Similarity Part 4: PCA and LDA

Made by: Jonathan Blade

Data set used can be found here https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?resource=download (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?resource=download)

(Note: The original data set was divided into two files. I used the file “train.csv” and took a train and test sample from said file.)

This data set is used for classification where the goal is to determine whether a customer was satisfied with their flight.

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
planeData <- read.csv("C:\\Users\\18327\\Desktop\\Academics\\Academics Fall 2022\\Machine Learning\\Portfolio Similarity\\train.csv", header=TRUE)

planeData$satisfaction <- as.factor(planeData$satisfaction)


sapply(planeData, function(x) sum(is.na(x)))
```

```
##                                  X                                id
##                                  0                                 0
##                             Gender                     Customer.Type
##                                Age                    Type.of.Travel
##                                  0                                 0
##                              Class                   Flight.Distance
##                                  0                                 0
##              Inflight.wifi.service Departure.Arrival.time.convenient
##                                  0                                 0
##             Ease.of.Online.booking                     Gate.location
##                                  0                                 0
##                     Food.and.drink                   Online.boarding
##                                  0                                 0
##                       Seat.comfort            Inflight.entertainment
##                                  0                                 0
##                   On.board.service                  Leg.room.service
##                                  0                                 0
##                   Baggage.handling                   Checkin.service
##                                  0                                 0
##                   Inflight.service                       Cleanliness
##                                  0                                 0
##         Departure.Delay.in.Minutes          Arrival.Delay.in.Minutes
##                                  0                               310
##                       satisfaction
##                                  0
```

```
planeData <- planeData[(complete.cases(planeData)),]
sum(is.na(planeData))
```

```
## [1] 0
```

```
planeData <- planeData[(complete.cases(planeData)),]
sum(is.na(planeData))
```

```
## [1] 0
```

```
set.seed(12345)

sample <- sample(c(TRUE,FALSE), nrow(planeData), replace=TRUE, prob=c(0.8,0.2))
train <- planeData[sample, ]
test <- planeData[!sample, ]

summary(train)
```

```
##        X                id             Gender          Customer.Type
##  Min.   :     0   Min.   :     2   Length:82892       Length:82892
##  1st Qu.: 25823   1st Qu.: 32494   Class :character   Class :character
##  Median : 51829   Median : 64856   Mode  :character   Mode  :character
##  Mean   : 51887   Mean   : 64862
##  3rd Qu.: 77921   3rd Qu.: 97256
##  Max.   :103903   Max.   :129880
##       Age        Type.of.Travel        Class           Flight.Distance
##  Min.   : 7.00   Length:82892       Length:82892       Min.   :  31
##  1st Qu.:27.00   Class :character   Class :character   1st Qu.: 413
##  Median :40.00   Mode  :character   Mode  :character   Median : 844
##  Mean   :39.35                                         Mean   :1190
##  3rd Qu.:51.00                                         3rd Qu.:1744
##  Max.   :85.00                                         Max.   :4983
##  Inflight.wifi.service Departure.Arrival.time.convenient Ease.of.Online.booking
##  Min.   :0.000         Min.   :0.000                     Min.   :0.000
##  1st Qu.:2.000         1st Qu.:2.000                     1st Qu.:2.000
##  Median :3.000         Median :3.000                     Median :3.000
##  Mean   :2.731         Mean   :3.059                     Mean   :2.761
##  3rd Qu.:4.000         3rd Qu.:4.000                     3rd Qu.:4.000
##  Max.   :5.000         Max.   :5.000                     Max.   :5.000
##  Gate.location   Food.and.drink  Online.boarding  Seat.comfort
##  Min.   :0.000   Min.   :0.000   Min.   :0.00    Min.   :1.00
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00    1st Qu.:2.00
##  Median :3.000   Median :3.000   Median :3.00    Median :4.00
##  Mean   :2.978   Mean   :3.202   Mean   :3.25    Mean   :3.44
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:5.00
##  Max.   :5.000   Max.   :5.000   Max.   :5.00    Max.   :5.00
##  Inflight.entertainment On.board.service Leg.room.service Baggage.handling
##  Min.   :0.000          Min.   :0.000    Min.   :0.000    Min.   :1.000
##  1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000    1st Qu.:3.000
##  Median :4.000          Median :4.000    Median :4.000    Median :4.000
##  Mean   :3.359          Mean   :3.384    Mean   :3.354    Mean   :3.633
##  3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:5.000
##  Max.   :5.000          Max.   :5.000    Max.   :5.000    Max.   :5.000
##  Checkin.service Inflight.service  Cleanliness    Departure.Delay.in.Minutes
##  Min.   :1.000   Min.   :0.000    Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000   1st Qu.:3.000    1st Qu.:2.000   1st Qu.:   0.00
##  Median :3.000   Median :4.000    Median :3.000   Median :   0.00
##  Mean   :3.303   Mean   :3.641    Mean   :3.287   Mean   :  14.73
##  3rd Qu.:4.000   3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000   Max.   :5.000    Max.   :5.000   Max.   :1592.00
##  Arrival.Delay.in.Minutes                  satisfaction
##  Min.   :   0.00          neutral or dissatisfied:46937
##  1st Qu.:   0.00          satisfied              :35955
##  Median :   0.00
##  Mean   :  15.14
##  3rd Qu.:  13.00
##  Max.   :1584.00
```

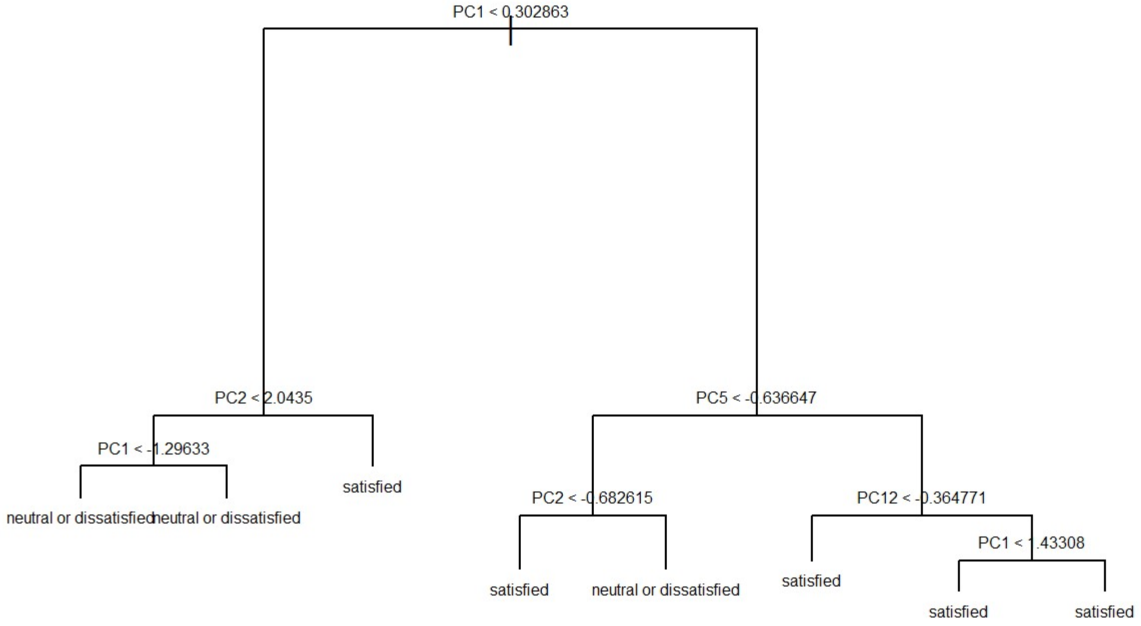
```
pca_out <- preProcess(train[,1:24], method=c("center", "scale", "pca"))

pca_out
```

```
## Created from 82892 samples and 24 variables
##
## Pre-processing:
##   - centered (20)
##   - ignored (4)
##   - principal component signal extraction (20)
##   - scaled (20)
##
## PCA needed 16 components to capture 95 percent of the variance
```

```
train_pca <- predict(pca_out, train[,1:24])
test_pca <- predict(pca_out, test[,])

train_df <- data.frame(train_pca$PC1, train_pca$PC2, train_pca$PC3, train_pca$PC4, train_pca$PC5, train_pca$PC6, train_pca$PC7, train_pca$PC8, train_pca$PC9, train_pca$PC10, train_pca$PC11, train_pca$PC12, train_pca$PC13, train_pca$PC14, train_pca$PC15, train_pca$PC16, train$satisfaction)

test_df <- data.frame(test_pca$PC1, test_pca$PC2, test_pca$PC3, test_pca$PC4, test_pca$PC5, test_pca$PC6, test_pca$PC7, test_pca$PC8, test_pca$PC9, test_pca$PC10, test_pca$PC11, test_pca$PC12, test_pca$PC13, test_pca$PC14, test_pca$PC15, test_pca$PC16, test$satisfaction)

library(class)

set.seed(12345)

pred <- knn(train=train_df[,1:16], test=test_df[,1:16], cl=train_df[,17], k=3)
meanPCA <- mean(pred == test$satisfaction)
print(paste("Accuracy with PCA: ", meanPCA))
```

```
## [1] "Accuracy with PCA:  0.901313882716646"
```

```
library(tree)
colnames(train_df) <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8", "PC9", "PC10", "PC11", "PC12", "PC13", "PC14", "PC15", "PC16", "Satisfaction")
colnames(test_df) <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8", "PC9", "PC10", "PC11", "PC12", "PC13", "PC14", "PC15", "PC16", "Satisfaction")

set.seed(12345)

tree1 <- tree(Satisfaction~., data = train_df)
plot(tree1)
text(tree1, cex=0.5, pretty=0)
```



```
pred2 <- predict(tree1, newdata=test_df, type="class")
pcaTree <- mean(pred2==test$satisfaction)
print(paste("PCA with tree: ", pcaTree))
```

```
## [1] "PCA with tree:  0.807216694039223"
```

The classification with PCA is slightly less accurate than without it. However, PCA's main function is to reduce the number of dimensions of the data. A consequence of this reduction of dimensions is a loss of interpretability. This can be seen in the above tree diagram, typically a plot that enhances interpretability, is now more difficult to understand.

```
library(MASS)

lda1 <- lda(satisfaction~., data=train)
lda1$means
```

```
##                                X       id GenderMale
## neutral or dissatisfied 51987.07 64311.77  0.4872063
## satisfied               51756.55 65579.43  0.4996245
##                         Customer.TypeLoyal Customer      Age
## neutral or dissatisfied                   0.7531798 37.52654
## satisfied                                 0.8995967 41.73870
##                         Type.of.TravelPersonal Travel  ClassEco ClassEco Plus
## neutral or dissatisfied                    0.49104118 0.6471867    0.09597972
## satisfied                                  0.07275761 0.1936031    0.04041163
##                         Flight.Distance Inflight.wifi.service
## neutral or dissatisfied         929.878              2.400026
## satisfied                      1529.356              3.164150
##                         Departure.Arrival.time.convenient
## neutral or dissatisfied                          3.126084
## satisfied                                        2.972160
##                         Ease.of.Online.booking Gate.location Food.and.drink
## neutral or dissatisfied               2.547010      2.975670       2.959542
## satisfied                             3.039216      2.980281       3.519455
##                         Online.boarding Seat.comfort Inflight.entertainment
## neutral or dissatisfied        2.655453     3.039223               2.896542
## satisfied                      4.026255     3.962258               3.963593
##                         On.board.service Leg.room.service Baggage.handling
## neutral or dissatisfied         3.017726         2.991329         3.375908
## satisfied                       3.861744         3.826811         3.969684
##                         Checkin.service Inflight.service Cleanliness
## neutral or dissatisfied        3.041375         3.387605    2.940686
## satisfied                      3.644723         3.972577    3.738173
##                         Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
## neutral or dissatisfied                   16.44617                 17.11869
## satisfied                                 12.47890                 12.55308
```

```
lda_pred <- predict(lda1, newdata = test, type="class")

meanLDA <- mean(lda_pred$class==test$satisfaction)
print(paste("Accuracy with LDA: ", meanLDA))
```

```
## [1] "Accuracy with LDA:  0.873538788522848"
```

With LDA, the accuracy is lower than PCA but the interpretability is maintained.