浙江工商大学

P2P借贷违约风险评价研究

——基于最大信息系数的模型构建及预测

CON ENE

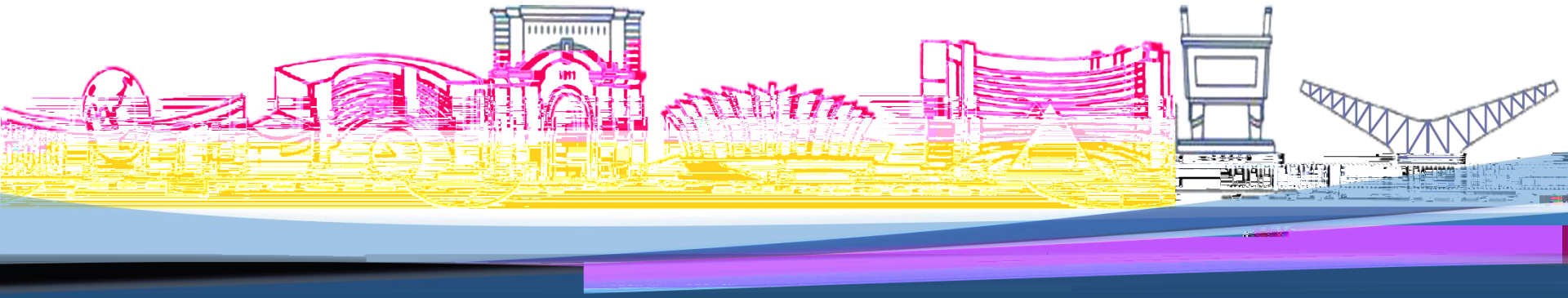
0

1

2

3

4



P2P网络借贷　　高违约风险

数据预处理　　最大

信息系数(MIC)　复杂网络模型

影响因素

LightGBM　　P2P借

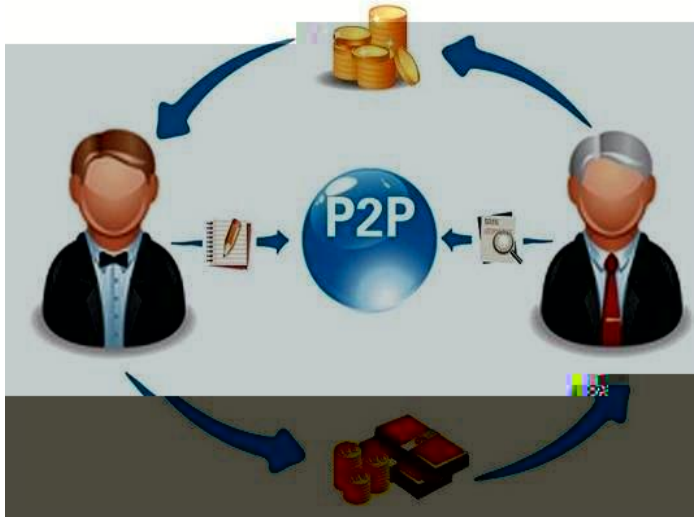
贷违约预测模型

最大信息系数　　复杂网络模型　　机器学习

MIC

P2P平台识别借贷违约风险提供有价值的参考。

1.



-
-
-
-

违约风险分析模型，以有效开

展P2P网络借贷的违约风险评估呢？

2.

2.1

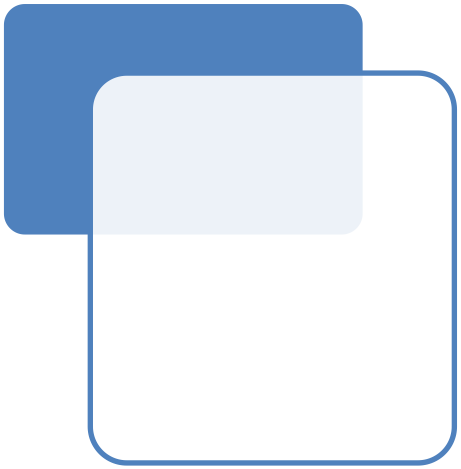
- 覆盖面广、申请数量多、交易规模大
- 高达109亿美元 上涨35%
- 公开透明 质量

2.

2.2

3.

3.1



3.1

3.1.1

（1）一致性检查

-

（2）冗余变量处理

-
-

3.1

3.1.1

（3）缺失值处理

-
-
-

（4）异常值处理

-
-
-

3.1.2

（1）维度归约

-

-

（2）数量归约

-

-

3.1.3

1

One-Hot Encoding(独热编码)：

-
-
-

Label Encoding(标签编码)：

-
-

2

- Label Encoding

3.

3.2 MIC 2 (IC)

普适性：

等价性：

3.2 MIC 2

3.2.1

1

- 不同形式
-
- 无向网络

2 IC

minepy

3.2 MIC 2

3.2.1

3

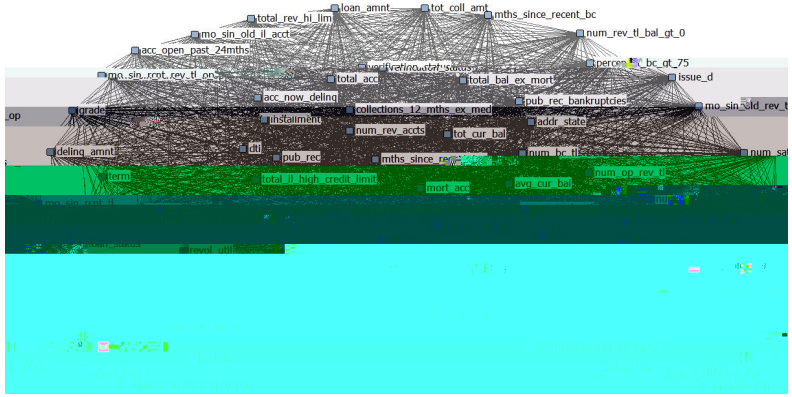
网络节点

65×65对称矩阵

设定阈值θ

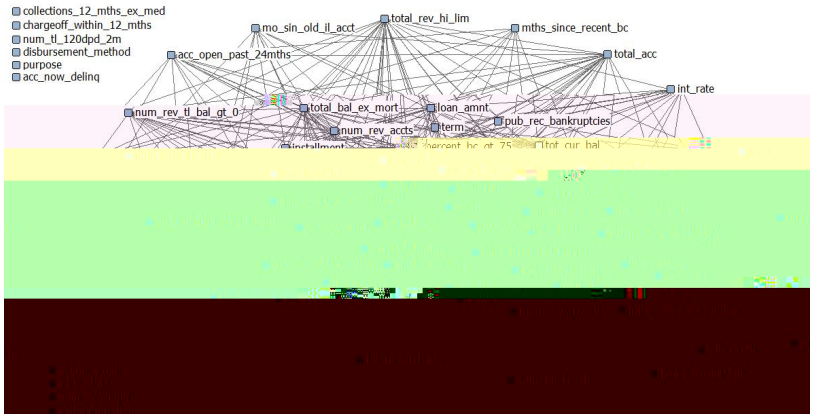
3.2 MIC

3.2.1

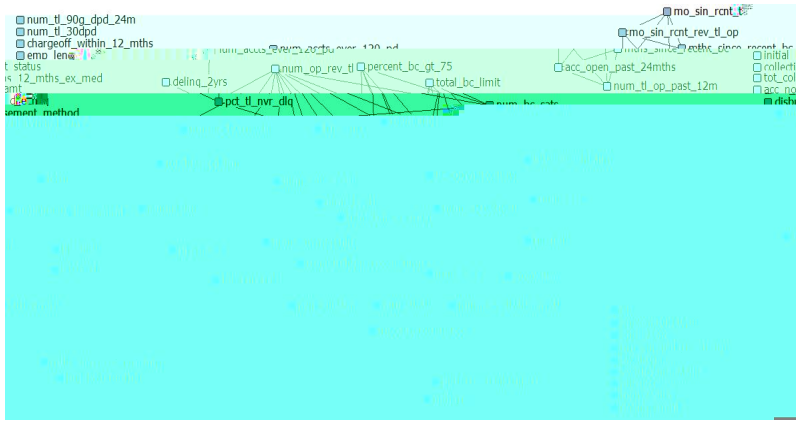


( ) $\theta = 0$

( ) $\theta = 0.013$



( ) $\theta = 0.05$



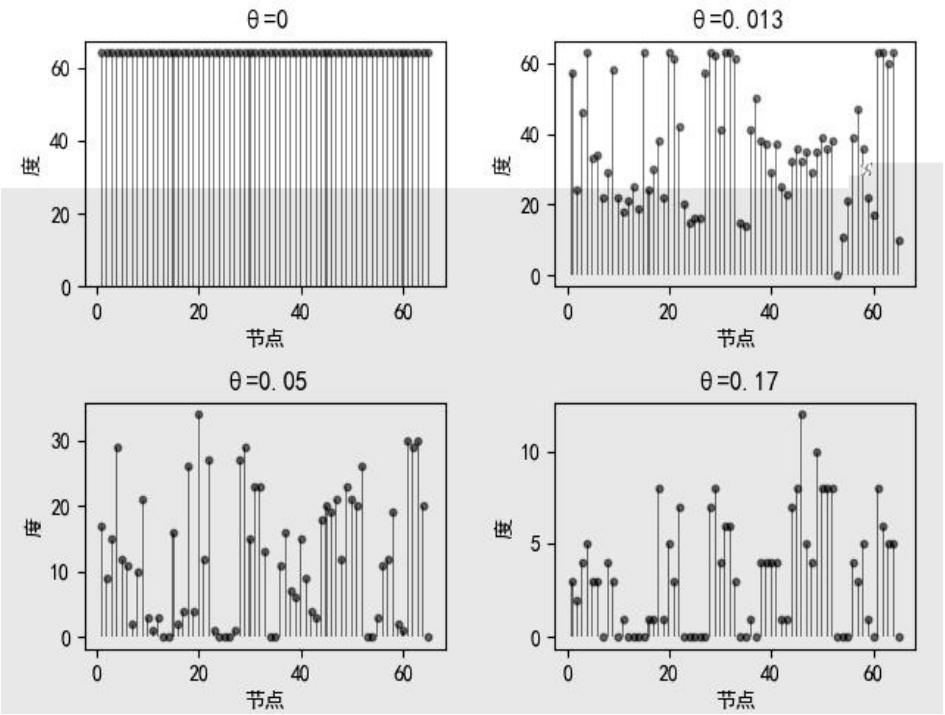
( ) $\theta = 0.17$

1 θ 2

3.2 MIC 2

3.2.2

1



- θ=0
- θ=0.013
- θ=0.05
- θ=0.17

2 θ

3.2.2

2

2 θ=0.013

k	0	1	2	3	4	5	6	7
$P(k)$	0.0154	0	0	0	0	0	0	0
k	8	9	10	11	12	13	14	15
$P(k)$	0	0	0.0154	0.0154	0	0	0.0154	0.0308

3.2 MIC

3.2.2

2

3 θ=0.05

k	0	1	2	3	4	5	6	7
$P(k)$	0.1538	0.0615	0.0462	0.0615	0.0462	0	0.0154	0.0154
k	8	9	10	11	12	13	14	15
$P(k)$	0	0.0308	0.0154	0.0462	0.0615	0		

3.2.2

2

4 θ=0.17

k	0	1	2	3	4	5	6	7
$P(k)$	0.2923	0.1231	0.0154	0.1077	0.1385	0.0923	0.0462	0.0462
k	8	9	10	11	12			
$P(k)$	0.1077	0	0.0154	0	0.0154			

结论： 随着阈值θ的增大，网络结构不断发生变化，网络中度较高的节点数量减少，度较低的节点数量增加。

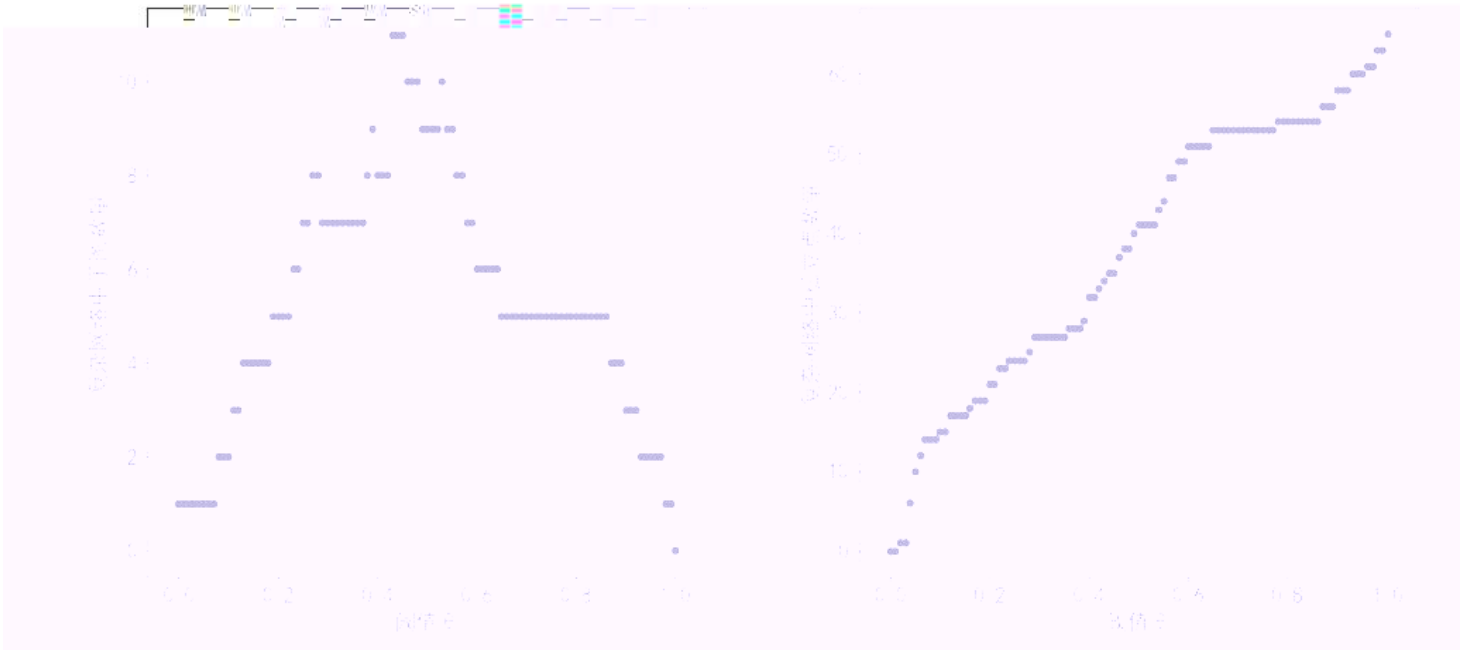
3.2 MIC 2

3.2.2

3

孤立点：

子图：



3 θ

3.2 MIC

3.2.2

4

借款者

信息

借款信息

借款者年收入(annual_inc)

借款者年收入是借款者信息相关因素和借款信息相关因素的桥梁

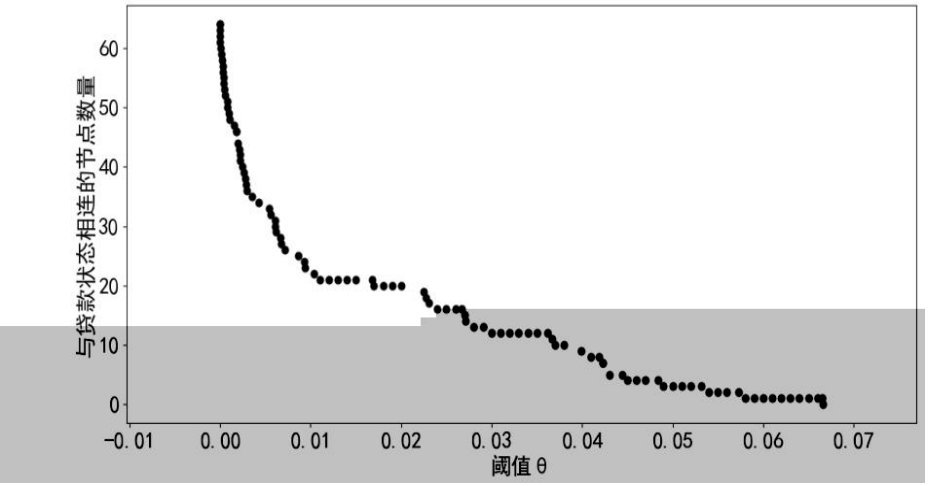
不同收入水平的借款者自身情况存在一定的差异

不同收入水平的借款者的借款需求也有所不同

3.2 MIC 2

3.2.2

5

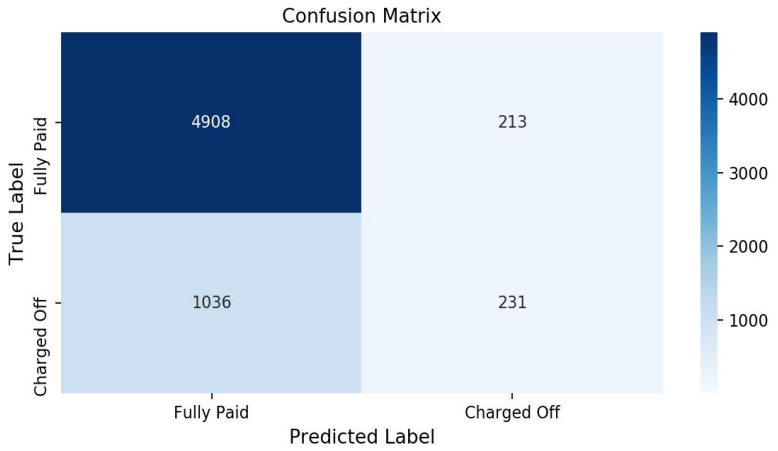


4 θ

3.3 MIC 2 L GBM

3.3.1

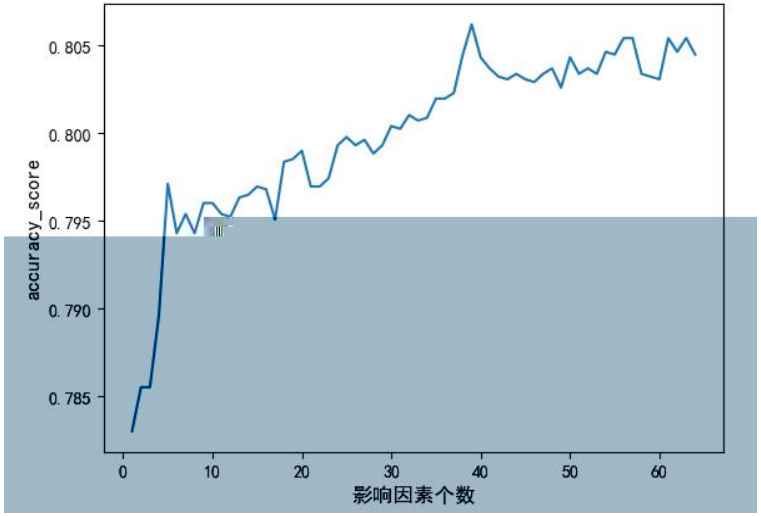
- 准确率
- 精确率
- 召回率
- F1-score
-
-



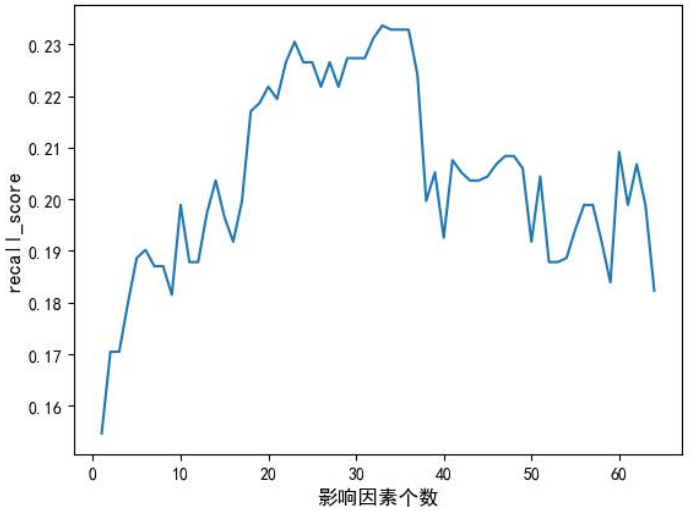
5 64

3.3 MIC 2 L GBM

3.3.1



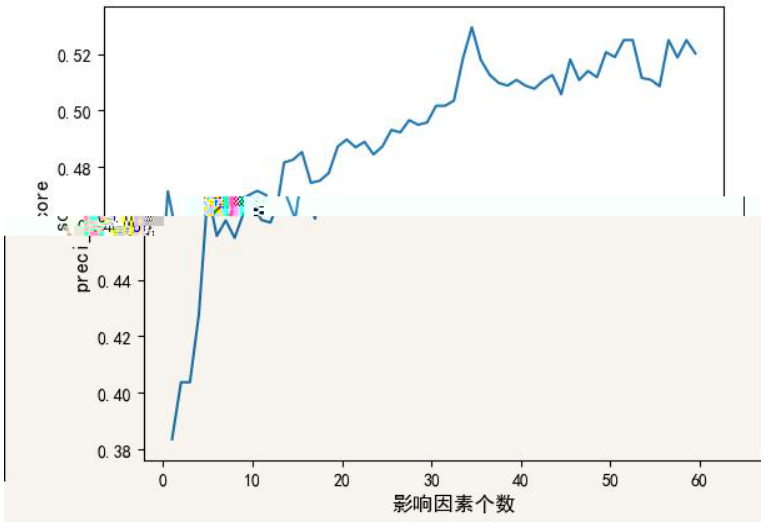
6



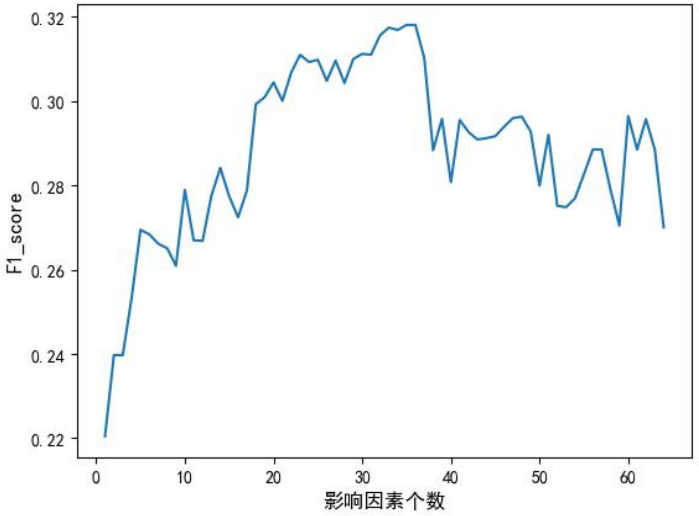
7

3.3 MIC 2 L GBM

3.3.1



8

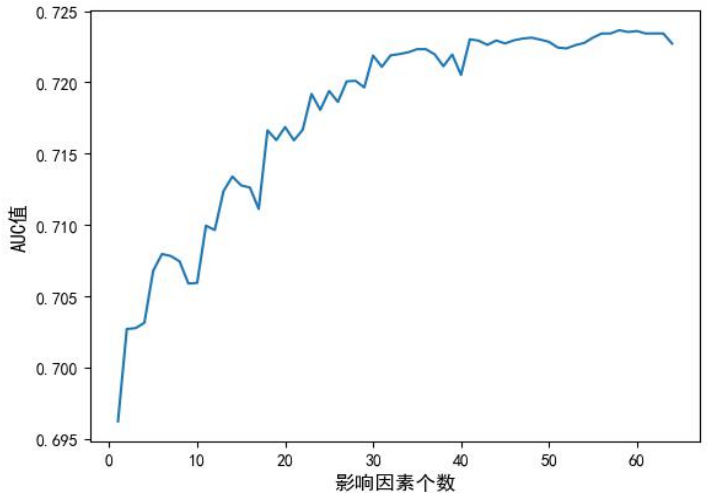


9

F1-

3.3 MIC 2 L GBM

3.3.2



-
-
-

10

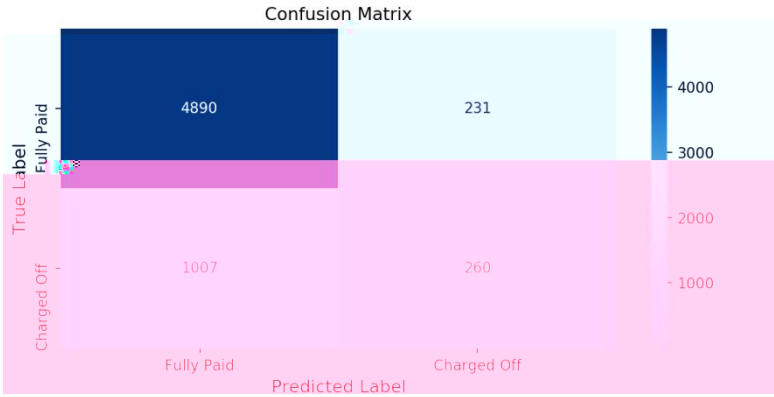
A C

研究目的：

选择影响因素数量为39的模型作为最终预测模型。

3.3 MIC 2 L GBM

3.3.2



11 39

4.

可利用最大信息系数(MIC)来识别

被解释变量的重要影响因素

4.

-
-

2

-
-

-
-

谢谢观看！