基于决策树的房地产市场数据挖掘新思路研究

王庆德 乔 夫①

摘 要：由于房地产消费者与销售人员间的信息不对称，许多“大数据”在交易完成前都无法获得，造成“实质贫数据”。以某知名房地产企业某商品住宅项目2014～2015年的销售成交记录（801条）和访问接待记录（865条）为样本，通过剔除“实质贫数据”的客户人口统计信息改进数据集，同时通过对已成交购房者按“消费状态”（即“刚需”“首改”“再改”和“升级”）进行决策树分类，得到“利益相关者”和“以投资为导向”两个稳定的关键变量。以此指导对访谈数据集的分类，发现购房者在与销售人员接触时，首先关注的是房源的面积大小，而非如价格、户型等其他信息，在兼顾精度的同时，发现可以直接指导人工销售的知识规则。

关键词：房地产 实质贫数据 决策树

一、引言

近年来，我国住房消费市场不断发育成熟，房地产行业也逐渐积累了大量的客户信息数据。这些房地产交易数据与客户数据，与宏观经济数据、房地产供给数据以及交易后服务数据一道构成了房地产市场的数据体系。这些数据对房地产企业经营的各个环节，尤其是后端销售环节具有重要意义（李海洋，2017）。但由于房地产交易不同于简单的网上购物，消费者必须借助专业销售人员的帮助才能完成交易，而消费者与销售人员之间的信息不对称，许多“大数据”在交易完成前都无法获得，因此针对历史数据的预测方法就无法使用。如何从这种名义上是大数据，实质上却是“贫数据”的房地产客户数据中获得有价值的信息，使数据真正可以应用与指导实践就成了一个重要的问题。本文在现有研究和决策树的成熟模型基础上，试图改善这两个问题，让机器学习不仅获得数据分析结果同时也获得知识，并可以将知识直接服务于人工房地产销售业务。

二、基于决策树的市场数据挖掘模型构建思路

决策树是一种重要的预测型数据挖掘技术，这种算法主要通过贪婪算法递归实现分类与预测功能。其系列算法起源于Hunt、Marin和Stone在1966年提出的单概念学习系统。Quinlan（1987）提出的ID3算法正式建立了决策树的算法框架。决策树分类预测的实现流程大致如下：（1）在理解问题的基础上，对数据进行清洗、赋值、标准化等预处理；（2）进一步使用具体决策树分类方法，利用训练样本构建决策树，并通过测试集样本检验决策树的效果；（3）根据结果调试相关参数与方法改进模型。

决策树在标准选择、改进思路和效果评价上有着多种不同规则。目前在许多理论与应用研究上，对模型效果的评价主要聚焦于预测精度。通过设置代价系统、改变抽样机制等方式，追求相关问题预测精度的最大化。但本文参考决策树在多个领域的应用研究（Liang等，2015；Kretser等，2015；Dhurandhar等，2015）认为，除此之外，研究方向还应包括：分析样本数据集与生成的数规则的关系、规则复杂度与预测准确性的折衷等方面。前一个问题是联系决策树算法与现实中具体应用的桥梁，而后一个问题则是将单纯的分类算法提升到系统科学的角度进行多目标的决策体系构建。本文即从这两个问题出发，首先从研究房地产销售数据的特性，站在从房

① 王庆德，国家信息中心，博士后，主要研究方向：广义虚拟经济学、国企改革、房地产金融和房地产经济、大数据研究等；乔夫，中国科学院大学经济与管理学院，博士研究生，主要研究方向：机器学习和数据挖掘。

地产交易的具体情境上分析数据的可得性及稳定性，以建立适当的数据集。此外，鉴于房地产交易的复杂性，消费者必然会借助专业销售人员的帮助完成交易，因此，房地产数据挖掘更应该对房地产销售人员的服务过程形成协助。房地产销售人员在服务过程中，对单个客户的精准把握自然是一方面，但在同样的时间里用尽可能少的交流，掌握切中购房需求的要害问题，对更多客户形成基本正确的判断，也是其拓展渠道提升业绩的方式。因此本文希望通过数据挖掘，得到分类标准稳定并易于解释的决策树规则。

三、数据分析

（一）数据说明与预处理

本文采用的数据包含两部分，均源自某综合性知名房地产企业的数据库，一部分是该企业在环渤海地区某城市一处住宅小区销售中心 2014 年成立以来的商品房住宅销售成交记录共 801 条；另一部分为该销售中心自成立以来的访问接待记录共 865 条。对这些一手数据做进一步清洗，在成交数据中剔除值缺失、明显输入错误记录 24 条，剩余有效记录 777 条；以同样的标准剔除访问接待记录中的无效记录 43 条，剩余有效记录 823 条。成为本文用于挖掘的数据集。

其中成交数据中包括 50 个字段，剔除没有信息含量的“序号”“组别”“职业顾问”“楼号”等字段后，剩余字段如表 1 所示。

表 1　成交数据特征分析

数据内容	数据特点	包含字段
消费者人口统计信息	数据稳定性强 隐瞒成本低	购房者年龄、职业、家庭构成、教育程度、家庭年收入、现居住社区、交款人年龄
消费者其他消费行为信息	数据稳定性差 隐瞒成本低	购房者日常上班路线、日常交通工具、日常消费场所、业余爱好、餐饮地点、私家车品牌及价格、手表价格、家庭价值观念
消费者住宅消费状态	数据稳定性强 隐瞒成本高	购房者置业需求
消费者购房需求	数据稳定性差 隐瞒成本高	购房者置业目的、对社区关注点、看重项目价值、物业管理、商业配套、交通配套、医疗配套、娱乐设施、赠送面积

此处对数据分析的特征分析，是基于对购房行为的基本判断。所谓数据稳定性强，是指这类数据反映的信息是客观事实，一般不具有随意性。如购房人的身份信息家庭信息无法根据消费者的意愿改变，但购房需求的细节如户型、周边配套设施等，属于消费者的主观意愿，本身没有客观事实与之对应，也就无稳定性可言。数据的隐瞒成本，是指消费者不披露该信息对于购房交易而言，是否影响交易的完成与完成质量。房地产的销售人员相对于消费者始终处于信息不对称的劣势地位。消费者具有选择是否披露自身相关信息的主动权。消费者在交易完成前，出于各种原因（家庭安全、隐私保护等）选择不披露与房产交易无直接关联的信息，就不会造成交易上的损失，因而隐瞒成本低。但如果一个消费者在与销售人员交流时隐瞒其对于价格、户型或周边设施等消费需求，这样的隐瞒将会对交易造成直接的影响。基于这一分析，容易发现：尽管房屋成交后的数据集由于产权登记这一流程，拥有上表中所有字段的信息，但在房地产交易完成之前，销售人员难以获得与消费者本次购房无直接关联的人口统计数据和其他消费习惯数据。因此，在数据集中应将两个门类的数据剔除，只保留后两个类型做进一步分析。

对于访谈接待数据，可整理如表 2 所示。

表 2　成交数据特征分析

数据内容	数据特点	包含字段
消费者信息来源	数据稳定性差 无法隐瞒	客户事件、认知渠道
消费者询价信息	数据稳定性差 隐瞒成本高	需求面积、承受单价、承受总价、户型选择、付款方式
消费者需求信息	数据稳定性差 隐瞒成本高	置业目的、购房主因、如未成交原因、如未成交转向

注：客户事件是指销售中心接待客户的方式，依据电访还是面谈以及是否首次接触分为四个选项。

由表 2 可知，在成交数据集中，各个字段均隐瞒成本较高，或如消费者信息来源这类信息由于导流渠道的确定性而无法隐瞒，因此均可以作为数据挖掘的数据集适用。

（二）实验结果及分析

基于上述数据处理后，在 SPSS16.0 平台上进行决策树分析。基于前章节的论述，本文以二项分枝，能生成较为简明的树规则 CRT 分类方法为主要方法，以 CHAID 为参考方法进行分析。针对所研究的问题，不同于类似银行信用、航班延误等预测，房地产的不同类型客户各有其价值与风险，刚需消费者潜在价值可能偏低，但需求迫切成交的几率也大，高端

升级型消费者潜在价值更高，但需求更多元，完成交易的几率较小。很难说哪类误判有更大的代价。因此，本文的分析中不设置成本偏好的不对称性。

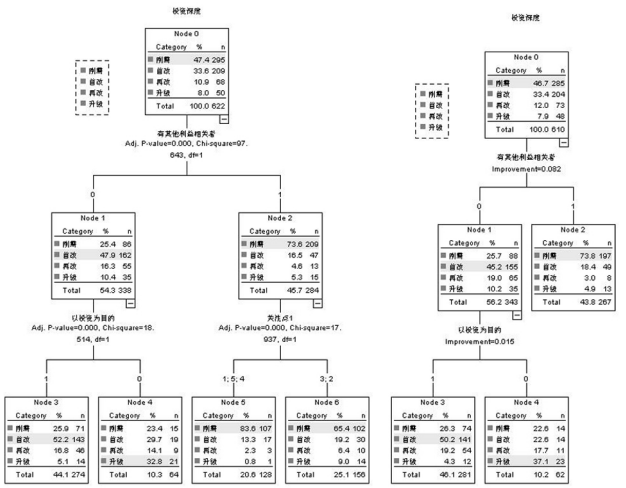
1. 对访谈客户记录的分析

通过观察访谈数据集可以发现，这一数据集建立决策树模型并分类的问题在于：缺少一个核心指标或组合规则将消费者加以区分，并以此为“类”对数据集进行分类，这与访谈数据没有结构化的设计、信息获得也比较零散有关。因此，本文首先对成交数据进行分析，试图解决对客户需求进行分类的“类”以何种规则确定的问题。

2. 对成交记录分析

通过对交易数据的观察发现，消费者的住宅消费状态这一信息具有优良特性：首先由于全国建立产权信息登记制度，消费者这一信息基本上属于确定信息；其次，根据现行法规，处于不同住宅消费状态的消费者在所适用的税收、贷款等交易内容会有很大不同，在这一信息上选择保留，将使自己在交易中蒙受损失。此外，房地产交易是大宗交易，客户的消费次数不可以直接反映出其财富水平、交易经验等相关信息，可以作为客户细分的代理变量。因此本文拟对消费状态（即刚需、首改、再改、升级）作为客户细分的标准，进行决策树分类。

我们随机选择占样本 80% 的数据点作为训练集，其展示的规则如图 1 所示。



（a） CRT 规则下的决策树　　（b） CHAID 规则下的决策树

图 1　训练集为 80% 时的成交数据决策树

进一步具体解释图（a）中右枝第二层的叶子集，关注点 1 的选项为：①建设质量，②价格低廉，③投资价值，④配套设施，⑤服务与环境。不难发现，在这一设置下，数据在不同的分类规则中，呈现出类似的树结构。首先，划分不同消费阶段客户的规则是是否有利益相关者，这一属性将刚需客户分离出来，在进一步分类时均表现为无法再分（CRT 规则下，尽管按关注点 1 属性进一步分类，但分类后两类预测值均为“首改”）。进一步，“是否以投资为目的”这一属性又将“首改”“再改”与“升级”客户加以区分。

这一分类规则有一定的经济含义，首先婚房与给父母买房，属于有利益相关者，这一类需求往往迫切缺乏考虑的余地，因此往往是刚需用户。而其他客户中，又以“关注住宅质量”还是“单纯为投资而购房”加以区分。已有一套住房的客户，首次购买改善性住房时往往更注重房屋升值空间，愿意升值牺牲居住质量；而已有多次购房经历的消费者，再次购房往往更关注生活质量；而已经购买过改善性住房，但其购买住房还没有达到升级标准的再改客户在分类中并不明显，是因为其兼具了首改客户与升级客户的特征。此外通过观察在 CRT 规则的右枝第二层可以发现，刚需客户关注设施环境等问题的比例远大于关注价格与升值空间，这一点与刚需客户往往购房自住或给家人居住，同时该笔交易对其家庭财富影响巨大有关。基于以上分析可知，利益相关人与投资导向是对房地产客户细分的关键规则，其他的规则并不重要，这一点对房地产的销售人员如何进行沟通有一定指导意义。将此规则简化为波士顿矩阵如图 2 所示。

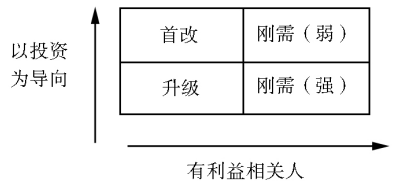
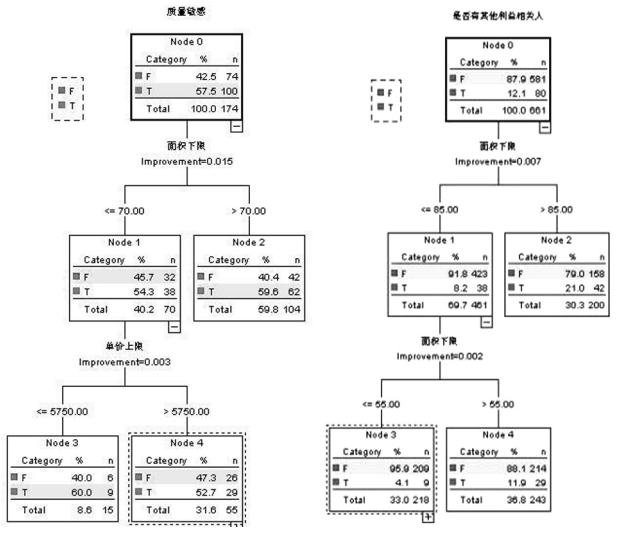


图 2　成交客户细分波士顿矩阵

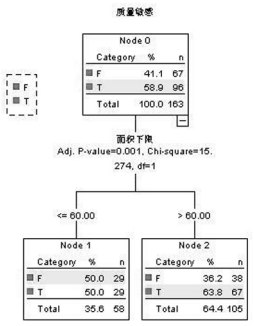
3. 投资深度重分类与访谈记录再验证

以消费状态对客户细分的主要规则是，是否有利益相关者和是否以投资为导向。基于这一认识，进一步对访谈数据集进行分析，结果如图 3 所示。

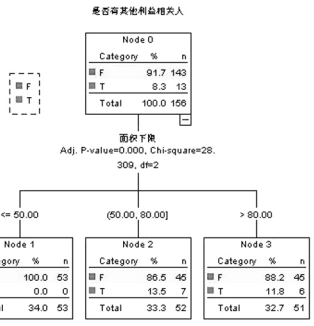


(a) CRT 方法下
“质量敏感”分类

(b) CRT 方法下
“利益相关人”分类



(c) CHAID 下
“质量敏感”分类



(d) CHAID 下
“利益相关人”分类

注：训练集样本比例为 80%，图示为测试集的分类结果

图 3 基于客户细分关键因素的访谈数据决策树

如图 3 所示，两种方法的相互验证表明这样一个观点：无论是通过“利益相关人”还是“质量敏感”来分类，对客户区分度最大的因素都是住房面积。这一点不难解释：当客户主动联系具体销售中心接受访谈时，可以断定其对于自身经济条件和诸如学区等关键因素均已形成判断，而所联系的楼盘，则恰好是其通过对自身信息判断后找出的选项，这一点可以启发销售人员，在同客户沟通时，首先根据上述判断，沟通房源本身的户型面积等问题，可以提高沟通效率。同时应注意，分枝的叶子集形式也有一定的信息含量，应注意到：倾向于购买大户型住宅的客户主要是“中间层”，“刚需”客户与“升级”客户则有购买小户型的倾向。

四、结论与政策建议

由于房地产消费者与销售人员间的信息不对称，许多在交易完成前都无法获得的“大数据”被用于数据挖掘和预测，造成“实质的贫数据”。本文通过对房地产销售数据稳定性与隐瞒成本的判断，剔除客户人口统计信息，改进了数据集。同时通过对已成交购房者按“消费状态”进行决策树分类，得到“利益相关者”和“以投资为导向”两个稳定的关键变量。以此指导对访谈数据集的分类，发现购房者在与销售人员接触时，首先关注的是房源的面积大小，而非如价格、户型等其他信息。本文在兼顾预测精度的同时，发现了可以直接用于指导人工销售业务的知识规则，房地产销售人员可借助以上规则，提高沟通效率、改善销售业绩。

根据以上结论，地方在制定住房政策时，也应充分利用数据挖掘的技术及其发现的知识：一方面，地方政府应与房地产企业建立更加系统的信息共享机制，建立认识住房需求的信息抓手，以便对已经存在的住宅交易进行市场调节；另一方面，地方政府可进一步善用拥有的户籍等信息，加强对居民房产需求的理解，进而改善区域内房地产开发的事前规划，使市场更加平稳有序地发展。

参考文献：

[1] LIANG C，ZHANGY，SHI P，et al. Learning Accurate Very Fast Decision Trees from Uncertain Data Streams [J]. International Journal of Systems Science，2014，46（16）：1～19.

[2] MINGERS J. Expert Systems—Experiments with Rule Induction [J]. Journal of the Operational Research Society，1986，37（11）：1031～1037.

[3] RAILEANU L E，STOFFEL K. Theoretical Comparison between the Gini Index and Information Gain Criteria [J]. Annals of Mathematics and Artificial Intelligence，2004，41（1）：77～93.

[4] 李海洋. 大数据在房地产营销中的应用研究 [J]. 房地产导刊，2015（24）：797～811.

[5] 梁循. 数据挖掘算法与应用 [M]. 北京：北京大学出版社，2006.

[6] 刘枏，轩朵，陈蕴. 大数据下房地产信息服务的挑战及对策研究 [J]. 建筑经济，2017，38（2）：77～81.

[7] 陶妍艳. 基于知识发现的房地产企业客户信息分析研究 [D]. 武汉：武汉理工大学，2006.

[8] 张崇，熊烨明. 基于海量网站日志数据的房地产需求指数研究 [J]. 数学的实践与认识，2017（5）：125～133.