基于机器学习的二手车价格评估方法

谢 杨,温 华,张 洁

(西南交通大学 机械工程学院,四川 成都 610031)

摘 要:目前,我国每年乘用车二手车市场交易规模已经高达300万辆以上,二手车交易中如何对车辆进行合理的估值已经成为消费者和经销商最为关注的问题。通过利用数据挖掘技术并结合评估师经验,建立了一种新的二手车价格评估模型,该模型在高达百万样本数据的基础上利用机器学习中的聚类、多元回归等方法将车辆的上牌时间、表征里程、所属地区等因子作为自变量,成新率(保值率)作为因变量而建立。通过实际的评估验证,该模型具有较好的评估效果。

关键词:二手车评估;二手车保值率;机器学习;多元回归

中图分类号:F406.4 文献标识码:A 文章编号:1006-8937(2015)11-0116-03

1 背景概述

我国汽车保有量近年来也实现了快速的增长,截止2014年底,我国乘用车保有量达到了8 307万辆,并且以每年近10%的速度快速增长。2014年乘用车销量达到了1 970.06万辆,连续六年销量全球第一。同年全国共交易二手乘用车351.43万辆,同比增长15.25%。我国二手车增速接近于新车市场增速的两倍。二手汽车取代新车市场地位、成为汽车消费市场的主体是汽车产业发展的必然趋势。美国二手车是新车交易量的3.3倍,德国为2.3倍。保守预测,如果我国二手车与新车交易量达到1:1的水平,市场规模也在千万辆以上。

目前二手车的评估还主要是由评估师根据自己的经验进行,通过数据挖掘技术、经验或其他方法来建立二手车评估模型的研究才兴起不久,目前还没有一种能够具有高准确度,可操作性好的评估模型。不同的车型、配置、车主使用习惯与保养水平、使用年限、地区限购等因素,二手车的价格会有较大的不同。

本文基于高达百万的样本数据,并综合了主流观点和评估理论中所要考虑的因素,对二手车价格评估有主要影响的众多变量进行了分析,最后得到对二手车价格影响最大的多个变量,建立起能够较为合理清晰的反映和解释二手车交易价格的多元回归统计模型。在该模型的基础上,利用传统方法或经验值对其进行参数修正,使其能够较为准确的对大多数情况的二手车进行评估。

2 特征变量与关系模型

2.1 实验数据

实验采用的数据包含:车型,车系,车型配置,车身颜色,车辆用途,行驶里程,所属地区,使用年限,新车价,交易价等,总量在100万行左右。

2.1.1 数据分析

交易数据是对二手车市场交易最为直接的反映,通过数据分析可知:在二手车市场上交易比较活跃的车系有A6、宝马5系、凯越、凯美瑞、雅阁、A4L、福克斯、宝马3系、宝来、君威、锐志、迈腾、科鲁兹、朗逸、天籁、速腾等。可以看出B级车在二手车市场上较受欢迎,其次是A级车。在交易量中约50%集中在30个车系上,在我们统计的1 000个车系占3%;交易量的75%集中在约115个车系上,约占车系总量的11.5%;交易量的90%集中在约225个车系上,约占车系总量的22.5%。除去准新车(指还未上牌或车龄极小的车辆),交易量的70%都集中在车龄5年内的车辆上,车龄活跃程度排名依次为3,2,4,5,1年。车辆交易最为活跃的地区为华东区(江苏省和浙江省),其约占整个市场的30%。

2.1.2 建模思路

通过数据分析可知,市场交易的绝大部分车辆都集中在少部分的车系上,所以如果能够评估好这一百多个车系,便能满足市场评估的大部分需求。在交易集中的这部分车型或车系上,可以利用其丰富的样本数据,挖掘出一个合理的评估模型。对于车型样本数据足够的车型,可以为每个车型建立一个评估模型;然后再为样本数据足够的车系建立评估模型;最后结合数据挖掘和评估师经验为剩余约大部分车系建立评估模型。

2.2 特征变量分析

现行的二手车价格评估方法有多种,如现行市价法,重置成本法和清算价格等。这些方法大都是通过经验来进行评估,不能很好的反映市场因素对车价的影响,而二手车价格受到市场因素影响最大。影响二手车价格的主要变量有:车型(配置、排放、油耗等)、使用年限、车况、有无事故、行驶里程、车身颜色、交易地区(地方政策法规、消费者对不同品牌喜好度等)、新车市场情况(新车销量、后续车型折扣率)、车辆用途等。样本数据并未含所有上诉特征变量,其主要包含:车型及其配置与新车价等、后续车型新车价、车身颜色、车辆用途、交易地区、上牌时间、交易时间、交易价格。

从经验上讲,这些变量对二手车价格都有影响,但是并没有一个科学严谨的证明说明这一点。我们从统计意义上的"相关性"角度来分析。

统计学上的相关性是指两个变量因素的相关密切程度,两个变量的关系可以直观地用散点图表示,当其紧密地群聚于一条直线的周围时,变量间存在强相关性。

2.2.1 使用年限

二手车价格的最大影响因子便是使用年限,为便于利用散点图分析,令差价率=(新车价-二手车价)/新车价,得到的差价率与使用年限的散点图如图1所示,通过散点图分析可得到结论:二手车保值率与使用年限强相关,可通过二次多项式曲线进行关系拟合。

作者简介:谢杨,西南交通大学机械工程学院。

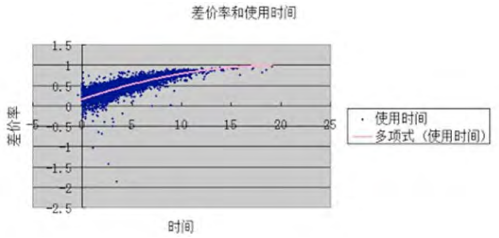

图1 差价率散点图

2.2.2 行驶里程

基于经验考虑,车龄和行驶里程一般存在很强的正相关性,而我们要建立的数学模型需要尽量避免这样的相关性(即多重共线性),因此对行驶里程做如下变换:年均里程=里程数/使用时间,再对年均里程和车龄进行检验,可知年均里程对于使用时间来说,kendall系数较小,可以认为二者无显著相关关系,可将年均里程加入模型中。

2.2.3 车身颜色

对其进行方差(ANOVA)分析,得到的Sig>0.05,认为不同颜色之间二手车交易价格没有显著差异。这个结论和我们的经验相悖,说明可能样本数据中不能提取出颜色对交易价格影响的因素。在二手车市场上大众色系(银、黑、灰、白)相对最为保值,这主要因为其受众群体最高。同时在每个车系推出时都会有一种主打色,主打色系车型的保值率通常高于其他颜色款。因此将颜色分为3类,第一类为某车系主打色A,第二类为大众色B,第三类为其他色C。在评估出了一辆车的保值率(0~1小数表示)后根据其颜色归类将A、B、C类分别乘以102%,100%,98%,对其进行修正。

2.2.4 车辆用途

通过样本数据的方差分析同样不能得出车辆用途对价格有明显的影响,因此同样利用经验值对保值率进行修正。车辆用途分为非营运(保值率不变)、营运(按1~10 a保值率乘以90%~80%递减)、其他(按1~10 a保值率乘以95%值85%递减)。

2.2.5 交易地区

通过对交易地区进行散点图与方差分析,Sig<0.001,可知地区对交易价格有显著的影响。交易地区的值属于属性变量,在模型中可转换为虚拟变量,便于做回归分析。

2.2.6 新车价格

对于建立在车系上的一个具体的评估模型,其可适用该车车系下所有车型。根据经验,即使是同车系的车型,其保值率率也会存在细微的差异。如一两年车龄的同车系的低配和高配车型,新车价差价部分为10万,在二手车交易时这部分的差价往往低于5万,所以建立车系模型时也需将其作为特征变量加入。

2.3 建立关系模型

通过以上的数据和特征变量分析,可以根据不同车型在市场交易的活跃度建立起3个层次化的评估模型。

2.3.1 基于车型的评估模型

对于样本数据量达到200个以上单个车型,建立起以保值率r为因变量,使用年限cl_1,使用年限的平方cl_2,交易地区dq,年均使用里程lc作为自变量的多元线性回归模型。其中因为车辆的保值率和使用年限是二次项关系,所以通过引入自变量cl_2使模型更加准确。交易地区属于定性数据,通过转换为虚拟变量引入多元回归模型:

$$Dj=\begin{cases}1 & \text{地区取值为}j\\ 0 & \text{其他}\end{cases}\quad j=1,2\ldots,9$$

其中,地区取值为9(即最后一个地区)时,用D_1~D_9都取值为0来表示。到保值率后,引入车辆用途修正系数a,车身颜色修正系数b对保值率进行修正,模型用数学公式表述为:

$$r=ab(b_0+b_1cl_1+b_2cl_2+b_3D_1+b_4D_2+b_5D_3+b_6D_4+b_7D_5+b_8D_6+b_9D_7+b_{10}D_8+b_{11}D_9+b_{12}l_C)$$

2.3.2 基于车系的评估模型

利用同样的方法为样本数据量达到500个以上的单个车系建立评估模型,与车型唯一不同的是,在车系模型中,新增车型的新车价xcj作为自变量。车型的新车价能够反映一个车系里不同配置的车型二手车价格的不同。同样也利用车辆用途和车身颜色修正系数a、b对模型进行修正。模型用数学公式表述为:

$$r=ab(b_0+b_1cl_1+b_2cl_2+b_3D_1+b_4D_2+b_5D_3+b_6D_4+b_7D_5+b_8D_6+b_9D_7+b_{10}D_8+b_{11}D_9+b_{12}l_C+b_{13}xcj)$$

2.3.3 通用评估模型

通过将车型分为11类,从样本数据中提取出一个能够覆盖大部分车型通用评估模型,具体的分类方法见表1,根据分类分别为每个类别建立一个评估模型,模型用数学公式与基于车型的评估模型相同。

表1 通用模型车型分类

类别	价格区间	分类编号
进口	15万以下	1
	15~35万	2
	35万以上	3
合资	8万以下	4
	8~15万	5
	15~30万	6
	30万以上	7
国产	8万以下	8
	8~15万	9
	15~25万	10
	25万以上	11

3 多元线性回归

在建立的模型中,存在不同量纲的变量,量纲不同,也会造成模型各变量的系数缺乏直接的含义,不能直观反映每个变量的重要性,即对因变量的解释能力。为了消除量纲影响和变量自身变异大小和数值大小的影响,故将数据标准化。对于评估模型中的使用年限采用离差标准化,将因变量中的观察值减去该变量的最小值,然后除以该变量的极差,其数学公式表述为:

$$x_{ik}'=[x_{ik}-Min(x_k)]/R_k,$$

使用年限:

$$cl_1'=(cl-0.5)/9.5,$$

对于年均行驶里程,新车价采用标准差标准化,将某变量中的观察值减去该变量的平均数,然后除以该变量的标准差,数学公式为:

$$x_{ik}'=[x_{ik}-u_k)]/S_k,$$

年均行驶里程:

$$l_C'=(lc-1.51)/0.79,$$

新车价:

xcj'=(xcj- 29.41) /30.82。

3.1 基于车型的评估模型

单个车型样本数据在200个以上的有1 200个左右,通过多元线性回归,可得到每个模型的参数,选取其中一个车型“ A4L2013款35TFSI无级变速舒适型三厢”,其回归后的模型为:

r=ab(0.814- 0.679cl$_1$'+0.097cl$_2$'- 0.07lc'- 0.011D$_{华东区}$
- 0.006D$_{华南区}$- 0.015D$_{上海区}$- 0.017D$_{西北区}$- 0.007D$_{西南区}$
- 0.019D$_{华北区}$- 0.015D$_{东北区}$- 0.006D$_{华南区}$- 0.007D$_{华中区}$)

3.2 基于车系的评估模型

单个车系样本数据在500个以上的有280个左右,通过多元线性回归,可得到每个模型的参数,选取其中一个车系 别克凯越”,其回归后的模型为:

r=ab(0.365- 0.658cl$_1$'+0.151cl$_2$'- 0.06lc'- 0.0654xcj'
- 0.009D$_{华东区}$+0.012D$_{华南区}$- 0.019D$_{上海区}$- 0.005D$_{西北区}$
- 0.001D$_{西南区}$- 0.004D$_{华北区}$+0.002D$_{东北区}$- 0.015D$_{华南区}$
- 0.008D$_{华中区}$)

3.3 通用评估模型

对分类后的11个类别多元线性回归运算,可得到每个分类模型的参数,选取第6个分类,进口品牌并且新车价介于15~35万,其回归后的模型为:

r=ab(0.796- 0.739cl$_1$'+0.136cl$_2$'- 0.02lc'- 0.022xcj'
- 0.015D$_{华东区}$+0.003D$_{华南区}$- 0.019D$_{上海区}$- 0.009D$_{西北区}$
- 0.011D$_{西南区}$- 0.018D$_{华北区}$+0.008D$_{东北区}$- 0.011D$_{华南区}$
- 0.016D$_{华中区}$)

衡量回归模型优劣的统计量见表2。R为复相关系数,它表示模型中的所有变量与因变量之间的线性回归关系的密切程度大小。它的取值介于0~1之间,R越大说明线性回归关系越密切。调整R^2为重点关注的统计量,它的值越大,模型拟合效果越好,表中调整的R2分别为0.752,0.926,0.883。最后给出标准估计的误差,它的大小反映了建立模型预测因变量的精度,值越小说明所建模型越好。模型方差分析结果中概率P值0.000<0.001,所以该模型是有统计意义的。

表2 模型验证

模型	R	R^2	调整后的R^2	标准估计误差	F	Sig
车型模型	0.868	0.753	0.752	0.0360	642.02	0.000
车系模型	0.962	0.926	0.926	0.0427	948.41	0.000
通用模型	0.940	0.883	0.883	0.0514	863.94	0.000

4 结 语

建立一个精确的二手车评估模型是一项非常困难的工作,因为每一二手车辆车的价格除了受其具体的车况、车主使用习惯等之外,还很大程度上受市场供求关系,品牌知名度以及国家政策等因素的影响。本文利用机器学习的方法,通过挖掘历史交易数据建立了一个能够覆盖大部分车型的评估模型,能够较准确的评估出一辆普通车况的二手车价格,具有较好的使用价值。

参考文献:

[1] 国家统计局.2014年国民经济和社会发展统计公报[EB/OL].http://society.people.com.cn/n/2015/0226/c1008- 26599463.html,2015- 02- 26.
[2] 中国报告大厅.2014年1- 11月中国二手车销量分析:增长率近新车三倍[EB/OL].http://www.chinabgao.com/stat/stats/39670.html,2014- 12- 23.
[3] 侯江丽,赵飞.基于AHP算法的二手车评估方法的研究[J].邢台职业技术学院学报,2013,(3).
[4] 郭振江.旧机动车评价方法的建立与体系研究[D].西安:长安大学,2011.

(上接第89页) 入分析故障原因,制定相关的标准规范,以促进电压互感器的合理使用。

5 结 语

220 kV变电站电压互感器故障千变万化,作为维修检查人员,必须有丰富的经验,具备专业知识与技术能力,并能够灵活应用,准确了解故障原因,才能为有效处理故障打好基础。

参考文献:

[1] 汪晓明,何萍,童军心,等.一起220 kV电容型电压互感器故障的原因分析及防范措施[J].江西电力,2014,(1).
[2] 霍思敏,汤吉鸿.一起复杂事故的保护动作行为分析[J].电力自动化设备,2012,(3).
[3] 沈靖.基于110 kV变电站的电压互感器故障原因研究[J].中国科技信息,2011,(22).
[4] 吴雷锋.220 kV母线电压互感器二次反送电原因分析[J].电力学报,2013,(4).
[5] 侯明哲.220 kV母线电容式电压互感器故障处理及原因分析[J].技术与市场,2013,(12).