**员工离职数据回归预测**

在Kaggle网站上下载一份数据。然后在skleam选择一个适合的回归算法进行训练、评估和预测。

1.数据来源

（1）员工离职数据：

数据集中包含14999行，10列。

网址：<https://www.kaggle.com/jiangzuo/hr-comma-sep/version/1>

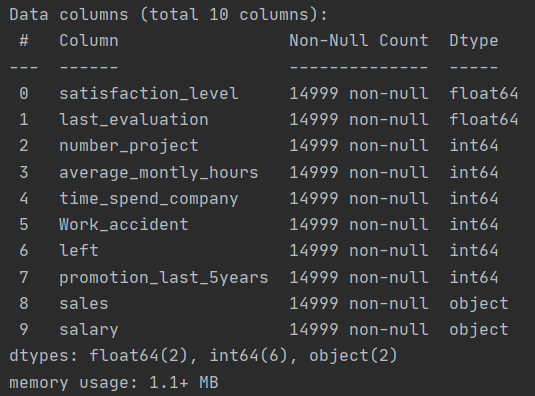
（2）数据特征说明：

|  |  |
| --- | --- |
| satisfaction\_level | 对公司的满意度 |
| last\_evaluation | 绩效评估 |
| number\_project | 参加的项目个数 |
| average\_montly\_hours | 平均每月工作时长 |
| time\_spend\_company | 工作年限 |
| Work\_accident | 是否发生过工作差错：0未发生，1已发生 |
| left | 是否已经离职：0未离职，1已离职,这是目标预测值 |
| promotion\_last\_5year | 五年内是否升职：0未升职，1已升职 |
| sales | 职业：定性变量，10个取值 |
| salary | 薪资水平：定量变量，3个取值 |

2.数据处理

(1)读取数据，列出数据的前十行。

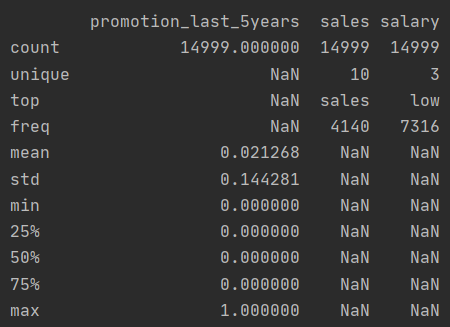
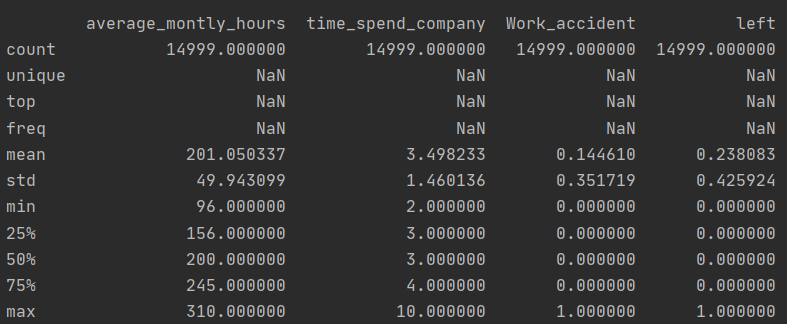
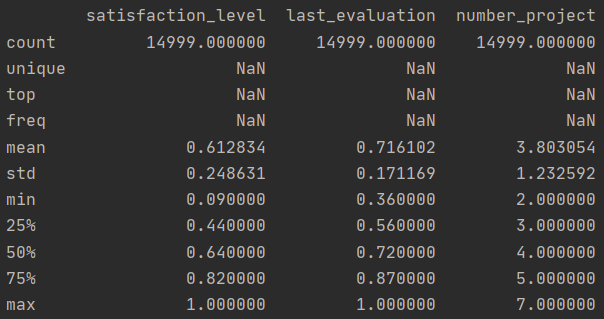
(2)通过data.info()查看数据基本信息。



(3)做数据简要分析。

print(data.describe(include='all'))

观察可发现: average\_montly\_hours特征与其他特征值量级有差距，需要在数据预处理时进行处理。



(4)数据预处理

i.缺失值处理

在查看数据基本信息中已经看到数据没有缺失值，所以不需要进行这个操作。

ii.离散化处理

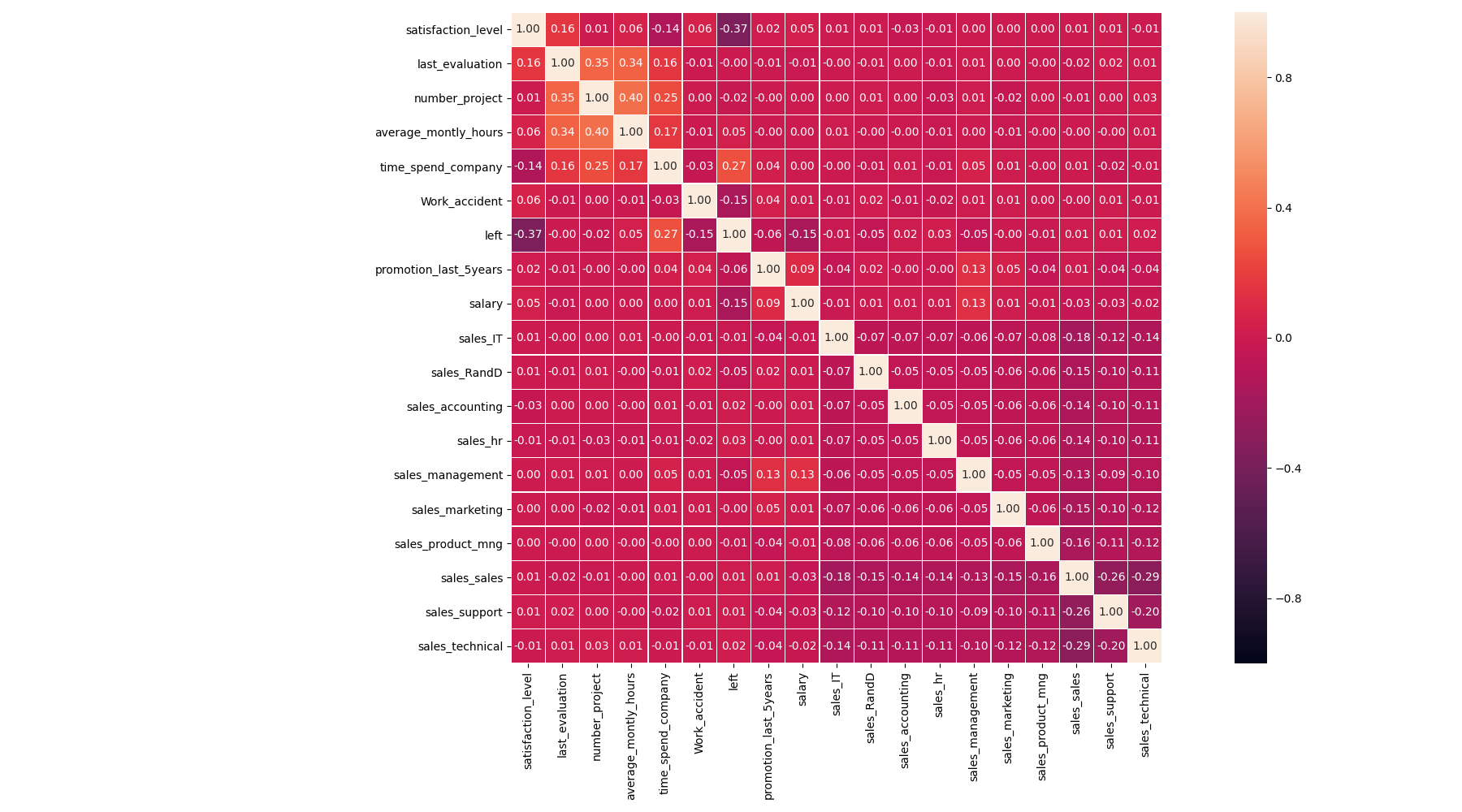
salary为定序变量，将其三个字符类型值转换为数值型。sales是定类变量，对其进行one-hot编码。

iii.归一化处理

average\_montly\_hours与其他特征差异较大，所以进行归一化处理，减小影响。

iiii.计算相关系数绘制热力图

使用斯皮尔曼相关系数计算相关系数，通过绘制出的热力图可以发现变量之间的关联性。越接近1表示正相关性越高，越接近-1表示负相关性越高，越接近0表示相关性越低。



(5)数据集划分

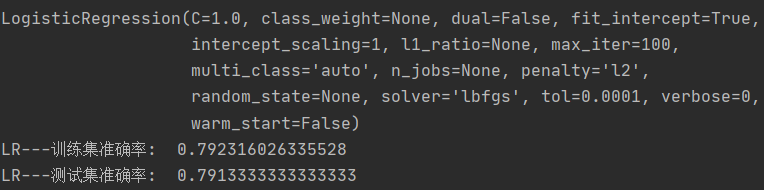
x=data.drop(['left'],axis=1)  
y=data['left']  
x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2,stratify=y,random\_state=1)

3.进行回归模型训练和评估，使用模型对数据进行回归预测。

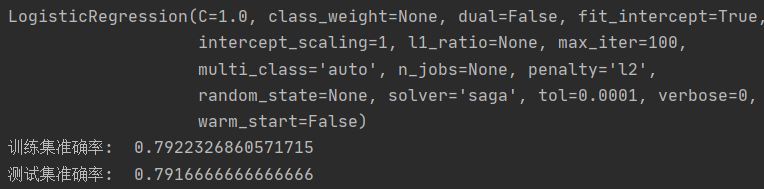
通过多个模型对比，选出最佳模型。

i. 线性回归模型

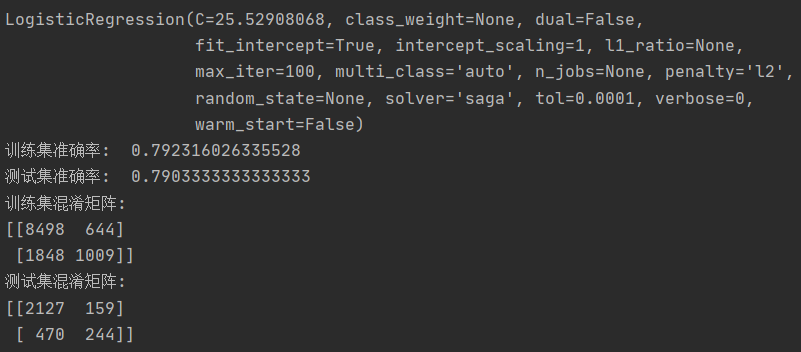
先训练逻辑回归模型的训练，输出准确率。



使用随机梯度下降优化算法，测试集准确率相比提高了一点。

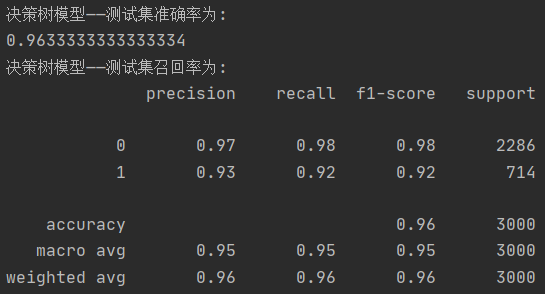


用准确率进行10折交叉验证选择合适的参数C再次预测。



ii. 决策树模型

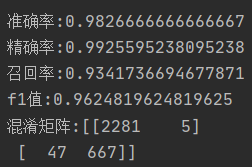
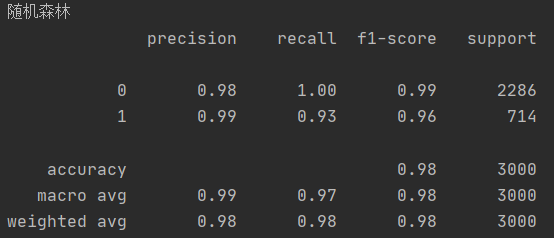
使用sklearn的DecisionTreeClassifier来预测。该模型在测试集上有0.97的精准率和0.98的召回率。



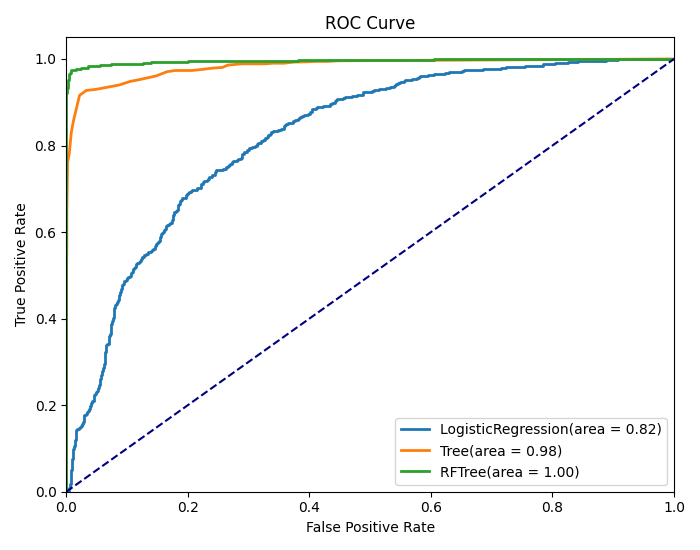
iii. 随机森林模型

使用RandomForestClassifier模型预测：

rf=RandomForestClassifier(  
 criterion="entropy",  
 n\_estimators=1000,  
 max\_depth=None,  
 min\_samples\_split=10  
)  
rf=rf.fit(x\_train,y\_train)



iiii.模型评估：ROC曲线



ROC曲线越靠近左上角说明分类效果越好, 与之对应的auc的值就越大。由图可知，随机森林模型>决策树>逻辑回归。

4.将预测结果写入S3存储。

将随机森林模型预测的结果保存为.csv文件，保存并上传到S3。

# 输出结果  
path = 'predict\_result.csv'  
np.savetxt(path,predict\_rf,delimiter="\n")

