“房子是租来的，但生活不是。”

——重庆租房价格水平研究报告

复旦第三区交通委小组

李泽君、赵雅滢、谢炳辉、王维实、钟诚

2019.5.7

目录

一、 背景介绍

“房子是租来的，但生活不是。”

党的十九大报告提出，“坚持房子是用来住的、不是用来炒的定位，加快建立多主体供给、多渠道保障、租购并举的住房制度”。在这种政策导向和当下越来越紧张的住房市场的双重影响下，“租房”这一概念应运而生，并迅速的受到广大初入社会的青年所亲睐。对于刚刚大学毕业的青年人来说，租一套房子既可以在忙碌的城市中迅速拥有一个温馨的小窝，也可以把不多的收入投入到更有回报的项目上而不是成为“房奴”。数据显示，2018 年重庆整体租金上浮超过 20%，正体现了目前房地产租赁市场的火爆。

同时，在房地产租赁市场中，房东与租客双方存在严重的信息不对称现象。部分房地产中介机构，隐瞒交易双方真实交易（租赁）价格，从中截流交易款项，或者炒作规划概念，虚构楼盘升值空间。因此，我们如果可以通过房源的相关特征对其在市场上的价值进行评估，就可以消除信息不对称对双方造成的经济损失，减少黑心中介的获利。

重庆，作为国内人口最多的城市，占地广，人口多，又受到城市多山地的影响，宅建用地本就紧张，在边缘区域人口流入城区时，势必造成房价的上升和租房市场的大幅发展，故我们以重庆的租房市场为例，一窥租房价格水平的全貌。

二、 数据预处理

2.1 概述

数据集中由以下几栏组成：基本信息、房源地段、房源平台、发布时间、是否近地铁、是否免中介费、房源优势和房租，其中房租水平分为高低两项，为需要研究的因变量；其于为自变量。基本信息由三部分组成：房屋名、户型大小与租赁方式，我们采用文本提取的方法提取出相关信息，形成新的特征。

2.2 缺失值处理

因为本数据集中的数据多为水平变量，故缺失值较少；但在处理户型大小时，我们发现有 434 个样本没有形似“X 室 X 厅”的规则表述，而为“套间”，“X 房”，“X 床”，占

总样本数约 20%，通过对装修相关资料的查阅，最终决定采用如下对应关系：

开间——一室一厅，X 居——X 室——X 房

以这种方式进行处理后，缺失值下降到 179 例，占总样本数约 13%，以“未标记”进行处理。同样，对于其他特征中的缺失值，因其总数不多，统一以“未标记”水平对其进行处理。

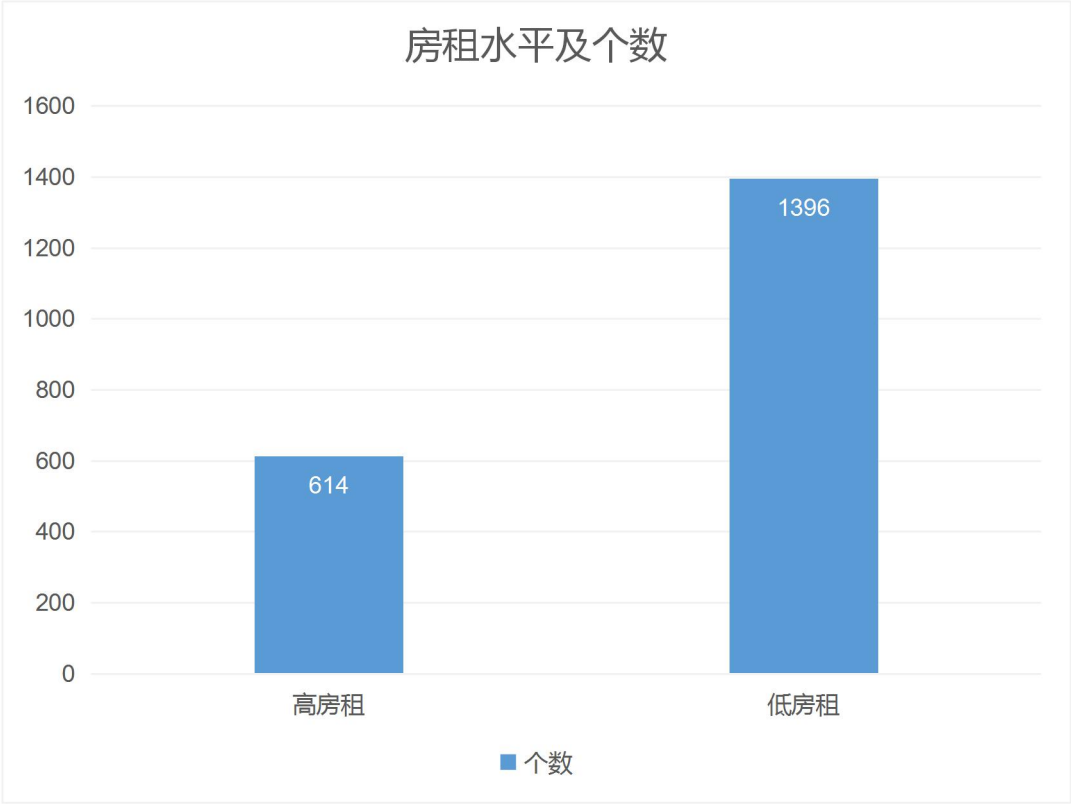
2.3 预处理结果

变量类型		变量名	取值水平	备注
因变量		房租水平	高、低	
自变量	时间数据	发布时间	X 天前、X 月前	
	分类数据	房源平台	链家、小租乐时尚长住公寓等	
		房源优势	月租、新上、随时看房等	
		房源地段	渝北、江北等	
	0-1 变量	是否近地铁	是/否	
		是否免中介费	是/否	
	分类数据	户型大小	X 室 X 厅	缺失值单独分类
	分类数据	租赁方式	合租、整租、其他	缺失值单独分类

三、 变量分析

3.1 因变量

首先对因变量水平数进行分析并绘制柱状图，从图中我们可以发现，数据集中有 1396 个样本为低房租，614 个样本为高房租，比例约为 2:1。

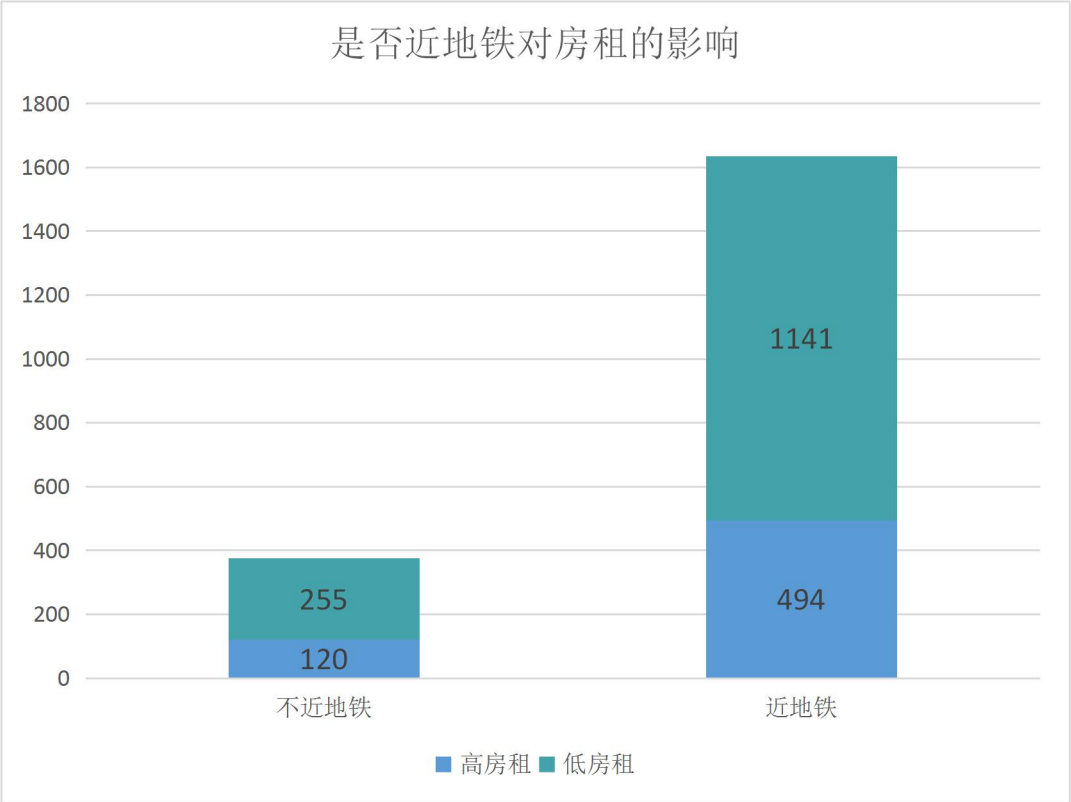


图一 房租水平及其个数

3.2 是否近地铁

对于住房来说，交通条件是非常重要的一方面，便捷的交通条件意味着其能为住客省下一大笔交通费用。如果在房屋周围有地铁等公共交通工具，房屋的价值也会随之水涨船高。

以是否进地铁对房租的高低水平作图，可以发现在房屋样本中，大部分的房屋都是近地铁的，可能与重庆城区山地环境所造就的发达城市交通网络有关。而无论是否近地铁，高房租与低房租样本比例都约为 1:2，与总体比例相同，可以推断，是否近地铁并不是影响房租水平的主要因素。



图二 近地铁——房租水平堆积柱状图

3.3 是否免中介费

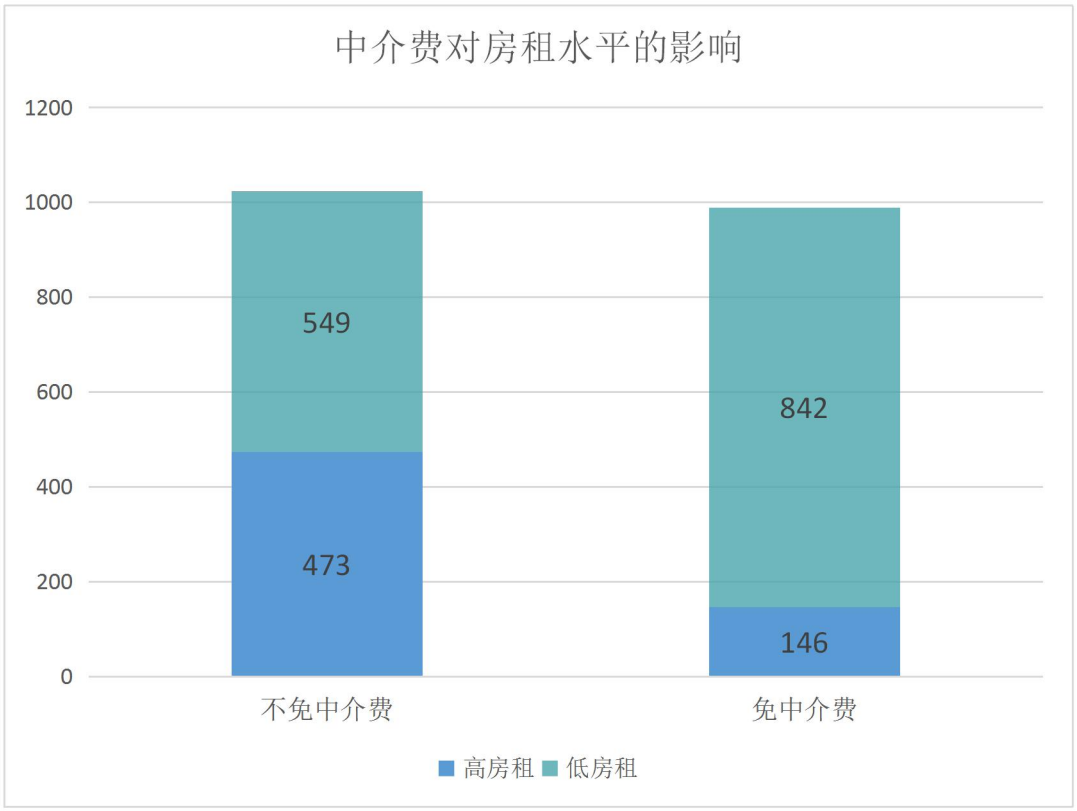
租过房子的人都知道，中介在租赁双方的议价过程中起到了十分重要的作用，中介费用的收取也会被计算在房租价格之内，故我们对中介费用与房租水平的关系进行研究。

从图中可以看出，收取中介费与不收中介费的样本量近似相同，而在不免中介费的样本中，高房租的比例要高于免中介费样本中高房租的比例。与我们对于该特征的预测相同。

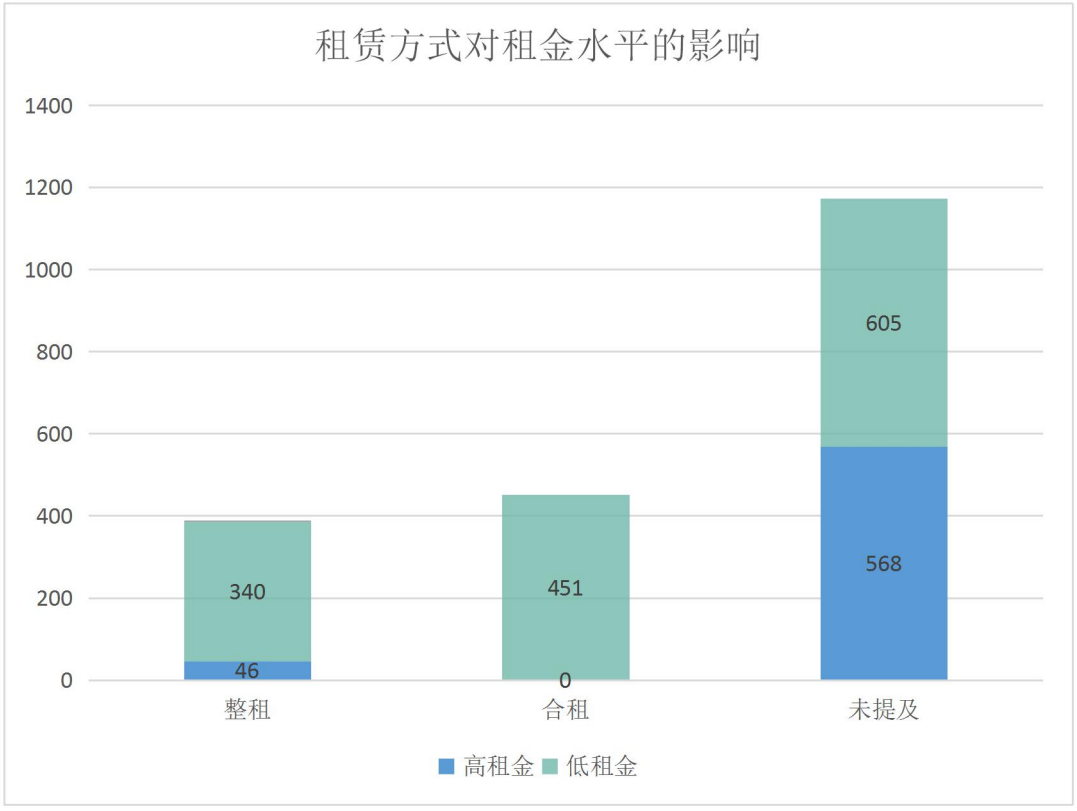
3.4 租赁方式

在现有的租赁市场中，有两种主要的租赁方式——整租和合租。合租是指至少两人一起租住一间或一套住房，是比较常见的居住方式。特别是在北京、上海等一线城市。合租者多为外地的年轻人，刚刚工作，还没有买房子的能力，于是几个人一起租一套房子。整租是指一个房子只有自己住，自己承担费用。显然，整租的费用要高于合租的费用。

考察原始数据，我们发现，大部分的租赁方式并没有对整租和合租作出说明，但是对于作出说明的样本特征较为明显，合租方式的样本均为低租金，而整租方式的样本低租金与高租金之比也达到了近 9:1，故推断租赁方式可能并不会对租金水平造成显著影响。



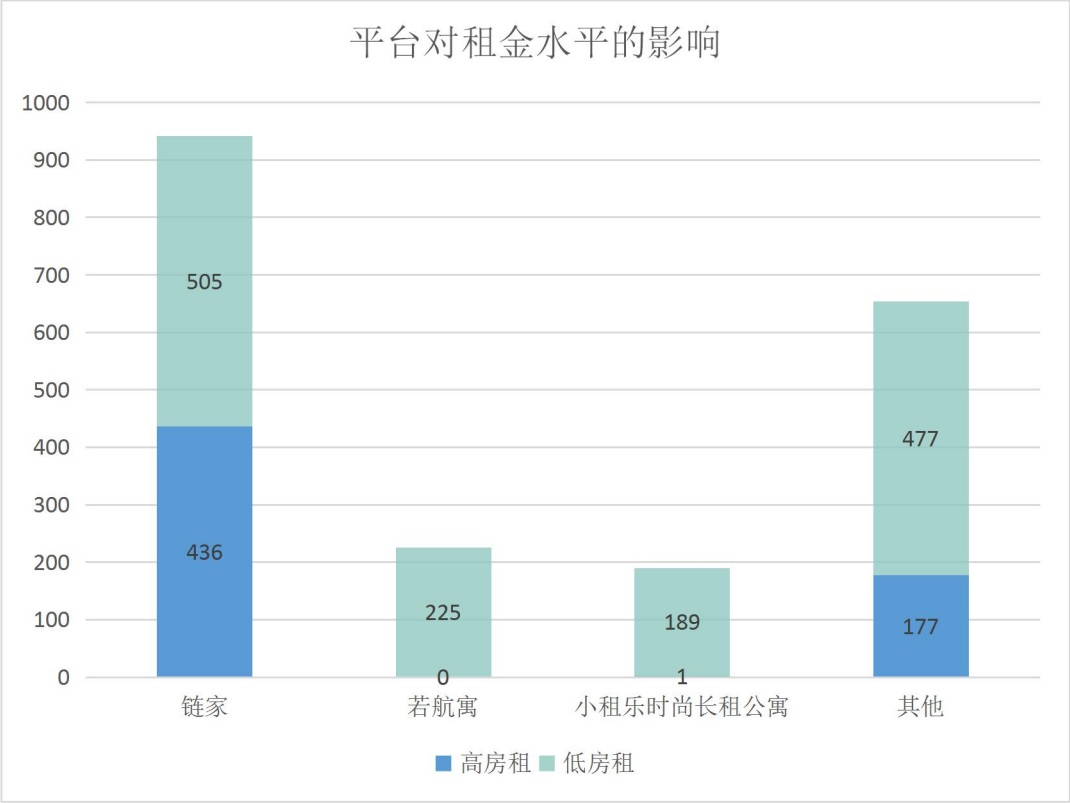
图三——中介费--房租水平堆积柱状图



图四---租赁方式--房租水平堆积柱状图

3.5 租赁平台

不同的租赁平台面向不同层次的顾客，自然面向高端顾客的平台租金会比中低端顾客的高。在我们选择租赁平台的时候，也会根据房屋的质量选择不同的租赁平台，我们对不同平台上租金价格水平进行研究，把样本数较高的链家、若航寓与小租乐等大平台和其他样本数较低的租赁平台进行比较。



图五——租赁平台--租金水平堆积柱状图

观察柱状图可以发现，不同平台面向的客户与租金水平具有较大差别，链家平台中高低房租样本量接近，而若航寓和小租乐则全部为低房租产品。不同租房平台的租金水平差异明显。

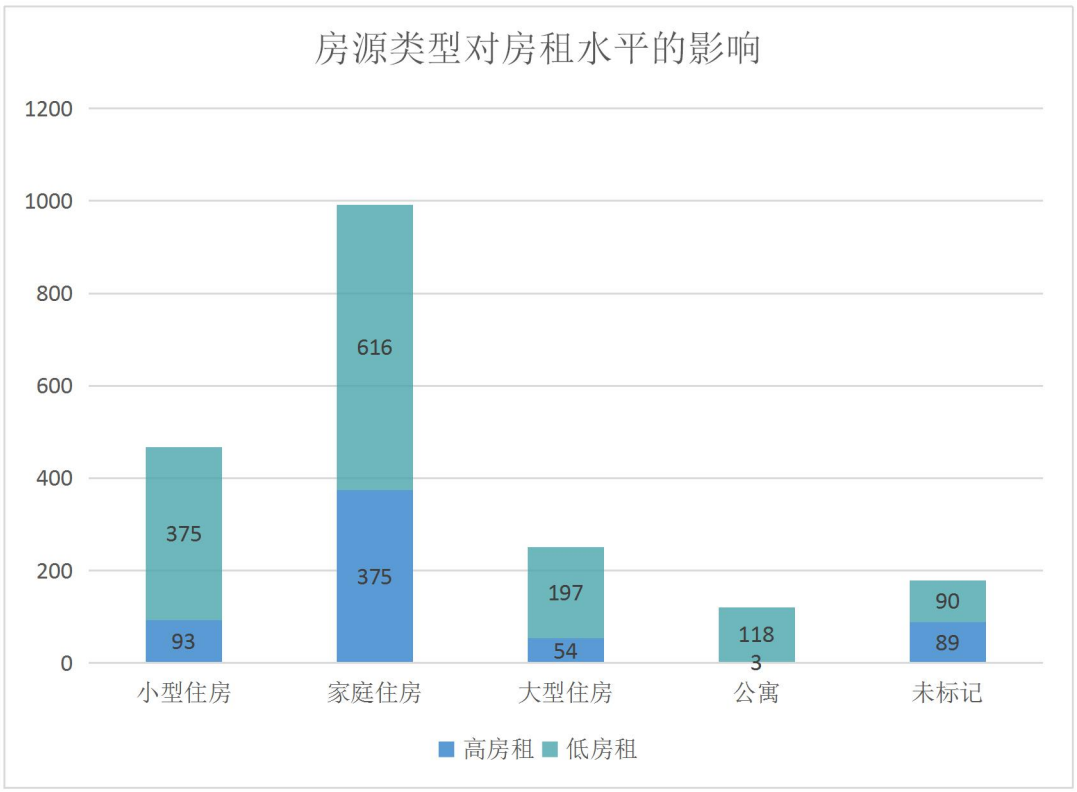
3.6 房型

显然，房屋的大小对房租的租金有较大影响，而样本中的房屋类型过多，故考虑对其进行归类处理。原始数据如下：

类型	样本量	高房租比例
1 室 0 厅	107	0.08
1 室 1 厅	359	0.23
1 室 2 厅	2	0
2 室 1 厅	497	0.30
2 室 2 厅	108	0.33
3 室 1 厅	268	0.39
3 室 2 厅	117	0.71
3 室 3 厅	1	1
4 室 0 厅	4	0
4 室 1 厅	207	0.1
4 室 2 厅	40	0.65
5 室 0 厅	7	0
5 室 1 厅	77	0
5 室 2 厅	14	0.07
6 室 1 厅	9	0.11
6 室 2 厅	6	0.17
7 室以上	8	0
未标记	179	0.49

表一——房型--房租水平表

通过查询表格与相关重庆租房资料可以发现，三室两厅的高房租比例相当之高，更深入的研究发现，3 室 2 厅的豪华装修占比相对其他户型来说较高，这在很大程度上提高了该户型的平均价格。另一方面，装修上我们将 3 室 2 厅以上的称为大户型，故根据房型的大小将其分为 1 居室的小型住房、2 至 3 居室的中型住房、4 居室的大型住房和 5 居室以上的公寓，结果如下

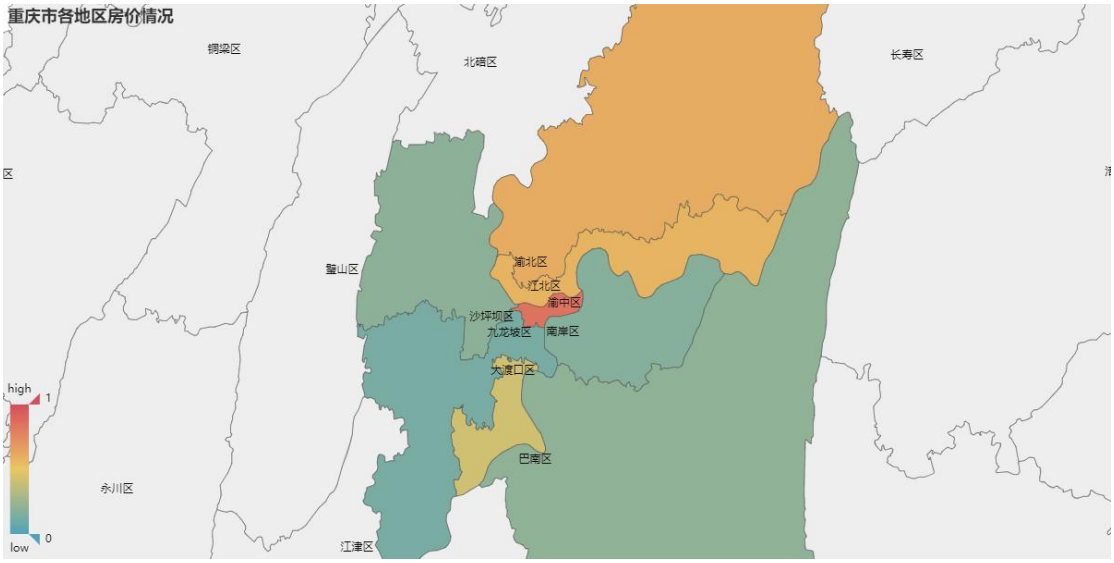


图六——房型--房租水平堆积柱状图

此柱状图在一定程度上体现了房型对租金的影响，家庭住房的高租金比例要高于单人居住的小型住房，也高于可能由多人合租的大型住房和公寓，体现了家庭租住往往能承担更高的租金。

3.7 地段

对于房源来说，最大的优势就是地段，无论是学区房或是邻近地铁都能给房子带来价值上的增长，所以我们探究了地段对于房屋价格的影响：

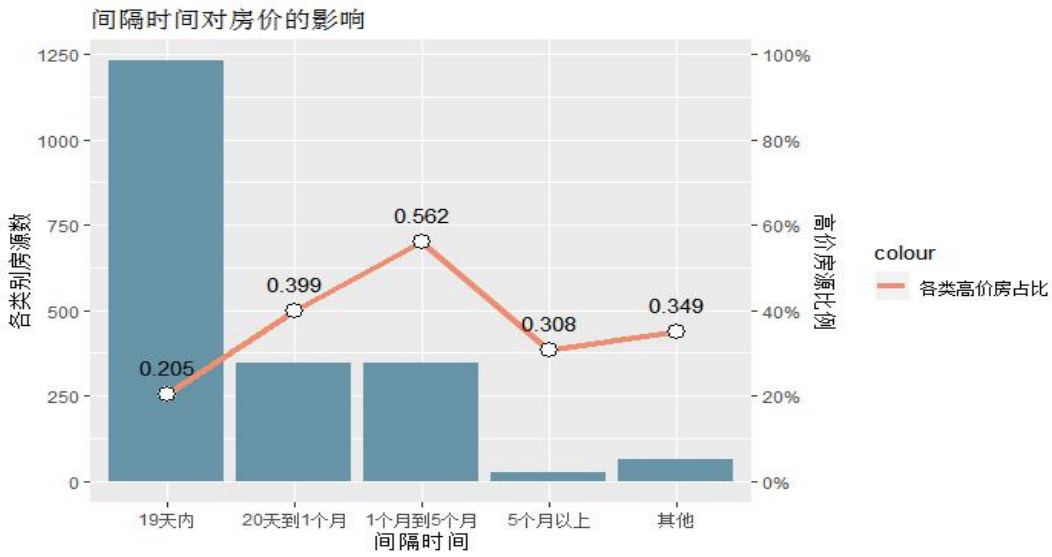


图七——地段对于房屋价格的影响

这里我们用了一个地段内部高房价房源的占比作为衡量一个地段总体放假的指标，地图中颜色越红的代表比例越高，而蓝绿色代表比例越低，从表格和图中我们都能看出来，整体而言重庆呈现了一个北方房价贵南方房价便宜的趋势，因为重庆的区域划分和长江息息相关，所以这个趋势可能与江北江南的气候和历史原因有关，北部相对而言发展更久，学校和企业的原因使得房价较高。当然和每个城市一样，“渝中”作为城市的中心，作用各大商圈（比如解放碑购物圈），寸土寸金，自然是房价最高的区域。对于大渡口地区，虽然其中高房价房源占比较周围较高，但是因为本身样本量较小，所以代表性不强。

3.8 间隔时间

因为数据里不同的房源的发布时间不同，本部分讨论发布时间对于价格的影响：

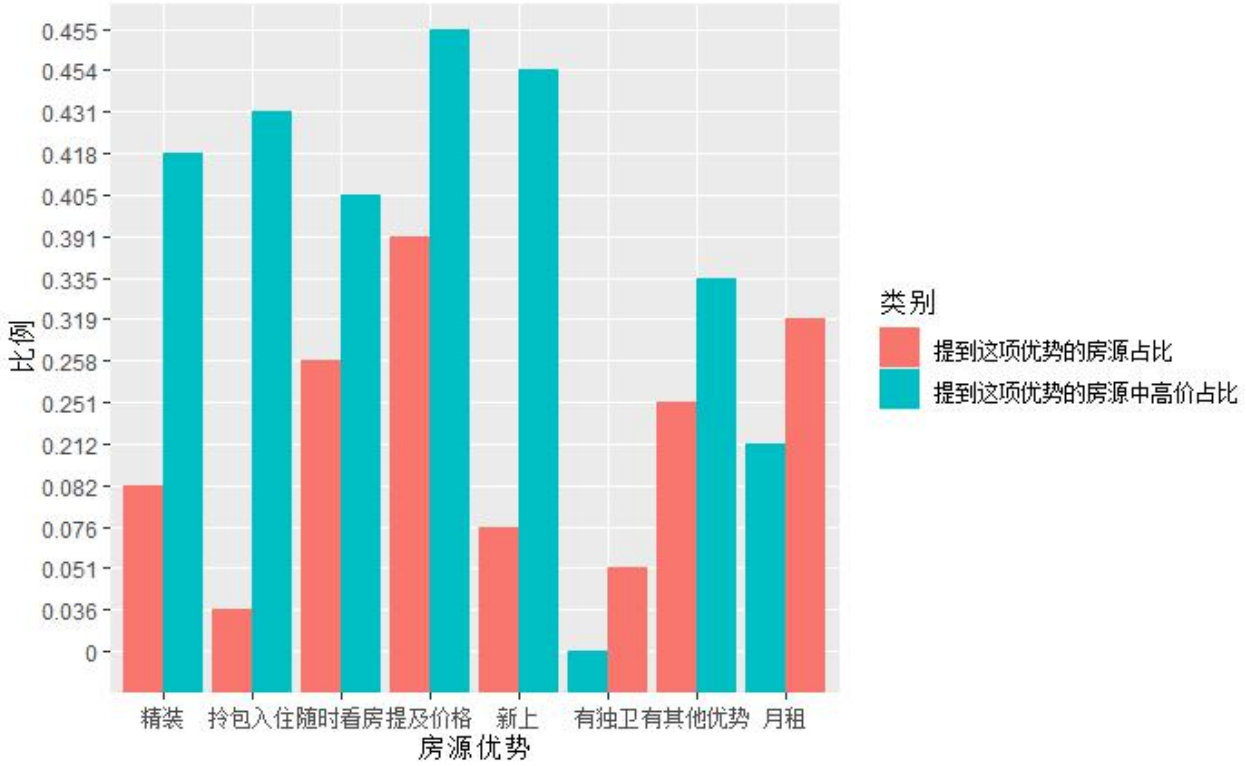


图八——间隔时间对房价的影响

从图中可以看出，随着间隔时间的上升，高价房所占的比率有一个先上升后下降的趋势，说明在我们统计本数据集的时候，最近 20 天内有很多的房子登录平台（而且最近 20 天内的房源占比最多），说明重庆的主流房源市场还是普遍低于平均水平的，新上的房源能在一定程度上代表这一段时间内的总体房价水平，而因为一些较贵的房源在租赁时需要深思熟虑，所以通常需要 1 个多月的的考察时间，但是如果时间太长，比如 5 个月以上的，几乎剩下的就只是一些因为质量确实不好的房源，高价的自然也很少。

3.9 房源优势

对于房源的几种优势，我们发现大部分房源的优势都比较分散，如独卫，拎包入住等等，但不包括地铁和中介费，因为数据集里把各种优势划分为了这两个优势和其他优势，我们故推断几乎每个房源都需要填写是否近地铁和是否有中介费，但是其他的优势的填写更自由一些，所以这些优势应当理解为是否提及。



图九——房源优势--高房租比例柱状图

红色柱状图代表所有房源里，提及到各种优势的房源占比，蓝色柱状图代表了提及到这几种优势的房源里高价房源所占的比重。在红色柱状图中比较我们可以看出，提及独卫，新上和拎包入住的房源数都很少，说明这几项优势的房源数较少或者房东在上架平台时很少提及，而其他的几种优势相对更多更普遍；而对比蓝色柱状图可以看出独卫和月租这两个优势相对而言对对于提高房价的作用有限，提及了这两种优势的房源中高价房占比都比较低，而其他的优势都能在一定程度上对提高房价有作用。

同时比较红色和蓝色柱状图我们还可以看出，新上，拎包入住，精装这三个优势其对于增加房价的作用都不错，但是其实提及到这三项优势的房源比率都较低，说明这三种优势都普遍存在于高房价房源里，不过愿意提及这些优势的房东比较少。

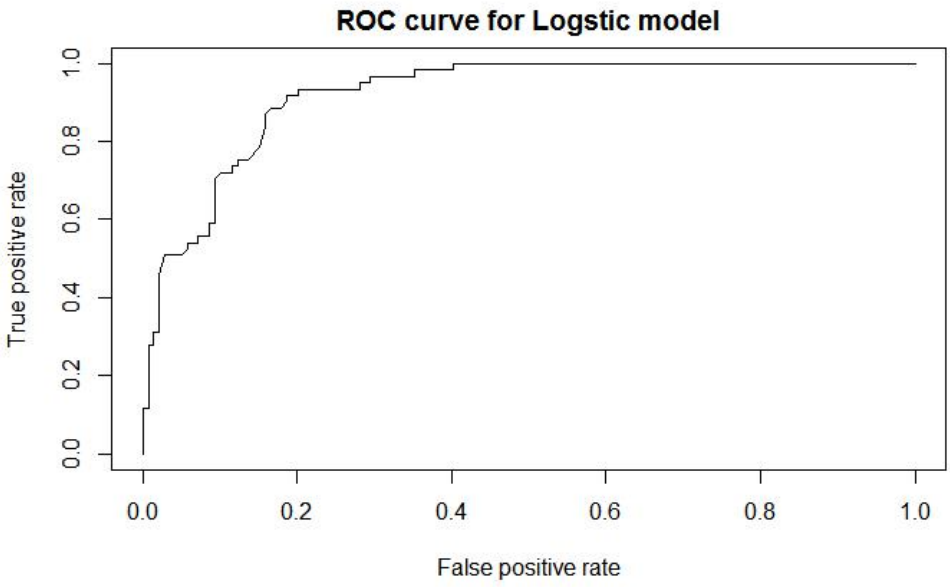
四、 模型拟合

在对数据集中各特征进行描述分析，并根据数据集对其进行分类后，我们可以对其使用不同的模型进行拟合，比较其优劣之后取得更好的拟合效果。

4.1 逻辑斯蒂建模

我们先以团购数为因变量，使用上文所述各变量对其进行线性拟合。通过十折交叉验证，从模型拟合的结果来看，在随机选取的验证集上,模型的查准率、查全率、准确率均值在 0.77，0.65，0.80 左右，绘制 ROC 曲线，计算出 AUC 值也在 0.89 左右。一些影响较显著的变量与 ROC 曲线如下：

变量名	估计值	标准误差	P 值
1 居室小型住房	-0.10	0.03	0.006
2 至 3 居室家庭住房	0.21	0.03	<0.001
4 居室大型住房	0.39	0.05	<0.001
5 居室以上大型公寓	0.32	0.06	<0.001
间隔时间 1-5 个月	0.14	0.02	<0.001
间隔时间 20 天-1 个月	0.07	0.02	0.003
租赁方式整租	0.39	0.05	<0.001
南部地区	-0.14	0.03	<0.001
其他地区	0.37	0.07	<0.001
沙坪坝地区	-0.30	0.04	<0.001
免中介费	-0.34	0.09	<0.001



图十一—逻辑斯蒂模型 ROC 曲线

通过逻辑斯蒂回归模型的拟合，我们可以发现，户型、间隔时间、租赁方式、房屋所处地区和中介费都会对房屋的租金水平起到明显影响，其中大户型、上线时间长、整租方式、位于重庆外城区的免中介费房屋更有可能获得较高的租金水平，逻辑斯蒂模型能取得不错的拟合效果。

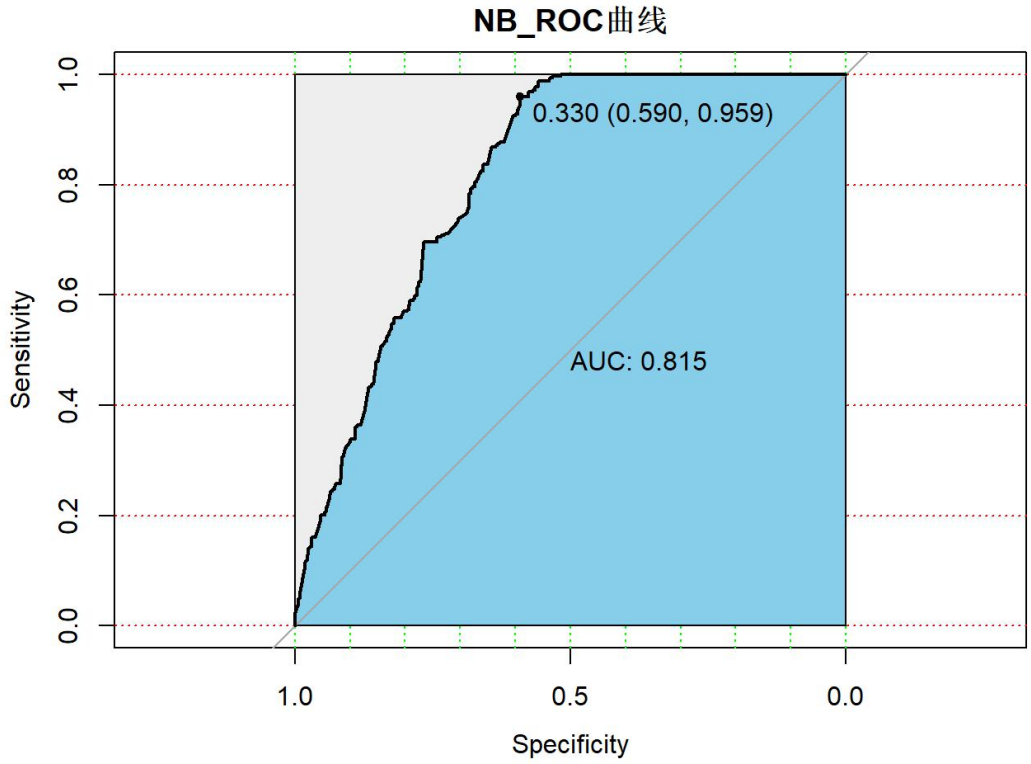
4.2 LDA 模型

使用 LDA 模型对重庆房源进行建模。若仅使用十折交叉验证，准确率仅精确到小数点后一位，小数点后第二位会浮动。因而做十次十折交叉验证后，平均准确率，可以精确到小数点后两位。

LDA 来对重庆房源的数据建模，对训练集的准确率为 82%，其 ROC 的 AUC 均值为 90%；对测试集的准确率为 81%，其 ROC 的 AUC 均值为 89%。

4.3 朴素贝叶斯模型

我们再使用朴素贝叶斯模型对数据集进行拟合，结果如下图



图十一——朴素贝叶斯模型

用朴素贝叶斯模型进行拟合时，我们发现朴素贝叶斯模型的拟合效果并不理想，AUC 值仅为 0.716，不如其他模型。

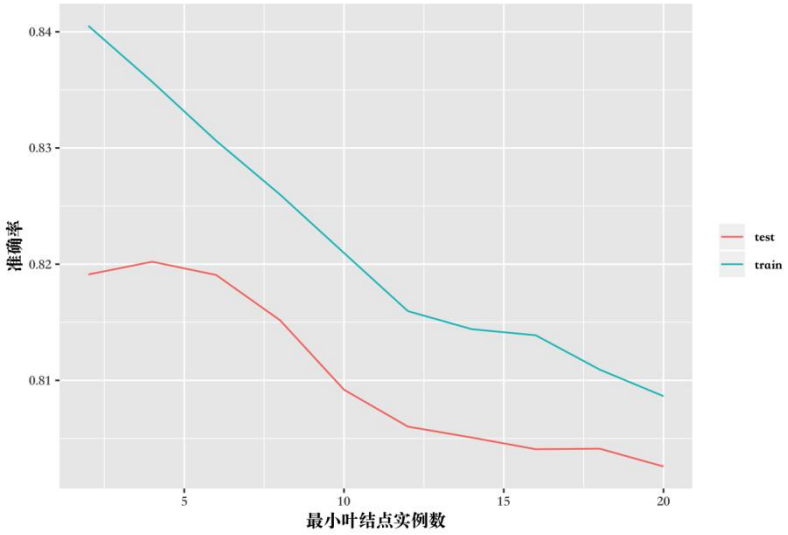
4.4 k-NN

我们使用 k-近邻算法试图对数据集进行分类，但是我们发现，当使用十折交叉验证并分别取 k 值为 1-10 时，模型 AUC 值均在 0.79 左右浮动，准确率约为 0.69，并不如其他模型理想。

4.5 决策树

我们使用 RWeka 包中的 J48 函数，C4.5 算法构建多叉决策树模型。我们对叶结点的最小实例数进行考察，观察其不同对决策树模型预测准确率的影响。做法为对每一个最小实例数值，进行 10 次十折交叉验证，并取其准确率的均值。

根据多次实验，我们发现最小叶结点实例数为 4 或者 6 的时候，测试集的准确率比较好。我们最后选择最小叶结点实例数为 6，此时测试集的准确率比较优，同时决策树的大小也比较合适。此时的决策树模型为：



图十二——最小叶结点实例数对准确率的影响

```
租赁方式 = 合租: 0 (299.0)
租赁方式 = 未提及
|   HouseT = 未标记
|   |   间隔时间 = 19天内: 0 (31.0/5.0)
|   |   间隔时间 = 1个月以上到5个月内: 1 (47.0/14.0)
|   |   间隔时间 = 20天到1个月内: 1 (24.0/7.0)
|   |   间隔时间 = 5个月以上: 0 (7.0/2.0)
|   |   间隔时间 = 其他: 0 (7.0/2.0)
|   HouseT = 一居室小型住房
|   |   是否提及价格 = 是: 0 (146.0/11.0)
|   |   是否提及价格 = 否
|   |   |   是否精装 = 是: 0 (12.0/2.0)
|   |   |   是否精装 = 否
|   |   |   |   间隔时间 = 19天内: 0 (6.0)
|   |   |   |   间隔时间 = 1个月以上到5个月内: 1 (56.0/19.0)
|   |   |   |   间隔时间 = 20天到1个月内: 1 (7.0/2.0)
|   |   |   |   间隔时间 = 5个月以上: 0 (7.0/3.0)
|   |   |   |   间隔时间 = 其他: 1 (0.0)
|   HouseT = 2至3居室家庭住房
|   |   地区 = 江北: 1 (95.0/38.0)
|   |   地区 = 九龙坡
|   |   |   是否随时看房 = 是: 1 (9.0/3.0)
|   |   |   是否随时看房 = 否: 0 (6.0/2.0)
|   |   地区 = 南部地区: 0 (55.0/21.0)
|   |   地区 = 其他: 1 (8.0/3.0)
|   |   地区 = 沙坪坝: 0 (33.0/5.0)
|   |   地区 = 渝北: 1 (139.0/51.0)
|   |   地区 = 渝中: 1 (57.0/12.0)
|   HouseT = 4居室大型住房: 1 (33.0/1.0)
|   HouseT = 5居室以上大型合租公寓: 1 (2.0)
租赁方式 = 整租
|   是否免中介费 = 否
|   |   HouseT = 未标记: 0 (13.0/2.0)
|   |   HouseT = 一居室小型住房: 0 (7.0/1.0)
|   |   HouseT = 2至3居室家庭住房: 1 (29.0/9.0)
|   |   HouseT = 4居室大型住房: 1 (2.0)
|   |   HouseT = 5居室以上大型合租公寓: 0 (0.0)
|   是否免中介费 = 是: 0 (203.0/5.0)

Number of Leaves  :  29

Size of the tree :  39
```

图十三——决策树模型

从最小叶结点实例数为 6 的决策树模型中，我们可以考察自变量的重要程度。

第一个进行分裂的变量是租赁方式（整租，合租，未标记）。

我们发现合租的房源直接被标记为房价低，这也是与常识相符合的，与其他房客合租意味着共同承担租金，因而更便宜。

对于整租的房源，第二个分裂的变量为是否减免中介费。减免中介费的房源标记为房价低，这可能是由于客观条件不甚好的房源会通过减免中介费来试图增加房源的吸引力。

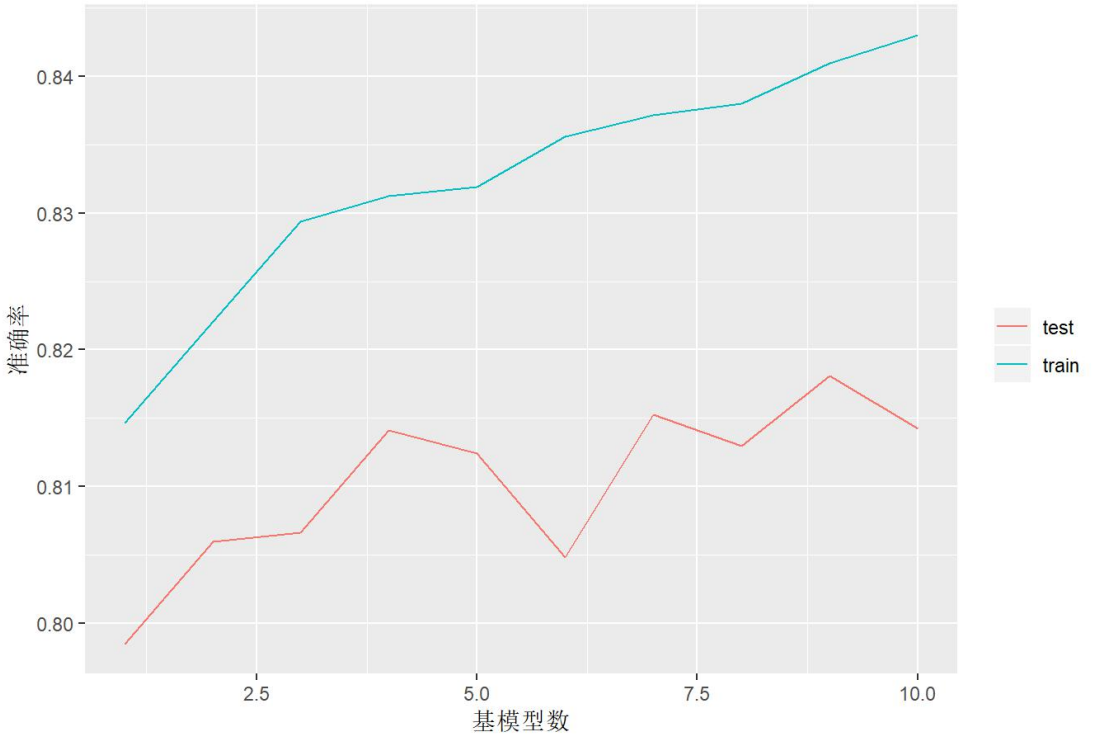
对于租赁方式未标记的房源，第二个分裂的变量为房屋大小。4 居室及以上的房源都被标记为房价高，这也是易于理解的，房子越大价格越高。2-3 居室作为主力房源，下一层分类的最主要变量为其地段，我们发现江北，渝北，渝中地区的 2-3 居室的房价较高，而南部地区，沙坪呗的房价偏低。1 居室小型住房的下一层分裂变量为是否在信息中提及价格，我们发现提及价格的房源一般是低价的，这可能是由于低价房源是有价格优势的，因而会明显地将其价格标明以增加其房源的吸引力。

另外，观察未标记租赁方式以及一居室的间隔时间叶结点，我们发现其呈现出一致的规律性：0-19 天前发布的房价偏低，20 天-5 个月内发布的房价偏高，5 个月前发布的房源少且房价偏低。5 个月还卖不出去的房源可能会通过降价来尽快出售，20 天-5 个月内发布的房源是售卖的主力。

我们使用的决策树模型的训练集准确率为 83%，测试集准确率为 82%。

4.6 Boosting

在使用 Boosting 模型的过程中，我们对算法使用基模型的数量进行考察，看不同基模型数量对模型预测准确率的影响。做法为对每一个 M（数量）值，进行 3 次五折交叉验证，并取其准确率的均值。



图十四——基模型数量对准确率的影响

可以看到我们最后基模型数为 8 时，准确率最高，我们最后取基模型数 8 位的最后模型为

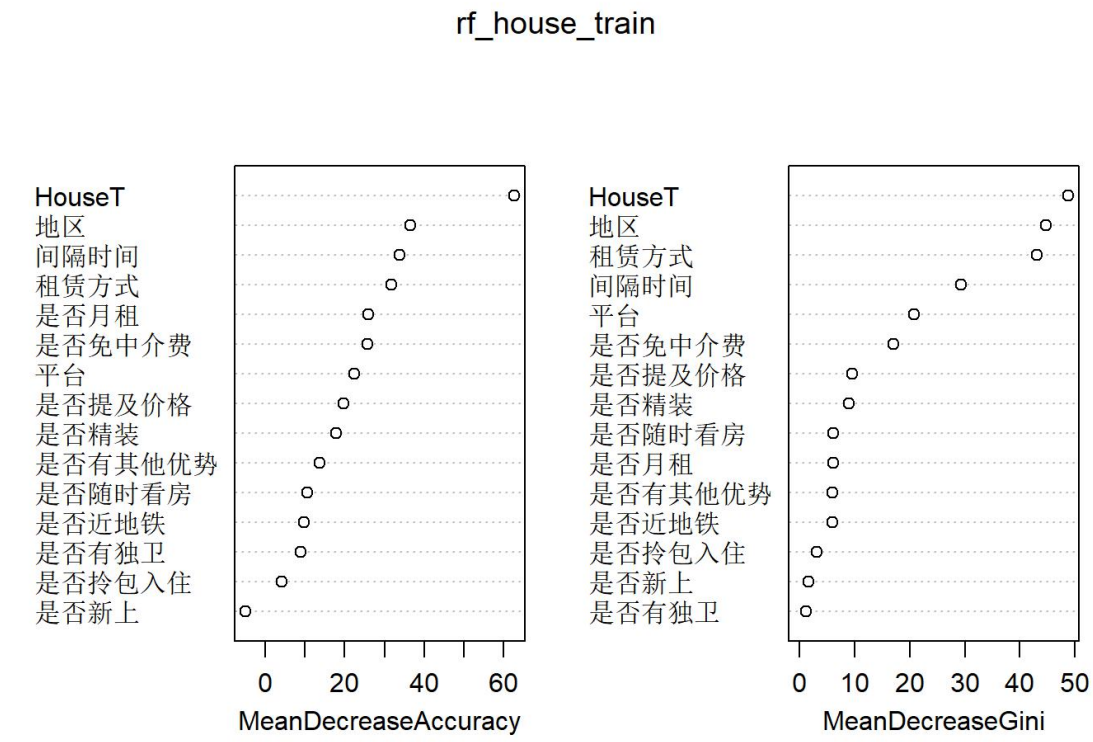
```
1) root 2010 825 0 (0.58955224 0.41044776)
  2) 租赁方式=合租 180   0 0 (1.00000000 0.00000000) *
  3) 租赁方式=未提及,整租 1830 825 0 (0.54918033 0.45081967)
    6) 平台=若航寓 93   0 0 (1.00000000 0.00000000) *
    7) 平台=链家,其他,小租乐时尚长租公寓 1737 825 0 (0.52504318 0.47495682)
     14) HouseT=未标记,一居室小型住房,2至3居室家庭住房 1705 797 0 (0.53255132 0.46744868)
       28) 地区=南部地区,沙坪坝,渝北,渝中 957 406 0 (0.57575758 0.42424242) *
       29) 地区=江北,九龙坡,其他 748 357 1 (0.47727273 0.52272727)
         58) 是否精装=是 68  25 0 (0.63235294 0.36764706) *
         59) 是否精装=否 680 314 1 (0.46176471 0.53823529)
          118) 间隔时间=其他 12   1 0 (0.91666667 0.08333333) *
          119) 间隔时间=19天内,1个月以上到5个月内,20天到1个月内,5个月以上 668 303 1 (0.45359281 0.54640719)
            238) 是否新上=否 623 291 1 (0.46709470 0.53290530)
              476) 间隔时间=19天内 189  84 0 (0.55555556 0.44444444)
                952) 平台=链家,小租乐时尚长租公寓 150  57 0 (0.62000000 0.38000000) *
                953) 平台=其他 39  12 1 (0.30769231 0.69230769) *
              477) 间隔时间=1个月以上到5个月内,20天到1个月内,5个月以上 434 186 1 (0.42857143 0.57142857) *
            239) 是否新上=是 45  12 1 (0.26666667 0.73333333) *
     15) HouseT=4居室大型住房,5居室以上大型合租公寓 32   4 1 (0.12500000 0.87500000) *
```

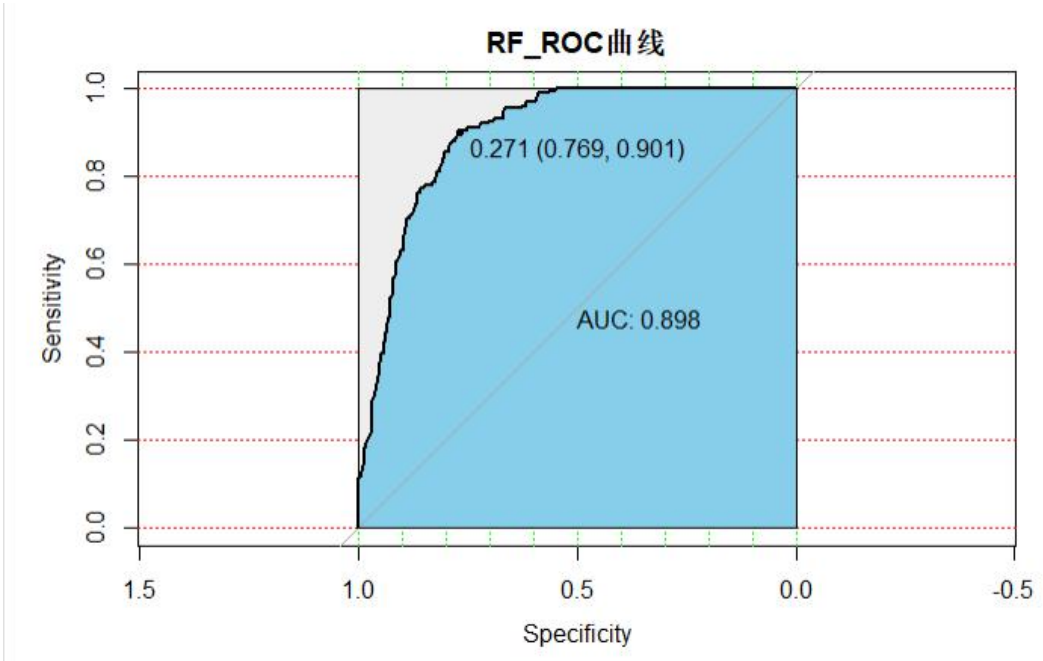
图十五——Boosting 建模结果

建模结果显示，根节点第一次分裂依据租赁方式，随后根据平台、房屋大小和所属地区进行分裂，与其余模型揭示规律相近。通过验证，我们使用的 Boosting 模型的训练集准确率为 84%，测试集准确率为 82%

4.7 随机森林

最后，我们使用随机森林的方法对数据集进行建模。结果如下：





图十六——随机森林建模结果

从随机森林建模结果可以发现，对房租水平影响最大的因素有：户型、租赁方式、房屋所处位置和发布时间。与决策树、boosting 模型形成相互印证，AUC 值为 0.898。

五、 外部资料与模型的改进方向

在对已知数据集的给定数据集进行处理时，我们发现数据集的缺失比较严重，在租赁方式和户型上都存在相当数量的缺失值。另一方面，我们下一步可以探寻房屋名字和房屋楼层对于房租水平的影响程度，一般来说楼层对于房屋的租金也有一定程度的影响。

在网上搜索重庆市房屋租赁情况时，我们发现，之所以渝中区房屋租赁价格偏高，是因为渝中区是重庆的行政、经济中心，商业的繁荣使得渝中区的写字楼更为集中，故产生了房租水平高的现象。故租房的租金与重庆区域经济发展水平也有关系。

另一方面，在房屋的特征之间仍有很多值得探寻的方面，如户型和装修的关系，地区与交通情况的关系等，也可以为房屋出租提供建议。

六、 结论

重庆受到地势局限，区域扩张成为一大问题，随着不断有人口流入，所以未来城市租房竞争会更加激烈，而且租房的人口也会越来越多。故我们小组通过研究重庆市租房市场的现状，试图发现租客与房东之间关于价格的平衡点，让房屋租赁市场更加合理。

通过调查研究，我们发现，对于房租水平来说，房屋的户型、区位、租赁方式是影响房屋租金的首要因素，位于城市中心、整租的 3-4 房一厅的家庭式住房更容易获得较高的租金，同时，通过选取合适的租房平台也能使“门当户对”的租客和房东迅速匹配。

对于租客来说，我们提出以下建议：

（1）选择商圈附近的单人公寓，或与他人合租，因为重庆城区的公共交通十分便捷，即使在商圈旁边也能迅速到达目的地，同时能省下不少房租。

（2）在租房时，根据自身经济情况选择合理、安全的租房平台，如“链家”，“若航寓”等，通过有针对性的查找选择适合自己的房屋。

对于房东来说，我们提出以下建议：

（1）选择合适的平台对自己的房屋进行出租，抓住房屋定位进行推销。

（2）尽可能多的陈述自己房屋的优势，如拎包入住、独卫等，可以增加自己的议价能力。

在建模过程中，我们在决策树、Boosting 模型和随机森林模型中取得了较好的拟合效果，AUC 值均在 0.80 上下浮动，相比起来，其他模型往往只能取得 0.70 左右的 AUC 值。在决策树、Boosting 模型和随机森林模型中，数据集的分裂方式接近，都是从户型、区位、租赁方

式等特征进行分裂，三种建模方式互相印证，使得结论更为可信。

“房子是租来的，但生活不是”，愿每一位房东和租客都能获得良好的租住体验。

七、 小组分工

李泽君：数据预处理、作图

谢炳辉：k-NN、Boosting 建模，代码整理汇总

赵雅滢：LDA、决策树建模

王维实：朴素贝叶斯、随机森林建模

钟诚：数据作图、逻辑斯蒂建模、查找资料及撰写报告

复旦大学第三交通委小组

2019.5.12