广东金融学院实验报告

课程名称：数据分析与数据挖掘

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验编号  及实验名称 | 实验三 数据挖掘案例实践 | | | 系 别 | 互联网金融 |
| 姓 名 | 钟子期 | 学号（完整学号） | 18154A151 | 班 级 | 18154A1 |
| 实验地点 | 实验楼403 | 实验日期 | 2020/12/10 | 实验时数 |  |
| 指导老师 | 侯昉 | 同组其他成员 | 无 | 成 绩 |  |
| 一、实验目的及要求  1、进一步熟悉基本的数据获取、整理和分析流程及其操作。  2、掌握基本关联规则挖掘、分类、聚类和相关回归分析的算法并应用到实际数据中。 | | | | | |
| 二、实验环境及相关情况（包含使用软件、实验设备、主要仪器及材料等）  1．使用软件：Python开发工具  2．实验设备：个人计算机 | | | | | |
| 三、实验内容及步骤（包含简要的实验步骤流程）  1.自行搜集数据，使用本学期学习的某个算法对其进行数据整理、分析和挖掘。并对得到的结果进行分析总结。  部分数据下载网站：  <http://archive.ics.uci.edu/ml/datasets.php>  <https://www.kesci.com/home/dataset>  <https://www.zhihu.com/question/342295029>  import pandas as pd  import numpy as np  from sklearn.preprocessing import LabelEncoder  from sklearn.impute import SimpleImputer  from sklearn.model\_selection import train\_test\_split  from sklearn.preprocessing import StandardScaler  from sklearn import ensemble  import matplotlib.pyplot as plt  from matplotlib import rcParams  plt.rcParams['font.sans-serif'] = ['SimHei']  plt.rcParams['font.family'] = 'fangsong'  # 清洗数据集方法  def brush(workerDF):  # print(workerDF.head())  # 删除Employee ID和Date of Joining列  workerDF.drop(['EmployeeID'], axis=1, inplace=True)  workerDF.drop(['Date of Joining'], axis=1, inplace=True)  # print(workerDF.head())  # 将字符串类别的列转换为数字类型  workerDF[['Gender']] = workerDF[['Gender']].apply(LabelEncoder().fit\_transform)  workerDF[['Company Type']] = workerDF[['Company Type']].apply(LabelEncoder().fit\_transform)  workerDF[['WFH Setup Available']] = workerDF[['WFH Setup Available']].apply(LabelEncoder().fit\_transform)  # 找出有缺省值的行  sample\_incompletedata = workerDF[workerDF.isnull().any(axis=1)]  #print(sample\_incompletedata.head())  # 用列的中位数填充缺失值  # imputer = SimpleImputer(strategy="median")  # workerDF = imputer.fit\_transform(workerDF)  for i in workerDF.columns:  workerDF[i] = workerDF[i].fillna(workerDF[i].mean())  # print(workerDF.head())  # 转化为浮点数  workerDF = pd.DataFrame(workerDF, dtype=np.float)  # print(workerDF.head())  return workerDF  # 旧数据  #导入员工工作信息  data = pd.read\_csv('train.csv')  workerDF = brush(data)  # 分离训练集和测试集  wkDFdata = workerDF.drop(['Burn Rate'], axis=1)  wkDFrate = workerDF['Burn Rate']  # random\_stat随机种子  Train\_x, Test\_x, Train\_y, Test\_y = train\_test\_split(wkDFdata, wkDFrate, test\_size=0.2, random\_state=42)  # 数据标准化  ss = StandardScaler().fit(Train\_x)  Train\_x = pd.DataFrame(ss.transform(Train\_x), columns=Train\_x.columns)  Test\_x = pd.DataFrame(ss.transform(Test\_x), columns=Test\_x.columns)  # print(Train\_x.head())  # 随机深林回归  model\_rf = ensemble.RandomForestRegressor(n\_estimators=20)  model\_rf.fit(Train\_x, Train\_y)  rf\_score = model\_rf.score(Test\_x, Test\_y)\*100  # print(rf\_score) # 准确度  # 预测  rf\_pred = model\_rf.predict(Test\_x)  # print(rf\_pred) # 预测值  # print(Test\_y.values) # 真实值  # 可视化  fig = plt.figure(figsize = (10, 6))  T1 = Test\_y.values[:100]  P1 = rf\_pred[:100]  x1 = [i for i in range(len(T1))]  x2 = [i for i in range(len(P1))]  plt.plot(x1, T1, 'g', label='实际值', linewidth=3)  plt.plot(x2 ,P1, 'r--', label='预测值', linewidth=2)  fig.tight\_layout()  plt.xlabel('前100条样本')  plt.ylabel('疲劳消耗速度')  plt.legend()  plt.title('疲劳消耗速度预测与实际对比模型（准确率：{:.2f}%）'.format(rf\_score))  plt.show()    # 消耗速度划分的方法  def cul\_percent(data):  percent1,percent2,percent3,percent4,percent5 = 0, 0, 0, 0, 0  for i in data:  if i <= 0.2:  percent1 += 1  elif i <= 0.4:  percent2 += 1  elif i <= 0.6:  percent3 += 1  elif i <= 0.8:  percent4 += 1  else:  percent5 += 1  percent\_list = [percent1, percent2, percent3, percent4, percent5]  return percent\_list  # 饼图  #预测疲劳程度的比例  p\_percent = cul\_percent(rf\_pred)  #实际疲劳程度的比例  t\_percent = cul\_percent(Test\_y.values)  # 状态标签  labels = ['消耗极慢', '缓慢', '消耗正常', '加快', '消耗极快']  plt.pie(p\_percent, labels=labels,  autopct="%.1f%%",  shadow=True,  pctdistance= 0.7,  labeldistance=0.9,  wedgeprops=dict(width=0.4))  plt.pie(t\_percent, labels=labels,  autopct="%.1f%%",  pctdistance= 0.5,  shadow = True,  labeldistance=0.8,  radius=0.6,  wedgeprops=dict(width=0.4, edgecolor='w'))  plt.axis('equal')  plt.title('测试集与训练集的疲劳燃尽速度占比')  plt.legend(ncol=2, bbox\_to\_anchor=(1.2,0), title='训练集比例 测试集比例 ')  plt.savefig('饼图')  plt.show()    **# 用预测模型对新数据进行预测**  # 导入数据并清洗  data = pd.read\_csv('test.csv')  pre\_data = brush(data)  # 数据标准化  pre\_data = pd.DataFrame(ss.transform(pre\_data), columns=pre\_data.columns)  # print(pre\_data.columns)  girl\_data = []  # 总数据预测  total\_pre = model\_rf.predict(pre\_data)  # print(pre\_data.head())  # 划分成男女两个数据集  sex = pre\_data['Gender'].unique()  girl\_data = pre\_data[pre\_data['Gender'].isin([sex[0]])]  boy\_data = pre\_data[pre\_data['Gender'].isin([sex[1]])]  # 对男女数据集进行预测  pre\_girl = model\_rf.predict(girl\_data)  pre\_boy = model\_rf.predict(boy\_data)  # print(len(pre\_girl), len(pre\_boy))  # 调用疲劳消耗速度划分方法，对男生和女生前5800条数据按状态进行比例划分  girl\_per = cul\_percent(pre\_girl[0:5800])  boy\_per = cul\_percent(pre\_boy[0:5800])  # print(girl\_per, boy\_per)  # 直方图  plt.hist(total\_pre, 30, edgecolor='k', facecolor='c', alpha=0.6)  plt.title('疲劳燃尽速度分布')  plt.xlabel('疲劳燃尽速度')  plt.ylabel('人数')  plt.show()  # 条形图  plt.subplot(121)  x = np.arange(5)  bar\_width = 0.35  plt.bar(x, girl\_per, bar\_width, color='k', alpha=0.5)  plt.bar(x+bar\_width, boy\_per, bar\_width, color='r', alpha=0.5)  for i in range(5):  plt.text(x[i], girl\_per[i], '%d' % girl\_per[i], ha='center', va='bottom')  plt.text(x[i]+bar\_width, boy\_per[i], '%d' % boy\_per[i], ha='center', va='bottom')  plt.xticks(x, labels)  plt.xlabel('疲劳消耗速度')  plt.ylabel('人数')  plt.legend(['男生', '女生'])  plt.title('男女疲劳燃尽程度与人数对比')  plt.subplot(122)  plt.bar(x, girl\_per, bar\_width, color='k', alpha=0.5)  plt.bar(x, boy\_per, bar\_width, bottom=girl\_per, color='r', alpha=0.5)  for i in range(5):  plt.text(x[i], girl\_per[i]+boy\_per[i], '%d' % (girl\_per[i]+boy\_per[i]), ha='center', va='bottom')  plt.xticks(x, labels)  plt.xlabel('疲劳消耗速度')  plt.ylabel('人数')  plt.legend(['男生', '女生'])  plt.title('各燃尽程度人数分布')  plt.show()  print('结论：\  \n 1、大部分人疲劳燃尽速度保持在0.6以下，属于正常消耗速度，\  少部分成员疲劳燃尽速度大于0.6，对这部分人要减少工作分配和工作时间，\  过高的疲劳燃尽速度会导致员工过快进入疲劳程度，在疲劳程度下工作效率极低，\  因此分配工作是要考虑疲劳消耗程度，让员工保持良好的工作状态。\  \n 2、男性和女性的在同一工作量和工作时间下，女性的疲劳消耗速度会大于男性，\  在分配工作量时可以适度减少女性的工作量，保持合理疲劳消耗程度。') | | | | | |
| 六、教师评语  1、完成所有规定的实验内容，实验步骤正确，结果正确；  2、完成绝大部分规定的实验内容，实验步骤正确，结果正确；  3、完成大部分规定的实验内容，实验步骤正确，结果正确；  4、基本完成规定的实验内容，实验步骤基本正确，所完成的结果基本正确；  5、未能很好地完成规定的实验内容或实验步骤不正确或结果不正确。  评定等级：  签名： | | | | | |