机器学习在票房预估中的实战

李明辉

猫眼电影高级技术专家

WOTD

World Of Tech
2017年12月1-2日

全球软件开发技术峰会

[深圳站]

报名咨询： 010-68478816
议题提交： wot@51cto.com
市场合作： yangxh@51cto.com
商务合作： songjc@51cto.com
媒体合作： yankk@51cto.com
在线咨询（微信）：18401576051

团·购·享·受·更·多·优·惠

5折 优惠（截止8月31日）
现在报名，立省1400元 / 张

内容提纲

- 行业背景
- 技术体系
- 预测算法
- 工作展望

内容提纲

- 行业背景
- 技术体系
- 预测算法
- 工作展望

行业背景

机遇与风险

- 中国电影市场高速发展
- 高风险、投资回报率难以预测
- 大投入未必有大产出
- 预测工具的缺失

票房预测的意义

- 投资：预估投资回报率，控制风险
- 宣发：营销成本控制
- 上映：智能排片，利益最大化

行业背景


FROM THE DIRECTOR OF FACE/OFF AND M:I-2
NICOLAS CAGE
WINDTALKERS
HONOR WAS THEIR CODE.


马特·达蒙 景甜 佩德罗·帕斯卡
威廉·达福 张涵予 鹿晗 彭于晏 林更新 郑恺 黄轩 陈学冬 王俊凯 特别出演 刘德华
张艺谋导演作品
THE GREAT WALL
12月16日 颠覆想象
3D IMAX 3D 中国巨幕3D 4DX DBOX

行业背景


亏损1.2亿美元
FROM THE DIRECTOR OF FACE/OFF AND M:I-2
NICOLAS CAGE
WINDTALKERS
HONOR WAS THEIR CODE.


马特·达蒙 景甜 佩德罗·帕斯卡
威廉·达福 张涵予 鹿晗 彭于晏 林更新 郑恺 黄轩 陈学冬 王俊凯 特别出演 刘德华
亏损8000万美元
张艺谋导演作品
THE GREAT WALL
12月16日 颠覆想象

行业背景

票房预测的难点

- 中国电影处于野蛮生长期
- 信息量很大，但是垃圾更多
- 可供参考或学习的样本量少
- 有些感性特征难以量化

影响票房的因素

- 题材（受众群范围）
- 卡司阵容（粉丝群范围）
- 影片质量（口碑效应）
- 档期（同行竞争）
- 宣传力度（主动传播）
- 非市场因素

行业背景

票房预测的难点

- 中国电影处于野蛮生长期
- 信息量很大，但是垃圾更多
- 可供参考或学习的样本量少
- 有些感性特征难以量化

影响票房的因素

- 题材（受众群范围）
- 卡司阵容（粉丝群范围）
- 影片质量（口碑效应）
- 档期（同行竞争）
- 宣传力度（主动传播）
- 非市场因素

行业背景

- 票房预测的发展阶段

萌芽阶段(1915~1960)

对变量（因素）的摸索

- 高昂的拷贝价格
- 制片需要了解影片的质量
- 盖洛普研究影响票房的因素

初级阶段(1980~2006)

复杂因素分析模型的建立

电视的发展、刺激电影行业控制风险

巴里•利特曼票房回归模型

斯格特·苏凯的竞争市场预测模型

机器学习模型BPNN

发展阶段(2006~2013)

采用单一数据源为核心的分析

- 基于博客的票房预测模型
- 基于新闻报道的票房预测模型
- 基于Twitter进行票房预测的模型
- 基于google搜索引擎的预测模型
- 基于维基百科的预测模型

行业背景

- 国内的票房预测服务

猫眼

淘宝

腾讯

百度

艺恩

时光

一起拍电影

票房透视镜

行业背景

- 票房预测服务的分布

	猫眼	淘宝	腾讯	百度	艺恩	时光	一起拍	犀牛娱乐
全国实时票房	✔	✔	✔	✔	✔	✔		
影院实时票房	✔							
次日票房预测	✔		✔	✔			✔	✔
总票房预测	✔			✔			✔	✔

内容提纲

- 行业背景
- 技术体系
- 预测算法
- 工作展望

技术体系

实时票房

- 每部影片在全国、各影院、城市的票房、排片、人次、平均票价等

天级票房

- 每部影片在全国次日票房预测
- 前一日17点、21点

总票房

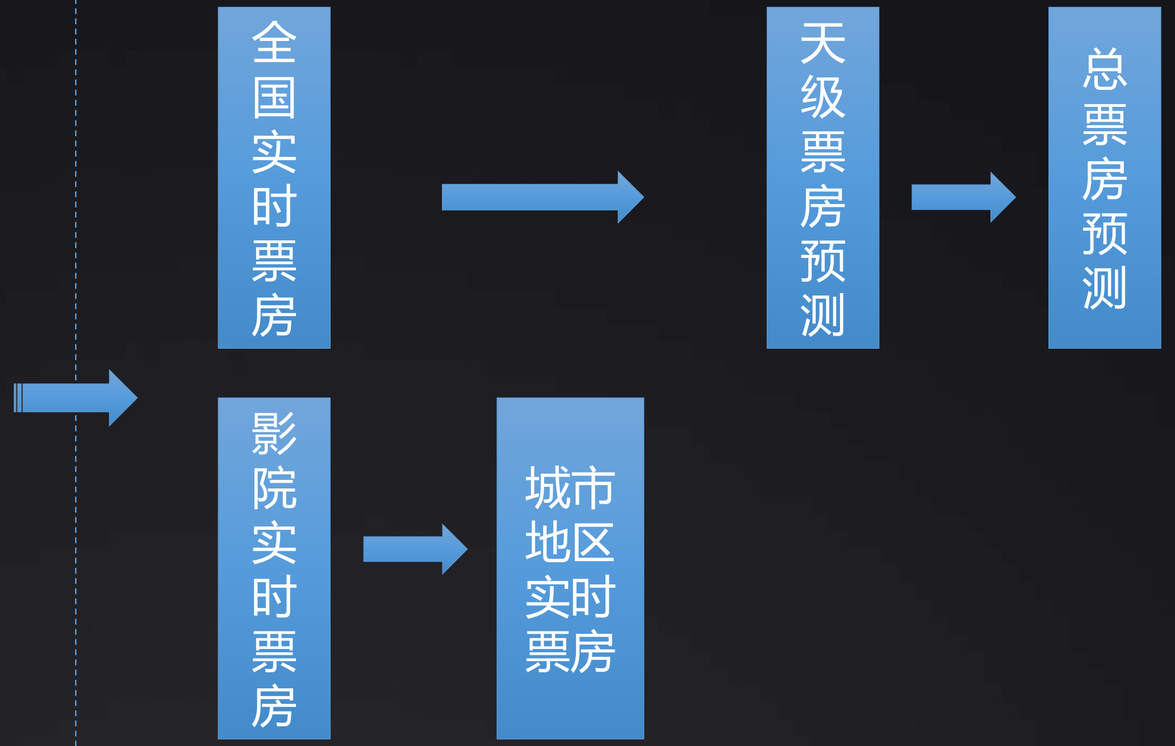
- 每部影片在整个上映周期的全国票房预测
- 首映日、首周末时间节点

预售票房

天级票房预测结果

技术体系

基础数据



预测服务

猫眼专业版

技术体系

实时票房

总票房

天级票房

内容提纲

- 行业背景
- 技术体系
- 预测算法
- 工作展望

实时票房

案例：《我不是潘金莲》

11.18首映：万达排片低于其他院线

冯小刚《潘金莲至王健林的一封信》

王思聪反击

曝光度激增

票房

实时票房

- 猫眼的数据优势
  - 与90%以上的影院合作，对接18种售票系统
  - 拥有详细的影院场次及座位状态

影院

售票系统

座位图

实时票房

- 猫眼的数据优势
  - 与90%以上的影院合作，对接18种售票系统
  - 拥有详细的影院场次及座位状态

影院

售票系统

座位图

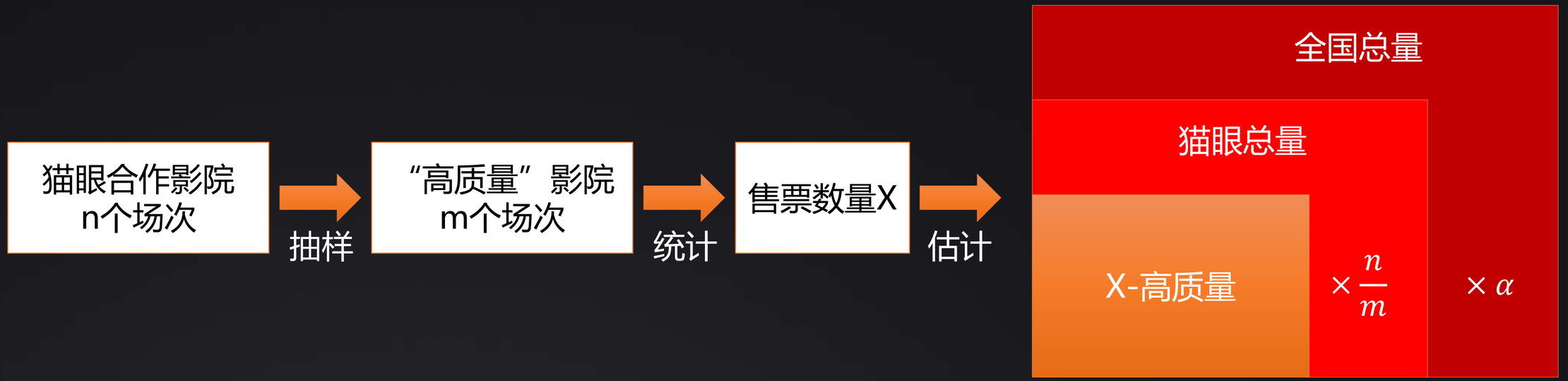
问题：不可选 ≠ 已售

实时票房

- 问题定义
  - 已知猫眼合作影院（占全国约90%）每部电影、每个场次的实时座位图数据
  - 预测每部电影在全国全部影院的实时票房总和
- 难点
  - 数据噪音：座位图中状态不明确
  - 数据不完备：非全部影院
- 解决方案
  - 数据抽样：以部分样本为基础，估计全量

实时票房

- 统计模型



$$g_t = X_t * \frac{n_t}{m_t} * \alpha_{t-1}$$

实时票房

修正

数据清洗

票价处理

历史alpha值计算

实时票数计算

实时票房计算

实时大盘计算

影院黑白名单

影院实时票房计算

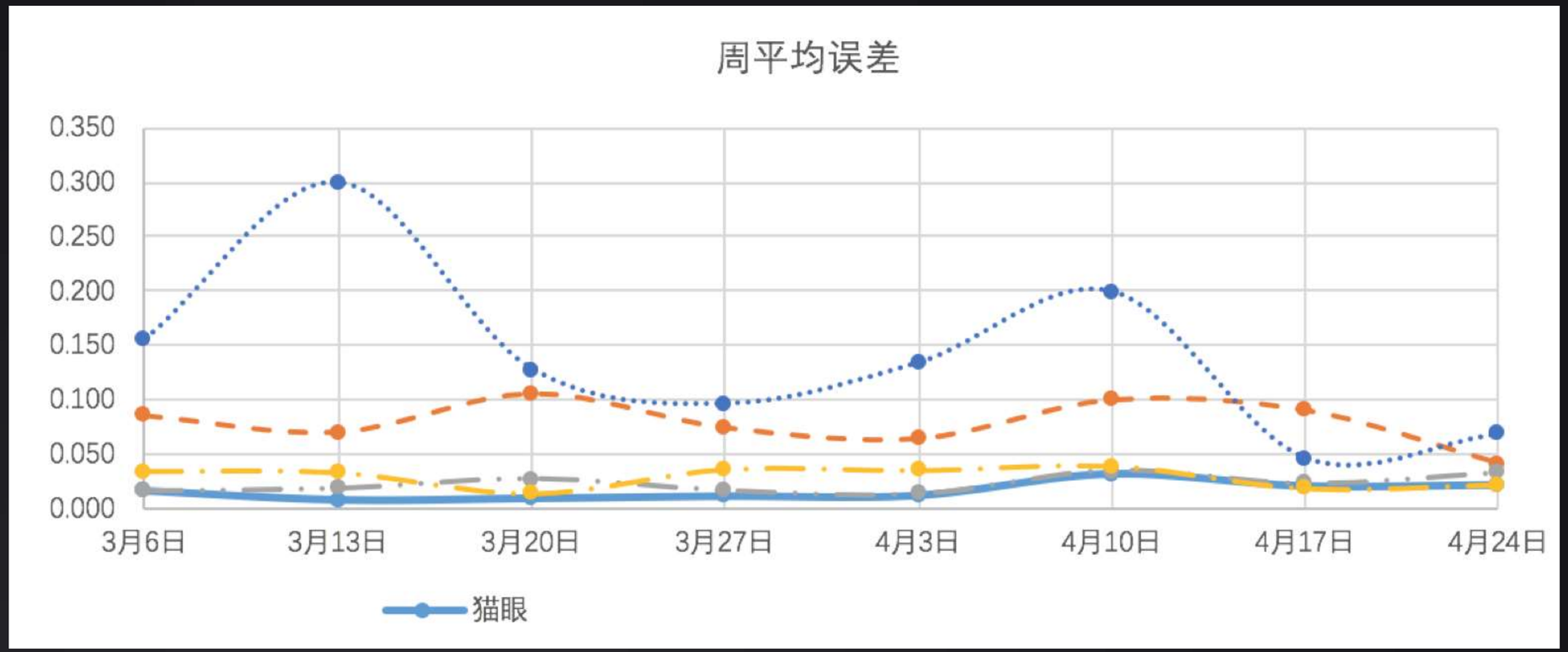
地区实时票房计算

离线计算

在线计算

实时票房

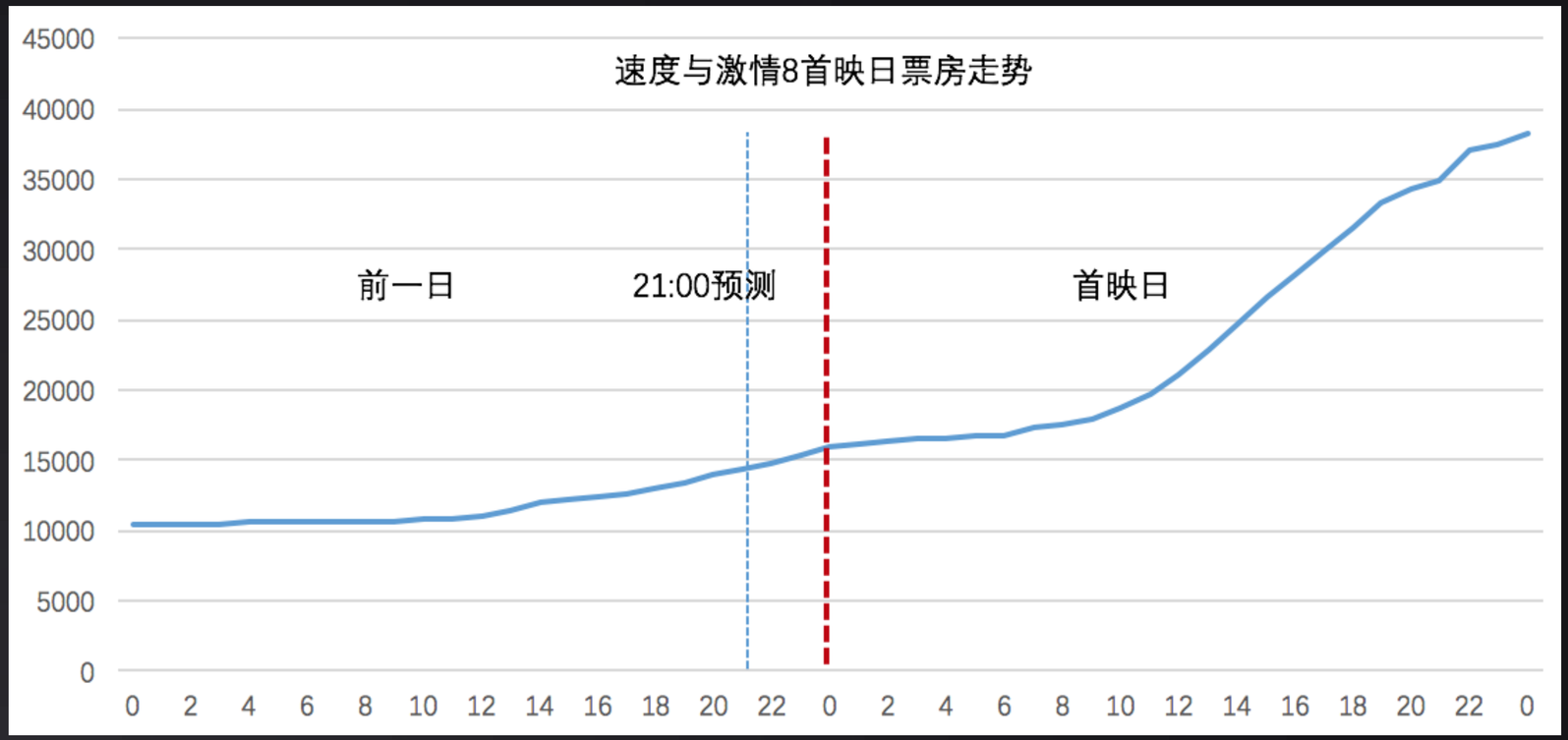
- 效果对比


周平均误差
0.350
0.300
0.250
0.200
0.150
0.100
0.050
0.000
3月6日
3月13日
3月20日
3月27日
4月3日
4月10日
4月17日
4月24日
猫眼

天级票房预测

天级票房预测

- 猫眼的优势
  - 在线售票系统：直接的交易数据，预售票房、排片场次等
  - 精准的实时票房为天级票房提供可靠的特征
- 问题定义
  - 次日票房预测
  - 前一日21:00
  - 前一日17:00
- 难点
  - 影片数量少
  - 异常数据

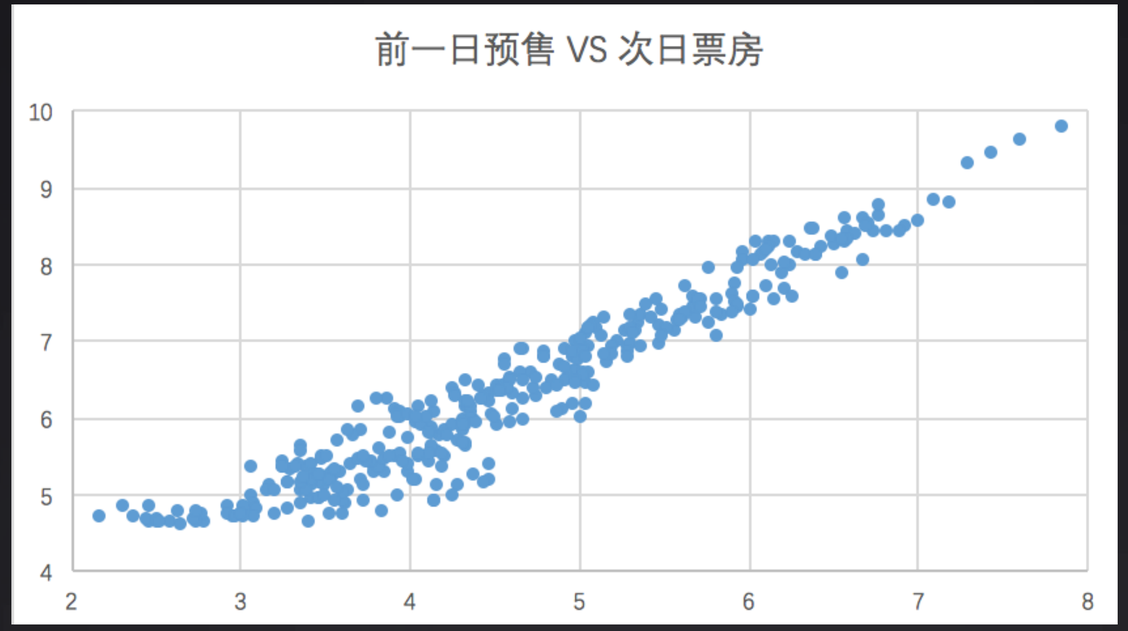


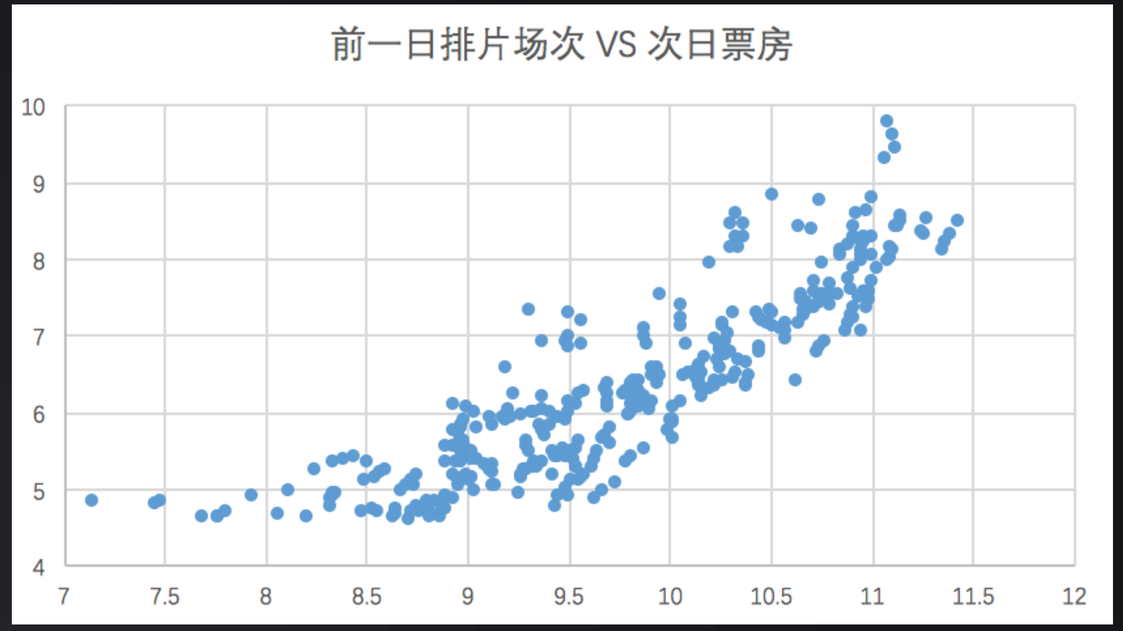
天级票房预测

- 基本假设
  - 票房 = 票价*人次 = 票价*场次*场均人次
  - log(票房) = log(票价)+log(场次)+log(场均人次)
  - y=w1*x1+w2*x2+...

天级票房预测

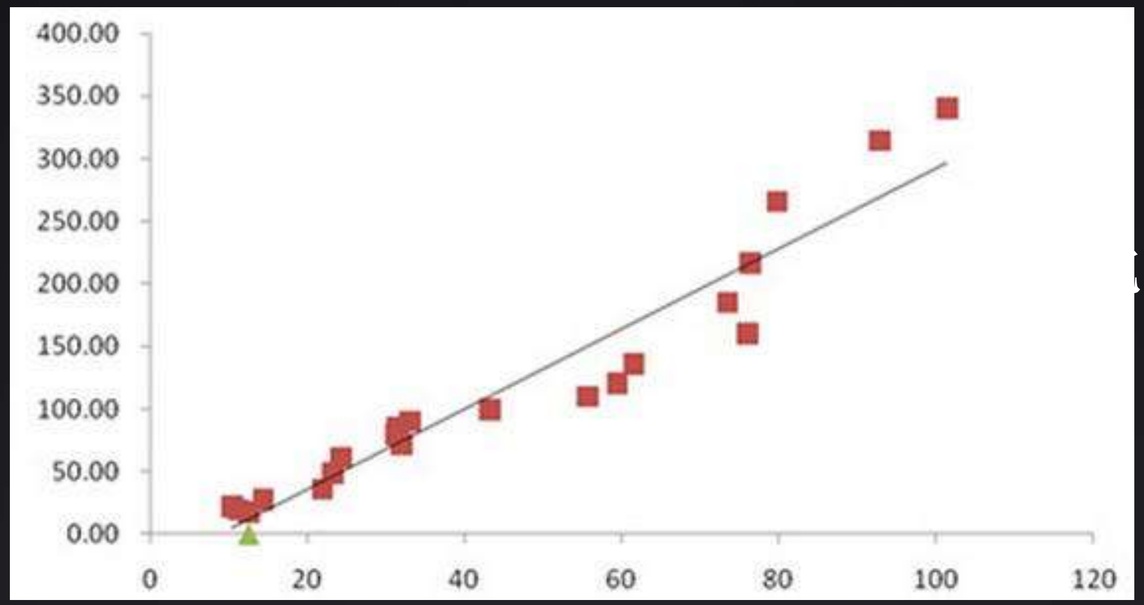
- 基本假设
  - 票房 = 票价*人次 = 票价*场次*场均人次
  - log(票房) = log(票价)+log(场次)+log(场均人次)
  - y=w1*x1+w2*x2+...

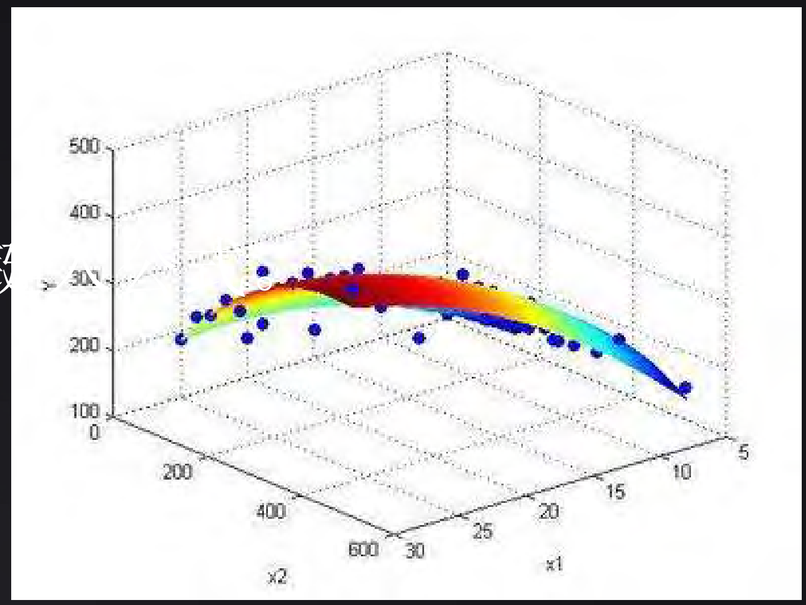




天级票房预测

- 线性回归模型





$$f(X) = W^T X + b = w1 * x1 + w2 * x2 + \cdots + wm * xm + b$$

$$\text{loss} = \frac{1}{n} \|Y - W^T X\|^2$$

天级票房预测

- 模型算法
  - 多元线性回归（LR）
  - GA特征选择算法

特征大类	特征小类	特征			
实时票房	日票房	dailyBox(t-1, t)	dailybox(t-2, t)	dailybox(t-2, t-1)	dailybox(t-1, t-1)
	平均票价	avgPrice(t-1, t)	avgPrice(t-2, t)	avgPrice(t-2, t-1)	
	上映场次	totalShow(t-1, t)	totalShow(t-2, t)	totalShow(t-2, t-1)	totalShow(t-1, t-1)
	人次	totalView(t-1, t)	totalView(t-2, t)	totalView(t-2, t-1)	totalView(t-1, t-1)
	大盘票房	sumDailyBox(t-1, t)	sumDailyBox(t-2, t)	sumDailyBox(t-2, t-1)	
	大盘人次	sumTotalView(t-1, t)	sumTotalView(t-2, t)	sumTotalView(t-2, t-1)	
猫眼订单	日票房	maoyanOrder(t-1,t)	maoyanOrder(t-1,t-1)		
黄金时间	上映上次	hotShow(t-1,t)			
	座位数	hotSeat(t-1,t)			
节假日特征	节假日	holiday(t-1)	holiday(t)	holiday(t+1)	
组合特征	...				

天级票房预测

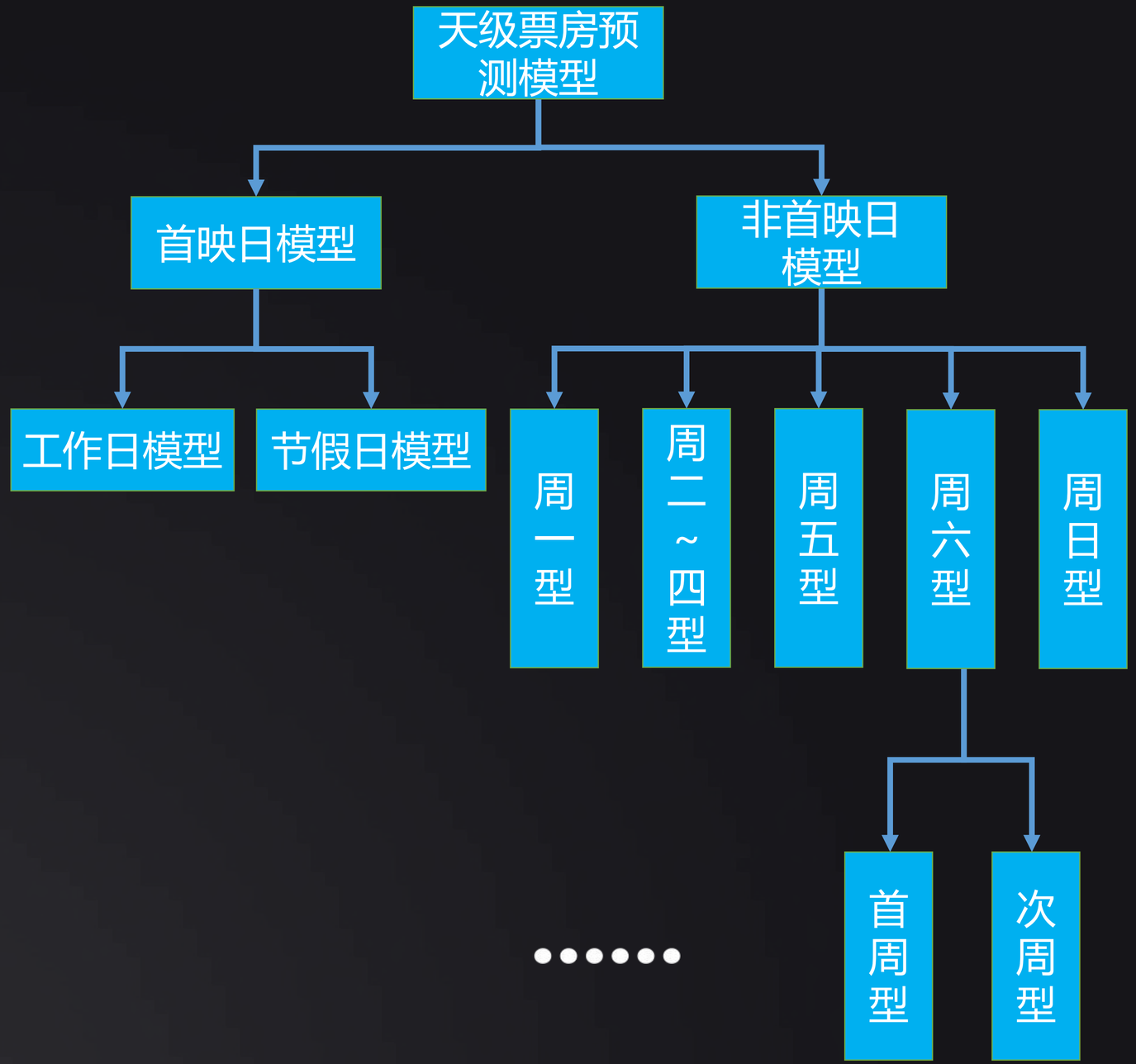
- 模型拆分


票房衰减规律
1.5
1
0.5
0
-0.5
-1
周五首映
周六高票房首周
周日首周
周一高票房首周
周二
周三
周四
周五非首映
周六高票房次周
周日次周
周一高票房次周
周二
周三
周四
标准衰减
票房系数

天级票房预测

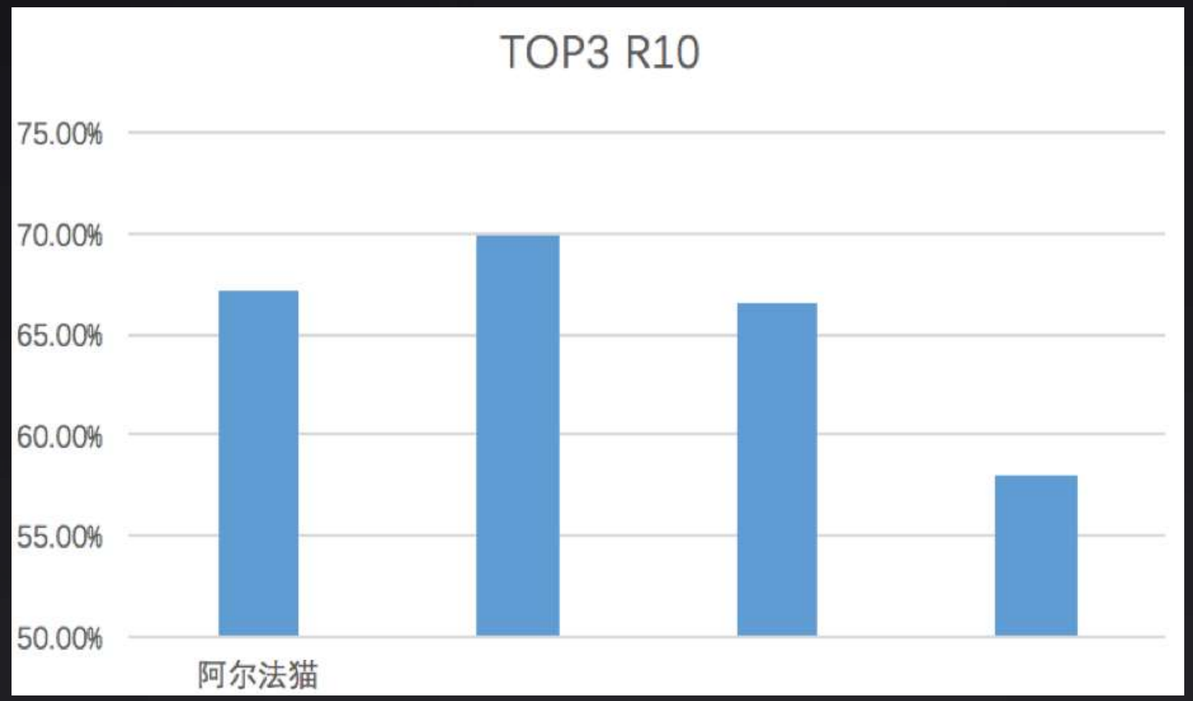
- 模型拆分

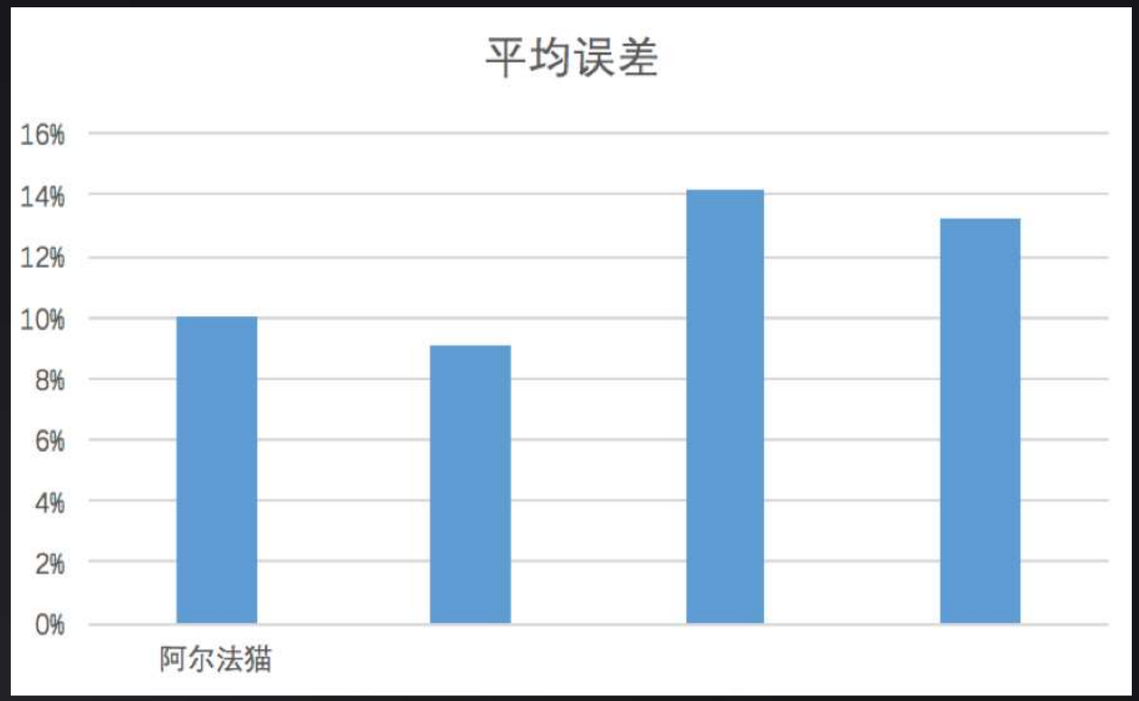

票房衰减规律
1.5
1
0.5
0
-0.5
周五首映
周六高票房首周
周日首周
周一高票房首周
周二
周三
周四
周五非首映
周六高票房次周
周日次周
周一高票房次周
周二
周三
周四
标准衰减
票房系数


天级票房预测模型
首映日模型
非首映日模型
工作日模型
节假日模型
周一型
周二~四型
周五型
周六型
周日型
首周型
次周型
……

天级票房预测

- 效果对比



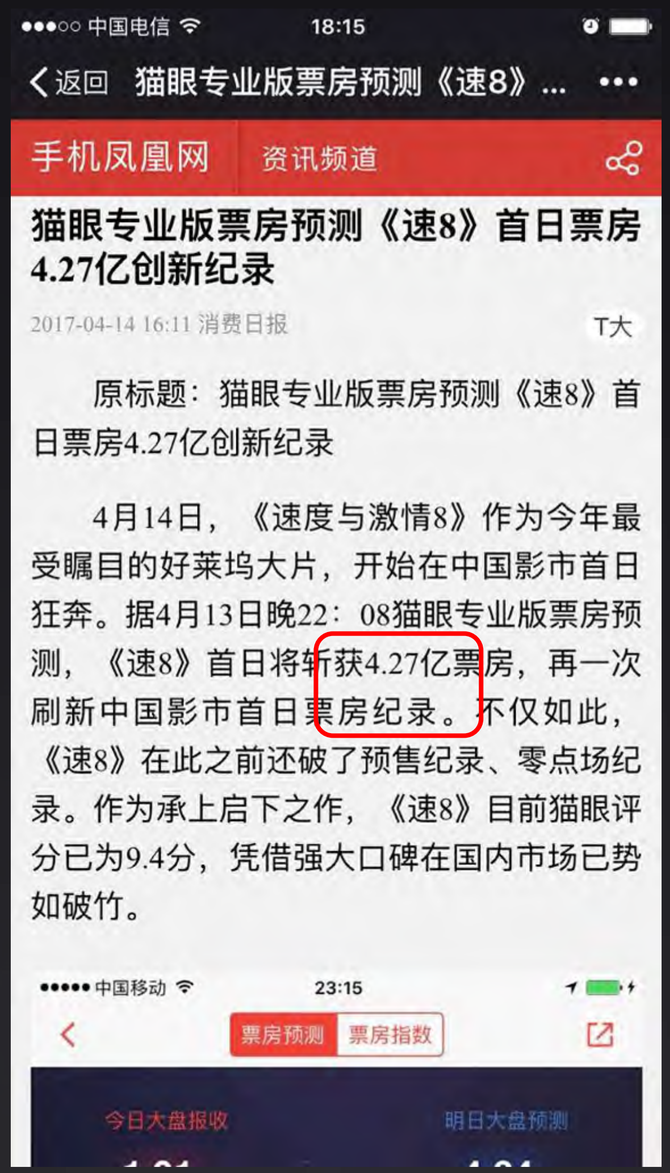


猫眼比竞对提前发布12个小时

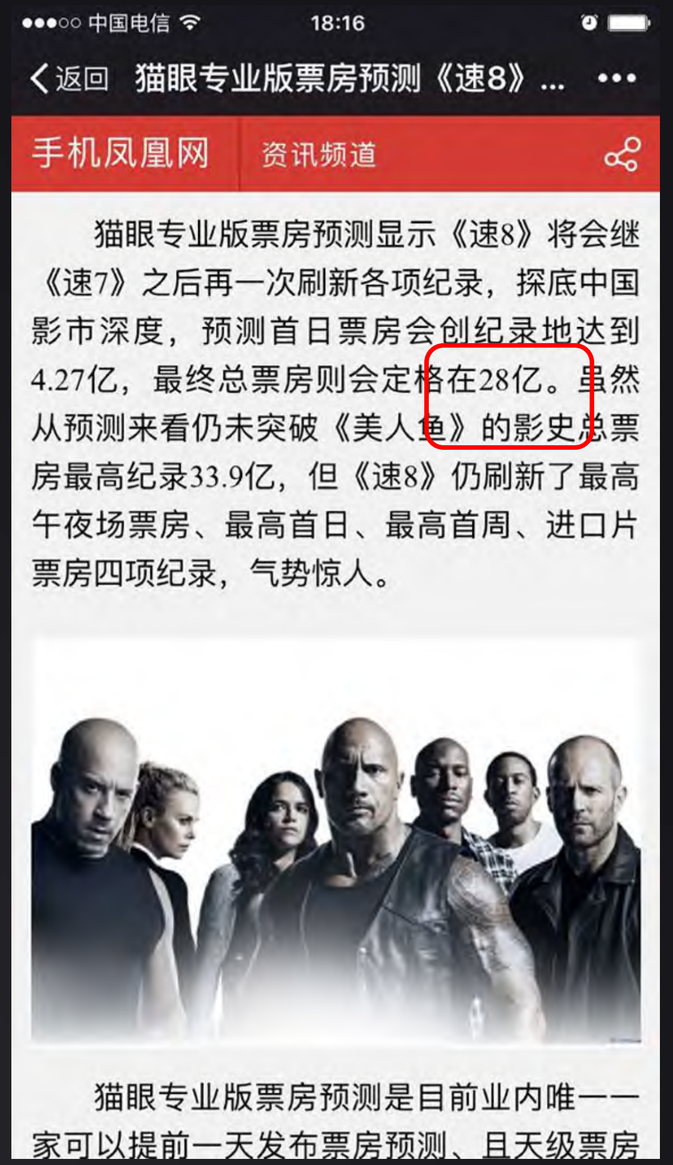
R10=预测误差在10%以内的样本占比

天级票房预测

首映日准确度98%



总票房准确度95%



总票房预测

总票房预测


天级票房-21点预售比
天级票房-阿尔法猫
天级票房-21点预售
天级票房-专资票房
60,000
50,000
40,000
30,000
20,000
10,000
0
20170413
20170416
20170419
20170422
20170425
20170428
20170501
20170504
20170507
20170510
20170513
20170516
20170519
20170522
dt.maoyan.com

FAMILY IS MORE
THE FAST AND THE FURIOUS
IN CINEMAS APRIL 12


天级票房-21点预售比
天级票房-阿尔法猫
天级票房-21点预售
天级票房-专资票房
12,000
10,000
8,000
6,000
4,000
2,000
0
20170505
20170507
20170509
20170511
20170513
20170515
20170517
20170519
20170521
dt.maoyan.com


DANGAL
摔跤吧！爸爸
5月5日 先棒为敬

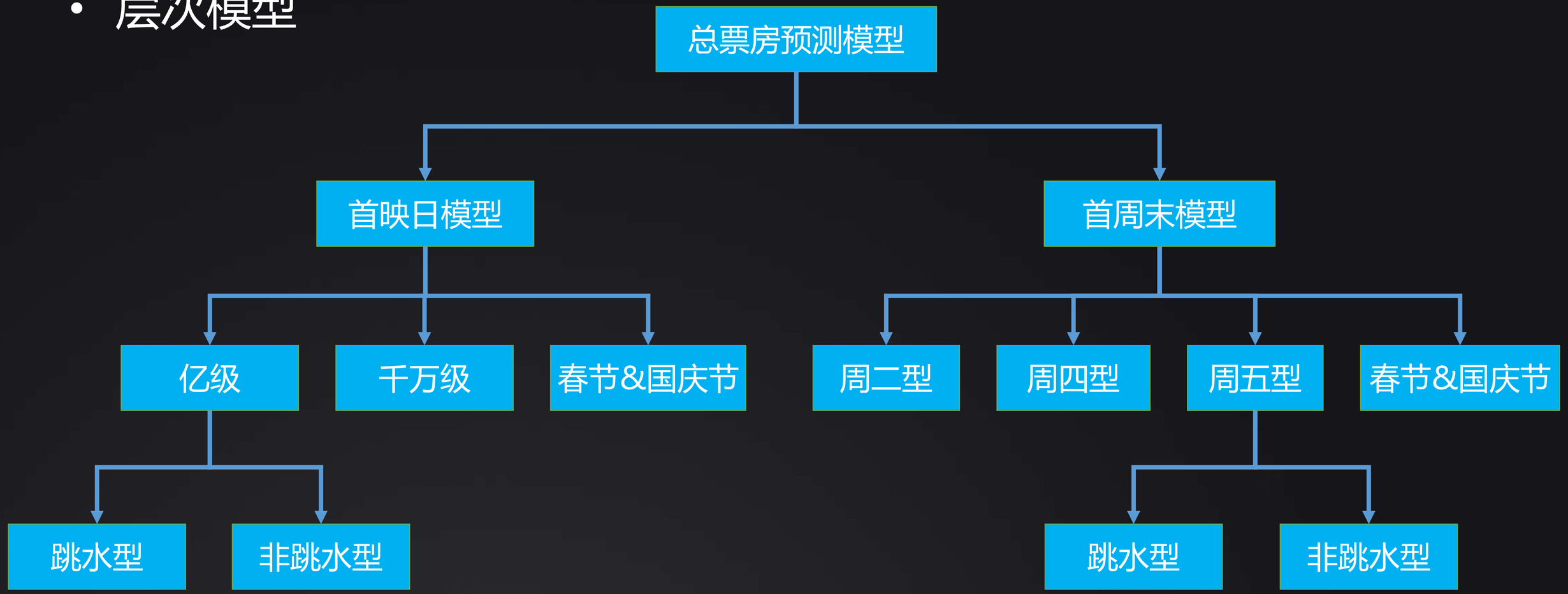
总票房预测

- 问题定义


首映日
首周末
天级票房-21点预售比
天级票房-阿尔法猫
天级票房-21点预售
天级票房-专资票房
60,000
50,000
40,000
30,000
20,000
10,000
0
20170413
20170415
20170417
20170419
20170421
20170423
20170425
20170427
20170429
20170501
20170503
dt.maoyan.com

总票房预测

- 层次模型


总票房预测模型
首映日模型
首周末模型
亿级
千万级
春节&国庆节
周二型
周四型
周五型
春节&国庆节
跳水型
非跳水型
跳水型
非跳水型

总票房预测

- 模型算法

特征大类	特征小类	特征				影响因素
天级票房	票房	box(t=1)	box(t=2)	box(t=3)	box(t=4)	票房的体量
	票房走势	(box(t=3)-box(t=4)) / box(t=3)	(box(t=1)-box(t=4)) / box(t=1)	(box(t=3)-box(t=1)) / box(t=1)		口碑
	排片比	showRate(t=1)	showRate(t=2)	showRate(t=3)	showRate(t=4)	票房的体量
	票房比	boxRate(t=1)	boxRate(t=2)	boxRate(t=3)	boxRate(t=4)	票房的体量
	场均人次	avgViewer(t=1)			avgViewer(t=4)	票房的体量
影片属性	猫眼想看数	wishNum				口碑
	猫眼评分	score				口碑
	影片类型	type				
档期	上映年月	month	year			档期
竞争	竞争影片想看数	rivalWishNum				其他影片的影响
	竞争影片数量	rivalNum				其他影片的影响

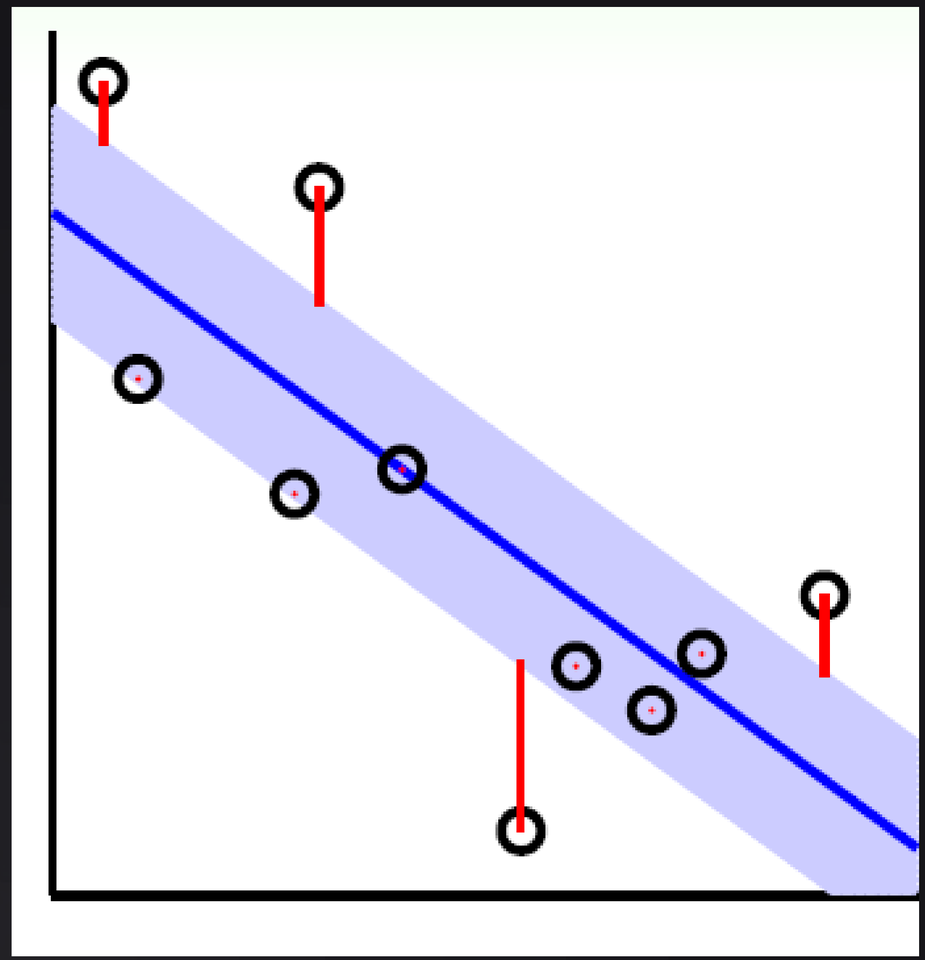
总票房预测

- 支持向量回归模型（SVR）
  - 损失函数：

$$err(x_i, y_i) = \begin{cases} 0 & |y_i - w \bullet \phi(x_i) - b| \leq \epsilon \\ |y_i - w \bullet \phi(x_i) + b| - \epsilon & |y_i - w \bullet \phi(x_i) - b| > \epsilon \end{cases}$$

  - 目标函数：

$$min \ \ \frac{1}{2}||w||_2^2 \ \ s.t \ \ |y_i - w \bullet \phi(x_i) - b| \leq \epsilon (i = 1, 2, \ldots m)$$



总票房预测

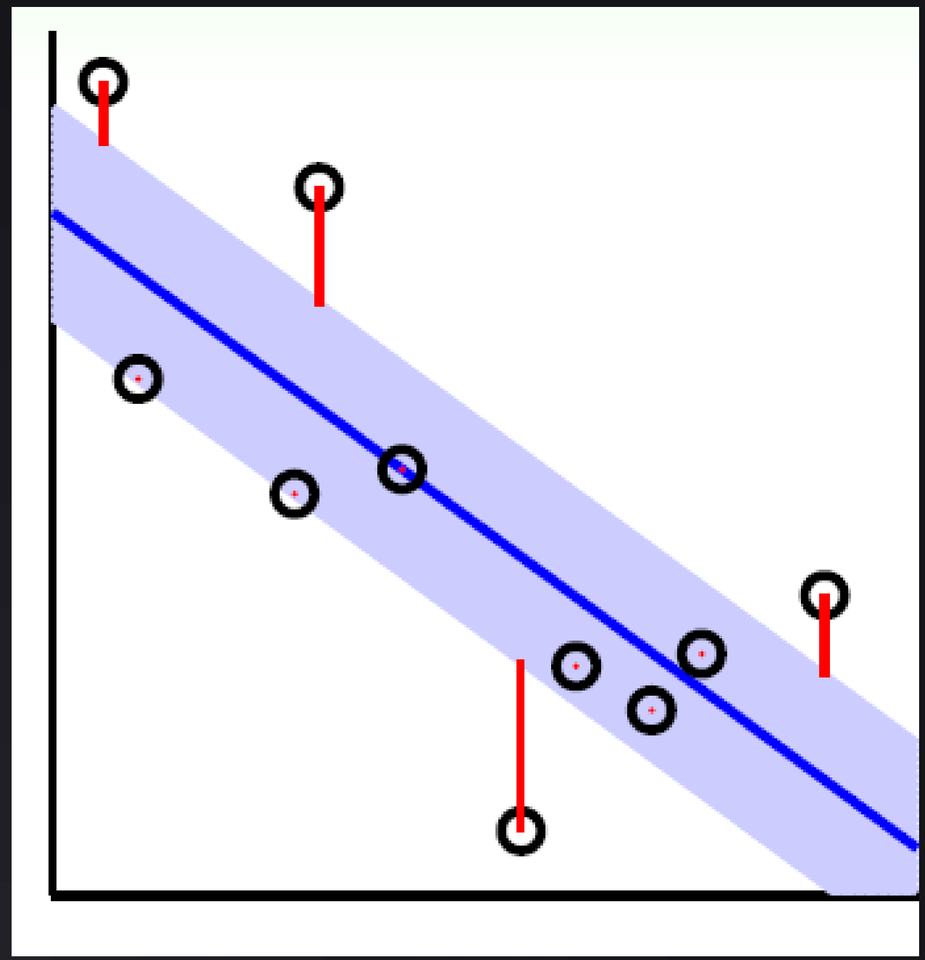
- 支持向量回归模型（SVR）
  - 损失函数：

$$err(x_i, y_i) = \begin{cases} 0 & |y_i - w \bullet \phi(x_i) - b| \le \epsilon \\ |y_i - w \bullet \phi(x_i) + b| - \epsilon & |y_i - w \bullet \phi(x_i) - b| > \epsilon \end{cases}$$

  - 目标函数：

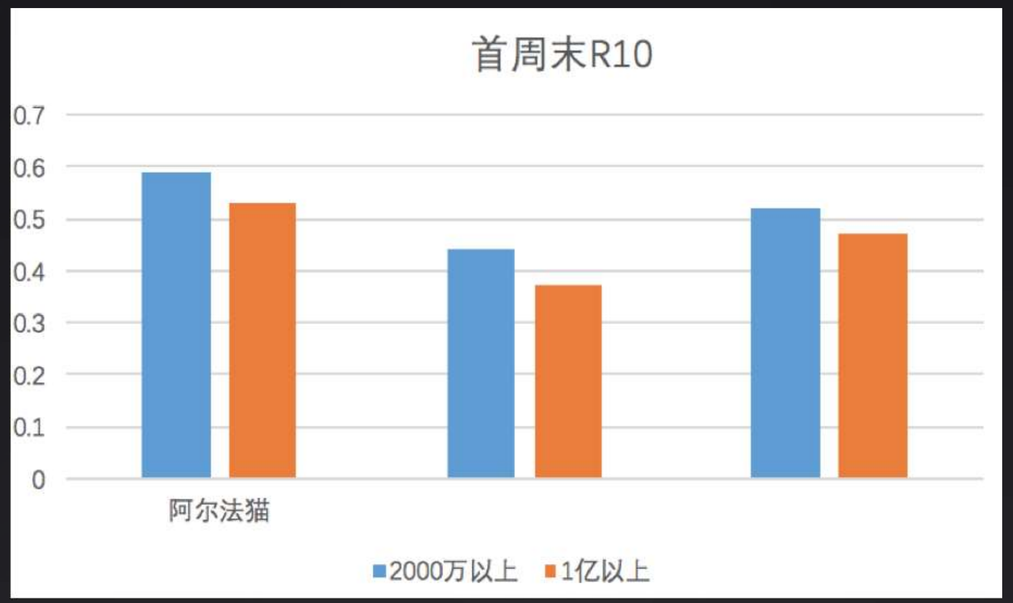
$$min \ \ \frac{1}{2}||w||_2^2 \ \ s.t \ \ |y_i - w \bullet \phi(x_i) - b| \le \epsilon (i = 1, 2, \ldots m)$$

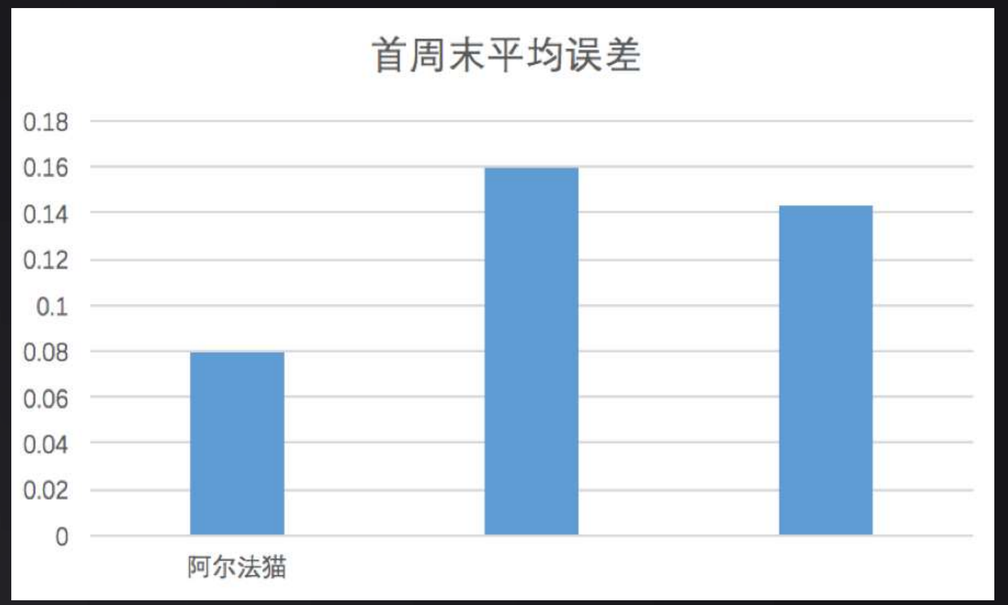
  - 优点：
    - 特征维度大于样本数时，仍然适用
    - 小样本情况下，模型泛化性强
    - 非线性核函数，可解决非线性的回归问题



总票房预测

- 效果对比
  - 数据统计时间：2017春节~2017.4月底下线，共计27部过2000万，19部过亿

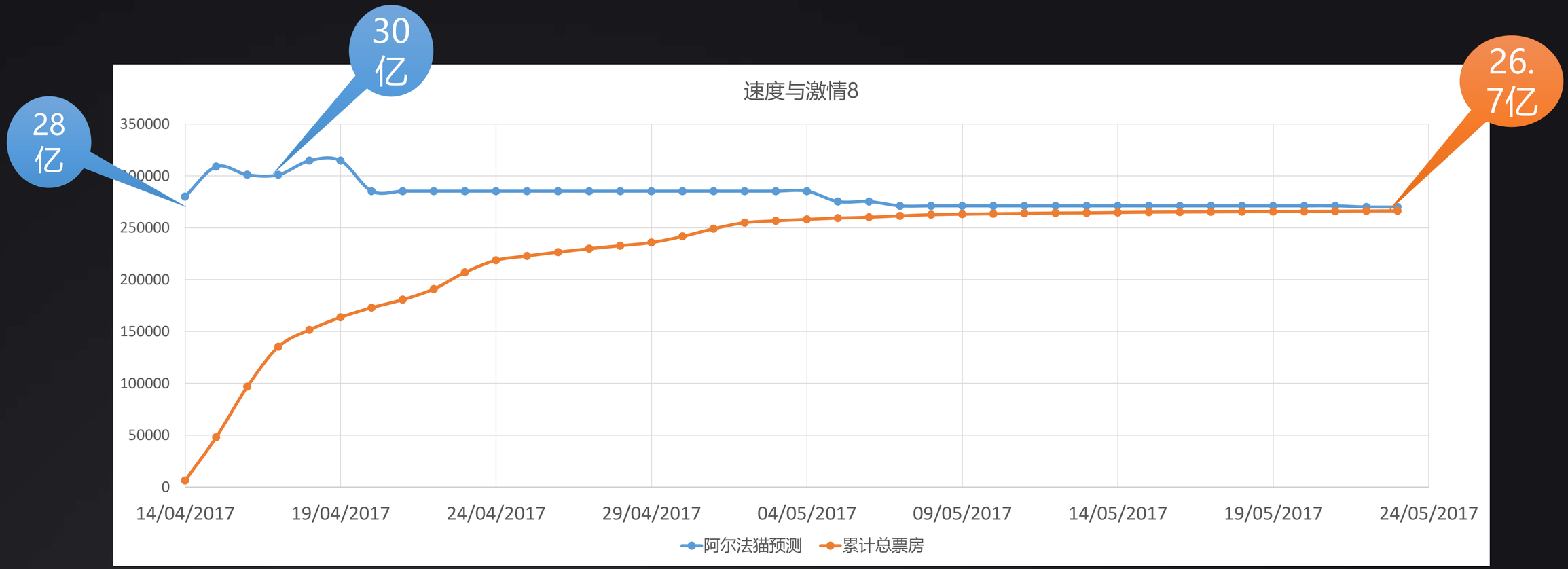




猫眼比竞对提前发布12个小时

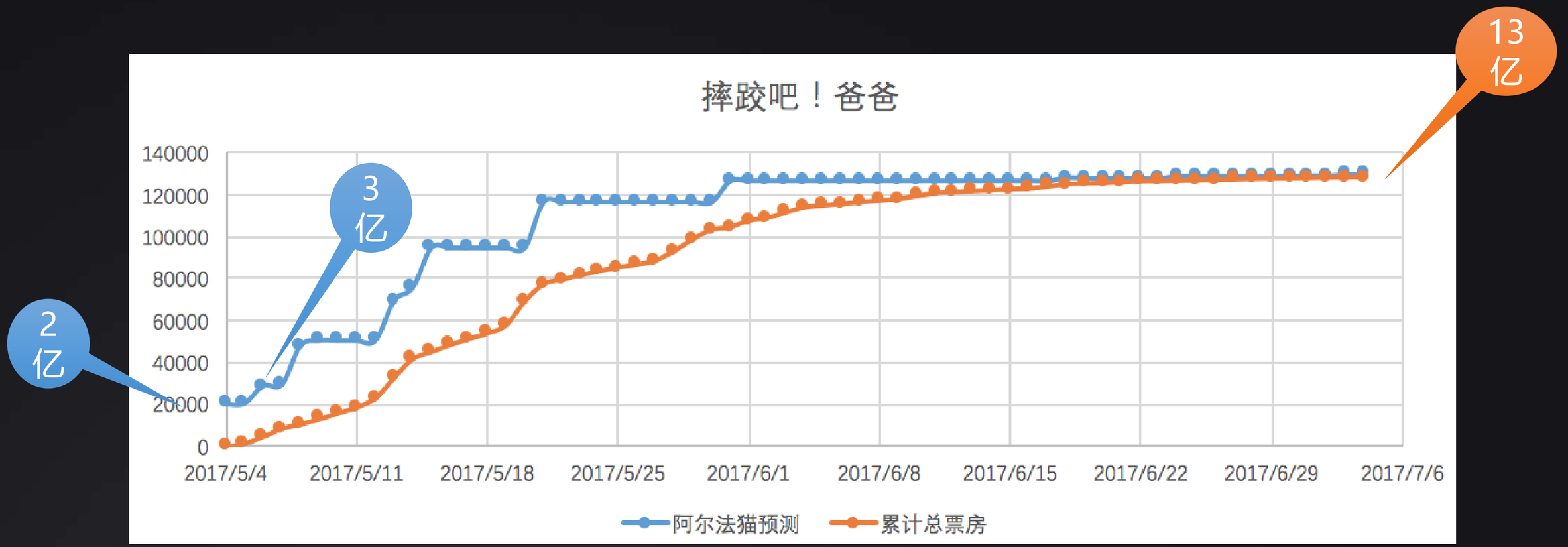
总票房预测

- 效果示例


30亿
28亿
26.7亿
速度与激情8
350000
250000
200000
150000
100000
50000
0
14/04/2017
19/04/2017
24/04/2017
29/04/2017
04/05/2017
09/05/2017
14/05/2017
19/05/2017
24/05/2017
阿尔法猫预测
累计总票房

总票房预测

- 效果示例


摔跤吧！爸爸
140000
120000
100000
80000
60000
40000
20000
0
2017/5/4
2017/5/11
2017/5/18
2017/5/25
2017/6/1
2017/6/8
2017/6/15
2017/6/22
2017/6/29
2017/7/6
阿尔法猫预测
累计总票房
3亿
13亿
2亿

内容提纲

- 行业背景
- 技术体系
- 预测算法
- 工作展望

工作展望

票房预测的基础：提高效果、提前时间点、方法创新

- 洞察票房内在规律，不断探索新的方法
- 正在尝试预测票房走势、预售比
- 交互式预测系统：总票房、天级票房
- 天级票房再提前1~2天的小目标
- 总票房预测提前1个月的大目标

应用拓展

- 影片排片：排片助手=>智能排片
- 发行营销：参与营销计划，票补的优化
- 树立票房预测的行业标杆

THANK YOU