基于 Python 的数据可视化方法和系统实现

黄 琪

（诺丁汉大学 计算机科学学院，浙江 宁波 315100）

摘 要：数据可视化分析能够让人们从纷繁复杂的数据中获取有价值的信息，同时，利用机器学习方法能让人们利用已有数据，科学、合理预测未知数据。基于 Python 的数据可视化方法和机器学习进行设计，运用数据清洗和可视化等技术，对预处理后的数据进行数据集划分、特征工程、预估器流程和模型评估，利用 Scikit-learn 机器学习库和 LightGBM 库分析房价，得到房价规律。

关键词：Python；机器学习；Anaconda；Scikit-learn

中图分类号：TP311.5-4 **文献标识码：**A **文章编号：**1003-9767（2019）14-137-04

Data Visualization Method and System Implementation Based on Python

Huang Qi

(School of Computer Science, University of Nottingham, Ningbo Zhejiang 315100, China)

Abstract: Data visualization analysis can enable people to obtain valuable information from complex data. Meanwhile, machine learning can enable people to use existing data to predict unknown data scientifically and reasonably. Based on Python's data visualization method and machine learning, data cleaning and visualization technology are used to divide the pre-processed data into data sets, feature engineering, predictor flow and model evaluation. Scikit-learner machine learning library and LightGBM library are used to analyze house price and get the law of house price.

Key words: Python; machine learning; Anaconda; Scikit-learn

0 引言

随着大数据和人工智能而衍生机器学习技术，是当今人们研究的热门话题。机器学习是人工智能的一个分支，倾向于从数据中学习模式。机器学习算法包括三大类，监督学习、无监督学习和强化学习。机器学习领域，Python 占据半壁江山。Python 中的机器学习库非常多，例如 Tensorflow、LightGBM、SciPy 和 Scikit-learn 等。

监督学习（Supervised Learning）是人们常说的分类，通过已有训练样本（即已知数据及对应输出）训练得到一个最优模型（这个模型属于某个函数的集合，最优表示在某个评价准下最佳），利用这个模型将所有输入映射为相应输出，简单判断输出，实现分类，从而具有分类或预测未知数据的能力。

无监督学习（Unsupervised Learning）与监督学习的不同之处在于未事先训练任何样本，直接对数据建模。人们生活中经常用到无监督学习，比如参观一个画展，对艺术一无所知，但欣赏多幅作品后，可以将其分成不同派别。

对于强化学习（Reinforcement Learning），某些应用中，系统输出是动作序列。这种情况下，单个的动作不重要，重要的是策略，即达到目标的正确动作序列。如果一个动作是有效策略的组成部分，该动作则有效。这种情况下，机器学习程序应评估策略的优劣程度，学习以往好的动作序列，以便产生有效策略。这种学习方法称为增强学习算法。

1 相关技术介绍

1.1 Python 数据分析及可视化相关库

Numpy 是 Numerical Python 的简称，是一个 Python 科学计算的基础包。它不但能够完成科学计算任务，而且能够作为高效多维数据容器，存储和处理大型矩阵。Pandas 是基于 Numpy 的一种工具，几乎贯穿整个数据分析流程。Pandas

作者简介：黄琪（1998—），男，安徽马鞍山人，本科。研究方向：人工智能。

是 Python 的数据分析核心库，不但为时间序列分析提供支持，还提供了一系列能够快速、便捷处理结构化数据的数据结构和函数。Matplotlib 是一个用 Python 语言编写的绘图库。它基于 Numpy 的数组运算功能，可以轻易画出各种统计图形，比如散点图、柱状图、折线图等。

Seaborn 是一个基于 Matplotlib 的 Python 数据可视化库。它提供了一个高度交互式界面，用于绘制有吸引力且信息丰富的统计图标。Seaborn 在 Matplotlib 的基础上进行更高级的 API 封装，从而使制图更加简单。大多数情况下，使用 Seaborn 能绘制具有吸引力的图，使用 Matplotlib 能制作具有特色的图。人们应将 Seaborn 视为 Matplotlib 的补充，而不是替代物。Seaborn 能高度兼容 Numpy 与 Pandas 数据结构和 SciPy 与 Statsmodels 统计模式。

1.2 Scikit-learn 机器学习库

自 2007 年发布以来，Scikit-learn 已成为 Python 重要的机器学习库。Scikit-learn 简称 Sklearn，支持包括分类、回归、降维和聚类四大机器学习算法，还包含特征提取、数据处理和模型评估三大模块。

1.3 LightGBM 算法介绍

LightGBM（Light Gradient Boosting Machine）是微软亚洲研究院研发的一个开源、快速、高效基于决策树算法的提升框架（GBDT、GBRT、GBM 和 MART），被用于排序、分类、回归等多种机器学习，支持高效率的并行训练。

2 软件总体设计

从链家网站上获取近 30 万条二手房的地理位置、交易时间、交易价格、建筑类型、梯户比和交易总价格等共 23 个字段的数据。建立房屋价格分别与建筑类型、装修情况、有无电梯、有无地铁、房屋所有权是否超过 5 年以及二手房所在地区结合的箱线图、散点分布图，用户可以清楚看到不同因素影响下二手房的价格分布，为不同需求的购房者提供更直观的选择依据。

2.1 数据预处理及可视化功能

数据预处理作为任何数据分析的首要步骤，是最复杂和重要的一个环节。数据预处理包括数据合成、数据清洗、数据标准化和数据变换。数据可视化作为数据分析的成果展示，主要功能是为分析人员提供数据背后隐藏的信息和规律，为数据分析报告提供依据。

2.2 房价预测功能

房价预测功能的主要任务是模型调参和模型训练。程序需要提前为预处理后的数据建立机器学习模型，经过调参提高准确率后进行模型训练，每次训练评估训练准确率，最后得出准确率较高的模型并保存。

3 软件的实施

3.1 可视化分析以及分析报告

从 kaggle（https://www.kaggle.com/ruiqurm/lianjia）上获取到近 32 万条北京二手房房价数据中的脏数据，分析并发现脏数据，人为改正或者删除。采用 DOM 数据中的中值填补缺失值，通过 fillna() 方法 housing['DOM'] = housing['DOM'].fillna(median) 将提前获取的 median 中值填补到缺失记录，通过 split() 方法提取楼层数据，并替换原有数据。经过处理，还保留 297 367 条记录。

3.2 数据可视化

采用 Series 的 replace() 方法，事先准备替换的字典，字典的 key 是要替换的数字，value 是替换后的真实含义。将字典作为 replace() 方法的参数传入，具体实现过程如下：

```
def replace_value(data, repalce_index):
    return data.replace(repalce_index)
buildingType = {1:" 塔楼 ", 2:" 平房 ", 3:" 板楼 / 塔楼 ", 4:" 板楼 "}
housing['buildingType'] = replace_value(housing['buildingType'], buildingType)
```

本文展示 buildingType 的替换过程，其余字段替换过程与之类似。可视化准备工作基本完成后，实现数据的视化。

第一，查看 Price（每平米价格）的分布情况，具体实现如下：

```
housing['price'].hist(bins = 30)
plt.xlabel('Price')
plt.ylabel('Count')
```

北京二手房每平方米价格的分布不是一个正态分布，而是右偏分布，且价格大多数分布在 2 万 ~ 6 万，价格越高，价格下降频率越平稳，最终几乎达到 15 万，如图 1 所示。

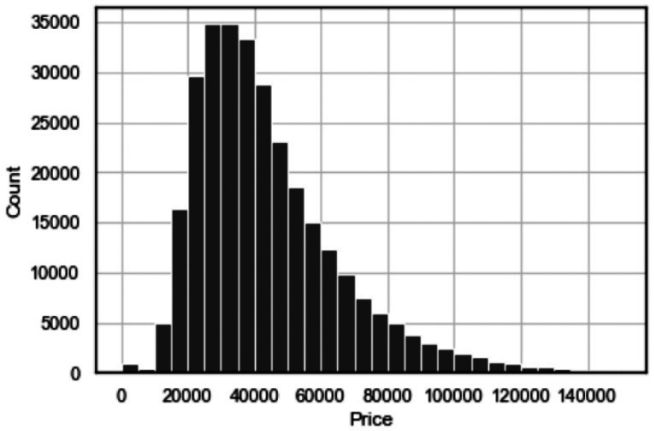

图 1 二手房每平方米的价格分布图

第二，交易时间和房屋面积的数据可视化分析。根据 tradeTime（交易年月日），将二手房价格按照年或月分组，得出每个组的平均价格，从而得到一个基于时间的二手房价

格折线图。折线图能反映二手房价格在一个时间区间的走势，对分析房价未来走势有帮助。根据 square（面积），将各种二手房的面积向下取整为 10、20、30…，将每组的平均价格和交易数量进行可视化，将不同面积分组的数据按照年份再分组，得到时间和面积双重变化下的二手房房价变化走势。使用交互性更强的 pyecharts 进行数据可视化，导入 pyecharts 包：

```
import pyecharts.options as opts
from pyecharts.charts import Bar,Line
from pyecharts.globals import ThemeType
```

第一，绘制折线图。

```
line = Line(‘init_opts: Union[pyecharts.options.global_options.InitOpts, dict] = <pyecharts.options.global_options.InitOpts object at 0x0000028AA7887C18>’)
line.add_xaxis(‘xaxis_data: Sequence’)
line.add_yaxis('series_name: str', 'y_axis: Sequence')
```

参数 init_opts 默认为 None，也可以设置折线图主题。参数 xaxis_data 需要传入一个作为图标横坐标的数据序列（Sequence）。Python 中序列有两种，元组（tuple）和列表（list）。参数 series_name 可以理解为图标的名字。参数 y_axis 和参数 xaxis_data 类似，y_axis 需要传入一个作为纵坐标的元组或列表。pyecharts 的折线图可以进行很多设置，比如设置标记线、设置横纵坐标缩放轴、设置折线图填充等。

第二，绘制柱状图。

```
bar = Bar('init_opts: Union[pyecharts.options.global_options.InitOpts, dict] = <pyecharts.options.global_options.InitOpts object at 0x0000028AA7FD55F8>')
bar.add_xaxis(xaxis_data: Sequence)
bar.add_yaxis('series_name: str', 'yaxis_data: Sequence')
```

柱状图的绘制方法参数与折线图的绘制方法参数基本相同，只是创建对象有所不同。将二手房数据根据年份分组时发现，2010 年之前的二手房交易数据都只有个位数。鉴于样本过少，去除 2010 年之前的数据，只分析 2010 年 1 月到 2018 年 1 月的数据，如图 2 所示。

2010 年，平均房价约为 1.2 万元人民币。之后七年，可以观察到平均价格大幅增长和轻微下降。2017 年 3 月和 4 月，平均价格创历史最高，每平方米 7 万多元，之后价格突然下降。2010 年 1 月到 2018 年 1 月北京市不同面积二手房的交易量柱状图如图 3 所示。

从柱状图中发现，二手房面积大于 30 m2 和小于 220 m2 的交易量比较多，面积 50 ~ 60 m2 房子的购买数量最多。2010 年 1 月到 2018 年 1 月北京市不同面积二手房交易每平方米平均价格折线图如图 4 所示。

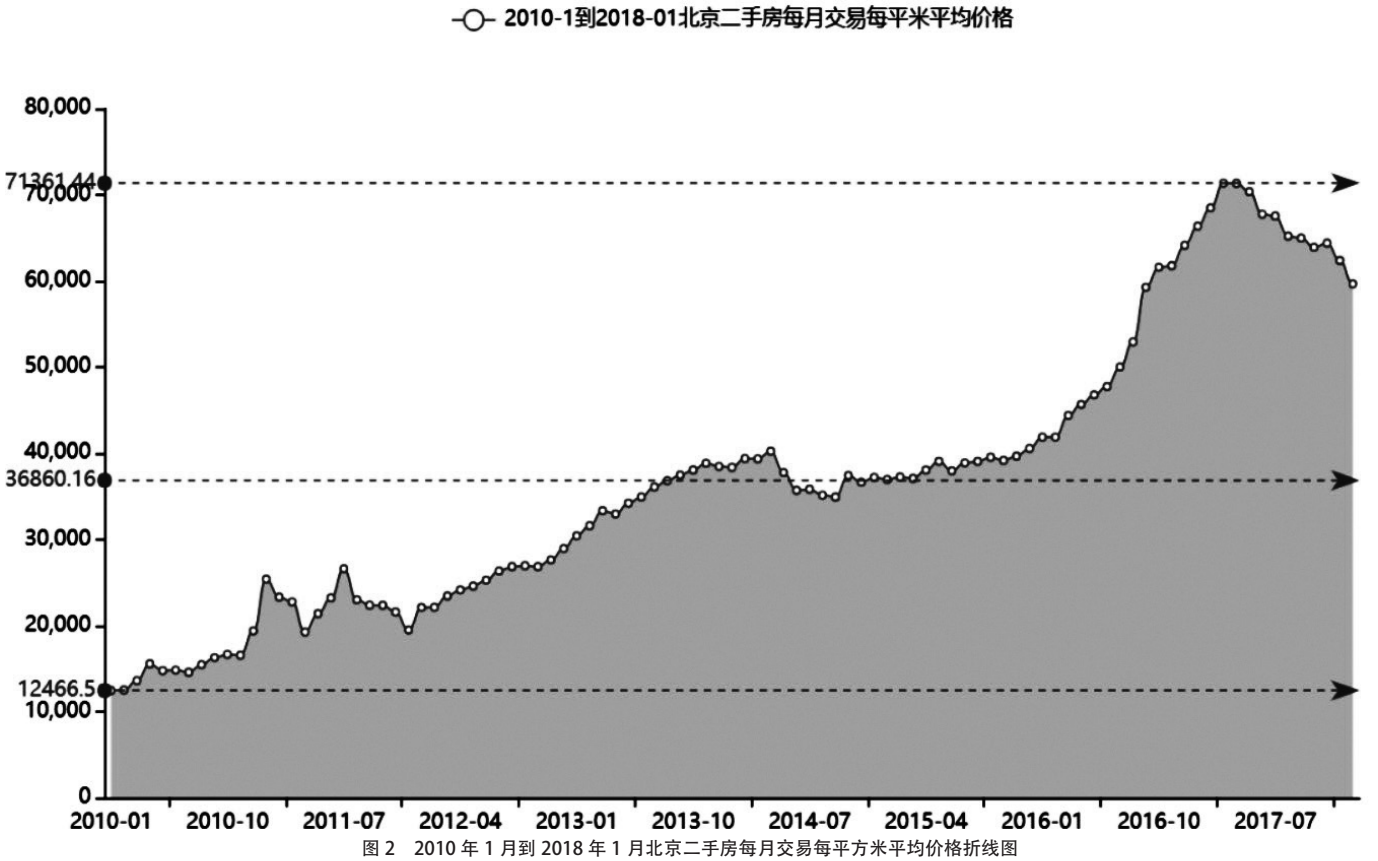

图 2 2010 年 1 月到 2018 年 1 月北京二手房每月交易每平方米平均价格折线图

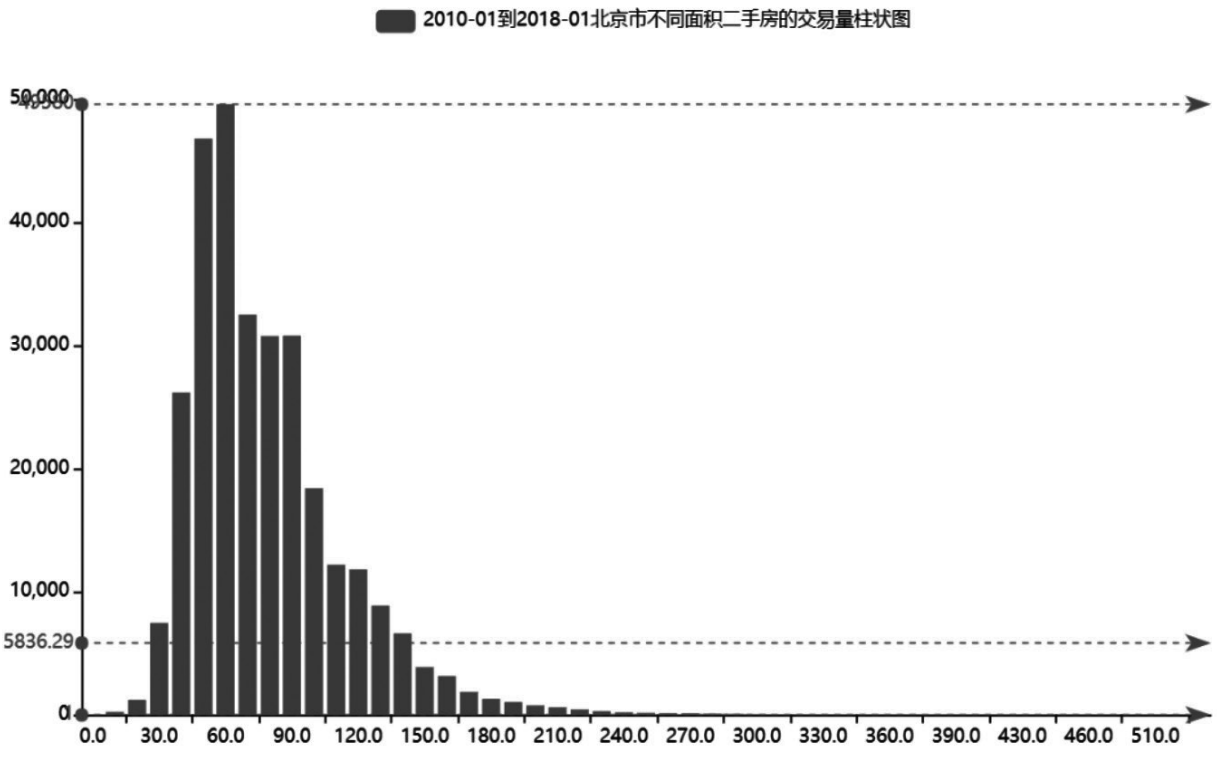

图 3　2010 年 1 月到 2018 年 1 月北京市不同面积二手房交易量柱状图

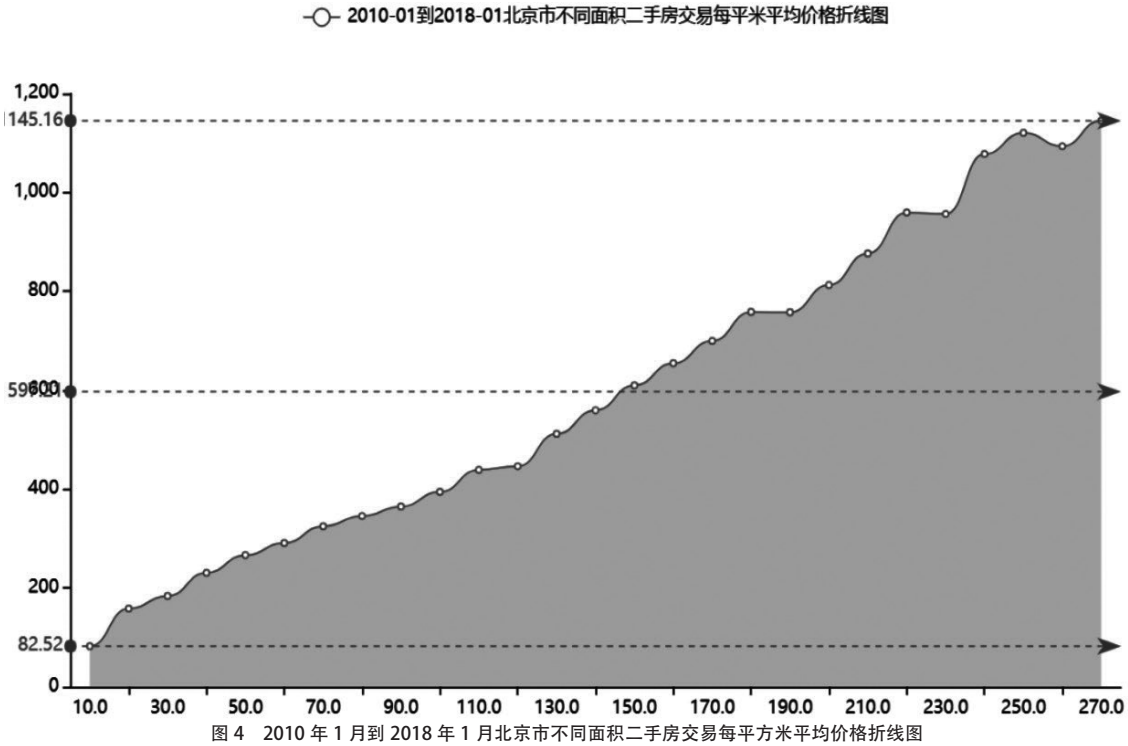

图 4　2010 年 1 月到 2018 年 1 月北京市不同面积二手房交易每平方米平均价格折线图

以上折线图基本符合正常房价走势，即面积越大的房子价格越高，50 ~ 60 m2 的房子平均价格在 280 万元左右。

4　结　语

从机器学习结果来看，预测房价在 ± 50 万以内的比例只有 44% 左右。从客观角度来说，结果不理想，但从实验整个流程来看，该结果通过系统参数调优得出，很难再通过简单参数调节提升算法准确度。要想进一步提高算法准确性，需进一步处理原数据，剔除离群数据样本，减少离群点对预测模型的干扰，从而提高算法准确性。

本软件运用 Python 强大的数据处理工具、数据可视化工具以及机器学习工具，完成北京市 30 万条二手房的房屋数据和房价数据的预处理、数据可视化，得出数据可视化分析报告，并构建房价预测模型，基本实现了既定目标，可为购房者提供一定帮助。