Seminar research methodology for data science
CS4125

Delft University of technology

COURSEWORK B
-STATISTICAL TESTING-

Nathan Buskulic, 4947916, n.m.m.buskulic@student.tudelft.nl
Mitchell Deen, 4396340, M.R.Deen@student.tudelft.nl
Þórunn Arna Ómarsdóttir, 4917499, omarsdottir@student.tudelft.nl
Team 5

April 3rd, 2019

1 RQ1: Improvement of transfer learning

The first question we want to answer is : "How much does transfer learning improve over typical non-transfer learning?" To answer this question we try to figure out whether transfer learning from multiple training datasets achieves better performance than simple models trained for the specific target task without transferring.

First it is good to take a look at the data to see how the distribution of data points is. Figures 1 and 2 the distribution for the baselines combined (B) and for the transfer learning models combined(M). By looking at them we can see that both groups have a very similar distribution. But since we have much more samples for the transfer learning that distribution is more detailed. The B subset looks like it is a mixture of two Gaussian distributions that have a peak around 0.1 and 0.6. but the M subset seems to have an extra Gaussian included in between the other two. By just looking at the means for the classes we get the idea that the groups yield different results, since the mean for B is 0,378 and the mean for M is 0,5048. The standard deviation for both groups is fairly similar, B: 0.2328 and M:0.2606, so the means are both as reliable. Moreover it is important to keep in mind the size difference of B versus M, B is 21 data points while the size of M is 2639. So M has 125 times as many data points. These numbers raise some questions on the validity of the comparisons.

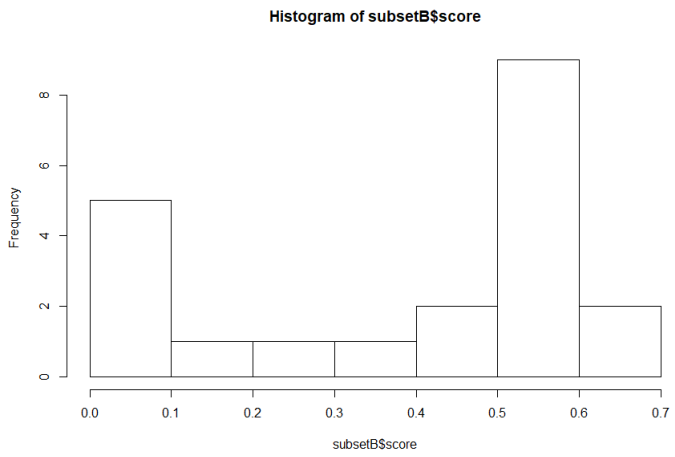


Figure 1: Histogram showing the distribution of data points for the baselines combined.

To understand better what the data tells us and make sure we do not miss characteristics hidden in a single model, we calculated the means and standard deviation for each of the models. The results from that can be found in table 1 below. We can see that the means are overall higher for the transfer learning models, but that is not always the case, since the mean for B1 is higher than the mean for MF. It is also interesting to see standard deviation is rather similar for all models, since there are only 7 scores for each of the B models but 490-560 scores for each of the M models. Since the data sample is so small for the B models the data obtained from then is rather small to be reliable as a

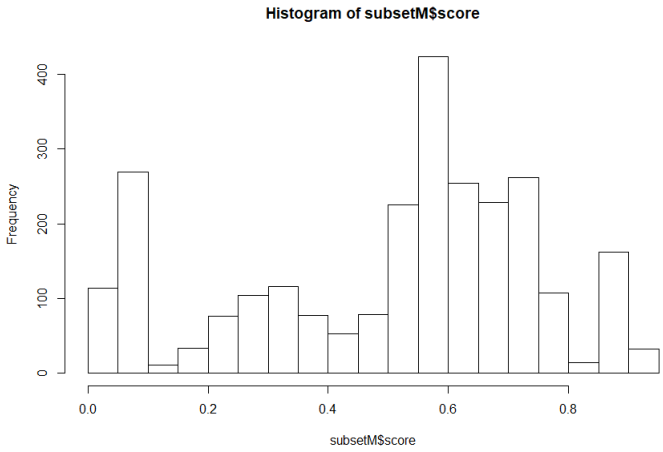


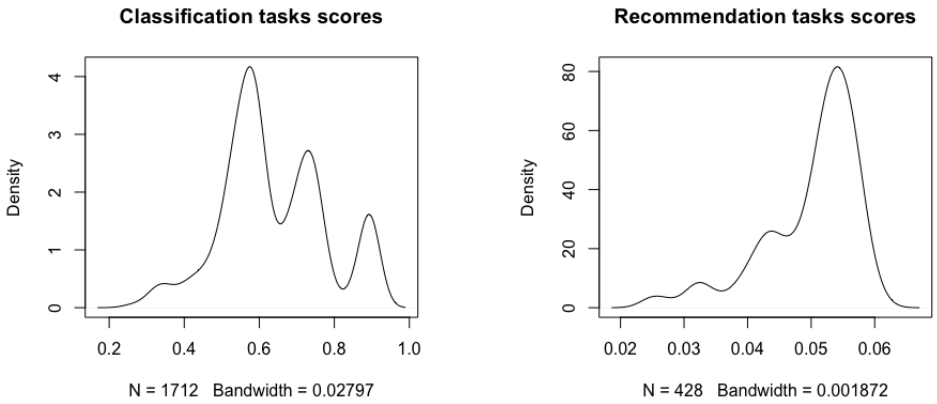
Figure 2: Histogram showing the distribution of data points for the transfer learning models combined.

good representation of the baselines.

Table 1: Mean and standard deviation for each B and M model

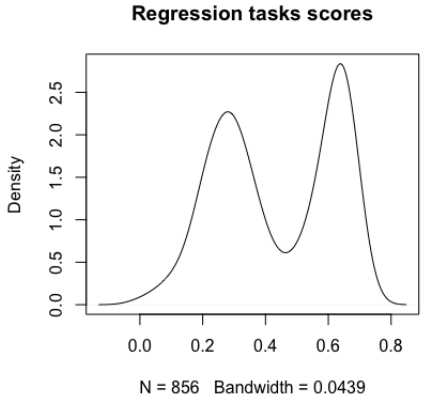
	B1	B2	B3	MN	M1	M2	M3	MF
Count	7	7	7	490	546	539	504	560
Mean	0.495	0.328	0.353	0.531	0.509	0.515	0.520	0.455
Std. dev.	0.248	0.265	0.220	0.247	0.250	0.247	0.252	0.244

The histograms of Figure 1 and 2 clearly shows two peaks. This incited us to explore what could be the main effect for the creation of these peaks. In order to further our analysis we fitted a linear model on the data by adding every parameters to the model (test datasets, training datasets, models) and check the result of an ANOVA on this model. We found that the model had a R-squared value of 0.9487 which means that 94% of the variance of the score can be explained by the parameters. We also saw that the test datasets seems to have the biggest importance. Therefore we fitted a new model with only the test datasets as parameters. This new model yielded a R-squared value of 0.9159. This mean that by only knowing the testing dataset, 91% of the variance in the model can be explained, or in other words, knowing the model and the training datasets can only help to explain 3% of the variance. Thus, our analysis for the different questions should take this important information into account.
Our next step was to try to explain the two peaks in the score density by the difference in tasks in the dataset. We thus plotted the density of the scores for each task and see if there is clear groups. This results can be found in Figure 3. It appears that there is not one clear gaussian per group and that each test dataset yield different distributions of results. This was to be expected since that even if the tasks were the same, the datasets changed and thus, the distribution shift as well. It however good to notice the difference



(a) Density of the score for the classification test datasets

(b) Density of the score for the recommendation test dataset



(c) Density of the score for the regression test datasets

Figure 3: Score density for each group of test datasets tasks

in term of general trends between groups. For example the recommendation task produce very low results (<0.1) while the two other groups are more spread between 0 and 1.

We also wanted to see what values each training set yields in general to see if there is something going on there that needs to be inspected further. No training sets are used to train the baselines so we cannot compare the results between the baselines and transfer learning models. But in table 2 we have the overall mean for all models in the experiment, grouped by training sets. They are all quite similar, the maximum variation between any pair is 0.02.

It is also curious to see if the outcome per training differs a lot for each test set. In figure 4 each line represents one training set. We see the score for each training set in

Table 2: Mean of all results for per training set

	TrD1	TrD2	TrD3	TrD4	TrD5	TrD6	TrD7	TrD8
Mean over all M models	0.498	0.516	0.509	0.509	0.515	0.510	0.516	0.518

relation to each test set set. We can see that for each test set all of the training set yield very similar results. Therefore we do not consider it necessary to look into that further.

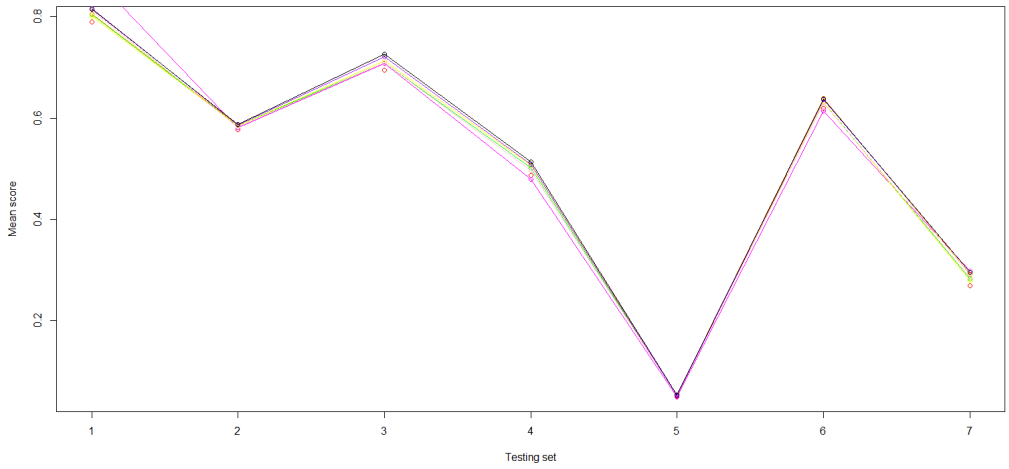


Figure 4: Lines show the scores for one training set, the x axis represents the different test sets

We decided to apply two versions of t-tests on the data to understand the difference between the two model categories better, that is between the baselines and the transfer learning models.

1.1 Global mean t-test

We first wanted to see if there were a significant difference in the global mean between the baselines and the Mx models. In order to do that we combined the mean of all baselines together and we did the same for the Mx models. The t-test produces a p value of 0.02478. This value would indicate that there is a difference between the mean score of the baselines and the Mx models. However since each system heavily depends on the testing dataset, it would be more interesting to compare the means that we got for each dataset.

1.2 Paired t-test on means

We calculated the mean of the results for each testing dataset for the baselines(B1,B2,B3) combined and for the transfer learning models combined(MN,M1,M2,M3,MF). In figure 5

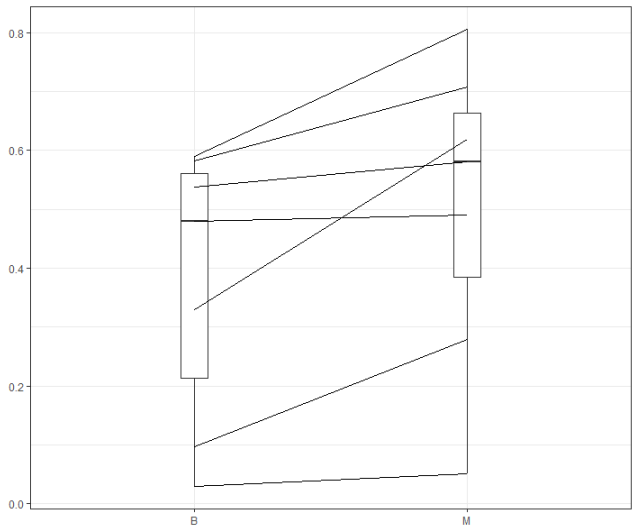


Figure 5: Plot that shows the connection between each pair

we can see each mean pair connected with a line. We can also see a clear interaction effect, that the mean is always at least slightly higher for M than B. But since the data points what we have for the baselines are very few it is good to check whether the differences of the pairs follow normal distribution. To do that we perform a Shapiro normality test on the difference of each pair which gives us a p-value of 0.466 which implies that the differences are not significantly different from normal distribution, so we can assume normality.

Since the slopes are so different in Figure 5, we decided to explore more the effect of the test dataset on the effect of the difference between the baselines and the M models. In Table 3 are presented explicitly the means of the baselines and the M models for each test dataset. We can see that for TeD2, TeD4 and TeD5 the difference is very small while on the other TeD, the M models achieve a big leap in performance compare to the baselines.

Table 3: Mean of the baselines and the M Models for each dataset

	TeD1	TeD2	TeD3	TeD4	TeD5	TeD6	TeD7
Baselines Mean	0.590	0.538	0.583	0.479	0.030	0.330	0.096
Models Mean	0.806	0.581	0.708	0.490	0.051	0.619	0.279

However, by applying a paired t-test, where the pairs are the two measures per testing dataset, we get the p-value of *0.02044*. This value also indicates that there is a significant difference between the mean score of the baselines and the transfer learning models.

When looking at the results from both t-test we see that the results are very similar but not the same, but both show that the difference of methodology has significant effect

on the results. We therefore consider that we have managed to show that the transfer learning models achieve better performance than the simple models.

2 RQ2: Effect of learning strategy

In this section we will study the effect of different strategies to simultaneously learn one model from multiple TrD's. The experiments have resulted in five different models that use multiple TrD's: MN, M1, M2, M3 and MF. We are interested in their performance compared to a simple model where only one TrD is used, as well as in differences between the methods. Additionally, we consider both an overall performance, based on all TeD's combined, and performance on specific test data. As some methods might be more suitable for certain types of test data.

We will start by comparing the overall mean scores of the different groups. The results are shown in Table 4 and Figure 6.

Table 4: Mean and standard deviation of the scores for each model, for all test data.

Model	Count	Mean	Std. dev.
M1	546	0.509	0.250
M2	539	0.515	0.248
M3	504	0.520	0.252
MF	560	0.455	0.244
MN	490	0.531	0.247
S	336	0.429	0.231

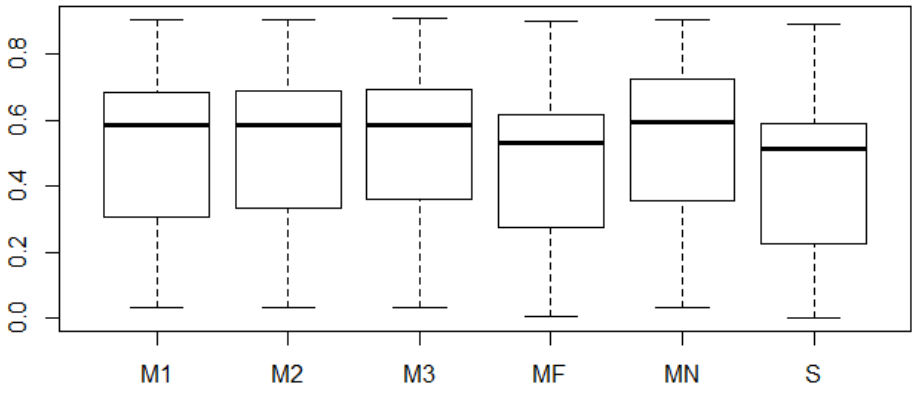


Figure 6: Boxplot of scores for each model, for all test data.

First inspection of the results reveals that, based on all test data, the performance of most of the multiple TrD models might improve over the single TrD model. To check whether there is any statistically significant difference in mean scores between the groups at the 0.05 significance level, we performed an Anova test ($F = 12.05, p < 0.001$) and found that there are indeed differences between the groups. To refine this result, checking for pairwise comparisons using Tukey's HSD statistic confirmed the visual interpretation: M1, M2, M3 and MN score better than S when considering all test data ($p < 0.001$ for all four pairwise tests). The magnitude of this effect is an improvement in score of $0.08-0.10$, which can be considered a meaningful improvement, given that one is interested in scoring well on all different types of test data. The difference between model S and MF is much smaller (0.02) and not significant ($p = 0.65$). These pairwise test are however corrected for multiple testing, so we might not have enough power to find a significant difference at this point. Either way, since the other multiple TrD models have shown much more convincing results, we can conclude that these methods should be preferred over MF, when good performance on all data sets is desired.

Next we would like to point out a few interesting details that came up when analyzing the performance of the models on specific test data. We know that the average score varies greatly per test data set; only drawing conclusions based on the average score over all test data might result in missing some important characteristics of the model's performance.

Figure 7 does however seem to confirm our earlier conclusions. The pattern of M1, M2, M3 and MN outperforming MF and S seems to be consistent over all types of test data. We do not think it is meaningful to go into pairwise comparisons for each type of test data; instead we use this result to restate our belief that the multiple TrD models achieve higher scores than the single TrD model, with added confidence.

3 RQ3: Effect of training datasets

The third and final topic that we would like to investigate is the effect of the TrD's on the model performance. The experiment included eight different sets of training data. Each TrD was used ot train the single TrD model six times, for each TeD. This allows us to compare the mean scores of using each of the TrD's for training (see Table 5 and Figure 8). We will for now focus on the effects of training with individual TrD's rather than with combinations of them.

Table 5: Mean and standard deviation of the scores for each TrD used, for all test data.

TrD	Count	Mean	Std. dev.
1	42	0.356	0.194
2	42	0.387	0.266
3	42	0.401	0.220
4	42	0.396	0.228
5	42	0.489	0.234
6	42	0.442	0.215
7	42	0.471	0.235
8	42	0.488	0.232

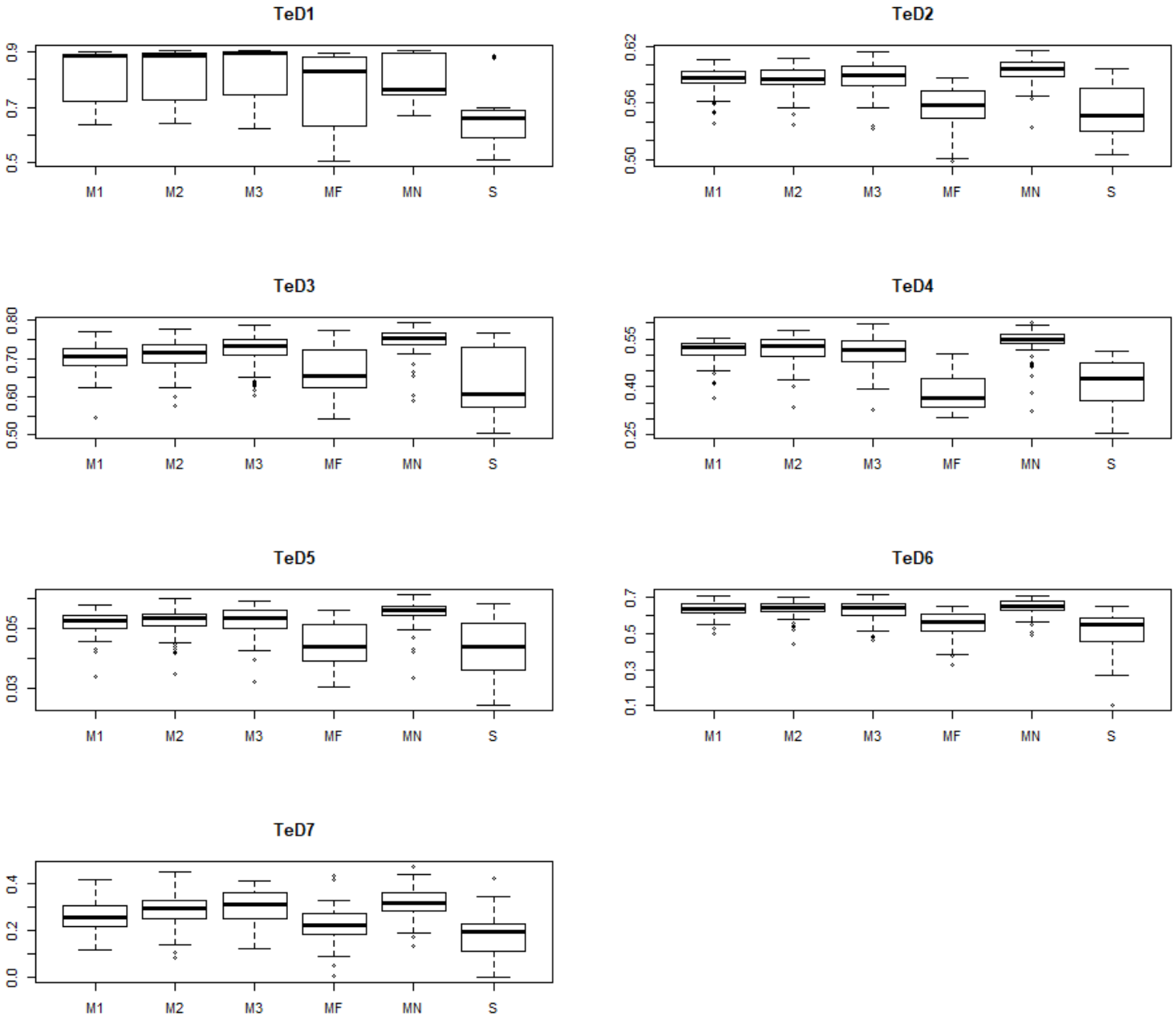


Figure 7: Boxplot of scores for each model for each TeD.

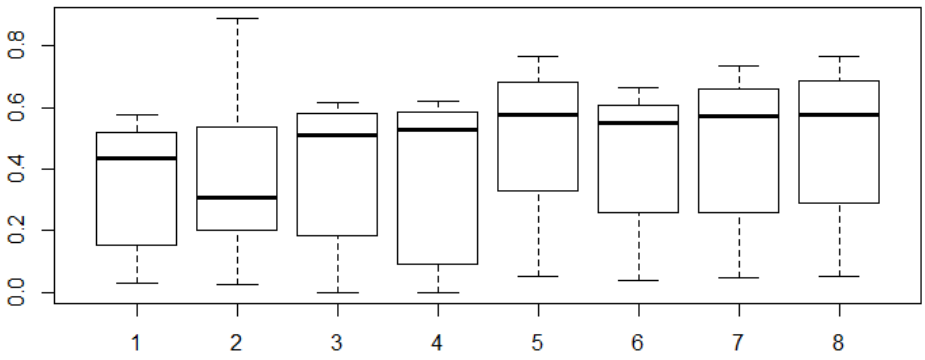


Figure 8: Boxplot of scores for each TrD used, for all test data.

Due to the large difference in scores between the TeD's, we again see a large range of values for each TrD, similar to the comparison of models from Figure 6. The distribution does however vary more here. This makes sense intuitively; we expect the effect of the specific use of a training data set to be more dependent on the test data used than the use of a specific model. We perform another Anova comparison of group means and find it to be significant at the 0.05 level ($F = 2.044, p = 0.049$). Even though the differences between the group means grow as large 0.13 - which is more than we found when comparing models - our test has less power in this case due to a reduced number of data samples. Hence a series of pairwise tests, such as Tukey's HSD, will not be able to distinguish individual differences. We are still interested in differences for individual TeD's. We can for example see that TrD2 is able to achieve very high scores for some TeD; no other TrD manages to reach these values.

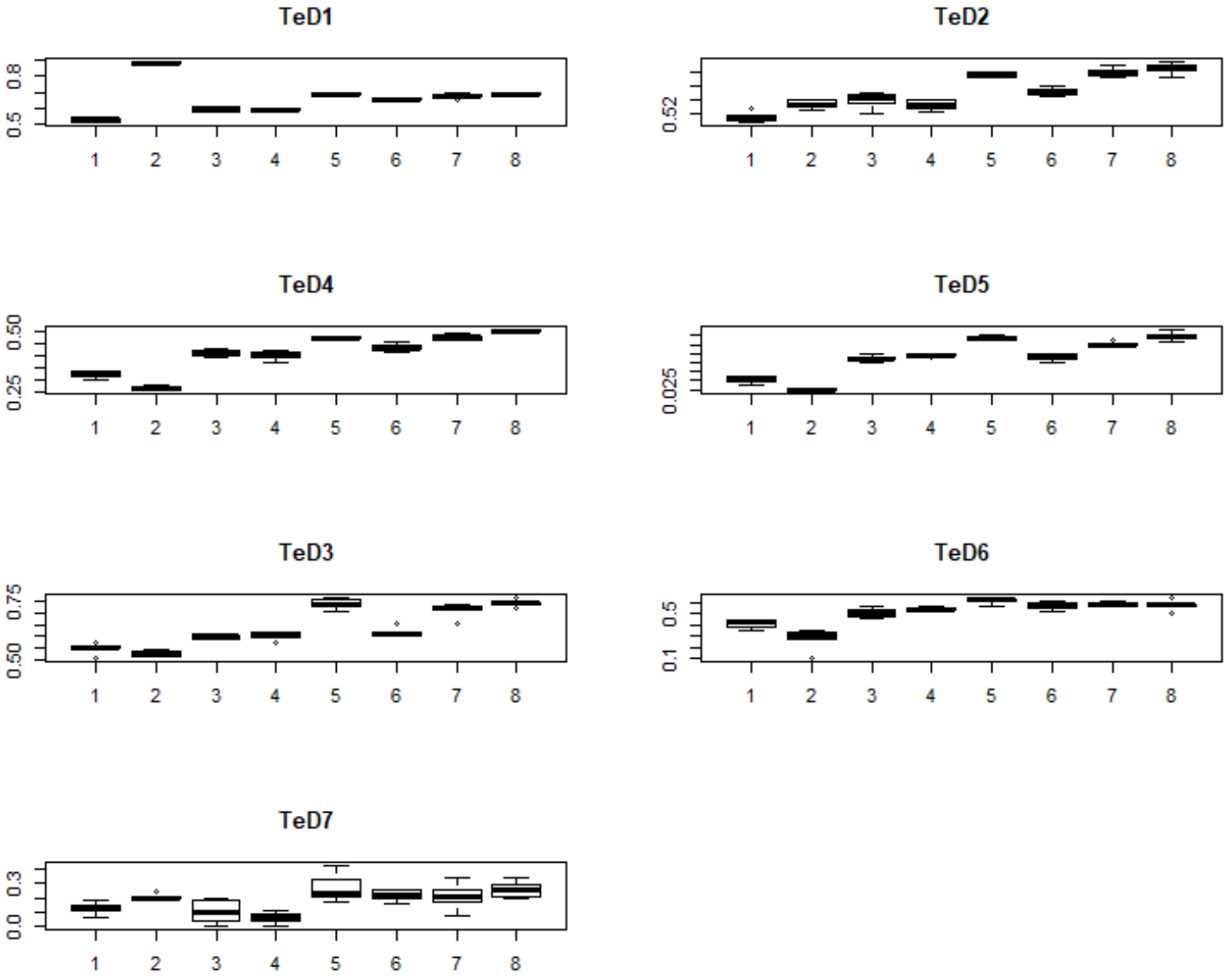


Figure 9: Boxplot of scores for each TrD used and for each TeD.

In Figure 9 we can clearly see a difference in performance in our sample data for each TeD. When retesting for differences between the training sets for each TeD individually, we find significant differences ($p < 0.001$) for all of them, confirming our expectations.

Overall we can conclude that the effect of using different TrD's highly depends on which TeD is used for testing. If we are only interested in performance on all types of data, the differences in performance are not convincing enough to formulate a clear recommendation on which training set to use. When diving deeper into the individual types of test data, we do however find clear indications of differences between data sets.