Deconstructing Data Science

Part 2 - Anubhav Gupta

The current human process for coming up with the nominees for a category in the Oscars, there a voting mechanism that is managed by "an accounting team at Price Waterhouse Coopers. The firm mails the ballots of eligible nominees to members of the Academy in December to reflect the previous eligible year with a due date sometime in January of the next year, then tabulates the votes in a process that takes some 1700 man-hours."[1] The voting is performed by a select 6028 members of the academy who aside from having unusually high level of quality and distinction is their respective field also meet a strict set of quantitative standards that are laid down by the academy. Each member belongs to a particular branch for example a director or actor or producer and they can only vote for the candidates that are running for that branch.This is a pretty robust model with a good variety of academy members voting in order to come up with a list of nominees for a category.

In order to come up with the list of nominees from a pool of candidates for a particular category, this first algorithmic approach that I think I would be decision trees. Based on the features that had a big weightage in making a candidate a winner, we can build a decision tree and then classify the candidates into nominated or not nominated based on those features. In the category for predicting the nominations for the best film, a decision tree model could be build with the following features and if a movie satisfies all these features, then that movie would be classified as a nominee else it would not make it to the list of nominees

- Oscar nominations from that movie in other categories (like best actor, best actress and best director) exceeds 8
- Revenue generated by the movie exceed $abc
- The movie either won the award for the best feature film either in Golden Globe or in Directors Guild of America and so on.

Since decision trees are really easy to understand and operate as white boxes, the model for selecting the nominations can be made available publicly. the features that are used as decision points can be refined based on the feedback that may be provided.

With this case, there is a probability of an under-representation that might occur. If the features that are used as decision variables introduce a bias in such a way that a certain category/class of movies can never be classified as nominees, that would lead to an under representation of that category. Lets take a very simple example, if there is

[1] http://mentalfloss.com/article/54560/how-are-oscar-nominees-chosen

a feature that says only movies with a budget greater than $x can be classified as nominated, then all the lower budget movies will be under represented. So its all a matter of what features are selected in order to classify a movie as nominated or not nominated.