CS4125 Coursework B, Group 10

Thomas Bos
Delft University of Technology
Delft, Netherlands
t.c.bos@student.tudelft.nl

Daniël van Gelder
Delft University of Technology
Delft, Netherlands
d.vangelder-1@student.tudelft.nl

Jessie van Schijndel
Delft University of Technology
Delft, Netherlands
j.vanschijndel@student.tudelft.nl

I. Results

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

In order to identify the difference in performance between transfer learning and non-transfer learning, we will investigate the following sub-questions:

1) What is the overall difference in scores between non-transfer and transfer learning models?
2) Is there a significant effect of the test dataset on the score?
3) Is there a significant difference in scores between transfer learning models that train on a single dataset and on two or more datasets?
4) Does refining a transfer learning model improve the model score? How much refinement is necessary for a significant improvement? What is the influence?

We will first explore the data obtained by running the experiments. In Figure 1 the score distribution for the model types across the three different tasks is visualized. It is immediately clear that performance on the recommendation task (test dataset `TeD5`) is very low, while the scores for classification (test datasets `TeD1` - `TeD4`) and regression (`TeD5` and `TeD6`) vary greatly. In order to gain better insight we let go of the analysis per task and perform analysis per test data set (`TeD`). In Figure 2 we can clearly make out Gaussian distributions of scores with varying means for each `TeD`, indicating that `TeD` can have an effect on the score. To visualise the difference across the different Transfer Learning models, in Figure 3 and 4 the scores of the Transfer Learning models across all `TeDs` is visualised. We can see that it is likely that using multiple training sets and refinement improves the scores of the models. Testing should show whether these effects are significant.

To test the effects of the test datasets on the score a linear mixed effects analysis is performed. Our null model (m_0) has no fixed effects apart from a fixed intercept for each test dataset fitted to the score. The effects model (m_1) adds the model type (non-transfer learning vs. transfer learning) as a fixed effect. The effects model has a significantly better fit ($p < 0.0001$). The random effect of the test dataset explains 76.9% (0.238 out of 0.309 total Std. Dev.) of the variance left after the fixed effects and thus has a strong effect on the test score.

Having concluded that the Test Dataset does show a significant effect on the score. We evaluate if there is still a significant difference in mean score for each test set. To account for other factors, an Estimated Marginal Means model was created and the outcome of the tests are shown in table I. It shows that for 3 out of the 7 test datasets (`TeD2`, `4` and `5`) the difference is not significant ($p > 0.05$).

Next, we analyse whether transfer learning achieves a better mean score on all test sets through a pairwise Estimated Marginal Means analysis accounting for varying test datasets. Results show a significant score increase ($t(2982) = 7.587,\ p < 0.01$) of 0.118 when transfer learning is used.

Now, we will examine the differences in mean scores between transfer learning models. Comparing the mean scores

All authors contributed equally to this work.

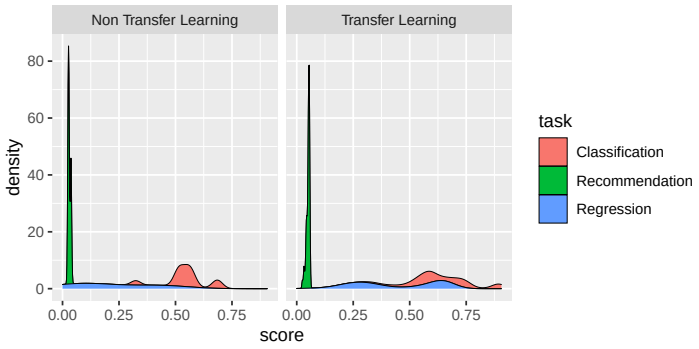

Fig. 1. Histogram plots that show the distribution of the model scores across the three different tasks.

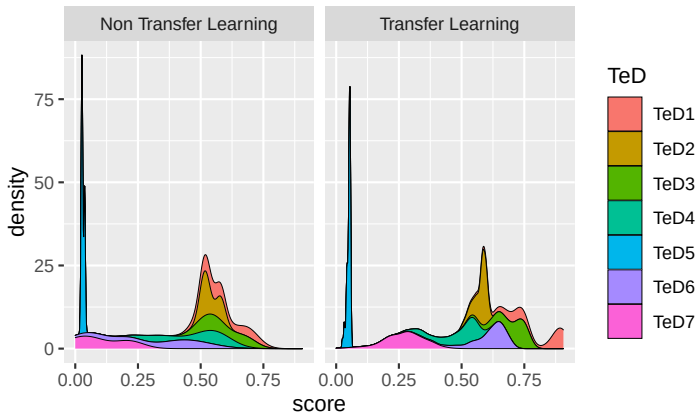

Fig. 2. Density plots of the scores per TeD for transfer and non-transfer learning models.

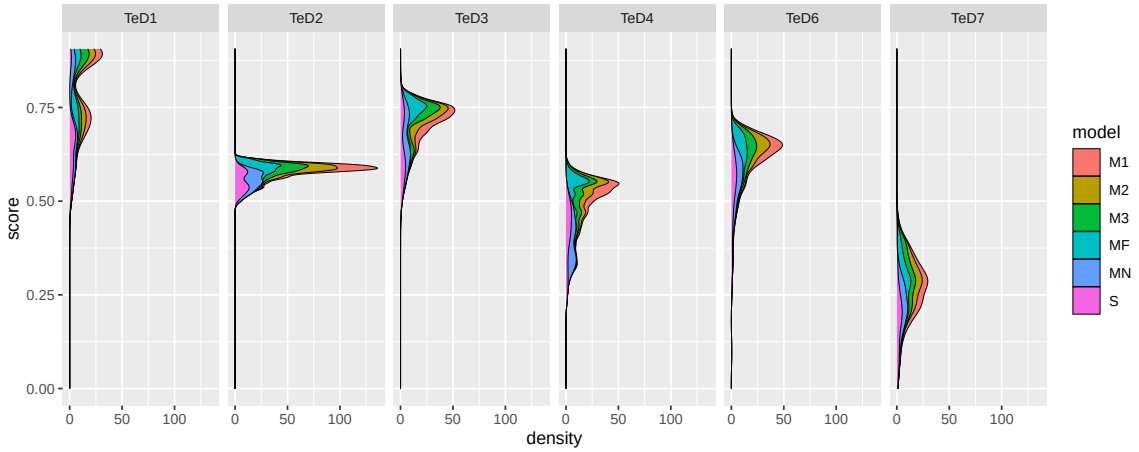

Fig. 3. The scores of the Transfer Learning models across the different TeDs. TeD5 has been put in a different plot (Figure 4) as it has a very small and high density peak.

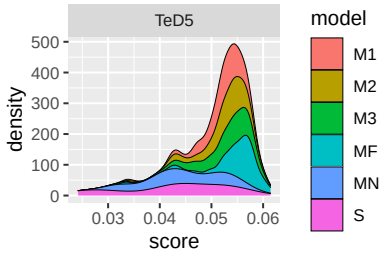

Fig. 4. The scores of the Transfer Learning models on TeD5. The density is concentrated around 0.05 which means all models perform badly on this set. It still looks like refinement and multiple training sets improve the score but tests should show whether this is significant.

Task	Test Dataset	Estimated Difference	t-ratio	p-value
Classification	TeD1	0.199	4.84	< 0.0001
Classification	TeD2	0.039	0.96	0.34
Classification	TeD3	0.117	2.84	0.005
Classification	TeD4	0.002	0.06	0.95
Recommendation	TeD5	0.020	0.48	0.63
Regression	TeD6	0.278	6.73	< 0.0001
Regression	TeD7	0.172	4.17	< 0.0001

TABLE I
SIGNIFICANCE TESTS ON DIFFERENCE IN MEANS OF SCORE OF TRANSFER LEARNING AND NON-TRANSFER LEARNING MODELS ON SPECIFIED TEST SET. IN ALL CASES THE TESTS HAD 2982 DEGREES OF FREEDOM AND A STANDARD ERROR OF 0.0412

on the test datasets of the transfer learning models that train on either one or more than one training sets shows that using additional training sets significantly increases the mean score ($t(2961) = 20.11,\ p < 0.0001$). Comparing the mean score of transfer learning models shows that refining one part of the model on the target test set significantly increases the score over performing no refinement ($t(2933) = 15.5,\ p < 0.0001$). The differences between different amounts of refinement are small. Doing refinement on the full model yields a significantly higher mean score over refining one ($t(2933) = 6.114,\ p < 0.0001$) and two parts ($t(2933) = 4.515,\ p = 0.0001$), but is not significantly better ($t(2933) = 2.852,\ p = 0.05$) than doing refinement on three parts of the model only. We can conclude that, over all test datasets, refining a transfer learning model to the test dataset is beneficial but suffers from diminishing returns as larger parts of the model are refined. Figure 6 shows the means per model type.

We will now investigate the influence of refinement of the models on the mean scores on the different test sets. From figure 5 it looks like as more refinement is done, the mean score seems to increase on all test data sets, although on `TeD1` the median has decreased with full model refinement. Carrying out pairwise comparisons of each model while faceting by test datasets shows that on all test sets except `TeD5` there is a significant improvement ($p < 0.01$) in doing any level of refinement. However, refining more than one component of the model gives only significantly better results ($p < 0.05$) on some of the test sets (`TeD1`, `TeD2`, `TeD4`, `TeD7`).

Overall, we can thus conclude that in this experiment transfer learning models improve over typical non-transfer learning models by an average mean score of 0.118. However, a significant difference could not be found for the scores 3 out of the 7 test datasets. When relating the test datasets to the Machine Learning tasks, we observe that for the regression task the difference is significant (see Table I), for the recommendation task no significant difference could be found and for the classification task a significant difference could only be found for 2 of the 4 test datasets. The improvement that the transfer learning models make can be further strengthened by refining parts of the transfer learning model, seen in figure 6. Training a transfer learning model on more than one dataset is beneficial as well. In the next section we will further quantify this improvement as well.

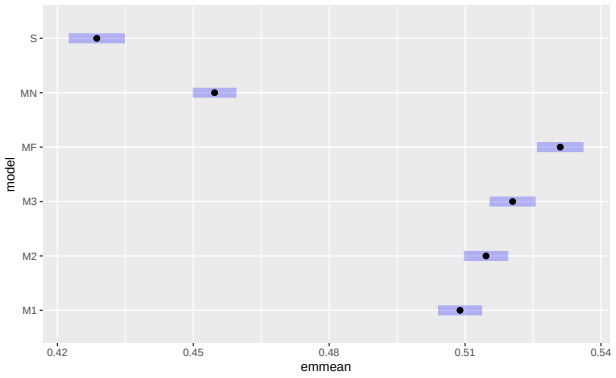

Fig. 6. Estimated Marginal Means of score for refining transfer learning models. S represents a model trained on a single training dataset with no refinement. Models MN, M1, M2, M3 and MF represent models trained on multiple training datasets with no refinement and one, two and three parts refinement and full refinement, respectively.

RQ2: What is the effect of the `TrD`*'s on the final model performance?*

To investigate the effect of the `TrD`'s on the final model performance, we ask ourselves the following questions:

1) What is the effect of the inclusion of each `TrD` on the final model performance?
2) What is the effect of the number of `TrD`'s used on the final model performance?

We then create four models, m_0, m_1, m_2 and m_3. Model m_0 contains a fixed intercept and a random intercept for each `TeD`. This model will serve as our baseline. It is the same model as m_0 in RQ1. The intercept of this model has an estimated value of 0.495 (95% CI [0.319, 0.672]). The random effect of the `TeD` variable is estimated to explain 76% (0.238 out of 0.309 total SD) of the variance left and thus has a strong effect on the test score.

Model m_1 contains a fixed intercept for each of the `TrD`s and a random intercept for each `TeD`. This model can be used to assess the influence of each individual `TrD`. The estimated parameter values are shown in Table II. From these we see that most training data sets have a significant effect on the score as 0 is not included in the 95% confidence intervals. Only `TrD1` shows a negative effect. All other effects are positive. The parameter for `TrD4` is not significant. The random effect of the `TeD` variable is estimated to explain 76% (0.238 out of 0.309 total SD) of the variance left and thus has a strong effect on the test score. Model m_1 has a significantly better fit than m_0 ($p < 0.0001$) as can be seen in table III.

	lower	est.	upper
(Intercept)	0.2631648548	0.439589418	0.61601398
TrD1	-0.0254206479	-0.020523354	-0.01562606
TrD2	0.0250630806	0.029688450	0.03431382
TrD3	0.0053079824	0.010264103	0.01522022
TrD4	-0.0002136489	0.004862280	0.00993821
TrD5	0.0186913890	0.023800344	0.02890930
TrD6	0.0043191379	0.009069989	0.01382084
TrD7	0.0167247598	0.021697763	0.02667077
TrD8	0.0249501291	0.029825458	0.03470079

TABLE II
APPROXIMATE 95% CONFIDENCE INTERVALS FOR m_1

Model Name	Df	AIC	BIC	Log. Lik.	p-value
$\mathbf{m_0}$	3	-7180	-7162	3593	
$\mathbf{m_1}$	11	-7943	-7877	3983	$< .0001$

TABLE III
ANOVA TEST BETWEEN m_0 AND m_1.

Model m_2 contains a fixed intercept for the number of training sets used and a random intercept for each `TeD`. This model can be used to assess the influence of the number of `TeD`s used. The intercept of this model has an estimated value

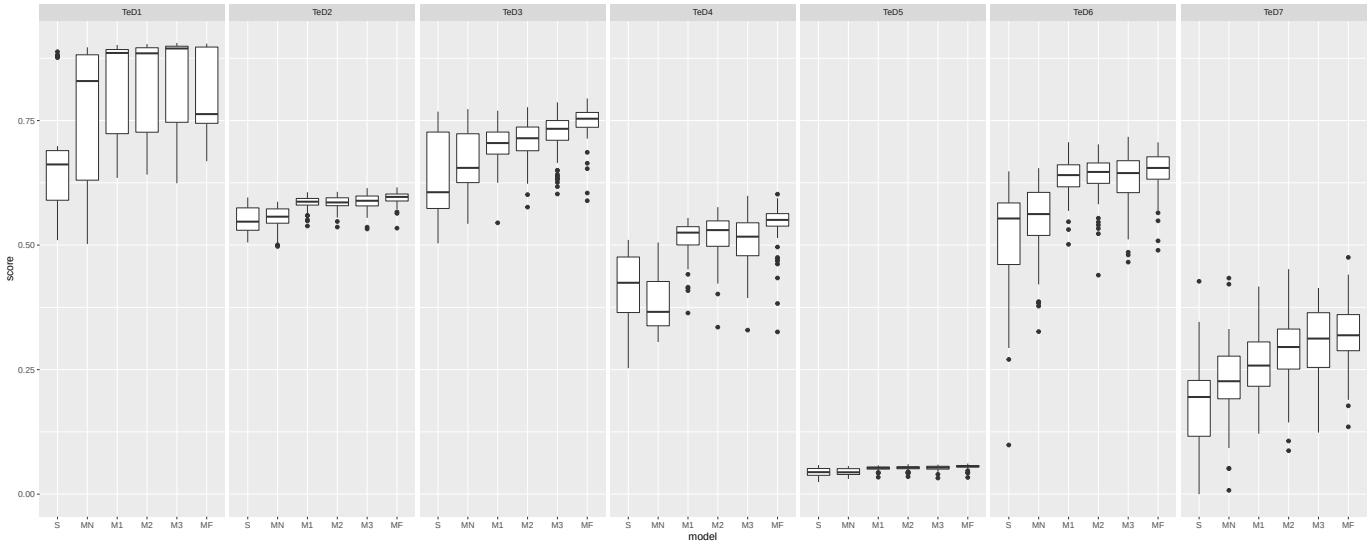

Fig. 5. Boxplots of the transfer learning models (`S` and `Mx`) over varying test datasets (`TeDx`)

of 0.442 (95% CI [0.266, 0.618]). The effect of the number of TeDs used has an estimated value of 0.013 (95% CI [0.012, 0.014]). The random effect of the TeD variable is estimated to explain 76% (0.238 out of 0.309 total SD) of the variance left and thus has a strong effect on the test score. To further investigate this effect, we look at figure 7. The median scores go up as the number of TrDs increases. This supports our previous findings. The increase in score seems to become smaller for a higher trdcount. From the image, we see that there might be an logistic relation.

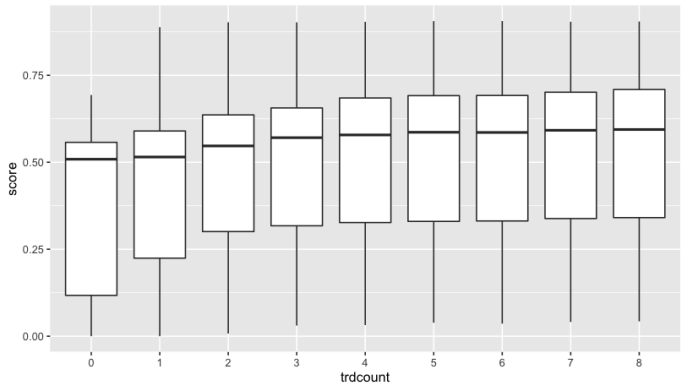

Fig. 7. Boxplots visualizing test scores for different number of TrDs.

Model m_2 has a significantly worse fit than m_1 ($p < 0.0001$) as can be seen in table VI. Therefore, using the amount of TrDs as a parameter is not better than accounting for each individual TrD.

Model Name	Df	AIC	BIC	Log. Lik.	p-value
$\mathbf{m_1}$	11	-7943	-7877	3983	
$\mathbf{m_2}$	4	-7633	-7609	3820	$< .0001$

TABLE IV
ANOVA TEST BETWEEN m_1 AND m_2.

Model m_3 extends model m_1 by adding interaction terms between TeD1 and TrD2 and between TeD5 and TrD2. The resulting confidence intervals are shown in V. Model m_3 has a significantly better fit than m_1 ($p < 0.0001$) as can be seen in table VI.

We conclude that the TrDs have a significant effect on the score. We also conclude that measuring the individual TrDs is a better predictor of test score than incorporating simply the number of TrDs used. For most TrDs, their inclusion result in a higher test score. For TrD1 the opposite is true, and for TrD4 the positive effect is not certain. The interaction between TrD2 and TeD1 and between TrD2 and TeD5 has a significant effect and including them in the model results in the best fit.

	lower	est.	upper
(Intercept)	0.372245840	0.481142898	0.590039956
TrD1	-0.023543879	-0.017708777	-0.011873675
TrD2	0.002091422	0.008206857	0.014322293
TrD3	0.005931187	0.010186652	0.014442118
TrD4	0.000456431	0.004813831	0.009171232
TrD5	0.019378674	0.023764173	0.028149672
TrD6	0.005034349	0.009112627	0.013190906
TrD7	0.017513401	0.021783687	0.026053974
TrD8	0.025893886	0.030096680	0.034299475
Ted.TeD1	-0.209868077	0.167579602	0.545027280
Ted.TeD5	-0.846524897	-0.469077218	-0.091629540
TrD1:TrD2	-0.013307642	-0.005458648	0.002390347
TrD2:Ted.TeD1	0.171622855	0.182927234	0.194231613
TrD2:Ted.TeD5	-0.024662455	-0.013358076	-0.002053697

TABLE V
APPROXIMATE 95% CONFIDENCE INTERVALS FOR m_3

Model Name	Df	AIC	BIC	Log. Lik.	p-value
$\mathbf{m_1}$	11	-7943	-7877	3983	
$\mathbf{m_3}$	16	-8856	-8759	4444	$< .0001$

TABLE VI
ANOVA TEST BETWEEN m_1 AND m_3.