Regression Analysis on Movies Dataset

The central problem presented by the movies dataset to which I sought to apply machine learning techniques was the question of whether a given movie was likely to win an Academy Award for Best Picture, and, secondary to that, whether an assesment as to the overall quality of a movie could be made based on the objective factors given in the dataset. Although the nominations variable has numbers indicating the number of awards for which a movie was nominated, all of the movies from the original acad_awd dataset were Best Picture Oscar winners. Therefore, the real issue here is simply whether a movie won or not. The number of awards won is irrelevant. Accordingly, after creating the training set and testing set, I converted nominations to a simple binary "Yes/No" variable.

Since nominations, now binary, is the response variable, I chose logistic regression as the method for making predictions here. It remained only to choose the best predictor variables for the model.

Constructing a Model

I began by uniting the previously divided datasets (based on the year the films were released) into a single dataset called all_movies. I then divided this set again for training and testing purposes; 80% of all_movies became a training subset called movies_train, while the remaining 20% was set aside as the testing set called movies_test.

I then created a number of preliminary models with different combinations of predictor variables and assessed each one to determine which was optimal. The variables that appeared to be of primary siginificance were duration (length of a film measured in minutes), gross and rating (a film's imdb rating, on a scale of 1 to 10, with 10 being the highest). Based on my prior statisitical analysis of the movies dataset that indicated the prevalence of some genres over others among Oscar-winning movies, I also determined that the genres should also be included as predictor variables in the model.

After proceeding through a number of models both with and without the genre variables and one with an interaction term between gross and budget that proved to be of little significance, I ultimately settled upon two models. The first model, called movies.mod7, included duration, gross, rating, the genre variables as well as a language variable (indicating the language in which the movie was made). The language variable was not significant in itself, but appeared to have positive synergistic effects on the model in conjunction with the other variables. Movies.mod7 had one of the lowest residual deviance scores (at 408.53) among the other models with high signficance scores for duration, gross and particularly rating, with a probability score ($Pr(>|z|)$) on the order of 10^{-16}.

I then ran a forward step model (called fwd_model) on the variables duration, gross, rating, budget, language, country and the genre variables to find an optimal model. The step model ultimately chose as predictor variables rating, country, duration, budget, gross, country (the country in which the film was produced), and the genre variables Drama, Mystery, Sci.Fi, Animation, Romance, Fantasy, Documentary, Horror, War and Family. Rating remained the most significant variable, though not as significant as in movies.mod7. $Pr(>|z|)$ was on the order of 10^{-13}, rather than 10^{-16}, while budget, gross and duration also appeared highly significant as well. The residual deviance for fwd_model was much lower, however, at 304.01 (a difference of 104.49) as revealed by running the anova function on both models. As these two models appeared to be the strongest, I decided to evaluate them both further.

Evaluating the models

I ran the predict functions on both models over the training set with an initial threshold of .5 and then ran ROC curves for both using the pROC package. The optimal thresholds were shown to be 0.045 and 0.055 with Areas Under the Curve (AUC) of 0.947 and 0.971, respectively. I then re-ran the predictions using these new optimal thresholds and obtained the following confusion matrices.

```
##
## pred1_train   No  Yes
##         No  3243   12
##         Yes  288   61
```

```
##
## pred2_train   No  Yes
##         No  3353    5
##         Yes  178   68
```

There was a lower overall accuracy for the first model (91.7% vs. 94.9%), with higher False Negative and higher False Positive rates in the first model. Movies.mod7 had a Sensitivity of 83.6%, a Specificity of 91.8%, a False Negative rate of 16.4% and a False Positive rate of 8.2%, while fwd_model had a Sensitivity of 93.2%, a Specificity of 95.0%, a False Negative rate of 6.8% and a False Positive rate of 5.0%.

Initially, it appeared that fwd_model was completely superior, but this changed once applying the predctions to the test set. The confusion matrices for the test set were as follows (using the new optimal thresholds of 0.034 and 0.046, respectively):

```
##
## pred1  No Yes
##   No  774   2
##   Yes  99  12
```

```
##
## pred2  No Yes
##   No  812   4
##   Yes  61  10
```

There was again a lower overall accuracy for the first of the models (88.6% vs. 92.7%, respectively). For movies.mod7 there was a Sensitivity of 85.7%, a Specificity of 88.7%, a False Negative of 14.3% and a False Positive of 11.3%, while for fwd_model there was a Sensitivity of 71.4%, a Specificity of 93.0%, a False Negative rate of 28.6% and a False Positive rate of 7.0%. Thus, in this instance movies.mod7 had a higher Sensitivity and lower False Negative rate than fwd_model. (However, it should be noted that the proportion of positives ("Yes" values for nominations) in the test set population was lower than that in the training set population, 1.6% vs 2.0% respectively, so the probability of getting a "Yes" hit for nominations was greater in the training set population.) Therefore, although fwd_model is superior in most respects, the first model may be preferred for this dataset where greater sensitivity is desired. It was for this reason that I ultimately chose movies.mod7 to test its ability to make predictions based on specific movies contained in the test set.

Applying the model

I chose two films from the dataset that both had equal values for rating in order to control for the strength of the rating variable by itself. The films I chose were Hellraiser, a film I doubted would be Oscar-worthy, and Seven Years in Tibet, a film I believed could be Oscar-worthy. Both films had ratings of 7. Applying movie.mod7 to a dataframe containing these films' values for the variables in the model, I obtained the following:

```
##                       name          fit       se.fit rating
## 3970           HELLRAISER 7.825771e-11 1.378491e-07      7
## 1340 SEVEN YEARS IN TIBET 3.886629e-02 2.770828e-02      7
```

As can be seen, neither movie was predicted to win a Best Picture Oscar, but Seven Years in Tibet had a much higher likelihood of winning (approximately 4% vs. less than 1 millionth of 1%, with very small standard errors), despite both movies having imdb ratings of 7. (It should also be noted that the results obtainable from fwd_model ultimately proved to be of the same order, so either model would have given accurate results here.) Seven Years in Tibet did not win an Oscar, but was nominated for many other awards,

including a Golden Globe, as revealed by the imdb profile for the film available on imdb.com. Hellraiser was nominated for a Saturn, but nothing as prestigious as a Golden Globe. Thus, imdb rating is not the only determining factor in predicting whether a film will win a Best Picture Oscar. Indeed, this model correctly predicted that neither of these films would win an Oscar and was also able to make (or at least confirm) a "subjective" assessment as to the quality of both films, if such quality can be measured by the relative probability of winning a Best Picture Oscar and by the other awards won by each film.