Will this movie get an Oscar ?

Project Proposal - Group 5

Alice Bizeul & Gaia Carparelli & Hugues Vinzant & Antoine Spahr December 3, 2019

Abstract

On the 9th of February 2020, the Academy awards ceremony, also known as the Oscars, will take place and reward some of the best movies of 2019. The Oscars are one of the most prestigious movie-related award one can obtain. Because of this renown, multiple people would be interested in getting insights on which film would potentially compete or even win an Oscar. For example, a movie director could gauges whether the casting he chose will increase his chances of being awarded an Oscar. Or a film corporation could gather information on whether the movies it is funding are worth the effort and investment.
In this context, it would be needed to have an Oscar predictor, a statistical model that could predict the awards results. Hence, the goal of our project is to build models able to predict if a movie would be awarded or nominated for an Oscar, but also to estimate the number of nominations and awards obtained. Hence, this project would be composed of two regression models: one predicting the number of Oscar nominations and the other predicting the number of Awards.

Dataset

In order to develop these models, the IMDB subset from Kaggle[1] will be used as it is a lighter version of the full IMDB dataset suited for local computation. This dataset contains information for around 5000 movies between 1980 and today. The features available are presented on table 1. Those data will be used either to construct the graph either as the features for the models.
The Oscar's data have already been scrapped from the Oscar's website[2]. This dataset will provide the labels for the training. See table 1 for the columns.
We will probably add the Golden Globes awards as features for the model as the ceremony precedes the Academy award and might be decisive for the Oscar results. This dataset would be scrapped from the Golden Globes website[3].

Procedure plan

The movie data will be structured over a network. Each movie is a node and they are connected based on their similarity in casting, crew, genre, and keywords. Four weighted adjacency similarity matrices will be built from the four similarity features (casting, crew, genre and keywords). These matrices will then be fused together through a weighted sum. The weights will be chosen to maximize some graph's property (ex. the clustering coefficient or smoothness). There will be five signals on top of this network, namely : the budget, the revenues, the duration, the vote average and the popularity (and maybe the Golden Globes results). The number of nominations will also be added to the feature vector in the case of award winner predictors. Then an exploration analysis of the created network will be performed to better understand the data.
The model construction will take advantage of the network structure. Indeed the signals will be filtered on the graph and the model will use the filtered signal to make its guess about a movie.

[1] https://www.kaggle.com/tmdb/tmdb-movie-metadata#tmdb_5000_movies.csv
[2] https://www.oscars.org/oscars/ceremonies/
[3] https://www.goldenglobes.com/winners-nominees/

Annexes

Dataset	Variable	Content for one movie	Purpose
IMDB - credits	movie id	id to link the movies	
	title	the movie title	for correspondence between label and node
	cast	all the casting for the movie (actor and actress with their character) as a list of dictionnaries : each dictionnary correspond to one casting member	for node connection
	crew	all the crew members for the movie (with their job) as a list of dictionnaries : each dictionnary correspond to one crew member	for node connection
IMDB - movies	budget	the cost of production of the movie in dollar	node signal
	genre	list of genre of the movie	for node connection
	homepage	link to the movie web page	
	id	movie id	
	keywords	list of keywords characterizing the movie content	for node connection
	original language	language of the movie	
	original title	title of the movie	
	overview	a short description of the movie	
	popularity	a popularity score	node signal
	production company	list of company who produced the movie	
	production country	list of countries where the movie has been produced	
	release date	date of the movie release	
	revenue	revenue generated by the movie	node signal
	runtime	the movie duration in minutes	node signal
	spoken language	list of language spoken in the movie	
	status	state if the movie has been released	
	tagline	a short catchy sentence for the movie	
	title	the movie title	
	vote average	the average vote for this movie	node signal
	vote count	the number of vote for this movie	node signal
Oscars	film	the movie title	
	year	the year for which the ceremony is awarding	
	Oscars	the number of Oscars won by the movie	node label
	Nominations	the number of Nomination for an Oscars of the movie	node label (or signal when predicting the Oscars)

Table 1: Dataset and Variables Overview. The columns *Purpose* indicate for what purpose the variable will be used in the network.