# The Assessment of Dress Sale Classifiers

Patrick C pcosta

April 28, 2021

## Contents

**Introduction** **1**

**Exploratory Data Analysis** **2**

**Modeling** **12**

**Discussion** **15**

```
set.seed(151)
library("knitr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("MASS")
library("klaR")
library("tree")
library("rpart")
library("rpart.plot")
```

```
dress_train <- readr::read_csv("http://stat.cmu.edu/~gordonw/dress_train.csv")
dress_test <- readr::read_csv("http://stat.cmu.edu/~gordonw/dress_test.csv")
```

## Introduction

As the fashion tastes continue to involve at an increasingly rapid pace, it is important for companies to look at and quantify trends to better predict sales outcomes of specific pieces. The factors that lead to a successful clothing article, then, are varied. Within the scope of this paper, we will train and evaluate machine learning classification techniques for determining if a clothing piece sold well or not based on a number of variables acquired by Air University students Muhammad Usman & Adeel Ahmed.

# Exploratory Data Analysis

The sample data was gathered over the past year from dress sales at an individual clothing store, with a number of individual predictor variables being compared to the sale success of the dress as follows:

Style: dress style (cute, work, casual, fashion, party), Price: price range (low, average, high), Rating: average customer rating from dress factory market survey (average of stars, 0-5) , Season: which season is the dress appropriate for (summer, fall, winter, spring), NeckLine: type of neckline (O-neck, V-neck, other), Material: if it is a cotton dress or not, Decoration: if it has any decoration or not, Pattern: if the fabric has a pattern (yes) or of it's a solid color (no), Sleeve: if the dress has a sleeve, Waistline: type of waistline (other, empire, natural)

Our response label that we will work to predict with the given classifiers, then, is: Recommendation: binary outcome if the dress sells well (1) or not (0).

**Summary of Response Label**

In the training set, we have 347 observations, with 189 dresses that did not sell well (comprising 54.47%) and 158 dresses that did sell well (comprising 45.53%).

```
table(dress_train$Recommendation)
```

```
##
##   0   1
## 189 158
```

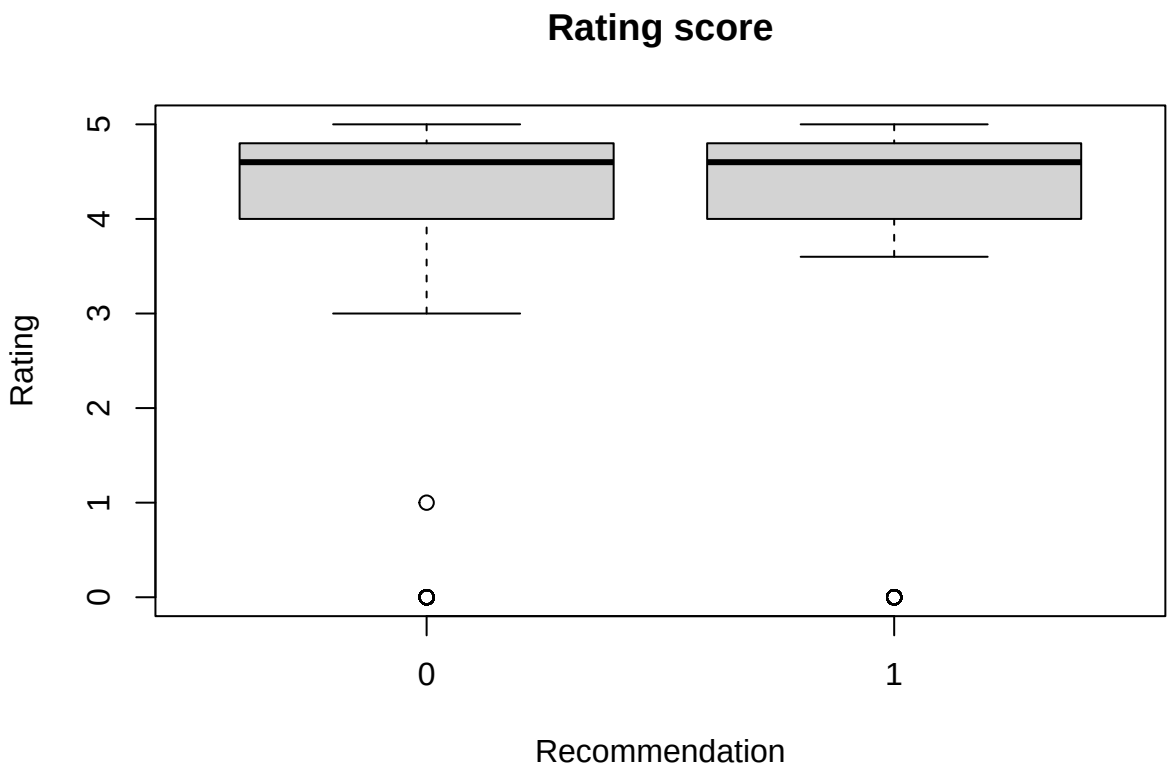
```
prop.table(table(dress_train$Recommendation))
```

```
##
##         0         1
## 0.5446686 0.4553314
```

**EDA on relationship between Recommendation and quantitative variable**

Next, we turn to look at visualizing the relationship between the response variable (recommendation) and the quantitiative predictor variable (Rating). To see whether it will be useful in determining recommendation, we will view the boxplot as follows:

```
boxplot(Rating ~ Recommendation, data = dress_train, main = "Rating score")
```

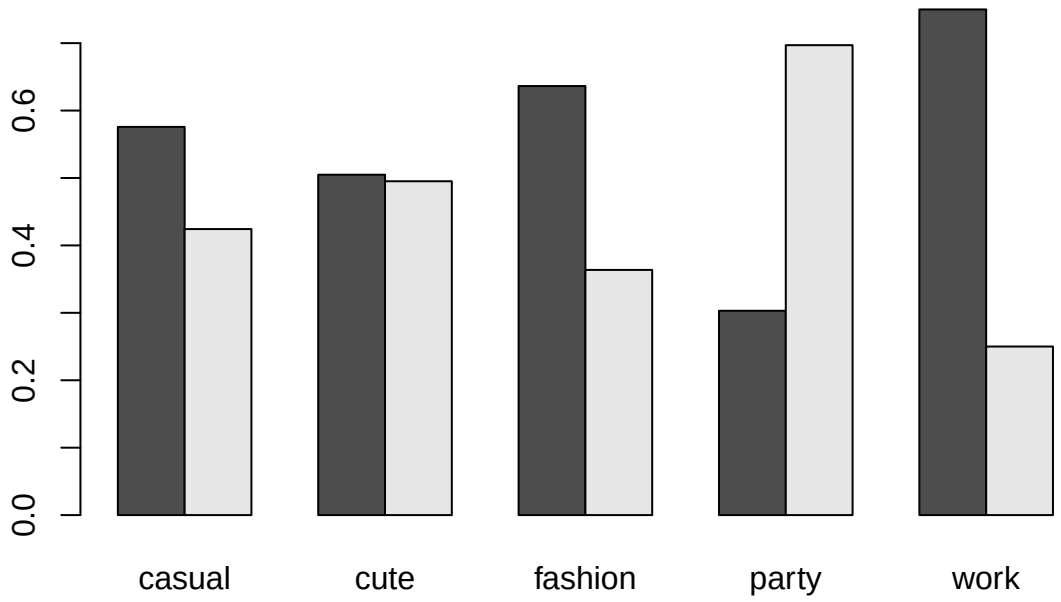



Within the above boxplot, we can see that poorly selling dresses have a wider range of rating scores, with more outliers towards 0, but overall the two recommendation levels have nearly identical IQRs and medians. As such, there does not appear to be a noteworthy difference between the two groups and no strong evidence of a relationship.

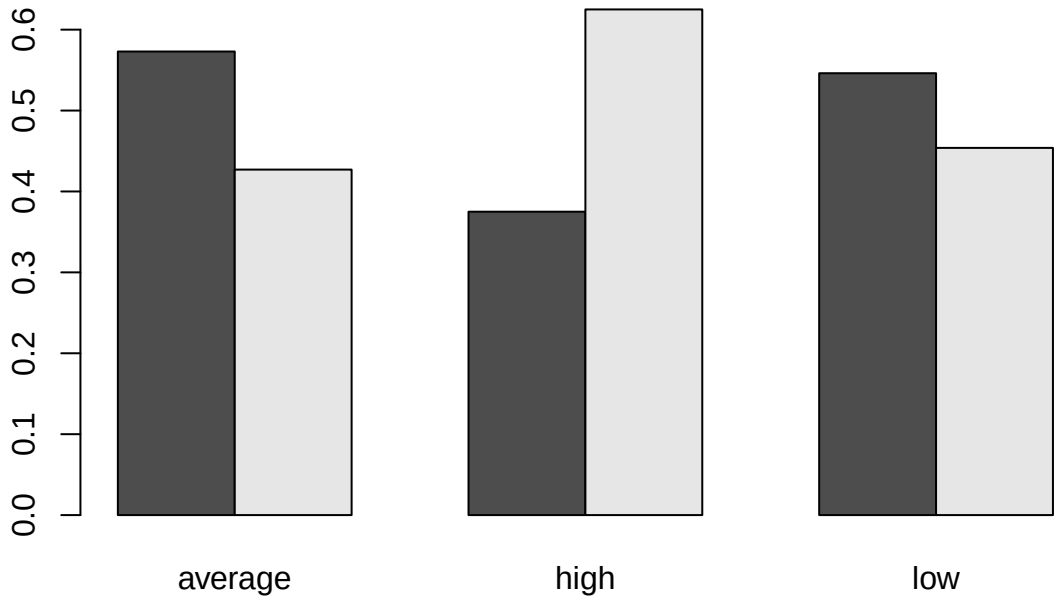
**EDA on relationship between Recommendation and quantitative variable**

We then move on to looking at each categorical predictor variable in comparison to the respond variable (Recommendation). To accomplish this, we will use proportional bar graphs as follows:

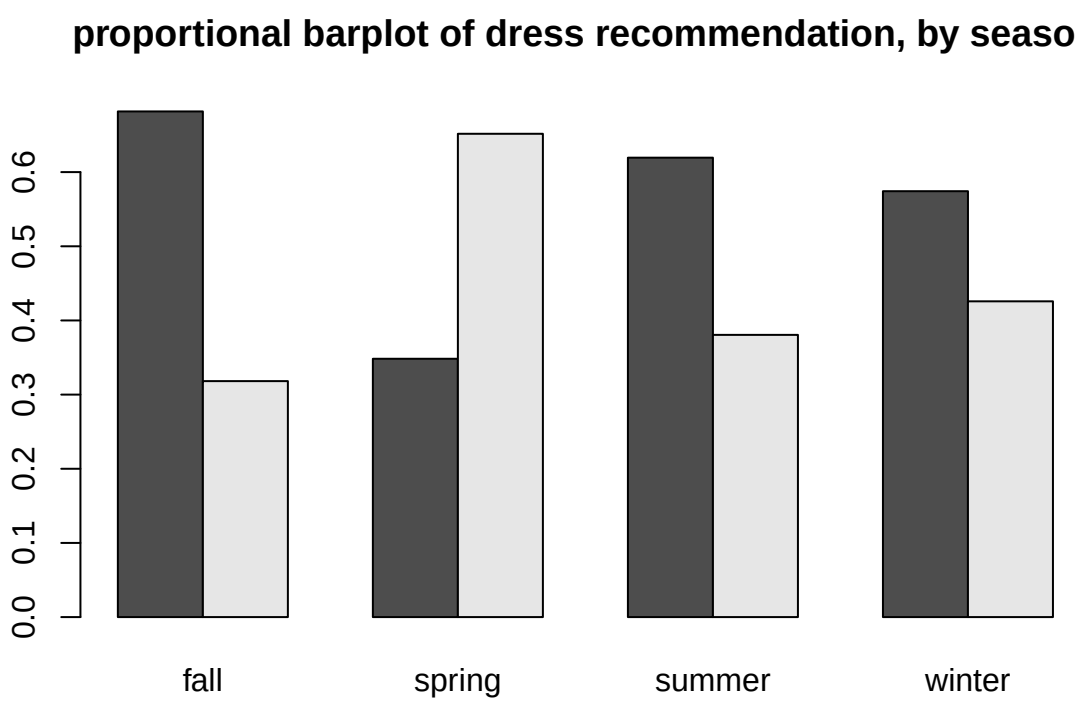
```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Style),
          margin = 2), beside = TRUE,
    main = "proportional barplot of dress recommendation, by style")
```



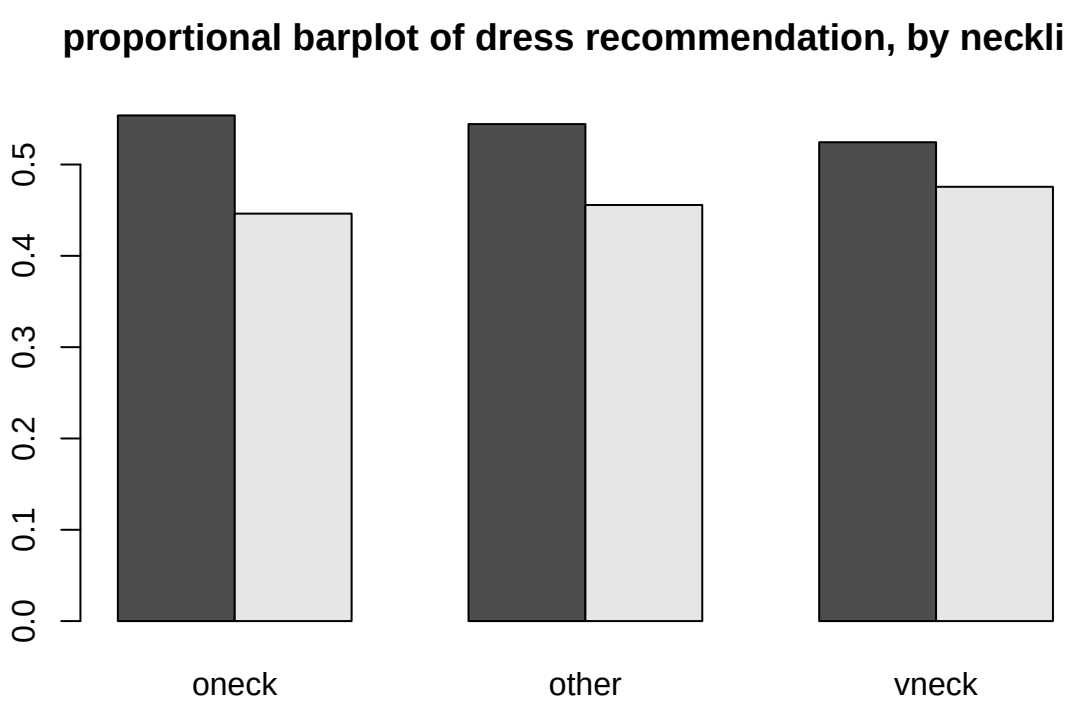

```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Price),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by price")
```



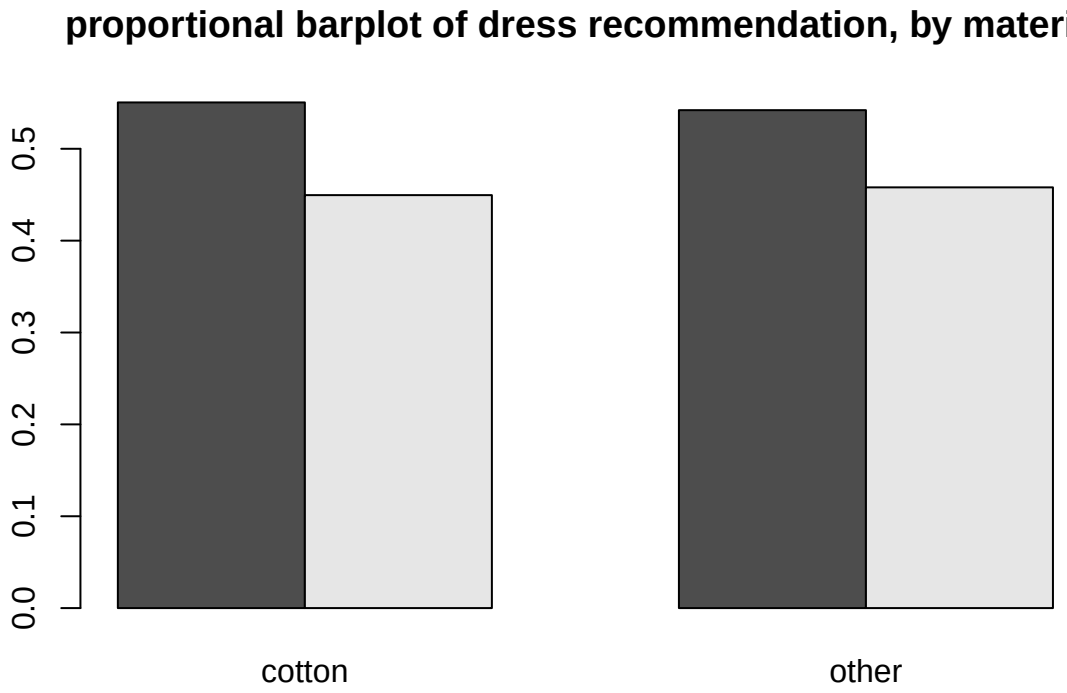

```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Season),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by season")
```



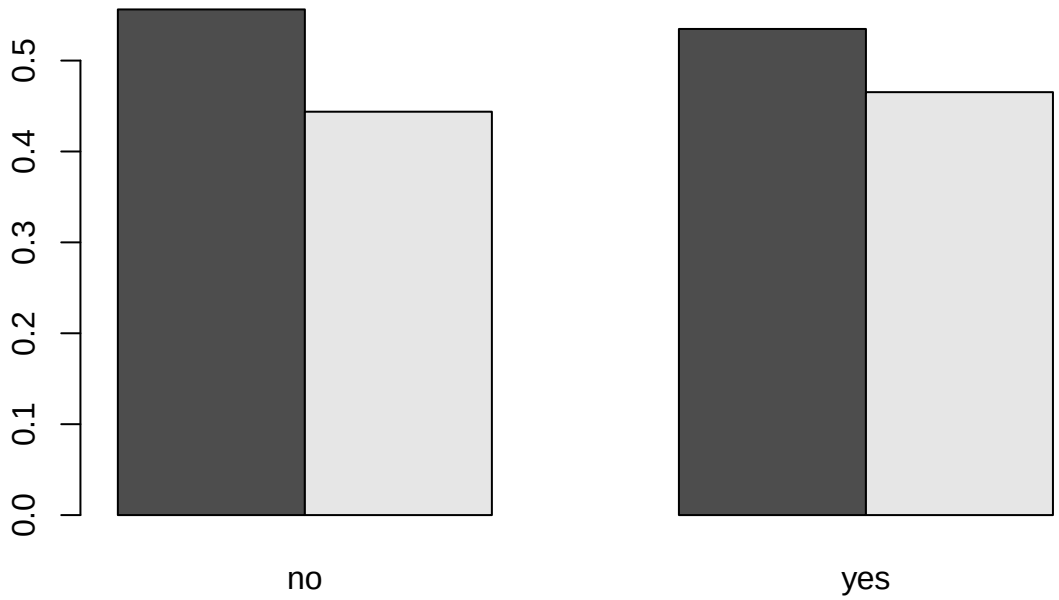

```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$NeckLine),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by neckline")
```



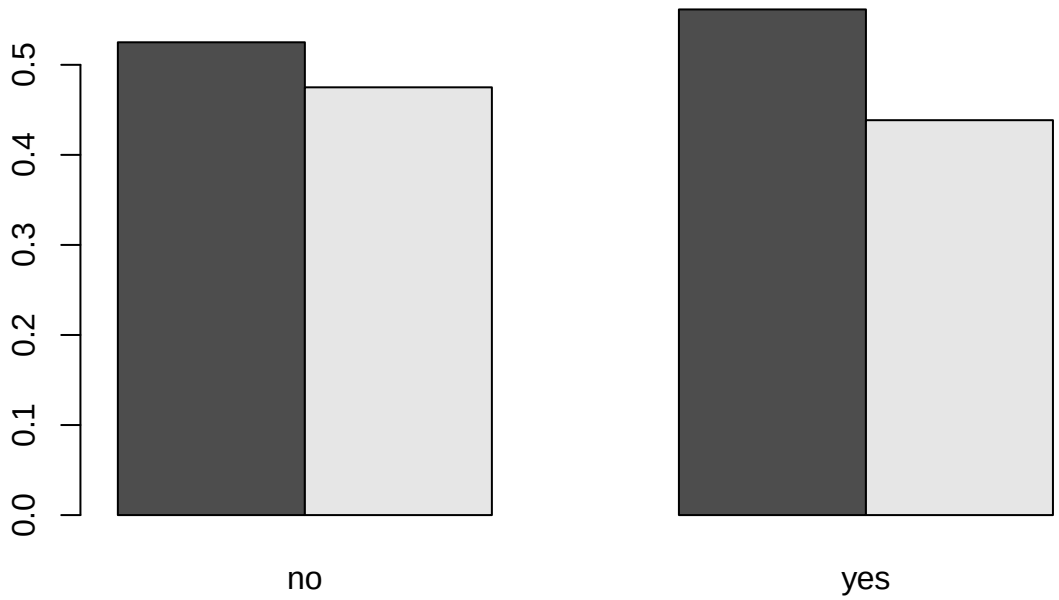

```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Material),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by material")
```



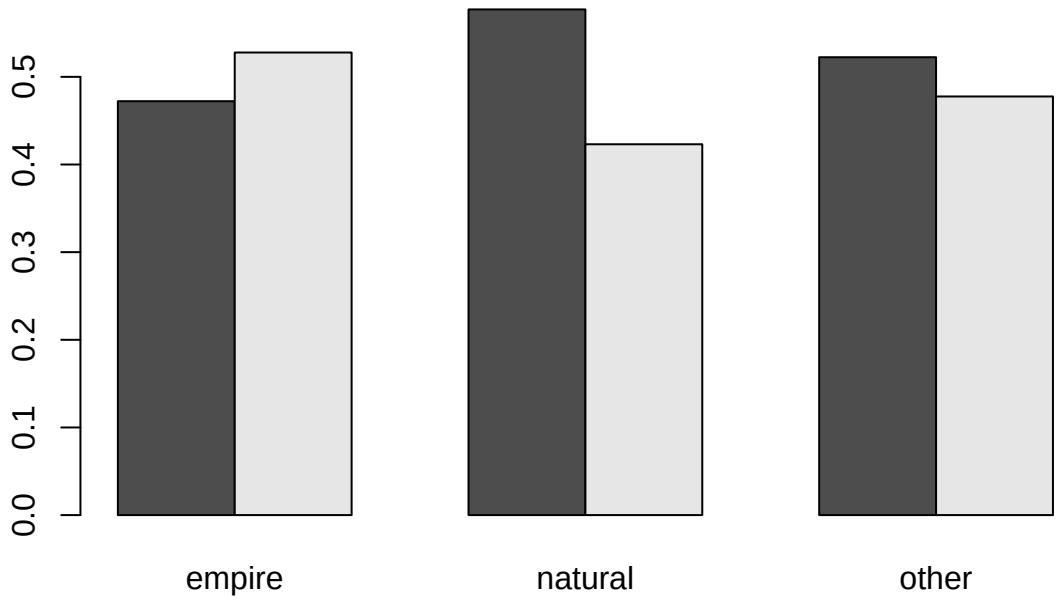

```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Decoration),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by decoration")
```



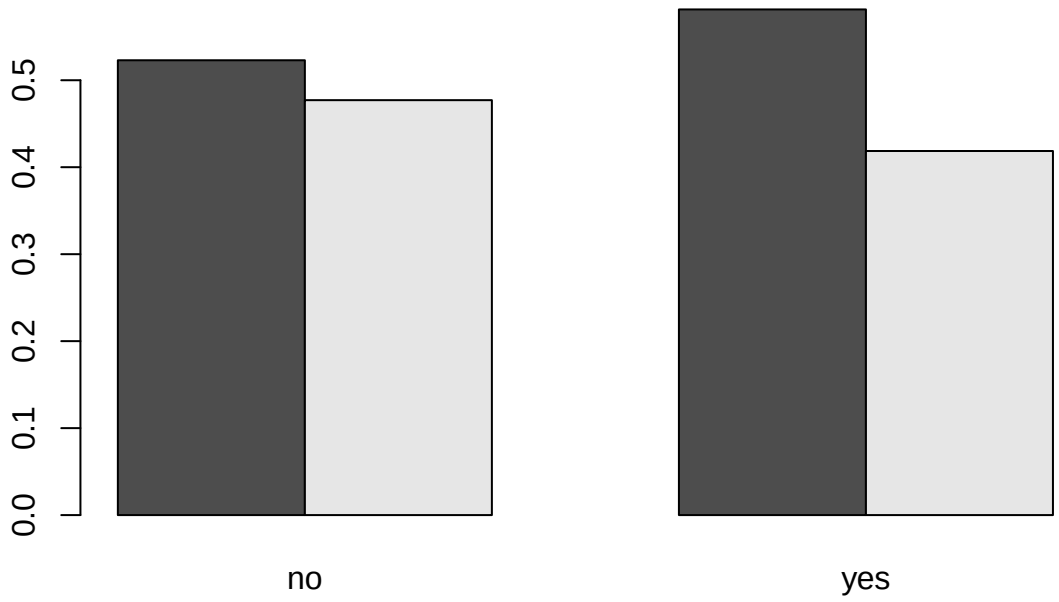

```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Sleeve),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by sleeve")
```




```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Waistline),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by waistline")
```




```
barplot(
  prop.table(
  table(dress_train$Recommendation, dress_train$Pattern),
            margin = 2)
      , beside = TRUE,
      main = "proportional barplot of dress recommendation, by pattern")
```




From the above bar graphs, it appears that party dresses appear to have the highest proportion of individual pieces that sell well, while work dresses have the lowest proportion of the same. Cute dresses appear to have a 50/50 proportion of those that sell well vs. do not, while casual and fashion have more dresses that do not sell well. Additionally, higher priced dresses appear to have a larger proportion of individual pieces that sell well. Spring also looks to be the season with the higher proportion of well-selling dresses, while Fall is the season with the higher number of poorly-selling dresses by a considerable margin. Summer and Winter also look to have more dresses that do not sell as well. The 6 other predictor variables (neckline, material, decoration, sleeve, waistline, and pattern) all appear to have similar proportions of individual pieces that sell well vs. do not sell well, with "poorly-selling" having a higher proportion for nearly all categories.

Also, as our dataset only features one quantitative variable, we cannot create a pairs plot for comparison.

# Modeling

We now look to modeling dress success through the use of LDA, QDA, classification trees, and binary logistic regression. To prevent overfitting, we randomly split our data into training and test datasets. All four models were built using the same training observations and assessed on the same set of test observations.

**Linear Discriminant Analysis (LDA)**

For both LDA and QDA models, we only have one continuous variable (rating), so the LDA classifier is built on the training data as follows:

```
dress.lda <- lda(Recommendation ~ Rating, data = dress_train)
```

We then assess the performance of the LDA classifer on our test data.

```
dress.lda.pred <- predict(dress.lda, as.data.frame(dress_test))
table(dress.lda.pred$class, dress_test$Recommendation)
```

```
##
##       0   1
##   0 100  49
##   1   0   0
```

On the test data, LDA gave an overall error rate of (0+49/149) = 0.328 which is somewhat low. The classifier performs best when looking at poorly-selling dresses (error rate of 0/100 = 0%), but performs poorly when looking at well-selling dresses (error rate of 49/49 = 100%).

**Quadratic Discriminant Analysis (QDA)**

Next, we look to view the QDA classifier and assess its performance as follows:

```
dress.qda <- qda(Recommendation ~ Rating, data = dress_train)
dress.qda.pred <- predict(dress.qda, as.data.frame(dress_test))
table(dress.qda.pred$class, dress_test$Recommendation)
```

```
##
##       0   1
##   0 100  49
##   1   0   0
```

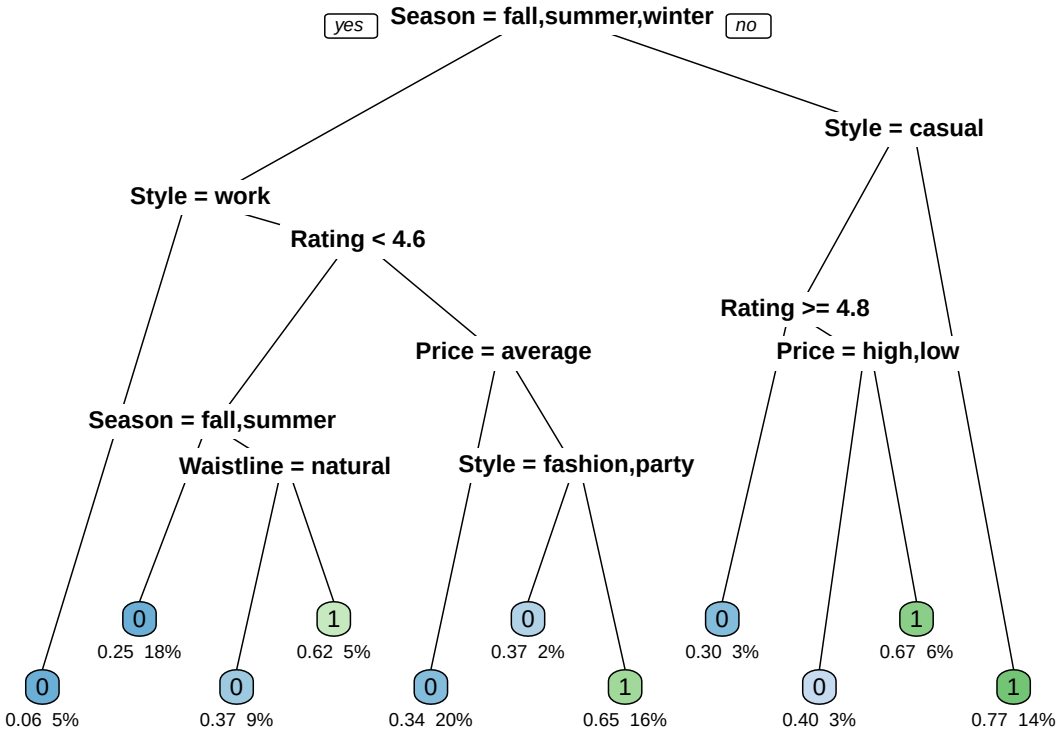
The QDA classification yielded the same results as our LDA classification. On the test data, QDA gave an overall error rate of (0+49/149) = 0.328 which is somewhat low. The classifier performs best when looking at poorly-selling dresses (error rate of 0/100 = 0%), but performs poorly when looking at well-selling dresses (error rate of 49/49 = 100%).

**Classification Trees**

While we previously could only account for Rating within the LDA and QDA classifiers, we can also use a classification tree to account for our categorical variables. We fit a classification tree on the training data and plot it, as follows:

```
dress.tree <- rpart(Recommendation ~ Style + Price + Rating + Season + NeckLine + Material + Decoration
                    data = dress_train, method="class")

rpart.plot(dress.tree,
           type = 0,
           clip.right.labs = FALSE,
           branch = 0.1,
           under = TRUE)
```




The created classification tree used Season, Style, Rating, Price, and Waistline to classify dress sales. We then investigate the tree's performance on our test data as follows:

```
dress.tree.pred <- predict(dress.tree, as.data.frame(dress_test), type="class")
table(dress.tree.pred, dress_test$Recommendation)
```

```
##
## dress.tree.pred  0  1
##               0 71 31
##               1 29 18
```

This classification tree gave an overall error rate of (29+31)/149 = 0.40, which is moderate. It performed best at predicting poorly-selling dresses (error rate of 29/100 = 0.29), but performed somewhat poorly on predicting well-selling dresses (error rate of 31/49 = 0.63).

### Binary Logistic Regression

Last, we look to creating a binary logistic regression model to predict the sale performance of a specific dress. In this example, we can also include our categorical variables and fit a binary logistic regression to the data as follows:

```
dress.logit <- glm(factor(Recommendation) ~ factor(Style) + factor(Price) + Rating + factor(Season) + fa
                     data = dress_train,
                     family = binomial(link = "logit"))
```

Next, we assess the model's performance on our test data by obtaining the test classification using a threshold probability of 0.5 and viewing a confusion matrix as follows:

```
dress.logit.prob <- predict(dress.logit, as.data.frame(dress_test), type = "response")
dress.logit.pred <- ifelse(dress.logit.prob > 0.5, "1", "0")
table(dress.logit.pred, dress_test$Recommendation)
```

```
##
## dress.logit.pred  0  1
##                0 77 27
##                1 23 22
```

Our binary logistic regression model has an overall error rate of (27+23)/149 = 0.3355, which is fairly low. It performs best at predicting poorly-selling dresses (error rate of 23/100 = 0.23), but performs moderately at predicting well-selling dresses (error rate of 27/29 = 0.551).

One final note is that both LDA and QDA performed better on poorly-selling clothes by a considerable margin.

**Final Recommendation**

Of the four classifiers tested, binary logistic regression performed best on the test data. While the classification tree performed almost as well at binary logistic regression, both LDA and QDA performed poorly at accurately predicting well-selling dresses (despite having the lowest overall error rates). Both LDA and QDA performed exactly the same, though.

Our final recommendation is binary logistic regression, as it featured the lowest error rates while including more than just one quantitative variable (as seen within both the LDA and QDA classifiers). Despite this, the classification tree is at a close second due to relatively similar error rates between the two separate classification methods of BLR and the tree. LDA and QDA are not recommended for use, as they merely feature one quantitative variable and only are strong when predicting poorly-selling dresses.

# Discussion

Overall, our binary logistic regression and classifier tree models did fairly well at predicting dress sales based on the given predictor variables. It is important to note, however, the faults with both LDA and QDA classifiers on the data. Gathering additional quantitative data would likely benefit the outcomes of both classification methods and improve well-selling dress error rates overall.

For future research, it could be beneficial to the company to create models that predict sales of pieces within each season, as sale proportions appeared to vary wildly from season-to-season. This additional research would also aid the company in determining the desired price range, style, etc. for each fiscal year.