Homework 1: Representation

Shishido, Juan
juanshishido

February 18, 2016

1 Ideal Features

1.1 Movies

In this subsection, I discuss the ideal features for the Academy Award for Best Picture. Of the six categories we are trying to predict, this is the only one that is specifically about the film and not individual performances.

According to a data visualization by Bloomberg Business[2], there are two features that have, historically, been most predictive for winning Best Picture. These are genre and release date.

An overwhelming number of winners in this category have been dramas. Bloomberg lists seven genres: drama, biography, musical, romance, thriller, comedy, and adventure. Each genre would be represented as a binary feature and each film could belong to a single genre.

The release date data is quite interesting. Bloomberg groups films by the quarter of the year in which they were released. They note that a majority of Oscar-winning films are released in the last quarter—September through December—around the time that nominations are decided. This data would be represented as a set of four binary features, each corresponding to a quarter of the year.

Bloomberg also list a film's budget and box office earnings as important factors. However, their binning is, in my opinion, quite arbitrary. For budget, they group films that spent \$4 million or less, between \$4 and \$16 million, between \$16 and \$64 million, more than \$64 million, and films where data was unavailable. I believe that cost and earnings data is important. However, I would represent it as a continuous variable. Depending on the observed distribution of the data, it might make sense to transform by, for example, taking the *log*.

Another interesting set of features are those related to the results of other awards. FiveThirtyEight's Walt Hickey notes that, their model, relies on data from awards that historically predict the Oscars[1]. They use information on both nominations and winners. Some examples include the Golden Globes, the Critic's Choice Movie Awards, and the Producers Guild of America awards ceremony. The idea is that there is some overlap in the individuals who vote for the Oscars and those that vote for the other awards. In fact, "many members of the Screen Actors Guild will vote for both the SAG awards and for the Oscars"[3]. For each award, there would be two binary features, one representing whether the film was nominated and another representing whether the film won.

Another potentially important set of features relate to the invididuals involved with the particular film. For example, a well-known or well-respected director may be more likely to produce high-quality films. This does not have to only include directors who have won the Academy Award for Best Director. In fact, according to Wikipedia[4], most directors with more than one win have only won twice. Studios, similarly, may play a role in producing Oscar-worthy films. The

Bloomberg visualization mentioned previously notes that Columbia Pictures has the most studio wins. For directors and studios, the features would be continuous. The value would be determined by the total number of awards won, normalized by the total number of films created. Most of the time, this value would be between 0 and 1, but it is possible, for some directors or studios, that this number be higher.

Movie ratings are also potentially important. This feature would most likely be a value between 0 and 5. It would be an average of several ratings. The challenge in this case would be to make sure that every film in the data set is represented by all of the rating institutions. Otherwise, the numbers may be skewed if some films are only rated by institutions who, on average, rate higher (or lower) than others. Time would be an important consideration here, too. The way critics or individuals think about a "4-star" rating, for example, may change over time.

It may also be useful to include information on the other Oscar categories that particular films are associated with. For example, films whose directors, actors, screenplays, scores, etc. are also nominated may have better chances of winning the Academy Award for Best Picture. This feature would simply be a count variable.

Finally, while popularity or sentiment may not reflect the opinions or predispositions of the Oscar's voter-base, it could be a potentially important feature. There are two ways this could be done. First, the number of mentions that films receive in the popular media—news, music, etc.—could be a proxy for "buzz." For a given film, this value would be normalized by the total number of film references (in a specified time period). This could be further enhanced by tracking the sentiment associated with those mentions.

1.2 People

For "people-based" awards, Bloomberg again identifies important features. They distinguish between male and female awards. The most historically predictive feature, for either sex, is race. According to their data source, only one non-white female and only seven non-white males have won Best Actress and Best Actor, respectively. This data might be represented as two binary features—white and non-white.

It turns out prior nominations and prior wins are negatively associated with Oscar wins. Most winners have never been nominated. For both prior nominations and prior wins, the features would have three levels—none, one, or more than one. These features would then be binarized.

Age also seems to be an important factor. In this case, there is a difference between females and males. Most women who have won Best Actress have been between 20 and 40. For Best Actor, this value is between 30 and 50. For this feature, it might make sense to bin into a few categories rather than having a continuous variable. Bloomberg's categories are: 20-29, 30-39, 40-49, and 50 and over.

Several of the features described in the previous subsection could also be used for modeling the "people-based" categories.

2 Subset to Instantiate

In this section, I'll describe where an "ideal" subset of aforementioned features may come from.

A majority of the data can be found on Wikipedia. The `data/wiki/movies` and the `data/wiki/people` subdirectories include HTML files for the corresponding entries. For films, for example, there is information on the director, actors, budget, box office earnings, etc. For people, on the other hand, data on age, filmography, and awards is available.

Ratings data can come from Rotten Tomatoes or IMDB.

The nominations information could come from could come from The Official Academy Awards Database.

Data on mentions could be found by scraping news sites—perhaps focusing on aggregator sites—or Twitter.

References

[1] W. Hickey, *Fivethirtyeights guide to predicting the oscars*, 2016.

[2] L. Meisler, Y. Romero, K. Collins, and A. Pearce, *How to build an oscar winner*, 2015.

[3] N. Silver, *Oscar predictions, election-style*, 2013.

[4] Wikipedia, *List of directors with two or more academy awards for best director*.