二手汽车交易价格预测——基于机器学习技术

摘要：

随着我国汽车市场的发展，汽车保有量不断增加。个体消费观念的转变也让购买使用二手车的行为逐步被消费者接受，二手车市场也随之更加繁荣。然而由于没有一个科学的二手车价格评估体系，影响了二手车产业链的健康发展。为了使得二手车电商、车辆制造企业以及购买者进行快速准确的给二手车定价、评估资产风险、达成交易共识，迫切需求一种更科学更准确的估价模型。

二手汽车的交易价格受到多方面的影响，从微观上来讲，主要从车辆本身的属性，即车型参数、维修情况、区域因素、行驶路程等等分析对二手车价格的影响。因此，文本选自 B2C 二手车电商交易网站的公开数据集，本着数据可得且能够反映影响因素的原则，最终选取了三十一个数据字段作为二手车价格评估模型的输入数据，其中十五个数据因为涉及到隐私，进行了脱敏化处理。

本文基于机器学习方法建立了良好的二手车价格预测模型，采用 LinearRegression，Ridge，Lasso 三种线性回归模型和随机森林非线性回归模型对二手车交易价格进行回归预测，总结出影响交易价格最重要的四个特征为 used_time（车龄），notRepairedDamage（是否尚未修复），kilometer（里程数），power_bin（汽车功率），并将线性与非线性模型进行比较，得出最适合的模型是随机森林回归模型，为二手车市场交易提供一个很好的价格指导，为规范二手车市场、保护消费者权益贡献了力量。

1.研究目标

从二手车本身性质来看，由于其属于非标品，具有“一车一价”的特点。二手车的价格，会受到车辆使用强度、养护情况、使用区域、品牌溢价、多方面的影响，交易价格也会有很大的波动。而二手车市场本身存在交易双方信息不对称的问题，不同的交易机构采取不同的交易模型，导致市场二手车定价混乱，消费者难以相信二手经销商，买卖双方交易效率低下，对双方都造成了损失。因此，构建一套科学完善的二手汽车交易价格预测模型对于维护双方利益，促进市场蓬勃发展有重要的意义。

因此，文本的研究目的是通过输入一辆二手车的相关信息，以此尽可能准确

的预测出这辆二手车价格，作为对消费者乃至经销商的参考，能够做到搭建一个模型仅在线上就完成对“一车一价”的精准预测，从而推进二手汽车市场蓬勃发展。

2. 研究问题

问题分析旨在明确研究问题，为后续方法选取提供依据，通过查找文献资料，明确研究问题，理清研究思路。

➢明确问题

本题属于真实工业场景的二手汽车交易问题，目的在于对于给定的二手车成交时的信息，来预测二手车成交时的价格，通过翻阅资料，发现影响二手车成交价格的因素主要来自宏观政治，社会以及微观车型，车况，区域等因素。

➢变量分析

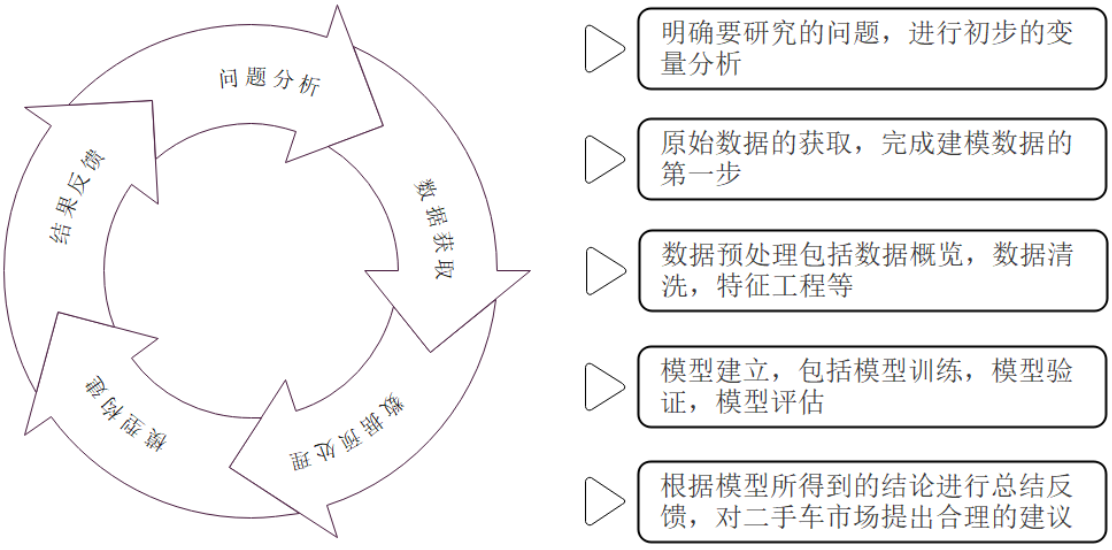
本课题的因变量是二手车成交价格，自变量是车况，车型等因素，可以初步确定数据中所给出的变量与预测价格之间的关系：

表 1 因素初步分析结果

因素变量	车身类型	汽车功率	已行驶公里
对成交价格影响	正向	正向	负向
因素变量	汽车有尚未修复的损坏	使用时间	品牌
对成交价格影响	负向	负向	正向

3. 研究思路

本文的研究思路明确，研究过程见下图。



4.数据来源与收集

本课题采用的是真实工业场景数据，数据来源是 B2C 二手车电商交易网站。数据集一共有三十一列，其中一列为最终二手车成交价格，其余三十列均为二手车的特征数据，包括：

- SaleID - 销售 ID，唯一编码
- name – 汽车交易名称
- regDate - 汽车注册时间
- model - 车型编码
- brand - 品牌
- bodyType - 车身类型
- fuelType - 燃油类型
- gearbox - 变速箱
- power - 汽车功率
- kilometer - 汽车已行驶公里，单位万 km
- notRepairedDamage - 汽车有尚未修复的损坏
- regionCode - 看车地区编码
- 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）

因为数据来源是真实数据场景，所以该网站已经初步对包括汽车交易名称，车型编码，汽车品牌等涉及交易隐私的数据做了脱敏处理，还有十五列包含二手车辆的评价，概况等描述性文本信息，该网站已经进行了简单的处理，将其做了词向量 word embedding 处理，都为 label encoding 形式，最后一列是因变量即要预测的二手车成交价格。

5.数据预处理

5.1 探索性数据分析

所谓探索性数据分析，也就是 EDA，其价值主要在于熟悉数据集，了解数据集，这个过程将涉及数据质量分析（如缺失值，重复值，异常值，正负样本不平衡等），数据统计量分析（集中趋势，分散程度，描述统计），数据分布分析。对数据集进行验证来确定所获得数据集可以用于接下来的机器学习或者深度学习使用。当了解了数据集之后我们下一步就是要去了解变量间的相互关系以及变量与预测值之间的存在关系。引导本文进行数据处理以及特征工程的步骤,使数据集的结构和特征集让接下来的预测问题更加可靠。

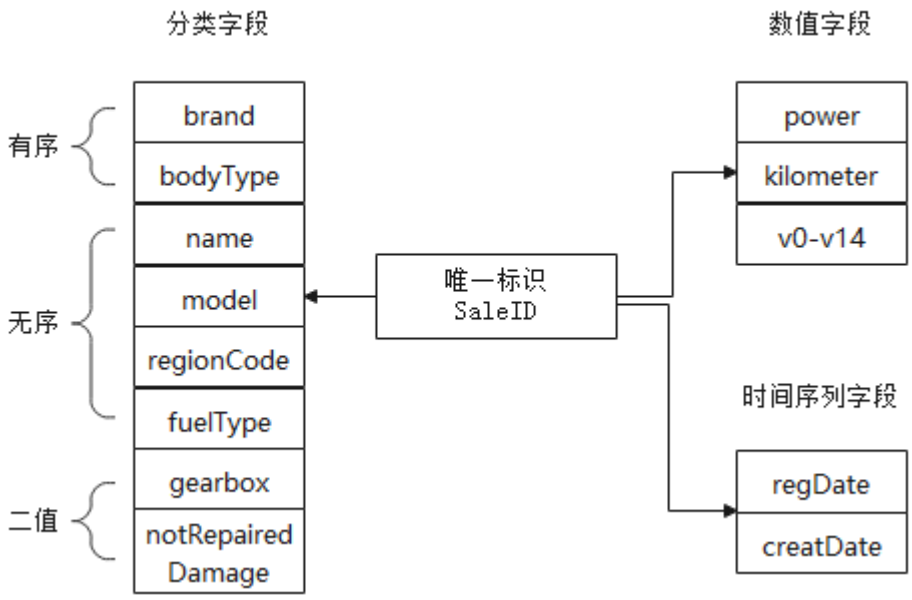
5.1.1 原始数据概览

	SaleID	name	regDate	model	brand	bodyType	fuelType	gearbox	power	kilometer	...	v_5	v_6	v_7	v_8	v_9	v
0	0	736	20040402	30.0	6	微型车	汽油	手动	60	12.5	...	0.235676	0.101988	0.129549	0.022816	0.097462	-2.881
1	1	2262	20030301	40.0	1	厢型车	汽油	手动	0	15.0	...	0.264777	0.121004	0.135731	0.026597	0.020582	-4.900
2	2	14874	20040403	115.0	15	微型车	汽油	手动	163	12.5	...	0.251410	0.114912	0.165147	0.062173	0.027075	-4.846
3	3	71865	19960908	109.0	10	豪华轿车	汽油	自动	193	15.0	...	0.274293	0.110300	0.121964	0.033395	0.000000	-4.509
4	4	111080	20120103	110.0	5	微型车	汽油	手动	68	5.0	...	0.228036	0.073205	0.091880	0.078819	0.121534	-1.896
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
149995	149995	163978	20000607	121.0	10	敞篷车	汽油	自动	163	15.0	...	0.280264	0.000310	0.048441	0.071158	0.019174	1.988
149996	149996	184535	20091102	116.0	11	豪华轿车	汽油	手动	125	10.0	...	0.253217	0.000777	0.084079	0.099681	0.079371	1.839
149997	149997	147587	20101003	60.0	11	微型车	柴油	手动	90	6.0	...	0.233353	0.000705	0.118872	0.100118	0.097914	2.439
149998	149998	45907	20060312	34.0	10	大巴车	柴油	手动	156	15.0	...	0.256369	0.000252	0.081479	0.083558	0.081498	2.075
149999	149999	177672	19990204	19.0	28	商务车	汽油	自动	193	12.5	...	0.284475	0.000000	0.040072	0.062543	0.025819	1.978

由上表可知，'name','model','brand','bodyType','fuelType','gearbox','notRepairedDamage','regionCode'等字段均为分类变量，'power', 'kilometer','v_0','v_1','v_2','v_3','v_4','v_5','v_6','v_7','v_8','v_9','v_10','v_11','v_12','v_13','v_14'等字段为数值变量，'regDate','creatDate'为时间序列变量，因变量 price 为连续性数值变量，因此，本研究问题是一个多因素回归预测问题。

通过对问题的整体把握，下面将所有特征整理成下表 2 的形式，以便于理解。

表 2 字段初识




- ```
  Train_data.shape
  ```

```
(150000, 31)
```
```

通过 Train_data.shape 可以看出训练数据一共有 150000 行，31 列，通过训练数据对最终的模型进行训练，将测试数据在训练好的模型上进行测试可得最终的预测结果。

```
Train_df.describe()
```

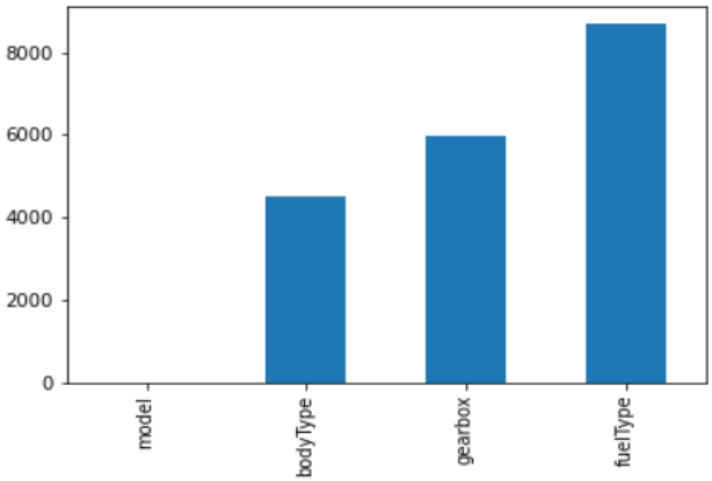
	SaleID	name	regDate	model	brand	bodyType	fuelType	gearbox	power	kilometer
count	150000.000000	150000.000000	1.500000e+05	149999.000000	150000.000000	145494.000000	141320.000000	144019.000000	150000.000000	150000.000000
mean	74999.500000	68349.172873	2.003417e+07	47.129021	8.052733	1.792369	0.375842	0.224943	119.316547	12.597160
std	43301.414527	61103.875095	5.364988e+04	49.536040	7.864956	1.760640	0.548677	0.417546	177.168419	3.919576
min	0.000000	0.000000	1.991000e+07	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000
25%	37499.750000	11156.000000	1.999091e+07	10.000000	1.000000	0.000000	0.000000	0.000000	75.000000	12.500000
50%	74999.500000	51638.000000	2.003091e+07	30.000000	6.000000	1.000000	0.000000	0.000000	110.000000	15.000000
75%	112499.250000	118841.250000	2.007111e+07	66.000000	13.000000	3.000000	1.000000	0.000000	150.000000	15.000000
max	149999.000000	196812.000000	2.015121e+07	247.000000	39.000000	7.000000	6.000000	1.000000	19312.000000	15.000000

➢通过 describe()函数，对训练集中的所有特征进行了简单的统计分析描述，可以初步得到一些有趣的信息，bodyType 字段的中位数是 1，可见出售的大部分汽车还是较为平民化的二手汽车，kilometer 的里程数平均数为 12.59 万 km，平均使用程度足够老化等等，为接下来的实证分析提供了初步的验证[1]。

5.1.2 数据质量分析

➢缺失值分析[2]

	model	bodyType	fuelType	gearbox
数据类型	float64	float64	float64	float64
列是否缺失	True	True	True	True
缺失个数	1	4506	8680	5981
缺失比例	6.66667e-06	0.03004	0.0578667	0.0398733



[1] 因为 describe()不便对分类数据做数值化统计，所以这里的统计过程是在字符串型分类数据 label encoding 之后做的分析

[2] 同脚注 1

图 1 缺失特征统计

从直观的表格和图形的形式展示了训练数据集的缺失数据，本数据集存在四个缺失字段，其中 model 字段仅缺失一个值，用众数填充是很合理的建议，bodyType，fuelType，gearbox 三个字段均有不同程度的缺失值，数据缺失规模均不超过 6%，缺失规模不大，不建议删除，考虑直接填充。（这边对于缺失数据的处理也要依据所选机器学习模型的特性，基于决策树的集成模型 LightGBM 对缺失数据自训练，所以可以考虑不填充）

➢异常值分析

```
Test_df.info()
```

```
10  notRepairedDamage  50000 non-null  object
```

通过 info()函数发现，其中有一个字段 notRepairedDamage 是 object 类型，其余均为 float 浮点数类型，说明其中除了包含数字，还有别的字符串型符号，很有可能存在缺失数据。

```
Train_df['notRepairedDamage'].value_counts()
```

```
0.0     111361
-        24324
1.0      14315
```

“-”以此符号填充，表示数据缺失，缺失比例过高。分析至此，总结一共出现五个缺失字段：model，bodyType，fuelType，gearbox，notRepairedDamage。

```
Train_data["seller"].value_counts()
```

```
0     149999
1          1
Name: seller, dtype: int64
```

```
Train_data["offerType"].value_counts()
```

```
0     150000
Name: offerType, dtype: int64
```

“seller”和“offerType”两个字段存在严重的正负样本不平衡的问题，考虑本研究问题的背景，之后删除这两个字段。

通过以下箱线图观察异常值分布，可以看出异常值主要分布在 power 和 kilometer 字段，为接下来的特征工程提供了依据。

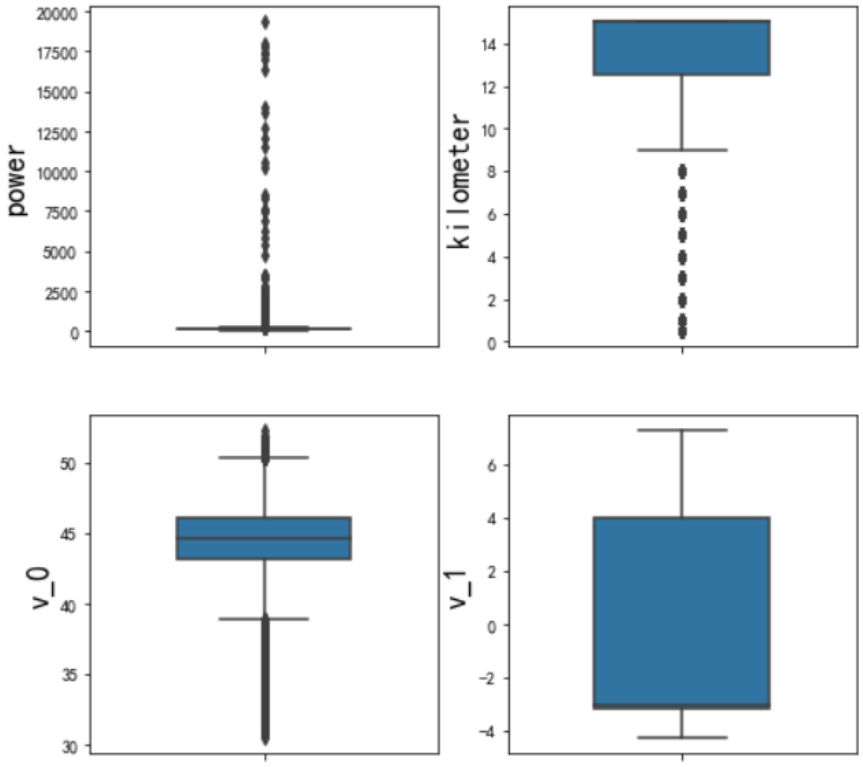


图 2 异常值分析

5.1.3 数据分布分析

➢ 预测值分布

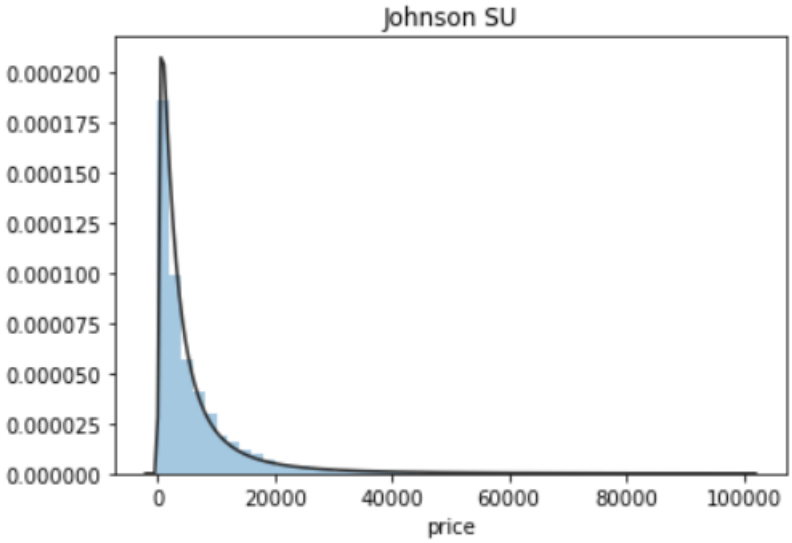


图 3 预测值分布

由图可以看出二手汽车的交易价格符合无界约翰逊分布，价格不服从正态分布，所以在进行回归之前，它必须进行转换。虽然对数变换很合适，但最佳拟合还是无界约翰逊分布。

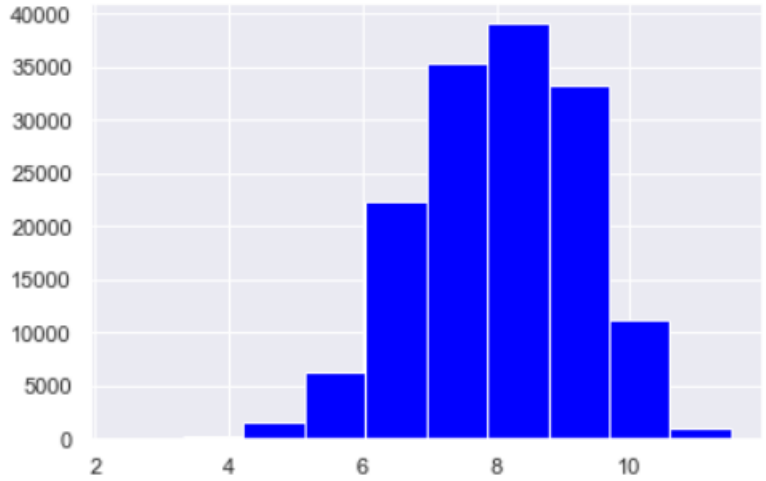


图 4 交易价格对数变换后分布

进行简单的对数变换后，使得预测值的分布较为均匀，接近正态分布，考虑用于最终的回归。

➢ 分类数据[3]

分类变量可以分为两类类别变量。类别变量可以细分成有序变量和无序变量，典型的有序变量就是学历，如博士研究生，硕士研究生，本科生等在业务含义上本身是有高低之分的。而无序变量在业务含义上是无序的，如班级等。

在本课题中，依据各变量的含义，将 brand 和 bodyType 归为有序变量，其余分类变量归为无序变量。

而分类变量运用于机器学习时，分为两类模型：（1）数值大小有意义；（2）数值大小无意义，只在乎顺序。虽然大多数模型要求输入项是数值型变量，但其对数值的处理方式是完全不同的。有些模型的损失函数对数值大小是敏感的，即变量间的数值大小本身是有比较意义的，如逻辑回归，SVM 等，我们暂将其称为 A 类模型；有些模型本身对数值变化不敏感，数值存在的意义更多的是为了排序，即 0.1,0.2,0.3 与 1,2,3 是没有区别的，这部分模型绝大部分是树模型，暂将其称为 B 类模型。

分类变量只有转化为数值变量才能被机器所识别。对于分类变量的数值型处理有 label encoding 和 one hot encoding 两种方法。

label encoding 是将类别变量中每一类别赋一数值，从而转换成数值型。比如有一列 [dog,cat,dog,mouse,cat]，我们把其转换为[1,2,1,3,2]。这里就

[3] 本部分在数据处理一开始就完成，方便之后的数据查看，详见代码文件

产生了一个奇怪的现象： dog 和 mouse 的平均值是 cat，所以 label encoding 最直观的缺点就是赋值难以解释。

one hot encoding 的优点就是它的值只有 0/1，不同的类型存储在垂直的空间。缺点就是，当类别的数量很多时，特征空间会变得非常大。

本课题分类变量较多，遂考虑采用 label encoding 编码进行数值型转化。

	SaleID	name	regDate	model	brand	bodyType	fuelType	gearbox	power	kilometer	...	v_5	v_6	v_7	v_8	v_9	v
0	0	736	20040402	30.0	6	1.0	0.0	0.0	60	12.5	...	0.235676	0.101988	0.129549	0.022816	0.097462	-2.881
1	1	2262	20030301	40.0	1	2.0	0.0	0.0	0	15.0	...	0.264777	0.121004	0.135731	0.026597	0.020582	-4.900
2	2	14874	20040403	115.0	15	1.0	0.0	0.0	163	12.5	...	0.251410	0.114912	0.165147	0.062173	0.027075	-4.846
3	3	71865	19960908	109.0	10	0.0	0.0	1.0	193	15.0	...	0.274293	0.110300	0.121964	0.033395	0.000000	-4.509
4	4	111080	20120103	110.0	5	1.0	0.0	0.0	68	5.0	...	0.228036	0.073205	0.091880	0.078819	0.121534	-1.896
149995	149995	163978	20000607	121.0	10	4.0	0.0	1.0	163	15.0	...	0.280264	0.000310	0.048441	0.071158	0.019174	1.980
149996	149996	184535	20091102	116.0	11	0.0	0.0	0.0	125	10.0	...	0.253217	0.000777	0.084079	0.099681	0.079371	1.839
149997	149997	147587	20101003	60.0	11	1.0	1.0	0.0	90	6.0	...	0.233353	0.000705	0.118872	0.100118	0.097914	2.439
149998	149998	45907	20060312	34.0	10	3.0	1.0	0.0	156	15.0	...	0.256369	0.000252	0.081479	0.083558	0.081498	2.075
149999	149999	177672	19990204	19.0	28	6.0	0.0	1.0	193	12.5	...	0.284475	0.000000	0.040072	0.062543	0.025819	1.978

转化后的数据见上图，其中：

- 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
- notRepairedDamage - 汽车有尚未修复的损坏：是：0，否：1

5.1.4 数据关系分析

➢ 数值特征

★数值特征之间的关系可视化

	power	kilometer	v_0	v_1	v_2	v_3	v_4	v_5	...	v_8	v_9	v_10	v_11	v_12	v
power	1.000000	-0.019631	0.215028	0.023746	-0.031487	-0.185342	-0.141013	0.119727	...	0.155956	-0.140203	-0.092717	-0.122107	0.161990	-0.103
kilometer	-0.019631	1.000000	-0.225034	-0.022228	-0.110375	0.402502	-0.214861	0.049502	...	-0.407686	-0.149422	0.083358	0.066542	-0.370153	-0.285
v_0	0.215028	-0.225034	1.000000	0.245049	-0.452591	-0.710480	-0.259714	0.726250	...	0.514149	-0.186243	-0.582943	-0.667809	0.415711	-0.136
v_1	0.023746	-0.022228	0.245049	1.000000	-0.001133	-0.001915	-0.000468	0.109303	...	-0.298966	-0.007698	-0.921904	0.370445	-0.087593	0.017
v_2	-0.031487	-0.110375	-0.452591	-0.001133	1.000000	0.001224	-0.001021	-0.921857	...	0.180285	-0.236164	0.274341	0.800915	0.535270	-0.055
v_3	-0.185342	0.402502	-0.710480	-0.001915	0.001224	1.000000	-0.001694	-0.233412	...	-0.933161	0.079292	0.247385	0.429777	-0.811301	-0.246
v_4	-0.141013	-0.214861	-0.259714	-0.000468	-0.001021	-0.001694	1.000000	-0.259739	...	0.051741	0.962928	0.071116	0.110660	-0.134611	0.934
v_5	0.119727	0.049502	0.726250	0.109303	-0.921857	-0.233412	-0.259739	1.000000	...	0.010686	-0.050343	-0.440588	-0.845954	-0.258521	-0.162
v_6	0.025648	-0.024664	0.243783	0.999415	0.023877	-0.000747	-0.011275	0.091229	...	-0.294956	-0.023057	-0.917056	0.386446	-0.070238	0.000
v_7	-0.060397	-0.017835	-0.584363	-0.110806	0.973689	0.191278	-0.054241	-0.939385	...	0.028695	-0.264091	0.410014	0.813175	0.385378	-0.154
v_8	0.155956	-0.407686	0.514149	-0.298966	0.180285	-0.933161	0.051741	0.010686	...	1.000000	-0.063577	0.094497	-0.369353	0.882121	0.250
v_9	-0.140203	-0.149422	-0.186243	-0.007698	-0.236164	0.079292	0.962928	-0.050343	...	-0.063577	1.000000	0.026562	-0.056200	-0.313634	0.880
v_10	-0.092717	0.083358	-0.582943	-0.921904	0.274341	0.247385	0.071116	-0.440588	...	0.094497	0.026562	1.000000	0.006306	0.001289	-0.000
v_11	-0.122107	0.066542	-0.667809	0.370445	0.800915	0.429777	0.110660	-0.845954	...	-0.369353	-0.056200	0.006306	1.000000	0.006695	-0.001
v_12	0.161990	-0.370153	0.415711	-0.087593	0.535270	-0.811301	-0.134611	-0.258521	...	0.882121	-0.313634	0.001289	0.006695	1.000000	0.001
v_13	-0.103430	-0.285158	-0.136938	0.017349	-0.055376	-0.246052	0.934580	-0.162689	...	0.250423	0.880545	-0.000580	-0.001671	0.001512	1.000
v_14	-0.023808	-0.120389	-0.039809	0.002143	-0.013785	-0.058561	-0.178518	0.037804	...	0.030416	-0.214151	0.002244	-0.001156	0.002045	0.001
price	0.219834	-0.440519	0.628397	0.060914	0.085322	-0.730946	-0.147085	0.164317	...	0.685798	-0.206205	-0.246175	-0.275320	0.692823	-0.013

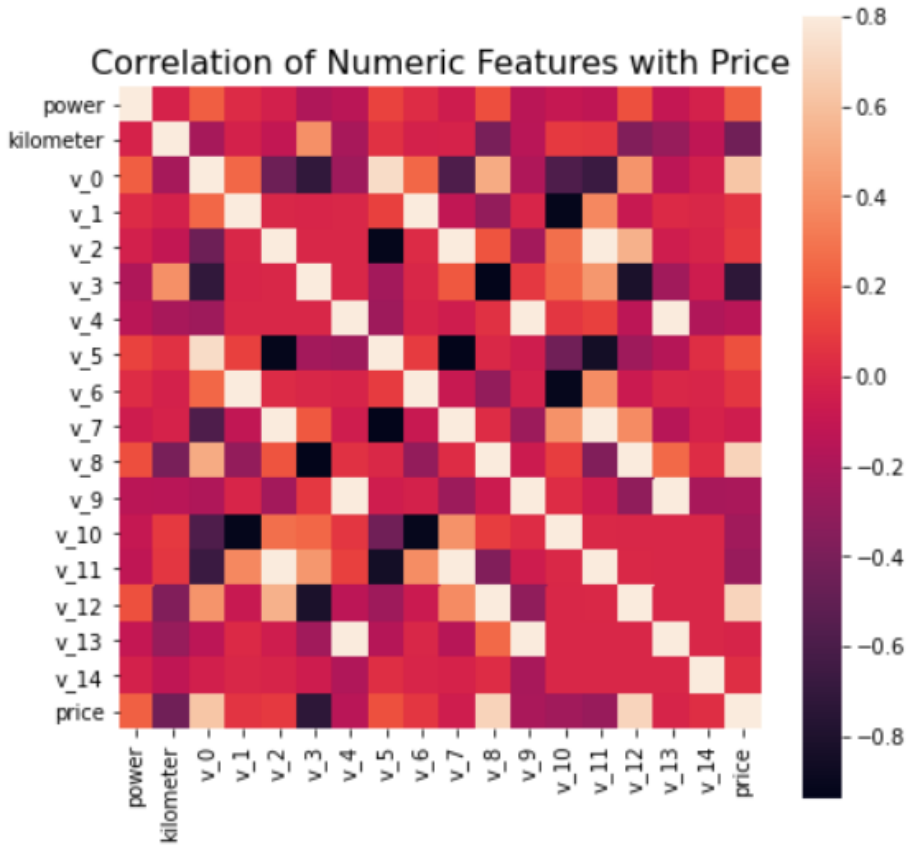
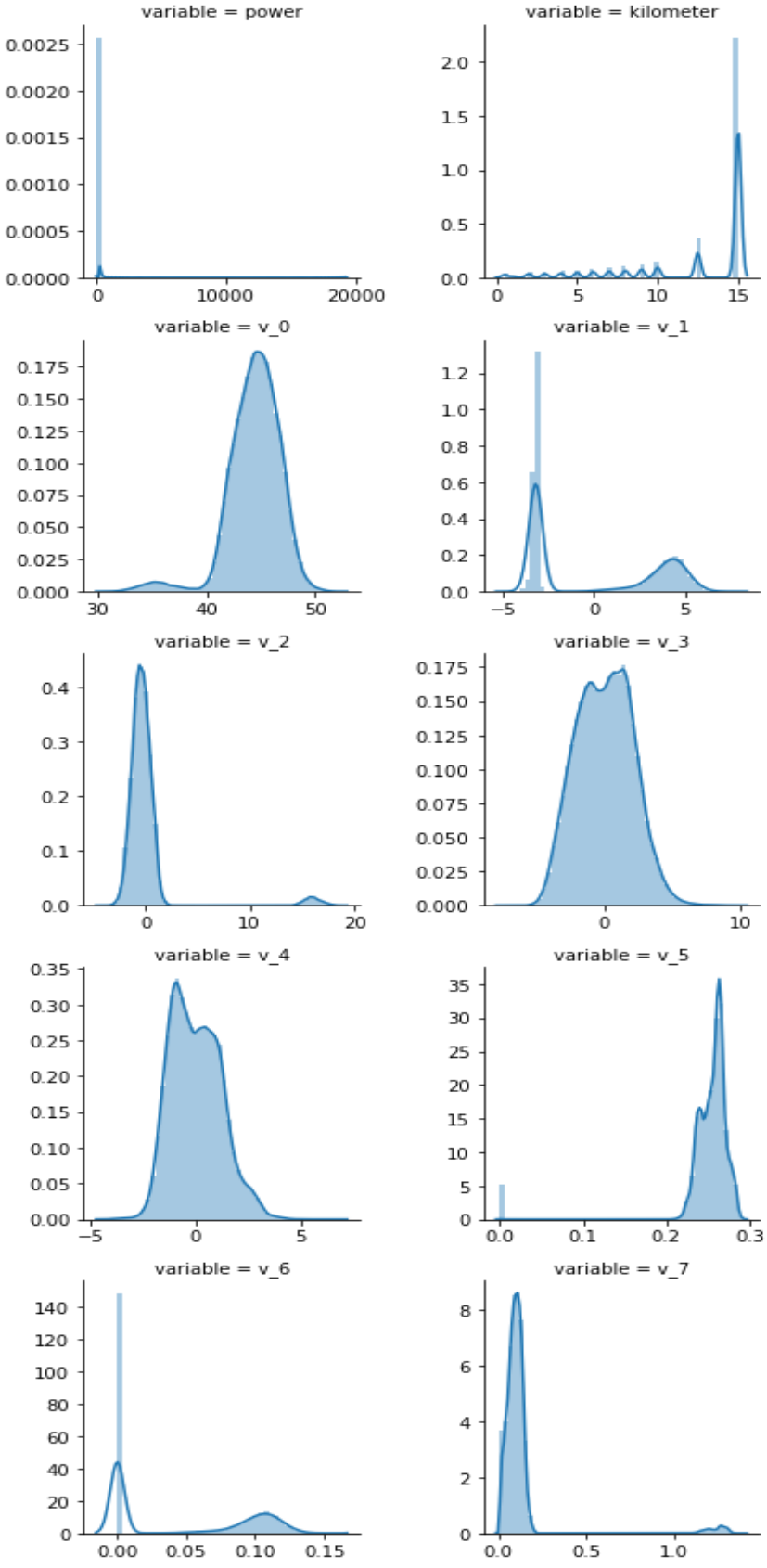


图 5 数值型特征之间关系热力图

由表格数字和相关系数可视化可以清楚的看出两个重要特征 power 和 kilometer 之间呈现出微弱的负相关关系，可以认为汽车行驶路程与汽车功率之间并无显著关系，其余 word embdding 特征因为匿名原因难以解释。

★数值分布特征


variable = power
variable = kilometer
variable = v_0
variable = v_1
variable = v_2
variable = v_3
variable = v_4
variable = v_5
variable = v_6
variable = v_7

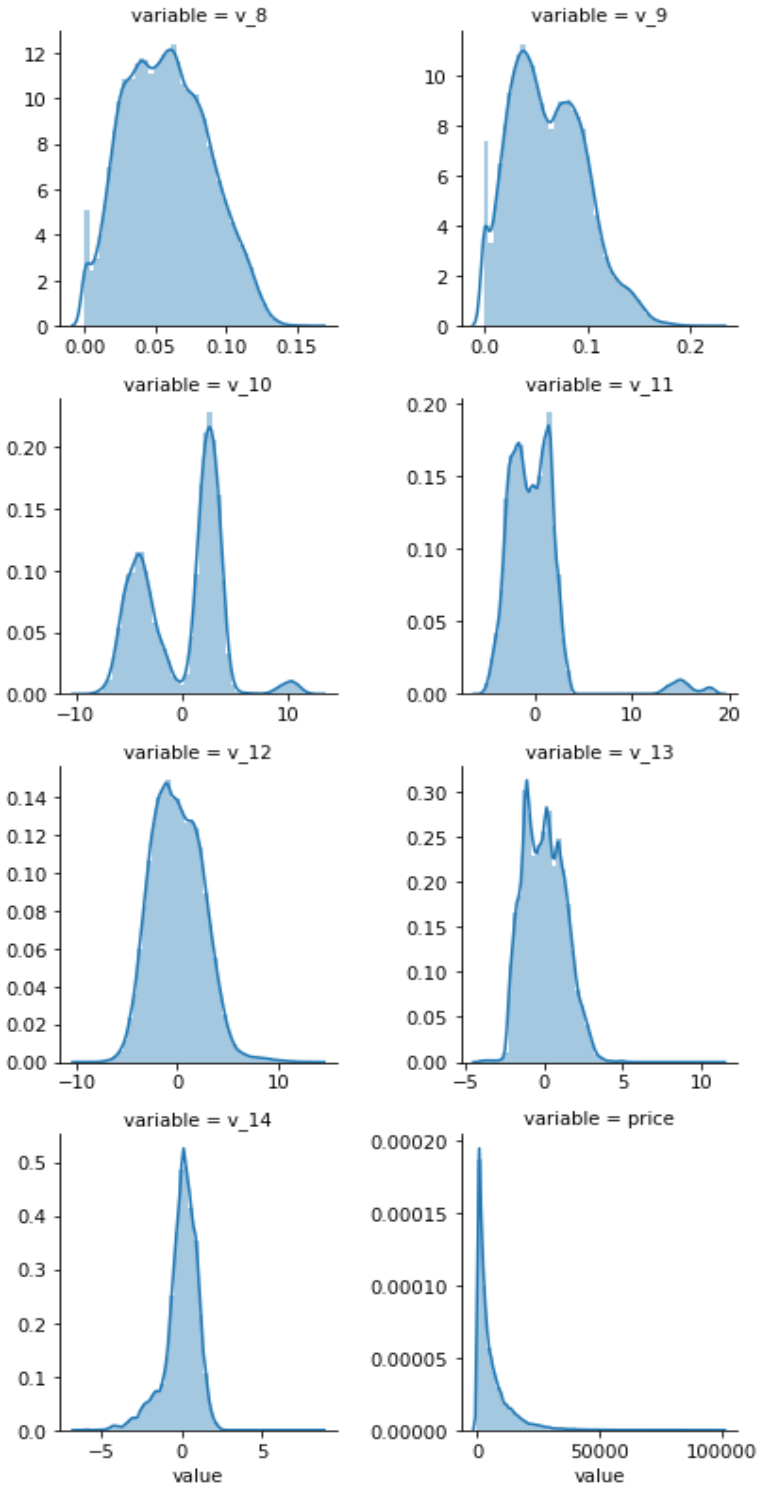


图 6 数值特征核密度直方图分布

从上图可以清晰的看出匿名特征分布较为良好，大多接近正态分布，而 power 和 kilometer 属于偏态分布，查看其偏度和峰值：

```
power          Skewness: 65.86     Kurtosis: 5733.45
kilometer      Skewness: -1.53     Kurtosis: 001.14
```

Power 特征是典型的右偏分布，且 SK 值很大，属于高度偏态数据，平均数大于中位数大于众数，K 值远大于 0，属于尖态分布，数据非常集中。

kilometer 特征是高度左偏分布，K 值接近于 0，相比与正态分布，略有一些尖峰。

★数值特征与预测值之间的关系

数值特征与预测值 price 之间的关系见下表，其中 power 与 price 正相关，kilometer 负相关，且 kilometer 相关性强于 power，对于接下来的特征工程提供了依据。

表 3 数值特征与预测值之间的相关系数

变量	power	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7
price	0.220	0.628	0.061	0.085	-0.731	-0.147	0.164	0.069	-0.053
	v_8	v_9	v_10	v_11	v_12	v_13	v_14	kilometer	
price	0.686	-0.206	-0.246	-0.275	0.693	-0.014	0.036	-0.441	

➢ 分类特征

★分类特征分布

分类特征频数分布直方图见下图，其中 model 和 brand 特征分布较为分散，可能是由于汽车型号和品牌众多的原因，所以并没有展示。

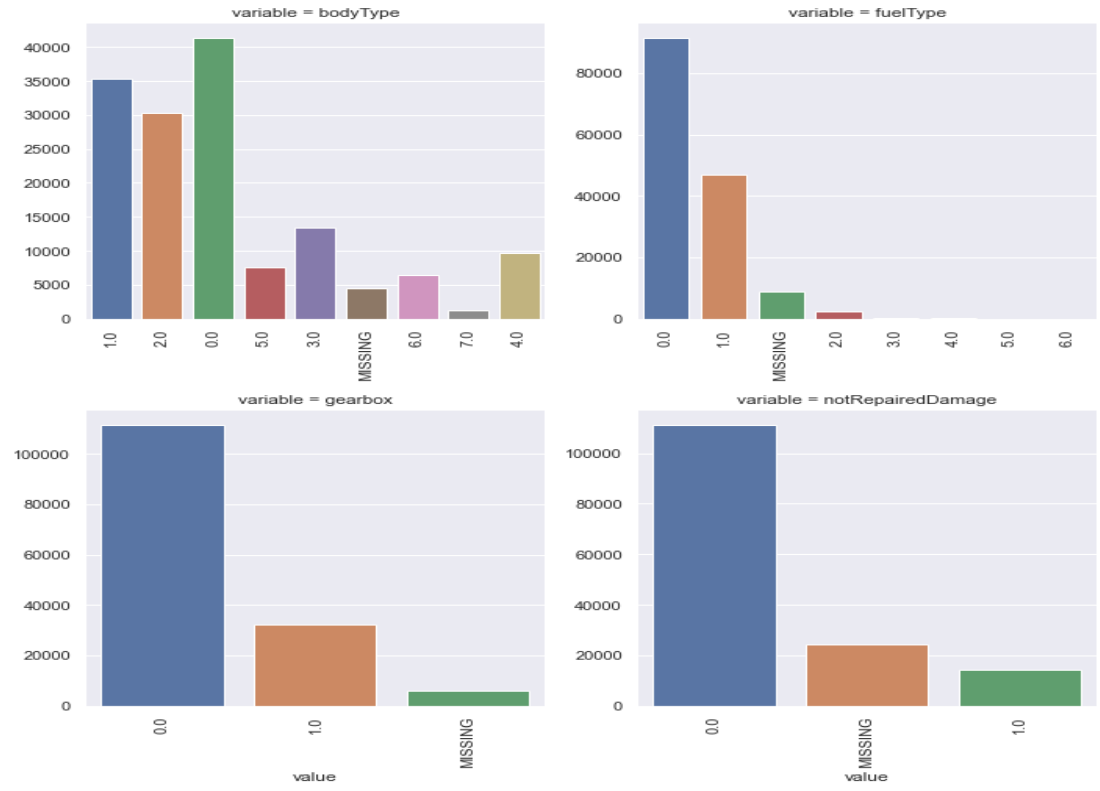


图 7 分类特征直方图

分类特征 bodyType 主要集中于 0，1，2 三个值，说明豪华轿车，微型车，厢型车三个种类的汽车成交量较多，gearbox=0 表示手动挡汽车，起出售量较自动挡更多，而在交易成功的汽车中，大部分都是有尚未修复的损坏，可见汽车质量并不理想。

★分类特征与预测值之间的关系

表 4 分类特征与预测值之间的关系

	notRepairedDamage		gearbox	
	0	1	0	1
mean(price)	6979.78	2285.09	4704.03	10645.26

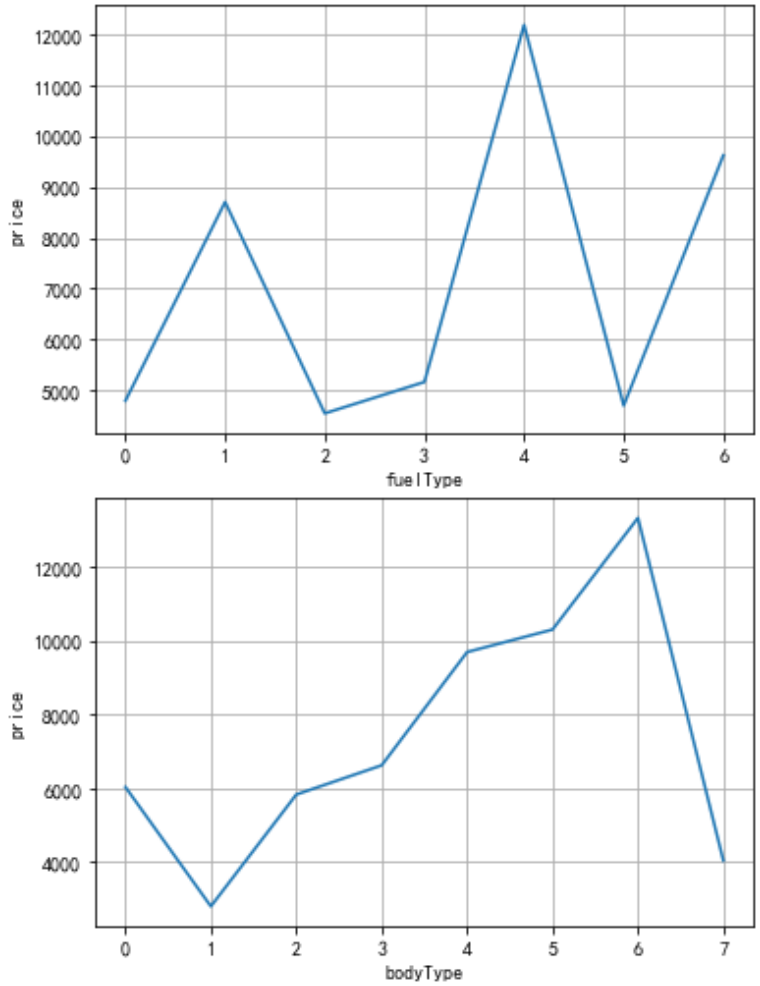


图 8 分类特征 fuelType，bodyType 与预测值之间的关系

由上图表可知，分类特征 notRepairedDamage，gearbox，fuelType，bodyType 与预测值之间关系紧密。

5.2 特征工程

“数据决定了机器学习的上限，而算法只是尽可能逼近这个上限”，这里的数据指的就是经过特征工程得到的数据。特征工程指的是把原始数据转变为模型的训练数据的过程，它的目的就是获取更好的训练数据特征，使得机器学习模型逼近这个上限。特征工程能使得模型的性能得到提升，有时甚至在简单的模型上也能取得不错的效果。特征工程在机器学习中占有非常重要的作用，一般认为包括特征构建、特征提取、特征选择三个部分。特征构建比较麻烦，需要一定的经验。特征提取与特征选择都是为了从原始特征中找出最有效的特征。它们之间的区别是特征提取强调通过特征转换的方式得到一组具有明显物理或统计意义的特征；而特征选择是从特征集合中挑选一组具有明显物理或统计意义的特征子集。两者都能帮助减少特征的维度、数据冗余，特征提取有时能发现更有意义的特征属性特征选择的过程经常能表示出每个特征的重要性对于模型构建的重要性。

➢异常值处理

利用之前各个特征的箱型图可视化，作为异常值处理的基础。power 特征的数据分布较为异常，属于高度右偏分布，是统计学中典型的长尾分布，考虑删除极端异常值，极端异常值判断公式是：

$$X > Q3 + 3(Q3 - Q1) \text{或} X < Q1 - 3(Q3 - Q1)$$

$Q1$和$Q3$表示下四分位点和上四分位。

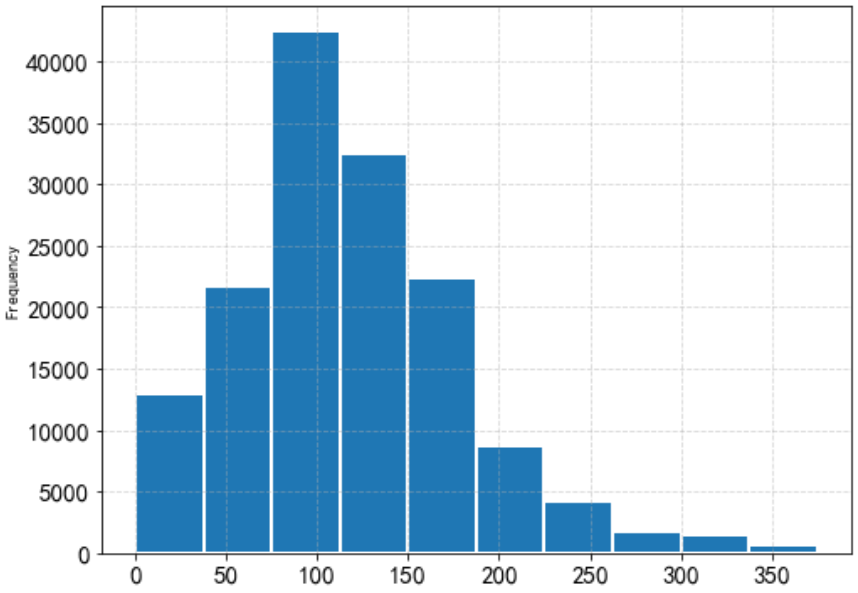


图 9 power 列删除异常值之后的分布

删除异常值之后可以看出，power 特征近似接近正态分布，效果显著。

➢填充缺失值

将'model','bodyType','fuelType','gearbox','notRepairedDamage'这五列均用众数填充缺失值。

➢特征构造

★使用时间

在之前的数据处理中，未提及两个时间序列特征，分别是 regDate 和 creatDate。regDate 表示汽车的注册日期，即购买日期，creatDate 表示汽车的售卖日期，两者的差即为汽车使用时间。利用 to_datetime 函数（其中 errors=coerce，以避免得到时间出错的格式）得到汽车的使用时间，然而存在 1.5w 条数据为空值，暂且放置不管，处理方式随模型而定。（xgboost 模型本身可以处理缺失值）

★构造城市变量

RegionCode 特征是地区编码，其中第一个字符到倒数第四个字符代表的是城市编码，因此，利用 apply 函数可以将城市信息提取出来。

★分类变量的销售统计

最好的特征构造方法就是利用 groupby 函数对分类特征进行分类汇总，选取本问题最具代表性的 brand 特征，在二手汽车的销售中，往往人们更看重的是车子的品牌 brand 和车型 bodyType（张远森，2018），使用过的二手汽车总是会产生各样的问题，而 brand 和 bodyType 就保证了车子的质量。基于此，本文对 brand 和 bodyType 分别统计各品牌，各车型的销售量，最大/小销售价格，平均销售价格，销售总价等，形成影响最终成交价格的新特征。

★数据分箱

一般在建立模型时，需要对连续变量离散化，特征离散化后，模型会更稳定，降低了模型过拟合的风险。比如在建立 logsitic 模型时就需要对连续变量进行离散化，离散化通常采用分箱法。其原因可以概括为以下几点：离散后的特征对异常值更具鲁棒性，如 power>30 为 1 否则为 0，对于 power 为 200 的也不会对模型造成很大的干扰；特征离散后模型更稳定，如用户年龄区间，不会因为用户年龄长了一岁就变化，特征离散化以后，起到了简化了逻辑回归

模型的作用，降低了模型过拟合的风险。还有很重要的一点，就是可以将缺失作为独立的一类带入模型。

对以上重要数值特征即 power 的分析可知，前 99%的数据都集中在 0-350 之间，所以对此连续型数值的分箱依据采取分频数分箱，每间隔十为一组，共分为 30 组，数据范围是 0-300。下图是数据在 0-300 范围内的分布，用正态分布拟合良好。

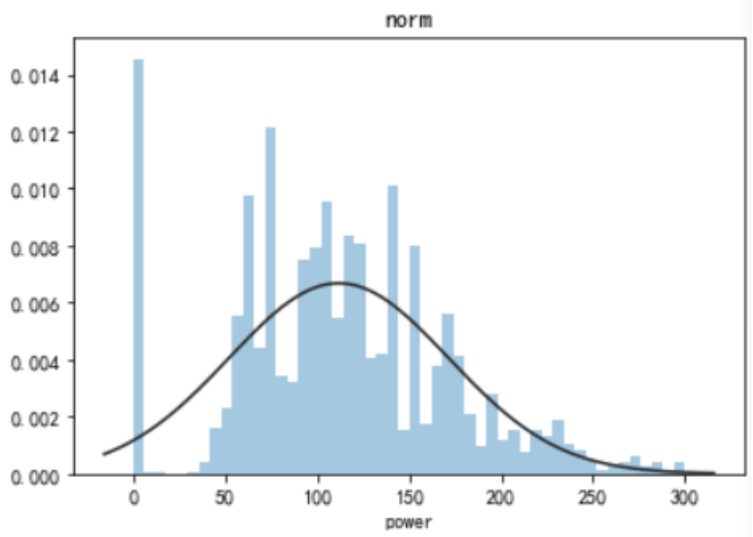


图 10 power 小于 300 时的分布

★数值特征归一化

将'power','kilometer','v_0','v_1','v_2','v_3','v_4','v_5','v_6','v_7','v_8','v_9','v_10','v_11','v_12','v_13','v_14'这 15 个数值特征以及新生成的 brand 和 bodyType 分类汇总的数值特征一并进行归一化处理，以消除数值特征带来的量纲影响。

至此，数据处理部分到此结束，清洗好的数据可以用于接下来的建模，将数据输出为 Train_data_clear.csv。

6.模型构建

本文的问题是一个预测回归问题，解决回归问题的典型方法有多元线性回归，岭回归，lasso 回归和随机森林回归。前三者都属于线性回归，后者属于非线性回归，根据本文的研究问题，将测试这四个模型之间的优劣，并选取其中最优的模型作为二手汽车的价格预测模型。

6.1 线性回归

➢线性回归原理

线性回归模型主要使用传统线性回归和岭回归，因为传统多元线性回归模型较简单，将不再做过多讨论，主要讨论岭回归在二手车价格预测模型中的原理。使用多项式回归时，如果多项式最高次项比较大，模型就容易出现过拟合[4]。正则化是一种常见的防止过拟合的方法，一般原理是在损失函数（衡量预测值与真实值之间差距的函数）后面加上一个对参数的约束项，这个约束项被叫做正则化项(regularizes)。在线性回归模型中，通常在岭回归中的正则化项为:加上所有参数(不包括常数项)的平方和。岭回归与传统多元线性回归不同的是岭回归有代价函数：

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\left(y^{i}-\left(wx^{i}+b\right)\right)^{2}+\lambda\|w\|_{2}^{2}=MSE(\theta)+\lambda\sum_{i=1}^{n}\theta_{i}^{2}$$

线性模型中还有个 lasso，其区别主要在正则化项不同。Lasso 回归模型是 L1 正则化，而岭回归是 L2 正则化。L2 正则化在拟合过程中通常都倾向于让回归系数（weights）尽可能小，最后构造一个所有参数都比较小的模型。因为一般认为参数值小的模型比较简单，能适应不同的数据集，也就是泛化性能突出，也在一定程度上避免了过拟合现象。可以设想一下对于一个线性回归方程，若参数很大，那么只要数据偏移一点点，就会对结果造成很大的影响；但如果参数足够小，数据偏移得多一点也不会对结果造成什么影响，也可以理解成模型的抗扰动能力强。而 L1 正则化有助于生成一个稀疏系数矩阵，进而可以用于特征选择。

岭回归与传统的多元线性回归主要用于研究线性关系，选用这两个模型与随机森林模型做对比分析，探究二手车价格评估中的特征变量的关系，以及线性回归在二手车估价模型中的效果。

[4] 过拟合指模型的训练效果很好，但泛化能力较差

➢线性回归模型构建

在之前对预测值的分布的测试中，预测值属于无界约翰逊分布，因此这边对其做对数处理，以消除非正态分布对模型的影响。

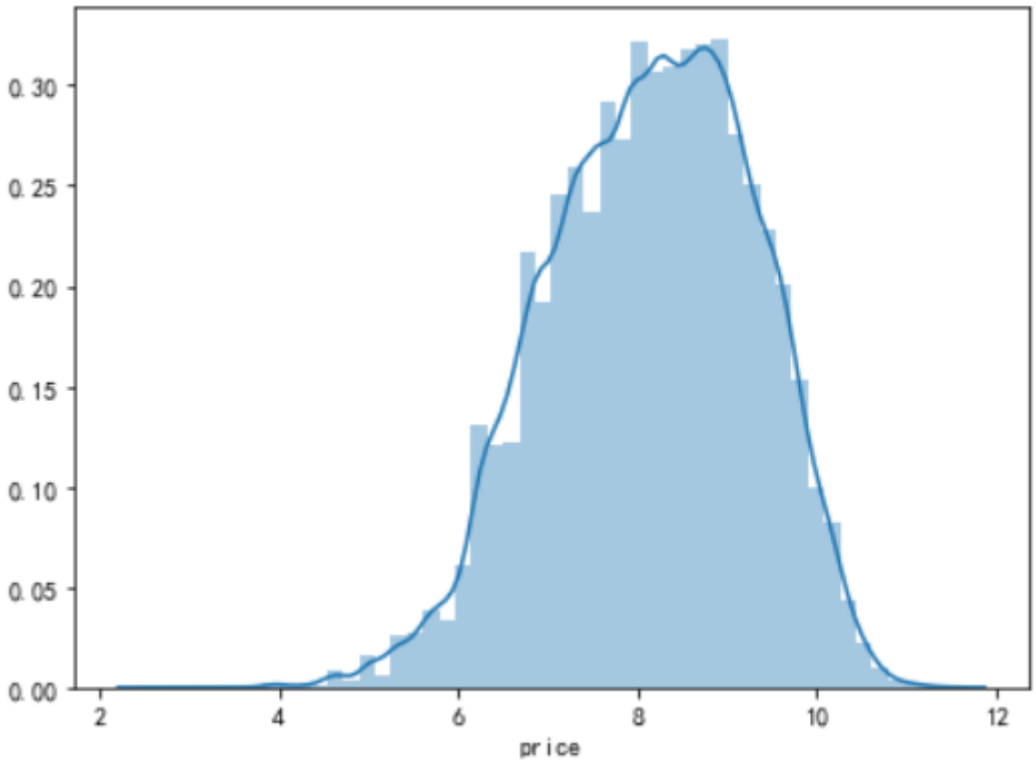


图 11 price 取对数后预测值分布

在使用训练集对参数进行训练的时候，习惯的做法是将一整个训练集分为两个部分，一般分为：训练集（train_data）和测试集（test_data）。这其实是为了保证训练效果而特意设置的。其中测试集很好理解，其实就是完全不参与训练的数据，仅仅用来观测测试效果的数据。而训练集则牵涉到模型拟合优度的问题了。

因为在实际的训练中，训练的结果对于训练集的拟合程度通常还是挺好的（初始条件敏感），但是对于训练集之外的数据的拟合程度通常就不那么令人满意了。因此我们通常并不会把所有的数据集都拿来训练，而是分出一部分来（这一部分不参加训练）对训练集生成的参数进行测试，相对客观的判断这些参数对训练集之外的数据的符合程度。这种思想就称为交叉验证（Cross Validation）

所以为了评价模型的好坏，我们可以采用交叉验证的方法。

表 5 线性模型评估[5]

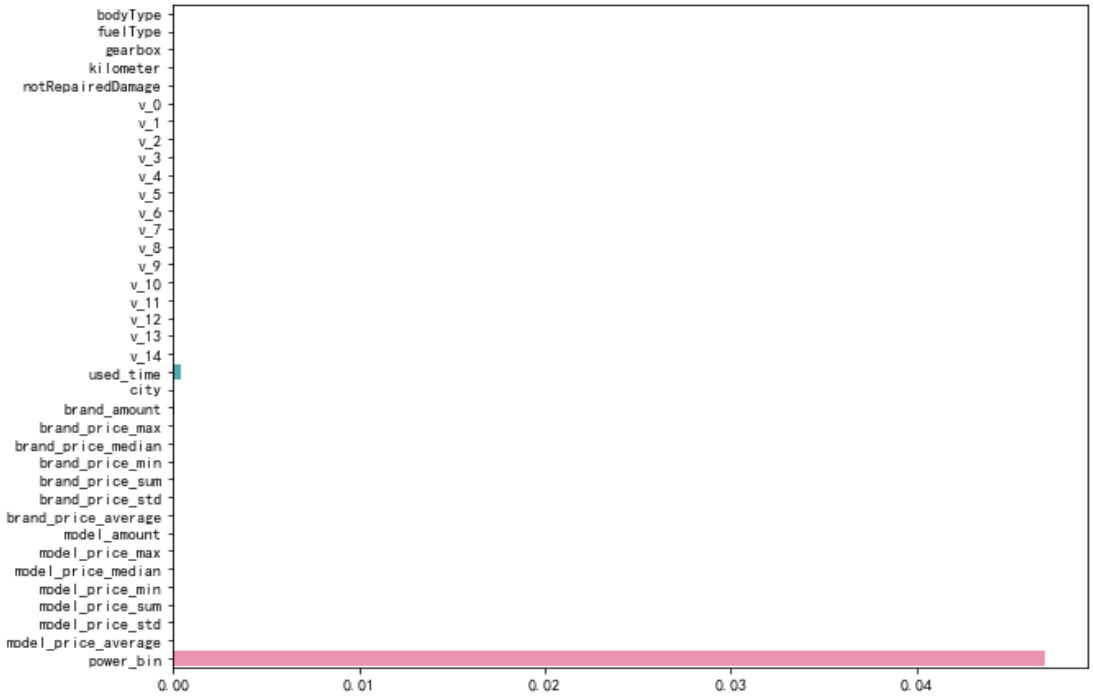
	LinearRegression	Ridge	Lasso
cv1	0.190288	0.193661	0.440650
cv2	0.193954	0.196980	0.437916
cv3	0.193906	0.197078	0.442738
cv4	0.191450	0.194576	0.434396
cv5	0.195767	0.198953	0.442082

上表是对三个线性模型进行评估的得分，其中模型评估标准选取 MAE (Mean Absolute Error)，即平均绝对误差。其计算公式为

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

其中，y_i为真实值，$\hat{y}_i$为预测值，MAE 就是真实值与预测值之间的误差的绝对值平均。所以根据 MAE 的定义，得分越小模型越好。从上表可以看出，LinearRegression 多元线性回归模型在各折交叉验证上均表现最优。而 Lasso 回归模型显然不适合做模型预测，但其提供的 L1 正则化可以用来做特征选择。

由下图可以看出影响最终交易价格最重要的特征是 power（使用功率）和 used_time（车龄）。



[5] 因为对预测值做了对数处理，所以此处的 MAE 值较小，真正的 MAE 值为 933.06

图 12 Lasso 特征选择

6.2 随机森林回归

➢模型理论

随机森林先通过 bootstrap 抽样方法从原始样本中抽样，然后对抽取的样本进行 CART 决策树建模，不同的决策树之间是独立同分布的。各个决策树的结果不完全一致，随机森林组合各棵树的结果，通过投票得出最终的预测结果。不同于传统的决策树模型经常对训练数据进行过拟合，由于随机森林集成了众多决策树，每一棵决策树跟其他决策树都不同，都会以不同方式的过拟合，那么对这些树的结果取平均值，就可以降低过拟合。除此之外，随机森林对异常值和噪声也具有良好的容忍度。通过投票决策，随机森林最终的分类模型为：

$$H(x) = argmax_Y \sum_{i=1}^{k} I(h_i(x) = Y)$$

上式中，$H(x)$表示组合分类模型，h_i表示单个决策树分类模型，Y 表示目标变量，即输出变量，$I(\cdot)$为示性函数。文本使用的是回归模型。

由于随机森林是通过 bootstrap 对原始数据进行抽样，当数据集很大时并不是所有数据均被抽取，只有约 63.2%会出现在样本集中，且剩余没有被抽取的数据被称为袋外（out-of-bag）数据。我们可以使用这部分数据作为测试集来估计随机森林模型的泛化能力，即 OOB 估计。Breiman 在其研究中证明了 OOB 估计的无偏性，因此 OOB 可以度量随机森林中变量的重要性，这样可以对原始变量进行筛选从而建立合理的指标体系。OOB 方法度量变量重要性的基本思路为对某个变量加入一定噪声后，其 OOB 数据测试随机森林性能得到的准确率与加入噪声之前的准确率相比较，两者之间差值越大说明该变量越重要。利用随机森林这点特性，可以对特征变量进行重要性排序。

➢模型实现

利用 python 机器学习库 sklearn 中的 RandomForestRegressor 对数据进行回归，得到以下结论。

```
Accuracy on training set: 0.994
Accuracy on test set: 0.957
```

利用 RandomForestRegressor().score()函数对模型进行评分，可见模型在训练数据上的拟合度高达 0.994，在训练数据集上的拟合度也达到了 0.957，所

以模型的泛化能力超过了三个线性模型（线性模型拟合度只有 0.7），说明随机森林模型在此问题上预测效果最好，也说明之前的特征工程效果良好。

★特征重要度排序

随机森林模型会输出二手车数据集中，每个特征重要性的度量结果，如表所示。

表 6 随机森林特征重要性

	V_12	V_3	V_0	used_time	notRepairedDamage	kilometer	power_bin
特征重要度	0.716	0.114	0.075	0.013	0.004	0.004	0.003

除去一些不可解释特征的影响，分别是 used_time，notRepairedDamage，kilometer，power_bin 这四个特征最重要。下图是特征重要性可视化。

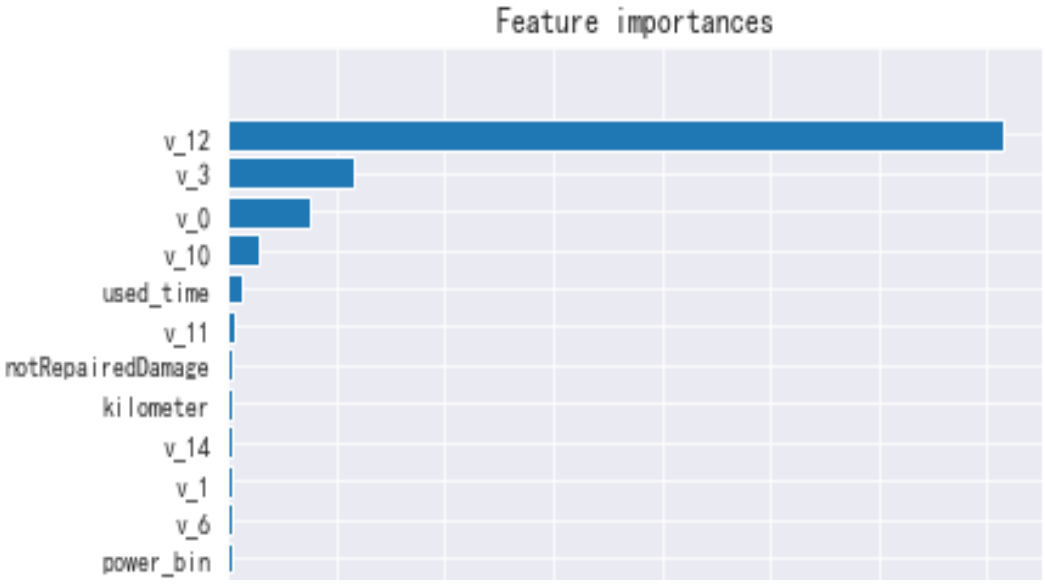


图 13 随机森林特征重要性可视化

7.结论

本文构建了预测二手汽车交易价格的线性和非线性回归预测模型。通过对随机森林回归、多元线性回归、岭回归以及 lasso 回归模型的预测效果评价，可以分析出随机森林模型的预测效果最佳，同时也分析了特征变量的重要性。发现了随机森林抗干扰能力强、不容易出现过度拟合、预测准确率高、学习速度快等各

方面的优点，充分了解到随机森林算法是值得我们研究学习和推广的，在预测二手汽车交易价格等此类问题上，可以推广其成为预测标准模型。

此外，文本得出了以下主要结论：

➢随机森林算法对每个变量在模型中的重要性进行了度量，其中 used_time（车龄）的重要程度最高，其次为 notRepairedDamage（是否存在未修复的损坏），kilometer（汽车已行驶里程），power_bin（汽车功率）等。在研究二手车保值率的影响因素和保值率的计算方法时，可以结合随机森林算法对变量重要性度量这一优势，继续对车辆保值率深入的研究。消费者在购买新车或者二手车时，也可参考特征变量重要性的结果来选车。

➢文本比较了线性模型和非线性模型对于此问题的模型优劣，发现随机森林算法精度远超线性模型。具体分析其原因，虽然回归分析法在理论上可以很好的预测二手车价格而且也是当前领域运用比较广泛的方法，但是二手车价格与很多因素相关，客观上受到车辆保有量、车龄、损坏未修复、道路环境、燃油价格、车辆使用情况、和区域特点等众多因素的影响，又由于数据中存在大量的多分类数据，采用多元线性回归方法分析会使得问题维度陡然增高，问题更加复杂。因此，并不适宜使用多元线性回归方法来做二手车价格的预测。

相比之下，随机森林方法能够很好的解决非线性问题，具有良好的容错性、泛化能力、自适应能力。二手车价格预测问题中，输入的数据维度较高，且为非线性问题，随机森林能够很好的处理本问题，构建多决策树进行了较好的拟合，得到了较为良好的模型。