Analysis on Sales of Dresses from Clothing Store X

Xuduo Victor Wen

Contents

Introduction 1

Exploratory Data Analysis 1

Modeling 11

Discussion 14

Introduction

A dress is a garment traditionally worn by women or girls consisting of a skirt with an attached bodice. Top selling dresses often have stand out qualities that distinguish them from the ordinary ones. In particular, we will examine data collected by Muhammad Usman & Adeel Ahmed from Air University of a specific clothing store (referred to clothing store X in the title). The clothing store dress sales have not been very successful in the last year. Although some styles sold out very fast, others were still there after sales events. The store wants to know which dresses they should order for next year so they have more successful sales.

In this paper, we will train and evaluate machine learning classification technique for predicting whether the dress sells well or not, based on ten different variables (most are categorical and one is quantitative) . We will take a closer look at the explanatory and response variables involved in the section below.

Exploratory Data Analysis

Background and Variables

In data from Air University, there were ten predictors that determined whether the dress sells well or not. The determining predictors are listed below:

- **Style**: dress style (cute, work, casual, fashion, party)
- **Price**: price range (low, average, high)
- **Rating** (this is the only quantitative predictor): 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)

Next, we have our response labels that we want to classify (with 1 or 0) with our predictor:

- **Recommendation**: binary outcome if the dress sells well (1) or not (0)

Summary of the Response Labels in the Training Dataset

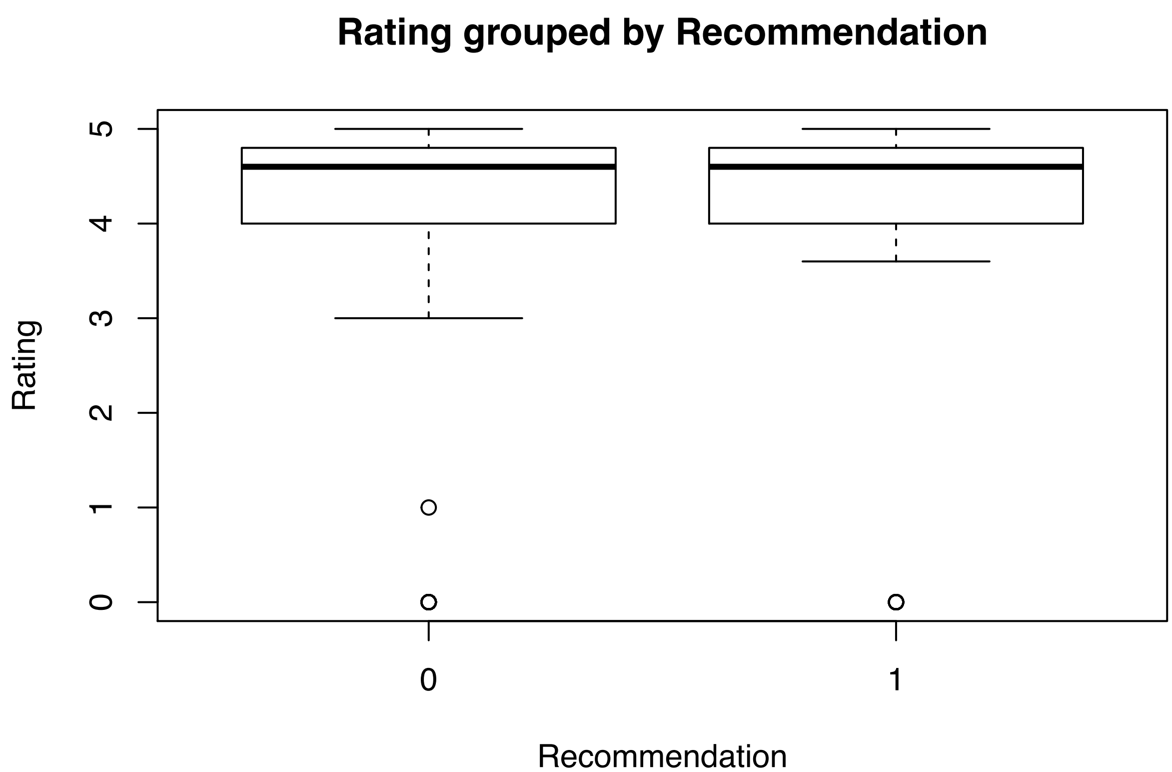
We first note that in the traning set, we have 347 observations. There are 189 dresses that do not sell well, comprising 54.47% of all dresses, and there are 158 well sold dresses, comprising of 45.53% of all adresses in the data. These statistics are also shown in the tbale below.

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

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

** EDA (exploratory data analysis) on relationships between type and the quantitative variable**

We then move toward visualizing the relationship between the response (recommendation) and the various predictors (characteristics of the dress). For visually exploring whether we expect the quantitative predictor to be useful in helping to classify the recommendation, we show boxplot, which appear as follows:

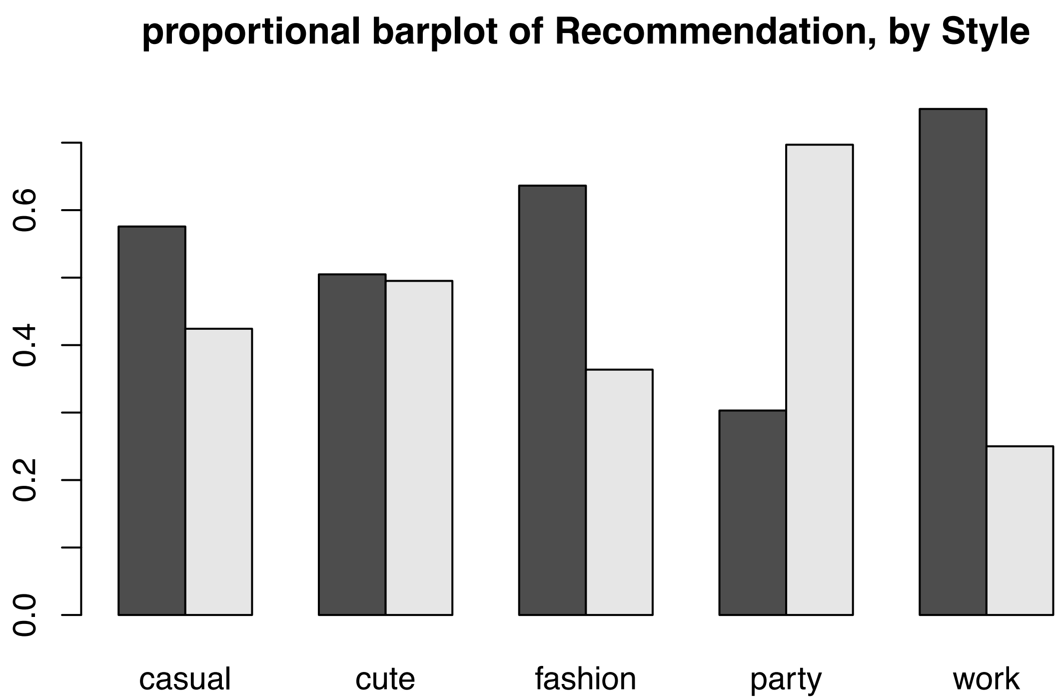


In the above boxplot, we note that if the boxplot shows difference between recommendations due to rating, then we have evidence of a relationship and a variable that might be useful in our classifier (although note that this is not the same as a statistically significant relationship). With that in mind, we note that a well sold dress tends to have noticably higher rating than a not well sold dress. The maximum, IQR and median of ratings for well sold dresses and not well sold dresses are about the same. However, the minimum rating given by customers for a well sold dress is also noticably higher than the minimum rating given by customers for a not well sold dress, about 0.5 difference between the two. There are two outliers for the not well sold group and one outlier for the well sold group.

EDA on relationships between type and the categorical variable

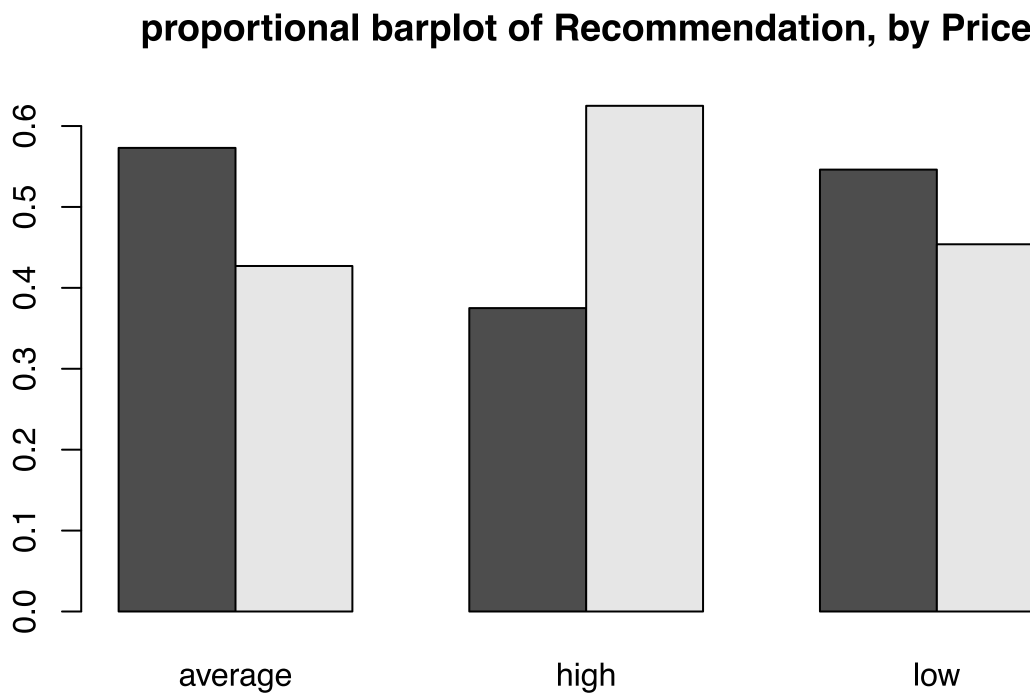
To explore the relationship between recommendation and the categorical predictors, we can look at the conditional proportions of recommendation, conditioned on the categorical predictors, shown as follows:

```
##
##        casual      cute   fashion     party      work
##   0 0.5757576 0.5048544 0.6363636 0.3030303 0.7500000
##   1 0.4242424 0.4951456 0.3636364 0.6969697 0.2500000
```



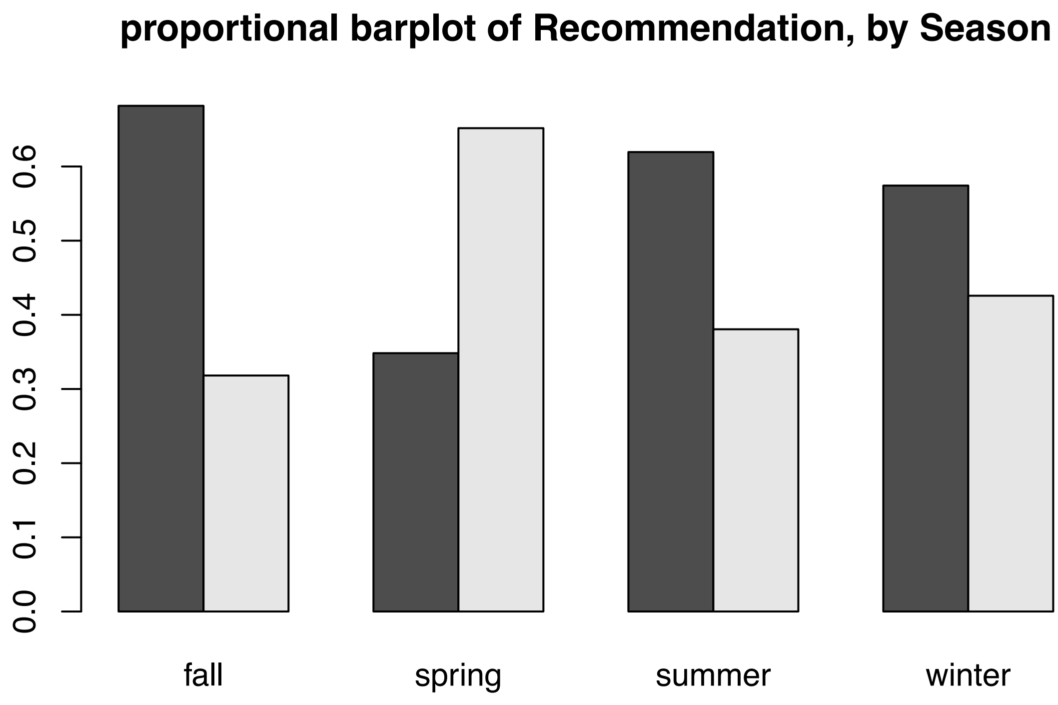
The first variable we'll take a look at is style. It seems like work dresses had the highest probability of not being well sold, while party dresses had the lowest probability of not being well sold. It seemed like more "formal" the dress is, the higher the probability that it is not well sold.

```
##
##       average      high       low
##   0 0.5729730 0.3750000 0.5461538
##   1 0.4270270 0.6250000 0.4538462
```



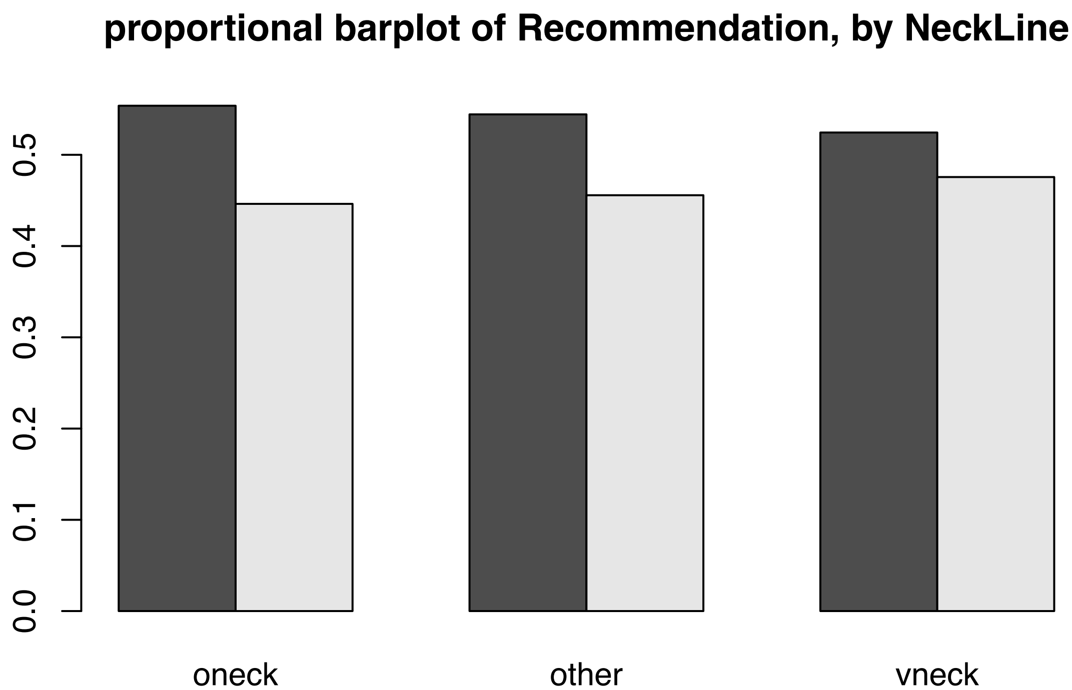
Next, we have the variable Price. High priced dresses had the highest probability of being well sold. However, the difference between probabilities of being well sold or not for low and average priced dresses is relatively small. There are no obvious/significant trend between the low and averaged priced dresses.

```
##
##         fall    spring    summer    winter
##   0 0.6818182 0.3483146 0.6194690 0.5742574
##   1 0.3181818 0.6516854 0.3805310 0.4257426
```



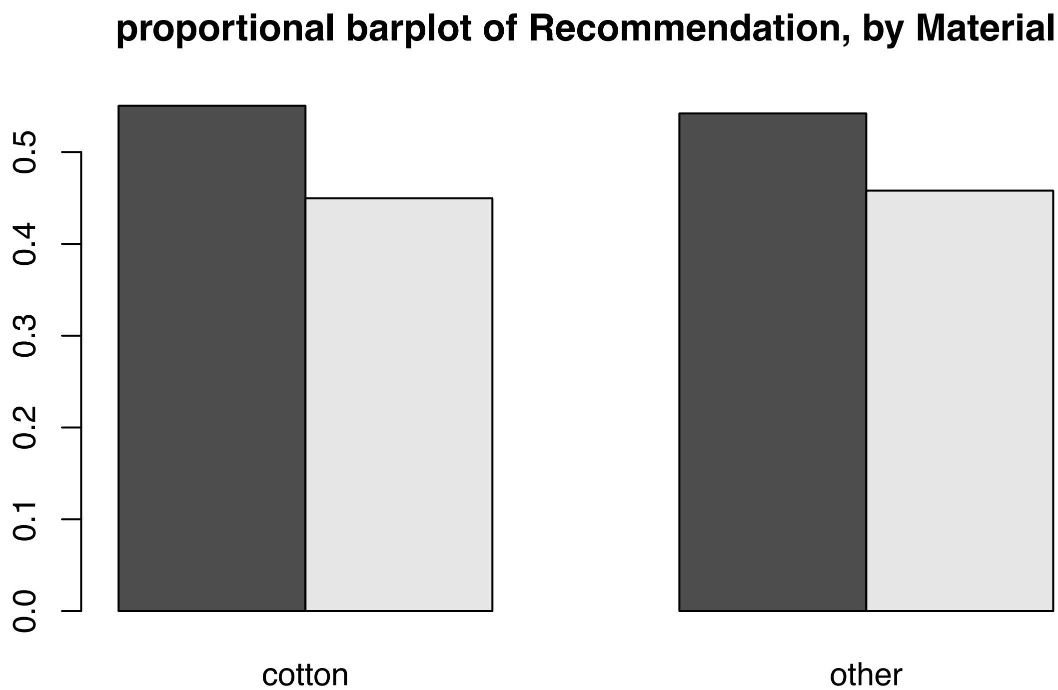
Next, we have the variable season. It looks like that Spring dresses had the highest chance of being sold well while fall dresses had the lowest relative probability. The difference between summer and winter are close.

```
##
##         oneck     other     vneck
##   0 0.5537634 0.5443038 0.5243902
##   1 0.4462366 0.4556962 0.4756098
```



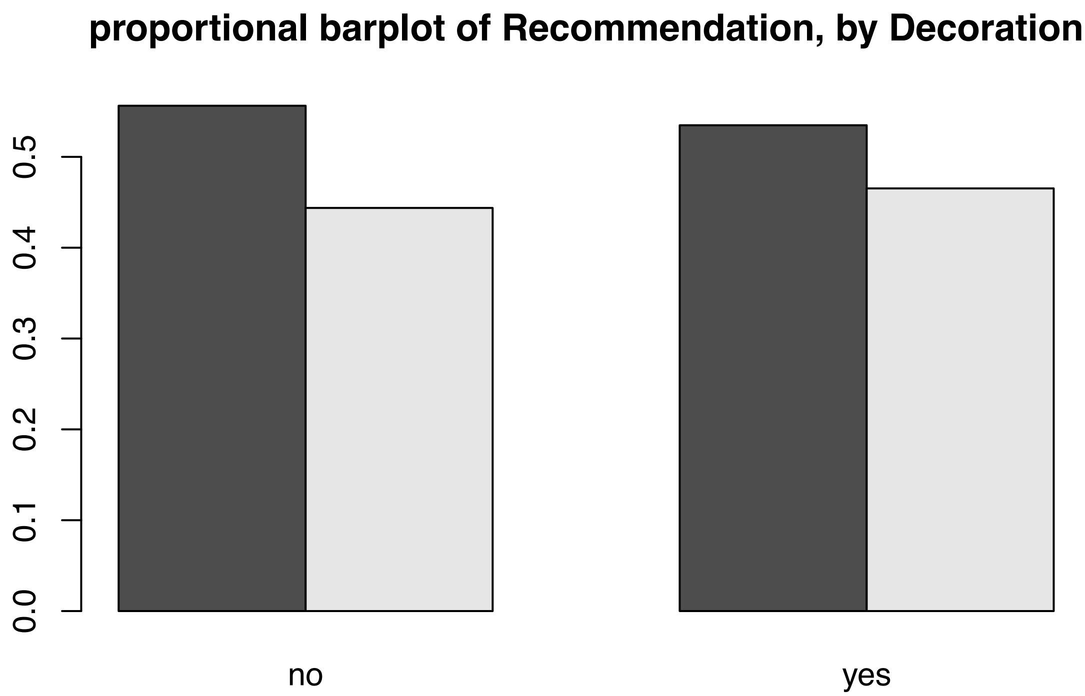
Next, we have the variable NeckLine. Surprisingly, the relative probability of dresses with each neckline being well sold had only minor differences. Vneck had the highest probability while Oneck had the lowest probability. However, the difference between the two are only about 3%.

```
##
##        cotton     other
##   0 0.5504587 0.5420168
##   1 0.4495413 0.4579832
```



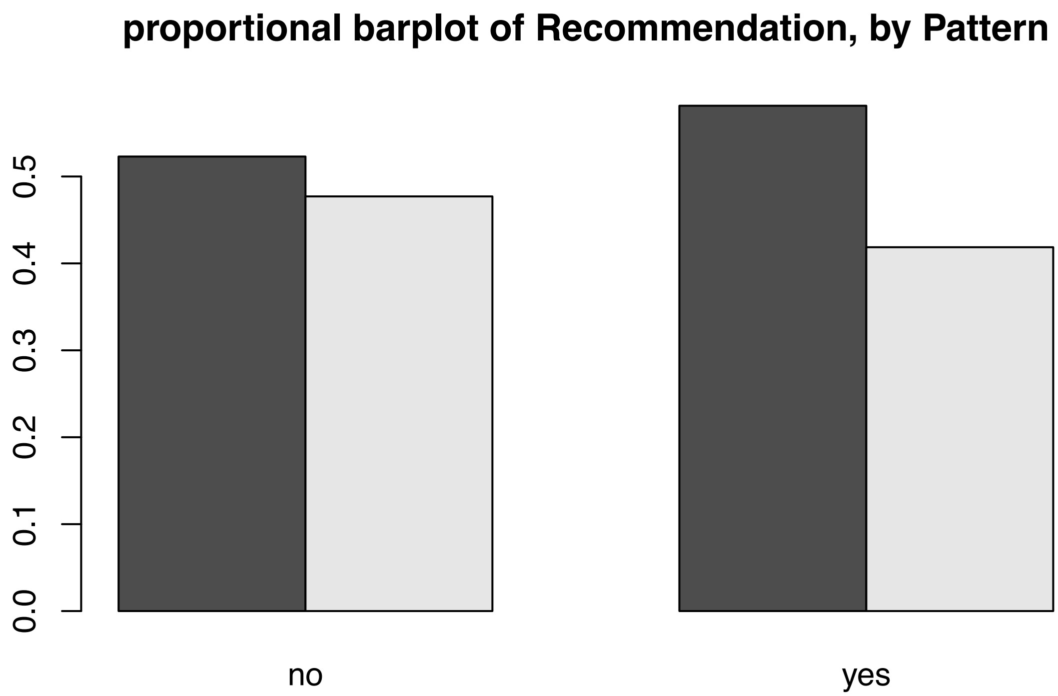
Next, we have the variable Material. Surprisingly, whether a dress is made out of cotton or other do not infleunce the probability of it being well sold by very much. Dresses made of other materials had slightly higher probability of being well sold than cotton dresses, by only about 1%.

```
##
##            no       yes
##   0 0.5562500 0.5347594
##   1 0.4437500 0.4652406
```



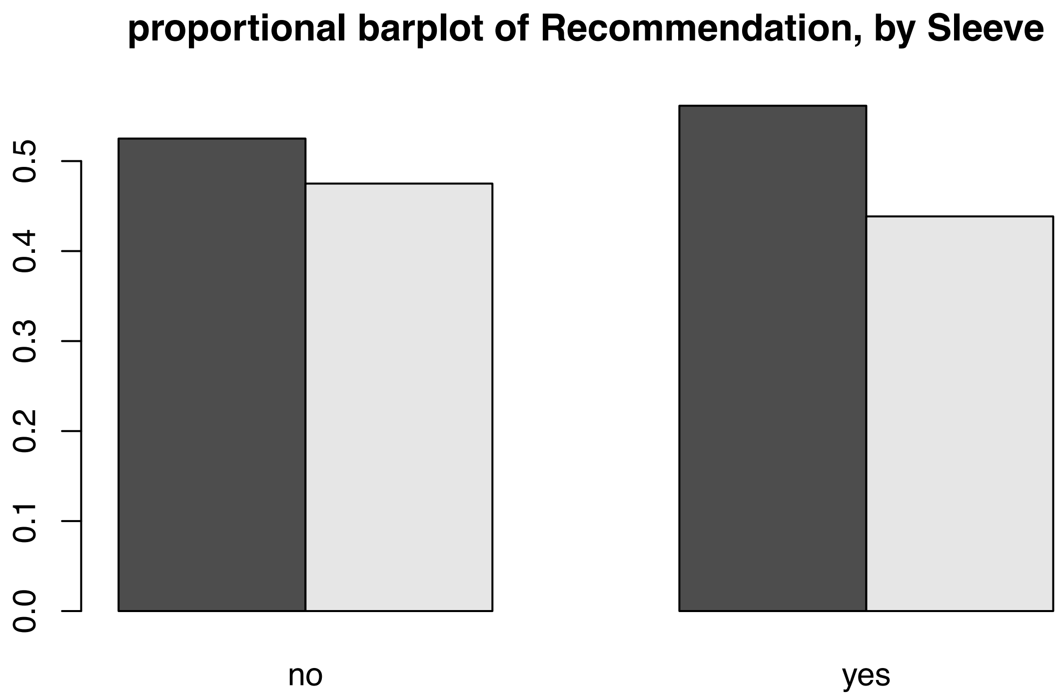
Next, we have the variable Decoration. Surprisingly, whether a dress has decoration or not do not infleunce the probability of it being well sold by very much. Decorated dresses had slightly higher probability of being well sold than undecorated dresses, by only about 2%.

```
##
##             no       yes
##   0 0.5229358 0.5813953
##   1 0.4770642 0.4186047
```



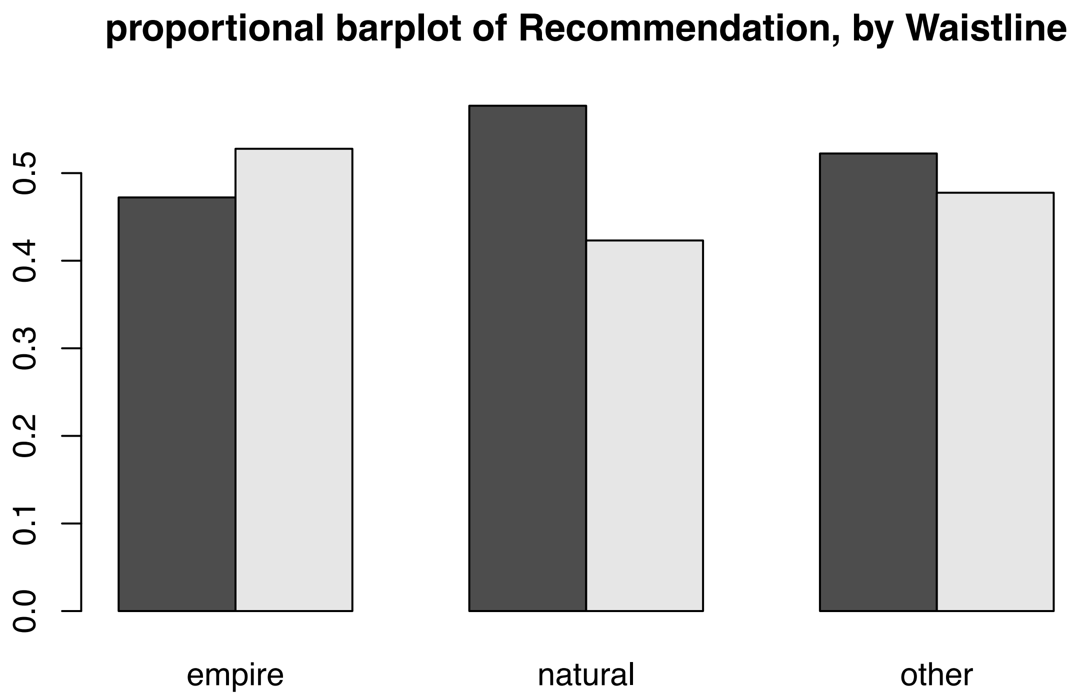
Next, we have the variable Pattern. Patterned dresses had lower probability of being well sold than solid color dresses.

```
##
##             no       yes
##   0 0.5250000 0.5614973
##   1 0.4750000 0.4385027
```



Next, we have the variable Sleeve. Sleeveless dresses had higher probability of being well sold than sleeved color dresses.

```
##
##        empire   natural     other
##   0 0.4722222 0.5769231 0.5223881
##   1 0.5277778 0.4230769 0.4776119
```



Lastly, we have the variable Waistline. Empire waistlined dresses had the highest probability of being well sold while natural waistlined dresses had the lowest probability of being well sold. Interestingly, the difference among the three probabilities are about 5% each.

It is important to note that categorical variables with minor differences might not be the most appropriate choice as predictors, since there are not a lot of difference between the categories (such as Decoration, Materials, Sleeve, Pattern, Neckline).

Since there is only one quantitative predictor, we will not create bivariate/pairs plot on quantitative predictors.

Modeling

We now turn to building and assessing our classifiers for predicting the recommendation. Our four classifiers are: linear discriminant analysis (lda), quadratic discriminant analysis (qda), classification trees, and binary logistic regression.

To ensure that our models are not overfitting to our sample, we randomly split our observations into training and test sets. All four models were built using the same training observations and assessed on the same set of test observations.

Linear Discriminant Analysis (LDA) For our LDA and QDA models, we use only the quantitative variables (Rating). The LDA classifier is built on the training data as follows:

Then we investigate the performance of the LDA classifier on our test data as follows:

```
##
##          0    1
```

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

On the test data, LDA gave an overall error rate of (49+0)/149 = 0.328859/32.89% which is quite low. In particular, we do best at finding the not well sold dresses (error rate of a perfect 0/100 = 0). Our LDA has a much higher (100%) error rate for classifying well sold dresses (49+0/49).

Quadratic Discriminant Analysis (QDA)

Similarly, we use our one quantitative variable for training a QDA classifier as follows:

And we investigate the performance of the QDA classifier on our test data as follows:

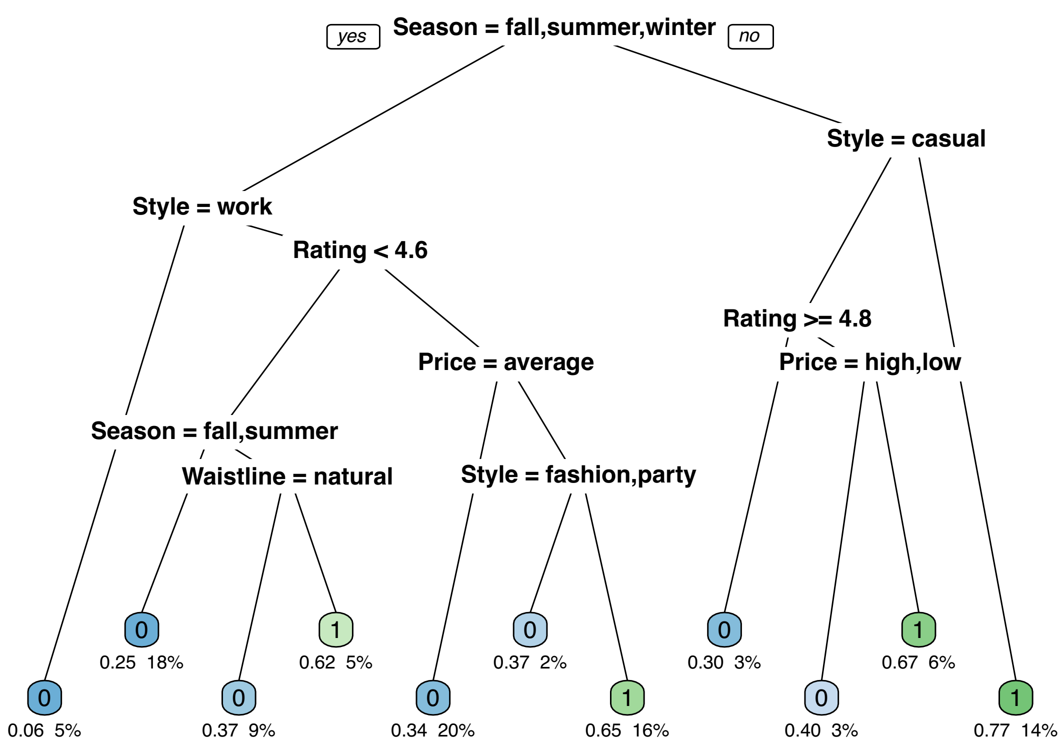
```
##
##         0    1
##    0 100   49
##    1   0    0
```

Because we only had one quantitative predictor, surprisingly, our results from the QDA is identical as the LDA. With the exact same all three error rates. With a 100% error rate for classifying true well sold dresses and 0% error rate for classifying not well sold dresses, our LDA and QDA model could be underfitting for one classification while overfitting for the other at the same time. Overall, our model has an error rate of near 33% for both the QDA and LDA. This error rate is workable until further modeling (the next two models), but we would generally want to reduce it but not to the extent of possibly overfitting.

Classification Trees

While we could only take into account the quantitative variables in the LDA and QDA classifiers, we can also account for the nine categorical variables in a classification tree.

We fit a classification tree on the training data and plot it, as follows:



We note that the classification tree selected **Season, Style, Rating, Price, and Waistline** to use to classify the wine type. Season is the most important variable to start our classification tree off with, based on a computer calculated Gini index. [In general, the "most important" variables that the tree determines for classification will be indicated from top down on the tree.] We then investigate the performance of the tree classifier on our test data as follows:

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

The results from the classification tree is not better than our results from the QDA and LDA. Our classification tree model has an overall error rate about 40.27%, which is a bit higher than our error rates from the previous two models. The error rate for not well sold dresses, ($29/100 = 29\%$) are better than the error rate for well sold dresses ($31/49 = 63.27\%$). Based on the overall error rate on test data for the dresses, the LDA and QDA are still slightly better models.

Binary Logistic Regression

Finally, we consider binary logistic regression to model the recommendation of dresses. Similarly to the classification trees, a logistic classifier can use all the variables including all the quantitative and categorical variables.

We train a logistic classifier on the training data, and then inspect the resulting confusion matrix from the test data, as follows:

We first fit a binary logistic regression to the data as follows:

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

We then apply the logistic model to the test data:

```
dress.logit.prob <- predict(dress.logit, as.data.frame(dress_test),
                            type = "response")
```

Since the logistic model applied to the test data yields probabilities (not [well sold/not well sold] classification), we will convert the logistic probabilities into classification predictions by thresholding the probability, so that if prob>0.5 we will classify it as one type of Recommendation (else, classify as the other type).

In order to associate the correct direction of probability with the appropriate dress recommendation, we need to see how "Recommendation" is default ordered. We do that by running "levels" on the factored response variable, as follows:

```
levels(factor(dress_test$Recommendation))
```

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

We then obtain test classification from the logistic model using a threshold probability of 0.5, as follows:

```
dress.logit.pred <-ifelse(dress.logit.prob > 0.5,"1","0")
```

We then evaluate how the the logistic classifier performed on our test data with a confusion matrix as shown:

```
table(dress.logit.pred, dress_test$Recommendation)
```

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

The logistic model as a classifier (using threshold probability of 0.5) performs nearly as well as QDA and LDA, with overall error rate of 0.33557 (27+23/149 = 33.557%). For not well sold dresses, it gives an error rate of 0.23 (23/100 = 23%), and for well sold dresses, it gives an error rate of of 0.551 (27/49 = 55.1%).

We note that as for classification tree, the logistic regression classifier performed better on both well sold and not well sold dresses.

Final Recommendation

Of the four classifiers we tested, the QDA and LDA performed the best . Logistic regression performed nearly as well as QDA & LDA overall. The classification tree model had the highest overall error rate on the test data. There is seemingly issue with overfitting for the classification tree and the binary logistic regression model, since the error rates are not nearly as perfect.

We note that LDA and QDA performed identical to each other, with a perfect error rate for not well sold dresses (0%) and a completely flawed error rate for well sold dresses (100%). Because of this extremity (overfitting for one and underfitting for the other) that both LDA and QDA have, we will take these two models out of the consideration.

Our final recommendation is the binary logistic regression model. Of the four classifier, the binary logistic regression model an error rate of that is almost identical to our LDA and QDA model. Since we ruled out those two models, then the binary logistic regression model is the most appropriate model for us. It had the lowest error rate out of all four models for well sold dresses, and a lower error rate for not well sold dresses than the classification tree model. Even though it is not perfect (error rate of 33% is a bit high), but it is our best model to predict whether the dress sells well or not based on our data from Air University for the current store.

Discussion

Overall, our models were satisfactory at classifying dress recommendations. We note the surprising perfect performance of QDA and LDA on not well sold dresses, and remark that an overfitting issue may be remedied by a random forest, if enough data can be found. The overall error rate might not be perfect, but it is workable since our real world data might contain flaws that could influence the calculated error rates through the collection process.

It would probably be helpful to collect more data, for example, from the previous years also. It would also be helpful to have some of the categorical variables as quantitative variables, for example, price could be in dollars instead of categories. If we were to collect more quantitative variables, then the results from the LDA and QDA would be more different (possibly in a good way since QDA and LDA only consists of quantitative variables).

Other areas for future research that could be of greater interest to the industry would be to build models to predict the multinomial recommendation variable. The response variable could be also modified since well sold or not is a subjective measurements. Different store owners from different backgrounds would have different perception on how well selling is well. We could extend our sample to other stores as well instead of focusing on one store. This combination would allow data to be collected from store that had a majority of well sold dresses or a store that had extremely bad records in dress selling. Other areas of interests could be also explored in the future as well.