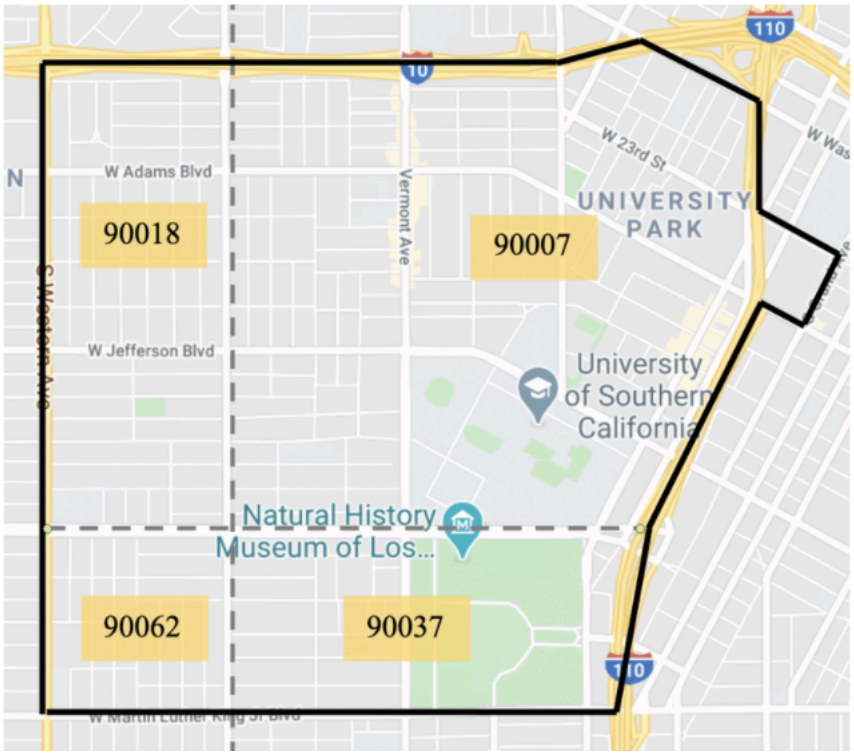
USC 周边定价模型 1.0 报告

Machine Learning

By LA Investment & Analyst

目录

1.数据概况



1.1 数据采集

上图为本次报告划分的 USC 周边地区范围，所有数据均为该范围内市场上正在出租的公寓，数据来源于 Apartments 和 Zillow，总共采集到 192 条数据，22 个特征。特征分别为房间数，卫浴数，房间大小，建造时间，是否带家具，有无 shuttle，总楼层数，品牌效应，总 unit 数等等。其中是否有暖气，车位，宠物，泳池，健身房，家具，是否在安全区内都已做了预处理，1 为有，0 为无。下面为数据特征详细介绍和示例：

Apt Name	Address	zip code	Rent	Bed	Bath	Size	Cooling/Heat	Pets	Parking	Laundry	Pool	Gym	furnished
Mardigras Apartments	720 W 27th St	90007	1600	0.5	1	400	1	0	1	share	1	0	0
Mardigras Apartments	720 W 27th St	90007	1800	0.5	1	475	1	0	1	share	1	0	0
Mardigras Apartments	720 W 27th St	90007	2200	1	1	650	1	0	1	share	1	0	0
Mardigras Apartments	720 W 27th St	90007	3200	2	2	956	1	0	1	share	1	0	0

Year Built	Within Safe	walk to USC	distance	location	Walk Score	Transit Score	Shuttle	Brand	Story	Total Units
old	1	10	0.5	N	88	82	1	0	3	59
old	1	10	0.5	N	88	82	1	0	3	59
old	1	10	0.5	N	88	82	1	0	3	59
old	1	10	0.5	N	88	82	1	0	3	59

1.2 特征介绍

Rent：每月整套单元的总租金；如出现一房间两床位的情况，则取所有床位租金的和（例如 2b2b 有 4 个床位每个床位 1000，则租金按照整套 4000 计算）

Bed：整套单元总房间数；其中 studio 为 0.5 个房间

Bath：整套单元总卫浴数

Size：整套单元面积

Cool/Heat：是否带有空调暖气；有为 1，无为 0

Zip Code：该房屋所在位置的 zip code

Pets：是否允许养宠物；允许为 1，不允许为 0

Laundry：洗衣机烘衣机在 unit 内部还是为公用；在 Unit 内为 in，公用为 share

Pool：是否带有泳池；有为 1，无为 0

Gym：是否配有健身房；有为 1，无为 0

Furnishied：是否带家具；有为 1，无为 0

Year Build：公寓老旧程度；新为 new，旧为 old

Within Safe：是否在安全区内；是为 1，否为 0

Walk to USC：步行至 USC 时间；数据来源 google map，目的地统一为 USC

Distance：与 USC 的路径距离；数据来源 google map，目的地统一为 USC

Location：公寓位于 USC 的哪个方位；如在学校北面为 N，位于学校西面为 W

Walk Score：步行分数；数据来源 zillow

Transit Score：交通便利分数；数据来源 zillow

Shuttle：是否有 Shuttle bus 往返于学校

Brand：是否有品牌效益；如 Loranzo/Gateway 大型公寓为 2，Tripalink 中小型公寓为 1，其余为 0

Story：总楼层数

Total Unit：总单元数

1.3 数据问题

- **数据量太少**。目前最主要问题是数据量太少，模型容易过拟合，误差也较大。
- 原额外考虑特征：是否包水电网煤。这方面可能天然对整个单元租金存在 $100 左右的误差影响，但是由于缺失过多遂放弃。
- 缺失值：在某些特征上数据采集比较困难，无法得知准确数字。以上特征中单元面积，总层数，总单元数均存在缺失值。缺失值可以用统计方法记性填补，但是仍会影响模型准确性。
- 建造年份仅分为 New，Old 两种情况，不够细化，应将最近新建公寓 remodel 的公寓分开不能统一归为一类。
- Distance：使用直线距离会比路径距离更有效。

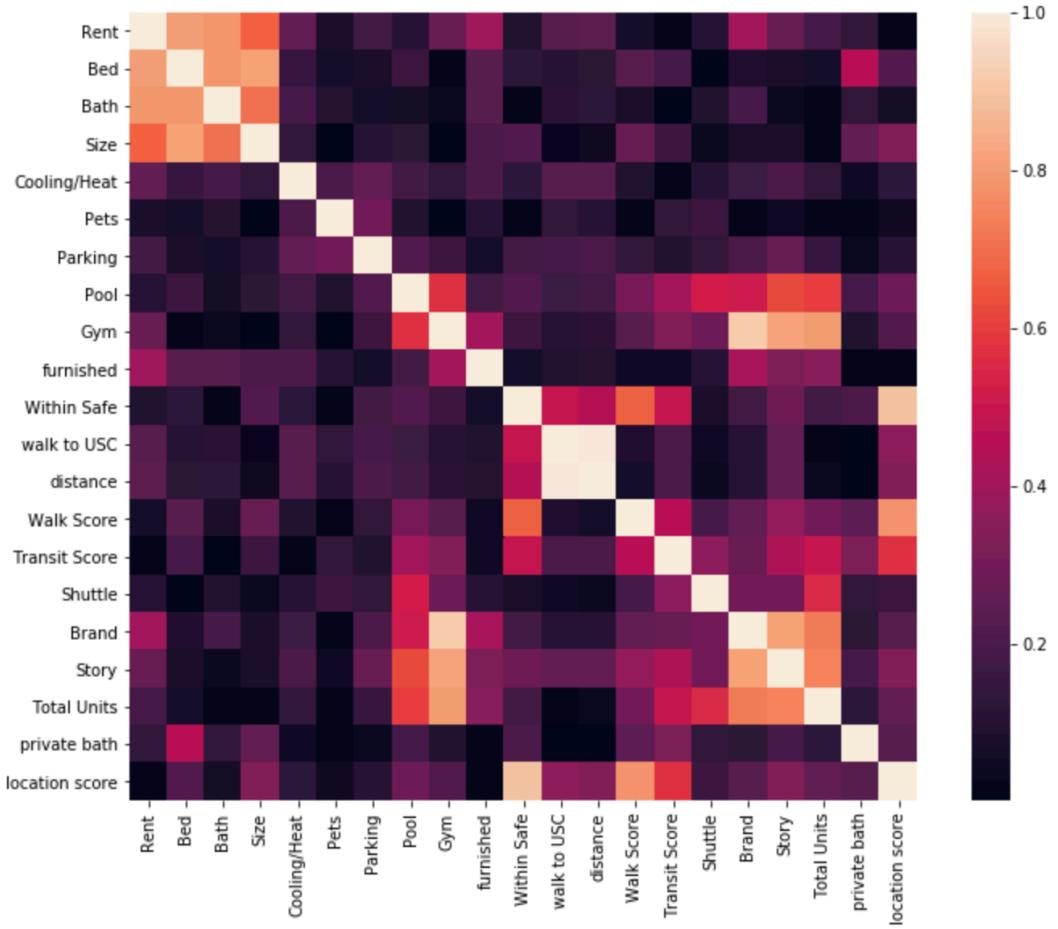
1.4 新增数据

- Private Bath：定义$Private\ Bath = \frac{Bed}{Bath}$，代表卫浴的分享程度。“1”代表独立卫浴，“<1”代表有 share，“>1”代表有独立卫浴同时又有 share 的情况。
- Location Score：结合 zip code，location，within safe 三个方面基于市场经验对 USC 周边地区位置进行打分并相加；打分表见下图，在安全区内 1 分反之 0 分。Location Score 为三者之和。

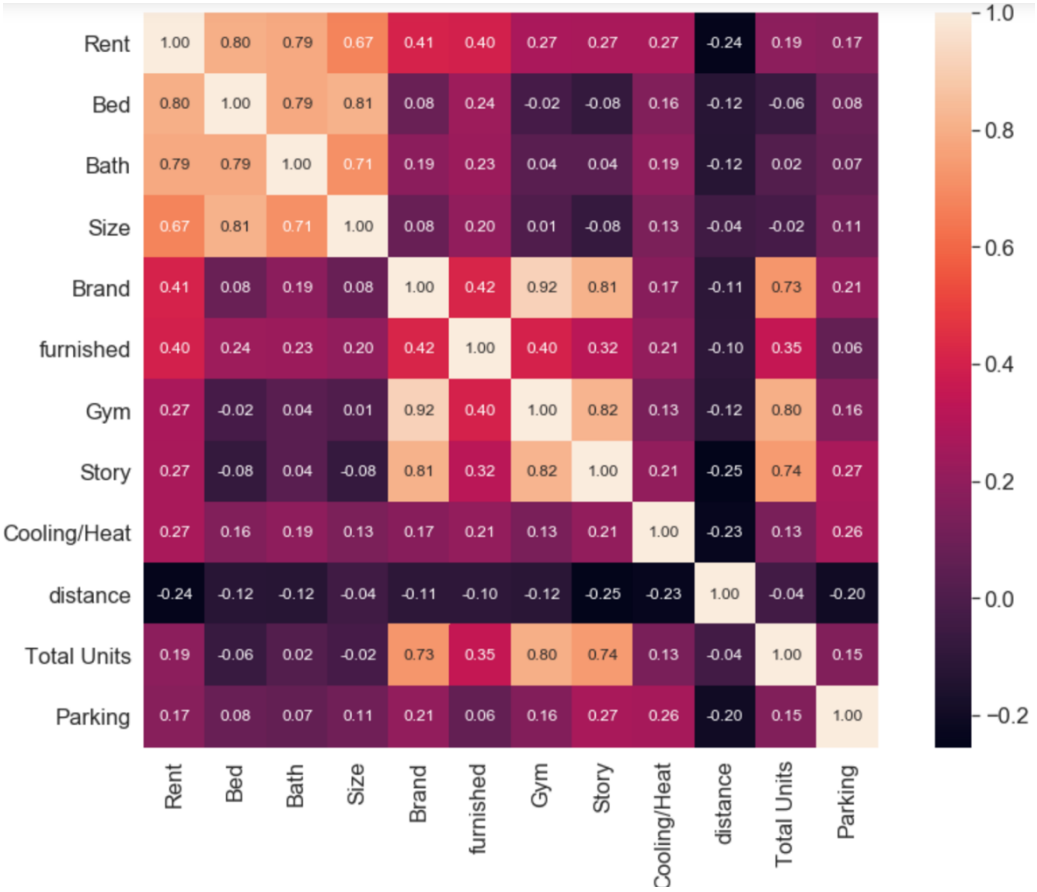
Features	Variable	Score
location	NE	6
	N	5
	SE	4
	W	3
	NW	2
	SW	1
zip code	90007	4
	90018	3
	90037	2
	90062	1

2.数据分析

2.1 数据相关性



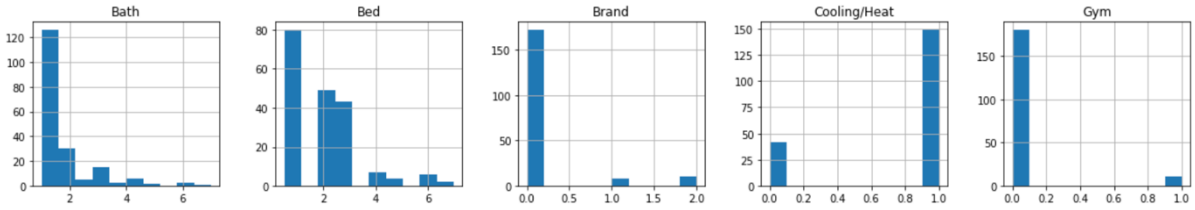
上图为所有特征之间的相关性热图，颜色由浅至深代表相关性由高到低，负相关也取绝对值考量。最上方第一行为租金与其他特征的相关性情况，我们可以明显的看出 Rent 与 Bed，Bath 和 Size 有很高的相关性，意味着房间数卫浴数越多单元面积越大，租金越高。其次与 Rent 比较相关的是 furnished 和 Brand，有带家具的公寓或者有品牌的公寓租金也比较高。除此之外我们还能看出 walk to USC 和 distance 之间是完美相关，完全可以去掉其中一个特征。还有 Brand 和 Story，Total Unit 三者直接也有很高的相关性，也可以适当去掉一两个特征。



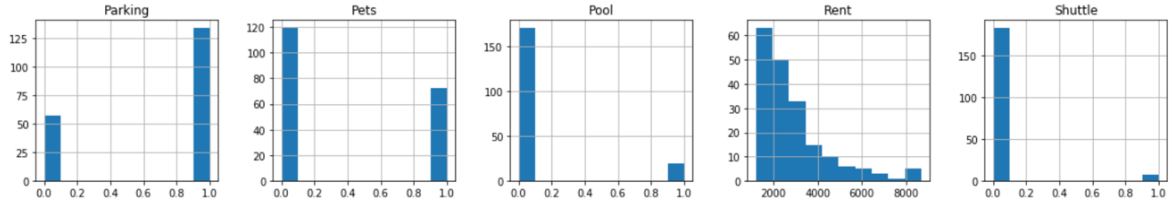
上图为与 Rent 相关性排名前十一的特征，可以清楚的看到每两个特征之间的相关系数。Rent 除了与 Bed，Bath，Size 高度相关之外，与 Brand 和 furnish 也有较高的相关性。Parking 排在第十一位。Pool 和 Gym 也和 Brand，Story，Total Unit 有较高的相关性。

2.2 数据柱状图

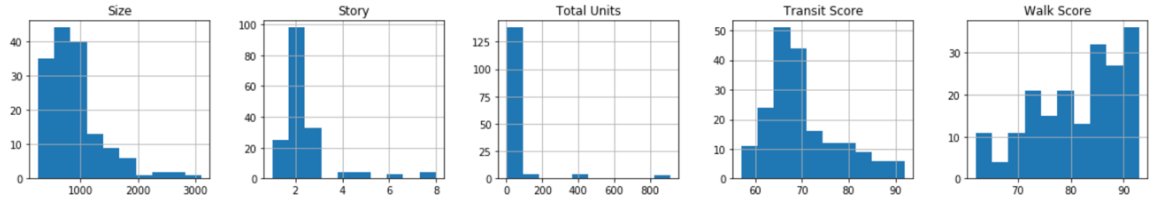
以下针对每个特征进行观察：



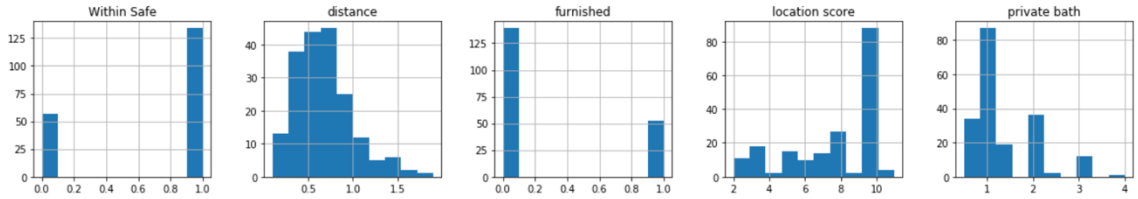
从左至右分别代表卫浴数，房间数，品牌，是否带空调暖气，是否有健身房。Bath 是以 1 个厕所为主，也反映了 studio 或者 1b1b 居多；房间数上也大体相同。品牌上大多为无品牌的为主，少量小品牌公寓和大品牌公寓。大部分公寓都带有空调暖气。极少数公寓带有健身房。



大部分公寓带有车位，不允许养宠物。极少数公寓带有游泳池和配备 shuttle bus。租金来说集中在 3000 以下，少数几个高于 8000。



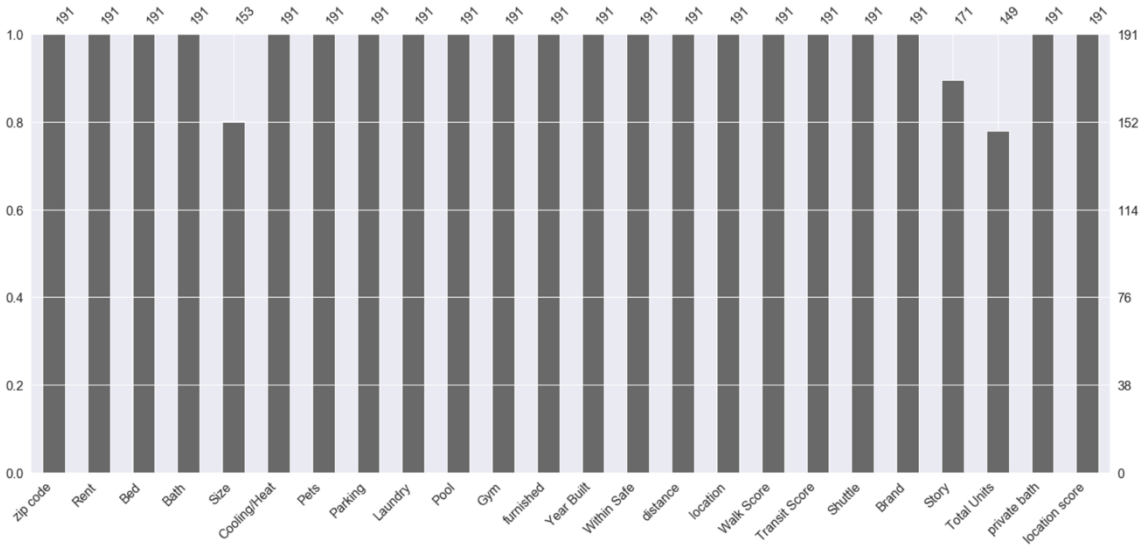
Size 集中在 1000 sqft 以内，大面积的单元不多。Story 都以 2 层为主，其次是 3 层的小公寓，4 层以上偏少，对应的是大公寓偏少。Total Unit 与 Story 类似。交通便利分数主要集中在 60-70 分这区间。步行分数则以 90 分以上最多。



大部分公寓都在安全区内。路径距离与步行时间相同，主要都在 1mile 范围内。绝倒数公寓都不带家具。地理位置分数多为 10 分。独立卫浴还是集中在 1 个，少数有房间多厕所少的情况。

3.数据处理

3.1 缺失值处理

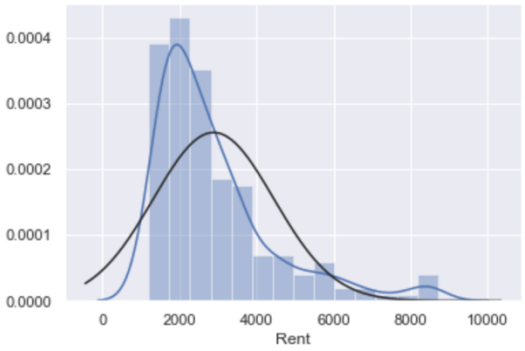


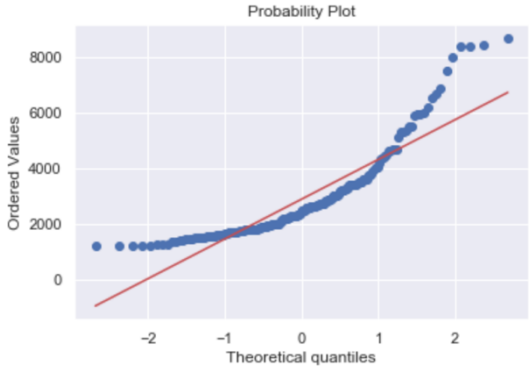
上图为所有特征缺失情况，可以明显看到在 Size，Story 和 Total Unit 都有数量不等的缺失。Total Unit 缺失最多约 22%，与 Gym，Story，Brand 相关度高，所以删去该特征。其次是 Size 缺失约 20%，尽管 Size 缺失值多，但是考虑到与 Rent 高度相关，这里按照具有相同 Bed 数量的 Unit Size 平均数(mean)进行填充。用平均数而不是中位数(median)是因为在面积上并没有什么 outlier 且用平均数模型拟合效果更好。最后是 Story 缺失约 10.5%，考虑大部分缺失楼层数据都是 house，所以取众数(mode)即 2 层填充。

3.2 类别型变量处理

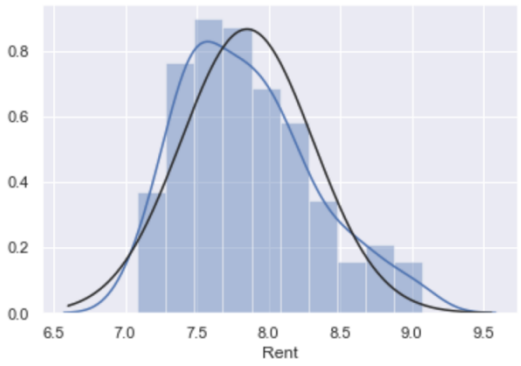
- 对于类别型变量进行了相应处理以便模型可以使用。采用的两种处理方法分别是 Label Encoder 和 One-Hot。

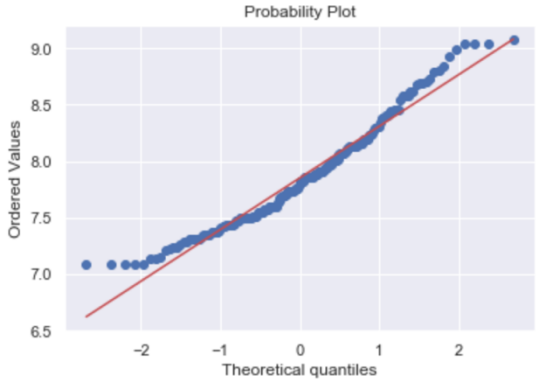
3.3 租金（Y 值因变量）平滑处理





上图可以看出 Rent 并不服从正态分布，那么在做线性回归中就会影响最后整个模型的准确性，所以在做线性回归之前对 Rent 进行了平滑处理，让调试过后的租金尽可能的服从正态分布，从而使线性模型变得有更高的准确性。下图为平滑处处理过后的 Rent 分布情况，相比之前已明显更接近正态分布。





4.建模分析

4.1 线性模型

本报告测试了三种不同的线性回归模型分别是：Linear Regression, Lasso Regression, Elastic Net。

右边三图从上至下分别是使用不同数量的重要特征情况下的平均绝对误差（Mean Absolute Error）对比；样本外准确值对比；交叉验证下准确值对比。可以明显的看到不论使用多少个重要特征，Elastic Net 模型比 Linear 和 Lasso 两个模型表现都要优秀。在使用前 20 个重要特征时，平均绝对误差值最小，样本外准确度和交叉验证时的准确度最高。

Elastic Net 模型拟合最优结果为：

MAE = 0.148751

Train Score = 0.848055

Test Score = 0.685784

CV Score = 0.752133

训练集的得分远高于测试集，模型欠拟合，并且 cross validation 得分不高，弃用此模型。

4.2 基于决策树的模型

决策树是一种机器学习的方法。决策树是一种树形结构（可以是二叉树或非二叉树），其中每个非叶节点表示一个特征属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。相比线性模型不能拟合非线性数据，决策树可以不考虑数据是否线性可分，并且可以处理带有缺失值数据，运行速度较快。本报告测试了三种基于决策树的模型，分别是 Random Forest，Gradient Boost，XGBoost。

4.2.1 RANDOM FOREST

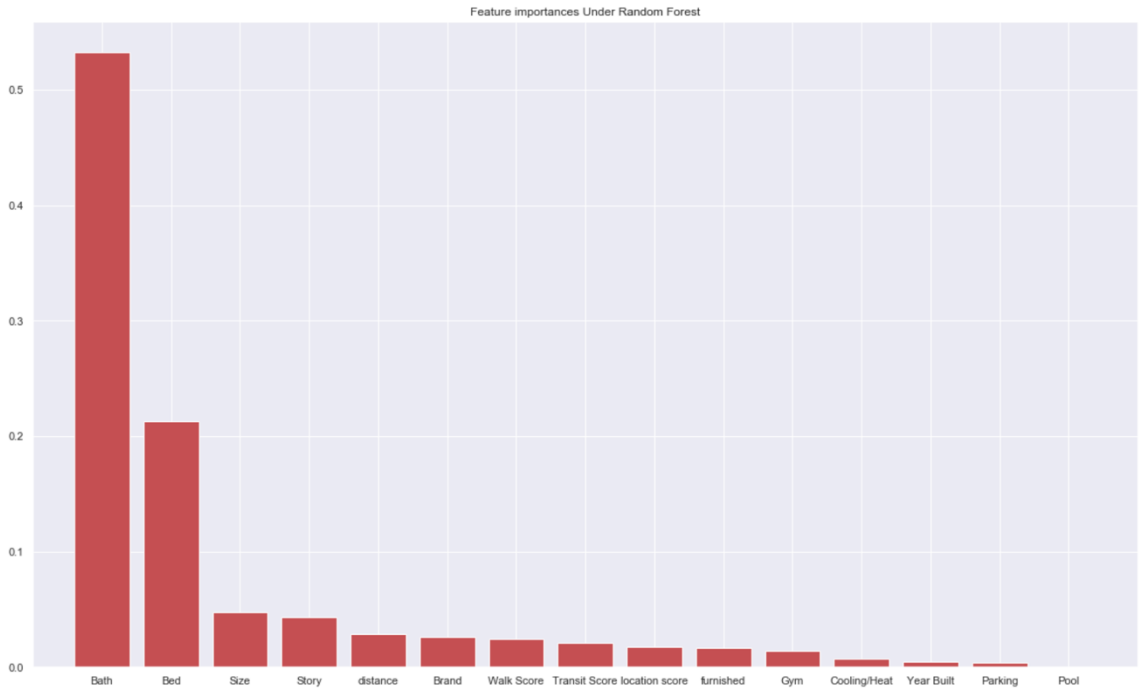
目前最优模型选用特征是：'Bed', 'Bath', 'Size', 'Cooling/Heat', 'Parking', 'Pool', 'Gym', 'furnished', 'Year Built', 'distance', 'Walk Score', 'Transit Score', 'Brand', 'Story', 'location score'。

模型拟合最优结果为：

MAE: 334.67　　Train Score: 0.960892

Cross Validation Score: 0.775924　　Test Score: 0.824243

训练集和测试集的得分说明了模型的拟合情况。目前来看测试集的得分并不高，模型明显存在过拟合的情况，cross validation 低得分也证明了这一点。MAE 代在租金预测上存在$334.67 的误差。



上图为 Random Forest 模型按照特征的重要程度绘制的柱状图。可以明显看到 Bath 有超过 50%的重要性位列第一，往后依次为 Bed，Size，Story，distance，Brand， Walk Score 等等。以上特征为模型训练后选出，存在一定随机性，可供参考。

4.2.2 GRADIENT BOOST

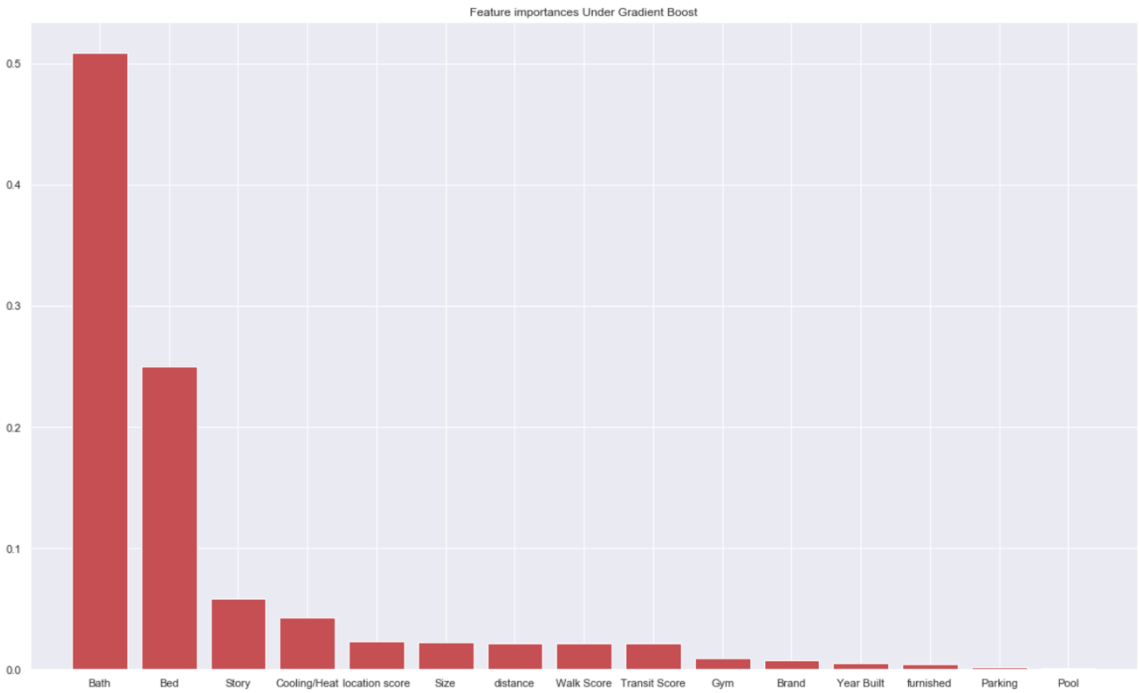
目前最优模型选用特征是：'Bed', 'Bath', 'Size', 'Cooling/Heat', 'Parking', 'Pool', 'Gym', 'furnished', 'Year Built', 'distance', 'Walk Score', 'Transit Score', 'Brand', 'Story', 'location score'。

模型拟合最优结果为：

MAE: 257.77　　Train Score: 0.991029

Cross Validation Score: 0.798001　　Test Score: 0.91847

训练集和测试集的得分都高于上一个模型，但是 cross validation 得分反应该模型仍然存在过拟合情况，还需进一步调试。MAE 代表在租金预测上存在$255.77 误差，还是比较大的，但是已小于 Random Forest 模型，目前可以先保留。



上图为 Gradient boost 模型按照特征的重要程度绘制的柱状图。与 Random Forest 相同的是 Bath 有超过 50%的重要性位列第一，往后依次为 Bed，Story，Cooling/Heat，location score 等等，排序与上一个模型略有不同。

4.2.3 XGBOOST

目前最优模型选用特征是：'Bed', 'Bath', 'Size', 'Cooling/Heat', 'Parking', 'Pool', 'Gym', 'furnished', 'Year Built', 'distance', 'Walk Score', 'Transit Score', 'Brand', 'Story'

模型拟合最优结果为：

MAE: 256.06　　Train Score: 0.985107

Cross Validation Score: 0.861382　　Test Score: 0.935802

整体表现优于前两个模型，测试集的得分都高于上两个模型，cross validation 得分也达到了 86%，相比前两个过拟合程度有所降低。MAE 代表在租金预测上存在$256.06 误差，略高于 gradient boost，可以保留。

4.3 模型融合

模型融合可以减少单个模型可能因误选使泛化性能不佳的风险，降低了模型的偏差，得到比单个模型更好的拟合结果。

选用 Gradient Boost 和 XGBoost 两个最优模型利用 voting 方法进行组合。

目前最优模型选用特征是：'Bed', 'Bath', 'Size', 'Cooling/Heat', 'Parking', 'Pool', 'Gym', 'furnished', 'Year Built', 'distance', 'Walk Score', 'Transit Score', 'Brand'

模型拟合最优结果为：

MAE: 223.29　　Train Score: 0.988987

Cross Validation Score: 0.835919　　Test Score: 0.939669

4.4 模型运用

1. 实测：随机选择市场上正在出租的一个户型，位于 1239 W 30th St, LA 90007。该户型为 2 个 bed，2.5 个 bath，面积是 1125 sqft，有带 cooling/heat 和 parking，所以两项值为 1，不带家具 furnished 为 0，Year Built 为 1，距离 USC 0.3 miles，Walk Score 为 88 分，Transit Score 为 67 分，无品牌。详情如下：

Bed	Bath	Size	Cooling/Heat	Parking	Pool	Gym	furnished	Year Built	distance	Walk Score	Transit Score	Brand
2.0	2.5	1125.0	1	1	0	0	0	1	0.3	88	67	0

模型预测整个 unit 租金为$3324.02，平均单个房间租金为整个 unit 租金除去房间数，即$1662.01 每房间。考虑到模型有$223.29 的平均绝对误差，所以整套 unit 预测租金上限为$3547.31，预测租金下限为$3100.73。单个房间预测租金在$1550.37 到$1773.66 范围内。

下图为实际情况，真实租金为$3395，平均单房间租金为$1697.5。

2 BRs　　2½ Bathrooms　　$3,395　　1　　1,125 Sq Ft

（来源 apartments.com）

整套 unit 租金预测值与真实值仅存在约$71 的误差。真实值在预测范围内。

2. 利用模型对 1664 单房间租金进行预测：

Bed	Bath	Size	Cooling/Heat	Parking	Pool	Gym	furnished	Year Built	distance	Walk Score	Transit Score	Brand
6.0	6.0	1408.00	1	1	0	0	1	1	1.0	83	60	1

6b6b 整套租金预测结果为$7154.57，通过上述方法计算后，6b6b 单房间租金定价范围在 $1155.21 到 $1229.65 之间。

5.模型问题和提升

1. 模型问题：首先当前模型的平均绝对误差还是过高，虽然是对一个单元租金进行预测，MAE 平摊到每个房间值还是相对比较高，误差还是不够小。其次，目前交叉验证得分不够高，模型尚存在过拟合的情况，主要原因是数据量太少，很容易导致模型过拟合。

2. 解决办法：尽可能多的收集数据。从多种渠道上如问卷调查或者公司内部的 leasing 收集数据。数据收集后在进一步的完善模型，可以尝试不同模型的组合找到最优模型。

3. 模型实用：若用于 Tripalink 自己建造的公寓，可以根据所需要的特征罗列一个表，根据不同地址，户型，面积查找对应的租金范围。