基于机器学习对电动汽车价格预测

欧阳汉　廖文琪　田秋红

（贵州财经大学大数据统计学院　贵州　贵阳　550025）

摘　要：使用比赛数据，通过挖掘原始数据并结合背景知识创建了7个特征，使用随机森林和XGBoost模型对特征重要性排名前十五的特征中共同出现的特征进行选取，得到9个特征；然后建立KNN、集成学习类、神经网络等模型，选取三组数据集分别建立模型，最终选取了筛选后的数据集和使用Adam优化器的神经网络作为最终的价格预测模型，其准确率达到了96.6%。

关键词：电动汽车；特征工程；KNN；集成学习；神经网络

一、引言

随着中国经济的快速发展，我国汽车工业将面临未来的大变革，汽车行业的良性发展改善消费经济结构。同时中国新能源汽车在全球的占比超过44%，新能源的出现使得汽车行业稳步前进，其中电动汽车消费市场潜力巨大，在汽车行业发挥着重要作用。对于电动汽车价格的影响因素，Hasishi Ishitani（2007）认为电动汽车研发、企业组织关系对新能源汽车发展体系的建立有益[1]；黄振邦等（2007）对不同种类车型的混合动力电动汽车特点进行研究，并对电动汽车行业发展前景做了预期[2]；闫兆炜（2012）从技术因素、市场因素、产业化发展因素方面对中国新能源汽车产业发展进行了研究[3]。在方法研究上，李宝胜（2020）先对数据进行预处理，选取的方法是Pearson相关系数法和主成分分析法，得到有价值的样本数据后使用支持向量机模型对电动汽车进行了研究[4]；林倩玉（2019）运用“学习曲线”的方法对特斯拉和比亚迪的电动汽车价格预测，并与传统典型燃油车进行了对比[5]。

不同品牌的电动汽车有着不同规格的汽车属性和价格，本文基于上海财经大学首届研究生工业与金融大数据建模与计算邀请赛的比赛数据，通过挖掘属性与价格之间的关系，创建合适的特征并根据所有的特征重要性进行筛选，最后建立合适的模型对这批未知价格电动汽车的价格进行预测[6]。

二、数据预处理

（一）数据描述

本文所使用的数据是比赛数据，共有20个属性，可以将其划分为硬件类、基础功能类和外观类三类。硬件类包括电池容量（mAh）、单个电池充电时长、前置摄像头像素、主要摄像头像素、屏幕高度和宽度（cm）、像素分辨率高度和宽度、处理器核心个数、处理器执行速度、随机存取储存器、内存（GB）；基础功能类包括是否支持4G、是否支持3G、是否支持双SIM卡、是否支持Wi-Fi、是否支持蓝牙、是否支持触摸屏；外观类包括移动深度（cm）、重量。

对电动汽车来说，电池总能量（mAh）和电池的个数决定了其续航里程，而汽车的重量会极大影响电动汽车每公里的能耗，这两个对消费者来说，都是必须考虑的属性，而充电速度是能够提升消费者后续体验的重要属性；前置摄像头和主要摄像头的配置，是属于电动汽车智能业务方面的属性，主要应用是障碍物识别等，对于普通消费者来讲，这可能是一项非必需属性；触摸屏幕是人与电动汽车智能交互的媒介，其尺寸大小和屏幕分辨率能够大幅影响驾驶者的驾驶体验，对于一般消费者来说，也是非必需的属性；值得注意的是，支持越多功能的电动汽车，其耗电量就越大，而越好的处理器核心数和处理器执行速度能够通过降低功耗改善其电量消耗问题，并且支持更多的功能，这对消费者来说，也是提升体验的非必须属性；内存（GB）的大小和随机存取储存器个数对电动汽车智能化性能的影响很大，内存大小决定着是否能拥有更多功能，随机存取储存器大小决定着能否同时运行这些功能的上限；在网络制式方面，一般能够有了支持4G的网络对3G的需求就不会很大，没有3G网络就意味着该电动汽车的娱乐软件更新等问题需要Wi-Fi（只包含有屏幕的汽车），因此，这三种属性一定程度上代表了电动汽车的智能化程度，对消费者来说是属于提升体验的属性，能否支持蓝牙和双SIM卡也是同样的作用。

（二）分类型变量分析

通过查看关于价格等级的标签分布状况，发现关于价格的标签分类还是较为均衡，总共1500个数据，每类标签的数量分布都在350~400之间，继续查看每个价格标签下各个分类特征的分布情况：

作者简介：欧阳汉（1979.8-），男，贵州贵阳人，贵州财经大学大数据统计学院副教授，厦门大学经济学博士，研究方向：数据挖掘，宏观经济模型；
廖文琪（1996.8-），男，贵州遵义人，贵州财经大学大数据统计学院研究生，研究方向：数据挖掘、经济社会统计；
田秋红（1996.7-），女，贵州遵义人，硕士研究生，贵州财经大学大数据统计学院研究生，研究方向：数据挖掘、经济社会统计。

表 1　不同分类特征在不同价格下的分布情况

Price	4G	3G	D - SIM	Wi - Fi	Bluetooth	Tscreen
3	0. 545946	0. 770270	0. 521622	0. 502703	0. 513514	0. 486486
2	0. 485411	0. 779841	0. 496021	0. 509284	0. 490716	0. 445623
1	0. 542005	0. 750678	0. 517615	0. 514905	0. 523035	0. 542005
0	0. 536458	0. 765625	0. 494792	0. 492188	0. 481771	0. 523438

在表 1 中能够观察得到，3G 网络支持在各个价位的汽车中，都占据了较大的比例，全都占据了 75% 以上，说明 3G 网络支持在电动汽车的普及率比较高，可以看作是电动汽车的标准配置；4G 配置的普及率没有 3G 高，高价车、中低价车都占据了 50% 以上，只有价位 2 没有，这是比较奇怪的，实际上，在价位 2 的双 SIM 卡支持、蓝牙支持和触屏支持占据的比例都偏低；对于双 SIM 卡的支持和蓝牙的支持，占据 50% 以上的价位都为 1 和 3；Wi - Fi 的支持在各个价位占据的比例都在 50% 左右，相差不是很大；而触摸屏更多的出现于低价车中，在高价位车中占据的比例不到一半。

（三）特征构造

由于电池的区间和价格有关，因此，将电池容量划分为五个等级，分别是 0 至 4，分别代表低、中低、中、中高、高，并命名为电池等级（Battery capacity level）。

由于缺少电池个数的属性，不能构建关于续航里程的特征，但通过电池容量、充电时间和车身重量构建了两个特征，一个是电池充电速率的评估特征，BAspeed 表示充电速率，BA 为电池容量，Ctime 为充电时间，具体计算为:

$$BAspeed = BA \div Ctime$$

另一个是电池固定消耗的评估特征，Carconsumption 表示车身消耗，具体计算为:

$$Carconsumption = BA \div Vweight$$

由于处理器的核心个数和处理速度与价格没有明显的线性关系，选择将其整合，成为评估其性能的特征，P performance 表示处理器性能，NOP 表示处理器核心数，Pspeed 表示处理速度，具体计算为:

$$Pperformance = NOP * Pspeed$$

一般说屏幕的尺寸指的是它的对角线距离，将已有的屏幕宽度和屏幕长度通过计算得出尺寸数据，并划分为 5 个等级，分别是 0 至 4，代表大屏幕（22 ~ 30）、中大屏幕（17 ~ 21）、中等屏幕（12 ~ 16）、中小屏幕（7 ~ 11）、小屏幕或无屏幕（2 ~ 6cm），用 size 表示，具体计算为：

$$Size = \sqrt{(2\&Sheight\hat{}2 + Swidth\hat{}2)}$$

由于 4G 的等级在 3G 之上，且支持 4G 的大多数都支持 3G，因此，将网络制式支持划分为 3 等级，分别为 0 至 2，代表最高支持的网络制式为 4G、3G 和无网络制式支持，并命名为网络支持 networksupport。屏幕的分辨率一般也是组合来作为评估屏幕的特征，将划分为 5 个等级，分别是 0 至 4，代表高分辨率（1600 * 1200）、中高分辨率（1300 * 800）、一般分辨率（1000 * 500）、中低分辨率（700 * 200）、低分辨率（700 * 200）以下，并命名为 SRlevel。

（四）特征选择

在创建完 7 个特征之后，加上原来的 20 个特征，想知道哪些特征是真正对价格有着影响的特征，因此，选择基于 XGBoost 的特征重要性和基于随机森林排序的特征重要性筛选[1]，可以观察到两个模型的特征重要性分布有相同的特征重要性选择，也有着不同的特征重要性选择。在对具体权重进行对比后，发现两个模型对于重要性排名前五的特征选择只有一个不同，相同的是 RAM）、BA、R - width、Car - consumption；对于后五位的不重要特征的选择有两个不同，相同的是 3G、4G、Bluetooth。

在通过两个模型的特征重要性对比之后，决定选取两个模型特征重要性的前十五中都出现的特征作为筛选之后的特征，筛选后的特征为 9 个，具体是 RAM、BA、R - width、R - height、Car - consumption、BA spend、Memory、V - weight、FontCP。

（五）特征重要性分析

筛选之后的特征中，有 7 个是原始特征，2 个是创造的特征。7 个原始特征中，属于硬件类的有 6 个，分别是随机存取储存器（RAM、电池容量（BA）、像素分辨率宽度（R - width）、像素分辨率高度（R - weight）、内存（Memory）、前置摄像头像素（FrontCP），外观类的是汽车重量（V - weight）。说明在 1500 个数据集中，汽车的硬件类属性对于价格的影响是决定性的，而基础功能类的属性在筛选的特征中一个都没有出现，外观类的两个特征出现了一个，由此得出，在市场上，硬件属性和外观属性对于价格的影响更为明显。创建的特征出现了两个，分别是车身消耗（Car - consumption）和充电速率（BA - level），说明对于消费者来说，电动汽车最重要的还是续航能力及影响续航能力的因素，它们都对消费市场有着重要的影响。

三、模型评估与选择

为了找到更好的模型对验证集进行预测，选取了随机森林、KNN、XGBoost、多层感知机等几种模型来进行效果对比，为了验证创建特征和筛选特征是否真的有效，还做了三组数据集作为对比，ecar 表示原始数据集，ecarf 表示创建特征之后的数据集，ecarfs 表示创建特征后进行筛选的数据集，将最终模型表现和数据集表现好的模型和数据集作为最终的模型和数据集。在接下来所有的模型中，现将数据集划分为 1050 个训练集和 450 个测试集，随机数种子设为 1，使每次数据划分保持稳定，并将其标准化。

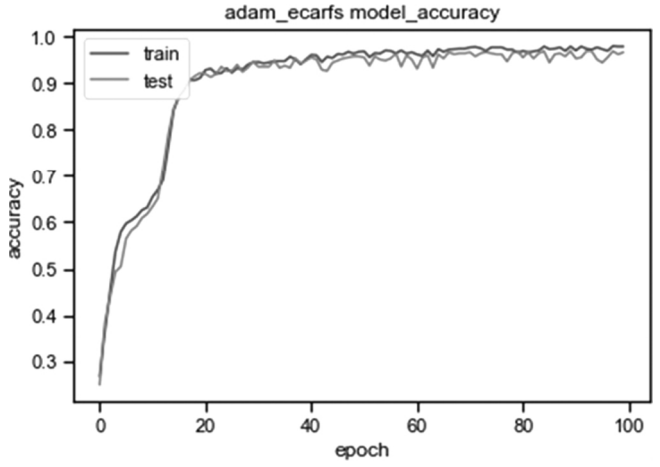
随机森林、KNN、XGBoost、多层感知机等得到训练集准确率测

试集准确率如表 2 所示。

表 2　多种方法结果

方法	数据集	训练集准确率	测试集准确率
KNN	ecar	0. 931	0. 924
	ecarf	0. 931	0. 938
	ecarfs	0. 931	0. 938
随机森林	ecar	0. 894	0. 893
	ecarf	0. 899	0. 887
	ecarfs	0. 898	0. 880
XGBoost	ecar	0. 999	0. 931
	ecarf	0. 995	0. 916
	ecarfs	1	0. 911
SGD	ecar	0. 924	0. 178
	ecarf	0. 926	0. 163
	ecarfs	0. 942	0. 140
Adam	ecar	0. 924	0. 192
	ecarf	0. 909	0. 204
	ecarfs	0. 966	0. 096

从表 2 可以看出，在使用 SGD 优化器时，每个模型的精准度都达到了 92% 以上，损失函数均小于 0. 18，准确率最高且损失函数最低的数据集是 ecarfs；在使用 Adam 优化器时，ecarf 的精准度最低、损失函数最大，但 ecarfs 的准确率最高，达到了 96. 6% ，损失函数最小，为 0. 09。两个优化器综合来看，ecarfs 数据集、使用 Adam 优化器的模型准确率最高、损失最小，它的效果是最好的，Adam 优化器的 ecarfs 数据集神经网络准确率和损失函数如图 1 所示。



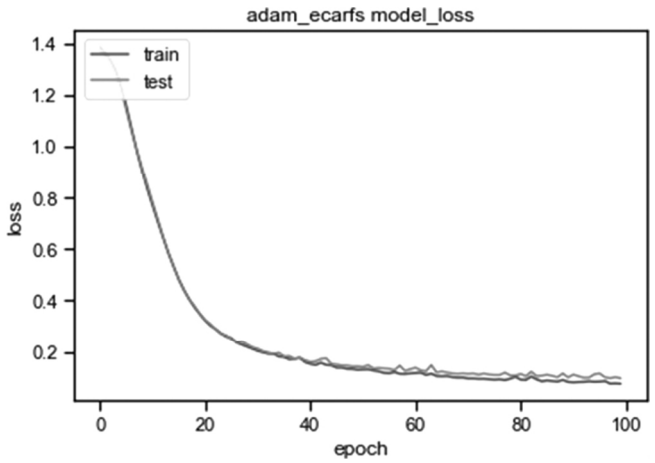


图 1　神经网络准确率和损失函数图

从准确率图 1 可以看出模型的准确率收敛的还是比较快，在 20 次左右的迭代就已经达到了 90. 0% 以上，而损失函数 40 次左右的迭代达到了 0. 2 以下。综合来看，训练集和测试集在准确率图像、损失函数图像上的表现差不多，因此模型不存在过拟合的情况。

通过建立三组不同的数据集和 4 种不同的模型，根据准确率和损失函数，最终选取了 ecarfs 数据集和神经网络 Adam 优化器的模型作为最终的数据集与模型，其在测试集的准确率达到了 96. 6% ，损失函数达到了 0. 096。

参考文献:

[1] Hasishi Ishitani. Overview of Japan’s Effort on Plug – in Hybird，Vehicle，EVS – 23 Plug – in Hybrid Electric Vehicle workshop [J]. California USA. December 2007.

[2] 黄振邦，吴森. 混合动力电动汽车研究开发及前景展望[J]. 城市车辆. 2007 (7): 34 – 36.

[3] 闫兆炜. 中国新能源汽车产业发展研究 [D]. 东北财经大学，2012.

[4] 李宝胜，秦传东. 基于粒子群优化的 SVM 多分类的电动车价格预测研究 [J]. 计算机科学，2020，47 (S2): 421 – 424.

[5] 林倩云，邱国玉，曾惠，等. 基于“学习曲线”的我国纯电动汽车价格补贴及其可持续性研究 [J]. 管理现代化，2019，39 (03): 39 – 43.

[6] 王众. 基于电动汽车用电行为的电池预测研究 [D]. 青岛理工大学，2019.

[7] 卢泓宇，张敏，刘奕群，等. 卷积神经网络特征重要性分析及增强特征选择模型 [J]. 软件学报，2017，28 (11): 2879 – 2890.