Predicting movies' domestic gross

Mauro Gentile

01/26/18

Features and data sources

Movie characteristics

- Gener
- Run time
- Mpaa
- Release month
- Distributor
- Oscar nominations
- Oscars won
- …

Financials

- Adj. budget
- Opening we adj. gross
- (Number of theatres)
- …

Cast and Director celebrity

- Whtd. no of movies made
- Whtd. avg. dom. gross
- Tot. oscar nominations
- Tot. oscars won
- …

\+ interactions

Sources:

10.000 movies from “Box Office Mojo”
6.000 movies from “The Numbers”
list of nominations and Oscar prizes

Celebrity level indicator

"The Beguiled" example

Director

No of movies: 5
Oscar won: 1
Cum. dom. gross: 65 M$
Cum. WW gross: 190 M$

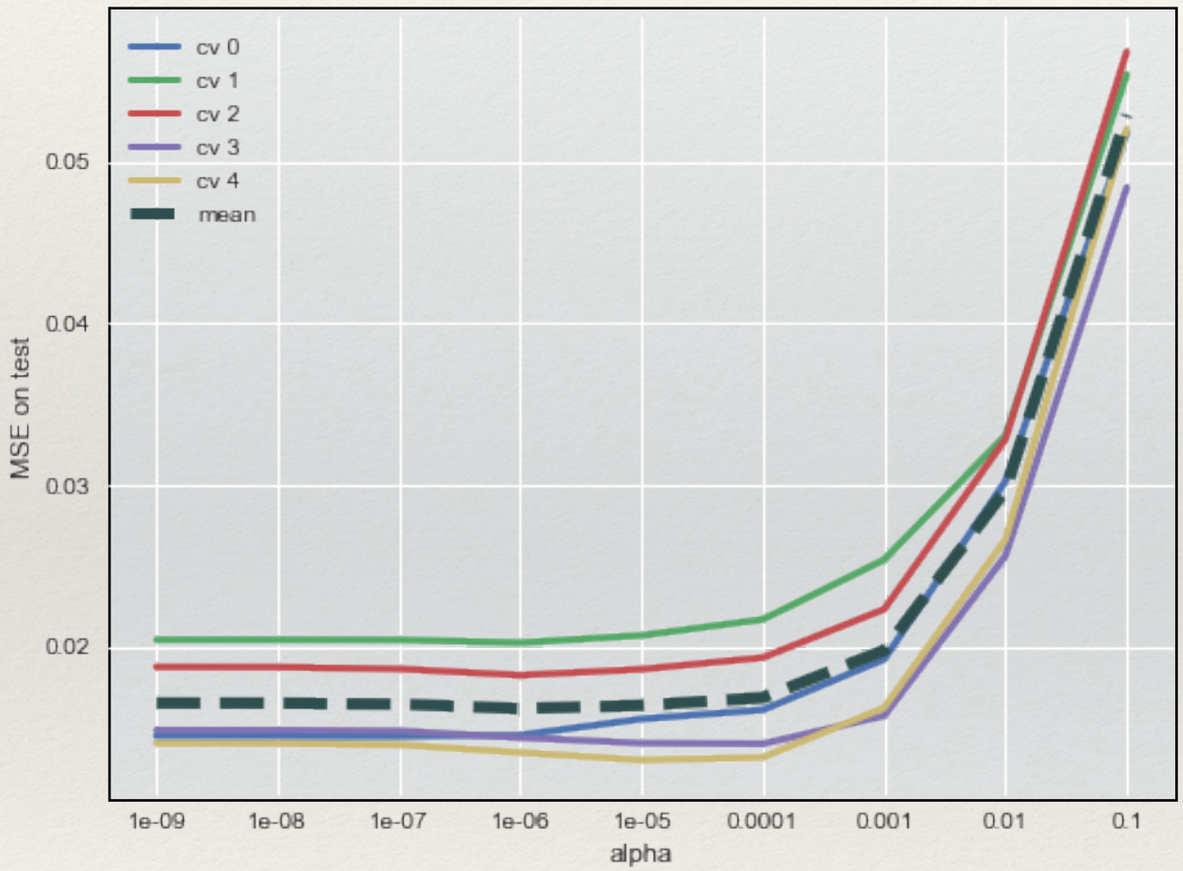
Cast

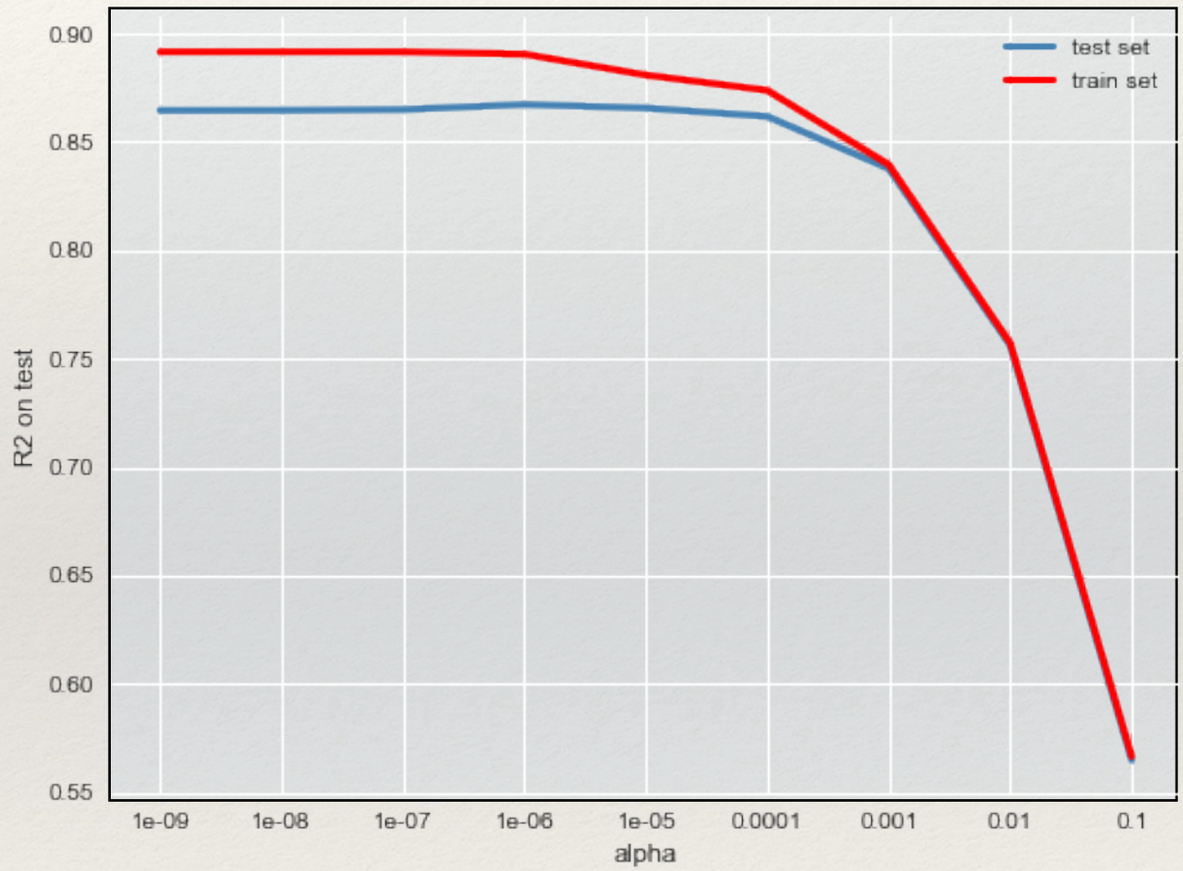
Actor	Total no of prev. movies	Oscars nomin.	Prizes won	Total gross of prev. movies
Nicole Kidman	31	4	1	3,7B$
Elle Fanning	14	0	0	700M$
Kirsten Dunst	26	0	0	1.7B$
Colin Farrell	26	0	0	1.2B$

Weighted total Cast's gross: 2,06B$
Average no of movies: 24.25
Cumulative oscars nomination: 4
Cumulative oscars win: 1

Model results

Lasso, optimised on 5 k-Folds

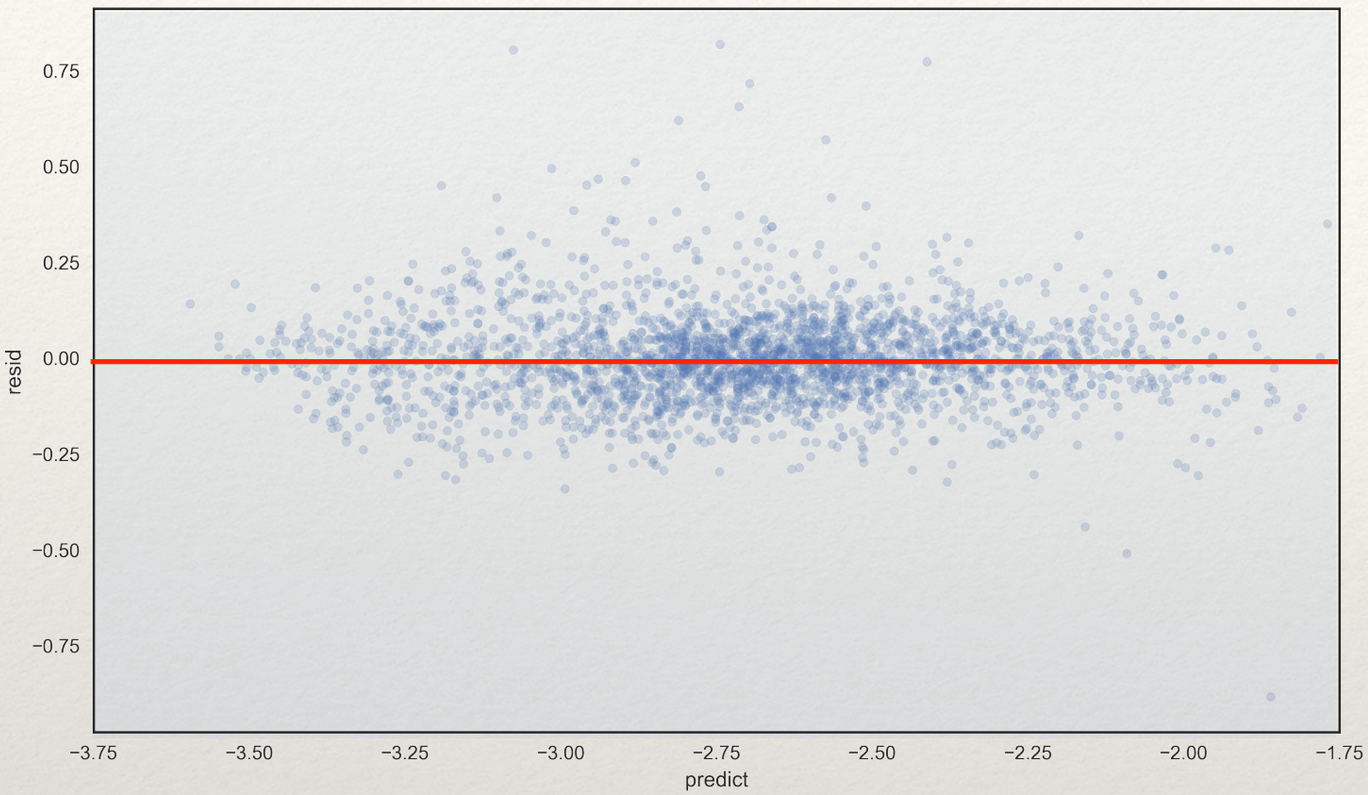


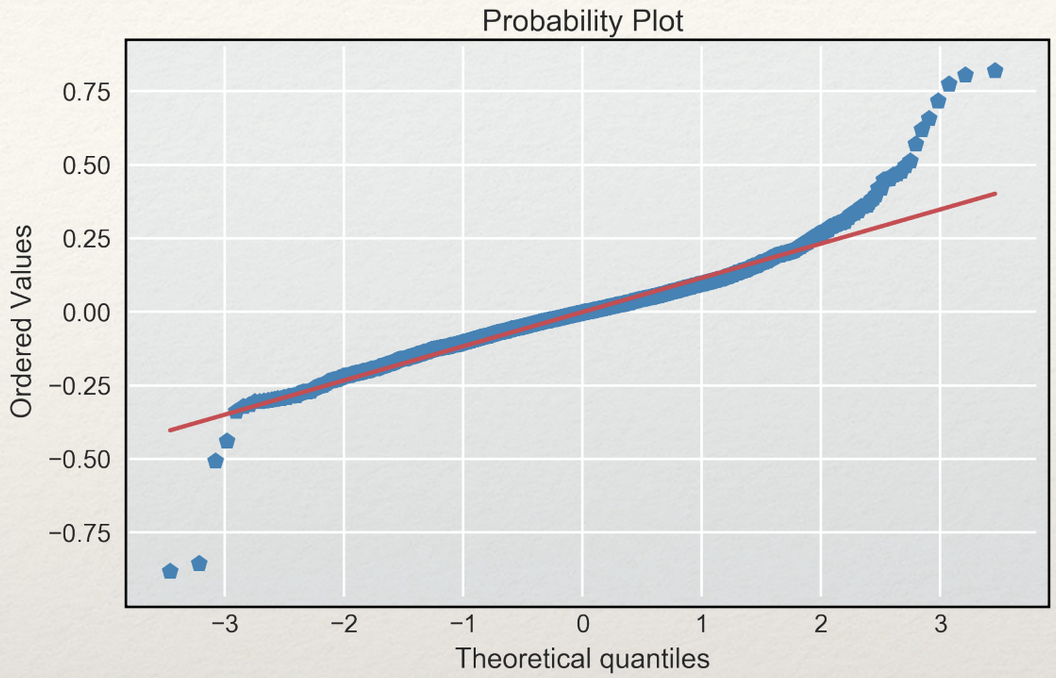


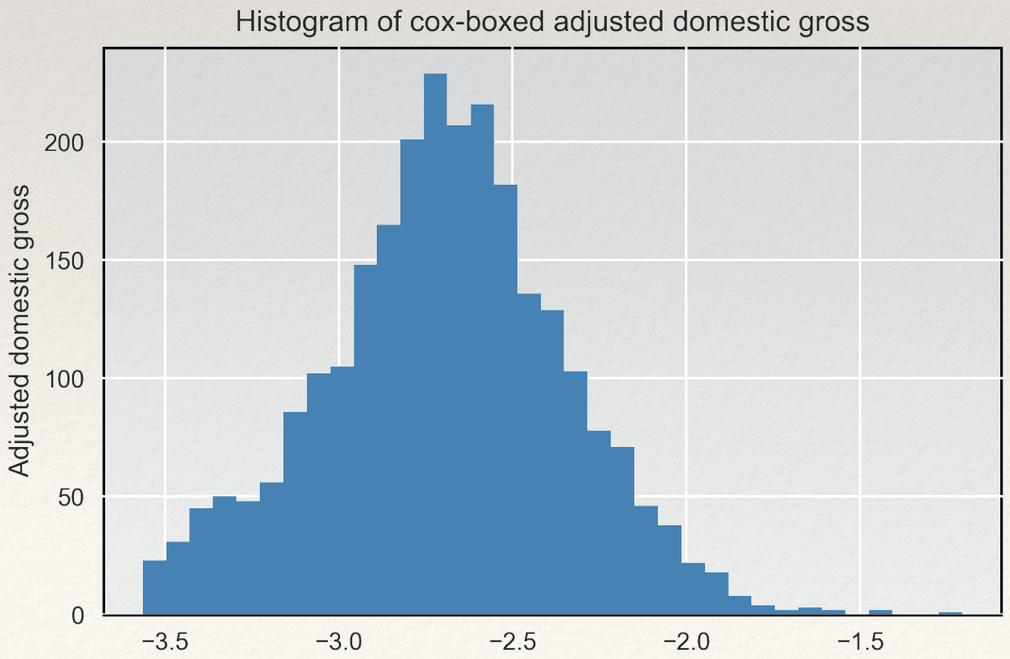
Applied to a cleaned dataset of 2.500 movies described by 312 features

5-kF avg R2 on Test: 0.87

Model assessment







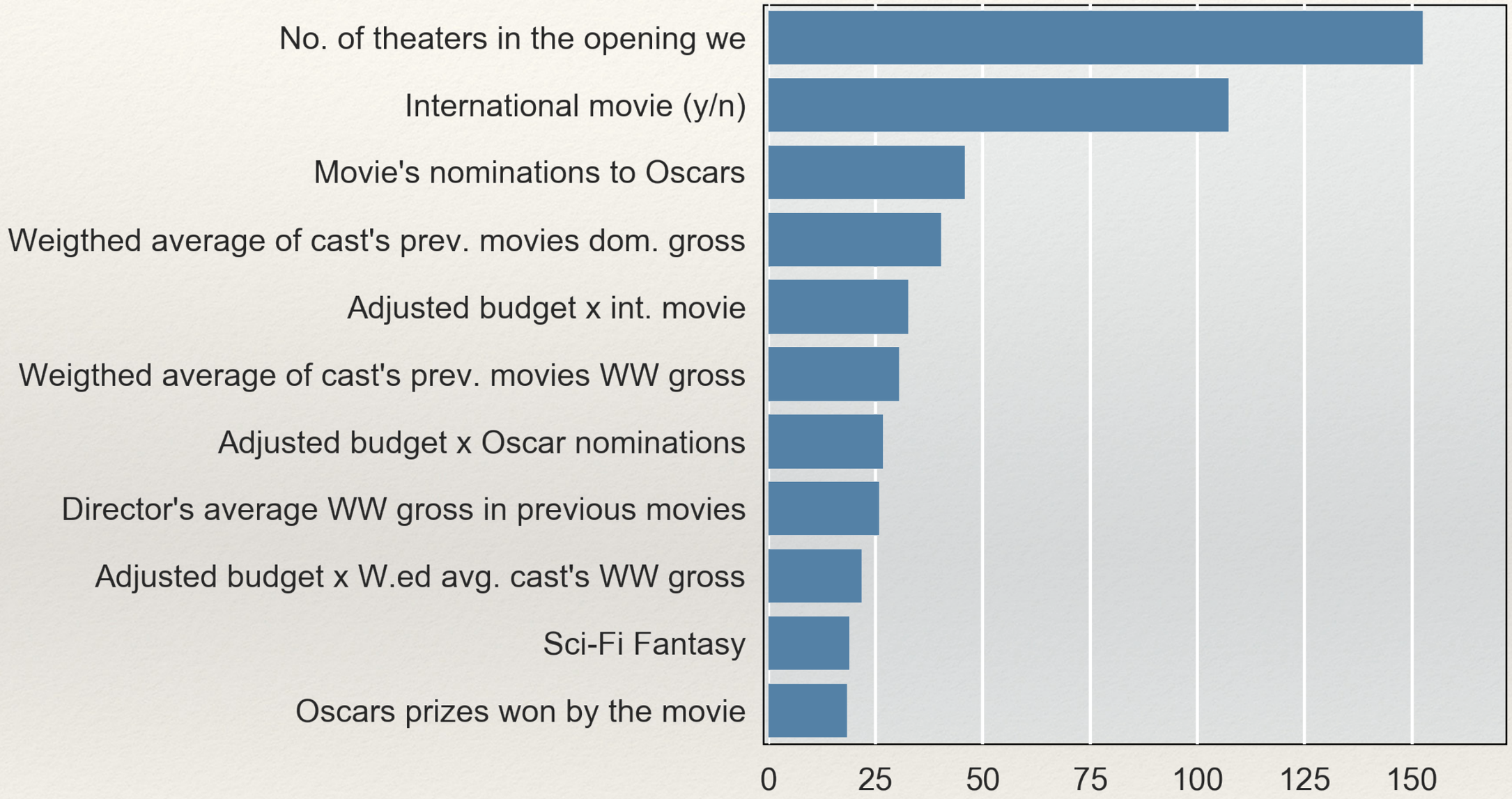
Residuals are:

- nearly normal
- centered in 0
- homoscedastics

Dependent variable is:

- nearly normal in box cox space

Features importance



- Engineered features, especially those on celebrity
- Oscars nomination and prize & no of theatres at the opening we

Model applicability in movie lifetime

Model applicability in movie lifetime

R2:

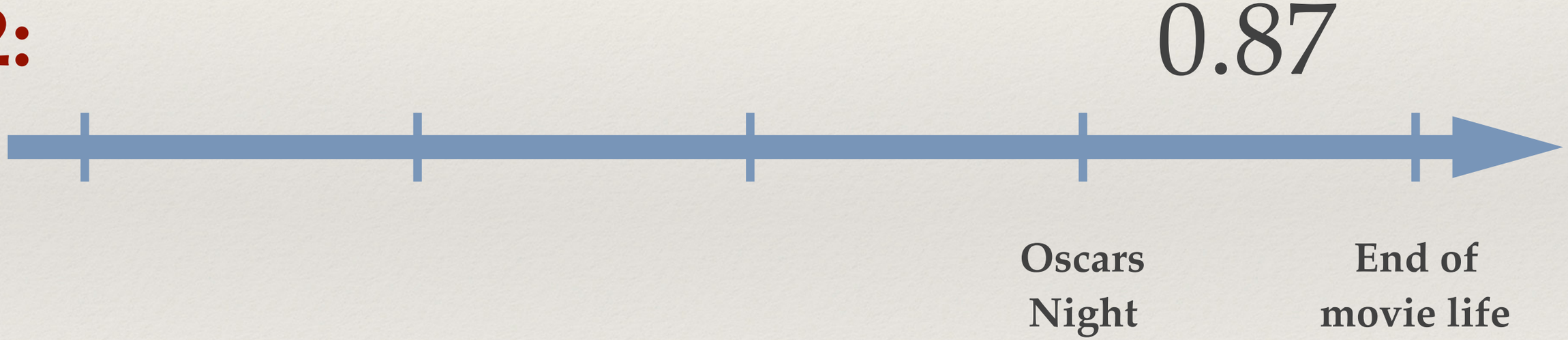
0.87

Oscars Night

End of movie life

Model applicability in movie lifetime

R2:



FULL MODEL

Model applicability in movie lifetime

R2:

FULL MODEL

Model applicability in movie lifetime

R2: 0.87

Release date | **Oscars Night** | **End of movie life**

DROPPED:

- Nominations
- Oscars prizes

FULL MODEL

Model applicability in movie lifetime

R2:

DROPPED:

- Nominations
- Oscars prizes

FULL MODEL

Model applicability in movie lifetime

R2: 0.82 0.87

Production start | **Release date** | **Oscars Night** | **End of movie life**

DROPPED:

- Nominations
- Oscars prizes

FULL MODEL

Model applicability in movie lifetime

R2:

DROPPED:

- Theatres op. we
- Gross op. we

DROPPED:

- Nominations
- Oscars prizes

FULL MODEL

Model applicability in movie lifetime

Model applicability in movie lifetime

R2: 0.68 0.82 0.87

Idea | **Production start** | **Release date** | **Oscars Night** | **End of movie life**

DROPPED:

- Theatres op. we
- Gross op. we

DROPPED:

- Nominations
- Oscars prizes

FULL MODEL

Model applicability in movie lifetime

R2: 0.68 0.82 0.87

Idea | **Production start** | **Release date** | **Oscars Night** | **End of movie life**

DROPPED:

- Mpaa
- Run time
- Rel month
- +others

DROPPED:

- Theatres op. we
- Gross op. we

DROPPED:

- Nominations
- Oscars prizes

FULL MODEL

Model applicability in movie lifetime

R2: 0.64 0.68 0.82 0.87

Idea | **Production start** | **Release date** | **Oscars Night** | **End of movie life**

DROPPED:

- Mpaa
- Run time
- Rel month
- +others

DROPPED:

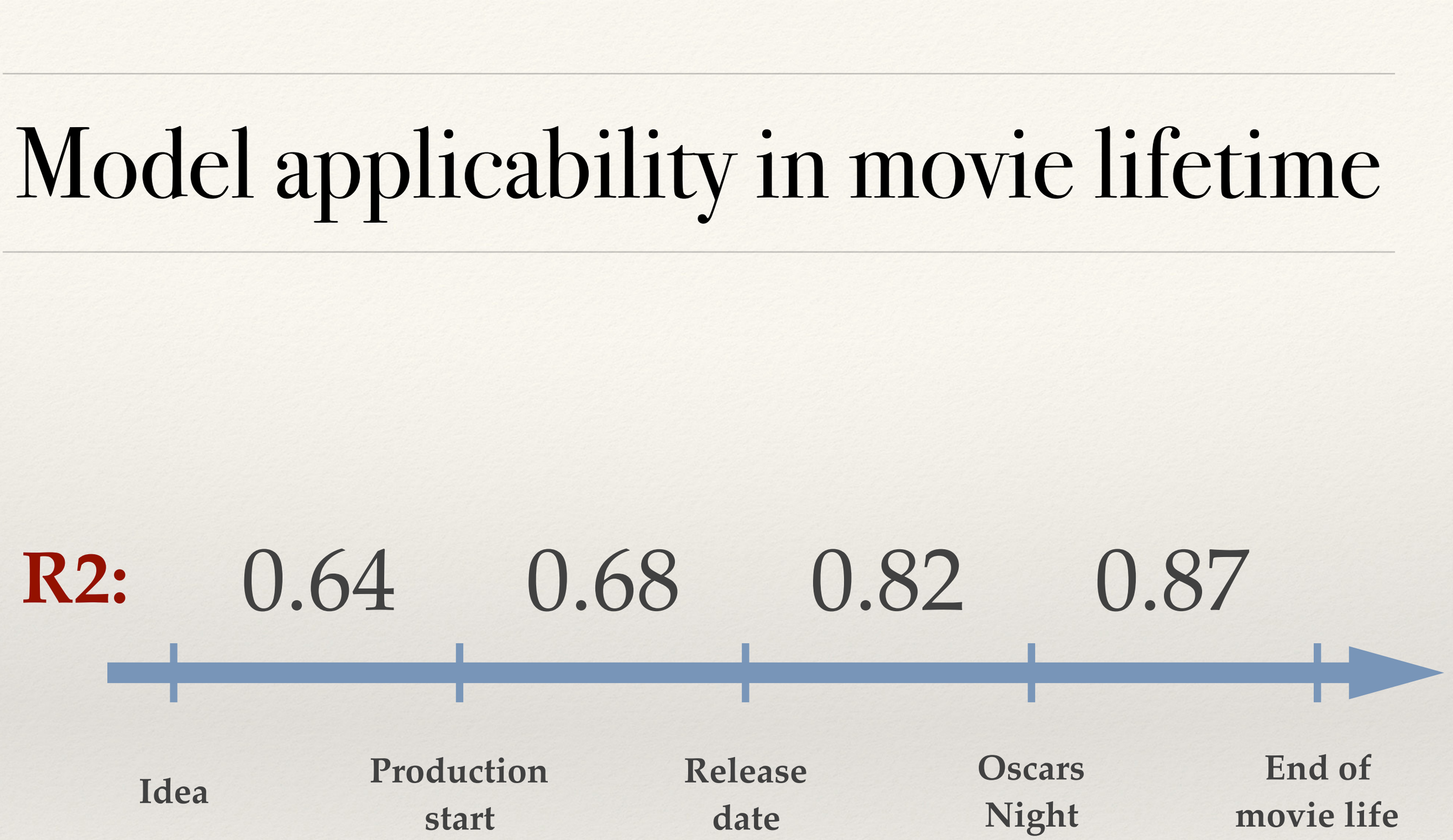
- Theatres op. we
- Gross op. we

DROPPED:

- Nominations
- Oscars prizes

FULL MODEL

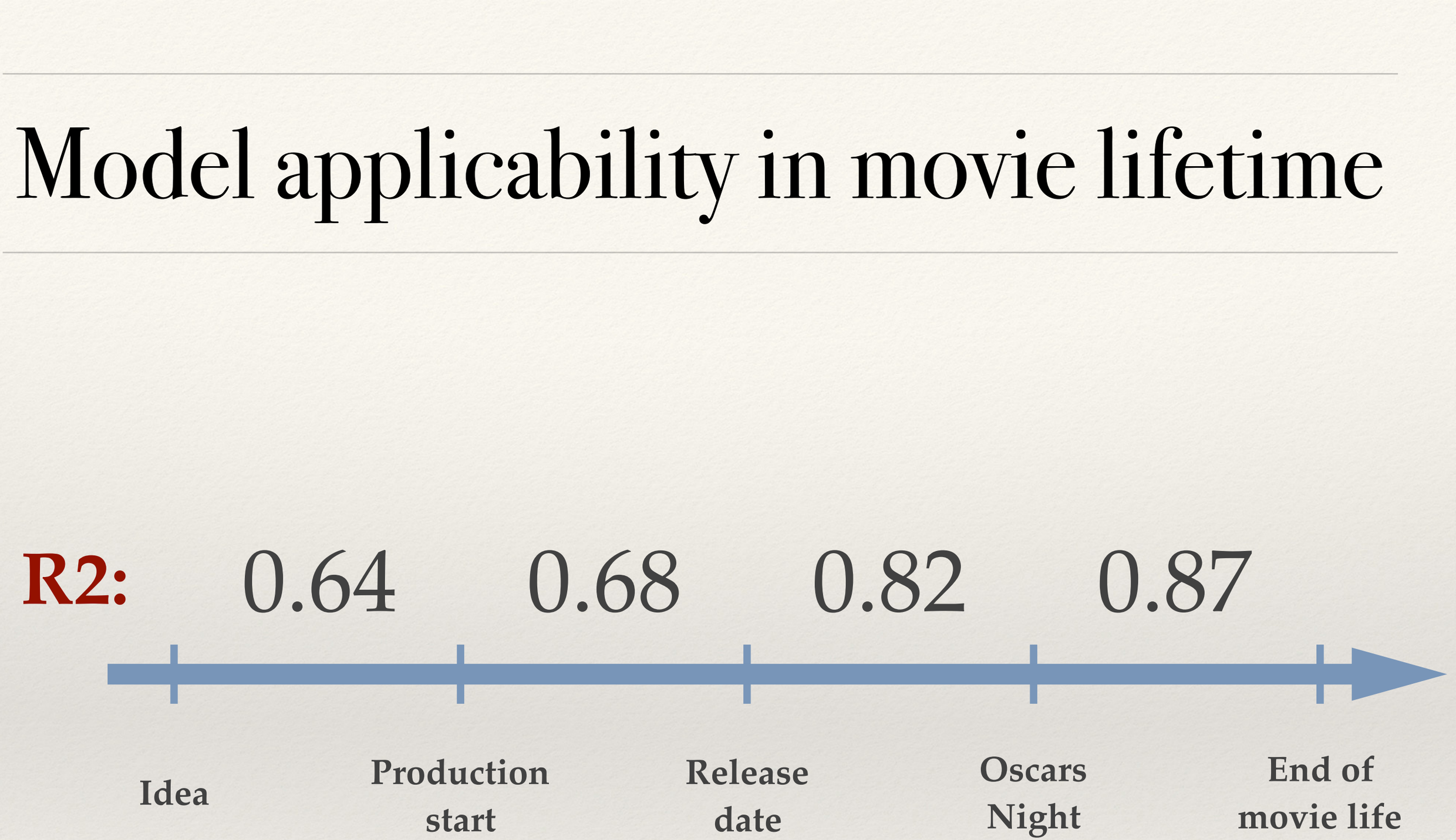
Model applicability in movie lifetime



Model applicability in movie lifetime

The lower the available information, the higher the uncertainty

Model applicability in movie lifetime



The lower the available information, the higher the uncertainty

4 different models that answer to different business questions

Model applicability in movie lifetime

Use:	Production decision	Dimensioning & planning		
R2:	0.64	0.68	0.82	0.87

The lower the available information, the higher the uncertainty

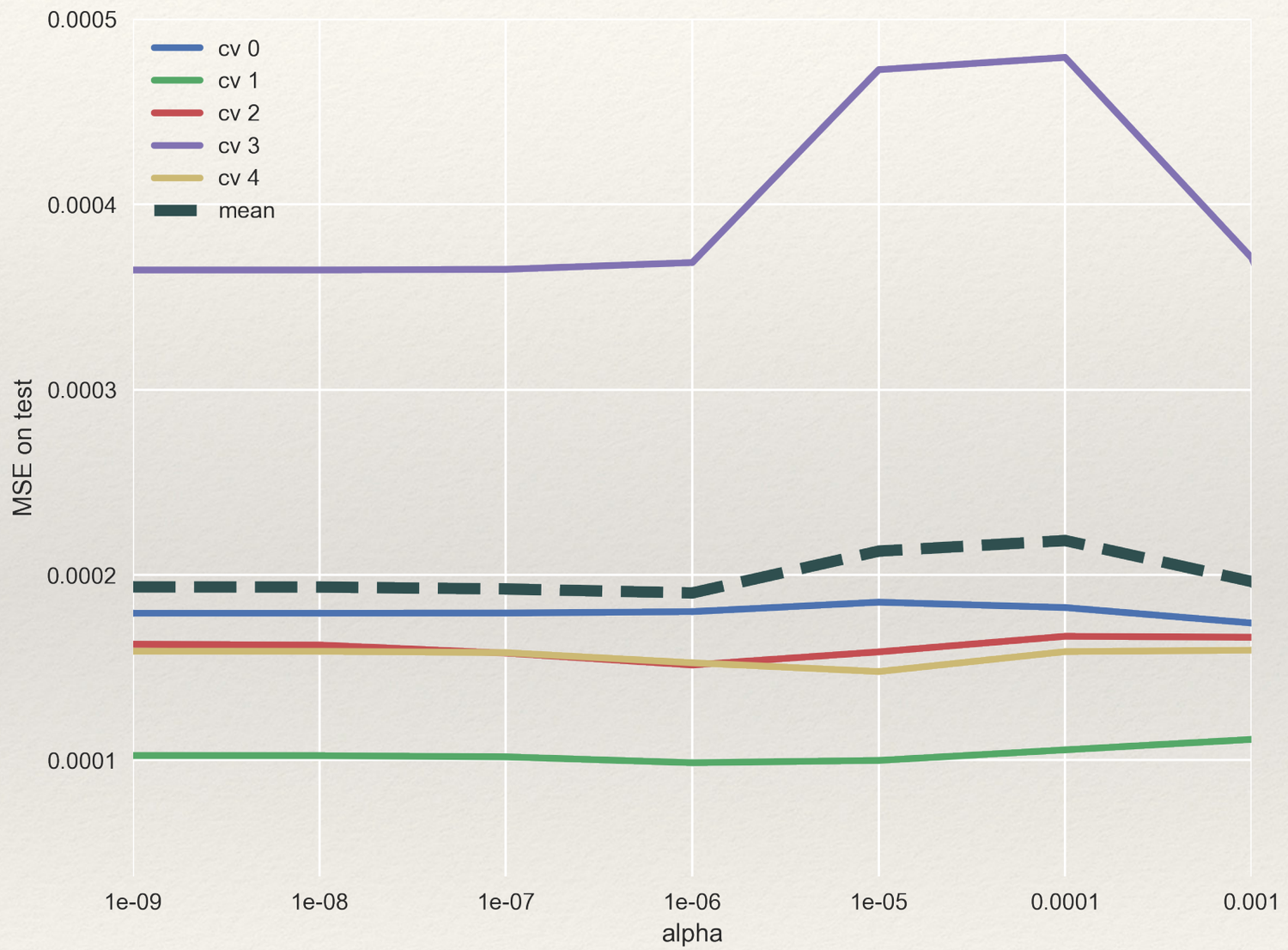
4 different models that answer to different business questions

Would have loved to do

- Second thought on features that could improve the early stage' model
  - *actors' pairs?*
  - *actor/director pairs?*
- Better investigating the reason why budget is not in the top features
- Check if I can partially recover the 75% of dropped rows
  - *Despite the text normalisation, I lost many rows during the db merging*

Thank you

Ineverted


0.0005
0.0004
0.0003
0.0002
0.0001
cv 0
cv 1
cv 2
cv 3
cv 4
mean
MSE on test
1e-09
1e-08
1e-07
1e-06
1e-05
0.0001
0.001
alpha