Predicting Academy Award Winners - Repository

Problem statement: Our idea is to use movie attributes (month of release, actor in lead role, reviews from leading social media etc) to predict the movie's relevance and chance for awards and recognition in leading award shows. We are using Academy awards data to test our approaches due to its longevity and public interest.

Data: To start with we have been using movie reviews from leading web forums, newspaper columns as data points for our classifier. We realize that extracting results out of text isn't going to be enough for a classifier to predict with sufficient confidence, but we are starting with the raw text reviews as starting point. We intend to add other movie metadata features once we are at a point where we are struck with a performance point and not able to improve further.
We had modeled the problem as binary classification in which either an entry has won the award or not under particular category. We had a skewed distribution of positives and negatives in quantity. We added weight to balance the distribution once, then used equal number of data points by throwing out less quality negative samples. This had helped to deviate the model from getting wrongly biased. Ideally we would like to model the classification not as simple binary approach but as a selection from a nomination set where the classifier need to assign weightage to each nomination like a regression fashion. The weights should sum up to one. We haven't quite figured out how to construct the model, there are few examples scenarios like anomaly detection where the data set is generally skewed in numbers. We will look to take inspiration from such scenarios.
Another challenge is, our data set is pretty limited sized one where we have 88 set of award results to run our analysis one. By modern data science range, this is going to be a factor limiting our options in terms of techniques applicable.
Approach: For baseline results we aren't tuning the features in any significant way. We applied the data to vectorizer to turn raw text into vectors. Once we have the vectors passed them into classifiers and collected the accuracy results. We had used two linear classifiers for baseline. They give just above half accuracy for our small data set.

Baseline Results:

Categories	SVM	Stochastic Gradient Descent
Best Picture	0.67	0.78
Best Director	0.76	0.65
Best Actor	0.67	0.56
Best Actress	0.53	0.59
Best Cinematography	0.73	0.73