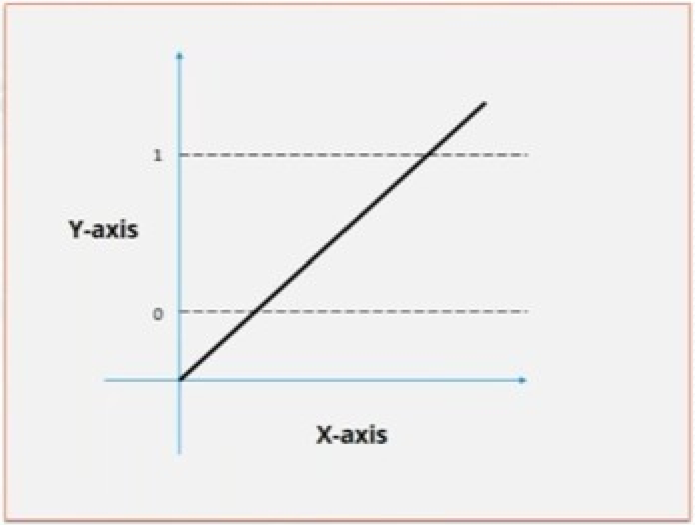


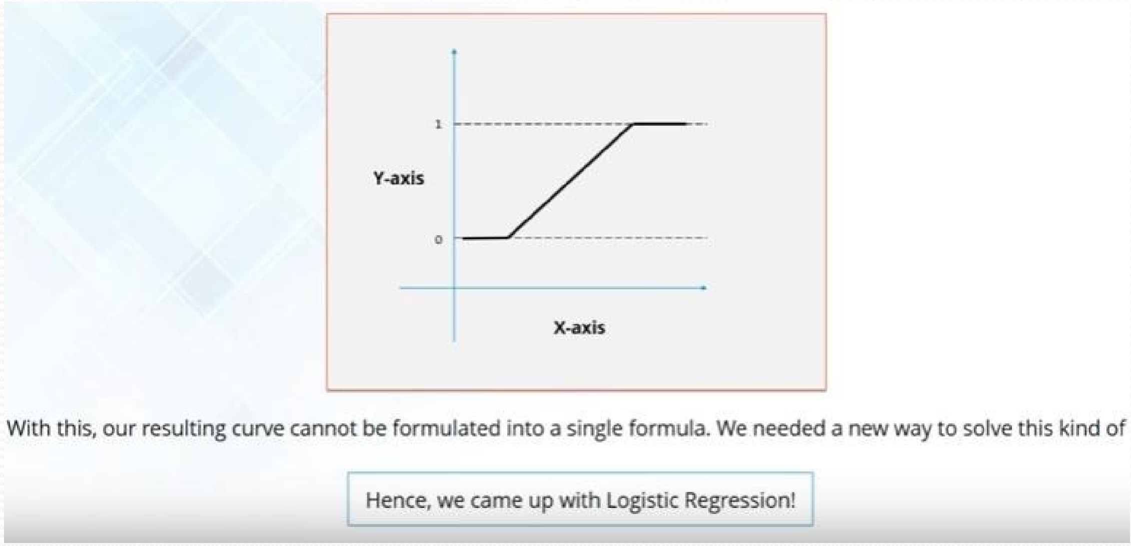
Real-life Challenges

- Gaming - Win vs Loss
- Sales - Buying vs Not buying
- Marketing – Response vs No Response
- Credit card & Loans – Default vs Non Default
- Operations – Attrition vs Retention
- Websites – Click vs No click
- Fraud identification –Fraud vs Non Frau
- Healthcare –Cure vs No Cure

Why Logistic Regression/Logit

- Whenever the outcome of the Dependent variable is discrete, like 0/ 1, True/ False, Yes/ No, A, B, C
- The model guarantees the probability of an observation belonging to a particular group is 0 or 1.
- No matter how hard we try, OLS regression probability will be outside 0-1 range in all likelihood.





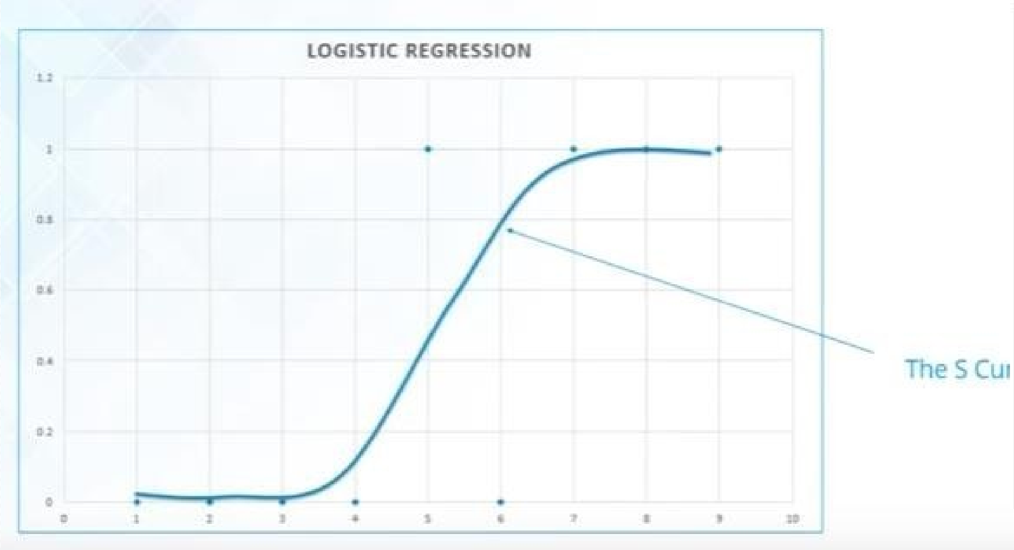
With this, our resulting curve cannot be formulated into a single formula. We needed a new way to solve this kind of

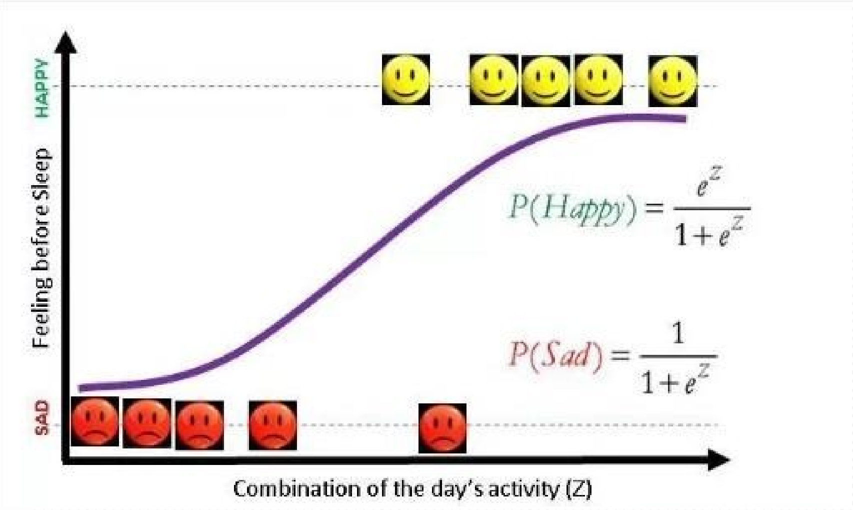
Hence, we came up with Logistic Regression!

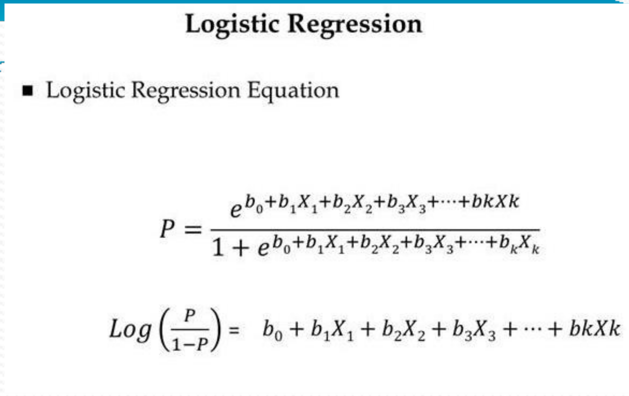
Some concepts

- Binary logit model commonly deals with the issue of classifying an observation into one of two groups. In this sense it is similar to two group discriminant analysis.

- Sigmoid Curve also known as S Curve.
- Odds are used to counteract the fact that linear regression produces probability outside the range of 0 and 1. Going with Odds forces the upper bound on the probability. Lower bound is achieved by taking natural Log of regression value.




HAPPY
Feeling before Sleep
SAD
$P(Happy) = \frac{e^Z}{1+e^Z}$
$P(Sad) = \frac{1}{1+e^Z}$
Combination of the day's activity (Z)



Logistic Regression

- Logistic Regression Equation

$$P = \frac{e^{b_0+b_1X_1+b_2X_2+b_3X_3+\cdots+bkXk}}{1+e^{b_0+b_1X_1+b_2X_2+b_3X_3+\cdots+b_kX_k}}$$

$$Log\left(\frac{P}{1-P}\right) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + bkXk$$

- One way to conceptualize (non-technically) the **probability** of an event is the number of ways that an event can occur divided by the total number of possible outcomes. The odds for an event is the ratio of the number of ways the event can occur to the number of ways it does not occur.
- An odds ratio (OR) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.
- Logit is successful due to Odds Ratio.
- When a logistic regression is calculated, the regression coefficient (b1) is the estimated increase in the log odds of the outcome per unit increase in the value of the exposure.
- The 95% confidence interval (CI) is used to estimate the precision of the OR.
- Maximum Likelihood Estimation decides the slopes/individual coefficients.
- Logit and Probit forms the Bell curve.

Steps involved in Logistic Regression

1. **Log Likelihood Ratio Test :** A statistical test used for comparing the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model).

2. **Pseudo R Square:** Indicates the Model robustness, R2 reduces the uncertainty produced by the intercept model.

3. **Individual Coefficients:** Indicates the statistical importance of the predictor variables.

4. **Odds – EXP:** Practical Importance of the predictor variables.

5. **Predictor variable Importance:** Methods to know the Predictor variables importance.

6. **Confusion/Classification Matrix/Table:** A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

7. **ROC Plot:** A receiver operating characteristic **curve**, i.e. **ROC curve**, is a graphical **plot** that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

8. **Adjusting the Cutoff threshold values.**

Case Study : Diabetes Dataset

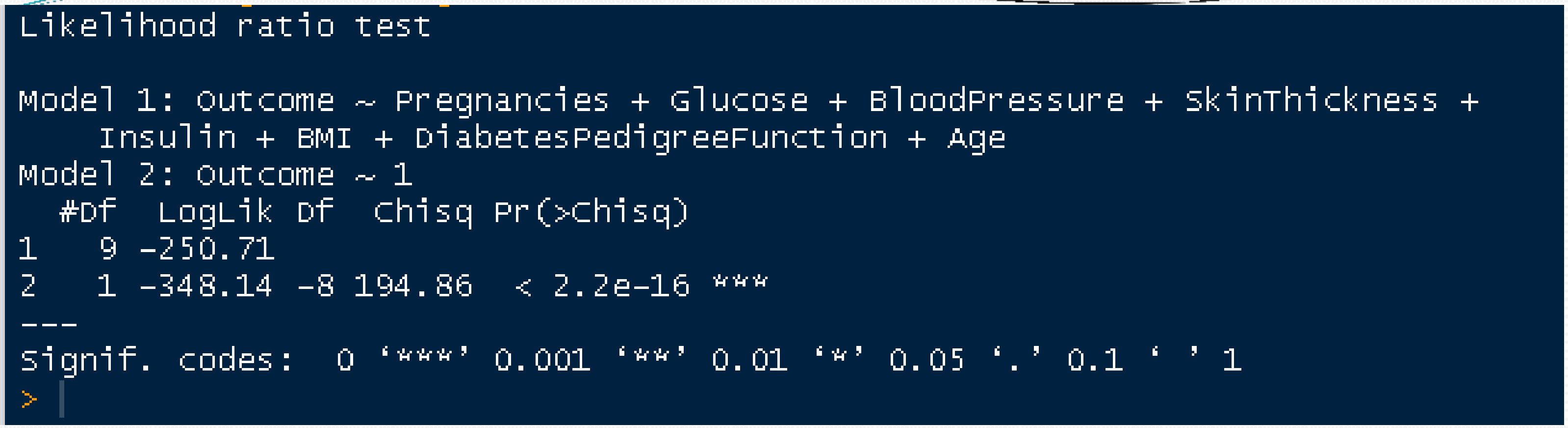
Here is an interesting Diabetes Dataset, In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- Attribute Information:
  1. Number of times pregnant
  2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  3. Diastolic blood pressure (mm Hg)
  4. Triceps skin fold thickness (mm)
  5. 2-Hour serum insulin (mu U/ml)
  6. Body mass index (weight in kg/(height in m)^2)
  7. Diabetes pedigree function
  8. Age (years)
  9. Class variable (0 or 1)

Exercise

- Split the datasets into Training and Test (70:30).
- Develop a logistic regression model, obtain the output.
- Interpret the results and predict Diabetes occurrence.

Log Likelihood Ratio Test



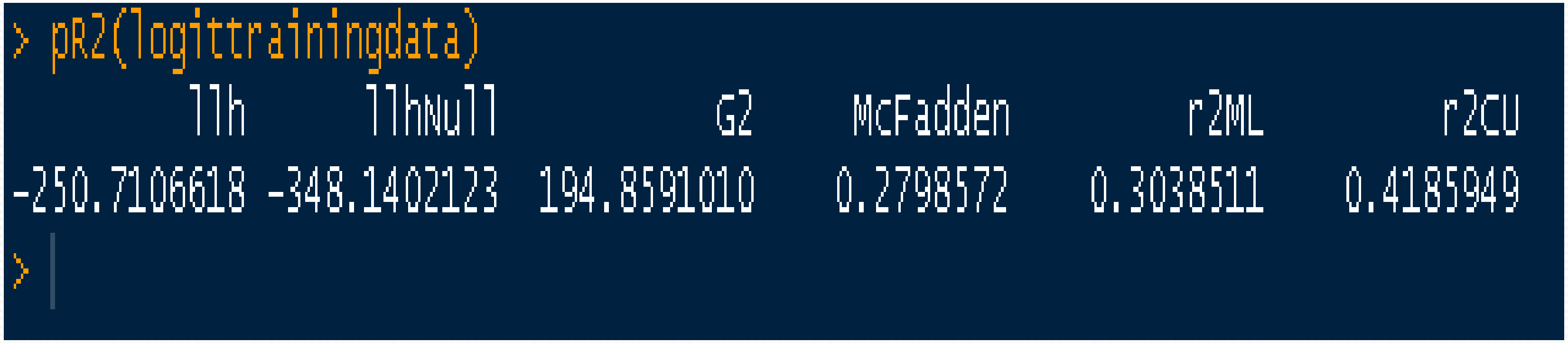
```
Likelihood ratio test

Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
    Insulin + BMI + DiabetesPedigreeFunction + Age
Model 2: Outcome ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -250.71
2   1 -348.14 -8 194.86  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Interpretation

- The overall test of model significance based on the ChiSq test above is indicating the likelihood of outcome depends on predictor variables.

- Using the statistical Language this implies that the null hypothesis is of all Betas are Zero
  is rejected and we conclude that at least one Beta is on zero.

Pseudo R Square



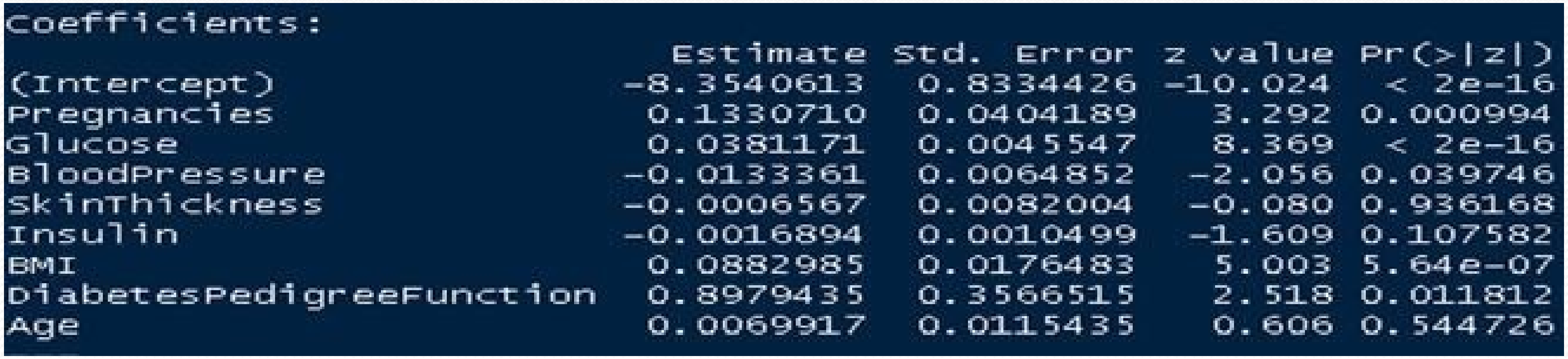
```
> pR2(logittrainingdata)
         llh      llhNull           G2     McFadden         r2ML         r2CU
-250.7106618 -348.1402123  194.8591010    0.2798572    0.3038511    0.4185949
> 
```

Interpretation

- Based on McFaden R Square, we conclude that 27.98 percent of the uncertainity of the
  Intercept only Model (Model2) has been explained by the Full model (Model 1).

- Thus the goodness of fit is reasonably robust.

Individual Coefficients

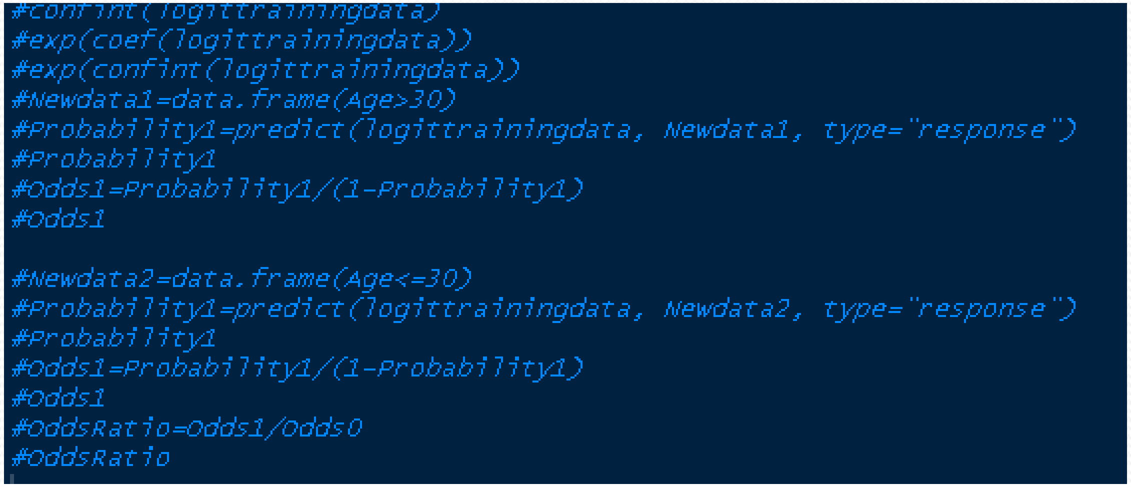


```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.3540613  0.8334426 -10.024  < 2e-16
Pregnancies               0.1330710  0.0404189   3.292 0.000994
Glucose                   0.0381171  0.0045547   8.369  < 2e-16
BloodPressure            -0.0133361  0.0064852  -2.056 0.039746
SkinThickness            -0.0006567  0.0082004  -0.080 0.936168
Insulin                  -0.0016894  0.0010499  -1.609 0.107582
BMI                       0.0882985  0.0176483   5.003 5.64e-07
DiabetesPedigreeFunction  0.8979435  0.3566515   2.518 0.011812
Age                       0.0069917  0.0115435   0.606 0.544726
```

Interpretation

- Only Attitude towards Glucose/Pregnancies is significant in explaining Outcome.
- Does this imply that focus should be on fostering more positive attitude on Glucose levels
  and not worry too much about the Skin thickness/Insulin??
  - Yes as per statistical significance however No as per practical
    significance (ODDS)

ODDS – EXP



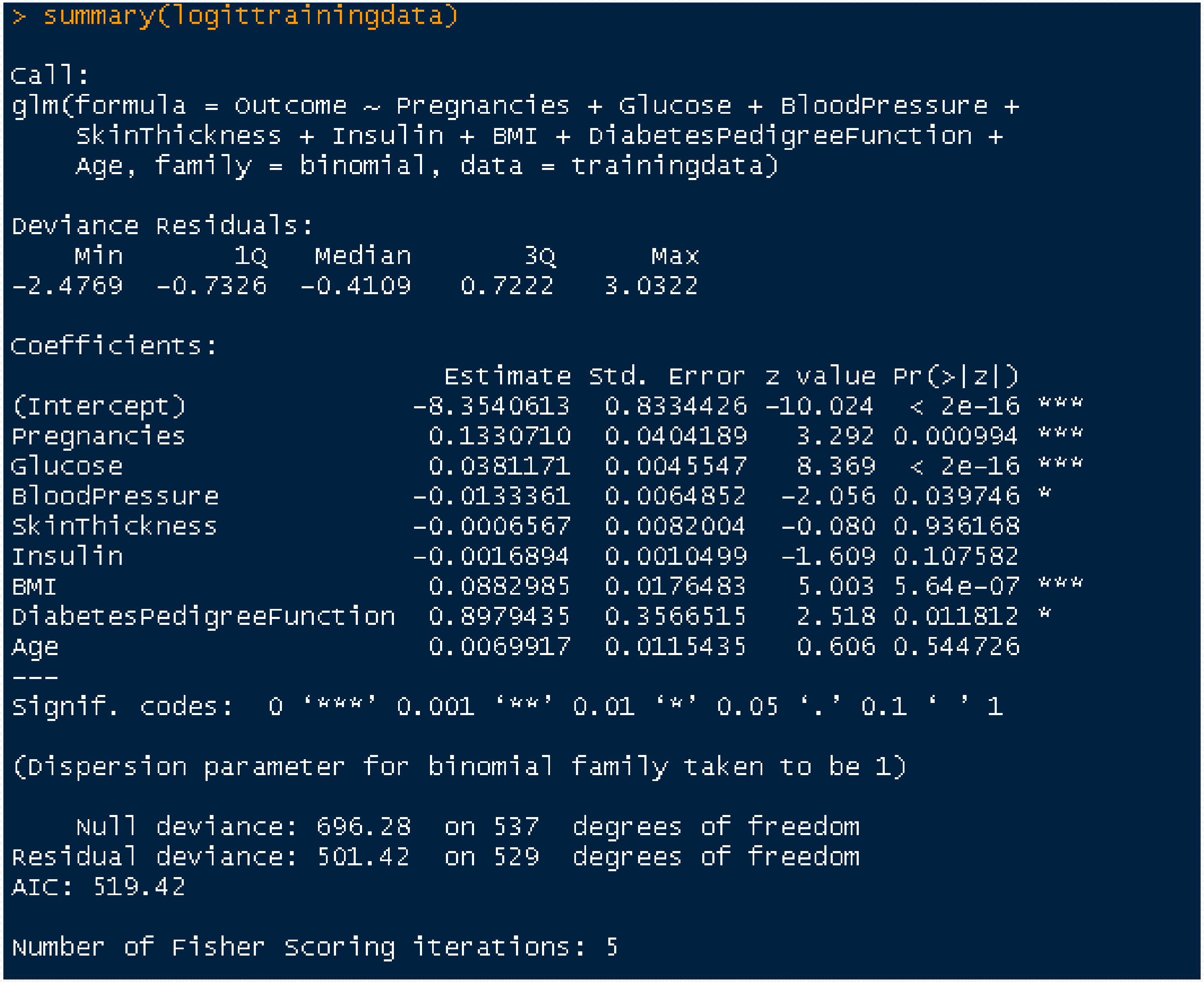
```
#confint(logittrainingdata)
#exp(coef(logittrainingdata))
#exp(confint(logittrainingdata))
#Newdata1=data.frame(Age>30)
#Probability1=predict(logittrainingdata, Newdata1, type="response")
#Probability1
#Odds1=Probability1/(1-Probability1)
#Odds1

#Newdata2=data.frame(Age<=30)
#Probability1=predict(logittrainingdata, Newdata2, type="response")
#Probability1
#Odds1=Probability1/(1-Probability1)
#Odds1
#OddsRatio=Odds1/Odds0
#OddsRatio
```

Interpretation

- Basket 1 : Statistically significant and Odds is more than 1.
- Basket 2 : Statistically NOT significant and Odds is more than 1
- Basket 3 : Statistically significant and Odds less than 1
- Basket 4 : Statistically NOT significant and Odds also less than 1

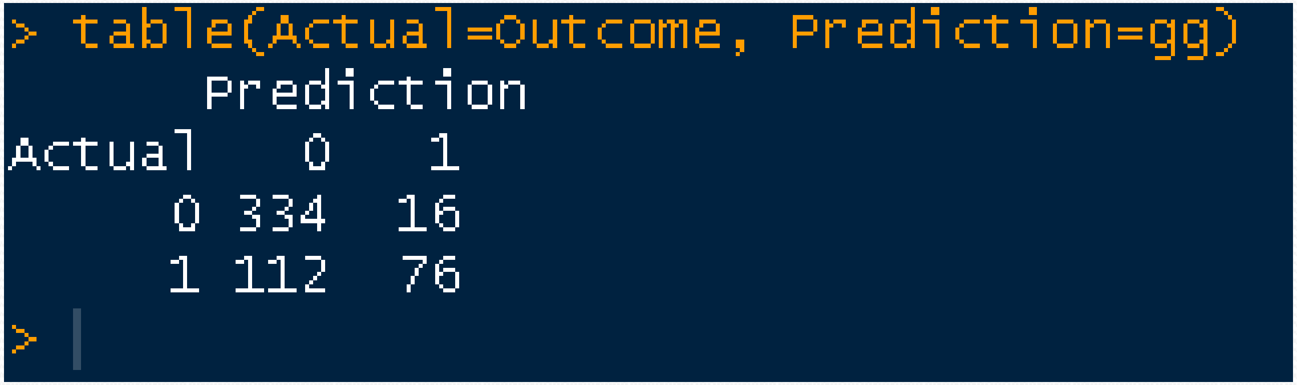
	Statistically	ODDS RATIO
	Significant	>1
Basket1	1	1
Basket2	0	1
Basket3	1	0
Basket4	0	0

Predictor variable Importance



```
> summary(logittrainingdata)

Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure + 
    SkinThickness + Insulin + BMI + DiabetesPedigreeFunction + 
    Age, family = binomial, data = trainingdata)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4769  -0.7326  -0.4109   0.7222   3.0322  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -8.3540613  0.8334426 -10.024  < 2e-16 ***
Pregnancies               0.1330710  0.0404189   3.292 0.000994 ***
Glucose                   0.0381171  0.0045547   8.369  < 2e-16 ***
BloodPressure            -0.0133361  0.0064852  -2.056 0.039746 *  
SkinThickness            -0.0006567  0.0082004  -0.080 0.936168    
Insulin                  -0.0016894  0.0010499  -1.609 0.107582    
BMI                       0.0882985  0.0176483   5.003 5.64e-07 ***
DiabetesPedigreeFunction  0.8979435  0.3566515   2.518 0.011812 *  
Age                       0.0069917  0.0115435   0.606 0.544726    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 696.28  on 537  degrees of freedom
Residual deviance: 501.42  on 529  degrees of freedom
AIC: 519.42

Number of Fisher Scoring iterations: 5
```

Classification Table

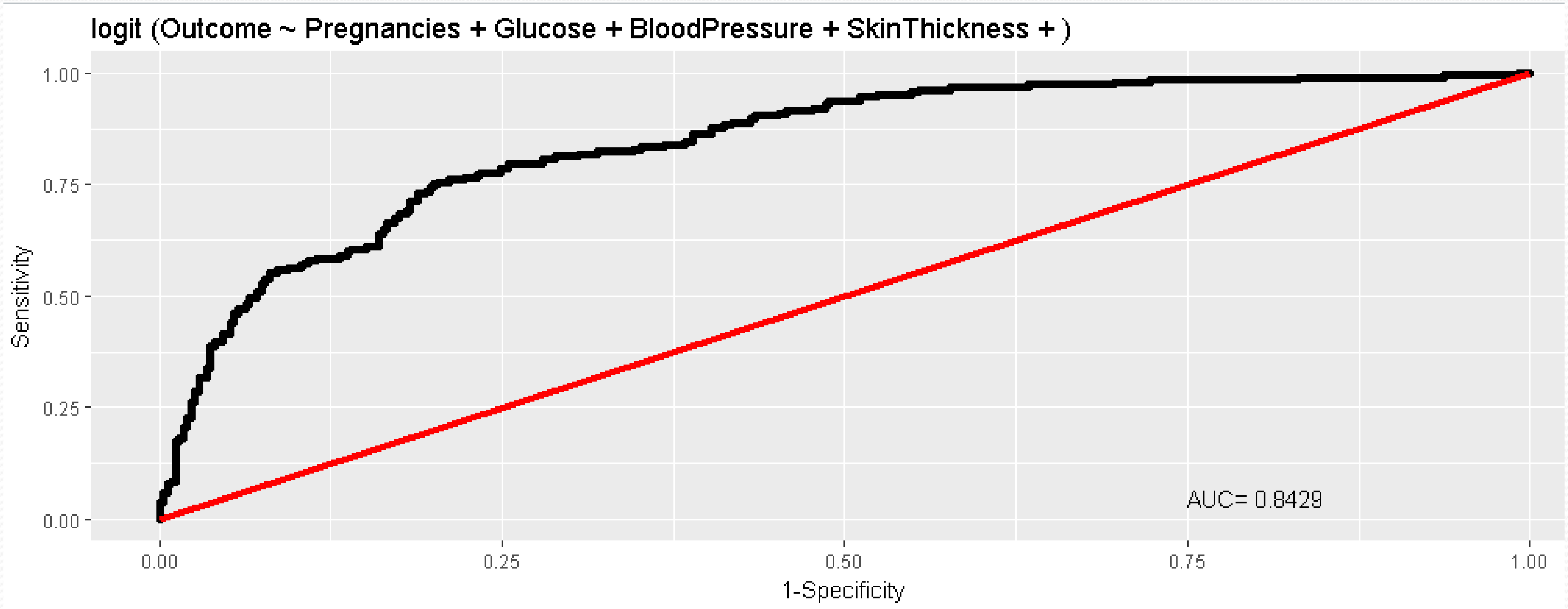


```
> table(Actual=Outcome, Prediction=gg)
      Prediction
Actual   0   1
     0 334  16
     1 112  76
> 
```

Interpretation

- **410 out of 538 observations are correctly classified indicating overall accuracy of 76.2 percent.**
- **Also out of 350 No diabetic cases, model has correctly predicted 334 cases and gone wrong in 16 cases.**
- **Likewise, out of 188 Diabetic cases, model correctly predicted 76 cases and gone wrong in 112 cases.**
- **There is an issues between Type1 (Sensitivity) and Type2 (1-specificity) errors.**

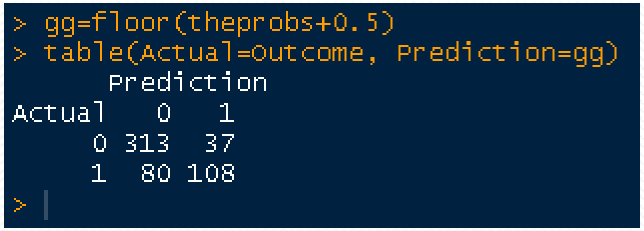
ROC Plot



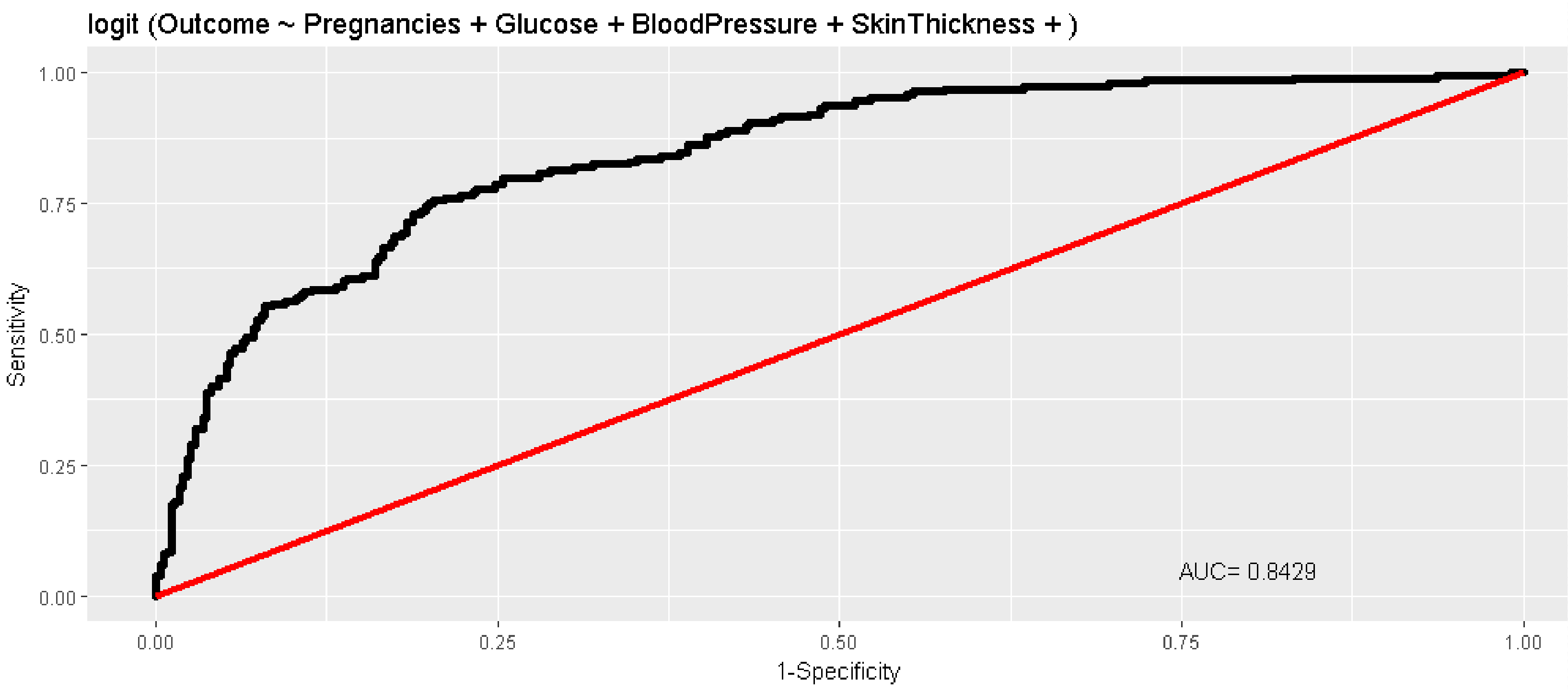
Interpretation

- **Usually curve should not be near to the fit line.**
- **AUC > 70% is reasonably good.**

Adjusting the Cutoff/Threshold values



```
> gg=floor(theprobs+0.5)
> table(Actual=Outcome, Prediction=gg)
      Prediction
Actual   0   1
     0 313  37
     1  80 108
> |
```



Rcodes

```
# Set the working Directory
#setwd("D:/DataSscience/10_wipro/Talks/LogisticRegression")
setwd("F:/0_1_Trainings/Wipro")

# Import the Data Set
diabetesdata=read.csv("diabetes.csv", header=TRUE)

# Split the Data set into Training and Test in 70:30 proportion
# First run the model on Traininf Data and then validate it with Test data.
library(caret)
set.seed(123)
index <- createDataPartition(diabetesdata$Outcome, p=0.70, list=FALSE)
trainingdata <- diabetesdata[ index,]
testdata <- diabetesdata[-index,]

# List the Dimensions
dim(trainingdata)
dim(testdata)

### Predict and check on the Training Data
attach(trainingdata)
#Model Fitting
logittrainingdata=glm(Outcome~Pregnancies+Glucose+BloodPressure+SkinThickness+Insulin+BMI+DiabetesPedigreeFunction+Age, data=trainingdata, family=binomial)
#logittrainingdata=glm(Outcome~., data=trainingdata, family=binomial)
#logittrainingdata=glm(Outcome~.-SkinThickness, data=trainingdata, family=binomial)
```

R codes

- #**Step 1 Log Likelihood Ratio Test **#
- library(lmtest)
- lrtest(logittrainingdata)
- #**Step 2 Pseudo R Square **#
- library(pscl)
- pR2(logittrainingdata)
- #**Step 3 Individual Coefficients**#
- summary(logittrainingdata)
- #**Step 4 ODDS RATIO - EXP ***#
- #**Step 5 Predictor Variable Importance***************************************#
- #confint(logittrainingdata)
- #exp(coef(logittrainingdata))
- #exp(confint(logittrainingdata))
- #Newdata1=data.frame(Age>30)
- #Probability1=predict(logittrainingdata, Newdata1, type="response")
- #Probability1
- #Odds1=Probability1/(1-Probability1)
- #Newdata2=data.frame(Age<=30)
- #Probability1=predict(logittrainingdata, Newdata2, type="response")
- #Probability1
- #Odds1=Probability1/(1-Probability1)
- #OddsRatio=Odds1/Odds0
- #**Step 6 Confusion Matrix/Classification Table******************************#
- theprobs=fitted(logittrainingdata)
- gg=floor(theprobs+0.3)
- table(Actual=Outcome, Prediction=gg)
- #**Step 7 ROC Plot

THANK YOU !!!