The Science of Dress Profitability

Wenting Xu, Yukun Tan, Chuan Du

December 21, 2017

Abstract

Designing stylish and comfortable dresses has been challenging to adapt to a more competitive clothing industry. To investigate what significantly affect consumers' decisions to purchase dresses and what kind of dresses are more marketable, statistical models including KNN, elastic net, random forest and boosted tree are applied to analyze 500 dresses, using their Sale status (good or bad) as response and 14 variables as predictors. Based on the comparison in test accuracy, the elastic net model performs the best, implying that dress producers and designers could use the elastic net model to predict the profitability of their newly designed dresses or alter values of essential attributes to improve sales. Codes and tables are displayed to illustrate the data and analysis.

Introduction

With the advancement of technology and capitalization, the clothing industry is becoming more competitive by the minute. In a market that promotes creativity as well as comfortable lifestyles, making a stylish dress could appear challenging. What are the significant factors affecting consumer's decisions when making a purchase? Dresses of which material, pattern or design are more marketable? In this analysis, we are going to use statistical models to investigate the profitability of dresses based on their attributes. From a dress producer or designer's perspective, this information is crucial to making strategic decisions regarding which kind of dresses to produce to potentially maximize profit and mitigate loss.

Materials and Methods

Data

The dataset used in this analysis contains attributes of 500 dresses and their recommendations according to their sales (either good sale or poor sale). Recommendations are monitored by alternate days. The dataset was found in the UCI Machine Learning Repository. The source of the dataset is from Muhammad Usman & Adeel Ahmed, Air University, students at Air University. There are 500 observations in the dataset, with a few missing values. There are 14 variables in the dataset, 11 of which will be used as predictors and the variable `Recommendation` is used as the response, which is a categorical variable with 0 representing the dress is the poor sale and 1 representing the good sale. In this analysis, We will construct models in order to distinguish the significant factors that affect consumer's purchasing decisions, and predict the profitability of dresses before they are released in the market.

We dropped two variables, `Dress_ID` and `Rating`. Variable `Dress_ID` represents the unique ID of each dress, which will not be used in any analysis. Variable `Rating` will also not be used in any analysis, since customer rating is not an intrinsic characteristic of dresses. Besides, since the goal of this analysis is to predict the profitability of dresses before they are released in the market, customer rating, which is made after the dress is purchased (and indeed after the dress is released), is not a factor to be considered.

```
# handling missing values
dress = na.omit(dress_raw)

# dropping Rate and ID variables
dress = subset(dress, select = -c(Rating, Dress_ID))
```

Table 1: Name of Attributes

Style	Price	Size	Season	NeckLine	SleeveLength	WaistLine	Material	FabricType	Decoration	PatternType

This dataset required a significant amount of data cleaning. First, missing values were omitted with `na.omit`. In the original dataset, some dresses' price levels are in both upper case and lower case (e.g. "low", "Low"), so they were made all consistent with the first letter in upper case (e.g. "Low"). Also, some dresses' sizes were in full name (e.g. "small"), while the others used the abbreviation (e.g. "s" or "S"), so they were all changed into uppercase abbreviation (e.g. "S"). Also, wrong spellings were corrected, such as "Automn" into "Autumn" and "wollen" into "woolen." Moreover, for dresses' Fabric Type, Decoration, and Pattern Type, some of the values are shown as "null." Based on real-world experience, for a dress, the fabric type has to exist while decoration and patterns don't have to exist. For example, there can be a white dress with no decoration and patterns, but it has to have its natural fabric type. Thus, for dresses with "null" in Fabric Type, they were categorized into "other", while those with those with "null" in Decoration and Pattern Type were categorized into "none", where "other" and "none" have existed groups in the original dataset. The code below shows the data cleaning process mentioned above, which serves to illustrate the types of data cleaning applied to the dataset. Other similar cleaning processes are done with code omitted.

```
# handling missing values
dress = na.omit(dress_raw)

# making upper and lower cases consistent
dress_raw$Price[dress_raw$Price == 'low'] = 'Low'

# making full name and abbreviation consistent
dress_raw$Size[dress_raw$Size == 'small'] = 'S'
dress_raw$Size[dress_raw$Size == 's'] = 'S'

# correcting wrong spellings
dress_raw$Season[dress_raw$Season == 'Automn'] = 'Autumn'
dress_raw$FabricType[dress_raw$FabricType == 'wollen'] = 'woolen'

# handling "null" values in Fabric Type, Decoration and Pattern Type
dress_raw$FabricType[dress_raw$FabricType == 'null'] = 'other'
dress_raw$Decoration[dress_raw$Decoration == 'null'] = 'none'
dress_raw$PatternType[dress_raw$PatternType == 'null']  = 'none'
```

Variable `Dress_ID` represents the unique ID of each dress, which will not be used in any analysis. Variable `Rating` will also not be used in any analysis. The goal of this analysis is to predict the profitability of dresses before they are released in the market. Thus customer rating, which is made after the dress is purchased (and certainly after the dress is released), is not a factor to consider. The response `Recommendation` is a categorical variable with 0 representing the dress is bad-selling and 1 representing good-selling.

The variables used as predictors are summarized in Table 1. For a detailed description of the predictors, refer to Table 4 in the Appendix.

Finally, we set up the training and testing data, and define two helper functions for later use. One is `get_best_result` function to get the best model among different models with different tuning parameters. The other is `accuracy` function to get the accuracy of the selected model.

```
# train & test split
dress_idx = sample(1:nrow(dress), size = round(0.7 * nrow(dress)))
dress_train = dress[dress_idx, ]
dress_test = dress[-dress_idx, ]
```

Models

Four models were used to analyze the data, which are KNN, elastic net, random forest and boosted tree. To properly estimate model prediction performance, the resampling method used is 5-fold cross-validation. Since the attributes of our dataset are all categorical variables, the distribution of the data is not observable. Thus discriminant analysis is not considered in this analysis. The model training process and the tuning grid used for each model is shown in the subsections below. The best model in each case and its test accuracy is also recorded using the helper functions defined above.

```
#use 5-fold cross validation
cv_5 = trainControl(method = "cv", number = 5)
```

K-Nearest Neighbors Model

We choose 26 numbers to be the number of neighbors for KNN model. The risk of overfitting would decrease as the number of neighbors increase.

```
dress_knn = train(
  Recommendation ~ .,
  data = dress_train,
  method = "knn",
  trControl = cv_5,
  tuneGrid = expand.grid(k = seq(1, 51, by = 2))
)
```

Elastic Net Model

Instead of using logistics regression without regularization, the elastic net method is considered to reduce the risk of overfitting. `Tunelength` is set to be 20, which would provide a large result.

```
dress_elastic = train(
  Recommendation ~ .,
  data = dress_train,
  method = "glmnet",
  trControl = trainControl(method = "cv", number = 5),
  tuneLength = 20
)
```

Random Forest Model

'Out of Bag' resampling solution is used instead of 'Cross-Validation' here, since OOB is more computationally efficient and yield similar results as would CV.

```
# resampling solution
oob = trainControl(method = "oob")
```

```
dress_rf = train(
  Recommendation ~ .,
  data = dress_train,
  method = "rf",
  metric = "Accuracy",
  trControl = oob,
  tuneGrid = expand.grid(.mtry = seq(1, 30))
)
```

Boosted Tree Model

We set up a training grid with different tuning parameters for the boosted tree to achieve a more accurate result.

```
gbm_grid =  expand.grid(interaction.depth = 1:5,
                        n.trees = (1:6) * 500,
                        shrinkage = c(0.001, 0.01, 0.1),
                        n.minobsinnode = 10)
```

```
boosted = train(Recommendation ~ ., data = dress_train,
                        method = "gbm",
                        trControl = cv_5,
                        verbose = FALSE,
                        tuneGrid = gbm_grid)
```

Results

Test accuracy is used as a metric to compare the performance of models. The final results of the models are summarized in the table below.

Table 2: Model Performance and Tuning Parameters

Method	Tuning Parameter	Parameter Value	Resampled Accuracy	Test Accuracy
KNN	"k"	23	0.6495506	0.527027
Elastic Net	"alpha"	1	0.6290464	0.6418919
	"lambda"	0.039152		
Random Forest	"mtry"	0.6405797	0.2052168	0.6081081
Boosted Tree	"shrinkage"	0.01	0.6408537	0.6148649
	"interaction.depth"	1		
	"n.trees"	500		

Discussion

Based on table 2, the elastic net model performs the best with a test accuracy around 64%. This result implies that the probability of the classifier correctly classifies of dresses based on the attributes is 64%. The elastic net model is linear, parametric, and discriminative. The performances of the random forest model and the boosted tree model are also close to 64%, and those models are nonparametric. Finally, the performance of the KNN model is not good and is very close to a random guess. Since the predictors in the model are all categorical with many levels, this might imply that the KNN model suffered from the curse of high dimensionality.

Thus using the elastic net model, dress producers and designers could predict the profitability of their new dresses before they release the dress into the market. If a dress is predicted to be not profitable, its producer or designer might reconsider its design, and potentially alter the values of some attributes to increase the profitability of the dress. A table is made below to investigate which attributes are significant in the model, showing the first few crucial variables and their respective importance.

	variables	importance

Table 3: Top Ten Variable Importance

	variables	importance
8	Styleparty	100.00000
20	SeasonSpring	94.90191
79	FabricTypepoplin	88.46561
23	NeckLineboat-neck	64.95628
41	SleeveLengthshort	61.55542
60	Materialnull	37.86713
3	Stylecute	37.27876
96	Decorationlace	34.20980
64	Materialrayon	28.91685
36	NeckLinev-neck	22.66252

From table 3, we can see that attributes `Style`, `Season`, `FabricType`, `NeckLine`, `SleeveLength`, `Material`, `Decoration` have some levels that are significant. This observation might indicate that changing the values of these attribute might have a greater impact on dresses' profitability. On the other hand, attributes with no significant levels are not important in the model. For example, `Size` and `Waistline` might not affect the profitability of dresses. This information makes sense because, from real-world experience, it is always the case that all clothes, including dresses, are of different sizes and waistlines.

Conclusion

Since the dress market is highly competitive, a way to predict the profitability of the newly designed dress is significant for the clothing industry to make a profit. In our analysis, we used KNN, elastic net, random forest, and boosted tree models to predict the profitability of the new dresses according to the attributes. Based on the result, the elastic net model was found with the highest accuracy among the four models we built. Also, seven attributes played significant roles in the dresses' attributes, which means that factors, such as `Style`, `Season`, `FabricType`, `NeckLine`, `SleeveLength`, `Material`, `Decoration`, will mostly affect customers' decisions when purchasing dresses. Overall, the result of the analysis is intended to improve the sales of dress for the clothing industry.

Appendix

Table 4: Description of Attributes

Attribute	Level	Description
Style	12	Categorical variable, describe the style of the dress. The style can be sexy, casual, vintage, brief, cute, bohemian, novelty, flare, party, work, OL, fashion.
Price	5	Categorical variable, describe the price of the dress. Price can be low, high, average, medium and very-high
Size	5	Categorical variable, describe the size of the dress. It contains five size, S, M, L, XL, Free
Season	4	Categorical variable, describe the season for the dress. Four seasons listed, Spring, Summer, Autumn, Winter

Attribute	Level	Description
NeckLine	15	Categorical variable, describe the neck line of the dress. It contains 15 levels, which are backless, boat-neck, bowneck, halter, mandarin-collor, o-neck, open, peterpan-collor ruffled, Scoop, slash-neck, sqare-collor, Sweetheart, turndowncollor,v-neck
SleeveLength	9	Catgorical variable, describe the sleeve length of the dress. The values are butterfly, capsleeves, full, halfsleeve, petal, short, sleeveless, threequarter, turndowncollor
WaistLine	5	Categorical variable, describe the waist line of the dress. It can be dropped, empire, natural, null, princess.
Material	21	Categorical variable, describe the material of the dress. It has 21 levels, which are acrylic, cashmere, chiffonfabric, cotton, knitting, lace, linen, lycra, microfiber, milksilk, mix, model, null, nylon, other, polyster, rayon, silk, spandex, viscos, wool
FabricType	18	Categorical variable, describe the fabric used to make the dress. It contains 18 values, which are batik, broadcloth, chiffon, Corduroy, dobby, flannel, jersey, knitting, lace, organza, other, poplin, satin, sattin, terry, tulle, woolen, worsted
Decoration	23	Categorical variable, describe the decoration attach on the dress, which can be applique, beading, bow, button, cascading, crystal, draped, embroidary, feathers, flowers, hollowout, lace, none, pearls, plain, pockets, rivet, ruched, ruffles, sashes, sequined, tassel, Tiered
PatternType	13	Categorical variable, describe the pattern on the dress. It can be animal, character, dot, floral, geometric, leopord, none, patchwork, plaid, print, solid, splice, striped
Recommendation	2	Categorical variable. It only has two levels, 0 and 1.

References

Usman, Muhammad, and Adeel Ahmed. "Dresses Attribute Sales Dataset." UCI Machine Learning Repository, 19 Feb. 2014. Retrieved from http://archive.ics.uci.edu/ml/datasets/Dresses_Attribute_Sales#