二手车价格预测的总结-Task2

参考链接：https://blog.csdn.net/sliceoflife/article/details/105003927

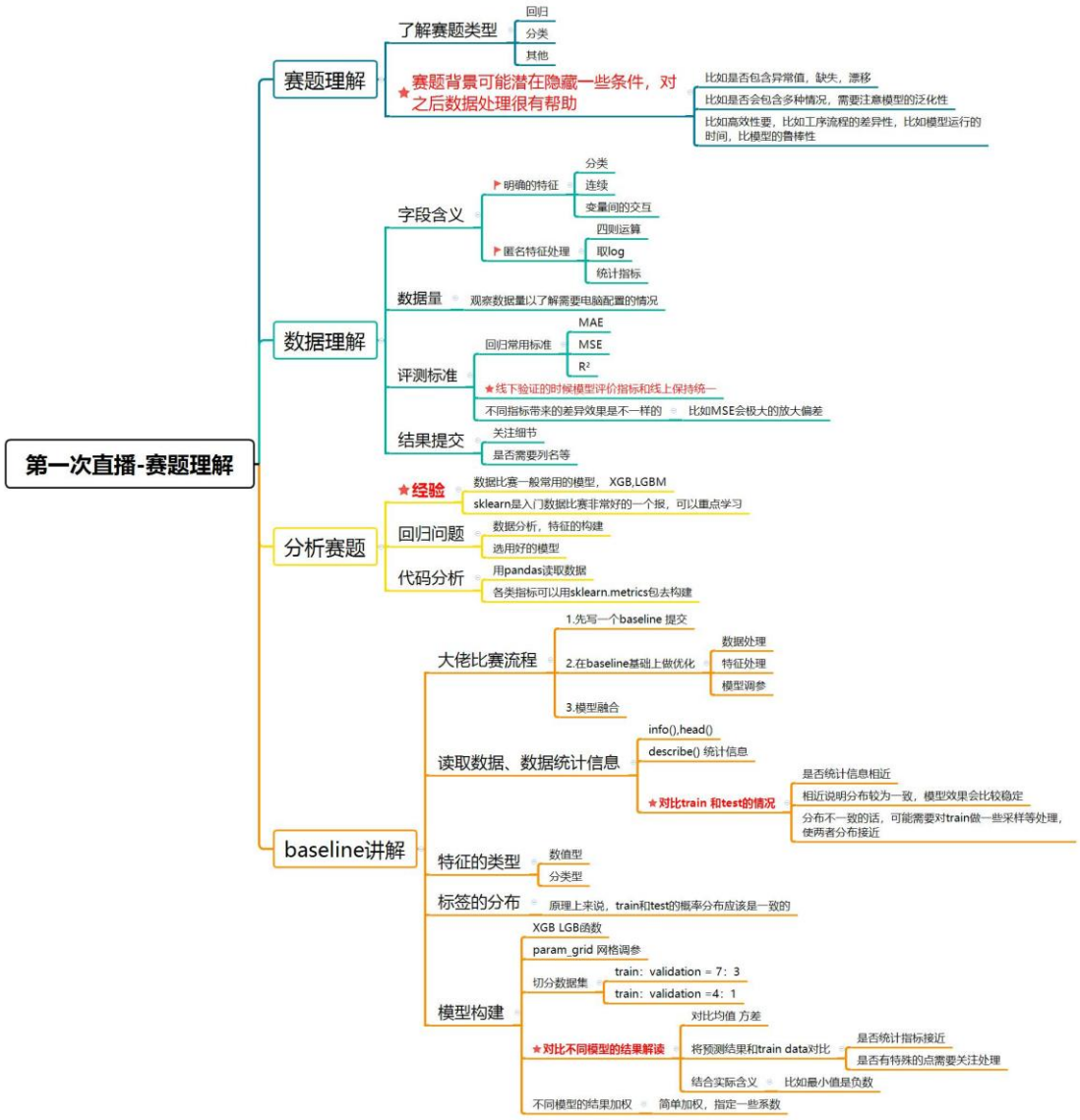
参考链接：https://blog.csdn.net/qq_41023769/article/details/105040287

参考链接：https://blog.csdn.net/weixin_43289424/article/details/105052609

参考链接：https://shimo.im/docs/RhQKcjD8qtw9XQYt/read

1.赛题理解

首先对赛题的了解，包括了先对赛题有一个直观地认知，然后要对数据进行观察分析，进行赛题的分析，之后先自己写一个 baseline 然后提交到网站上，进行对比分析，定位自己当前的水平，然后思考是否能提高。进一步的，组队调整模型，集合大家的模型开展研究分析。



目标：价格预测 -> price

数据来源：某交易平台的二手车交易记录

数据量：总数据量超过 40w，包含 31 列变量信息，其中 15 列为匿名变量，从中抽取 15 万条作为训练集，5 万条作为测试集 A，5 万条作为测试集 B

数据处理：已对 name、model、brand 和 regionCode 等信息进行脱敏。

评测标准：MAE(Mean Absolute Error)，MAE 越小说明模型预测得越精确。

结果提交：csv 格式（SaleID,price），与 sample_submit.csv 中的格式一致

————————————————————

关于二手车的详细的分析，可见 https://shimo.im/docs/RhQKcjD8qtw9XQYt/read
这个链接里，分析了影响二手车价格的

（1）硬性指标：车型，配置，车况，里程，上市时间

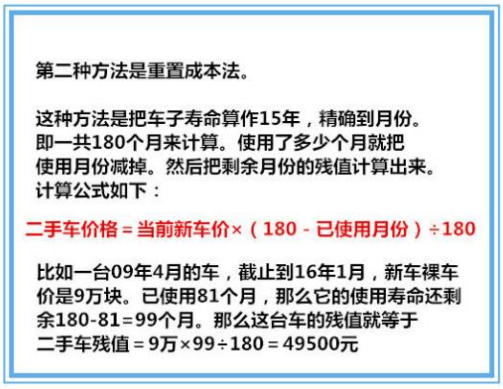
（2）软指标：违章次数、地域、新车价格、购车时机、购车渠道。

二手车常用的估价方法：

（1） 残差方法：计算公式为：评估价=市场现行新车售价×[15%（不动残值）+85%（浮动值）×（分阶段折旧率）]+评估值。

（2） 折旧的方法：

（3） 里程的方法：具体为：一部车有效寿命 30 万公里，将其分为 5 段，每段 6 万公里，每段价值依序为新车价的 5/15、4/15、3/15、2/15、1/15。假设新车价 12 万元，已行驶 7.5 万公里（5 年左右），那么该车估值为 12 万元×（3+3+2+1）÷15=7.2 万元



（4）

对汽车行业的统计数据进行了分析以及对二手车市场的销售进行了分析。

2 数据概况

- 数字全都脱敏处理，都为label encoding形式，即数字形式
- 数据概况介绍，描述列的性质特征

Field	Description
SaleID	交易ID，唯一编码
name	汽车交易名称，已脱敏
regDate	汽车注册日期，例如20160101，2016年01月01日
model	车型编码，已脱敏
brand	汽车品牌，已脱敏
bodyType	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	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	变速箱：手动：0，自动：1
power	发动机功率：范围 [0, 600]
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
regionCode	地区编码，已脱敏
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
price	二手车交易价格（预测目标）
v系列特征	匿名特征，包含v0-14在内15个匿名特征

赛题的相关变量分析：

车型：私用和商用

私用：基本型 >> SUV > MPV（7-8 人的车）

商用：客车 >> 货车

交易地区：广东 > 浙江 > 山东

车龄：3-6 > 3 > 7-10

车型：A >> B > A0 > A00 > C > D （后两者可理解为豪华轿车）

https://baike.baidu.com/item/A00%E7%BA%A7%E8%BD%BF%E8%BD%A6

价格区间：3 > 3-5 > 5-8 > 8-12

交易均价：17-19：6.53 - 6.22 - 6.30

新能源汽车的分类：https://baike.baidu.com/item/%E6%96%B0%E8%83%BD%E6%BA%90%E8%BD%A6

交易车型：A00 >> A > SUV(D?) > A0

价格：3-5 > 5-8 > 3 > 8-12

车龄：2 >> 2-4 > 4-6

其他

作者给出的结论：

1. 私用车和商用车可以拆开分析（题目中主要区分开车型）
2. 车型中，小型车如 B 以下的可单独分析（题目中为微型车）
3. 新能源车和燃油车可拆开分析（题目中主要用以区分燃油类型）
4. 华东地区是二手车交易的主要地区
5. 一些有联系的字段：车型-发动机功率-价格，行驶公里-注册日期

评测标准

平均绝对误差 MAE

极值带来的误差影响会非常大。**对于 price 的离群点可能要做专门的分**析

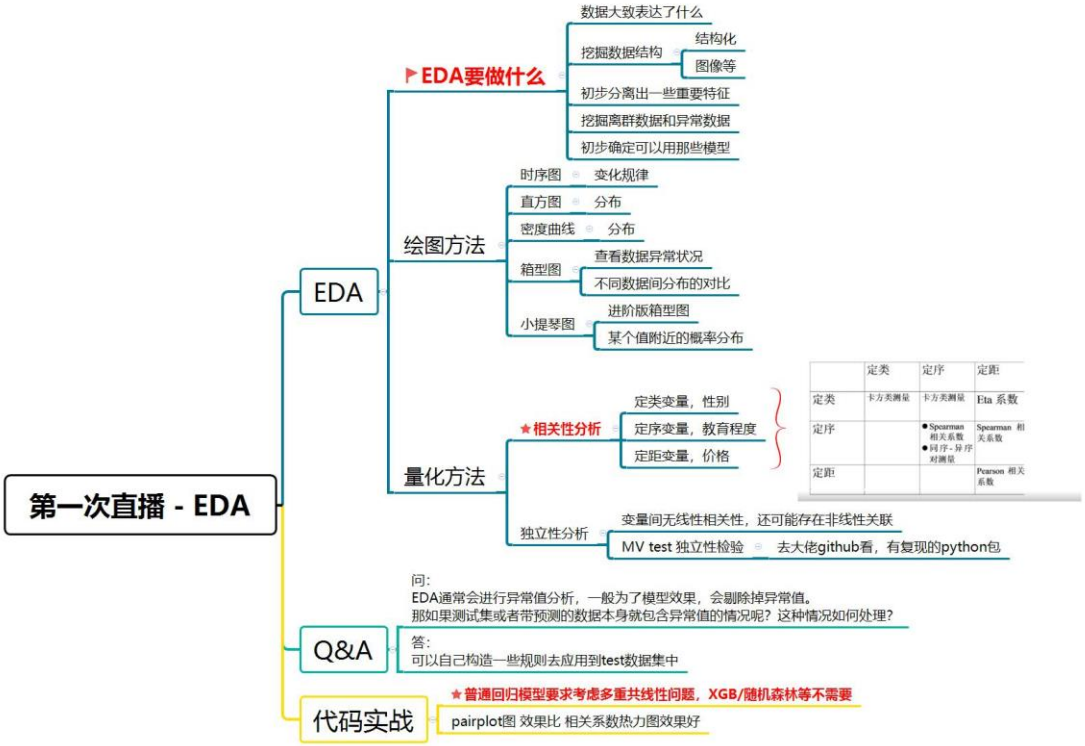
回归一般需要注意的问题

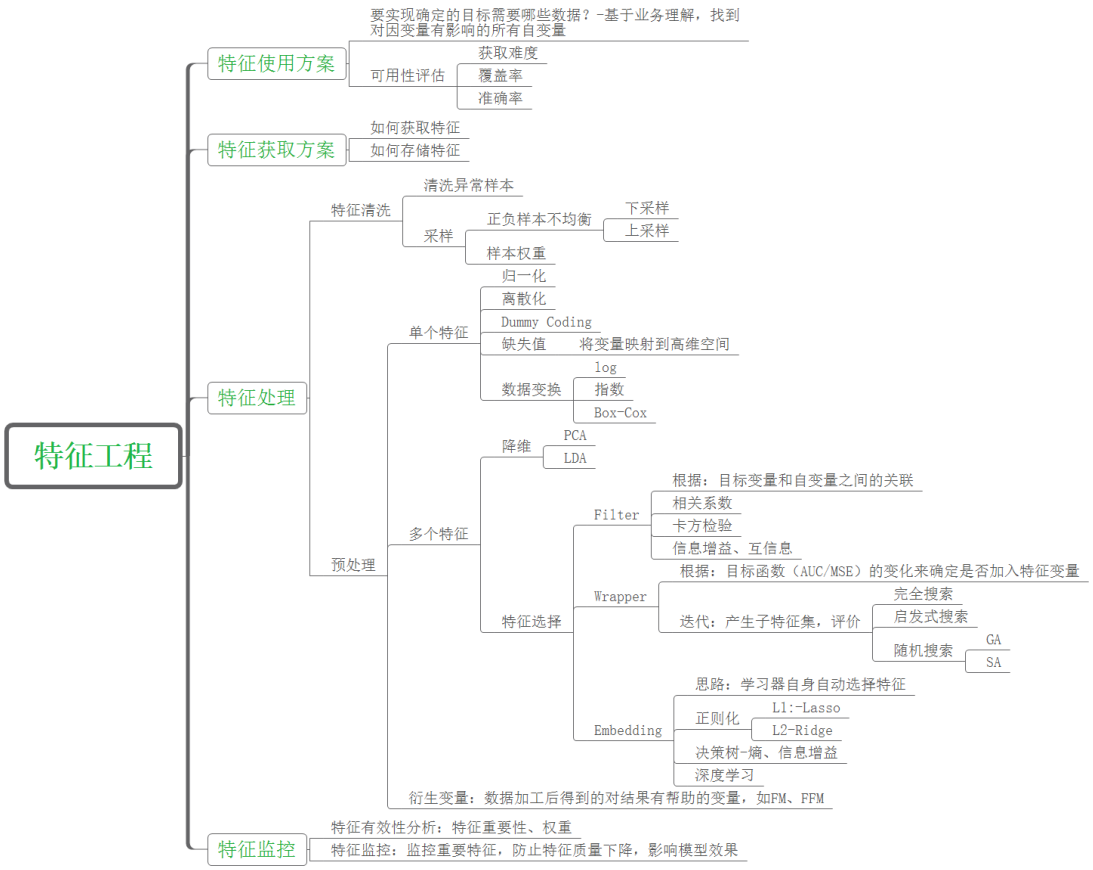
极值的处理

多重共线性的问题

问题：对于 creatdata 变量是汽车开始售卖的时间还是汽车被卖出去的时间呢？因为我想知道，这辆车从开始买入到卖出时的时间间隔，才可以有效地对汽车的年限进行评估。

2.数据的探索性分析：


第一次直播 - EDA
EDA
EDA要做什么
数据大致表达了什么
挖掘数据结构
结构化
图像等
初步分离出一些重要特征
挖掘离群数据和异常数据
初步确定可以用那些模型
绘图方法
时序图
变化规律
直方图
分布
密度曲线
分布
箱型图
查看数据异常状况
不同数据间分布的对比
小提琴图
进阶版箱型图
某个值附近的概率分布
量化方法
★相关性分析
定类变量，性别
定序变量，教育程度
定距变量，价格
独立性分析
变量间无线性相关性，还可能存在非线性关联
MV test 独立性检验
去大佬github看，有复现的python包
定类
定序
定距
Eta 系数
Spearman 相关系数
Pearson 相关系数
Q&A
问:
EDA通常会进行异常值分析，一般为了模型效果，会剔除掉异常值。
那如果测试集或者带预测的数据本身就包含异常值的情况呢？这种情况如何处理?
答:
可以自己构造一些规则去应用到test数据集中
代码实战
★普通回归模型要求考虑多重共线性问题，XGB/随机森林等不需要
pairplot图 效果比 相关系数热力图效果好


特征工程
特征使用方案
要实现确定的目标需要哪些数据？-基于业务理解，找到对因变量有影响的所有自变量
可用性评估
获取难度
覆盖率
准确率
特征获取方案
如何获取特征
如何存储特征
特征处理
特征清洗
清洗异常样本
采样
正负样本不均衡
下采样
上采样
样本权重
预处理
单个特征
归一化
离散化
Dummy Coding
缺失值
将变量映射到高维空间
数据变换
log
指数
Box-Cox
多个特征
降维
PCA
LDA
特征选择
Filter
根据：目标变量和自变量之间的关联
相关系数
卡方检验
信息增益、互信息
Wrapper
根据：目标函数（AUC/MSE）的变化来确定是否加入特征变量
迭代：产生子特征集，评价
完全搜索
启发式搜索
随机搜索
GA
SA
Embedding
思路：学习器自身自动选择特征
正则化
L1:-Lasso
L2-Ridge
决策树-熵、信息增益
深度学习
衍生变量：数据加工后得到的对结果有帮助的变量，如FM、FFM
特征监控
特征有效性分析：特征重要性、权重
特征监控：监控重要特征，防止特征质量下降，影响模型效果

https://www.zhihu.com/question/29316149/answer/110159647

2.1 什么是 EDA：

探索性数据分析（Exploratory Data Analysis，简称 EDA），是指对已有的数据（特别是调查或观察得来的原始数据）在尽量少的先验假定下进行探索，通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。特别是当我们面对各种杂乱的“脏数据”，往往不知所措，不知道从哪里开始了解目前拿到手上的数据时候，探索性数据分析就非常有效。探索性数据分析是上世纪六十年代提出，其方法有美国统计学家 John Tukey 提出的。

其价值在于：熟悉数据、了解数据，对数据集进行验证，以确定可以使用。如在预测中：了解变量与预测值之间关系；了解数据中的模式、趋势、异常值、相关性等；使用可视化和定量方法来了解数据所具有的特性；为之后的数据预处理和特征工程提供必要的理论依据。

2.2 EDA 分析的过程：

（1）导入数据的时候，注意数据的格式，因此要灵活应用 pandas。
（2）数据总览: info()、describe()，isnull() 通过对数据的总览，有每列的统计量，个数 count、平均值 mean、方差 std、最小值 min、中位数 25%　50%　75% 、以及最大值 看这个信息主要是瞬间掌握数据的大概的范围以及每个值的异常值的判断，比如有的时候会发现 999 9999 -1 等值这些其实都是 nan 的另外一种表达方式，有的时候需要注意下。

其次，对均值，方差，中位数等参数的理解，来分析数据是否存在无用的序列或者也包括是否存在数据的倾斜等。

其次，对于 nan 的值可以进行展示，并可以运用替换的方法进行替换。
（3）对预测值的倾斜分布进行探索，如果数据分布倾斜的厉害，则一般对预测没什么用。常常对数据结果的分布情况进行查看，观察其是否符合总体分布概况，如无界约翰逊分布，正常分布，对数变换后的分布等。
（4）对预测值的偏度与峰度进行探索。偏度：Skewness > 0 ，正偏差数值较大，为正偏或右偏。长尾巴拖在右边，数据右端有较多的极端值。Kurtosis>0 比正态分布的高峰更加陡峭——尖顶峰。Skewness < 0 ，负偏差数值较大，为负偏或左偏。长尾巴拖在左边，数据左端有较多的极端值。数值的绝对值越大，表明数据分布越不对称，偏斜程度大。峰度描述某变量所有取值分布形态陡缓程度的统计量，简单来说就是数据分布顶的**尖锐程度**。峰度是四阶标准矩计算出来的。（1）Kurtosis=0 与正态分布的陡缓程度相同。（2）Kurtosis>0 比正态分布的高峰更加陡峭——尖顶峰（3）Kurtosis<0 比正态分布的高峰来得平台——平顶峰。
（5）查看预测值的频数，对于大于某个值很少的数据，当做是特殊值予以填充或删除。
（6）特征分布的探索，unique 分布，箱型图，小提琴图，柱形图，可视化每个类别的频数，
（7）相关性分析的分析，特征的偏度与峰度，数字特征分布可视化，数字特征的关系可视化，多变量互相关系可视化。

2.3 代码函数：

（1）常用的各种数据科学以及可视化库:
　　数据科学库 pandas、numpy、scipy；
　　可视化库 matplotlib、seabon、missingno；
　　其他；
（2）载入数据：
　　载入数据集：读取 csv 文件使用 pd.read_csv()，注意 read_csv 默认是逗号分割，这里要设置 sep= ’ ’
　　简略观察数据：head()、tail()、shape、dtypes
（3）数据总览: info()、describe()
　　通过 info()了解数据信息：数据列、数据缺失情况、数据类型
　　通过 info 有助于了解是否存在除了 nan 以外的特殊符号异常
　　通过 describe()来熟悉数值型数据的相关统计量：count、std、min、25%、50%、75%、max
　　通过 describe(include=[‘O’])来熟悉非数值型数据的相关统计量：count、unique、top、freq
　　看 describe()这个信息主要是瞬间掌握数据的大概的范围以及每个值的异常值的判断
　　有时候，会发现 999 9999 -1 等值这些其实都是 nan 的另外一种表达方式，需要注意下
　　有时候，只有两类的类别，却统计出三类…
（4）判断数据缺失和异常

查看每列的存在 nan 情况：isnull()、isnull().any() 、isnull().sum() --True-1、False-0

异常值检测：info()、values_counts()

缺失值和异常值处理：replace()、fillna()、del

（5）了解预测值的分布:scipy.stats

总体分布概况：正态分布、无界约翰逊分布、对数正态分布等

查看 skewness and kurtosis：偏度系数、峰度系数，一般与正态分布 0 比较

查看预测值的具体频数：hist

长尾型数据处理：log、Box-Cox

（6）数字特征分析

相关性分析：corr()-相关系数

查看数字特征的偏度系数和峰度系数

数字特征可视化：单个数字特征分布可视化、数字特征相互之间的关系可视化（热图-heatmap()）

（7）类型特征分析

unique 分布：nuinque()可以查看有多少个不同值

类别特征可视化：箱形图、小提琴图、柱形图、直方图

（8）时间特征分析：

（9）生成数据报告：

https://www.zhihu.com/question/24590883/answer/782584888

```
import pandas_profiling
pfr = pandas_profiling.ProfileReport(Train_data)
pfr.to_file("example.html")
```

特征工程中常用的目标对象所注意的事项:

（1） 时间戳的处理，往往数据中包含了年月日时分秒等信息，此时要注意一下时间的选取要满足模型的需要，而不是一味的追求运用所有变量。对于跨时区的时间，要注意标准化时间。

（2） 分解类别属性，独热编码，可以很好的降低运算量。

（3） 分箱与分区，有时候，将数值型属性转换成类别呈现更有意义，同时能使算法减少噪声的干扰，通过将一定范围内的数值划分成确定的块。要注意进行分区的特点是能保证落入一个分区的数据能表现出共同的特征。

（4） 特征的交叉，通过将单个特征组合成交叉特征，确保交叉特征能表现出协同的作用、

（5） 特征的选择，例如用到评分的方法或相关性来搜索出特征子集。

（6） 特征的缩放，如一个人的年龄和他对应的收入，当如岭回归要求必须将特征缩放到相同的范围内时，通过缩放可以避免某些特征比其他特征获得大小非常悬殊的权重值。

（7） 特征的提取：

2.4 EDA 分析结果：

（1）3 个特征类别：日期特征、类别特征（含字符型 notRepairedDamage）、数值特征

（2）5 个类别含缺失值：model、bodyType、fuelType、gearbox 、notRepairedDamage

（3）distplot()画图看下 price 数据情况，发现存在长尾形式（考虑做下 log 转换）

（4）offerType 报价类型 就一个值，可以删了
（5）seller 销售方 虽然有两个类别，但最大次数占比达到 0.99，另一个类别就一个值，样本太不均衡了，考虑删除
（6）notRepairedDamage 本来两个类别 0、1，但是含有-类别，要处理

（7）creatData_year 基本都是 2016 年的,那这个特征基本没什么用了
（8）creatDate_month: 基本都是 3\4 月份,可先留着
（9）热图查看各特征之间关系：
（10）跟 price 相关性比较高(颜色深)的有汽车注册年份（regDate_year）、V_0 、V_3、 V_8、V_12、 汽车已行驶公里数（kilometer）；
（11）注册年份,应该可以理解为车越新，价格越高；应该可以理解为跑的路程越多，车就越旧，价格就越低；
（12）而公里数，则直接影响汽车的寿命。
（13）除了跟 price 的相关性，有些特征相互之间的相关性也很高，这些特征之间可能存在冗余现象，训练的时候可以依据效果尝试去掉一部分，或者拆分成两部分，做模型融合。

比如:

V_1 跟 V_6、V_10
V_2 跟 V_5、V_7、V_11
V_3 跟 V_8
V_4 跟 V_9、V_13 等
匿名数据分布都还可以，先不考虑