天池：零基础入门数据挖掘 - 二手车交易价格预测

杨福康*

2020 年 3 月 18 日

1 赛题介绍

1.1 赛题目标

根据二手车交易记录，预测一辆车子的成交价格

1.2 赛题评价

MAE(Mean Absolute Error)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{1}$$

绝对值，感觉不是很常见的赶脚。

1.3 赛题数据

1. 汽车信息
   - 车子的原始价值属性
     - model 车型编码，已脱敏
     - brand 汽车品牌，已脱敏
     - bodyType 车身类型：

*1766084780@qq.com

  - fuelType 燃油类型
  - gearbox 变速箱
  - power 发动机功率：范围 [0, 600]
- 车子的损害程度属性
  - name 汽车交易名称，已脱敏
  - kilometer 汽车已行驶公里，单位万 km
  - notRepairedDamage 汽车有尚未修复的损坏：是：0，否：1
  - regDate 汽车注册日期，例如 20160101，2016 年 01 月 01 日
  - offerType 报价类型：提供：0，请求：1
  - 上述，可合成车子的已使用时间。

2. 售卖地区与车主信息
   - seller 销售方：个体：0，非个体：1
   - offerType 报价类型：提供：0，请求：1
   - creatDate 汽车上线时间，即开始售卖时间
   - regionCode 地区编码，已脱敏
3. 匿名变量 15 种

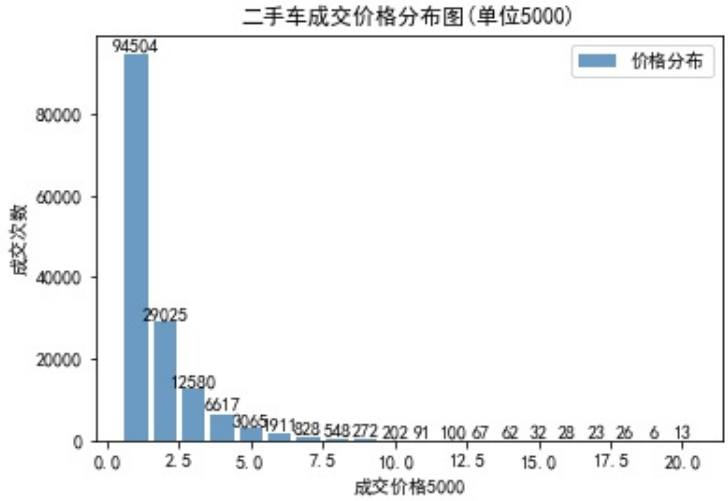
1.4 解题关键

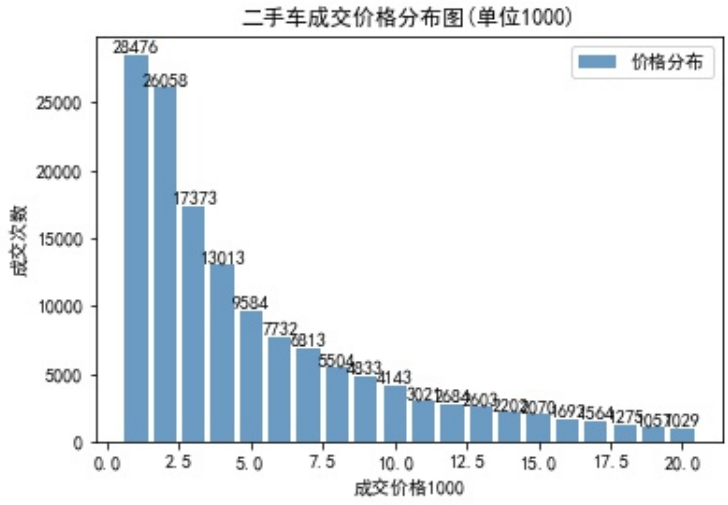
1. 我打算用 LGB，由于 LGB 采用决策树的方式，并不会组合特征，而是按照属性分类。所以，我需要找到合适的组合属性
2. 我打算自己按照赛题理解组合一些特征。比如，车子的使用年限
3. 用深度学习，组合一些特征。

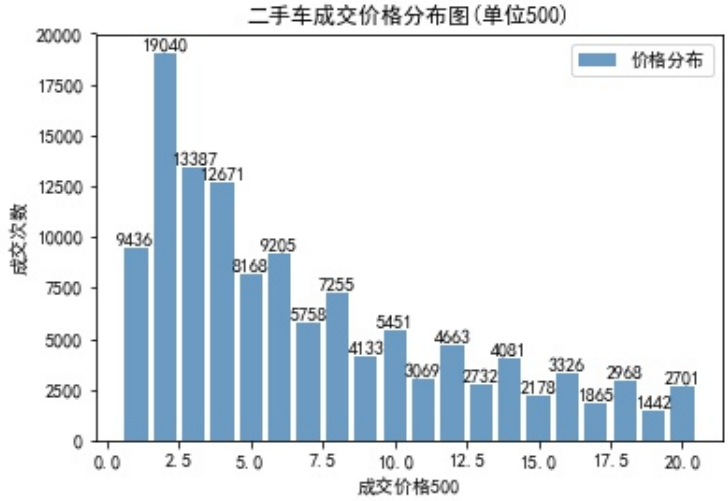
2 赛题理解

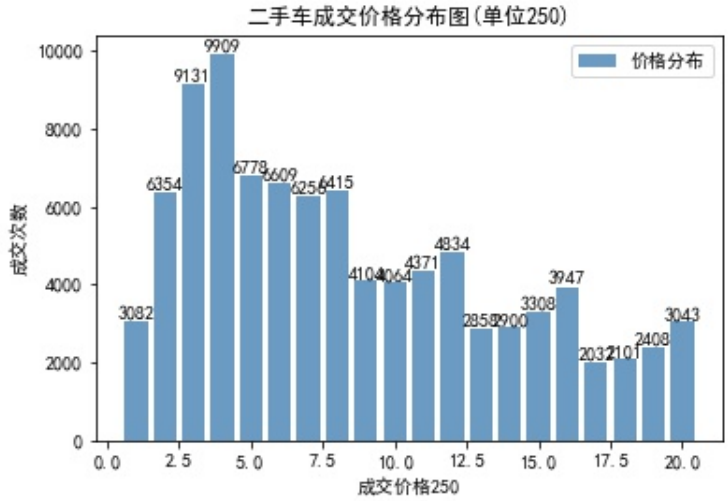
2.1 探索性数据分析 (EDA)

2.1.1 成交价格分布



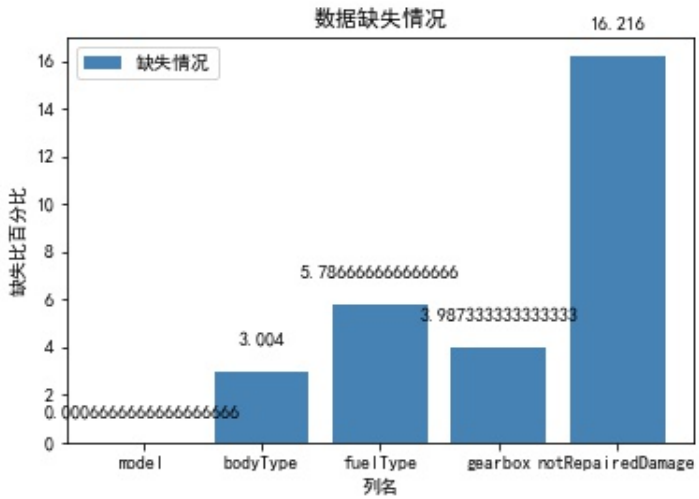






可以发现成交价格绝大多数小于 5000 5000 以内可以发现每 1000 是一个梯度，并且在最后一点卡在 9 的地方会稍微高一点点。大致可以认为是因为，国人喜欢整数预算买车费用。

2.1.2 数据缺失情况



可以发现数据缺失不是很严重，“汽车有尚未修复的损坏”项是缺失最严重的，有 16

2.1.3 地区分析

一共有 7905 个 regionCode。

销售的二手车数量：

```
 1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
91  92  93  94  95  96  97  98  99 101 102 103 104 105 106 107 109 110
111 112 113 115 116 117 118 120 125 126 129 130 132 134 136 137 258 369
```

有多少个地区销售这么多的二手车：

```
454 461 504 454 416 366 320 306 245 241 204 171 184 163 171 146 162 113
119 115  96  98 119  92  99  93  88  77  70  71  86  75  67  69  55  56
54  49  58  53  48  52  65  44  41  46  37  29  26  28  28  33  28  31
24  22  28  20  22  19  20  16  26  15  22  12   8  16  14  10   5  12
```

11	9	13	8	6	12	12	13	5	5	6	4	10	5	7	2	6	6
3	7	4	9	5	2	2	5	2	2	2	1	3	3	1	5	1	1
1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	1