数据与价值

数据挖掘与产业实践

目录

01 数据

什么是数据?

“凡是可以记录的都是数据!”

数字

金融投资

运动健康

天气污染

图像

医学成像

指纹识别

人脸识别

声音

SIRI

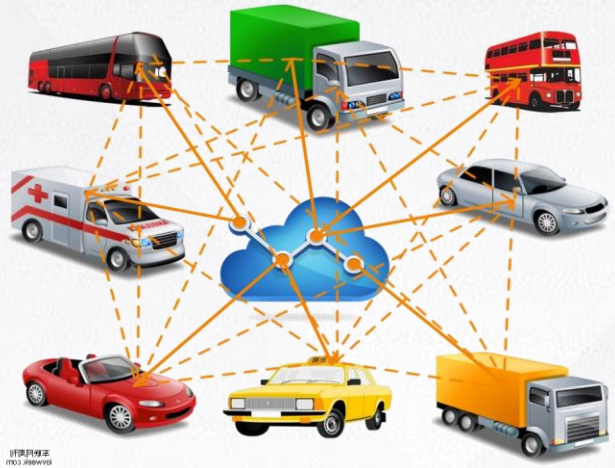
搜狗语音输入

微信语音转文字

数据时代的特征

手机

可穿戴设备



车联网

02 价值

什么是价值?

“

价值就是业务的核心诉求!

”

价值的三个表现：企业

收入

支出

风险

收入：个性化推荐

金融投资

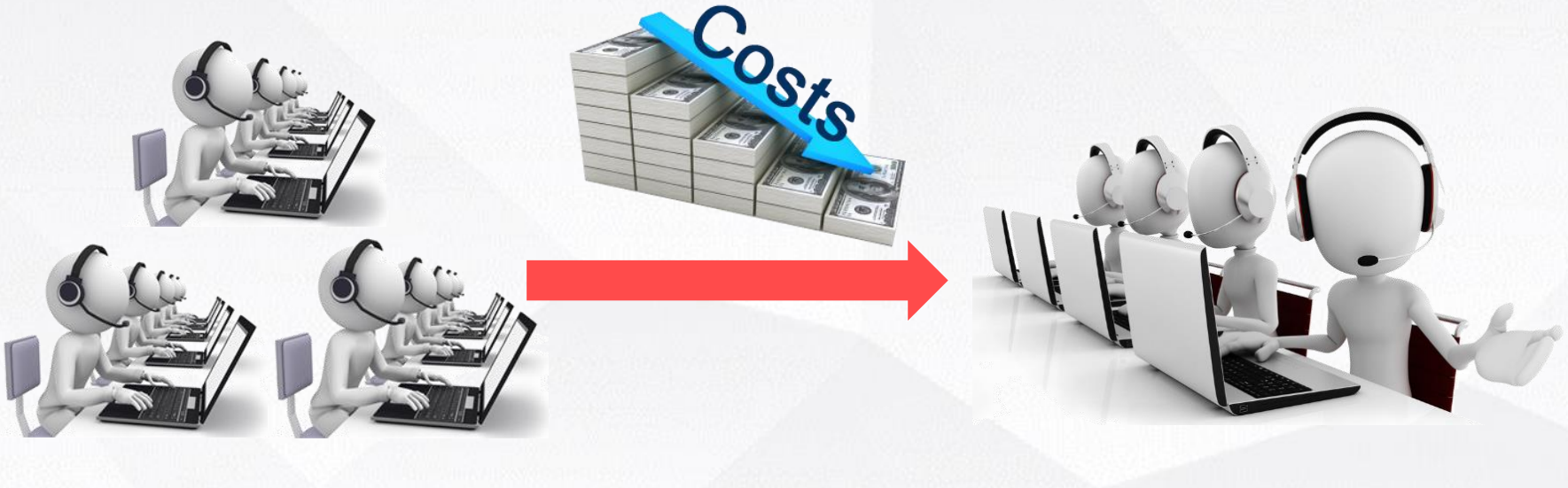
商品推荐

购买记录

收入：广告销售


Bai du 百度
美容医院
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约54,600,000个
搜索工具
为您推荐: 整形美容医院 美容 医疗美容医院 医疗美容 整形美容医院排行榜
美容医院北京伊美尔狂欢10天限时抢! - 美容医院哪个好
热点: 美容医院 优势: 量身定制美丽方案 | 一流的整形设备
美容医院伊美尔5月限时特惠，汇聚国际医师团队，塑美超过300万，19年专注整形美容，北京整形医院创新品牌，伊美尔"领跑中国医美行业"!
www.ymexf.com.cn 2016-05 - 推广 - 评价
「北京八大处VIP门诊部」专注整形59年!
热点: 美容医院 优势: 深受爱美人士的好评 | 专业为您服务
美容医院 强大的医疗美容团队，五星级整形美容服务，深受爱美人士的好评!面部年轻化..
介绍: 激光美容的介绍 治疗: 常见的乳房美容整形手术
专题: 医疗美容中的健康美白法 热门: 冰点脱毛美容的特点和效果
www.pinganzhengxing.com 2016-05 - 推广 - 评价
美容医院 亚洲连锁整形品牌 美莱领衔整形18年。
美容医院 医疗美容连锁高端品牌-美莱，汲取中韩整形精髓，整合国际医疗整形美容医生资源，提供更保障整形效果.18年26家亚洲连锁品牌，豪华舒适的诊疗环境，广受爱美女性..
www.bjmylike.com 2016-05 - 推广 - 评价
激光美容医院伊美尔专利祛斑 19年品牌保障
激光美容医院，专业祛斑国际行家亲诊，北京伊美尔特效去雀斑，黄褐斑，色斑。激光美容医院量身定制治疗方案，让求美者享受一对一美肤治疗!
www.evercarebj.com 2016-05 - 推广 - 评价
医院北京整形伊美尔整形总院!全国连锁整形美容医院!
医院北京整形，医院北京整形-优选北京伊美尔整形医院，伊美尔整形品牌发源地!伊美尔专注整形19年，手术精细无痕个性定制!医院北京整形哪家好?
百度搜索 有V有保障
相关机构
展开
北京凯尔整形美容医院
瑞韩医疗美容医院
广州美莱医院
北京叶子整形美容医院
杭州维多利亚医疗美...
北京丽都医疗美容医院
深圳希思医疗美容医院
韩国原辰整形外科医院
网友在讨论
黄金交易网
人民币外汇汇率查询
女装连衣裙批发
1.95刺影合击
给百度提建议

支出：呼叫中心



支出：制造业

风险

信用评估

收入 ↑

成本 ↓

价值的三个表现：政府

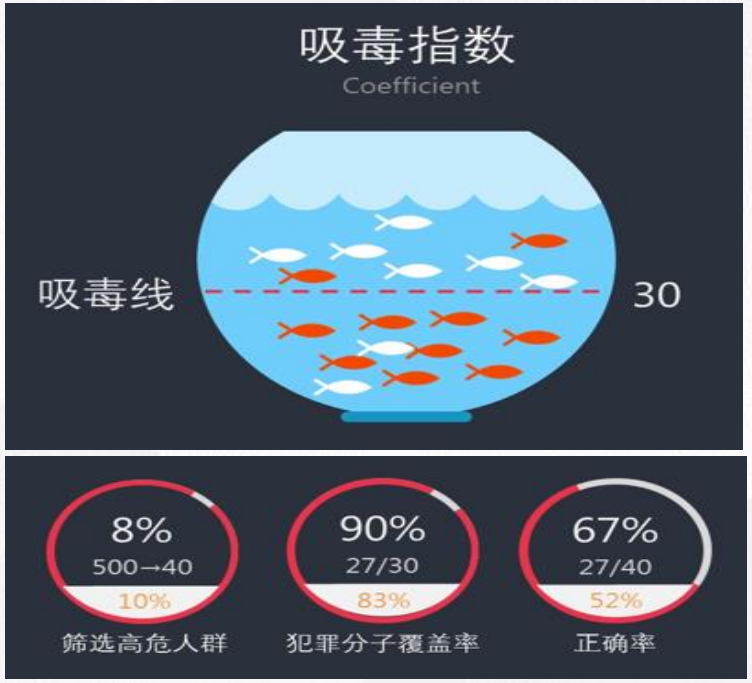
收入

支出

风险

政府：安全

2014明星吸毒队



政府：医疗

疾病诊断

新药研发

政府：就业

人尽其职

政府：教育

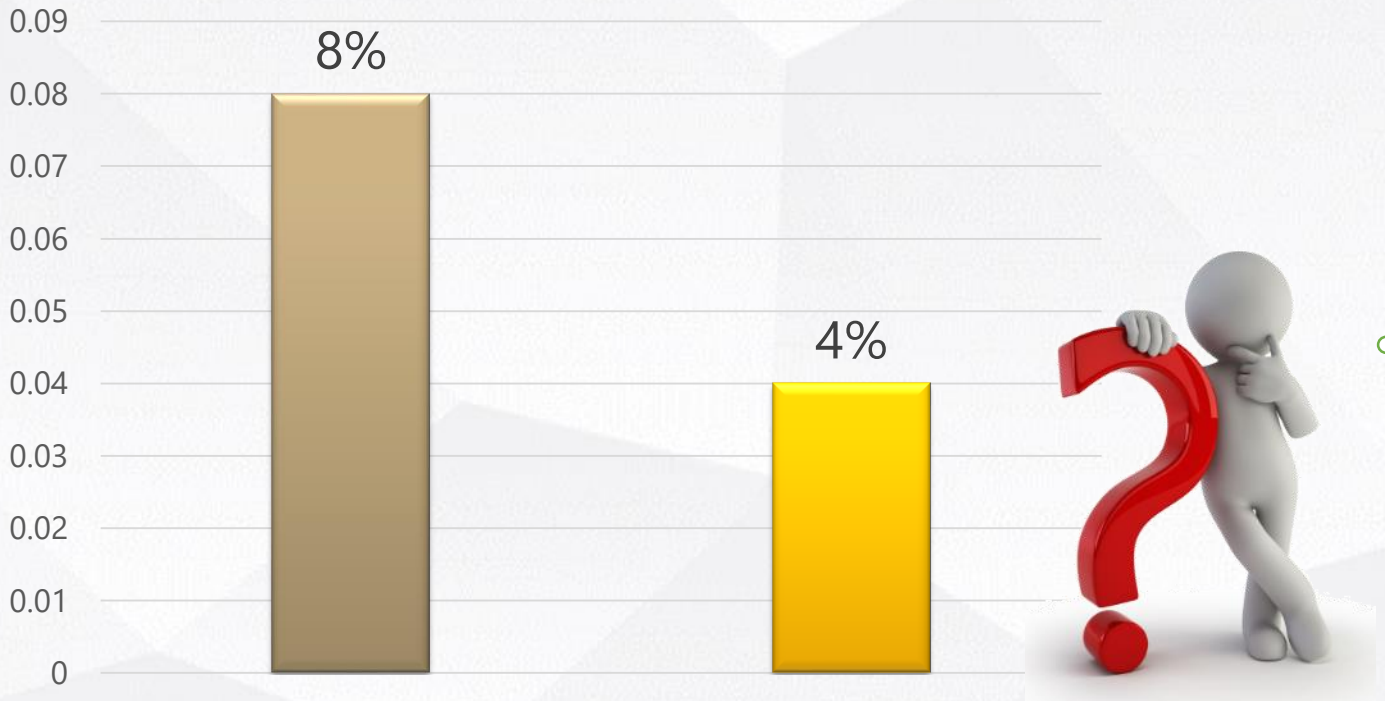
政府：环境

雾霾

人口健康、经济发展

可以量化的参考系

个性化推荐转化率



效果如何？？

03 数据挖掘

数据分析、挖掘?

“业务问题 vs. 数据可分析问题”

Y是什么？价值！

客户流失：Y = 流失与否

征信：Y = 是否逾期

保险：Y = 赔付金额

X是什么？业务知识！

客户流失：X =

在线时长

活跃程度

朋友个数

...

征信：X =

消费记录

工作背景

教育程度

...

保险：X =

出险记录

驾驶习惯

汽车状况

...

老王的淘宝店

隔壁老王开了一家淘宝店，他有1万元广告预算，他应该把钱投到哪呢?

他有三个选择

Baidu 百度

CCTV.
中国中央电视台
CHINA CENTRAL TELEVISION

And 一个问题

销售业绩

?

Baidu 百度

CCTV.
中国中央电视台
CHINA CENTRAL TELEVISION

And 一个问题

因变量：连续型

自变量

?

明确因变量

Y = UBI车险

Y = 消费价格

Y = 房价

目标（有监督学习）

- 因变量：Y
- 自变量：$X = (\mathrm{X}_1, \mathrm{X}_2, \cdots, X_p)'$

$$\mathrm{Y} = f(X) + \epsilon$$

误差项

为何要估计f？

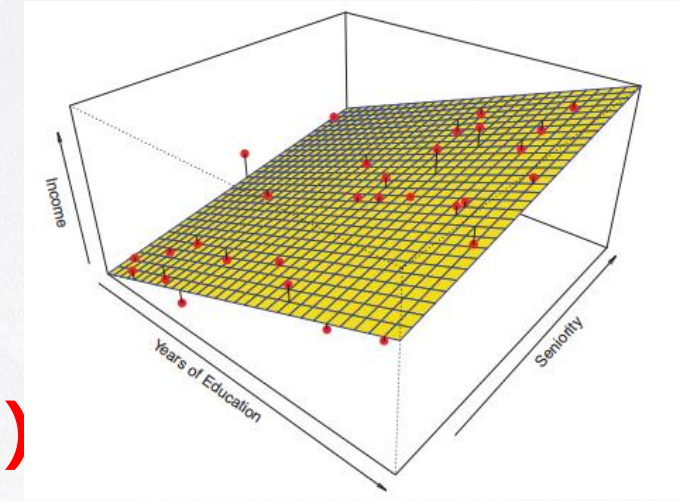
- 目标1：**预测（Prediction）** $\hat{Y} = \hat{f}(X)$

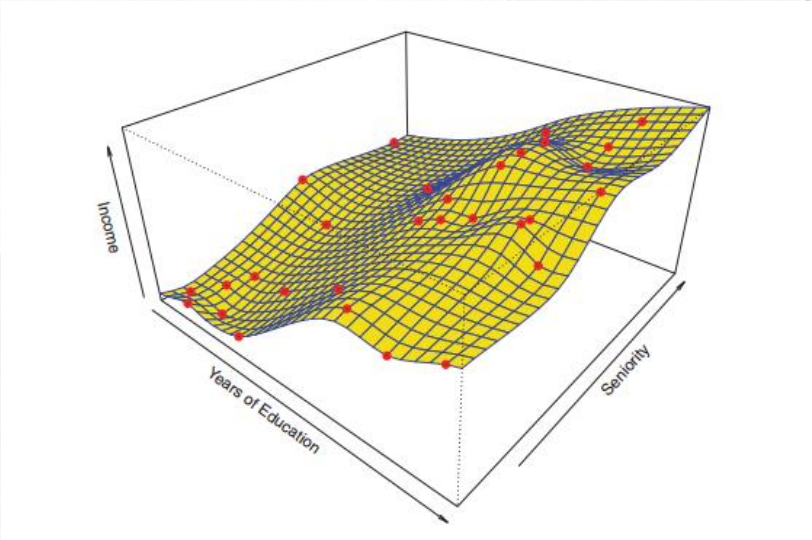
$$\mathrm{E}\left(Y - \hat{Y}\right)^2 = \underbrace{\left[f(X) - \hat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}}$$

- 目标2：**推断（Inference）**
  - 哪些自变量与Y有关？
  - 具体是什么关系？能否线性表达？

如何估计f？

- 方法1：**参数方法（Parametric Methods）**
  - 例如：线性回归模型
- 方法2：**非参数方法（Non-parametric Methods）**
  - 不假设关于f的参数形式
  - 较好的拟合性和光滑性：Spline，Kernel Smoothing
  - 缺点：容易过拟合





模型评价

- 连续型因变量：MSE（Mean Squared Error）

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

- Training MSE：在训练集评估
- **Test MSE**: 在预测集评估

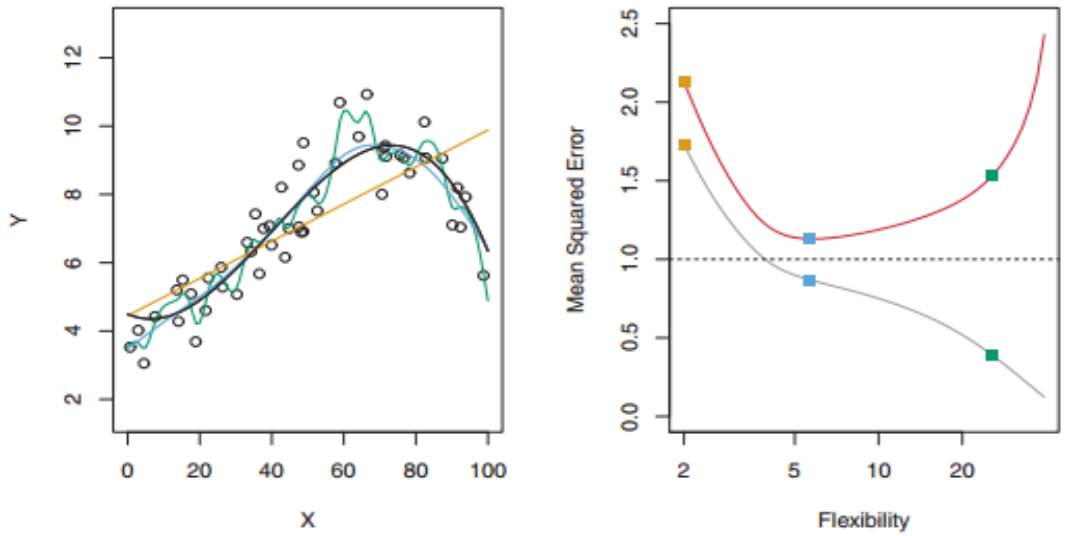


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Bias-Variance Trade-off

- Expected test MSE:

$$\mathrm{E}\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon)$$

Y：回归问题

Y：个人收入

教育程度
性别
年龄
…

Y：登录时长

好友发帖数
粉丝数
转发数
…

Y：汽车保养花费

汽车品牌
价格
车型
…

Y：分类问题

Y：是否流失

当月花费
好友个数
满意度
…

Y：是否延误

天气状况
机型
目的地
…

Y：是否被ST

资产规模
资产周转率
资产收益率
…

Y：计数

购买行为

Y = 购买次数

品牌
价格
收入
…

生育行为

Y =子女数目

收入
父母是否独生
初育年龄
…

社交行为

Y =微博数目

好友发帖数
粉丝数
在线时长
…

Y: 定序

Y：顾客满意度

品牌
功能
价格
…

Y：幸福感

收入
孩子个数
工作性质
…

Y：英语程度

受教育程度
年龄
性别
…

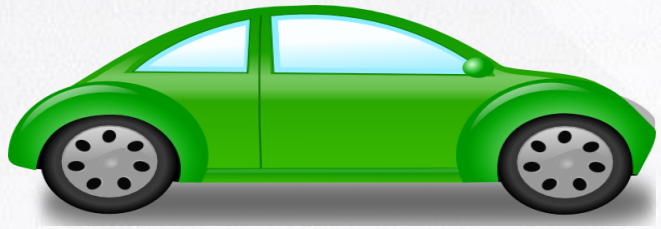
案例：某品牌轿车客户价值分析

客户价值?

Y=客户保养价值

下一期的保养花费

X：车



- 车价
- 车型
- 车龄
- 里程
- ……

X：人

- 年龄
- 性别
- 保养次数、花费
- ……

X：店

- 店内环境
- 店内促销
- 定期维系
- ……

回归分析结果

变量名	估计值	P-value	变量名	估计值	P-value
常数项	-0.162	<0.001	当期保养总花费	0.258	<0.001
车型-A	0.001	0.067	当期保养总次数	0.129	<0.001
车型-B	0.260	<0.001	当期新增里程数	0.198	<0.001
车型-C	0.113	0.159	累积购车数量	0.055	<0.001
车型-D	0.176	0.035	车价	0.143	<0.001
车型-其它	-0.094	0.273			
调整R方：36.37%					

X：店

1. 哪些变量重要？车型？保养？里程？

2. 有多重要

学习思考：客流量

学习思考：信用卡


小积分
大乐趣
699积分兑换 万达3D电影票
，您好！
以下是您昨天的信用卡消费明细：
卡号末4位 交易日期 时间 币别 交易金额 商户名称
Last 4 card numbers
Transaction date
Transaction time
Currency
Amount of original currency
Merchant name
12/11/10 09:27:06 美元 12.92 消费 一般刷卡消费
账单分期，让购物更畅快！ 点击了解详情
轻点鼠标，精彩立即展现！
299积分起兑
麦当劳套餐
amazon.cn
亚马逊网购选分期
0手续费+5倍积分
推荐亲友办卡
1000积分轻松拿，多推多送！

谢谢！