Naïve Bayes Classifier
ML ALGORITHM

DATA SCIENCE
INSTITUTE

MACHINE LEARNING METHODS

There is lot of overlapping between statistical modeling and machine learning. The Regression Models are used extensively in ML applications.

Regression and Machine Learning Algorithms

- Linear and Logistic Regression
- Naive Bayes Classifier
- Support Vector Machines
- K Nearest Neighbours Method (KNN)
- Decision Trees
- Random Forest
- Neural Networks

Naive Bayes is a probabilistic classification technique

Naive Bayes Classifier

- A **naive Bayes classifier** is a simple probabilistic classifier based on Bayes' theorem.
- It can be used as an alternative method to Logistic Regression
- It is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

DATA SCIENCE INSTITUTE

Conditional Probability

The **conditional probability** of an event B is the probability that the event B will occur given the knowledge that an event A has already occurred.

This probability is written $P(B|A)$.

If A and B are independent events then $P(B|A)=P(B)$

Example: An unbiased die with numbers 1-6 is tossed.

A: Getting a number greater than 1

B: Getting an even number

$P(A)= 5/6$ and $P(B)=3/6=1/2$

$P(B|A)= 3/5$ since sample space is reduced to 5 points given that A has already occurred.

DATA SCIENCE INSTITUTE

Bayes Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Where

P(A) is the prior probability or marginal probability of A.

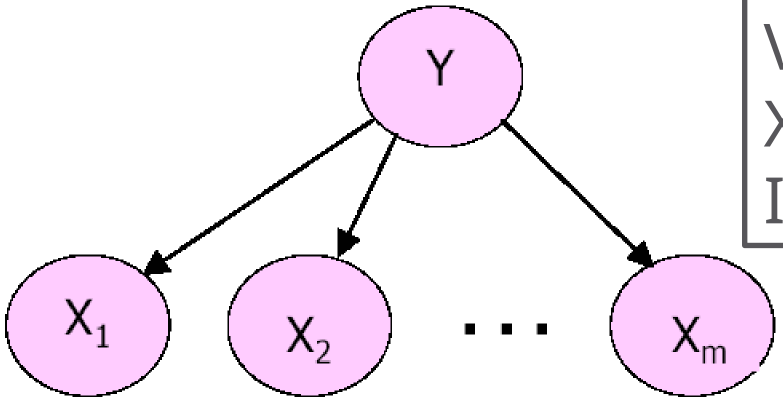
P(A|B) is the conditional probability of A, given B.

P(B|A) is the conditional probability of B given A.

P(B) is the prior or marginal probability of B.

DATA SCIENCE INSTITUTE

Naïve Bayes Framework



Y : Categorical Dependent Variable
X_i : Categorical/Continuous Independent Variable

<u>Objective</u>: To estimate Y given the values of X_i's or $P(Y|X_1\ X_2\ ....X_m)$ using the Naïve Bayes Classifier.

<u>Assumption</u>: All X_i's are conditionally independent of each other.

Example

Consider a simple example where dependent variable Y is binary variable and there are 2 independent

variables X1 and X2.

We classify Y = 1 as potential buyer of a certain product

Y = 0 otherwise

Let X1 denote age of the individual

X1 = 0 for age group 25-30 yrs.

= 1 for age group 31-40 yrs.

Let X2 denote gender

X2 = 0 if Gender=female

= 1 if Gender=male

Classification Rule

For the given values of age(X1)and gender(X2), we want to classify the customer as potential buyer or not.

Using Naïve Bayes Classifier we estimate following 2 conditional probabilities:

$$P(Y=0/X1=a1, X2=a2) \text{ and } P(Y=1/X1=a1,X2=a2);$$

here a1 and a2 are values of X1 and X2 for a particular customer.

We classify $Y = 0$ if $P(Y=0/X1=a1, X2=a2) > 0.5$

OR $Y = 1$ if $P(Y=1/X1=a1,X2=a2) > 0.5$

Expected Output

Once the classification rule is applied the output can be shown as follows:

Case#	X1	X2	P(Y=1/X1,X2)	P(Y=0/X1,X2)	Y classified as
1	1	0	0.29	0.71	0
2	1	1	0.65	0.35	1
240	0	0	0.51	0.49	1

DATA SCIENCE INSTITUTE

Advantages of the Method

- Classification rule is simple to understand.
- The method requires a small amount of training data to estimate the parameters necessary for classification.
- The evaluation of the classifier is quick and easy.
- The method can be a good alternative to logistic regression.

Naïve Bayes Method
Few Comments

- Assumption of conditional independence of the independent variables is highly impractical.
- In case of continuous independent variables the density function must be known or assumed to be normal.
- In case of categorical independent variables the probabilities cannot be calculated if the count in any conditional category is zero. For ex.: if there are no 'Y=1' in the age group 25-30 yrs. then
  $P(X1=0/Y=1) = 0$.

  In that case we replace 0 by 0.5/n (n = sample size) so that the probability expression does not reduce to zero.

  The function in R for Naive Bayes classifier performs Laplace smoothing to handle this problem.

Example

Employee churn model.

Independent variables are:

- Gender
- Experience Level (<3, 3-5 and >5 years)
- Function (Marketing, Finance,Client Servicing)
- Source (Internal or External)

Dependent variables is "status" (=1 if employee left within 18 months from Joining date)

Data Snapshot

sn	status	function	exp	gender	source
1	1	CS	<3	M	external
2	1	CS	<3	M	external
3	1	CS	>=3 and <=5	M	internal
4	1	MARKETING	<3	M	external
5	1	FINANCE	<3	F	internal
6	1	CS	>=3 and <=5	F	internal
7	1	MARKETING	<3	F	internal
8	1	FINANCE	<3	F	external
9	1	CS	<3	M	internal
10	1	CS	>5	M	external
11	1	CS	>5	F	external
12	1	CS	<3	F	external
13	1	MARKETING	<3	M	external
14	1	FINANCE	<3	M	external
15	1	FINANCE	<3	M	internal
16	1	CS	<3	M	external
17	1	CS	<3	F	internal
18	1	CS	<3	F	internal
19	1	MARKETING	>=3 and <=5	M	internal
20	1	FINANCE	<3	M	external

DATA SCIENCE INSTITUTE

Naive Bayes Method in R
Employee Churn Data

```
empdata<- read.csv(file.choose())
library(e1071)
churnmodel<-naiveBayes(status~function. + exp + gender + source,
    data=empdata)
churnmodel
```

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.6024096 0.3975904 

Conditional probabilities:
   function.
Y          CS   FINANCE MARKETING
  0 0.1600000 0.4000000 0.4400000
  1 0.5454545 0.2424242 0.2121212

   exp
Y          <3 >=3 and <=5        >5
  0 0.2000000   0.2000000 0.6000000
  1 0.7575758   0.1212121 0.1212121

   gender
Y           F         M
  0 0.4600000 0.5400000
  1 0.4242424 0.5757576

   source
Y    external  internal
  0 0.3400000 0.6600000
  1 0.5454545 0.4545455
```

Output shows a list of tables, one for each predictor variable.

For each categorical variable the conditional probabilities for each class.

Eg. For gender, conditional probabilities for Females are 0.46 for no churn and 0.4242 for churn

DATA SCIENCE INSTITUTE

Naive Bayes Method in R…
Predicted Probabilities and ROC

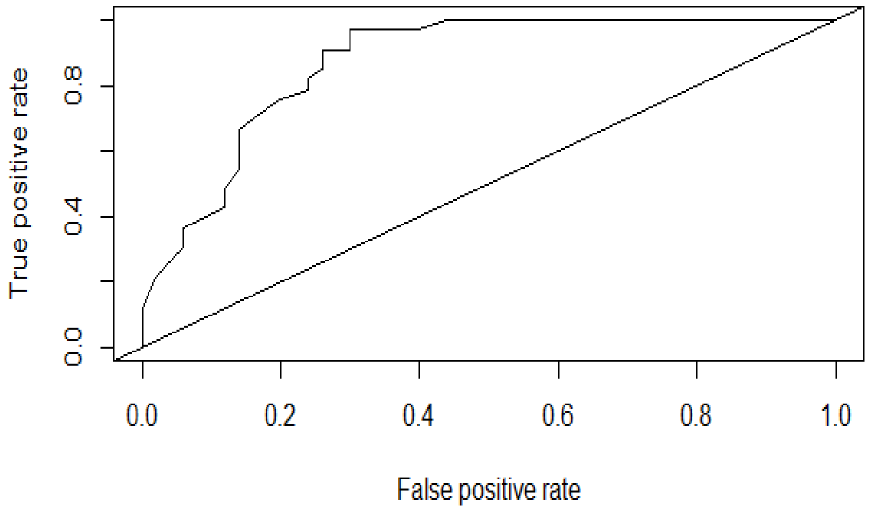
```
# Predicted probabilities and ROC Curve
prednb<-predict(churnmodel,empdata,type='raw')

library(ROCR)
pred<-prediction(prednb[,2],empdata$status)
perf<-performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)

# Area under ROC Curve in R (AUC)
auc<-performance(pred,"auc")
auc@y.values
```

[1] 0.8706061

```
> prednb
               0          1
 [1,] 0.06419224 0.93580776
 [2,] 0.06419224 0.93580776
 [3,] 0.49966736 0.50033264
 [4,] 0.53586540 0.46413460
 [5,] 0.13777290 0.86222710
 [6,] 0.56258650 0.43741350
 [7,] 0.59789573 0.40210427
 [8,] 0.07347548 0.92652452
 [9,] 0.06419224 0.93580776
[10,] 0.15592410 0.84407590
```



Predicted Probability for 1st Case

Naive Bayes Calculations (1st row)

① $P(Y=1) = \dfrac{\text{Number of values with } Y=1}{\text{Total number of values}} = \dfrac{33}{83}$

$P(Y=0) = \dfrac{50}{83}$

② $P(X_1 = \text{CS} \mid Y=1) = \dfrac{18}{33}$ $\quad$ $P(X_1 = \text{es} \mid Y=0) = \dfrac{8}{50}$

③ $P(X_2 = \text{"<3"} \mid Y=1) = \dfrac{25}{33}$ $\quad$ $P(X_2 = \text{"<3"} \mid Y=0) = \dfrac{10}{50}$

④ $P(X_3 = M \mid Y=1) = \dfrac{19}{33}$ $\quad$ $P(X_3 = M \mid Y=0) = \dfrac{27}{50}$

⑤ $P(X_4 = \text{external} \mid Y=1) = \dfrac{18}{33}$ $\quad$ $P(X_4 = \text{external} \mid Y=0) = \dfrac{17}{50}$

Predicted Probability for 1st Case

$$P(y=1 \mid \underline{X}) = \frac{\frac{33}{83} \times \frac{18}{33} \times \frac{25}{33} \times \frac{19}{33} \times \frac{18}{33}}{\frac{33}{83} \times \frac{18}{33} \times \frac{25}{33} \times \frac{19}{33} \times \frac{18}{33} + \frac{50}{83} \times \frac{8}{50} \times \frac{10}{50} \times \frac{27}{50} \times \frac{17}{50}}$$

$$= \boxed{0.9358}$$

Case Study: Predicting Defaulters

Objective : To predict whether the customer applying for the loan will be a defaulter.

Predictors: Age group, Years at current address, Years at current employer, Debt to Income Ratio, Credit Card Debts, Other Debts.

Dependent Variable: Defaulter (=1 if defaulter ,0 o.w.)

Sample Size:700

DATA SCIENCE INSTITUTE

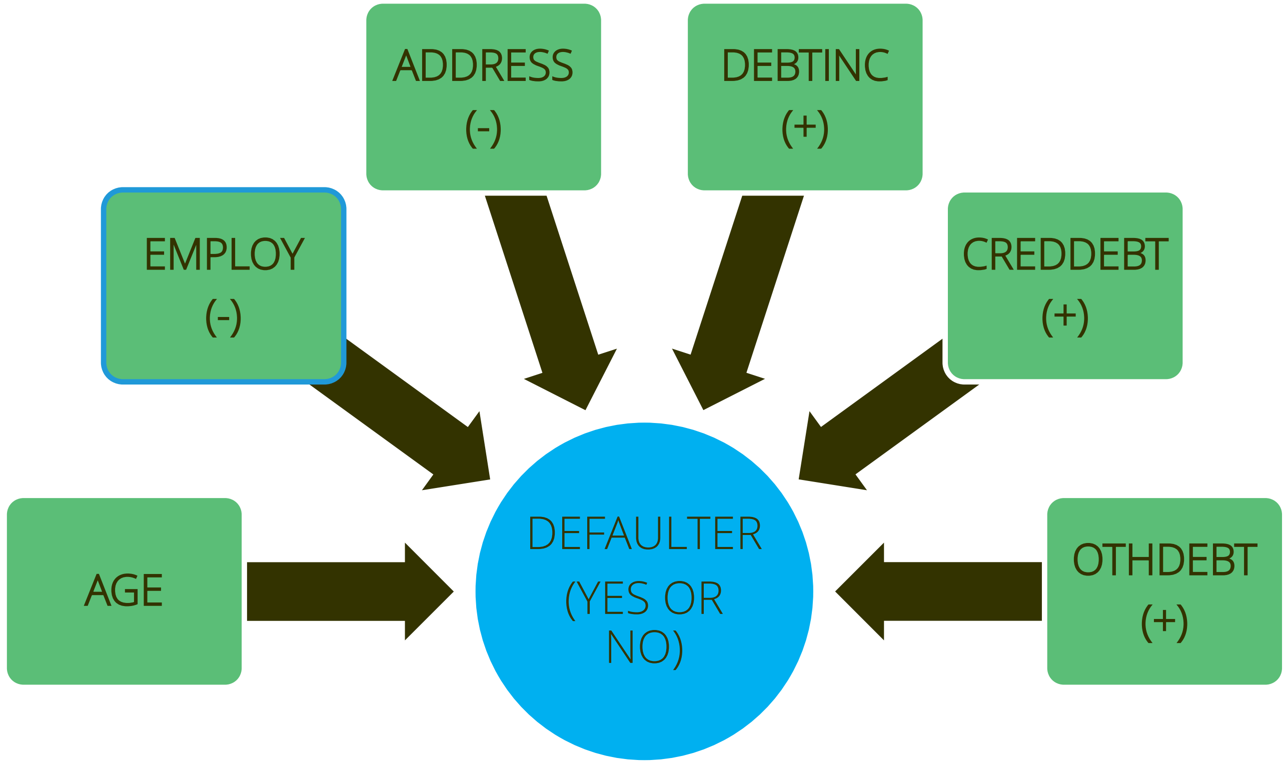
Snapshot of the Data

AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTER
3	17	12	9.30	11.36	5.01	1
1	10	6	17.30	1.36	4.00	0
2	15	14	5.50	0.86	2.17	0
3	15	14	2.90	2.66	0.82	0
1	2	0	17.30	1.79	3.06	1
3	5	5	10.20	0.39	2.16	0
2	20	9	30.60	3.83	16.67	0
3	12	11	3.60	0.13	1.24	0
1	3	4	24.40	1.36	3.28	1
2	0	13	19.70	2.78	2.15	0
1	0	1	1.70	0.18	0.09	0
1	4	0	5.20	0.25	0.94	0

Age groups: 1 (<28 years), 2(28-40 years), 3 (>40 years)

DATA SCIENCE INSTITUTE

Case Study: Predicting Defaulters…



Logistic Regression in R

```
bankloan$AGE<-as.factor(bankloan$AGE)
riskmodel<-
    glm(DEFAULTER~AGE+EMPLOY+ADDRESS+DEBTINC+CREDDEBT+OTHDEBT,fa
    mily=binomial,data=bankloan)
summary(riskmodel)
```

GLM: Generalized Linear Models

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.78821	0.26407	-2.985	0.00284 **
AGE2	0.25202	0.26651	0.946	0.34433
AGE3	0.62707	0.36056	1.739	0.08201 .
EMPLOY	-0.26172	0.03188	-8.211	< 2e-16 ***
ADDRESS	-0.09964	0.02234	-4.459	8.22e-06 ***
DEBTINC	0.08506	0.02212	3.845	0.00012 ***
CREDDEBT	0.56336	0.08877	6.347	2.20e-10 ***
OTHDEBT	0.02315	0.05709	0.405	0.68517

EMPLOY,ADDRESS,DEBTINC and CREDDEBT are statistically significant.

ATA SCIENCE INSTITUTE

Excluding 'AGE' and 'OTHDEBT'

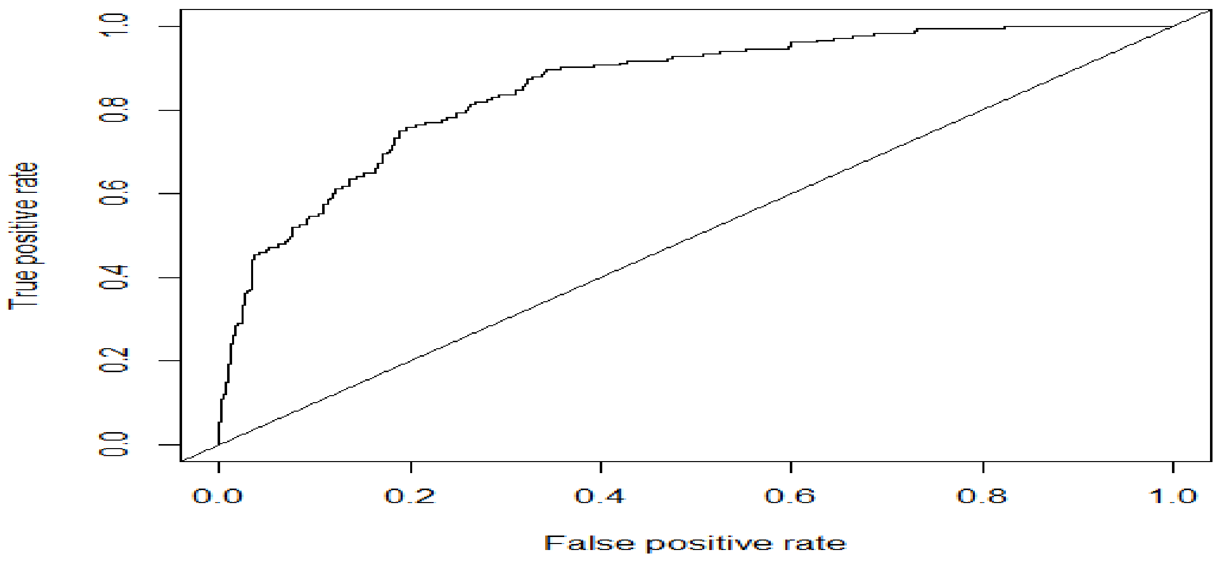
```
riskmodel<-glm(DEFAULTER~EMPLOY+ADDRESS+DEBTINC+CREDDEBT,
                    family=binomial,data=bankloan)
summary(riskmodel)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.79107	0.25154	-3.145	0.00166 **
EMPLOY	-0.24258	0.02806	-8.646	< 2e-16 ***
ADDRESS	-0.08122	0.01960	-4.144	3.41e-05 ***
DEBTINC	0.08827	0.01854	4.760	1.93e-06 ***
CREDDEBT	0.57290	0.08725	6.566	5.17e-11 ***

EMPLOY,ADDRESS,DEBTINC and CREDDEBT are statistically significant.

ROC Curve in R

```
library(ROCR)
bankloan$predprob<-fitted(riskmodel)
pred<-prediction(bankloan$predprob,bankloan$DEFAULTER)
perf<-performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)
```

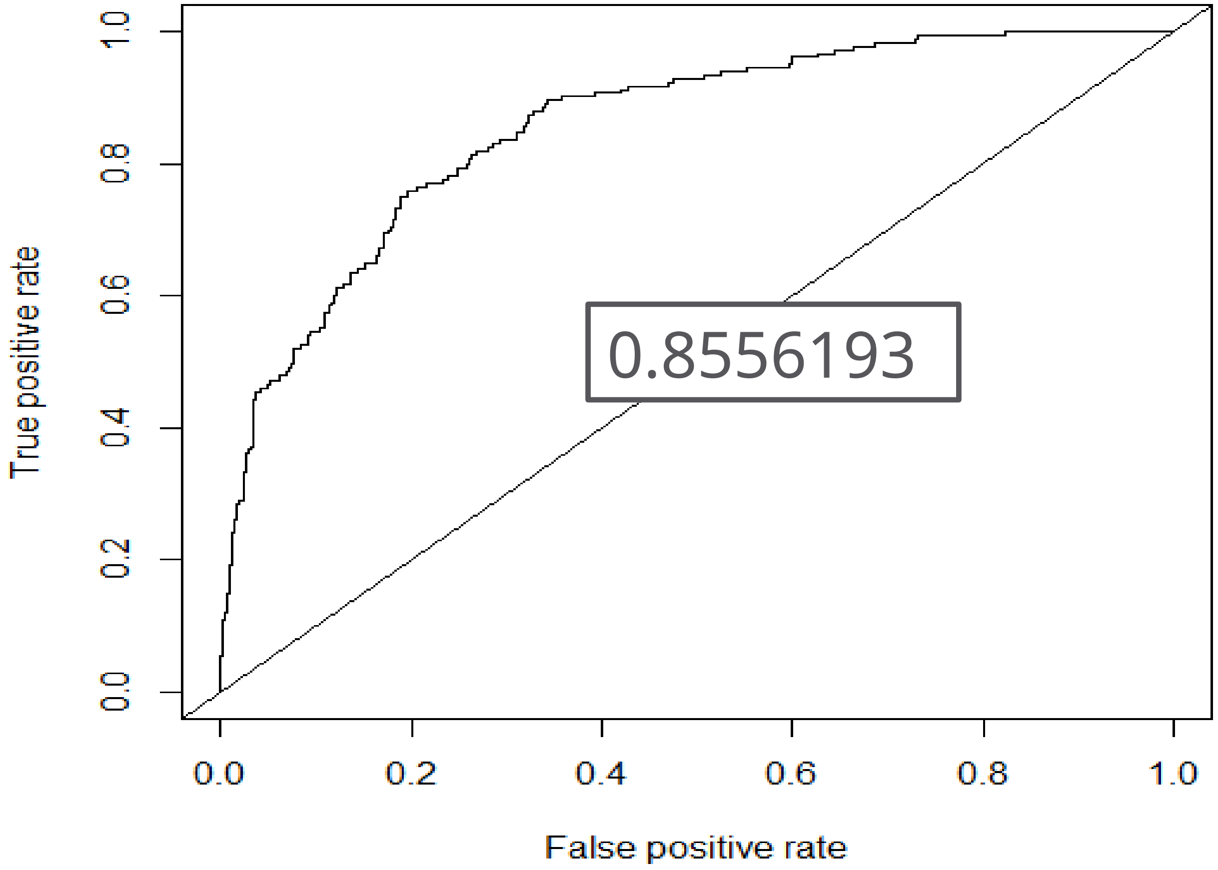


DATA SCIENCE INSTITUTE

Area Under ROC Curve in R

```
auc<-performance(pred,"auc")
auc@y.values
```

0.8556193



Naive Bayes Method in R…

library(e1071)

riskmodel2<-naiveBayes(DEFAULTER~EMPLOY + ADDRESS+ DEBTINC + CREDDEBT,data=bankloan)

riskmodel2

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.7385714 0.2614286 

Conditional probabilities:
   EMPLOY
Y       [,1]     [,2]
  0 9.508704 6.663741
  1 5.224044 5.542946

   ADDRESS
Y       [,1]     [,2]
  0 8.945841 7.000621
  1 6.393443 5.925208

   DEBTINC
Y        [,1]     [,2]
  0  8.679304 5.615197
  1 14.727869 7.902798

   CREDDEBT
Y       [,1]     [,2]
  0 1.245397 1.422238
  1 2.423770 3.232645
```

Output shows a list of tables, one for each predictor variable.

For each categorical variable the conditional probabilities for each class.

For each numeric variable, for each target class, mean and standard deviation

Eg. For EMPLOY, mean for churn status = 0 is 9.58 and sd is 6.66

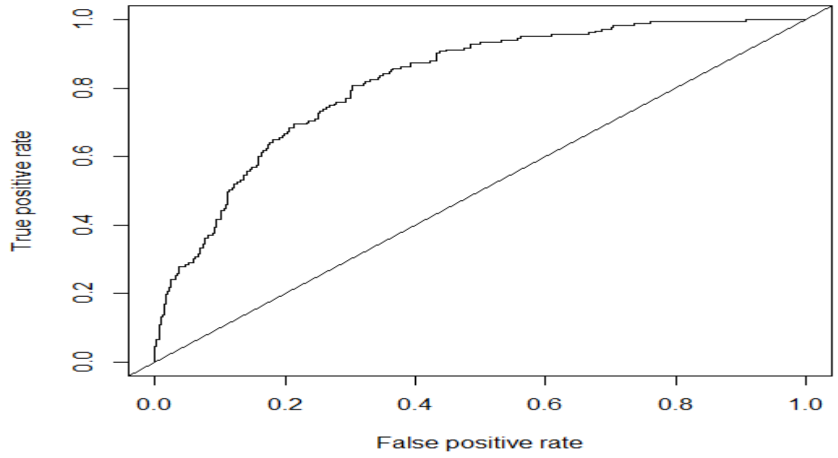
Naive Bayes Method in R…
Predicted Probabilities and ROC

```
# Predicted probabilities and ROC Curve
prednb<-predict(riskmodel2,bankloan,type='raw')

library(ROCR)
pred<-prediction(prednb[,2],bankloan$DEFAULTER)
perf<-performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)

# Area under ROC Curve in R (AUC)
auc<-performance(pred,"auc")
auc@y.values

[1] 0.8188266
```



DATA SCIENCE INSTITUTE

Naive Bayes Method in R…
Confusion Matrix

```
# Generating a Confusion Matrix
library(caret)
prednb<-predict(riskmodel2,bankloan,type='raw')
bankloan$predY<-ifelse(prednb[,2]>0.25,1,0)
confusionMatrix(factor(bankloan$predY),factor(bankloan$DEFAULTER))
```

```
Confusion Matrix and Statistics

                Reference
Prediction     0     1
         0   444    83
         1    73   100

        Accuracy : 0.7771
```

Try different cut-off values to get best model.

DATA SCIENCE INSTITUTE

Read More ....

Naïve Bayes Method in R…
Laplace Smoothing

- If a given class and feature value never occur together in the training data, then the frequency-based probability estimate will be zero.
- This is problematic because it will wipe out all information in the other probabilities when they are multiplied.
- Therefore, it is often desirable to incorporate a small-sample correction, called pseudo-count, in all probability estimates such that no probability is ever set to be exactly zero.
- This way of regularising naive Bayes is called Laplace Smoothing when the pseudo count is one, and Lidstone Smoothing in the general case.

```
model<-naiveBayes(Y~X1 + X2+ X3,data=…, laplace=2)
```

Note: naiveBayes in R replace '0' probability by '0.001' for out of sample prediction.

DATA SCIENCE INSTITUTE

Naïve Bayes Method in R…
Laplace Smoothing

This prob will be 0 if numerator count () is 0

Laplace smoothing will replace this probability with a value obtained by the formula:

$$\hat{\theta}\, i = \frac{f_i + a}{N + a\, d}$$

where

: Smoothing Parameter

: Number of observations for

: Number of classes of

Naïve Bayes Method in R…
Laplace Smoothing

Y	X1	X2	X3
0	1	M	A
0	2	M	A
0	2	M	A
0	1	M	A
0	2	F	A
1	2	F	A
1	2	M	B
1	2	M	B
1	2	M	B
1	2	M	B
1	2	F	B
1	2	F	B
1	2	M	B
1	2	M	A
0	2	M	A
0	2	F	A
0	1	F	A
0	1	F	B
0	1	F	B
1	2	M	B
1	2	M	A
1	2	M	A
1	2	M	A
1	2	F	B
1	2	F	B
1	2	F	B
1	2	M	B
1	2	M	B

Variable X1 is a factor with two levels, 1 & 2. There is no observation in the data with X1 =1 when the dependent variable Y =1. Hence, P(X1=1|Y=1) =0. We thus introduce smoothing, to avoid loss of information.

```
data1<-read.csv(file.choose(),header=T)
head(data1)
data1$X1<-as.factor(data1$X1)      #X1 is a factor variable
```

DATA SCIENCE INSTITUTE

Naïve Bayes Method in R…
Laplace Smoothing

```
# First run the default Naive Bayes model
model<-naiveBayes(Y~X1+X2+X3,data=data1)
model         # Check model output, conditional probability for Y=1, X1=1 is 0.0
A-priori probabilities:
Y
      0         1
0.3571429 0.6428571
Conditional probabilities:
        X1
Y       1   2
  0.    0.5 0.5
  1.    0.0 1.0
……
laplacemodel<-naiveBayes(Y~X1+X2+X3,data=data1,laplace=2)
```

laplace= tells R the value of pseudo-count to be used to smoothen the model. There is no rule for choosing the appropriate pseudo-count. The number should be low, but not so low that the resulting probability is almost 0.

DATA SCIENCE INSTITUTE

Naïve Bayes Method in R…
Laplace Smoothing

laplacemodel

```
Naive Bayes Classifier for Discrete Predictors
Call: naiveBayes.default(x = X, y = Y, laplace = laplace)
A-priori probabilities:
Y
        0         1
0.3571429 0.6428571
Conditional probabilities:
   X1
Y            1          2
  0 0.50000000 0.50000000
  1 0.09090909 0.90909091
   X2
Y           F         M
  0 0.5000000 0.5000000
  1 0.3636364 0.6363636
   X3
Y           A         B
  0 0.7142857 0.2857143
  1 0.3181818 0.6818182
```

$$\hat{\theta}\, i = \frac{f_i + a}{N + a\, d} = \frac{2}{18 + 2 \times 2} = 0.09090$$

Since there are no observations for X1=1 and Y=1, x_i=0

=2

Total no. of Y=1 is 18

X1 is a factor with two classes, hence d=2

THANK YOU!!

DATA SCIENCE
INSTITUTE