Analysis & Model Evaluation

Like what we anticipated, Logistic regression is an absolute beast in binary classification. The model using either BoW and TF-IDF performs consistently across the validation set and the final test set, achieving an average classification accuracy around 88%.

If we look at the performance of the two feature-selection processes in the logistic model in terms of classification accuracy, the TF-IDF beats BoW by a strikingly 3% margin. One rational behind could be TF-IDF’s great performance in pinpointing significant features thorough messy, long-winded text. There are on average 100-200 words in each film review, and it is very likely that only a few sentences or words in each text that practically express the author’s sentiment toward the film while the rest are purely noise. Hence, by leveraging a weighted approach such as TF-IDF, the model is more capable in parsing the text to identify the true sentiment of the author.

Logistic regression not only is one of the best in classification accuracy but also tells us a ton of information except just 0 and 1. Thus, we calculated the p-value and the logistic regression weights for each feature we extracted, hoping to get a better sense of which words are most likely to tip off the model to gear towards one side to the other for a given review.

From the output, we successfully identified words that distinguish positive/negative sentiments. A review that contains positive words like “well worth,” “funniest” or popular film genre such as “marvel,” “bond,” “Japanese” are more likely to be classified as positive sentiment whereas a review that has negative words such as “worst,” “waste,” “terrible,” “bore” are more prone to be classified as negative sentiment by the model.