Comparing ML Models to Predict Movie Age-Certifications

Elements of Data Processing (COMP20008_2023_SM2)

Word Count: 2974

Du-Simon Nguyen, 1352062

Bill Tran, 1355921

Shiyoun Kim, 943020

Executive Summary

This study aimed to discover which features lead to the most effective model in predicting the age certifications in a dataset of movies. Model 1 was developed using movie descriptions by preprocessing into a bag of words, applying TF-IDF, reducing features using TruncatedSVD and using Random Forest to train the dataset. Model 2 used many different features (genres, country, year, type and director) which was reduced using PCA and KNN to construct the prediction model. It was found that movie descriptions in Model 1 did not provide a convincing prediction of age certifications, whereas Model 2 showed better performance. Neither model were able to consistently predict 'G' age certifications which limited performance metrics. From this investigation, it is recommended that: more complex techniques and hyperparameter tuning should be utilized for a more rigorous model, a larger dataset should be acquired for better training and testing, and further investigation conducted into ways to improve accuracy of the 'G' age certification category.

Introduction

This report explored the relationships between a variety of given features and age-certifications of movies and TV-shows in order to predict missing age-certification values using two classification models: Random Forest and k-Nearest Neighbours (KNN). Random Forest is a supervised classification ML algorithm that uses a collection of decision trees to predict our target label. KNN is another classification model which operates on the premise that similar instances will be located close to each other in space and classifies instances according to the k nearest data points as determined in the training dataset. The driving motivation behind this research comes from the importance of providing accurate content advisory information to the audience to enhance their experiences. Furthermore, watching each movie and then deciding age certification of that film could be costly and time consuming. Thus, by using ML we hope to build a model that can accurately predict age certification of movies without watching them.

Data

Two datasets, “titles.csv” and “credits.csv” were used to conduct our analysis. These contained detailed information about movies and TV shows on Netflix. “titles.csv” contained 5,850 different films, represented as a combination of 15 different features such as id, description, runtime and imdb score. “credits.csv” contained 77,801 rows where each record was an actor or director with their respective role in a film in “titles.csv”. Furthermore, in this dataset, there were over 50,000 different actors and over 3,000 different directors.

Methodology

Data preprocessing:

As both of our models used different features to classify unknown age certifications in the dataset, we first had to shrink our dataset such that every title record contained a valid value in each of these features (i.e. was not missing or unknown). For “credits.csv”, this involved removing any records where the person’s role was not “DIRECTOR”, as well as only extracting the relevant columns (i.e. “person_id”, “id” and ”name”) for space and time efficiency. For “titles.csv”, preprocessing involved removing any titles that do not have all the relevant features that will be used in our modelling stage; these features included "description", "genres", "production_countries", "release_year", and "type". This process was necessary to ensure that we use the same datasets on our two models, or else the different data distributions may potentially lead to inaccurate comparisons. We then joined two datasets on movie ID. Additionally, in “production_countries”, we found that all countries were represented as abbreviations except for one instance of “Lebanon”. Thus, we manually changed it to “LB”.

Finally, we standardised the age_classification values into more general labels based on domain knowledge. For example, PG-13 was equivalent to PG, and lesser-known classifications such as TV-14 were also mapped to PG. Other than standardisation purposes, this step also aimed to generalise our model. This is because having too many similar classifications may result in an overly complex model. This standardisation is represented as a dictionary called “cert_map” with keys being the set of unique age_classifications and values the more general label we map the key to.

```
certs_map = {'R': 'R', 'G': 'G', 'NC-17': 'R', 'PG-13': 'PG', 'TV-14': 'PG', 'TV-Y7': 'PG',
             'TV-Y': 'G', 'TV-MA': 'MA', 'TV-G': 'G', 'PG': 'PG', 'TV-PG': 'PG', 'MA': 'MA'}
```

Both model 1 and model 2 used cross-validation with n=10 folds to evaluate their model.

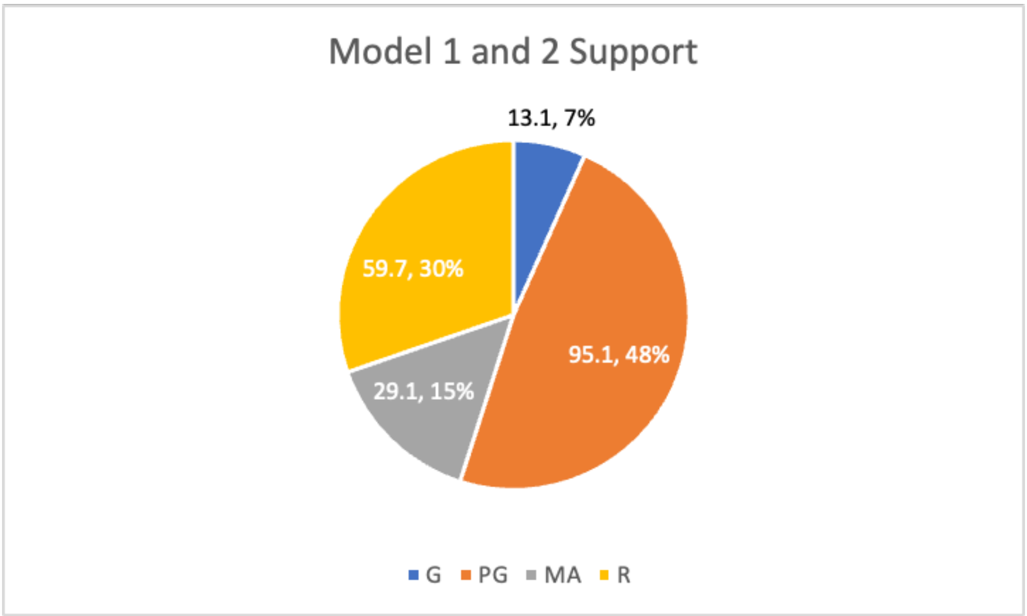

Figure 1. Average support proportions for the 10 test folds (n=197 instances per fold) used in cross-validation to test train respective models.

Model 1: Random Forest

Model 1 is a supervised model that aims to predict the age certifications in our dataset using the "description" feature. Using various nltk and sklearn libraries, we first processed "description" into a bag of words (BoW), implementing various text preprocessing techniques including lower-case-folding, removing stop words, lemmatizing and tokenization. These text preprocessing steps were aimed to reduce the sparsity of the words as well as the occurrence of unhelpful closed-class words for more accurate classification. Instead of analysing the raw counts of the tokens in each description, we transformed the BoW into TF-IDF values. This involved assigning a normalised weight to each token which would assist in classifying descriptions based on the importance of its tokens. With this sparse TF-IDF matrix, we perform feature extraction and reduce our features to three components using TruncatedSVD - a dimensionality reducing technique that works well with sparse matrices, especially on TF-IDF matrices (1). Initially, we considered PCA to be standardized with model 2, however, our TF-IDF matrix was simply far too sparse (9,678 features) for it to be a valid method.

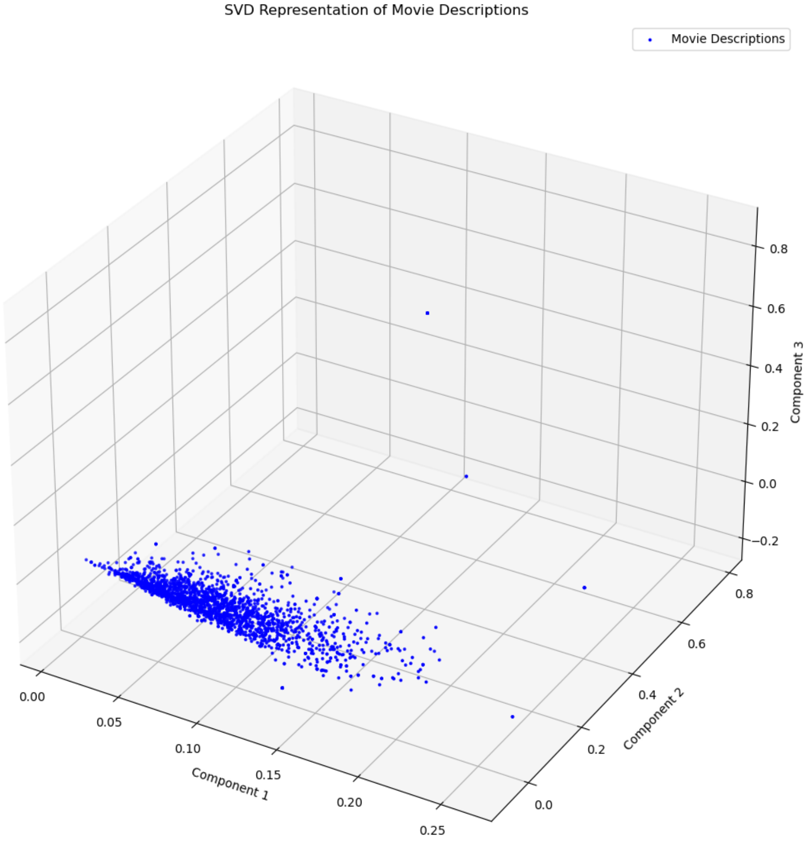

Figure 2. A scatter plot of the SVD representation of Movie Description using k=3 components taken from 1 of the n=10 folds in cross validation. Variance explained by each component=[0.00165258 0.00435035 0.00398565]. Total variance explained = 0.01.

With our processed SVD representation of title descriptions, we used RandomForest to evaluate our data using cross-validation with n=10 splits. In order to prevent feature leakage, we processed the description of each train and test dataset into their corresponding SVD matrices within each training step iteration. This is because, if we pre-compute it all before evaluation, then our train datasets will have access to specific words/descriptions that only appear in the test dataset, leading to feature leakage.

Model 2: KNN

In Model 2 we used Principal Component Analysis (PCA) to reduce the dimensionality and KNN method to construct a model that predicts age_certification. We first, split the dataset into our merged (train + validation) and test data using a simple holdout 80:20 split, in order to prevent feature leakage. We chose the number of neighbours to include in our model by implementing the 10-fold cross validation method on our merged dataset with three principal components. We chose this method because it allowed us to utilise our training dataset efficiently by allowing the merged dataset to train and validate our model. We then calculated the accuracy of each potential number of neighbours and selected the one with the highest accuracy level.

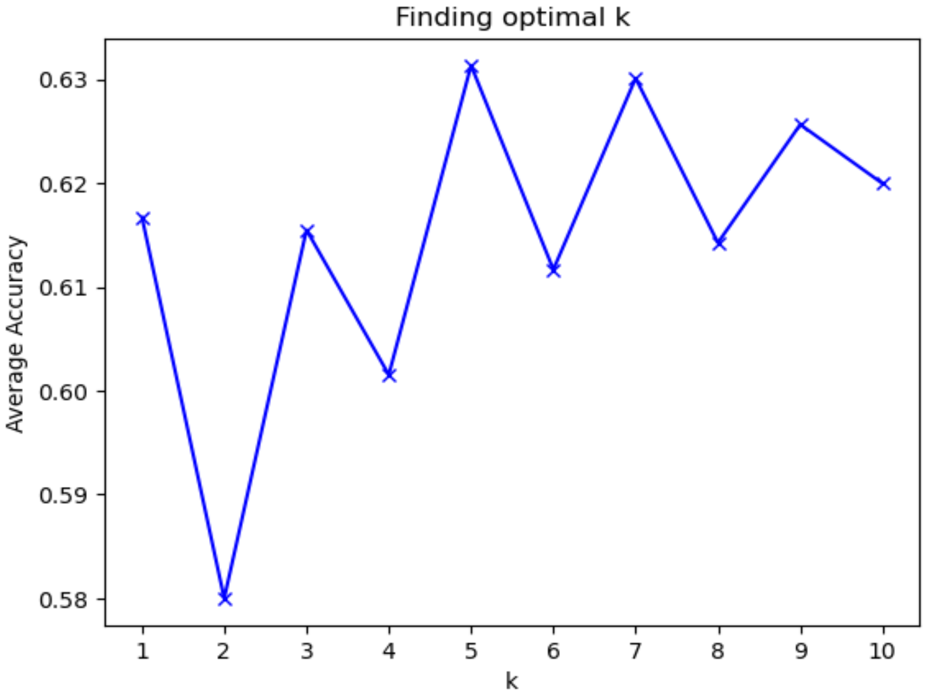

Figure 3: Hyperparameter tuning to find optimum number of neighbours with k=3 components

From Figure 3, we can observe that the accuracy of our model is at the highest (63.1%) when the number of neighbours for our model is at 5. Therefore, in our Model 2, to predict age_certification of a movie, our KNN looks at age_certification of the 5 closest data points based on their principal component values and uses majority rules to predict the age certification of the movie.

In detail, at each fold we created dummy variables for each category of genres, production countries, release year, type and director in the training data at that fold (to prevent feature leakage). It is important to note that we used the director's IDs instead of their names to reduce the possibility of identifying a director as two separate persons due to spelling mistakes. After this, we standardised each feature value and calculated principal components. It was crucial for us to standardise the values because the difference in magnitude of scales between features was significant. For example, in our training/validation dataset, all the dummy variables are measured in 0 and 1 but the values for release years were between 1966 and 2022. Thus, if we did not standardise the values then we would have overweighted towards the release years in our principal component calculation. Also, at each fold three components captured more than 95% of variations.

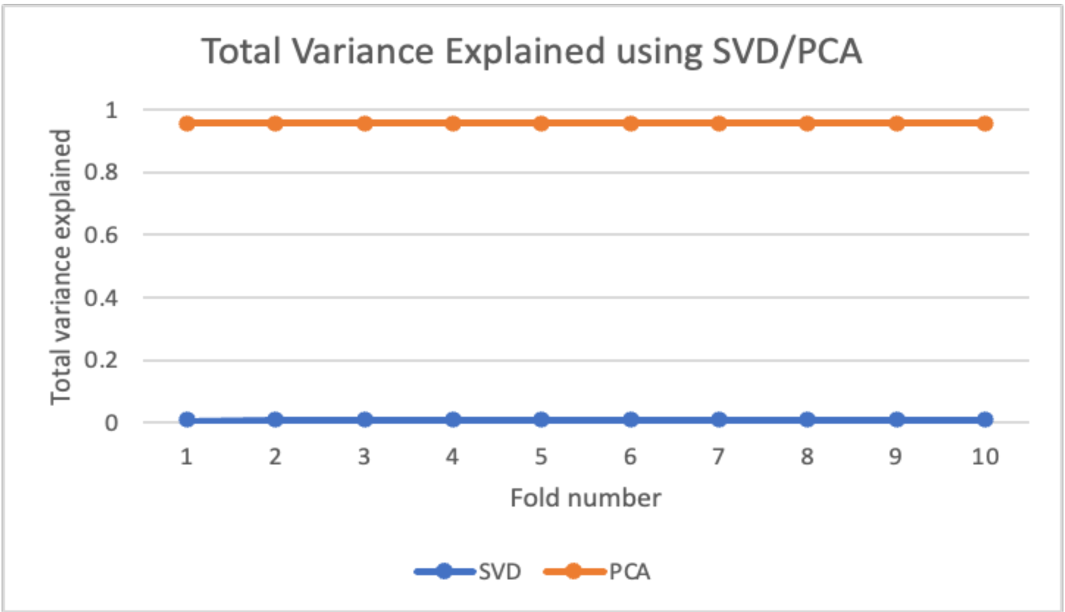

Figure 4. Variance explained for each n=10 folds. SVD was used to reduce dimensions for Model 1, PCA was used to reduce dimensions for Model 2. Both were reduced to k=3 components. Average variance explained PCA=0.9563. Average variance explained SVD=0.0102.

Results - Interpretation and Findings (Test dataset)

The two models had varying levels of success. Following the testing and training of each model, classification reports and confusion matrices were generated per test fold, providing metrics such as recall, precision and accuracy for analysis. An aggregate measurement was then taken from these 10 folds for each model for analysis.

Looking solely at accuracy, Model 1 finds a mean accuracy of 0.41 while Model 2 has an average accuracy of 0.6426. This is better than the baseline Zero-R between the four potential age ratings which would yield 0.25 accuracy.

Model 1 tends to predict 'PG' for its test cases, which also happens to be the most populous category in the dataset as seen from Figure 1. From Figure 5, 'PG' has an average precision and recall of 0.491 and 0.662 respectively, which is significantly higher than the other certifications in Figure 5. In fact, classification 'G' has a low average recall and precision of 0.024 and 0.055 respectively. As such, it is questionable whether this model would perform well for another dataset with lower proportions of 'PG' movies. The model also gives a very low average variance explained across its folds of 0.0102, indicating that the 'description' feature by itself may not be enough to produce a strong model.

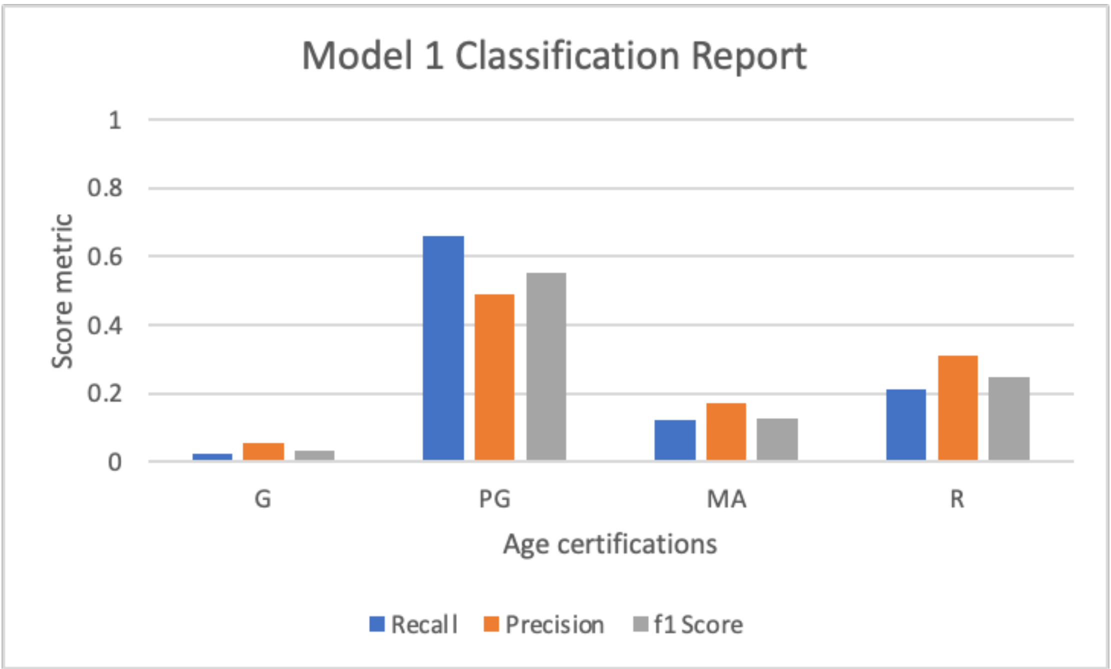

Figure 5. Average recall, precision and f1 scores for the 10 folds used in cross-validation to train our RandomForest model

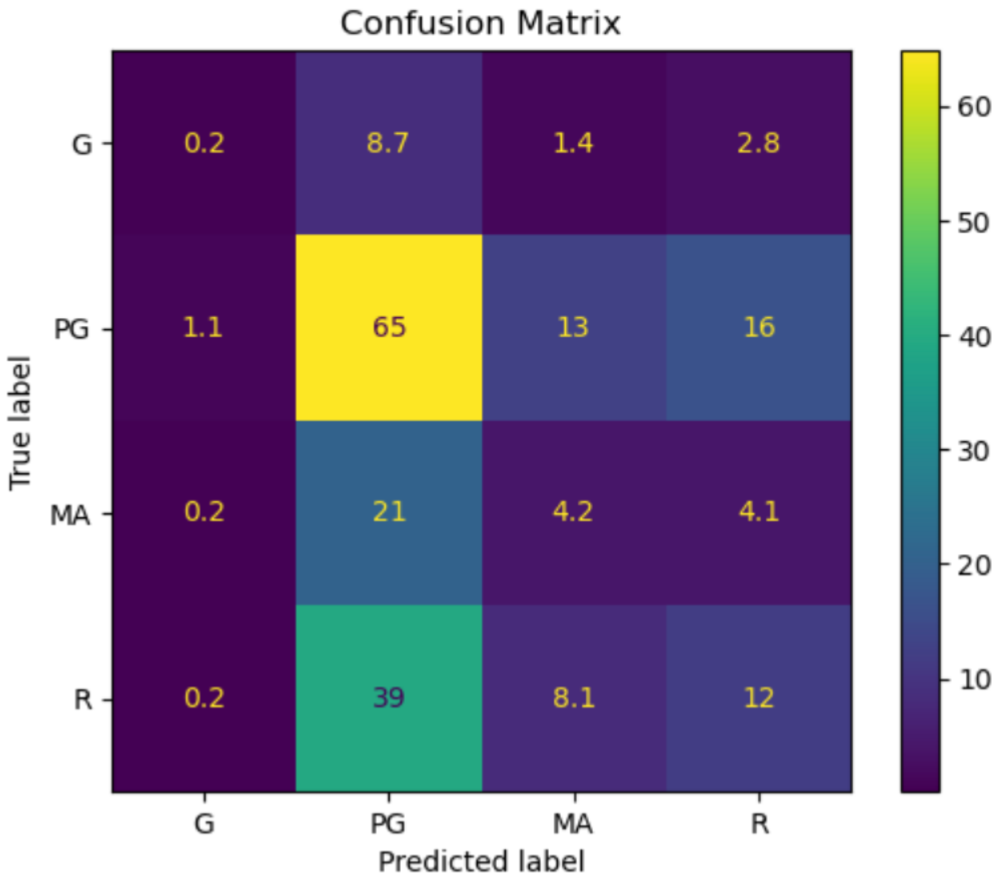

Figure 6. Average confusion matrix for the 10 folds used in cross-validation to train our RandomForest model

In Model 2, precision, recall and f1-score are high for each age category with the exception of classification ‘G’, shown in Figure 7. The model has an average variance explained of 0.9563, an indicator that the model might have good predictive power. This may be because more features are being used, so more information is available to explain the variability of the target variable. However, any accuracy for movies with classification ‘G’ remains low. This is because for movies with ‘G’ certification, the model has a tendency to predict ‘PG’ instead, as seen in Figure 8. This is more desirable than the inverse case as there is less risk of children being exposed to mature themes. For this model, ‘MA’ tends to have the highest predictive power despite being the second least populous in the test set, indicated in Figure 1.

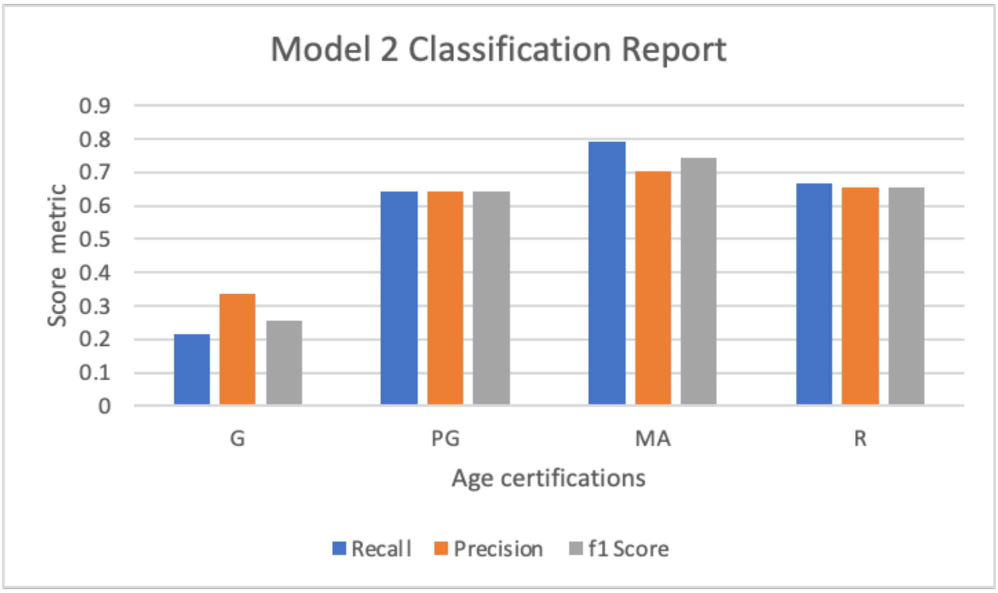

Figure 7. Average recall, precision and f1 scores for the 10 folds used in cross-validation to train our KNN model.

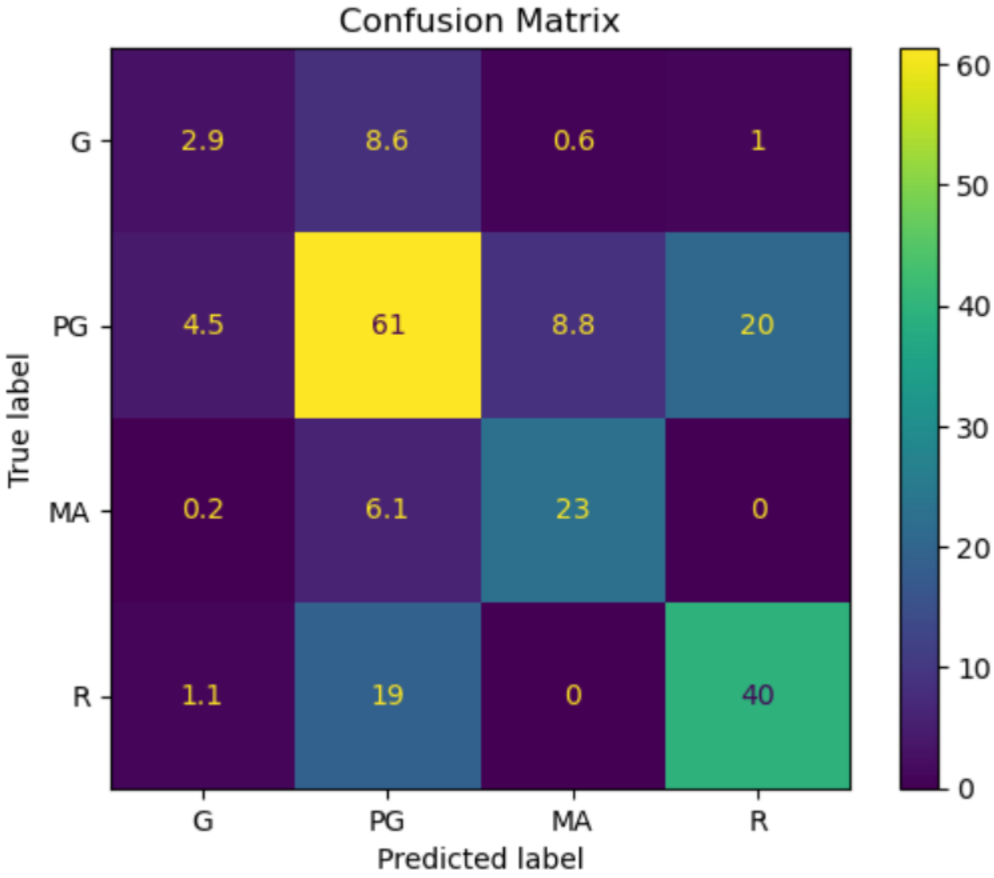

Figure 8. Average confusion matrix for the 10 folds used in cross-validation to train our KNN model.

A common point of error for both models is with age certification 'G'. One possible cause of this is an insufficient number of movies with 'G' rating in the dataset, so training cannot develop a strong enough model for predictions of this classification, as shown in Figure 1 where the average support for 'G' are 13.1 instances. Another reason may be because there are not many clear distinctions between different age ratings. The 'G' rating in general does not have many discriminative features that help distinguish it from other groups. For example, using a movie description (as in Model 1) a human evaluator would not have a list of words that help define the 'G' category, but rather investigate the absence of keywords which are more common with higher age ratings such as 'guns' or 'crime'. Our Model 1 can't capture this method of evaluation, because for every description there will be an almost endless number of words that aren't present in other movie descriptions. However, even this method is not fail-safe since words such as 'suicide' or 'terrorist are sometimes found in movies with 'G' certification as well. This lack of distinction is also reflected in Model 2, where 'G' movies often get incorrect predictions of 'PG', as seen in Figure 8.

From the results and metrics of each method, we conclude that Model 2 has greater reliability in predicting the age certification of movies. As 'PG' is the most populous age certification at 48.27%, Model 1's tendency to skew its prediction towards 'PG' casts doubts on its functionality if it were to be used in other datasets. As such, the actual accuracy of the model may be lower than shown in the provided metrics. On the other hand, Model 2's predictions are more evenly distributed, mainly across the highest three age ratings of 'PG', 'MA' and 'R'. The major drawback for both models is the inability to predict 'G' classifications, which influences accuracy.

Limitations and Improvements Opportunities

There are a considerable number of limitations with our methodology. In our preprocessing steps, we shrink the dataset by a large amount due to the absence of many age_certification values as well as values of the features used to predict our target label. Therefore, we may not have enough information for our classifiers to build accurate prediction models. For example, the raw dataset initially contains 5850 titles, after preprocessing we were left with 1970 valid entries which we must further split into training and testing data, as the rest of the titles have either unknown age_certifications, or have invalid values for the features we use to build the model. The former shrinkage is inevitable, as it is the label we are trying to predict. However, the latter reason can be addressed by using imputation techniques which would help retain more valid instances.

The biggest limitation of our research would have to be its credibility, which is demonstrated by low performance metrics for both Model 1 and 2. Model 1's poor performance could have been foreseen in Figure 3, with the SVD scatter plot representation of the Movie Descriptions as well as the extremely low variability explained by the three components. We can observe that the majority of the data points are clustered into a group with high overlap, even with small dot size, and there is not much distribution. This may explain Random Forest Classifer's poor ability to classify each instance into the correct age_certification. Hence, giving an average of 43% accuracy and low recall and precision for each category. This may be due to the sheer sparsity and high dimensionality of the TF-IDF matrix. Which resulted in high information loss when we reduced it to three components. Thus, we could potentially improve this method by simply increasing the number of components, but this would be harder to visualize. Additionally, rather than simply transforming movie descriptions into BoW->TFIDF->SVD, we apply more complex techniques. For example, we could extract phrases instead of tokens and use sentiment analysis of the descriptions for better classification. Finally, we could also hyperparameter tune the Random Forest classifier to find the optimum number of decision trees, as well as the depth of each tree, similar to how in Model 2 we hyperparameter tuned kNN to find the optimum number of neighbours.

On other hand, in Model 2, with just three components of PCA we captured over 95% of the variation in the datasets but the accuracy of our prediction was around 65%. This could be a result of overfitting the model to our training data. In other words, implementing PCA was not enough to overcome noises created from including features like director and production countries that have high variations within each feature. For example, in our training dataset there were 1,576 different movies/TV shows and 1,355 different directors. This shows that each director only filmed one or two movies, meaning that most directors in our training dataset were different to the directors in our testing dataset. In detail, only 84 out of 394 films in our testing data were filmed by directors in our training data. As a result of this, by including the variable director as one of our variables, it overfits the model to our training dataset and creates noise in our prediction. Thus, we potentially could improve Model 2 by simply removing features that have high variations. Otherwise, we could gather additional data that are still relevant in predicting age_certification and that have a lower variance (e.g. production company).

Conclusion

In this report we analysed the effectiveness of using different characteristics and features of a movie to develop machine learning models that predict a film's age certification. We found that a description of the movie, although intuitively powerful, cannot make a strong prediction on its own using our techniques and preprocessing. Alternatively, a combination of many features (i.e. genre, country, release year, type and director) showed much more promising predictions. However, an unresolved issue is that we could not accurately predict age certifications 'G', a factor which strongly hampered the performance of the two models presented. Furthermore, there are limitations in our methodology in preprocessing and model development which further affected our prediction accuracy, and it remains to be seen if our models could work in other datasets with different age certification distributions. Despite this we were able to establish that there is merit to using machine learning models to predict unknown age certifications and with further improvements, as suggested in our limitations, stronger models could be developed in future investigations.

References

1. sklearn.decomposition.TruncatedSVD — scikit-learn 1.3.1 documentation
2. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
5. Ed Workshops
6. Lecture slides