德国信用风险数据分析

1 课题背景

银行信用风险的大小和质量决定着银行盈利水平的高低，对银行证券稳定、长远的发展有着至关重要的影响，银行使用数据挖掘方法建立目的明确、层次分明的信用风险分析模型有着重要价值。

早期的信用风险研究寻求数学解决方法，Z分数模型等都是比较具有代表性的方法。随着银行信贷的大规模增长及客户信用信息的迅速变化，形成了复杂的数据资源，信用风险的形式与日俱增。因此，Hashemi and Blanc、Guilherme Barreto Fernandes、謝宇等分别采用神经网络和粗糙集成分集合、logistic模型作为解释变量、改进BP人工神经网络模型对银行信用风险进行预测得到了较好改进。但以上的方法在预测精度和准确性上还有待提高。

本文结合<华泰证券人工智能系列,人工智能选股函数的改进> 引入XGBoost（eXtreme Gradient Boosting)算法建立信用风险分析优化模型，基于UCI上德国信用数据集与决策树、GBDT、支持向量机等模型进行对比分析，验证了改进后的xgboost损失函数应用于信用风险分析具有更好的性能。

2 数据维度分析

数据集描述

数据集包含1000条观测，每个观测有20个categorial(类别型，可重编码为numeric)或symbolic(符号型)属性。每条观测代表某人在银行开卡的记录，并且每条观测根据其属性值被打上了good或bad信用风险的标签。

数据集内容

原数据集内容比较复杂,整理出来的特征字段包括

字段名称	字段类型	字段描述
Age	numeric	
Sex	text	male, female
Job	numeric	0 - unskilled and non-resident,1 - unskilled and resident, 2 - skilled, 3 - highly skilled
Housing	text	own, rent, or free)
Saving accounts	text	little, moderate, quite rich, rich
Checking account	numeric	in DM - Deutsch Mark
Credit amount	numeric	in DM
Duration	numeric	in month
Purpose	text	car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/
target	text	- Good or Bad Risk

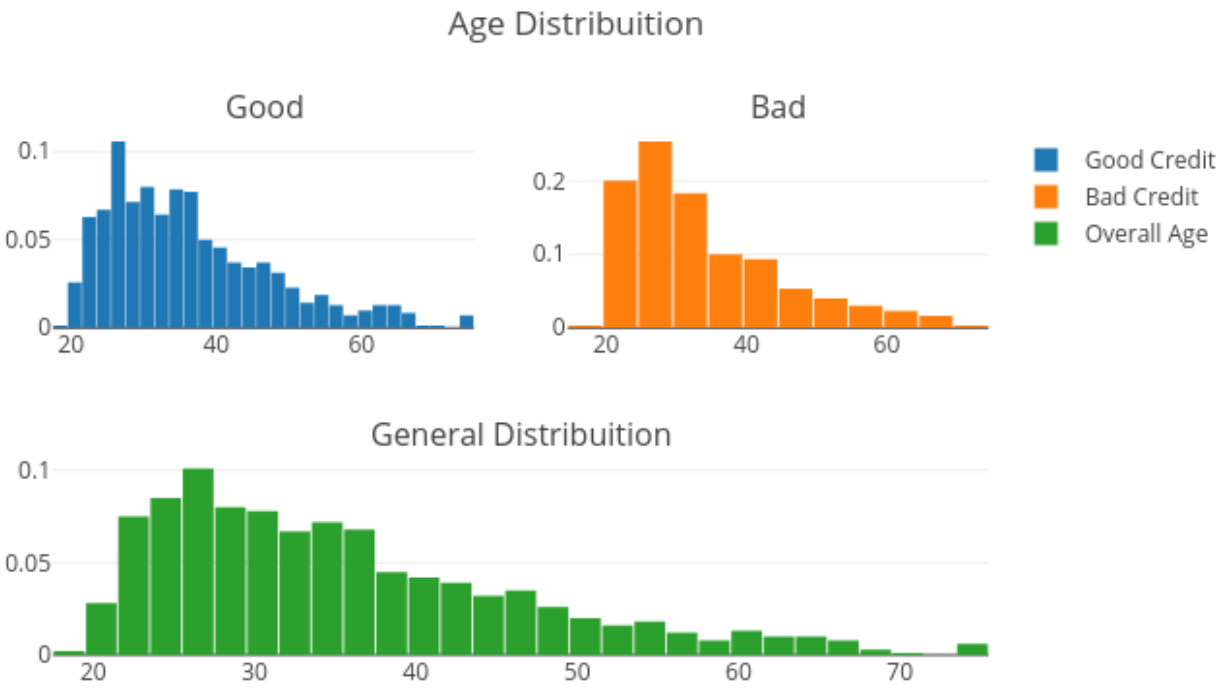
3 探索性分析

3.1 信用样本正负比例图

图[1]数据样本正负比例,可以看到数据中,正样本的比例70%,负样本比例30%

3.2 信用样本与年龄分布图

图[2]信用样本与年龄分布图,可以看到真负样本的各个年龄段的统计情况,可以看到样本中,数据年龄段区间集中在20到40岁.在信用良好的的年龄区间中,27岁信用良好占比最高,占比10.57%,在信用不良的年龄区间中,25-29岁之间的不良占比最高,占比25.33% 在所有样本空间中,年龄区间跨越0到70岁,其中27岁样本最高,占比10.1%,70岁样本最低,占比1%.



3.3 年龄区间与存款关系图

图片详情页

将样本按照如下区间进行切分

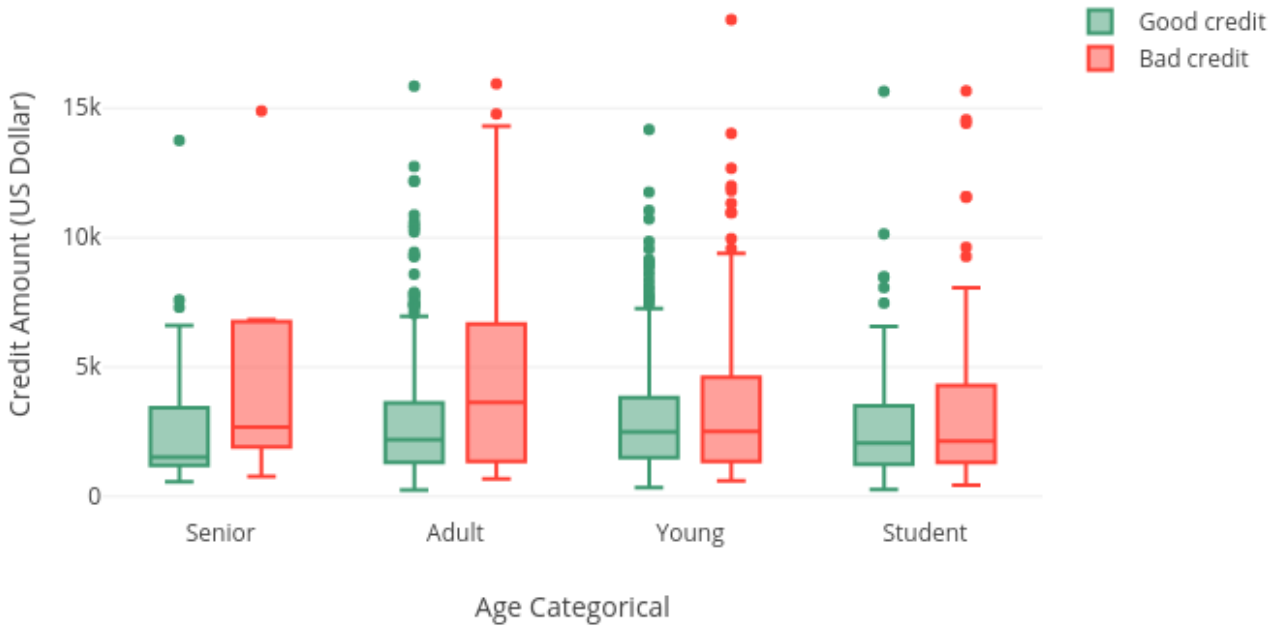
年龄段	范围
student	0~18
young	18~25
Adult	25~35
Senior	35~60

箱状图	解释
上四分位数	75%处
中位数	50%处
下四分位数	25%处
异常值	大于上四分位数1.5倍四分位数差的值，或者小于下四分位数1.5倍四分位数差的值
四分位数差	计算上四分位数和下四分位数之间的差值，即四分位数差

图[3] 画出年龄与存款之间的箱状图

年龄段	信用状况	min	下四分位数	中位数	上四分位数	max	平均值	四分位区间
student	good	276	1236	2055	3509	15653	2372.5	2273
student	bad	433	1313	2134.5	4294.5	15672	2803.75	2981.5
young	good	343	1493.5	2495.5	3815	14179	2654.25	2321.5
young	bad	609	1345	2509	4611	18424	2978	3266
Adult	good	250	1319	2197	3607.5	15857	2463.25	2288.5
Adult	bad	684	1351	3625	6670	15945	4010.5	5319
Senior	good	571	1209	1520	3434.5	13756	2321.75	2225.5
Senior	bad	766	1908	2683.5	6761	14896	4334.5	4853

可知，在所有的样本中，学生段的存款能力最弱，中年人段的存款能力最强。年龄区间在35~60之间的存款均值最大，0-18区间的存款均值最小。在所有的好信用分类中，25-35年龄区间的存款均值最大，但是其存款数额的范围也是好信用中最大的。

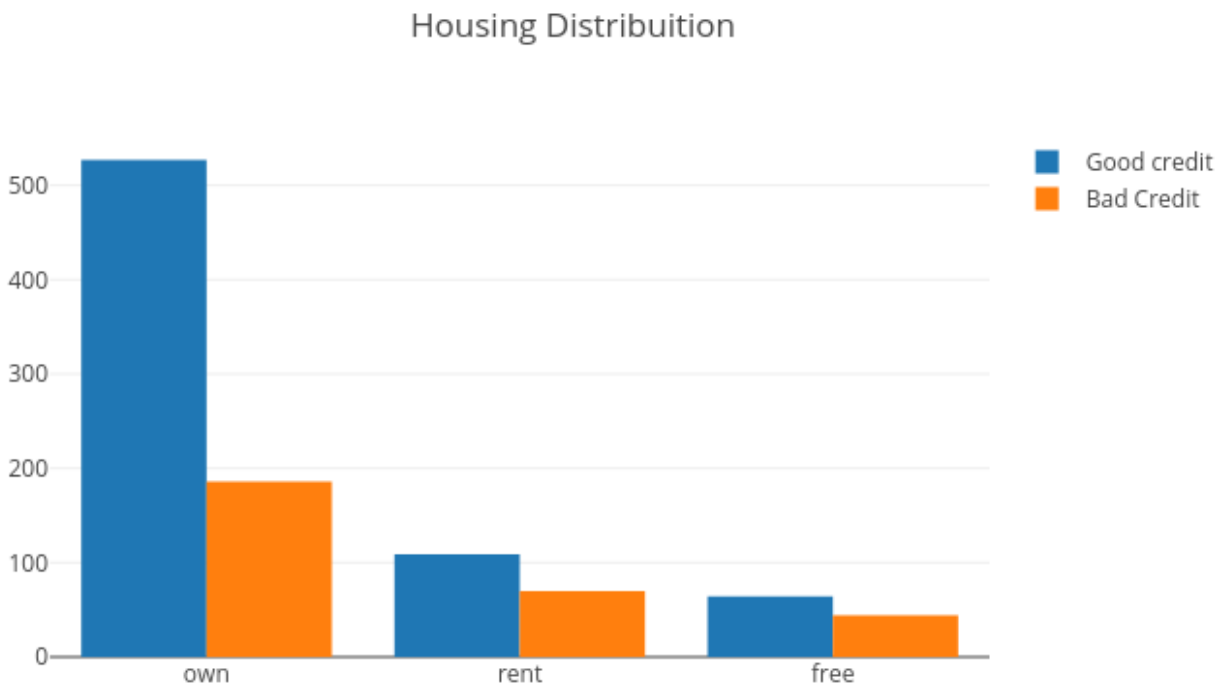


3.4 信用与house关系

house详情

房屋情况	信用状况	样本数	所占该类比例	总样本占比
own	good	527	73.9%	52.7%
own	bad	186	26.1%	18.6%
rent	good	109	60.9%	10.9%
rent	bad	70	38.1%	7%
free	good	64	59.3%	6.4%
free	bad	44	40.7%	4.4%

由图可知，数据中大部分客户的信用均是好，约占70%。大部分的客户都有自己的房子，而拥有自住房的存款客户信用好的比例也是最大。在rent和free中，信用好坏的样本比例大致相等。

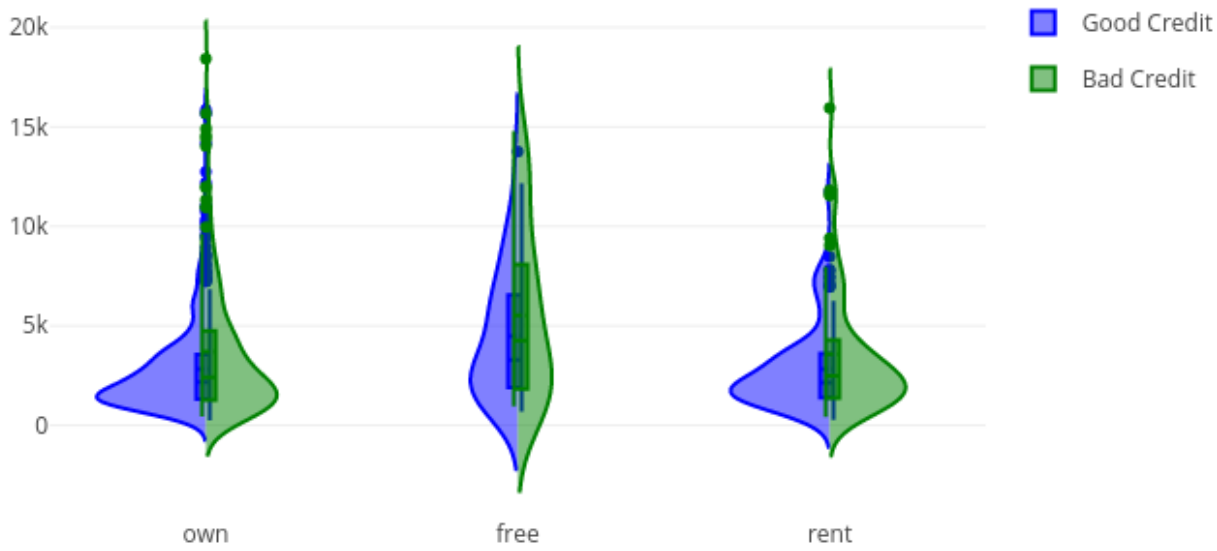


3.5 信用和housing以及存款之间的关系

链接

房屋情况	信用状况	min	下四分位数	中位数	上四分位数	max	平均值	四分位区间
own	good	250	1332	2171	3562.75	15857	2837.577	2230.75
own	bad	448	1282	2418	4746	18424	3693.8	3464
free	good	700	1906	3296	6553	13756	4472.9	4647
free	bad	947	1828	4268	8091.5	14782	5536.5	6263.5
rent	good	276	1402.75	2146	3603.5	11760	2827.1	2200.75
rent	bad	433	1371	2488	4280	15945	3582.7	2909

由图表可知，房屋情况为free的客户其存款区间最为分散，房屋情况为own的客户存款范围最为稳定。own和rent房子的客户信用好坏比例大致相等。房屋情况为free的存款均值最大，但是其信用也为bad。纵观上表数据，信用为bad的存款均值均比good高。拥有房的bad credit客户存款最大额最大。



3.6 信用与性别之间的关系

左图为不同性别正负样本比例，男性正样本约占男性总样本的72.3%，女性正样本约占女性总样本的64.8%。（男性比女性信用好？）

右图为不同性别的信贷金额，男性正例贷款金额中位数是2346，75%的人在3845.5以下；负例中位数金额2820，75%的人在6148.75以下。女性正例贷款金额中位数是1927，75%的人在3457.3以下；负例中位数金额2039，75%的人在4383.75以下。明显看出女性比男性信贷金额低，正样本比负样本信贷金额低。（不管什么性别信用不好的往往比信用好的贷款高）

画出性别与信用之间的箱状图

性别	信用状况	min	下四分位数	中位数	上四分位数	上边缘	max
male	good	276	1413.25	2346	3845.5	7476	15.857k
male	bad	639	1516.5	2820	6148.75	12.68k	15.945k
female	good	250	1258	1927	3457.25	6419	10.722k
female	bad	433	1234.75	2039	4383.75	8318	18.424k

链接

3.7 信用与工作之间的关系

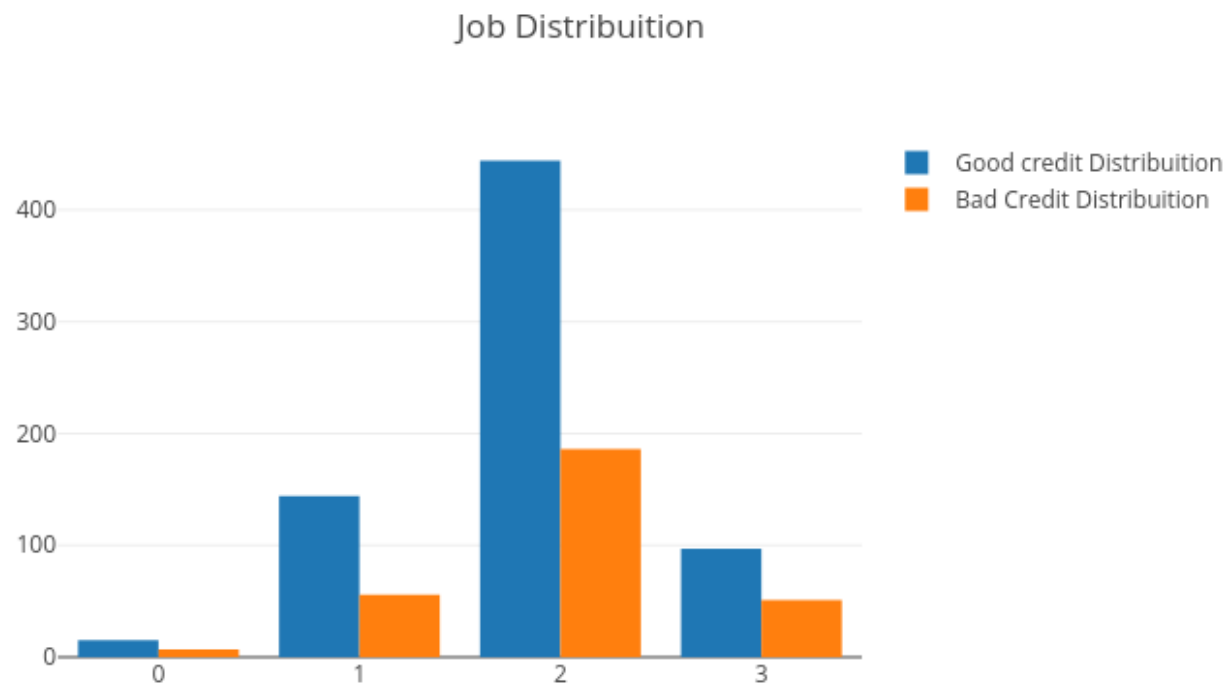
上图为不同工作正负样本统计数据，其中非熟练非居民占总样本2.3%，非熟练居民20%，熟练者占63%，高熟练者占14.8%。非熟练非居民中50%信用良好，非熟练居民和熟练者中70%左右，而高熟练者只有65%。

下图为不同工作正负样本信贷金额统计数据，正样本中75%非熟练非居民信贷金额3716以下，75%非熟练居民3091.5以下，75%熟练者3545.5以下，75%高熟练者6781.25以下。负样本中75%非熟练非居民信贷金额2299.25以下，75%非熟练居民2742.5以下，75%熟练者4843以下，75%高熟练者9307以下。（说明工作越好信贷额度越高，信用不好的往往比信用好的贷款高）

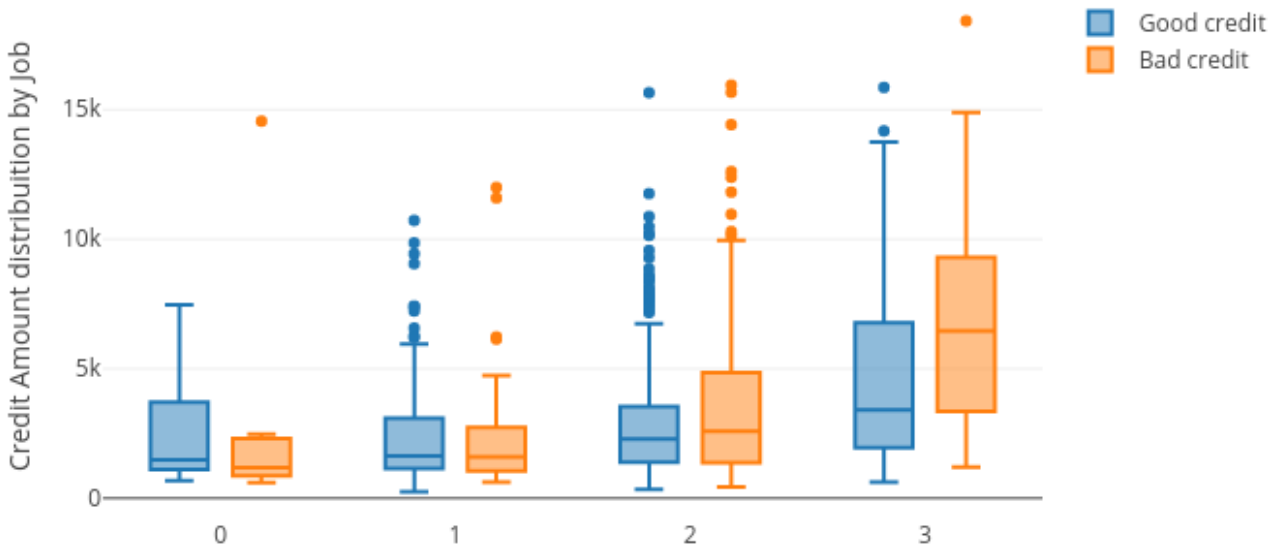
画出信用与工作之间的箱状图

工作	信用状况	min	下四分位数	中位数	上四分位数	上边缘	max
0（非熟练非居民）	good	672	1117.25	1480	3716	7472	7472
0（非熟练非居民）	bad	609	860	1193	2299.25	2473	14.555k
1（非熟练居民）	good	250	1146	1626	3091.5	5954	10.722k
1（非熟练居民）	bad	626	1049.5	1596.5	2742.5	4746	11.998k
2（熟练者）	good	338	1392	2281.5	1545.5	6742	15.653k
2（熟练者）	bad	433	1371	2585	4843	9.96k	15.945k
3（高度熟练者）	good	629	1955.75	3416	6781.25	13.756k	15.857k
3（高度熟练者）	bad	1209	3346	6458	9307	14.896k	18.424k

链接



链接



3.8 信用跟年龄和工作之间的关系

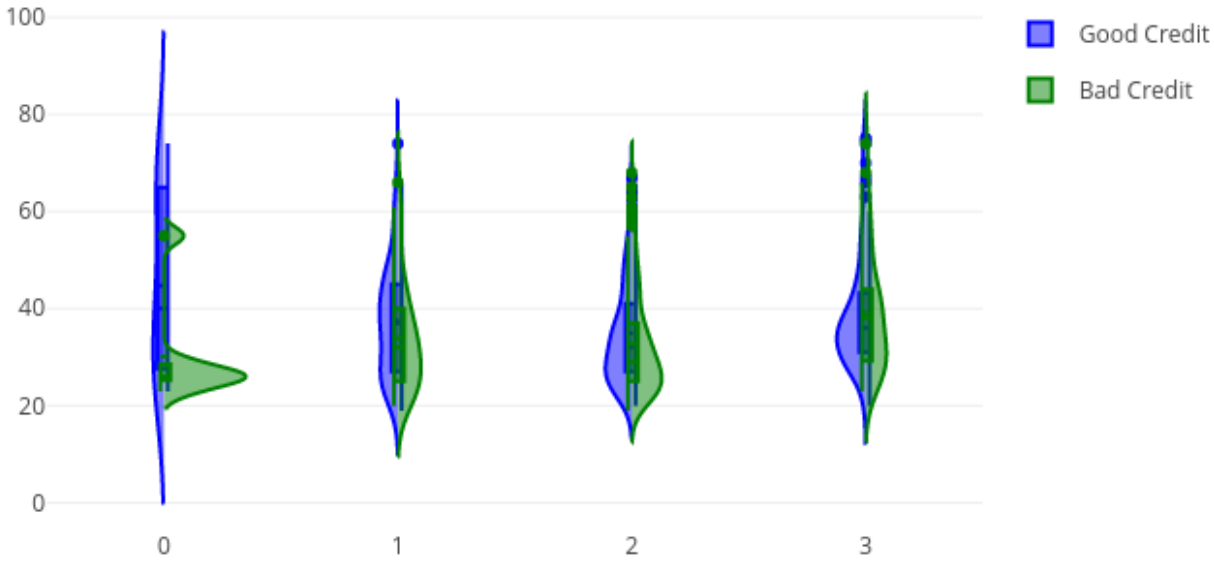
非熟练非居民中，负样本大多年龄在25-28岁之间，正样本年龄27-65岁之间；非熟练居民中，负样本大多年龄在25-40岁之间，正样本年龄27-45岁之间；熟练者中，负样本大多年龄在25-37岁之间，正样本年龄27-41岁之间；高度熟练者中，负样本大多年龄在29-44岁之间，正样本年龄31-43岁之间.（工作一般或者不太好的情况下，年龄稍大的人更被认为可能有良好的信用，工作好信用也好，不过工作太好也可能存在风险吧）

小提琴图则展示了任意位置的密度，通过小提琴图可以知道哪些位置的密度较高,主要在需要观察分布密度时使用.

画出信用跟年龄和工作之间的小提琴图

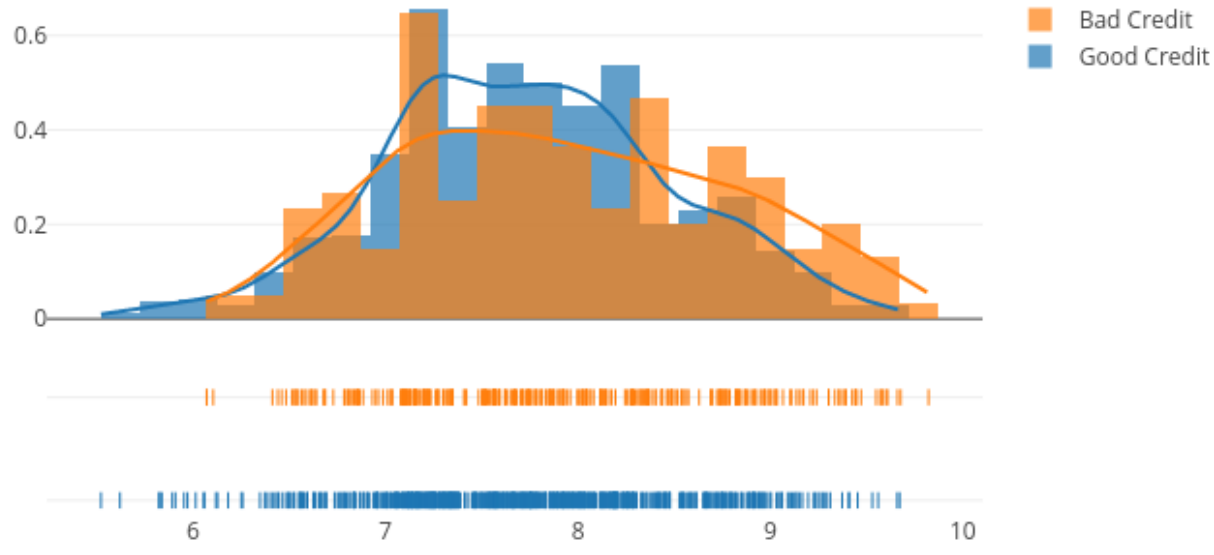
工作	信用状况	min	下四分位数	中位数	上四分位数	上边缘	max	平均数
0	good	23	27.5	40	65		74	44.7
0	bad	23	25.25	26	28.5	29	55	30
1	good	19	27	37	45	66	74	37.5
1	bad	20	25	32	40	61	66	40
2	good	20	27	32	41	62	67	34.9
2	bad	19	25	29	37	55	68	32.7
3	good	20	31	36	43.25	60	75	39
3	bad	23	29.25	38	44	66	74	39

链接

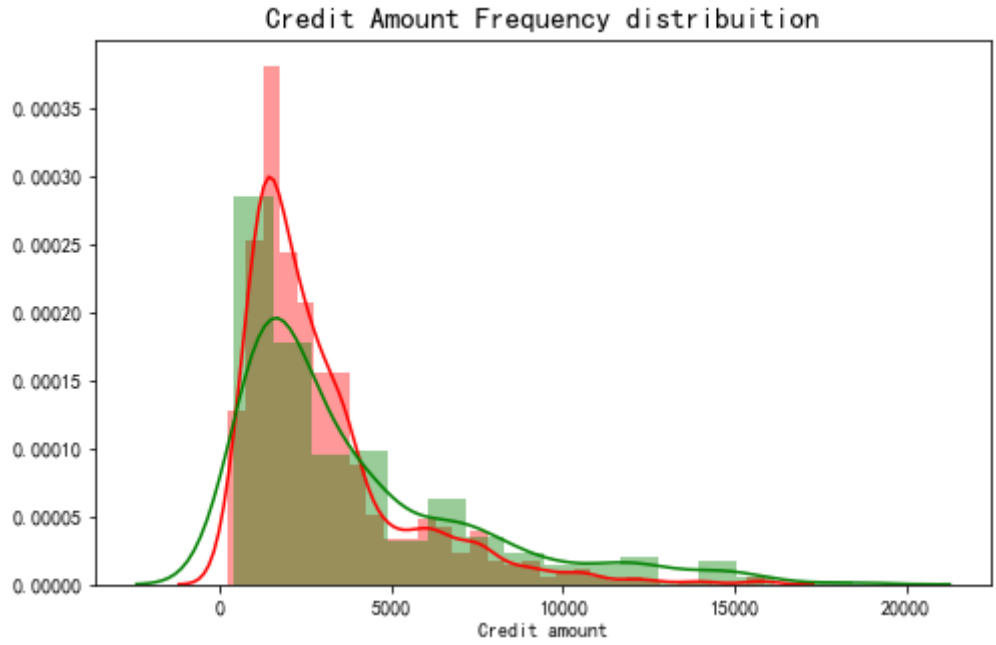


3.9 信用和银行存款之间的关系

信用额度取对数之后做的分布情况 链接

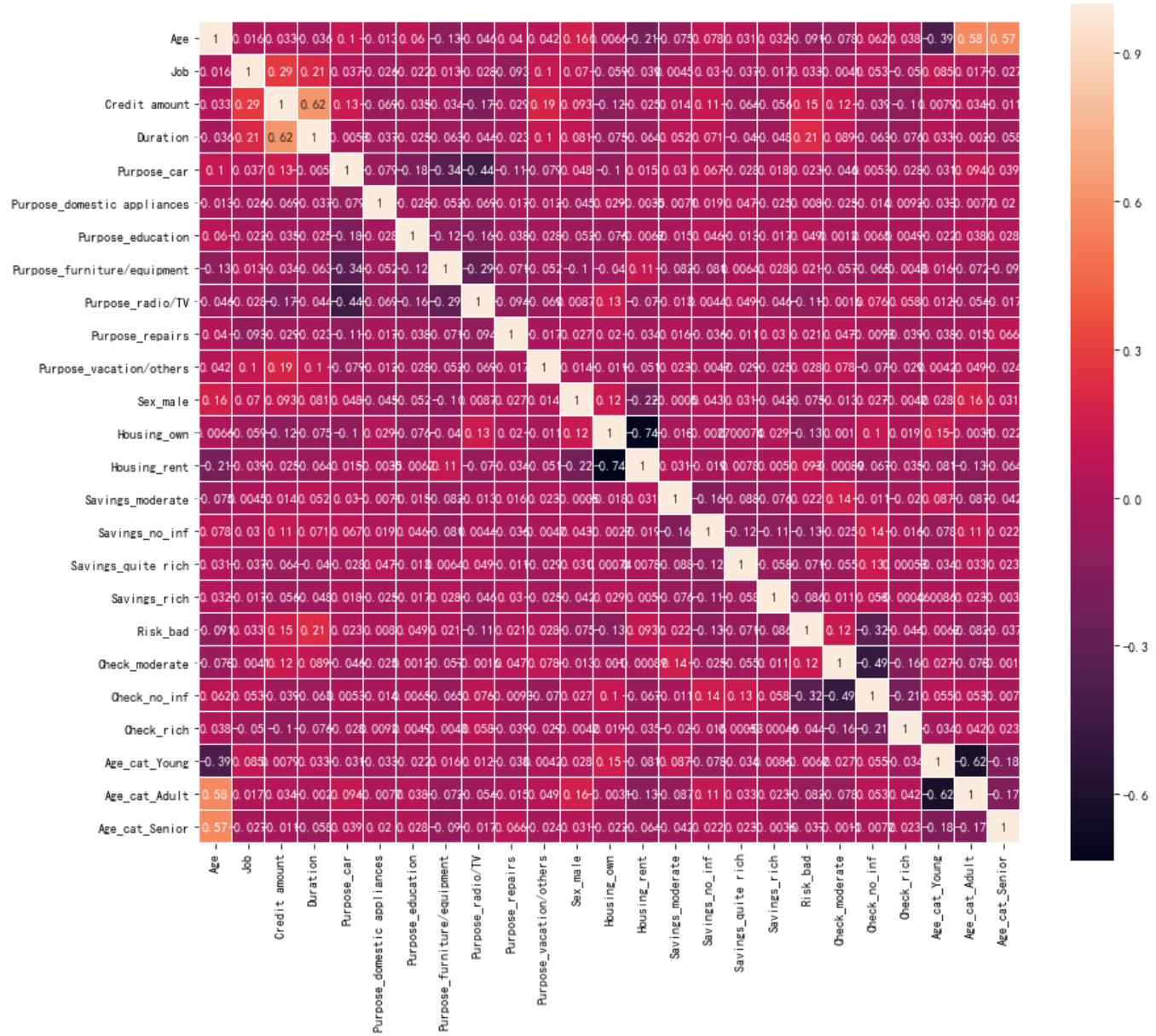


信用额度频率分布，大多数人贷款金额在0-5000



储蓄金额和信用关系（70%的人几乎没存款，越有钱信用越良好）储蓄金额和信贷金额和信用的关系（信用差的人贷的多，因为贷的多所以还不上所以信用差）

看的是所有特征两两的相似度，例如age和Age_cat_Adult的相似度就比较高



4 模型

4.1 xgboost

xgboost 官方文档 https://xgboost.readthedocs.io/en/latest/

4.2 xgboost和其他算法对比

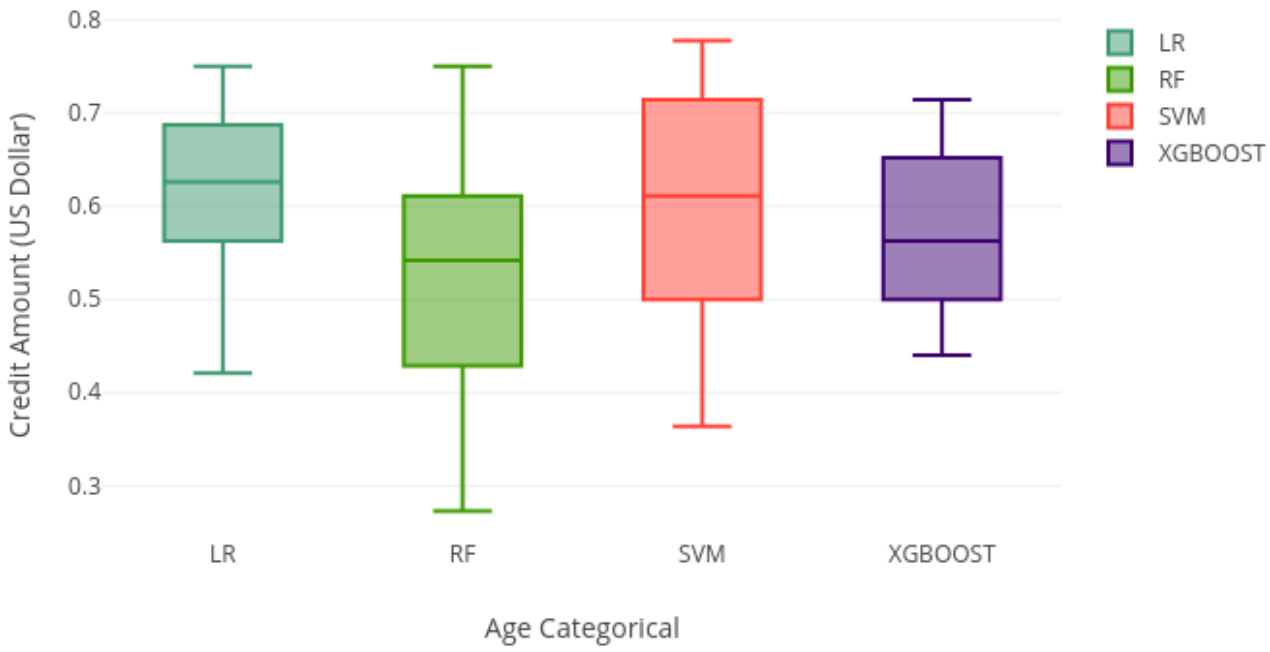
- 算法:SVM,LR,随机深林,xgboost
- 评价:精确率,召回率,f1scor

实验方法: 将数据按照10折交叉验证,分别计算svm,LR,随机深林以及xgboost的精确率,召回率,和f1score,实验结果如下:

精确率

index	LR	RF	SVM	XGBOOST
0	0.692308	0.545455	0.714286	0.652174
1	0.750000	0.750000	0.500000	0.636364
2	0.562500	0.437500	0.400000	0.560000
3	0.571429	0.357143	0.666667	0.541667
4	0.529412	0.600000	0.555556	0.500000
5	0.421053	0.272727	0.363636	0.440000
6	0.615385	0.666667	0.666667	0.680000
7	0.636364	0.611111	0.750000	0.565217
8	0.687500	0.538462	0.777778	0.444444
9	0.666667	0.428571	0.500000	0.714286
avg	0.6132618	0.5207636	0.589459	0.5734152

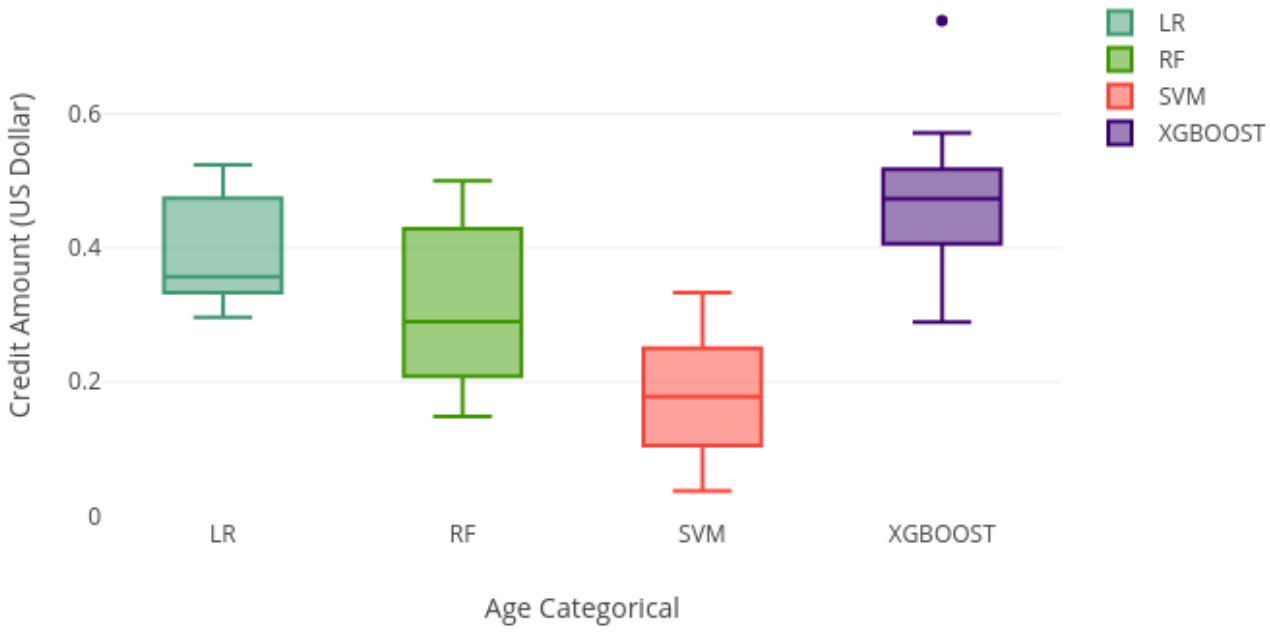
https://plot.ly/~fengweijie/23/



召回率

index	LR	RF	SVM	XGBOOST
0	0.333333	0.185185	0.185185	0.517241
1	0.473684	0.263158	0.105263	0.466667
2	0.391304	0.304348	0.173913	0.437500
3	0.363636	0.318182	0.181818	0.481481
4	0.310345	0.275862	0.172414	0.379310
5	0.500000	0.500000	0.250000	0.289474
6	0.333333	0.208333	0.083333	0.739130
7	0.350000	0.450000	0.300000	0.406250
8	0.523810	0.428571	0.333333	0.480000
9	0.296296	0.148148	0.037037	0.571429
avg	0.3875741	0.3081787	0.1822296	0.4768482

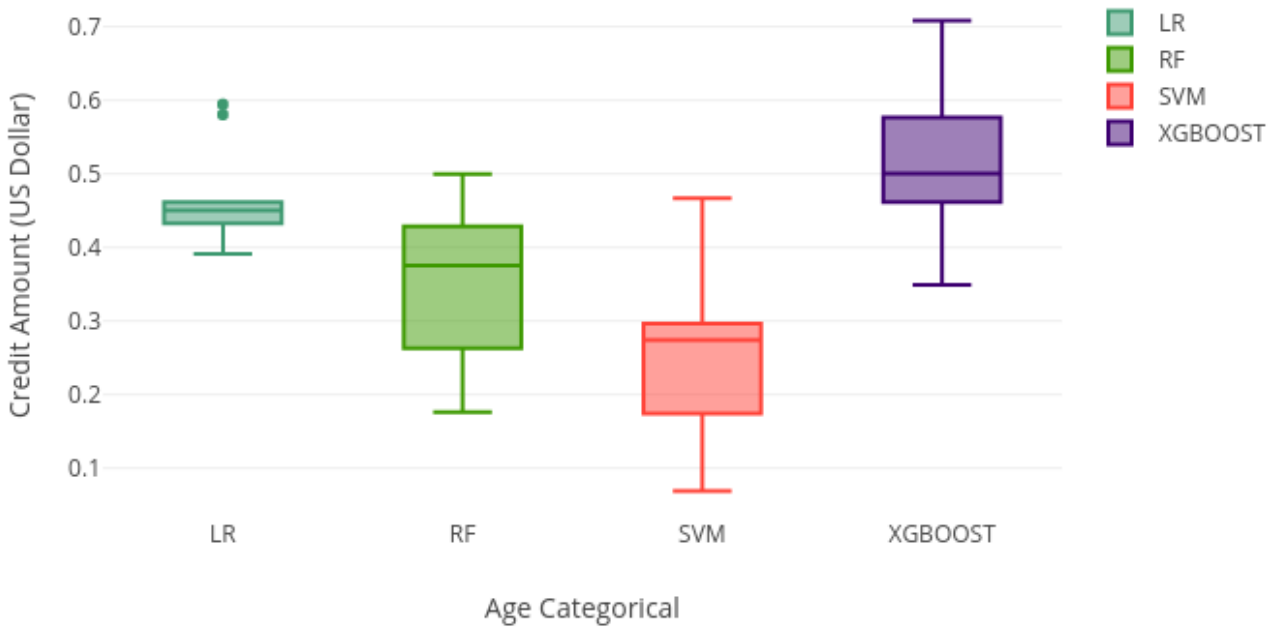
https://plot.ly/~fengweijie/25/



f1score

index	LR	RF	SVM	XGBOOST
0	0.450000	0.176471	0.294118	0.576923
1	0.580645	0.275862	0.173913	0.538462
2	0.461538	0.263158	0.242424	0.491228
3	0.444444	0.378378	0.285714	0.509804
4	0.391304	0.372093	0.263158	0.431373
5	0.457143	0.200000	0.296296	0.349206
6	0.432432	0.451613	0.148148	0.708333
7	0.451613	0.500000	0.428571	0.472727
8	0.594595	0.424242	0.466667	0.461538
9	0.410256	0.428571	0.068966	0.634921
avg	0.467397	0.3470388	0.2667975	0.5174515

https://plot.ly/~fengweijie/27/



综合模型的精确率和召回率,以及F1score的实验表现,xgboost相对其他算法,在数据的能够同时兼顾召回率和准确率,在风险评估预测上有比较大的优势.

4.3 改进损失函数后的xgboost

对于改进xgboost中的自定义损失函数接口,改进的损失函数为

$$\text{Weighted\_Loss} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log\left(f(x_i)\right) + \beta(1-y_i)\log\left(1-f(x_i)\right)\right)$$

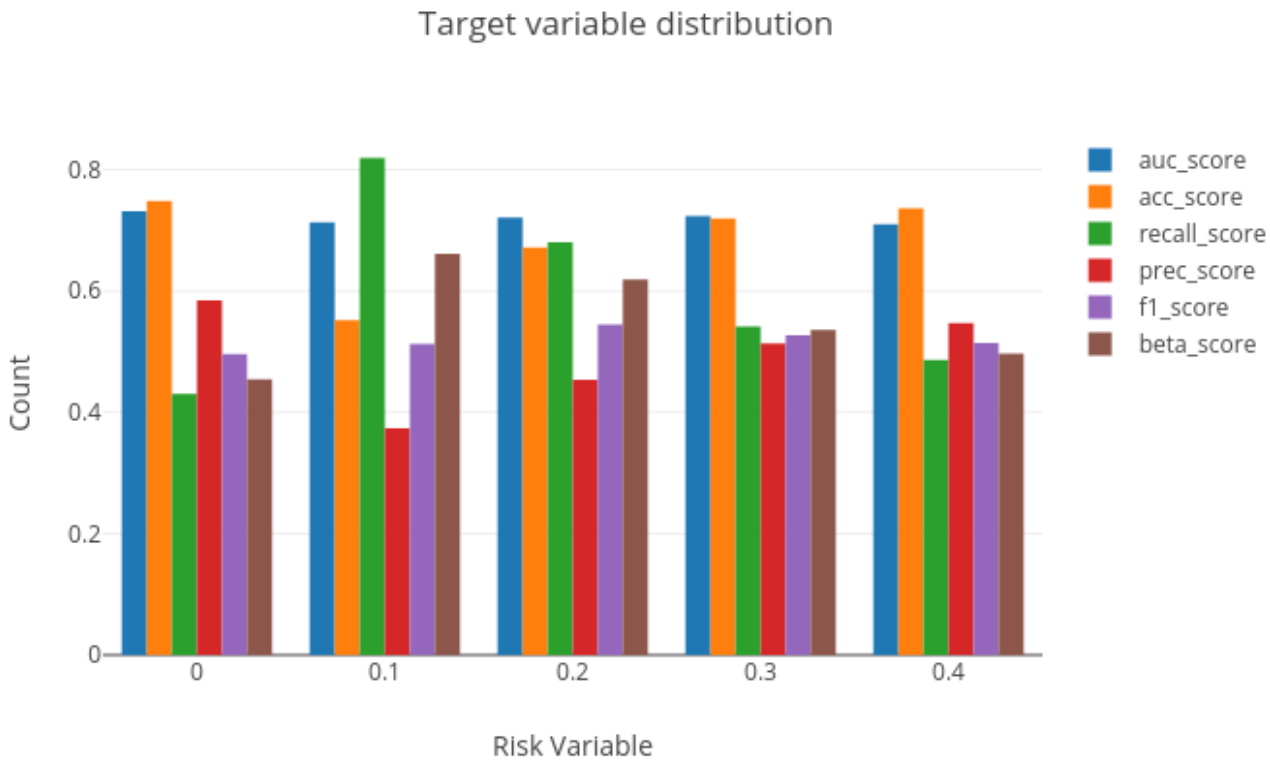
关于β值的确定

- 假设正样本个数为n 1 ,负样本个数为n 2 ,n 1 + n 2 = n,则β = n 1 /n 2 。
- 当n 1 ≫ n 2 时,说明正样本数目远多于负样本,此时β取值很大,模型会重点针对Loss2进 行优化。
- 当n 1 ≪ n 2 时,说明负样本数目远多于正样本,此时β取值较小,模型会重点针对Loss1进 行优化。

二阶损失函数推到见<华泰人工智能选股第16页> 实现代码如下:

```
import numpy as np
beta = 0.2
def weightloss(preds,dtrain):
    y = dtrain.get_label()
    p = 1.0 / (1.0 + np.exp(-preds ))
    grad = p * (beta + y - beta*y) - y
    hess = p*(1-p)*(beta + y - beta*y)
    return grad,hess
```

beta	auc_score	acc_score	recall_score	prec_score	f1_score
1	0.731976	0.748	0.430556	0.584906	0.496000
0.1	0.713561	0.552	0.819444	0.373418	0.513043
0.2	0.721403	0.672	0.680556	0.453704	0.544444
0.3	0.723237	0.720	0.541667	0.513158	0.527027
0.4	0.709699	0.736	0.486111	0.546875	0.514706



5 总结

本文我们首先介绍了德国信用风险评估的背景,之后对数据与风险评估做了深入的分析,接着对对比LR,SVM,决策树以及随机森林算法性能的比较,最后根据样本不平衡情况修改了损失函数,初步得出以下结论:

- 在本次信用风险评估中,通过对比LR,svm,随机深林以及xgboost,证明xgboost比其他算法性能,在预测模型上,有比较好的准确率
- 在本次风险评估中,对数损失函数是机器学习中最常用的二分类模型损失函数,由逻辑回归的极大似然估 计过程推导而来。对数损失函数可以被分解为两项,分别代表二分类的假阳性误差和假阴性误差,在普通的对数损失函数中,两类误差的权重是相等的。
- 针对分类模型中两类样本不均衡的问题,我们引入了加权损失函数并给出了具体的形式,该损失函数能增大数量较少一类样本的损失项权重。进一步提升了xgboost在预测模型上的表现.

6 参考文献

1. 《统计学习方法》 李航
2. 《机器学习》 周志华
3. 机器学习——深度学习(Deep Learning)