

二手车交易价格预测

背景描述及需求

自2010年开始，中国的二手车市场开始步入了快速发展阶段，并开始成为汽车市场重要的组成部分，二手车市场已经具备了实现跨越式发展的基本条件，为符合国际通行的运营模式的应用打下了坚实的市场基础。面对市场上激烈的二手车价格竞争，为了针对二手车高效的销售，对此根据二手车历史数据进行分析价格的涨幅情况。

本次对二手车历史销售数据进行合理的价格预测，本次数据挖掘数据建模目标如下：

1. 借助二手车历史销售数据分析对价格的影响的相关因素；
2. 建立模型预测二手车销售价格，掌握二手车价格的情况；
3. 针对二手车价格预测情况做好活动方便合理销售。

- 背景描述及需求
- 实施过程
 - 数据接入
 - 数据探索
 - 数据预处理
 - 相关性分析
 - 建立模型
 - 模型评估
 - 业务分析
- 总结



参数

属性

帮助

归一化

方法选择

标准化



参数

- ☐ 单位标准差归一化
- ☐ 平均数据中心化

分裂特征的数量 范围是 ≥ 2 的整数 ②

150

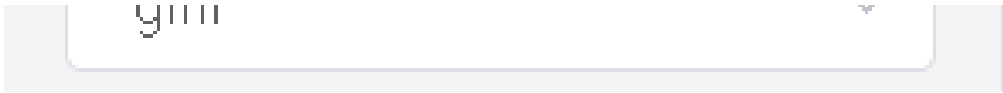
树的深度 范围是 $[1, 30]$ 的整数 ②

10

计算信息增益的方式 ②

gini





- 1) 获取数据，数据来源于赛题二手车交易价格预测；
- 2) 对获取的数据进行基本的处理操作，相关性分析输入特征；
- 3) 根据统计特征数据建立二手车交易价格预测模型；
- 4) 对模型结果进行评估。

实施过程

本案例共收集到15万条二手车历史数据，字段详细说明如下：

字段名称	类型	字段说明
SaleID	整型	交易ID，唯一编码
name	整型	汽车交易名称，已脱敏
regDate	整型	汽车注册日期，例如20160101，2016年01月01日
model	浮点型	车型编码，已脱敏
brand	整型	汽车品牌，已脱敏
bodyType	浮点型	车身类型：豪华轿车：0，微型车：1，厢型车：2，大巴车：3，敞篷车：4，双门汽车：5，商务车：6，搅拌车：7
fuelType	浮点型	燃油类型：汽油：0，柴油：1，液化石油气：2，天然气：3，混合动力：4，其他：5，电动：6
gearbox	浮点型	变速箱：手动：0，自动：1
power	整型	发动机功率：范围 [0, 600]
kilometer	整型	汽车已行驶公里，单位万km
notRepairedDamage	字符串	汽车有尚未修复的损坏：是：0，否：1
regionCode	整型	地区编码，已脱敏
seller	整型	销售方：个体：0，非个体：1
offerType	整型	报价类型：提供：0，请求：1
creatDate	整型	汽车上线时间，即开始售卖时间
price	整型	二手车交易价格（预测目标）
v系列特征	整型	匿名特征，包含v0-14在内15个匿名特征

数据接入

在实验中添加 [数据源](#) 节点，将二手车交易价格数据读取进来，部分数据如图所示：

🕒 当前显示 100 条 / 总共有 150000 条数据 提示: 点击单元格可查看超出的内容

×

# SaleID	# name	# regDate	# model	# brand	# bodyType	# fuelType	# gei
0	736	20040402	30.0	6	1.0	0.0	
1	2262	20030301	40.0	1	2.0	0.0	
2	14874	20040403	115.0	15	1.0	0.0	
3	71865	19960908	109.0	10	0.0	0.0	
4	111080	20120103	110.0	5	1.0	0.0	
5	137642	20090602	24.0	10	0.0	1.0	
6	2402	19990411	13.0	4	0.0	0.0	
7	165346	19990706	26.0	14	1.0	0.0	
8	2974	20030205	19.0	1	2.0	1.0	
9	82021	19980101	7.0	7	5.0	0.0	
10	18961	20050811	19.0	9	3.0	1.0	

注意: 表头中📌表示特征列, *表示标签列

表头真名 ☒ 表头别名

数据探索

本案例的探索分析是对数据进行缺失值分析与数据分布分析, 分析出数据的缺失和分布情况。通过 [聚合](#) 节点查看原表数据中notRepairedDamage字段类型有三种。需要将“-”的数据设置为空值。

🕒 当前显示 3 条 / 总共有 3 条数据 提示: 点击单元格可查看超出的内容

×

# Group_notRepairedDamage	# Count_notRepairedDamage
1.0	14315
0.0	111361
-	24324

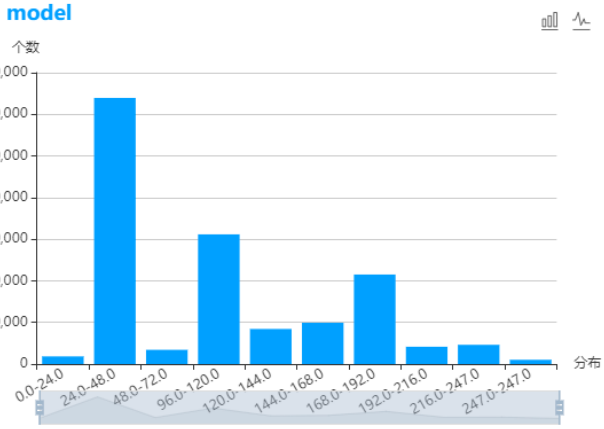
注意: 表头中📌表示特征列, *表示标签列

表头真名 ☒ 表头别名

通过 [全表统计](#) 节点统计所有特征, 查看各指标的分布情况, 发现部分数据含有缺失情况, 如图:

查看全量统计

方差	总和	行数	唯一值	缺失值	偏度	峰度
1.87501...	1124992...	150000	150000	0	0.0	-1.2000...
3.73368...	1025237...	150000	99662	0	0.55760...	-1.0399...
2.87830...	3005125...	150000	3894	0	0.02849...	-0.6973...
2453.81...	7069306.0	150000	248	1	1.48437...	1.74038...
61.8575...	1207910	150000	40	0	1.15074...	1.07612...
3.09985...	260779.0	150000	8	4506	0.99151...	0.20688...
0.30104...	53114.0	150000	7	8680	1.59546...	5.87979...
0.17434...	32396.0	150000	2	5981	1.31750...	-0.2641...
31388.6...	17897482	150000	566	0	65.8625...	5733.25...
15.2303...	1898355	150000	12	0	-1.5473...	1.12992...
3554594...	3874615...	150000	7905	0	0.68887...	-0.3408...



数据预处理

我们通过 [派生列](#) 节点对notRepairedDamage字段中“-”值转换为空值处理，如图：

派生列配置

STRING

notRepairedDamage

DOUBLE

model

bodyType

fuelType

函数

累积

字符拼接(拼接符)

字符拼接

字符截取

小数点保留

添加/编辑表达式

```
1 case [notRepairedDamage] when '0.0' then 0
2 when '1.0' then 1
3 else null end
```

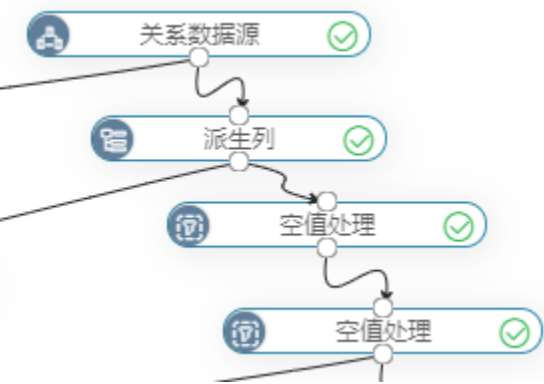
派生列名: notRepairedDamage_ 操作: 确定 取消

notRepairedDamage_ case [notRepairedDamage] when

确定

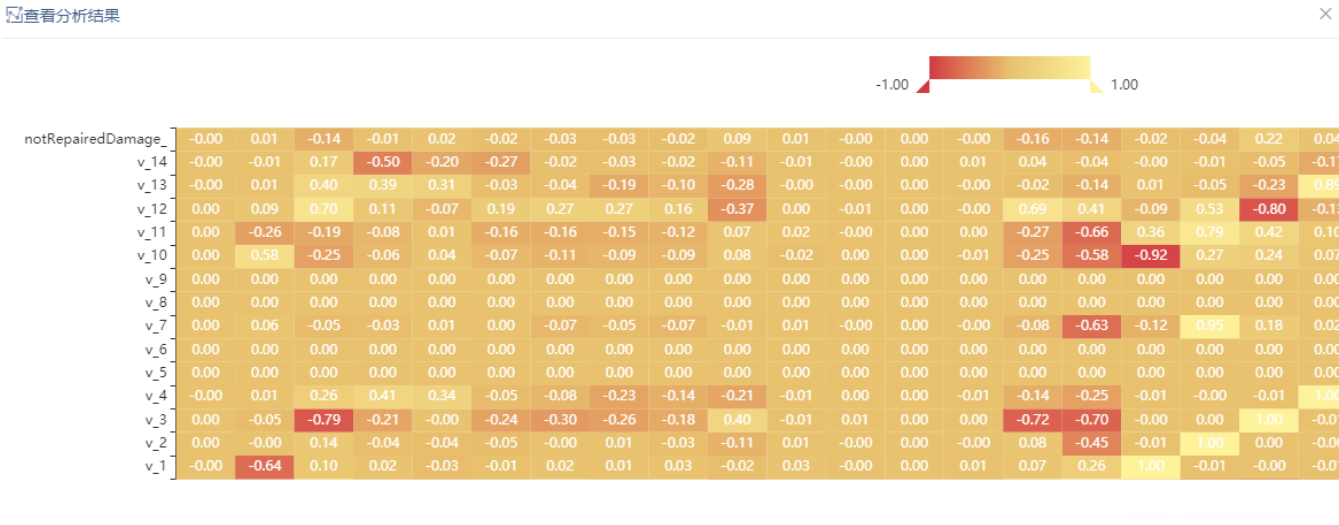
取消

根据全表统计查看缺失数据分布情况，对正态分布数据使用平均值填补，右偏分布的数据使用中位数填充处理。整个的数据预处理流程图如图：



相关性分析

通过 [相关性分析](#) 节点，相关性分析部分结果如图：



我们 [特征选择](#) 相关性系数较高的特征输入模型，特征选择如图：

选择特征列

☐ 源数据列表

0/14

☐ #SaleID

☐ # name

☐ # regDate

☐ # model

☐ # brand

☐ # bodyType

☐ 已选字段列表

0/18

☐ # gearbox

☐ # power

☐ # kilometer

☐ # v_0

☐ # v_1

☐ # v_2

< 到左边

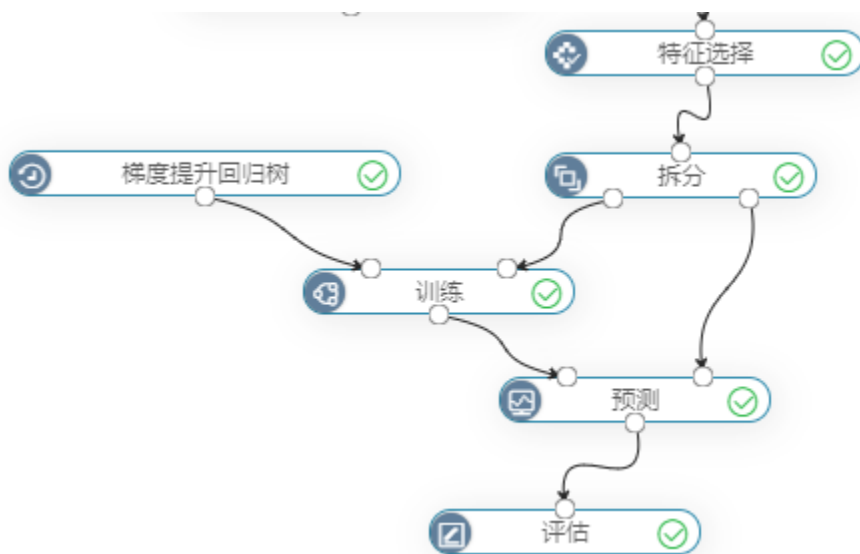
到右边 >

确定

取消

建立模型

本案例采用 [梯度提升回归树](#) 算法对模型进行训练，使用 [拆分](#) 节点按照7:3的比例将数据集拆分为训练集和测试集。整体的流程图如图所示：



算法参数配置如图：

» 参数

属性

帮助

归一化

方法选择

标准化

参数

☐ 单位标准差归一化

☐ 平均数据中心化

分裂特征的数量 范围是 ≥ 2 的整数 ?

150

树的深度 范围是 $[1, 30]$ 的整数 ?

10

计算信息增益的方式 ?

gini

模型训练后预测的结果如图：

	As featuresNormalized	# prediction
.0]	[0.0,68.0,5.0,2.0,3.0,1.0,1.0,-2.0,0.0,0.0,0.0,0.0,2.0,-2.0,1.0,2.0,44.0]	5924.658705689071
	(18,[1,2,5,6,7,16,17],[101.0,15.0,-1.0,1.0,-4.0,4.0,45.0])	3344.0900629776697
48.0]	[1.0,239.0,13.0,-2.0,0.0,3.0,-1.0,-4.0,0.0,0.0,0.0,0.0,1.0,-2.0,1.0,2.0,48.0]	13368.243024203146
])	(18,[1,2,3,5,7,14,15,16,17],[45.0,15.0,1.0,-3.0,4.0,2.0,-1.0,-3.0,42.0])	864.632965756282
2.0])	(18,[1,2,3,5,6,7,14,15,16,17],[54.0,15.0,1.0,-3.0,2.0,-1.0,2.0,-1.0,2.0,42.0])	837.4526134059266
.0]	[1.0,54.0,5.0,1.0,1.0,2.0,0.0,-2.0,0.0,0.0,0.0,0.0,1.0,-2.0,1.0,2.0,46.0]	7561.855827842441
45.0])	(18,[0,1,2,4,6,7,13,15,16,17],[1.0,218.0,10.0,-1.0,-3.0,2.0,-1.0,-1.0,-3.0,45.0])	3888.9037893882983
15.0])	(18,[1,2,4,5,6,7,13,14,16,17],[170.0,15.0,-2.0,-2.0,2.0,-4.0,-1.0,2.0,4.0,45.0])	2567.928089995176
46.0])	(18,[1,2,4,5,6,7,13,15,16,17],[163.0,15.0,-1.0,1.0,-3.0,2.0,-1.0,-1.0,-3.0,46.0])	4753.850295059237
46.0]	[1.0,163.0,13.0,0.0,-1.0,2.0,-3.0,2.0,0.0,0.0,0.0,0.0,-1.0,-1.0,0.0,-3.0,46.0]	7698.818363190019
47.0])	(18,[1,2,4,5,6,7,13,15,16,17],[133.0,10.0,-2.0,1.0,-3.0,2.0,-1.0,-1.0,-3.0,47.0])	5219.9202281859625

注意：表头中📌表示特征列，*表示标签列

表头真名 ☒ 表头别名

模型评估

接入 评估 节点对模型进行评估，评估结果如图所示（R2大概为0.94）：

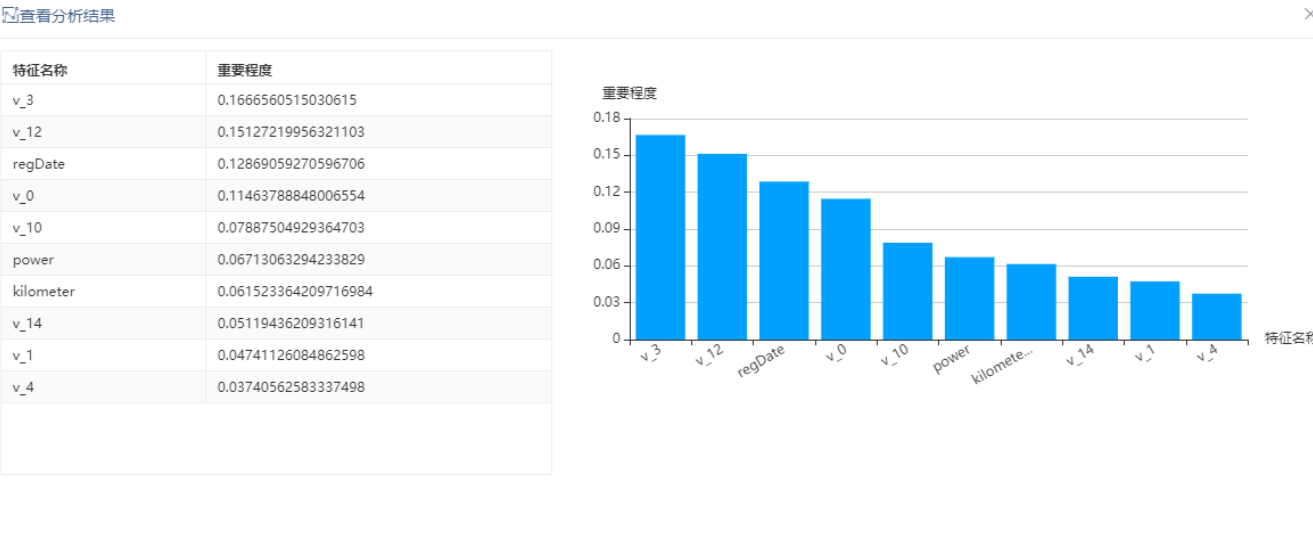
📊查看分析结果

×

指标	值
rmse(均方根误差)	1862.4855432429028
mse(均方误差)	3468852.398788811
r2(决定系数)	0.9378128158553526
mae(平均绝对误差)	918.3668907246604
adjusted r2(校正决定系数)	0.9377879718986186

业务分析

我们通过 GBDT特征选择 节点输出重要性较高的10个特征，结果如图：



分析发现v_3、v_12、regDate等几个特征是影响二手车交易价格的因素。

总结

本案例结合二手车交易价格预测案例，重点介绍了梯度提升回归树预测分析在实际案例中的应用。本案例借助二手车交易历史记录建立模型预测二手车交易未来价格，二手车价格市场；针对影响二手车交易价格因素，可以做好相关的优惠活动。