信用卡评分模型 Give Me Some Credit

联系邮箱：chenyy@stu.pku.edu.cn, 2000013149@stu.pku.edu.cn, chang18102101836@pku.edu.cn

指导教师：童云海

北京大学，人工智能引论课程
2020-2021，春季学期

摘要

信用评分卡模型在国外是一种成熟的预测方法，尤其在信用风险评估以及金融风险控制领域更是得到了比较广泛的使用。本小组利用Kaggle上的15万条数据，对数据加以清洗，利用随机森林法填补缺失值，采用模型变量WOE编码方式离散化，并运用logistic回归模型，建立并训练二分类变量的广义线性模型，最终建立具有良好评估效果的信用评分卡。

关键词： 信用卡 随机森林 逻辑回归 信用评分卡

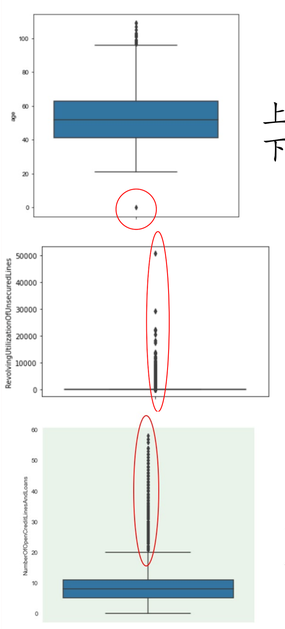
引言

银行在银行在决定是否向客户发放贷款时，会首先对贷款者的信用进行评分。一类简单的信用评分算法是根据贷款者的各项特征行计算，得到贷款者的违约概率，最终银行倾向于贷款给那些违约概率较低的贷款者。我们的任务是，使用某种自己感兴趣的方法，分析数据不同特征对于用户信用评分的重要程度，以及不同特征之间的相关关系，并完成信用评分（300 – 900分）算法的建模以及算法实现。

建模准备

①数据导入——十个特征，一个结论

S/n	Variable Name
1	SeriousDlqin2yrs
2	RevolvingUtilizationOfUnsecuredLines
3	age
4	NumberOfTime30-59DaysPastDueNotWorse
5	DebtRatio
6	MonthlyIncome
7	NumberOfOpenCreditLinesAndLoans
8	NumberOfTimes90DaysLate
9	NumberRealEstateLoansOrLines
10	NumberOfTime60-89DaysPastDueNotWorse
11	NumberOfDependents

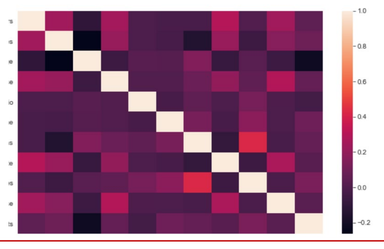


年龄
上界连续可保留
下方零值需删除

可用额度比值
比值必然小于一
上方大数均异常

信贷数量
上界连续无极端
所有数据均保留

②数据分析——异常较多，需要筛选
（以三个特征为例展示）



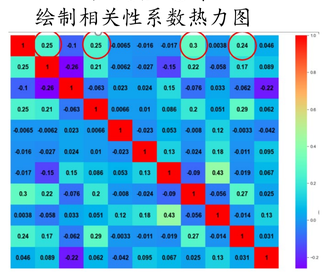
随机森林是一种新兴的机器学习算法，而且近些年他在随机森林领域的运用十分广泛，当然也是因为他本身非常灵活实用。它是一种以决策树作为基本单元，将多棵树集成的算法，具有当下算法中非常高的准确度，对于大数据的适用性很好

③数据预处理——参考前期分析，理性进行处理

④特征选择——选择合适特征，准备建模分析

- WOE 编码：$WOE_i = \ln(\frac{P(y_i)}{P(y_n)}) = \ln(\frac{y_i/y_T}{n_i/n_T})$
- IV 值计算：$IV_i = (P(y_i) - P(n_i)) \times WOE_i = \left(\frac{y_i}{y_T} - \frac{n_i}{n_T}\right) \times \ln(\frac{y_i/y_T}{n_i/n_T})$

相关性选择
绘制相关性系数热力图



通过相关性系数热力图可以看出，自变量与因变量相关性并不强，说明仅通过相关性不能做出较好的特征选择。

两个维度特征选择

IV值筛选
计算并绘制IV值柱状图

IV值大于0.1即代表该特征有中等的预测价值，大于0.3即代表该特征预测的能力很强。因此通过上图可以看出，有5个特征IV值较高，可以挑选出来建立模型。

模型建立

⑤模型建立——WOE转换、丢弃变量并进行逻辑回归

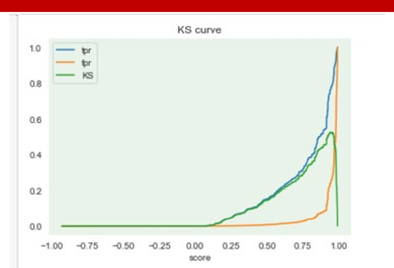
X = 自变量
Y = 因变量

```
Y=training['SeriousDlqin2yrs']    #因变量
#剔除对因变量影响不明显的变量
X=training.drop(['SeriousDlqin2yrs','DebtRatio','Monthl
X=training.iloc[:,-5:]
X.head(5)
```

```
#5.2 利用STATSMODEL包来建立逻辑回归模型得到回归系数，后面可用于建立标准评分卡
X1=sm.add_constant(X)
logit=sm.Logit(Y,X1)
result=logit.fit()
print(result)
print(result.summary())
```

	coef
const	-2.7251
RevolvingUtilizationOfUnsecuredLineswoe	0.6565
agewoe	0.5041
NumberOfTime30-59DaysPastDueNotWorsewoe	0.5567
NumberOfTimes90DaysLatewoe	0.5965
NumberOfTime60-89DaysPastDueNotWorsewoe	0.4276

“coef”一栏即为逻辑回归所求得的回归系数

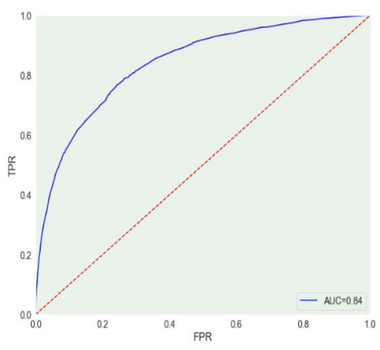


<Figure size 1440x1440 with 0 Axes>

0.5241713995900088

KS值高达0.524，说明模型区分效果不错

⑥模型评估——AUC值与KS值



AUC值高达0.84，说明模型拟合效果不错

⑦建立评分卡——构造评分函数，进行评分

$$Score = \sum_1^n \left(WOE_i \times coef_i + \frac{coef_0}{n} \right) \times factor + offset$$

⑧对测试集进行预测——经典再放送

	predict	BaseScore	ilizationOfUns	age	e30-59DaysPastDu	rOfTimes90Days	e60-89DaysPast	Score
1	0.076739446	435	20	4	14	34	23	530
2	0.027417396	435	-5	-3	14	34	23	498
3	0.015523241	435	-22	-7	14	34	23	477
4	0.073192089	435	-5	5	28	34	23	520
5	0.08639633	435	20	7	14	34	23	533
6	0.023695521	435	-5	-7	14	34	23	494
7	0.076383066	435	20	2	14	34	23	528
8	0.022138259	435	-22	-14	28	34	23	484
9	0.010910252	435	-23	-12	14	34	23	471
10	0.019593406	435	-23	5	14	34	23	488
11	0.018344733	435	-22	2	14	34	23	486
12	0.010910252	435	-23	-12	14	34	23	471
13	0.03572072	435	-5	7	14	34	23	508
14	0.06207429	435	20	-3	14	34	23	523
15	0.022034074	435	-22	7	14	34	23	491
16	0.042777829	435	20	-14	14	34	23	512
17	0.020351458	435	-22	5	14	34	23	489
18	0.186177157	435	20	7	14	34	33	543
19	0.076739446	435	20	4	14	34	23	530
20	0.019322574	435	-22	3	14	34	23	487
21	0.023695521	435	-5	-7	14	34	23	494
22	0.010910252	435	-23	-12	14	34	23	471
23	0.615888459	435	20	-12	28	47	33	551
24	0.069681822	435	-5	2	28	34	23	517
25	0.018344733	435	-22	2	14	34	23	486
26	0.076739446	435	20	3	14	34	23	529

在我们的模型中，1表示坏客户，0表示好客户，因此评分卡中评分（Score）愈高，说明是坏客户的概率愈大。我们给出评分的同时，计算了每位客户是坏客户的概率（Predict）。在左图中，红框部分的一位客户的预测概率高达0.62，则极有可能为坏客户，这就给予了银行决策的依据。

总结

本次建模完整实现了数据读入、数据分析、数据预处理（即数据清洗）、特征选择、模型建立、模型评估、创建信用评分卡、预测测试集等八个步骤，建立起了一个具有良好预测效果的信用评分模型。通过本次建模过程，我们可以深刻地感受到，身处一个数据爆炸的时代，对数据进行合理的筛选、划分、建模显得尤为重要。我们的工作由于时间原因也有不足之处，比如我们在模型的选择上比较保守，后期我们将继续尝试梯度上升法等方式进行更好地分类。

参考文献

[1] Stuart J. Russell,Peter Norvig:《Artificial Intelligence—A Modern Approach》
[2]基于python的信用卡评分模型，https://blog.csdn.net/gxhzoe/article/details/80428560

AIPKU