Activity – Logistic Regression

Learning outcomes: Logistic Regression

After completing this exercise, you should be able to understand and perform below tasks.

1. Building classification model using logistic regression technique.
2. Validating the model results.
3. Handling multicollinearity and dimensionality reduction.
4. Evaluation of error metrics.
5. Applying the models on un-seen data
    a. Splitting data into train and test data sets
    b. Comparing the error metrics
6. Interpretation of the results

Assignment – Logistic Regression

Problem Statement

The “Bank.txt” file consists of the data related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be (or not) subscribed. The data and attribute description are in the folder.

Perform following Steps on the data:

1. Import the data into R
2. Understand the data and preform required preprocessing steps. Explain the reason for each step.
    - Structure and summary of the data
    - Dealing with missing values
    - Changing and recoding the factor levels for the following attributes
3. Split the data into train and test datasets
4. Implement the logistic regression model using all attributes and predict the results

```
#################Logistic regression model################

log_reg <- glm(y ~ ., data = train_data, family = binomial)

# Predicting on the train data

prob_train  <- predict(log_reg, type="response")
```

5. Identify appropriate error metric for this problem, and compute the values for these metrics on both train and test data.
   - Precision
   - Recall
   - Accuracy
6. Identify the important attributes using VIF and stepAIC

```
## Variable selection

library(car)

vif(log_reg)

library(MASS)

stepAIC(log_reg)
```

7. Use ROCR curve to obtain, reasonable cutoff for probabilities and using that probability as a threshold to obtain best set of predictions

```
##ROCR curves..Loading the required libraries

library(ROCR)

library(ggplot2)

# Predicting on the train data

prob_train  <- predict(log_reg, type="response")

#  The prediction object takes the probability scores and the original levels for theses data as input

prob <- prediction(prob_train , train$y)

# Extract performance measures (True Positive Rate and False Positive Rate) using the “performance()” function from the ROCR package

Perf  <- performance(prob, "tpr", "fpr")

# Plotting the true positive rate and false negative rate based on the threshold value

plot(Perf , col=rainbow(10), colorize=T, print.cutoffs.at=seq(0,1,0.05) )

str(Perf )
```

```
# For different threshold values identifying the tpr and fpr

cutoffs <- data.frame(cut= Perf@alpha.values[[1]], fpr= Perf@x.values[[1]],

                      tpr= perf@y.values[[1]])

# Sorting the data frame in the decreasing order based on tpr

cutoffs <- cutoffs[order(cutoffs$tpr, decreasing=TRUE),]

head(subset(cutoffs, fpr < 0.2))

# Plotting the true positive rate and false negative rate based based on the cutoff
# increasing from 0.1-1

plot(Perf, colorize = TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))
```