Introduction to machine learning
Report II: Classification and regression

Kristin Anett Remmelgas s203129
Kristo Karl Aedma s205738
Andreas Lau Hansen s194235

November 17 2020

Contents

1 Regression

1.1 Regression part a - linear regression

1.1.1 Introduction and feature transformations

The task in the regression part is to predict the average vote of a movie using all other attributes except metascore. As described in the previous report the average vote and metascore are highly correlated and we decided not to use one score to predict another. The used attributes are year, duration, number of votes, number of reviews from users, number of reviews from critics and the movie genre. All other attributes except the genre are continuous. The genre attribute has been one-out-of-K coded adding 20 new columns to the data matrix representing different genres. A movie can also have multiple genres. Since we will be using regularization in the next tasks the data was standardized so that each column of the data matrix $\mathbf{X}$ has mean 0 and standard deviation 1.

1.1.2 Regularized linear regression

In this task a regularization parameter λ was introduced and the generalization error for different values of λ was estimated. The range of λ-values was chosen by doing a few test runs and decided to be from $10^{-5}...10^{9}$. Ideally the range should be chosen so that the generalization error would first drop and then start to increase but as can be seen from Figure 1 this could not be found. For estimating the generalization error K = 10 fold cross-validation (algorithm 5) was implemented and the result can be seen in Figure 1.

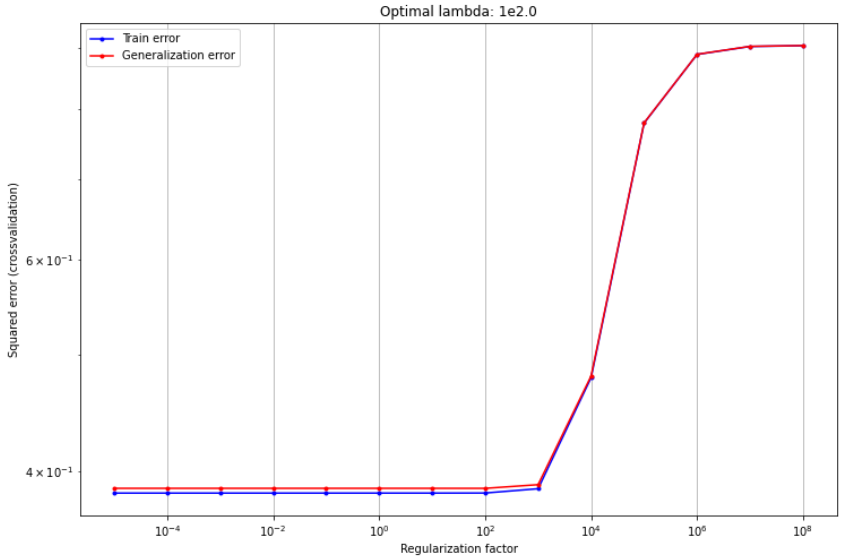


Figure 1: Generalization error as a function of λ

The results show that for every λ value before 10^2 the generalization errors are very similar. This shows that the weights for the attributes are almost the same no matter the λ-value in that range. When taking a closer look at the graph it can be seen that the generalization error values are similar but not exactly the same and there is a small decrease around $\lambda = 10^2$ which has been estimated to be the optimal value.

1.1.3 Attribute weights

The linear regression model predicts the output by using the calculated weights for different attributes. The weights for the model with the optimal value $\lambda = 100$ can be seen in Table 1. It can be seen from the table that the attribute "reviews from users" (number of reviews) has the largest positive effect on the value predicted by the model. Therefore, the more users have reviewed the movie, the higher the average vote is predicted to be. This might be because movies with a higher vote have more views and thus are more likely to be reviewed. It can also be seen that the continuous attributes affect the result more than the genres. Attributes drama, horror, animation and biography have the largest weights out

of all the genres. All four except horror genre have a positive affect on the prediction. All in all the results seem realistic.

Table 1: Weights calculated by the best linear regression model

Offset	6.22
year	-0.11
duration	0.13
votes	0.13
reviews_from_users	0.49
reviews_from_critics	0.12
Action	-0.01
Adventure	-0.03
Animation	0.08
Biography	0.05
Comedy	-0.02
Crime	0.02
Drama	0.09
Family	-0.01
Fantasy	-0.03
History	0.01
Horror	-0.09
Music	0
Musical	-0.02
Mystery	-0.01
Romance	0.01
Sci-Fi	-0.03
Sport	0.02
Thriller	-0.03
War	0.03
Western	-0.01

1.2 Regression part b - 2-level cross-validation

The task in this part is to compare three different models: regularized linear regression, baseline and artificial neural network using two level cross-validation with $K_1 = K_2 = 10$. In order to reuse the train/test splits for all three models a random seed was added when splitting the data. The same seed was then used for all methods.

1.2.1 Linear regression and baseline

The linear regression was implemented as in part a (with the same range of values for λ) but now adding an outer K-fold loop $K_1 = 10$. The results for the last outer fold can be seen in Figure 2. The figure on the right includes the test and train errors for the linear regression model which are similar to the result from part a. The figure also shows the test error for the baseline model which is higher than the linear regression models'. As a baseline model a linear regression model with no features was used (computes the mean of y on the training data, and uses this value to predict y on the test data). The figure on the left depicts the mean coefficient values (excluding the offset) for each λ. It can be seen that most of the attributes affect the result minimally. Also the mean value is rather stable for all λ-values. The optimal λ-value for this specific fold is 10.

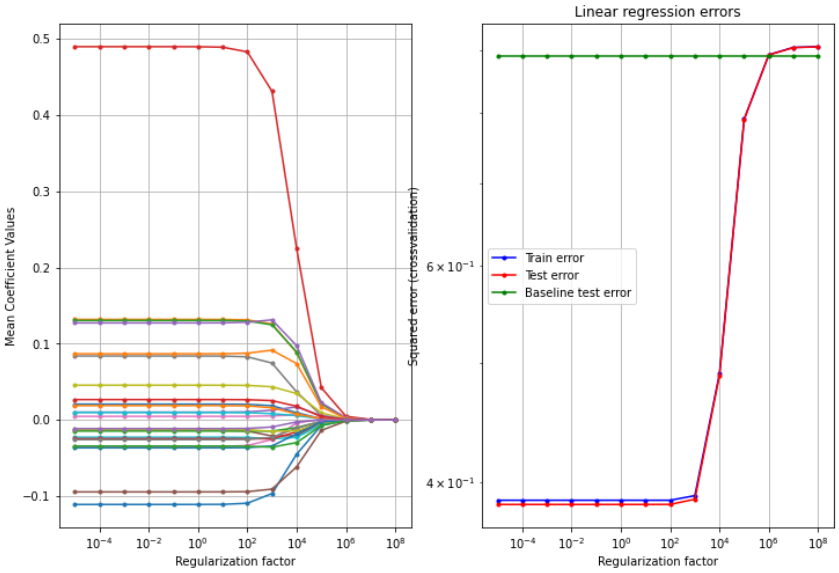


Figure 2: Linear regression last outer fold results

1.2.2 Artificial neural network

An ANN model was fitted to the data. As complexity-controlling parameter for the ANN, the number of hidden units h was used. Based on a few test-runs, a reasonable range of values for h was chosen to be 1-15 units in one hidden layer. The results from the last fold can be seen in Figure 3.

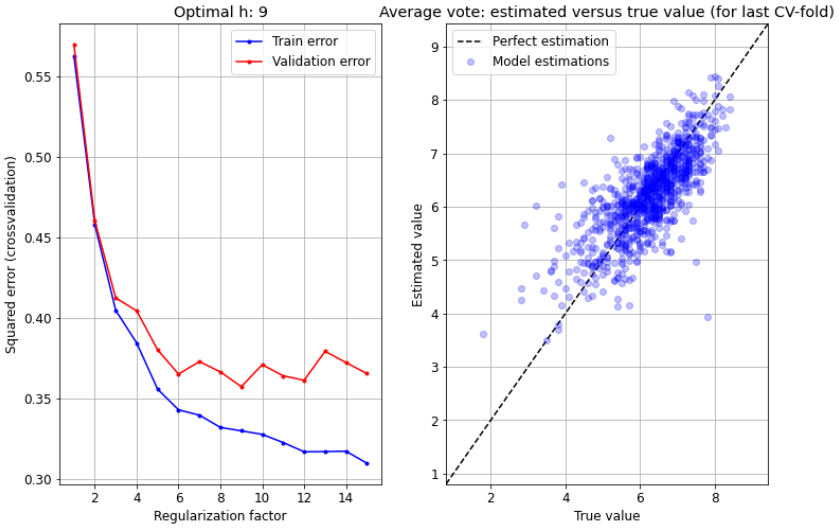


Figure 3: Last outer fold results for ANN

In the specific fold shown in Figure 3 the optimal number of hidden units was 9 but as can be seen from the figure on the left the validation error fluctuates quite much therefore the difference in error compared to using 6 hidden units (for example) is not much. The figure on the right shows the estimated versus true value of the data for this fold. It can be seen that most of the results are rather close to the line depicting the perfect estimation. It can also be said that for movies with lower true average vote values the model tended to predict higher values than the real ones.

1.2.3 Two-level cross-validation table

The results from the two-layer cross-validation were put into a table. The table shows for each of the outer folds $K_1 = 10$ the optimal value of the hidden units and regularization factor as found in each

inner loop as well as the estimated generalization errors. It also includes the baseline test error. Mean square loss per observation (MSE) was used as the error measure.

Table 2: Two-level cross-validation table used to compare the three models

Outer fold	ANN		Linear regression		Baseline
k	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	14	0.339	100	0.385	0.896
2	8	0.366	100	0.379	0.872
3	13	0.334	10	0.380	0.965
4	10	0.318	10	0.360	0.893
5	11	0.385	10	0.418	0.904
6	13	0.320	100	0.351	0.868
7	12	0.442	100	0.424	0.920
8	15	0.371	10	0.391	0.918
9	12	0.372	100	0.395	0.909
10	9	0.383	10	0.381	0.889

The table shows that, as mentioned before, the optimal value of hidden units fluctuates quite much in each outer fold ranging from 8-15 units. 12 and 13 being the only values represented more than once. The value for the optimal regularization factor varies between two values: 10 and 100 which is in accordance with regression part a results. The lowest errors for each method are: 0.318 for ANN, 0.351 for the linear regression model and 0.868 for the baseline method. From this table it can be estimated that the artificial neural network performed the best, followed closely by the linear regression model and the baseline having by far the largest error but to compare the three models adequately statistical comparisons were carried out in the next part.

1.2.4 Statistical evaluations

The task in this part is to compere the best of each three methods statistically. The paired t-test from setup 1 was chosen for this. Hold-out cross-validation method was used, the number of hidden units for the ANN model was chosen to be 12 and the regularization factor for linear regression was 100. The comparisons were made pairwise (linear regression vs baseline, baseline vs ANN, ANN vs linear regression). The results of the statistical evaluations are as follows:

- The linear regression VS baseline confidence interval is: $(-0.6109, -0.4712)$ and the p-value is $4.27 \cdot 10^{-47}$.
- The baseline VS ANN confidence interval is: $(0.4984, 0.6508)$ and the p-value is $5.95 \cdot 10^{-45}$.
- The ANN VS linear regression confidence interval is: $(-0.0552, -0.0119)$ and the p-value is 0.00118.

The significance level α was chosen to be 0.05 which means that if p falls below this threshold it can be said that the test showed a significant result and the hypothesis that the models perform equally can be rejected. It can be seen from the results that when comparing the other two models with the baseline model the p-value is extremely small and the confidence interval does not include zero, meaning that the artificial neural network and the linear regression model both perform significantly better than the baseline model. When comparing the ANN model to the linear regression it can be seen that the confidence interval is closer to zero and the p value is also larger which means the models perform more similarly but p is still small enough to conclude that the ANN model is better.

2 Classification

2.1 Problem

The goal of the classification problem is to predict the genre of the movie by using all other attributes: year, duration, average vote, number of votes, metascore, number of reviews from users and number of reviews from critics. It is a multiclass problem as there are 20 genres. In the regression part of this report every movie could have more than one genre but for the classification a data transformation was

made so that each movie belongs to exactly on genre (a random one from the multiple) similar to the PCA plot in report 1.

2.1.1 Setup

We will be using a classification tree, a logistic regression classifier, and a baseline. The baseline method simply predicts the class of the test set to be in the same class as the largest class in the training set (in our case most likely the drama genre). For the classification tree, we control the complexity by changing the tree depth in the range $[2, 3, .., 9]$. For the logistic regression classifier, which is actually a multinomial regression classifier due to the multiple labels, we chose the range $[10^{-12}, 10^{-11}, ..., 10^{-4}]$. These values were selected in a trial run.

2.2 Model comparison

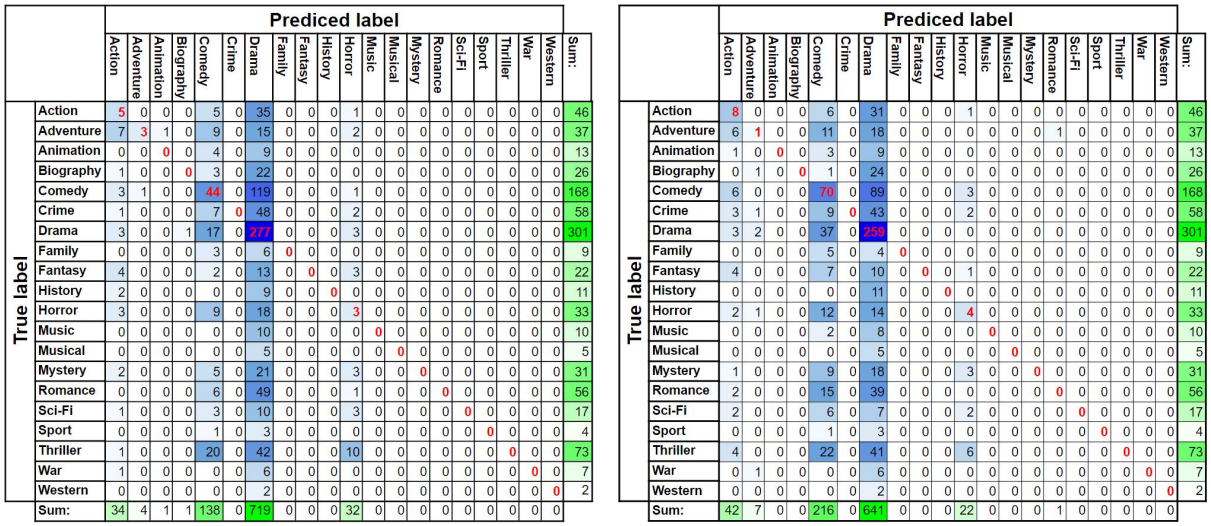
To compare the models we used a 2 layer cross-validation structure $K_1 = K_2 = 10$ as in the regression part b. The results for each outer fold are shown in Table 3. From the table, we can conclude that the best tree classifier depth is $x_i^* = 5$ and the best regularization factor is $\lambda_i^* = 10^{-6}$. However, it should be noted that during some trial runs and analysis of the outputs from the inner folds, the different values of regularization factors had diminutive differences. The error rate is calculated as a simple misclassification rate.

Table 3: Two-level cross-validation table used to compare the three models

Outer fold	Tree		Logistic regression		Baseline
k	x_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	6.0	0.647	1.0e-12	0.626	0.648
2	6.0	0.620	1.0e-12	0.606	0.627
3	6.0	0.661	1.0e-12	0.646	0.661
4	5.0	0.639	1.0e-12	0.642	0.656
5	5.0	0.636	1.0e-12	0.630	0.664
6	5.0	0.642	1.0e-06	0.623	0.647
7	5.0	0.649	1.0e-06	0.646	0.673
8	5.0	0.644	1.0e-06	0.632	0.659
9	5.0	0.611	1.0e-06	0.611	0.639
10	5.0	0.632	1.0e-06	0.643	0.676

Overall the performance of all models is extremely poor.

From now on the models were trained with the optimal parameter from Table 3 using all of our data leaving 10% for testing purposes (hold-out method). Now when we can train the best models for logistic regression and tree classifiers we can also test them and look at the confusion matrices to see how well they perform (Figure 4). Color coding is showing higher values as stronger colors. The green sides are indicating the sum of the specified genres for true values on the right and predicted values at the bottom. Needless to say both classifiers tend towards bigger classes and have zero guesses on several smaller classes. But then again 2 out of 4 of the largest valued cells belong to the correctly predicted ones.



True label \ Predicted label	Action	Adventure	Animation	Biography	Comedy	Crime	Drama	Family	Fantasy	History	Horror	Music	Musical	Mystery	Romance	Sci-Fi	Sport	Thriller	War	Western	Sum:
Action	5	0	0	0	5	0	35	0	0	0	1	0	0	0	0	0	0	0	0	0	46
Adventure	7	3	1	0	9	0	15	0	0	0	2	0	0	0	0	0	0	0	0	0	37
Animation	0	0	0	0	4	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	13
Biography	1	0	0	0	3	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	26
Comedy	3	1	0	0	44	0	119	0	0	0	1	0	0	0	0	0	0	0	0	0	168
Crime	1	0	0	0	7	0	48	0	0	0	2	0	0	0	0	0	0	0	0	0	58
Drama	3	0	0	1	17	0	277	0	0	0	3	0	0	0	0	0	0	0	0	0	301
Family	0	0	0	0	3	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	9
Fantasy	4	0	0	0	2	0	13	0	0	0	3	0	0	0	0	0	0	0	0	0	22
History	2	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	11
Horror	3	0	0	0	9	0	18	0	0	0	3	0	0	0	0	0	0	0	0	0	33
Music	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	10
Musical	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	5
Mystery	2	0	0	0	5	0	21	0	0	0	3	0	0	0	0	0	0	0	0	0	31
Romance	0	0	0	0	6	0	49	0	0	0	1	0	0	0	0	0	0	0	0	0	56
Sci-Fi	1	0	0	0	3	0	10	0	0	0	3	0	0	0	0	0	0	0	0	0	17
Sport	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Thriller	1	0	0	0	20	0	42	0	0	0	10	0	0	0	0	0	0	0	0	0	73
War	1	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	7
Western	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Sum:	34	4	1	1	138	0	719	0	0	0	32	0	0	0	0	0	0	0	0	0	

(a) Logistic regression classifier

True label \ Predicted label	Action	Adventure	Animation	Biography	Comedy	Crime	Drama	Family	Fantasy	History	Horror	Music	Musical	Mystery	Romance	Sci-Fi	Sport	Thriller	War	Western	Sum:
Action	8	0	0	0	6	0	31	0	0	0	1	0	0	0	0	0	0	0	0	0	46
Adventure	6	1	0	0	11	0	18	0	0	0	0	0	0	0	1	0	0	0	0	0	37
Animation	1	0	0	0	3	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	13
Biography	0	1	0	0	1	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	26
Comedy	6	0	0	0	70	0	89	0	0	0	3	0	0	0	0	0	0	0	0	0	168
Crime	3	1	0	0	9	0	43	0	0	0	2	0	0	0	0	0	0	0	0	0	58
Drama	3	2	0	0	37	0	259	0	0	0	0	0	0	0	0	0	0	0	0	0	301
Family	0	0	0	0	5	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	9
Fantasy	4	0	0	0	7	0	10	0	0	0	1	0	0	0	0	0	0	0	0	0	22
History	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	11
Horror	2	1	0	0	12	0	14	0	0	0	4	0	0	0	0	0	0	0	0	0	33
Music	0	0	0	0	2	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	10
Musical	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	5
Mystery	1	0	0	0	9	0	18	0	0	0	3	0	0	0	0	0	0	0	0	0	31
Romance	2	0	0	0	15	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	56
Sci-Fi	2	0	0	0	6	0	7	0	0	0	2	0	0	0	0	0	0	0	0	0	17
Sport	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Thriller	4	0	0	0	22	0	41	0	0	0	6	0	0	0	0	0	0	0	0	0	73
War	0	1	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	7
Western	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Sum:	42	7	0	0	216	0	641	0	0	0	22	0	0	0	1	0	0	0	0	0	

(b) Tree classifier

Figure 4: Logistic regression and tree classifiers' confusion matrices

2.3 Statistical evaluations

For statistical evaluations we used McNemar's test which is a part of the course's toolbox. The comparisons were made pairwise (baseline vs tree classifier, baseline vs logistic regression classifier, tree classifier vs logistic regression classifier). The results of the statistical evaluations are as follows:

- The baseline vs tree classifier confidence interval is: $[-0.0675, -0.0207]$ and the p-value is $3.0 \cdot 10^{-4}$.
- The baseline vs logistic regression classifier confidence interval is: $[-0.052, -0.0147]$ and the p-value is $6.394 \cdot 10^{-4}$.
- The tree classifier vs logistic regression classifier confidence interval is: $[-0.0314, 0.0099]$ and the p-value is 0.358.

The significance level has same meaning and value as it had in the regression part ($\alpha = 0.05$). From those results we can confirm that both, the tree classifier and the logistic regression classifier, have smaller errors than baseline and therefore are better classifiers. On the other hand 0 falls in the confidence interval for the tree classifier vs logistic regression classifier and therefore we can't reject the null hypothesis for that one.

2.4 Features for logistic- and linear regression

We can analyse which features the logistic regression model uses to make predictions similar to what we did with the linear regression model as seen in Table 1 to compare if the same features are used in the same way. For convenience the most important part of it is reprinted in Table 4

Table 4: Linear regression model attribute weights for one fold

Offset	year	duration	votes	reviews_from_users	reviews_from_critics
6.23	-0.11	0.13	0.13	0.49	0.12

The features that the logistic regression models uses are shown in Table 5. It is seen that the features it uses for classification is heavily dependent on the class. At large it looks more or less random. It's not really possible to compare it to the features used for linear regression since the linear regression model is only predicting one attribute.

Table 5: Coefficients for the logistic regression classifier

	year	duration	avg_vote	votes	metascore	reviews_from_users	reviews_from_critics
Action	0.061	0.283	-0.149	0.275	-0.432	0.166	0.244
Adventure	0.018	-0.09	-0.195	0.344	-0.22	0.212	0.272
Animation	0.295	-1.687	1.019	0.295	-0.147	0.036	0.068
Biography	0.247	0.36	0.828	-0.293	-0.131	-0.374	0.323
Comedy	-0.127	-0.426	-0.084	0.524	-0.067	-0.318	-0.265
Crime	-0.04	0.312	0.04	0.178	-0.135	-0.108	-0.082
Drama	0.018	0.162	0.239	0.086	0.389	-0.066	-0.636
Family	-0.338	-0.08	-0.185	0.216	-0.219	0.09	-0.008
Fantasy	-0.121	0.064	-0.293	0.081	-0.098	0.106	0.389
History	0.111	0.38	0.561	-0.265	0.252	0.174	-0.251
Horror	0.125	-0.759	-0.793	-0.876	0.11	0.336	0.973
Music	-0.164	0.29	0.017	0.131	0.43	-0.034	-0.505
Musical	-0.396	0.423	-0.211	-1.862	0.045	0.722	-0.403
Mystery	0.081	-0.105	-0.322	-0.052	0.059	0.22	0.205
Romance	-0.292	0.162	0.102	0.091	-0.022	-0.352	-0.283
Sci-Fi	0.032	0.018	-0.618	0.349	-0.047	0.237	0.385
Sport	-0.129	0.365	0.626	0.36	-0.382	-0.395	-0.262
Thriller	0.15	-0.137	-0.201	0.027	-0.177	0.158	0.138
War	0.052	0.352	0.911	-0.101	-0.162	-0.406	-0.39
Western	0.417	0.114	-1.292	0.491	0.956	-0.403	0.088

3 Discussion

3.1 Conclusions

From the regression part it can be concluded that both the artificial neural network and the linear regression model perform significantly better than the trivial baseline model. It can be said that the best performing model for predicting the average vote of a movie is the artificial neural network. Though if we take into consideration that it took the ANN 9 hours and the linear regression model under 10 seconds to implement the cross-validation model training it is arguable whether the improvement in results is worth it.

Regarding the classification, both models, a classification tree and a logistic regression model, were found to be (statistically significant) better than than a baseline. That is not particularly impressive as they all had an error rate of $60 - 65\%$ and they only surpassed the baseline by a few percent at best. Had we used stratification for the train/test splits, we would likely have seen a much higher error rate. Ideally we would also have liked to use the description of the movies, as it would likely have decreased the classification error. The poor performance might also have been caused by the fact that if a movie originally had multiple genres we randomly chose one of them, which might have made the genre more difficult to predict.

3.2 Comparison with previous analysis

There are 18 notebooks on Kaggle that have used this dataset. Some of them only implement data preprocessing and make different visualizations. 11 notebooks continue to making a movie recommendation system. 10 of the recommendation systems use natural language processing and cosine similarity to analyse text which is rather difficult to compare to our goal. One notebook was found that seemed to have used linear regression and a decision tree classifier on the data but as all the clarifying comments were in Turkish it was rather difficult to understand. It seems that the predicted value was the "votes" attribute and the conclusions suggest that the model which gave the best results was a linear regression model followed by a decision tree [1].

4 Contributions

Table 6: Contributions

Section	**40% or more responsible**
1.1 Regression part a	Kristin
1.2 Regression part b - linear regression	Kristin, Kristo
1.2 Regression part b - baseline	Kristin, Kristo
1.2 Regression part b - ANN	Kristin, Kristo
1.2 Regression part b - Statistical comparison	Kristin
2.1 Classification problem and setup	Andreas
2.2 Classification - logistic regression and baseline	Andreas
2.2 Classification - decision tree	Andreas
2.2 Classification - Model comparison	Andreas, Kristo
2.3 Classification - Statistical comparison	Andreas, Kristo
3.1 Conclusion	Kristin, Andreas
3.2 Previous analysis	Kristin

References

[1] URL: `https://www.kaggle.com/canseltan/moviebox`.