浙江工商大學

《机器学习》课程论文

题目：关于某金融产品认购情况的预测

学　院：统计与数学学院

专　业：应用统计 2202

学　号：2020040176

学生姓名：朱 晨

二〇二三 年 四 月

关于某金融产品认购情况的预测

摘 要

本文利用某金融产品认购情况数据集，其中的各字段描述主要包括客户的基本信息数据、与当前活动的最后一次联系有关的数据以及与社会和经济背景有关的一些指标。首先对数据进行基础的描述统计和数据分析，发现客户年龄、职业差距、婚姻状况、受教育程度及违约、贷款等因素均对最终购买情况产生影响，再分别采用 KNN 模型、决策树算法、Logistic 回归、支持向量机算法和随机森林模型，进行不同模型的测试与比较。最后根据各个模型对测试集的预测效果，综合比较各个模型的 precision、recall、f1-score、accuracy 以及 ROC 曲线，根据各项得分，发现随机森林模型（Random Forest）具有准确率高、运行高效的特点，即使不用降维也可以处理高维特征，最终决定采用随机森林模型进行认购情况预测。

关键词：随机森林模型；描述统计；模型训练；混淆矩阵

目 录

一、引 言

机器学习不同模型的特点及适应条件各不相同，通常需要根据具体的研究问题，选择合适的算法模型。分析一个与某金融机构的营销活动有关的数据集，这些营销活动是基于电话的。通常情况下，金融机构需要对同一客户进行一次以上的联系，以便了解该金融产品是否会被客户所认购。数据集中的各字段描述主要包括客户的基本信息数据、与当前活动的最后一次联系有关的数据以及与社会和经济背景有关的一些指标，输出变量则为客户是否认购该金融产品。根据所提供的数据，本文使用python利用不同模型进行训练，根据模型的准确度及其它数据，选择其中最优的机器学习模型，对另一个数据集中的人群是否认购该金融产品进行预测。

1.1 解决思路

围绕如何准确预测金融产品认购情况，本次研究主要分为三个部分：

（1）对原始数据集做可视化处理，并进行基础的描述统计和数据分析，以便于掌握数据的分布特征，了解数据的变量构成情况，进而为后续建模分析提供参考。

（2）将数据划分为训练集和测试集，并分别采用KNN模型、决策树算法、Logistic回归、支持向量机算法和随机森林模型，利用训练集和测试集的数据，进行模型的测试与比较。综合比较各个模型的precision、recall、f1-score、accuracy以及ROC曲线，根据各项得分，评价不同模型的预测效果。

（3）根据上文所进行的五种模型算法的测试结果，综合考虑各个评价指标得分，最终决定采用随机森林模型进行本次金融产品认购情况的预测工作。

二、数据的处理分析

2.1 数据处理

（1）首先在python中导入数据集，并进行可视化处理，结果如下图1所示，本数据集共有26645行，20列。

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	campaign	passed_days	previous	pre_outcome	emp_rate	cpi	cci	r3m	employed	y
0	27	admin.	single	university.degree	no	no	no	cellular	mar	tue	1	999	0	nonexistent	-1.8	93.369	-34.8	0.637	5008.7	yes
1	55	unemployed	married	basic.9y	no	no	yes	cellular	may	mon	1	999	0	nonexistent	-1.8	92.893	-46.2	1.264	5099.1	yes
2	25	blue-collar	single	basic.9y	no	yes	no	cellular	may	fri	3	999	1	failure	-1.8	92.893	-46.2	1.250	5099.1	yes
3	43	admin.	married	university.degree	unknown	yes	no	cellular	aug	tue	1	999	0	nonexistent	1.4	93.444	-36.1	4.968	5228.1	yes
4	33	admin.	married	high.school	no	no	no	telephone	jul	tue	1	999	0	nonexistent	-2.9	92.469	-33.6	1.044	5076.2	yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26640	41	admin.	married	high.school	no	yes	yes	telephone	jun	thu	4	999	0	nonexistent	1.4	94.465	-41.8	4.866	5228.1	no
26641	35	blue-collar	single	basic.6y	no	no	no	cellular	nov	thu	1	999	1	failure	-0.1	93.200	-42.0	4.076	5195.8	no
26642	53	housemaid	married	basic.4y	unknown	yes	no	telephone	aug	thu	7	999	0	nonexistent	1.4	93.444	-36.1	4.962	5228.1	no
26643	35	services	married	high.school	unknown	yes	no	cellular	apr	wed	2	999	0	nonexistent	-1.8	93.075	-47.1	1.445	5099.1	no
26644	48	blue-collar	married	basic.9y	no	yes	no	cellular	may	tue	2	999	0	nonexistent	-1.8	92.893	-46.2	1.344	5099.1	no

26645 rows × 20 columns

图 1 数据集导入

接下来对数据进行基本处理，检查数据是否存在空值，经过检查发现，本数据集不存在空值。并将输出变量y改为0，1变量，以便于后续的操作分析。

（2）对数值型变量进行描述统计，分析包含变量的个数、均值、标准差、最小值、最大值等，结果如下。

	age	campaign	passed_days	previous	emp_rate	cpi	cci	r3m	employed
count	26645.000000	26645.000000	26645.000000	26645.000000	26645.000000	26645.000000	26645.000000	26645.000000	26645.000000
mean	39.982361	2.559617	962.848039	0.169600	0.083231	93.574912	-40.482222	3.624934	5167.167198
std	10.409996	2.774588	185.999815	0.488594	1.573705	0.579165	4.633930	1.733887	72.212853
min	17.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.000000	43.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

图 2 描述性统计分析

2.2 数据分析

（1）查看数据中输出变量y的分布情况

由图可知，变量y的分布存在明显差异，y为0的个数超过23000，而y为1的个数仅为3000左右。即数据分布不平衡，后续需要采用过采样方法平衡数据集。

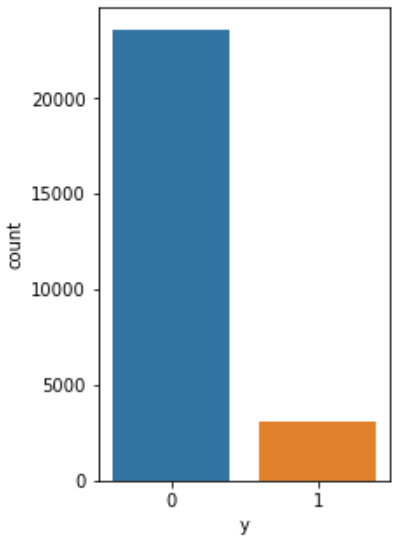

图 3 是否认购分布

（2）绘制是否购买金融产品与客户年龄的分布图

如下图所示，两者均表现为轻微右偏的近似正态分布，且分布特点差别不大，峰值均出现在30-40岁之间，可初步认为客户年龄对于是否购买该金融产品的影响较小。

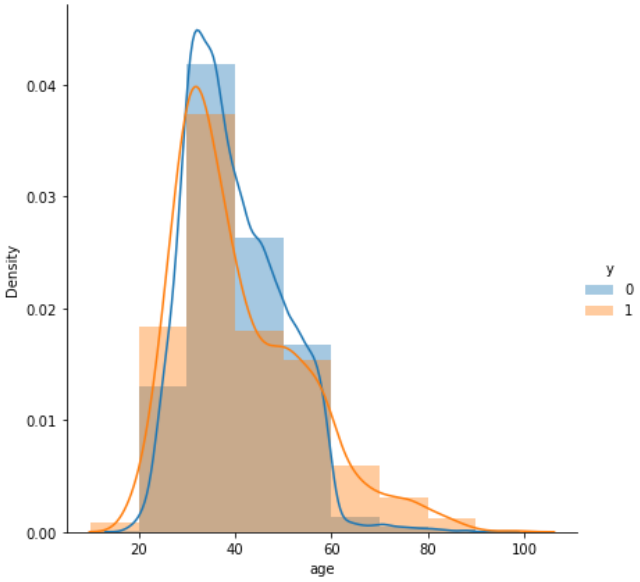

图 4 客户年龄分布

（3）查看不同职业的人群对于该金融产品的购买意向

根据下图所示结果，容易看出学生student和退休人群retired最倾向于购买，蓝领阶层对其购买意向最低，其余人群对于购买该产品的意向相差不大。初步分析可能是由于学生具备较高的知识水平，了解一定的理财知识，退休人群则由于面临的经济压力较小，且拥有较多精力来研究金融产品，因此两者的购买意向高于其他职业人群；蓝领是指那些从事体力劳动的工人，其人群特点为具备较强的体力劳动能力，接受的文化教育较为缺乏，尤其是金融理财方面的知识掌握不多，在这方面相对保守，且闲置的可支配资金不多，因此购买比例最低。

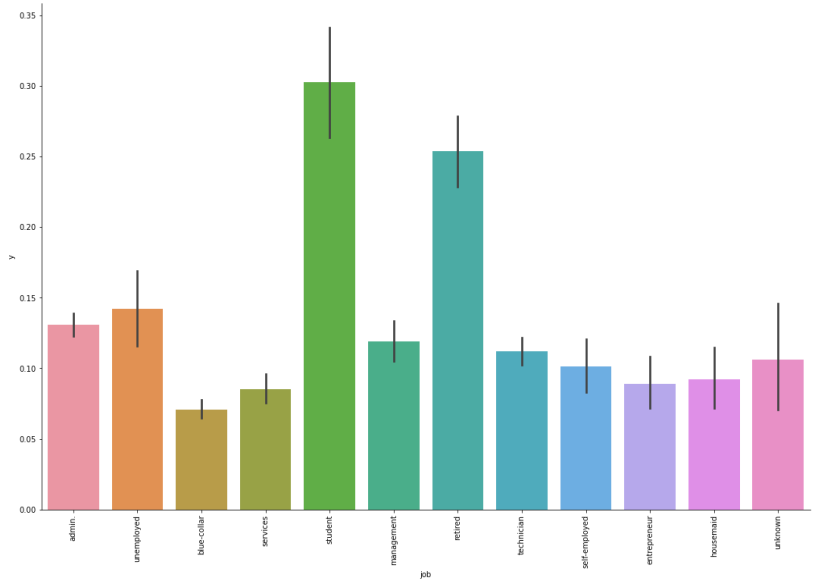

图 5 购买意向的职业分布

查看不同职业人群的购买人数，管理员admin、蓝领blue-collar以及技术人员technician购买的人数最多，与上图购买意向进行比较，可以发现购买人数多的人群与购买意向高的人群并不一致，由此说明该数据中各个职业的人群分布并不均衡，购买意向高不代表购买人数多。

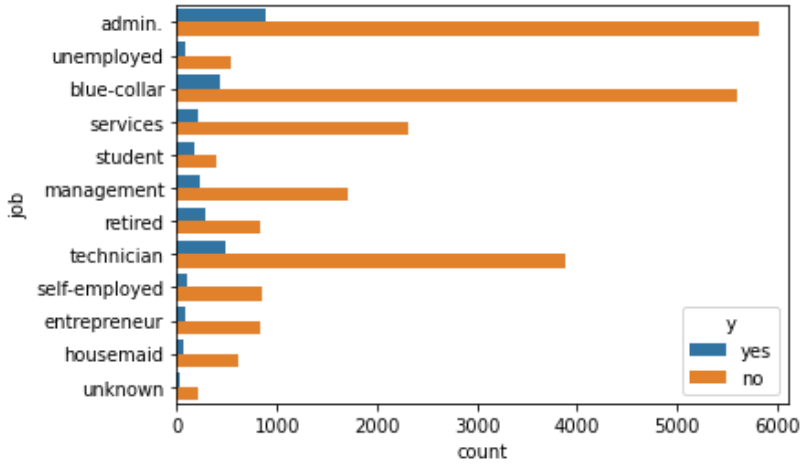

图 6 职业分布

（4）婚姻状况对购买意向的影响

排除婚姻状况未知的人群，可以发现单身人士对于产品的购买意向略高一些，已婚及离异人群购买意向则相对较低。初步分析可以认为是由于单身人群自身面临的经济压力较小，因此拥有较多的可支配资金用来购买金融产品，而已婚及离异人群由于需要面对家庭方面的一些经济压力，可能缺乏资金用来购买金融产品进行理财。

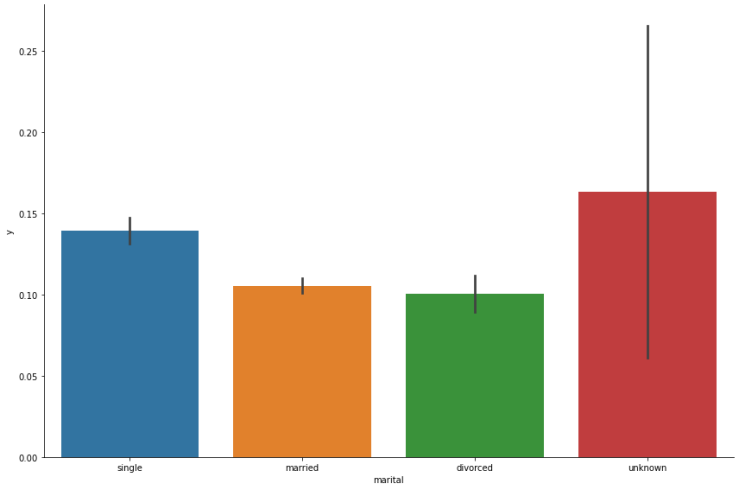

图 7 婚姻状况分布

（5）教育水平对购买意向的影响

图中信息表明，文盲和大学文凭的人群购买金融产品的比例较高，这两种学历处于两种极端，要么极高要么极低，而处于中等学历的人群购买的比例较低。主要原因是由于受过高等教育的人群具备一定的金融知识，能够帮助自己进行正确的决策，而文盲则是因为缺乏相关知识，容易受到金融产品销售的蛊惑，因此购买比例也处于较高水平，而中等学历人群则相对保守，害怕金融产品的风险，因此购买意愿不强。

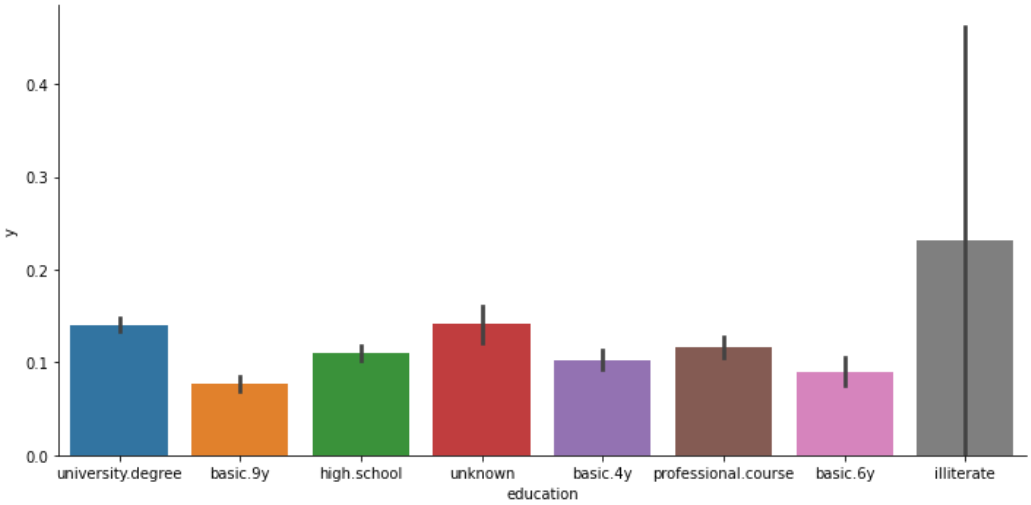

图 8 受教育程度分布

下图为购买金融产品人群在教育及职业的分布情况，由条形图反映的信息，排除未知状况。从横向看，管理员admin人群中大学文凭和职业教育人群的购买比例相对较高；失业人群unemployed中接受过4年基础教育的人群购买意向较高；蓝领blue-collar中高中文凭的人群购买金融产品比例最高；学生人群中，各类教育经历的人群购买金融产品比例显著高于其他职业人群，与上文学生人群购买意愿最高的结论相一致，其中接受6年基础教育的人群购买意向最强烈；对于退休人群，则表现为文盲的购买比例最高……

从图表的纵向信息可知，拥有大学文凭的购买人群中，学生和退休人群的占比最高，购买意向最强烈；接受9年基础教育的人群则是学生占比最高，其余大部分教育阶层人群，均呈现出学生购买意愿最强的现象；只有文盲人群中，企业家、退休人员、自主创业self-employed的占比显著，购买意向最为强烈。

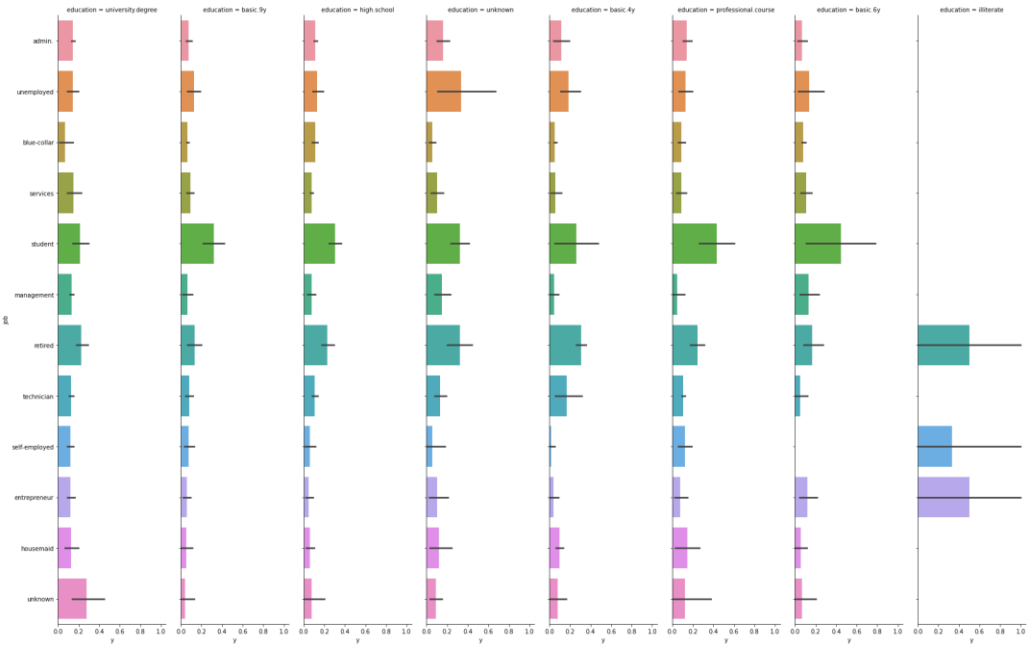

图 9 购买意向的受教育程度分布

（6）查看违约行为、住房贷款以及个人贷款对购买金融产品的影响

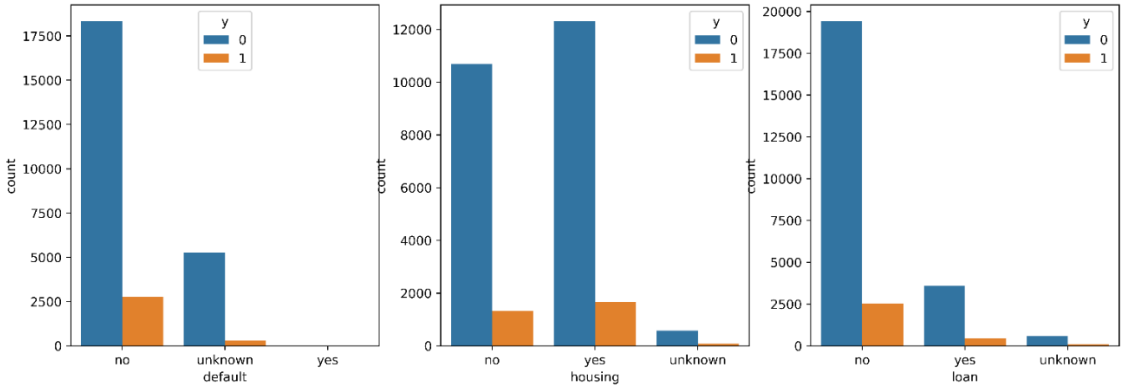

图 10 违约、贷款分布

违约行为、住房贷款、个人贷款对于购买金融产品的影响如图所示，上图为原始情况，由于包含未知项的干扰，因此去除未知项，进一步绘制如下所示条形图。由条形图可知，只有不存在违约行为的客户，才会选择购买该类金融产品；对于是否拥有住房贷款和个人贷款的居民，其购买意愿则相差不大，即住房贷款及个人贷款对于购买该金融产品的影响不大。

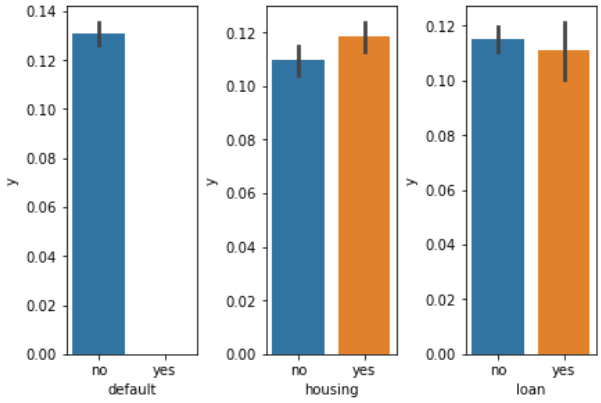

图 11 违约、贷款分布优化

（7）其他因素的影响

以下图表阐述其余变量对于是否购买金融产品的影响，由于各类影响错综复杂，故在此不再赘述，具体指标影响如下图所示。

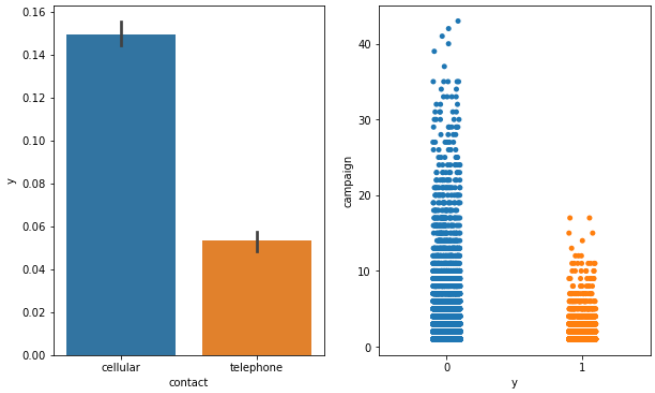

图 12 沟通方式对购买意向的影响

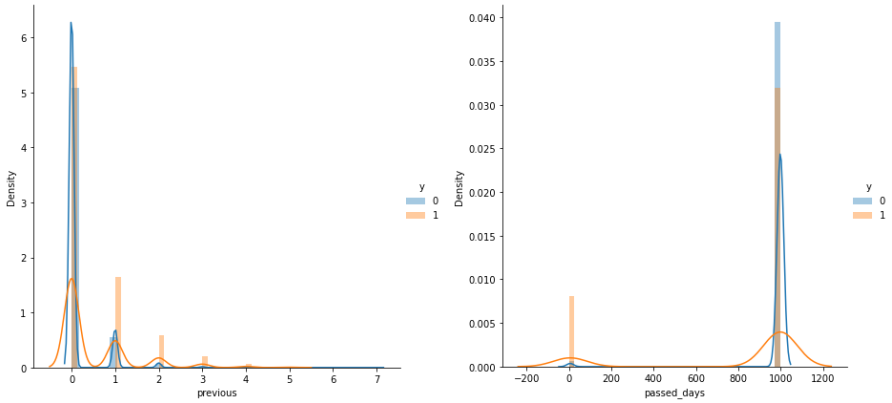

图13 在此之前联系客户的次数以及上次联系客户后的天数对购买意愿的影响

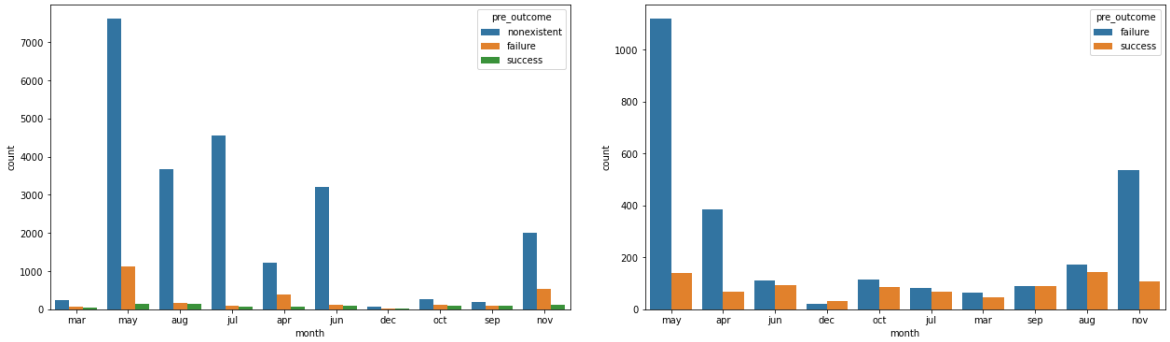

图14 前期营销活动以及去除干扰项的结果

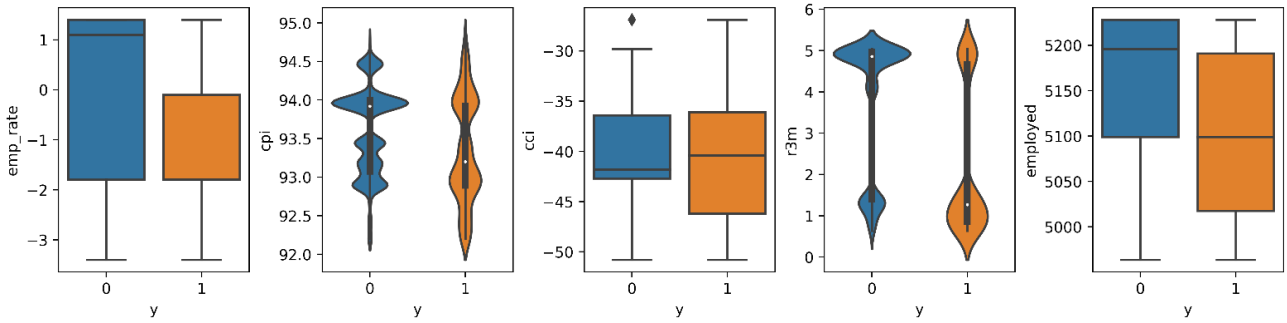

图15 其他社会经济相关指标与购买意愿的关系

三、模型的训练测试

进一步进行数据处理，对所给数据进行编码以及平衡数据集的操作，并对数值型变量进行标准化处理，进一步划分训练集和测试集，为后续采用不同模型进行建模求解工作做好前期准备。

接下来利用不同模型对训练集进行建模求解，并利用测试集中的数据进行效果评估。具体模型形式如下：

3.1 KNN模型

3.1.1 KNN模型训练

如图1-1所示，为KNN模型的输出模型评估结果，该模型准确率为0.8698262466450064。

	precision	recall	f1-score	support
0	0.98	0.76	0.85	7045
1	0.80	0.98	0.88	7113
accuracy			0.87	14158
macro avg	0.89	0.87	0.87	14158
weighted avg	0.89	0.87	0.87	14158

图 1-1 KNN模型评估结果

3.1.2 网格搜索最佳参数

如图1-2为网格搜索选取的最佳参数结果，根据输出结果可知，最佳参数的准确率为0.8698262466450064。

	precision	recall	f1-score	support
0	0.98	0.76	0.85	7045
1	0.80	0.98	0.88	7113
accuracy			0.87	14158
macro avg	0.89	0.87	0.87	14158
weighted avg	0.89	0.87	0.87	14158

图 1-2 网格搜索最佳参数结果

3.1.3 五折交叉验证

对该KNN模型进行五折交叉验证，验证结果如下：

```
array([0.8592402 , 0.84849402, 0.86120781, 0.85802936, 0.85740236])
```

图 1-3 五折交叉验证结果

3.1.4 KNN模型的ROC曲线

图1-4为本次KNN模型的ROC曲线，如图所示，该模型的AUC=0.874112，说明模型效果较好。

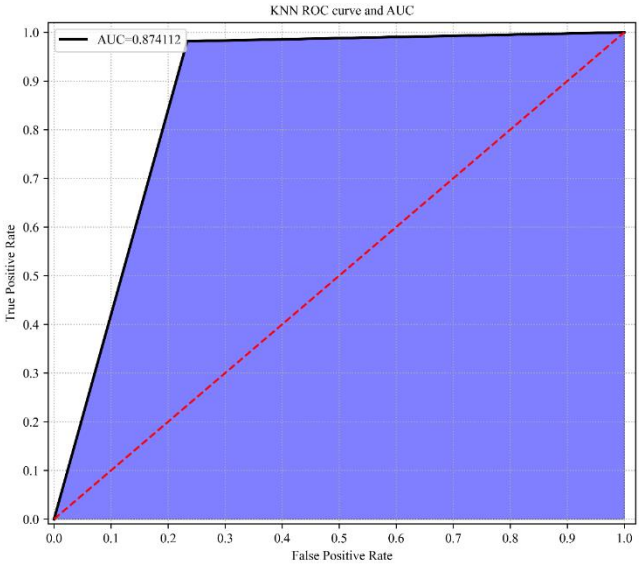

图 1-4 KNN模型的ROC曲线

3.1.5 KNN模型的混淆矩阵

绘制KNN模型的混淆矩阵如图1-5所示，根据主对角线颜色深度和数据个数可知，该模型的准确率、精确度、召回率依次为0.8746291849131234、0.8093416782568382、0.9817236046675102，整体预测效果较好。

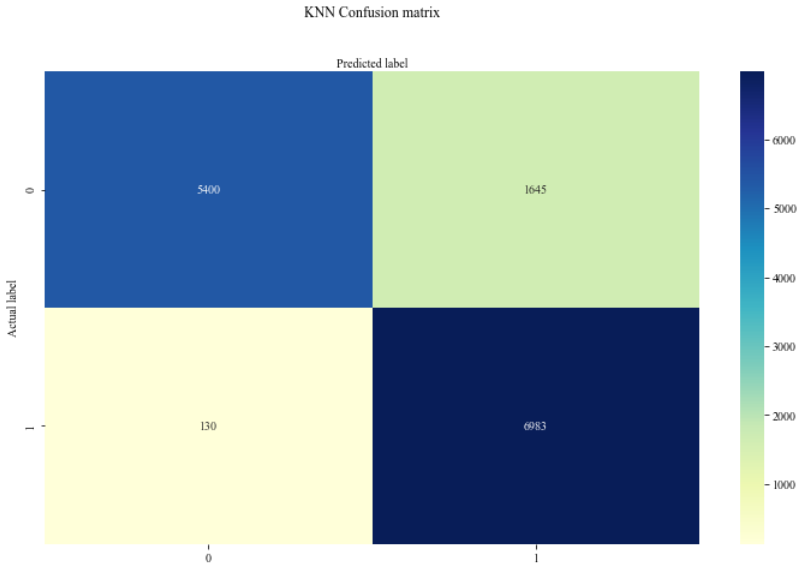

图 1-5 KNN 模型的混淆矩阵

3.2 决策树算法

3.2.1 决策树算法训练

如图2-1所示，为使用决策树算法的输出模型评估结果，根据2-1可知，该算法的预测准确率为0.9007628196072892。

	precision	recall	f1-score	support
0	0.91	0.89	0.90	7045
1	0.89	0.91	0.90	7113
accuracy			0.90	14158
macro avg	0.90	0.90	0.90	14158
weighted avg	0.90	0.90	0.90	14158

图 2-1 决策树算法训练结果

3.2.2 网格搜索最佳参数

网格搜索最佳参数结果如图2-2所示，调参后的预测准确率为0.8959598813391721。

	precision	recall	f1-score	support
0	0.89	0.91	0.90	7045
1	0.91	0.88	0.89	7113
accuracy			0.90	14158
macro avg	0.90	0.90	0.90	14158
weighted avg	0.90	0.90	0.90	14158

图 2-2 网格搜索最佳参数结果

3.2.3 五折交叉验证

对决策树算法模型进行五折交叉验证，验证结果如图2-3所示：

```
array([0.89783563, 0.88890571, 0.88875435, 0.8952626 , 0.8982743 ])
```

图 2-3 五折交叉验证结果

3.2.4 决策树算法的ROC曲线

绘制决策树算法的ROC曲线，根据图2-4结果可以看出，AUC值=0.895808，说明本次模型的整体效果较好。

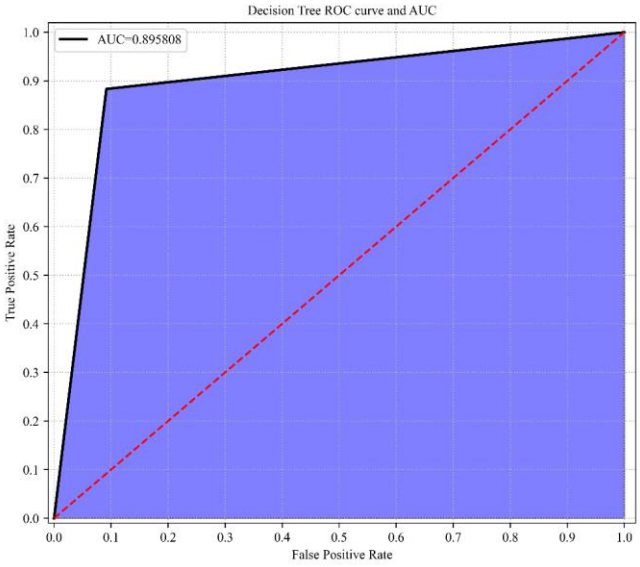

图 2-4 决策树算法的ROC曲线

3.2.5 决策树算法的混淆矩阵

如图2-5所示，为使用决策树算法所得到的混淆矩阵，观察矩阵可知，本次决策树算法的模型准确率为0.8959598813391721，精确度0.9073967061542907，召回率达到0.8830310698720653。

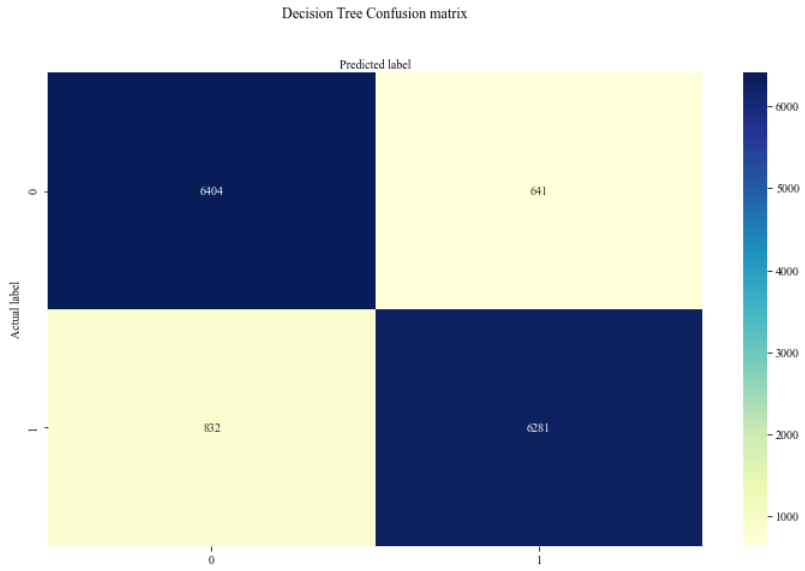

图 2-5 决策树算法的混淆矩阵

3.3 Logistic回归

3.3.1 Logistic回归模型训练

由图3-1所示，使用Logistic回归模型所得到的输出模型评估结果显示，该模型的预测准确率达到0.9363610679474502，具体输出结果如下：

	precision	recall	f1-score	support
0	0.90	0.98	0.94	7045
1	0.98	0.89	0.93	7113
accuracy			0.94	14158
macro avg	0.94	0.94	0.94	14158
weighted avg	0.94	0.94	0.94	14158

图 3-1 Logistic回归训练结果

3.3.2 网格搜索最佳参数

网格搜索最佳参数结果由图3-2所示，搜索出的最佳参数准确率达到0.9363610679474502，处于较高水平。

	precision	recall	f1-score	support
0	0.90	0.98	0.94	7045
1	0.98	0.89	0.93	7113
accuracy			0.94	14158
macro avg	0.94	0.94	0.94	14158
weighted avg	0.94	0.94	0.94	14158

图 3-2 网格搜索最佳参数结果

3.3.3 五折交叉验证

对Logistic回归模型进行五折交叉验证，验证结果见下图3-3：

```
array([0.93491751, 0.93370667, 0.93809596, 0.93764189, 0.93551317])
```

图 3-3 五折交叉验证结果

3.3.4 Logistic回归模型的ROC曲线

如图3-4所示为Logistic回归模型的ROC曲线绘制结果，使用该模型得到的AUC=0.936578，说明模型较为成功。

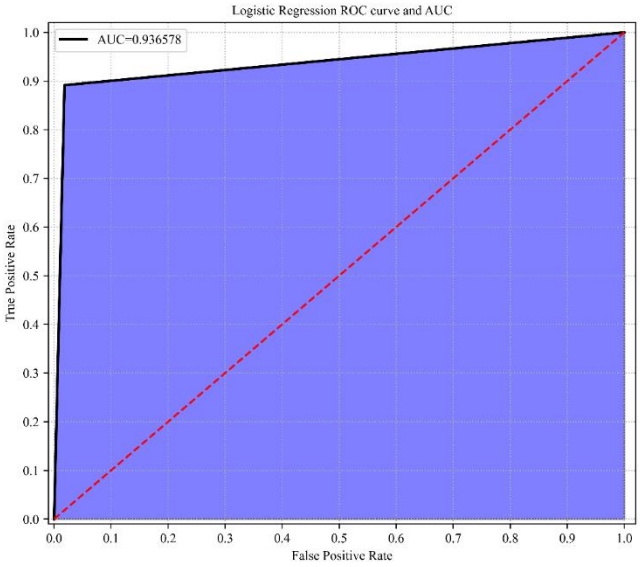

图 3-4 Logistic回归的ROC曲线

3.3.5 Logistic回归模型的混淆矩阵

Logistic回归的混淆矩阵如下图，根据图3-5可知，采用Logistic回归模型预测的准确率达到0.9363610679474502，精确度达0.9800618238021638，召回率0.8914663292562913，模型预测效果较好。

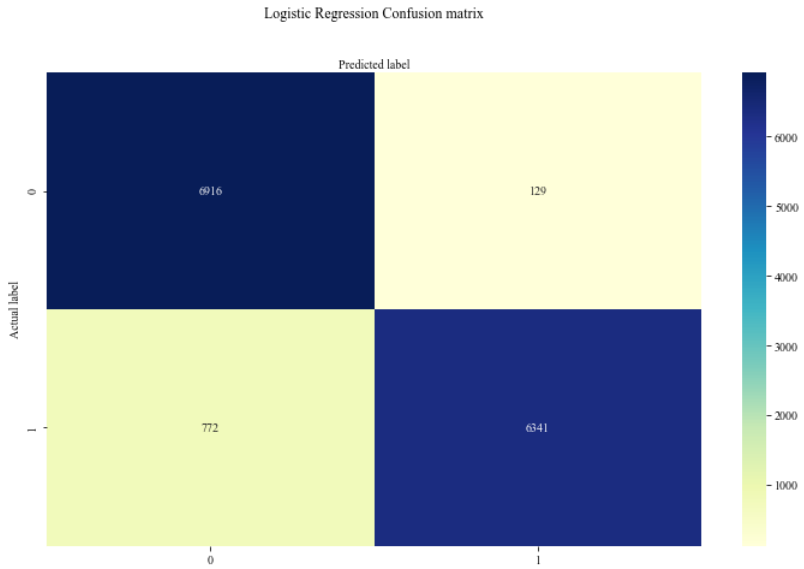

图 3-5 Logistic回归的混淆矩阵

3.4 支持向量机

3.4.1 支持向量机模型训练

采用支持向量机进行模型训练，并输出准确率及模型评估结果，其中模型准确率达到0.899773979375618，输出模型评估结果如图4-1所示：

	precision	recall	f1-score	support
0	0.87	0.94	0.90	7045
1	0.94	0.86	0.90	7113
accuracy			0.90	14158
macro avg	0.90	0.90	0.90	14158
weighted avg	0.90	0.90	0.90	14158

图 4-1 支持向量机训练结果

3.4.2 网格搜索最佳参数

利用网格搜索选取最佳参数，搜索结果如图4-2所示，调参后的准确率为0.9258369826246645，输出模型评估结果如下：

	precision	recall	f1-score	support
0	0.90	0.96	0.93	7045
1	0.96	0.89	0.92	7113
accuracy			0.93	14158
macro avg	0.93	0.93	0.93	14158
weighted avg	0.93	0.93	0.93	14158

图 4-2 网格搜索最佳参数结果

3.4.3 五折交叉验证

进行五折交叉验证，验证结果如下图4-3：

```
array([0.92689572, 0.92492811, 0.92598759, 0.92417133, 0.92264608])
```

图 4-3 五折交叉验证结果

3.4.4 支持向量机的ROC曲线

如图4-4所示为支持向量机的ROC曲线绘制结果，使用该模型得到的AUC=0.926014，说明模型整体效果较好。

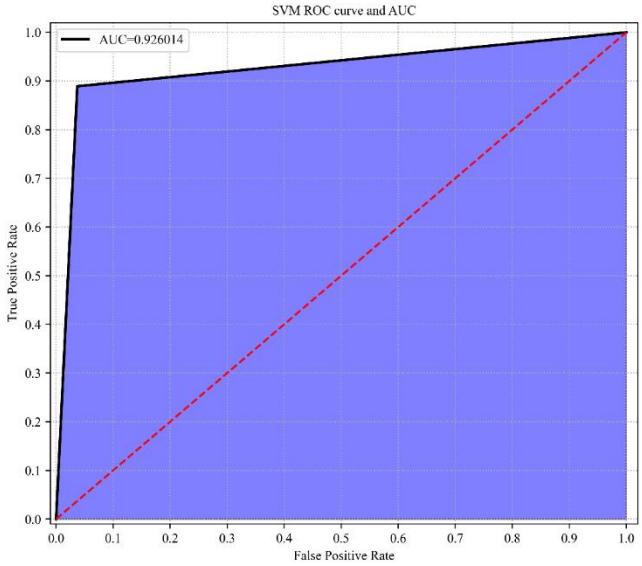

图 4-4 支持向量机的ROC曲线

3.4.5 支持向量机的混淆矩阵

支持向量机的混淆矩阵如下图4-5所示，该模型的准确率为0.9258369826246645，精确度为0.9602246849855777，召回率为0.889216926753831。

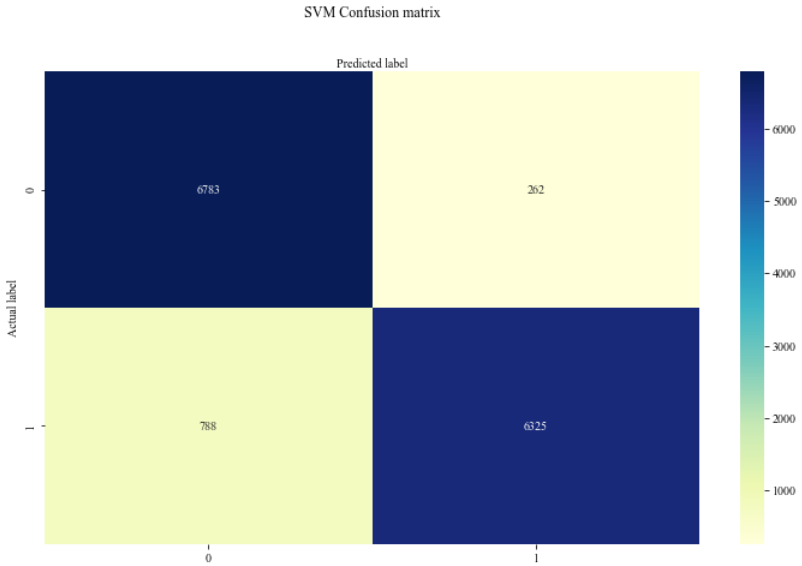

图 4-5 支持向量机的混淆矩阵

3.5 随机森林模型

3.5.1 随机森林模型训练

如图5-1所示，使用随机森林模型所得到的输出模型评估结果显示，该模型的预测准确率达到0.9407402175448509，具体输出结果如下。

	precision	recall	f1-score	support
0	0.93	0.95	0.94	7045
1	0.95	0.93	0.94	7113
accuracy			0.94	14158
macro avg	0.94	0.94	0.94	14158
weighted avg	0.94	0.94	0.94	14158

图 5-1 随机森林模型训练结果

3.5.2 网格搜索最佳参数

利用网格搜索随机森林模型的最佳参数，预测准确率达到0.9378443282949569，调参后的输出模型评估结果如图5-2：

	precision	recall	f1-score	support
0	0.93	0.95	0.94	7045
1	0.95	0.93	0.94	7113
accuracy			0.94	14158
macro avg	0.94	0.94	0.94	14158
weighted avg	0.94	0.94	0.94	14158

图 5-2 网格搜索最佳参数结果

3.5.3 五折交叉验证

对随机森林模型进行五折交叉验证，验证结果如下图5-3：

```
array([0.93355532, 0.93355532, 0.93718783, 0.93219313, 0.932637  ])
```

图 5-3 五折交叉验证结果

3.5.4 随机森林模型的ROC曲线

随机森林模型的ROC曲线如图5-4所示，由图表可知，随机森林模型的AUC值=0.937886，表示该模型预测效果较好。

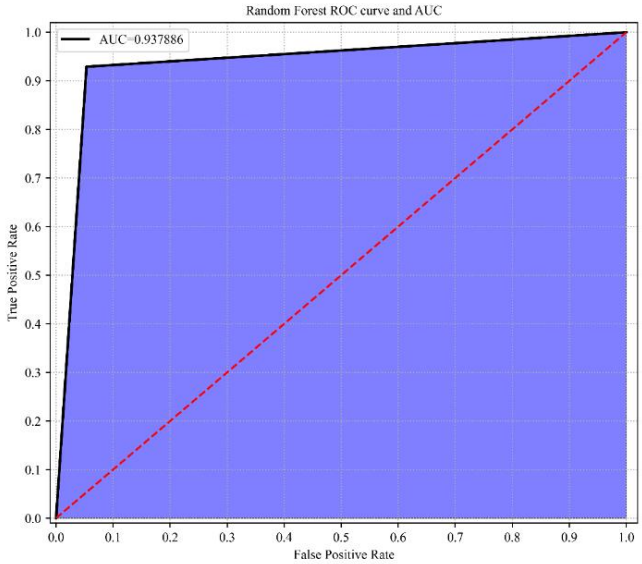

图 5-4 随机森林模型的ROC曲线

3.5.5 随机森林模型的混淆矩阵

图5-5表示使用随机森林模型所得到的混淆矩阵，观察矩阵可知，本次随机森林模型准确率为0.9378443282949569，精确度0.9461703650680029，召回率达到0.929143821172501，模型整体效果令人满意。

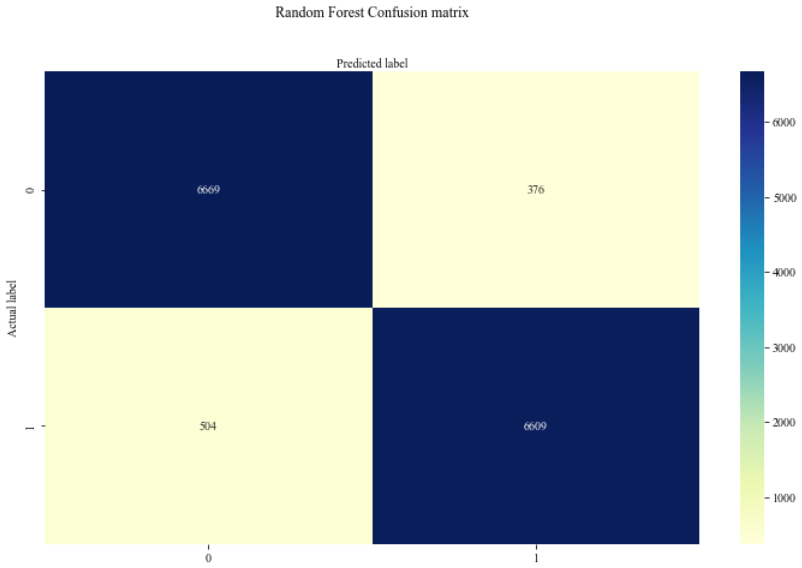

图 5-5 随机森林模型的混淆矩阵

四、模型的确定与预测

4.1 预测模型的确定

本次分析所采用的五种模型评估结果如下表1所示，综合比较各个模型的准确率、精确度、召回率、F1得分以及ROC曲线中的AUC值。准确率最高的是Logistic回归和随机森林模型；精确度最高的为Logistic回归；召回率最高的为KNN模型；F1得分最高的模型为随机森林模型；AUC值最高的模型是Logistic回归和随机森林模型。

	accuracy	precision	recall	f1-score	AUC
KNN模型	0.87	0.80	0.98	0.88	0.87
决策树算法	0.90	0.91	0.88	0.89	0.90
Logistic回归	0.94	0.98	0.89	0.93	0.94
支持向量机	0.93	0.96	0.89	0.92	0.93
随机森林模型	0.94	0.95	0.93	0.94	0.94

表 1 模型综合比较

其中Logistic回归和随机森林模型得分最优的次数最多，但由于Logistic回归模型在召回率方面存在明显短板，综合比较认为随机森林模型的预测效果最好，因此决定采用随机森林模型进行金融产品购买的最终预测。

4.2 随机森林模型预测

（1）数据的导入与处理

首先导入数据，进行可视化处理，如图16所示，需要预测的数据集中共有2000条数据，19个变量。并进一步进行数据的编码与标准化，测试模型时所用数据有62列，此时仅剩61列。去除上文数据集中多余的“default_yes”列，并重新进行训练。

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	campaign	passed_days	previous	pre_outcome	emp_rate	cpi	cci	r3m	employed
0	48	technician	married	high.school	unknown	no	yes	cellular	aug	tue	1	999	0	nonexistent	1.4	93.444	-36.1	4.963	5228.1
1	46	blue-collar	married	professional.course	unknown	no	yes	telephone	may	tue	2	999	0	nonexistent	1.1	93.994	-36.4	4.856	5191.0
2	43	admin.	married	high.school	no	yes	no	telephone	jul	mon	10	999	0	nonexistent	1.4	93.918	-42.7	4.960	5228.1
3	49	services	married	high.school	unknown	yes	no	cellular	jul	wed	1	999	0	nonexistent	1.4	93.918	-42.7	4.957	5228.1
4	39	technician	married	professional.course	no	no	no	cellular	aug	wed	14	999	0	nonexistent	1.4	93.444	-36.1	4.964	5228.1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1995	34	technician	single	professional.course	no	no	no	telephone	jun	thu	8	999	0	nonexistent	1.4	94.465	-41.8	4.866	5228.1
1996	34	management	married	university.degree	no	no	no	cellular	aug	thu	1	999	0	nonexistent	-2.9	92.201	-31.4	0.873	5076.2
1997	25	blue-collar	married	basic.9y	unknown	yes	no	cellular	jul	thu	12	999	0	nonexistent	1.4	93.918	-42.7	4.968	5228.1
1998	30	technician	single	professional.course	no	yes	no	cellular	aug	tue	6	999	0	nonexistent	1.4	93.444	-36.1	4.963	5228.1
1999	28	management	single	university.degree	no	yes	no	cellular	jul	wed	9	999	0	nonexistent	1.4	93.918	-42.7	4.963	5228.1

2000 rows × 19 columns

图 16 预测数据导入

（2）数据的预测

使用随机森林模型对新导入数据进行预测，该操作步骤与上文训练模型时的步骤相同，故在此不再赘述。

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	campaign	passed_days	previous	pre_outcome	emp_rate	cpi	cci	r3m	employed	y
0	48	technician	married	high.school	unknown	no	yes	cellular	aug	tue	1	999	0	nonexistent	1.4	93.444	-36.1	4.963	5228.1	no
1	46	blue-collar	married	professional.course	unknown	no	yes	telephone	may	tue	2	999	0	nonexistent	1.1	93.994	-36.4	4.856	5191.0	no
2	43	admin.	married	high.school	no	yes	no	telephone	jul	mon	10	999	0	nonexistent	1.4	93.918	-42.7	4.960	5228.1	no
3	49	services	married	high.school	unknown	yes	no	cellular	jul	wed	1	999	0	nonexistent	1.4	93.918	-42.7	4.957	5228.1	no
4	39	technician	married	professional.course	no	no	no	cellular	aug	wed	14	999	0	nonexistent	1.4	93.444	-36.1	4.964	5228.1	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1995	34	technician	single	professional.course	no	no	no	telephone	jun	thu	8	999	0	nonexistent	1.4	94.465	-41.8	4.866	5228.1	no
1996	34	management	married	university.degree	no	no	no	cellular	aug	thu	1	999	0	nonexistent	-2.9	92.201	-31.4	0.873	5076.2	yes
1997	25	blue-collar	married	basic.9y	unknown	yes	no	cellular	jul	thu	12	999	0	nonexistent	1.4	93.918	-42.7	4.968	5228.1	no
1998	30	technician	single	professional.course	no	yes	no	cellular	aug	tue	6	999	0	nonexistent	1.4	93.444	-36.1	4.963	5228.1	no
1999	28	management	single	university.degree	no	yes	no	cellular	jul	wed	9	999	0	nonexistent	1.4	93.918	-42.7	4.963	5228.1	no

2000 rows × 20 columns

图 17 预测数据导出

将预测结果添加到数据集，预测时采用的变量为0，1变量，在此根据是否认购金融产品进行还原，将其对应还原为“yes”“no”。输出预测结果，可知对于本次所需要预测的2000个数据，其中认购该金融产品的预测个数为248人，不购买的人数为1752人。

参考文献

[1]Siemers Friederike Maite,Bajorath Jürgen. Differences in learning characteristics between support vector machine and random forest models for compound classification revealed by Shapley value analysis[J]. Scientific Reports,2023,13(1).

[2]Prasojo Rahman Azis,Putra Muhammad Akmal A.,Ekojono ,Apriyani Meyti Eka,Rahmanto Anugrah Nur,Ghoneim Sherif S.M.,Mahmoud Karar,Lehtonen Matti,Darwish Mohamed M.F.. Precise transformer fault diagnosis via random forest model enhanced by synthetic minority over-sampling technique[J]. Electric Power Systems Research,2023,220.

[3]Gelbard Rondi B,Hensman Hannah,Schobel Seth,Stempora Linda,Gann Eric,Moris Dimitrios,Dente Christopher J,Buchman Timothy,Kirk Allan,Elster Eric. A Random Forest Model Using Flow Cytometry Data Identifies Pulmonary Infection after Thoracic Injury.[J]. The journal of trauma and acute care surgery,2023.

[4]Shen Feng,Yang Zhiyuan,Zhao Xingchao,Lan Dao. Reject inference in credit scoring using a three-way decision and safe semi-supervised support vector machine[J]. Information Sciences,2022,606.

[5]王智立.基于主成分分析的随机森林信用卡违约预测[J].金融文坛,2023(01):49-52.

[6]Zou Dexu,Xiang Yongjian,Zhou Tao,Peng Qingjun,Dai Weiju,Hong Zhihu,Shi Yong,Wang Shan,Yin Jianhua,Quan Hao. Outlier detection and data filling based on KNN and LOF for power transformer operation data classification[J]. Energy Reports,2023,9(S7).

[7]Li Yang,Ercisli Sezai. Data-efficient crop pest recognition based on KNN distance entropy[J]. Sustainable Computing: Informatics and Systems,2023,38.

[8]郜燕群.基于粒化 SVM 的互联网金融产品大数据回测分析——以“掌柜钱包”为例[J].现代商贸工业,2019,40(16):104-106.DOI:10.19311/j.cnki.1672-3198.2019.16.046.

[9]马卫民,许卫华. 数据挖掘在预测金融机构发行个人理财产品中的应用[C]//中国运筹学会智能计算分会.第三届中国智能计算大会论文集.Global-Link Publisher,2009:66-70.

[10]Chen Haiqing,Zhao Xu,Zhu Leilei,Cheng Weihu,Xu Lu. Fitting generalized logistic distribution by least squares based on the logistic transformation of order statistics[J]. Communications in Statistics - Theory and Methods,2023,52(2).

[11]Lydersen Stian. Logistic regression with more than two categories.[J]. Tidsskrift for den Norske laegeforening : tidsskrift for praktisk medicin, ny raekke,2022,142(10).