## Research Question 3

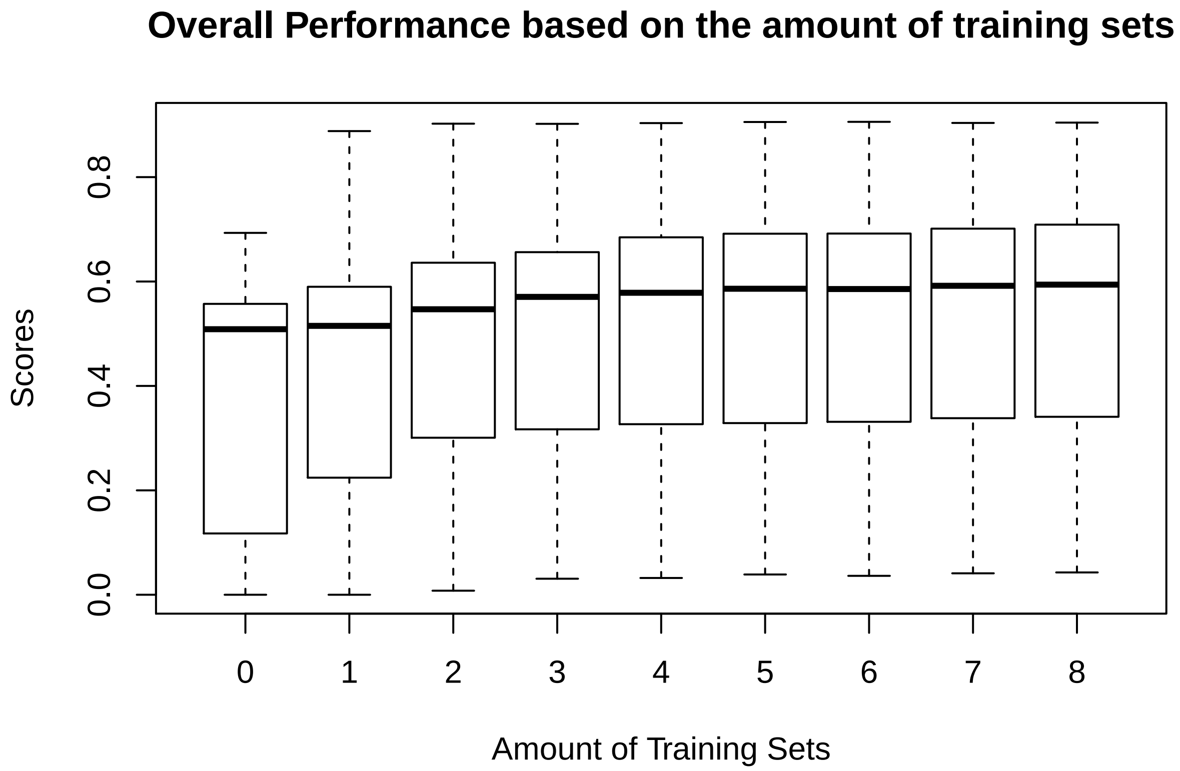
### Observing Data

We perform a simple look at the means of the test score of each combination of TrDs. We can from this look at which combination of training sets achieves the maximum test score.

```
##     TrD1 TrD2 TrD3 TrD4 TrD5 TrD6 TrD7 TrD8     score
## 140    0    1    1    0    1    0    1    1 0.5660503
## 119    0    1    1    0    1    1    0    1 0.5624479
## 101    0    1    1    0    1    0    0    1 0.5618264
## 148    0    1    1    1    1    0    1    1 0.5605164
## 128    1    1    1    0    0    0    1    1 0.5561881
## 131    0    1    0    1    0    0    1    1 0.5524732
## 122    0    1    0    1    1    1    0    1 0.5498406
## 107    1    1    1    1    1    0    0    1 0.5495461
## 72     0    1    0    1    0    1    1    0 0.5490720
## 163    0    0    1    0    1    1    1    1 0.5489510
## 65     1    1    1    1    1    0    1    0 0.5485297
## 56     0    1    0    0    1    0    1    0 0.5464158
## 76     0    1    0    0    1    1    1    0 0.5462193
## 153    1    1    1    0    0    1    1    1 0.5454987
## 116    0    1    0    0    1    1    0    1 0.5453343
## 151    0    1    0    0    0    1    1    1 0.5450413
## 141    1    1    1    0    1    0    1    1 0.5435354
## 99     0    1    0    0    1    0    0    1 0.5427647
## 106    0    1    1    1    1    0    0    1 0.5415554
## 71     1    1    1    0    0    1    1    0 0.5411909
## 61     0    1    0    1    1    0    1    0 0.5406027
## 142    0    0    0    1    1    0    1    1 0.5401571
## 136    0    1    0    0    1    0    1    1 0.5397787
## 172    0    1    1    1    1    1    1    1 0.5397223
## 127    0    1    0    0    0    0    1    1 0.5394829
```

From this we can see that most of the combinations with the highest performance scores contain the Training Sets TrD 2 and 8. Further we can see that the top 3 combinations also contain TrD 3 and 5 potentially making these training sets also relevant. However this can also be a result of simply the use of more training sets improving the score.

Thus we want to analyse the effect of the amount of training sets used on the performance.

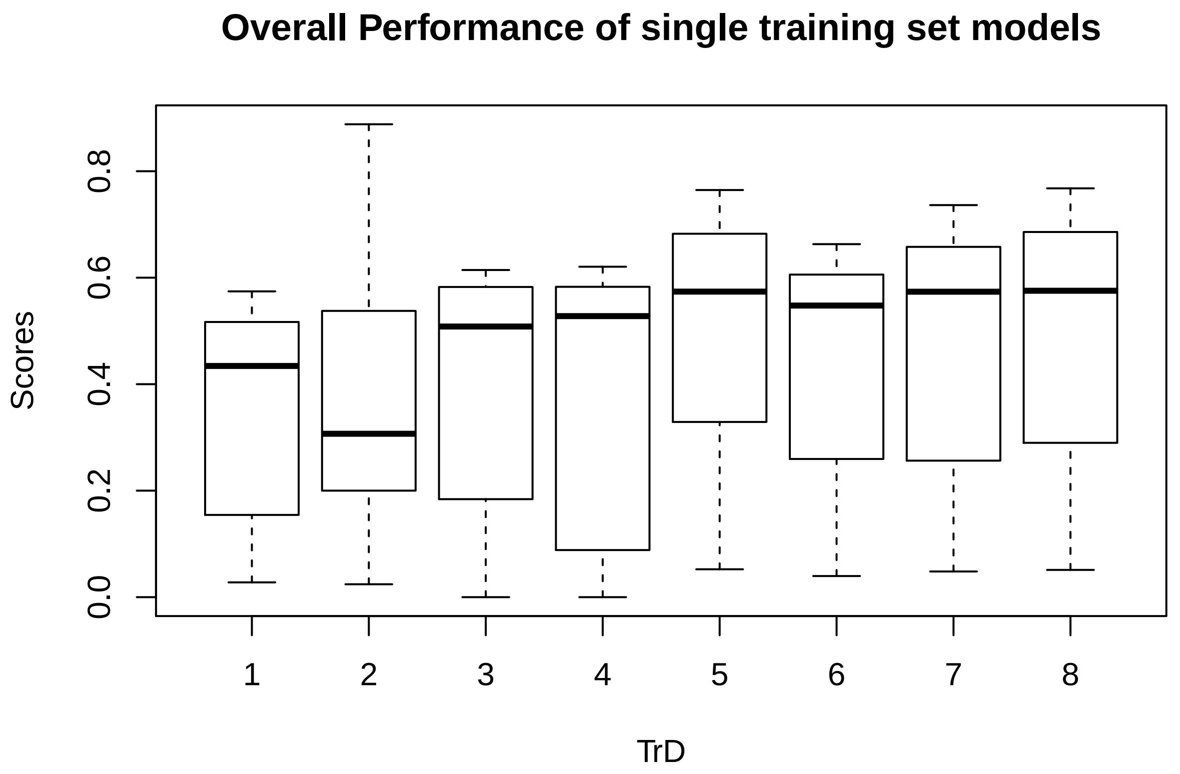


```
##
## Call:
## lm(formula = score ~ TrainCount, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50363 -0.18329  0.06978  0.16911  0.43422
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.442061   0.009987  44.264  < 2e-16 ***
## TrainCount  0.013036   0.002175   5.995 2.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2473 on 2994 degrees of freedom
## Multiple R-squared:  0.01186,    Adjusted R-squared:  0.01153
## F-statistic: 35.93 on 1 and 2994 DF,  p-value: 2.285e-09
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.422478910 0.46164286
## TrainCount  0.008771912 0.01729971
```

Looking at the boxplot we can see that the amount of training sets used in general indeed has an effect on the score, although this effect seems to be minimal after 3 training sets used. Additionally also the linear model shows a sufficiently low p value, thus we can reject the null-hypothesis, meaning that there is a relevant relation between the score and the amount of training sets used. In this case the relation is one where the score increases with an increase in training sets.

We can also observe the single Training data set models and see how they perform



From this analysis we can see that for the simple models, those trained on either TrD 5,7,8 have the best average general performance for all testdatasets. Important to note is that the scores when using TrD 2 show a observably larger variance when compared to the other sets.

### Linear models

Now we make a linear model with the score and all the training data sets.

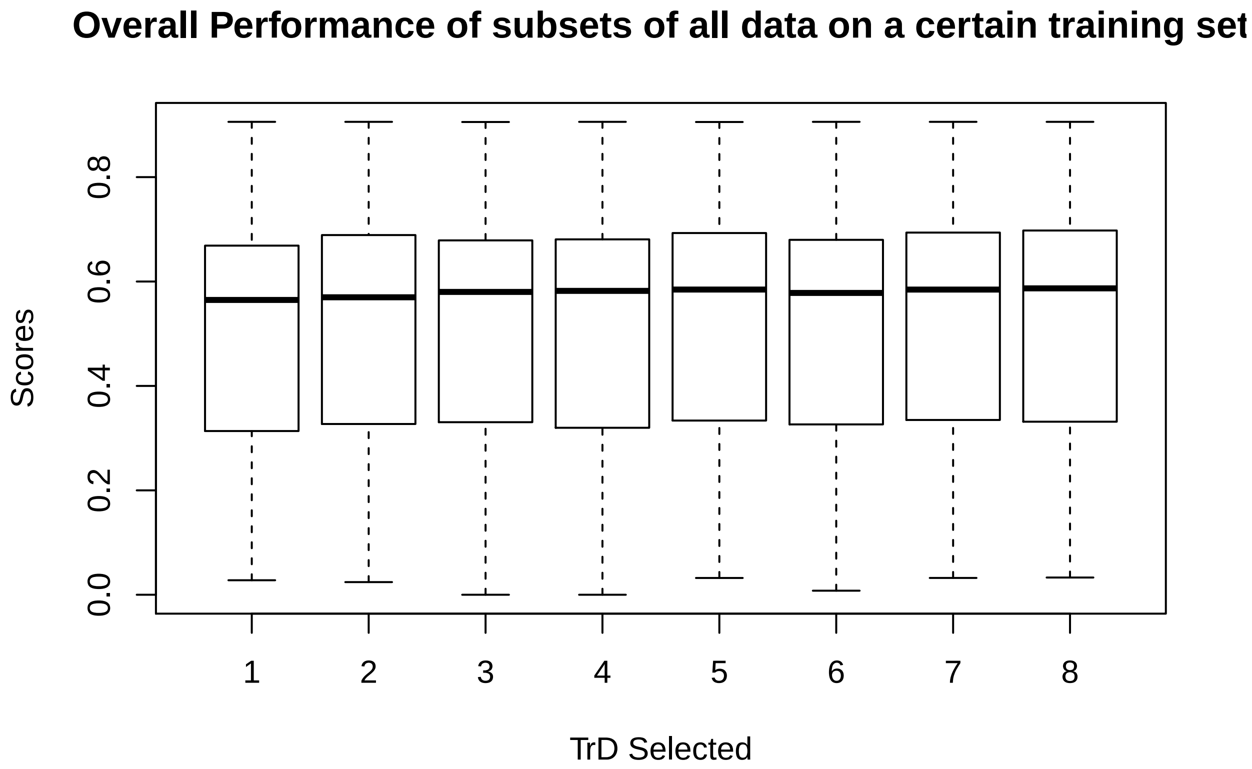
```
##
## Call:
## lm(formula = score ~ TrD1 + TrD2 + TrD3 + TrD4 + TrD5 + TrD6 +
##     TrD7 + TrD8, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51582 -0.18532  0.06918  0.17583  0.43854
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.439589   0.010191  43.133  < 2e-16 ***
## TrD1        -0.020523   0.009717  -2.112  0.03476 *
## TrD2         0.029688   0.009178   3.235  0.00123 **
## TrD3         0.010264   0.009834   1.044  0.29668
## TrD4         0.004862   0.010072   0.483  0.62929
## TrD5         0.023800   0.010137   2.348  0.01895 *
## TrD6         0.009070   0.009427   0.962  0.33604
## TrD7         0.021698   0.009867   2.199  0.02796 *
## TrD8         0.029825   0.009674   3.083  0.00207 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2466 on 2987 degrees of freedom
## Multiple R-squared:  0.0193, Adjusted R-squared:  0.01667
## F-statistic: 7.348 on 8 and 2987 DF,  p-value: 1.018e-09
```

```
##                   2.5 %       97.5 %
## (Intercept)  0.419606599  0.459572237
## TrD1        -0.039576313 -0.001470396
## TrD2         0.011693416  0.047683484
## TrD3        -0.009017722  0.029545928
## TrD4        -0.014885661  0.024610222
## TrD5         0.003923916  0.043676772
## TrD6        -0.009413234  0.027553212
## TrD7         0.002350257  0.041045269
## TrD8         0.010857953  0.048792964
```

From this we can see that the addition of the training data sets TrD 3,4,6 is generally not of relevance for the performance score. Interestingly 4 and 6 also don't appear in the combination of sets with the maximum average test score.

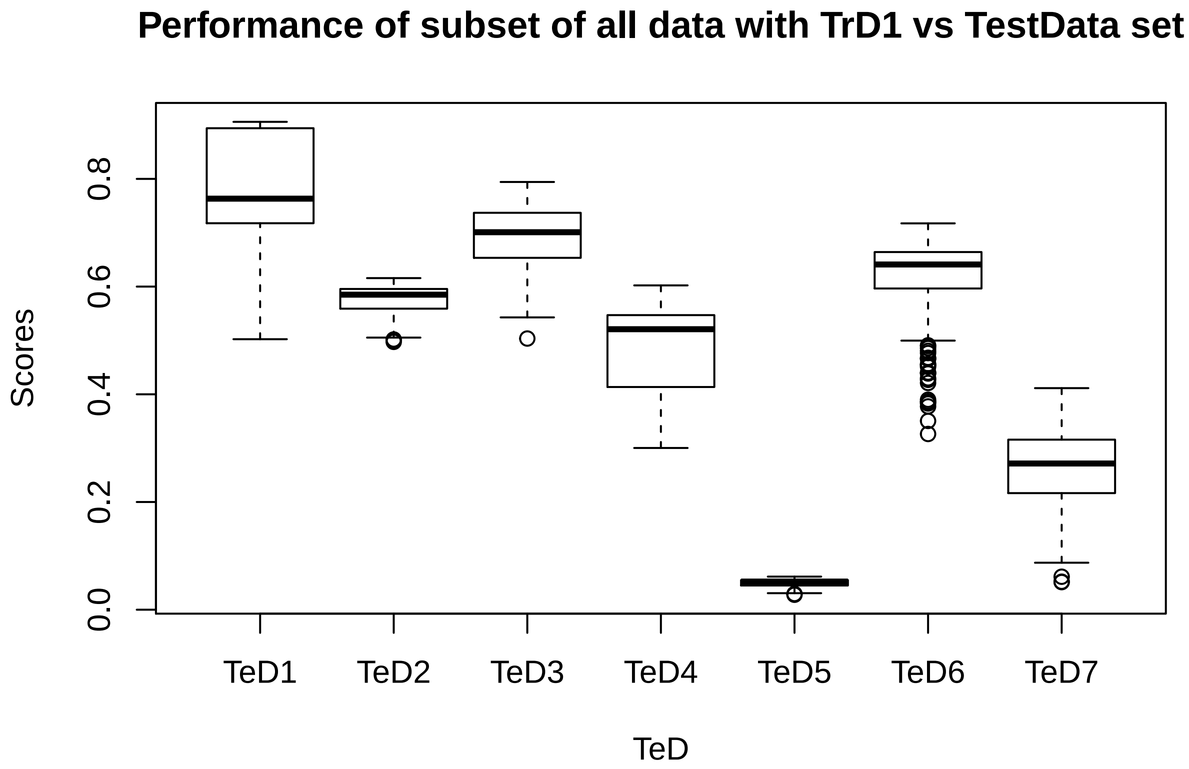
### Subset analysis

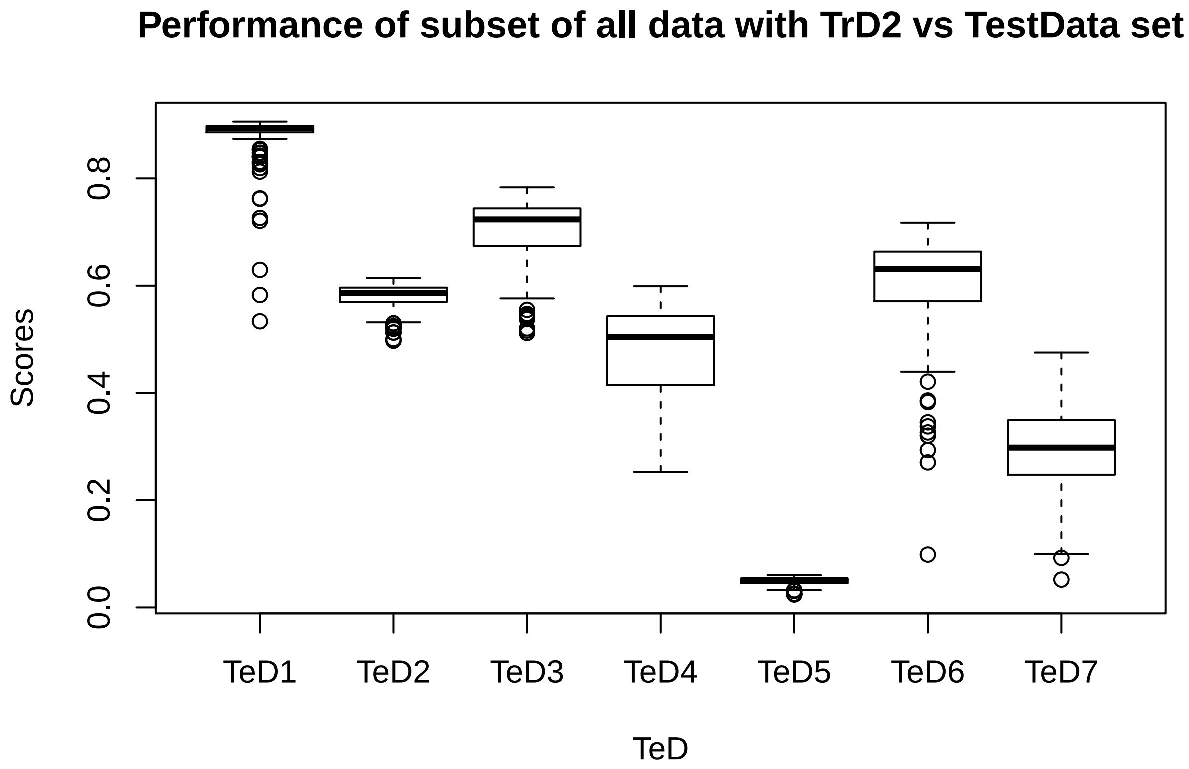
To analyze the performance that one subset adds to the overall performance we take the subset of all data where a certain training set is used and look if the performance has an observable difference.



From this we can see that the addition of a single training set has no effect on the overall average performance when considering all Test data sets and models.

We visualize the subsets of all combinations of TrD1 and TrD2 and plot their scores against the test scores. Overall, for most training sets the scores are similar. Although for TrD2 there does seem to be a higher score when performed on TeD1, leading to believe that TrD2 performs particularly well on TeD1. For the other Training sets very similar results were observed as for TrD1 thus we choose not to visualize these.


Performance of subset of all data with TrD1 vs TestData set
Scores
0.0
0.2
0.4
0.6
0.8
TeD1
TeD2
TeD3
TeD4
TeD5
TeD6
TeD7
TeD



## Conclusion

We find that the combinations of training sets that give the best average performance contain the datasets TrD2 and TrD8. We observe a positive effect($p<0.001$) on the score with an increase in the amount of sets used. Analysis of the performance of the single training set models we observe that some single training set have a higher average performance in general. TrD 5,7 and 8 provide the best performance. Notably 2, has a low mean performance but also very high variance. Only TrD1($p = 0.03476$), TrD2($p=0.00123$), TrD5($p=0.01895$), TrD7($p=0.02796$), TrD8($p=0.00207$) are significant for the score using a threshold of $p<0.05$. The analysis of the subsets shows that over all models and combinations of training data sets the performance averages out to the same.