Adam Fletcher: N Brown Data Science Recruitment Exercise Submission

1. Examination of Factors Influencing Womenswear Size

The dataset contains a large amount of high quality data detailing information from multiple sources such as customer sizing data, buying habits and demographic data. The dataset is complete with no missing values. To aid with the prediction of womenswear size, customer sizing variables were assessed for their correlation to womenswear size to form the basis of a predictive model, this data was chosen over information such as demographic data to avoid overfitting of the model.

2. Distribution and Correlation of Sizing Data Against Womenswear Size

Initially, the distribution of womenswear sizes was assessed over the training dataset to assess the normality of the distribution (Figure 1). The data is normally distributed with a slight positive skew showing the most common size to be 16. The distribution of the other sizing metrics; corsetry briefs size, corsetry cup size, footware size and age were assessed in addition to the correlation to womenswear size (Figure 2). The global distribution of coresetry briefs (Figure 2A, black line) is relatively normal with a slight positive skew, of interest is the correlation to womenswear size (Figure 2A, coloured distribution), where the distribution mode for each womenswear size falls as a specific corsetry briefs size. However, there are two shoulders around this distribution where womenswear size is above or below the mode. Additionally, there is a large amount of overlap between the distributions, which increases the risk of misclassification within the model utilsing only this variable. The data is further limited by relatively few customers with a corsetry brief size of over 24 and a high proportion of size 12 womenswear customers whose corsetry briefs size is higher than their womenswear size. Specifically, 29.3% of womenswear size 12 customers have a corsetry briefs size of 16, compared with 19.4% womenswear size 16 customers with a cosetry brief size of 20. This reduces the predictive power of the model for these womenswear sizes. The Spearmans Rho of womenswear against corsetry briefs size was 0.817 (p < 2.2x10^{-16}) highlighting a very strong positive correlation between corsetry briefs size and womenswear size.

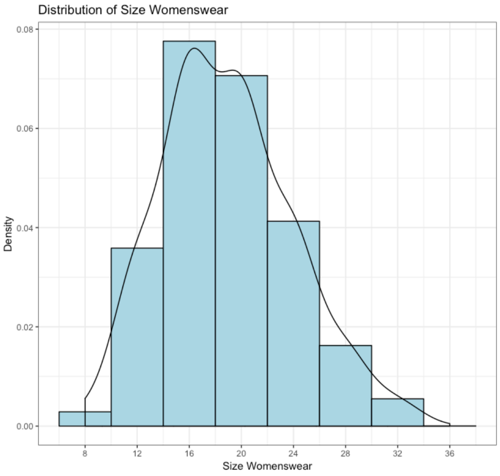


Figure 1: The Global Distribution of Womenswear Size across the Training Dataset.

The global distribution of corsetry cup size (Figure 2B, black line) shows a heavy weighting towards a cup size of 4 with a slight negative skew. The distribution of corsetry cup size doesn't change with womenswear size (Figure 2B, coloured distributions), and this lack of correlation is corroborated by the Spearman's Rho of 0.135 (p < 2.2x10^{-16}) indicating a weak/no correlation. The global distribution of footware size is regular (Figure 2C black line) with a weak positive correlation between footware size and womenswear size (Spearman's rho = 0.322). However, it can be seen that women with womenswear size above 20 tend to have larger footware sizes (7 and above) than women under size 20, this is caveated as the sample size of customers with womenswear size above 24 is small (8.9% of population).

The age of the customers is normally distributed with the most frequent age of the customer being 56, there are however clusters of customers in their late 30's and ealy 70's, potentially representing a targetted product-line or advertising campaign. Looking across all the womenswear sizes there is no correlation between age and womenswear (Spearman's rho = -0.054).

In conclusion, coresetry brief size has the highest correlation to womenswear size and thefore may have the highest predictive potential, with footware size providing an additional but weaker predictive factor.

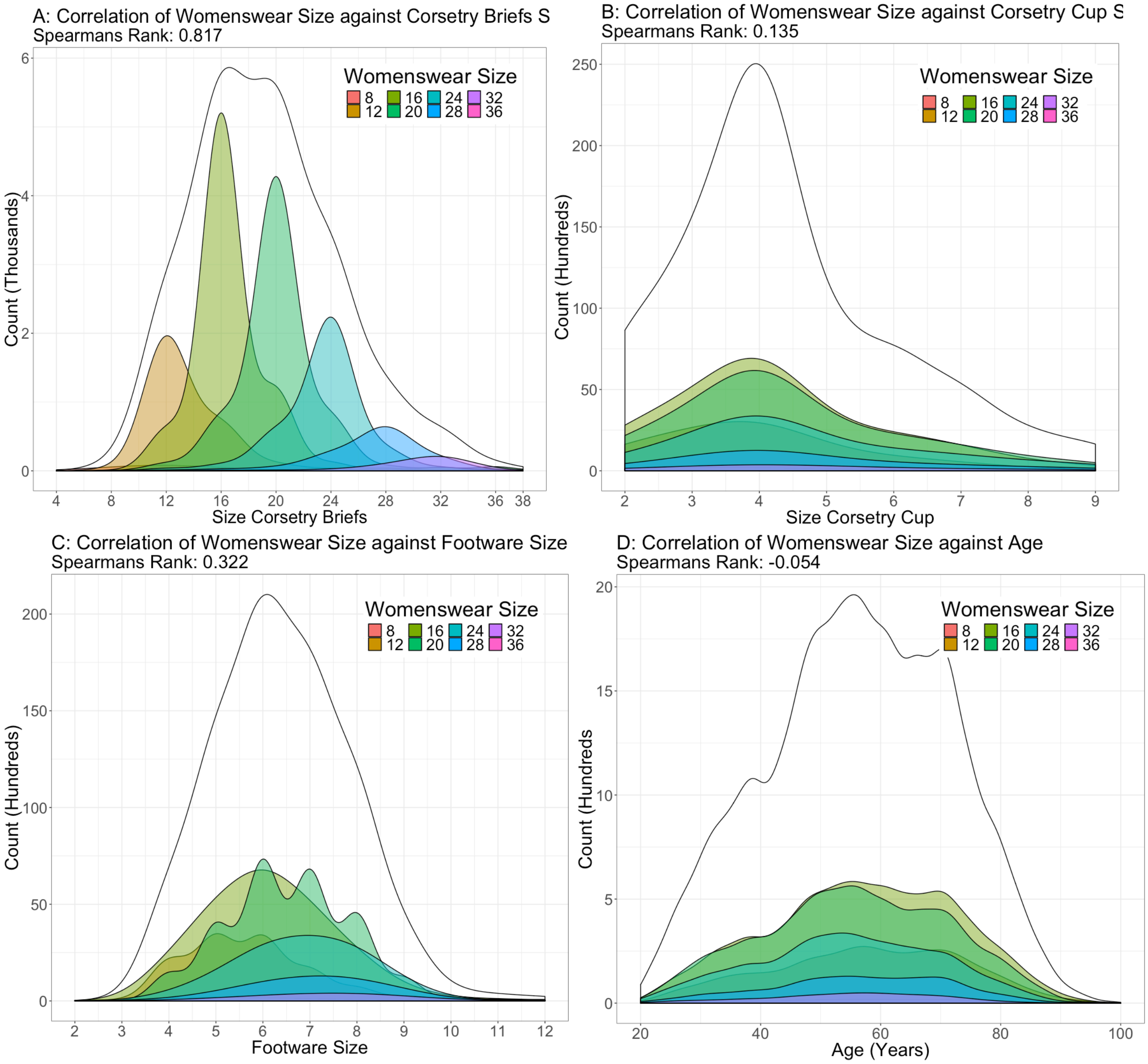


Figure 2: Distribution of Sizing Metrics and Their Correlation to Womenswear Size

3. Correlation of Corsetry Briefs Size and Footware Size Against Womenswear Size

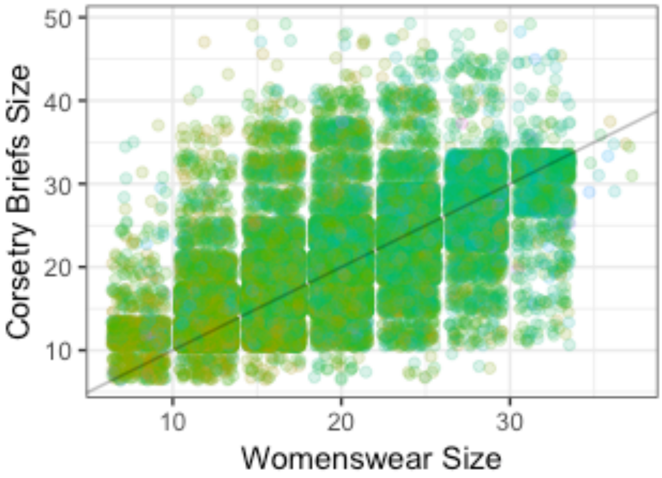


Figure 3: Correlation of Womenswear size against Corsetry Briefs Size and Footware Size

As corsetry briefs size and footware size had the highest correlation to womenswear size, the dual correlation between these factors was assessed (Figure 3). Both individually correlated variables can be viewed independently on the figure with the majority of the data following the womenswear size equals corsetry brief size correlation already established (Figure 2A) additionally the depth of green in the data points increases as both brief size and womenswear size.

In conclusion corsetry briefs size and footware size are co-correlated, and combining both these factors could result in a predictive model with higher accuracy than individually.

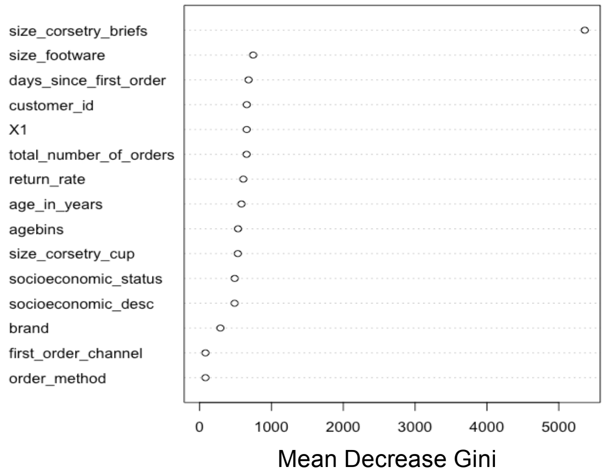
4. Testing an Initial Random Forest Model to Predict Womenswear Size

In order to train and validate a model to predict womenswear size, the provided training data was further split into a random forest training model (containing 70% of the data) and a random forest validation model (containing the remaining 30% of the data). The hyperparameters of the random forest model were tuned such as; the number of variables selected in each decision tree (mtry) and the number of trees to be used (ntree).

An initial random forest model was generated using all provided variables. As predicted the most important variable to the random forest model was corsetry briefs size, then footware (Figure 4A), the out of bag (OOB) for this model is 52.4% resulting in the model being 56.4% accurate at predicting womenswear size against the random forest training dataset and an accuracy of 47.3% against the random forest validation dataset. An important observation is that customer ID is the 4th most important variable in the model. This is a random variable, indicating that this variable and everything below it are not beneficial to the model. Additionally, including days since first order will likely lead to overfitting the model although it is deemed important, a hypothesis supported by a Spearmans Rho correlation of 0.05.

Therefore the model will be repeated by only including corsetry briefs size and footware size. The confusion matrix of the initial model (Figure 4B) further highlights that 99.7% of the dataset was predicted to be either size 16 or 20, the two most common womenswear sizes. The predictive power of this model is lacking at other sizes due to overfitting the model.

A: Importance of Each Variable to the Random Forest Model



B: Confusion Matrix of Initial Model

		Actual Size							
		8	12	16	20	24	28	32	36
Predicted Size	8	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	0	0
	16	258	3276	6196	1666	935	394	121	1
	20	18	166	1249	5105	2983	1160	405	2
	24	0	1	2	10	45	5	1	0
	28	0	0	0	0	0	0	0	0
	32	0	0	0	0	0	0	0	0
	36	0	0	0	0	0	0	0	0
Model Accuracy		0	0	83.2	75.3	1.1	0	0	0

Figure 4: Metrics and Outputs From the Initial Random Forest Model

5. Training a Model to Predict Womenswear Size from Corsetry Briefs and Footware Size

		Actual Size							
		8	12	16	20	24	28	32	36
Predicted Size	8	102	46	12	12	5	1	0	0
	12	429	5583	1384	289	105	36	19	0
	16	63	1938	12605	1971	241	81	22	0
	20	28	28889	2855	11255	1625	182	55	4
	24	14	105	312	2017	6304	757	59	0
	28	6	36	93	165	835	2096	263	1
	32	2	37	117	114	133	484	813	3
	36	0	0	0	0	0	0	0	0
Model Accuracy		15.8	69.5	72.5	71.1	68.2	57.7	66.0	0
Current Accuracy		73.7	70.4	70.1	70.1	70.9	72.7	74.0	84.9

Figure 5: Confusion Matrix of the Random Forest Training Dataset.

The initial random forest model had accuracy against the validation dataset of 47.3%. The model was then retrained using only the corsetry briefs and footware size, which results in a model with an OOB of 30.9%. The accuracy of the improved model against the random forest training dataset is 69.21% with an accuracy of 69.47% against the validation dataset. The confusion matrix of the random forest training dataset (Figure 5) and validation dataset (Figure 6) shows that for sizes 12 through 32 the majority of the

		Actual Size							
		8	12	16	20	24	28	32	36
Predicted Size	8	32	27	9	0	1	0	0	0
	12	202	2459	645	117	44	21	5	0
	16	19	759	5364	819	111	33	15	0
	20	13	131	1217	4882	686	72	28	1
	24	6	34	114	819	2731	333	28	0
	28	1	19	36	91	332	870	118	1
	32	1	14	62	53	58	230	333	1
	36	0	0	0	0	0	0	0	0
Model Accuracy		12.3	71.4	72.0	72.0	69.0	55.9	63.2	0
Current Accuracy		73.7	70.4	70.1	70.1	70.9	72.7	74.0	84.9

Figure 6: Confusion Matrix of the Random Forest Validation Dataset.

samples are correctly predicted. In comparison to the initial model the new model is less accurate at predicting womenswear sizes of 16 and 20, however, the new model is more accurate at predicting other sizes. Of note is that 0% of customers were predicted to be size 36, this is likely to be due to the lack of data at this size. This is also observed with customers with a womenswear size 8, where the lack of measurements incorrectly categorises them as size 12 more often that it correctly categorises them. This model was then used to predict the womenswear size of the provided test dataset (uc_data_test_AFletcher.csv) as column 'predicted_size_womenswear'.

6. Assessing the Accuracy of the Model

The accuracy of the trained model differs with each womenswear size (Figures 5 and 6: Model Accuracy). The companies current accuracy of delivery is calculated as the total number of returns for each womenswear size as a percentage (Figure 5 and 6: Current Accuracy) it can be seen that for the most common sizes (sizes 12 to 20) the model can more accurately predict these dress sizes, leading to a lower return rate for these dress sizes. At less common dress sizes, the model is less predictive. This is indicative that the model although useful for the more common womenswear sizes, requires more information be able to more accurately predict the wide variety of womenswear sizes.

7. Further Metrics for Predicting Womenswear Size and Deploying the Model

In order to generate future models with improved predictive power, extra variables that display higher correlation for womenswear size need to be collected. I hypothesise that additional variables included should be waist and chest measurements. As corsetry briefs size was highly predictive, this measures the underwear size and therefore indicative of hip size, collecting waist measurements could further refine the model by accounting for the customer's body shape.

The model might also be improved by collecting chest measurements. Although corsetry cup size was collected in the initial dataset, the majority of women had a cup size of 4 severely reducing its correlation with womenswear size. However, chest size might be a more accurate measurement. I hypothesise that combining; waist measurement, chest measurements and corsetry size (or hip measurement) will further improve the model for predicting womenswear size.

Collection of this new data and deploying the model for online purchases could be conducted by providing an online dress-sizing tool. This online tool would request the previously requested metrics to feed into the established model for predicting womenswear size. In addition, the new metrics would be requested which will help build a new database to form the basis of a new training dataset for subsequent releases of the dress-sizing tool. For in-store purchases, the placement of secure tablets which have the tool displayed will allow women to predict their dress size before going to a fitting room, in this case an additional metric of whether the prediction was accurate could be asked providing a metric for validating the model.