用KNN预测Airbnb房价的报告

胡煜彬 汪乐涛

2019年3月

前言

在这份报告里，我们尝试用KNN算法对Airbnb在华盛顿附近出租房屋的租金进行预测。

我们希望探究如何能产生最好的预测，并为为房东按市场成交价格给出对房屋租金制定的建议，为房客按照房东出价规律给出他对房价的合理期待。

方法

1. 数据预处理——胡煜彬

   我们首先把有用的数据都处理成了float格式，去掉了“$”“%”“，”等符号（见下表）

host_response_rate	87	
host_acceptance_rate	70	为空表示没有成交
host_listings_count	190	
accommodates	5	
room_type	3	
bedrooms	2	
bathrooms	2	
beds	2	
price	629	
cleaning_fee	100	
security_deposit	NaN	
minimum_nights	3	
maximum_nights	1125	
number_of_reviews	0	
latitude	38.9077	
longitude	-77.0502	
city	Washington	删除，包含于经纬度
zipcode	20037	删除，包含于经纬度
state	DC	删除，包含于经纬度

2. 评价标准（sklearn.score）200次取平均——汪乐涛

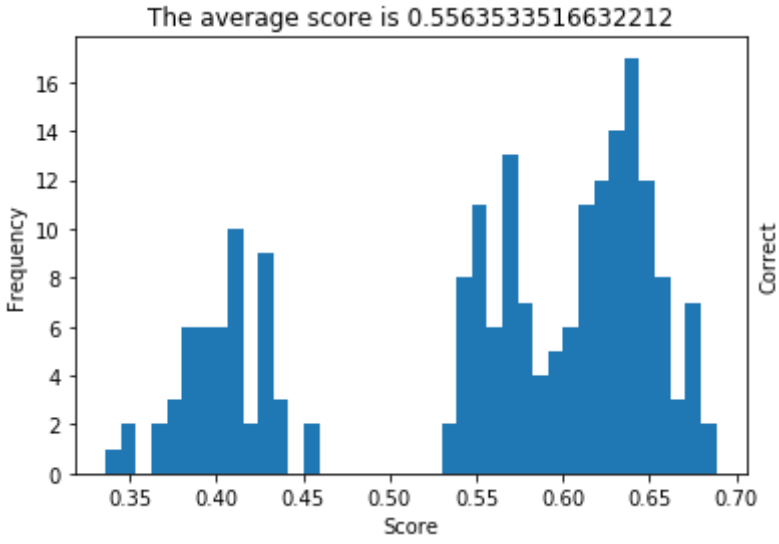
结果与讨论

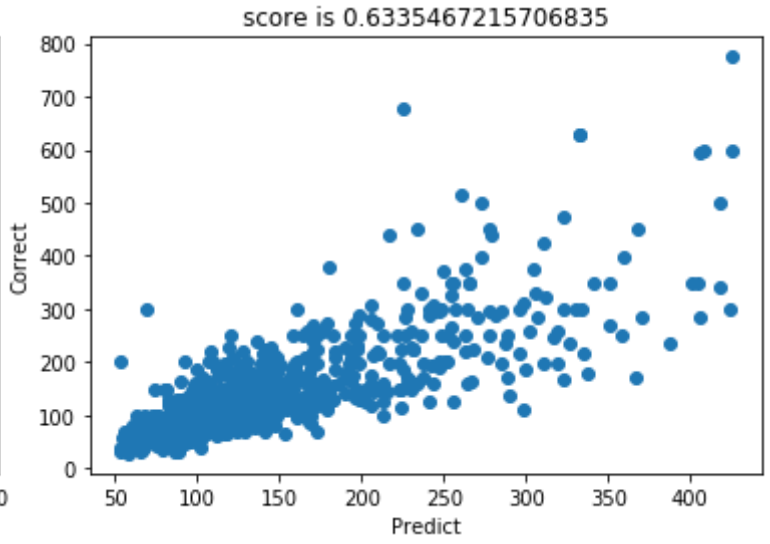
3. 数据选择探究——胡煜彬

我们认为host_acceptance_rate反映房屋是否成交。host_acceptance_rate为零的房屋信息，因为这些房间没有真正成交，不能反映市场价格。仅在对房东出价规律的研究中考虑。

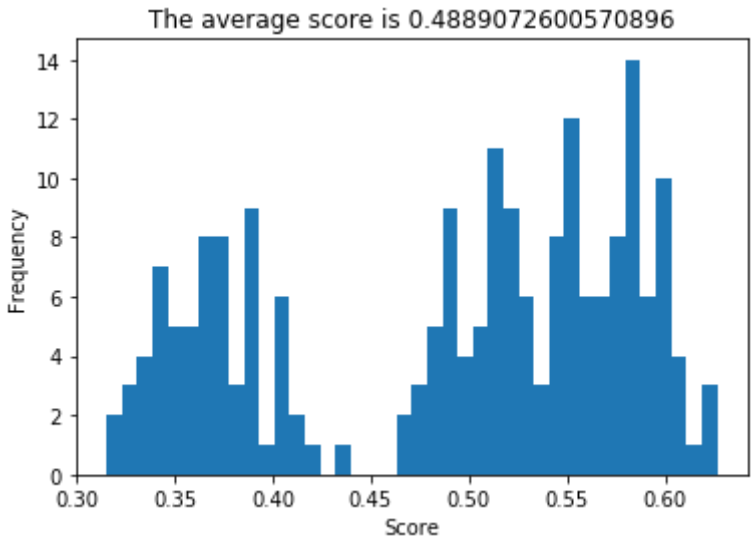
另外我们逐一探究了其他feature是否影响预测结果，产生最佳预测的选择包含以下特征：accommodates，room_type，bedrooms，bathrooms，beds，cleaning_fee，number_of_reviews，latitude，和longitude。

最佳预测200次测试的分数分布和其中一次预测的示例见下面两图：

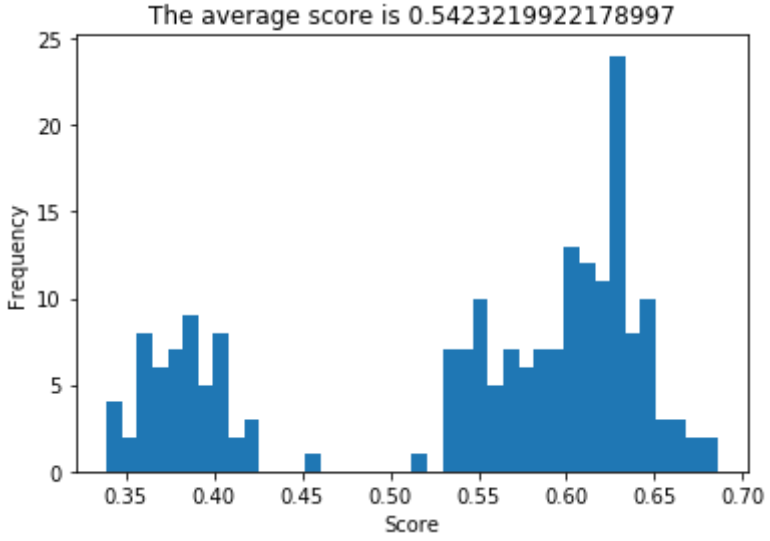




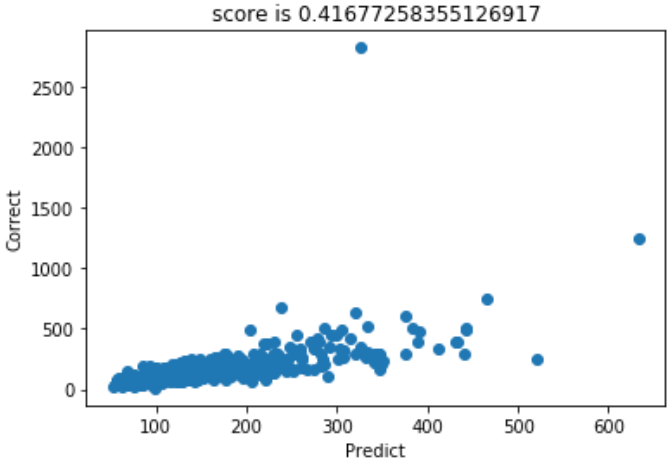
当经纬度改成距市中心距离时，分数显著降低（见下图），因此我们保留了经纬度两个特征。



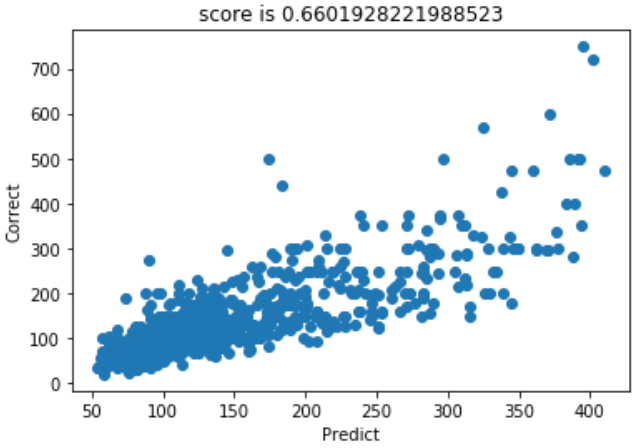
当删除了可能与居住人数相关的特征（bedrooms， bathrooms&beds）时，分数没有升高（见下图），所以我们没有只保留相互独立的特征。

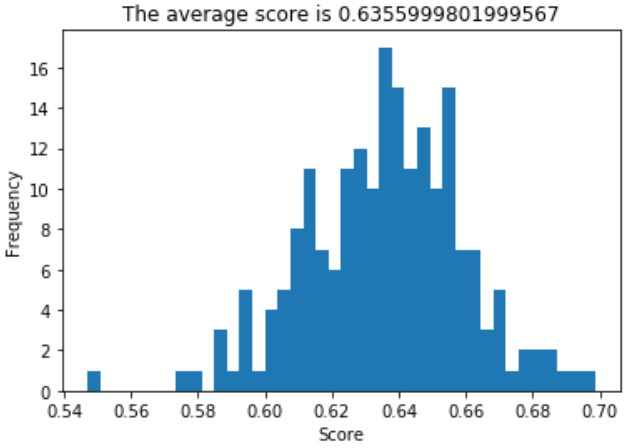


另外，我们注意到了直方图中很明显的两个峰。我推测这很可能是因为我们的预测不适合价格较高的房屋（见下图，$2822的房屋使得预测分数变得很低）。



所以我尝试添加了一个价格过滤，滤掉了所有不低于$1000的房屋（在约3700条数据中仅有20个），发现预测效果非常明显地变好（见下两图），并且直方图中左峰也消失了。这说明$1000美元以上的房屋与更便宜的房屋的价格规律不一样，而我们的模型更适合预测低于$1000美元Airbnb房屋的价格。高价房屋相比廉价房屋可能有更多数据中没有囊括特色因素，比如奢侈装修、名人居所、广阔庭院、艺术价值等等。





所以在后续的实验中我们对数据进行了选择，仅保留<$1000的房屋

4. 归一化方法研究——汪乐涛

   我们用了scikit-learn里的preprocessing包里的预设的许多用来做数据预处理的类进行的模型训练前的归一化。我选用了其中的一些类：Binarizer, FunctionTransformer, Imputer, KernelCenterer, MaxAbsScaler, MinMaxScaler, Normalizer, OneHotEncoder, QuantileTransformer，通过比较KNN模型训练后的预测准确率来选择最优秀的归一化方法。

5. 最佳预测结果——胡煜彬
6. 对买方与卖方策略的影响——汪乐涛

我把我的代码推到github上了
（最后附上我们的代码