COMP30027 Report

Anonymous

1 Introduction

2 Methodology

2.1 Understanding the Data

2.1.1 Class Distribution

By counting the frequency of each class and understanding the distribution of the classes in the training data, we can gain valuable insights when designing and evaluating models.

2.2 Data Preprocessing

2.2.1 Normalisation

For most of the numeric values in the dataset, the important information is how it compares to other values, rather than the actual value itself. These values should be normalised. Since some models prefer 0 mean data, such as logistic regression and support vector machines, standardisation will be the normalisation method of choice. However, attributes like `title_year` or `average_degree_centrality` are more valuable as the value itself, and so these will not be standardised.

2.2.2 Removing Highly Correlated Attributes

There are a few reasons to remove all but one attribute from a set of highly correlated attributes. Not only does removing the attributes decrease dimensionality and makes the models faster, but having multicollinearity among independent variables can result in less reliable statistical inferences. By removing them, we should be able to increase the models' ability to generalise and not overfit to the training data.

2.2.3 Dropping Unuseful Values

Considering that the data is likely to be biased towards certain regions of the world, there will likely be certain attributes the exhibit an overwhelming majority of one variation, with many other unique but less prevalent values. These attributes would not be very useful to the data, and the best way to approach them is dropping them from the dataset. This also has the additional advantage of decreasing the dimensionality of the data, and hence increasing the speed of the models.

2.2.4 One Hot Encoding

2.3 Baseline Models

To provide a baseline for benchmarking, we can develop a few simple models. The simplest that can be developed is just a model that guesses the most common class. Following this, a One Rule model could be developed on the categorical attributes, such as film categories.

2.4 Model Design

2.4.1 Base Models

Three base models were picked, these were:

- Support Vector Machine
- Logistic Regression
- Random Forest

For each of the models, the below steps were followed to find the best model:

1. The training data was split into a training set and validation set.
2. For each hyperparameter model, a few reasonable options were selected.
3. Iterating over all combinations of hyperparameters, the combination that resulted in the highest accuracy over the training set was picked.
4. The model is evaluated over the validation set.
5. Using sequential forward selection and the best hyperparameters for the model, the best combination of features is selected for the training of the model.
6. Using the best hyperparamters and best features, the model is evaluated against

2.4.2 Ensemble Model

3 Results

3.1 Class Distribution

Figure 1 shows the distribution of the classes in the training dataset. It's clear that a rating of 2 is by far the most common, followed by 3. However, it's only a 61.2% majority, which means a model that always guesses 2 won't have the best accuracy.

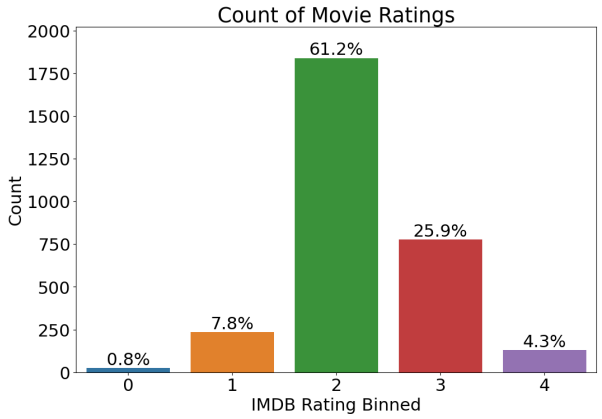

Figure 1: Count plot showing the distribution of classes in the training data

3.1.1 Zero Rule Model

Based of this class distribution, a 0R model should have approximately a 60% accuracy, and this will be used as a baseline for the rest of the models.

3.2 Preprocessing

3.2.1 Unhelpful Features

Since most films are in english and there are many other unique values with very low frequency, the language feature is not very useful and should be dropped. A similar occurence also exists with the country feature, where the USA is an overwhelming majority.

Similarly, a large majority of movies are either R or PG-13 rated, and so the rating of a movie would not be very helpful and can lead to overfitting. As such `content_rating` was also dropped from the dataset.

3.2.2 Names

Through filtering out all of the text attributes, we are also filtering out the names of actors and the movie director. Whilst the quality of film can be correlated to its director, and people often enjoy certain movies with certain actors more, the names alone are not enough to generalise the IMDB rating of a movie. This could instead result in the model memorising which directors and actors are better rated, and since there's a limited number of names, would result in overfitting to the training data.

3.3 Model 1: Support Vector Machine

The best hyperparameters were found to be:
C: 10
Gamma: 0.01
Kernel: RBF
The performance of the model was as below:
Accuracy on validation set: 74.3%
Cross validation score on all data: 69.8%

Class	Prec.	Recall	F1-Score	Support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	48
2	0.75	0.93	0.83	377
3	0.72	0.57	0.63	152
4	0.75	0.47	0.58	19

Table 1: *Classification Report of the Best SVM Model without Feature Selection*

	Prec.	Recall	F1-Score
Macro Avg	0.44	0.39	0.41
Weighted Avg	0.68	0.74	0.70

Table 2: *Summary of Classification Report of the Best SVM Model without Feature Selection*

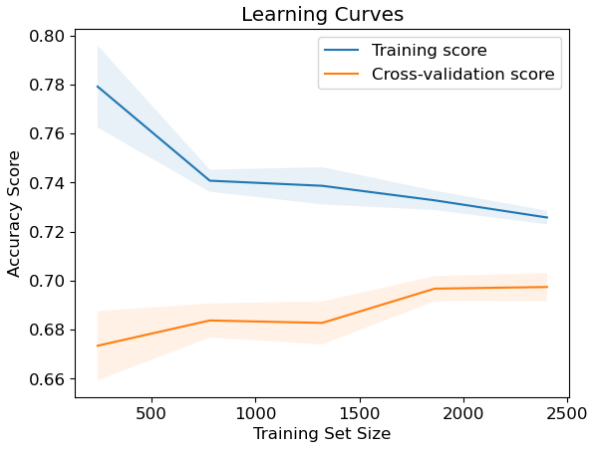

Figure 2: *Learning Curve for the Best SVM Model*

The best feature set found was:

- Number of Critics for Reviews

- Duration
- Director Facebook Likes
- Gross
- Number of Voted Users
- Cast Total Facebook Likes
- Face Number in Poster
- Title Year
- Movie Facebook Likes
- Genres

The performance of the model was as below:
Accuracy on validation set: 73.4%
Cross validation score on all data: 73.8%

Class	Prec.	Recall	F1-Score	Support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	48
2	0.74	0.93	0.83	377
3	0.70	0.54	0.61	152
4	0.69	0.47	0.56	19

Table 3: *Classification Report of the Best SVM Model without Feature Selection*

	Prec.	Recall	F1-Score
Macro Avg	0.43	0.39	0.40
Weighted Avg	0.67	0.73	0.69

Table 4: *Summary of Classification Report of the Best SVM Model without Feature Selection*

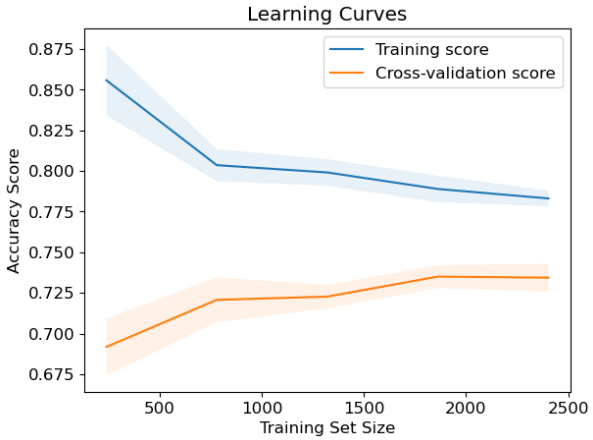

Figure 3: *Learning Curve for the Best SVM Model with Feature Selection*

3.4 Model 2: Logistic Regression

The best hyperparameters were found to be:
C: 0.1
Max Iterations: 1000
Penalty: L2
Solver: Saga
The performance of the model was as below:
Accuracy on validation set: 65.2%
Cross validation score on all data: 62.7%

Class	Prec.	Recall	F1-Score	Support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	48
2	0.65	0.99	0.79	377
3	0.69	0.12	0.20	152
4	0.00	0.00	0.00	19

Table 5: *Classification Report of the Best Logistic Regression Model without Feature Selection*

	Prec.	Recall	F1-Score
Macro Avg	0.27	0.22	0.20
Weighted Avg	0.58	0.65	0.54

Table 6: *Summary of Classification Report of the Best Logistic Regression Model without Feature Selection*

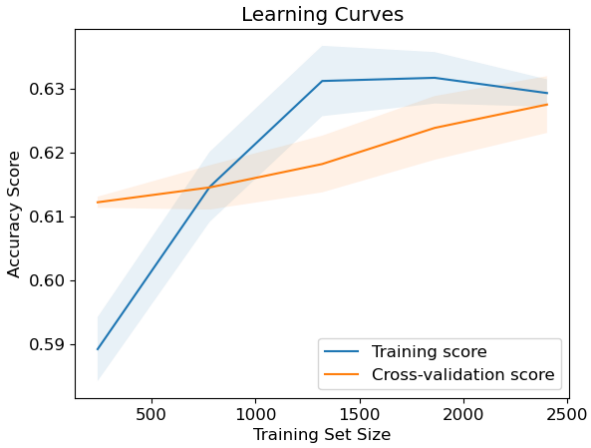

Figure 4: *Learning Curve for the Best Logistic Regression Model*

The best feature set found was:

- Number of Critics for Reviews
- Duration
- Director Facebook Likes
- Gross

- Number of Voted Users
- Cast Total Facebook Likes
- Genres

The performance of the model was as below:
Accuracy on validation set: 72.0%
Cross validation score on all data: 62.9%

Class	Prec.	Recall	F1-Score	Support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	48
2	0.72	0.95	0.82	377
3	0.71	0.44	0.54	152
4	0.89	0.42	0.57	19

Table 7: *Classification Report of the Best Logistic Regression Model without Feature Selection*

	Prec.	Recall	F1-Score
Macro Avg	0.46	0.36	0.39
Weighted Avg	0.66	0.72	0.67

Table 8: *Summary of Classification Report of the Best Logistics Regression Model without Feature Selection*

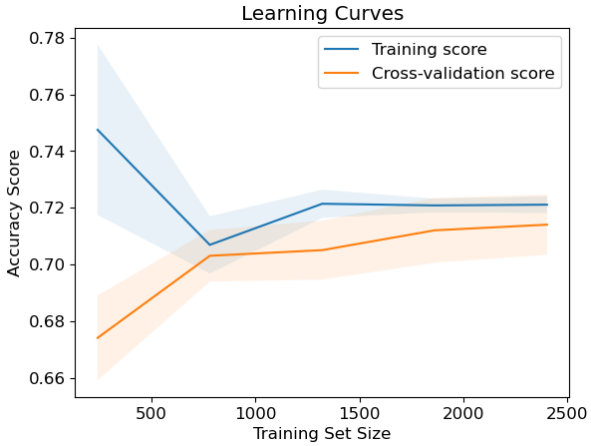

Figure 5: *Learning Curve for the Best Logistic Regression Model with Feature Selection*

3.5 Model 3: Random Forest

The model was seeded using the seed 42.
The best hyperparameters were found to be:
Max Depth: 10
N Estimators: 100
Max Features: Auto
The performance of the model was as below:
Accuracy on validation set: 70.7%
Cross validation score on all data: 71.2%

Class	Prec.	Recall	F1-Score	Support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	48
2	0.72	0.92	0.81	377
3	0.64	0.45	0.53	152
4	0.62	0.42	0.50	19

Table 9: *Classification Report of the Best Random Forest Model without Feature Selection*

	Prec.	Recall	F1-Score
Macro Avg	0.40	0.36	0.37
Weighted Avg	0.64	0.71	0.66

Table 10: *Summary of Classification Report of the Best Random Forest Model without Feature Selection*

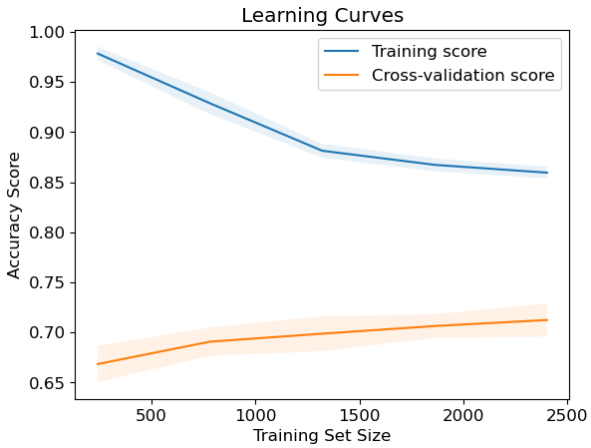

Figure 6: *Learning Curve for the Best Random Forest Model*

The best feature set found was:

- Number of Critics for Reviews
- Duration
- Director Facebook Likes
- Gross
- Number of Voted Users
- Cast Total Facebook Likes
- Title Year
- Movie Facebook Likes
- Average Degree of Centrality

- Genres

The performance of the model was as below:
Accuracy on validation set: 70.4%
Cross validation score on all data: 72.0%

Class	Prec.	Recall	F1-Score	Support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	48
2	0.74	0.94	0.83	377
3	0.69	0.51	0.59	152
4	0.67	0.32	0.43	19

Table 11: *Classification Report of the Best Random Forest Model without Feature Selection*

	Prec.	Recall	F1-Score
Macro Avg	0.42	0.35	0.37
Weighted Avg	0.66	0.73	0.68

Table 12: *Summary of Classification Report of the Best Random Forest Model without Feature Selection*

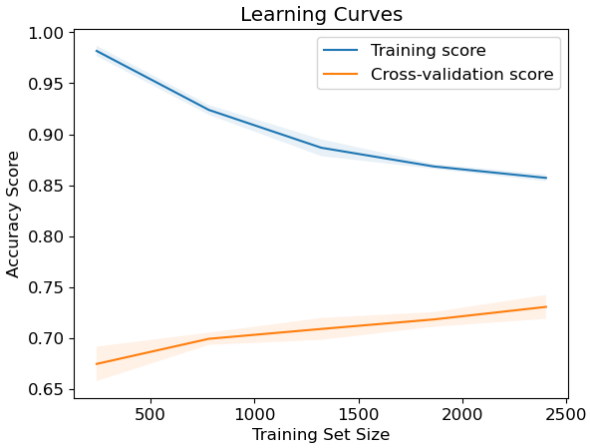

Figure 7: *Learning Curve for the Best Random Forest Model with Feature Selection*

4 Discussion & Critical Analysis

4.1 Overview of Base Models

None of the 3 base models were able to correctly classify movies with rating of 0, and most failed to classify movies with rating of 1. The biggest issue is the distribution of the classes in the training data, since these classes only make up less than 9% of the training data. It is also a lot easier to determine if a movie does really well (rating of 4), as opposed to a movie doing terribly from the provided feature set. Looking at the distribution of all numerical attributes for example, they are all very heavily skewed. Since most if not all of the features are not negatively correlated with the class, the models will have an easier time determining a higher class, whereas with lower ratings they are likely to group it with the most common class. This can be empirically noticed in all the models when looking at the statistic in predicting a rating of 2; the models had high recall, they were able to correctly classify most instances of the rating, but they also had significantly lower precision, indicating that the class was predicted when it was incorrect. This contrasts other class predictions, where the precision was higher than the recall.

4.2 Suppport Vector Machine

However, it seems that with feature selection, the model does not change much. During the feature selection process the only feature that was removed was `average_degree_centrality`, and so it makes sense that that it performs relatively similary.

5 Conclusion

References

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Use **bold** for **emphasis**, but use sparingly.

Short quotations "*are included in the main text, in normal paragraph style, between double quotes and italicized*" All quotes should be properly referenced. Note that the citation style is defined in the accompanying style file; it is similar to AAAI house style. But you may use other (formal) citation styles if you prefer.

You can cite related papers or books like this (Bishop and Nasrabadi, 2006). Text[1] with footnotes at bottom of page.

5.0.1 Subsubsection

Figures should be placed in the text, not at the end. Figures must be captioned and explicitly mentioned in the text .

[1]Footnote text