Comparing ML Models to Predict Movie Age-Certifications

Du-Simon Nguyen, 1352062

Bill Tran, 1355921

Shiyoun Kim, 943020

Introduction

- Aim
  - Find best Machine Learning algorithm model to accurately predict age certification of movies and TV-Shows
- Motivation
  - Providing accurate content advisory information
- Models:
  - Random Forest & K-Nearest Neighbours

Methodology

Data preprocessing

- Feature selection
  - Data shrinkage

- Data linkage
  - movies.csv to titles.csv on MOVIE_ID

- Standardising Age Certifications

```
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'}
```

Testing datasets

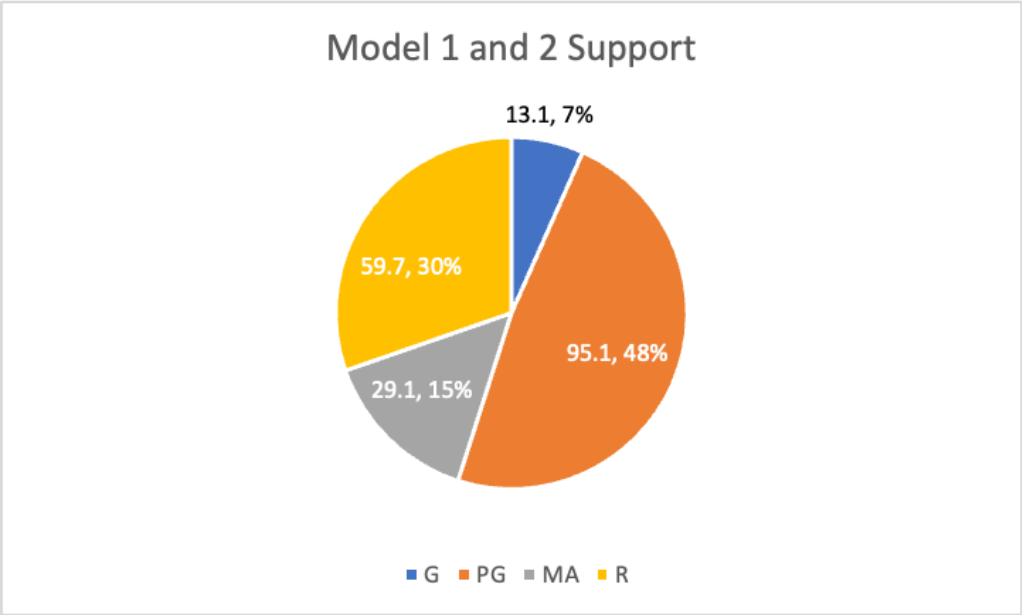

Figure 3. Average support proportions for each testing fold in the 10 partitions (n=197 instances per fold) used in cross-validation to train respective models.

Model 1 - Random Forest

- Description -> Text Preprocessing -> BoW -> TF-IDF -> SVD -> Random Forest to evaluate

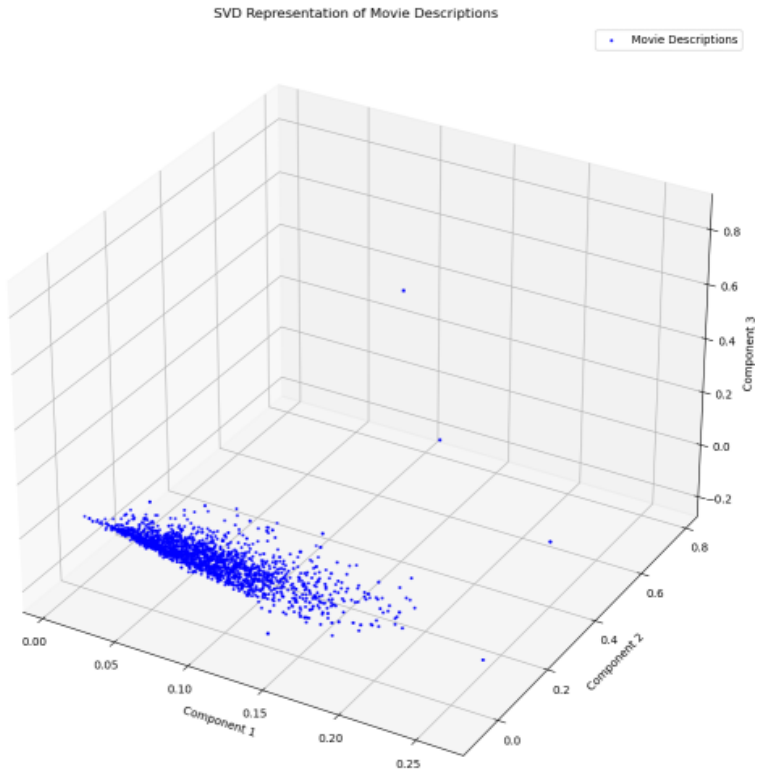

Figure 1. A scatter plot of the SVD representation of Movie Description using k=3 components taken from 1 of the n=10 folds in cross validation. Variance explained by each component = [0.00165258 0.00435035 0.00398565]. Total variance explained = 0.01.

Model 2 - k Nearest Neighbours

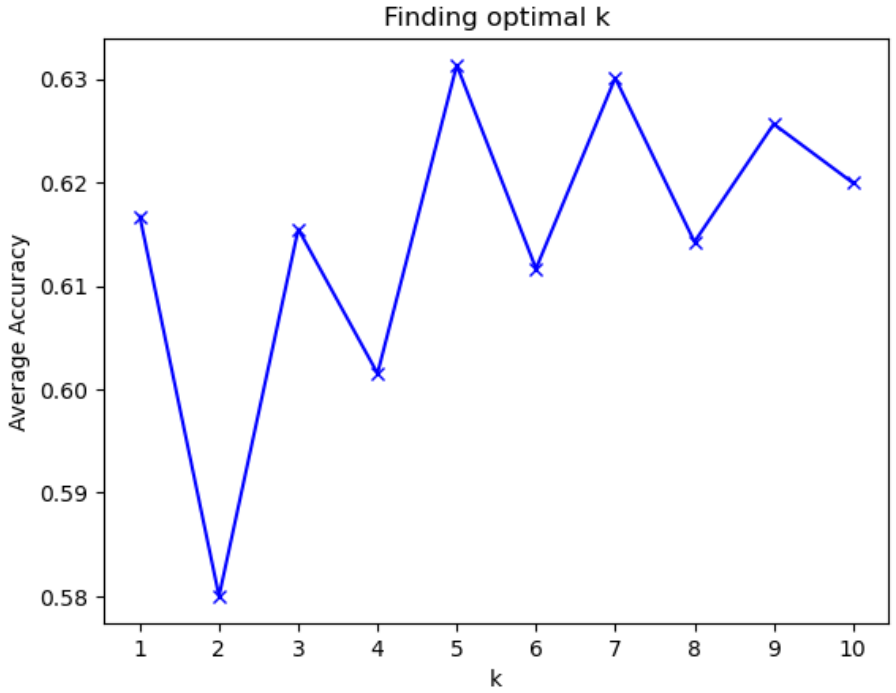

Figure 2: Hyperparameter tuning to find the accuracy of each potential number of neighbours with k=3 principal components.

- Hyperparameter tuning using 10 - fold Cross Validation
- PCA -> kNN to evaluate
- Features used:
  - Genres, Production Countries, Release year, Type and Director

Feature Extraction

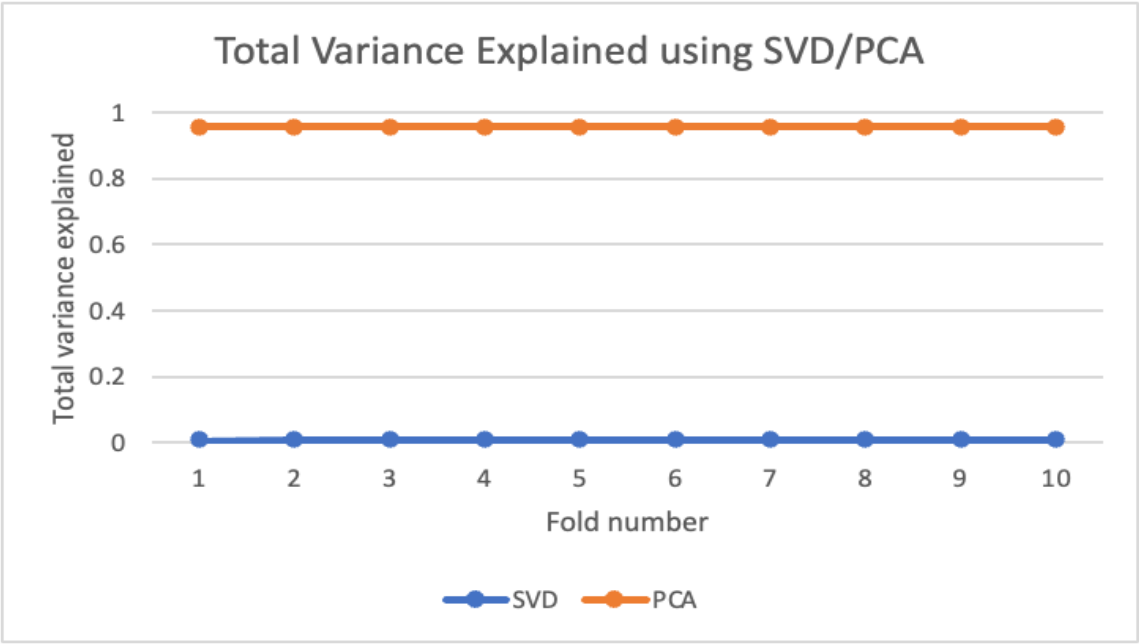

Figure 4. Variance explained for each n=10 folds. SVD was used to reduce dimensions for Model 1, PCA was used to reduce dimensions for Model 2. Both were reduced to k=3 components. Average variance explained PCA = 0.9563. Average variance explained SVD = 0.0102.

Results and Analysis

Model 1 vs Model 2

Accuracy: 0.41

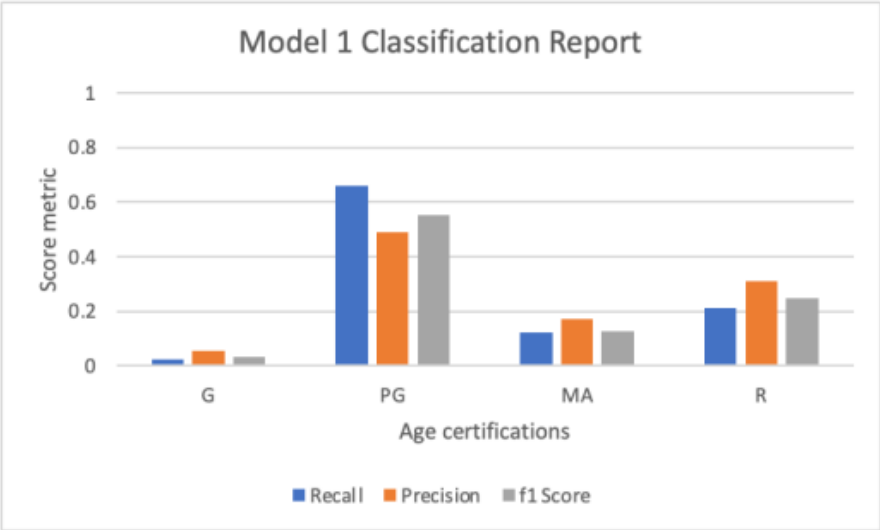

Figure 5. Average recall, precision and f1 scores for the 10 folds used in cross-validation to train our Random Forest model

Accuracy: 0.6426

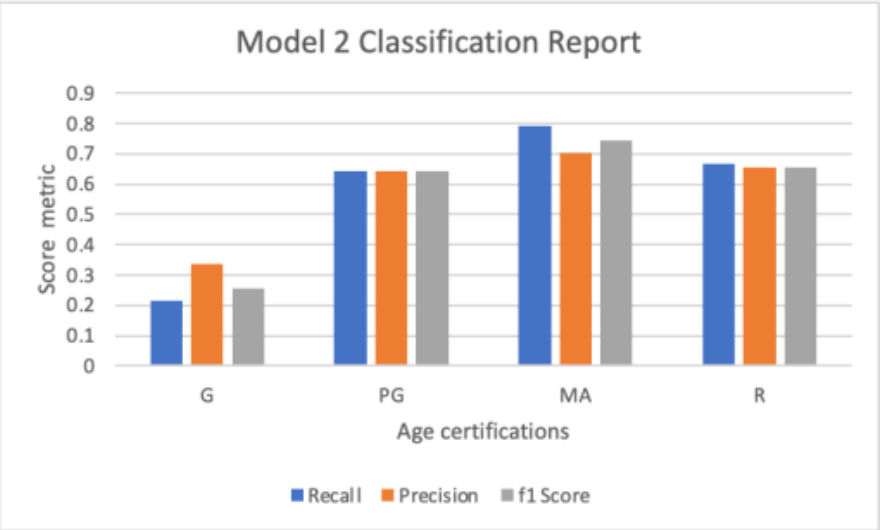

Figure 7. Average recall, precision and f1 scores for the 10 folds used in cross-validation to train our kNN model

Model 1 vs Model 2

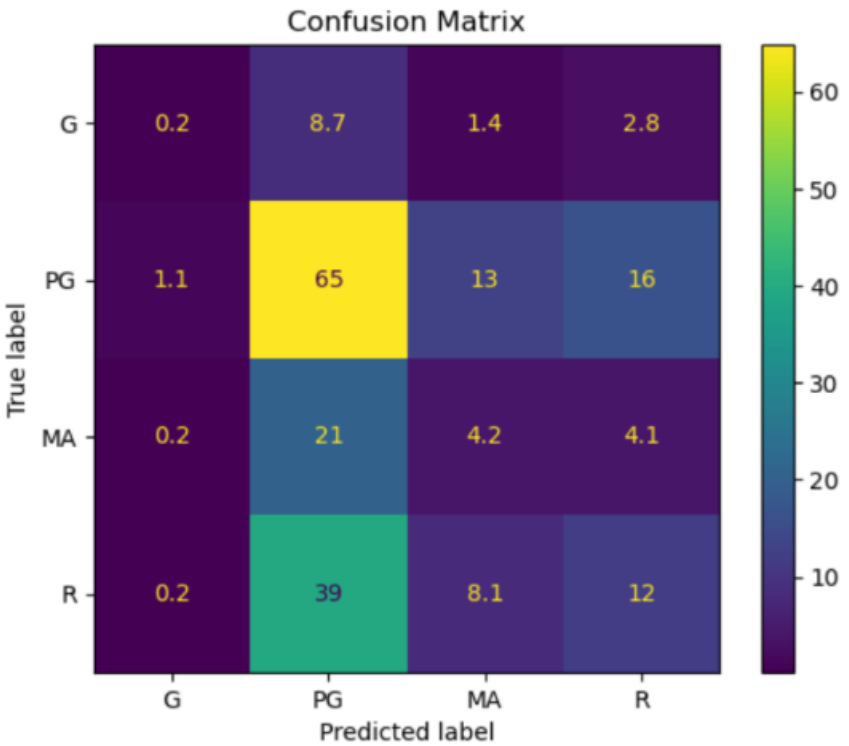

Figure 6. Average confusion matrix for the 10 folds used in cross-validation to train our Random Forest model. n=197 instances.

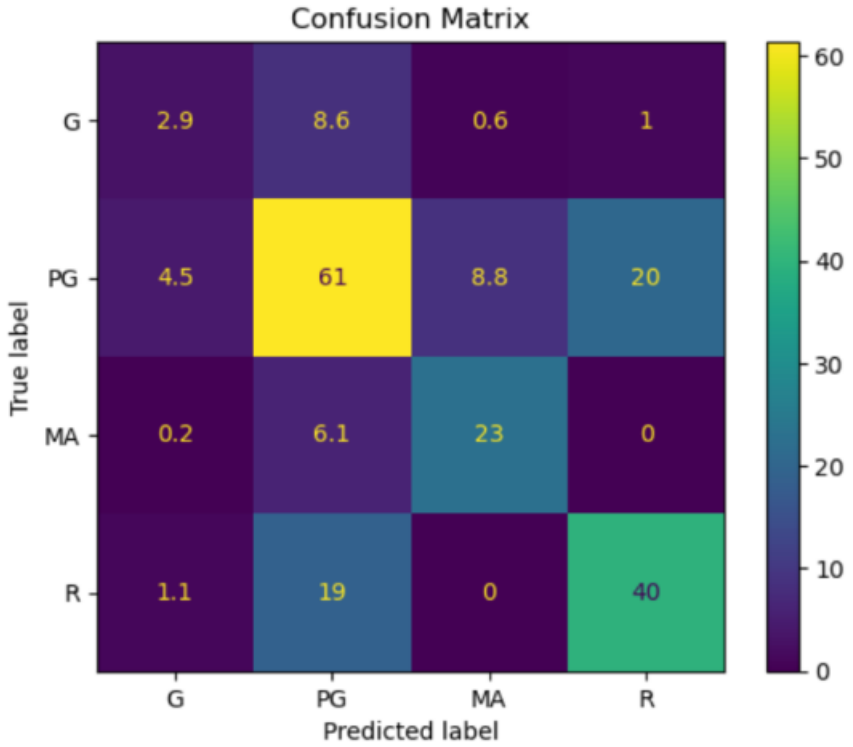

Figure 8. Average confusion matrix for the 10 folds used in cross-validation to train our kNN model. n=197 instances.

Why is ‘G’ accuracy so low?

- Not enough ‘G’ certifications in the dataset
  - Model underfitting

- Difficulty in finding discriminative features that distinguish from other age ratings
  - What words in description would indicate that the movie/show is rated ‘G’?
  - Differentiate between ‘G’ and ‘PG’?

Conclusion

- Limitations
  - Imbalanced distribution
  - Preprocessing - too much data shrinkage
  - Model 1 - virtually no variability explained
  - Model 2 - too many features leading to noise
- Improvements
  - Using another dataset to balance distribution
  - Preprocessing - various imputation techniques
  - Model 1 - hyperparameter tuning our Random Forest
  - Model 2 - removing features with high variations

Thanks for listening :)

Any questions?