1 Predicting the Academy Award Winners (Oscars)

Tool: Multinomial Logit

The Analytics Edge: Each year the Oscars awards create huge interest in the movie industry, fans and box-office. There is tremendous consequences of taking an Oscar home, be it future earnings or fame. Using a simple multinomial logit model with data available before the awards are given such as other awards (Golden Globe), other nomination in the Oscars, it is possible to obtain simple models that can predict the winners in four categories - Best Picture, Best Director, Best Leading Actor and Best Leading Actress. This provides an alternate prediction model to expert opinions.

Oscars (Academy Awards)

The Academy Awards (Oscars) is an annual American award honouring cinematic achievements in films. The awards were first presented in 1929 and is now widely recognised as the most prestigious cinema awards in the world.

1.1 Nominations for Oscars

Currently the Oscar nominations are announced to the public in late January and the awards are presented in late February or early March.

Academy of Motion Picture Arts and Sciences (AMPAS), a professional honorary organization with about 6000 members from different disciplines in film production (such as actors, writers, designers, directors, make-up artists) vote on the nominees and final winners of the Oscar awards.

For most categories, members from each of the branches of AMPAS vote to determine nominees in their categories (for example, directors vote for directors). In special cases such as the Best Picture, all voting members are eligible to select the nominees. From these votes, the top 5 nominees are typically selected to be nominated as Oscar nominees.

The winners are determined by a second round of voting in which all the members are then allowed to vote in most categories.

1.2 Viewership of Oscars ceremony

1998 (70th Academy Awards): 57 million viewers
Best Picture: Titanic

2018 (90th Academy Awards): 26.5 million viewers
Best Picture: Birdman

2020 (92nd Academy Awards): 23.6 million viewers (fell from 29.5 million in 2019)
Best Picture: Parasite

1.3 Impact of Oscars on Movie Performance

`www.boxofficemojo.com`: Box office performance

In 2016, the weekly earnings for Birdman went up from 1.3 million to 2.5 million in the week following the Oscars. The other big rise in weekly gross for the movie came after it was nominated for the Golden Globe awards.

Similarly in 2014, the Best Picture winner 12 Years A Slave went from 800,000 to 1.5 million to 2.9 million in the weeks following the oscars.

Similarly winning an Oscar can help significantly increase the salaries and quality of scripts that actors and actresses receive.

Question: Is it possible to predict the winners of Oscars with any reasonable degree of accuracy?

Many people in the media make predictions on the winners of Oscars.

- For example, they might use the buzz of Awards season to make predictions.
- Develop data driven models to make predictions.

In this, we will consider an approach based on discrete choice models used by I. Pardoe and D. K. Simonton in their paper "Applying discrete choice models to predict Academy Award winners".

Note that an important aspect of the prediction here is to predict the winner from one of several nominees (typically the winner from 5 nominees).

2 Choice modeling

2.1 Binary Choice

We have

- $y =$ observed response variable, taking values 0 or 1 (No or Yes).
- $x_1, \ldots, x_p$ attributes of choice, or, explanatory variables

We model the responses as

$$y^* = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon = \boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon.$$

Here y^* is an unobserved latent variable, ϵ is the error term and we denote $\boldsymbol{x} = (1, x_1, \ldots, x_p)$ (column vector) and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ to be the (column) vector of $p+1$ parameters. The relationship between y and y^* is:

$$y = \begin{cases} 1 \text{ if } y^* \geq 0 \\ 0 \text{ otherwise} \end{cases}.$$

Now considering Y to be the random variable representing the response,

$$\begin{aligned}\Pr(Y=1|\boldsymbol{x}) &= \Pr(Y^* \geq 0|\boldsymbol{x}) = \Pr(\boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon \geq 0|\boldsymbol{x}) \\ &= \Pr(\epsilon \geq -\boldsymbol{\beta}^\top \boldsymbol{x}|\boldsymbol{x}) \\ &= \Pr(\epsilon \leq \boldsymbol{\beta}^\top \boldsymbol{x}|\boldsymbol{x}) \text{ (assuming } \epsilon \text{ is symmetric)} \\ &= F(\boldsymbol{\beta}^\top \boldsymbol{x})\end{aligned}$$

where $\epsilon \sim F(\cdot)$.

With a choice of $F(z) = \frac{e^z}{1+e^z}, z \in \mathbb{R}$ we get,

$$\Pr(Y=1|\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}}}{1+e^{\boldsymbol{\beta}^\top \boldsymbol{x}}}.$$

This is the logistic regression model we have seen in the previous class.

2.2 Multinomial Choice

Given $k = 1, \ldots, K$ choices (alternatives), $i = 1, \ldots, n$ consumers (observations), the utility of consumer i for alternative k is,

$$U_{ik} = \boldsymbol{\beta}' \boldsymbol{x}_{ik} + \epsilon_{ik}$$

Here $\boldsymbol{\beta}$ is the weight given to the observable information $\boldsymbol{x}_{ik}$ that includes aspects specific to the individual i and choice (alternative) k. The term ϵ_{ij} captures the noise term that models aspect of choice making not captured by attributes included in $\boldsymbol{x}_{ik}$.

For example x_{ik} might include demographic information, information such as price of a product.

$$\underbrace{\Pr(Y_i = k)}_{\text{Probability that the } i^{th} \text{ individual chooses k}} = \Pr(U_{ik} > U_{il} \forall l \neq k)$$

In some cases, you might have alternate specific constants ASC_k for each alternative.

$$U_{ik} = ASC_k + \boldsymbol{\beta}' \boldsymbol{x}_{ik} + \epsilon_{ik}$$

Assuming that the $\epsilon_{ik} \sim$ are independent and identically distributed with Gumbel (type 1 extreme value) distributions,

$$\begin{aligned}\text{CDF}: &\quad F(y) = e^{-e^{-y}}, \quad y \in \mathbb{R}, \\ \text{PDF}: &\quad f(y) = e^{-y} e^{-e^{-y}}, \quad y \in \mathbb{R}.\end{aligned}$$

It was shown by McFadden (1974) that,

$$\Pr(Y_i = k) = \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_{ik}}}{\sum_{l=1}^{k} e^{\boldsymbol{\beta}' \boldsymbol{x}_{il}}}.$$

This model is known as the **conditional logit model** (also referred to as multinomial logit often).

Special Case of the model (Multinomial Logistic Regression):

Assuming only individual specific data and suppose we want to categorize the individual into one of several categories it gives rise to a multinomial logistic regression model.

$$\Pr(Y_i = k|\boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta}_k' \boldsymbol{x}_i}}{\sum_{l=1}^{k} e^{\boldsymbol{\beta}_l' \boldsymbol{x}_i}}$$

Here $\boldsymbol{\beta}_k$ is a vector of weights corresponding to category k.

2.2.1 Advanced: derivation of logit probability

Denote $v_{ik} := \boldsymbol{\beta}' \boldsymbol{x}_{ik}, \forall k = 1, \ldots, K$ and $i = 1, \ldots, n$.

$$\begin{aligned}
&\Pr(\text{individual i chooses } k^{th} \text{ altenative}) = \Pr(U_{ik} > U_{il}, \forall l \neq k) \\
&= \Pr(v_{ik} + \epsilon_{ik} \geq v_{il} + \epsilon_{il} \forall l \neq k) \\
&= \Pr(\epsilon_{il} \leq v_{ik} - v_{il} + \epsilon_{ik} \forall l \neq k) \\
&= \int_{-\infty}^{\infty} \Pr(\epsilon_{il} \leq v_{ik} - v_{il} + \epsilon_{ik} | \epsilon_{ik} = y) \underbrace{e^{-y} e^{-e^{-y}}}_{\text{Density of } \epsilon_{ik}} \, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \underbrace{\prod_{l \neq k} \Pr(\epsilon_{il} \leq v_{ik} - v_{il} + y)}_{\text{From independence of } \epsilon_{il}} e^{-y} e^{-e^{-y}} \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \prod_{l \neq k} e^{-e^{-(v_{ik} - v_{il} + y)}} e^{-y} e^{-e^{-y}} \mathrm{d}y \\
&= \int_{-\infty}^{\infty} e^{-y} \prod_{l=1}^{K} e^{-e^{-(v_{ik} - v_{il} + y)}} \mathrm{d}y \\
&= \int_{-\infty}^{\infty} e^{-y} e^{-\left(\sum_{l=1}^{K} e^{-(v_{ik} - v_{il} + y)} \right)} \mathrm{d}y \\
&= \int_{-\infty}^{\infty} e^{-y} e^{-e^{-y} \left(\sum_{l=1}^{K} e^{-(v_{ik} - v_{il})} \right)} \mathrm{d}y \\
&= \int_{-\infty}^{0} -e^{-t \sum_{l=1}^{K} e^{-(v_{ik} - v_{il})}} \mathrm{d}t \quad (\text{define } t = e^{-y}, \mathrm{d}t = -e^{-y} \mathrm{d}y) \\
&= \frac{e^{v_{ik}}}{\sum_{l=1}^{K} e^{v_{il}}}
\end{aligned}$$

2.2.2 Properties of Multinomial Logit Model

Independence of irrelevant alternatives

$$\frac{\Pr(Y_i = k)}{\Pr(Y_i = l)} = \left(\frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{ik}}}{\sum_{t=1}^{K} e^{\boldsymbol{\beta}'\boldsymbol{x}_{it}}}\right) \frac{1}{\frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{il}}}{\sum_{t=1}^{K} e^{\boldsymbol{\beta}'\boldsymbol{x}_{it}}}} = e^{\boldsymbol{\beta}'(\boldsymbol{x}_{ik} - \boldsymbol{x}_{il})},$$

$$\log \frac{\Pr(Y_i = k)}{\Pr(Y_i = l)} = \boldsymbol{\beta}'(\boldsymbol{x}_{ik} - \boldsymbol{x}_{il})$$

This property of the MNL model is known as the Independence of Irrelevant Alternatives wherein a choice set consisting of two alternatives k and l, adding in a third alternative does not change the ratio of $\Pr(Y_i = k)/\Pr(Y_i = l)$. Namely the new alternate gains share proportionately from the choice share of existing alternatives in the set.

Example: Blue bus/ Red bus.

Alternatives: Car, Red Bus

$\Pr(C) = \Pr(R) = 0.5$ (Suppose a commuter chooses between a car and a red bus with equal probability)

Assume a new blue bus is added and reasonably commuters don't care about the color of the bus.

Alternatives: Car, Red Bus, Blue Bus

$\Pr(C) = 0.5, \Pr(R) = \Pr(B) = 0.25$: Reasonable prediction

$\Pr(C) = \Pr(R) = \Pr(B) = 0.33$: MNL prediction

This happens because blue bus and red bus are perfect substitutes here and not captured by MNL model.

Note this might be a smaller issue in the case of Oscars predictions since it is not clear if nominees are likely to be close substitutes except if an individual receives multiple nominations in a category in the same year or nominated movies might be from similar genres making them closer substitutes.

3 Maximum Likelihood Estimation

Given the observations:

$\boldsymbol{x}_{ik}$ (attributes of individual i and alternative k).

$$z_{ik} = \begin{cases} 1 & \text{if individual i chooses k} \\ 0 & \text{otherwise. (Note } y_i = k \iff z_{ik} = 1) \end{cases}$$

The problem of estimating the β (weights) that maximises the likelihood of the observations is given as:

$$\max_{\boldsymbol{\beta}} \underbrace{L(\boldsymbol{\beta})}_{\text{Likelihood given } \beta} = \max_{\boldsymbol{\beta}} \prod_{i=1}^{n} \prod_{k=1}^{K} \Pr(Y_i = k)^{z_{ik}}$$

Taking logarithms, this problem can be reformulated as maximizing the log-likelihood.

$$\max_{\boldsymbol{\beta}} \underbrace{LL(\beta)}_{} = \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\Pr(Y_i = k)) = \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\frac{e^{\boldsymbol{\beta} \boldsymbol{x}_{ik}}}{\sum_{l=1}^{k} e^{\boldsymbol{\beta}' \boldsymbol{x}_{il}}} \right)$$

This problem can be efficiently solved as the objective function is concave. This is one of the few formulas for choice probabilities where the objective function is known to be concave.

Testing quality of fit

1. Akaike's Information Criterion:

   $$AIC = -2LL + 2(p + 1)$$

   where LL: Log-likelihood, p : no. of parameters. Smaller the AIC, the better.

2. Likelihood ratio index (McFadden's index):

   $$\rho := 1 - \frac{LL(\hat{\beta})}{LL(0)}$$

   where $\hat{\beta}$ is the estimated value of the parameters. LL(0) refers to the log-likelihood when all the parameters are set to 0 (no model).

   This compares the quality and fit of the model in comparison to a model in which all the parameters are equal to 0. The likelihood ratio index ranges from 0 (estimated model is no better than zero parameters) to 1 (estimated model perfectly predicts the choice observed).

3. Percent correctly predicted

   This identifies the alternative with the highest probability for each individual observation and determining whether or not this was what the actual choice was.