



AI选房中深度学习的实践及优化

周玉驰

贝壳找房 - 数据智能中心 - 策略算法部

想做团队的领跑者 需要迈过这些“槛”

成长型企业，易忽视人才体系化培养
企业转型加快，团队能力又跟不上

VS

从基础到进阶，超100+一线实战
技术专家带你系统化学习成长

团队成员技能水平不一，
难以一“敌”百人需求

VS

解决从小白到资深技术人所遇到
80%的问题

寻求外部培训，奈何价更高且
集中式学习

VS

多样、灵活的学习方式，包括
音频、图文 和视频

学习效果难以统计，产生不良循环

VS

获取员工学习报告，查看学习
进度，形成闭环



课程顾问「橘子」

回复「QCon」
免费获取
学习解决方案

极客时间企业账号 # 解决技术人成长路上的学习问题

自我介绍

周玉驰

- 硕士毕业于中科院
- 先后就职于华为，百度和医渡云
- 目前就职于贝壳找房
- 主要负责两个方向
 - ✓ 房源策略算法
 - ✓ 房客人关系图谱



扫一扫二维码图案，加我微信

目录

- 为什么要做AI选房
- 如何做AI选房
- 模型演变历程
- 实践应用
- 总结&思考

为什么做AI选房？

贝壳找房发展&挑战



1.87亿
房屋



3000万
月活跃用户



20万
经纪人



2.1万
门店

挑战

- 找到好房难度大，成本高
- 需要强大的房源质量盘点工具



200万
贝壳全部房源



98
门店平均房源

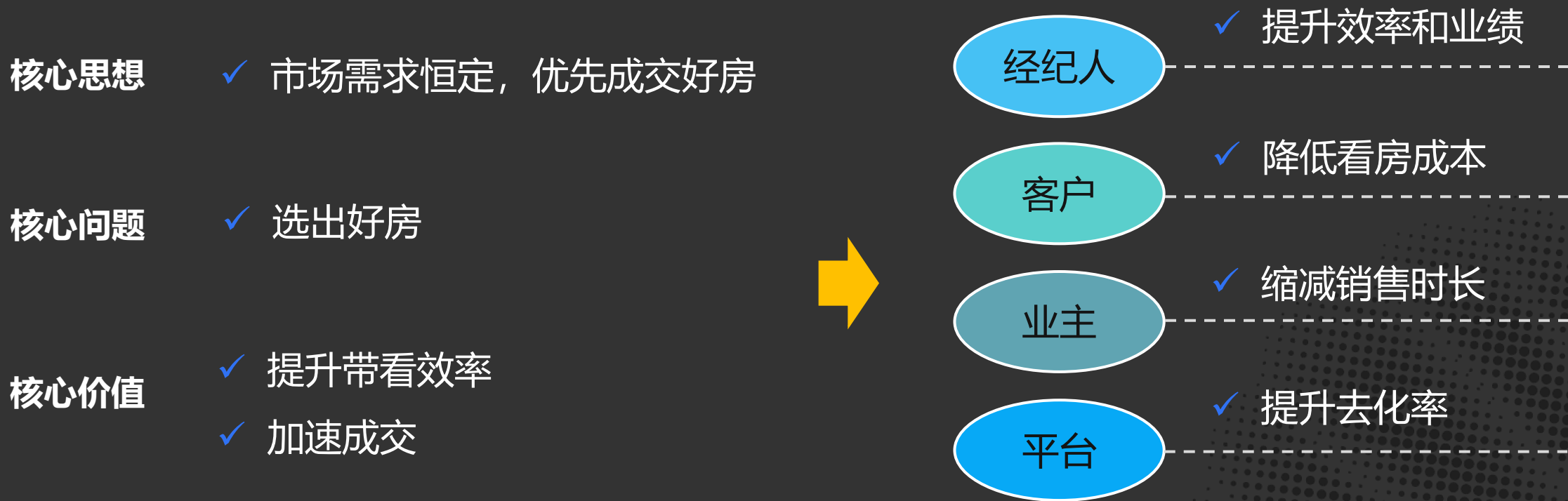


10-25
经纪人熟悉房源



70%
跨店成交占比

目标&价值



人工选房方法

人工选房流程

- 每周举行周例会
- 讨论并投票选出好房

存在问题

- 选房成本高
- 选房带有主观性
- 无法盘点所有房源质量

人工选房标准



AI选房本质上是TopN排序问题

AI选房 - 房源质量打分

好房定义

以成交为导向

- ✓ 近期能够成交的房子就是好房
- ✓ 近期成交概率越高，房源越好

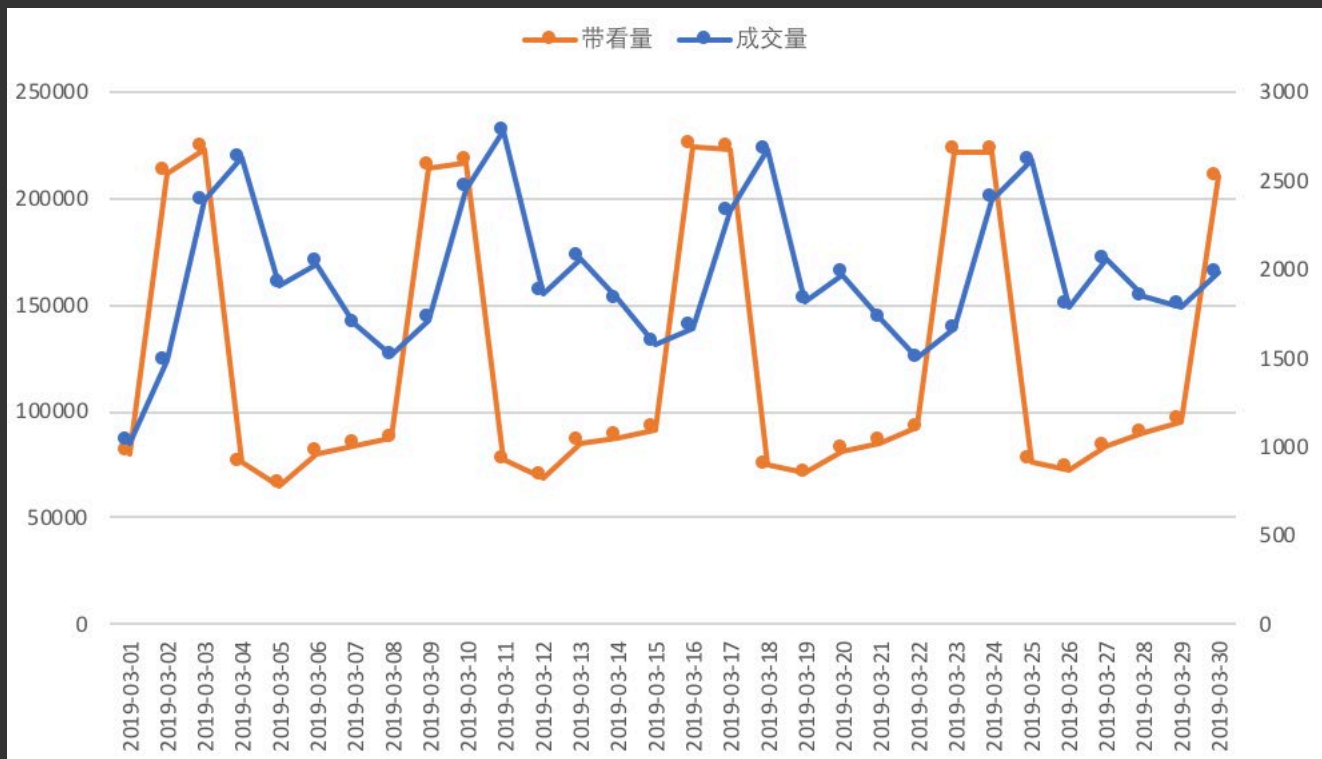
AI选房建模

$$Y = f(X)$$

- ✓ Y: 未来? 天能否成交
- ✓ X: 最近? 天房源产生的所有行为
- ✓ 样本: 挂牌满? 天的房源

AI选房建模

作业周期性分析



- ✓ 成交/带看具有周期性
- ✓ 周期性单位：周

时间选择：周的倍数

2019年3月每天的成交量和带看量

AI选房建模

$$Y = f(X)$$

- ✓ Y: 未来? 天能否成交
- ✓ X: 最近? 天房源产生的所有行为
- ✓ 样本: 挂牌满? 天的房源

why?

why?

- 时间太短: 信息传递不充分
- 时间太长:
 - 中间出现其他原因导致成交
 - 无法及时反馈效果
- 综合考虑, 并对比测试: 选择2周
- 时间太短: 行为信息不足
- 时间太长: 浪费机器资源
- 对比测试: 选择2周

AI选房建模

$$Y = f(X)$$

- ✓ Y: 未来? 天能否成交
- ✓ X: 最近? 天房源产生的所有行为
- ✓ 样本: 挂牌满? 天的房源

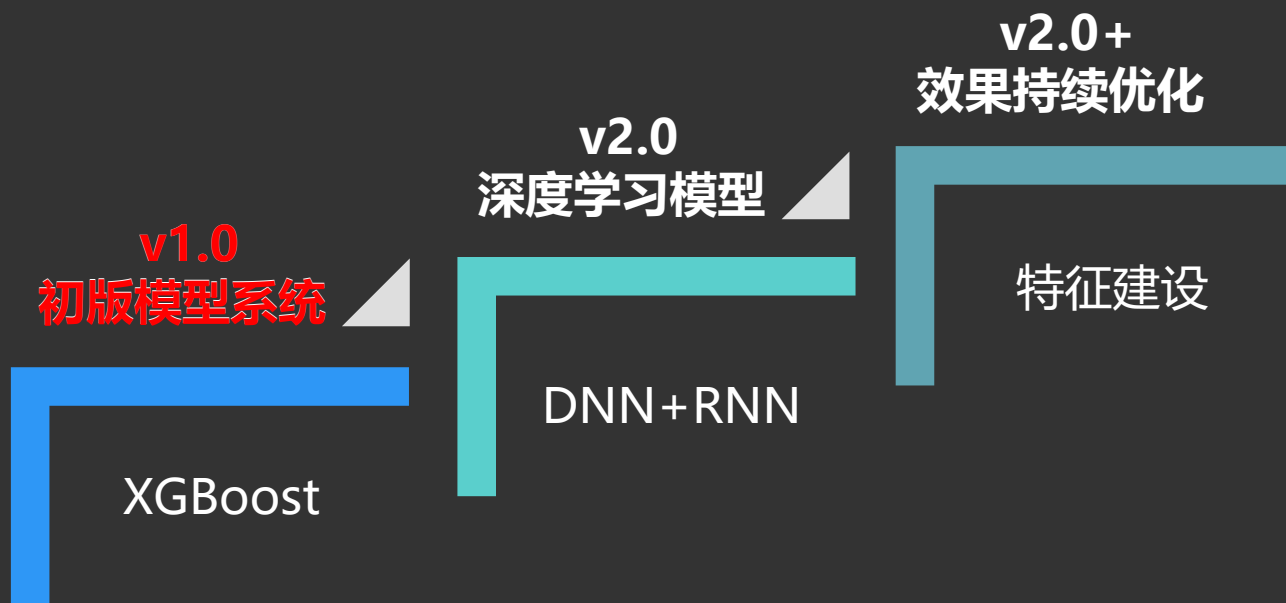
why?

- 行为特征选择14天进行聚合
- 挂牌不足14天房源, 行为特征信息不足
- 结论: 选择挂牌满14天的房源

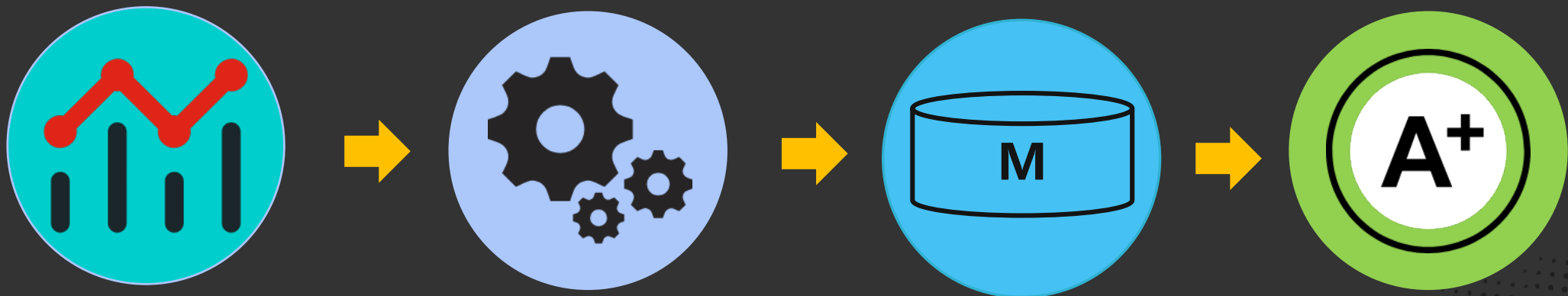
模型演变历程



模型演变历程



v1.0 - 初版模型系统概览



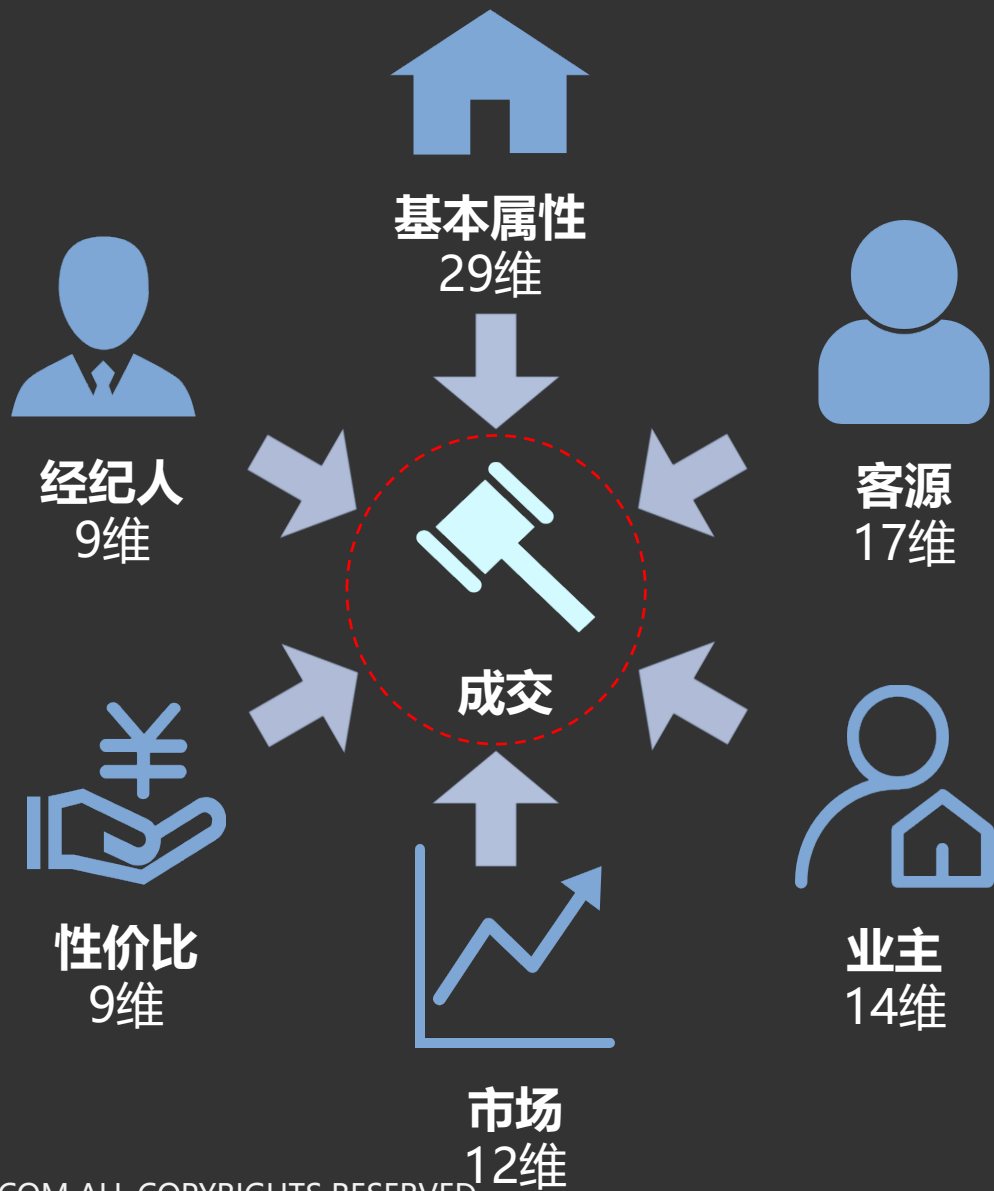
- 房源特征
 - ✓静态特征
 - ✓时序特征

- 特征处理
 - ✓特征提取
 - ✓特征组合
 - ✓离散化

- 模型预测
 - ✓XGBoost

- 分数映射
 - ✓房源质量分数

房源特征

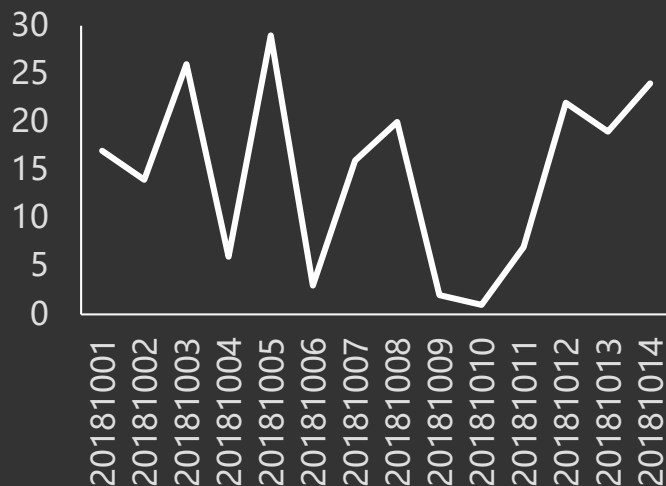


- 一套房源能否成交同很多因素相关
- 6大方向设计了90维特征
- 静态特征：69维
- 时序特征：21维

时序特征提取



- ✓ 浏览
- ✓ 关注
- ✓ IM聊天
- ✓ 电话
- ✓ 带看
- ✓ 跟进
- ...



- 均值
- 方差
- 极值
- ...



- 最近14天浏览量均值
- 最近7天浏览量均值
- ...

v1.0 - 小结

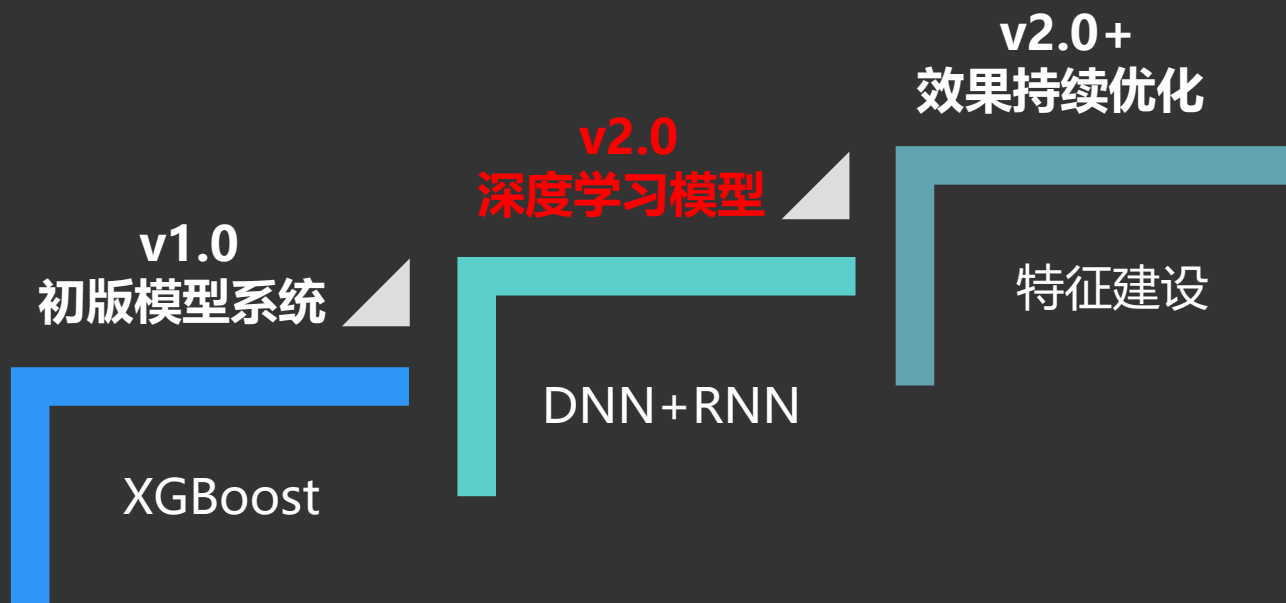
解决的问题

- 人工 -> 机器
- 解决了人工选房的问题
 - ✓ 选房成本低
 - ✓ 选房没有主观性
 - ✓ 可以盘点所有房源质量

存在的问题

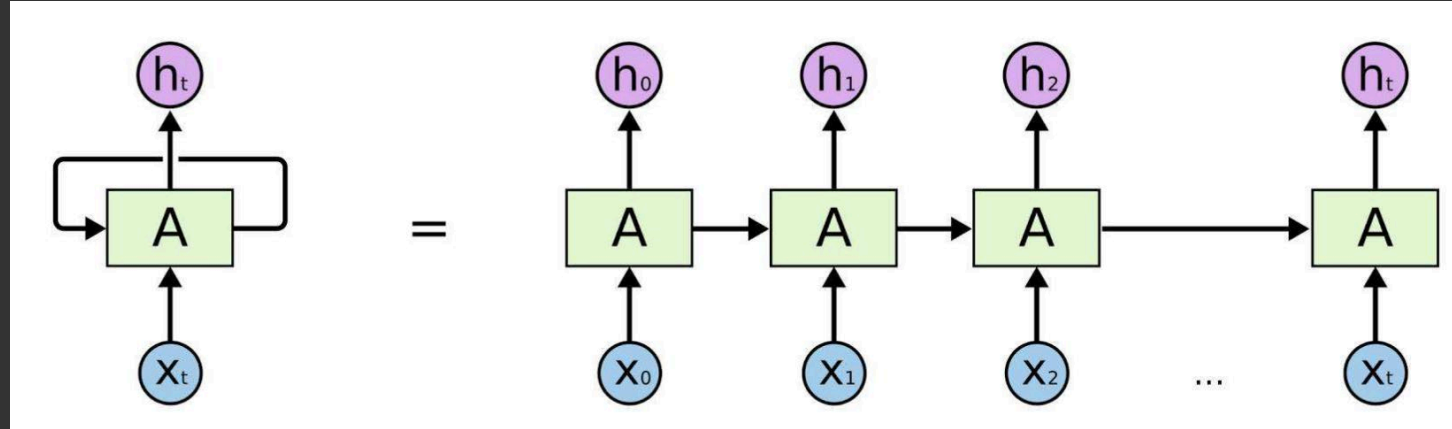
- 很难兼容新上房源
 - ✗ 新上房源与库存房源在行为特征上差异巨大
 - ✗ 引入新上房源，会严重干扰模型
- 时序数据特征爆炸
 - ✗ 时序特征进行特征提取，得到的特征数量庞大
 - ✗ 随着迭代的进行，新加入特征边际效应递减，但是成本高

模型演变历程

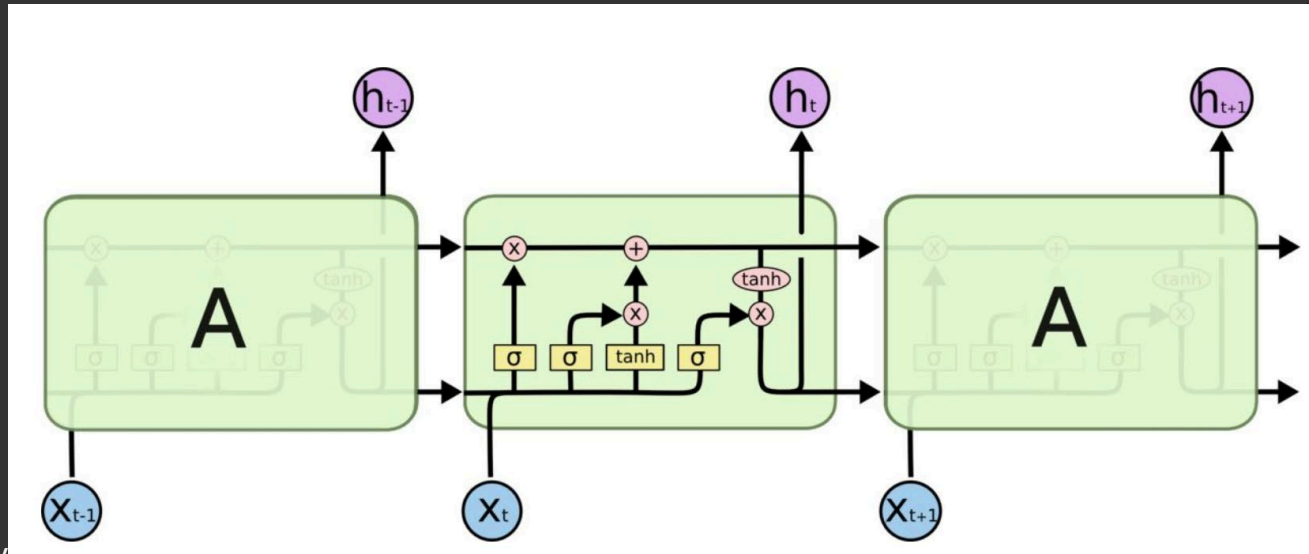


RNN

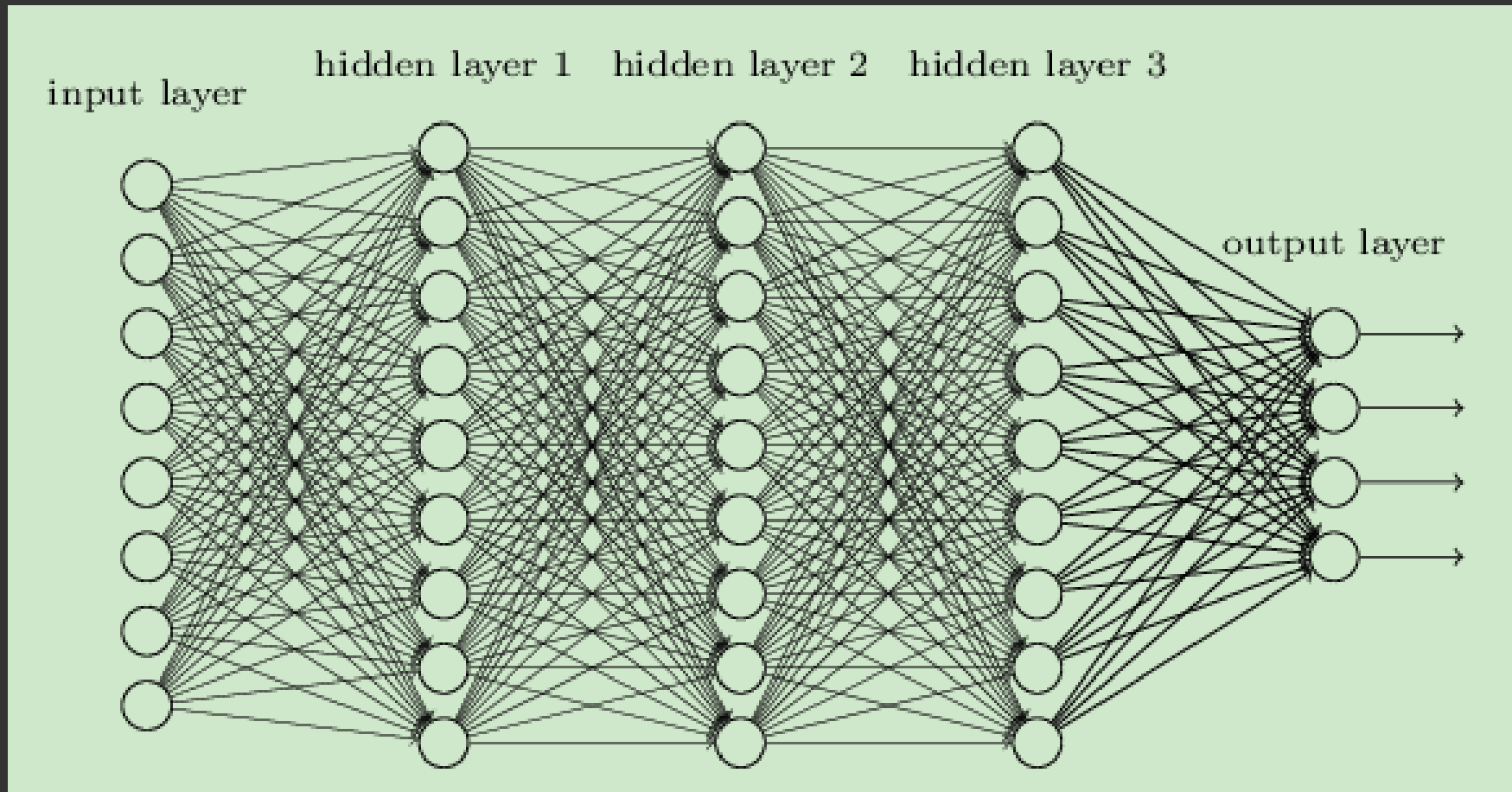
RNN



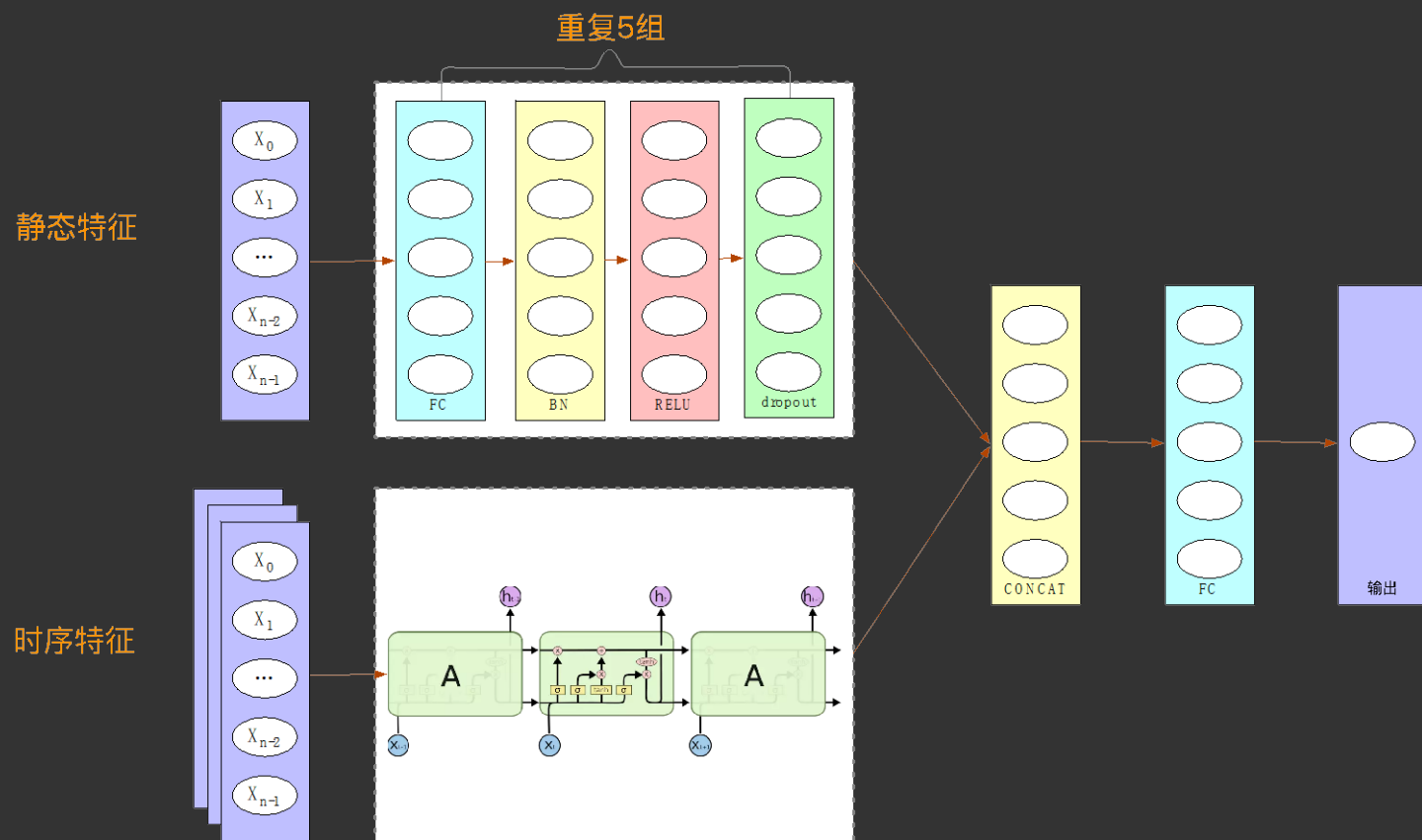
LSTM



DNN



深度学习模型结构



- 混合模型: DNN + RNN
- Deep neural networks (DNN)
 - 全连接的多层感知机
 - BatchNormalization
 - 激活层 (RELU)
 - dropout正则化
- Recurrent neural networks (RNN)
 - LSTM

模型系统对比

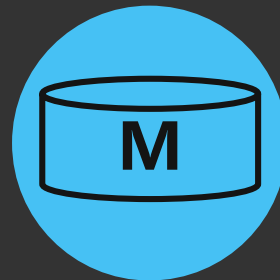
v1.0



房源特征



特征处理



XGBoost



分数映射

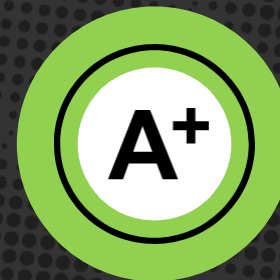
v2.0



房源特征



DNN + RNN

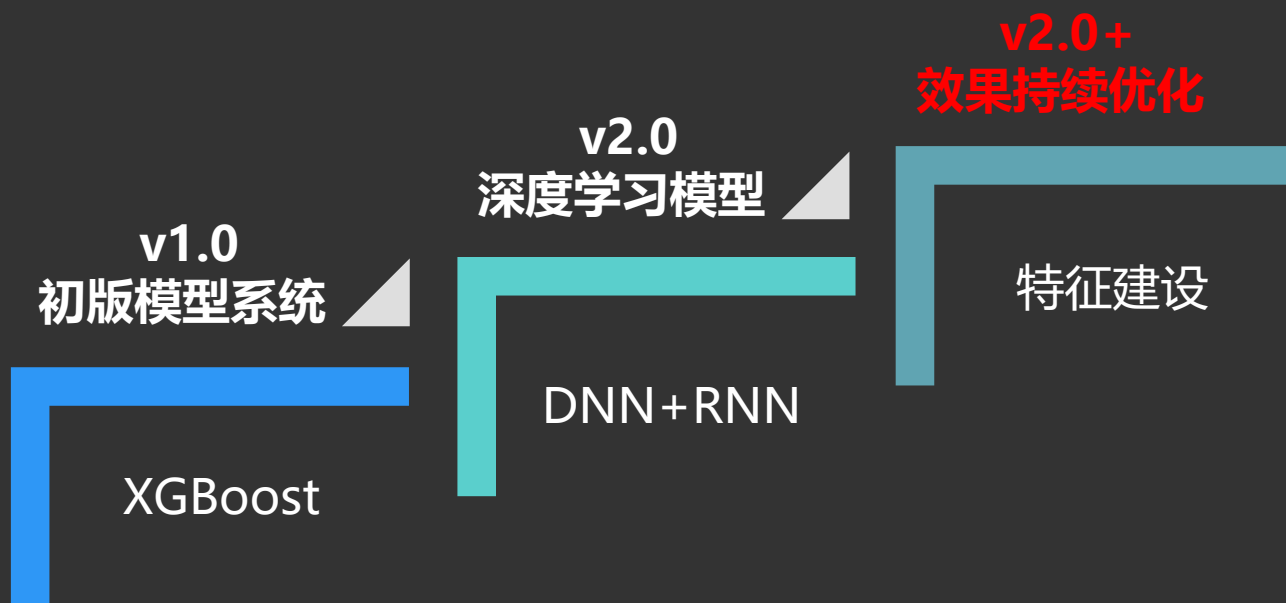


分数映射

模型指标对比

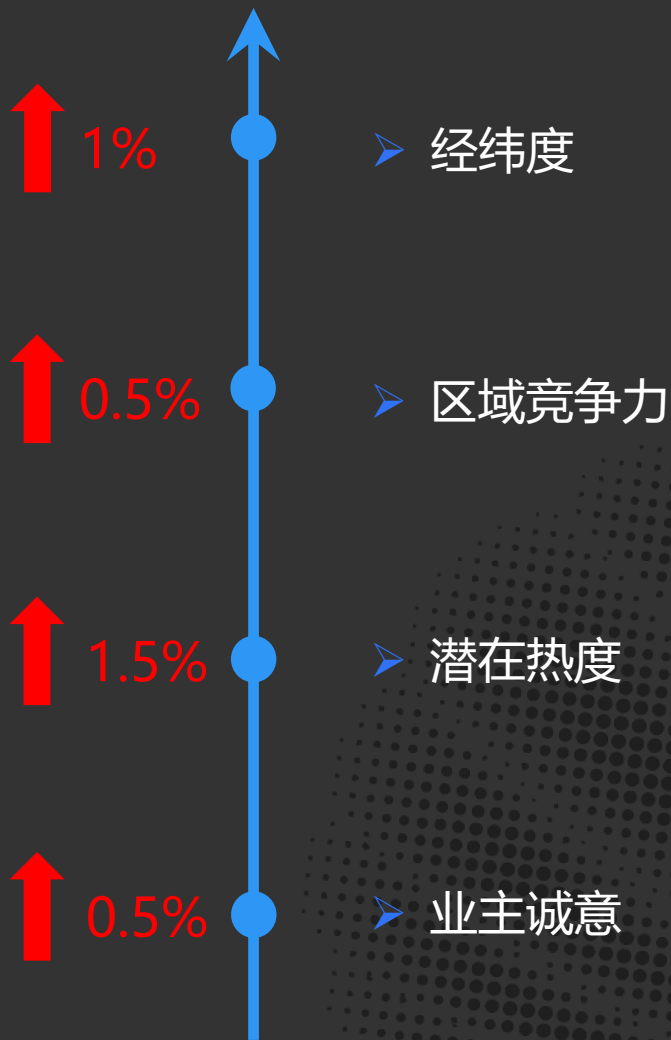
	v1.0	v2.0
AUC	0.814	0.831
Top1000去化率	30.72%	+0.83%
Top2000去化率	25.28%	+1.2%
Top3000去化率	22.13%	+1.24%

模型演变历程



v2.0+：持续优化

特征维度	现状分析
房源基本属性	? 可以完善补充
客户	? 可以挖掘
市场	? 可以挖掘
业主	? 体现不完善
经纪人	✓ 考虑完整
性价比	✓ 考虑完整

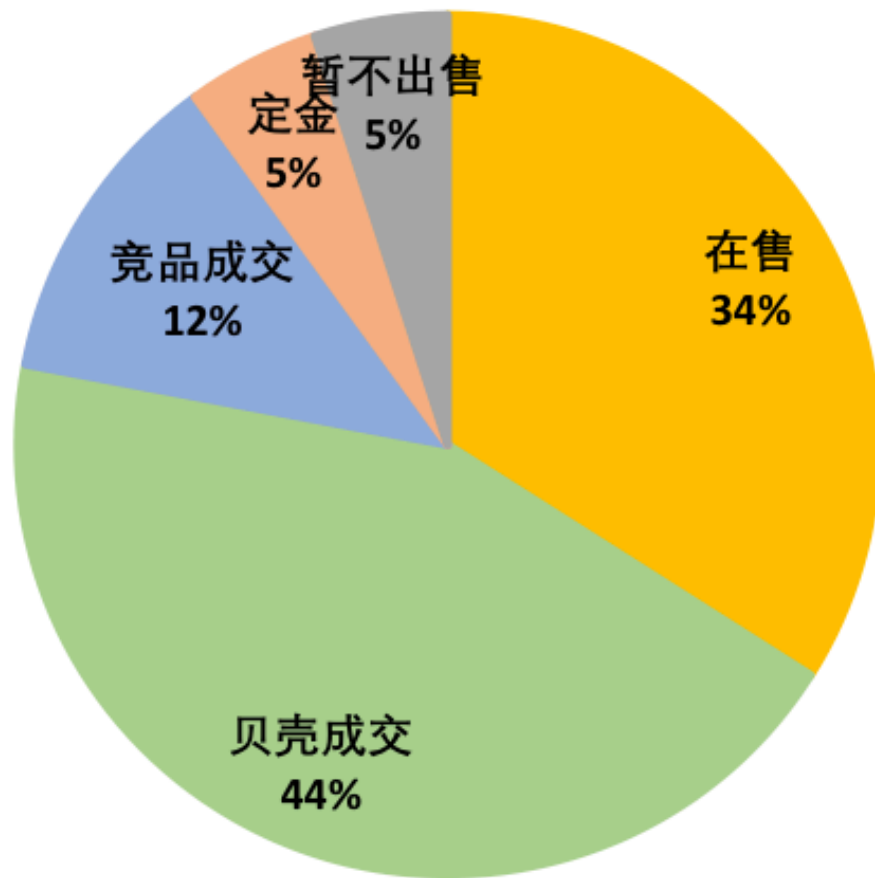


v2.0+：持续优化

业主诚意

类型	占比
成交	56%
在售	34%
定金	5%
暂不出售	5%

TOP100高分房源分析



v2.0+: 持续优化

业主诚意

- 巧妇难为无米之炊：行为稀少
- 能做什么？
 - 挖掘：经纪人对业主态度的描述

经纪人点评

合格

核心卖点：朝阳公园 枣营南里 精装 南北 带客厅两居室 诚意出售 看房方便

户型介绍：南北两居室，进门左手边是客厅，正规客厅可以放电视沙发，侧排不像其他老房子卫生间有台阶，南边是主卧室带阳台，附是厨房，明亮整洁，

装修描述：这套房子的装修是当时两年前装的到现在空调没开过，基本没住过，换房阿姨还有点不舍得，整体橱柜留出了洗衣机冰箱的地方，卫生间的防水用的双层大理石，窗户大小，包括阳台都是精心设计过的，当时是为了自己用的所以比较用心，

小区介绍：小区地处朝阳公园西路西侧，与朝阳公园只有一路之隔，从北门进车南门出车，人车分流，井井有条。小区建筑一般在85-96年左右，内有花园、水系、亮马河支流，健身器材等。

03-15 09:21 链家-苗宝存(维护人)

不着急出售，价钱就这样

02-27 16:37 链家-苗宝存(维护人)

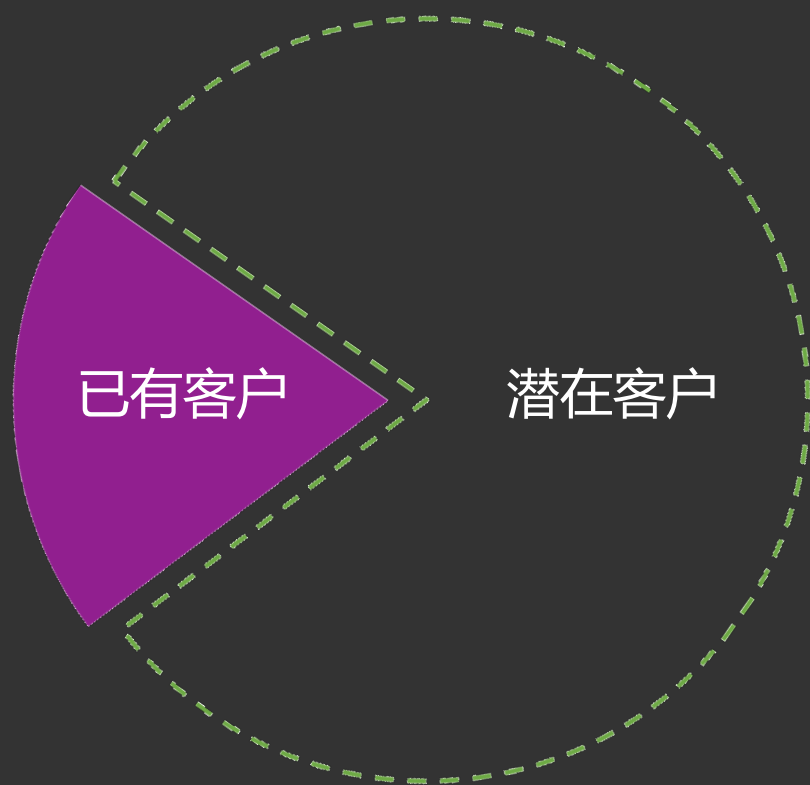
业主诚心出售，但心里价位比较高，看房也不方便，能接受价位的客户可随时签约

02-12 17:11 链家-苗宝存(维护人)

还是诚信卖，但价格不降，看房也不方便

v2.0+：持续优化

潜在热度



一套房源的用户组成结构

- 客户潜在的热度，反映市场偏好
- 影响因素
 - 潜在客户对房源的偏好
 - 潜在客户的购房意愿强度

v2.0+：持续优化

潜在热度

➤ 对房源偏好

- 商圈偏好
- 小区偏好
- 居室偏好
- 面积偏好
- 价格偏好

...

➤ 购房意愿强度

➤ 单套房源的客户潜在热度

$$\sum_{\text{所有潜在客户}} \left(\begin{array}{c} \text{潜在客户} \\ \text{对房源偏好} \end{array} \times \begin{array}{c} \text{潜在客户} \\ \text{购房意愿强度} \end{array} \right)$$

v2.0+：持续优化

区域竞争力



区域内排名
(门店/商圈)

- 是否成交与周围竞争者有关
- 体现竞争力的特征
 - ✓ 价格：总价、单价
 - ✓ 行为：浏览、带看 ...

v2.0+：持续优化

经纬度



房源地理位置信息

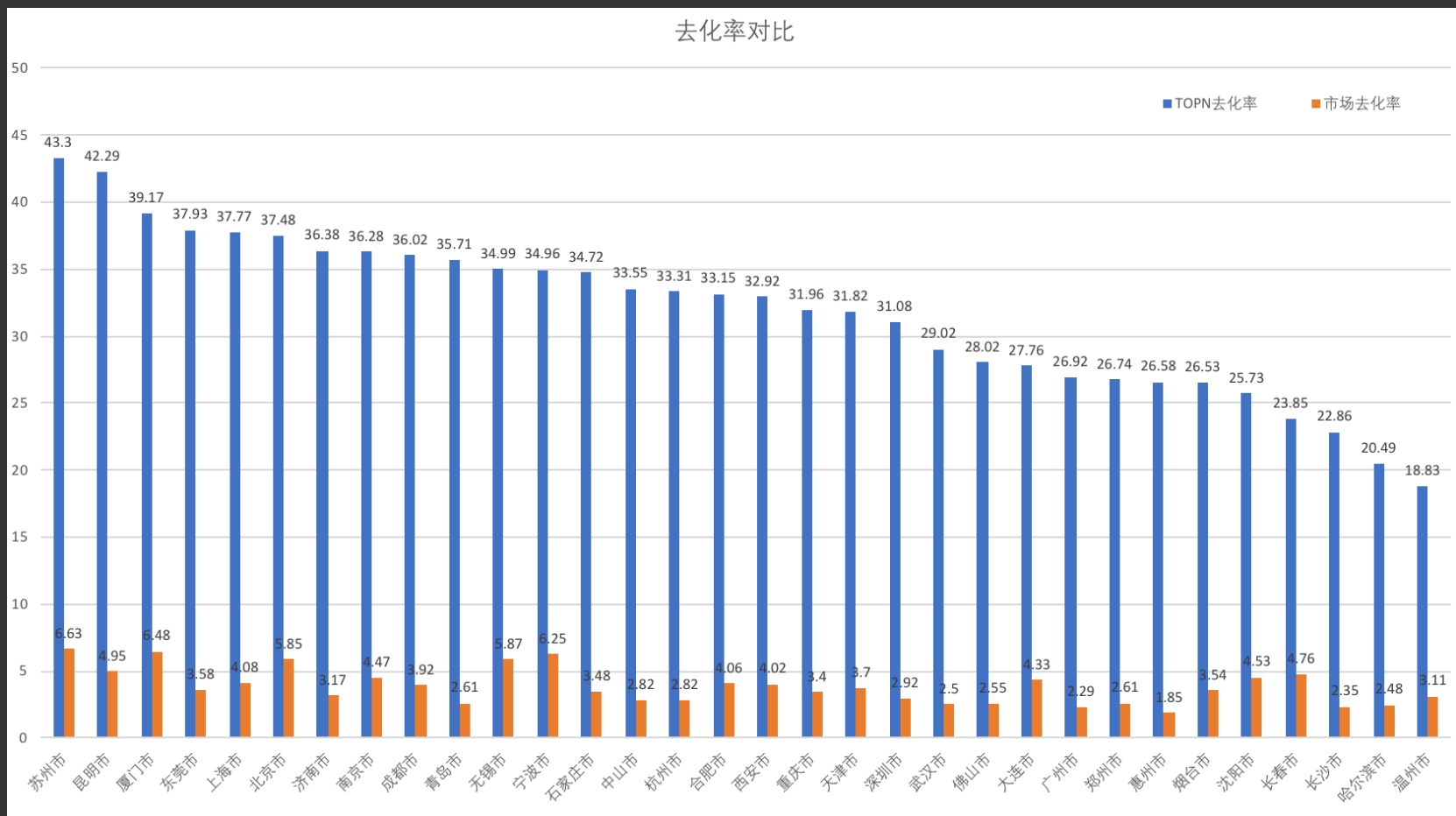
➤ 市场偏好

- ✓ 反映地段偏好

➤ 区域竞争力

- ✓ 结合体现竞争力的特征
- ✓ 反映某一距离范围内的竞争力

效果评估



去化率(一周平均值)

➤ 指标

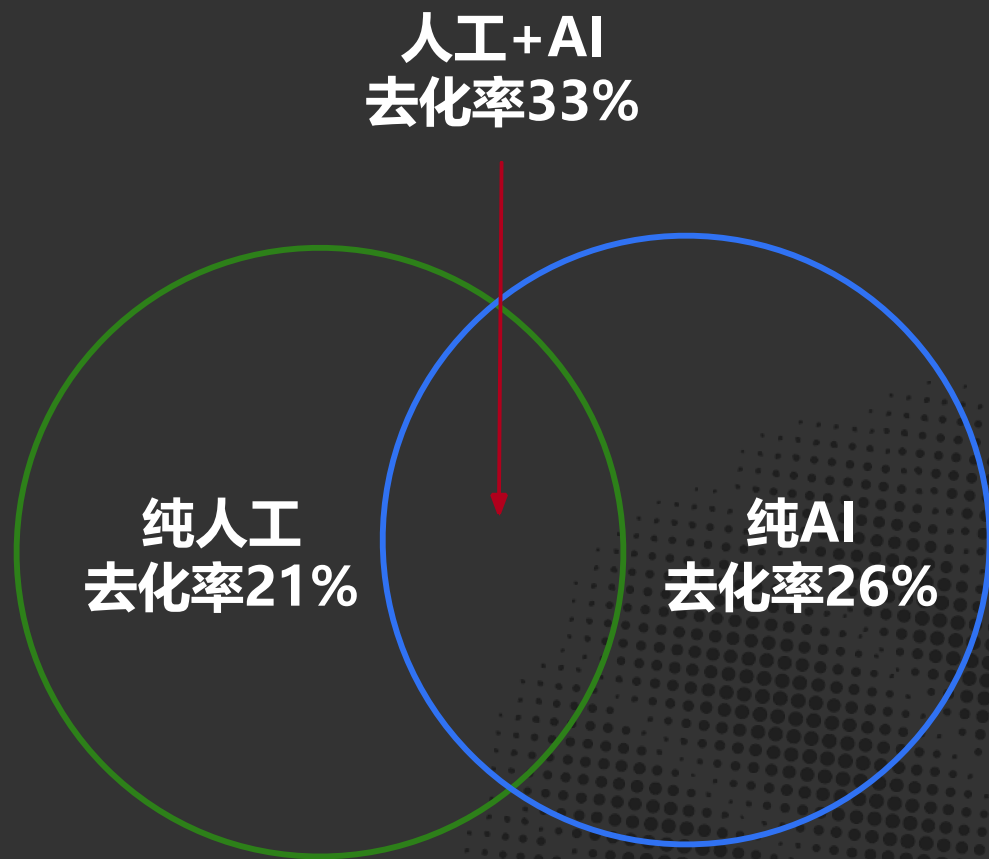
- TopN去化率
- $N = 2.5 * \text{周成交量}$

➤ 32个城市平均值

- TopN去化率: 31.7%
- 自然去化率: 3.8%

人工选房 VS AI选房

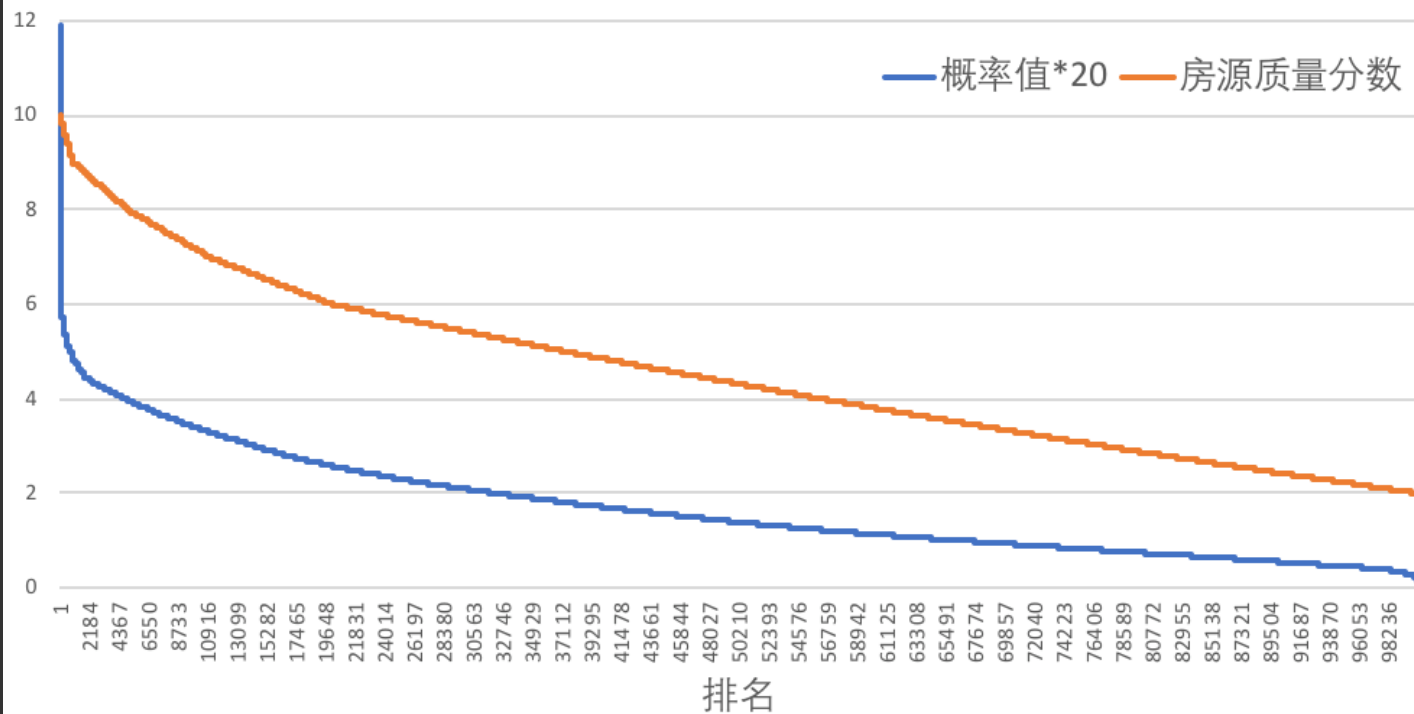
- 条件：基于相同的名额下进行比较
- 人工选房和AI选房重合率**48%**
- 三种模式下的去化率
 - 人工+AI：33%
 - 纯AI：26%
 - 纯人工：21%



实践应用

分数映射

房源质量分数 VS 概率值



➤ 模型输出

- 每天不稳定，范围波动大
- 分数分布不合理
- 不易于业务使用

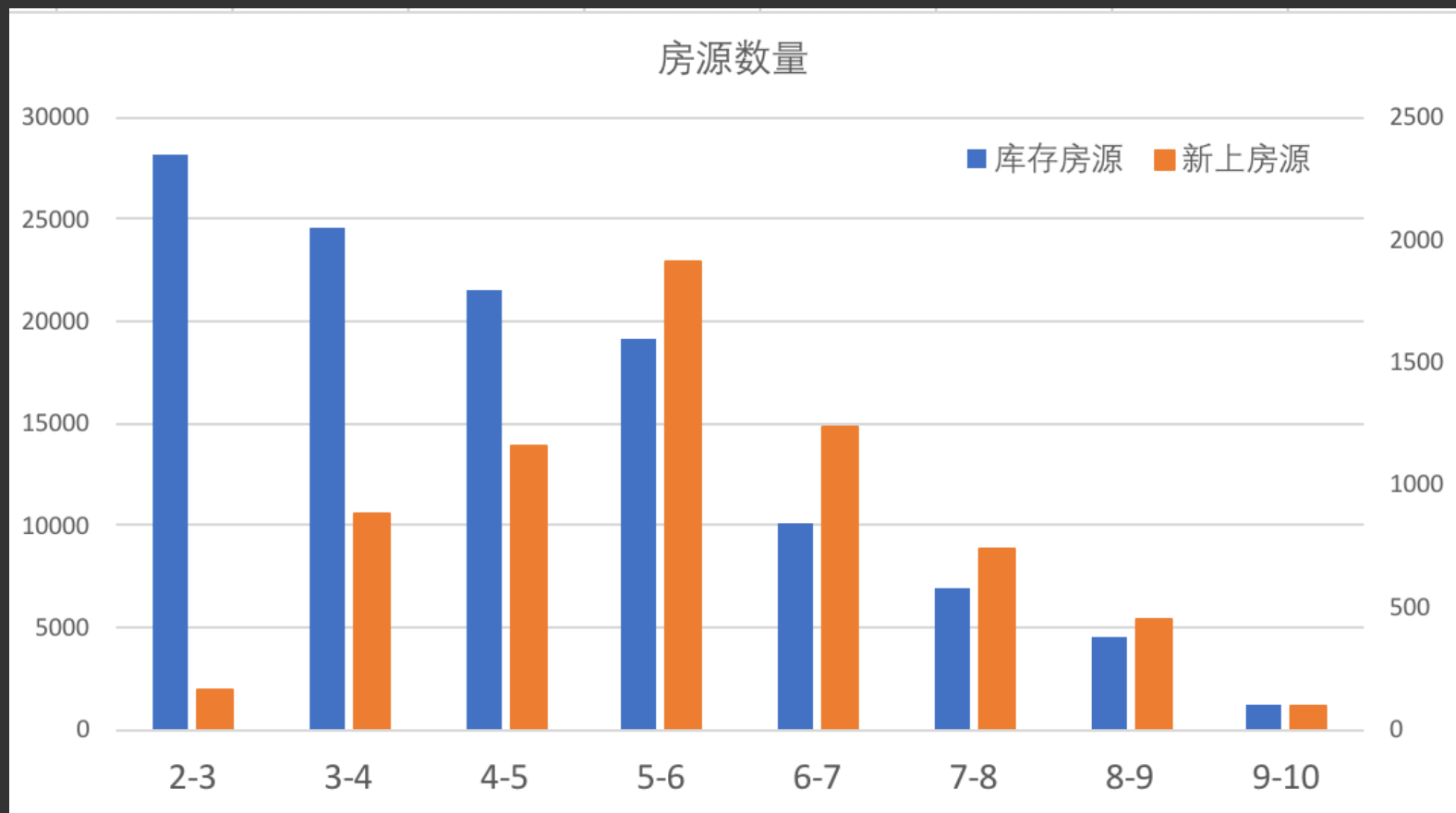
➤ 房源质量分数

- 根据概率值排名进行映射
- 分数分布比较稳定
- 10分制易于业务使用

➤ 分数映射公式

$$\text{MinMaxScaler} \left(\frac{1}{e^{ax}} \right)$$

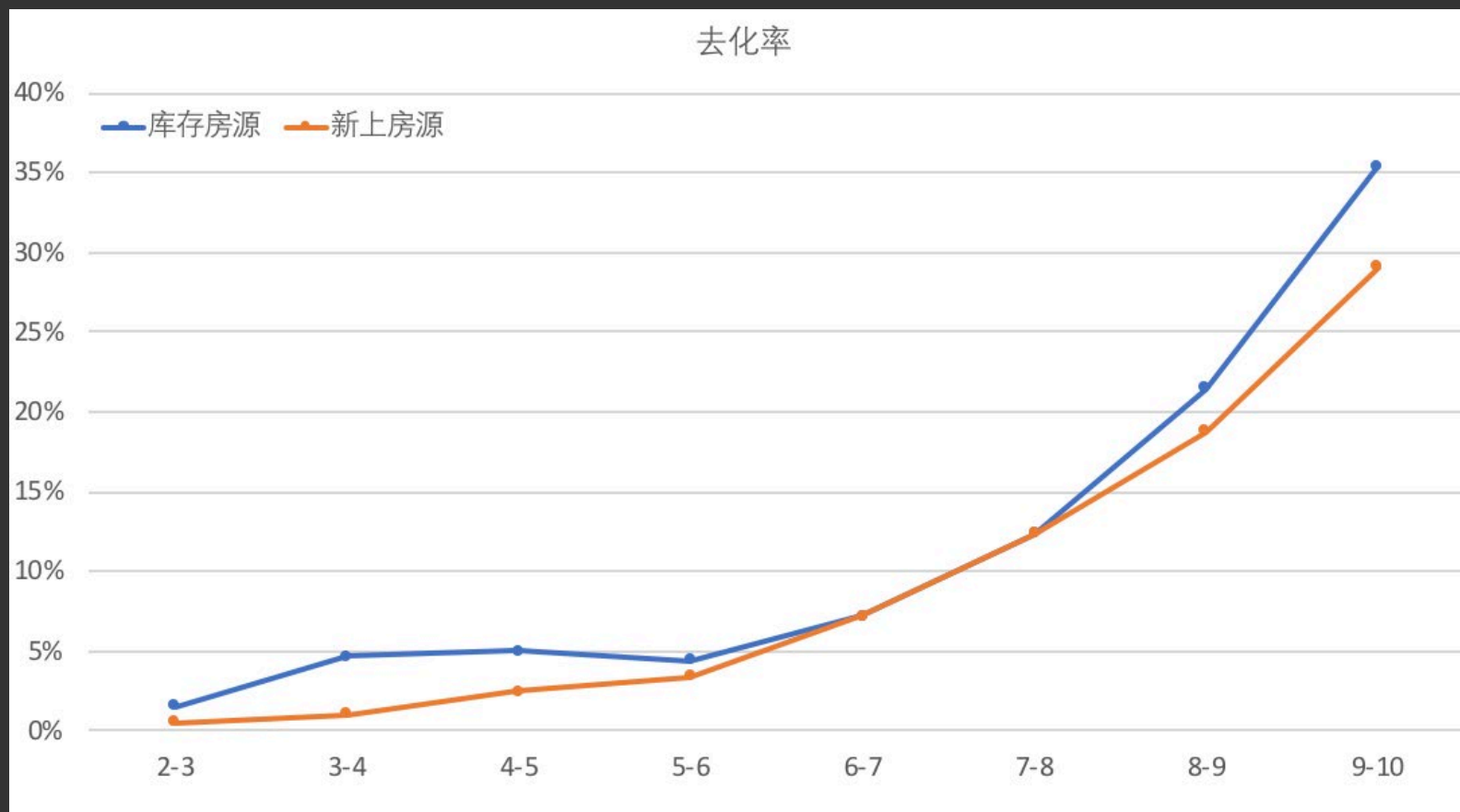
库存&新上房源



数量分布

✓ 新上房源分数略高

库存&新上房源



去化率

✓ 分数越高，质量越好

了解分

经纪人的疑问

- 分数解释：打分是怎么计算的
- 如何操作可以提升打分？

质量分数

- 具有排序意义
- 很难引导经纪人



优质房 (A)

次优房 (B)

一般房 (C)

雷达图

雷达图

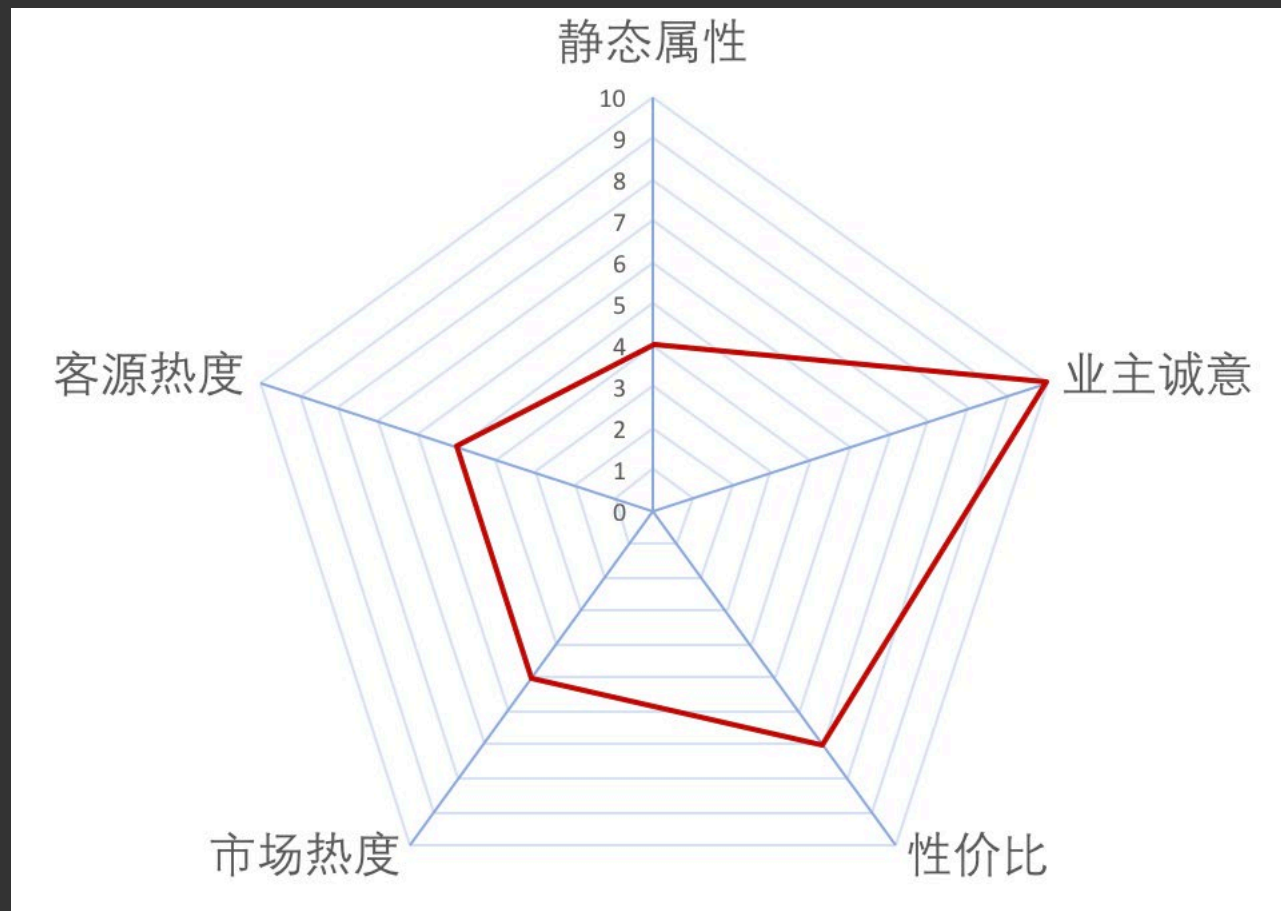
- 明示数据的核心打分维度
- 每个维度展示特征的优缺点
- 引导经纪人，提高分数

举例：

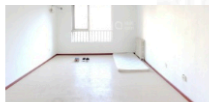
- 业主诚意
- 性价比

维度选择

- 正相关的核心维度



房源质量分数 - B端场景



清河 101104489391

学区房

满五唯一

VR房

236

万

单价:58114元/平米

户型	面积	朝向	楼层
1-1-1-1	40.61平	东	高楼层/6

挂牌时间: 10天 (2019-04-09)

备件信息: 齐全

房屋来源: 人际开发-老客户主动委托

历史委托: 5次

房源等级: A

维护人: 高立艳

VIP信息: 暂无

系统评分: 9.9

房源状态: 链家网展示

贝壳网展示

1000971/十

AI选房 - B端场景

辅助经纪人选房

<

好房候选池

本店全部房源

需审核的房源

8

24

300

本期好房名额

参与房源数

已选出好房数

门店内房源按照大数据打分排序，详情可查看 [门店房源盘点](#)



领航新硅谷待审核

4-2-2-2 · 145平 · 06/25层 · 南北

1400万 6.5万/平

维护人：张筱雨

9.5分 推荐选 超过90%的房源

选为好房



领航新硅谷未申请

4-2-2-2 · 145平 · 06/25层 · 南北

1400万 6.5万/平

维护人：张筱雨

8.4分 可考虑 超过90%的房源



领航新硅谷已选为好房

4-2-2-2 · 145平 · 06/25层 · 南北

1400万 6.5万/平

维护人：张筱雨

3.0分 不建议 超过10%的房源

撤销好房

辅助经纪人盘房

<

天拓店房源盘点

房源打分说明：
1. 房源按照系统打分排序，系统打分代表未来两周成交概率，分数越高，成交概率越大。
2. 分值含义：6分以上代表全城TOP20%排名，7分及以上代表全城10%排名；8分代表全城TOP1%；排名越靠前成交概率越高。

房源总数：215

发送表格到邮箱

排名	房源编码	维护人	楼盘	建筑面
1	101103326578	张三	融泽嘉园	86.01
2	101103326578	张三	融泽嘉园	86.01
3	101103326578	张三	融泽嘉园	86.01
4	101103326578	张三	融泽嘉园	86.01
5	101103326578	张三	融泽嘉园	86.01
6	101103326578	张三	融泽嘉园	86.01
7	101103326578	张三	融泽嘉园	86.01
8	101103326578	张三	融泽嘉园	86.01
9	101103326578	张三	融泽嘉园	86.01
10	101103326578	张三	融泽嘉园	86.01
11	101103326578	张三	融泽嘉园	86.01

高分房源直接推为好房



AI选房 - C端场景



总结&思考

总结&思考

- AI选房解决的是房地产领域的TopN排序问题
- AI选房采用了DNN + RNN的混合网络结构
 - DNN, 静态数据; RNN, 时序数据
 - DNN+RNN的混合模型, 提供了静态数据+时序数据的解决方案
- 模型输出值并不能直接适用于业务, 需要做一些转换
 - 为了便于经纪人理解和指导经纪人, 采用分数映射和雷达图两种方式

极客邦科技 会议推荐2019

5月

QCon 北京

全球软件开发大会

大会: 5月6-8日
培训: 5月9-10日

QCon 广州

全球软件开发大会

培训: 5月25-26日
大会: 5月27-28日

6月

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE

上海

技术领导力峰会

时间: 6月14-15日

GMTC 北京

全球大前端技术大会

大会: 6月20-21日
培训: 6月22-23日

7月

ArchSummit 深圳

全球架构师峰会

大会: 7月12-13日
培训: 7月14-15日

10月

QCon 上海

全球软件开发大会

大会: 10月17-19日
培训: 10月20-21日

11月

GMTC 深圳

全球大前端技术大会

大会: 11月8-9日
培训: 11月10-11日

AiCon 北京

全球人工智能与机器学习大会

大会: 11月21-22日
培训: 11月23-24日

12月

ArchSummit 北京

全球架构师峰会

大会: 12月6-7日
培训: 12月8-9日



JOIN US

zhouyuchi001@ke.com