|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Predictive Modelling  Group Assignment  Section B, Group 7   * Gowri R. Varadhan * Jaya Garg * Jayaram K * Pradeep T * Saurabh Arora * Vivek Sahoo |
| |  |  |  | | --- | --- | --- | | Great Learning, Bangalore | 1/28/18 | PGP - BABI | |

Table of Contents

[Question: 2](#_Toc504941960)

[Solution: 2](#_Toc504941961)

[**Data Sufficiency:** 2](#_Toc504941962)

[**Loading Packages:** 3](#_Toc504941963)

[**Data Analysis:** 3](#_Toc504941964)

[**Assumption Test** 8](#_Toc504941965)

[**Discriminant Analysis** 10](#_Toc504941966)

[File Attachments: 16](#_Toc504941967)

# **Question:**

**Books by Mail from Paul Green**

Books by Mail company is interested in offering a new title called The Art History of Florence to 1000, existing customers. Of these, 83 actually purchased the book, a response rate of 8.3 percent. Hence, the company sent a test mailing to them in this regard. The company also sent out an identical mailing to another 1000 customers to serve as holdout sample. The scope of the study primarily confined to predicting whether a customer will buy the new book or not is based on two input variables namely months since last purchase and number of art books purchased.

Perform Discriminant Analysis using R and Interpret the results using training data. You should incorporate all statistical tests associated with Discriminant and test your accuracy with the test data. Critique the cutoff point probability and suggest what should be the right cutoff.

# **Solution:**

The problem statement requires building a model to better understand whether a particular Customer ID will make the purchase the new book or not, based on the no. of Months since last purchase and number of art books purchase.

**Data Sufficiency:**

**Data Sets**:

* PaulBooks1: Considered as Testing Dataset
* PaulBook2: Considered as Training Dataset
* X Variable: Months, NoBought
* Y Variable: Purchase

**PaulBooks1**

| **Variable** | **Min.** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max.** |
| --- | --- | --- | --- | --- | --- | --- |
| Months | 1 | 7 | 12 | 12.46 | 15 | 35 |
| NoBought | 0 | 0 | 0 | 0.389 | 1 | 5 |

Number of IDs, purchased the book: 917

Number of IDs, did not purchase the book: 83

**PaulBooks2**

| **Variable** | **Min.** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Max.** |
| --- | --- | --- | --- | --- | --- | --- |
| Months | 1 | 7 | 12 | 12.91 | 16 | 35 |
| NoBought | 0 | 0 | 0 | 0.373 | 1 | 3 |

Number of IDs, purchased the book: 919

Number of IDs, did not purchase the book: 81

RCode

data\_df <- read.table("F:/BABI/Group Assignments/Predictive Modelling/PaulBooks2.csv",sep = ",", header = T)

data\_df[,4]<-as.factor(data\_df[,4])

summary(data\_df)

df\_test <- read.table("C:/Jaya/GL/Predictive Modelling/GA/PaulBooks1.csv",sep = ",", header = T)

df\_test <- df\_test[,c(2,3,4)]

df\_test[,3]<-as.factor(df\_test[,3])

summay(data\_test)

**Loading Packages:**

Before we start building models, we start with loading necessary packages. Following are R packages used in building Discriminant Analysis model:

library(MASS)

library(caret)

library(biotools)

library(ggpubr)

library(dplyr)

library(doBy)

library(ggplot2)

library(scales)

library(gtools)

library(ROCR)

library(psych)

library(reshape2)

library(knitr)

library(MVN)

library(Hotelling)

library(gridExtra)

**Data Analysis:**

Next, we start analyzing the training dataset provided:

1. **Finding Missing Values:**

Using R, we find out the missing values in the training dataset and we found out that there are no missing values.

RCode:

cat("\n Variables with number of missing values \n")

sapply(data\_df[,c(2:4)], function(x) sum(is.na(x))) # To report missing values

plot\_missing(data\_df, "Missing Values")

1. **Finding Constant values for the Variables:**

We found out that there are 35 constant values for Months & 4 constant values for NoBought in the training Dataset

RCode:

cat("\n Variables with constant values \n")

sapply(data\_df[,c(2:4)], function(x) length(unique(x)))

1. **Finding Variables having Near Zero Variance:**

In the test conducted, it was observed that there are no variables of near zero variance.

RCode:

nzv <- nearZeroVar(data\_df[,c(2:4)])

nzv

1. **Omitting NA Values**

There were no NA values identified in the training dataset provided.

Now, using Psych library we can further summarize the dataset

**PaulBooks2 (Training dataset)**

| **Purchase** | **Variables** | **n** | **mean** | **Standard Deviation** | **Median** | **trimmed** | **mad** | **range** | **skew** | **kurtosis** | **se** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Months | 919 | 13.26 | 8.19 | 12 | 12.42 | 5.93 | 34 | 0.86 | 0.16 | 0.27 |
| 0 | NoBought | 919 | 0.32 | 0.56 | 0 | 0.21 | 0 | 3 | 1.79 | 3.11 | 0.02 |
| 1 | Months | 81 | 8.93 | 7.25 | 7 | 7.89 | 5.93 | 29 | 1.19 | 1.05 | 0.81 |
| 1 | NoBought | 81 | 1.02 | 0.95 | 1 | 0.92 | 1.48 | 3 | 0.56 | -0.68 | 0.11 |

**PaluBooks1 (Testing dataset)**

| **Purchase** | **Variables** | **n** | **mean** | **Standard Deviation** | **Median** | **trimmed** | **mad** | **range** | **skew** | **kurtosis** | **se** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Months | 917 | 12.73 | 8.11 | 12 | 11.84 | 5.93 | 34 | 0.94 | 0.55 | 0.27 |
| 0 | NoBought | 917 | 0.33 | 0.61 | 0 | 0.21 | 0 | 5 | 2.04 | 5.58 | 0.02 |
| 1 | Months | 83 | 8.41 | 5.95 | 9 | 8.90 | 5.93 | 29 | 0.96 | 1.31 | 0.65 |
| 1 | NoBought | 83 | 1.00 | 1.06 | 1 | 0.85 | 1.48 | 4 | 0.97 | 0.18 | 0.12 |

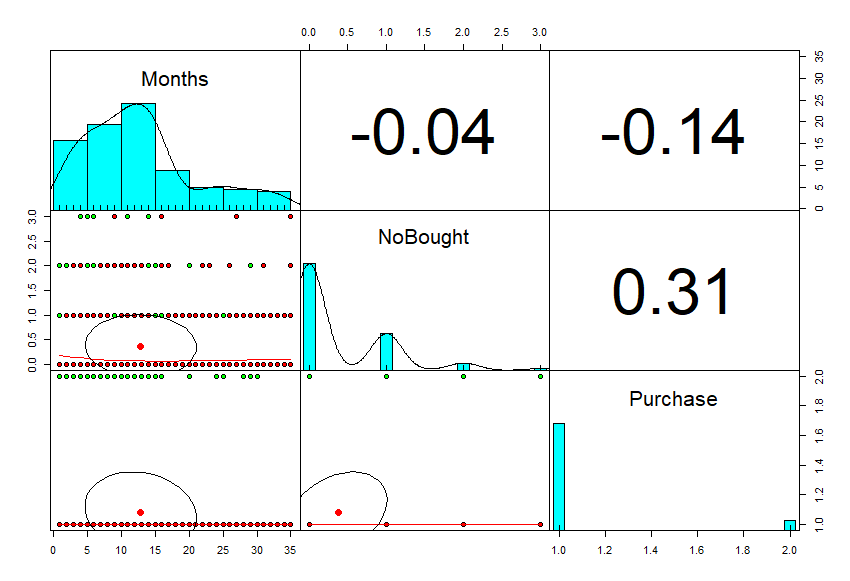
Within the groups “0” and “1” have differences in means for the two independent variables namely Months and NoBought. Means are sharply differentiated. Also, across the groups, for both the variables, means are differentiable.

RCode:

describeBy(data\_df[,c(2,3)], data\_df$Purchase)

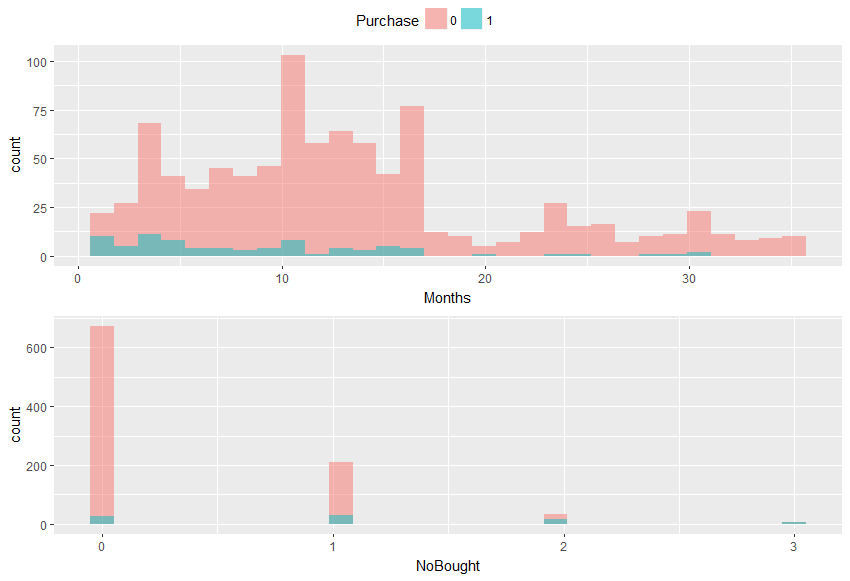
describeBy(df\_test[,c(2,3)], df\_test$Purchase)

1. **Analyzing Correlation**

Next, we analyze correlation between X & Y variables graphically and try to understand the relationship.

We observe that the Number of months since the last visit is negatively correlated with the Purchase, stating that the more the no. of months since last visit lesser is the chances of purchasing the last book. Similarly, the Number of books purchased is correlated positively with the Purchase, stating the more no. of art books purchased in the past, more are the chances of the Customer ID to purchase the said book.

Also, at an individual level, the No. of art books purchased seems to be the variable that most differentiates between Purchase (less overlap between populations)



* As Months since last purchase increases, the incidence of Purchase cases decreases.
* As no of art books purchased increases incidence of Purchase cases increases.

RCode:

pairs.panels(data\_df[2:4],gap = 0, bg = c("red","green")[data\_df$Purchase], pch=21)

p1 <- ggplot(data = data\_df, aes(x = Months, fill = Purchase)) +

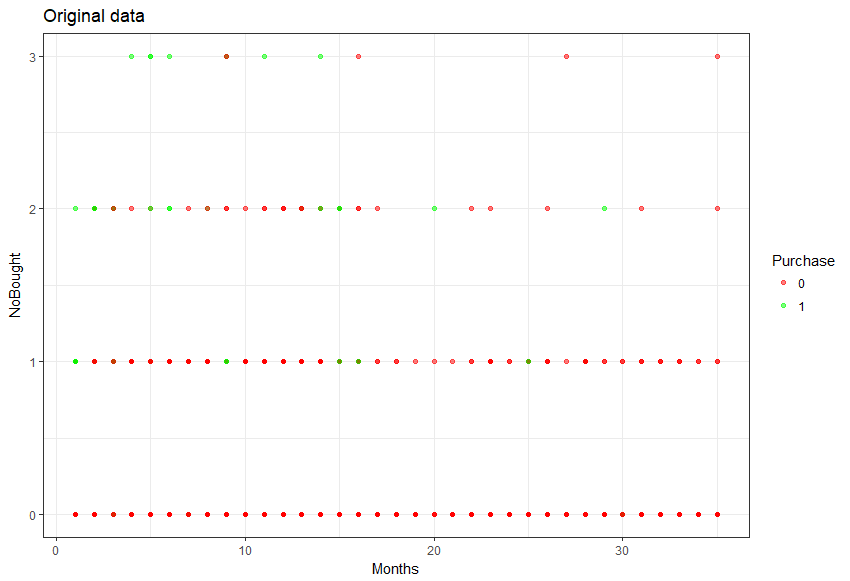
geom\_histogram(position = "identity", alpha = 0.5)

p2 <- ggplot(data = data\_df, aes(x = NoBought, fill = Purchase)) +

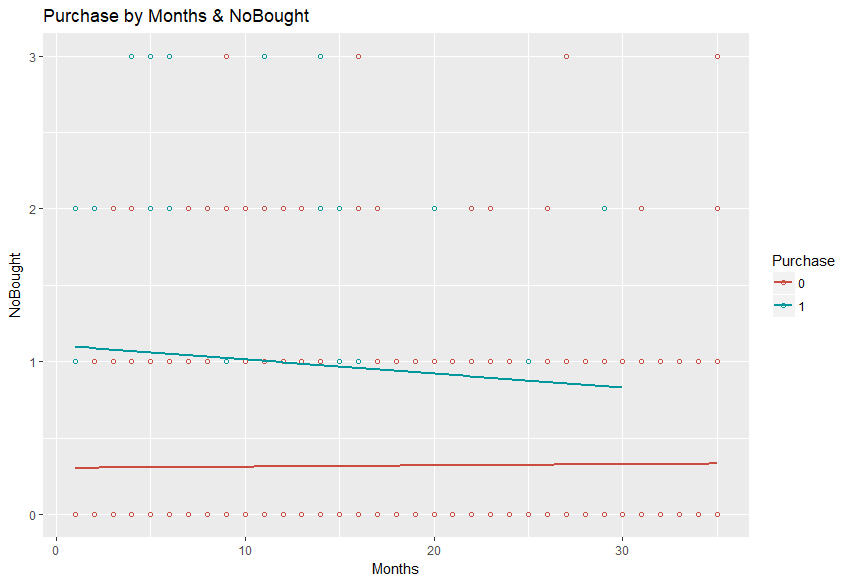
geom\_histogram(position = "identity", alpha = 0.5)

ggarrange(p1, p2, nrow = 2, common.legend = TRUE)

Similarly, we can plot the data using scatter plot to understand Months & NoBought relationship



* From the scatterplot matrix we could see the linear relationship and the distribution between observed variables.
* Months and NoBought at glance seemed to have non-normal distribution.
* Relationship between Months-Purchase suggest have negative linear relationship.
* Relationship between NoBought-Purchase suggest have positive linear relationship.



RCode:

plotData <- ggplot(data = data\_df, mapping = aes(x = Months, y = NoBought, color = Purchase)) + geom\_point(alpha = 0.5) + scale\_color\_manual(values = c("0" = "Red", "1" = "Green")) + theme\_bw() + theme(legend.key = element\_blank()) + labs(title = "Original data")

plotData

ggplot(data\_df, aes(x=Months, y=NoBought, color=Purchase)) +

geom\_point(shape=1) +

scale\_colour\_hue(l=50) + # Use a slightly darker palette than normal

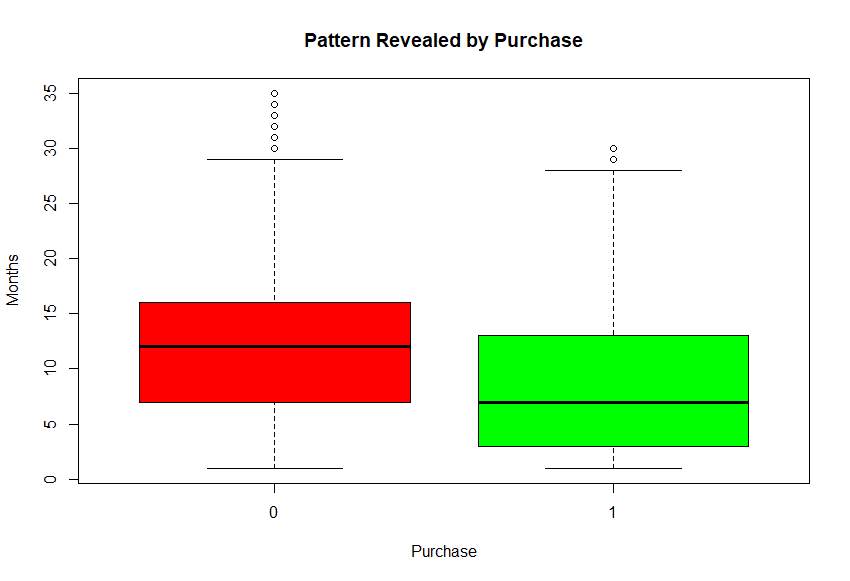
geom\_smooth(method=lm, # Add linear regression lines

se=FALSE) + # Don't add shaded confidence region

labs(title = "Purchase by Months & NoBought")

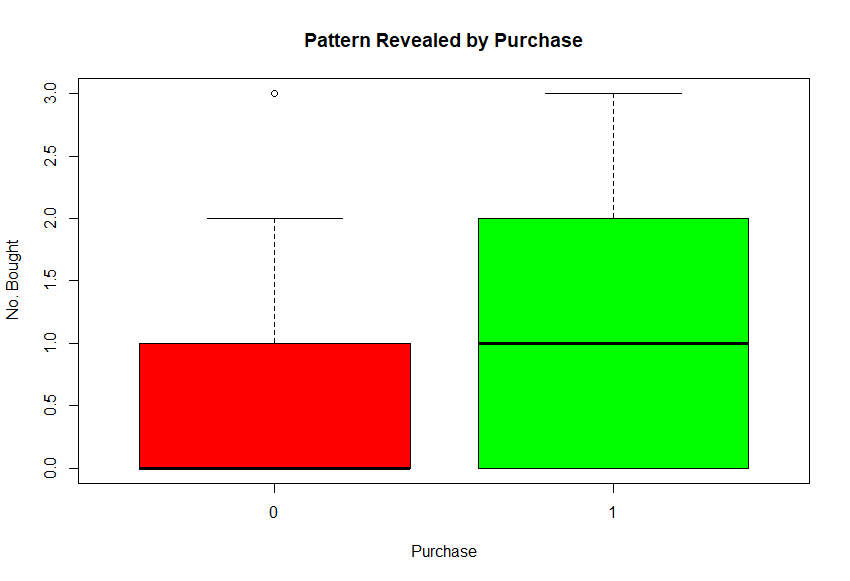
1. **Profiling**

Let us understand the pattern of Purchase from the reference of Months since last Purchase



Although Purchase shows difference with regard to Months since last Purchase but it is also overlapping. It shows slight differentiation of the groups.

Similarly, we try understanding the pattern of Purchase from the reference of No. of art books purchased by the Customer ID



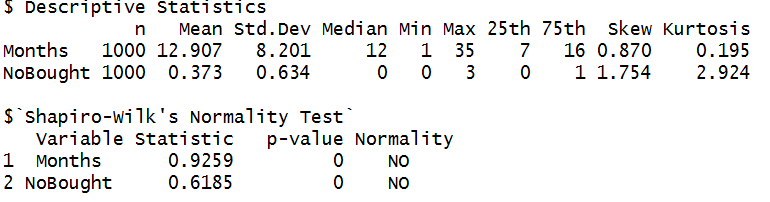
Purchase equals “0” and “1” shows differences with regard to number of art books purchased. Prima facie, NoBought seems like a differentiator of the groups.

**Assumption Test**

**Normality Test:**

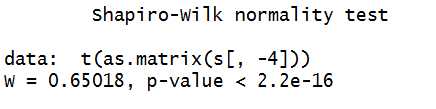
* Linear Discriminant Analysis (LDA) require normality of the independent variables and equal dispersion and covariance structures for the groups defined by the dependent variables.
* In graphical tests above, Months and NoBought was suggested to have non-normal distribution.
* Shapiro-Wilk test and Kolmogorv-Smirnov test employed to confirmed the previous presumption.

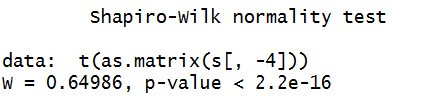
1. **Shapiro-Wilk’s Normality Test:**



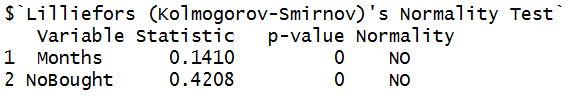
**Shapiro-Wilk’s Normality Test for each Group:**

**Group 0:**





1. **Kolmogorov-Smirnov Normality Test:**



**Interpretation:**

* The hypothesis of normality is rejected for all variables.
* Both tests show significant evidence of lack of multivariate normality.
* The LDA has some robustness against the lack of multivariate normality, but it is important to take it into account in the conclusion of the analysis.

Rcode:

data\_df\_tidy <- melt(data\_df[-1], value.name = "valor")

kable(data\_df\_tidy %>% group\_by(Purchase, variable) %>% summarise(p\_value\_Shapiro.test = shapiro.test(valor)$p.value))

hzTest(data = data\_df[,2:3], qqplot = FALSE)

roystonTest(data = data\_df[,2:3], qqplot = TRUE)

uniNorm(data\_df[2:3], type = "SW", desc = TRUE) # Shapiro-Wilk Test

#Group 0

s <- subset(data\_df, data\_df$Purchase == 0)

shapiro.test(t(as.matrix(s[,-4])))

#Group 1

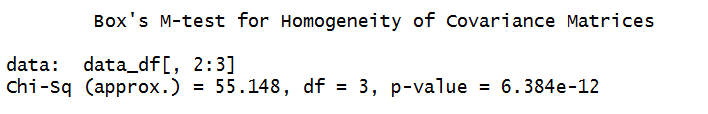
s <- subset(data\_df, data\_df$Purchase == 1)

shapiro.test(t(as.matrix(s[,-4])))

uniNorm(data\_df[2:3], type = "Lillie", desc = FALSE)

1. **Similarity of Dispersion (Homoscedasticity):**

Box’s M test to assess the similarity of the dispersion matrices of the independent variables among groups.



**Interpretation:**

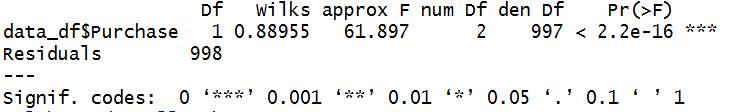
* p−value in Box’s M test is <0.01. Thus, we REJECT the H0 hypothesis of equality of covariance matrices and conclusion is that the dispersion matrices of the independent variables among group is NOT equal.
* However, since the Box's M test is very sensitive to the lack of multivariate normality, it is often significant not because the covariance matrix is not constant but because of the lack of normality, which is the case for the data. For this reason, it is assumed that the covariance matrix is constant and that LDA can achieve a good classification accuracy. In the evaluation of the model, it will be seen how good this approach is. In addition, the conclusions should explain the assumption made.

RCode:

hom<-boxM(data\_df[,2:3],group = data\_df$Purchase)

hom

1. **MANOVA**



Interpretation: Discriminant analysis is significant.

RCode:

manova2 <- manova(as.matrix(data\_df[,2:3])~data\_df$Purchase, data = data\_df)

summary(manova2, test = "Wilks")

1. **Hotelling test:**

Performs a two-sample Hotelling's T-squared test for the difference in two multivariate means.

Interpretation: p-value is 0, which means that both groups differ significantly

RCode:

hotel<-hotelling.test(.~Purchase, data = data\_df, pair = c(1,2))

hotel$pval

1. **Forward Greedy Wilks:**

Let's choose the attributes that affect the quality of the separation.

Interpretation: Variable NoBought render significant influence on result.

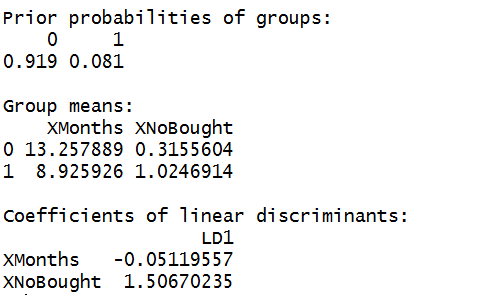
RCode:

PurchaseType <- as.numeric(data\_df$Purchase)

greedy.wilks(data\_df[,2:3], PurchaseType)

**Discriminant Analysis**

**Step 1: Differentiation-Fisher’s Discriminant Functions**



Interpretation:

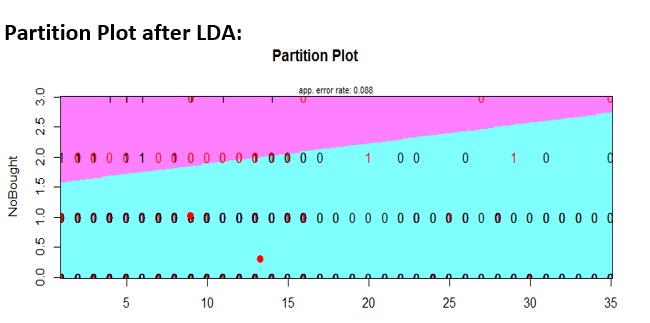
* LDA generates 1 discriminant function for group 0 and 1.

The function is: Z = -0.05119557 \* Months + 1.50670235 \* NoBought

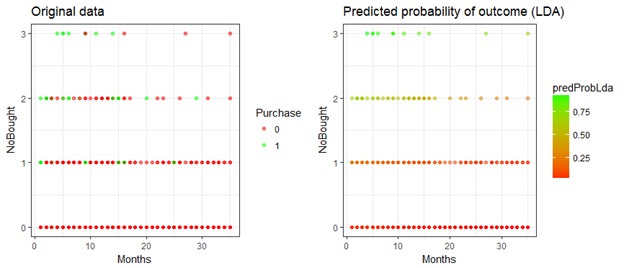
* If Z is large, then the LDA classifier will predict that the customer will Purchase, and if it is small, then the LDA classifier will predict the customer will not Purchase.
* The LDA output indicates that our prior probabilities are π1=0.919 and π2=0.081; in other words, 91.9% of the training observations are customers who did not purchase and 8.1% represent those that purchased.
* It also provides the group means; these are the average of each predictor within each class, and are used by LDA. These suggest that customers that tend to Purchase have, on average, 8.9 months since last purchase and 1 number of art book purchased than non-purchasers.

**Step 2: Graphical Display of Discriminant Functions:**

Partition Plot after LDA:



Distribution of Outcome in Original Data Vs LDA:



Interpretation:

The Partition Plot shows a slight separation between the Purchase “0” and “1”. The same can be seen from the comparison of original vs LDA outcomes too.

Rcode:

partimat(Purchase~NoBought+Months, data=df\_train,method="lda",prec = 200, #image.colors = c("Orange", "Blue"), col.mean = "Red")

plotData <- ggplot(data = data\_df, mapping = aes(x = Months, y = NoBought, color = Purchase)) +

geom\_point(alpha = 0.5) + scale\_color\_manual(values = c("0" = "Red", "1" = "Green")) +

theme\_bw() + theme(legend.key = element\_blank()) + labs(title = "Original data")

plotData

df\_train$predProbLda <- pred$posterior[,"1"]

df\_train$predClassLda <- pred$class

library(gridExtra)

plotLdaProb <- ggplot(data = df\_train, mapping = aes(x = Months, y = NoBought, color = predProbLda)) + geom\_point(alpha = 0.5) + #layer(geom = "point", alpha = 0.5) + scale\_color\_gradient(low = "Red", high = "Green") + theme\_bw() + theme(legend.key = element\_blank()) + labs(title = "Predicted probability of outcome (LDA)")

grid.arrange(plotData, plotLdaProb, ncol = 2)

**Step 3: Correlation between Discriminant Functions and Input Variables:**

DF1

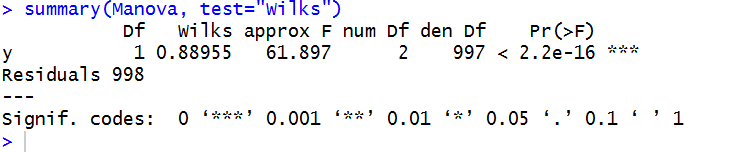
Months -0.4339

NoBought 0.9183

Interpretation:

* There is only one discriminant function DF1. The variable NoBought is much strongly related in terms of performance with Purchase than Months.
* It can also be seen that, NoBought has a positive relation while Months have a negative one. Hence the Purchase probability increases when Months since last purchase is less and Number of art books purchased is more.

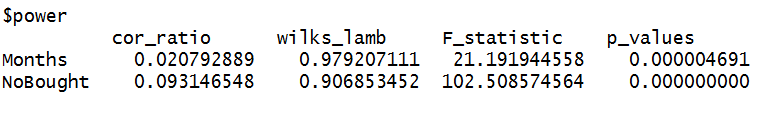
**Step 4: Differentiation using Wilks Lambda**



Interpretation:

* The p value(2.2e-16) is highly significant indicating that the Independent variables on an overall basis are significant predictors of the group separation.
* In other words, the two groups “0” and “1” are exhibiting differences in means that are statistically highly significant.

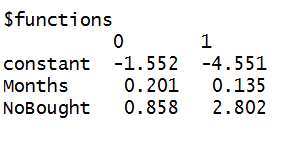
**Individual Independent Variables**



**Interpretation:**

* The p value is overwhelmingly statistically significant for the Independent variable NoBought indicating that NoBought is an excellent predictor of the group.
* The p value for the Independent Variable Months also shows statistical significance indicating that this is also a good predictor of the group.
* The relative importance could be inferred from the correlation ratio for each variable. NoBought is rank 1, followed by Months (rank 2) for their ability to differentiate the groups.

**Step 5: Mahalanobis-Classification Functions**

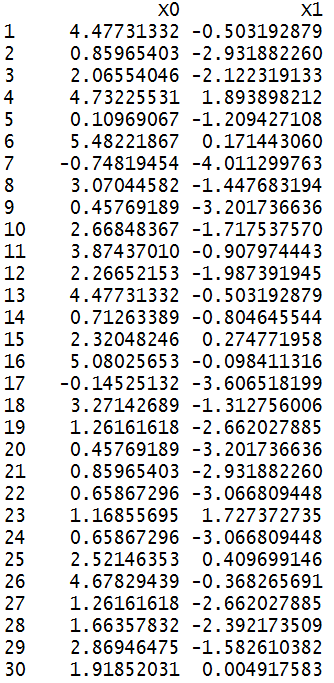


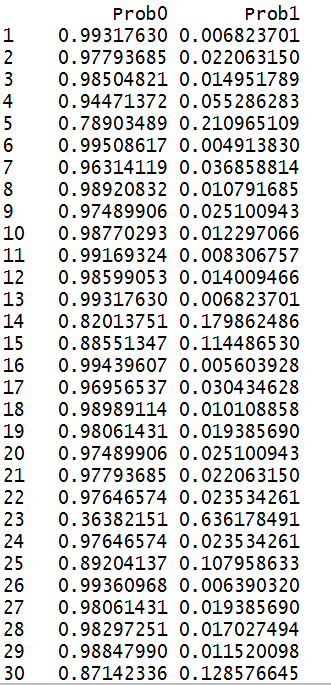
Interpretation:

* Compute the scores for both the groups(“0” and “1”) for each record.
* The largest of all the scores determine group membership.

**Step 6: Classification-Discriminant Scores**

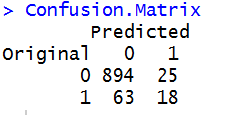
Discriminant Scores





**Step 7: Confusion Matrix using Mahalnobis Cutoff**

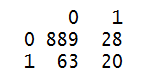
* **Training Data**



Accuracy: 91.2%

Error Rate: .088

* **Testing Data**

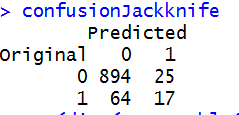


Accuracy: 90.9%

Error Rate: .091

**Step 8: Model Tunning using Jack Knife Classification**

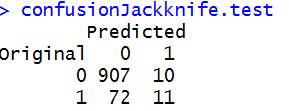
* Jackknife (Cross Validation on Train Data):



Accuracy: 91.1%

Error Rate: .089

* Jackknife (Cross Validation on Test Data):



Accuracy: 91.8%

Error Rate: .083

**Step 9: Evaluation of the Model**

| **Measure** | **Training dataset Value** | **Testing dataset Value** | **Jackknife(Training Data)** | **Jackknife(Testing Data)** |
| --- | --- | --- | --- | --- |
| Overall Accuracy | 0.912 | 0.909 | .911 | .918 |

**Step 10: Evaluation of Result using different Cut-off Points**

* Fisher’s Cutoff Point

Using desDA function in R the cutoff value is coming out to be 0.09878

Using this value we get following results in the confusion matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Actual | | Grand Total |
|  |  | 0 | 1 |  |
| Predicted | 0 | 693 | 30 | 723 |
| 1 | 226 | 51 | 277 |
|  |  | 919 | 82 | 1000 |

Accuracy: 74.4%

Error: 0.256

Comparing the accuracy of Mahalnobis Cut-off Point and Fisher’s Cut-off Point. We found out that the accuracy of the Mahalnobis’ is better.

We can however assess the operational significance of the model by using it to determine a mailing strategy based on the cost of mailing an offer to purchase the book (The Art History of Florence) and associated profit in case the book is purchased.

Let us assume that:

* Cost of mailing an offer as $x
* Profit in case the customer responds and purchases the book as $y

The profit maximization strategy is to mail to all customers with expected profitability greater than zero. Thus, we will mail to households for whom the following inequality holds

# **File Attachments:**

1. Predictive Modelling\_PaulBooks1-Using Solver\_Group 7
2. Predictive Modelling\_PaulBooks2-Using Solver\_Group 7
3. Predictive Modelling\_RCode\_Group 7\_January 28, 2018