Dress Popularity: Predicting Which Dresses At a Clothing Shop Will Sell

Colin Yip
cdyip

Due Wed, April 20, at 8:00PM

Contents

Introduction

As a struggling clothing store, it is vital that for next year the store orders dresses that will sell well. Predicting which dresses sell well is also important to customers, competitors, and others for the following reasons:

1. Competitors want to know which dresses sell well to beat out other stores in the area.
2. Individuals want to know which dresses sell well to determine whether they are getting a fair price or not.
3. Banks want to know which dresses sell well to determine which stores they should give loans to. The bank would be more inclined to give a loan to a store that sells popular dresses than one which sells unpopular ones.

We seek to answer the following research question: "What factors determine wheter a dress will sell well?" We perform quantitative analysis to determine which factors are most important and classify dresses as sells well or does not sell well.

Exploratory Data Analysis

Overview/Background of Dataset

The data was collected from records of dress sales from last year at the clothing store. For each dress, we have the following explanatory variables:

- Style: style of the dress (cute, work, casual, fashion, or party)
- Price: price range (low, average, or high)
- Rating: average rating from dress factory market survey (average of stars, from 0-5)
- Season: the season the dress is appropriate for (summer, fall, winter, or spring)
- NeckLine: neckline type (O-neck, V-neck, or other)
- Material: wether the material is cotton or not (coton/other)
- Decoration: whether the dress has any decoration or not (yes/no)

- Pattern: whether the fabric has a pattern or not (yes/no)
- Sleeve: whether the dress has a sleeve or not (yes/no)
- Waistline: waistline type (other, empire, or natural)

We also have the following response variable that we aim to predict:

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

Exploratory Data Analysis on Response Variable

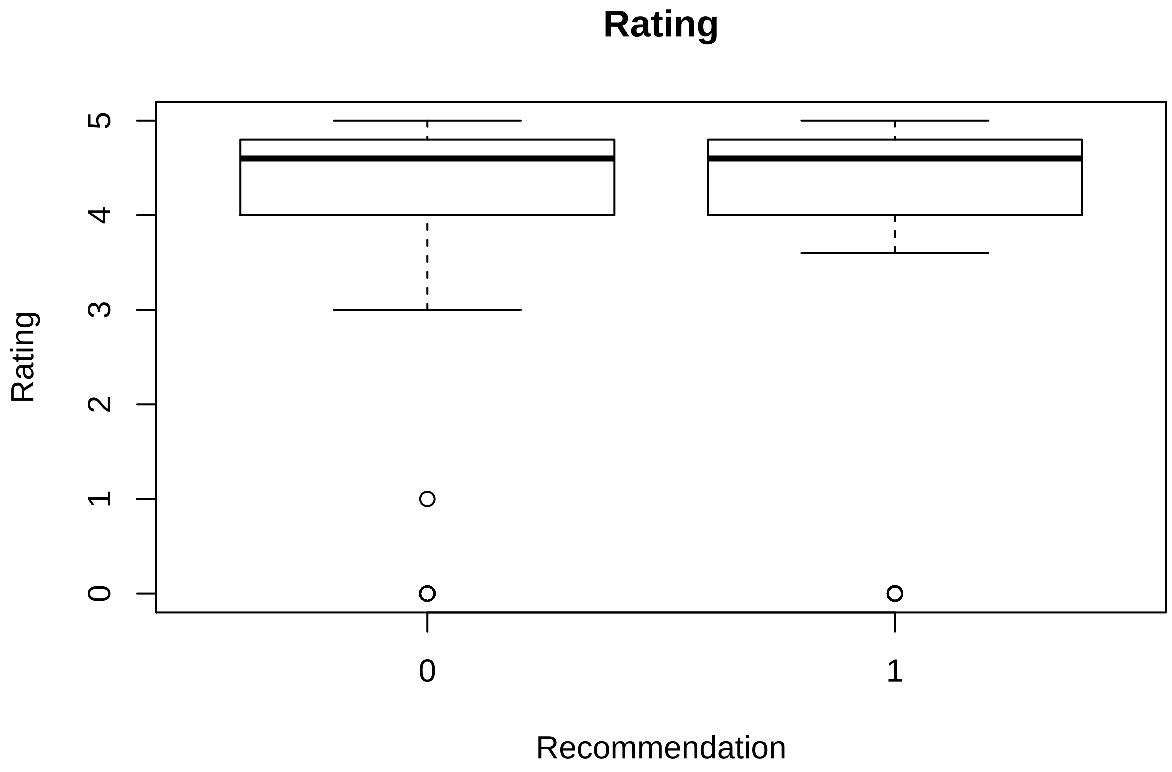
From the following tables, we observe that our training dataset has 347 total dresses. Of those, 158 dresses (about 45.5% of the sample) sell well and 189 (about 54.7% of the sample) do not.

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

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

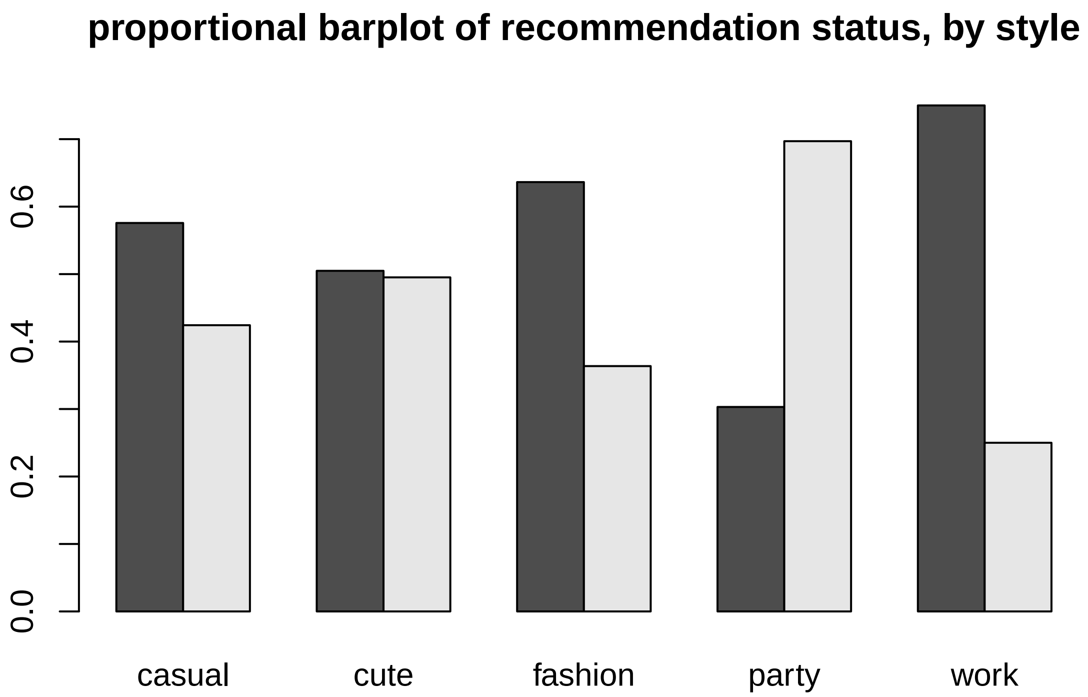
Exploratory Data Anlysis on Response Variable and Each Explanatory Variable

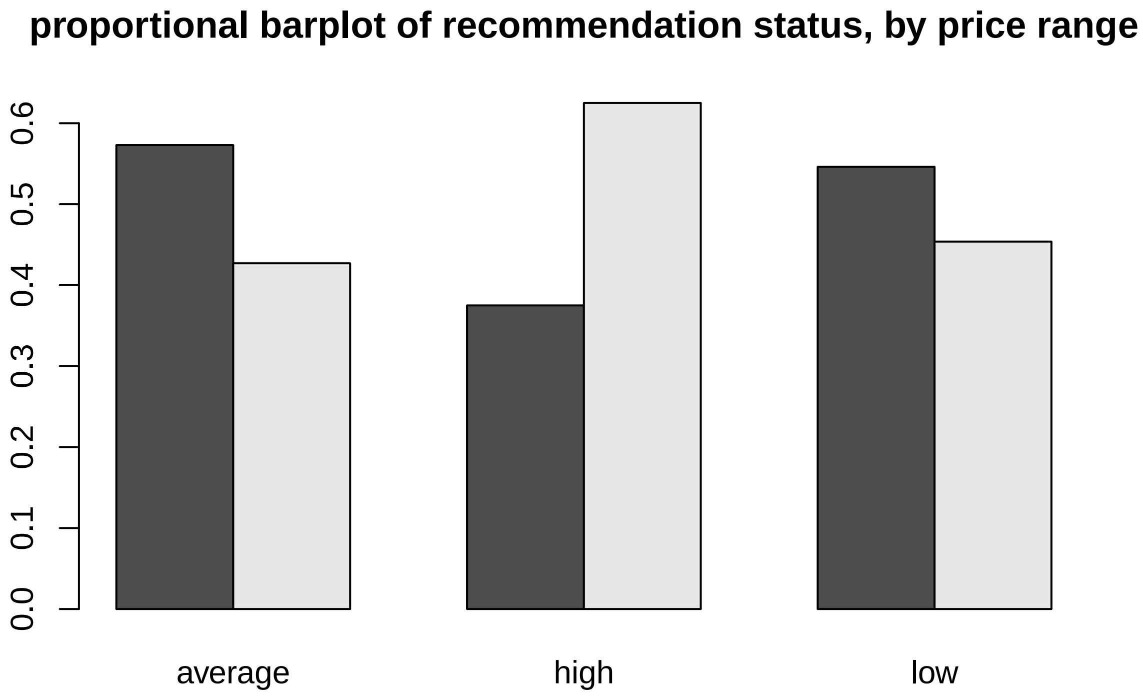
To determine which explantory variables will be most useful to predict the response variable (Recommendation), we perform Exploratory Data Analysis (EDA) on the response variable and each explanotary variables. We start by making boxplots to display the relationship between Recommendation and the only quantitative explanatory variable (Rating).

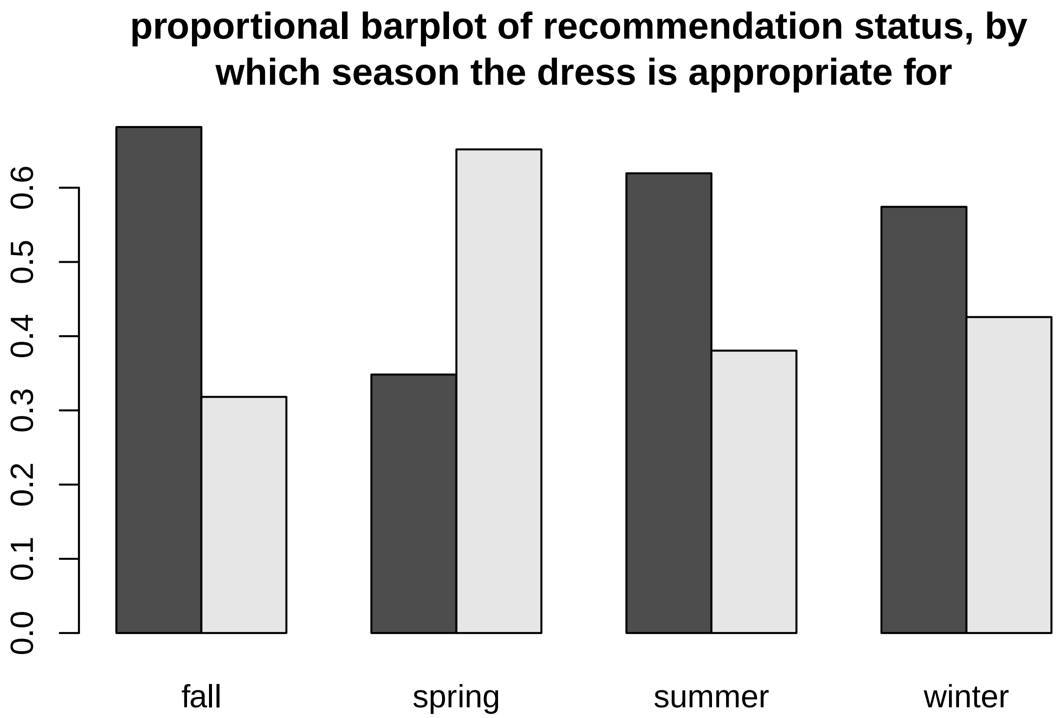


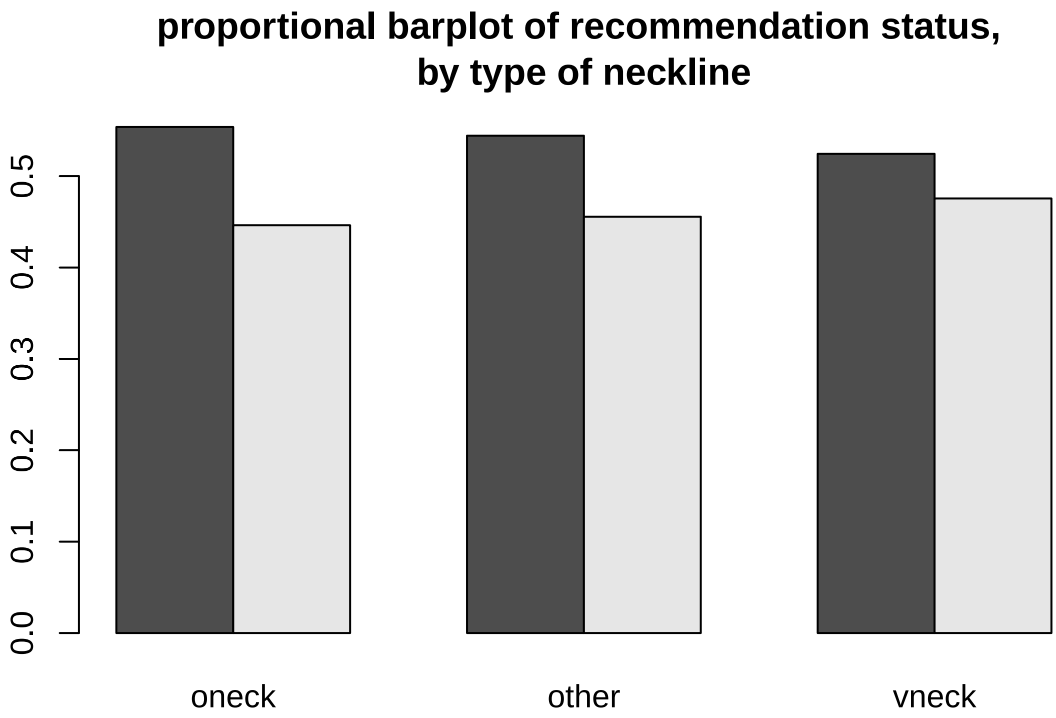
We observe that there does not appear to be a significant relationship between average customer rating of a dress and whether the dress sells well. The spread of Rating for dresses that did not sell well (Recommendation = 0) is larger than the spread Rating for dresses that did sell well (Recommendation = 1).

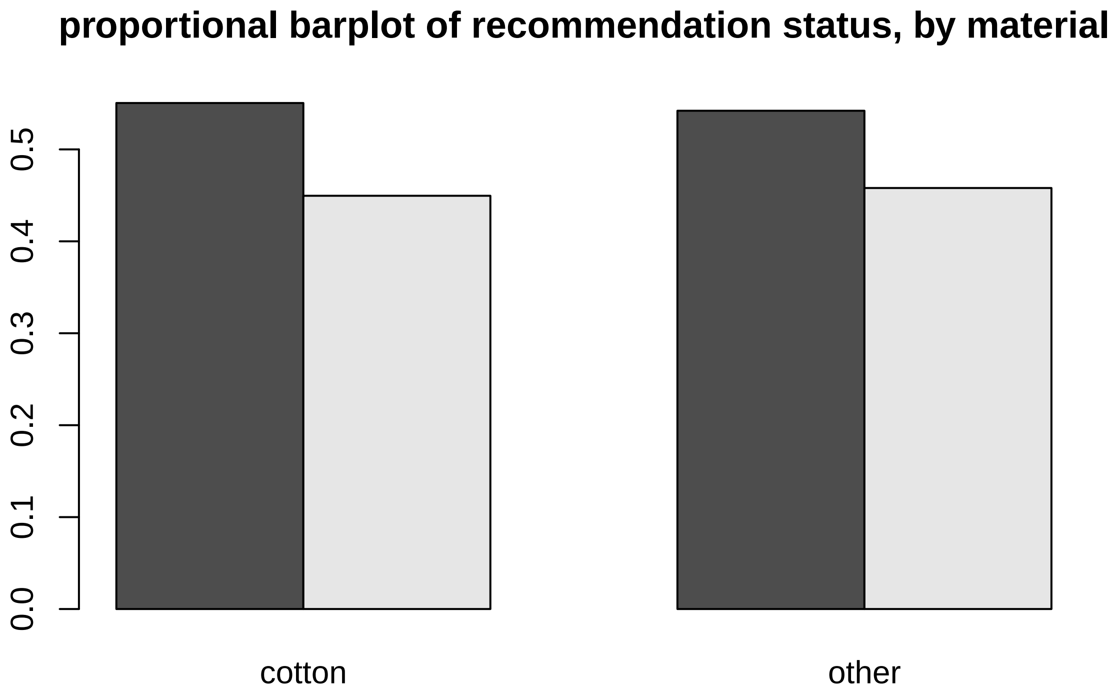
We next create bar charts to analyze the relationship between Recommendation and each of the categorical explanatory variables.

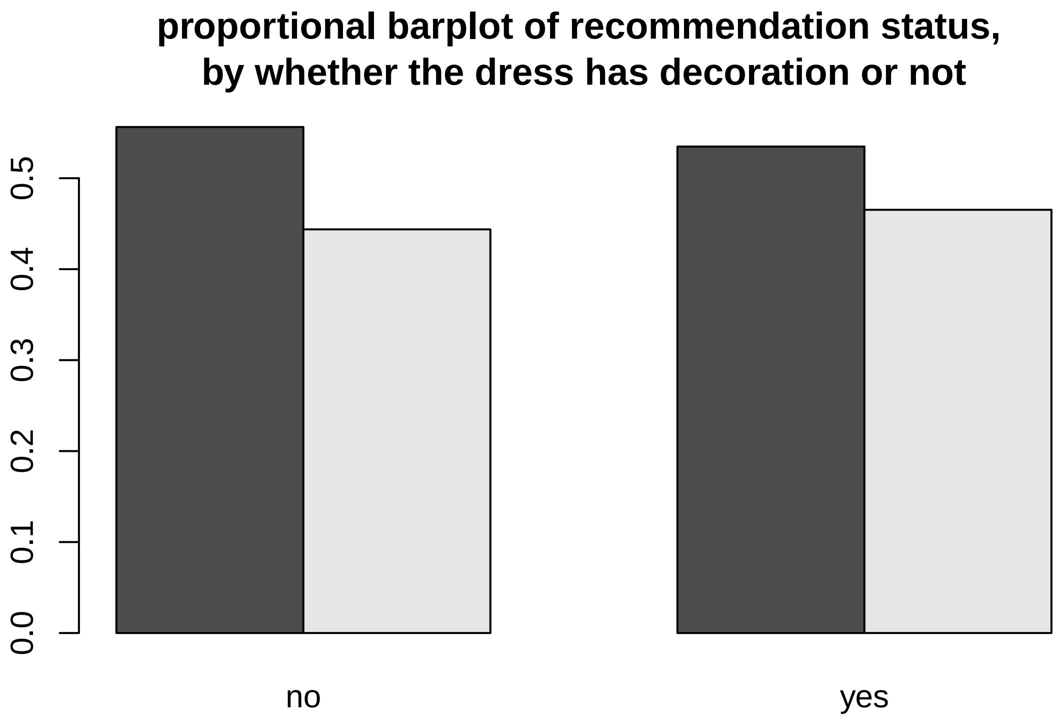



proportional barplot of recommendation status, by price range
0.0
0.1
0.2
0.3
0.4
0.5
0.6
average
high
low

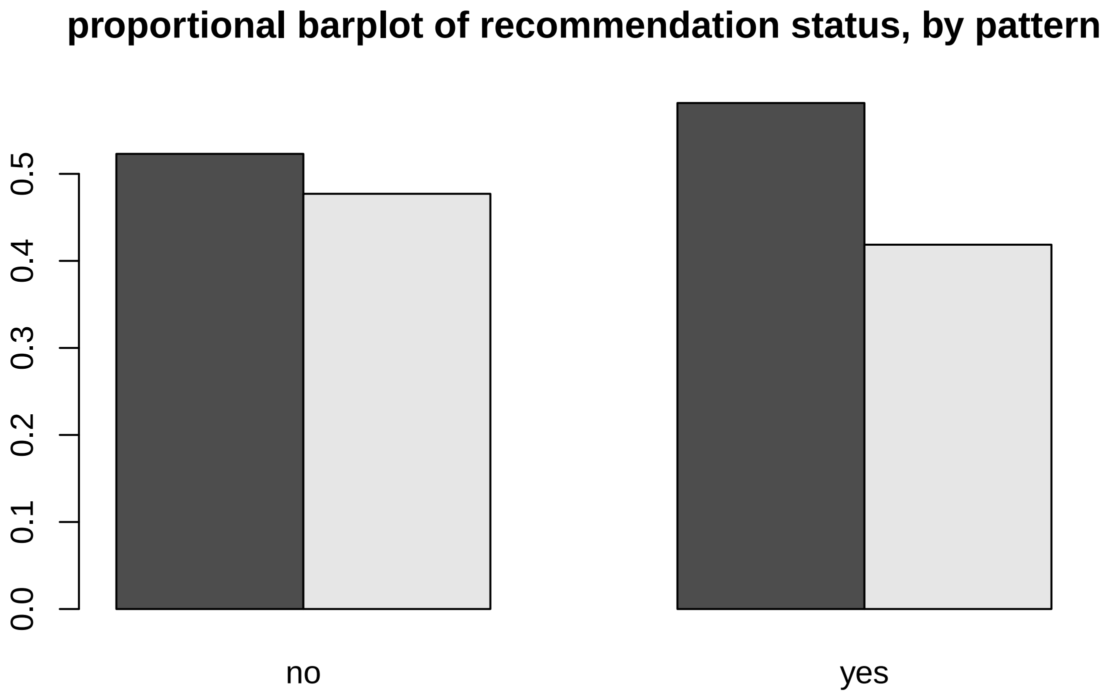

proportional barplot of recommendation status, by
which season the dress is appropriate for
0.0
0.1
0.2
0.3
0.4
0.5
0.6
fall
spring
summer
winter

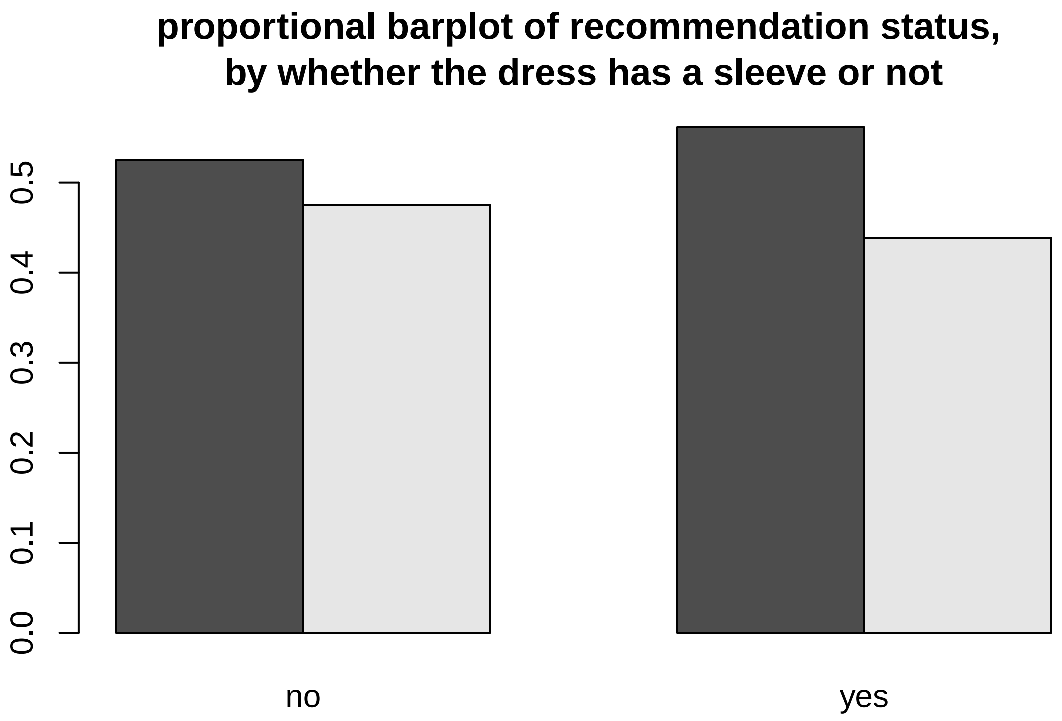

proportional barplot of recommendation status,
by type of neckline
0.0
0.1
0.2
0.3
0.4
0.5
oneck
other
vneck

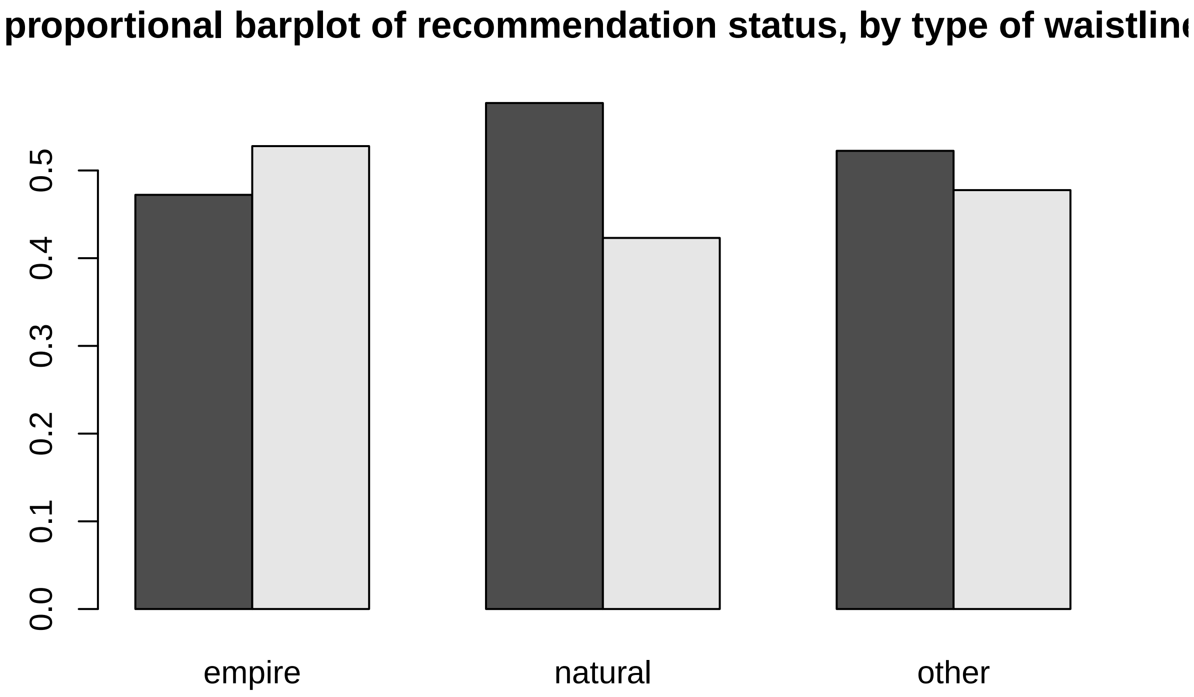

proportional barplot of recommendation status, by material
0.0
0.1
0.2
0.3
0.4
0.5
cotton
other



proportional barplot of recommendation status, by whether the dress has decoration or not


proportional barplot of recommendation status, by pattern
0.0
0.1
0.2
0.3
0.4
0.5
no
yes


proportional barplot of recommendation status,
by whether the dress has a sleeve or not
0.0
0.1
0.2
0.3
0.4
0.5
no
yes



proportional barplot of recommendation status, by type of waistlin

We are interested in explanatory variables which have a significant relationship with Recommendation because those may be useful to classify which dresses will sell well. It appears that dresses that sell well make up a larger percentage of party dresses than work or fashion dresses. Dresses that sell well represent a larger proportion of spring dresses than dresses of any other season. Dresses that do not sell well make up a lower proportion of low and average priced dresses than high priced dresses. Overall, based on difference in the heights of bars, these three variables appear to be the most significant out of the nine categorical explanatory variables.

Modeling

We now build our model to classify whether a dress sells well or does not sell well. We first create an LDA model using all of the predictors. We only use continuous variables (Rating).

We build the model using our training data and then evaluate its performance using the test data. Below is the confusion matrix for the LDA model.

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

We observe that the LDA model has an overall error rate of (0+149)/149 = 0.329. The model classify every dress as Recommendation = 0 (does not sell well). The inaccuracy of the model may be caused by the fact that there is only one explanatory variable (since there was only one quantitative, continuous variable in this dataset).

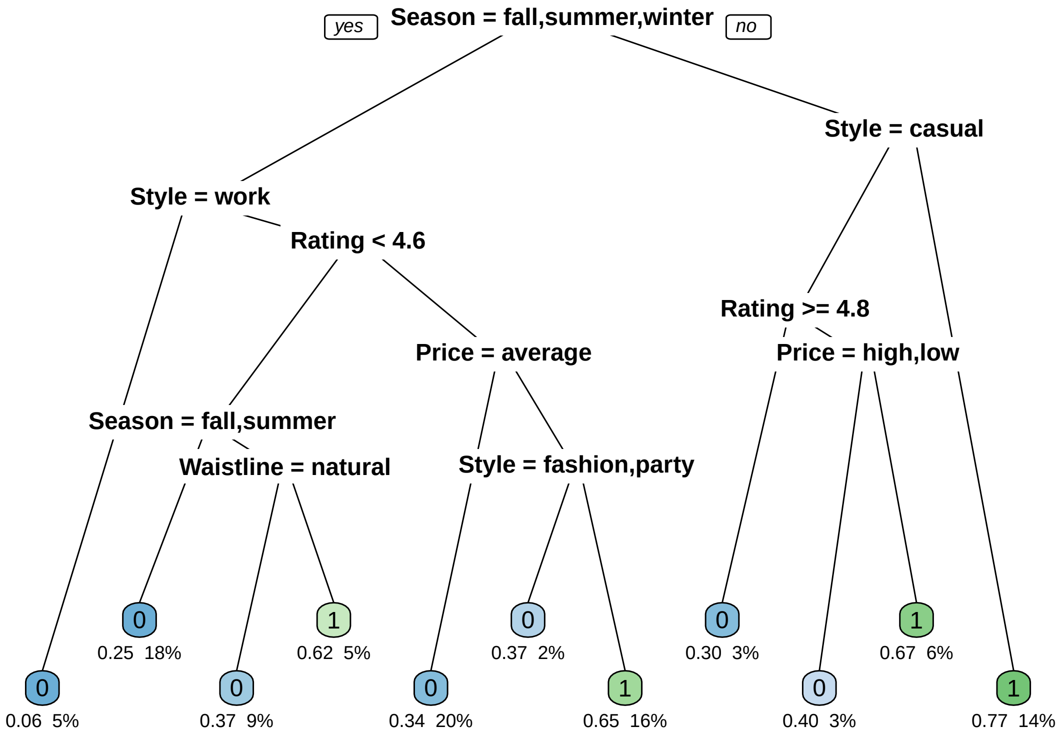
We next build the QDA model using all of the quantitative, continuous variables and evaluate its performance using the training data. Below is the confusion matrix for the QDA model.

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

We observe that the QDA model has an overall error rate of (0+149)/149 = 0.329. This model also classifies every dress as Recommendation = 0 (does not sell well). The QDA model has the same issue as the LDA model: there is only one explanatory variable since only one of the variables in the dataset was quantitative and continuous.

Classification Tree

We next build a classification tree, which can include categorical variables. The following is the classification tree and its confusion matrix.



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

We observe that the overall error rate is (31+29)/149 = 0.403. The season a dress is appropriate for was judged to be the most important variable since it is at the top of the classification tree. Due to the high number of variables used in the tree, the model may have been overfitted to the training data.

Binary Logistic Regression

Finally, we will use a binary logistic regression model to predict whether a dress will sell well or not. We first fit the model to the training data. Since the logistic model produces probabilities and not classifications, we will classify a dress as Reccomendation = 1 (sells well) if the probability that it sells well > 0.5. If not, we classify it as Recommendation = 0 (does not sell well). We then evaluate its performance using the confusion matrix of the model with the test data.

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

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

The logistic regression model has an overall error rate of (23+27)/149 = 0.336. We observe that the logistic regression model performs better for dresses that don't sell well than dresses that do sell well. The logistic model has a lower error rate than the classifcation tree (0.336 to 0.403) and a higher error rate than the LDA and QDA models (0.336 to 0.329).

Final Decision

We will use the logistic regression model as our final model.

Although they had lower error rates, the LDA and QDA models both classified all dresses as does not sell well, which suggests they do not predict recommendation status well. They also only used one out of the available ten explanatory variables.

We choose the logistic regression model over the classification tree because it has a lower error rate. Classification trees, especially when using a high number of explanatory variables (ten, in this study) are susceptible to overfitting. A pruned classification tree could fix this issue and would be a viable alternative to the logistic regression model.

Discussion

We recommend the logistic regression model as the final model because it can use categorical variables and has a lower error rate than the classification tree.

Overall, the model did well at classifying which dresses sell well. The store, competing stores, and investors can use this model to predict which dresses to buy and which stores will be successful.

One limitation of the model is that it is difficult to explain to nonexperts; the interpretation of parameters and arbitrary probability threshold are complicated topics. Therefore, a pruned classification tree could be an option for future research. The current classification tree may be overfitted to the data and with pruning, the error rate could be decreased.

For future analysis of which dresses sell well, more quantitative variables could be added. For example, change the Price variable from categorical price ranges to quantitative dollar amounts. This would make LDA and QDA more viable options, which may produce a better model.