Categorical Data Analysis Using R in Machine Learning

K. UlagaPriya*

Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies, Chennai, India.

K. Kalaivani

Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies, Chennai, India.

Abstract

Data analysis is an important aspect in Machine learning. Analysis can be done using several analytical tools like Excel, Python, R, Octave, SPSS etc. It is important to know how to perform data analysis before creating any model in Machine Learning. Data analysis helps in understanding the data better before applying the Machine Learning algorithm. Analysis can be performed numerically and visually which would help the analyst to understand the basics and relationship between data. Normally in data science projects many data analysts directly do the modeling without understanding the data, which leads to rework or additional effort. In this paper we will see how to understand the data better and how to analyse the data using R.

1. Introduction

R is a programming language which is used for statistical computing and data analysis in Machine Learning.

R is a open source language which has around 7800 packages for various computational tasks. It is an easy and flexible language where beginners can also easily adapt to the language. R is not only entrusted by academicians, but it is also one of most popular language used by Data analyst, Researcher, Statistician etc.

The data which is used in this package is from ISLR Package with default item which is named as Credit Card Default Data.

Install the Software

The integrated Development Environment available for developing R application is RStudio(R Version 3.5.1). This tool is available in the open source link which can be downloaded and installed.

Install the Packages

There are predefined datasets available in R. This paper uses the ISLR Package

```
library(ISLR)
```

Load the Data

Load the predefined data Smarket.

```
data=Smarket
```

Summarize the Dataset

Print the summary details of the loaded data.

Received: 25 Oct 2018/Accepted: 25 Nov 2018

```
summary(data)
```

```
'data.frame':  1250 obs. of  9 variables:
 $ Year     : num  2001 2001 2001 2001 2001 ...
 $ Lag1     : num  0.381 0.959 1.032 -0.623 0.614 ...
 $ Lag2     : num  -0.192 0.381 0.959 1.032 -0.623 ...
 $ Lag3     : num  -2.624 -0.192 0.381 0.959 1.032 ...
 $ Lag4     : num  -1.055 -2.624 -0.192 0.381 0.959 ...
 $ Lag5     : num  5.01 -1.055 -2.624 -0.192 0.381 ...
 $ Volume   : num  1.19 1.3 1.41 1.28 1.21 ...
 $ Today    : num  0.959 1.032 -0.623 0.614 0.213 ...
 $ Direction: Factor w/ 2 levels "Down","Up": 2 2 1 2 2 2 1 2 2 2 ...
```

Visualising the Dataset

Density Plot

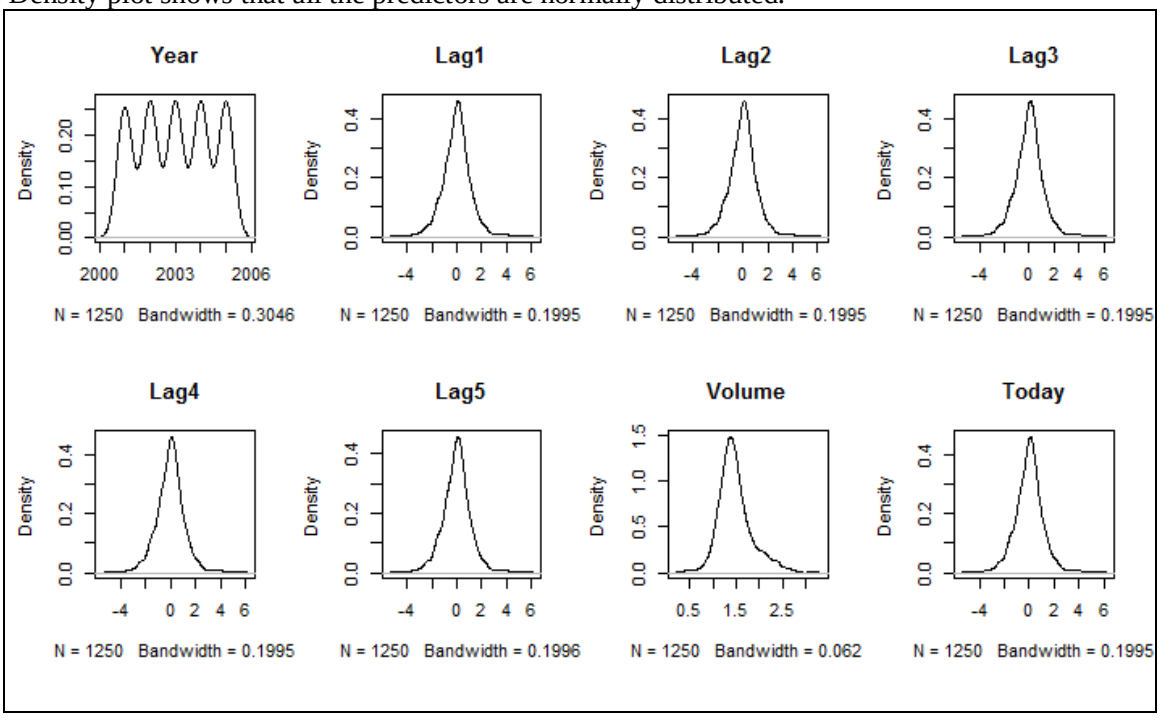
```
par(mfrow=c(2,4))

for(i in 1:8) {

 plot(density(data[,i]),main=names(data[i]))

}
```

The Density plot shows that all the predictors are normally distributed.



Received: 25 Oct 2018/Accepted: 25 Nov 2018

Box Plot

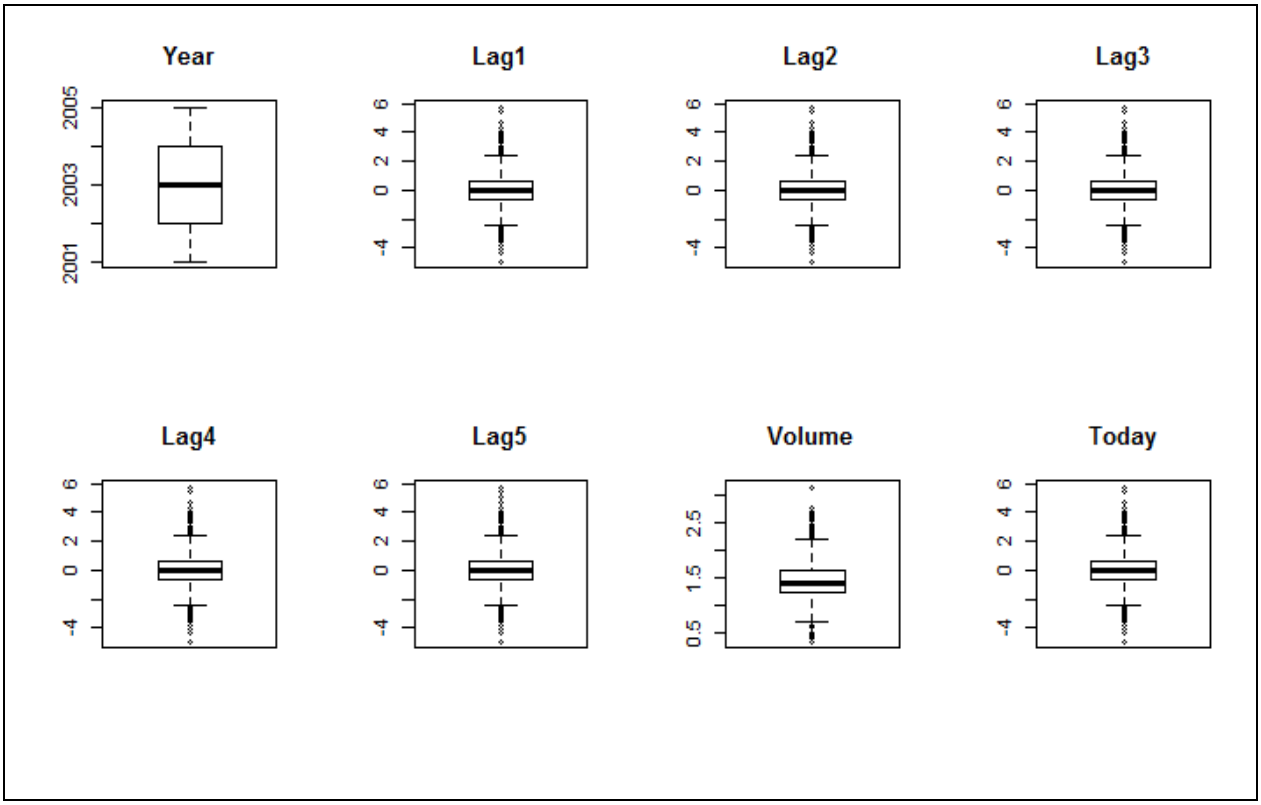
```
par(mfrow=c(2,4))

for(i in 1:8) {

 boxplot(data[,i], main=names(data)[i])

}
```

The box plot shows that all the data is around the median.



Build the Training and Test Set

The dataset is splitted into training set and test set with 70% of data in training set and 30% of data as test set.

```
##############Splitting the data set to test and Train####

trainindex=createDataPartition(data$Direction,p=0.70,list=FALSE,times=1)

train=data[trainindex,]

test=data[-trainindex,]
```

Received: 25 Oct 2018/Accepted: 25 Nov 2018

Build the Data Model

Different algorithms which works on Qualitative data are listed here.

- Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discrimnant Analysis
- K-Nearest Neighbour

```
logisticmodel=train(Direction~. , data=data,method="glm",metric="metric")

ldamodel=train(Direction~. , data=data,method="lda",metric="metric")

qdamodel=train(Direction~. , data=data,method="qda",metric="metric")

knearestmodel=train(Direction~. , data=data,method="knn",metric="metric")
```

Select the Best Model

```
# Summarize accuracy of models

result=resamples(list(logreg=logisticmodel,lda=ldamodel,qda=qdamodel,knn=knearestmodel))

summary(result)
```

```
Call:
summary.resamples(object = result)

Models: logreg, lda, qda, knn
Number of resamples: 25

Accuracy
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
logreg 0.9889135 0.9933333 0.9954649 0.9945792 0.9977876 0.9978814    0
lda    0.9279476 0.9432314 0.9545455 0.9520643 0.9585062 0.9763948    0
qda    0.8800000 0.9148472 0.9234043 0.9245369 0.9311828 0.9568966    0
knn    0.8303571 0.8395604 0.8556263 0.8559690 0.8649886 0.8933054    0

Kappa
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
logreg 0.9778066 0.9865856 0.9907970 0.9891346 0.9955657 0.9957595    0
lda    0.8555757 0.8858162 0.9089476 0.9037361 0.9169824 0.9525238    0
qda    0.7600691 0.8299084 0.8463494 0.8484090 0.8618384 0.9136889    0
knn    0.6606602 0.6774460 0.7100174 0.7110859 0.7297531 0.7860340    0
```

Received: 25 Oct 2018/Accepted: 25 Nov 2018

The Best Model is

The logistic regression model provides better result when compared to other models.

```
#summarise the best model

print(logisticmodel)
```

```
> print(logisticmodel)
Generalized Linear Model

1250 samples
   8 predictor
   2 classes: 'Down', 'Up'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1250, 1250, 1250, 1250, 1250, 1250, ...
Resampling results:

  Accuracy   Kappa
  0.9935797  0.9871231


>
```

Make Predictions

As Logistic regression produce good result, the same model is applied on test data for predictions. The accuracy is 100%.

```
#estimate of Logistic Regression on Test Set

predictions=predict(logisticmodel,test)

confusionMatrix(predictions,test$Direction)
```

```
Confusion Matrix and Statistics

          Reference
Prediction Down  Up
      Down  180   0
      Up      0 194

               Accuracy : 1
                 95% CI : (0.9902, 1)
    No Information Rate : 0.5187
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.4813
         Detection Rate : 0.4813
   Detection Prevalence : 0.4813
      Balanced Accuracy : 1.0000

       'Positive' Class : Down
```

Received: 25 Oct 2018/Accepted: 25 Nov 2018

2. Conclusion

This paper clearly explains the step by step process for analyzing categorical variables in R. It explains from Loading the data and building different models with training set. The test set is predicted with the best model and the metrics were evaluated.

Received: 25 Oct 2018/Accepted: 25 Nov 2018