中央财经大学

Central University of Finance and Economics

学年学期：　　2023 - 2024 学年第一学期

课程名称：　　人工智能原理与金融应用

课程代码：　　_____

任课教师：　　张宁、王靖一、王忏、韩东力

姓名：　董心诣(组长) 邓羡韵 谢妤晴

学号：　2021310447、2021310352、2021310341

班级：　金融科技 21

总　　分：_____

评 分 人：_____

基于人工智能原理的银行客户认购产品的预测模型

目录

图表目录

1. 问题背景

用户购买预测是数字化营销领域中的重要应用场景。我们用人工智能的方法对前期客户调研信息进行数据分析，预测客户是否会购买银行的产品，从而为企业提供销售策略，提高营销效率，也为消费者提供更适合的商品推荐，精准对接需求。

该数据集由阿里云天池大赛提供，其中包括客户前期沟通的信息（和客户联系的次数，上一次联系的时长，上一次联系的时间间隔）、客户的基本信息（年龄、职业、婚姻、之前是否有违约、是否有房贷等）以及当前市场的情况（就业、消费信息、银行同业拆解率等）。

本文选取某银行的客户营销信息作为模型训练和验证数据集，建立？模型，进而基于不同的指标进行比较。

2. 数据和变量含义

数据集包括 22500 条购买记录，其中 19548 条购买记录为“是”（87%），2952 条索赔记录为“否”（13%）。19 个变量含义及类型如下表：

表 1：变量含义及类型

字段	说明	类别
age	年龄	离散
job	职业：admin, unknown, unemployed, management…	离散
marital	婚姻：married, divorced, single	类别
default	信用卡是否有违约: yes or no	类别
housing	是否有房贷: yes or no	类别
contact	联系方式：unknown, telephone, cellular	类别
month	上一次联系的月份：jan, feb, mar, …	类别
day_of_week	上一次联系的星期几：mon, tue, wed, thu, fri	类别
duration	上一次联系的时长（秒）	离散
campaign	活动期间联系客户的次数	离散
pdays	上一次与客户联系后的间隔天数	离散
previous	在本次营销活动前，与客户联系的次数	离散

poutcome	之前营销活动的结果：unknown, other, failure, success	类别
emp_var_rate	就业变动率（季度指标）	连续
cons_price_index	消费者价格指数（月度指标）	连续
cons_conf_index	消费者信心指数（月度指标）	连续
lending_rate3m	银行同业拆借率 3 个月利率（每日指标）	连续
nr_employed	雇员人数（季度指标）	连续
subscribe	客户是否进行购买：yes 或 no	类别

3. 特征工程和采样

在数据预处理阶段得到的对客户购买影响有重要影响的 19 个特征中，描述数据点的类型有连续（float）、离散（int）和类别（object）。由于数据表示方式会对机器学习模型的性能产生巨大影响，需要进行数据标准化、文本编码以及数据降维，生成新的特征。

将连续型和离散型的数值变量进行标准化后的描述性统计表和频数直方图如下。根据图 1：数值变量的频数直方图我们可以直观地看到无法根据单独的某个变量来区分是否购买产品的两类客户，因此需要进行建模。

表 2：描述性统计表

	Mean	Std	Min	Max
age	0	1	-2.01	5.01
duration	0	1	-0.80	2.79
campaign	0	1	-0.47	7.42
pdays	0	1	-2.36	0.84
previous	0	1	-0.69	2.44
emp_var_rate	0	1	-2.21	0.84
cons_price_index	0	1	-2.10	2.11
cons_conf_index	0	1	-2.31	2.46
lending_rate3m	0	1	-1.68	1.22
nr_employed	0	1	-2.47	2.06

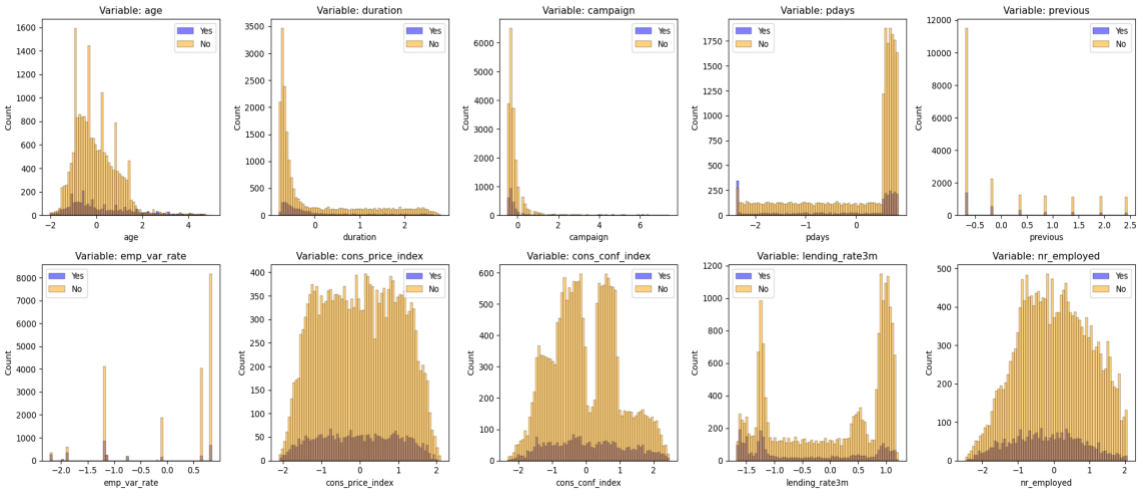

图 1：数值变量的频数直方图

之后，对类别型变量：job, marital, education, default, housing,loan, contact, month, day_of_week, poutcome 进行独热编码。根据独热编码方法，我们总会获得了对购买有影响的独立特征。

同时，我们注意到如图？所示，数据集中购买产品的客户为少数类，不购买产品的客户为多数类。这种情况在机器学习中被称为数据不平衡问题。处理不平衡问题的方法主要有两类，一种是在数据层面进行下采样（Undersampling）或过采样（Oversampling）,减小或增加某个类别的数量。本文选择 SMOTE（Synthetic Minority Over-sampling Technique）过采样的方法，处理不平衡数据集中的分类问题。其基本思想是通过创建少数类别的合成样本来增加少数类别的样本数量，从而平衡类别分布。最终购买与不够买产品的样本数量分别为 5864（23%）和 19548（77%）。

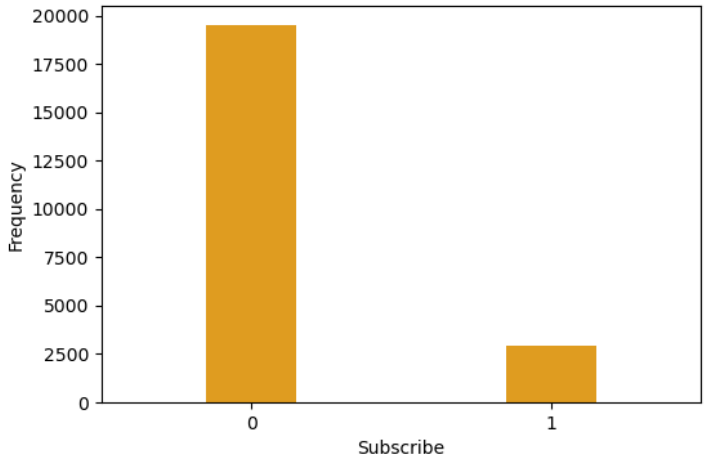

图 2：原始 Subscribe 变量的频数柱状图

接下来剔除一些信息重叠的特征，即进行数据的降维。在机器学习中，主流的降维处理技术有主成分分析（PCA）、线性判别分析（LDA）和核主成分分析（KPCA），其中前两种属于线性降维，最后一个属于非线性降维。本文将采用 PCA 技术进行数据的降维。根据累计解释方差比例超过 60%的标准，我们选择 8 个主成分。将来自两类样本的不同主成分变量配对画图，我们得到了图 3。可以看出经过主成分降维后，两类样本各自聚类效果较好。

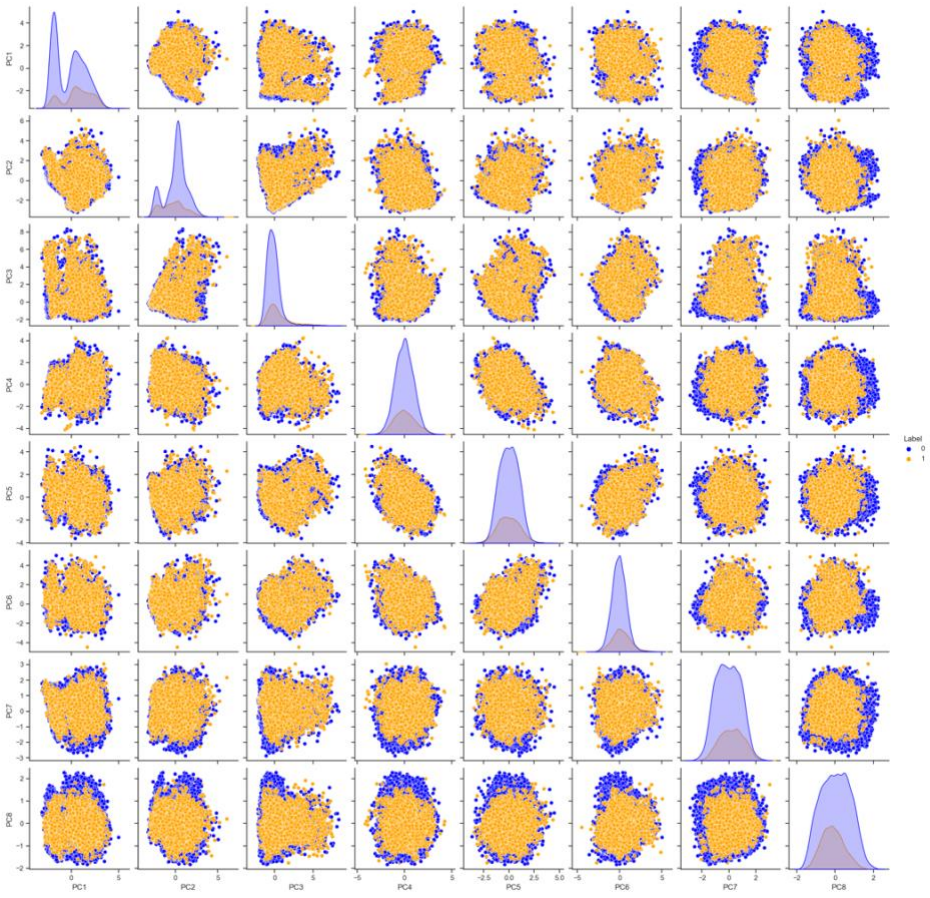

图 3：8 个主成分的 Pair Plot

4. 模型分析与比较

4.1. 聚类分析

聚类分析是一种无监督学习方法，它旨在根据样本数据之间的相似性将样本分成多个群集。由于聚类分析无法直接解决二分类问题，小组采取的思路是先对训练集中的数据进行 K-means 聚类，把银行客户划分成若干类，再以每一类客户的认购比例作为该类客户购买银行产品的意愿，进而对预测客户是否会购买银行产品。

我们先通过肘方法确定聚类数量。根据绘制的碎石图（图 4）可以得知，在 k=4 处误差平方和的下降率突然变缓，因此最佳分组数量为 4，即 k 取 4。

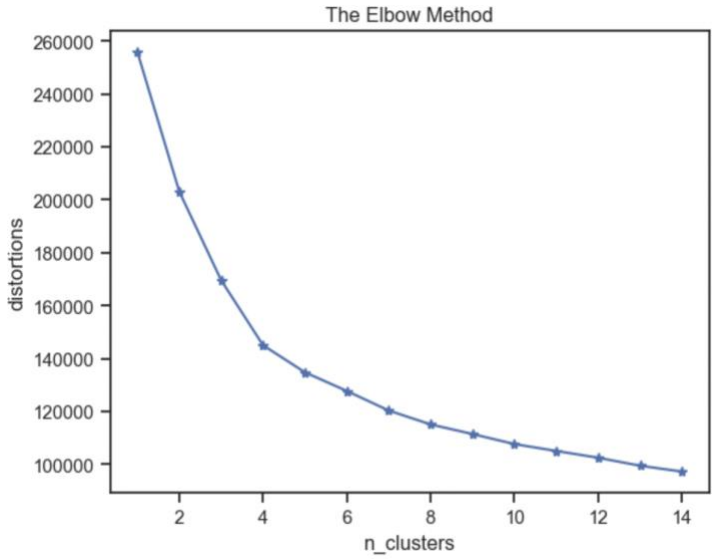

图 4：碎石图

接着我们对训练集的客户数据进行 K-means 聚类分析，获得银行客户的聚类结果。根据上述结果，我们以 PCA 降维得到的前几个主成分为坐标绘制了聚类结果的 2d 散点图和 3d 散点图，如下所示。可以看出 K-means 还是能较好地将客户分成了 4 类。

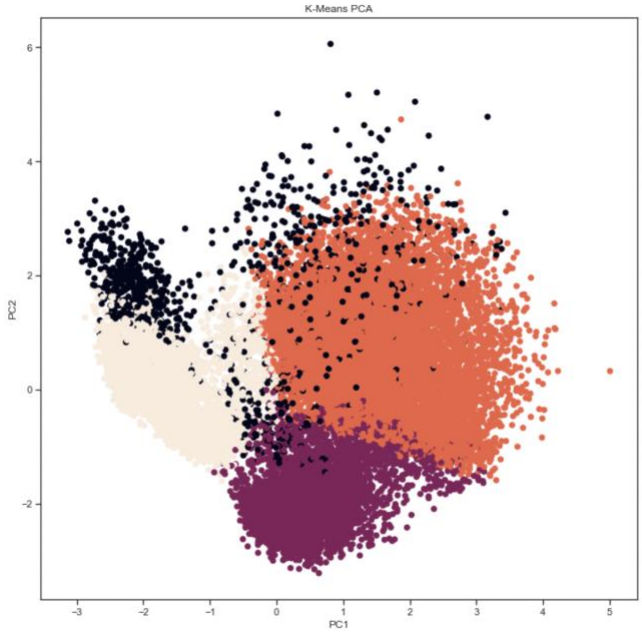

图 5：二维散点图

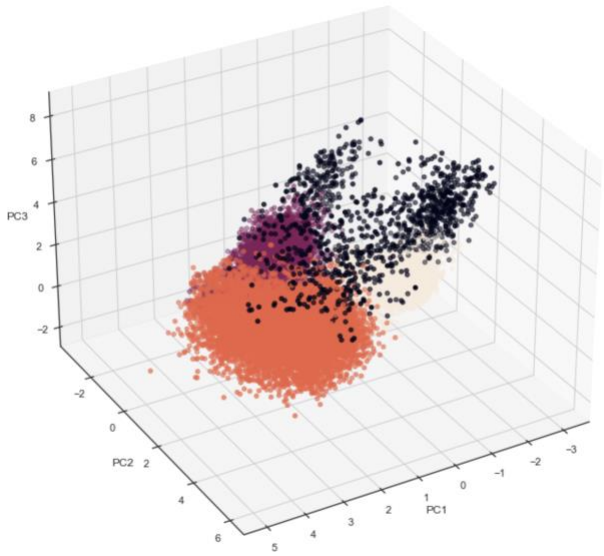

图 6：三维散点图

我们对测试集的数据也同样进行标准化、独热编码、降维等处理，并对这些样本进行分组，得到测试集中的客户的类别结果，如下图所示。

	客户类别
0	2
1	2
2	2
3	2
4	2
...	...
7495	3
7496	3
7497	3
7498	3
7499	3

[7500 rows x 1 columns]

图 7：测试集的客户类别

在整理了训练集每位客户的所属类别以及是否购买产品后，我们运用 Excel 计算得出每类客户的认购比例，并将比例作为该类客户购买银行产品的概率。通过下图可以看出第一类客户的购买比例大于 60%，因此在一定程度上可以预测该类客户倾向于购买银行产品，而第二、三、四类客户则不会购买银行产品。

	客户类别	认购比例
0	1	0.618538
1	2	0.347713
2	3	0.246801
3	4	0.105100

图 8：每类客户的认购情况

4.2. Logistic 回归

我们先将客户数据集划分成训练集和测试集，构建 Logistic 回归模型对训练集数据进行训练，接着对测试集数据进行回归分析，得到预测值，如下图所示。

	y_pred
0	0
1	0
2	0
3	0
4	0
...	...
5078	0
5079	1
5080	0
5081	0
5082	0

[5083 rows x 1 columns]

图 9：Logistic 回归结果

为了评估逻辑回归的预测效果，我们输出了模型的准确率并绘制了混淆矩阵，模型在测试集中的准确率达到 79.1%。

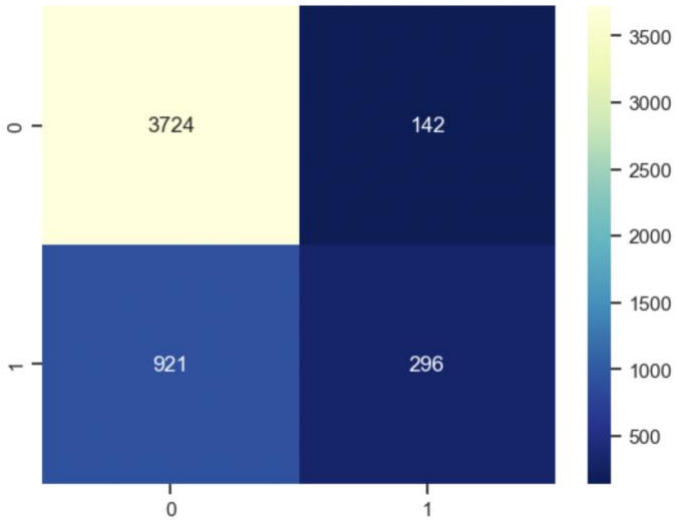

图 10：Logistic 回归的混淆矩阵

同时，我们计算了 AUC 值并绘制了 ROC 曲线，如下图所示，AUC 值为 0.603。

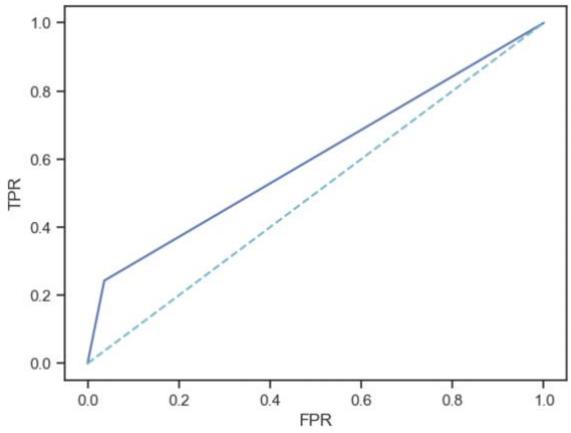

图 11： ROC 曲线

4.3. K 最近邻算法（KNN）

KNN 算法也是一个常用于分类问题的方法，它是指对一个未知样本的预测取决于与其接近的 k 个邻居的数据，这个 k 值的选择会直接影响到预测结果。在银行产品预测这一问题中，我们先构建 k 取默认值 5 的初始模型，对测试集进行预测得到如下图所示的结果。

	y_pred
0	0
1	0
2	0
3	0
4	0
...	...
5078	0
5079	0
5080	0
5081	0
5082	0

[5083 rows x 1 columns]

图 12：KNN 预测结果

同样地，我们绘制了对应的混淆矩阵和 ROC 曲线，该初始 KNN 模型在测试集的准确率为 82.4%，AUC 值为 0.747，预测结果明显优于聚类分析和 Logistic 回归。

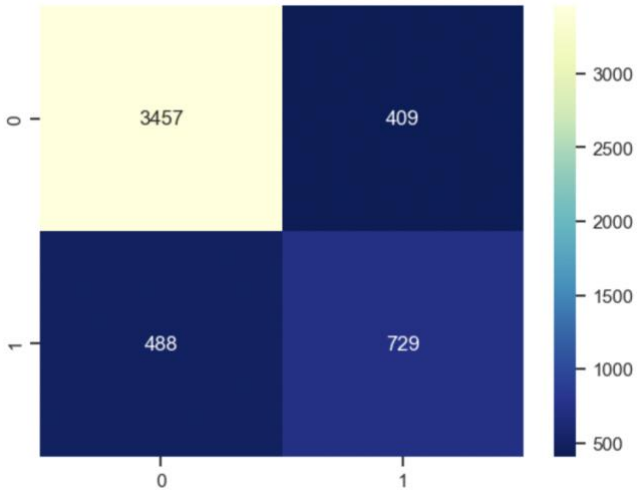

图 13：KNN 混淆矩阵

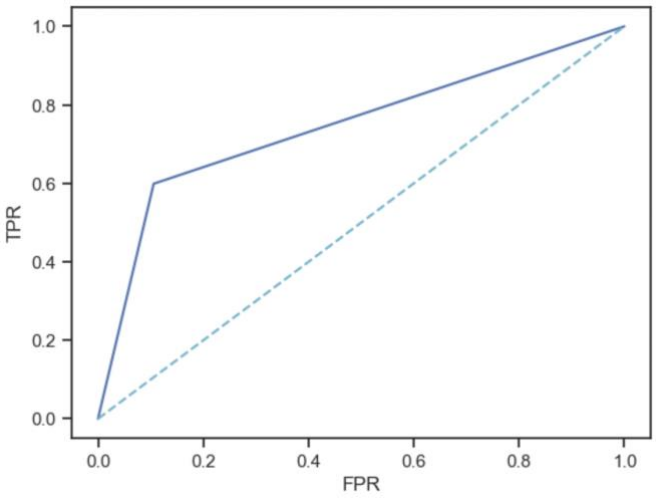

图 14：KNN ROC 曲线

为了选择预测效果最好的 k 值，我们绘制了 k 值与模型准确度之间的关系图，可以看出随着 k 值的增大，模型的准确性呈现出递减的趋势，因此可以认为初始的 KNN 模型已是预测效果较好的选择。

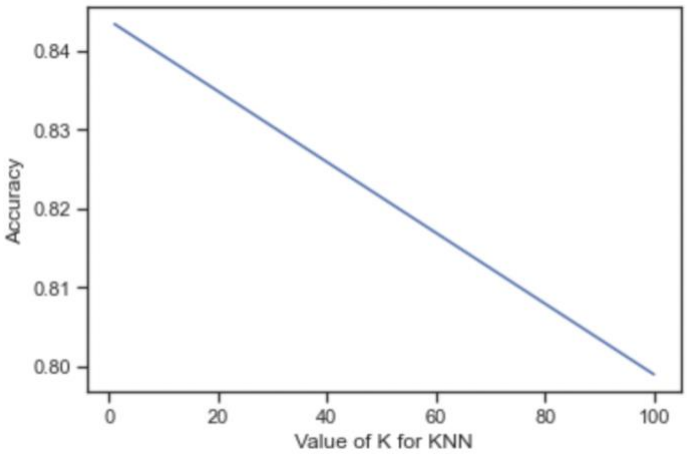

图 15：k 值与模型准确度的关系

4.4. 决策树

决策树是一种常用的分类方法，属于监督学习方式。代码运行结果显示，使用决策树，测试集准确率为 0.791，AUC 值为 0.646。混淆矩阵和 roc 曲线如图所示。

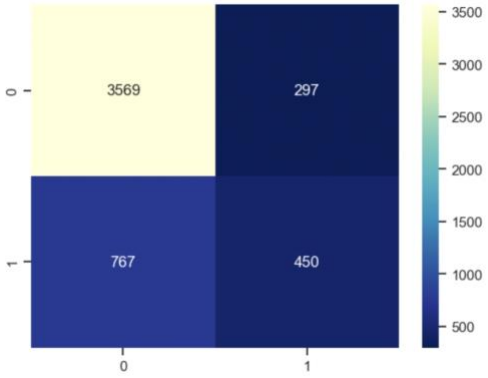

图 16：决策树混淆矩阵

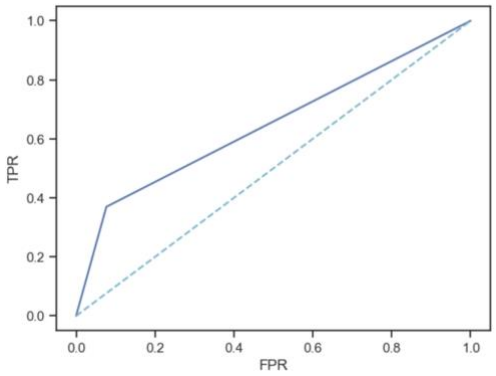

图 17：决策树 roc 曲线

```
|--- feature_2 <= 1.09
|   |--- feature_1 <= -0.38
|   |   |--- feature_0 <= 0.31
|   |   |   |--- class: 0
|   |   |--- feature_0 >  0.31
|   |   |   |--- feature_0 <= 1.95
|   |   |   |   |--- class: 0
|   |   |   |--- feature_0 >  1.95
|   |   |   |   |--- class: 1
|   |--- feature_1 >  -0.38
|   |   |--- class: 0
|--- feature_2 >  1.09
|   |--- class: 1
```

图 18：文字形式输出的决策树

	特征名称	特征重要性
2	PC3	0.484412
0	PC1	0.258604
1	PC2	0.256984
3	PC4	0.000000
4	PC5	0.000000
5	PC6	0.000000
6	PC7	0.000000
7	PC8	0.000000

图 19：决策树特征重要性

从文字形式输出的特征树和特征重要性表格可知，仅有三个特征参与分类，由于决策树模型的弊端，其他五个特征包含的信息在分类中遗漏。树类分析是对一个一个特征进行处理，采用分割的方法。虽然能够深入数据细部，但失去了对全局的把握。分层一旦形成，它与其他节点的关系就被切断了，以后的挖掘只能在局部中进行。

4.5. 随机森林

随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定。其用随机的方式建立一个森林，森林里面有很多的决策树，并且每棵树之间是没有关联的。

代码运行结果显示，使用随机森林，测试集准确率为 0.840，AUC 值为 0.726。两项指标上，随机森林均优于决策树，这与理论相符。它可以处理大量的输入变量，对于多种变量，产生更高准确度的分类器。

混淆矩阵和 roc 曲线如图所示。

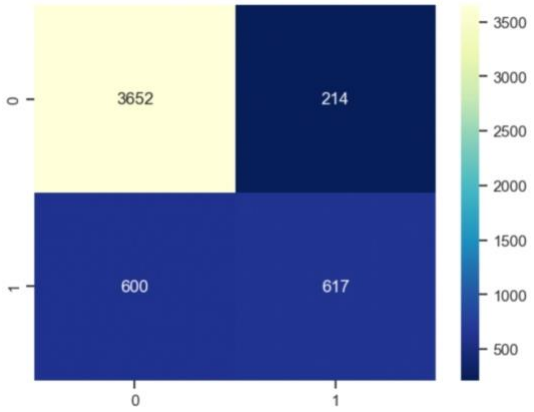

图 20：随机森林混淆矩阵

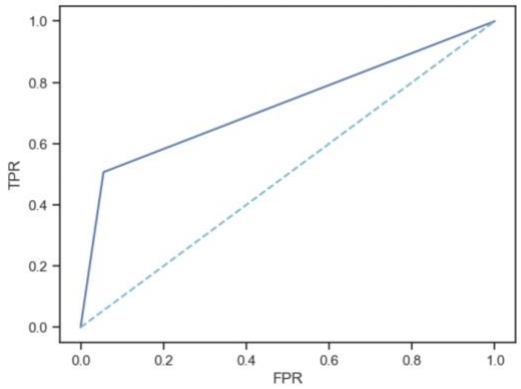

图 21：随机森林 roc 曲线

特征名称	特征重要性
PC3	0.153913
PC2	0.151368
PC1	0.150350
PC6	0.115240
PC8	0.112115
PC7	0.108368
PC4	0.106751
PC5	0.101897

图 22：随机森林特征重要性

与决策树相比，随机森林的特征重要性发生较大变化。八个特征都参与分类，提供了更多信息，更少的信息被浪费。PC3 仍然是最重要的特征，但重要性大幅下降。前三名特征的重要性相差不大，PC2 的重要性超过 PC1。

但随机森林在该实验中也存在一定弊端。本次实验样本量较大，为获得更准确的结果，随机森林模型里树的数量也取得较大（在本次实验中为 500），这导致代码运行较慢。

4.6. 支持向量机

支持向量机（SVM）是一种监督学习方式，是对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。支持向量机中的核函数通常用于映射输入数据到高维特征空间，从而将原始的非线性问题转化为高维空间中的线性问题函数，进行非线性分类、聚类和降维等任务。

在本次报告中，我们的核函数是径向基函数（RBF），即 sklearn 的 SVM 分类算法中使用的默认核函数。结果显示，这种核函数优于线性核（Linear Kernel）。

代码运行结果显示，使用支持向量机，测试集准确率为 0.803，明显高于决策树；AUC 值为 0.645，略低于决策树。这两项的表现都不如随机森林。

支持向量机对参数较敏感，存在多个参数需要调优，不同的参数组合效果差异大。而在本次报告中仅简单地选择了核函数，这对支持向量机的准确度造成一定影响。

混淆矩阵和 roc 曲线如图所示。

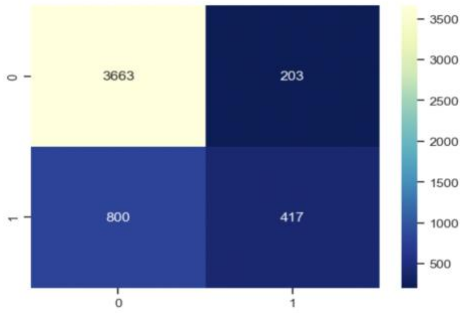

图 23：支持向量机混淆矩阵

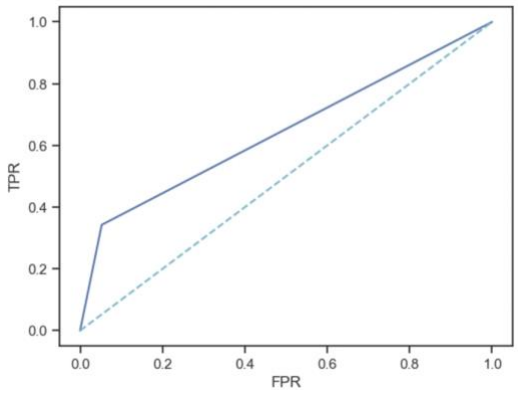

图 24：支持向量机 roc 曲线

5. 总结

通过比较各类机器学习模型的准确度和 AUC 值，可以看出 KNN 算法和随机森林的预测效果明显优于其他三种模型，准确度都大于 80%，支持向量机和决策树的预测效果基本一致，Logistic 回归效果最差。

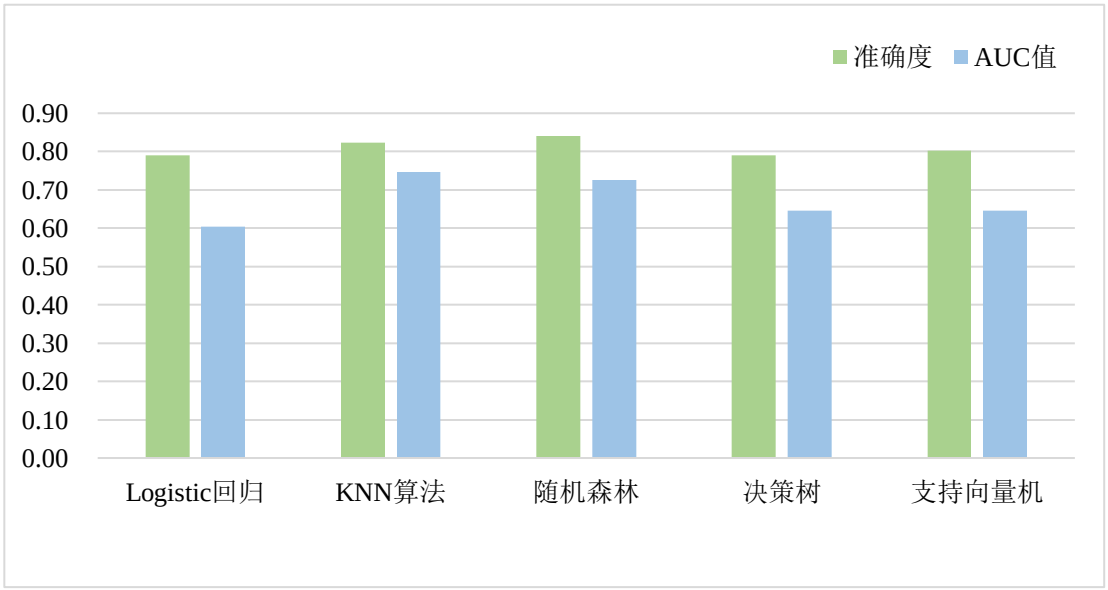

图 25：不同模型预测效果对比