N Brown Data Science Recruitment Exercise – Size Prediction

1. Important information to predict customer's dress size

Detailed body measurements are important to fit the customer's body perfectly. To give an example, circumference of waist, hips, breast, shoulder, so forth are needed to provide a perfect size of clothes. However, it might be impractical to get all the details on clothing industry, especially on online method.

As can be seen on table 1, several factors are listed as important information for predicting the size of womenswear. Several sizes like corsetry cup, corsetry brief, footwear and tummy shape can help to figure the actual size of the customer's body. However, preference, age and socioeconomic are also needed to provide additional information of size.

No	Factors	Description
1	corsetry cup	identify the breast circumference
2	corsetry brief	identify the hips circumference
3	size of footwear	identify approximately height
4	tummy shape	approximate waist circumference (usually an option like average, flatter, curvier)
5	preference	the actual customer body's size can be up-size or down-size depends with their preference (such as fitted, average, loose)
6	age	age can help to predict the size
7	socioeconomic	location and economic condition can influence their body's size
8	brand	different brands might have slightly different measurement

Table 1 Factors which can predict customer's size

All factors except age and socioeconomic must be directly asked to the customer. These factors might be slightly more difficult than age and socioeconomic. Not all the customer can correctly provide the answers. However, age and socioeconomic can be obtained when a customer fulfills the registration form. As the previous information is usually needed to be completed on a registration form, therefore it is easier than acquiring their body's sizes.

2. Exploratory analysis

In terms of womenswear's size distribution, the most common is medium sizes like 16 and 20, while the least common are smallest and biggest such as 8, 36. This distribution also occurrs in size of corsetry brief with the similar range of size. Moreover, size footwear has same pattern like the previous features, there are high numbers on medium sizes and small numbers of smallest and biggest size. Age in years shares similar pattern like the former scale, the smaller numbers of age are 18 and below 100. However, a distinct trend on size corsetry cup such as high numbers on ranging from 2 to 6, after that the number is decreasing. There are three features such as days since first order, return rate, total number of orders which have similar distribution, a decreasing pattern.

In order to examine relationships between features and size womenswear, all factors have been calculated with correlation matrix. The strongest correlation is between size womenswear and size corsetry briefs (0.78), as bigger size tends to have bigger corsetry briefs. Total number of orders and days since first order have a strong association (0.73), as the number of orders are correlated with the time of customers have been signed up. Weak association occurrs between size footwear and size womenswear (0.30). Since size of corsetry brief has similar trend to size of womenswear, it also has a weak correlation with size footwear. The rest of features only have negligible association between size of womenswear.

Several rules can be found from the exploratory data analysis as follow. In general, number of customers per size on all brands, have most common measurements between 16 and 20 as well as least common sizes on size 8, 32, and 36. However, the trend on brand 2 looks different because the data is not much as the others. Otherwise, it might be similar to other brands. After age 60, all sizes (womenswear, corsetry briefs, corsetry cup, and footwear) have tendency to become smaller. However, there is a huge volatility after age 90 because the number of the data for that range is limited, therefore it seems unstable.

In terms of average return rate on size per brand, in general, the higher rate ranging from size 12 to 28 for all brands with different peaks of rate. While on brand 1, the highest rate is on size 12, the highest rates on brand 2, 3, and other are 24, 20 and 20 respectively. Regarding the socioeconomic groups which have tendency to higher return rate are prestige positions, suburban stability, and senior security, whereas the lower return groups are municipal challenge, family basics, and transient renters. The top four of socioeconomic groups which possess high return rate (prestige positions, suburban stability, so forth) tend to have lower average size of womenswear than others, whereas the low return rates' group tend to have higher average size womenswear than other groups.

3. Prediction of customer's womenswear size

To ensure the selected model is not overfitting to the training data, cross validation is primarily employed. K-Fold Cross Validation with 10-fold is also usually chosen. Each fold was measured with micro f1 score. The reason is this dataset has imbalanced data. There are 8 labels, however, the distribution of labels is not balanced. To give an example, size 16 has 24.825 customers (31%) whereas size 36 possess only 11 customers (0.014%). Not only size 36 which has low number of data, size 8, 32, 28 also suffer from the limited data. By utilizing micro f1, the evaluation metric focus on the balance of precision and recall. It also considers the small number of classes. If macro f1 is used, it will tend to give high accuracy as dominant classes have a good f1 score. After all, the mean and standard deviation of micro f1 scores were obtained, to give any information regarding stability of the model, as can be seen on a small number of standard deviation.

In terms of performance of the selected model, the mean of micro f1 score on cross validation on the training data was 0.69 with 0.004 standard deviation. To check the result of prediction on the training data, the confusion matrix was used. As can be seen on the code, average precision, recall and f1-score were 0.77. The model tends to predict bigger on size 12, 16. Moreover, the classifier has tendency to predict size 20 to become either size 16 or 24. Also, size 24 and 28 tend to be predicted as smaller size. In terms of size 8, 32 and 36, the model correctly predicts the actual label.

Random forest has been opted to become a classifier for this task. Several machine learning algorithms such as Multi Layer Perceptron, Logistic Regression, Bagging, Gradient Boosting were tried. Random Forest was stable, fast and less overfitting the dataset. The importance of features was measured by Random Forest. As the result, only 9 features were fitted into the classifier. Applying Principal Component Analysis (PCA) and Synthetic Minority Over-sampling Technique (SMOTE) was not able to outperform the current classifier.

4. Implementation of the prediction

Some recommendation regarding implementation of this prediction system are embedding this prediction to the website and periodically check the prediction system as it might need retraining new data. If the customer's size has been identified, she can fill the form to input some details like age, size of corsetry briefs and so on. The form can be embedded near size chart button. Also, if the existing customer wants to check their current size, they can refill the form again. The data might change from time to time, as new data grows every day and it would be beneficial to retrain the model with the new data. Therefore, it would be nice to examine its effectiveness on a certain period.

This model still needs to be optimized as some classes still have some misclassification. Moreover, customer might answer with wrong details of their measurement. The possible reasons are they do not exactly know their cup or briefs size, they measure with a wrong method, their body has been changed (because of a health condition or something else).

The possible risks might come on this model are several classes which still misclassified far away from their actual sizes and sometimes different brands or material of clothes can be different in measurement.