Workshop 4.11 Solution

Veerasak Kritsanapraphan

2/1/2020

library(stats)  
library(caret)  
library(e1071)  
library(fastDummies)  
library(dplyr)  
telco <- read.csv("WA\_Fn-UseC\_-Telco-Customer-Churn.csv", header=TRUE)

### Remove NA

str(telco)

## 'data.frame': 7043 obs. of 21 variables:  
## $ customerID : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...  
## $ gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...  
## $ SeniorCitizen : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Partner : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...  
## $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...  
## $ tenure : int 1 34 2 45 2 8 22 10 28 62 ...  
## $ PhoneService : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...  
## $ MultipleLines : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...  
## $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...  
## $ OnlineSecurity : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...  
## $ OnlineBackup : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...  
## $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...  
## $ TechSupport : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...  
## $ StreamingTV : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...  
## $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...  
## $ Contract : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...  
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...  
## $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...  
## $ MonthlyCharges : num 29.9 57 53.9 42.3 70.7 ...  
## $ TotalCharges : num 29.9 1889.5 108.2 1840.8 151.7 ...  
## $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...

telco <- na.omit(telco)

### Dummy Column

telcodum <- select(telco, Churn, gender,SeniorCitizen,Partner,tenure,PhoneService, MultipleLines,OnlineBackup,Contract,PaperlessBilling,PaymentMethod, MonthlyCharges ,TotalCharges)  
telcodum <- dummy\_cols(telcodum)

### K-Mean Clustering

output\_telco\_train <- telcodum[,'Churn']  
input\_telco\_train <- telcodum[, c(5,12:36)]  
set.seed(123)  
km <- kmeans(input\_telco\_train, 2)

### See how it cluster

table(km$cluster, telcodum$Churn)

##   
## No Yes  
## 1 1770 322  
## 2 3393 1547

## Second Round

set.seed(500)  
km <- kmeans(input\_telco\_train, 2)   
table(km$cluster, telcodum$Churn)

##   
## No Yes  
## 1 3393 1547  
## 2 1770 322

## Try Parameter

set.seed(123444)  
km <- kmeans(input\_telco\_train, 2, iter.max = 100, nstart = 25)  
table(km$cluster, telcodum$Churn)

##   
## No Yes  
## 1 1770 322  
## 2 3393 1547