

Forecasting Film Box Office Performance Using Naïve Bayes Analysis

Benjamin Atwater and A.E. Rodriguez, Ph. D

Department of Economics, University of New Haven

Introduction

Figures

Testing the Model

-Films are risky investments for studios which invest hundreds of millions of dollars for films' production.

-In 2016, there were more underperformers than overperformers (McClintock, 2016). The first two quarters of 2017 saw each month's box office in decline (Wilkinson, 2017).

-This project intends to give studio executives a risk management tool in determining if a pitched film will gain return on investment and by how much

Method

-Bayesian analysis combines prior information about a population parameter with evidence from information contained in a sample to guide the statistical inference process (Encyclopedia Britannica, 2017).

-Film sample was randomly generated from list of all films released (Figure 2)

-The dependent variable was box office performance, while the external variables were originality, studio, live action or not, MPAA rating, social relevance, Dow Jones at time of release, and time of release. (Figure 1)

Figure 1: Sample of database

	numbers	release dt	box office	disney-0,fox-1,sony-	original-0,	social rele	DJ at time	summer b	G-0,PG-1,1	true story	american	animated-0,
1981												
The Last Metro	87	13-Feb	3007436		6	0	5		1	1	1	1
Endless Love	22	17-Jul	3.1E+07		5	1	2		0	2	1	0
The Fan	86	15-May	3082096		7	1	4		1	3	1	0
Comin' at Ya!	60	24-Jan	1.2E+07		8	0	1		0	3	1	1
Eyewitness	76	13-Feb	6400000		1	0	3		1	3	1	0
1982												
I'm Dancing as Fast As I Can	123	5-Mar	291390		7	1	6		1	3	0	0
Still of the Night	89	19-Nov	5979947		6	0	3		1	1	0	0
Airplane II	28	10-Dec	2.7E+07		7	1	3		1	1	1	0
Death Wish II	48	19-Feb	1.6E+07		8	1	2		1	3	1	0
Raiders of the Lost Ark Reissue	34	16-Jul	21437879		7	1(reissue)	3		0	1	1	0
1983												
Vertigo	99	28-Oct	5343461		5	1(reissue)	4		1	1	1	0
Over the Brooklyn Bridge	148	2-Mar	837914		8	0	5		1	3	1	0
Valley Girl	41	29-Apr	1.7E+07		8	1	5		1	3	1	0
Strange Invader	138	16-Sep	1362303		8	0	5		1	1	1	0
The Rescuers	36	16-Dec	2.1E+07		1	1	3		1	0	1	0

Figure 2: Randomizer

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
79	44	123	108	30	90	130	221	217	100	211	193	26	226	118	63	199	285	91	163	159	203	176	431	234	107	127	482	580	236	14	
31	109	89	24	113	68	55	189	104	111	164	157	179	230	249	139	96	55	178	116	224	308	59	340	31	33	442	290	83	105	280	
112	51	28	7	8	105	14	178	195	78	43	191	11	85	24	43	140	185	61	22	138	267	196	426	423	361	536	468	489	263	197	
31	95	48	10	161	8	133	78	139	200	34	230	233	158	17	46	244	115	181	123	76	334	345	407	65	63	378	459	1	254	466	
55	89	34	149	151	85	167	211	42	238	147	207	168	145	73	175	185	258	236	93	256	261	438	63	135	21	577	371	122	121	294	

Figure 3: Conversion to numerical classifiers

FILM	year	numbers	studio	something	relevance	dowjones	bb season	rating	notrue	internat	liveaction	boxoffice
When A Man Loves a Woman	1994	Low	disney	no	D	Low	no	R	no	yes	yes	High
In Custody	1994	High	other	no	D	Low	yes	PG13	no	yes	no	Low
Mother's Boys	1994	High	other	yes	G	Low	no	G	yes	no	yes	Low
Murder in the First	1995	Low	wb	no	E	High	no	R	no	yes	yes	High
New Jersey Drive	1995	High	other	yes	B	High	yes	G	yes	yes	no	Low
Jury Duty	1995	Low	sony	no	E	High	yes	PG	no	yes	no	High
Vampire in Brooklyn	1995	Low	paramount	no	F	Low	yes	PG13	no	yes	no	High
Miami Rhapsody	1995	High	disney	no	E	Low	yes	PG	yes	yes	no	Low
Mary Reilly	1996	High	sony	yes	C	High	yes	G	yes	yes	no	Low
Open Season	1996	High	other	no	F	Low	yes	PG13	yes	yes	no	Low
Secrets and Lies	1996	Low	other	no	G	Low	yes	G	no	no	yes	High
Hate	1996	High	other	no	G	Low	no	PG	yes	no	no	Low
Swingers	1996	High	other	yes	B	Low	yes	PG13	no	yes	yes	Low
187	1987	High	wb	no	F	High	no	G	yes	yes	no	Low
G.I. Jane	1987	Low	disney	yes	C	High	yes	PG13	yes	no	yes	High
The Myth of Fingerprints	1987	High	other	yes	A	High	no	R	yes	no	no	Low
Midnight in the Garden of G	1987	Low	wb	no	C	Low	no	PG13	no	yes	no	High

Figure 5: Blade Runner 2049 within the model

	SELECT	High	Low
something	yes	0.727	0.468
dowjones	High	0.545	0.468
bbseason	no	0.455	0.309
notrue	no	0.364	0.319
internatstudio	no	0.818	0.596
liveaction	yes	1.000	0.723
studio	wb	0.273	0.128
relevance	C	0.182	0.160
rating	PG13	0.273	0.170

Based on a past pro	yes	no
High	0.45	0.55
Low	0.52	0.48

Dow Jones	High	Low
High	0.55	0.45
Low	0.47	0.00

BB season	yes	no
High	0.64	0.36
Low	0.72	0.28

Not True	yes	no
High	0.64	0.36
Low	0.61	0.39

International Studio	yes	no
High	0.18	0.82
Low	0.02	0.62

Live Action	yes	no
High	0.91	0.00
Low	0.71	0.29

Numerator	7.6006E-05	2.886E-05
	HIGH	LOW
Naive Bayes Estimate	72%	28%



Courtesy of Warner Bros. Pictures

-Consider the release of *Blade Runner 2049* for testing within the Bayesian model (Figure 5)

-The prior information (past external variables of films compared to their box office success) combined with compared to the new variables for *Blade Runner 2049* indicate a 72% chance that the film will perform within the high bracket of revenue and a 28% of underperforming.

Sources

Box Office Mojo. N.p., n.d. Web.

The Editors of Encyclopædia Britannica. "Bayesian Analysis." *Encyclopædia Britannica*. Encyclopædia Britannica, Inc., 01 Feb. 2016. Web.

IMDb. IMDb.com, n.d. Web.

McClintock, Pamela. "Summer Box-Office Slump: Revenue Down a Steep 22 Percent." *The Hollywood Reporter*. THR, 13 June 2016. Web.

N.d. *YouTube*. Warner, 19 Dec. 2016. Web.

Rodriguez, Armando E., Dr. "Best Picture Prediction: Experts vs the Machine." *UNH Economics Collective*. University of New Haven Economics Department, 24 Feb. 2017. Web.

Wilkinson, Alissa. "The Terrific and Terrible Summer 2017 Movie Season, Explained." *Vox*. Vox, 29 Aug. 2017. Web.

Acknowledgements

Without the SURF program, the College of Business, the Board of Governors, and especially Dr. Rodriguez, this project would not have been possible.