COMP30027 Report

1. Introduction

This project aims to develop machine learning models for predicting IMDB movie ratings based on various features including director, genres, language, duration, country, facebook likes, and others. The task involves classifying movies into one of five rating bins from 0 (lowest) to 4 (highest). Several supervised learning algorithms are explored: 0R (baseline), Decision Trees and Random Forests. Hyperparameter tuning was performed for Decision Trees and Random Forests. This report presents the methodology, including feature selection, model selection and comparison. It analyzes results using evaluation metrics, provides a critical discussion grounded in theoretical concepts, and conducts error analysis to gain insights into model strengths and weaknesses.

2. Methodology

2.1 Exploratory Data Analysis

Exploratory data analysis was performed on the training data to better understand its characteristics. This included inspecting the distribution of the label variable (IMDB score bins) to check for class imbalance, as well as analyzing the distributions of the various predictor features.

2.1.1 Label Distribution

The table shows a skewed distribution towards rating 2.

Rating	Count	Percentage
0	24	0.8
1	235	7.8
2	1839	61.2
3	777	25.9
4	129	4.3

Table 1- Distribution of IMDB Ratings in the Training Dataset

2.1.2 Feature Distribution

The figure shows that the USA has the most occurrences, followed by the UK.

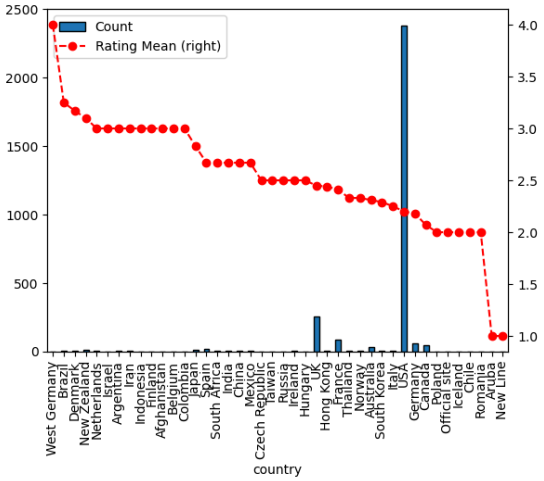

Figure 1- Count and Rating Mean of Movies by Country.

2.2 Feature Selection

We used mutual information in conjunction with a Random Forest classifier to identify key predictors and lower dimensionality. Mutual information is a suitable technique because it measures the dependency between a feature and the label variable. In this case, it was used to compute and rank scores for both numeric and label-encoded categorical features. The effectiveness of these features was assessed using nested cross-validation on the training data.

After feature selection, we retained all top-ranked numeric features shown in Figure 2, and the top three categorical features: country, content rating and language, shown in Table 2.

2.2.1 Mutual Information Score of Top Numeric Features

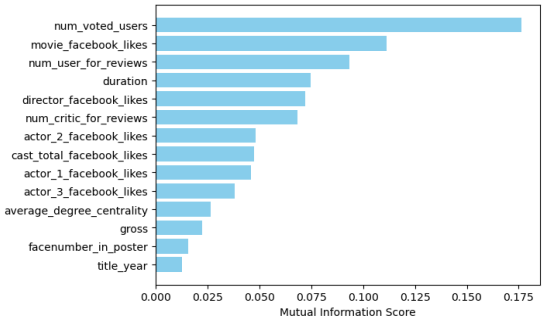

Figure 2- Top 14 Numeric Features Ranked by Mutual Information.

2.2.2 Mutual Information Score of Top Categorical Features

Features	Mutual Information Score
country	0.022
content_rating	0.019
language	0.018

Table 2- Top 3 Categorical Features Ranked from MI.

2.3 Feature Engineering

Some categorical features like country, content rating, and language contained many infrequent values. As a preprocessing step, we replaced infrequent values (those below 5% occurrence) with a new category called "other" to reduce noise and potential overfitting issues.

2.4 Train-evaluation Phase

We conducted a 75/25 holdout split on the preprocessed training data for training and model validation. The train and validation datasets were then standardized for models that were sensitive to the scale of the data.

2.5 Model Selection

To tackle the multiclass movie rating prediction task, we employed a diverse set of machine learning models, each with its unique strengths and characteristics.

2.5.1 Zero Rule (0R)

The 0R classifier serves as a simple baseline method that predicts the majority class for all instances. Its inclusion provides a reference point to assess the performance gains achieved by more sophisticated models.

2.5.2 Decision Trees

Decision Trees are well-suited for movie rating prediction because they naturally handle multiclass classification problems without needing to split the task into multiple binary classifications. This simplifies the modeling process. Additionally, Decision Trees provide clear, interpretable results and can capture non-linear relationships in the data, making them advantageous for understanding the factors influencing movie ratings.

2.5.3 Random Forests

A Random Forest classifier is ideal for movie rating prediction due to its ensemble approach, combining multiple decision trees to reduce overfitting and enhance generalization. It effectively handles both numerical and categorical data, capturing complex relationships within the data. Despite being more computationally intensive than single decision trees, its robustness and ability to provide feature importance insights make it a strong choice for this task.

2.6 Hyperparameter Tuning

Hyperparameter tuning involves selecting the best parameters that control the learning process of machine learning algorithms, balancing learning efficiency and overfitting prevention to enhance movie rating predictions. Although it can improve performance, results are not always guaranteed. We optimized Decision Trees and Random Forests using validation curves.

For Decision Trees, we tuned the maximum depth parameter to control tree complexity and prevent overfitting. Validation curves helped identify the optimal depth, maximizing validation accuracy while minimizing the gap between training and validation scores.

For Random Forests, we optimized the number of estimators (trees) in the ensemble with validation curves. The optimal number was where additional trees did not significantly improve the validation score, indicating the ensemble had sufficiently captured the data's relevant information.

3. Results

3.1 Model Performance

Table 3 summarizes the accuracy of different models on the training, validation, and test sets. Accuracy measures the proportion of correctly classified instances across all classes, making it a suitable choice for this multiclass classification task.

Classifier	Training Accuracy	Validation Accuracy	Test Accuracy
0R	0.6121	0.6125	0.5904
Decision Trees	1	0.5885	0.6383
Decision Trees (**after** tuning)	0.7656	0.6844	0.6383
Random Forests	1	0.7111	0.6835
Random Forests (**after** tuning)	1	0.7204	0.6941

Table 3- Performance of Classifiers.

The 0R model achieved an accuracy of around 61.2% on training and validation datasets, indicating the presence of a majority class in the label distribution, and an accuracy of 0.5904 on the test set. Decision Trees showed perfect training accuracy but lower validation (0.5885) and test accuracy (0.6383). Hyperparameter tuning improved the validation accuracy of Decision Trees to 0.6844, but test accuracy remained unchanged. Random Forests performed better, with a test accuracy of 0.6941 after tuning, up from 0.6835.

3.3 Model Evaluation

Table 4 presents the precision, recall, and F1-score for the classifiers.

Classifier	Precision	Recall	F1-score
0R	0.38	0.61	0.47
Decision Trees (**after** tuning)	0.69	0.68	0.65
Random Forests (**after** tuning)	0.66	0.72	0.68

Table 4- Evaluation Metrics of Classifiers.

The 0R model achieved a precision of 0.38, recall of 0.61, and an F1-score of 0.47. Decision Trees, after hyperparameter tuning, showed improved performance with a precision of 0.69, recall of 0.68, and an F1-score of 0.65. The Random Forest model, also after tuning, demonstrated the best balance between precision (0.66), recall (0.72), and F1-score (0.68).

3.4 Hyperparameter Tuning

3.4.1 Decision Trees

The hyperparameter tuned was the maximum depth. The validation curve in Figure 3 indicates that a depth of 7 was optimal.

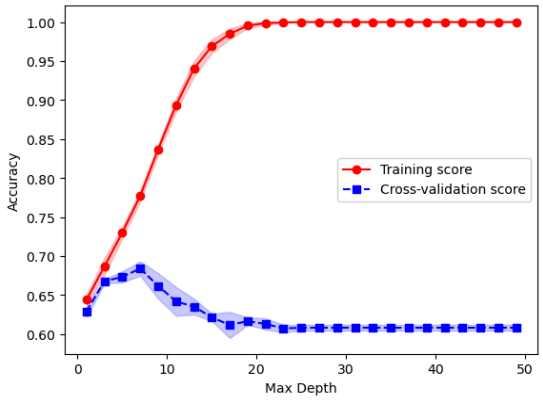

Figure 3- Validation Curve for Decision Trees.

3.4.2 Random Forests

The primary hyperparameter of interest was the number of estimators (trees in the ensemble). The validation curve in Figure 4 suggests that the optimal number of estimators was 330.

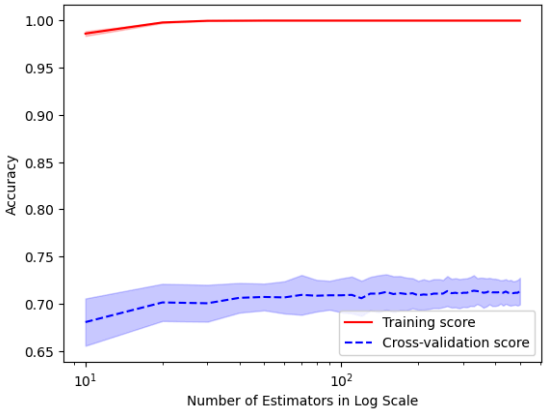

Figure 4- Validation Curve for hyperparameter n_estimators in Random Forests.

4. Discussion and Critical Analysis

4.1 Imbalanced Label Distribution

In a multiclass label distribution, imbalance causes the classifier to favor the majority class, leading to poor performance on minority classes. This results in low precision and recall for underrepresented labels, skewing overall accuracy. The classifier may fail to learn distinguishing features of minority classes, reducing its effectiveness and generalization across all classes (Tarekegn et al., 2021).

4.1 Hyperparameter Tuning

4.1.1 Decision Trees

The validation curve for the Decision Trees show overfitting. As max depth increases, the training accuracy quickly reaches 1.0, indicating perfect fitting of the training data. However, cross-validation accuracy peaks around a depth of 7 and then declines, stabilizing around 0.6. This suggests that deeper trees capture noise in the training data, reducing generalization performance. The optimal depth is around 7 for balanced accuracy.

4.1.2 Random Forests

The validation curve for the Random Forests indicates stable performance across various numbers of estimators. The training accuracy quickly reaches near 1.0, showing the model

fits the training data well. The cross-validation score, however, remains around 0.7, suggesting limited generalization improvement beyond a certain number of estimators. The chosen number of estimators is 330, which appears reasonable, but additional estimators offer diminishing returns. The gap between training and cross-validation scores indicates some overfitting, but it is not excessively severe. Balancing computational cost with performance, fewer estimators might be sufficient.

4.2 Feature Importance

The feature importance analysis from the Random Forest model, which offers valuable insights into the attributes influencing movie ratings, is presented in Figure 5.

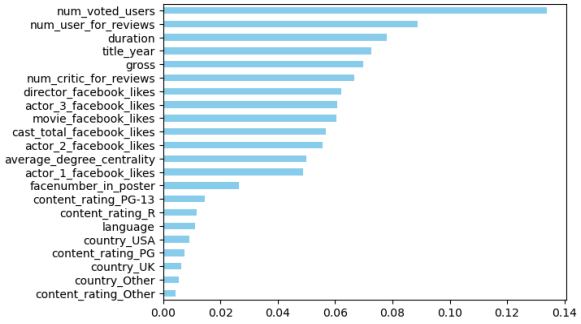

Figure 5- Features Ranked by Feature Importance from Random Forest.

The most influential features are media engagement metrics, such as the number of users who voted and user reviews, indicating audience interest. Financial success indicators, like gross revenue and critic reviews, also significantly impact ratings. Social media popularity, including likes for directors and actors, highlights the influence of online engagement. Moderately influential features include movie characteristics like duration and release year, reflecting audience preferences and rating trends. Less influential features, such as the number of faces in the movie poster, content ratings and countries, still contribute by reflecting certain aspects of audience appeal and cultural preferences. These insights suggest that a predictive model should mainly incorporate social media metrics and consider time-related changes to get a more accurate prediction.

4.3 Error Analysis

4.3.1 0R

The weighted average score of the evaluation metrics was relatively low, indicating that the model did not perform satisfactorily when considering class imbalance. The model was biased towards the majority rating (rating 2) while struggling with the minority ratings (0, 1, 3, 4), implying that the features did not make an effect on this model.

4.3.2 Decision Trees & Random Forests

Decision Trees and Random Forests were compared to understand their performance in predicting movie ratings. From the confusion matrix in Figure 6, the Decision Trees model shows reasonable performance for class 2 but struggled significantly with classes 0, 1, 3, and 4. Most misclassifications occurred between classes 2 and 3, as well as between classes 3 and 4.

An interesting observation was that Decision Trees showed the same test accuracy (0.6383) before and after hyperparameter tuning, despite a validation accuracy improvement from 0.5885 to 0.6844. This suggests tuning enhanced validation performance but did not improve test generalization, possibly due to overfitting on the validation set. Further exploration of tuning parameters or alternative validation techniques may be needed to improve test performance.

4.3.2.1 Confusion Matrix of Decision Trees

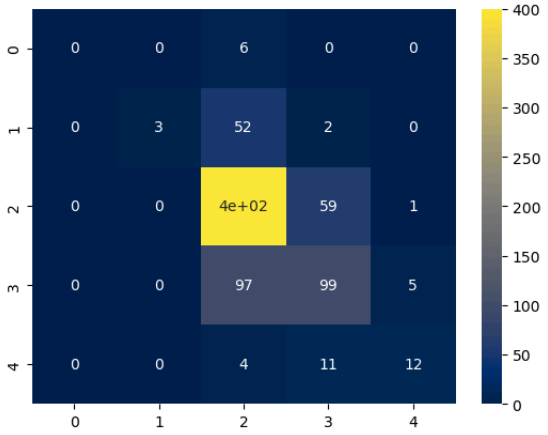

Figure 6- Confusion Matrix of Decision Trees after Tuning on Validation Set.

On the other hand, the Random Forests model, after hyperparameter tuning, achieved a higher

test accuracy of 0.6941. Unlike Decision Trees, Random Forests benefit from an ensemble approach, which combines multiple decision trees to reduce overfitting and enhance generalization. This ensemble method allows Random Forests to capture more complex patterns in the data, improving overall performance. However, similar to Decision Trees, it also faced issues with class imbalance. The classification report (Figure 7) shows precision and recall scores of 0 for classes 0 and 1, indicating a complete failure to identify these classes, likely due to their fewer samples. The confusion matrix (Figure 8) confirms significant misclassifications, especially with classes 0, 1, and 3 often incorrectly predicted as class 2, highlighting the model's tendency to overfit to the majority class. Applying techniques such as SMOTE or adjusting class weights might overcome imbalance in multi-label classification, and further hyperparameter tuning could enhance performance for minority classes.

4.3.2.2 Classification Report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	6
1	0.00	0.00	0.00	57
2	0.73	0.93	0.82	460
3	0.69	0.51	0.58	201
4	0.75	0.44	0.56	27
accuracy			0.72	751
macro avg	0.43	0.38	0.39	751
weighted avg	0.66	0.72	0.68	751

Figure 7- Classification Report of Random Forests after Tuning on Validation Set.

4.3.2.3 Confusion Matrix of Random Forests

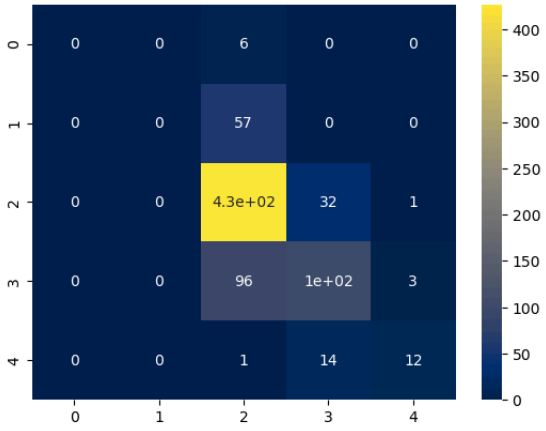

Figure 8- Confusion Matrix of Random Forests after Tuning on Validation Set.

5. Conclusion

In this project, we developed and analyzed three machine learning models: 0R, Decision Trees, and Random Forests, with the aim of predicting IMDB movie ratings. Among these models, Random Forests demonstrated superior performance in comparison to both Decision Trees and 0R. Our analysis identified social media metrics and time-related attributes as crucial predictors of movie ratings. However, we encountered significant challenges due to imbalanced label distribution, which notably affected the identification of minority classes. For future improvements, the focus should be on addressing class imbalance through advanced techniques such as SMOTE or by adjusting class weights. Additionally, further hyperparameter tuning could enhance the overall performance of the model.

6. References

Tarekegn, A., Giacobini, M., & Michalak, K. (2021). A Review of Methods for Imbalanced Multi-Label Classification. Pattern Recognition, 107965. https://doi.org/10.1016/j.patcog.2021.107965

[Pure Word Count: 1865]