**北京航空航天大学软件学院**

**《大数据处理与可视化分析》期末课程报告**

课题名称： 汽车价格预测分析

姓 名： 吴清杰

学 号： ZF2021347

专业方向： 人工智能

所属院系： 北京航空航天大学软件学院

任课教师： 王德庆

课程起止时间： 2020年 12月 至 2021年 1月

### 摘 要

本文通过对汽车咨询公司的不同汽车价格数据集做分析,旨在哪些变量对预测汽车价格具有重要意义,这些变量对汽车价格的描述程度如何

我们需要使用可用的独立变量对汽车价格进行建模。管理层将使用它来了解价格随自变量的确切变化情况，他们可以据此操纵汽车的设计，业务策略等

**关键词**:汽车价格,回归预测,数据分析

## 引言

该数据集从kaggle竞赛上平台获取，国内某汽车公司希望通过在美国设立生产部门并在当地生产汽车，从而在美国和欧洲同行中竞争来进入美国市场。他们与汽车咨询公司签约，以了解汽车定价所依赖的因素。具体来说，他们想了解影响美国市场汽车定价的因素，因为这些因素可能与中国市场有很大不同。该公司想知道：

哪些变量对预测汽车价格具有重要意义:

我们需要使用可用的独立变量对汽车价格进行建模。管理层将使用它来了解价格随自变量的确切变化情况，他们可以据此操纵汽车的设计，业务策略等，以达到一定的价格水平。此外，该模型将是管理层了解新市场定价动态的好方法

## 正文

对于汽车公司来说,一款汽车产品的定价关乎汽车产品在一个市场的竞争力,或许影响新兴汽车公司的承办,应该从市场价格数据中提取最关键的价格因素,从而获取消费趋势,进而定出强有力的价格,提高产品竞争力,使其成为明星产品,从而脱颖而出,使公司站稳脚根.

1. 数据预处理:
   1. 整理现有数据,其所有的特征分别为表1,表2,表3:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 1 | 2 | 3 | 4 | 5 | 6: | 7 | 8 | 9 |
| 列名 | Car\_ID | Symboling | carCompany | fueltype | aspiration | doornumber | carbody | drivewheel | enginelocation |
| 含义 | 自增长id | 风险等级(Categorical) | 汽车品牌(Categorical) | 燃油类型(Categorical) | 洗涤剂(Categorical) | 车门数量(Categorical) | 车体(Categorical) | 汽车轮胎(Categorical) | 发动机位置(Categorical) |

表1: 表列名第一部分

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 列名 | wheelbase | carlength | carwidth | carheight | carbweight | enginetype | cylindernumber | enginesize | fuelsystem |
| 含义 | 轴距 | 汽车长度 | 汽车宽度 | 汽车高度 | 汽车空重 | 引擎类型(Categorical) | 防止气缸位置(Categorical) | 引擎大小 | 燃油系统(Categorical) |

表2: 表列名第二部分

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 列名 | boreratio | stroke | compressionratio | horsepower | peakrpm | citympg | highwaympg | price(Dependent variable) |  |
| 含义 | 汽车镗孔率 | 冲程数 | 压缩比 | 马力 | 峰值 | 城市里程 | 高速里程 | 价格 |  |

表3: 表列名第三部分

* 1. 将现有数据中文字类别(Categorical)转为数字并去Car\_ID与数据无关的列,转换代码如下图1,图2:

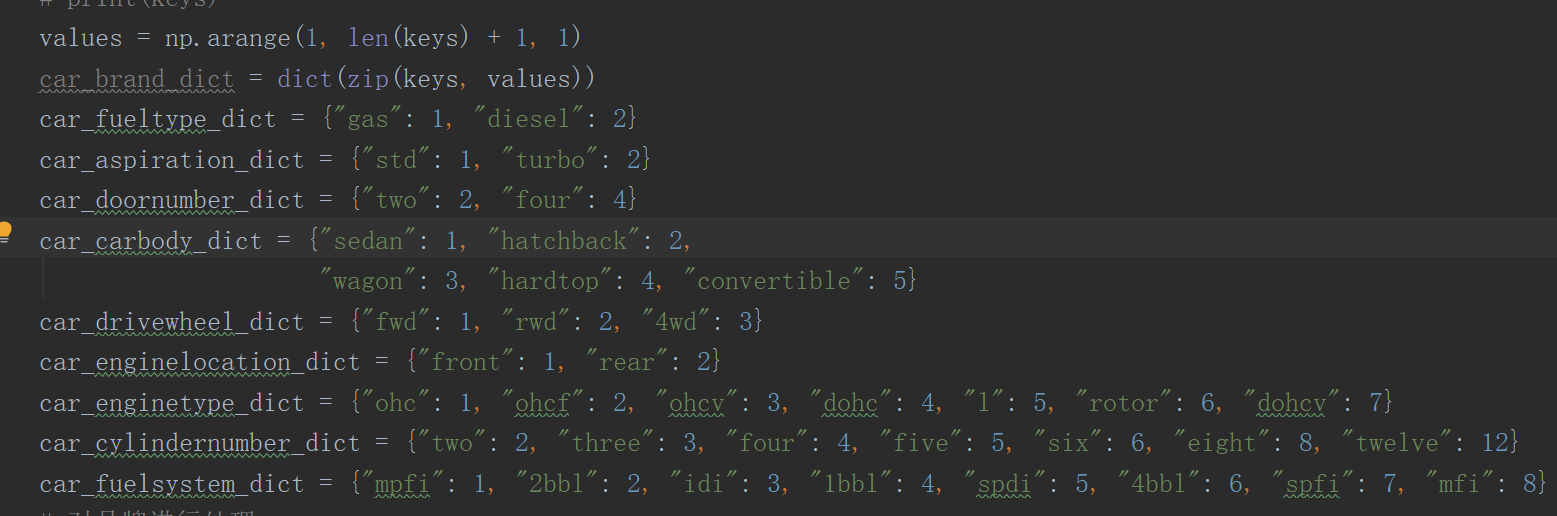


图1: 建立类型对应字典

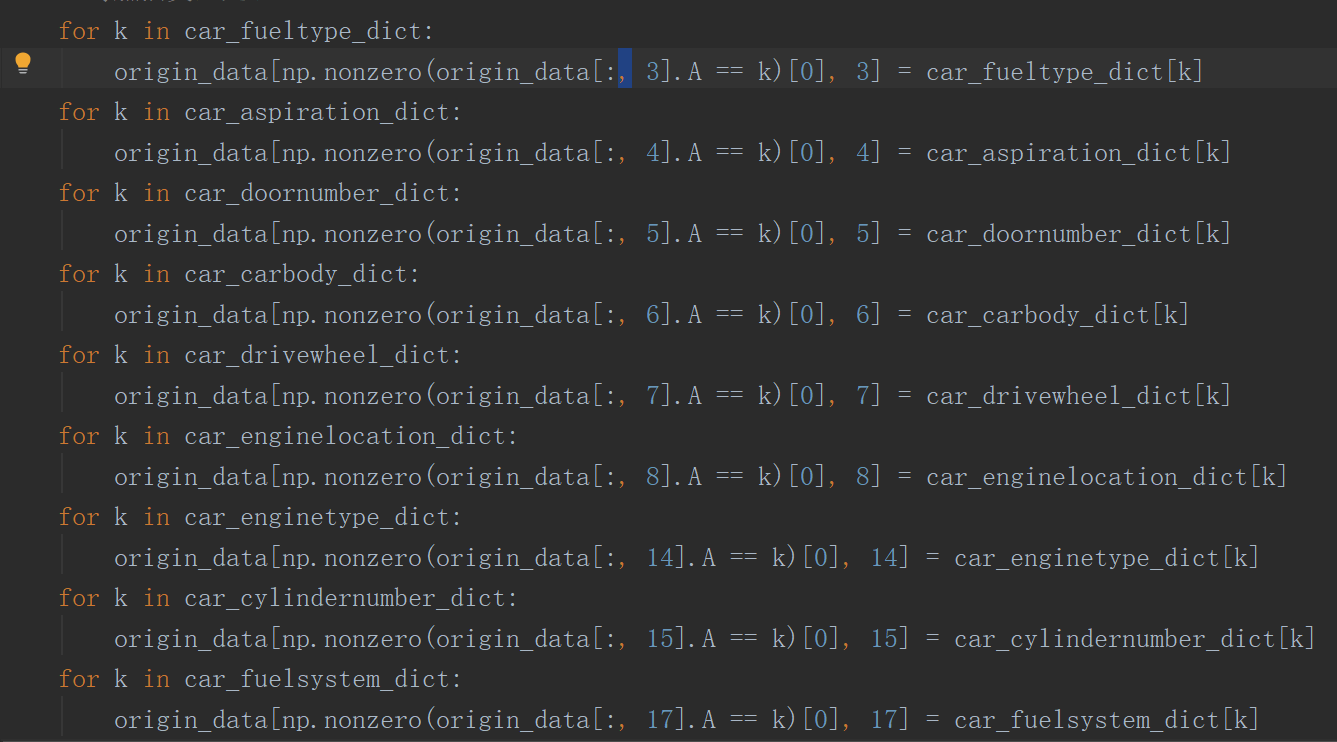


图2: 将数据的类型进行转换

转换之后对数据进行热图展示如图3所示:

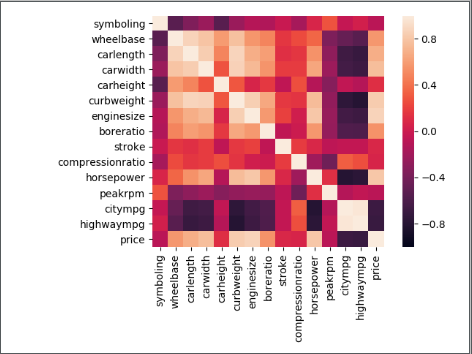


图3: 特征之间相关性的热图

1. 由图展示可得:某些数据相关性特别高,故对数据进行主成分分析,得出特征之中对价格影响最大特征,采用python的pandas库对数据首先进行标准化,然后再求协方差,再进行主成分分析(PCA).

主成分分析(PCA)的思想:设法将原来众多的具有一定相关的指标,重新组合成一堆无关的指标来表示原指标,其中尽中表示信息量(方差)最大的一组组合来尽可能多的反应原指标,如果第一组反应的不够理想,我们可以再加上第二组组合,直到达到可接受的程度,从而达到降维效果,使计算量大大降低,同时得出最重要的因素.

2.1对数进行标准化,代码实现如图4:

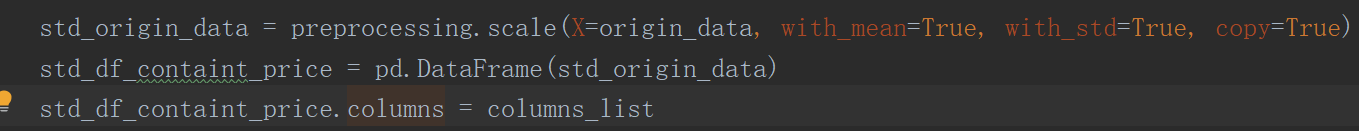


图4: 数据进行标准化

标准化结果为下图5:

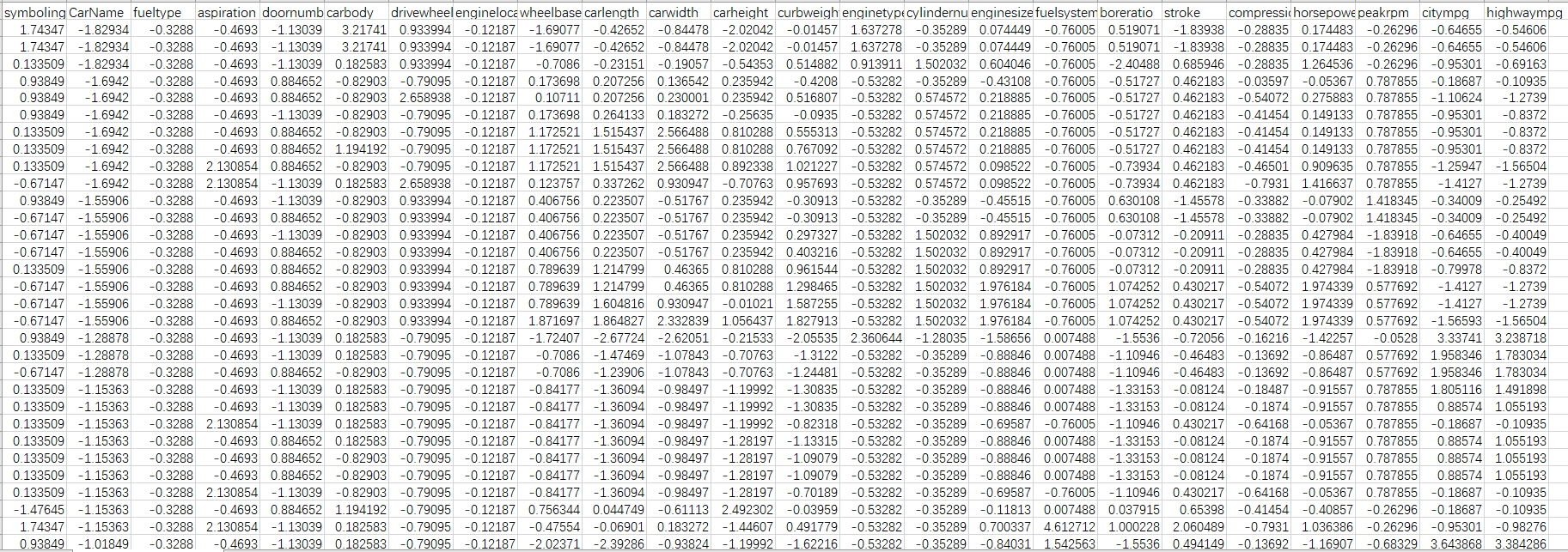


图5: 数据标准化结果

2.2获取数据的协方差矩阵,获取矩阵代码如下图6:

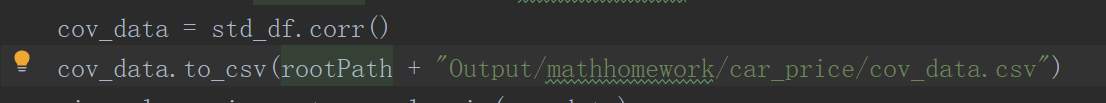


图6: 获取协方差矩阵

结果如下图7:

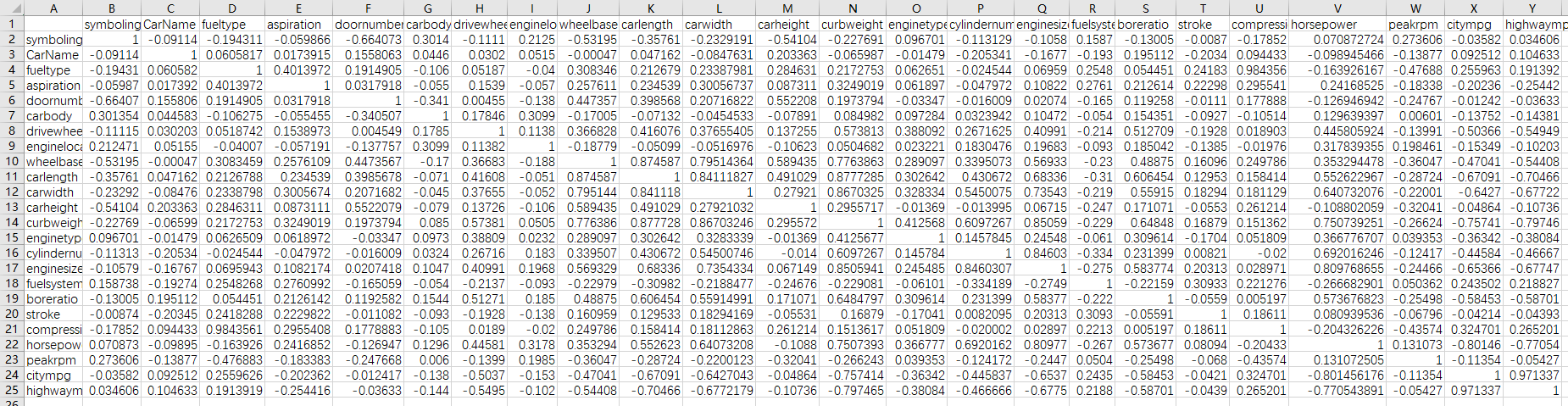


图7: 数据的协方差矩阵

2.3,对协方差矩阵进行主成分分析,获取解释度为前85%的因子,得到载荷矩阵,

转换代码如下图8,图9:

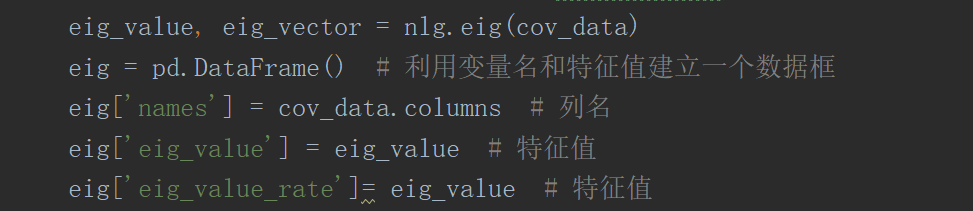


图8: 获取特征值

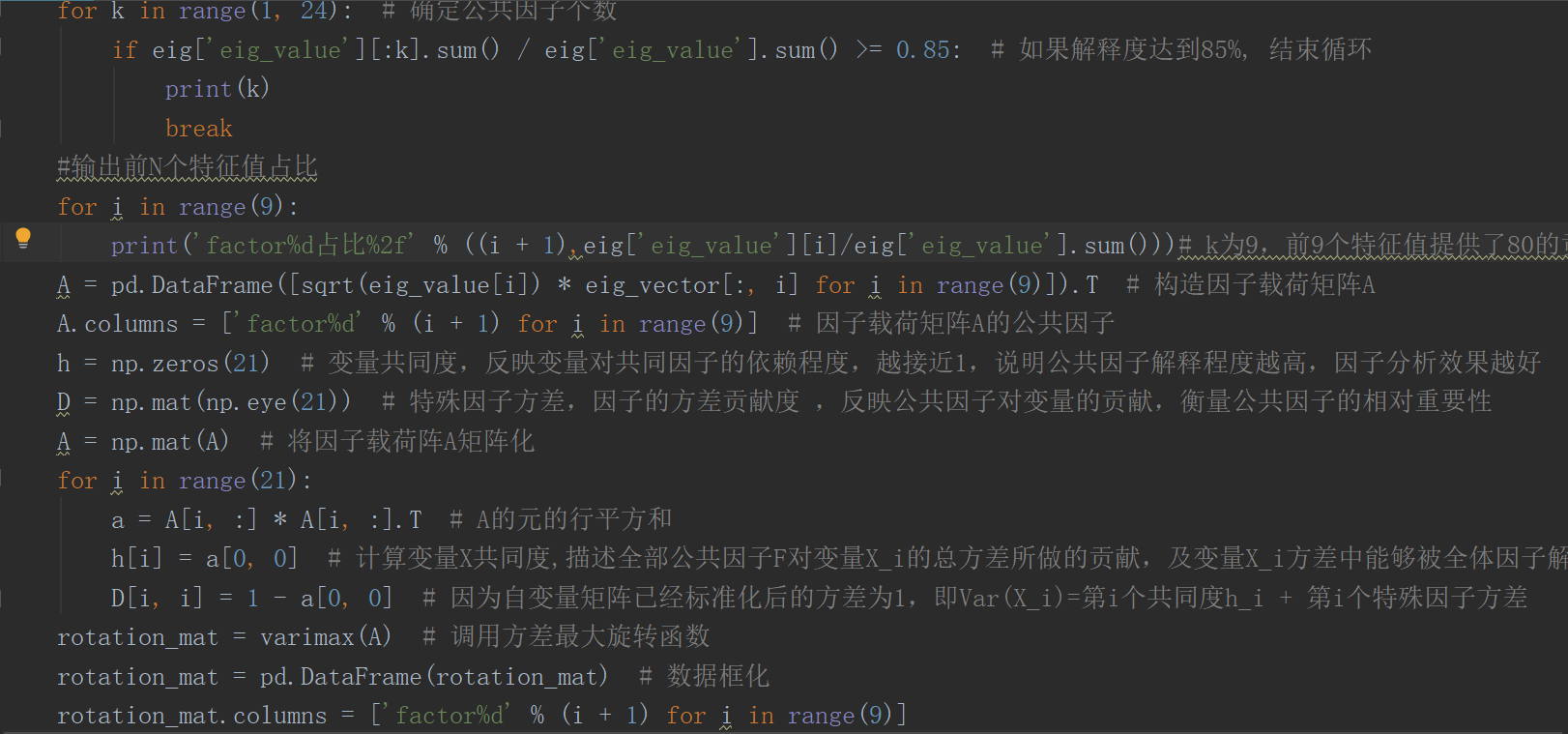


图9: 获取相关因子并旋转

,将因子解释程度控制在85%,并将其按照最大方差旋转结果如图10:

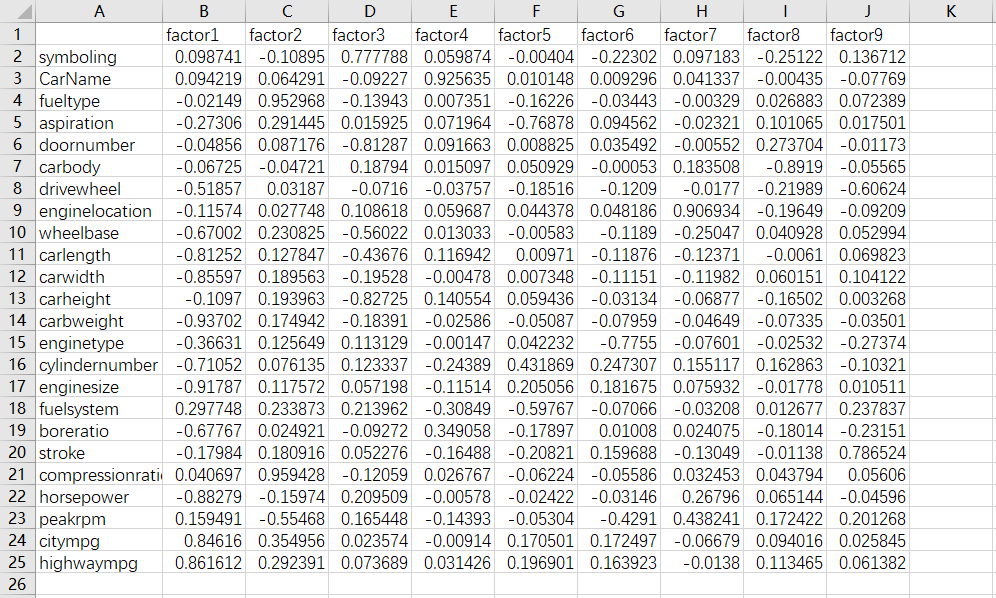


图10: 数据的载荷矩阵

每个特征值所占总的比例为图11:

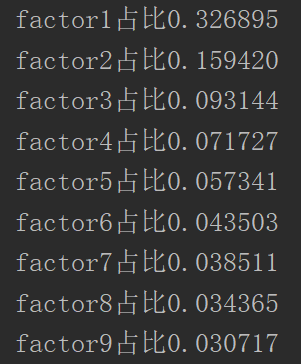


图11: 特征值所占比重

故可以得出:

我们采用对已有数据进行建模,从而可预测汽车价格,从而来制定合理的定价,对于数据中已有的类别数据,我们采用决策树或随机森林进行建模,便可以很好的来预测与制定合理的价格.

采用随机森林建模之后,我们对模型评估如下图6:

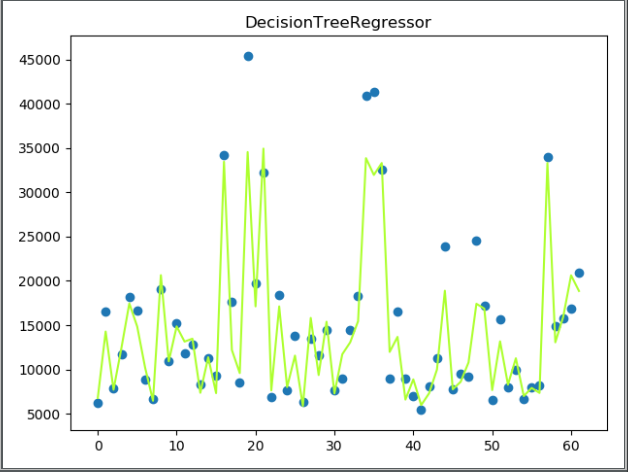


图6: 预测效果图

## 结论

从以上主成分分析的结果来说,第一特征值占比0.32,其比重最高,第二特征值0.15,与第一特征值相比差距悬殊,故从第一个特征值中carwidth, carbweight, enginesize, horsepower, citympg, highwaympg,主要是这几个因素影响价格,汽车公司在这几个关键因素点上提高质量,有利于公司在市场上扎稳脚跟,

同时利用现有数据进行数据建模,对于已生产的数据进行合理的定价,来提高公司的利润

## 参考文献

[1]孙海燕,周梦,李卫国,冯伟.数理统计[M].北京:北京航空航天大学出版社,2016.10.

[1]周志华.机器学习[M].北京:清华大学出版社,2018.5.