2019 年（第六届）全国大学生统计建模大赛

基于“人才引进”政策的
西安楼市现状探究及房价预测

参 赛 单位：西安理工大学

参赛者姓名：徐　征

洪 明 月

宋　玉

目 录

表格和插图清单

摘　要

历经两年零两个月的西安市人才引进政策效果显著，为更好地吸引人才留住人才，提升西安市的楼房宜居指数迫在眉睫。本文从横向纵向出发，立足整体与局部，就西安市房价的未来发展趋势以及周期规律作以分析，对比全市不同经济及行政区域的楼市交易情况，从而为决策者提供量化参考依据。

首先基于特定网站收集相关数据，并利用 Python 以及 R 语言进行数据预处理。随后针对房价进行分析，基于时间序列模型进行纵向探究，截止今年 6 月西安市房价未来四个月的平均值分别为 9776 元、9850 元、9902 元以及 9964 元。借助 Cox-Staut 趋势存在性检验认为房价持续增长无统计学意义。综合供需类指标对全国与西安市房价周期进行谱分析，得出计算相关指标并筛选变量建立卡尔曼滤波模型。分析看到，2002 年至 2018 年西安市房地产市场的供求周期时间长度大致相同。长远发展而言，市场将处于供求关系较匹配但需求更高的阶段。为探究影响房价的显著性因素，针对房价进行逻辑回归，横向分析结果可知，房屋总数、楼栋总数、房源、银行、开发商以及卫生服务变量较大程度的影响房价。

为进一步探究交易量情形，利用 GBM 梯度提升树算法对交易量残缺记录进行预测，并反馈得到针对全市交易量的边际效应分析结果，相对于浏览量的显著影响，房地产均价对全市交易量的影响并不显著，同时房屋建筑年代影响突出，“抢人大战”策略的进一步开展会显著影响西安楼市的交易量情况。西安市整体较多楼市呈现出中等温热状态，主城区楼盘交易情况则呈现疲软态势。考虑到随机森林筛选变量的稳健程度优于基尼指标，故运用随机森林对“三廊三带一通道”的产业空间格局下的曲江区、高新区、经开区以及长安区进行局部分析。长安区房价较低，但是越靠近大学，房价越高。曲江区邻近大雁塔等地标建筑群，基础设施完善，医院指标排名较前，房价排名第一。物业费及幼儿园等教育配套设施成为经开区不可忽略的指标。高新区一直居于西安市房价制高点，绿化问题成为该区域影响售楼行情的重要影响指标。

整体看来，西安市房价增长速度放缓且周期基本与全国保持一致，全市交易量呈现温热态势，居住环境成为关注热点。外来开发商一定程度上刺激西安市楼市发展，交易量地域差异性明显。如何进一步实现西安市产业如工业、文化产业的整体布局与居民基础设施建设的均衡化发展是留住人才的关键。总之，预期西安市房价虽仍有上涨趋势，但整体行情平稳，房源供给较为充足，为迎接更多优秀才子落户西安做好实质准备。

关键词：人才引进　多重插补　谱分析　卡尔曼滤波　梯度提升树

一、引言

人口，是特大中心城市的指标之一；人才，更是城市发展的蓄能电池，基于培育大格局视野下的西安来说，高层次人才的引进更是落实“人才强市”战略、以人才优势助力发展优势的关键因素[1]。从 2017 年放宽户籍准入以来，大西安“留”“引”结合加入了轰轰烈烈的“抢人大战”。数据显示，期间超过 115 万人成为“新西安人”，其中学历落户和人才引进占比高达 64.05%。“抢人大战”的背后是国家大力推进的大都市圈战略，是中国新一线城市有限名额之间最残酷的竞争[2]。人口的去留，不仅在于城市管理者的告白，更在于产业的转移、升级，就业、教育、城市宜居指数的提升等方方面面。户籍政策的再度放宽、针对人群的年龄降低、抢人时段更加密集是近三年来西安“留人”攻略的显著变化。

然而宽松人才引进政策降低房地产市场限购门槛后，西安本就供需结构紧张的房地产市场情况愈加引人注目。在国家去库存的政策影响下，区域内房源库存告急造成区域内房价升高。且西安市自出台“四限”（限购、限售、限贷、限价）和购房摇号政策后，西安各置业热区，房价整体偏高。据相关统计调查，应届毕业生毕业后首先考虑的是生存问题，同时也考虑长远发展问题。90.8%的应届毕业生对毕业后能获得的工资待遇水平很看重；84.3%的应届毕业生看重的是更多的发展机会[3]。“我们将用最优的创业政策、最好的发展环境，让每一位奋斗者都能在这里梦想成真，找到施展才华的舞台。”6 月 25 日晚，在“2019 西安大学生毕业盛典”上，面对即将走上社会的莘莘学子，西安市长李明远满怀诚意地说。因此西安市用一场国家级的送别仪式向全国人才传输积极打造人才发展环境，用更加开放友好宜居的西安不断吸引全国各地的人才。

于此同时根据中国社会科学院财经院创新工程重大成果《中国城市竞争力第 17 次报告》显示，西安宜居竞争力下滑 27 位，排全国第 51 名。据相关专家分析，归根结底在于西安市房价的暴涨，造成一定的购房恐慌。然而人口、人才、产业，是一个城市保持可持续竞争力的关键所在，如何科学合理对西安市的产业以及房产布局成为进一步吸引人才以及留住人才的关键。

改革开放看深圳，“一带一路”看西安。西安，如同正在膨胀的海绵源源不断的吸引全国各地的优秀人才。但如何把高素质人才真正聚集在一起，而不是成为来去匆匆的过客，基于“人才引进”策略下的西安市房价市场研究势在必行。

二、数据预处理

（一） 数据来源以及建模分析工具

我们主要从全国、西安地区入手进行数据调查分析，主要采用 Python、R、SPSS 以及 Matlab 进行综合分析。为确保数据来源的真实可信基于以下数据来源网站进行相关数据查询：

国家统计局：http://www.stats.gov.cn/

陕西统计局：http://www.shaanxitj.gov.cn/

房天下：http://fdc.fang.com/index/BaiChengIndex.html

安居客：https://xa.anjuke.com/

公积金：http://zfgjj.xa.gov.cn/xxgk/gjjsjtj.htm

国信房地产信息网：http://www.realestate.cei.gov.cn/tongji/allcity/tjss.aspx

（二） 完整数据相关性可视化结果

由于涉及交易量、成交量、绿化率等与房地产相关的共计 35 项指标，因此考虑到不同小区开发时间的差异以及对外信息的透明度不同，首先对 35 项指标中的连续性指标利用 R 软件中的 mice 包、VIM 包通过数据插值进行数据预处理。

通过计算其中 20 个连续变量之间的皮尔逊相关系数，通过选取相邻两变量未缺损的数据进行计算的到各变量间的线性相关系数，为方便说明对相关矩阵进行可视化：

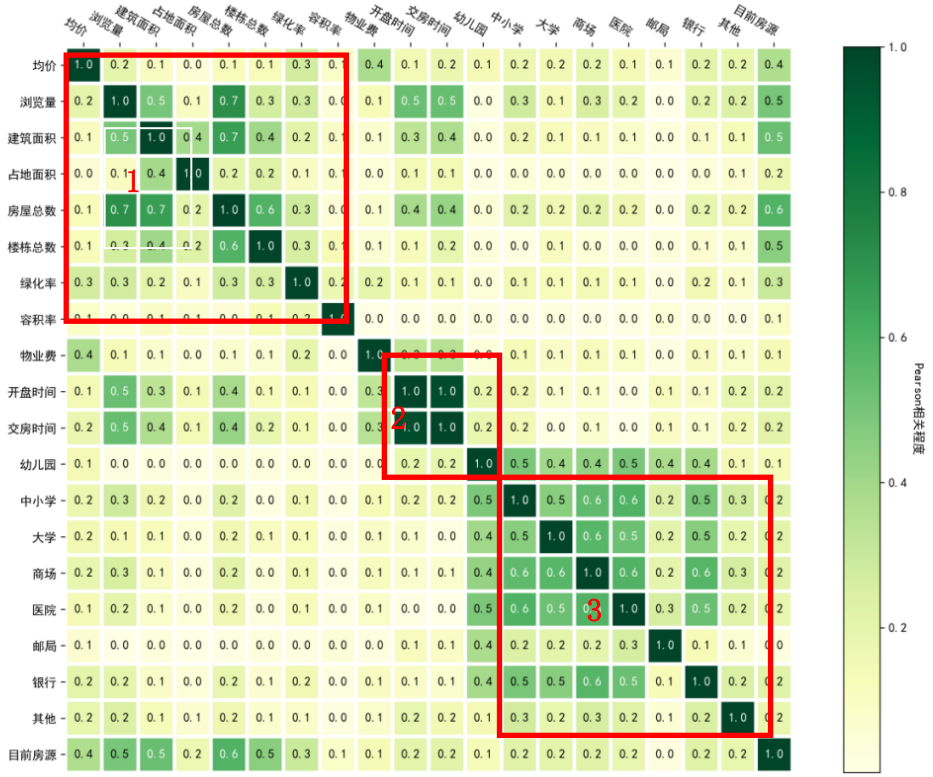


图 1　皮尔逊相关系数

根据图 1 所示，颜色越深指标之间的线性相关程度越高，分析可知其中三个区域的相关程度较高，为方便叙述标注为区域 1，2，3 进行分析。区域 1 主要涉及浏览量、建筑及占地面积、房屋总数以及绿化率指标；区域 2 主要涉及物业费、开盘时间以及交房时间这三项指标；区域 3 的特点更为明显，教育实力与商场、银行等金融聚点的两两线性相关程度极高。因此绝大多数变量间相关性较强且社会意义明显故基于该可视化图进行进一步数据回归插补。

（三） 残缺数据可视化结果

基于 R 给出部分残缺数据可视化结果，具体数据的残缺比例以及大致分布如下图所示：

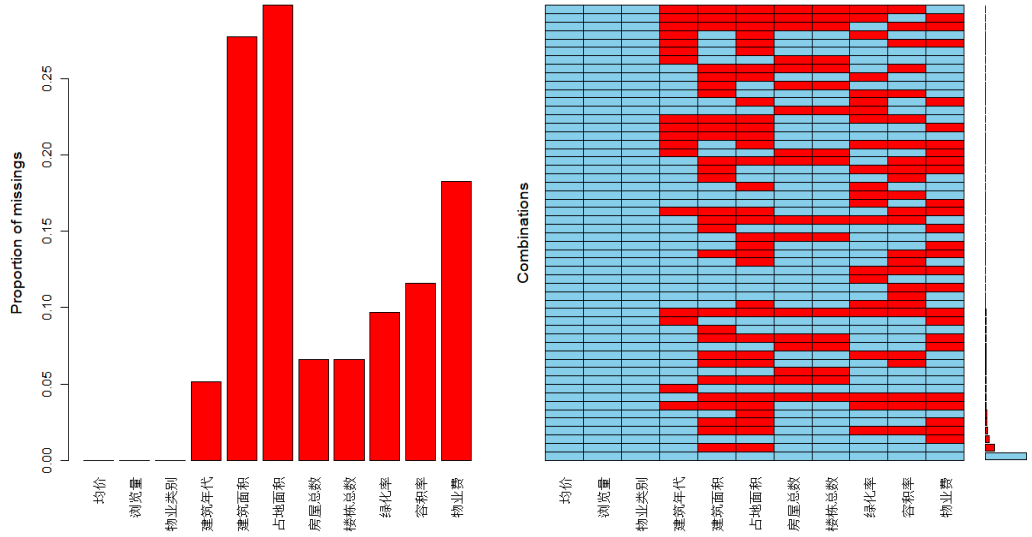


图 2　部分数据残缺比例以及分布图

根据图 2 所示，前 11 维数据中最大残缺指标为占地面积指标，约占据 1/4 的比例，其余大部分指标的残缺比例约在 10%以下并且数据维数较高。

（四） 数据集的多重插补

相对于处理方式较为简单的中心化趋势插补函数，本文采取多重插补降低残缺数据维数后，采用线性回归插补，此处给出多重插补的示意图：

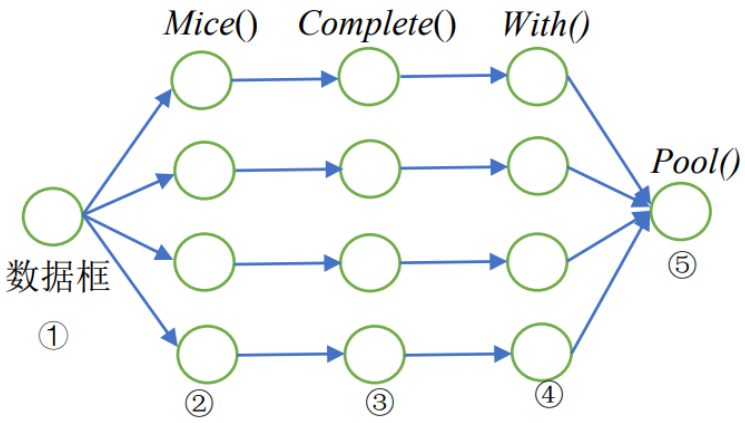


图 3　多重插补示意图

根据不断调试多重插补的参数使得插补后的数据指标在显著性水平为 0.1 的条件下通过检验数目最多进行多次试验，选择其中一个变量模型输出层返回共计 19 个指标，此处选取通过显著性检验个数较多的指标如下：

表 1　显著性检验通过个数较多指标

指标	浏览量	均价	目前房源	占地面积	房屋总数
通过检验个数	8	4	6	4	5

选择浏览量作为拟合对象，根据线性相关图反馈的结果进行模型初步拟合。目前房源、银行、幼儿园、开盘时间、绿化率、楼栋总数、占地面积以及房屋总数共计 8 项指标通过检验。作为预测数据集，考虑到其余残缺数据维度仍较大，同理基于通过显著性检验的指标再次进行多重插补，直到剩余两项指标大学、医院，根据图 1 返回的区域 3 结果可以看到医院、学校以及银行这些变量之间的线性相关程度较高，筛选线性相关程度高于 0.5 的变量进行插补。最终在进行多次多重插补以及线性回归插补的数据集的基础上进行建模分析。

三、西安市整体房价及周期波动情况的横纵向探究

随着西安经济不断发展，城市影响力不断上升，城市排名不断提高，“人才引进”政策的实施推广，也使得了更多人愿意留在西安。但生存问题和长远发展问题，房价的高低，很大程度上影响优秀人才的去留。作为正在汲取优秀人才的“海绵”如何优化自身产业结构，完善环境优化后期配套设施成为锁住“人才资源”的关键所在。

基于西安市房价不断增长的现状以及对未来楼盘市场的分析预测，从供需关系着手进行建模分析，具体流程图如下所示：

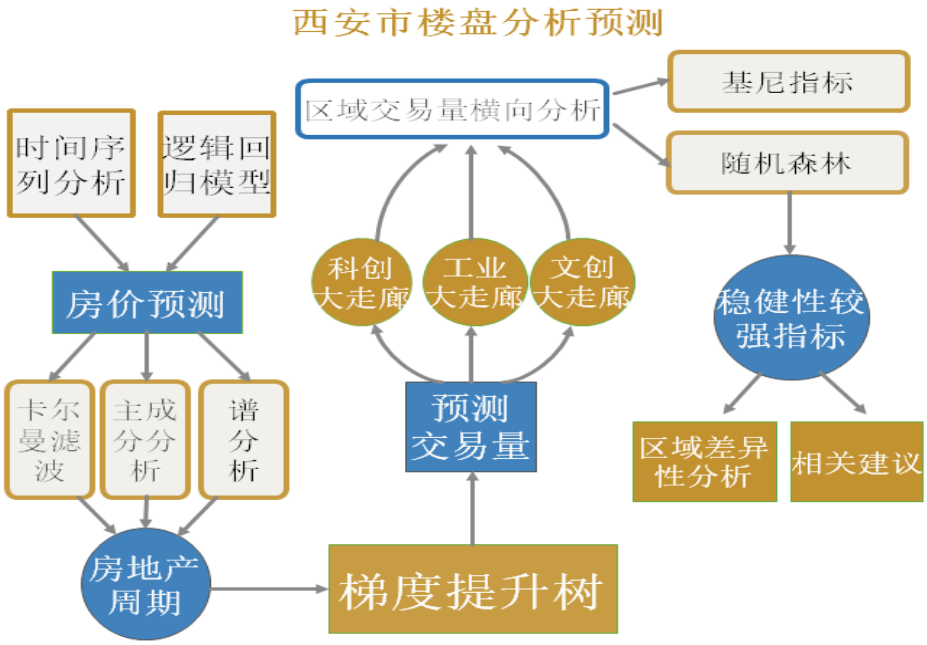


图 4　算法流程图

（一） 西安市房价的纵向分析

基于时间线，在 “人才引进”、“一带一路”以及“四限”政策的调控下，研究房价增长趋势、增长原因以及一系列可能影响房价的因素。

1. 基于一元时间序列以及COX-Staut对西安市未来房价增幅情况的预测

利用时间序列建立 ARIMA(1,2,1)模型，得到七月起未来四个月房价分别为 9776 元、9850 元、9902 元以及 9964 元，由此对西安市房价绘制 4 个月预期图：

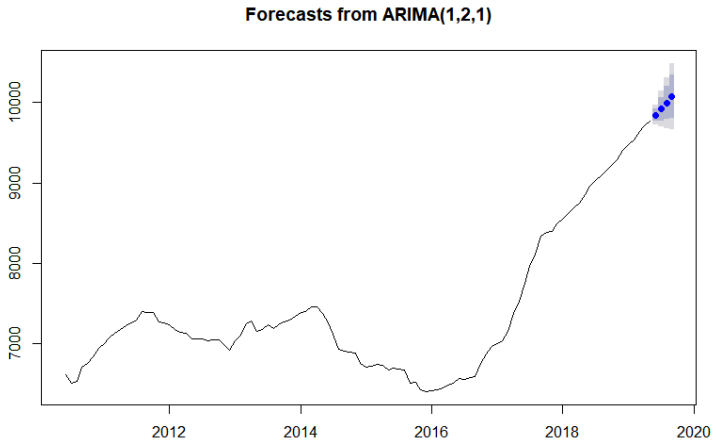


图 5　未来西安房价四个月预测图

相较于 2016-2019 年明显的房价增长趋势，未来 4 个月房价仍然在稳步上升，近年房价在国家调控下即将趋向于平稳状态。根据非参数统计方法 Cox-Staut 趋势存在性检验，在前期房价出现周期上下波动的现象的情况下，探究相关数据无正态分布假定时，西安市房价是否存在增长趋势并给出相关原假设：

H_0：西安市房价序列无趋势

H_1：西安市房价序列有增长趋势

依据 Cox-Staut 检验可得 $p = 0.1102 > 0.05$，无法拒绝原假设。据此可以看出西安市未来房价持续增长的趋势没有统计学意义。结合近期西安市陆续出台的限购以及保障政策，西安市房价的确有趋于缓和发展的态势，彰显政府宏观调控与市场自发调节对减轻房价负担的作用具有较大成效。

2. 基于谱分析对全国与西安市楼市周期情况的对比分析

考虑到房地产市场受多个因素的综合影响，仅用单一指标计算出的周期无法客观反映市场的变化规律。为此借助“主成分分析法”将实测的多个指标合成为独立的但保留了原信息的指标，进而借助“谱分析”方法识别房地产市场周期[4]。

首先利用“主成分分析法”分别计算房地产供给类和需求类的合成指标。将影响房地产供求关系的多个指标的原始数据标准化，以解决各指标的可综合性问题。随后进行因子分析，计算特征根和特征向量表，形成指标的主成分表达式，进而计算得到合成指标。

利用“谱分析法”对合成指标的周期进行判别。首先，检验合成指标时间序列的平稳性以保证谱分析的可信度和有效性；其次，确定截断点频率分量的个数及对应的频率与周期长度；随后，计算合成指标的谱密度值并绘制谱密度曲线；最后，根据曲线中的谱峰值确定合成指标的周期探讨房地产市场的发展趋势[5]。

根据 2001 年至 2018 年国信房地产信息网提供的相关数据，采用“主成分分析法”和“谱分析法”识别西安市房地产周期。根据周期理论和数据的可获得性，建立如下表所示的房地产市场供给和需求类的指标体系。

表 2　房地产市场供给和需求类的指标体系

类别	变量	符号
供给类指标	房地产固定资产投资额	X_1
	房地产开发投资额	X_2
	土地开发面积	X_3
	商品房施工面积	X_4
	商品房竣工面积	X_5
需求类指标	城镇居民人均可支配收入	Y_1
	城镇居民储蓄存款余额	Y_2
	商品房销售额	Y_3
	商品房销售面积	Y_4
	商品房平均售价	Y_5

应用“主成分分析法”确定合成指标，得到体现约 98.7%信息的供给类指标主成分 S_1, S_2 的表达式为：

$$S_1 = 0.341X_1 - 0.042X_2 + 0.339X_3 + 0.335X_4 \quad (1)$$

$$S_2 = 0.079X_1 + 0.999X_2 + 0.047X_3 - 0.003X_4 \quad (2)$$

同理需求类指标的主成分 D_1, D_2 的表达式为：

$$D_1 = 0.253Y_1 + 0.256Y_2 + 0.252Y_3 + 0.255Y_4 \quad (3)$$

$$D_1 = -2.295Y_1 - 0.039Y_2 + 2.790Y_3 - 0.436Y_4 \quad (4)$$

由方差贡献率可以看出，房地产固定投资额占比较大相应的城镇居民储蓄余额影响购房欲望。

经过一次差分 S 与二次差分 D 分别得到 S' 与 D'' 后，序列始终在 0 附近随机波动且没有明显趋势或周期，基本可以视为平稳序列。按照谱密度计算步骤与公式，对 S' 与 D'' 进行计算供给类和需求类的房地产市场周期。

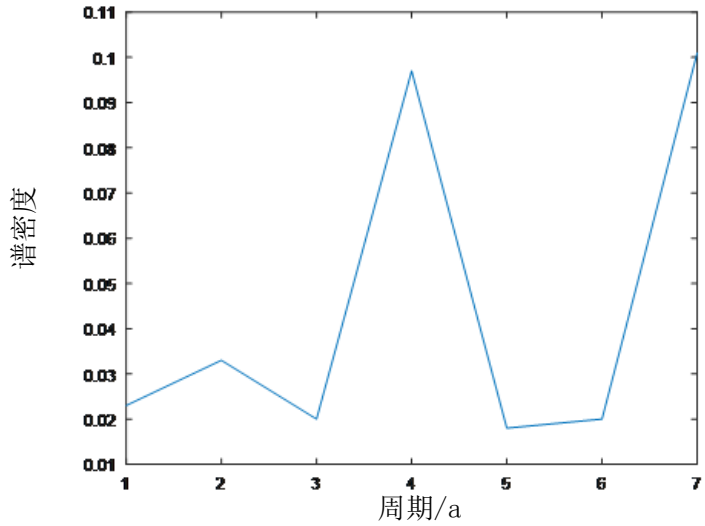


图 6　供给类的房地产市场周期

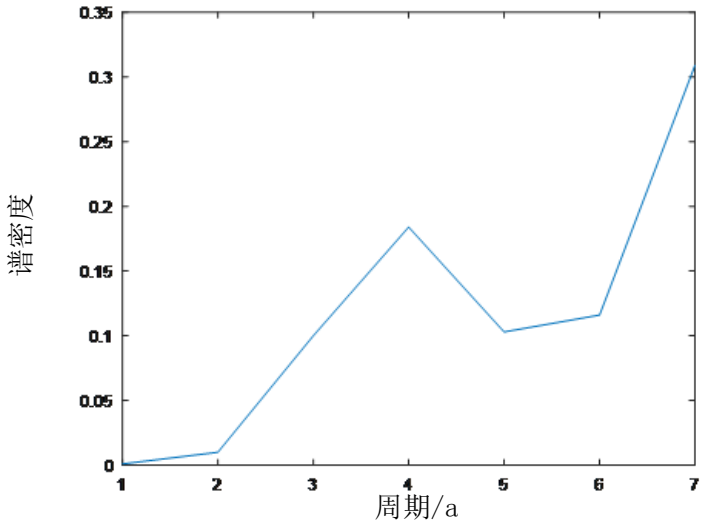


图 7　需求类的房地产市场周期

由图 6 和图 7 可知，2004 至 2017 年，全国房地产市场的供给类合成指标的谱密度曲线存在主谱峰和次谱峰；而需求类合成指标的谱密度曲线存在主谱峰。因此，相应的全国房地产市场的供给存在 3.5a 的主周期和 7a 的次周期。

具体而言，2004 至 2017 年全国房地产市场的供给和需求均存在明显的周期特征，其中供给与需求的主周期时间长度一致，表明全国房地产市场基本保持了供求关系的均衡发展。同时，供给还存在一个较明显的为期 7a 的次周期，而在同样的时间长度内，需求发展平稳且不存在周期性特征。长期看来，市场将处于供求关系不匹配且需求旺盛的阶段，即房地产市场价格存在长期上涨的趋势。

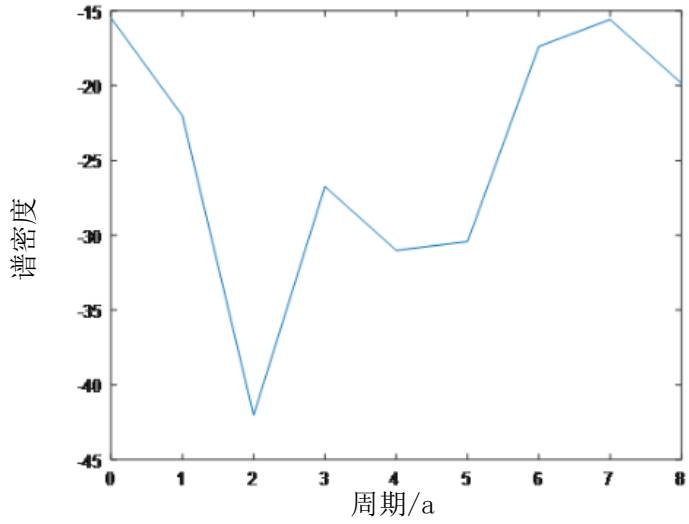


图 8　供给类的房地产市场周期

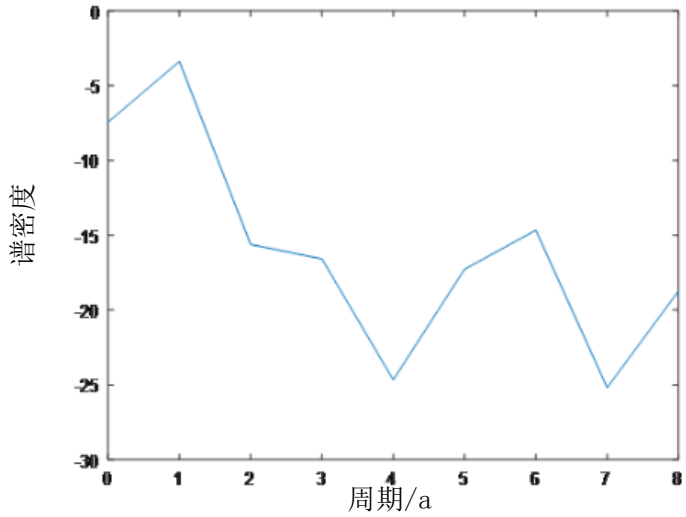


图 9　需求类的房地产市场周期

2002 至 2018 年 ，西安市房地产市场的供给类合成指标的谱密度曲线存在主谱峰(T=7)和次谱峰(T=3)；需求类合成指标的谱密度曲线存在主谱峰(T=1) 和次谱峰(T=6)。因此，相应的西安市房地产市场的供给存在 7a 的主周期和 3a 的次周期，需求存在 6a 的主周期和 3a 的次周期。

2002 至 2018 年西安市房地产市场的供给和需求均存在明显的周期特征，供给与需求的周期时间长度大致相等，表明西安房地产市场基本保持了供求关系的均衡发展，政策与需求同步稳进，房地产需求类指标更高。长期看来市场将处于供求关系较匹配但需求更高的阶段，即房地产市场价格仍存在长期上涨的趋势。

综上，西安市房价基本紧随全国房地产周期，但房价上涨趋势略高于全国。

3. 多变量间相似程度的量化分析

基于西安市房价近年来的持续增长的现状，从缴存单位户数、缴存职工户数、本月归集额、本月发放贷款额四个方面对西安市房价进行核心影响因素提取。

①灰色关联度法

灰色关联度的基本思想是根据各比较数列集构成的曲线簇与参考数列构成的曲线之间的几何相似程度来确定比较数列集与参考数列之间的关联度，比较数列构成的曲线与参考数列构成的曲线的几何形状越相似其关联度越大。

$$\xi_i(k)=\frac{\min\limits_{s}\min\limits_{t}\left|x_0(t)-x_s(t)\right|+\rho\max\limits_{s}\max\limits_{t}\left|x_0(t)-x_s(t)\right|}{\left|x_0(k)-x_s(k)\right|+\rho\max\limits_{s}\max\limits_{t}\left|x_0(t)-x_s(t)\right|} \qquad (5)$$

其中，参考序列为 x_0，比较序列为 x_i，k 为指标，$\rho \in [0,1]$ 为分辨系数，我们选选取 $\rho = 0.5$。为更直观说明，现绘制原始序列归一化后的序列曲线：

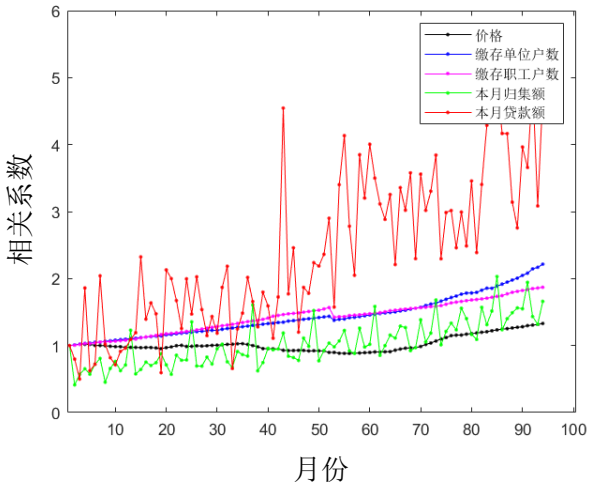


图 10　原始序列归一化曲线

进而基于序列间的关联度绘制图像有：

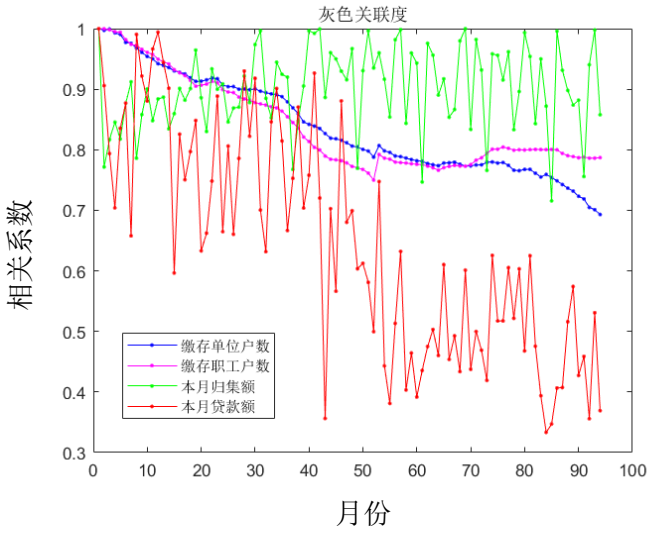


图 11　序列间关联度

根据图 11 可以看到本月归集额与房价的灰色关联度整体较大并且较为稳定，为量化不同影响因素的影响强度，汇总结果如下所示：

表 3　不同影响因素关联度

影响因素	缴存单位户数	缴存职工户数	本月归集额	本月贷款额
关联度	0.8386	0.8406	0.9018	0.6439

综上由图 10、图 11 和表 3 可以看出，运用灰色关联度法可知本月归集额对西安市房价影响程度最大，本月贷款额的影响程度最小。

②Spearman秩相关系数法

Spearman 秩相关系数是一个非参数性质的秩统计参数，由 Spearman 在 1904 年提出，用来度量两个变量之间联系的强弱。

$$r_s=\frac{\sum_{i=1}^{n}(R_i-\bar{R})(Q_i-\bar{Q})}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R})^2}\sqrt{\sum_{i=1}^{n}(Q_i-\bar{Q})^2}}=\frac{12\sum_{i=1}^{n}R_iQ_i-3n(n+1)^2}{n(n^2-1)} \qquad (6)$$

表 4　不同影响因素关联度

影响因素	缴存单位户数	缴存职工户数	本月归集额	本月贷款额
关联度	0.3025511	0.3135966	0.2771478	0.1340332

由表 4 可以看出，运用 Spearman 秩相关系数法可知对西安市房价缴存职工户数影响程度最大，本月贷款额的影响程度最小。

综上，贷款额与房价的关联度较小进一步说明了出台“四限”（限购、限售、限贷、限价）政策和公积金新政的发布，并没有影响西安房价的持续性增长，也没有影响购房热情。根据各楼盘反馈的信息，在按揭贷款的购房者中，90-95%是商贷，公积金贷款所占的比例在 5-10%之间。由此可以看出，在贷款买房的购房者中，使用公积金贷款所占的比例不到 1/10。由此首付的提高、贷款额度的倍数下降，只能影响到不足 10%的购房者，对于 90%以上的购房者无影响，从而对房价产生的影响更小。究其原因，较多单位没有缴纳住房公积金，且西安一半以上的楼盘不支持公积金贷款。

因此选取与房价关联度较高的缴存单位户数、缴存职工户数、本月归集额三个因素进行进一步多元时序分析研究。

4. 相关指标的卡尔曼滤波分析

卡尔曼滤波是根据上一状态的估计值和当前状态的观测值推出当前状态的估计值滤波方法，这里的滤波其实是指通过一种算法排除可能的随机干扰以提高检测精度的方法或手段。由于卡尔曼滤波是用状态方程和递推方法进行估计的，因而卡尔曼滤波对信号的平稳性和时不变性不做要求。这里不加证明地直接给出经典的离散系统卡尔曼滤波公式[6]，包括 5 个方程如下所示:

状态的一步预测方程：

$$\hat{x}_k = A\hat{x}_{k-1} \tag{7}$$

均方误差一步预测方程为：

$$P_k = AP_{k-1}A' + Q \tag{8}$$

滤波增益方程(权重)：

$$H_k = P_k^- H'(HP_k^- H' + R)^{-1} \tag{9}$$

滤波估计方程(k 时刻的最优值)：

$$\hat{x}_k = \hat{x}_k^- + H_k(z_k - H\hat{x}_k) \tag{10}$$

均方误差更新矩阵(k 时刻的最优均方误差)：

$$P_k = (I - K_k H)P_k^- \tag{11}$$

式中，x 表示状态向量，A 为状态转移矩阵， P 为误差协方差矩阵，Q 为系统噪声协方差矩阵， H为协方差矩阵，R为观测噪声协方差矩阵，K为卡尔曼增益矩阵，z 为观测向量。

通过关联度分析，选取房价、缴存单位户数、缴存职工户数、本月归集额四个指标，基于多元时间序列的卡尔曼滤波算法，通过每期拟合情况分析预测效果。

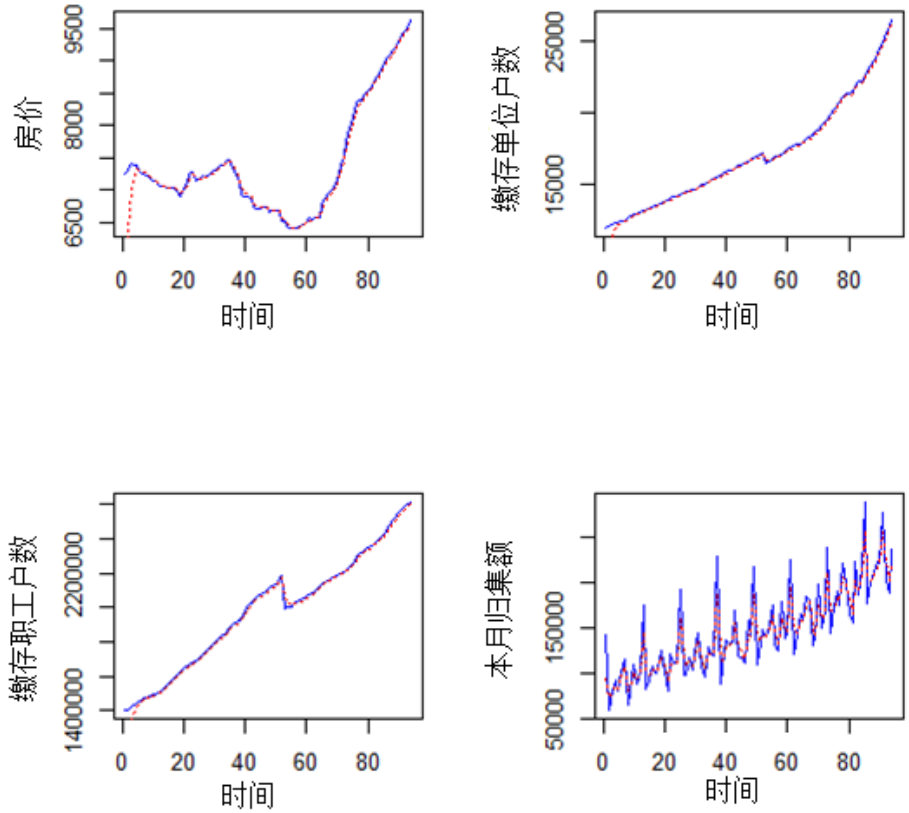


图 12　各指标时序图

由图 12 所示，四个图分别对应房价、缴存单位户数、缴存职工户数、本月归集额四个指标，可看到随着迭代次数的进行拟合效果越来越好，因此可以考虑做进一步分析。

随着单位和职工个人意识的提升，更多的职工群众纳入到公积金制度保障体系以减轻职工购房负担，也让更多的公积金缴存者享受到政策的优惠。归集公积金是住房公积金制度发展战略法的重要内容，公积金归集额的持续波动性增加，进一步扩大了住房公积金制度的覆盖面。卡尔曼滤波算法拟合图也侧面反映了西安公积金政策刺激房地产市场的积极意义，宏观政策调控的影响不可小觑。

（二） 西安市房价情况的横向分析

利用爬虫得到的横截面数据，探究影响西安市房地产价格的相关因素。

1. 基于示性函数的变量处理

考虑到相关指标中含有分类变量，因此根据二元逻辑回归对所有完整的数据集共 823 个指标进行分析，旨在探究在何种因素下会对房价的起伏有较大影响。

根据 5 月份西安市平均房价 9773 对横截面数据中的均价进行如下数据处理：

$$p_i=\begin{cases}1, p_i>9773\\0, p_i<9773\end{cases}$$

由于具体考虑房价的高低没有过于现实的研究意义，进而基于房价可能高于全市平均水平的概率进行进一步的建模分析。

2. 基于二元逻辑回归的西安市房价情况分析

借助示性函数在显著水平为 0.05 的条件下可以看到房屋总数、楼栋总数、目前房源、银行、开发商以及卫生服务变量通过检验。由于部分变量的回归系数十分靠近于 0，因此对上述回归模型进行简化有：

$$\hat{p}=0.19x_1+0.62x_2+1.1x_3-0.5x_4$$

其中，$\hat{p}$ 表示该楼盘可能超出平均房价的概率高低，x_1 表示银行个数，x_2 表示开发商类别为分类变量，具体含义如下：

$$x_2=\begin{cases}0,\text{开发商不确定}\\1,\text{其他}\\2,\text{陕西除西安市外}\\3,\text{西安市}\\4,\text{沿海}\end{cases}$$

同理，对于分类变量 x_3 以及 x_4 分别表示卫生服务以及环线位置情况，具体对应有：

$$x_3=\begin{cases}0,\text{卫生服务不确定}\\1,\text{其他}\\2,\text{包含在物业费内}\\3,\text{不包含在物业费内}\end{cases}$$

$$x_4=\begin{cases}1,\text{一环附近}\\2,\text{三环附近}\end{cases}$$

因此可以看到卫生服务费用是否包含会显著影响房价的价格，此外越是沿海投资商开发的房产相对均价更高。银行分布越密集，该地段的楼盘价格对应高出全市平均水平的可能性越大。

但仅在基于房价条件下，建筑年代变量并未通过检验，因此西安市的平均房价并未受到“抢人大战”策略的强烈冲击，紧随国家政策调控步伐。由于地域性差异十分明显，因此选择展开地域性楼市交易量分析具有显著意义。

四、全市与区域楼房交易量及影响因素的预测探究

（一） 西安市楼盘交易量情况的全局分析

1. 基于GBM梯度提升树的样本集训练及交易量态势预测

首先以西安市楼盘市场为整体，以影响西安市楼盘成交量为探究源头，对西安市整体楼市的变量重要性指标进行聚类分析。

对于开发商以及购房者而言，确切的预测交易量的连续数值意义不大，且由于阶段数据限制可能无法考虑到均衡开盘时间较晚的小区，因此通过 K-means 聚类转化为分类变量即高、中、低三个大类指标记为分类变量 y：

$$y=\begin{cases}0,\text{尚未开盘或者正在建设}\\1,\ \text{低交易量}\\2,\ \text{中等交易量}\\3,\ \text{较高交易量}\end{cases}$$

若将 0 作为训练数据则与实际情况不符，因为 0 代表该楼盘正在完善或者只是尚未对外销售，并非由于绝对原因导致数值为 0，因此训练样本时不进行考虑。进而基于处理后的数据集利用梯度提升树算法，将有交易量数据的记录共计 823 条构成的集合作为测试集，随后通过计算混淆矩阵检验模型预测精度。

①BOOSTED TREE

提升树是以分类树和回归树为基学习器的提升算法，将 $T(x;\Theta_m)$ 学习器定义为决策树，被认为是统计学性能最好的方法之一。

提升树模型如下：

$$f(x)=\sum_{m=1}^{M}\alpha_m T(x;\Theta_m) \qquad (12)$$

二分类问题 $T(x;\Theta_m)$ 为回归树，回归提升树使用的前向分布计算法：

$$f_0(x)=0 \tag{13}$$

$$f_m(x)=f_{m-1}(x)+T(x;\Theta_m) \tag{14}$$

$$f_M(x)=\sum_{m=1}^{M}T(x;\Theta_m) \tag{15}$$

前向计算到第 m 步时，需要求解目标函数：

$$\Theta_m^{'}=\arg\min_{\Theta}\sum_{m=1}^{N}L(y_i,f_{m-1}(x)+T(x_{-i};\Theta_m)) \tag{16}$$

$\Theta_m^{'}$ 为根据损失函数更新的第 m 棵树的参数。假设这里的损失函数我们定义为平方误差函数：

$$L=(y-f(x))^2 \tag{17}$$

损失函数为:

$$L=[y-f_{m-1}(x_i)-T(x_{-i};\Theta_m)]^2 \tag{18}$$

以上是基于平方误差函数的提升树模型。

②GBM改进梯度提升树

GBM（提升器）算法是基于梯度算法的改进提升树模型，它与之前提到的提升树关键性的差异在于残差更新的方式，随后给出具体模型：

$$g_m(x_i)=[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)} \tag{19}$$

$$\Theta_m^{'}=\arg\min_{\Theta,\beta}\sum_{i=1}^{N}[-g_m(x_i)-\beta_m\Theta(x_i)]^2 \tag{20}$$

进而最佳步长为：

$$\beta_m=\arg\min_{\beta}\sum_{i=1}^{N}L[y_i,f_{m-1}(x_i)+\beta_m\Theta_m^{'}(x_i)] \tag{21}$$

令 $f_m=\beta_m\Theta_m^{'}$ 更新则 GBM 模型为：

$$f(x_i) = f_{m-1} + f_m \tag{22}$$

在有监督机器学习中，为了确定该使得损失函数最小的模型，梯度下降算法的目标则是在每一轮迭代中求得当前模型的损失数的负梯度方向，乘以一定的步长学习速率，加到当前模型中形成此轮迭代产生的新模型，从而达到每一轮迭代后的模型，使得损失函数逐次减小的目的。

因此对于 Gradient Boost Machine 来说重要的变量有迭代次数 M、损失函数的形式 $\psi(y, f)$和基础学习器的形式 $h(x, \theta)$。

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi(y_i, f_{t-1}(x_i) + \rho h(x_i, \theta_t)) \tag{23}$$

调用 R 语言中的 gbm 包，在测试集数据上运行可以得到使得损失函数最小的模型迭代次数为 391 次，同时得到对应的误差平方随迭代次数的变化曲线图：

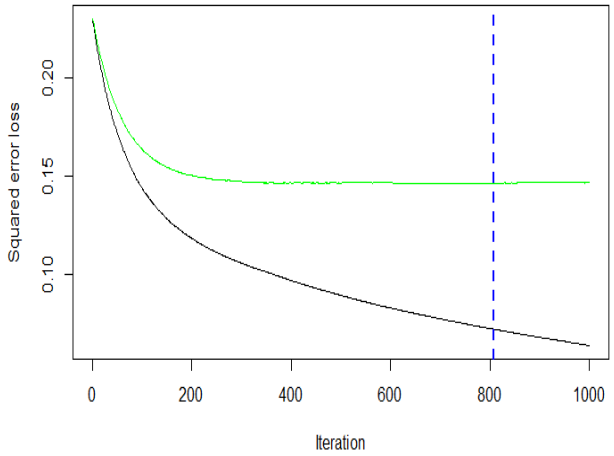


图 13　误差平方随迭代次数变化曲线

根据上图可以看到开始时随着迭代次数的增加，残差平方显著减少，但当迭代次数达到 800 次左右后，残差方差的减缓速度变缓，为了避免过拟合的情况确定最佳的迭代次数为 808 次。

随后基于建立好的梯度提升树对测试集进行预测，并将预测返回的结果与真实情况进行比较，计算其混淆矩阵进一步确定模型的预测精度：

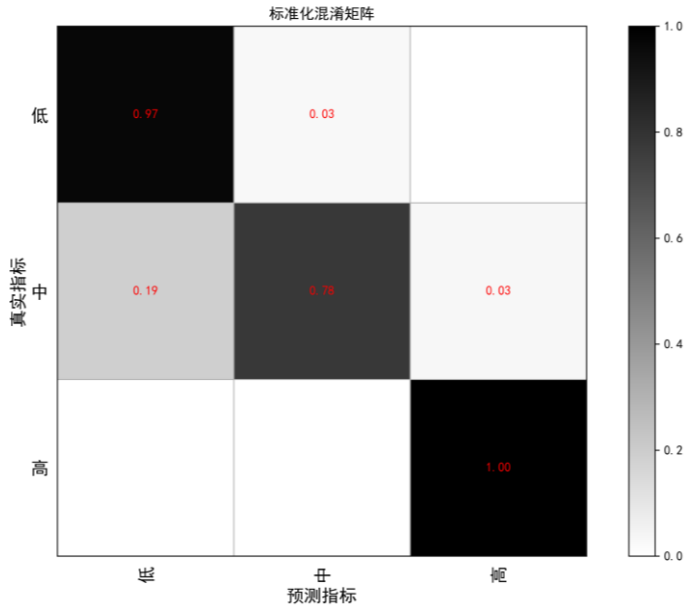


图 14　混淆矩阵

根据返回得到的混淆矩阵可以看到，各类交易类型的预测精度均在 0.8 左右，因此可以有理由认为该模型的预测精度较高，进而在预测集数据上进行预测，给出相关小区可能的交易行情，并且对上述所有变量进行边际效应分析。

2. 影响西安楼市交易量的边际变量分析

根据模型反馈得到针对西安市整体楼市的变量重要性指标：

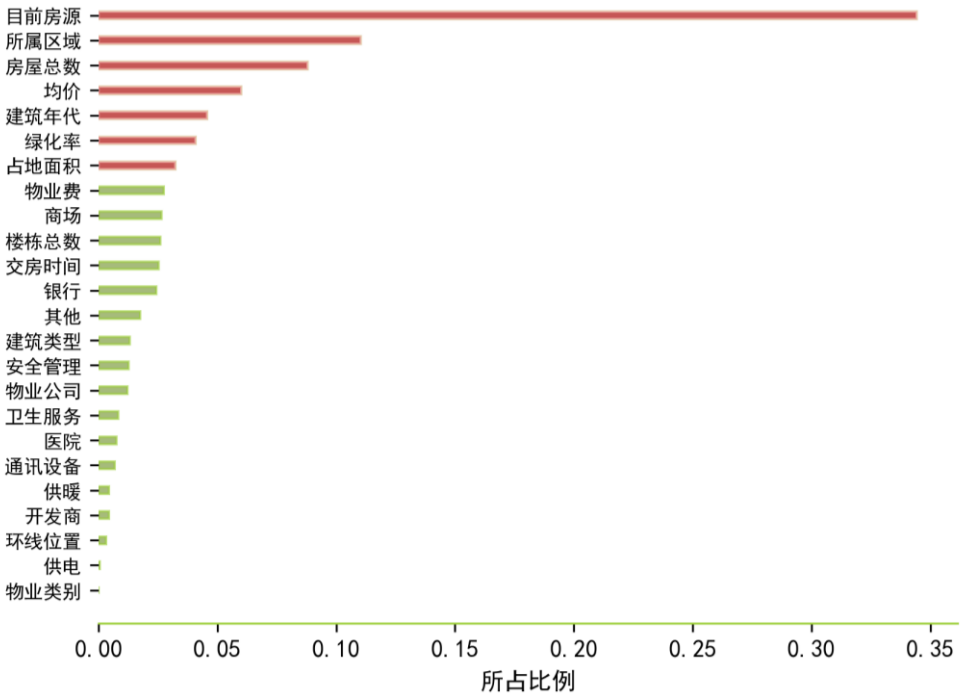


图 15　相关变量重要性图

根据上述返回得到的相关变量的重要性图可以较为直观的看到目前房源、房屋总数、浏览量、建筑年代、均价以及物业费是影响交易量较为显著的指标。值得注意的是建筑年代可以反映楼盘是否在进入“抢人大战”阶段，此外均价并没有十分显著的影响交易量。同时浏览量的作用更为明显，体现了开发商宣传的力度以及小区的知名度造成的品牌效益对顾客消费心理需求造成的较大影响，因此广告效应的作用不可小觑。同时通过对各楼盘的交易量 y 进行边际效应分析，分类汇总得到如下结果：

①国家调控政策因素

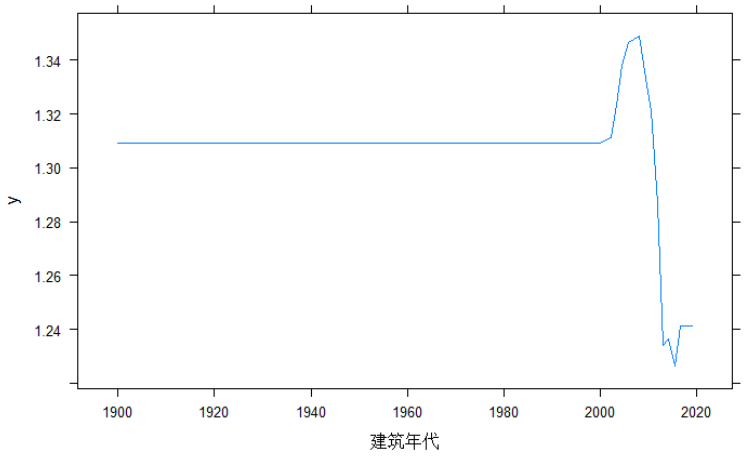
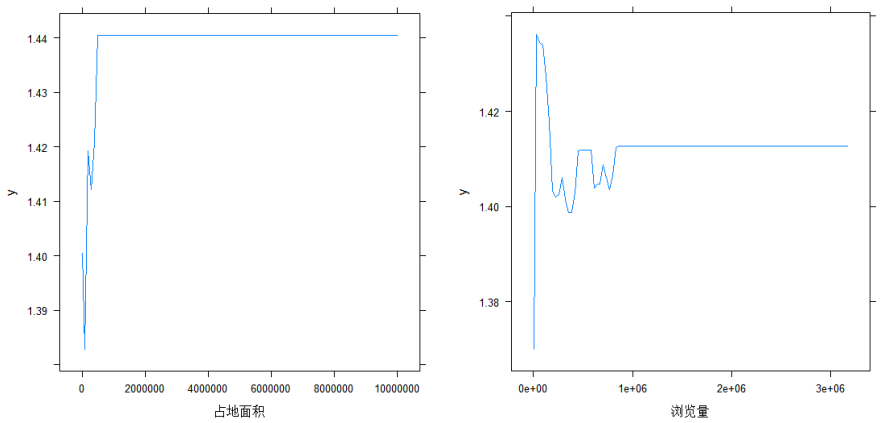
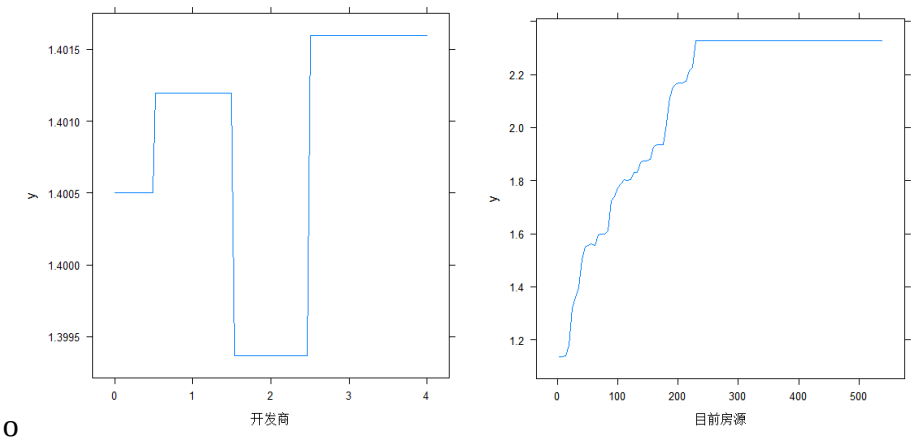


图 16　国家调控影响因素边际效应图

依据国家调控影响因素，2011 年 1 月“新国八条”出台之后，限购全面铺开，西安随大流也开始限购；在 2016 年 12 月底限购之后，西安已经走过整两年限购，销售量环比增速已经连续走低了 15 个月；2016 年之后政策进一步宽松，“抢人大战”愈演愈烈，房价开始抬头，随后房价持续性增长 34 个月。西安紧随国家政策调控，房地产交易量随房价变化及市场需求量变化明显。

②宏观影响因素探究





o

图 17　宏观影响因素边际效应图

据占地面积边际效应分析可知，购房者对于小区面积的敏感程度基本呈现出线性增长态势，随着楼盘占地面积的增加，成交量对应也较高。但是可以看到随着占地面积达到 2000000 平方米左右时，成交量基本呈现稳定状态。考虑到西安市占地面积较大的楼盘一般为别墅等高档小区，因此交易量对占地面积不再敏感的状态也符合实际情况。

根据浏览量边际效应可以较为直观的看到，初期交易情况随着浏览量的增加呈现非常快速的增长，楼盘前期的广告宣传投入力度很大程度上会影响房源的销售情况。但是随着较好的房源被购进审美疲软，虽然浏览量增加，但是大家的购房热情不再高涨呈现出持平的状态。

开发商边际效用图可以看到，购房者对于开发商的选择呈现出两端式的区别。一方面，西安房地产开发商市场近乎被沿海和西安本地开发商瓜分，除西安外的陕西其他城市开发商对西安楼盘的涉及程度并不深。另一方面，体现了购房者更大程度认可沿海和西安本地的房地产开发商。

此外，西安房源市场呈现持续增长后进入稳步阶段，前期西安的不断发展致使更多房地产开发商看好西安楼市行情，大量投资房地产因而房源得到不断扩充，成交量呈现持续性增长。后期过多开发商涌入房地产市场而常居人口数有限，限购政策的颁布为给更多人购房可能性的同时，也造成房价的不断提高，致使想要购房者买不起房，所以交易量呈现稳态。

③居住友好度影响因素探究

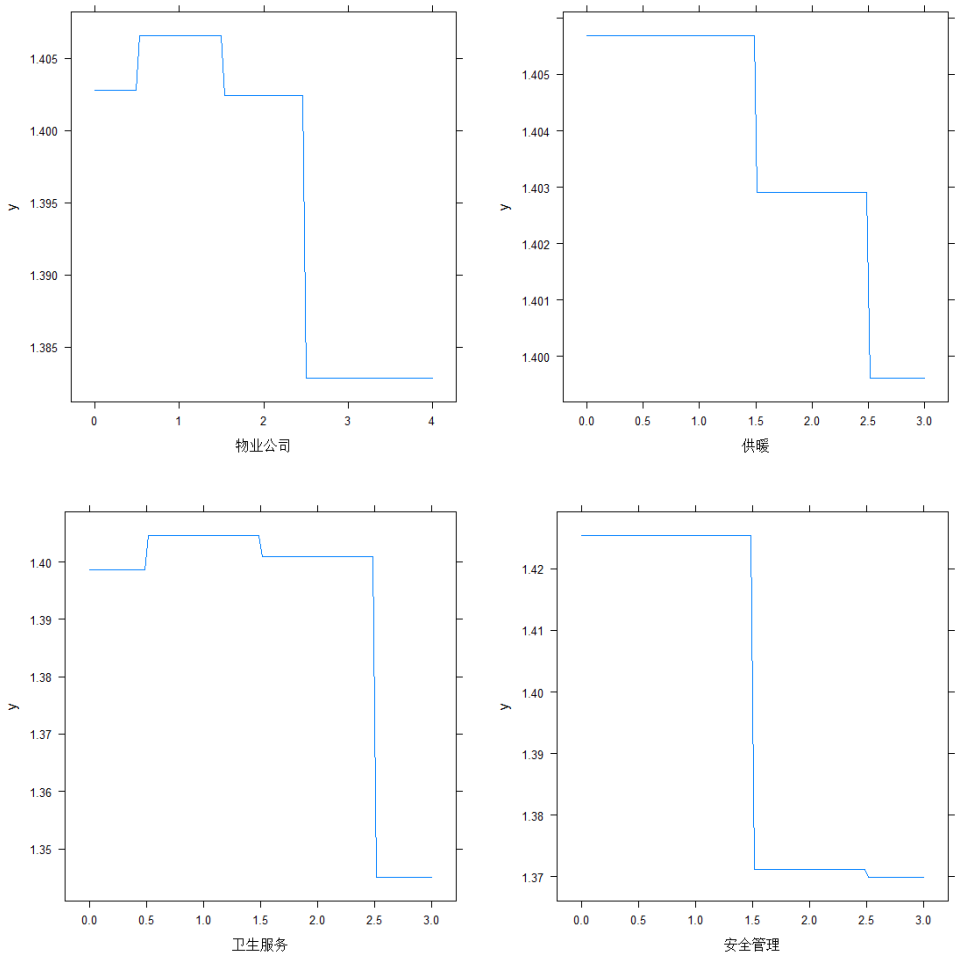


图 18　居住友好度边际效应图

通过与图 16 开发商边际效用图对比，图 17 物业公司图中可以看到，随着“一带一路”的引进和发展，更多外地资金流入西安本地市场，但整体而言沿海物业公司不太受西安本地购房者的欢迎，陕西本土公司更受购房者欢迎。

据供暖边际效用图显示，购房者对供暖选择的多样化发展，选择集中供暖的购房者数目减少。究其原因，空调等智能供暖设施的出现挤占供暖市场，同时也避免集中供暖时住户家中无人时仍在供暖的经济浪费和能源浪费问题的产生。

据卫生服务边际效用图可以知道，卫生服务也极大程度地影响着购房者的购房选择，说明了购房者对于住宅卫生环境也是较为看重的。卫生服务费包含在物业费中会节省住户开支，消费者倾向于更加经济明了的住房管理模式。

据安全管理边际效用图可以知道，随着各小区整体安全设施的不断完善，各楼盘没有显著差异性，住房安全隐患影响作用表现不显著。

④各类型学区房影响因素探究

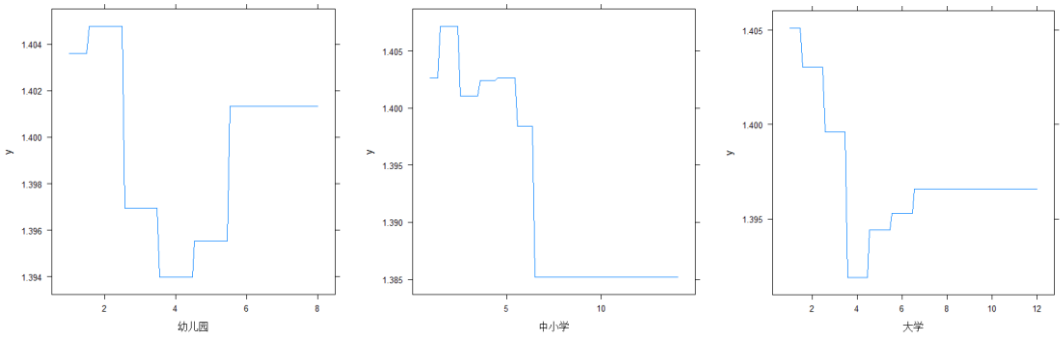


图 19　各类型学区房边际效应图

重点学校入学竞争日趋激烈，望子成龙的家长在选择购房时首先将目光投向了购买学区房源，觉得这是进入重点学校的便捷之路。房地产商们也纷纷用“学区房”这一特有名词提高楼盘的含金量。对比图 18，幼儿园的家长对于学区房的选择更多源于对孩子成长起于“赢在起跑线”的考虑。购房者对中小学数目的容忍程度最大，可以看到中小学生的家长会更大程度上愿意选择购买学区房。考虑到西安市大学分布较多的区域多位于郊区，因此一定程度上会抑制购房热情。

⑤基础设施影响因素探究

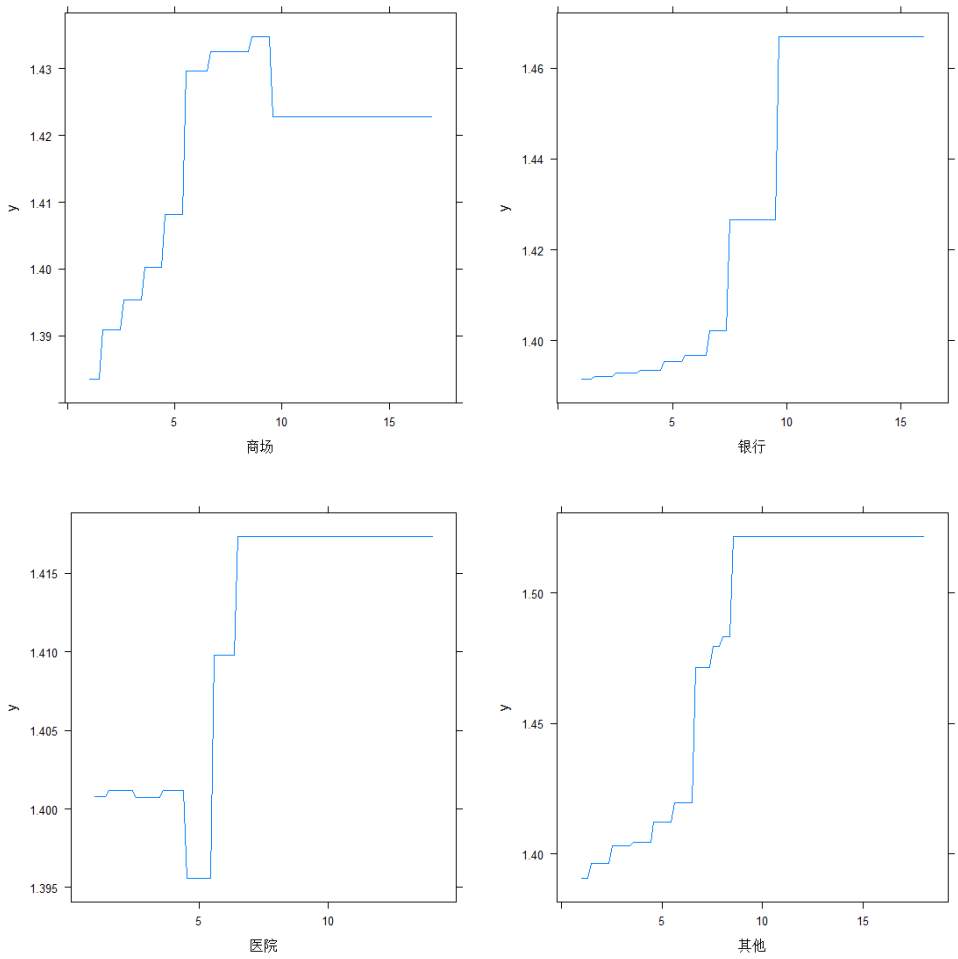


图 20　基础设施边际效应图

随着西安经济的快速发展，据图 19 显示，周围商场、银行、医院、其他娱乐设施的发展趋势基本相同，随着基础设施数目的增多呈现持续性增长至平稳状态。便捷的生活是住户选择的首要标准，因此对于房地产开发商来说，周围基础设施完善也是进行投资的首要选择。同时，周围房地产的开发布局也会影响商业娱乐设施的构建与开发，进而极力促成互利互惠、共同盈利的双赢局面。

⑥房地产位置与价格影响因素探究

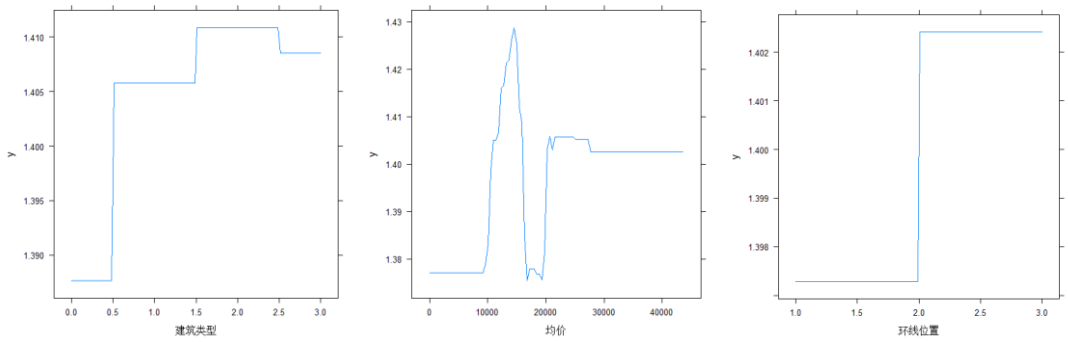


图 21　房地产位置与价格边际效应图

据建筑类型边际效用图显示，板楼、板塔结合、塔楼都是房地产开发商会选择的类型。相较而言，购房者更倾向于选择经济实惠又稳固的板塔结合的建筑方式，可以同时拥有板楼和塔楼的优势，达到经济实惠的效果。

据均价边际效用图可以知道，选择房价在 10000-18000 之间的购买者较多，20000 元左右的房产遇冷。销售状况较为低迷的楼盘多为洋房、别墅或商业中心区的房产。具体表现为曲江新区、高新区等经济发展较快区域的房价较高，且在持续性增长。同时这些地区也是上班族聚集地，更容易成为购房者的首位备选项，随着人才引进政策的发展，也推动了这些区域房价的不断增长。

据环线位置边际效用图可以知道，环线位置也会影响购房者选择。二环以内接近商业中心区，房价较高，不符合普通上班族的经济实力，因此购房者较小概率选择于此购房。二环以外，房价相对较低生活节奏相对较慢，符合当代年轻人的生活意愿，更多人选在在二环外购房。

3. 全市楼市交易量情况的预测

最后基于改进的提升树对交易量空缺的楼盘交易量进行预测，并汇总整体分析西安市楼盘交易情况：

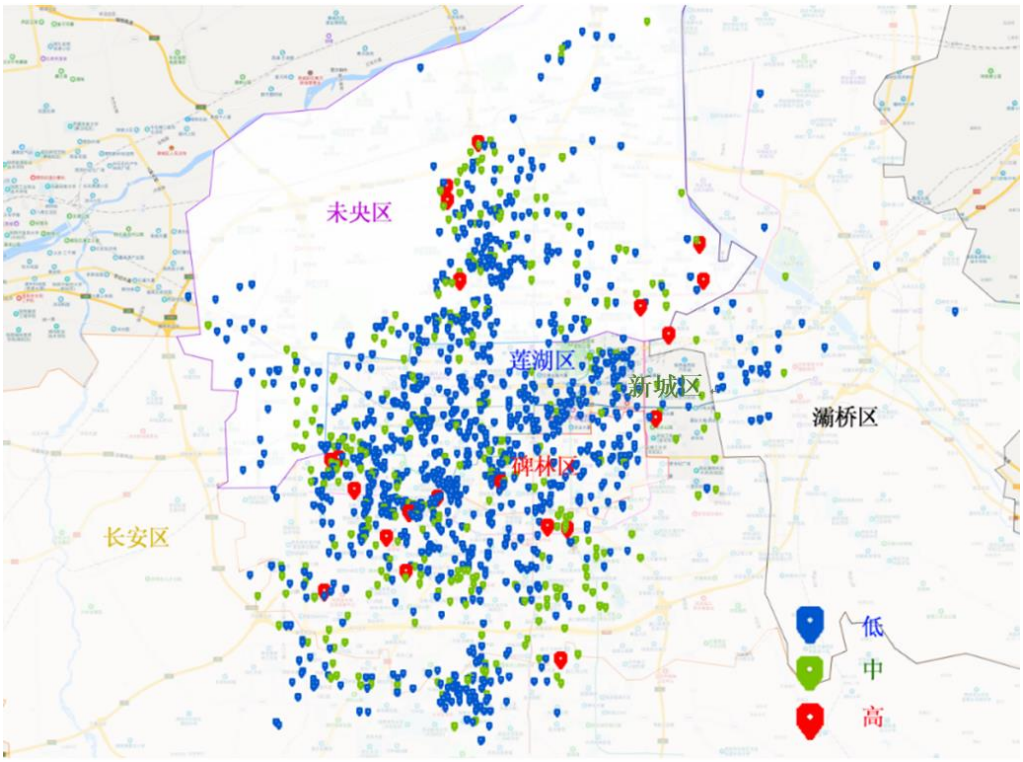


图 22　西安市楼盘交易情况图

根据预测图不难看出，交易量高的楼盘基本分布在高新区、未央区以及浐灞区附近，基本呈现出外散趋势，主城区楼盘交易情况呈现疲软态势。并且西安市较多楼市呈现出中等的温热状态，高交易量楼盘十分稀少，一方面由于政府调控“购房热”成效显著，另一方面从消费者对楼市保持观望的态度亦可以解释上述现象。因此可以看到西安市整体房源的供给较为充足，做好迎接更多优秀才子落户西安的实质准备。那么如何根据不同经济区域进行合理空间规划便成为进一步探究的重点，故针对西安市各重点区域进行针对性分析。

（二） 西安市重点区域楼盘交易量情况的局部分析

1. 基于Gini系数的特征变量筛选

由于交易量有高中低之分，而 Gini 系数可以较好衡量不平等性，由游皓麟著作的《R 语言预测实战》中可查阅得知基尼系数相关计算公式。在分类问题中，分类树节点 A 的 Gini 系数表示样本在子集中被错分的可能性大小，它通常记作这个样本被选中的概率 p_i 乘以它被错分的概率 $(1-p_i)$。假如响应变量 y 的取值有个分类，令 p_i 是样本属于 i 类别的概率，则 Gini 系数可以通过如下公式计算：

$$Gini(A)=\sum_{i=1}^{k}p_i(1-p_i)=1-\sum_{i=1}^{k}p_i^2 \quad (24)$$

对于连续型变量，可将数值排序，依次计算相邻值之间的平均值作为分支点，在产生的两类中（其中一类记为样本集合S，C_i是S中属于第i类的子集），计算S对响应变 量y的Gini 系数，公式如下：

$$Gini(S)=1-\sum_{i=1}^{k}\left(\frac{|C_i|}{|S|}\right)^2 \quad (25)$$

对于离散变量，可直接使用以上公式，计算样本集合S对应的 Gini 系数，其中S表示分类树的一个节点对应的子集合。当S中只有两类时，是经典的而二分类问题，此时的$Gini(p)=2p(1-p)$，其中 p 是样本点属于第一类的概率。

由于 Gini 系数可以表示样本在子集中被错分的可能性大小，其值越大，样本越有可能被错分，其值越小，样本越有可能不被错分，因此 Gini 系数越小越好。可以通过计算每个特征的 Gini 系数来进行特征选择。在分类树的前提下，每个特征F，都会有N各分类或区间，作为分类节点，通过以下公式计算该特征F的Gini 系数：

$$Gini(F)=\sum_{i=1}^{N}\frac{|F_i|}{|F|}Gini(F_i) \quad (26)$$

采用Gini 系数进行不平等性指标的分析，得到如下的返回结果：

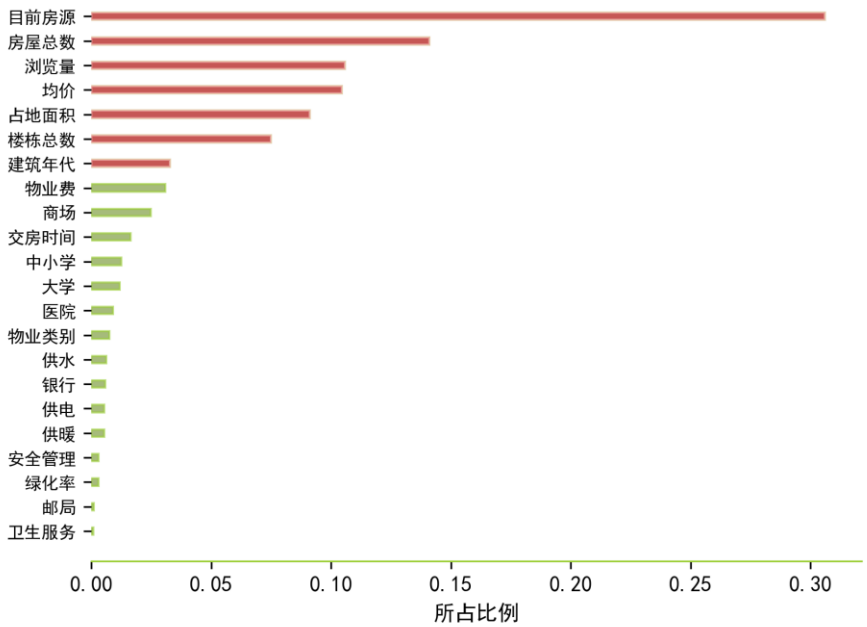


图 23　指标重要性图

返回得到的结果与图 15 进行对比可以较为明显的看到指标重要性较为稳健的是目前房源、房屋总数以及浏览量等指标影响仍较大，但是基尼指标对于绿化率的重要性程度排名差异较大，不排除具有一定随机性。

2. 基于随机森林的特征变量筛选

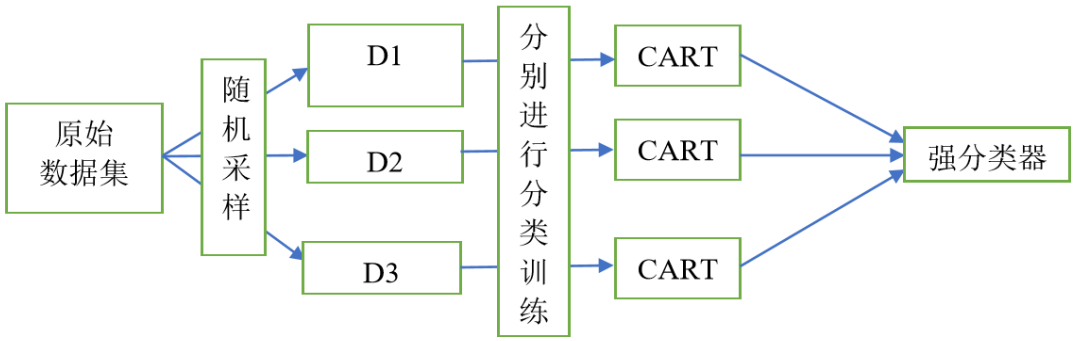


图 24　随机森林算法流程

随机森林(Random Forest，RF）是 Bagging 算法的一种，其实在介绍完 Bagging 算法之后，随机森林几乎是呼之欲出的，RF 相对于 Bagging 只是对其中一些细节做了自己的规定和设计。

首先，RF 使用了 CART 决策树作为弱学习器。我们只是将使用 CART 决策树作为弱学习器的 Bagging 方法称为随机森林。因此基于随机森林较好的泛化能力以及抗过拟合的能力返回结果如下所示：

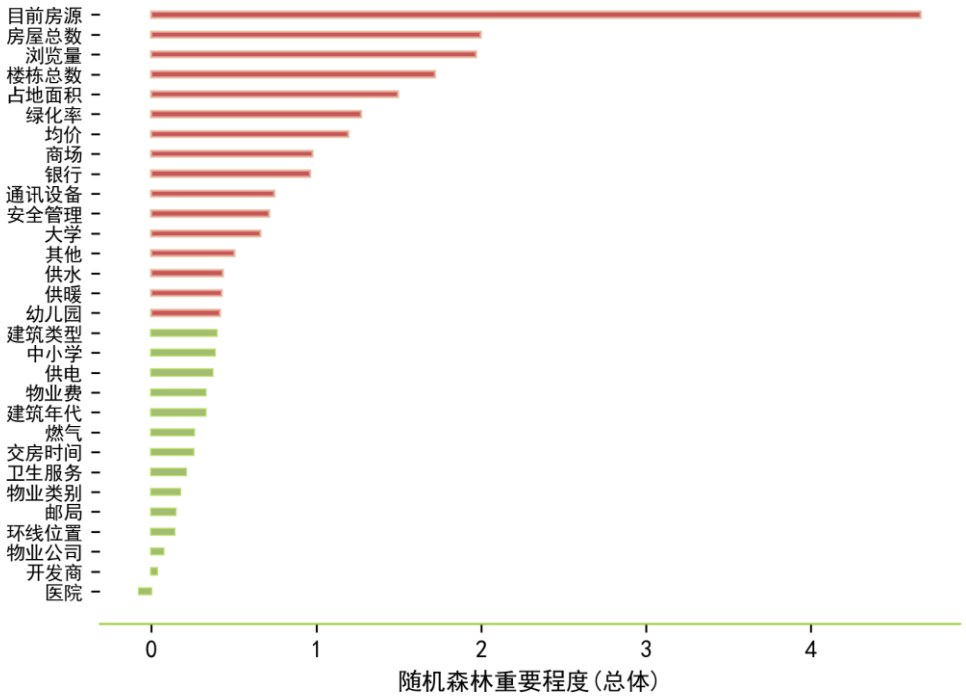


图 25　随机森林重要程度图（总体）

随机森林与梯度提升树的结论更加一致，为减少算法复杂度，针对其余各个区域的重要性指标选择随机森林方法进行变量重要性筛选，定性分析影响西安市各个区域发展的因素差异。

大西安产业结构规划图形成的“三廊三带一通道”空间格局中的三廊分别为以高新区为引领的科创大走廊、以经开区为引领的工业大走廊、以曲江区为引领的文创大走廊。

此外据 2019 年 6 月国家统计局西安调查队发布调研报告显示，西安正以蓬勃向上发展之力吸引应届毕业生的目光。因此特别针对西安市有名的大学城——长安区的房价情况进行分析。

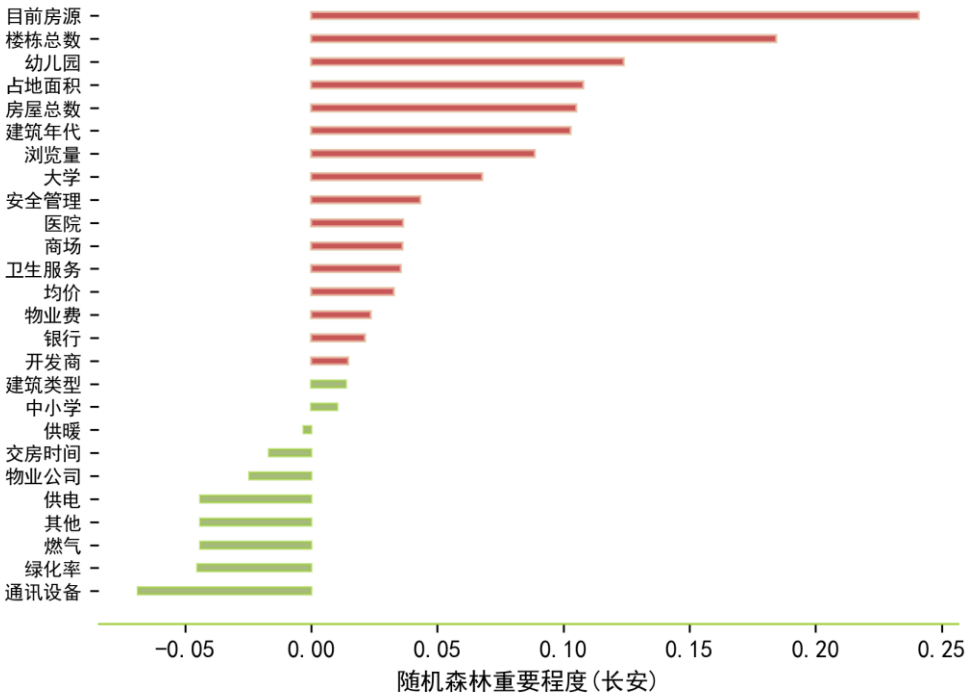


图 26　随机森林重要程度图（长安）

通过与西安市总体情况的对比不难看出，长安区平均房价靠后，并且由于长安区是西安市有名的“大学城”，越靠近大学的地段楼市房价越高，很好贴切了实际情况。同时值得注意的是供暖对于长安区的成交量影响较为显著，长安区地方政府的官方公布的数据文件侧面反映出长安区楼市入住率较低，无法满足硬性要求造成空房率较高的现状。因此通过提升居民便利服务设施，如增设银行、商场等聚点进一步带动以大学生为经济主体的长安区的房产交易情况。

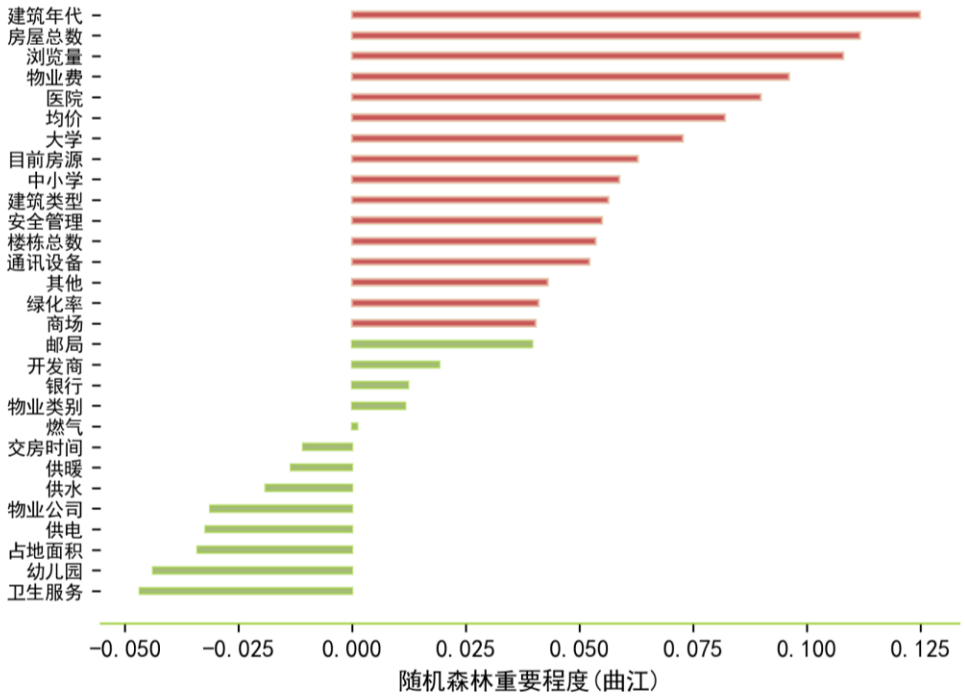


图 27　随机森林重要程度图（曲江）

曲江区的房价排名位于第一位，因此可以认为该区域居民的平均承受能力均较高。同时，医院变量排名显著靠前，通过分析曲江楼市特点不难看出，由于曲江环境较好因此相同地段内，绿化率的优势不再显著。曲江区邻近曲江遗址公园、大雁塔以及龙湖等地标建筑群可供休闲娱乐的选项较多，因此较多具有一定经济实力的退休人员选择在曲江地区进行购房，别墅群分布密集，因此建筑年代影响不容小觑。

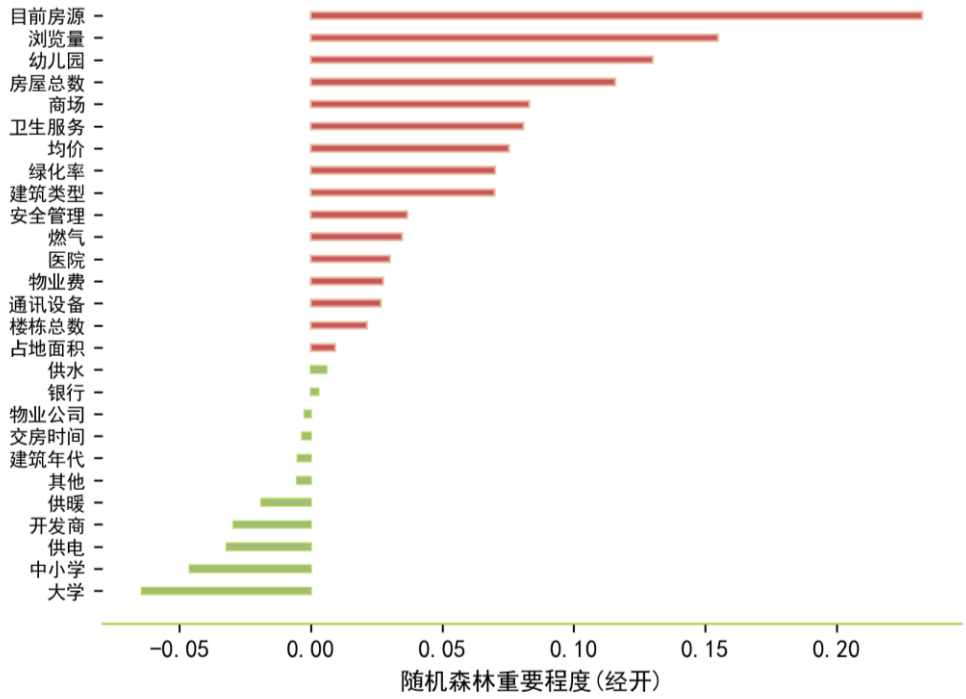


图 28　随机森林重要程度图（经开）

经开区平均房价排名较前，由于西安市经开区聚集了西安市咸阳机场以及各重大产业基地，截止 2019 年 1 月，经开区内共有企业 2800 多家，引进工业项目 500 余项。聚集大量新兴技术产业汇聚大量年轻新鲜血液，家政行业即卫生服务的重要性得以体现。此外，作为众多年轻人的就业点，幼儿园作为不可忽略的指标对该区的房地产行情其中举足轻重的作用。

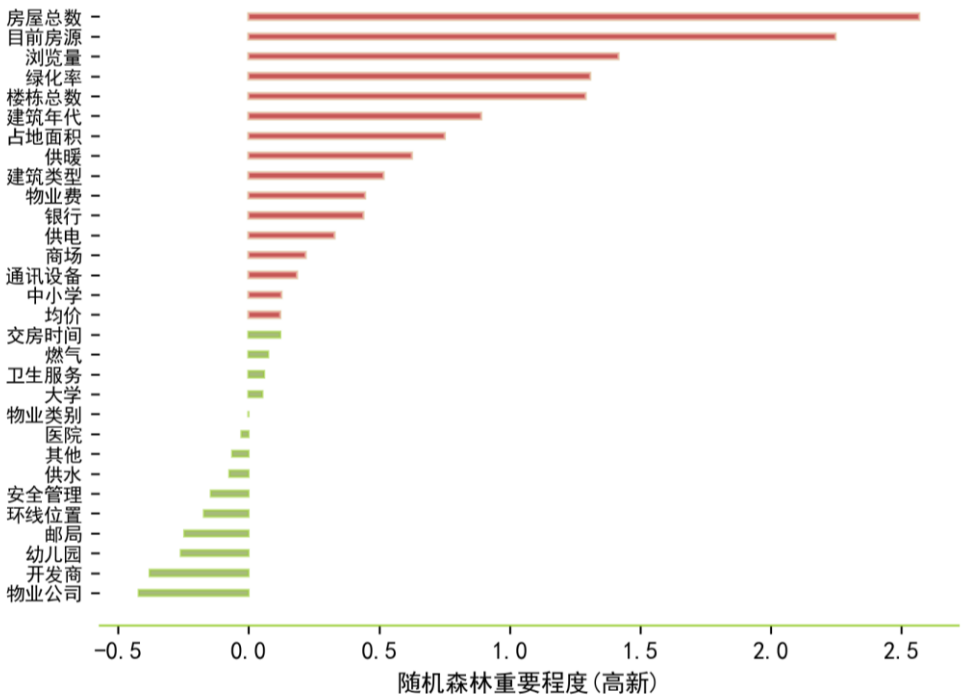


图 29　随机森林重要程度图（高新）

高新区一直居于西安市房价制高点，据统计最为西安面积仅 6.3%的区域却贡献超 15%的 GDP，作为西安经济最为活跃的地方。由于高新区的教育以及商业配套设施较为齐全，因此除了幼儿园影响显著外，开发商、开盘时间以及交房时间等楼盘本身的信息影响更为明显。特别的，在这个寸土寸金的地段，绿化率的影响更为显著。而与长安区由于入住率导致的供暖问题突出的原因不同，高新区由于存在“煤改气”以及“气改水”存在热力供应公司与住户之间的纠纷，进而造成供热对于高新区房价影响显著的现象。

因此针对不同区域进行影响交易量因素分析的结果可以看出，不同经济区域以及行政划分区域具有各自的发展格局以及楼市特色，因此合理发展经济稳控房价才能更好地留住毕业生为西安市各产业园区不断输送新兴的中坚力量。

五、结论与建议

（一）西安市房价纵向横向分析

1. 西安市房价增长速度放缓且周期基本与全国保持一致

①根据时间序列分析较为精准的预测值可以看到西安市房价在今年 9 月份很有可能逼近 1 万，整体增速放缓，调控房价作用显著。

②根据主成分分析可以看到房地产固定投资额占比较大，相应的城镇居民储蓄余额影响购房欲望。

③基于主成分结果对比分析西安市与全国房地产市场的周期情况，全国房地产市场的供给类合成指标的谱密度曲线存在主谱峰和次谱峰，而需求类合成指标的谱密度曲线仅存在主谱峰。西安市房地产市场的供给和需求均存在明显的周期特征，其中供给与需求的周期时间长度大致相等，表明西安房地产市场基本保持了供求关系的均衡发展，政策与需求同步稳进。然而，房地产需求类指标更高，这意味着从长期发展来看，市场将处于供求关系较匹配但需求更高的阶段，即房地产市场价格仍存在长期上涨的趋势。整体来看，西安市房价基本紧随全国房地产周期，但房价上涨趋势略高于全国。

④基于西安市大力宣扬“人才引入”政策，考虑房价、缴存单位户数、缴存职工户数、本月归集额四个指标之间的几何关联度较高，利用多元时间序列的 Kalman 滤波算法，从每期的拟合情况分析预测效果可看到住房公积金惠利性政策对于房地产市场有一定刺激作用，宏观政策调控影响不可小觑。

2. 外来开发商一定程度上刺激西安市楼市发展且地域差异明显

①通过横截面数据进行房价的多项逻辑回归分析，通过估计房价高于时间序列预测数据的概率值可以较为直观的看到卫生服务费用是否包含会显著影响房价的价格，此外外地开发商涌入西安房地产市场，一定程度上造成了房价的走高。

②不同经济区域的房价差距较为明显，越是远离城市中心的地域房价偏低的可能性也越高，西安市楼盘价格地域性差异显著。

因此相关政府仍需对现下的房地产市场进行合理监控，为更多的优秀人才解决住房问题，合理实现西安市产业如工业、文化产业的整体布局与居民基础设施建设的均衡化发展，为实现一系列的战略政策不断储备专业人才。

（二） 西安市楼盘交易量全局区域分析

1. 西安市全市交易量呈现温热态势且居住环境成为社会关注热点

针对全西安市而言，目前房源、房屋总数、建筑年代、浏览量、均价以及物业费是影响交易量较为显著的指标。自 2016 年底西安市拉开“抢人大战”序幕以来，全市交易量明显上升，均价对交易量的影响并不十分显著，但是浏览量的作用则更为明显，其体现了小区本身知名度以及开发商宣传力度。广告效应成为楼市出售前期吸引购买力的关键因素，居住环境成为购房者购房时的重要考核目标。

通过预测集以及真实交易量返回得到的预报图，交易量高的楼盘大致分布在高新区、未央区以及浐灞区附近，整体呈现外散趋势；主城区楼盘交易情况则呈疲软态势；较多楼市呈现出中等的温热状态，高交易量楼盘十分稀少。综上所述，西安市整体房源供给较为充足且交易态势平稳，为迎接更多优秀才子落户西安做好实质准备。

2. 西安市各区域交易量地域特色明显

依据“人才引进”下大西安产业结构规划图形成的“三廊三带一通道”的产业空间格局，重点分析四个区域楼盘大的交易情况。

“大学城”—长安区的房产均价靠后，但越靠近大学，房价相对越高。同时，针对长安区部分楼市入住率较低导致的无法满足硬性集中供暖的现状，考虑合理优化长安区住房环境成为留住西安本地人才的直接途径，提升居民便利服务设施也能够进一步带动以大学生为经济主体的长安区房产交易情况。

“别墅区”—曲江区邻近曲江遗址公园、大雁塔以及龙湖等地标建筑群，基础设施完善，医院指标排名较前，房价排名第一。较多具有一定经济实力的退休人员会选择在曲江地区进行购房，别墅区分布密度也体现了该区域居民的平均承受能力高于西安市平均水平，并且建筑年代一定程度上反映了“抢人大战”对楼盘销售的重要影响。

经开区，2800 多家企业，500 余项工业项目等大量新兴技术产业聚集地，同时也是高质量人才聚集地。考虑高收入人群对生活水平的高质量要求，楼盘物业费的重要性得以体现。此外，作为众多年轻人的就业园区，幼儿园作为不可忽略的指标对该区的房地产行情起着举足轻重的作用。

高新区一直居于西安市房价制高点，据统计占据西安面积仅 6.3%的区域却贡献超 15%的 GDP，是西安经济最为活跃的地方。该区的教育以及商业配套设施较为齐全，因此开发商、开盘时间以及交房时间等楼盘本身的信息影响更为明显。特别地，在这个寸土寸金的地段，绿化率的高低较为显著的影响购房者的选择倾向。与长安区入住率导致的供暖问题突出的问题不同，高新区由于“煤改气”以及“气改水”的问题存在热力供应公司与住户之间的纠纷，进而造成供热对于高新区房价影响显著的现象。因此如何进一步优化住户与承建商之间的服务协议，进一步改善住房周边环境，提升住户使用感与幸福感成为该区楼盘的破局点。

参考文献

引文文献

[1] 史策.基于培育大格局视野下创新西安市高层次人才引进机制研究[J].行政事业资产与财务,2019 年第 4 期：32-34
[2] 赵忠君，邹丽娜.人才引进政策实施效果评价[J].湖南财政经济学院学报，2019 年第 1 期：41-47
[3] 国家统计局西安调查队.就业签约进行时，选择意愿变化中[EB/OL].http://xadcd.xa.gov.cn/tjxx/rdzt/5d777a6ffd85082ba28539b2.html:2019 年 6 月 4 日/2019 年 6 月 18 日
[4] Hotelling H. Analysis of a complex of statistical variables into principal components[J].Education Psychology , 1933(3): 417-444
[5] 张红，谢娜.基于主成分分析与谱分析的房地产市场周期研究[J].清华大学学报，2008 年第 9 期：1404-1407
[6] 游皓麟.R 语言实战[M].北京：电子工业出版社，2016 年：251-411

阅读型文献

[7] 杨佃辉，陈轶，屠梅曾.基于聚类分析和非参数检验的房地产预警指标体系选择[J].东华大学学报，2006 年第 2 期：59-62
[8] Wernecke M and Rottke N.Holzmann C.Incorporating the real estate cycle into management decisions-evidence from Germany [J].Real Estate Portfolio Management , 2004(3): 171-186
[9] Wheaton W C. Real estate cycle Some fundamentals[J].Real Estate Economics , 1999，27(2): 209-230

附　录

1.主成分

首先对原始数据标准化对原始数据进行标准化处理，结果如下表所示：

表 5　标准化数据表

年份	供给类指标的标准值					需求类指标的标准值				
	X1	X2	X3	X4	X5	Y1	Y2	Y3	Y4	Y4
2001	-1.04444	-1.11288	1.17752	-1.23672	-.92329	-1.24057	-1.22108	-.96792	-1.25463	-1.19618
2002	-1.03050	-1.09180	.03507	-1.14124	-.89547	-1.20020	-1.16158	-.96074	-1.21548	-1.21886
2003	-.99632	-1.03420	.92385	-1.14648	-.87490	-1.15267	-1.10147	-.95807	-1.22048	-1.17349
2004	-.96648	-.99643	.09965	-1.08177	-.78988	-1.08559	-1.01482	-.95353	-1.27455	-.96976
2005	-.91727	-.93635	.12612	-.93782	-.82957	-.99423	-.91558	-.85729	-.96622	-.87247
2006	-.84667	-.85133	-1.90706	-.91125	-.75105	-.88661	-.79469	-.75578	-.75987	-.67317
2007	-.73068	-.72683	-.78044	-.79035	-.57828	-.73853	-.59503	-.65481	-.49457	-.64658
2008	-.59477	-.52570	.52183	-.61584	-.65952	-.52404	-.38656	-.63089	-.57786	-.42090
2009	-.40347	-.36482	-.82092	-.16299	-.45538	-.20750	-.19808	-.37883	.03260	-.42805
2010	-.17780	-.17359	1.70743	.05471	-.61886	.06902	-.02918	-.08724	.44699	-.18709
2011	-.01344	.04928	-.60241	.38533	-.27319	.38396	.13845	.42573	.68455	.53415
2012	.30314	.42603	-1.46919	.68019	.62036	.72116	.26405	.32753	.38592	.73770
2013	.65090	.78434	.86306	.72163	.06617	.98394	.47603	.45451	.54058	.77176
2014	.96283	.94949	.87426	1.03618	1.59101	.78294	.70588	.43828	.59674	.66586
2015	1.12787	1.00959	.37531	1.04851	.44057	.99135	.96922	.49979	.66664	.69014
2016	1.20063	1.22451	-1.15444	1.22036	1.64569	1.19716	1.26179	.76713	1.02132	.72282

2017	1.537 16	1.548 24	.7637 4	1.385 72	1.799 45	1.442 07	1.616 36	1.763 76	1.535 47	1.550 91
2018	1.939 32	1.822 43	-.733 38	1.491 86	1.486 18	1.458 34	1.986 29	2.528 36	1.852 85	2.113 20

随后计算因子的特征根和主成分的特征向量表对上表的数据进行因子分析。在对供给类指标和需求类指标数据分别进行因子分析时，先判断指标间是否有共线性存在。随后剔除部分导致共线性现象发生的变量，再次进行因子分析。计算因子的特征根和主成分的特征向量表标准化后的数据进行因子分析后，分别得到包含供给类和需求类原始指标所有信息的因子 S_n 和 D_n $(n=1,\ldots,4)$ 。可以得到各因子的特征根和方差贡献率如下表所示：

表 6　供给类因子特征根与方差贡献率

供给类因子	特征根	累计方差贡献率	需求类因子	特征根	累计方差贡献率
S1	2.897	72.429	D1	3.881	97.022
S2	.993	97.264	D2	.076	98.910
S3	.090	99.521	D3	.029	99.623
S4	.019	100.000	D4	.015	100.000

从上表可知， S_1,S_2 和 D_1,D_2 的累计方差贡献率达 97.264%、98.910%，可分别成为供给类指标与需求类指标的主成分。进一步计算得到 S_1,S_2 和 D_1,D_2 的特征向量表即主成分表达式的变量系数如下:

表 7　主因子特征向量表

供给类变量	S1	S2	需求类变量	D1	D2
X1	.341	.079	Y1	.253	-2.295
X2	-.04 2	.999	Y2	.256	-.039
X3	.339	.047	Y3	.252	2.790
X4	.335	-.003	Y4	.255	-.436

最后确定合成指标。合成指标值是主成分的加权平均数，权数为方差贡献率。2001-2018 年供给类与需求类的合成指标值 S 与 D 如下表所示：

表 8　合成指标表

年份	S1	S2	S	D1	D2	D
2001	-1.13	-.15	-1.12	-1.19	.74	-1.14
2002	-1.04	-.08	-1.02	-1.15	.65	-1.11
2003	-1.06	-.13	-1.04	-1.13	.55	-1.08
2004	-.97	-.08	-.95	-1.10	.43	-1.06

2005	-.91	-.08	-.90	-.95	.35	-.92
2006	-.77	.02	-.76	-.81	.29	-.78
2007	-.68	-.03	-.67	-.63	.11	-.61
2008	-.65	-.06	-.64	-.54	-.29	-.53
2009	-.31	-.02	-.31	-.19	-.59	-.20
2010	-.32	-.01	-.32	.10	-.60	.08
2011	.06	-.01	.06	.41	.00	.40
2012	.60	-.02	.59	.43	-.92	.40
2013	.45	.08	.45	.62	-1.24	.58
2014	1.18	.11	1.16	.64	-.86	.60
2015	.87	.11	.86	.79	-1.21	.74
2016	1.42	.02	1.40	1.08	-1.10	1.02
2017	1.56	.17	1.54	1.61	.88	1.60
2018	1.70	.10	1.67	1.99	2.82	2.01

2.谱分析

进行单位根检验的结果表明，S 与 D 均进行二次差分后的序列 S'' 与 D'' 可成为平稳时间序列，其中样本长度 $N=14$，谱分析所需样本数据如下：

表 9　谱分析所需样本数据

年份	S''	D''	年份	S''	D''
2004	0.28	0.00	2011	0.11	0.02
2005	0.07	0.12	2012	-0.07	-0.02
2006	-0.03	-0.08	2013	0.02	0.17
2007	0.05	0.11	2014	-0.06	-0.33
2008	0.02	-0.25	2015	-0.22	0.24
2009	0.02	0.39	2016	0.13	0.25
2010	0.06	-0.19	2017	-0.11	-0.20

确定截断点与频率分量的个数。为保证谱密度估计值的渐进无偏和一致性，需要合理确定截断点 M 及由 M 确定的窗函数所分辨的频率分量的个数 m 其中，$m=(1,2,\cdots,M)$ 通常，当 N<50 时，M=N/2。因此，这里的 $M=14/2=7, m=(1,2,\cdots,7)$ 根据频率的计算公式 $f=m/N$ 及周期计算公式 $P=1/f$，可以得到各频率分量对应的频率值与周期，如下表所示：

表 10　频率与周期汇总表

m	频率	周期/a
1	1/14	14.0
2	2/14	7.0
3	3/14	4.7
4	4/14	3.5
5	5/14	2.8
6	6/14	2.3
7	7/14	2.0

3.聚类算法

考虑到对于开发商以及购房者而言，确切的预测交易量的连续值意义不大，且由于阶段数据限制可能无法均衡开盘时间较晚的小区，因此通过聚类转化为分类变量即高、中、低三个大类指标。由于受到西安市各个区域差异的可能影响，因此采取加权聚类确定相关指标。收集得到的样本中房源来源分为城北、城南、城西、城东以及城内共计 5 个明确区别区域，为方便描述将标记为 A_1,A_2,A_3,A_4,A_5。对上述四个区域区域进行 K-means 聚类得到 k 为 2 时的四个聚类中心 $(x_i,y_i)(i=1,2,3,4,5)$，随后取四个区各自占总爬取得到的数据总数的占比进行加权聚类分析：

$$(X,Y)=\sum_{i=1}^{4}\frac{n_i}{N}(x_i,y_i)(i=1,2,3,4,5;N=1659)$$

其中,n_i 表示 4 个区域的记录总数，N 为 1659，$\frac{n_i}{N}$ 表示各个区占据总记录数的比例。

4.主要代码

①爬虫房地产数据代码（Python）

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver import ActionChains
from pyquery import PyQuery as pq
from time import sleep
import re,csv,requests
```

```
import pandas as pd
from bs4 import BeautifulSoup
import time
class Home:

    def __init__(self):
        self.headers = {}
        self.headers['User-Agent'] = 'Mozilla/5.0 (Windows NT 10.0;
Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/74.0.3729.131
Safari/537.36'
    def spider(self):
        for u in range(48,101):
            url =
'https://xian.esf.fang.com/housing/__0_0_0_0_'+str(u)+'_0_0_0/'
            res = requests.get(url,headers=self.headers)
            coding =
requests.utils.get_encodings_from_content(res.text)[0]
            res.encoding = coding
            soup = BeautifulSoup(res.text, 'lxml')
            b = soup.find_all(attrs={'class': 'plotTit'})
            res = []
            for i in b:
                res.append('https:'+re.findall(r'href="(.*?)/"
target="_blank"',str(i))[0])
            with open('fangtianxia2.txt','a') as f:
                for i in res:
                    f.write(i+'\n')
            print(res)
            print("第"+str(u)+"完成")
            time.sleep(5)

    def data(self):
        with open('fangtianxia2.txt','r') as f:
            url = []
            for i in f:
```

```
                url.append(i[:-1]+'/xiangqing/')
        with open('fangtianxia_xinxi.txt','a') as f:
            for i in url:
                f.write(i+'\n')

    def deal(self):
        urls = []
        with open('fangtianxia2.txt','r') as f:
            for i in f:
                urls.append(i[:-1])
        file = 'fangtianxia_deal.csv'
        title = ['名称','二手房源', '最近成交']
        page = 1211
        for url in urls:
            if urls.index(url) == 0:
                with open(file, 'a', newline='', encoding='utf-8-sig')
as f:
                    w = csv.writer(f)
                    w.writerow(title)
            if urls.index(url)<page:
                continue
            try:
                res = requests.get(url, headers=self.headers)

                res.encoding = 'gbk'
                soup = BeautifulSoup(res.text, 'lxml')
                b = soup.find_all('li','zygwbox')
                c = soup.find_all('b')
            except:
                b =[]
                c = []
            datas = []
            data = {}

            try:
```

```
        data['名称'] = c[0].text
    except:
        data['名称'] = url
    for i in b:
        datas.append(i.text)
    for i in datas:
        if "二手房源" in i:
            count = re.sub("\D", "", i)
            data['二手房源'] = count
        if "最近成交" in i:
            count = re.sub("\D", "", i)
            data['最近成交'] = count
    input_ = []
    for i in title:
        if i in data:
            input_.append(data[i])
        else:
            input_.append('暂无数据')
    with open(file,'a', newline='', encoding='utf-8-sig') as f:
        w = csv.writer(f)
        w.writerow(input_)
    print("第"+str(urls.index(url)+1)+"已完成 url="+url)
    time.sleep(1)
    # strs = ''
    # for i in b:
    #     strs = str(i.text)
    # strs = strs.split('\n')
    # while '' in strs:
    #     strs.remove('')
    # res = []
    # for i in range(0,len(strs),6):
    #     datas = []
    #     for j in range(i,i+6):
    #         datas.append(strs[j])
    #     res.append(datas)
```

```
            # print(res)
            # time.sleep(1000)
    def data_spider(self):
        urls = []
        file = 'fangtianxia_data.csv'
        with open('fangtianxia_xinxi.txt', 'r') as f:
            for i in f:
                urls.append(i[:-1])
        title = ['名称','均价', '浏览量', '小区地址', '所属区域', '邮编', '环线位置', '物业类别', '建筑年代', '开发商', '建筑类型', '建筑面积', '占地面积', '房屋总数',
                 '楼栋总数', '物业公司', '绿化率', '容积率', '物业费', '供水', '供暖', '供电', '燃气', '通讯设备', '卫生服务', '开盘时间', '交房时间', '开盘时间',
                 '交房时间', '电梯服务', '安全管理', '停车位', '幼儿园', '中小学', '大学', '商场', '医院', '邮局', '银行', '其他']
        page = 1211
        for url in urls:
            if urls.index(url) < page:
                continue
            if urls.index(url) == 0:
                with open(file, 'a', newline='', encoding='utf-8-sig') as f:
                    w = csv.writer(f)
                    w.writerow(title)
            try:
                res = requests.get(url, headers=self.headers)
                # coding = requests.utils.get_encodings_from_content(res.text)[0]
                # res.encoding = coding
                res.encoding='gbk'
                soup = BeautifulSoup(res.text, 'lxml')
                b = soup.find_all('dd')
                c = soup.find_all('dt')
                d = soup.find_all(attrs={'class':'org'})
```

```
        n = soup.find_all(attrs={'class':'tt'})
    except:
        b=[]
        c=[]
        d=[]
        n=[]
    data = {}
    all_data = []
    for i in b:
        string = i.text.replace('\xa0','').replace(' ','')
        s = string.split('：')
        all_data.append(s)
    for j in c:
        string = j.text.replace('\xa0', '').replace(' ','')
        s = string.split('：')
        all_data.append(s)
    if all_data != '':
        for i in n:
            data['名称'] = i.text
        try:
            data['均价'] = all_data[0][0]
        except IndexError as e:
            print(e)
        for i in d:
            data['浏览量'] = i.text
        for i in all_data:
            if i[0] in title:
                data[i[0]] = i[1]
    csv_text = []
    print(data)
    for i in title:
        if i in data:
            csv_text.append(data[i])
        else:
            csv_text.append('暂无数据')
```

```
            with open(file, 'a', newline='', encoding='utf-8-sig') as f:
                w = csv.writer(f)
                w.writerow(csv_text)
            print("第"+str(urls.index(url)+1)+"已完成 url = " + url)
            time.sleep(5)

home = Home()
home.deal()
```

②GBRT 梯度提升树代码（R）

```
library(gbm)
#基础数据准备
vdata<-read.csv(file.choose(),header=T)#数据完整的
#vdata<-read.csv(file.choose(),header=T)#缺失数据的
vdata=vdata[,2:32]
#建立 gbm 模型
gbm.obj<-gbm(交易量~.,data=vdata,distribution='gaussian',
             var.monotone=rep(0,30),
             n.trees=1000,
             shrinkage=0.01,
             interaction.depth=5,
             bag.fraction=0.5,
             cv.folds=10)
#用交叉验证确定最佳迭代次数
best.iter<-gbm.perf(gbm.obj,method="cv")
best.iter
#进行预测
vdata$pred=predict(gbm.obj,vdata,n.trees=best.iter)
write.csv(vdata,'D://统计建模//测试集预测效果.csv',row.names=F)
#查看前 6 行数据
head(vdata)
#统计残差平方和
sum((vdata$交易量-vdata$pred)^2)
#分析变量重要性
vital<-summary(gbm.obj,n.trees=best.iter)
#write.csv(vital,'D://统计建模//各指标重要程度.csv',row.names=F)
```

```
#绘制各变量的边际图
plot(gbm.obj,1,best.iter)
plot(gbm.obj,2,best.iter)
plot(gbm.obj,3,best.iter)
plot(gbm.obj,4,best.iter)
plot(gbm.obj,5,best.iter)
plot(gbm.obj,6,best.iter)
plot(gbm.obj,7,best.iter)
plot(gbm.obj,8,best.iter)
plot(gbm.obj,9,best.iter)
plot(gbm.obj,10,best.iter)
plot(gbm.obj,11,best.iter)
plot(gbm.obj,12,best.iter)
plot(gbm.obj,13,best.iter)
plot(gbm.obj,14,best.iter)
plot(gbm.obj,11,best.iter)
plot(gbm.obj,12,best.iter)
plot(gbm.obj,13,best.iter)
plot(gbm.obj,14,best.iter)
plot(gbm.obj,15,best.iter)
plot(gbm.obj,16,best.iter)
plot(gbm.obj,17,best.iter)
plot(gbm.obj,18,best.iter)
plot(gbm.obj,19,best.iter)
plot(gbm.obj,20,best.iter)
plot(gbm.obj,21,best.iter)
plot(gbm.obj,22,best.iter)
plot(gbm.obj,23,best.iter)
plot(gbm.obj,24,best.iter)
plot(gbm.obj,25,best.iter)
plot(gbm.obj,26,best.iter)
plot(gbm.obj,27,best.iter)
plot(gbm.obj,28,best.iter)
plot(gbm.obj,29,best.iter)
```

③关联度主要代码（MATLAB）

price=[7255,7294,7401,7386,7380,7270,7256,7236,7175,7138,7133,7055,7056,7058,7030,7045,7045,6992,6914,7030,7108,7242,7280,7153,7170,7232,7188,7233,7271,7296,7341,7380,7400,7456,7457,7373,7254,7121,6927,6904,6895,6878,6746,6700,6722,6737,6726,6666,6691,6679,6672,6510,6511,6426,6403,6406,6417,6446,6476,6508,6564,6551,6578,6589,6754,6871,6968,7004,7033,7149,7375,7531,7747,7966,8117,8333,8380,8393,8488,8552,8615,8691,8751,8855,8955,9032,9086,9160,9230,9289,9408,9475,9535,9638];
ucn=[11965,12098,12226,12351,12402,12545,12562,12727,12811,12929,13022,13107,13189,13286,13387,13479,13554,13647,13689,13870,13938,14068,14166,14249,14364,14461,14518,14583,14677,14677,14868,15002,15095,15226,15330,15452,15567,15657,15784,15881,15968,16073,16140,16339,16415,16525,16653,16757,16851,16977,17072,17182,16473,16651,16715,16943,17006,17140,17289,17418,17543,17676,17793,17876,17972,18147,18265,18463,18629,18807,19118,19369,19553,19898,20218,20544,20805,21149,21336,21359,21458,21848,22203,22198,22584,22934,23289,23669,24002,24489,24890,25655,25937,26497];
pcn=[1398349,1404901,1420847,1438259,1439501,1453096,1473272,1474938,1482144,1490293,1499128,1508453,1522543,1533775,1558653,1577779,1593423,1609344,1626921,1643616,1651992,1663248,1677797,1692130,1709004,1724228,1742479,1762102,1777494,1795670,1812061,1827123,1840868,1858768,1878931,1897506,1911436,1922873,1945548,1971691,2010985,2027025,2045702,2060980,2069463,2079444,2093480,2109620,2124822,2136046,2163014,2188814,1989104,1999936,2002322,2025672,2031926,2040684,2051971,2062400,2075605,2084087,2102775,2124680,2135800,2148650,2159396,2171171,2178528,2188225,2198186,2210319,2218333,2231069,2259844,2286878,2304614,2320209,2335647,2349610,2359517,2372125,2385416,2405662,2423399,2440782,2477480,2508857,2528768,2548621,2566912,2587058,2599069,2615625];
mc=[142693.360000000,58863.3700000000,81552.5100000000,93112.7000000000,81524.8600000000,102818.770000000,115171.800000000,64601.4600000000,93882.2200000000,108867.520000000,89113.7000000000,101264.120000000,175232.250000000,82186.7400000000,91614.5600000000,107309.680000000,100278.590000000,106128.370000000,125472.870000000,101491.440000000,81385.4600000000,122112.410000000,111375.630000000,112008.140000000,193088.160000000,99099.0300000000,98816.1900000000,117997.350000000,103324.950000000,135773.280000000,145386.160000000,107998.400000000,9621

2.5800000000,129888.360000000,123354.570000000,120201.270000000,22916
4.740000000,88793.6300000000,106324.500000000,136811.020000000,133400
.780000000,135442.610000000,169315.910000000,119959.350000000,117168.
260000000,110934.670000000,158611.510000000,140908.020000000,217268.4
30000000,110036.080000000,132261.200000000,147879.320000000,139962.97
0000000,152486.410000000,174787.140000000,131326.850000000,125742.760
000000,179757.680000000,139356.670000000,145257.810000000,225972.8200
00000,121773.210000000,142473.910000000,164847.550000000,158652.06000
0000,184104.910000000,181080.940000000,131855.370000000,138295.460000
000,197613.590000000,150306.070000000,169067.920000000,239704.0800000
00,144256.860000000,172779.700000000,190359.760000000,176145.54000000
0,222281.200000000,200046.210000000,166321.190000000,155695.500000000
,224087.680000000,187183.690000000,216087,289620.940000000,176413.890
000000,199821.560000000,212608.400000000,222787.890000000,220967.7000
00000,277328.230000000,204506.120000000,188162.990000000,236880.74000
0000];
mli=[37480.3000000000,29883.2000000000,18731.7000000000,69693,23359.6
000000000,27037.5000000000,76496.6000000000,38101.9000000000,30697.40
00000000,26677.5000000000,34232.8000000000,36005.1000000000,40903.200
0000000,44661.1000000000,87067.1000000000,52234.1000000000,61339.5000
000000,55240.5000000000,22310,79818.8000000000,74978.3000000000,62640
,47018.4000000000,74790.9000000000,55080.6000000000,75929.5000000000,
57608,43015.3000000000,53764.8000000000,44394.8000000000,70031.900000
0000,81866.5000000000,24586.8000000000,46744.8000000000,55610.2000000
000,75598.5000000000,62108,47954,67366,59626,41569,64671,170368,66356
,92153,44969,70017,66687,83863,81960,88532,108761,58974,127496,154929
,104228.300000000,76792,144259,119978,150127,131129,116693,108048,122
012,82784,125812,113191,134126,86085,133414,113099,123829,144049,8604
6,111908,113051,92180,112166,93178,129533.500000000,89494.5000000000,
127599,160660,195759,187423,156118,156090,117679,103295,148534,137106
,184703,115546,177943];
x0 = price;
x1 = ucn;
x2 = pcn;
x3 = mc;

```
x4 = mli;
x0 = x0 ./ x0(1);
x1 = x1 ./ x1(1);
x2 = x2 ./ x2(1);
x3 = x3 ./ x3(1);
x4 = x4 ./ x4(1);
global_min = min(min(abs([x1; x2; x3; x4] - repmat(x0, [4, 1]))));
global_max = max(max(abs([x1; x2; x3; x4] - repmat(x0, [4, 1]))));
rho = 0.5;
zeta_1 = (global_min + rho * global_max) ./ (abs(x0 - x1) + rho * global_max);
zeta_2 = (global_min + rho * global_max) ./ (abs(x0 - x2) + rho * global_max);
zeta_3 = (global_min + rho * global_max) ./ (abs(x0 - x3) + rho * global_max);
zeta_4 = (global_min + rho * global_max) ./ (abs(x0 - x4) + rho * global_max);
figure;
plot(x0, 'ko-' )
hold on
plot(x1, 'b*-')
hold on
plot(x2, 'g*-')
hold on
plot(x3, 'r*-')
hold on
plot(x4, 'y*-')
legend('price', 'ucn', 'pcn', 'mc','mli')
figure;
plot(zeta_1, 'b*-')
hold on
plot(zeta_2, 'g*-')
hold on
plot(zeta_3, 'r*-')
hold on
plot(zeta_4, 'y*-')
title('Relation zeta')
legend('ucn', 'pcn', 'mc','mli')
mean(zeta_1)
```

mean(zeta_2)
mean(zeta_3)
mean(zeta_4)

致　谢

爬取数据点点滴滴，方晓数据来之不易；插补寻漏林林总总，才知建模精益求精。时间如白驹过隙，但是统计建模比赛这次经历却给我们每位参赛队员留下深深烙印。

一路走来，感谢队友的相互支持，我们不断将理论知识应用于生活实际，学会运用统计学思维看待并解决问题，收获颇丰。同时，王金霞老师严谨的治学精神，精益求精的工作作风，深深地感染和激励着我们。在她的悉心教导和孜孜教诲下，我们顺利完成这次论文的结题工作，在此谨向王老师致以诚挚的谢意！此外，感激数据挖掘课程老师的倾力帮助，并向提供西安市楼市相关信息的同学、朋友以及师长表示由衷的感谢！