Table of Contents

评分卡建模模板

author:33

date:2020/3/23

数据读取&基本情况查看

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5456 entries, 0 to 5455
Data columns (total 43 columns):
id          5456 non-null int64
date        5456 non-null datetime64[ns]
v1          4324 non-null float64
v2          5456 non-null object
v3          4255 non-null float64
v4          5293 non-null float64
v5          5440 non-null float64
v6          3454 non-null float64
v7          2467 non-null float64
v8          5456 non-null int64
v9          5456 non-null object
v10         5456 non-null object
v11         5098 non-null float64
v12         4967 non-null float64
v13         5213 non-null float64
v14         5004 non-null float64
v15         5070 non-null float64
v16         2054 non-null float64
v17         2054 non-null object
v18         5176 non-null float64
v19         5021 non-null float64
v20         5155 non-null float64
v21         5081 non-null float64
v22         5205 non-null float64
v23         5207 non-null float64
v24         5034 non-null float64
v25         5450 non-null float64
v26         5456 non-null int64
v28         4317 non-null float64
v29         4424 non-null float64
v30         4424 non-null float64
v31         4424 non-null float64
v32         4424 non-null float64
v33         4427 non-null float64
v34         4061 non-null float64
v35         88 non-null object
v36         1456 non-null float64
v37         1457 non-null float64
v38         3892 non-null float64
v39         100 non-null object
v40         4324 non-null float64
v41         5456 non-null int64
target      5456 non-null int64
dtypes: datetime64[ns](1), float64(31), int64(5), object(6)
memory usage: 1.8+ MB
None
```

连续型变量分布

Search:

	id	v1
25%	1364.75	
50%	2728.5	
75%	4092.25	
count	5456	43
max	5456	3
mean	2728.5	18.5080
min	1	
std	1575.155865	25.5455

Showing 1 to 8 of 8 entries

离散型变量分布

Search:

	v2	v9
count	5456	5456
freq	4190	1688
top	A	E
unique	3	5

Showing 1 to 4 of 4 entries

数据预处理

异常值检测&处理

针对异常变量可根据对数据对理解进行极端值分数映射转换或者删除极端值对应对样本

离散型变量个数

6

连续型变量个数

37

Search:

	v2	v9
count	5456	5456
freq	4190	1688
top	A	E
unique	3	5

Showing 1 to 4 of 4 entries

特征衍生

这里主要是做特征各种交叉衍生的工作

缺失值检测&处理

有缺失值的变量个数：

```
34
```

各变量缺失率展示

Show 10 entries Search:

index	var_name	queshi_num
0	v35	5368
1	v39	5356
2	v36	4000
3	v37	3999
4	v16	3402
5	v17	3402
6	v7	2989
7	v6	2002
8	v38	1564
9	v34	1395

Showing 1 to 10 of 34 entries Previous 1 2 3 4 Next

缺失过多特征删除

删除的缺失值过高的特征：

```
  0    v35
1    v39
2    v36
3    v37
Name: var_name, dtype: object
```

直接填充默认缺失值

另外的缺失值填充方法

- 根据相关性填充
- 根据后续结果看是否缺失值要根据特征中位数或者均值来填补

缺失值填充后变量分布

Search:

	id	v1	
25%	1364.75	1	
50%	2728.5	5	
75%	4092.25	18	
count	5456	5456	
max	5456	300	
mean	2728.5	-192.602456	-21
min	1	-999	
std	1575.155865	413.264315	41

Showing 1 to 8 of 8 entries

Search:

	v2	v9
count	5456	5456
freq	4190	1688
top	A	E
unique	3	5

Showing 1 to 4 of 4 entries

特征筛选

训练集好坏用户，1表示坏用户：

```
  0    2480
1    1167
Name: target, dtype: int64
```

时间外验证集好坏用户，1表示坏用户：

```
 0    1016
1     237
Name: target, dtype: int64
```

psi筛选

```
psi筛选删除的特征：  ['v17']
psi筛选的特征：  Index(['v1', 'v2', 'v3', 'v4', 'v5', 'v6', 'v7', 'v8', 'v9', 'v1
0', 'v11',
       'v12', 'v13', 'v14', 'v15', 'v16', 'v18', 'v19', 'v20', 'v21', 'v22',
       'v23', 'v24', 'v25', 'v26', 'v28', 'v29', 'v30', 'v31', 'v32', 'v33',
       'v34', 'v38', 'v40', 'v41', 'target'],
      dtype='object')
各变量psi：
```

Show 10 entries Search:

	var_na
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

Showing 1 to 10 of 36 entries Previous 1 2 3 4 Next

```
psi筛选特征完成--------------------------------
```

随机森林筛选特征

随机森林删除的特征：

Search:

var_na
30
31
32
33
34

Showing 1 to 5 of 5 entries

随机森林筛选的特征：

Show 10 entries Search:

var_na
0
1
2
3
4
5
6
7
8
9

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

iv筛选

特征分箱完成----------

iv筛选的特征： ['v5', 'v12', 'v13', 'v23', 'v15', 'v22', 'v20', 'v19', 'v11', 'v16', 'v21', 'v25', 'v18', 'v4', 'v14', 'v9', 'v24', 'v10', 'v26', 'v30', 'v3', 'v1', 'v2', 'v32', 'v6']
iv删除的特征： ['v7', 'v28', 'v31', 'v8']
iv筛选特征完成----------

人工去除偏事后特征

之所以在这个步骤去除，是因为一定程度上做过特征筛选，特征没有那么多了，可以更好地对每个特征进行解读；为什么放在相关性筛选之前做这个动作，是因为相关性筛选里面涉及到不同变量间关系而对变量进行删除，可能存在事后变量的存在而导致删除其余非事后变量这种现象，因此人工剔除偏事后特征在这一步做相对比较合适；

相关性筛选，多重共线性筛选

相关矩阵

Show 10 entries Search:

	v5
v1	0.183739
v11	0.020501
v12	0.049118
v13	-0.010652
v14	0.076944
v15	-0.004716
v16	0.183819
v18	0.064886
v19	0.081738
v20	0.101116

Showing 1 to 10 of 20 entries Previous 1 2 Next

```
相关性筛选的特征： ['v5', 'v12', 'v13', 'v16', 'v25', 'v4', 'v24', 'v3', 'v32']
相关性筛选删除的特征： ['v23', 'v15', 'v22', 'v20', 'v19', 'v11', 'v21', 'v18', 'v1
4', 'v1', 'v6']
连续变量相关性筛选完成----------------------------
多重共线性筛选的特征： ['v5', 'v12', 'v13', 'v16', 'v25', 'v4', 'v24', 'v3', 'v32']
多重共线性删除的特征： []
连续变量多重共线性筛选完成------------------------
```

分箱调整

```
自动调整单调分箱完成--------------------
```

连续型变量分箱调整后分箱和woe情况

Show 10 entries Search:

col	bin	IV	
v12	(-inf, -999.0]	1.064377	-1
v12	(-999.0, 468.5]	1.064377	
v12	(468.5, 541.5]	1.064377	1
v12	(541.5, 651.5]	1.064377	0
v12	(651.5, inf]	1.064377	-0
v13	(-inf, -999.0]	1.032758	
v13	(-999.0, 568.5]	1.032758	1
v13	(568.5, 619.5]	1.032758	0
v13	(619.5, 685.5]	1.032758	-0
v13	(685.5, inf]	1.032758	-1

Showing 1 to 10 of 43 entries Previous 1 2 3 4 5 Next

iv筛选后的离散型变量 ['v9', 'v10', 'v26', 'v30', 'v2']
连续和离散型变量分箱调整后分箱和woe情况

Show 10 entries Search:

col	bin	IV	
v10	a	0.624386	0
v10	b	0.624386	-0
v10	c	0.624386	-0
v10	d	0.624386	-1
v10	e	0.624386	-2
v10	f	0.624386	-2
v10	g	0.624386	0
v12	(-inf, -999.0]	1.064377	-1
v12	(-999.0, 468.5]	1.064377	
v12	(468.5, 541.5]	1.064377	1

Showing 1 to 10 of 64 entries Previous 1 2 3 4 5 6 7 Next

woe编码后相关性筛选

Show 10 entries Search:

	v5	v12
v10	0.051899	0.1180
v12	0.747298	
v13	0.655111	0.7770
v16	0.699436	0.7392
v2	0.445659	0.2948
v24	0.350397	0.4000
v25	0.285626	0.3439
v26	0.0477	0.0653
v3	0.70092	0.4257
v30	0.194599	0.2042

Showing 1 to 10 of 14 entries Previous 1 2 Next

woe编码后相关性筛选的特征： ['v5', 'v13', 'v16', 'v25', 'v9', 'v24', 'v30', 'v2', 'v32']
woe编码后相关性删除的特征： ['v12', 'v4', 'v10', 'v26', 'v3']
相关性筛选完成----------
woe编码后多重共线性筛选的特征： ['v5', 'v13', 'v16', 'v25', 'v9', 'v24', 'v30', 'v2', 'v32']
woe编码后多重共线性删除的特征： []
多重共线性筛选完成----------

```
Optimization terminated successfully.
         Current function value: 0.511759
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 3647
Model:                          Logit   Df Residuals:                     3645
Method:                           MLE   Df Model:                            1
Date:                Mon, 23 Mar 2020   Pseudo R-squ.:                  0.1836
Time:                        18:23:48   Log-Likelihood:                -1866.4
converged:                       True   LL-Null:                       -2286.2
Covariance Type:            nonrobust   LLR p-value:                1.358e-184
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7538      0.041    -18.578      0.000      -0.833      -0.674
v5             1.0000      0.039     25.934      0.000       0.924       1.076
==============================================================================
 ------------------------------------------------
Optimization terminated successfully.
         Current function value: 0.490219
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 3647
Model:                          Logit   Df Residuals:                     3644
Method:                           MLE   Df Model:                            2
Date:                Mon, 23 Mar 2020   Pseudo R-squ.:                  0.2180
Time:                        18:23:49   Log-Likelihood:                -1787.8
converged:                       True   LL-Null:                       -2286.2
Covariance Type:            nonrobust   LLR p-value:                3.776e-217
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7493      0.042    -17.959      0.000      -0.831      -0.668
v5             0.6617      0.046     14.269      0.000       0.571       0.753
v13            0.5899      0.047     12.480      0.000       0.497       0.683
==============================================================================
 ------------------------------------------------
Optimization terminated successfully.
         Current function value: 0.488859
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 3647
Model:                          Logit   Df Residuals:                     3643
Method:                           MLE   Df Model:                            3
Date:                Mon, 23 Mar 2020   Pseudo R-squ.:                  0.2201
Time:                        18:23:49   Log-Likelihood:                -1782.9
converged:                       True   LL-Null:                       -2286.2
Covariance Type:            nonrobust   LLR p-value:                6.700e-218
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7441      0.042    -17.805      0.000      -0.826      -0.662
v5             0.5927      0.051     11.544      0.000       0.492       0.693
v13            0.5166      0.053      9.806      0.000       0.413       0.620
v16            0.1915      0.061      3.155      0.002       0.073       0.310
==============================================================================
 ------------------------------------------------
Optimization terminated successfully.
```

```
         Current function value: 0.455744
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 3647
Model:                          Logit   Df Residuals:                     3642
Method:                           MLE   Df Model:                            4
Date:                Mon, 23 Mar 2020   Pseudo R-squ.:                  0.2730
Time:                        18:23:49   Log-Likelihood:                -1662.1
converged:                       True   LL-Null:                       -2286.2
Covariance Type:            nonrobust   LLR p-value:                5.858e-269
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7162      0.043    -16.484      0.000      -0.801      -0.631
v5             0.6518      0.055     11.956      0.000       0.545       0.759
v13            0.3595      0.056      6.423      0.000       0.250       0.469
v16            0.0962      0.065      1.486      0.137      -0.031       0.223
v25            0.7595      0.050     15.166      0.000       0.661       0.858
==============================================================================
 --------------------------------------------------
Optimization terminated successfully.
         Current function value: 0.419069
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 3647
Model:                          Logit   Df Residuals:                     3641
Method:                           MLE   Df Model:                            5
Date:                Mon, 23 Mar 2020   Pseudo R-squ.:                  0.3315
Time:                        18:23:49   Log-Likelihood:                -1528.3
converged:                       True   LL-Null:                       -2286.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7454      0.046    -16.137      0.000      -0.836      -0.655
v5             0.8263      0.059     14.021      0.000       0.711       0.942
v13            0.2922      0.059      4.957      0.000       0.177       0.408
v16            0.0939      0.068      1.377      0.169      -0.040       0.228
v25            0.4948      0.053      9.276      0.000       0.390       0.599
v9             1.0153      0.067     15.110      0.000       0.884       1.147
==============================================================================
 --------------------------------------------------
Optimization terminated successfully.
         Current function value: 0.417005
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 3647
Model:                          Logit   Df Residuals:                     3640
Method:                           MLE   Df Model:                            6
Date:                Mon, 23 Mar 2020   Pseudo R-squ.:                  0.3348
Time:                        18:23:49   Log-Likelihood:                -1520.8
converged:                       True   LL-Null:                       -2286.2
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7412      0.046    -16.000      0.000      -0.832      -0.650
v5             0.8057      0.059     13.593      0.000       0.690       0.922
```

v13	0.2640	0.060	4.433	0.000	0.147	0.381
v16	0.0778	0.069	1.132	0.258	-0.057	0.213
v25	0.4655	0.054	8.630	0.000	0.360	0.571
v9	0.9652	0.068	14.118	0.000	0.831	1.099
v24	0.2452	0.063	3.873	0.000	0.121	0.369

Optimization terminated successfully.
Current function value: 0.406572
Iterations 7

Logit Regression Results

Dep. Variable:	target	No. Observations:	3647
Model:	Logit	Df Residuals:	3639
Method:	MLE	Df Model:	7
Date:	Mon, 23 Mar 2020	Pseudo R-squ.:	0.3514
Time:	18:23:49	Log-Likelihood:	-1482.8
converged:	True	LL-Null:	-2286.2
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7433	0.047	-15.817	0.000	-0.835	-0.651
v5	0.7819	0.060	13.009	0.000	0.664	0.900
v13	0.2305	0.061	3.807	0.000	0.112	0.349
v16	0.0758	0.070	1.086	0.277	-0.061	0.213
v25	0.4791	0.055	8.732	0.000	0.372	0.587
v9	0.9538	0.069	13.783	0.000	0.818	1.089
v24	0.2388	0.064	3.729	0.000	0.113	0.364
v30	0.7385	0.087	8.512	0.000	0.568	0.909

Optimization terminated successfully.
Current function value: 0.406568
Iterations 7

Logit Regression Results

Dep. Variable:	target	No. Observations:	3647
Model:	Logit	Df Residuals:	3638
Method:	MLE	Df Model:	8
Date:	Mon, 23 Mar 2020	Pseudo R-squ.:	0.3514
Time:	18:23:49	Log-Likelihood:	-1482.8
converged:	True	LL-Null:	-2286.2
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7438	0.047	-15.794	0.000	-0.836	-0.652
v5	0.7777	0.065	11.947	0.000	0.650	0.905
v13	0.2307	0.061	3.810	0.000	0.112	0.349
v16	0.0766	0.070	1.095	0.273	-0.060	0.214
v25	0.4794	0.055	8.733	0.000	0.372	0.587
v9	0.9539	0.069	13.782	0.000	0.818	1.090
v24	0.2395	0.064	3.732	0.000	0.114	0.365
v30	0.7386	0.087	8.513	0.000	0.569	0.909
v2	0.0222	0.131	0.170	0.865	-0.234	0.279

Optimization terminated successfully.
Current function value: 0.406565
Iterations 7

Logit Regression Results

Dep. Variable:	target	No. Observations:	3647
Model:	Logit	Df Residuals:	3638
Method:	MLE	Df Model:	8
Date:	Mon, 23 Mar 2020	Pseudo R-squ.:	0.3514
Time:	18:23:49	Log-Likelihood:	-1482.7
converged:	True	LL-Null:	-2286.2
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7432	0.047	-15.813	0.000	-0.835	-0.651
v5	0.7816	0.060	13.000	0.000	0.664	0.899
v13	0.2310	0.061	3.812	0.000	0.112	0.350
v16	0.0759	0.070	1.087	0.277	-0.061	0.213
v25	0.4787	0.055	8.721	0.000	0.371	0.586
v9	0.9540	0.069	13.783	0.000	0.818	1.090
v24	0.2380	0.064	3.712	0.000	0.112	0.364
v30	0.7549	0.116	6.501	0.000	0.527	0.982
v32	-0.0397	0.187	-0.212	0.832	-0.406	0.327

```
--------------------------------------------------
显著性筛选的变量： ['v5', 'v13', 'v16', 'v25', 'v9', 'v24', 'v30']
显著性筛选删除的变量： ['v2', 'v32']
显著性筛选完成-------------------------
```

剔除系数和其他系数符号不一致的特征

这里之所以没有说系数为正或者为负，是因为如果计算woe时为坏比好，那么逻辑回归系数就为正，如果计算woe时为好比坏，那么逻辑回归系数就为负

```
删除系数为负的特征： []
最终入模特征： ['v5', 'v13', 'v16', 'v25', 'v9', 'v24', 'v30']
```

模型训练和评价

截距： -0.7561470537100758
特征系数： {'v5': 0.8062973323178239, 'v13': 0.26774567830576895, 'v16': 0.04458792530230484, 'v25': 0.4419652612567421, 'v9': 0.9732424369109149, 'v24': 0.19840993320999567, 'v30': 0.7959596465531374}
建模完成----------

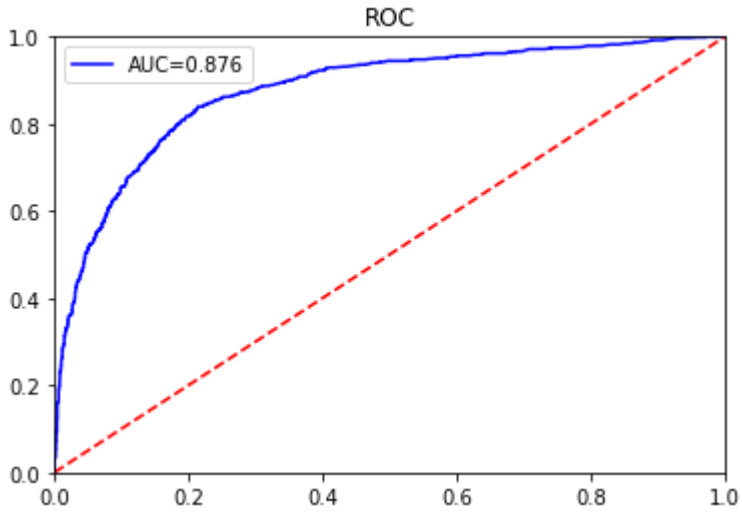
训练集好坏样本数：

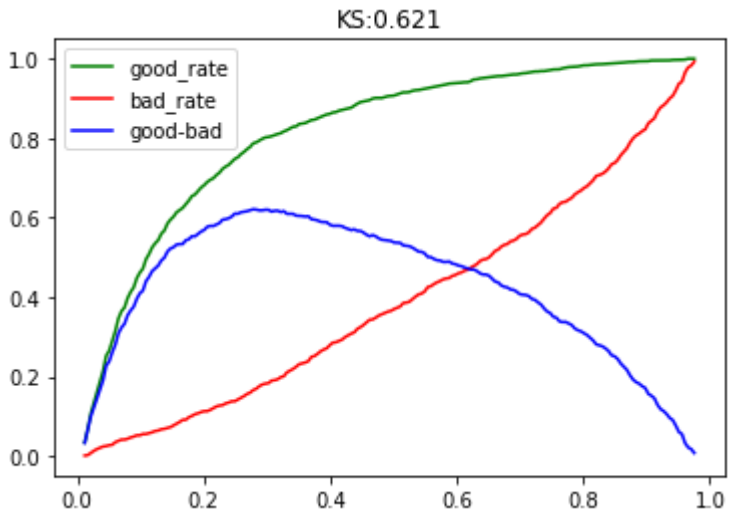
```
0    1995
1     922
Name: target, dtype: int64
```

验证集好坏样本数：

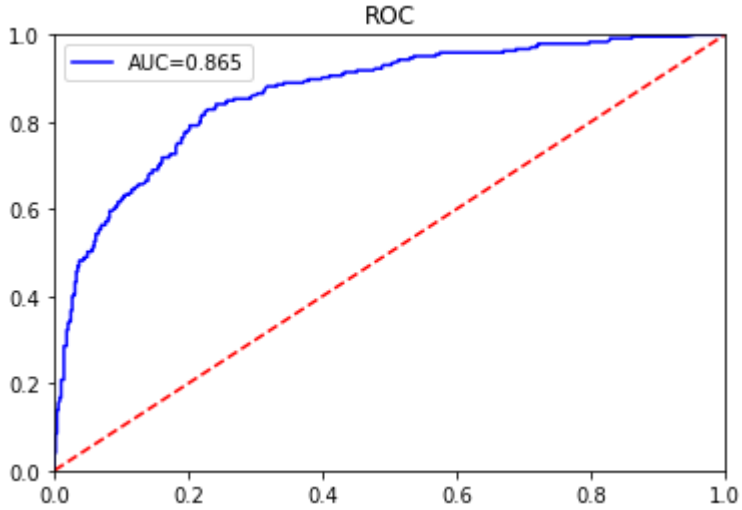
```
0    485
1    245
Name: target, dtype: int64
```

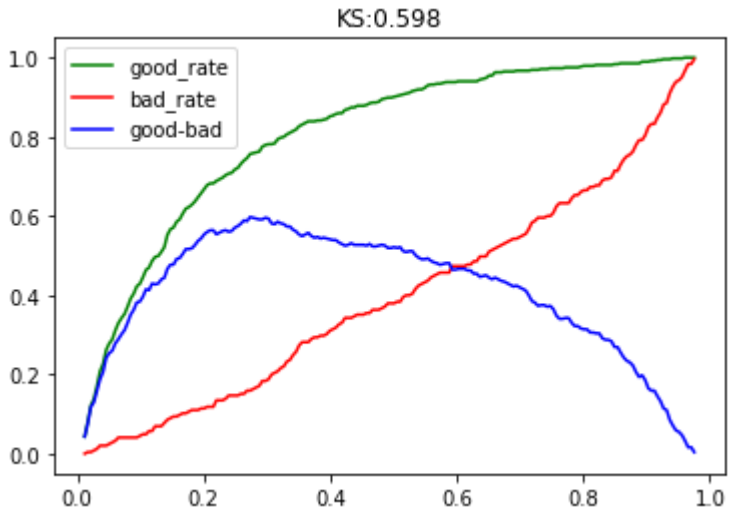
训练集的AUC，KS：





验证集的AUC，KS：

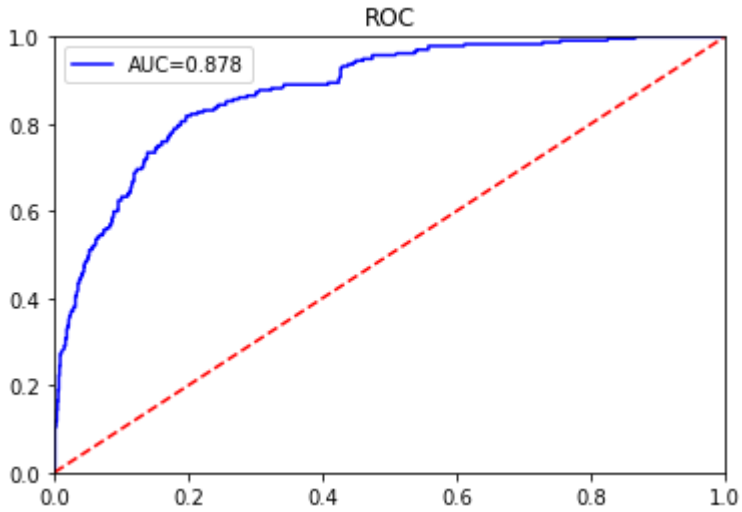

ROC
AUC=0.865
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

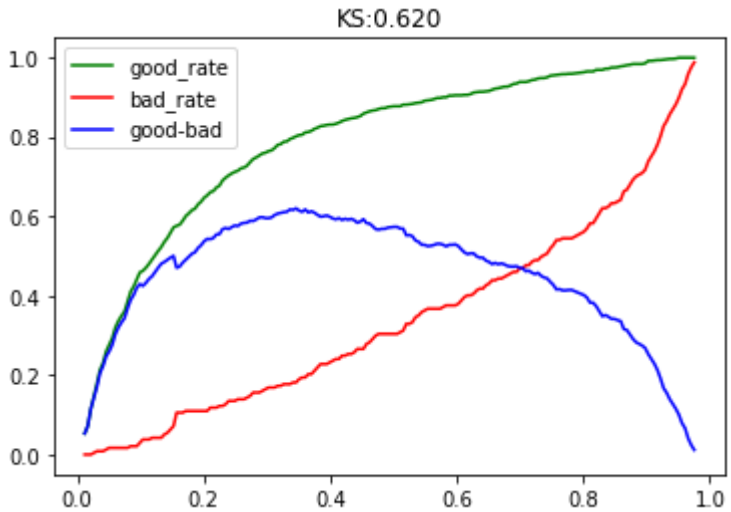

KS:0.598
good_rate
bad_rate
good-bad
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

时间外样本集好坏样本数

```
0    1016
1     237
Name: target, dtype: int64
```

时间外样本集的AUC，KS





分数映射&分数分布

分数刻度&各入模变量相应分箱得分

评分卡刻度

Search:

type
A
B
base_score

Showing 1 to 3 of 3 entries

变量各分箱对应的分数

Show 10 entries Search:

col	bin	IV
v13	(-inf, -999.0]	1.032758
v13	(-999.0, 568.5]	1.032758
v13	(568.5, 619.5]	1.032758
v13	(619.5, 685.5]	1.032758
v13	(685.5, inf]	1.032758
v16	(-inf, -999.0]	0.876177
v16	(-999.0, 3.5]	0.876177
v16	(3.5, 5.5]	0.876177
v16	(5.5, 6.5]	0.876177
v16	(6.5, inf]	0.876177

Showing 1 to 10 of 34 entries Previous 1 2 3 4 Next

训练集&验证集&时间外样本分数转换

评分转换完成----------

训练集&验证集&时间外样本分数分箱分布

训练集评分分箱分布

Show 10 entries Search:

	final_score	ks	pass_rate	total
0	(559.999, 589.0]	13.18%	89.58%	304
1	(589.0, 612.0]	24.65%	79.67%	289
2	(612.0, 632.0]	36.74%	69.56%	295
3	(632.0, 643.0]	41.26%	64.62%	144
4	(643.0, 654.0]	44.72%	59.55%	148
5	(654.0, 665.0]	48.44%	54.51%	147
6	(665.0, 675.0]	49.74%	49.61%	143
7	(675.0, 690.0]	48.12%	44.53%	148
8	(690.0, 703.2]	44.83%	39.73%	140
9	(703.2, 737.0]	34.23%	29.52%	298

Showing 1 to 10 of 12 entries Previous 1 2 Next

验证集评分分箱分布

Show 10 entries Search:

	final_score	ks	pass_rate	total
0	(559.999, 590.0]	12.80%	89.45%	77
1	(590.0, 618.7]	25.19%	79.59%	72
2	(618.7, 637.0]	34.51%	69.32%	75
3	(637.0, 648.0]	38.86%	64.38%	36
4	(648.0, 658.5]	43.01%	59.59%	35
5	(658.5, 671.0]	46.34%	54.11%	40
6	(671.0, 681.4]	45.77%	49.59%	33
7	(681.4, 689.0]	42.55%	44.38%	38
8	(689.0, 704.3]	39.32%	39.59%	35
9	(704.3, 740.2]	31.64%	29.59%	73

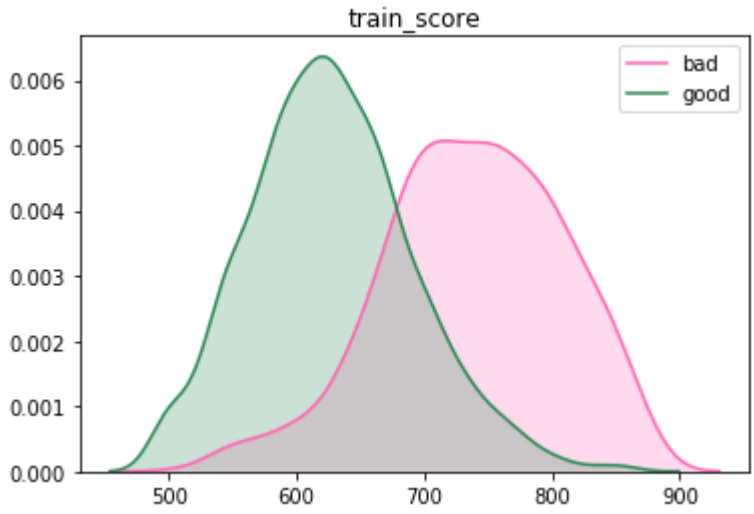
Showing 1 to 10 of 12 entries Previous 1 2 Next

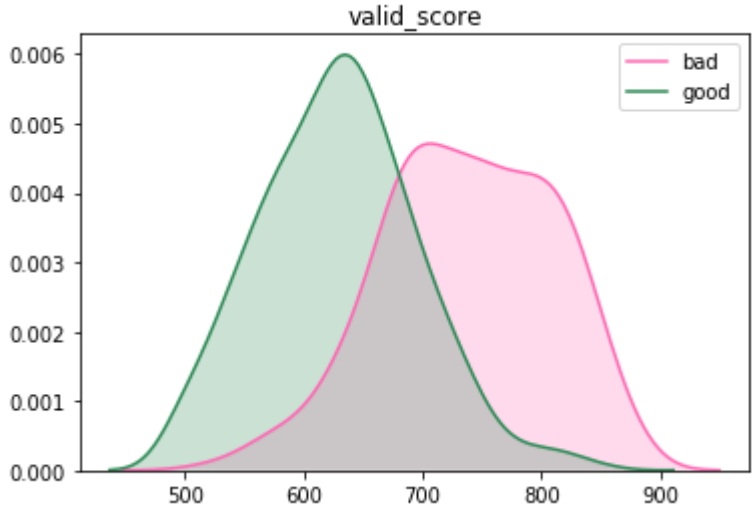
时间外验证集评分分箱分布

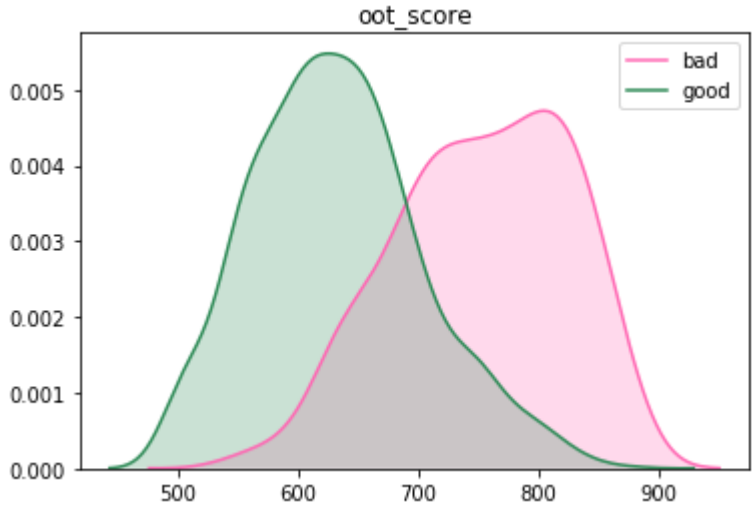
Show 10 entries Search:

	final_score	ks	pass_rate	total
0	(550.999, 581.0]	11.92%	89.07%	137
1	(581.0, 606.0]	23.71%	79.09%	125
2	(606.0, 626.0]	33.08%	68.95%	127
3	(626.0, 636.0]	37.72%	63.93%	63
4	(636.0, 645.0]	36.86%	59.14%	60
5	(645.0, 656.0]	42.97%	54.19%	62
6	(656.0, 666.0]	45.85%	49.32%	61
7	(666.0, 678.0]	48.21%	44.45%	61
8	(678.0, 691.0]	49.11%	39.51%	62
9	(691.0, 729.0]	41.41%	29.29%	128

Showing 1 to 10 of 12 entries Previous 1 2 Next






oot_score
bad
good
0.005
0.004
0.003
0.002
0.001
0.000
500
600
700
800
900