SupMat_3: Random Forest and Model Validation

Random Forest model

A single tree is likely to overfit the data and growing a forest based on resampling mitigates against this risk. On our data, a random forest of 500 trees reaches a very similar correct prediction rate of 94.57%. These results are outstanding, with only 10.8% of all zero instances (42/389), 5.2% of all *the* instances (41/795) and 2.2% of all *a* instances incorrectly classified (13/584) (Table 3.1).

Table (3.1): confusion matrix for article data based on a forest

	zero	**a**	**the**	**Total**
zero	**347**	14	28	
a	11	**571**	2	
the	37	4	**754**	
Total	395	589	784	**1768**

Again, each of the three articles is most frequently predicted as itself. Zero remains the article which is most often mispredicted (10.8%), followed by *the* (5.2%) and *a* (2.2%). In this forest, *a* and *the* are typically mispredicted as zero, while zero is most often mispredicted as *the*.

On the basis of the forest, the importance of each variable can be calculated. Using random permutation of the labels of each variable, the relative importance of the different predictors for the classification accuracy of the model is assessed. In addition to a decrease in accuracy, we can track the mean Decrease in Gini values, a forest-wide weighted average of the decrease in the Gini Impurity metric between the parent and daughter nodes that a variable is splitting. It can be defined as the total decrease in node impurity (weighted by the proportion of samples reaching a given node) averaged across all of the trees that make up the forest. A higher Mean Decrease in Gini indicates higher variable importance. Variables are sorted and displayed in the Variable Importance Plot created for the Random Forest by this measure. The most important variables to the model will be highest in the plot and have the largest Mean Decrease in Gini Values. Conversely, the least important variable will be lowest in the plot, and have the smallest Mean Decrease in Gini values.

The Gini plot in the right panel of Figure (3.1) reveals that Hearer Knowledge is the strongest predictor (Mean Decrease = 436), followed at a large distance by Number (Mean Decrease = 159), Referent Specificity (Mean Decrease = 122) and Count (Mean Decrease = 89), which are in turn followed at a large distance by Elaboration (Mean Decrease = 14) and Corpus (Mean Decrease = 9). Corpus and Elaboration are “fine-tuning variables” (Divjak 2015) and taken on their own do not contribute much to the correct classification of article use.

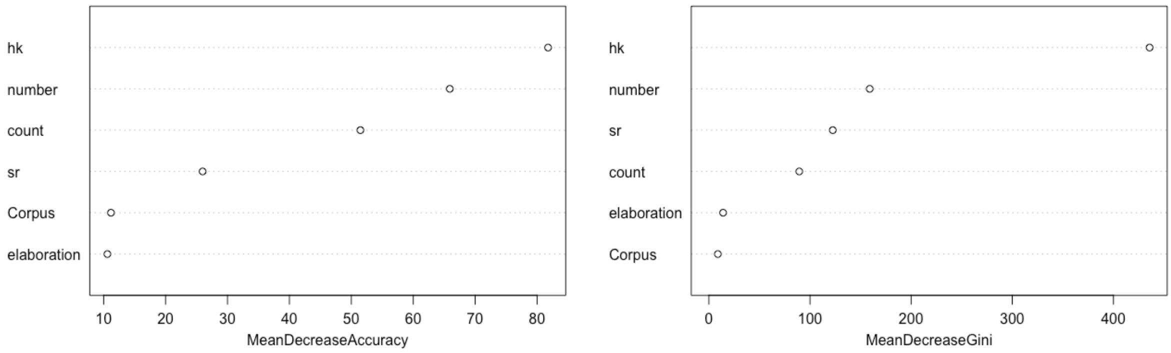

Figure (3.1): Mean Decrease in Gini (right panel)

Model validation

We validated the forest by training the same model on 70% of the data and testing it on 30% of the data. The correct prediction rate on the training sample was 94.99%, while the accuracy on the testing data was 94.2%; both values are marginally lower than the accuracy obtained on the full sample. Table (3.2) compares the percentages of predictions on the 70% of the data that served as training data (1237 datapoints) with predictions for the 30% of the data that served as testing data (531 datapoints)

Table (3.2): comparison of predictions on training and test samples (rounded)

TRAIN	zero	a	the	**TEST**	zero	a	the
zero	0.892	0.024	0.083		0.860	0.035	0.105
a	0.027	0.966	0.007		0.016	0.978	0.005
the	0.018	0.002	0.980		0.043	0.004	0.953

The variable importance plots for the training data, presented in Figure (3.2) shows that the relative variable importance is identical to that for the full dataset, with Hearer Knowledge coming out strongest, followed by Number, Count and Referent Specificity, and finally Elaboration and Corpus. Referent Specificity groups differently depending on measure used (Accuracy vs Gini). This confirms that we can be confident in our model.

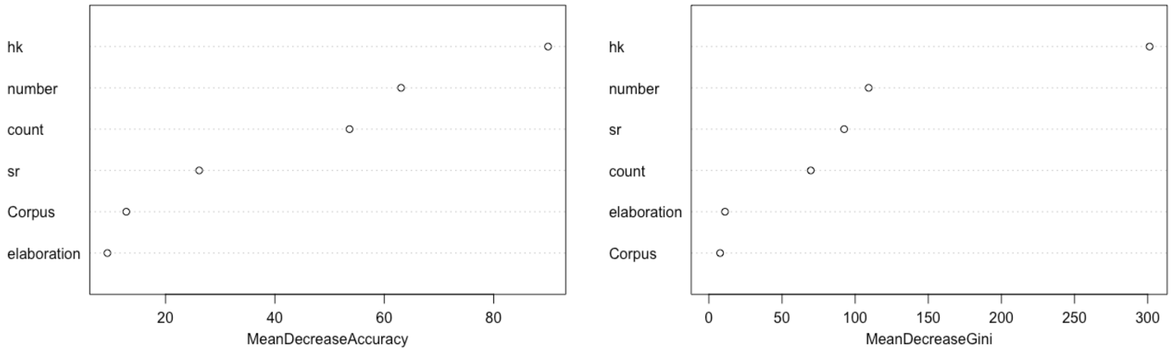

Figure (3.2): Mean Decrease in Gini for the training sample