二手车价格预测

2020年5月30日

21:20

Field 
SalelD 
name 
regDate 
model 
brand 
bodyType 
fuel Type 
gearbox 
kilometer 
notRepairedDamage 
regionCode 
seller 
offerType 
creatDate 
prlce 
Description 
*020160101, 
F.æhk*, 
2, *-æ: 3, 4, xqn;5$: 5, 6, 7 
[0, 600] 
Exe, 
GS)5: o, 1 
iÆ#t: 0, 1 

* 数据读取
* 训练集测试集合并
* 数据存储优化
* power ,v\_13，v\_14异常值处理 去掉对v\_13，v\_14的处理）
* 15个匿名特征间和交叉 去掉相关性不大的特征（‘v\_2','v\_6','v\_1','v\_13','v\_14','v\_7')
* 15个匿名特征和'model','brand', 'bodyType', 'fuelType','gearbox', 'power', 'kilometer', 'notRepairedDamage', 'regionCode'进行乘交叉

去掉相关性不大的特征（‘v\_2','v\_6','v\_1','v\_13','v\_14','v\_7')

* 提取'regDate', 'creatDate'两个日期信息
* “对类别少的'fuelType','gearbox','notRepairedDamage','bodyType','creatDate\_year'进行one-hot编码“
* 'regDate', 'creatDate', 'model', 'brand', 'regionCode','bodyType','fuelType','name','regDate\_year', 'regDate\_month', 'regDate\_day',

'regDate\_dayofweek' , 'creatDate\_month','creatDate\_day', 'creatDate\_dayofweek','kilometer'进行count编码 改成目标编码

* 'used\_time1'使用时间；'used\_time2'注册距离现在的时间；'used\_time3'开售距离现在的时间；并对三个时间进行分桶编码
* 数值特征对类别特征做统计刻画，随便挑了几个跟price相关性最高的匿名特征（'min''max''median'）

cross\_cat = ['model', 'brand','regDate\_year']

cross\_num = ['v\_0','v\_3', 'v\_4', 'v\_8', 'v\_12','power','v\_10','v11','kilometer']

* 去掉无用特征['SaleID','offerType','seller']
* 训练集测试集分离
* 'model','brand','name','regionCode'进行平均数编码
* 对'regionCode','brand','regDate\_year','creatDate\_year','kilometer','model'进行目标编码（'max','min','mean'）
* 丢掉没有用的特征（'regDate', 'creatDate','brand\_power\_min', 'regDate\_year\_power\_min'）
* MinMaxScaler特征归一化