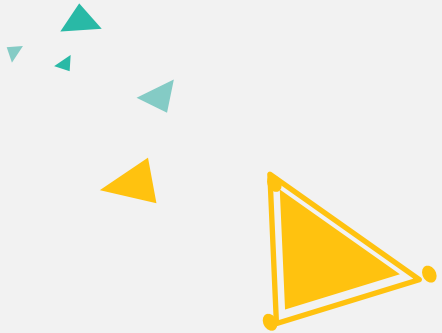
多边平台 –
理论框架 与因果学习的应用



About me

胡淏 @滴滴

- 数学、统计学 Columbia University graduate
- 多年互联网经验，现任滴滴 - 参谋部 - 资深数据专家
- 携程老粉丝，上次介绍 《模型优化不得不思考的几个问题》
- Machine Learning -> Data Analytics & Econometrics

摘要 & 大纲

我们通常会面临两个头疼的问题：项目目标应该是什么？这个项目在企业层面带来什么价值？这个看起来宽泛的**项目管理问题，实际上也是一个可以量化的问题**。

谈到机器学习，我们通常关注 *给定x，预测y* 的问题；本次内容讨论我们平时容易忽略掉的 *给定y，预测x* 一类问题。这是一类偏战略分析、运营规划的问题，**解决这类问题不仅仅需要技术能力，也需要对商业模式清晰的思考、对项目位置以及价值的正确定位** – 这构成了本次内容要涵盖的三个主体：

1. 【分析框架】 多边平台的经济学框架

2. 【项目定位】 模型的两类应用、价值、风险

3. 【技术选型】 因果推断与机器学习

我们身边无处不在的市场（多边平台）

多数互联网公司--无论是BATJ，独角兽，还是那些风口浪尖的企业，都可以类比成或大或小、或纷乱或简单的**市场**。在这个市场上都有卖家和买家双方，也是经济学中的供给端和需求端，而平台的作用就是去撮合买家和卖家之间的交易。

市场/多边平台

平台的作用：在供给端和需求端做**交易的撮合**

平台要解决传统行业 交易成本过高 的问题

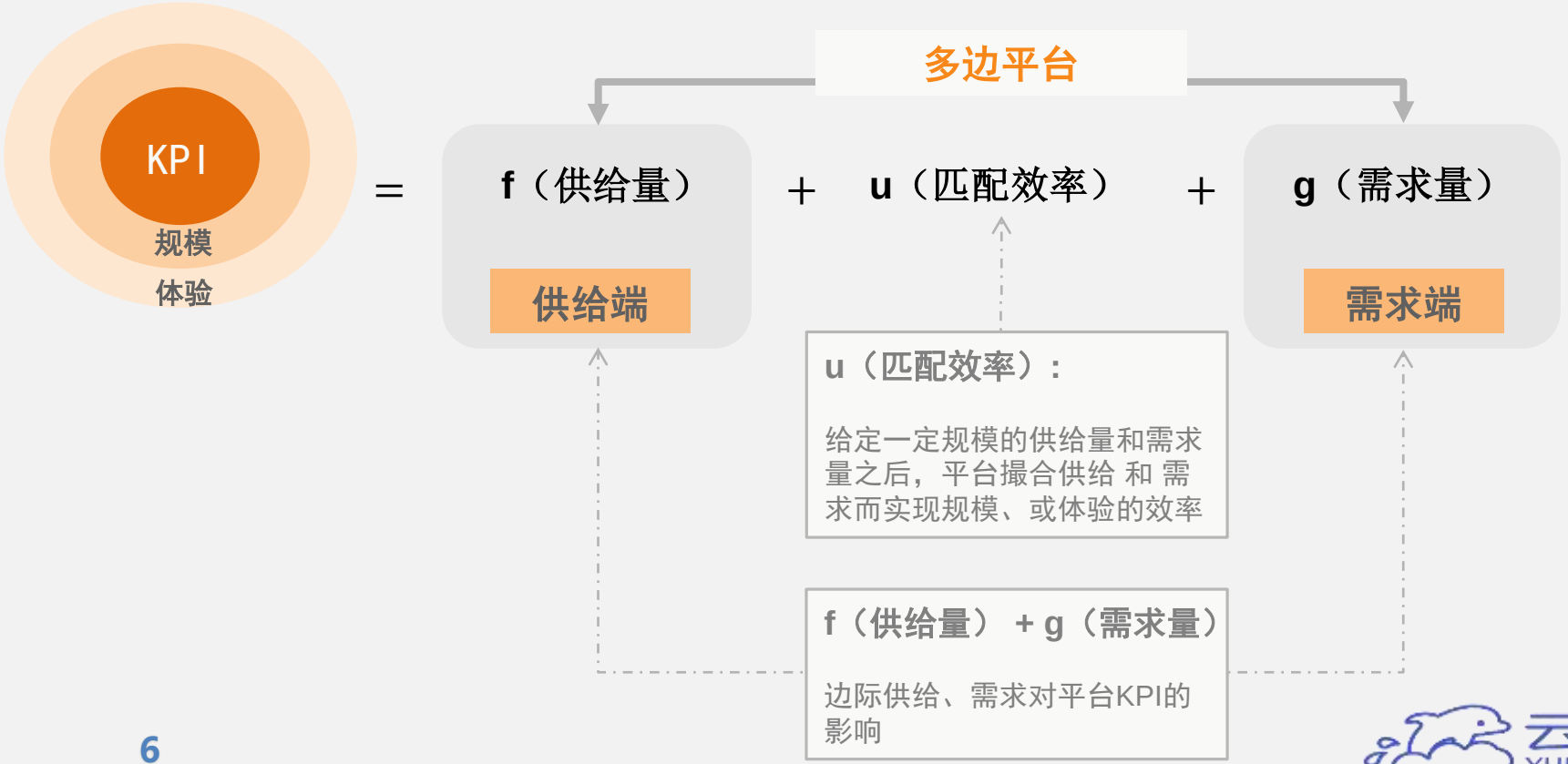
根据交易的三个阶段，**把交易成本划分为匹配成本，谈判、缔约成本，执行成本。** *Dahlman, Carl J*

交易过程

	出租车	滴滴
匹配成本	出门扬招； 视野范围内司乘匹配	根据司乘位置以及出行需求，线上匹配
谈判、缔约成本	加价； 拒载	平台统一定价； 快车、专车平台派单
执行成本	绕路； 中途拼人、违约	司机服务分； 出行安全、品质管控

- 平台在做供需之间交易撮合的同时，利用互联网+数据+算法来**降低交易成本**，这有利于供需规模进一步扩大，
- 平台通过收取平台费，对交易抽成等方式去获利.

多边平台的经济学框架 – 供 + 需 + 匹配效率



从 理论框架 看 日常项目

=

多边平台

f（供给量） +

- *满房自动开房模型*
- *库存管理*
- *反作弊*
- *补贴拉新*

u（匹配效率） +

- **搜索、推荐、分单调度**
- **定价、数据产品类**
- **识别类，行为预测类**
- **单量预测类、用户点评**
- **计算广告类**

g（需求量）

- *用户增长*
- *市场营销*
- *渠道分析*
- *流失分析*

难点：

1. 企业内部的项目是相互耦合、重复、甚至冲突的。如何量化一个项目、产品、业务线对平台的**边际价值**？
2. 是否可以通过机器学习的手段去学习：**KPI = F （ 算法$_A$，项目$_B$…，产品$_A$，抓手$_B$…，运营$_A$，活动$_B$…）**？
3. 如果这样的模型可以被训练出来，它的潜在价值和风险是什么？应该采用什么技术选型？如何去优化模型？

模型的两类应用 – 预测 & 优化

回归问题	Linear R	LASSO	glmnet ...
分类问题	Logistic R	CART	GBDT ...

$$Ground\ Truth:\ Y = f(x_1, x_2, \dots, x_p)$$
$$Estimated\ Function:\ \hat{Y} = f^*(x_1, x_2, \dots, x_p)$$

1 预测：给定x，预测y

- **应用价值：** 预测、识别本身存在价值，快速、高精度预测提高效率并降低人力成本
- **范围：** 绝大多数机器学习落入这个范畴。满房率预测，外卖配送时间预估，图像、语音识别、Google流感趋势预测、Walmart啤酒与尿布.....
- **优势：** 可容忍一个black box model，因此降低了对Y理论框架的要求 和 特征工程的难度。一个高精度的模型，$f(x_1, x_2, \dots, x_p)$ 与 $f^*(x_1, x_2, \dots, x_p)$ 可以相差很远，甚至系数的符号的相悖的！
- **难点：** 极差的业务可解释性和因果推断的能力！ and... 预测的准，so what?

2 优化：给定x的范围，寻找最优y下对应的x*

- **范围：** 通常 x 是一个可变动的产品、策略抓手，而y则通常是我们追求的业务指标/KPI。
  - case study a. 外卖配送时间算法精度 (12:30: 12.29>12.31) => 用户体验 ~ f(精度, 误差是否为正)
  - case study b. 定价 (GMV = f(price) s.t. price>=cost)
  - 我们正在调参的模型，参与的项目，服务的某个产品，所在的业务线，都可以通过特定的量化方式成为这类优化问题中的x，而业务目标就是我们要去优化的y
- **应用价值与优势：** **KPI = F （ 算法$_A$，项目$_B$...，产品$_A$，抓手$_B$...，运营$_A$，活动$_B$...），寻找X = x* 去最优化KPI**
- **难点与风险：...**

优化类应用的难点与风险 – 相关性 不等于 因果性

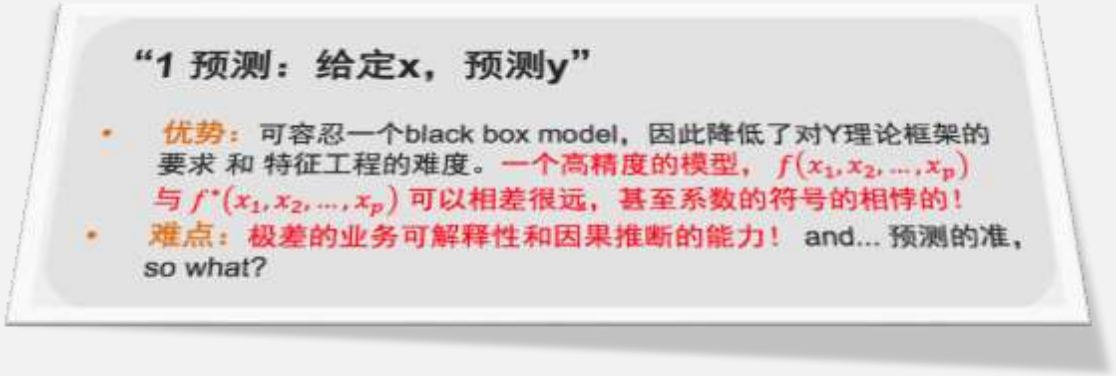
相关性 ≠ 因果性 !

Google 流感预测 无法预防流感

啤酒与尿布的质疑

身边的定价问题：

业务问题：如何通过定价来提升GMV （GMV = 总流水 = 单量*单均价）
Ground Truth：GMV = 1.2 – 0.1*价格
数据：每天的订单量和价格的信息
建模及结果：基于regression的相关性分析，发现，GMV* = 0.3 + 0.2*价格



难点与风险的本质：

模型通常在学习 *相关性*，而非 *因果性*。不过，优化类问题依赖因果关系

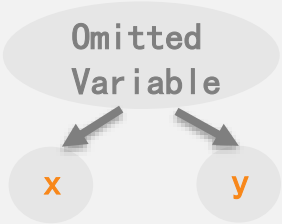
因果性与相关性差距的本质 – Omitted Variable

因果性

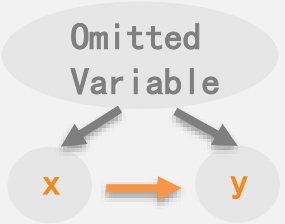
干预（x）纯粹的（排除其他所有因素的）作用效果（y）

相关性

y的发生通常伴随着x的发生



OV就是供需关系：
供不应求通常导致高价、交易高峰期



GMV = 0.3 + 0.2*价格

GMV = 1.2 – 0.1*价格

Advice:

1. 因果性 = 相关性 + omitted variable
2. 当你在解读模型中y与x之间的关系时，思考两个问题： a. x如何导致y？又是什么因素导致x？ b. 这些因素是否也会导致y？如果是，把这些因素引入到你的控制变量里面去。
3. 平台最上层的三个核心： f（供给量） + g（需求量） + u（匹配效率）

让模型开始学习因果性 – 机器学习 + 因果推断

	Omitted Var	技术选型的进步
同环比	除了时间因素被控制，其他都还是OV...	NA
AB testing	通过随机分流阻断了OV和x之间的关系	系统性评估AB两类干预的好坏；
Econometrics	通过引入OV，工具变量，或者PSM、DID、RDD等方法去量化因果关系	不依赖实验的方式对离散、连续类干预的作用；
Machine Learning + Causal Inference	不如计量的方法那么灵活，需要找到OV并引入模型作为控制变量	对异质性的探索，做细分人群、场景的因果推断

Reference:

1. Econometrics: 《mostly harmless econometrics》
2. Machine learning + causal inference （with R package~）:
   - Causal Forest: Estimation and Inference of Heterogenous Treatment Effects using random Forest
   - Causal Impact: Inferring causal Impact using bayesian structural time-series models

QA

Thank You!