I started with over 6000 features. I extracted 27 – approximately the square root of 700, the number of scripts – using SVD and tested three models.

The best hyperparameters for the random forest:

|  |  |
| --- | --- |
| number of trees | 120 |
| max-depth | None |
| criterion | ‘Gini’ |

The best hyperparameters for the gradient booster:

|  |  |
| --- | --- |
| learning rate | 0.05 |
| number of estimators | 250 |
| max. features | 10 |
| max. depth | 2 |

There are no hyperparameters for linear regression.

The RMSE results for the four models using 27 features extracted with SVD:

|  |  |
| --- | --- |
| simple model (average rating) | 13.923 |
| gradient booster | 15.761 |
| random forest | 15.156 |
| linear regression | 12.040 |

With SelectKBest and linear regression, 20 features gave the best RMSE of 11.685.