Machine Learning

The models selected for this classification prediction problem of predicting show rating were logistic regression and random forest. Data pre-processing included the tokenization or tagging of the script data from the *raw_text* column. Next we trained a Doc2Vec model using the tagged data, the resulting vectorized text was utilized in the training of our logistic regression and random forest models. The results indicate that both logistic regression and random forest models performed well, overfitting on the training set (1.0 accuracy) and 96% accuracy on the test set for random forest. Our logistic regression model performed similarly with an accuracy of 95% for both the train and test sets.

Next steps

Further refinement of the models include hyperparameter tuning as well as utilising the results of exploratory data analysis as well as statistical data analysis to inform which features might enhance model performance and thus prediction. Additionally, we could experiment with alternative text vectorization methods, such as bag-of-words.