二手车数据分析

尤俊浩(1900094810), 汪蔚涝(1900094813), 钟山(1700011623)

联系邮箱：1900094810@pku.edu.cn, 1900094813@pku.edu.cn, shanzhong@pku.edu.cn

指导教师：童云海

北京大学, 人工智能引论课程
2019-2020，春季学期

摘要

通过对Kaggle上得到的二手车数据集，进行数据分析，以找出影响价格的变量。同时进行其他的变量相关性分析。最后通过构建回归模型，试图通过测试集预测价格。

关键词：二手车，相关性分析，回归模型，价格预测，机器学习，数据智能

引言

二手车市场上的待售车呈现出售价、品牌、损耗程度、型号等多个方面上的丰富性。出于一些需求，如：对于二手车的价格和售卖情况，人们有一些先验猜测需要检验；卖方需要在挂牌前进行预测评估；规范和协调二手车市场需要对这些变量间联系的更深入的理解等等，有必要对二手车的销售情况和价值损失规律进行分析。

方法

对从Kaggle上有关二手车的数据集进行数据分析：

1. **进行数据预处理和数据清洗。**去除掉无用的维度，去除冗余排除异常点，填充空值，将德文翻译成中文增强可读性。
2. **进行数据分析，找出影响价格因素的变量，分析因子间的相关性。**为获得对数据整体的把握，我们考察了二手车市场上行驶里程和数量、行驶里程和价格、品牌和价格、车龄和价格等多组变量之间的联系。
3. **构建有关价格的回归模型。**各个回归模型的尝试，最后发现Random forest的效果最好，评分最高；对于Back fill, Front fill Drop na 和直接填充not-declared 差别不大。
4. **将项目成果进行多媒体展示**，包括海报、PPT、视频等形式。

上：车龄与数量和价格关联性分析的示意图. 价格先降后升，可能是车龄较大时车辆重新具有收藏价值。价格和数量整体呈反相关，符合价值供给关系规律。

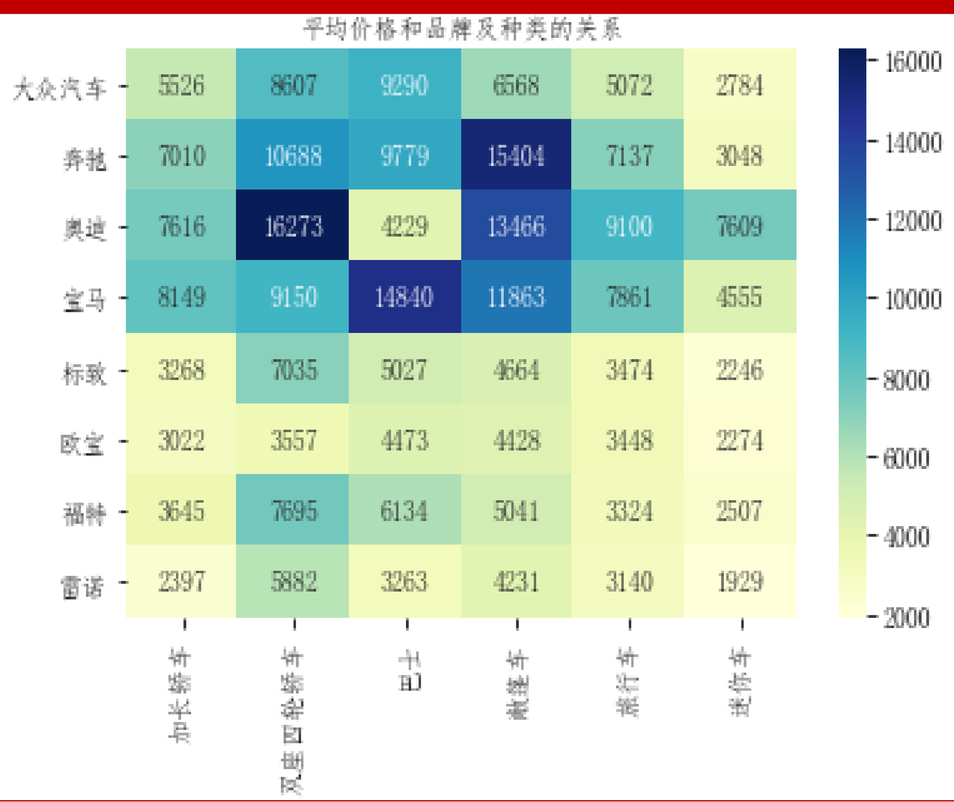
结论

二手车的价值与马力是正相关的，而与车龄和里程数是负相关的。

二手车的其他指标如燃料种类与变速箱种类也会影响其价格。

大多数车主选择在购入车子的十年后陆续转卖车辆。

物品的价值与其供给有关，越稀少的物品价值越高。



二手车的价格虽然与车龄是负相关的，但稀有的古董车却具有收藏价值，所以价格反而更高。

二手车市场的分布能够在一定程度上反应各品牌在当地的市场占有率。

左：平均价格和品牌及种类的关系的示意图. 采用热图形象化地展示三维数据（颜色深浅为第三维，表示平均价格的高低）便于观察变量间的关系。横向比较不同车型对价格的影响，纵向观察品牌对价格的影响。

反思和总结

我们意识到自己处理数据的手段还不够成熟。比如在填充空值时可以通过建模等方式降低对数据的损害、可能存在更多有意思的关系等待我们去挖掘、模型有待进一步优化等。

Python作为数据分析的主流工具集合了庞大的资源与库，使用起来非常方便，但想要用好各种工具需要很多的实践。分析过程中，及时进行可视化有助于检查操作的正确性和增加对当前数据的整体把握。

参考文献

[1] Dan—Trying to predict used car value
[2] Sentdex – Data Analysis with Python and Pandas
[3] Learning code with Chinese teacher — 5小时学会Python数据分析与展示

AIPKU