



Project Luther

*Predicting movie opening income
and best release season*

<https://github.com/gabll/Metis-Luther.git>



Project Luther

Project details

Client

Film distributor

Client's need

1. Predict the **movie opening income** in the first weekend
2. Find the **best season** for releasing a new movie



Why predict opening income?



On average,
**28% of domestic total gross
is generated during the first weekend**



Data source

<http://www.boxofficemojo.com/>

Box Office Mojo

Search Site > MOVIES

Search... Social Facebook Twitter Features News Release Sched Showtimes at **IMDb** Box Office Daily Weekend Weekly Monthly Quarterly Seasonal Yearly All Time Chart Watch International Indices **Movies A-Z** Studios

ALPHABETICAL INDEX

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

D-Da | De-De | Dh-Di | Dj-Do | Dr-Dr | Du-Dy

Title (click to view box office)	Studio	Total Gross / Theaters		Opening / Theaters		Open
D'Lucky Ones	ABS	\$64,352	3	\$31,820	3	5/5/2006
D-Day	Yash	n/a	52	n/a	52	7/19/2013
D.A.R.Y.L.	Par.	\$7,840,873	1,100	\$2,649,832	1,100	6/14/1985
D.C. Cab	Uni.	\$16,134,627	908	\$1,564,530	862	12/16/1983
D.E.B.S.	IDP	\$97,446	45	\$56,448	45	3/25/2005
D.O.A.	BV	\$12,706,478	908	\$3,751,432	875	3/18/1988
D2: The Mighty Ducks	BV	\$45,610,410	2,223	\$10,356,748	2,182	3/25/1994
D3: The Mighty Ducks	BV	\$22,955,097	2,060	\$6,170,358	2,056	10/4/1996
Da	FDal	\$644,532	1	\$11,085	1	4/29/1988
The Da Vinci Code	Sony	\$217,536,138	3,757	\$77,073,388	3,735	5/19/2006
Dabangg	Eros	\$1,288,549	68	\$628,137	62	9/10/2010
Dabangg 2	Eros	\$2,519,190	166	\$1,019,213	166	12/21/2012



Data source

<http://www.boxofficemojo.com/>

Box Office Mojo

> MOVIES

Search Site Search... Social Facebook Twitter Features News Release Sched. Showtimes at **IMDb** Box Office Daily Weekend Weekly Monthly Quarterly Seasonal Yearly All Time Chart Watch International Indices **Movies A-Z** Studios

A B C D E F

D-D

D'Lucky Ones
D-Day
D.A.R.Y.L.
D.C. Cab
D.E.B.S.
D.O.A.
D2: The Mighty Ducks
D3: The Mighty Ducks
Da
The Da Vinci Code
Dabangg
Dabangg 2

Features News Release Sched. Showtimes at **IMDb**

Box Office Daily Weekend Weekly Monthly Quarterly Seasonal Yearly All Time Chart Watch International Indices **Movies A-Z** Studios

Domestic Total Gross: **\$217,536,138**

Distributor: **Sony / Columbia** Release Date: **May 19, 2006**

Genre: **Thriller** Runtime: **2 hrs. 29 min.**

MPAA Rating: **PG-13** Production Budget: **\$125 million**

The Da Vinci Code

Summary **Daily** **Weekend** **Weekly** **Foreign** **Dvd / Home Video** **Similar Movies** **Images**

Total Lifetime Grosses

Domestic:	\$217,536,138	28.7%
+ Foreign:	\$540,703,713	71.3%
= Worldwide: \$758,239,851		

Domestic Summary

Opening Weekend: \$77,073,388 (#1 rank, 3,735 theaters, \$20,635 average)

% of Total Gross: 35.4% > View All 14 Weekends

Widest Release: 3,757 theaters Close Date: August 20, 2006

The Players

Director: Ron Howard
Writer: Akiva Goldsman
Actors: Tom Hanks, Audrey Tautou, Ian McKellen, Paul Bettany, Jean Reno, Alfred Molina
Producers: Brian Grazer, Ron Howard
Composer: Hans Zimmer

Images

> View All 37 Images

Related Stories

3/16/07 Around the World Roundup: 2006 Review
6/27/06 Around the World Roundup: 'Poseidon' Rises to the Top
6/19/06 Around the World Roundup: 'Da Vinci' / 'X-Men' Top Soccer



Data source

<http://www.boxofficemojo.com/>

Box Office Mojo

> MOVIES

Search Site
Search...

Social
Facebook
Twitter

Features
News
Release Sched.
Showtimes at [IMDb](#)

Box Office
Daily
Weekend
Weekly
Monthly
Quarterly
Seasonal
Yearly
All Time
Chart Watch
International

Indices
[Movies A-Z](#)
Studios

A B C D E F

D-D

Title (click to view box office)

D'Lucky Ones

D Day

D.C. Day

D.E.B.S.

D.O.A.

D2: The Mighty Ducks

D3: The Mighty Ducks

Da

The Da Vinci Code

Dabangg

Dabangg 2

Box Office
Daily
Weekly
Monthly
Quarterly
Seasonal
Yearly
All Time
Chart Watch
International

Indices
[Movies A-Z](#)
Studios

Box Office Mojo

Search Site
Search...

Adjuster: Actuals Go

The Da Vinci Code

Domestic Total Gross: \$217,536,138

DA VINCI CODE

MPAA Rating: PG-13

Production Budget: \$125 million

Runtime: 2 hrs. 23 min.

Summary Daily Weekend Weekly Foreign Dvd / Home Video Similar Movies Images

Total Lifetime Grosses

Domestic:	\$217,536,138	28.7%
+ Foreign:	\$540,703,713	71.3%
= Worldwide: \$758,239,851		

Domestic Summary

Opening Weekend: \$77,073,388 (#1 rank, 3,735 theaters, \$20,635 average)

% of Total Gross: 35.4% > View All 14 Weekends

Widest Release: 3,757 theaters Close Date: August 20, 2006

The Players

Director: Ron Howard
Writer: Akiva Goldsman
Actors: Tom Hanks, Audrey Tautou, Ian McKellen, Paul Bettany, Jean Reno, Alfred Molina
Producers: Brian Grazer, Ron Howard
Composer: Hans Zimmer

Images

> View All 37 Images

Related Stories

3/16/07 Around the World Roundup: 2006 Review

6/27/06 Around the World Roundup: 'Poseidon' Rises to the Top

6/19/06 Around the World Roundup: 'Da Vinci' 'X-Men' Top Soccer

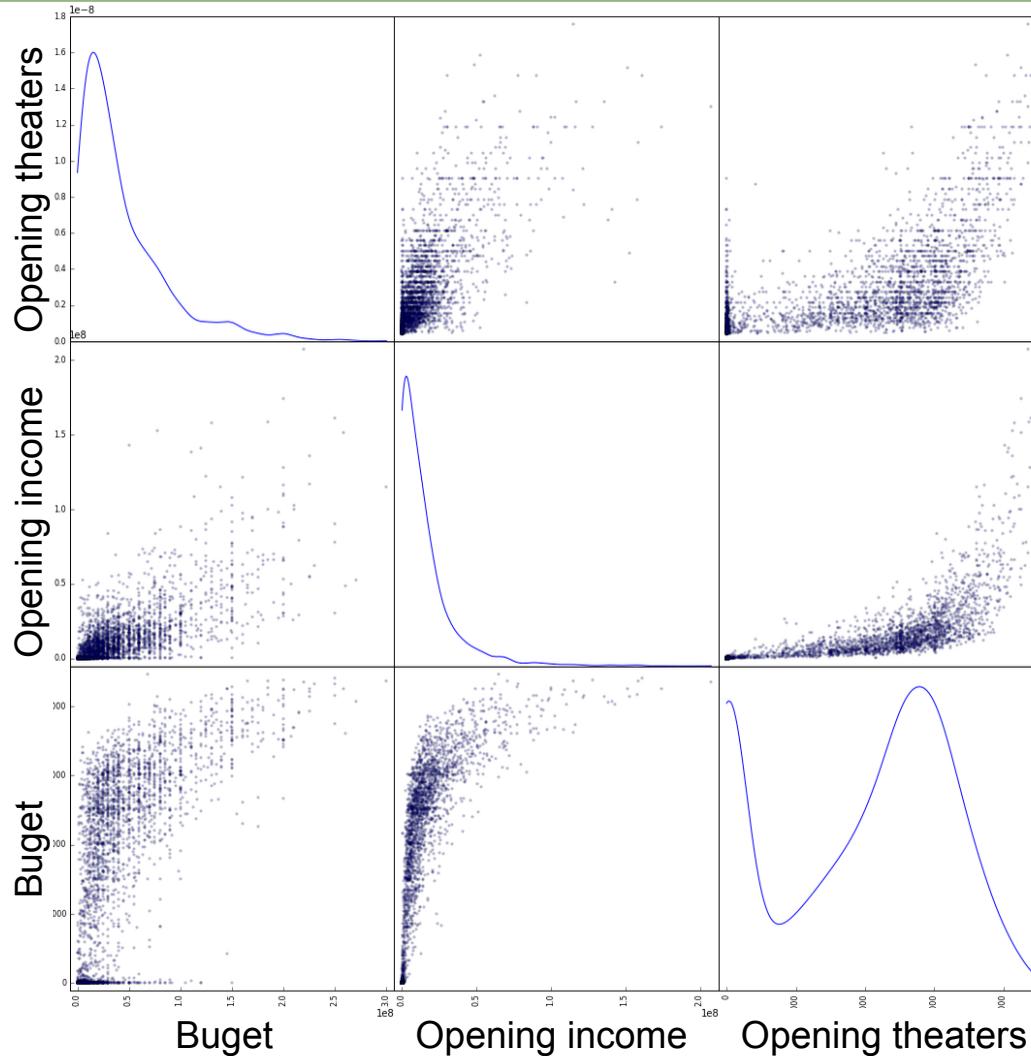
Python package: BeautifulSoup

Pages scraped: 15,230



1. Predict opening income

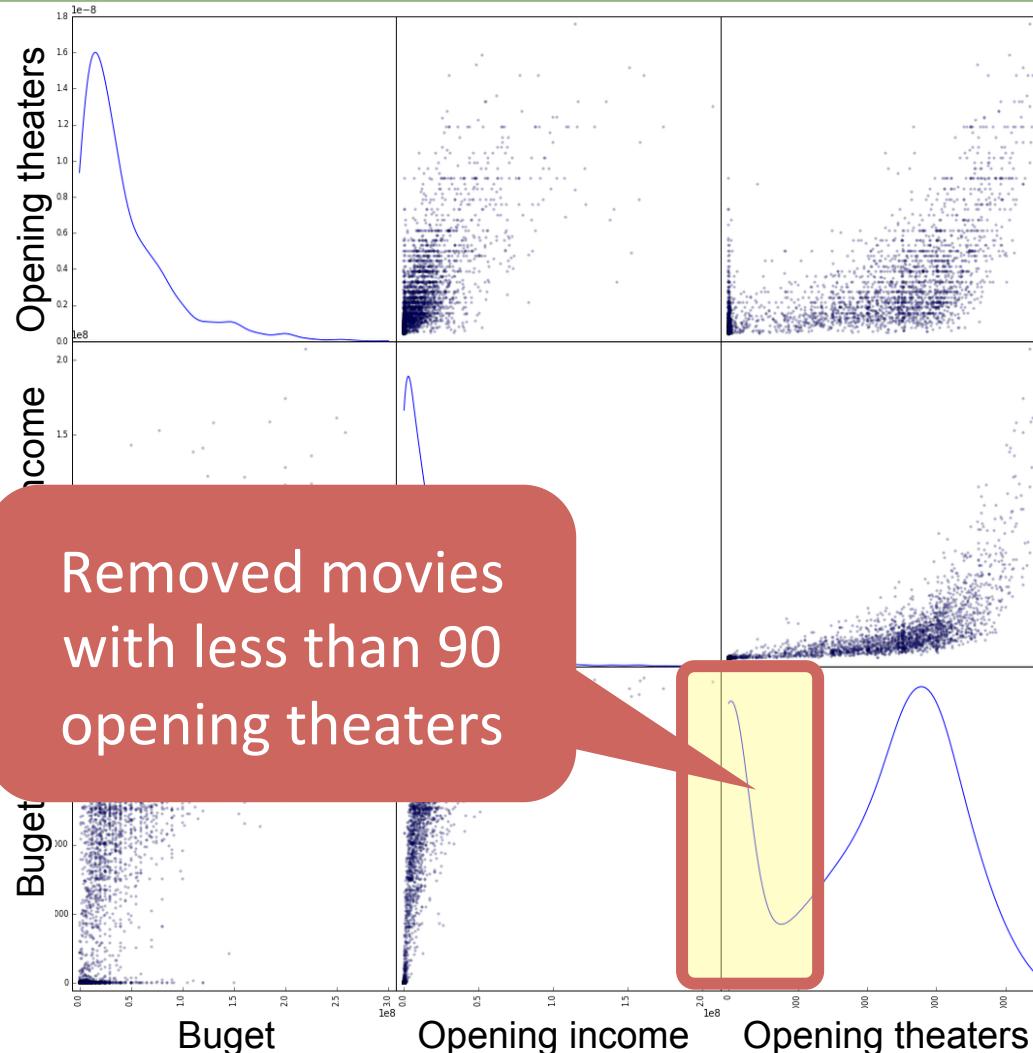
Exploratory Data Analysis





1. Predict opening income

Exploratory Data Analysis





1. Predict opening income

Model statement

$$\log(Y) = \beta_0 + \beta_1 \log(X_{\text{budget}}) + \beta_2 X_{\text{opening_theaters}}$$



1. Predict opening income

Linear Regression

$$\log(Y) = \beta_0 + \beta_1 \log(X_{\text{budget}}) + \beta_2 X_{\text{opening_theaters}}$$



11.8429



0.1261



0.0009

*Ordinary
Least
Squares*

$$R^2 = 0.69$$

≈70% of variation
explained by the
model

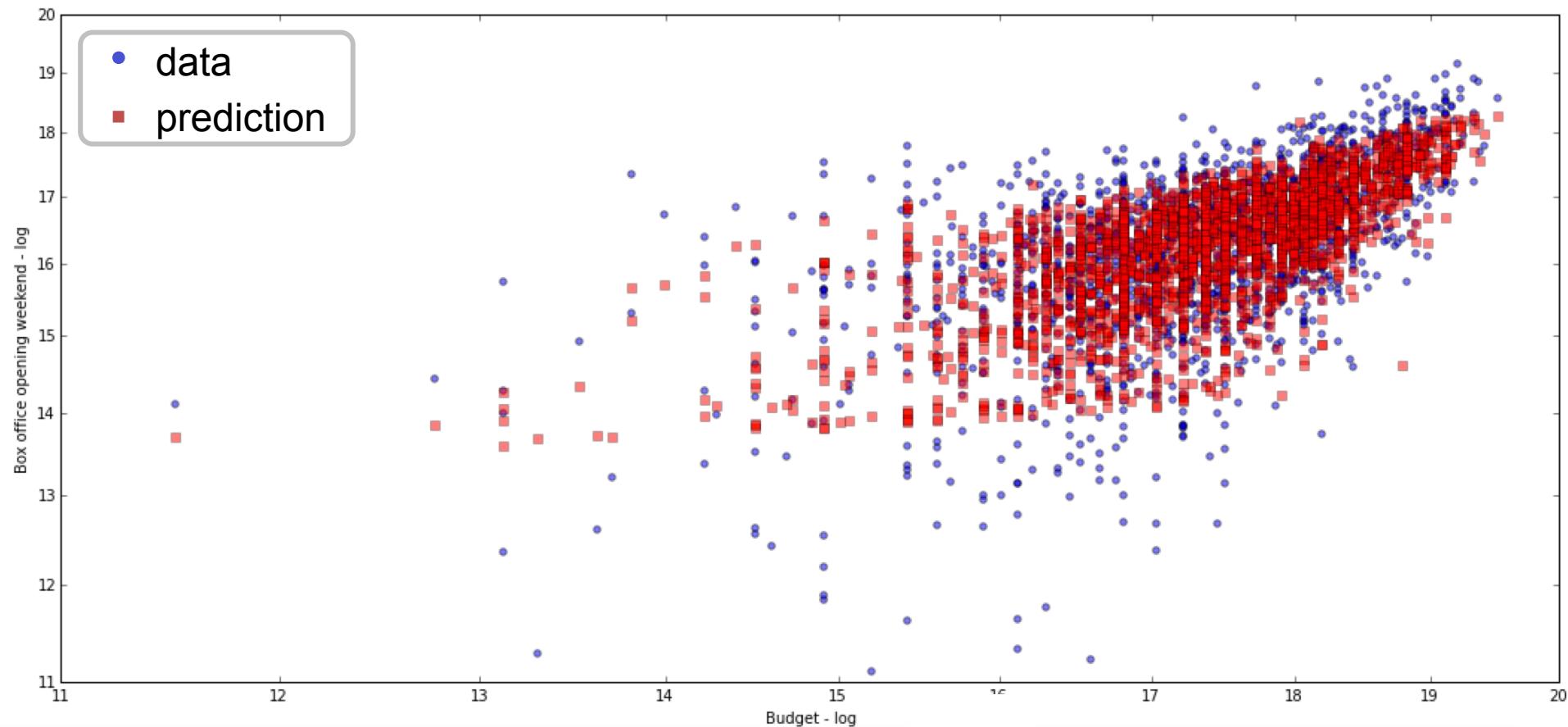
Other features (runtime, first week competition, MPAA rating, etc.)
don't improve significantly the value of R^2



1. Predict opening income

Linear Regression

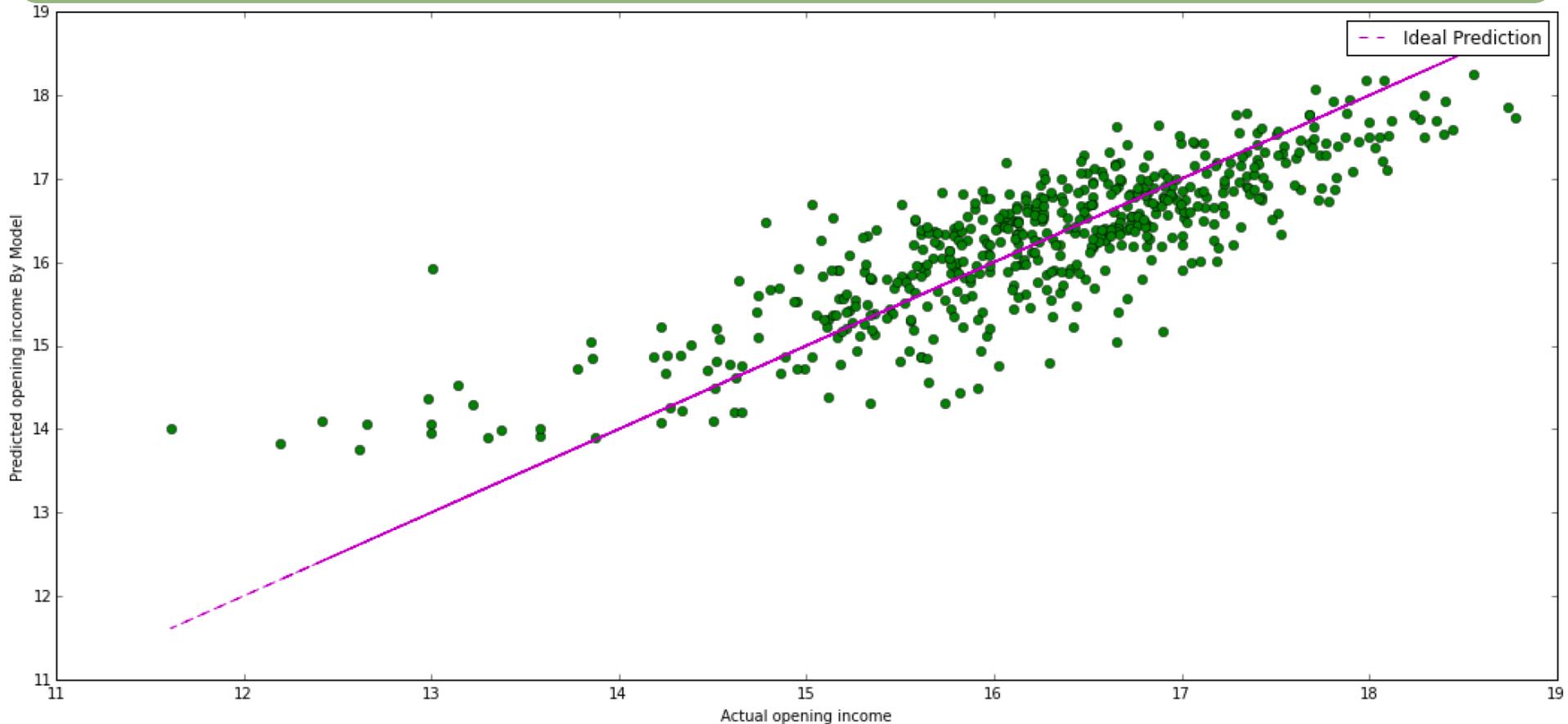
$$\log(Y) = \beta_0 + \beta_1 \log(X_{\text{budget}}) + \beta_2 X_{\text{opening_theaters}}$$





1. Predict opening income

Model evaluation



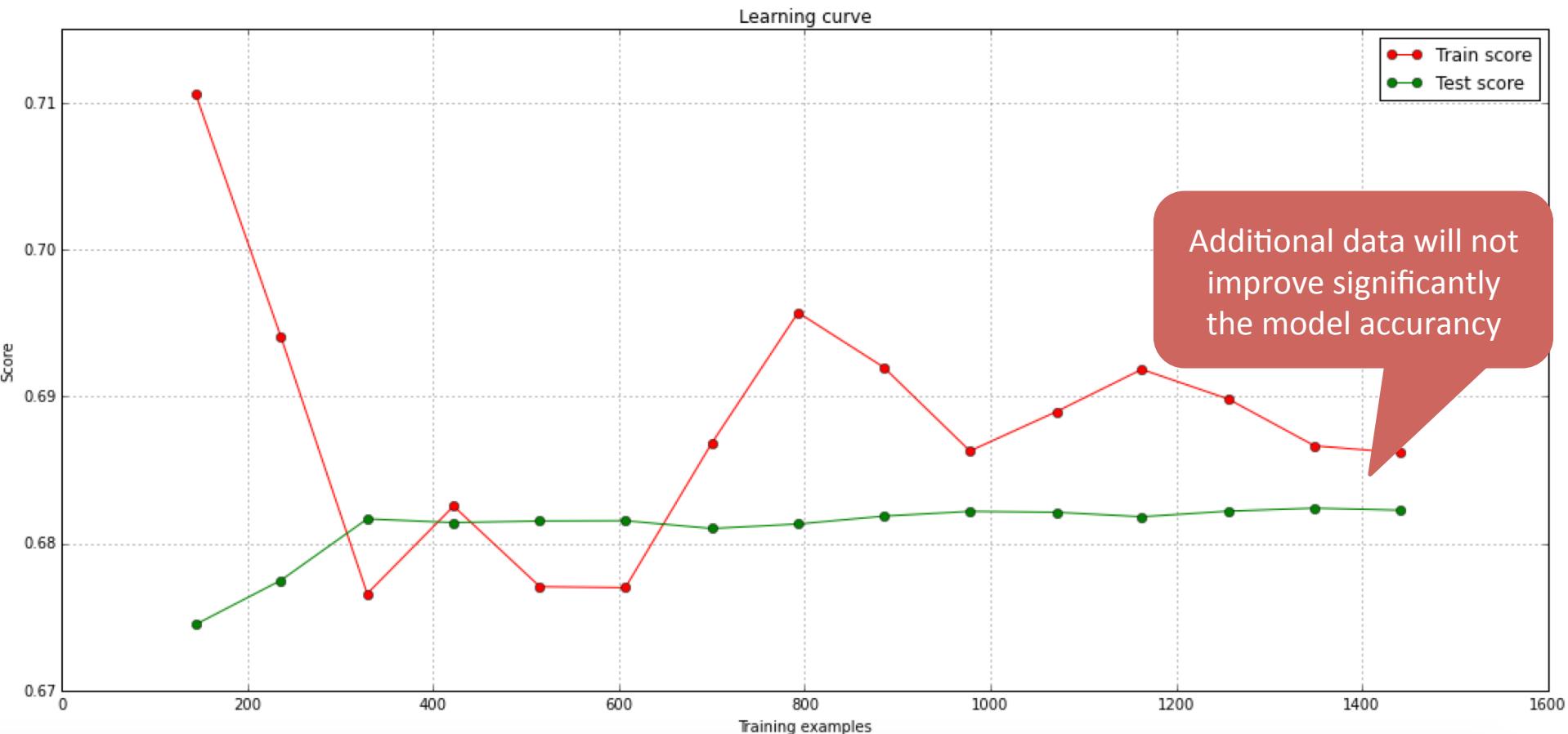
Training set = 75%
Testing set = 25%

Mean Squared Error = 0.34



1. Predict opening income

Model evaluation





2. Find the best season

Additional variables

	release_date	BOM_id	budget	domestic_total_gross	genre	movie_title	opening_income_wend	opening_theaters	rating	runtime_mins	release_weekend	release_season
10	2013-07-19	conjuring	20000000	137400141	horror	The Conjuring	41855326	2903	R	112	201329	Summer
23	2003-12-05	lastsamurai	140000000	111127263	war	The Last Samurai	24271354	2908	R	154	200349	Holiday
25	2012-09-21	houseattheendofthestreet	10000000	31611916	thriller	House at the End of the Street	12287234	3083	PG	101	201238	Fall
33	2007-06-08	hostel2	10200000	17609452	horror	Hostel Part II	8203391	2350	R	94	200723	Summer
37	2010-04-22	oceans	80000000	19422319	documentary	Oceans	6058958	1206	G	103	201016	Spring
39	2014-03-14	veronicamars	6000000	3322127	comedy	Veronica Mars	1988351	291	PG	107	201411	Spring
52	2010-05-14	nottingham	200000000	105269730	adventure	Robin Hood	36063385	3503	PG	148	201019	Summer
64	2011-08-05	tunnel11	135000	1532	horror	The Tunnel	507	1	None	90	201131	Summer
67	2000-01-28	instshegreat	44000000	2962465	comedy	Isn't She Great	1368705	750	R	96	200004	Winter
70	2006-08-11	pulse	20500000	20264436	horror	Pulse	8203822	2323	PG	90	200632	Summer

Movie genre
(binned from
boxofficemojo)

Release season
(binned from
boxofficemojo)

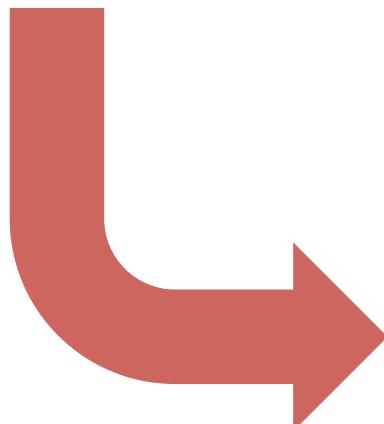
Dummy variables



2. Find the best season

Feature selection

$$\log(Y) = \beta_0 + \beta_1 (X_{\text{opening_theaters}}) + \beta_{2i} X_{\text{genre-season_i}}$$



OLS Regression Results

Dep. Variable:	y	R-squared:	0.731
Model:	OLS	Adj. R-squared:	0.726
Method:	Least Squares	F-statistic:	166.6
Date:	Thu, 29 Jan 2015	Prob (F-statistic):	0.00
Time:	23:50:15	Log-Likelihood:	-4525.6
No. Observations:	4372	AIC:	9193.
Df Residuals:	4301	BIC:	9646.
Df Model:	70		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[95.0% Conf. Int.]
opening_theaters	0.0011	1.08e-05	100.076	0.000	0.001 0.001
action-Fall	0.1113	0.068	1.625	0.104	-0.023 0.246
action-Holiday				0.000	0.277 0.593
action-Spring				0.001	0.093 0.367
action-Summer				0.000	0.385 0.579
action-Winter	0.2637	0.070	3.748	0.000	0.126 0.402
adventure-Fall	0.1149	0.144	0.800	0.424	-0.167 0.397
adventure-Holiday				0.000	0.218 0.559
adventure-Spring				0.027	0.025 0.401
adventure-Summer				0.000	0.235 0.476
adventure-Winter	0.3173	0.155	2.030	0.017	0.057 0.579
animation-Fall	0.0890	0.125	0.714	0.475	-0.155 0.333
animation-Holiday	0.0502	0.092	0.548	0.584	-0.129 0.230
animation-Spring	-0.1272	0.106	-1.194	0.229	-0.334 0.080
animation-Summer				0.583	-0.112 0.199
animation-Winter				0.001	-0.792 -0.203
comedy-Fall				0.058	-0.003 0.204
comedy-Holiday	0.3159	0.051	6.245	0.000	0.217 0.415
comedy-Spring	0.2416	0.046	5.216	0.000	0.151 0.332

Not relevant feature

Not relevant feature

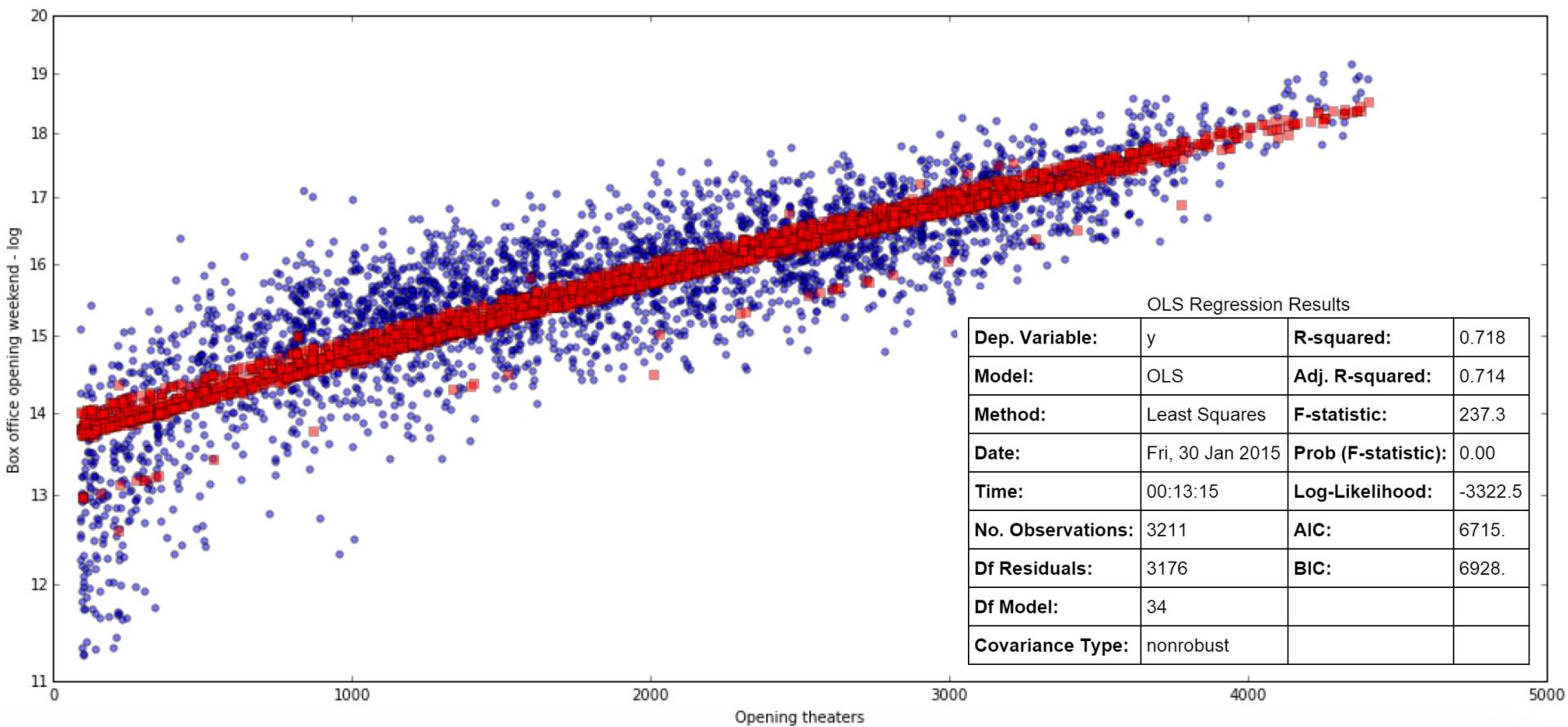
Not relevant feature



2. Find the best season

Feature selection – second iteration

$$\log(Y) = \beta_0 + \beta_1 (X_{\text{opening_theaters}}) + \beta_{2i} X_{\text{genre-season}_i}$$





2. Find the best season

Feature selection – second iteration

$$\log(Y) = \beta_0 + \beta_1 (X_{\text{opening_theaters}}) + \beta_{2i} X_{\text{genre-season}_i}$$

Action	coef
Holiday	0.5383
Winter	0.3658
Spring	0.3345
Summer	0.5894

Adventure	coef
Holiday	0.4964
Winter	0.4238
Spring	0.3226
Summer	0.4687

Drama	coef
Holiday	0.6453
Winter	0.4226
Spring	0.3570
Summer	0.4696

Comedy	coef
Holiday	0.4170
Winter	0.4329
Spring	0.3387
Summer	0.4107

Thriller	coef
Fall	0.2886
Holiday	0.4849
Summer	0.5799

Fantasy	coef
Holiday	0.5887
Spring	0.6123
Summer	0.5512



2. Find the best season

Feature selection – second iteration

$$\log(Y) = \beta_0 + \beta_1 (X_{\text{opening_theaters}}) + \beta_{2i} X_{\text{genre-season}_i}$$

Action	coef	Adventure	coef	Drama	coef
Holiday	0.5585	Holiday	0.4904	Holiday	0.6453
Winter	0.3658	Winter	0.4238	Winter	0.4226
Spring	0.3445	Spring	0.3226	Spring	0.3570
Summer	0.5894	Summer	0.4687	Summer	0.4696
comedy	0.4170	Fall	0.2886	Holiday	0.5887
Holiday	0.4329	Holiday	0.4849	Spring	0.6123
Spring	0.3387	Summer	0.5799	Summer	0.5512
Summer	0.4107				

Holiday doesn't represent the best season for a bunch of genres



2. Find the best season

Example



Genre: **action**

Released: **Winter 2014**

Actual opening

39.2 millions

Predicted opening

33 millions

Predicted opening in
summer season

54.7 millions



Project Luther

Thank you

<https://github.com/gabII/Metis-Luther.git>