Coursework Assignment B

CS4125 Seminar Research Methodology for Data Science

Navin Raj Prabhu - 4764722 , Javed Hassan - 4922573 , Pierpaolo Lucarelli - 4900626

4th March, 2019

#library load  
library(ggplot2)  
library(ggpubr)

## Loading required package: magrittr

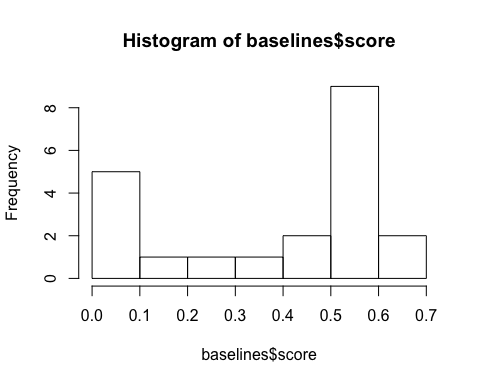
library(car) #Package includes Levene's test

## Loading required package: carData

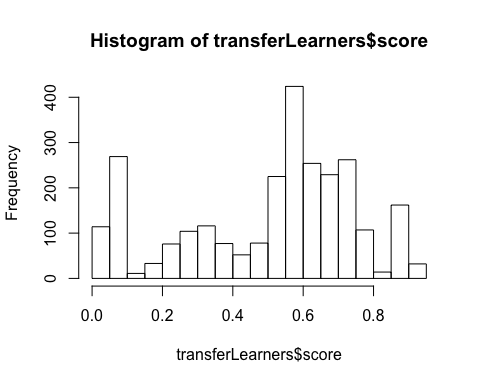
#Data Proc:::  
data = read.csv("data.csv")  
  
# baselines dont use transfer learning  
baselines <- data[data$model == 'B1' | data$model == 'B2' | data$model == 'B3',]  
  
# these models do use transfer learning - multiple training dataset  
transferLearners <- data[data$model != 'B1' & data$model != 'B2' & data$model != 'B3' & data$model != 'S',]  
  
# these models do use transfer learning - single training dataset  
singleDataTransferLearner <- data[data$model == 'S',]  
  
  
singleDataTransferLearner$train<- ifelse(singleDataTransferLearner$TrD1==1, 'TrD1',   
 ifelse(singleDataTransferLearner$TrD2==1, 'TrD2',   
 ifelse(singleDataTransferLearner$TrD3==1, 'TrD3',   
 ifelse(singleDataTransferLearner$TrD4==1, 'TrD4',  
 ifelse(singleDataTransferLearner$TrD5==1, 'TrD5',   
 ifelse(singleDataTransferLearner$TrD6==1, 'TrD6',  
 ifelse(singleDataTransferLearner$TrD7==1, 'TrD7',  
 ifelse(singleDataTransferLearner$TrD8==1, 'TrD8', 1))))))))

# 1. RQ1: How much does transfer learning improve over typical non-transfer learning?

#RQ1:  
  
# plot the means  
Bmean <- mean(baselines$score)  
TLmean <- mean(transferLearners$score)  
  
hist(baselines$score)



hist(transferLearners$score)



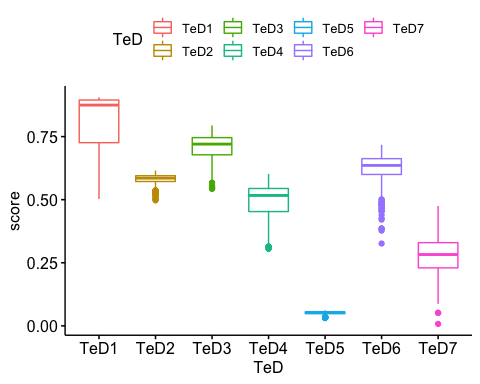
# H0: Is the mean of the transfer learners more than the mean of the baseline scores?  
# we test this hypothesis using a t-test  
significance <- t.test(transferLearners$score, baselines$score, alternative="greater")  
significance

##   
## Welch Two Sample t-test  
##   
## data: transferLearners$score and baselines$score  
## t = 2.4239, df = 20.349, p-value = 0.01239  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 0.03666383 Inf  
## sample estimates:  
## mean of x mean of y   
## 0.5048862 0.3780441

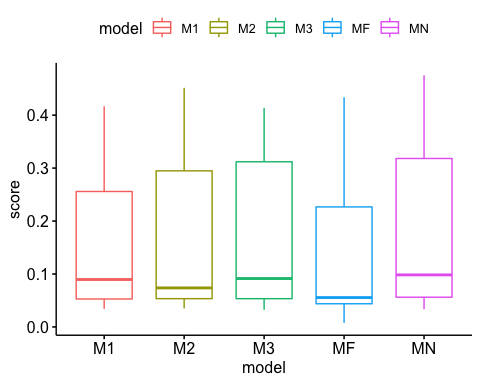
conclusion - The t-test returns a p-value of 1 which means that we do not need to reject H0. This gives an indication that the means on the methods that do make use of transfer learning are on average better than the ones that do not.

# 2. RQ2: What is the effect of different strategies to simultaneously learn one model from multiple TrD’s?

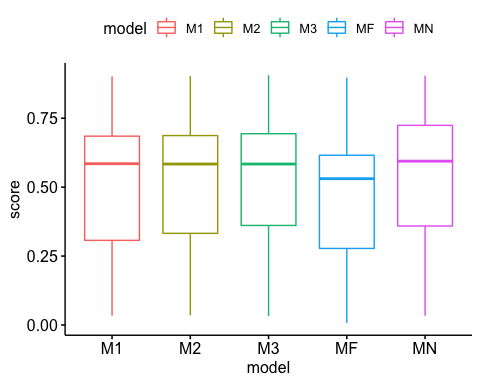
#RQ2:  
#Boxplot mean distribution - wrt Test Data  
ggboxplot(transferLearners,x = "TeD", y = "score", color = "TeD")



transferLearners\_Bad\_TeD = transferLearners[transferLearners$TeD == 'TeD5' | transferLearners$TeD == 'TeD7',]  
transferLearners\_Good\_TeD = transferLearners[transferLearners$TeD != 'TeD5' | transferLearners$TeD != 'TeD7',]  
  
ggboxplot(transferLearners\_Bad\_TeD,x = "model", y = "score", color = "model")



ggboxplot(transferLearners\_Good\_TeD,x = "model", y = "score", color = "model")



#leveneTest(score ~ model, data = transferLearners)  
  
# Compute the analysis of variance  
res\_aov <- lm(score ~ model + TeD, data = transferLearners)  
# Summary of the analysis  
summary(res\_aov)

##   
## Call:  
## lm(formula = score ~ model + TeD, data = transferLearners)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.253125 -0.022571 0.000525 0.037843 0.204602   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.809621 0.003672 220.505 < 2e-16 \*\*\*  
## modelM2 0.005737 0.003480 1.649 0.09933 .   
## modelM3 0.011600 0.003540 3.277 0.00106 \*\*   
## modelMF -0.054194 0.003447 -15.724 < 2e-16 \*\*\*  
## modelMN 0.022119 0.003566 6.203 6.43e-10 \*\*\*  
## TeDTeD2 -0.224968 0.004174 -53.898 < 2e-16 \*\*\*  
## TeDTeD3 -0.097859 0.004174 -23.445 < 2e-16 \*\*\*  
## TeDTeD4 -0.315307 0.004174 -75.542 < 2e-16 \*\*\*  
## TeDTeD5 -0.754505 0.004174 -180.766 < 2e-16 \*\*\*  
## TeDTeD6 -0.186175 0.004174 -44.604 < 2e-16 \*\*\*  
## TeDTeD7 -0.526284 0.004174 -126.088 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05731 on 2628 degrees of freedom  
## Multiple R-squared: 0.9474, Adjusted R-squared: 0.9472   
## F-statistic: 4736 on 10 and 2628 DF, p-value: < 2.2e-16

# Pairwise t-tests for - Effect of different Strategies  
pairwise.t.test( transferLearners$score , transferLearners$model)

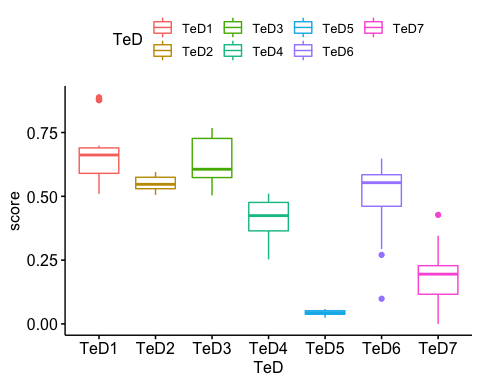
##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: transferLearners$score and transferLearners$model   
##   
## M1 M2 M3 MF   
## M2 1.00000 - - -   
## M3 1.00000 1.00000 - -   
## MF 0.00202 0.00052 0.00015 -   
## MN 0.91323 1.00000 1.00000 7.1e-06  
##   
## P value adjustment method: holm

Conclusion:

1. By plotting boxplot, we see that there is no apparent difference between the model’s accuracy.
2. From the output of the Levene’s Test results above, we can see that the p-value is greater than the significance level of 0.05. This means that there is evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the homogeneity of variances in the different treatment groups. So, we can use the ANOVA test. 3.In one-way ANOVA test, a significant p-value indicates that some of the group means are different, but we don’t know which pairs of groups are different. It’s possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of group are statistically significant.
3. From the results of the Pairwise and the Tuckey test, we can see that there no difference between the models. That is, given the data and eveidences of performance, we can say that there is no effect of the model strategies on the performance.
4. But, at the same time the model startegy MF is different from the other models, hence the strategy MF has an effect on the performance. But, from the boxplot, we can see that the mean distribution of MF model accuracies is less, hence we can stringly recommend against using strategy MF.

# 3. RQ3: What is the effect of the TrD’s on the final model performance?

#RQ3:  
  
#Boxplot mean distribution  
ggboxplot(singleDataTransferLearner,x = "TeD", y = "score", color = "TeD")



# Compute the analysis of variance  
res\_aov <- lm(score ~ TeD, data = singleDataTransferLearner)  
  
# Summary of the analysis  
summary(res\_aov)

##   
## Call:  
## lm(formula = score ~ TeD, data = singleDataTransferLearner)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.41416 -0.03311 0.00204 0.03605 0.24426   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.66248 0.01157 57.265 < 2e-16 \*\*\*  
## TeDTeD2 -0.11135 0.01636 -6.806 4.77e-11 \*\*\*  
## TeDTeD3 -0.02571 0.01636 -1.572 0.117   
## TeDTeD4 -0.25073 0.01636 -15.325 < 2e-16 \*\*\*  
## TeDTeD5 -0.61959 0.01636 -37.871 < 2e-16 \*\*\*  
## TeDTeD6 -0.14965 0.01636 -9.147 < 2e-16 \*\*\*  
## TeDTeD7 -0.47941 0.01636 -29.303 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08015 on 329 degrees of freedom  
## Multiple R-squared: 0.8821, Adjusted R-squared: 0.88   
## F-statistic: 410.3 on 6 and 329 DF, p-value: < 2.2e-16

pairwise.t.test(singleDataTransferLearner$score , singleDataTransferLearner$train)

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: singleDataTransferLearner$score and singleDataTransferLearner$train   
##   
## TrD1 TrD2 TrD3 TrD4 TrD5 TrD6 TrD7  
## TrD2 1.00 - - - - - -   
## TrD3 1.00 1.00 - - - - -   
## TrD4 1.00 1.00 1.00 - - - -   
## TrD5 0.23 1.00 1.00 1.00 - - -   
## TrD6 1.00 1.00 1.00 1.00 1.00 - -   
## TrD7 0.57 1.00 1.00 1.00 1.00 1.00 -   
## TrD8 0.24 1.00 1.00 1.00 1.00 1.00 1.00  
##   
## P value adjustment method: holm

Conclusion: The same analysis strategy of the RQ2 was used here, but before applying the annova model, the data was flattened to make data of format, TrainData and equivalent scores. The post flattened dataset, consisted of scores and its quivalent training data label. As a single model can use multiply training data, copies of scores with respective to all the training data used were generated. For Example, if Model M3 used TrD1 and TrD2 to give accuracy 0.5, the flattened data will consist of two rows with (TrD1, 0.5) and (TrD2, 0.5) as row values.

1. From boxplot we can notice an apparent difference in mean distribution.
2. This visuzlisation is backed up by the use of paired t-test. The t-test reveals that there no difference in the distribution between each of the groups. That means that the Training datasets have no effect on the performance of the final model.
3. One draw back of the system is that, the effect of combinations of TrDs cannot be explaine by the statistical analysis