医疗保险欺诈识别检测

小羊肖恩队

周鸿溢：数据概况 数据预处理
张羽鹏：数据可视化 模型构建

目录

一、项目背景

医保欺诈造成巨大损失

中华人民共和国最高人民检察院
The Supreme People's Procuratorate of the People's Republic of China

首页 | 机构设置 | 检察新闻 | 工作信息 | 检察业务 | 检察院建设 | 12309中国检察网 | 中国检察听证网 | 2024年05月22日 星期三

当前位置： 首页 > 检察院建设 > 理论研究

借助数字化手段监督整治医保诈骗

时间：2022-11-08 作者：张震宇 吴菊萍 来源：检察日报 【字体：大 中 小】

用科技手段识别医疗欺诈

●在数字化背景下，检察机关不能仅满足于个案的办理，而是应当以数字为牵引，通过数据建模、融合审查、类案监督的方式从个案办理中发现类案监督线索，从而精准打击医保诈骗犯罪，以检察监督促医保基金监管质效提升，为守护医保基金贡献检察力量。

●强化行刑衔接工作，与医保局、人社局开展数据分析技术协作，充分运用医保诈骗模型，定期为医保局提供数据分析支持，帮助医保局智能核查异常数据，并与医保局建立健全案件线索的移送制度，加强线索移送、接收衔接，完善案件处理信息通报机制，形成行刑整体共治，提升医保基金监管质效。

新闻频道 > 国际新闻

美国78人因涉嫌欺诈25亿美元的医疗保险遭起诉

来源：央视新闻 | 2023年06月29日 09:26:30

据路透社当地时间28日报道，美国司法部称，美联邦及地方政府针对一起涉及25亿美元的医疗保险欺诈计划，对来自16个州的78名被告提出刑事指控。据悉，该案件涉及的对象包括老年人、残疾人、艾滋病患者及孕妇。

据悉，面临指控的人员包括24名医生、护士和其他有执照的医疗专业人员及医疗保健高管，其中包括一家耐用医疗设备在线平台的首席执行官，该平台被指控欺诈索赔19亿美元。

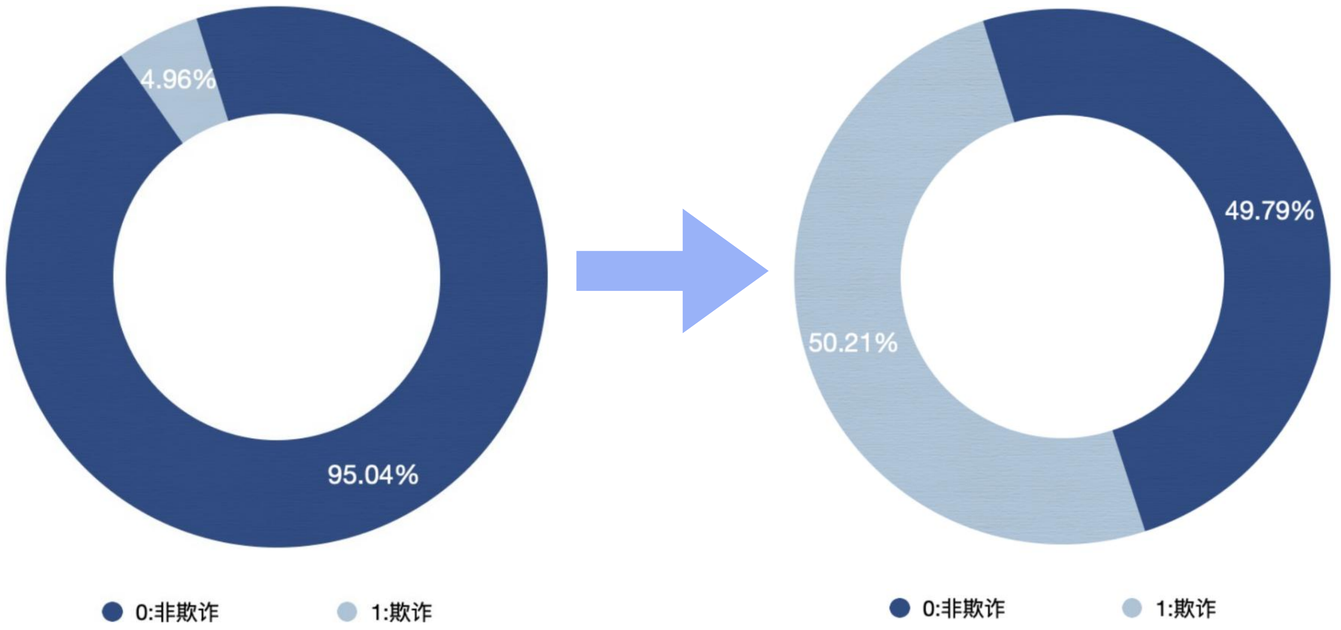
据官方人士称，该起医疗欺诈事件针对美国联邦医疗保险（Medicare）、州级医疗保险以及私人保险公司提供的补充医保项目，总金额达25亿美元，实际支付的金额约11亿美元。（总台记者 张颖哲）

二、数据探索

特征名	类型
个人编码	float
一天去两家医院的天数	int
就诊的月数	int
月就诊天数_MAX	int
月就诊天数_AVG	float
月就诊医院数_MAX	int
月就诊医院数_AVG	float
.....	.....
个人支付治疗费用占比	float
BZ_民政救助	int
BZ_城乡优抚	int
是否挂号	int
RES	int

数据总共有81个特征1个标签，16000条数据。

目标变量分析



非欺诈（0）：15207
欺诈（1）：793

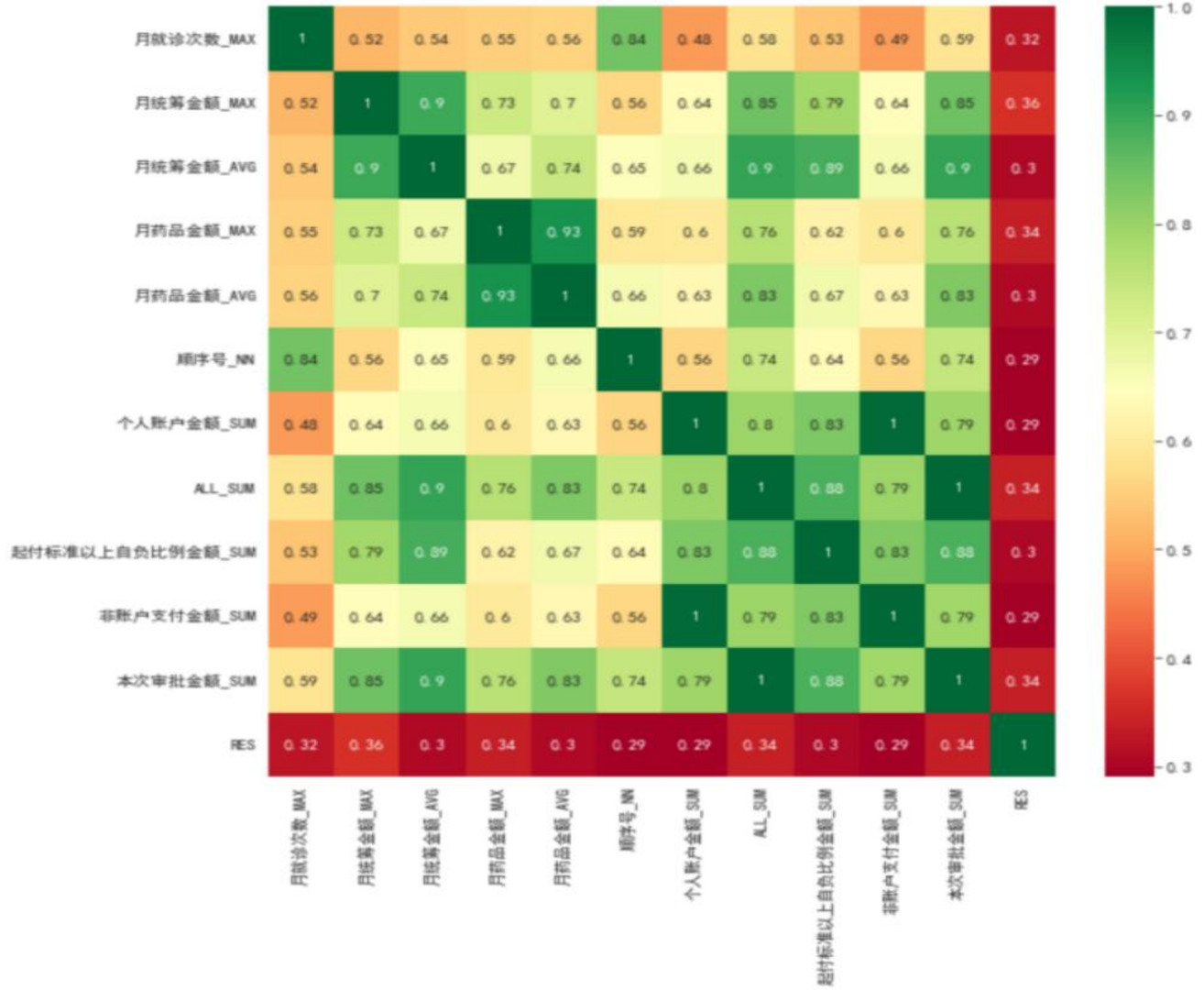
非欺诈（0）：15180
欺诈（1）：15180

二、数据探索-重要特征展示

特征名	类型
月就诊次数_MAX	int
月统筹金额_MAX	float
月统筹金额_AVG	float
月药品金额_MAX	float
月药品金额_AVG	float
起付标准以上自负比例金额_SUM	float
非账户支付金额_SUM	float
本次审批金额_SUM	float
顺序号_NN	int

初始化SelectKBest 选择器，计算特征与目标变量之间的相关性，选择排名前10的特征

特征相关性热力图



三、数据预处理

查看缺失值个数

	Total	Percent
出院诊断LENTH_MAX	355	0.022
个人编码	0	0.000
基本统筹基金支付金额_SUM	0	0.000
最高限额以上金额_SUM	0	0.000
医疗救助个人按比例负担金额_SUM	0	0.000

数据预处理

均值填充缺失值

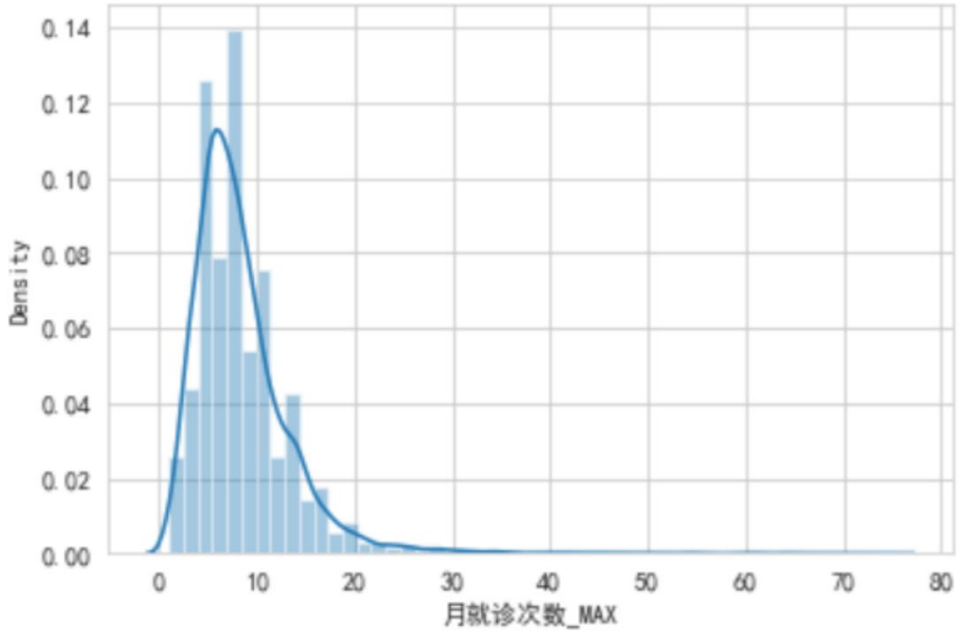
	Total	Percent
一天去两家医院的天数	0	0.000
治疗费申报金额_SUM	0	0.000
公务员医疗补助基金支付金额_SUM	0	0.000
基本统筹基金支付金额_SUM	0	0.000
最高限额以上金额_SUM	0	0.000

```
Length: 81, dtype: int64
一天去两家医院的天数        int64
就诊的月数               int64
月就诊天数_MAX           int64
月就诊天数_AVG         float64
月就诊医院数_MAX         int64
                       ...
个人支付治疗费用占比       float64
BZ_民政救助              int64
BZ_城乡优抚              int64
是否挂号                 int64
RES                   int64
```

数据类型转换

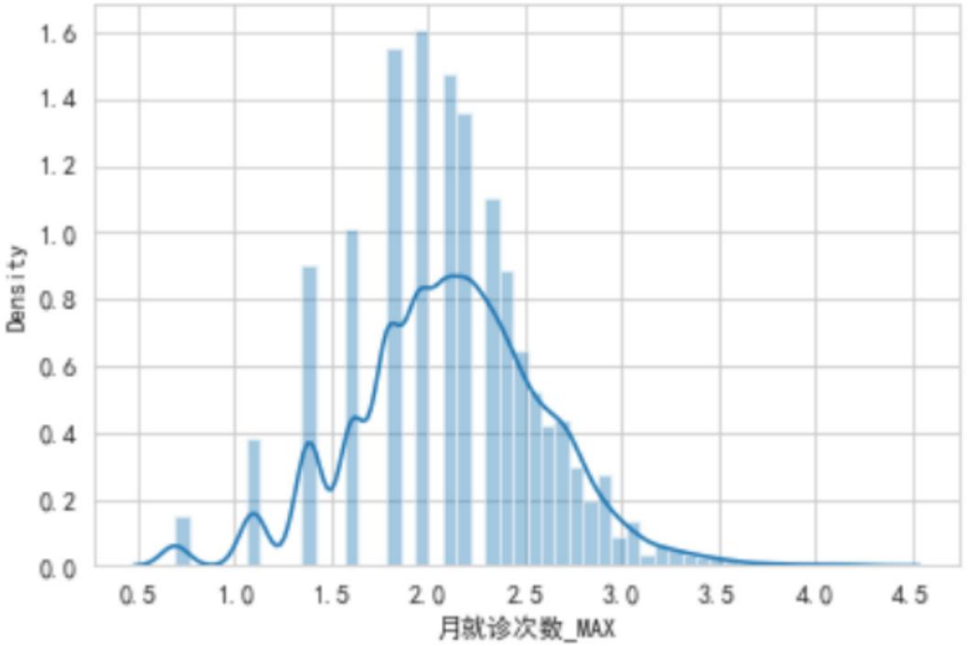
```
一天去两家医院的天数      float64
就诊的月数             float64
月就诊天数_MAX         float64
月就诊天数_AVG         float64
月就诊医院数_MAX       float64
                       ...
个人支付治疗费用占比      float64
BZ_民政救助            float64
BZ_城乡优抚            float64
是否挂号               float64
RES                  float64
Length: 81, dtype: object
```

三、数据预处理



峰度：16.866684
偏度：2.760366

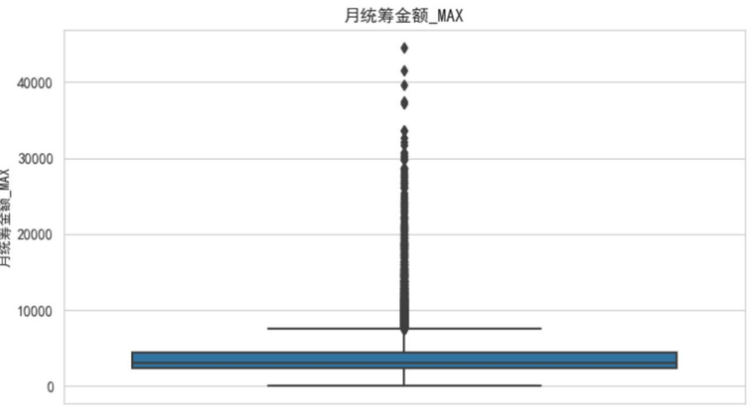
特征数据对数转换

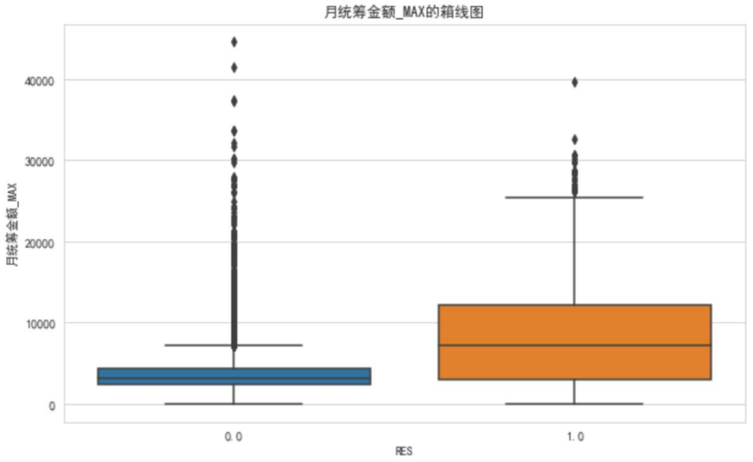


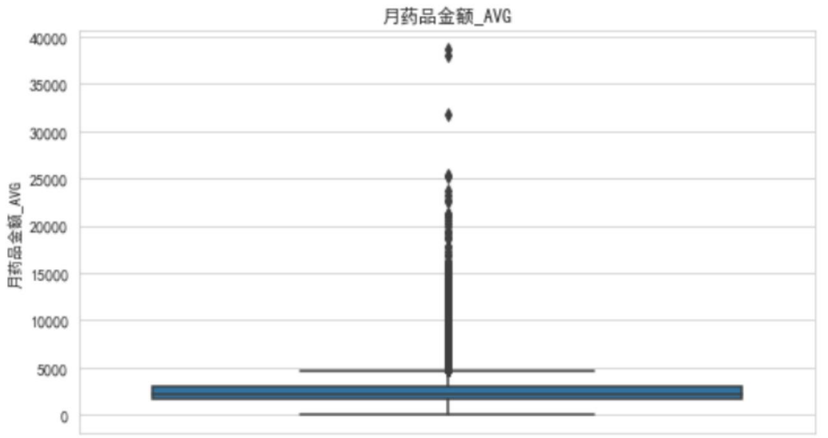
峰度：0.557231
偏度：0.001150

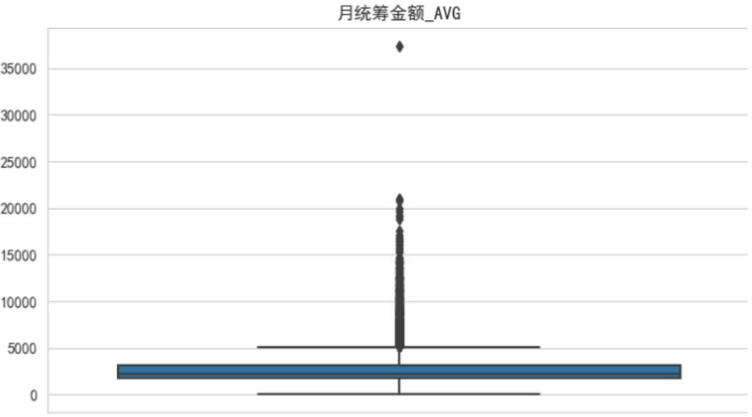
四、数据可视化

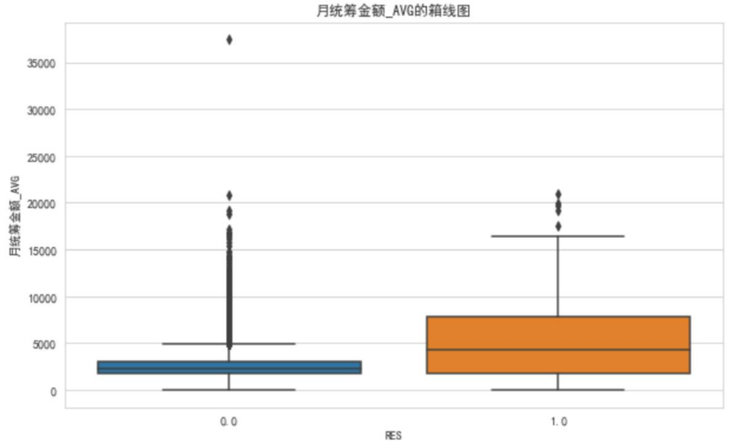
异常值检测及数据分布摘要

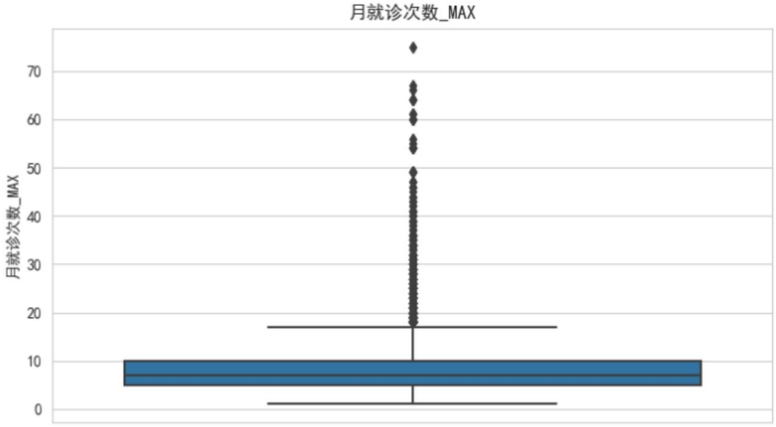












五、模型构建-决策树模型

决策树构建树形结构

Select Algorithm

Decision Tree

You Selected Decision Tree Algorithm

Select the size of Test Dataset (test train split)

0.20

0.00 1.00

Start Training

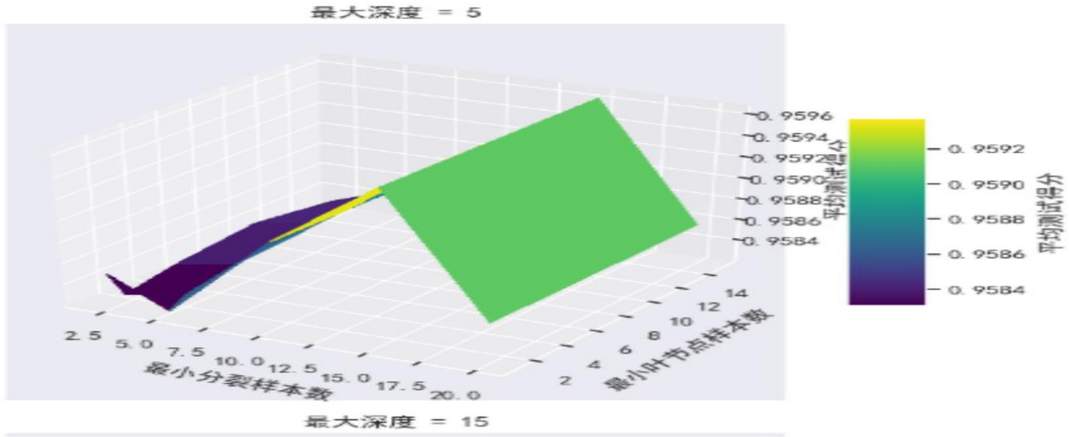
Model accuracy of Decision Tree: 94.1834451901566%

Confusion Matrix:

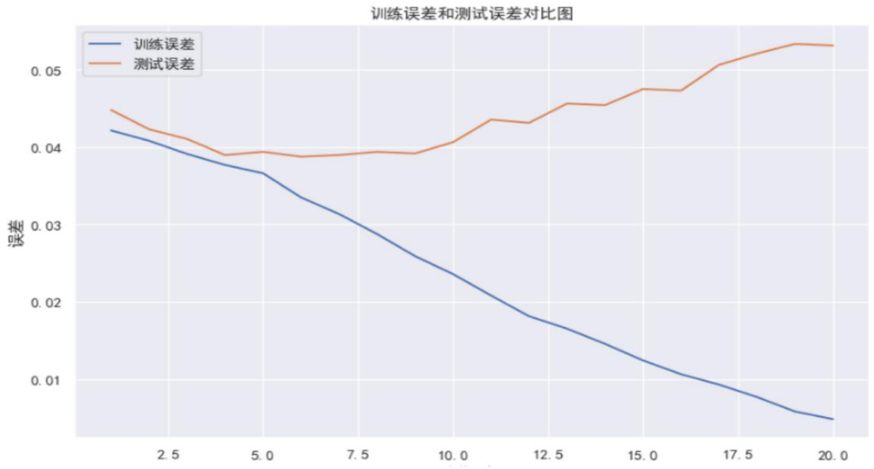
0	1
2,885	95
87	62

	Parameter	Value
0	Precision Score	0.9433
1	Recall Score	0.9418
2	F1 Score	0.9426

三维热力图对决策树模型进行解释



训练误差与测试误差对比图



五、模型构建-支持向量机模型

Select Algorithm

Support Vector Machine

You Selected Support Vector Machine Algorithm

Select the size of Test Dataset (test train split)

0.20

0.00 1.00

Start Training

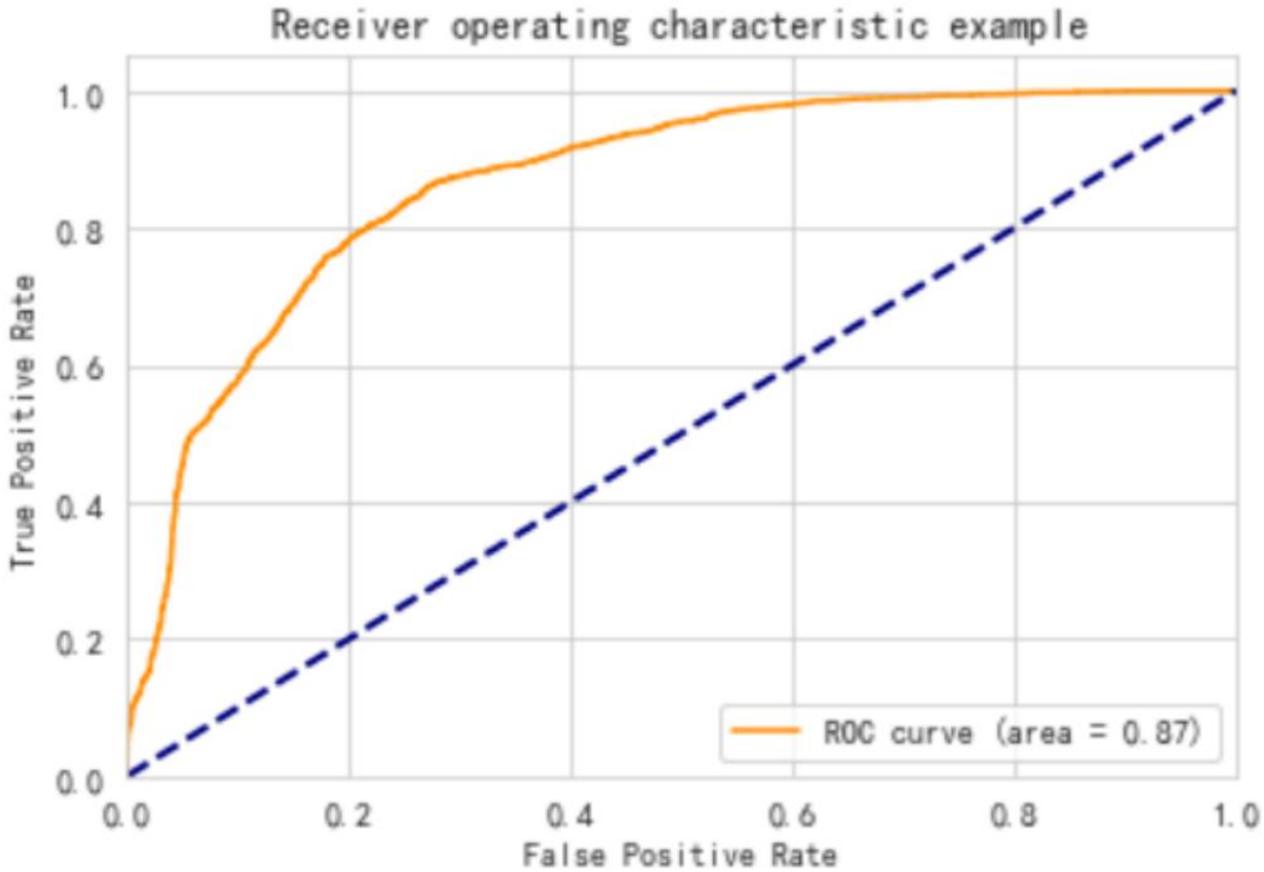
Model accuracy of Support Vector Machine: 96.132949824225%

Confusion Matrix:

0	1
2,964	16
105	44

	Parameter	Value
0	Precision Score	0.9547
1	Recall Score	0.9613
2	F1 Score	0.9534

ROC曲线对支持向量机模型进行解释



五、模型构建-随机森林模型

Select Algorithm

Random Forest

You Selected Random Forest Algorithm

Select the size of Test Dataset (test train split)

0.20

0.00 1.00

Start Training

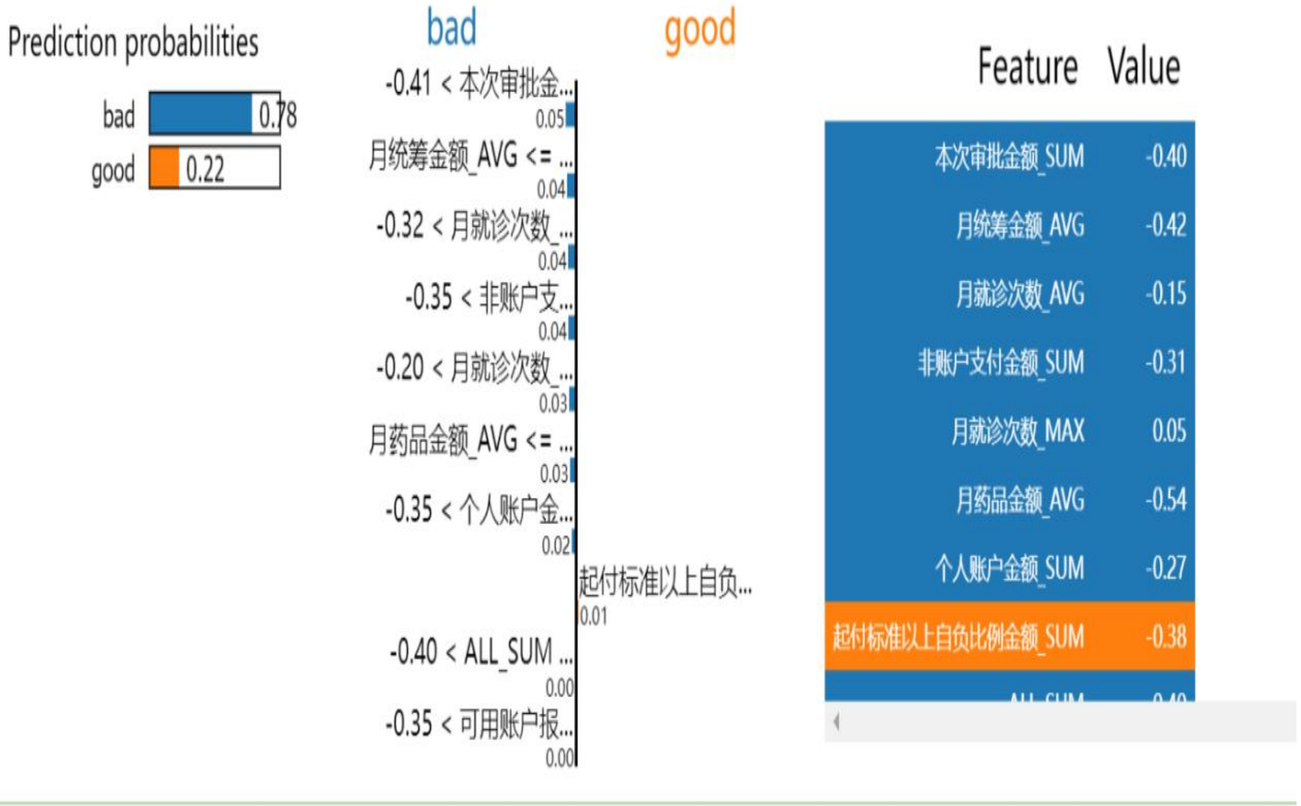
Model accuracy of Random Forest: 96.48449984020453%

Confusion Matrix:

0	1
2,966	14
96	53

	Parameter	Value
0	Precision Score	0.9602
1	Recall Score	0.9648
2	F1 Score	0.9584

LIME对随机森林模型进行解释



五、模型构建-特征交互

就诊平均账户成本 = 个人账户金额_SUM / 就诊次数_SUM

日均药品费用 = 月药品金额_AVG / 月就诊天数AVG

药疗联合发生额 = 药品费发生金额_SUM + 治疗费发生金额_SUM

就医_统筹交互指数 = 医院_统筹金_AVG / 月就诊医院数_MAX

时段强化药品费 = 交易时间YYYYMM_NN * 药品费发生金额_SUM

	月就诊次数_MAX	月统筹金额_MAX	月统筹金额_AVG	月药品金额_MAX	顺序号_NN	ALL_SUM	起付标准以上自负比例金额_SUM	非账户支付金额_SUM	本次审批金额_SUM	药疗联合发生额
0	7.000	3501.180	2541.293	3901.450	69.000	17218.750	1694.280	1742.190	16942.040	16299.750
1	4.000	2217.660	1637.358	2449.130	64.000	11195.720	1091.600	1172.070	10915.750	10848.920
2	9.000	3360.550	2583.053	3302.060	102.000	18135.520	2341.710	2403.200	17840.030	17749.170
3	6.000	3030.610	2057.720	1500.120	56.000	13719.380	1371.900	1299.720	13718.220	13719.380
4	5.000	2332.450	2196.315	2563.260	64.000	14747.540	1464.240	1505.650	14642.130	14747.540

```
count     16000.000
mean      17661.608
std       13314.162
min           3.000
25%       11055.950
50%       13999.460
75%       19782.033
max      239001.340
Name: 药疗联合发生额, dtype: float64
```

五、模型构建-模型预测

对于原始特征和交互后的特征，分别使用逻辑回归模型，Adaboost模型，决策树，随机森林等计算其准确率。

模型	未交互特征模型准确率	交互特征模型准确率
决策树模型	94.18%	93.51%
支持向量机模型	96.13%	96.21%
随机森林模型	96.48%	96.73%

可以看出特征交互后，大多数模型的准确率都有提升。

感谢聆听