**Python语言程序设计**

**大作业3：数据挖掘**

1. **问题陈述**

随着代步工具的普及，“买卖车”需求激增。但对于部分预算有限的个体或家庭而言，购置一辆二手车更为明智。二手车的巨大供给需求催生了近年来日益壮大的二手车市场，但二手车的售卖面临着价格漂浮的问题。



因此，我们的目标是根据卖家或买家提供的参数信息计算价格的合理区间，帮助用户判断二手车售卖价格的高低。

1. **方法**

**（一）动机**

在前面我们了解并学习了python部分人工智能库、机器学习等方面的内容，我们决定结合实际的问题场景，在人工智能库的帮助下选取相关的简单算法和大数据计算工具，通过案例的方式，去进一步实践机器学习和数据挖掘的大致流程，在此基础上建立模型并解决实际问题。

**（二）实现**

我们所实践的案例是二手车价格评估，选用的人工智能库为Scikits-Learn。我们选取了二手车平台上不同参数下的售卖价格数据集，目的是为了根据现有数据，将之经过数据预处理、数据清洗，选择必含数据项，筛除掉无效的或者缺失的数据；通过数据脱敏，保护用户的重要个人隐私，将之以随机值替代或直接覆盖；通过可视化，直观、形象地表达出数据之间内在的联系；在人工智能库Scikits-Learn的帮助下，经过机器学习和模型拟合创建模型，设计程序，帮助用户在输入相关车辆的参数信息后，输出理想价格区间，并依照此区间对该价格高低进行判断。

1. **数据探索**

**1.1. 数据整体情况介绍**

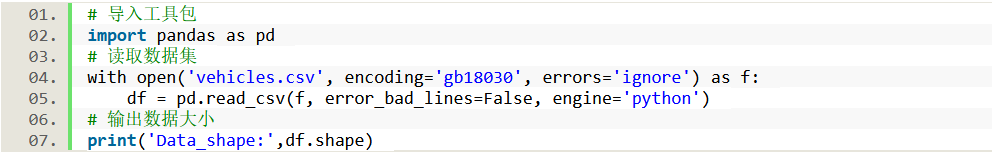
数据来源于外国某网站关于美国二手车交易记录的数据集，总数据量约为40万条，包括26列变量信息，数据集中的各字段如表1所示。

表1 二手车交易记录数据集字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Description** | **Field** | **Description** |
| **id** | 用户名 | **transmission** | 自动挡/手动挡 |
| **url** | 网址 | **VIN** | 车辆识别号码 |
| **region** | 地区 | **drive** | 驱动 |
| **region\_url** | 地区网址 | **size** | 车辆大小 |
| **price** | 售价 | **type** | 车辆类型 |
| **year** | 出厂年份 | **paint\_color** | 漆色 |
| **manufacturer** | 制造商 | **image\_url** | 虚拟网址 |
| **model** | 型号 | **description** | 二手车描述 |
| **condition** | 车辆状况 | **county** | 县 |
| **cylinders** | 汽缸数目 | **state** | 州 |
| **fuel** | 燃料种类 | **lat** | 纬度 |
| **odometer** | 行驶里程数 | **long** | 经度 |
| **title\_status** | 车辆改造经历 | **posting\_date** | 挂售时间 |

**1.2. 数据概况**

首先导入数据分析所用工具包，读取数据集，查看数据量大小。



输出结果：

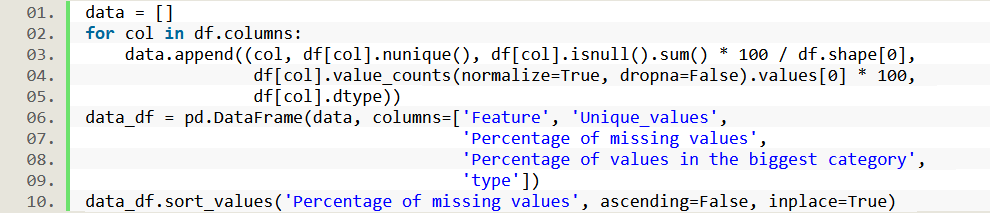
P6F3_LIUSN}Y($~`J9EL%UW

查看数据集前五行：

I63%2FZ(QI6]}T_{KO9D9CC



进一步查看数据概况：



得到数据概况如表2所示。

表2 二手车交易记录数据集数据概况

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **columns\_number** | **Feature** | **Unique\_values** | **Percentage of  missing values** | **Percentage of values in the biggest category** | **type** |
| 21 | **county** | 0 | 100 | 100 | float64 |
| 16 | **size** | 4 | 72.15472155 | 72.15472155 | object |
| 9 | **cylinders** | 8 | 40.15940159 | 40.15940159 | object |
| 8 | **condition** | 6 | 38.28938289 | 38.28938289 | object |
| 14 | **VIN** | 34773 | 36.03736037 | 36.03736037 | object |
| 15 | **drive** | 3 | 30.47430474 | 30.47430474 | object |
| 18 | **paint\_color** | 12 | 29.77629776 | 29.77629776 | object |
| 17 | **type** | 13 | 19.61119611 | 22.7402274 | object |
| 6 | **manufacturer** | 41 | 4.4200442 | 14.91914919 | object |
| 12 | **title\_status** | 6 | 2.27902279 | 95.03495035 | object |
| 11 | **odometer** | 38309 | 1.52101521 | 1.52101521 | float64 |
| 7 | **model** | 12613 | 1.28501285 | 1.4600146 | object |
| 10 | **fuel** | 5 | 0.60300603 | 82.91382914 | object |
| 23 | **lat** | 12791 | 0.53900539 | 2.03202032 | float64 |
| 24 | **long** | 12814 | 0.53900539 | 2.03202032 | float64 |
| 13 | **transmission** | 3 | 0.51400514 | 77.75577756 | object |
| 5 | **year** | 104 | 0.42800428 | 8.656086561 | float64 |
| 20 | **description** | 86284 | 0.03900039 | 0.08300083 | object |
| 19 | **image\_url** | 64480 | 0.03800038 | 1.53501535 | object |
| 25 | **posting\_date** | 94276 | 0.03800038 | 0.03800038 | object |
| 1 | **url** | 99999 | 0 | 0.00100001 | object |
| 4 | **price** | 6610 | 0 | 8.328083281 | int64 |
| 22 | **state** | 18 | 0 | 50.61450615 | object |
| 3 | **region\_url** | 84 | 0 | 2.98302983 | object |
| 2 | **region** | 84 | 0 | 2.98302983 | object |

根据整体的统计信息得知，county为无效字段（信息量为空）；size、cylinders、condition、VIN、drive、paint\_color、type字段缺失量较大，在后续分析过程中要考虑是否将其纳入影响范围；大部分数据类型为object，应通过赋值将其转化成方便建立模型的int64或float64类型。

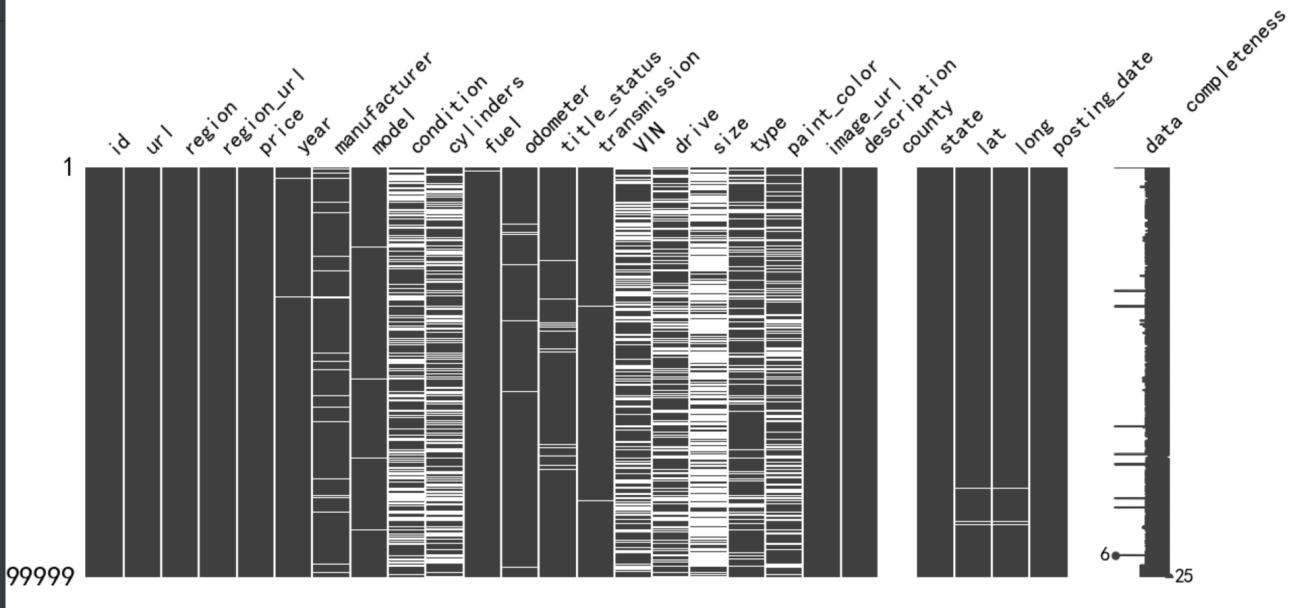
**1.3. 探索性数据分析**

**1.3.1 缺失值可视化**

在前面的数据信息统计表中发现有几个字段存在缺失值，接下来分析缺失值情况。



图1 二手车交易记录数据集缺失值矩阵图



将缺失值可视化后可以发现，size、condition、cylinders、drive等字段缺失值比较多，后续需要对这些字段缺失值进行处理。

**1.3.2 查看变量分布**

首先，查看二手车售卖价格的大致分布。

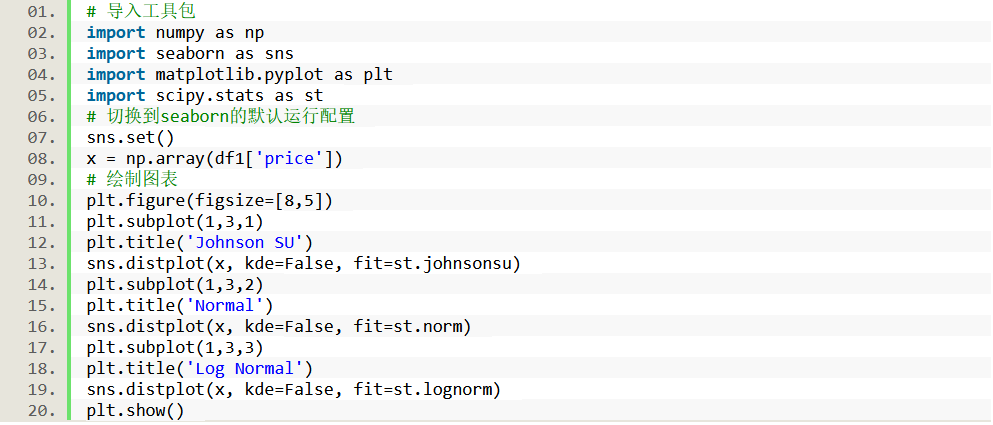
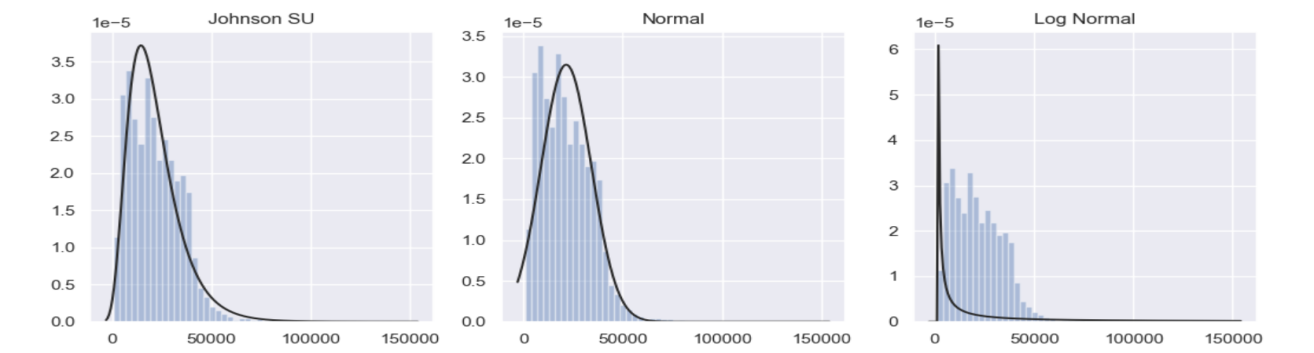


图2 二手车交易价格分布图



利用scipy模块下的johnsonsu、norm、lognorm将价格的总体分布画出来后，发现Johnson SU拟合效果较好，价格数据分布存在右偏，说明存在部分过大的极端值。需要对数据中的过大的价格值进行处理。

利用箱型图查看具体的分布划分，从箱型图中可以看出，价格大于60000则为离群值。

2V{64FFCD6%]ROI7)BLIOLE

图3 二手车交易价格箱型图



将价格大于60000的数据剔除后再重新画图，并对价格进行log取对数处理。由图可知，取对数后价格的分布相对集中，说明后续在特征工程中可对价格的数据进行log处理。

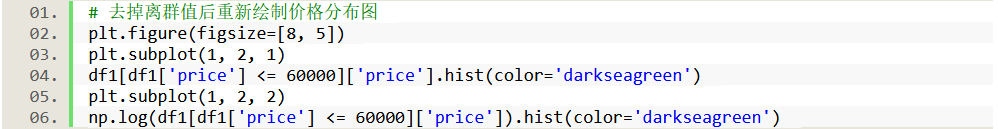
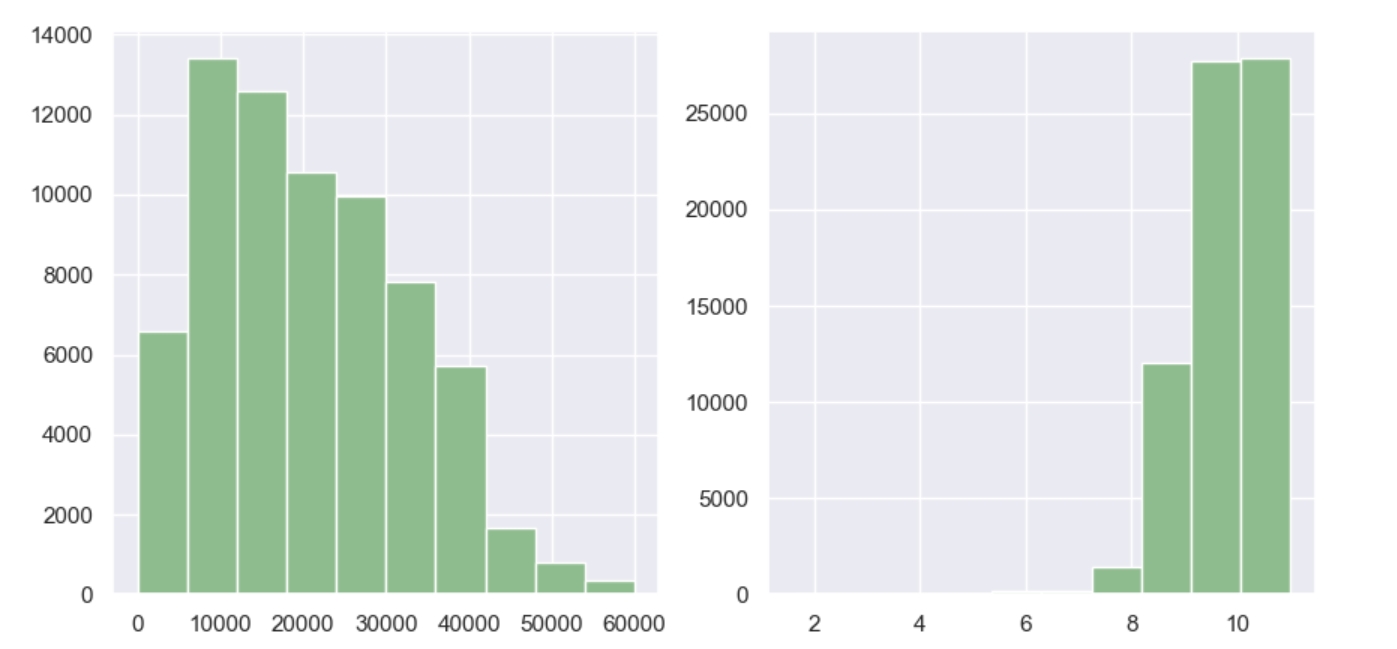


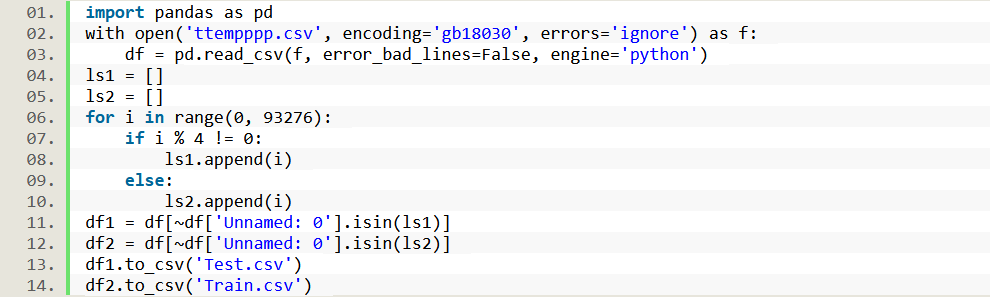
图4 二手车交易价格分布对比图



（左图为去掉离群值后，右图为左图取对数后）

1. **数据预处理**
   1. **划分测试集和训练集**

将原始数据按照 3 : 1 的比例，分成训练集和测试集，分别用来训练和测试模型。

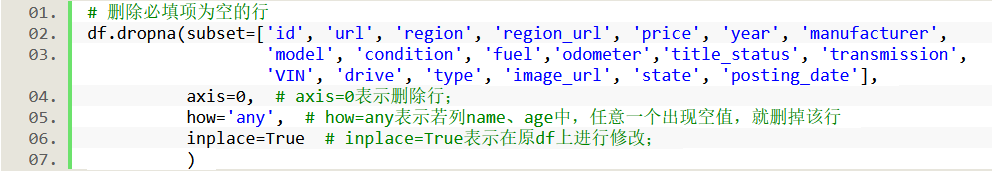


* 1. **处理缺失值**

由于数据集中部分行存在缺失值，会对后续预估分析造成影响，还有部分列字段对模型作用不大（例如description字段，因为具有主观性且篇幅较大，难以将其作为车辆价格预估的参考，lat、long对本次分析无意义，county、size缺失量过大），因此将部分字段直接删除。

AE@QY$0N(XGC9NN$RN]NV0B

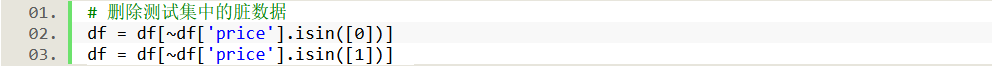
此外，有些缺失值可能会对预估模型造成影响，故设置一部分列为必填项，将必填项为空的行删除。



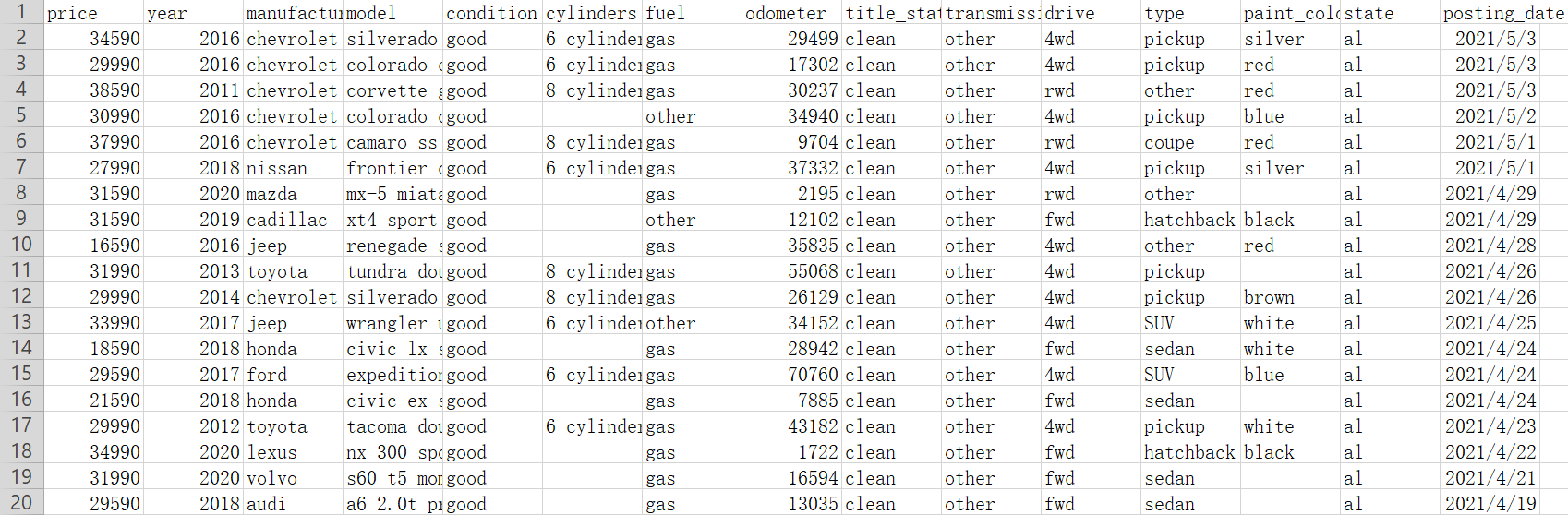
* 1. **数据清洗**

根据上面的分析，数据集中存在一些“脏数据”（脏数据是指不符合现实逻辑，且会对模型预测效果产生干扰的数据，脏数据的存在使得数据集的质量下降），因此要对数据集进行数据清洗的工作。

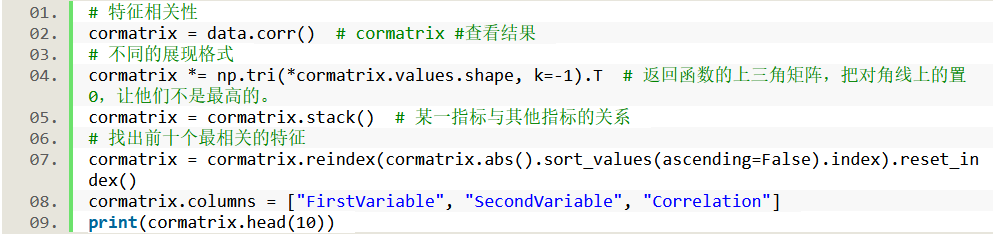
本数据集的price字段表示的是二手车的交易价格，应该是一个大于零的数，如果出现了price小于零或其他异常值的数据记录，则断定此数据是脏数据。

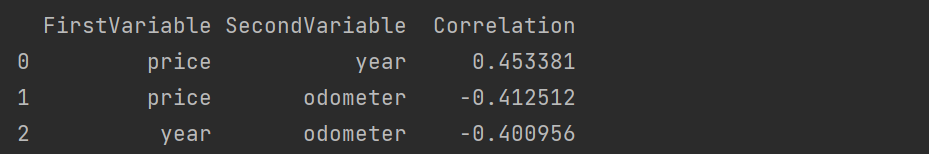


完成后部分数据如图所示。



1. **查看特征相关性**
   1. **相关性计算**





由图可以看出，特征字段之间相关性较低，这是由于大部分字段为object类型，故为了后续分析建立模型，为object类型特征量进行赋值处理。

* 1. **热力图展示**

对部分object类型数据进行赋值转换类型处理后，为了更直观地展示数据之间的相关性，绘制相关性热力图。

从相关性热度图表中，我们可以看到不同定距数据之间的相关性大小。从中可以挑选出与价格相关性较大的特征字段，剔除相关性为0的特征字段。此外后续的回归预测中还需要解决共线性特征。

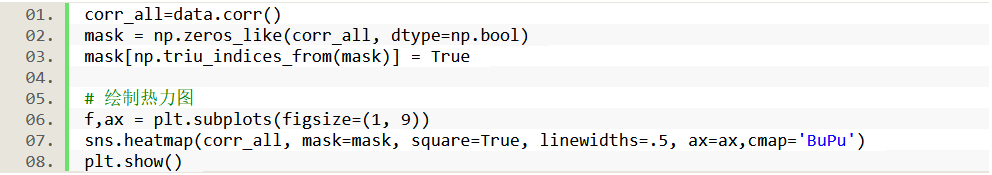
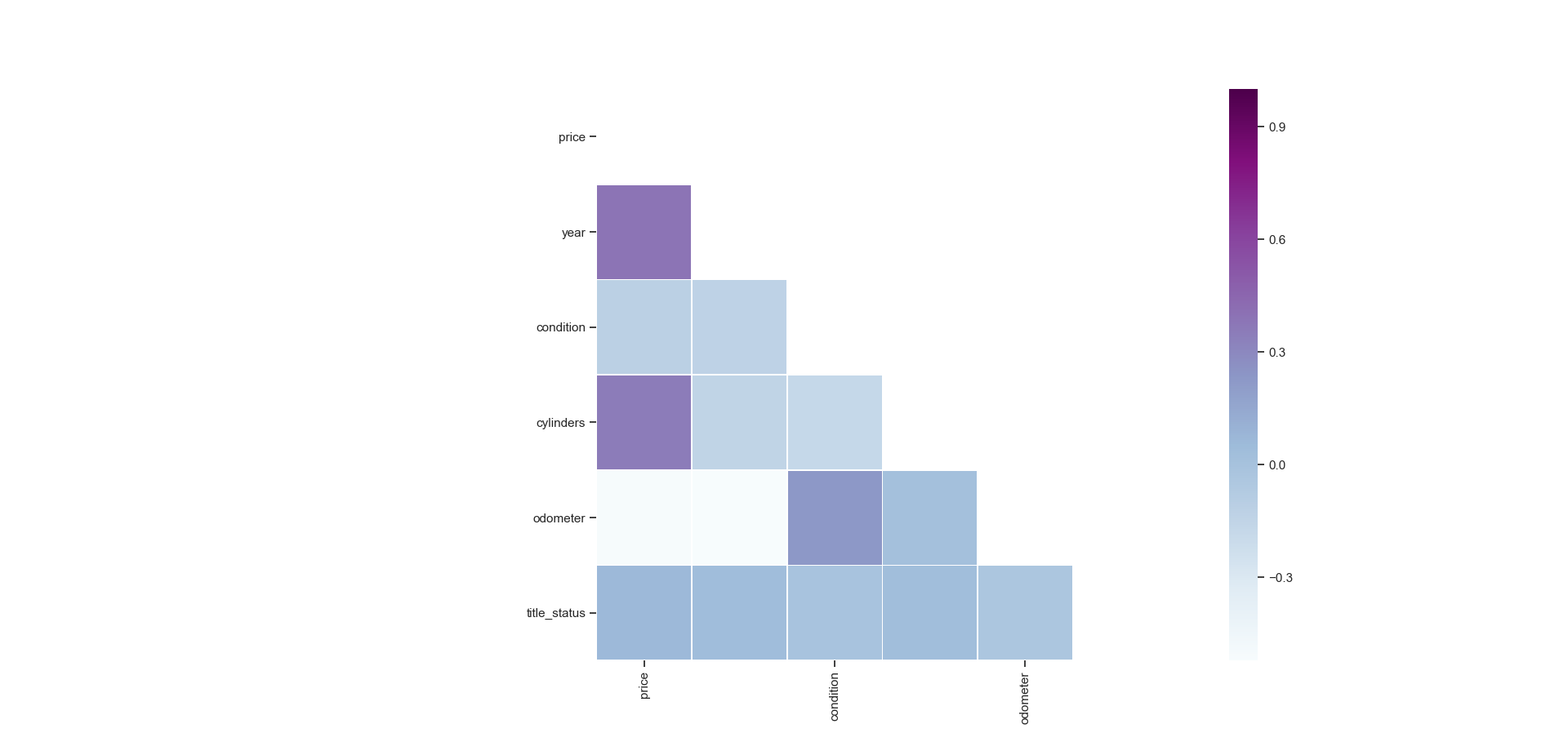


图5 部分字段相关性热力图



由图可以看出，price与year、cylinders、status的相关性较大。

* 1. **查看定类数据相关性**

**3.3.1. 分析年份与价格的关系**

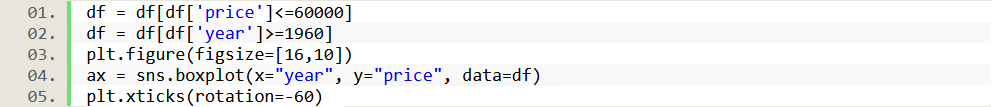
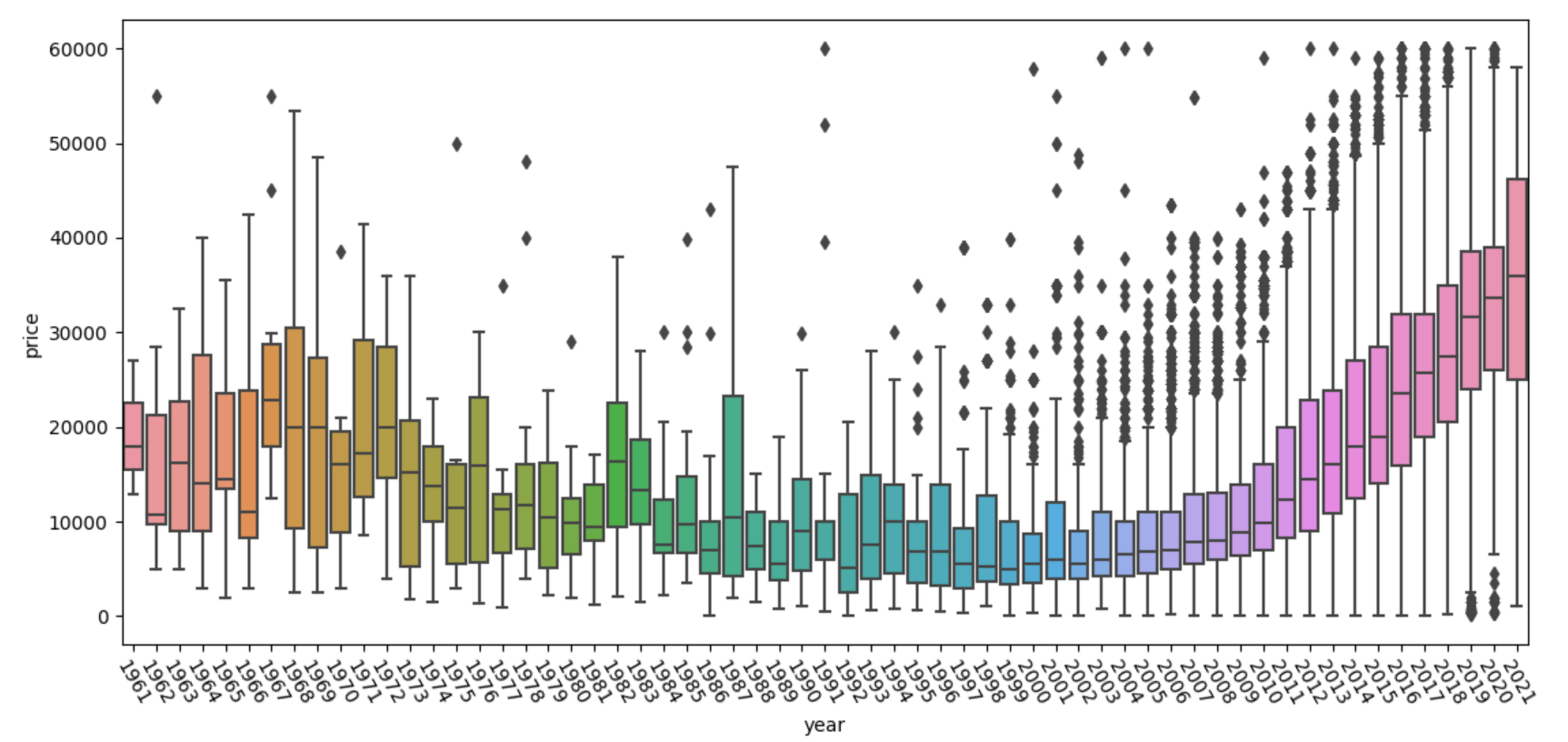


图6 不同汽车注册年份的价格分布箱型图



分析不同汽车注册年份的价格分布箱型图，上世纪80年代前由于数据分布较零散且数据量较小，不具有太大参考价值；而注册年份越往后，价格的跨度越来越大，推测二手车的车型越来越多，且价格有上升的趋势。

**3.3.2. 分析不同定类变量与价格之间的关系**

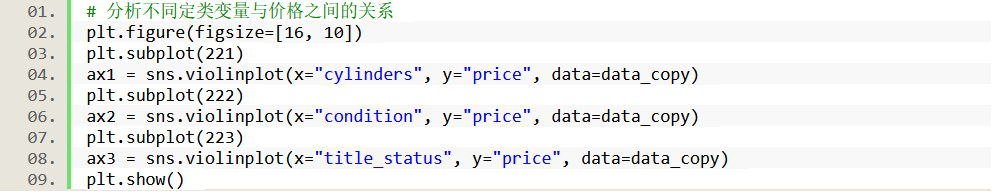
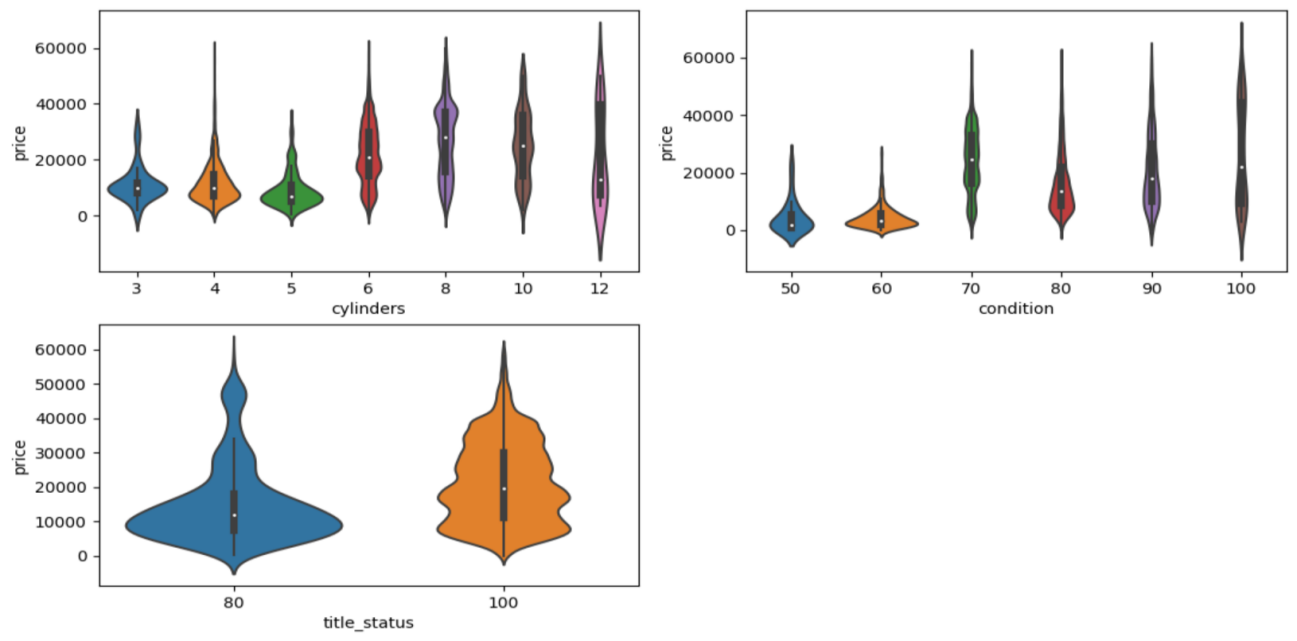


图7 不同定类变量与价格之间的关系图



查看不同类别字段下的价格分布，可以发现不同cylinder类型的车型，其价格跨度会有所不同；车辆状况的好坏程度对价格分布也有较大影响，越好的售价越高；没有改造过的车也比改造过的售价高一些。

* 1. **回归分析**

从上面热力图中，得出价格price和另外几个变量之间相关性很大。因此，接下来使用IMplot进一步查看两个因素之间的关系。Implot对所选数据集进行了一元线性回归，拟合出了一条最佳的直线。

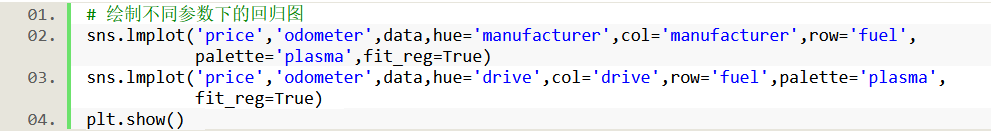
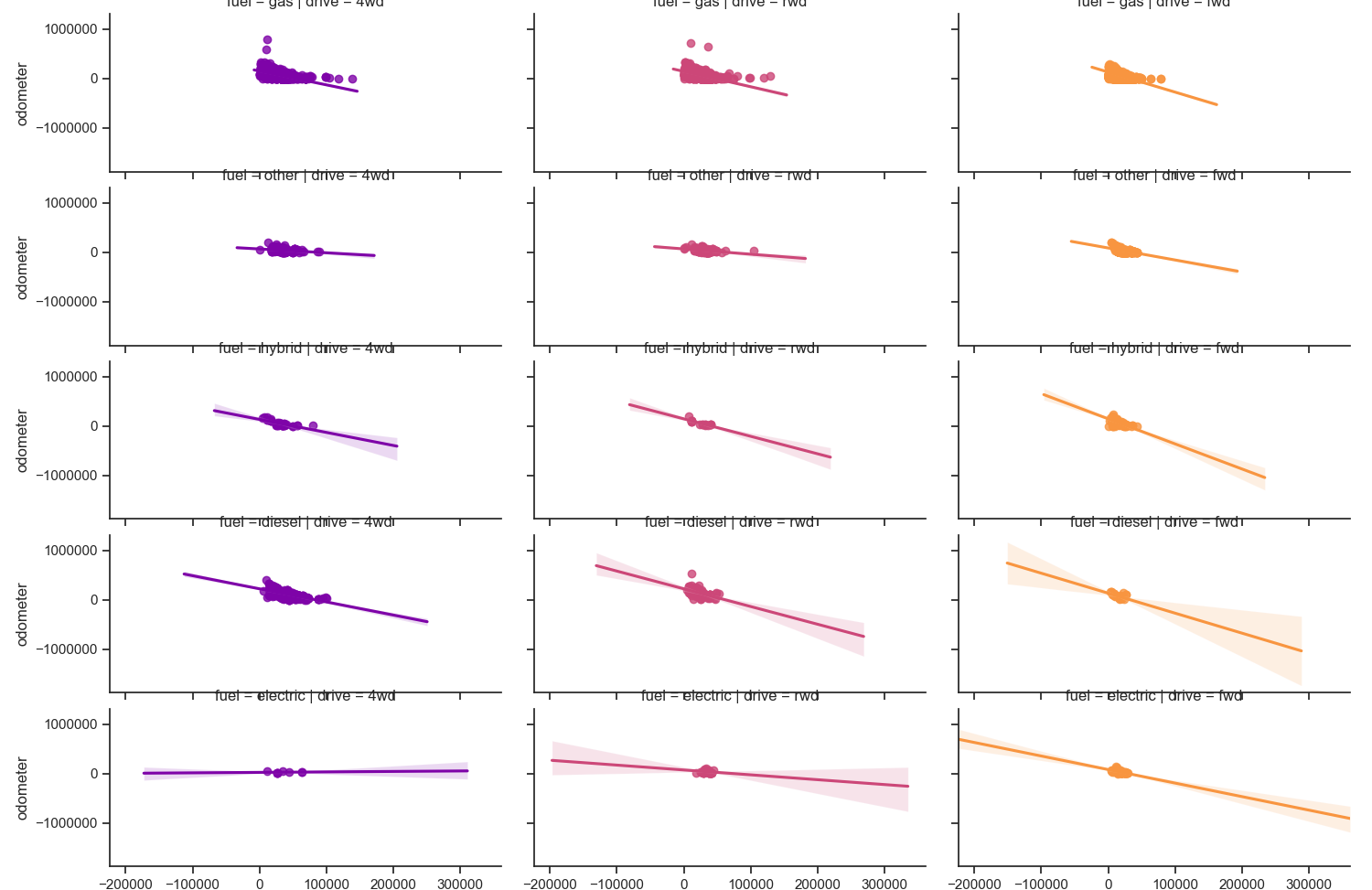
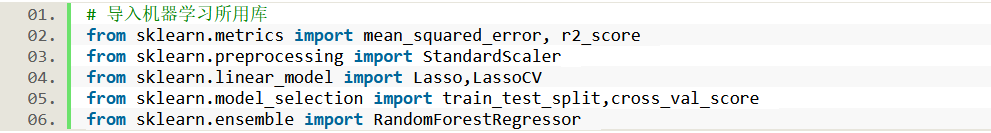


图7 一元线性回归拟合图

1. **模型建立**
   1. **Lasso回归**

首先，导入机器学习所用的人工智能库sklearn。



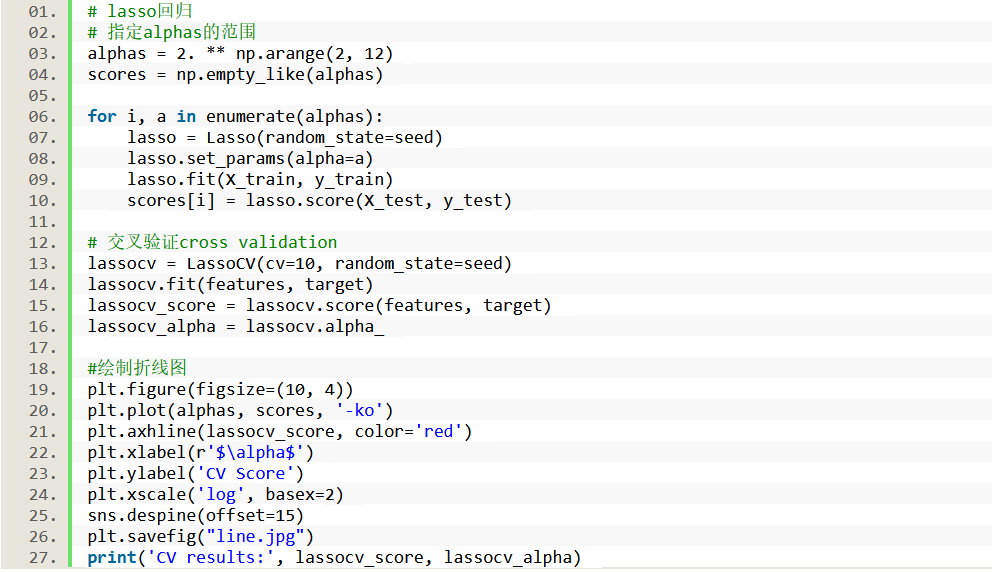
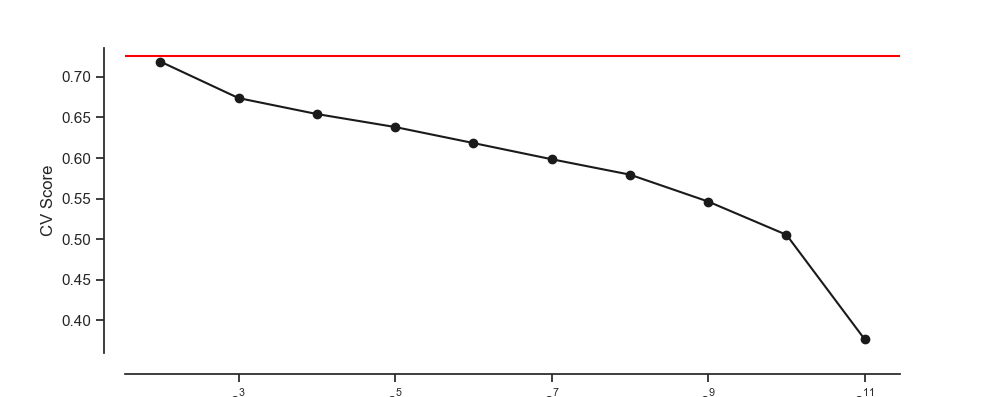


图8 Lasso回归折线图



* 1. **特征重要性分析**

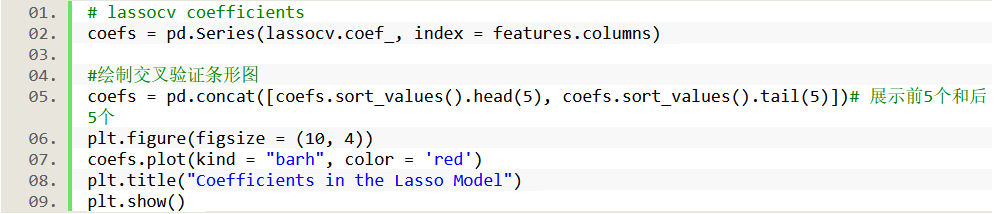
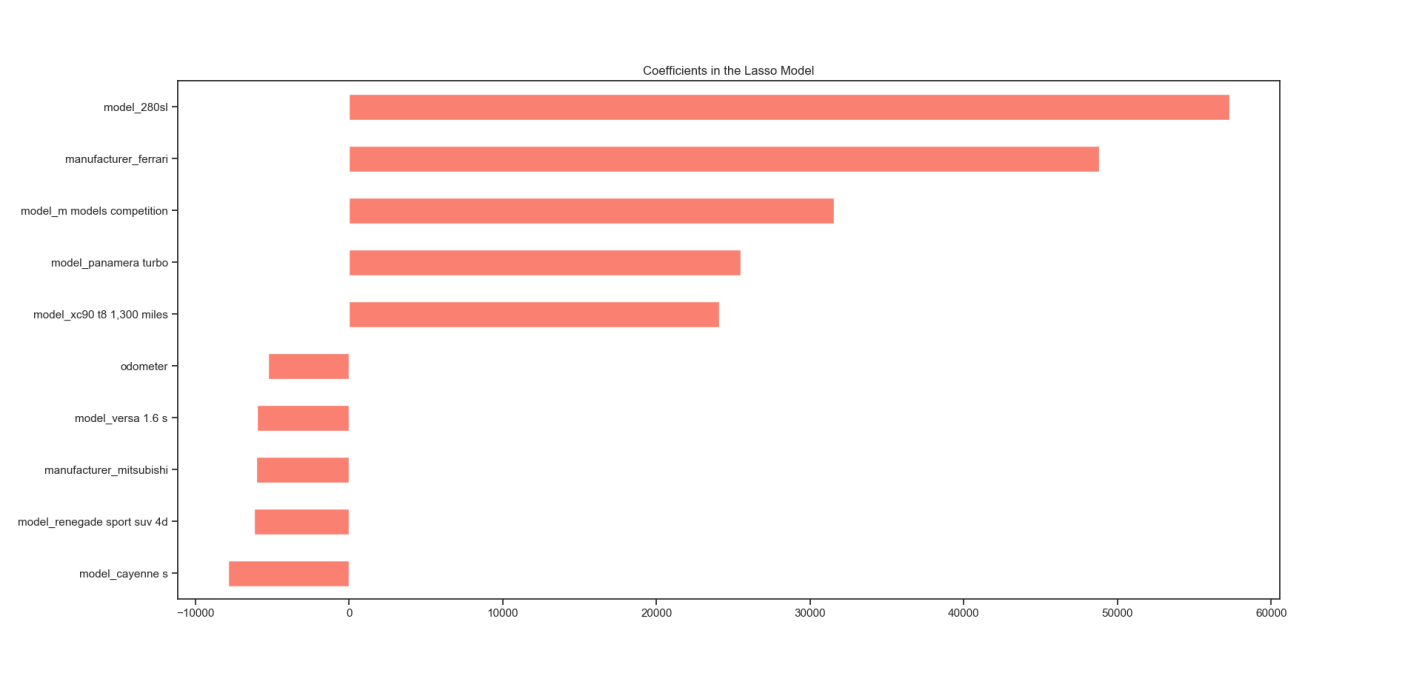
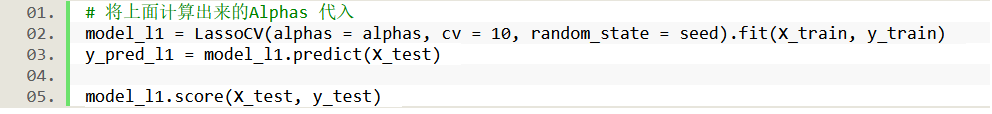
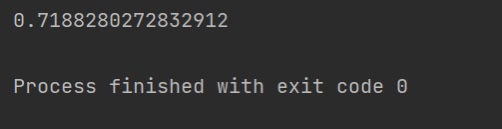


图9 特征重要性交叉验证图





运行结果为：



1. **改进和创新**
2. **观测可视化。**

积极使用matplotlib库、missingno库、seaborn库等可视化第三方库辅助数据进行特性分析及指标衡量，利用直观的可视图像对拟合效果及模型状况进行评估与检验，相比直接的数据。

1. **参数具象化。**

由于二手车数据的回归模型建立需要较多数值指标，通过参数运算获得非数指标的归整数据化，便于lasso回归模型建立。

1. **数据标准化。**

通过缩放手段避免部分特性的方差过大带来的高支配影响或部分特征值过小或过大带来的数据点集中在坐标轴某特定偏远区域现象。利用数据标准化可为分析数据带来带来方便。

1. **求解严谨化。**

通过增加惩罚系数alpha，强制影响系数绝对值之和小于某个固定值，避免单变量的支配影响。

1. **检验合理化。**

利用交叉检验、残差图、MSE和R²等方式对系数重要度及拟合效果进行检验评估，保障预测效果。

1. **结果与分析**
2. **residual plot 残差图**

画图表示实际值和预测值之间的差异。

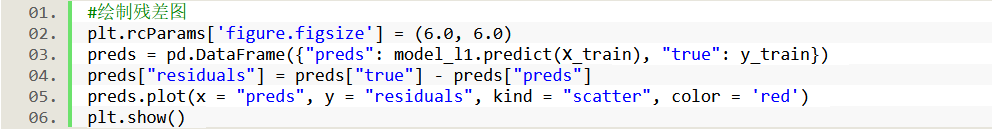
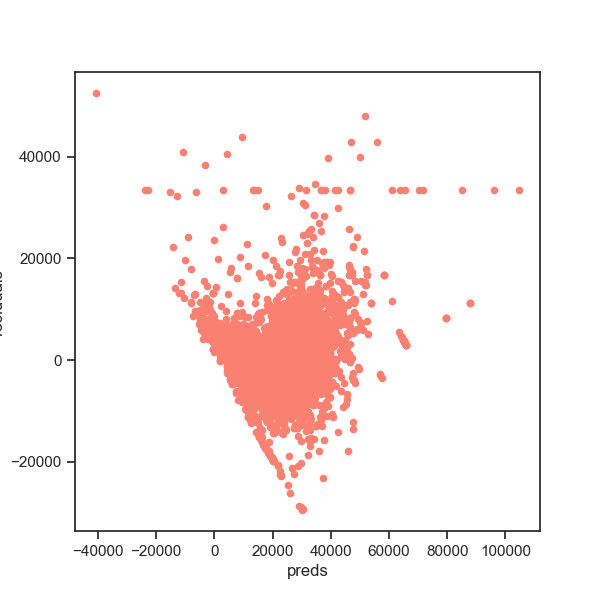
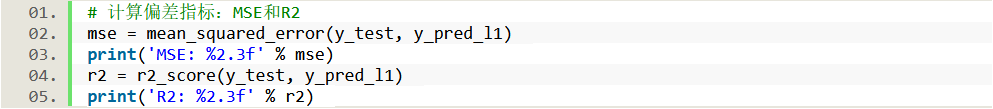


图10 residual plot 残差图



1. **MSE和R2**

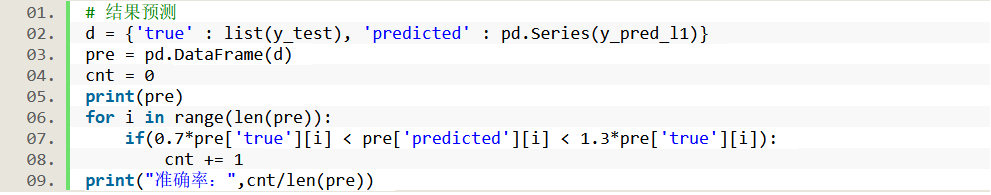


运行结果为：

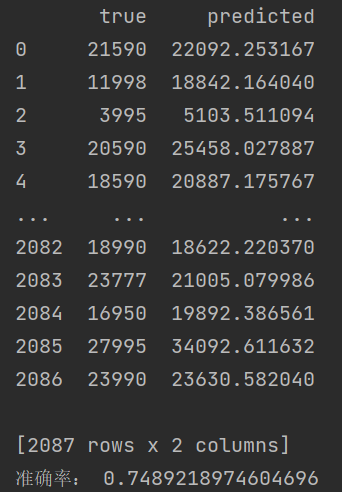
IMG_7454(20211113-193321)

1. **查看具体实际值和预测值**

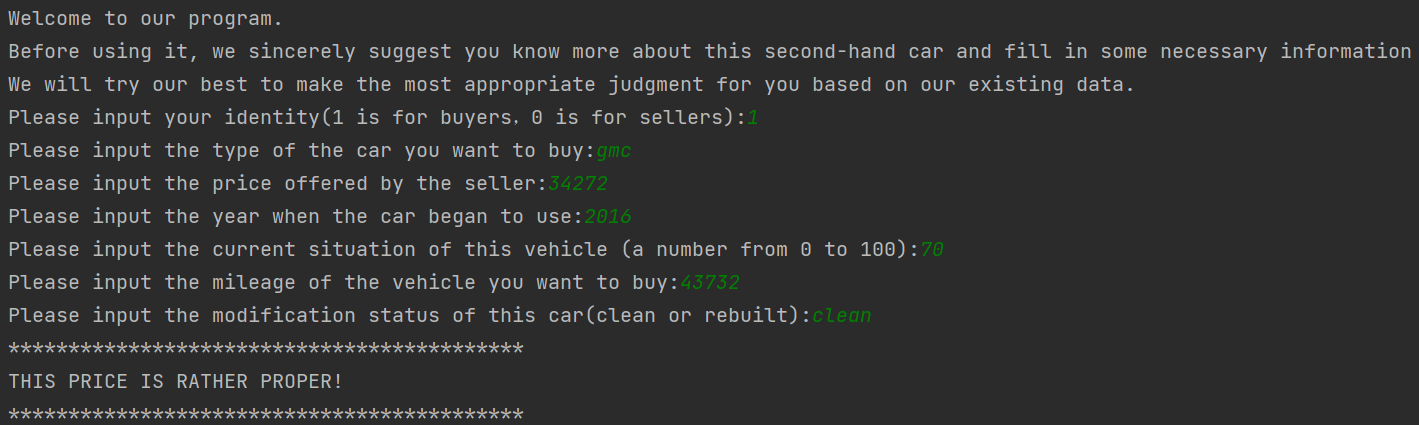
查看模型预测值与真实值的对比，并计算准确率。



结果如下：



面向客户（买家/卖家）的窗口：



1. **讨论与结论**
2. **优点：**

回归检验度高，通过多重检验保障回归准确性。

1. **缺点：**

拟合过程对于数值数据依赖性大，对于非数值数据的处理能力较差。

1. **改进方向：**

通过数据类别分析，进行非数值数据与数值数据的共同拟合。

1. **任务分配**

周一凝：（50%）

①数据信息获知(Information)

②数据清洗(Data Cleaning)

③缺失值处理及变量分布

④概况分析(Condition Analysis)

⑤模型建立(Model establishment)

⑥ppt制作及文档开发

史桠彬：（50%）

①数据信息获知(Information)

②概况分析(Condition Analysis)

③缺失值处理及变量分布

④模型建立(Model establishment)

⑤程序结果 (Data Analysis)

⑥ppt制作及文档开发