CV Homework 3

Code

This is an R script with the purpose of running logistic regression on dress sales data from Aliexpress to predict if they are recommended

Ramesh Kanakala

Step 1

Hide

```
#load the data and look at the first few rows
if(!require("readxl")) {
  install.packages("readxl")
}
library("readxl")
dress <- read_excel('DressSales.xlsx')
head(dress)
```

Step 2

Hide

```
#a. making Recommendation a factor
dress$Recommendation <- as.factor(dress$Recommendation)
#b. getting rid of the Dress_ID column
dress <- subset(dress, select = -Dress_ID)
#c. originally needed to change response into a factor using ifelse but not needed here; instead
making other variables num. and factors
names <- c(1:2,4:12)
dress[,names] <- lapply(dress[,names],factor)
dress$Rating <- as.numeric(dress$Rating)
library(stringr)
names(dress)<-str_replace_all(names(dress), c(" " = "." , "," = "" ))
#d. output column names
names(dress)
#e. output a summary of the data
dress <- dress[dress$Size != "s", ] #little cleaning
dress <- dress[dress$Size != "small", ]
dress <- subset(dress, select = -c(NeckLine, SleeveLength, Decoration))
dress <- dress[complete.cases(dress), ] #remove rows with NA
summary(dress)
```

f. Of the 500 observations in the data set, 58% (290) are classified as not recommended and 42% (210) are; I would consider this a balanced data set due to the closely even ratio.

Step 3

Hide

```
#plotting Recommendation against Rating and Price scores
par(mfrow=c(1,2))
plot(dress$Recommendation, dress$Rating, xlab = "Recommendation", ylab="Rating", main="Recommend
ation vs Rating", varwidth = TRUE)
plot(dress$Recommendation, dress$Price, xlab = "Recommendation", ylab="Price", main="Recommendat
ion vs Price", varwidth = TRUE)
```

Using the parameter varwidth = TRUE allows us to make the boxplot widths proportional to the square root of the sample sizes; in this case we can see that Not Recommended dresses is more common than Recommended. More importantly, we see that Recommended observations are associated with generally higher Ratings and Prices.

Step 4

Hide

```
#dividing into train/test, putting 75% in train
set.seed(1234)
i <- sample(1:nrow(dress), nrow(dress)*0.75, replace=FALSE)
train <- dress[i,]
test <- dress[-i,]
```

Step 5

Hide

```
#trying to build model predicting Recommendation from all predictors
glm1 <- glm(Recommendation~., family = "binomial", data = train)
```

Received a couple warnings and originally I thought because the training data is nearly perfectly linearly separable; the data is too easy to classify as it is separated too perfectly. However, after removing NAs, the error disappeared.

Step 6

Hide

```
#building another model with all predictors except Decoration (decided to drop it anyway in the
 end)
glm2 <- glm(Recommendation~., family = "binomial", data = train)
```

Decoration was originally causes the earlier warnings decided to remove it anyway.

Step 7

Hide

```
#predict on test data
#test <- test[!(test$Size=="s" | test$Season=="summer"),]
probs <- predict(glm2, newdata=test, type="response")
head(probs)
head(test$Recommendation)
#cor(probs, test$Recommendation) #couldn't explore correlation as Recommendation wasn't a percen
tage, rather a factor
```

Step 8

Hide

```
#binary predictions, output a table of predictions
pred <- ifelse(probs>0.5, 1, 0)
table(pred, test$Recommendation)
acc <- mean(pred==test$Recommendation)
print(paste("glm2 accuracy = ", acc))
```

Step 9

Hide

```
#ROCR graph and the AUC
library(ROCR)
pr <- prediction(probs, test$Recommendation)
# TPR = sensitivity, FPR=specificity
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
# compute AUC
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
print(paste("AUC = ", auc))
```

Hide

```
#plotting Recommendation v Season, Recommendation v Price
par(mfrow=c(1,2))
plot(dress$Recommendation, dress$Season, xlab = "Recommendation", ylab="Season", main="Recommend
ation vs Season", col=dress$Season)
plot(dress$Recommendation, dress$Price, xlab = "Recommendation", ylab="Price", main="Recommendat
ion vs Price", col=dress$Price)
```

Looking at the first plot, we can see that an most of the recommended fashion is from the Autumn style whereas Summer seems to be the greatest number in the for non-recommendations. Interestingly, in the second plot, the relative proportions of price levels are quite similar with a majority of clothes being in the average category and the least in the very-high category. Though average priced items make up a slightly larger proportion for non-recommended items, this may be due to the fact that most items are average prices anyway. Very high proportions are higher in the recommended side which also makes sense as these usually are of better quality and would be recommended.