数据驱动下房地产批量评估方法研究
——以大连市为例

杨智璇　张继瑜

摘要：随着城市面积的快速扩张，市中心土地市场供应的愈发紧缺，房地产交易市场已由新房向二手房转变，准确的房地产评估对于指导人们买卖房产，政府税收，制定经济发展战略都有十分重要的作用。21 世纪计算机技术和信息技术的发展，人们在房地产评估的方式上不再局限于传统的人工评估方式。在房地产评估方法研究中，伴随大数据技术和人工智能的进步，利用大数据技术来进行房地产评估逐渐成为了新的研究焦点。采用最新的机器学习算法结合房地产评估理论探索数据驱动房地产评估的可行性，经过实证证明相比于传统的机器学习方法，采用 CatBoost 算法进行房地产批量评估准确率高，具有极高的社会应用价值。

关键词：批量评估；大数据；机器学习；Catboost

中图分类号：F293　　**文献标识码**：A

文章编号：1001-9138-（2022）05-0054-06　　**收稿日期**：2022-04-03

DOI:10.13562/j.china.real.estate.2022.15.005

1 引言

在房屋交易中，二手房因配套较为完善以及价格更为低廉等优势，交易中愈发受到人们的青睐。现实中房地产转让、租赁、抵押、税收、征收、征用、司法拍卖、分家析产、损害赔偿、保险等活动对房地产评估均有需要。在这种情况下，政府、中介、买卖人对完善的二手房市场信息和准确的二手房价格需求越来越迫切。

传统的房地产估价方法以成本法、市场法、收益法为主，这些方法具有理论成熟，使用简单，应用案例广泛的优点。创新型的房地产评估方法如黄臻（2021）基于模糊实物期权法在房地产价格评估中的应用，也取得了较高的评估准确性。但上述研究方法适用于单宗评估，历史数据使用较少，对于批量评估工作来说评估效率低，难以满足大量业务的需求。

2 数据驱动房地产估价的理论框架

2.1 大数据驱动房地产估价的基本理论

大数据技术作为一种抽象的概念，简单来说就是对海量数据进行信息挖掘和数据分析来发掘数据的应用价值。21 世纪大数据技术在硬件的加持下飞速发展，在短短数年的时间里，大数据就实现了从概念到落地的过程，直接带动了全行业的技术变革。全球各行各业产生的数据总量已经呈现爆炸式增长，我们正快速经历着数字转

作者简介：杨智璇，东北财经大学投资工程管理学院硕士生导师。
张继瑜，东北财经大学投资工程管理学院硕士研究生。

基金项目：教育部产学合作协同育人项目“高校房地产本科专业体系改革与建设研究”（202102133002）；
辽宁省教育厅科学研究经费项目“空间信息协同下的智慧城市治理路径”（LN2020Q32）。

型，如何有效利用海量数据为社会带来贡献是大数据时代至关重要的事情。

大数据技术为各行各业的发展提供了新的动力和方法。房地产的相关数据形成了一定的规模并逐渐公开、透明，结合计算机技术、数据库技术的储存和运算能力，建立房地产自动评估模型进行房地产评估也越来越可行，利用自动评估模型可以实现房地产价格的更快速评估且评估的准确率更高。

刘辰翔等（2020）回溯了AVM的起源，并与传统评估模式进行对比分析，展现了房地产自动估价模型在中国广阔的应用前景。沈宏亮等（2021）得出研究结论，新冠疫情后房地产估价行业将会受到“互联网+大数据”前所未有的冲击。评估企业应主动出击，做好数字转型，迎接大数据时代的到来。

2.2 房地产大数据概述

传统的房地产评估方法存在较强的主观性，难以科学准确地揭示房地产市场的真实价值情况。大数据技术有利于改进传统理论和方法的不足，通过合理筛选和利用已成交房屋的交易数据，结合估价房地产所在区位特征，通过批量评估的方式，科学评估房地产在评估价值时点的价值。通过大数据技术与经典房地产估价理论和方法的融合，实现精准估价的目的。

尹延钧等（2021）对大数据挖掘中常用的分类算法进行分析，当前在大数据分析和数据挖掘阶段经典的分类算法主要有决策树、朴素贝叶斯、支持向量机（SVM）、神经网络分类算法等。关于大数据房地产估价技术，国内开展较晚，最新研究方法主要采用数学模型、回归分析、模糊数学、灰色预测、支持向量机、神经网络等梳理统计分析方法。究其原理，是将房地产属性因素进行量化，定量分析房地产价格的影响因素，并以此为依据进行房地产价格评估，这些方法的关键是依赖数据属性信息量化。由于受到行业数据的限制，难以应用和推广。

最早的批量评估研究始于Carbone和Longini（1977）基于多元线性回归方程建立AVMs（Automated Valuation Models）自动评估模型，进行房地产批量评估。近三年的研究主要集中于随机森林模式和Catboost等新兴算法研究，例如，李宇琪（2018）选用基于决策树的随机森林模型，对获取的房价数据进行清洗、归一等预处理，从信息增益等角度出发寻找影响房价的主要因素，从而训练获得了较为准确的房地产评估结果。CatBoost算法出现时间较晚，在各应用领域的研究还不够广泛。

在文献整理中发现，多数研究集中在运用机器学习算法模型进行房地产批量评估的研究，但是，对于众多机器学习算法，哪类或者哪种方法更加精准却鲜有研究。本文的研究将对比多元线性回归模型、随机森林模型和Catboost模型，通过实证检验方法揭示模型的精准性。

3 模型构建

3.1 数据驱动房地产批量评估模型构建

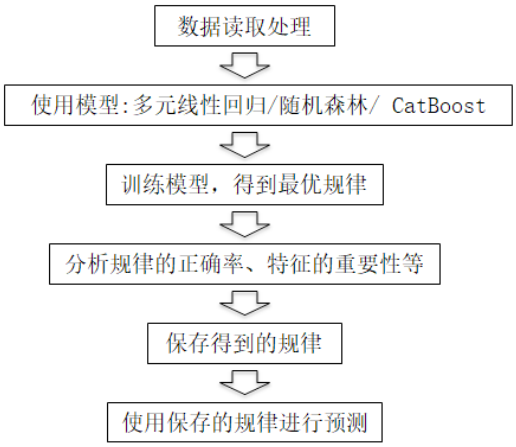
房地产批量评估是基于以往发生交易的价格数据，对尚未发生交易的房地产价格进行合理预测。在构建数据驱动房地产批量评估模型中，需要注重数据类型、数据集、数据拟合度和准确性等，此类因素将决定估算结果正确性（见图1）。

通常情形下，大数据建模应遵循以下步骤：一是房地产交易信息搜集和属性信息数据化，二是构建房地产批量评估模型，三是计算评估结果。

3.1.1 多元线性回归建模

唐文广等（2019）对多元线性回归模型在房

图 1：建模分析思路



地产评估中的应用展开研究，多元线性回归分析是指两个或两个以上的自变量的回归分析。

多元线性回归的优点：

（1）建模速度快，不需要很复杂的计算，在数据量大的情况下依然运行速度很快。

（2）与一元线性回归相比，更加接近现实，增强对因变量分析估计的贴合性。

多元线性回归进行房地产估价方法较简单，效率较快。但现实生活中房地产价格的影响因素较为复杂，各因素对房价的影响不尽相同，只简单进行回归评估其结果缺乏可靠性。

3.1.2 随机森林建模

随机森林是基于Bagging算法中的自助抽样技术将分类树进行组合，提出的一种使用灵活且精确度较高的机器学习算法。在实践应用中，随机森林算既可以应用于解决分类问题也可以应用于解决预测问题，随着研究深入，其在医学、管理学、经济学等众多领域应用均有优异的表现。

随机森林算法对多元线性不敏感，对缺失数据和非平衡的数据比较不敏感，同时还克服了决策树分类器产生过拟合现象，在特征维度较高解释变量多达几千条的时候，也能得到非常好的预测结果，因此被称为最好的机器学习算法之一。

随机森林有很多优点：

（1）相比于传统的机器学习算法，其在处理数据量较复杂时，也可以有较高的精准度。

（2）随机森林可以处理大量的变量问题，一个x类的名义变量可以用x-1分叉树来记录，并且可以随机选择部分数据来分类。

（3）其可以在决定类别时，通过编程调用来评估变量在分类中的重要性。

（4）在建造森林时，随机选择决策树节点划分特征，这样在样本特征特度较高时，仍然能较高效率的训练模型。

（5）随机森林建模速度快，对部分数据缺失不敏感。

随机森林算法相比于线性回归算法更加复杂，对影响因素的考察更为全面，是现今大数据估价研究中研究较深的一种算法。曾双（2021）研究表明采用随机森林模型评估相比于传统的房地产评估方法，预测精度更高，具有较大的适用性。

3.1.3 CatBoost 建模

2017年，俄罗斯搜索巨头Yandex首次公布了CatBoost算法，CatBoost算法是在GBDT算法框架下进行改进的，其相对于改进前的算法在准确率方面有了很大的提升，实际应用中表现也更加优秀。由于该算法目前较新，在各个领域的研究扩展深度尚浅，尚待探索。

CatBoost算法基于对称决策树为基础学习器，使用简单，调节参数较少，准确率极高。有别于其他的机器学习算法，CatBoost算法最大的特点是可以高效处理类别型特征。除此之外，它还对GBDT框架的机器学习算法进行了优化，解决了机器学习中常见的梯度偏差和预测偏移问题，

大大降低了模型过度拟合的发生，提高了算法的泛化能力。

CatBoost主要有三大创新之处：

（1）程序中嵌入了自动类别型特征数值化处理，不需要进行数据的预处理工作。

（2）CatBoost算法还利用了特征间的联系进行组合，丰富了数据之间的组合类型和维度。

（3）采用排序提升的方法对抗训练集中的噪声点，从而避免梯度估计的偏差，进而解决预测偏移的问题。

近来CatBoost算法在各个行业领域均展开了研究，且取得了不错的应用结果。但其在房地产评估领域的研究尚未展开，本次研究采用CatBoost算法训练出评估模型进行估价并与前两种算法进行准确率对比，探索该算法在房地产估价领域的应用价值。

4 实证检验

4.1 数据选取和预处理

4.1.1 样本选择

本研究以大连市2021年10月二手房网站公布的二手房挂牌信息为研究对象，通过Python爬虫技术来获取非典型的100个小区近6000条房屋挂牌信息为数据基础进行实证分析。由于网上公布的房屋属性数据种类繁多，因此在数据分析前要明确建模需要获得的房屋属性信息。

房价的影响因素主要可从个体因素、邻里环境、区位因素三个方面进行选择，在公开房屋价格属性的基础上，选取了房屋交易过程中，价格影响较大的属性作为目标属性。这些属性的选取符合房地产评估的原理，包含的属性有单价、房龄、卧室、客厅、卫生间、楼层、总层数、面积、类型、建筑类型、朝向、装修、梯户比、电梯、用途、绿化、容积率、物业费、区域共计20条属性信息。

4.1.2 数据预处理

由于获取的数据是从网页中抓取的，数据的分布和房屋的属性信息并没有完全公布，存在数据缺失和数据异常情况，因此在训练前要进行数据的预处理。数据的预处理内容主要包括补充缺失值、修正异常值和量纲一致处理等。

常用的数据缺失和异常处理的方法有删除法、替换法、插值法。而对于数据量纲不一致问题常用的处理办法有零均值标准化、最小最大规范法、对数变换法等。数据分析时应根据数据的特征来灵活选择合适的数据处理方法。

用计算机统计预处理后的数据，被评估房屋共计5208例，其中，自变量属性20条，因变量属性1条。接下来将80%的数据即有4166例房屋数据作为模型的训练数据集，剩余20%比例即有1042例房屋数据作为模型的测试数据集来验证模型的评估效果。

数据集划分完成后，将训练数据集带入算法中推出评估模型，再将测试数据集中的数据带入评估模型中输出预测结果，对预测结果进行指标分析（见图2）。

图 2：归一化处理后二手房属性图

	总价	行政区	单价	房龄	卧室	客厅	卫生间	楼层	楼高	面积	...	建筑类型	朝向	装修	梯户比	电梯	用途	抵押	绿化率	容积率	物业费用
0	235.00	1.0	0.27	0.58	0.17	0.50	0.0	0.33	0.33	0.13	...	0.5	0.67	1.0	0.50	1	0.5	0	0.47	0.17	0.62
1	178.00	1.0	0.29	0.58	0.17	0.25	0.0	0.67	0.33	0.09	...	0.5	0.67	1.0	0.50	1	0.5	1	0.47	0.17	0.62
2	456.75	1.0	0.28	0.63	0.83	0.50	0.5	0.33	0.33	0.28	...	0.5	1.00	1.0	0.75	1	0.5	1	0.47	0.17	0.62
3	190.00	1.0	0.32	0.58	0.17	0.25	0.0	0.67	0.33	0.09	...	0.5	0.67	1.0	0.50	1	0.5	0	0.47	0.17	0.62
4	270.00	1.0	0.31	0.63	0.17	0.50	0.0	0.67	0.33	0.14	...	0.5	1.00	1.0	0.75	1	0.5	1	0.47	0.17	0.62
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5203	273.00	0.0	0.37	0.21	0.17	0.50	0.0	0.67	0.41	0.12	...	1.0	1.00	0.5	0.75	1	0.5	0	0.73	0.10	0.78
5204	297.00	0.0	0.25	0.32	0.33	0.50	0.0	0.33	0.36	0.19	...	1.0	1.00	1.0	0.75	1	0.5	1	0.73	0.10	0.78
5205	49.80	0.0	0.07	0.32	0.00	0.25	0.0	0.33	0.17	0.04	...	1.0	0.67	1.0	0.75	1	0.5	1	0.73	0.10	0.78
5206	296.00	0.0	0.37	0.21	0.17	0.50	0.0	1.00	0.36	0.14	...	1.0	1.00	0.0	0.50	1	0.5	1	0.73	0.10	0.78
5207	320.00	0.0	0.37	0.21	0.33	0.50	0.0	0.67	0.36	0.15	...	1.0	1.00	1.0	0.75	1	0.5	0	0.73	0.10	0.78

5208 rows × 21 columns

4.2 结果分析

4.2.1 回归模型评价指标分析

本研究使用多元线性回归、随机森林回归、CatBoost回归这三个算法进行建模预测，需要对预测结果进行比较，来衡量三种方法评估结果的好坏，主要从模型的准确率和预测效果两部分来考虑。对于本研究房地产评估的问题属于机器学习中回归型问题，衡量模型的预测能力，主要是从两方面考察，首先考察三种模型的拟合度，即模型评估的准确率，其次通过对比测试集的测试指标，即平均绝对误差（Mean Absolute Error，MAE）、均方误差（Mean Squared Error，MSE）、均方根误差（Root Mean Squared Error，RMSE）来比较三种回归方法的优劣（见表1）。

表 1：回归评价指标对比表

	多元线性回归	随机森林	CatBoost
预测精度	0.929	0.953	0.99
平均绝对误差 MSE	1867.87	1142.56	338.42
方误差 RMSE	43.22	33.8	18.4
均方根误差 MAE	21.3	20.89	4.58

通过回归评价指标对比表可以看出，在二手房价格评估结果中，随机森林回归模型在各项评价指标中均优于多元线性回归模型，这说明相比于传统的多元线性回归模型，随机森林模型具有较强的评估优势，能够很好地解决非线性问题。其可以在回归拟合时表现良好，也可以在外展预测时确保较好的扩展能力。

同时也可以看出在二手房价格评估中CatBoost回归模型在各项指标表现上大幅优于多元线性回归模型和随机森林模型，这说明CatBoost回归模型具有更优的非线性拟合的优势，能够更好地解决房屋价格评估问题。其不光可以在回归拟合时使得模型表现更佳，并且在外展预测时可以确保较好的扩展能力，获得更优的二手房评估准确率。但同时在构建模型的过程中，CatBoost算法建模的工作量要明显低于随机森林模型。

4.2.2 回归模型评价指标分析

将预测集的房地产进行评估，通过运行程序，输出评估房价和真实挂牌房价的拟合图。

通过对比三种模型拟合优度图形，可以明显的看出，CatBoost模型评估预测值与真实值贴合

图 3：多元线性回归评估值拟合优度图

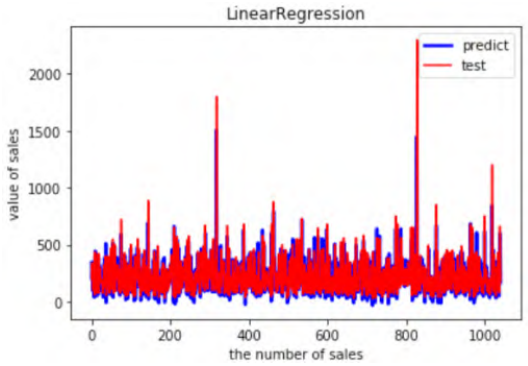


图 4：随机森林评估拟合优度图

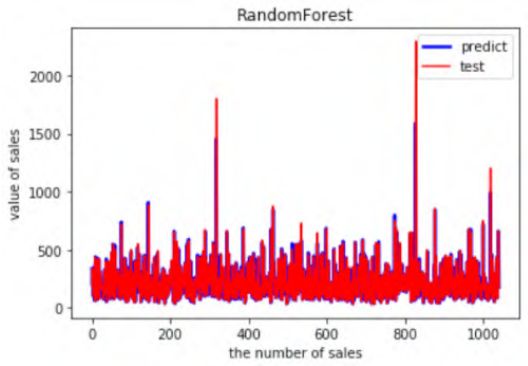
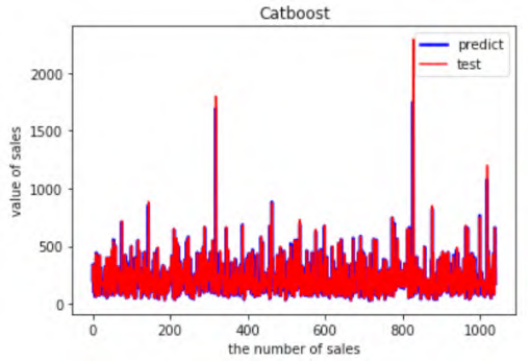


图 5：CatBoost 评估拟合优度图



度更高，具有更高的评估准确性（见图3、图4、图5）。

5 结论

通过对回归模型的预测精度、平均绝对误差、方误差、均方根误差等的比较，可以确定CatBoost回归模型相比其他的机器学习方法，评估效果更加优良。在CatBoost回归模型中，其预测精度达到了99%，是其他机器学习方法难以企及的，是当前研究中最优的以数据驱动的房地产批量评估方法。

参考文献：

1. 黄臻 . 基于模糊实物期权法在房地产价格评估的研究——以天津市为例 . 中国资产评估 . 2021.10

2. 刘辰翔 王卓 胡永强 . 大数据时代：从传统评估到自动估价系统 . 中国资产评估 .2020.04

3. 李宇琪 . 基于随机森林的房价预测模型 . 通讯世界 .2018.09

4. 苗丰顺 李岩 高岑 . 基于CatBoost算法的糖尿病预测方法 . 计算机系统与应用 .2019.28（9）

5. 沈宏亮 徐志革 . 后疫情时代房地产评估行业发展的应对策略探讨——基于波特五力模型的分析 . 中国资产评估 .2021.10

6. 唐文广 王梦茹 . 多元线性回归模型在房地产评估中的应用 . 科技和产业 .2019.19（06）

7. 尹廷钧 李灵慧 周蕊 . 大数据挖掘中的数据分类算法综述 . 数字技术与应用 .2021.39（01）

8. 曾双 . 随机森林模型在房地产评估中的适用性分析 . 中国管理信息化 . 2021.24（19）

9.Carbone R,Longini R L.A Feedback Model for Automated Real Estate Assessment.Management Science.1977.24（3）

10.Guomin Huang,Lifeng Wu,Xin Ma,et al.Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions.Journal of Hydrology.2019