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

极客时间 | 企业服务

这是一封正经“福利”贴

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

扫一扫二维码图案，加我微信

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

贝壳

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

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

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

目标&价值

核心思想

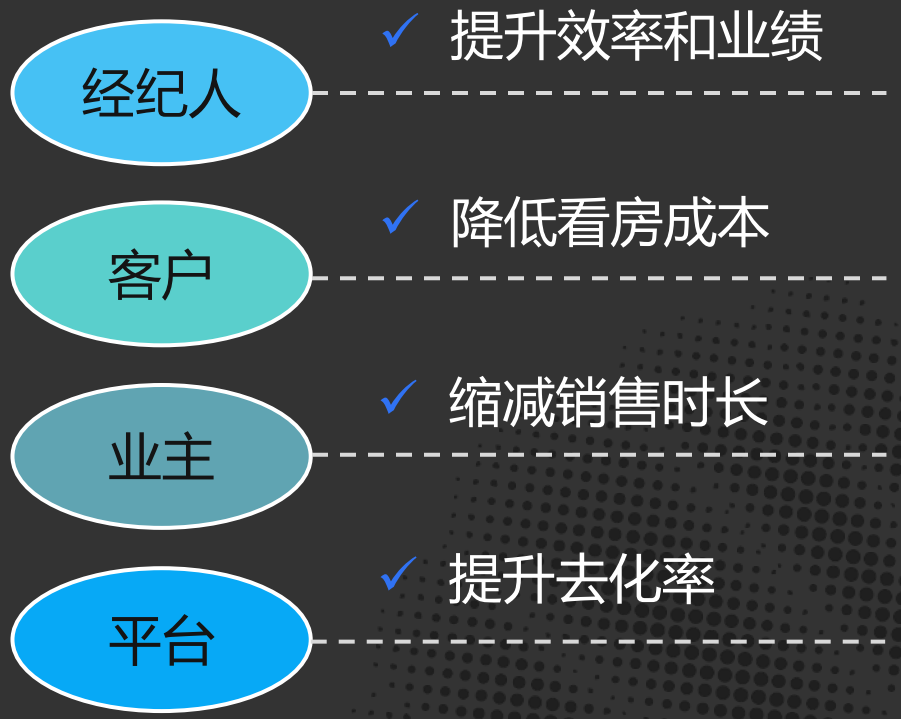
- ✓ 市场需求恒定，优先成交好房

核心问题

- ✓ 选出好房

核心价值

- ✓ 提升带看效率
- ✓ 加速成交



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

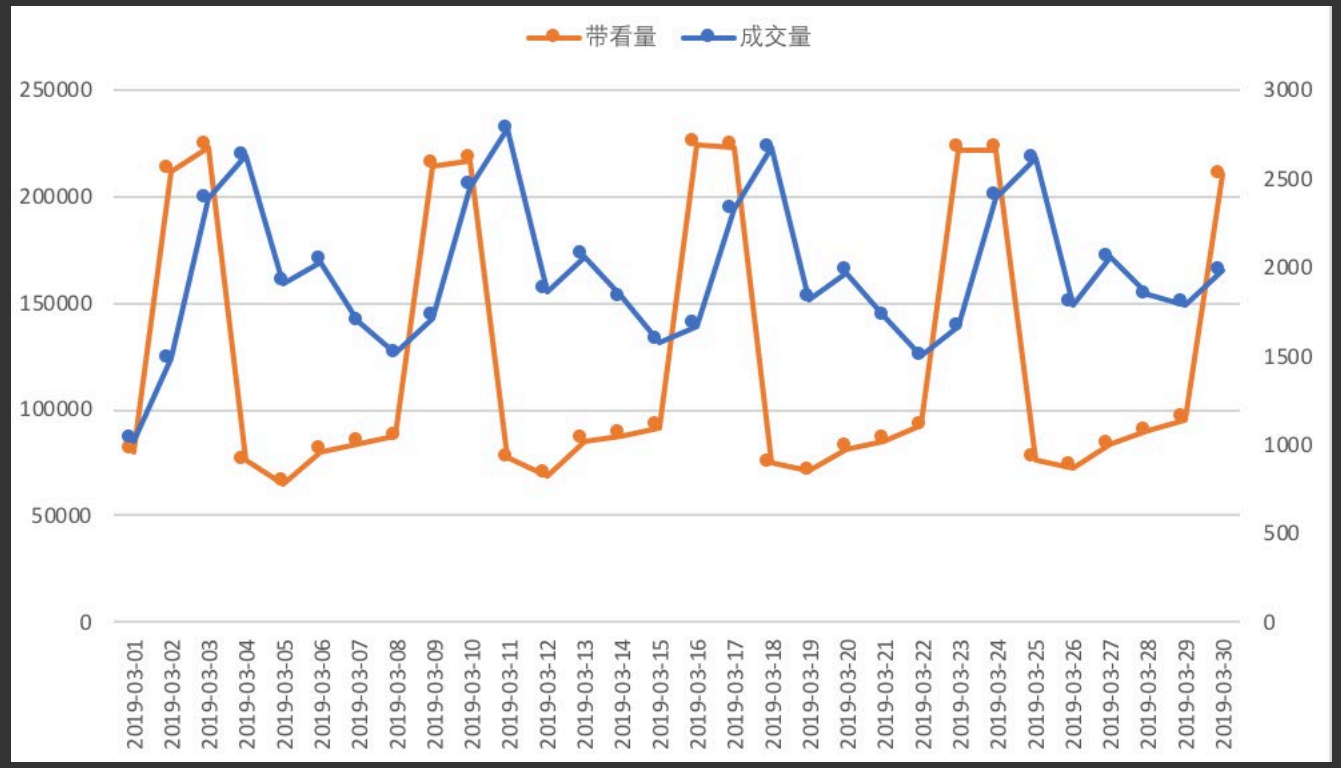
AI选房建模

$$Y = f(X)$$

- ✓ 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



2019年3月每天的成交量和带看量

- ✓ 成交/带看具有周期性
- ✓ 周期性单位：周

时间选择：周的倍数

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

- 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

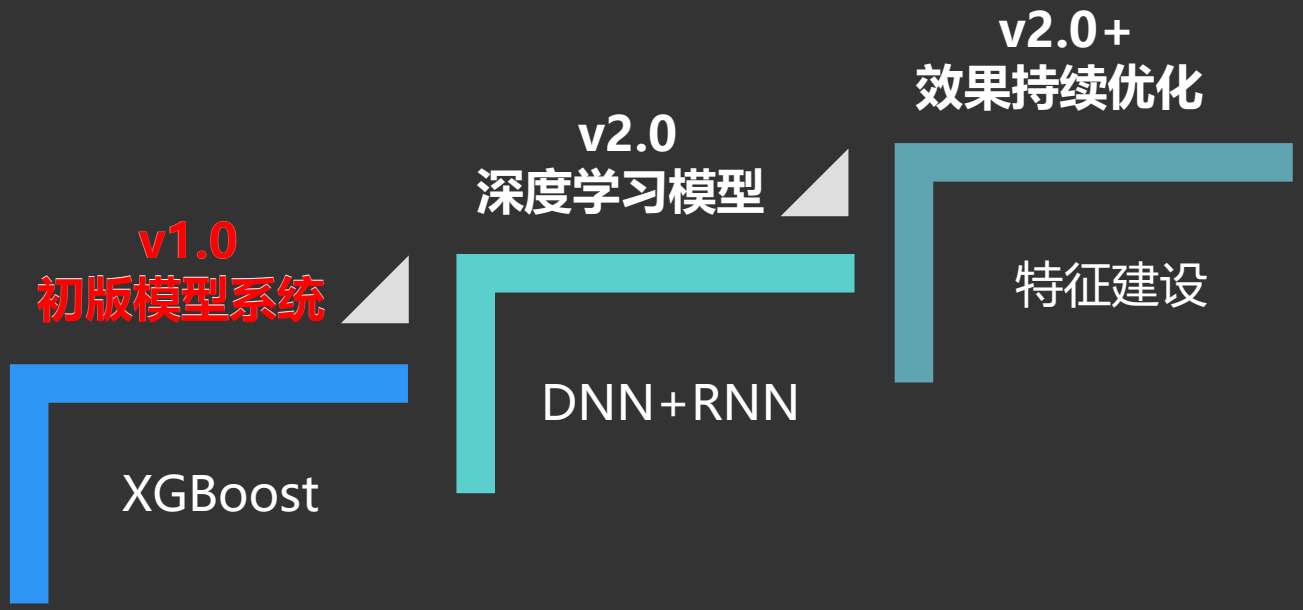
$$Y = f(X)$$

- ✓ 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

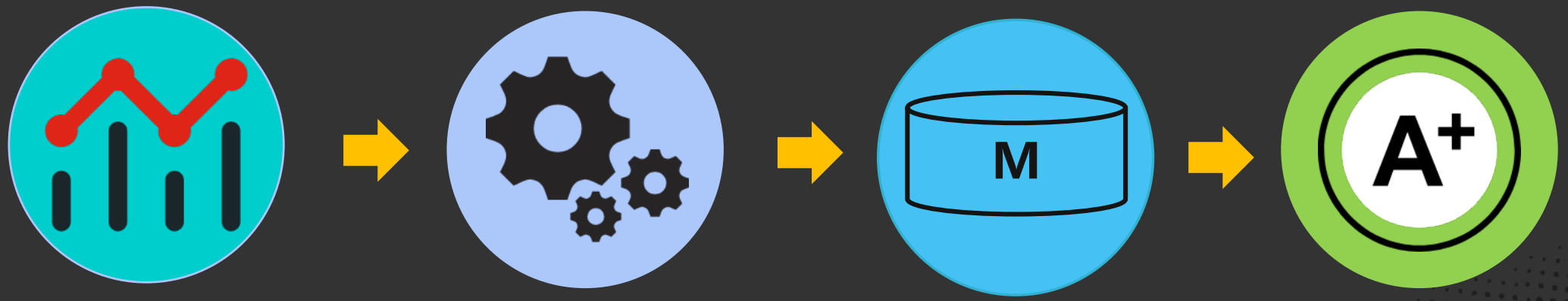
- ➢ 行为特征选择14天进行聚合
- ➢ 挂牌不足14天房源，行为特征信息不足
- ➢ 结论：选择挂牌满14天的房源

模型演变历程

模型演变历程

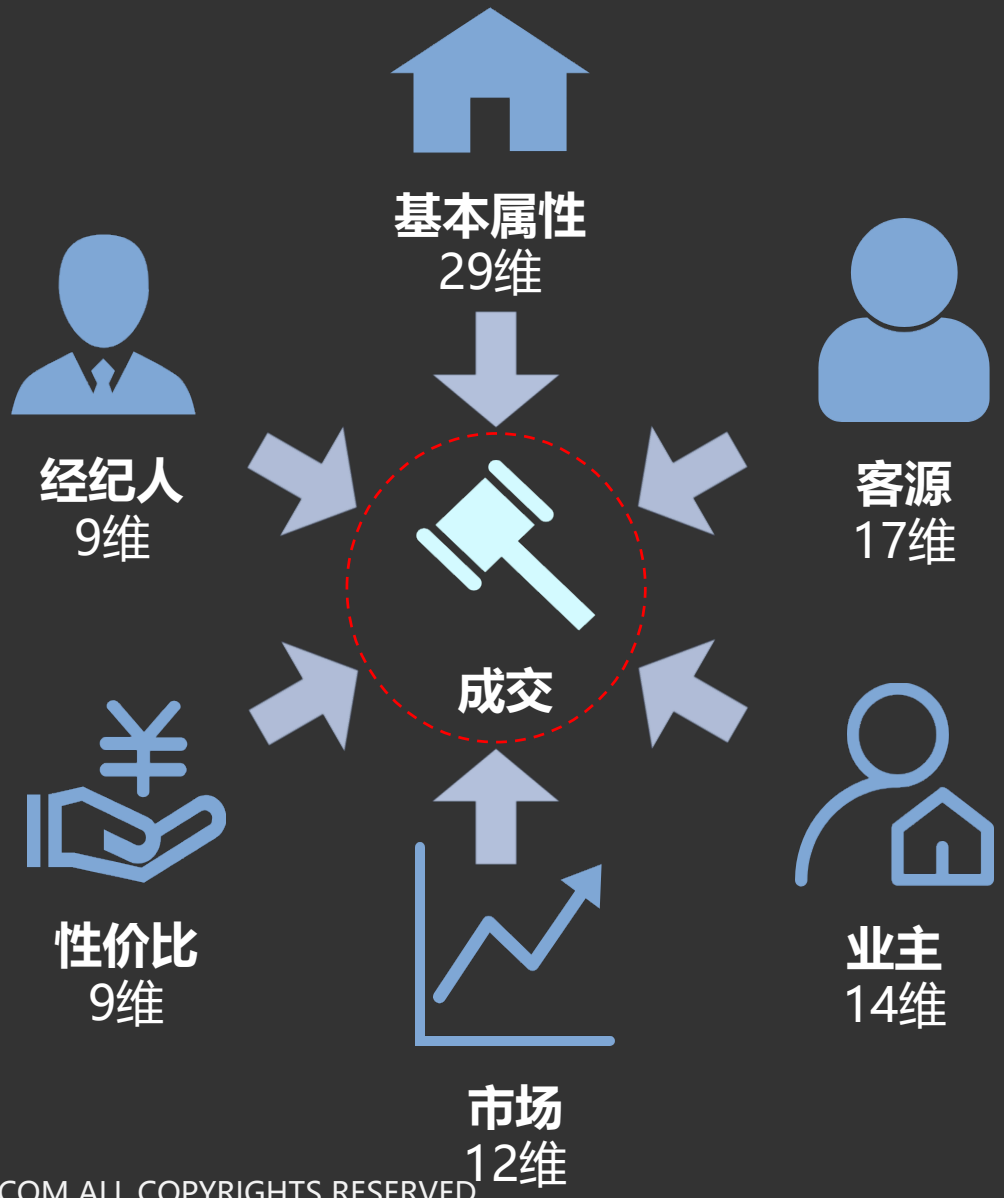


v1.0 - 初版模型系统概览



- 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

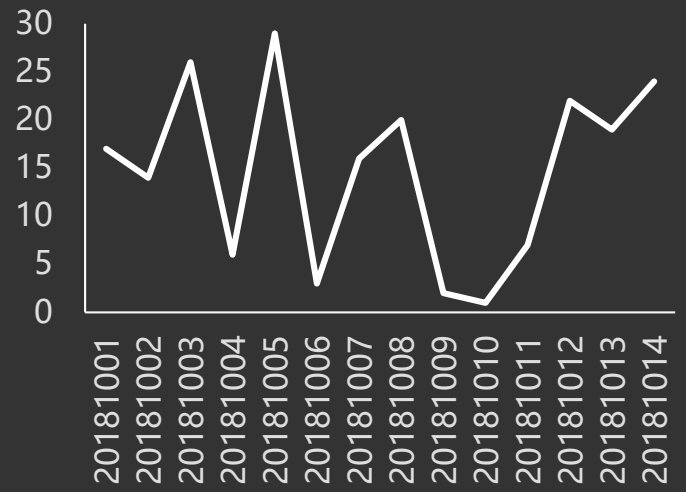
时序特征提取

时序特征
21维

提取函数
8个

提取特征
168维

- ✓ 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

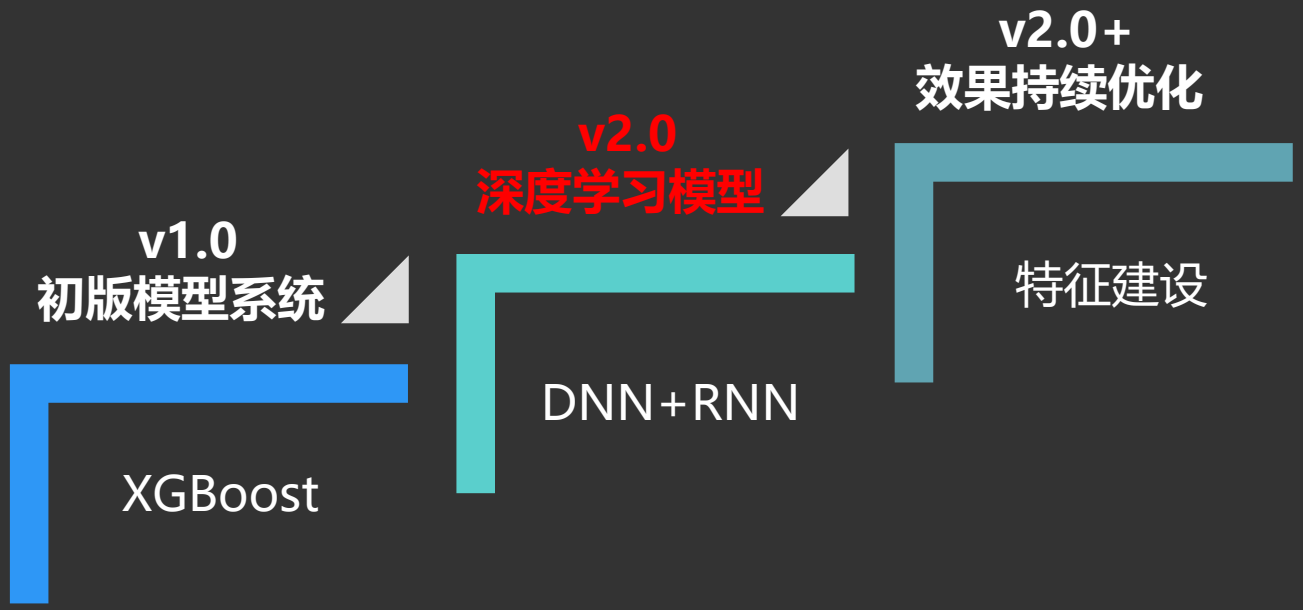
解决的问题

- **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

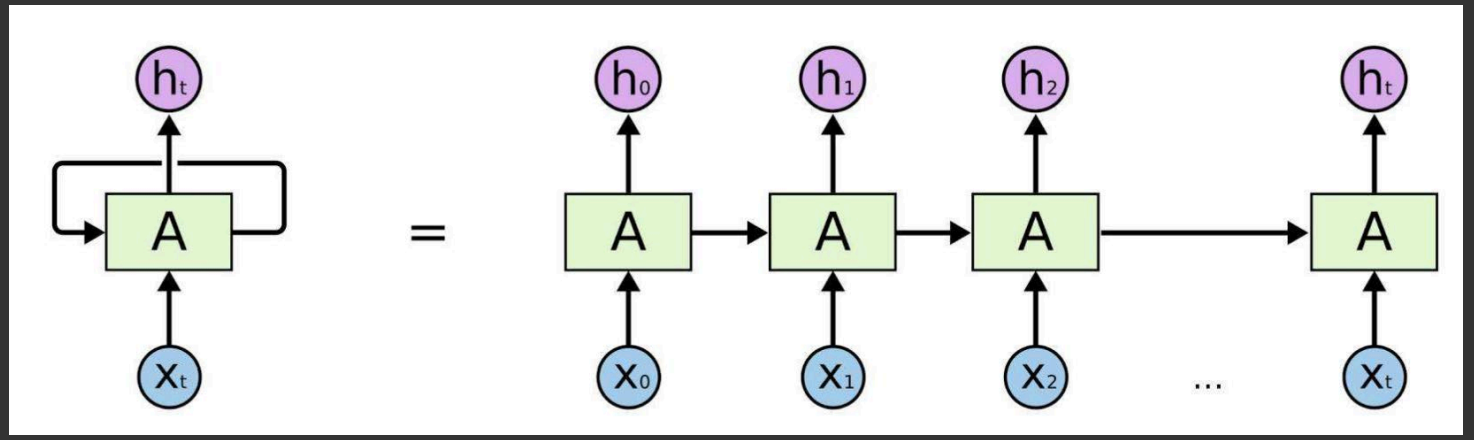
- **很难兼容新上房源**
  - ✗ 新上房源与库存房源在行为特征上差异巨大
  - ✗ 引入新上房源，会严重干扰模型
- **时序数据特征爆炸**
  - ✗ 时序特征进行特征提取，得到的特征数量庞大
  - ✗ 随着迭代的进行，新加入特征边际效应递减，但是成本高

模型演变历程

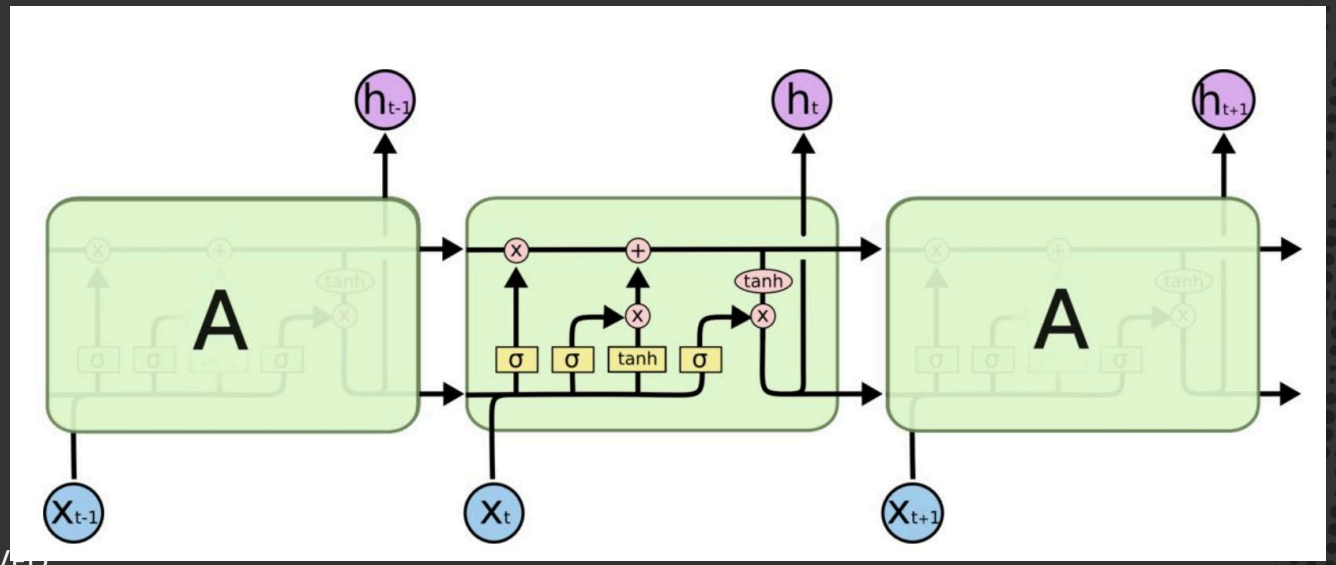


RNN

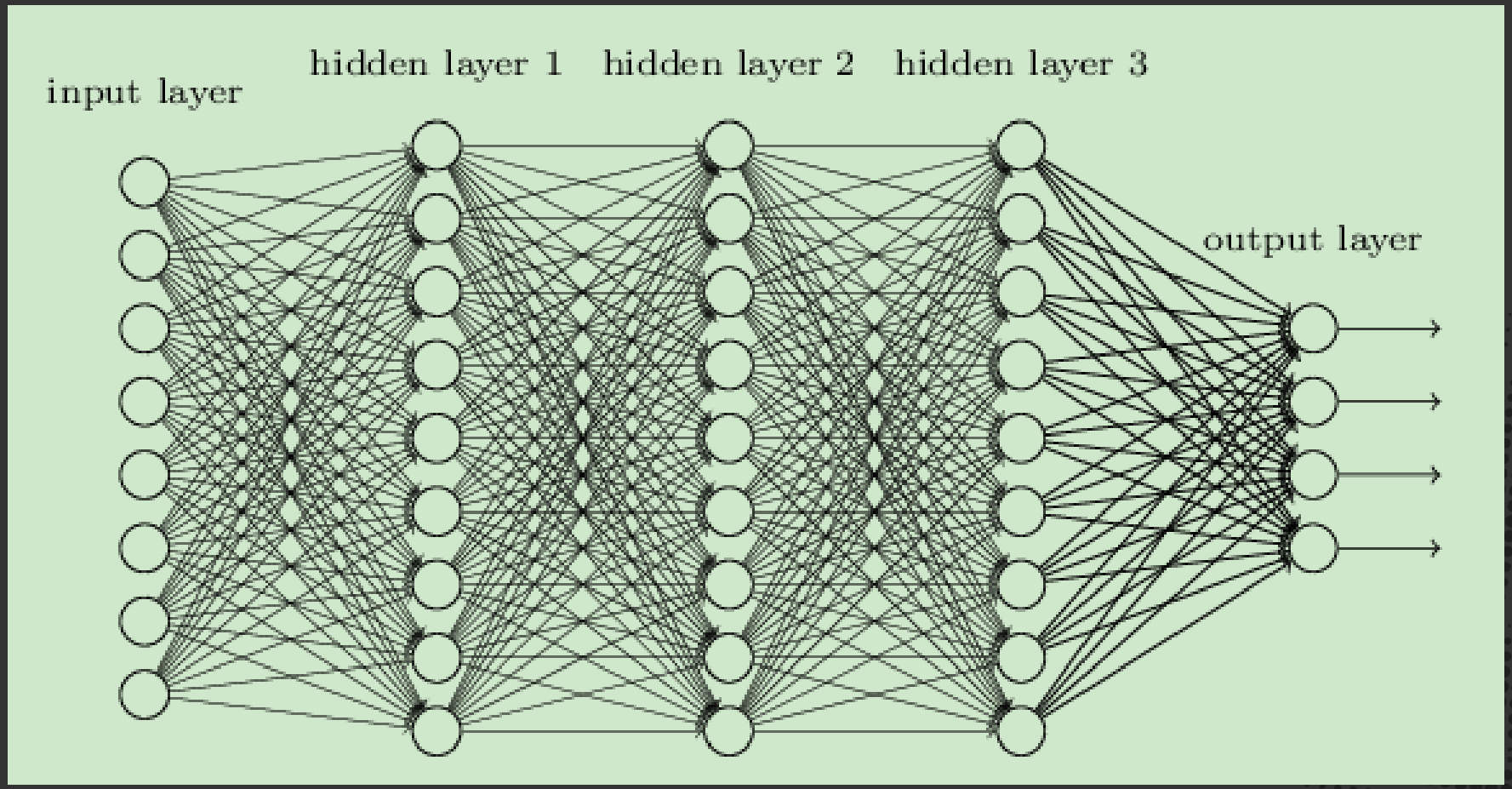
RNN



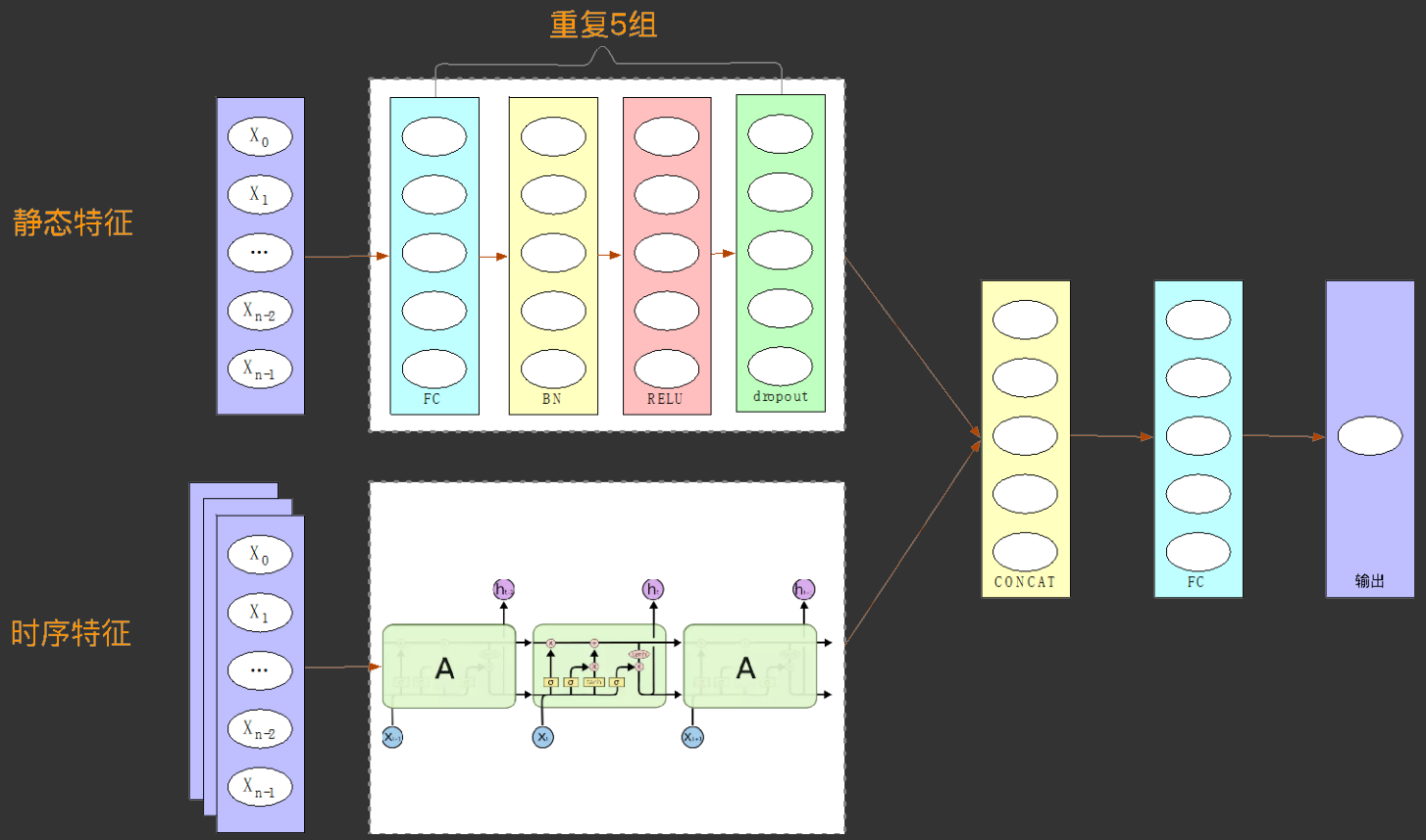
LSTM



DNN



深度学习模型结构



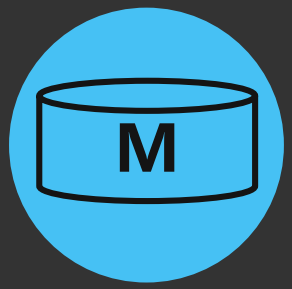
- 混合模型：DNN + RNN
- Deep neural networks (DNN)
  - 全连接的多层感知机
  - BatchNormalization
  - 激活层（RELU）
  - dropout正则化
- Recurrent neural networks (RNN)
  - LSTM

模型系统对比

v1.0

房源特征

特征处理



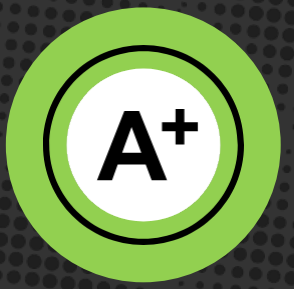
XGBoost

分数映射

v2.0

房源特征

DNN + RNN

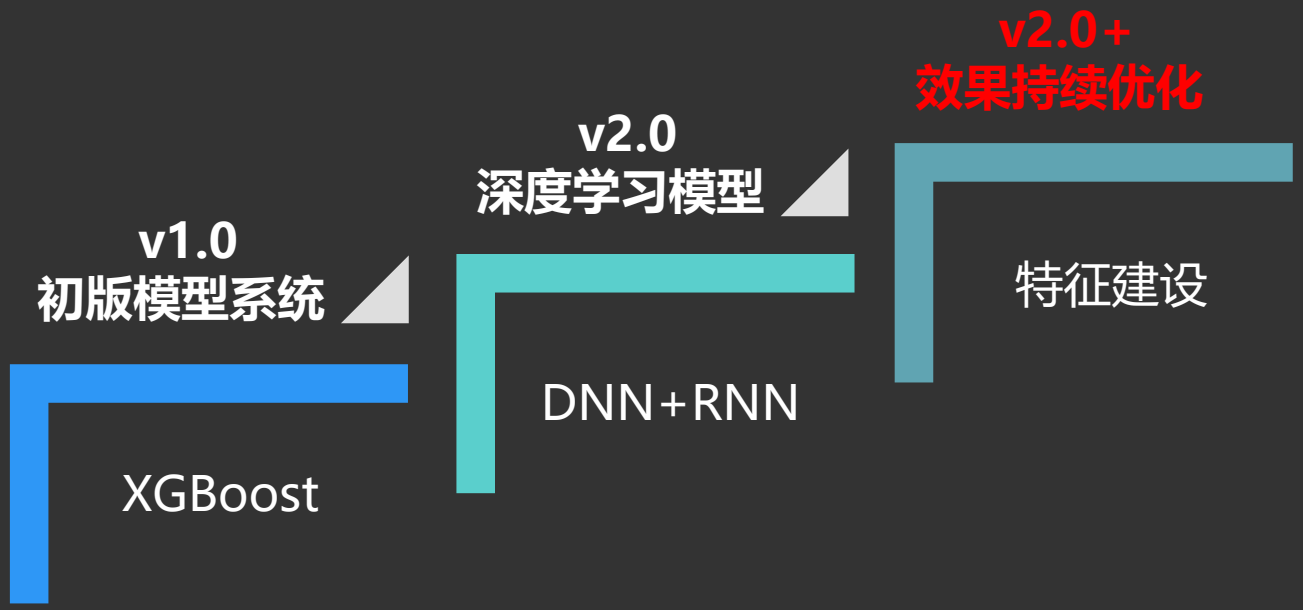


分数映射

模型指标对比

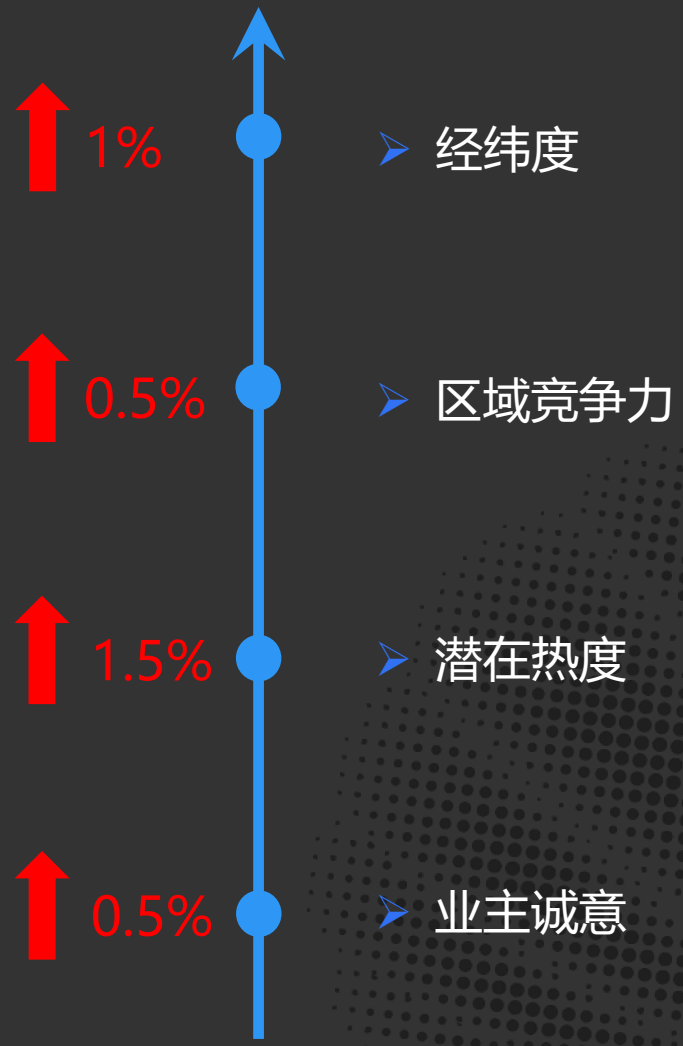
	v1.0	v2.0
AUC	0.814	**0.831**
Top1000去化率	30.72%	**+0.83%**
Top2000去化率	25.28%	**+1.2%**
Top3000去化率	22.13%	**+1.24%**

模型演变历程



v2.0+：持续优化

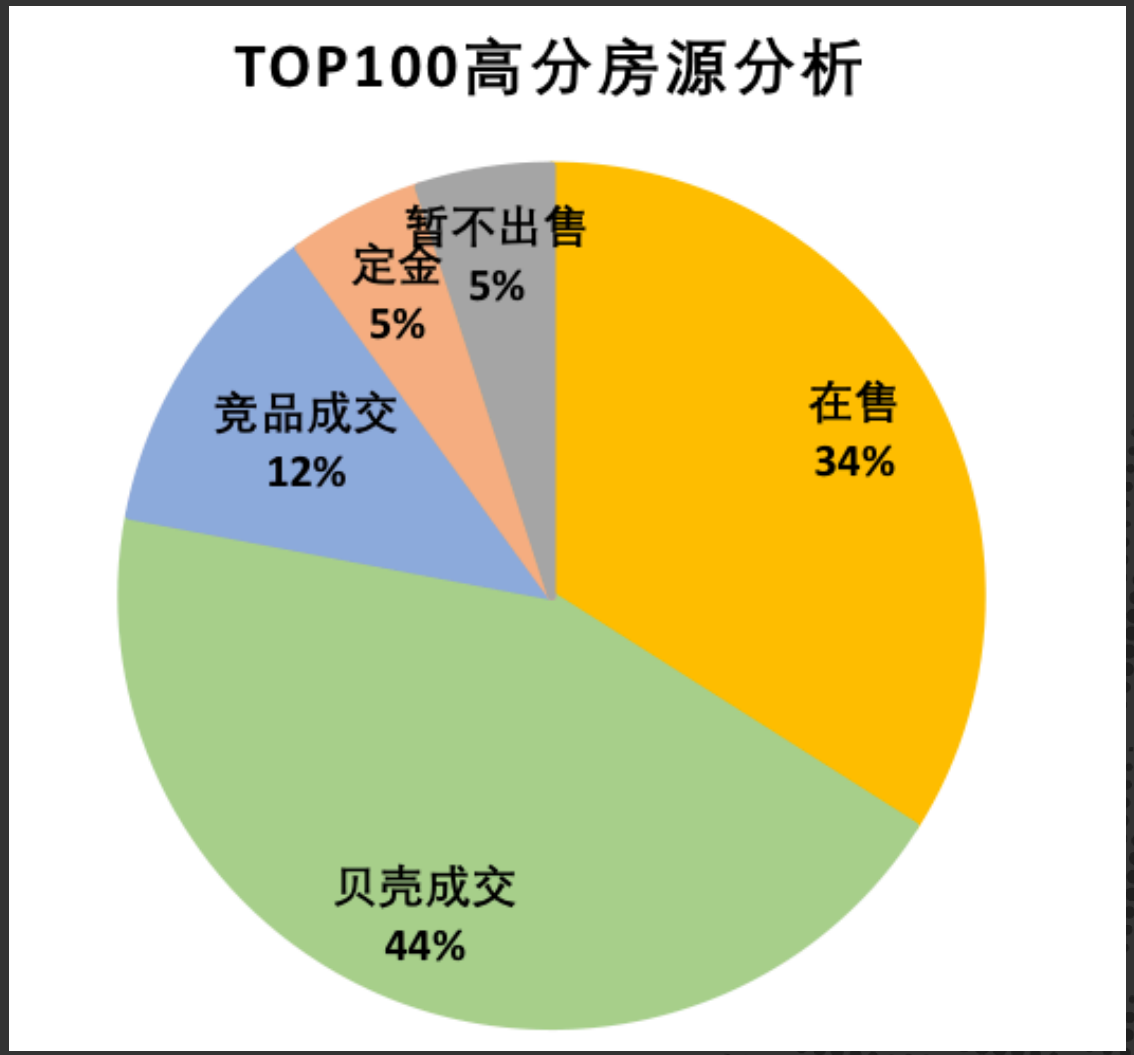
特征维度	现状分析
房源基本属性	？ 可以完善补充
客户	？ 可以挖掘
市场	？ 可以挖掘
业主	？ 体现不完善
经纪人	✓ 考虑完整
性价比	✓ 考虑完整



v2.0+：持续优化

业主诚意

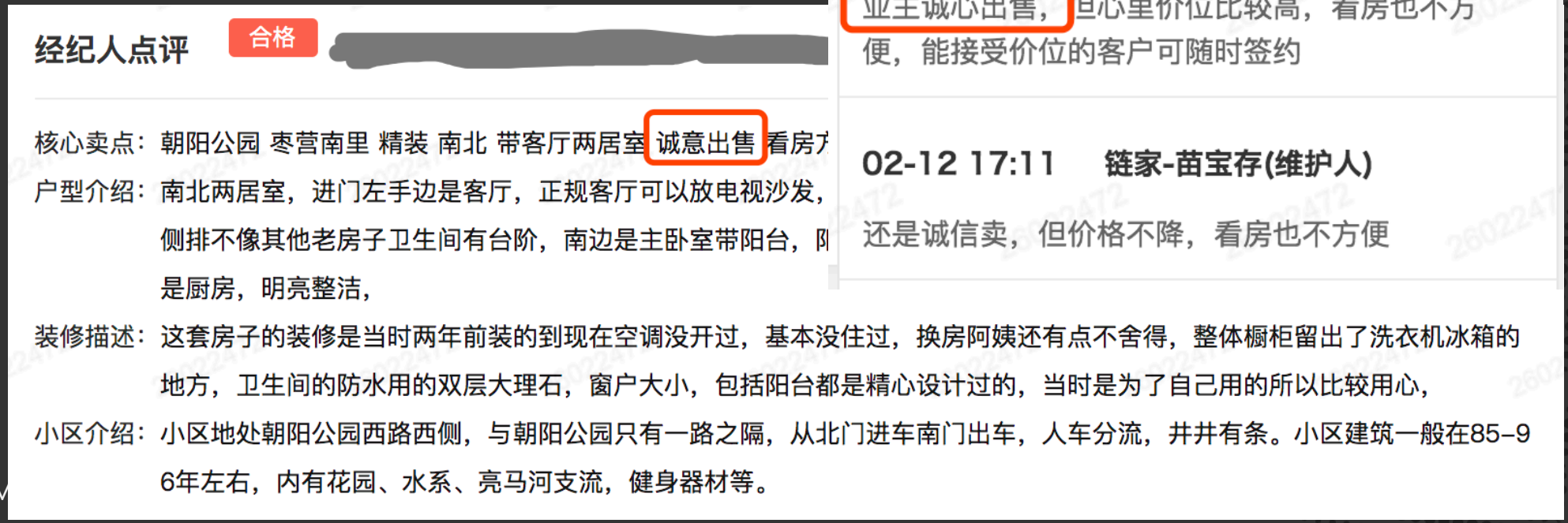
类型	占比
成交	56%
在售	34%
定金	5%
暂不出售	5%



v2.0+：持续优化

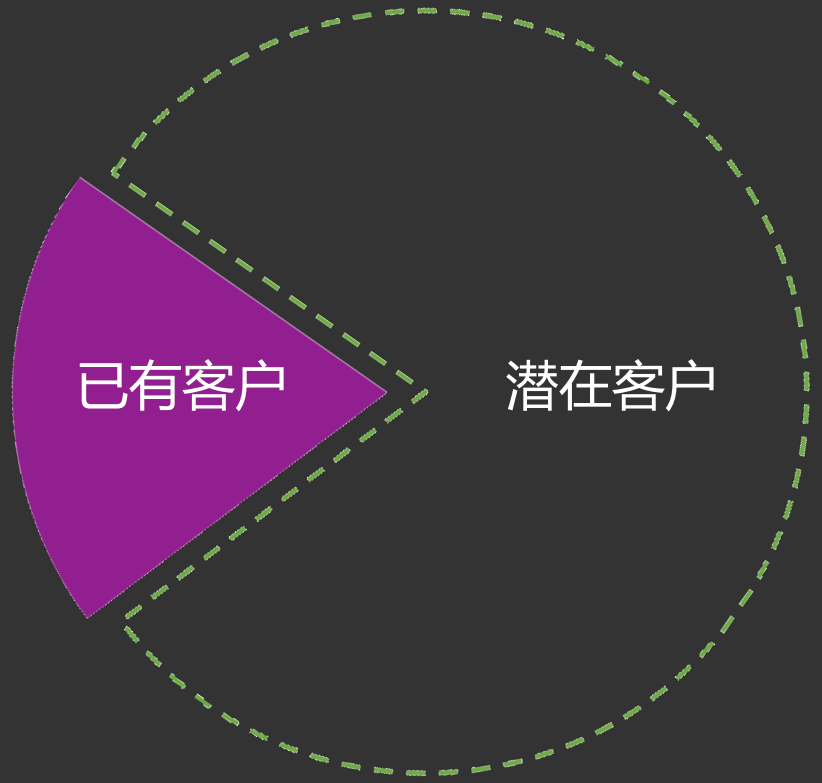
业主诚意

- 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



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
- 影响因子
  - 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 ...

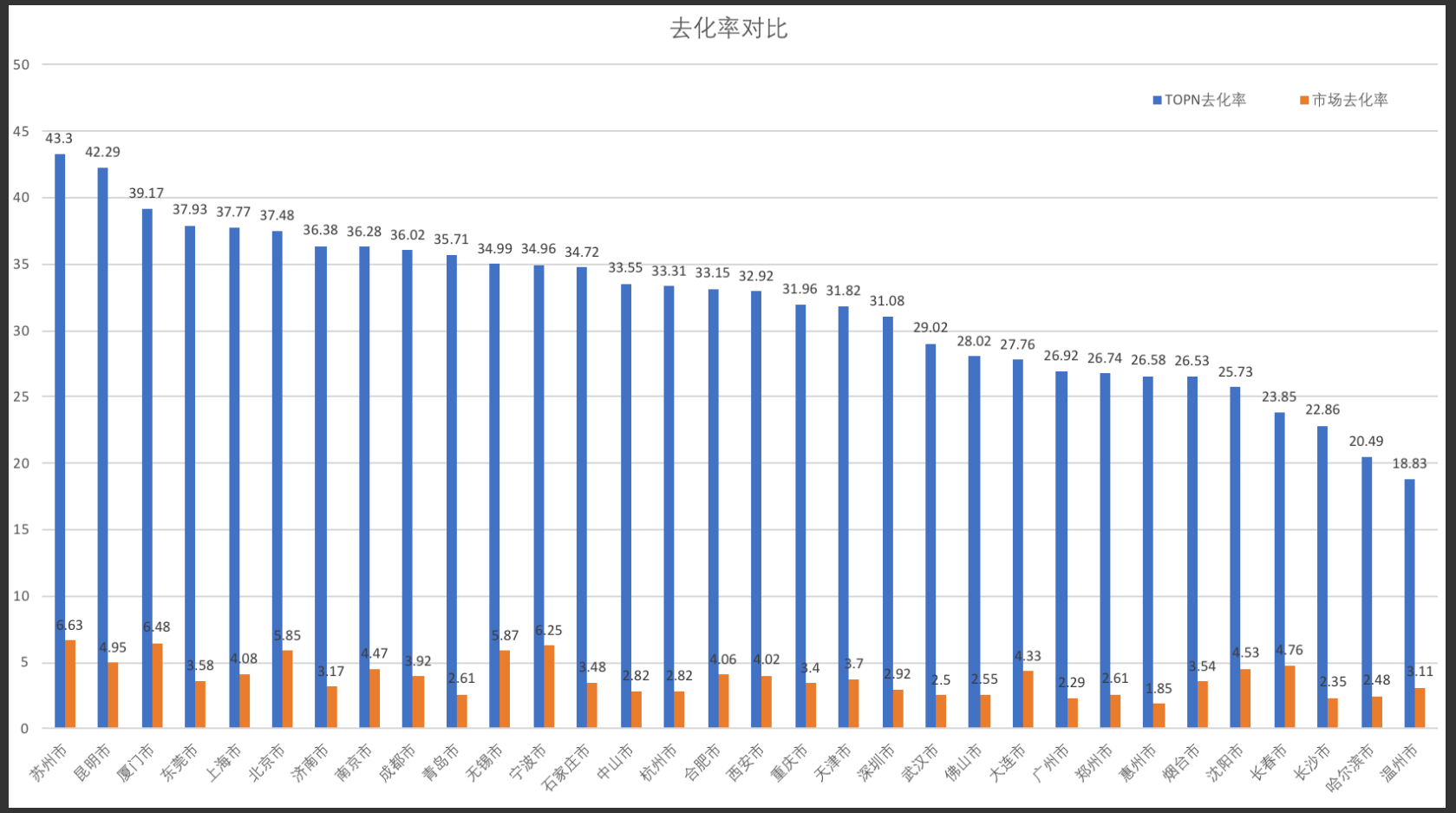
v2.0+：持续优化

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

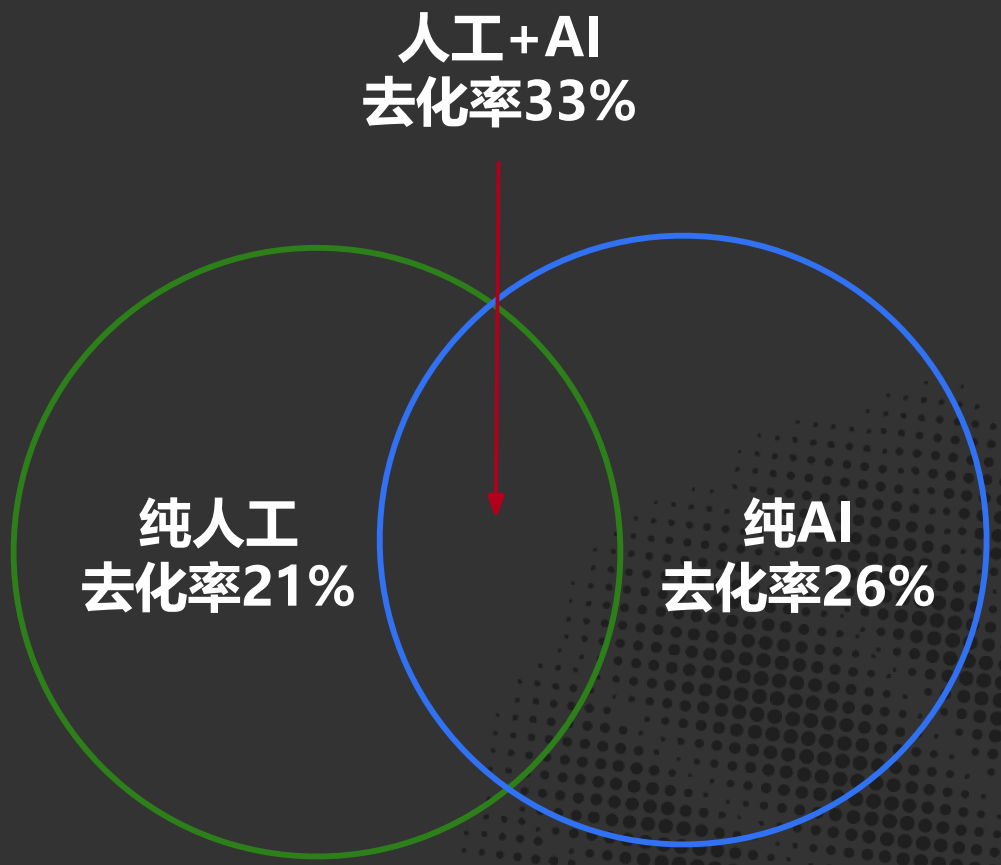


去化率(一周平均值)

- 指标
  - TopN去化率
  - N=2.5 * 周成交量
- 32个城市平均值
  - TopN去化率：31.7%
  - 自然去化率：3.8%

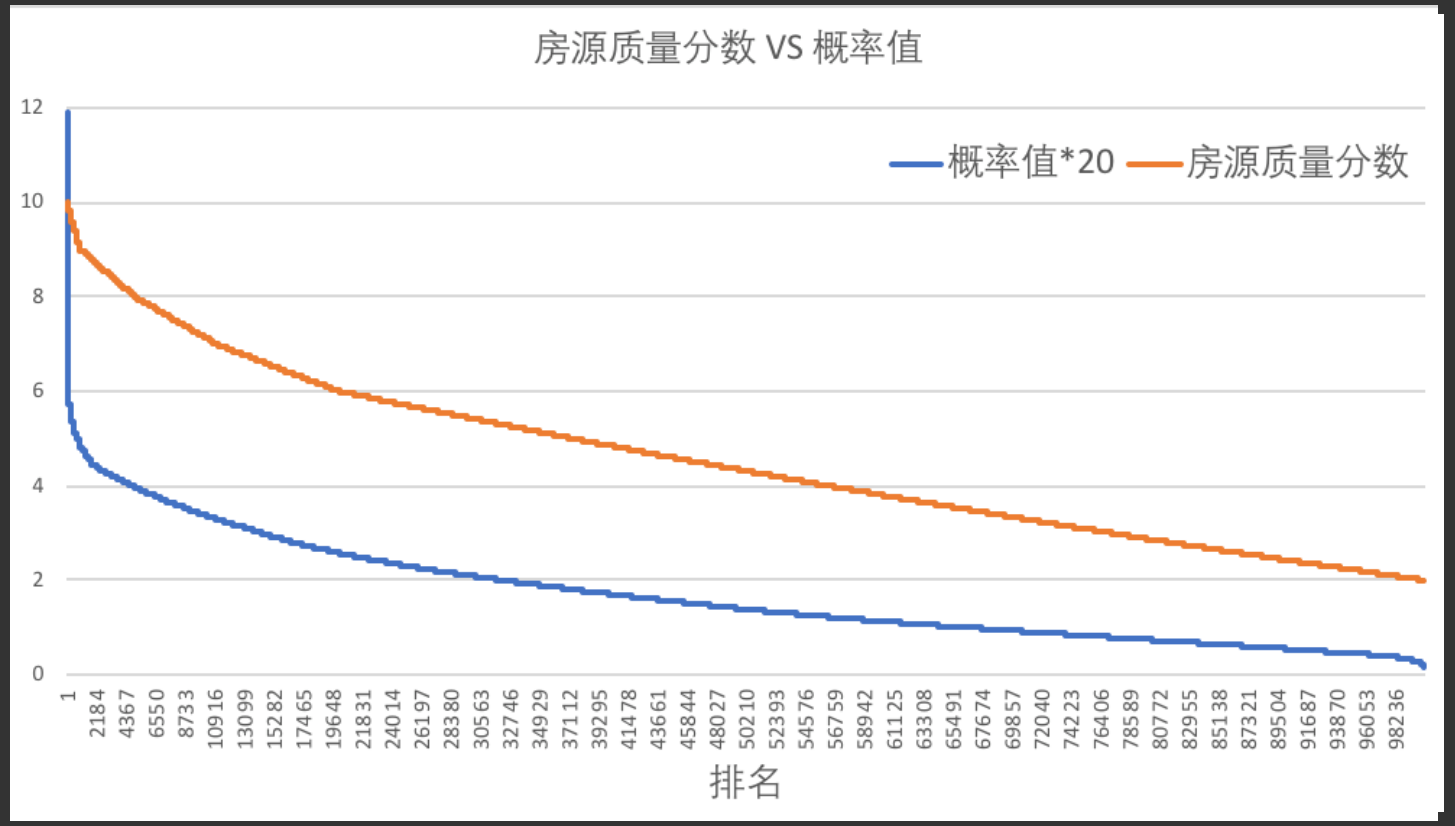
人工选房 VS AI选房

- 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

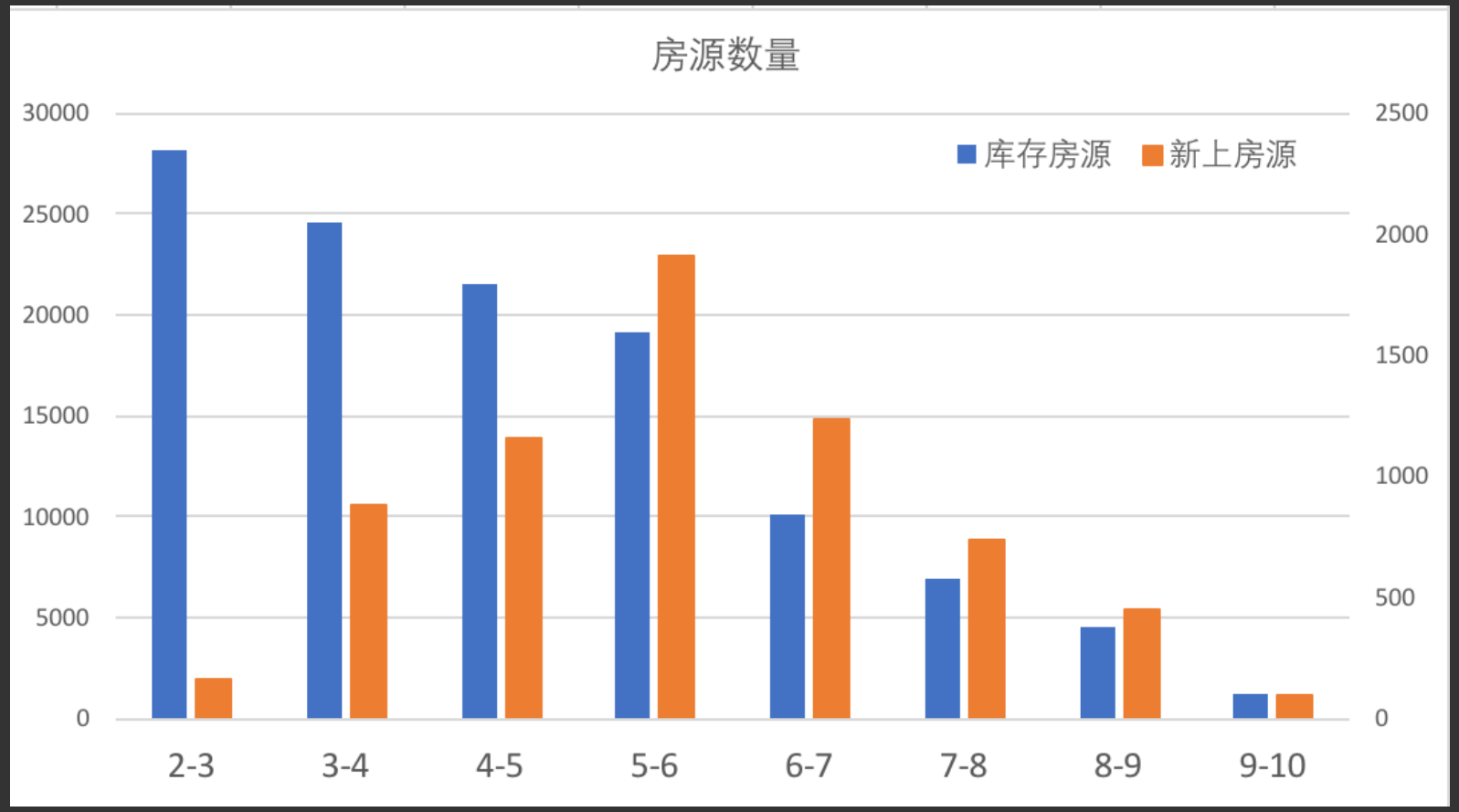
分数映射



- 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

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

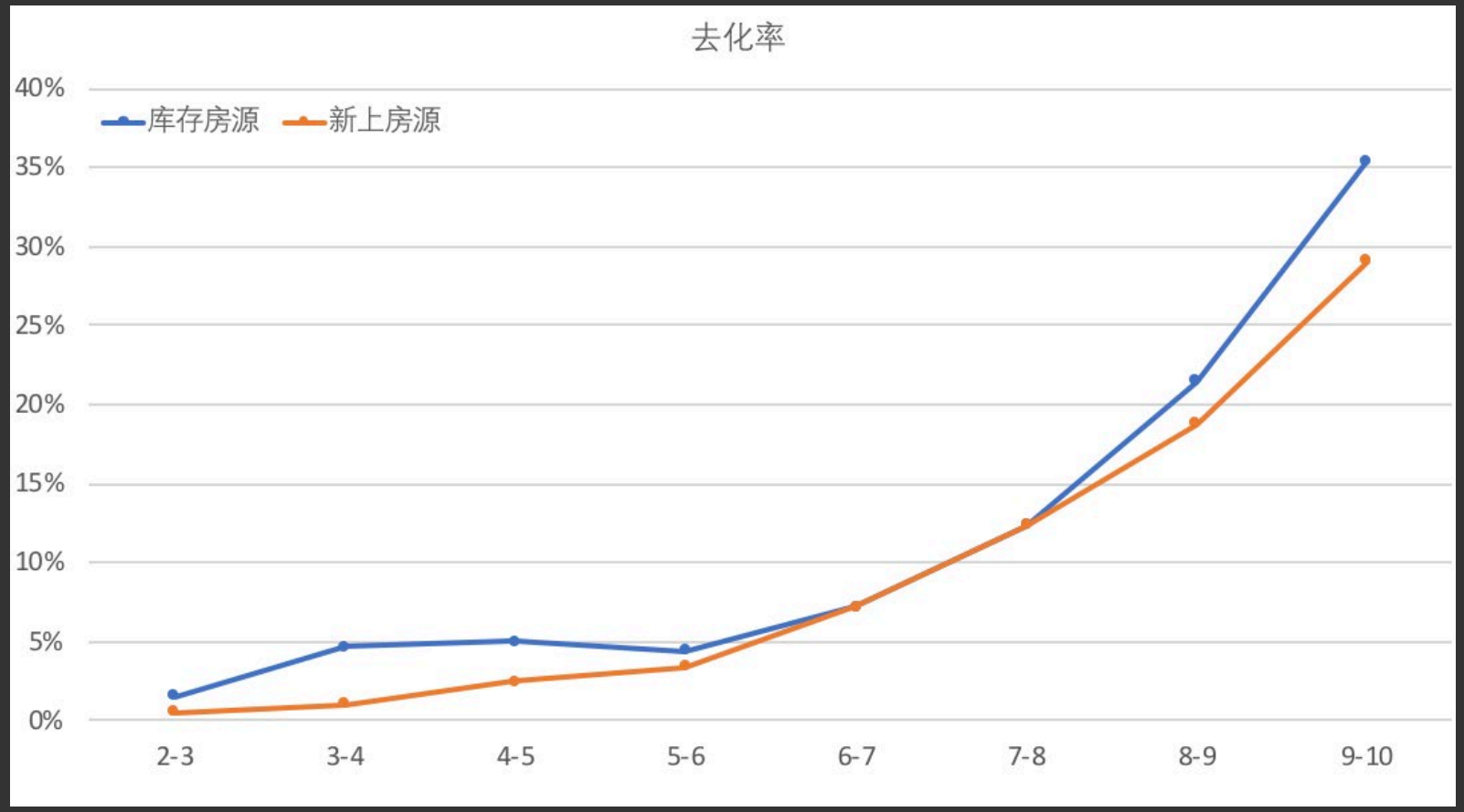
库存&新上房源



数量分布

- ✓ 新上房源分数略高

库存&新上房源



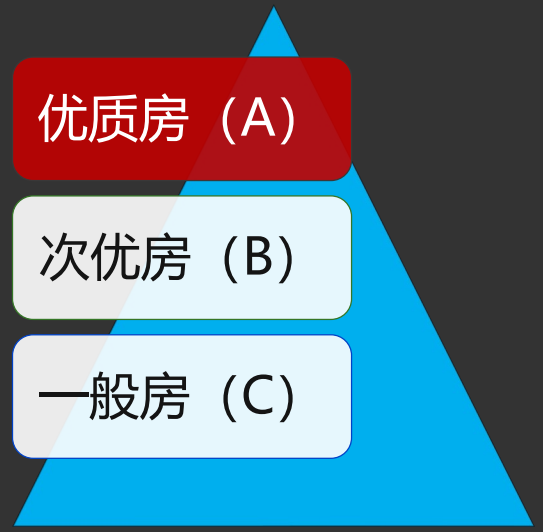
去化率

✓ 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

雷达图

雷达图

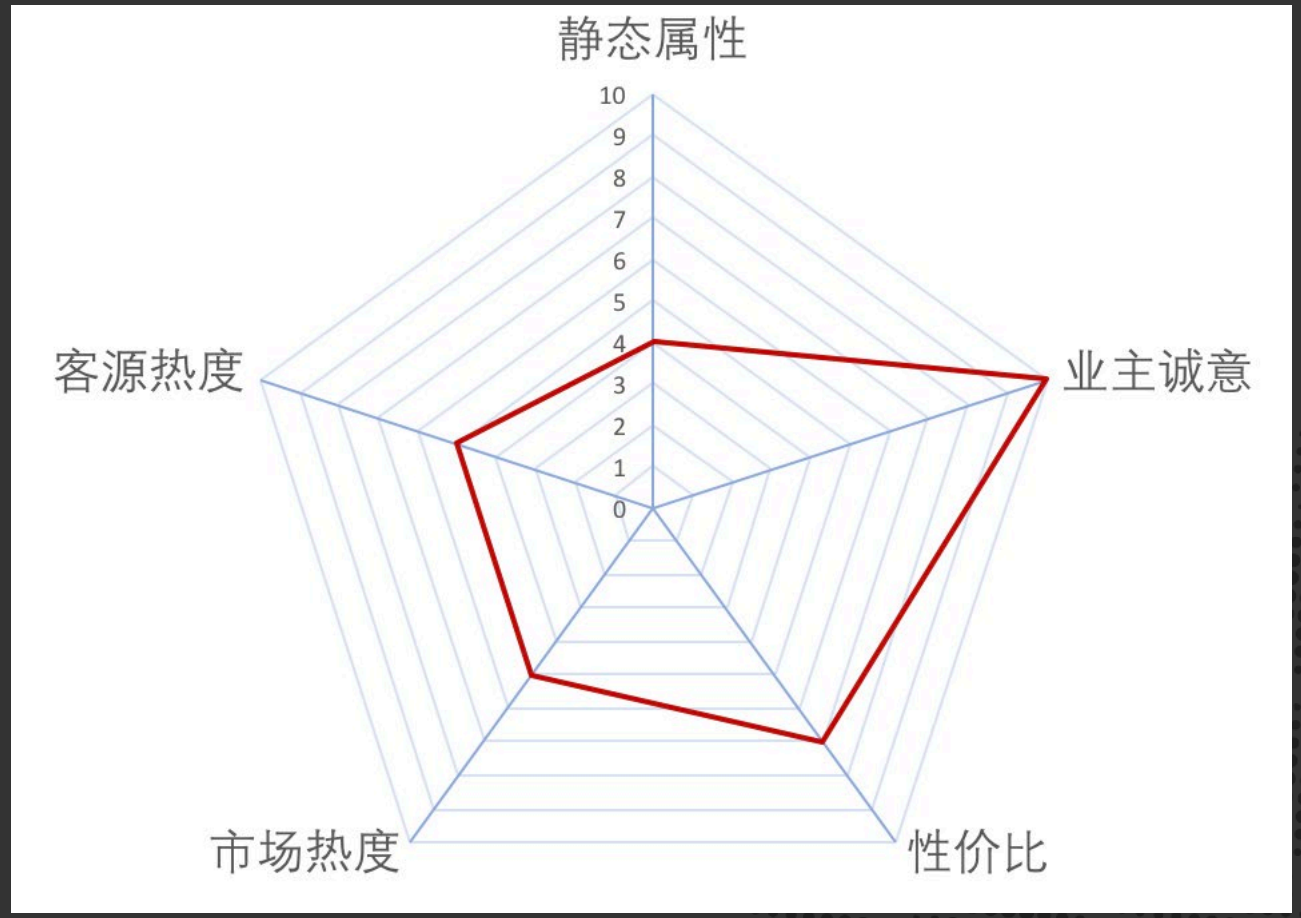
- 明示数据的核心打分维度
- 每个维度展示特征的优缺点
- 引导经纪人，提高分数

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

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）
房屋来源：人际开发-老客户主动委托
房源等级：A
VIP信息：暂无
房源状态：链家网展示
贝壳网展示

备件信息：齐全
历史委托：5次
维护人：高立艳
系统评分：9.9

清河 101104489391

学区房　满五唯一　VR房

AI选房 - B端场景

辅助经纪人选房

辅助经纪人盘房

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

Geekbang. 极客邦科技 InfoQ

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


贝壳
找房大平台
JOIN US
zhouyuchi001@ke.com