第一部分 本科生与研究生就业质量评分测量与分析

本项调查针对的对象是已就业的调查对象，包括已就业硕士生和已就业本科生。目的是基于已有研究得到能够衡量就业质量的指标体系，并利用主成分分析法确定各指标影响就业质量的权重，进而得到就业质量评分以便于对就业质量进行定量分析，从而探讨本科生与研究生的就业质量差异及其他相关结论。

一、确立就业质量测量指标

为了构建就业质量评分指标，我们首先需要确定哪些指标会影响就业质量。已有研究已经建立了相对完善的就业质量测量指标体系[大学毕业生就业质量测量指标体系建构]，可以反映毕业生就业质量的概貌，如表4-1所示。

表 4‑1 大学生就业质量测量指标体系

|  |  |  |  |
| --- | --- | --- | --- |
| 一级指标 | 二级指标 | 二级指标内涵 | 具体测量 |
| 就业机会A | 就业与否 | 是否获得就业机会 | 已经就业(落实工作与自主创业视为就业)和未就业两种，考研或出国的数据删除 |
| 工作特征B | 就业地区 | 就业所在区域 | 分城市、城镇和农村三种 |
| 单位类型 | 就业所在单位类型 | 分党政机关、事业单位、国有企业、三资企业、私营企业和集体企业六种 |
| 工作环境 | 工作物理环境与心理环境综合体 | 分“非常好”“很好”“一般”“很差”“非常差”五种情况 |
| 待遇水平C | 薪资水平 | 大学毕业生工作月薪 | 基本工资、奖金与福利等年总收入的月平均数。分6000元以上、5001~6000元、4001~5000元、3001~4000元和3000元以下五种情况 |
| 社会保障 | 单位是否为就业者购买了“五险一金” | 分“全部购买”“部分购买”“全部没买”三种情况 |
| 主观认同D | 人职匹配 | 专业、个性特征与职业性质匹配程度 | 分“非常匹配”“很匹配”“一般”“很不匹配”“非常不匹配”五种情况 |
| 社会地位感知 | 大学生对自己所属社会阶层的感知 | 基于陆学艺的十大阶层划分方法，划分为社会上层、社会中上层、社会中层、社会中下层和社会下层五大阶层 |
| 工作满意程度 | 大学生对自己工作的总体满意情况 | 分“非常满意”“很满意”“一般”“很不满意”“非常不满意”五种情况 |

基于上述就业质量测量指标，本项调查所设计的用于测量就业质量的指标如下表4-2所示：

表 4‑2 就业质量测量问卷

|  |  |
| --- | --- |
| 指标 | 具体测量 |
| 学校类型 | 普通院校、211重点院校、985重点院校 |
| 工作经验年限 | 1年内、2年内、3年内、3年以上 |
| 就业地区 | 县级以下地区、经济不发达的中小城市、经济较为发达的中等城市、省会城市及计划单列市、一线城市（北上广深港澳） |
| 工作环境 | 非常差、很差、一般、很好、非常好 |
| 薪资水平 | 3000以内、3000-5000、5000-8000、8000-10000、10000以上 |
| 单位是否购买五险一金 | 全部没买、部分购买、全部购买 |
| 其他职业福利 | 不享有、享有1种职业福利、享有2种职业福利、享有3种职业福利、享有3种以上职业福利 |
| 专业对口情况 | 非常不匹配、很不匹配、一般、很匹配、非常匹配 |
| 社会地位感知 | 社会下层、社会中下层、社会中层、社会中上层、社会上层 |

二、构建就业质量评分指标

为了定量评估调查对象的就业质量，我们需要建立一个评分指标。以往的研究通常使用层次分析法来确定就业质量指标体系中各个指标的权重。然而，该方法的缺点在于需要评价者的主观判断来进行指标选取和权重确定。本研究采用主成分分析法（Principal Component Analysis，简称PCA），通过多重线性变换和数据运算来获得指标权重。这种方法避免了主观因素的影响，同时也解决了指标间信息重叠的问题。利用有效问卷里的671个已就业调查对象进行分析。为了对定性指标进行定量分析，我们还对收集到的数据进行编码处理，例如学校类型里的“普通院校”用“1”代表、“211重点院校”用2代表、“985重点院校”用“3”代表。采用统计软件spss27.0进行统计分析。

**（一）就业质量测量指标主成分分析**

为了判断问卷调查数据是否适合进行主成分分析，我们进行了各种检验。其中，我们使用了pearson系数对各二级指标之间的相关性进行了分析，并生成了相应的热力图展示相关性结果，如图4-1所示：



图 4‑1 就业质量测量指标pearson相关系数图

根据对问卷调查所得数据的分析，发现各二级指标之间的Pearson相关系数均小于0.3，说明它们的相关性较低，相互之间独立性较强，因此适合进行主成分分析。此外，我们还进行了KMO检验来验证数据的适宜性，结果如下表4-3所示：

表 4‑3 就业测量指标KMO检验和Barlett检验

| KMO检验和Bartlett的检验 | | |
| --- | --- | --- |
| KMO值 | | 0.622 |
| Bartlett球形度检验 | 近似卡方 | 199.35 |
| df | 36 |
| P | 0.000\*\*\* |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | |

整体KMO值为0.622，表示数据适合做主成分分析。Bartelett检验值为199.35，显著性为0.000，表明数据适合进行主成分分析。在进行主成分分析之前，需要对数据进行标准化处理，以避免不同类别数据之间度量的不同性所带来的权重偏差问题。本研究采用极差法对数据进行标准化，并通过主成分分析得到了下表4-4所示的结果。

表 4‑4 主成分分析结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 成分 | 初始特征值 |  |  | 提取载荷平方和 |  |  | 旋转载荷平方和 |
|  | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % | 总计 |
| 1 | 2.943 | 32.7 | 32.7 | 2.943 | 32.7 | 32.7 | 2.448 |
| 2 | 1.198 | 13.313 | 46.012 | 1.198 | 13.313 | 46.012 | 1.527 |
| 3 | 1.001 | 11.117 | 57.129 | 1.001 | 11.117 | 57.129 | 1.167 |
| 4 | 0.892 | 9.915 | 67.044 |  |  |  |  |
| 5 | 0.771 | 8.566 | 75.611 |  |  |  |  |
| 6 | 0.657 | 7.305 | 82.916 |  |  |  |  |
| 7 | 0.571 | 6.341 | 89.257 |  |  |  |  |
| 8 | 0.503 | 5.587 | 94.844 |  |  |  |  |
| 9 | 0.464 | 5.156 | 100 |  |  |  |  |

通过主成分分析法提取了三个主成分。第一主成分至第三主成分的特征值分别为2.943、1.198和1.001,特征值均都大于1,3个主成分对就业质量总方差的解释率分别为32.7%、13.313%和11.117%,累积方差解释率达57.129%,表明主成分分析效果好。

**（二）构建就业质量评分公式**

首先，测量三个主成分对指标的载荷数，结果如下表4-5所示：

表 4‑5 主成分对各指标的载荷数

|  |  |  |  |
| --- | --- | --- | --- |
|  | 成分1 | 成分2 | 成分3 |
| 学校类型 | 0.735 | -0.139 | 0.122 |
| 工作经验年限 | 0.719 | 0.299 | -0.187 |
| 就业地区 | 0.672 | -0.328 | 0.074 |
| 工作环境 | 0.663 | -0.212 | 0.064 |
| 薪资水平 | 0.634 | -0.251 | 0.207 |
| 单位是否购买五险一金 | 0.07 | 0.669 | 0.286 |
| 其他职业福利 | 0.413 | 0.601 | -0.269 |
| 专业对口情况 | 0.487 | 0.012 | -0.661 |
| 社会地位感知 | 0.423 | 0.254 | 0.554 |

其次，继续计算指标在三个主成分线性组合中的系数，具体计算方法是用各指标对应各主成分的载荷数除以主成分对应的特征根的开平方，结果见下表4-6所示：

表 4‑6 各指标在各主成分线性组合中的系数

|  |  |  |  |
| --- | --- | --- | --- |
| 各指标线性组合系数 | | | |
|  | 主成分1 | 主成分2 | 主成分3 |
| 学校类型 | 0.283879368 | 0.010963591 | -0.660669748 |
| 工作经验年限 | 0.040804016 | 0.611220214 | 0.285857107 |
| 就业地区 | 0.240743694 | 0.549093197 | -0.268865601 |
| 工作环境 | 0.428442168 | -0.126994932 | 0.121939046 |
| 薪资水平 | 0.419115535 | 0.27317615 | -0.18690657 |
| 单位是否购买五险一金 | 0.246572839 | 0.232062682 | 0.553723208 |
| 其他职业福利 | 0.386472323 | -0.193690113 | 0.063968024 |
| 专业对口情况 | 0.369567802 | -0.229321784 | 0.206896578 |
| 社会地位感知 | 0.391718553 | -0.299671495 | 0.073963028 |

接着，对各指标在这三个主成分线性组合中的系数做加权平均，并进行归一化处理，计算出各指标的权重。其计算公式是表6中每列先乘以表4中对应主成分的方差贡献率，再每行相加，最后除以三个主成分的累计贡献率，并进行归一化处理。结果如下表4-7所示：

表 4‑7 就业质量测量指标权重表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学校类型 | 工作经验年限 | 就业地区 | 工作环境 | 薪资水平 | 单位是否购买五险一金 | 其他职业福利 | 专业对口情况 | 社会地位感知 |
| 0.0199 | 0.1206 | 0.1162 | 0.1303 | 0.1455 | 0.165 | 0.1027 | 0.108 | 0.0919 |

学校类型，工作经验，就业地区，工作环境，薪资水平，单位是否购买五险一金，其他职业福利，专业对口情况，社会地位感知分别编码为a,b,c,d,e,f,g,h,i。由此可算出就业质量评分公式：

每个指标都被赋予了一个权重系数，这些权重系数的范围在0至1之间，表示每个指标在总体评价中的相对重要性。这九个指标的权重系数之和为1，表明所有指标的重要性共同构成了就业质量的总体评价。例如，“薪资水平”指标的权重系数为0.1455，表示该指标在整个就业质量指标体系中的占比为14.55%。

利用就业质量评分公式计算出收集到的671个已就业调查对象的就业质量评分。

三、基于就业质量评分指标的数据分析

**（一）不同学历层次的就业质量分析**

图 4‑2 不同学历层次的就业质量评分

图4-2是不同学历层次的就业质量评分。学历为硕士的就业质量评分为3.202，学历为本科的就业质量评分为2.962，学历为大专的就业质量评分为2.816。说明随着学历的提升，其就业质量也随着不断提升，因此提升学历是能提高就业质量的。

**（二）不同学校层次的就业质量分析**

图 4‑3 不同学校层次的就业质量评分

图4-3是不同学校层次的就业质量评分。学校层次为985重点院校的就业质量评分为3.308，学校层次为211重点院校的就业质量评分为3.124，学校层次为普通院校的就业质量评分为2.808。说明随着学校层次的上升，其就业质量也随着不断提升，因此努力学习进入一所更高层次的院校是能提高就业质量的。

**（三）不同学科大类的就业质量分析**

图 4‑4 不同学科大类的就业质量评分

图4-4是不同学科大类的就业质量评分。从图中可以看出，不同学科大类的就业质量评分从高到低排名依次为理学工学>医学>经济管理>人文社科艺术>农学。因此假如从就业质量的角度来选择学科大类和专业的话，最推荐的是“理学工学”，最不推荐的是“农学”。

**（四）不同学科大类、不同学历的就业质量分析**

图 4‑5 不同学科大类、不同学历的就业质量评分

图4-5是不同学科大类、不同学历的就业质量评分。由图中可以看出：

1.“人文社科艺术”大类相关专业的学生，读研后所获得的就业质量提升幅度最大，因此最建议读研，以获得在就业质量方面的提升。

2.在就业质量提升幅度方面，不同学科大类读研后所获得的提升幅度从大到小依次为：人文社科艺术>医学>农学>理学工学>经济管理。

**（五）不同就业地区的就业质量分析**

图 4‑6 不同就业地区的就业质量评分

图4-6是不同就业地区的就业质量评分。由图中可以看出，最发达的一线城市其就业质量评分最高，为3.356；最不发达的县级以下地区，其就业质量评分最低，为2.428；两者相差悬殊。因此，如果想有更好的发展机会和职业生涯，应该去发达的城市寻找机会。

**（六）不同单位类型的就业质量分析**

图 4‑7 不同单位类型的就业质量评分

图4-7是不同单位类型的就业质量评分。由图中可以看出，在党政机关工作的就业质量评分最高，为3.175,这从侧面反映了现如今“考公热”、“宇宙的尽头是编制”等描述出现的原因。

不同单位类型的就业质量评分从高到低依次为：党政机关>三资企业>大型私企>国有企业>事业单位>中小型私企。

**（七）不同学历、不同工作年限的就业质量分析**

图 4‑8 不同学历、不同工作年限的就业质量评分

图4-8是不同学历、不同工作年限的就业质量评分。由图中可以看出：

1.工作年限相同，学历越高，其就业质量评分越高；

2.工作年限为“1年内”的硕士就业质量评分为2.937，工作年限为“3年内”的本科生就业质量分数为3.136，这表明，读研所花费的2-3年成本，其所获得的就业质量回报，不如本科生就业2-3年所获得的就业质量回报。但同时，也应该注意到，同样的工作年限，硕士在就业质量方面所能达到的上限，是远优于同样工作年限的本科生的。

因此，在就业质量方面，不是拿到硕士学位就能在刚毕业的时候超过同期毕业已经工作2-3年的本科生。但硕士学历群体的发展上限，要比本科学历的群体的发展上限高更多。

第二部分 硕士研究生读研原因分析

此分类模型是针对具有硕士学历的已毕业人士构建，目的是借助机器学习模型对样本进行训练预测，并且探讨哪些特征是影响硕士人群“是否后悔读研”的因素。

一、数据

**（一）特征集**

表 5‑1 分类模型所选的自变量量表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 第19题选择读研的原因（影响程度） | | | | | |
|  | 1 | 2 | 3 | 4 | 5 |
| 规避就业压力 | ○ | ○ | ○ | ○ | ○ |
| 父母和朋友的影响 | ○ | ○ | ○ | ○ | ○ |
| 提高自身学历和能力 | ○ | ○ | ○ | ○ | ○ |
| 跟随读研大趋势 | ○ | ○ | ○ | ○ | ○ |
| 看好专业未来就业发展前景 | ○ | ○ | ○ | ○ | ○ |
| 名校情结 | ○ | ○ | ○ | ○ | ○ |
| 对所选专业感兴趣 | ○ | ○ | ○ | ○ | ○ |
| 考虑个人自身今后发展 | ○ | ○ | ○ | ○ | ○ |

**（二）目标集**

表 5‑2 分类模型所选的因变量

|  |  |
| --- | --- |
| 第18题是否后悔读研？[单选题] | |
| ○是 | ○否 |

**（三）数据符合说明**

表 5‑3 第19题量表矩阵符号说明

|  |  |
| --- | --- |
| 符号 | 原题 |
| employment\_pressure | 规避就业压力 |
| Influence\_of\_relatives\_and\_friends | 父母和朋友的影响 |
| Improve\_ability | 提高自身学历和能力 |
| Postgraduate\_trend | 跟随读研大趋势 |
| Employment\_prospects | 看好专业未来就业发展前景 |
| Famous\_school\_complex | 名校情结 |
| Professional\_interests | 对所选专业感兴趣 |
| Personal\_development | 考虑个人自身今后发展 |

表 5‑4 第18题是否后悔读研符号说明

|  |  |
| --- | --- |
| 符号 | 原题 |
| 1 | 是 |
| 0 | 否 |

**（四）数据校验**

1．信度校验

表 5‑5 第19题量表矩阵信度检验表

|  |  |  |  |
| --- | --- | --- | --- |
| Cronbach'sα系数 | 标准化Cronbach'sα系数 | 项数 | 评价 |
| 0.823 | 0.824 | 8 | 不错 |

通常来说，问卷的Cronbach'sα系数在0.80以上代表该问卷设计良好，由表5-5中第19题的量表矩阵的Cronbach'sα系数为0.823，表明此特征集的可靠性较高，可纳入模型中。

2．效度校验

表 5‑6 第19题量表矩阵KMO和Bartlett的检验

|  |  |  |
| --- | --- | --- |
| KMO值 | 0.777 | |
| Bartlett球形度检验 | 近似卡方 | 263.817 |
| df | 28 |
| P | 0.000\*\*\* |

注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平

由表5-6知，KMO值为0.777，说明变量间的相关性较强；Bartlett球形度检验统计量观测值为263.817，概率p值接近于0，显著性水平α=0.05时，由于概率p值小于显著性水平α，应拒绝原假设，即认为相关系数矩阵与单位阵有显著差异。

3．平衡性校验

在实际应用中，分类问题中的类别型因变量可能存在严重偏差，即不同类别之间的比例严重失衡。在研究硕士是否后悔读研究时，后悔类别的观测样本占少数。

数据的不平衡可能导致预测结果存在偏见，即分类结果可能会偏向于观测较多的类别。因此，首先需要对目标集进行平衡性检验，统计“是/否”的比例，结果如下图1所示。

图 5‑1 硕士是否后悔读研的比例

由图5-1可知，样本类别比例是存在接近2:1的失衡的，故需要对其进行非平衡性处理。

**（五）非平衡数据的处理**

当解决分类问题中数据不平衡的情况时，最简单的方法是构建1:1的数据集，但这种方法存在许多问题。欠采样会导致某些隐含信息丢失，而过采样则容易导致模型过拟合。为了解决该问题，Chawla在2002年提出了SMOTE算法，它是基于随机过采样算法的改进，用于生成少数类别的一种过采样技术。该算法已被工业界和学术界广泛认可，是非平衡数据处理中常用的解决方法。

二、模型选择——XGBoost算法

XGBoost是一个开源的机器学习项目，是GBDT的一种高效实现。与GBDT相比，XGBoost使用了损失函数的一阶和二阶导数，并支持自定义损失函数，只要该损失函数具有可一阶和二阶求导的性质。XGBoost还支持并行计算，大大提高算法的运行效率。同时，它还能通过加入正则化项控制模型复杂度，避免过拟合。 XGBoost的核心算法思想是：通过不断添加树并进行特征分裂来生长一棵树，每次添加一棵树都是学习一个新函数f(x)，以拟合上次预测的残差。当训练完成得到k棵树后，预测一个样本的分数，其实就是根据该样本的特征，在每棵树中落到对应的一个叶子节点，每个叶子节点对应一个分数。最后将每棵树对应的分数加起来，就可以得到该样本的预测值。

三、模型微调——网格搜索（GridSearchCV）

在机器学习模型中，有些参数无法通过训练得到，需要人工选择，这些参数被称为超参数。选择不合适的超参数会导致欠拟合或过拟合等问题。为了解决这个问题，通常有两种选择方法。一种是根据经验进行微调，另一种是通过尝试不同大小的参数，将其带入模型中，选择表现最佳的参数。

GridSearchCV的名字其实可以拆分为两部分，GridSearch和CV，即网格搜索和交叉验证。网格搜索的作用是对一组超参数进行穷举搜索，以找到最优的超参数组合，从而提高模型的性能。具体而言，GridSearchCV会遍历所有可能的超参数组合，将它们传递给交叉验证模块，然后在每个超参数组合上运行交叉验证来评估模型的性能。这样做的好处是可以自动化地搜索最佳超参数组合，减少了人工调参的时间和精力成本。而交叉验证则是为了减少模型对数据的过度拟合，将数据划分成若干份，每次用其中一份作为验证集，其余作为训练集，然后交替使用这些数据子集进行多次训练和验证，从而得到更稳健的模型性能评估结果。

调用GridSearchCV类，设置参数组合，利用5重交叉验证的网格搜索法，选择最佳的参数组合，参数组合如下表5-7所示。

表 5‑7 模型超参数组合表

|  |  |  |
| --- | --- | --- |
| 参数 | 说明 | 数值范围 |
| n\_estimators | 基础模型的数量 | [40,50,70,90] |
| learning\_rate | 模型迭代的学习率或步长 | [0.01,0.03,0.05,0.1,0.2] |
| max\_depth | 每个基础模型包含的最大深度 | [2,3,4,5] |
| colsample\_bytree | 每个基础模型所需采样字段比例 | [0.4,0.6,0.8,1] |
| min\_child\_weight | 叶子节点中各样本点二阶导之和最小值 | [1,2,3,4] |

XGBoost最佳参数组合如下表5-8所示。

表 5‑8 XGBoost模型最佳参数表

|  |  |
| --- | --- |
| 参数 | 最佳值 |
| n\_estimators | 50 |
| learning\_rate | 0.1 |
| max\_depth | 2 |
| colsample\_bytree | 0.4 |
| min\_child\_weight | 1 |

四、模型评估

将测试集里的数据代入上述模型中，得到如下表5-9的数据统计。

表 5‑9 预测值与实际值对比

|  |  |  |
| --- | --- | --- |
| 数量 | 预测 | 实际 |
| 0（不后悔） | 12 | 12 |
| 1（后悔） | 12 | 12 |

由上表可知，并不能很好的看出模型的拟合效果，故通过混淆矩阵、ROC曲线来进行可视化评估。

**（一）混淆矩阵**

1．理论

在实际的数据集中，通常存在后悔和不后悔两种分类。使用XGBoost模型对样本进行分类预测后，可以得到两列数据，一列是真实的分类序列，另一列是模型预测的分类序列。通过比对这两个序列，可以得到一个汇总的列联表，也称为混淆矩阵。混淆矩阵对于评估模型的性能非常重要，它能够展示模型的分类表现，并且能够计算出多种评价指标，如准确率、召回率、F1值等。如下表5-10所示。

表 5‑10 混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际值 | | |
| 预  测  值 |  | 负例——0 | 正例——1 |
| 负例——0 | A，TrueNegative | B，FalseNegtive |
| 正例——1 | C，FalsePositive | D，TruePositive |

* A：表示正确预测负例的样本个数。
* B：表示预测为负例但实际为正例的个数。
* C：表示预测为正例但实际为负例的个数。
* D：表示正确预测正例的样本个数。
* 准确率：表示正确预测的正负例样本数与所有样本量的比值，即

，该指标用来衡量模型对整体数据的预测效果，用Accuracy表示。

* 正例覆盖率：表示正确预测的正例数在实际正例数中的比例，即，该指标反映的是模型能够在多大程度上覆盖所关心的类别，用Sensitivity表示。
* 负例覆盖率：表示正确预测的负例数在实际负例数中的比例，即，用Specificity表示。

评估模型的好坏，通常会使用混淆矩阵。混淆矩阵由真实分类和预测分类组成，用于比较模型的预测结果与实际结果的差异。在混淆矩阵的基础上，可以使用准确率（Accuracy）、正例覆盖率（Sensitivity）和负例覆盖率（Specificity）等指标来评估模型的性能。这三个指标越高，说明模型预测结果越准确，性能越好。

2．结果

表 5‑11 测试集混淆矩阵表

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际值 | | |
| 预测值 |  | 0（不后悔） | 1（后悔） |
| 0（不后悔） | 10 | 2 |
| 1（后悔） | 2 | 10 |
| 模型准确率 | 83% | | |
| 正例覆盖率 | 83% | | |
| 负例覆盖率 | 83% | | |

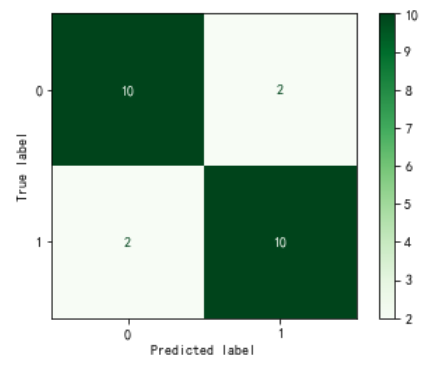


图 5‑2 测试集混淆矩阵可视化图

如上表5-11可知，就为模型简单的混淆矩阵，正确预测到正例在实际正例中占比83%，预测到负例在实际负例中同样也为83%，模型的整体准确率达到83%。总体来说，模型的预测准确率还是非常高的。如图5-2所示，通过将混淆矩阵转化为热力图，可以直观地展示不同类别之间的分类情况。在热力图中，颜色越深的区块代表该类别的样本数目越多。如果主对角线上的区块颜色比其他地方深很多，就说明模型在该类别上的正确预测数目很大。

**（二）ROC曲线**

1．理论

ROC曲线是一种通过绘制图形的方式来评估模型性能的指标。它利用两个度量值，横轴表示负样本错误率（即1-Specificity），纵轴表示正样本覆盖率（即Sensitivity），绘制出一条折线图。ROC曲线会考虑在不同阈值下，Sensitivity和1-Specificity之间的组合变化，从而描绘出一条曲线。评估模型性能时，计算ROC曲线下方的面积AUC也非常重要。通常情况下，当AUC在0.8以上时，模型的性能可以被认为是良好的。

2．结果

如下图5-3所示，在测试集上绘制了模型的ROC曲线，该曲线下方的面积为0.90，高于通常使用的评估标准0.8。因此，可以得出结论认为XGBoost模型的拟合是非常合理的，能够很好地反映数据的特征。

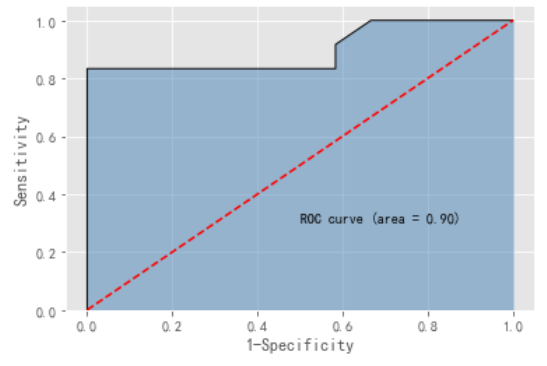


图 5‑3 测试集ROC曲线

**（三）分类报告**

表 5‑12 分类报告汇总表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.83 | 0.83 | 0.83 | 12 |
| 1 | 0.83 | 0.83 | 0.83 | 12 |
| accuracy |  |  | 0.83 | 24 |
| macroavg | 0.83 | 0.83 | 0.83 | 24 |
| weightedavg | 0.83 | 0.83 | 0.83 | 24 |

由上表5-12可知，各项指标均达到83%，达到了不错的分类评估结果。

五、XGBoost树状图解释

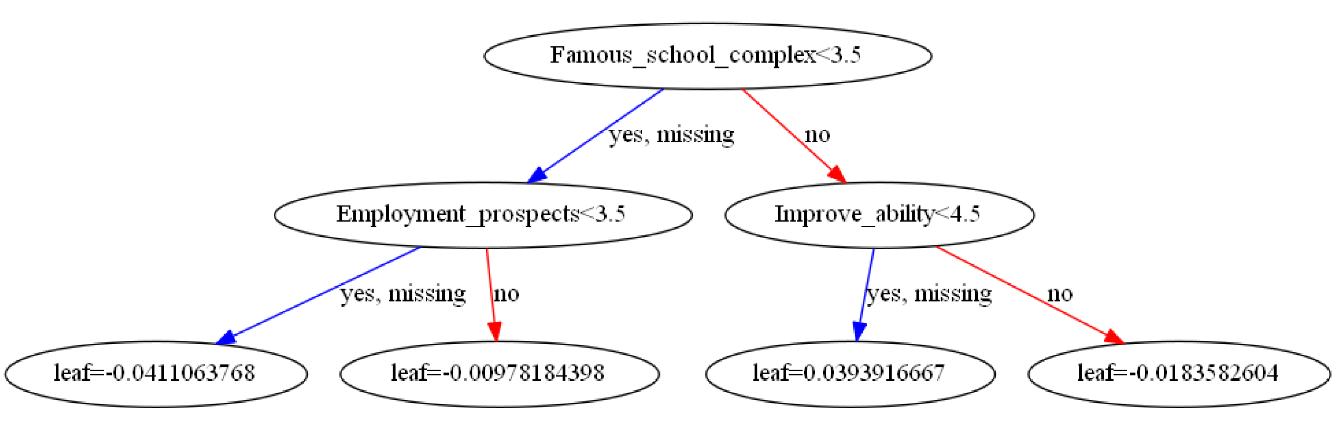


图 5‑4 XGBoost树状图

如上图5-4所示为XGBoost模型第50棵基础模型树的结构，其中根节点为“名校情结”，以阈值为3.5划分出左右子树；其中左子树为“看好专业未来就业发展前景”，以阈值为3.5继续划分左右叶子节点；右子树为“提高自身学历和能力”，以阈值为4.5继续划分出左右叶子节点，继而总共划分出4类叶子节点，leaf值为权重值，将多棵树权重求和后再做sigmoid得到的就是分类值结果。

六、模型的可解释性

对于模型可解释性，除了天生具有较好解释性的线性模型和决策树，sklearn中还有许多模型提供了importance接口，用于查看特征的重要性，这体现了模型解释性的理念。不过，传统的importance计算方法存在争议，且结果并不总是一致。

**（一）ELI5**

ELI5是一个能够帮助解释和调试机器学习分类器的Python包，其API统一且易于使用。除了提供内置支持多种机器学习框架的功能外，它还提供了一种解释黑盒模型的方法。相较于Yellowbrick，ELI5更注重于解释模型参数和预测结果，而Yellowbrick则更注重于特征和模型性能的解释。

1．特征重要性解释

在特征重要性图中，绿色的特征越靠近上方表示其对模型预测的影响越大。为了排除随机性，进行了多次洗牌并取结果的均值和标准差。某些特征可能会出现负值，这通常在特征重要性较低、数据集较小的情况下更为常见。 “+-”之后的数字反映了一次重新洗牌后的表现变化。一些特征重要性权重为负数，表示在随机重排这些特征后，模型的表现会更好。特征重要性图如下图5所示。

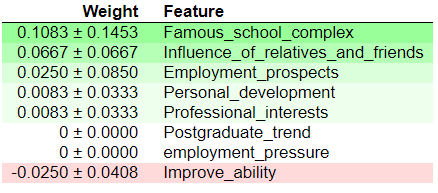


图 5‑5 ELI5特征重要性排序图

从图5-5可知，从整体上来看，“名校情结”重要性最大，其次是“父母和朋友的影响”，所以对于是否要读研的选择上，更应该关注这两方面；而“跟随读研大趋势”和“规避就业压力”反而对在读研上影响影响甚微，从而可以看出人们面对大环境下，特别是在疫情的影响下，呈现出更加理性的一面。

2．单样本预测结果

利用ELI5的show\_prediction函数可以对每个样本进行预测，并可视化出特征对预测值影响程度，按贡献值进行排序。

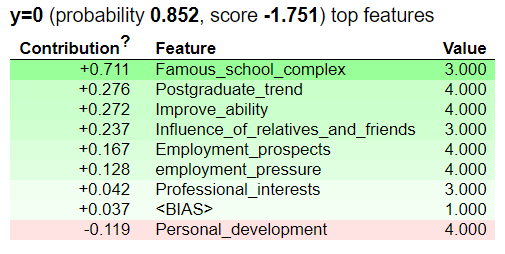


图 5‑6 第一个样本的预测结果

由图5-6可知，对第一个样本的预测结果影响最大的是“名校情结”，对第一个样本预测为“不后悔考研”，概率值达到85.2%。

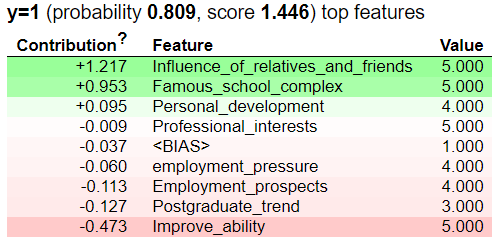


图 5‑7 第二个样本的预测结果

由图5-7可知，对第二个样本的预测结果影响最大的是“父母和朋友的影响”，其次是“名校情结”，对于第二个样本预测为“后悔考研”，概率值达到80.9%。

**（二）PDP(Partial Dependence Plot)**

部分依赖图（PDP）是一种可视化机器学习模型预测结果受特征影响的边际效应的图形。它可以帮助我们评估特征和目标之间的关系是否是线性、单调或更为复杂的。PDP显示了在固定其他特征的情况下，特定特征如何影响模型预测结果。在PDP图中，x轴表示该特征的取值范围，y轴表示该特征的边际影响值（通常为预测结果的平均值）。通过观察PDP图，我们可以了解该特征对于预测结果的影响趋势以及它与其他特征之间的相互作用。PDP图也被称为短PDP或PD图。

1．一维PDP

* y轴表示的是模型预测相较于基线值或最左边的值的变化。
* 蓝色阴影部分表示置信区间。

也就是说PDP在X1的值，就是把训练集中第一个变量换成X1之后，原模型预测出来的平均值，查看单个特征和目标值的关系。

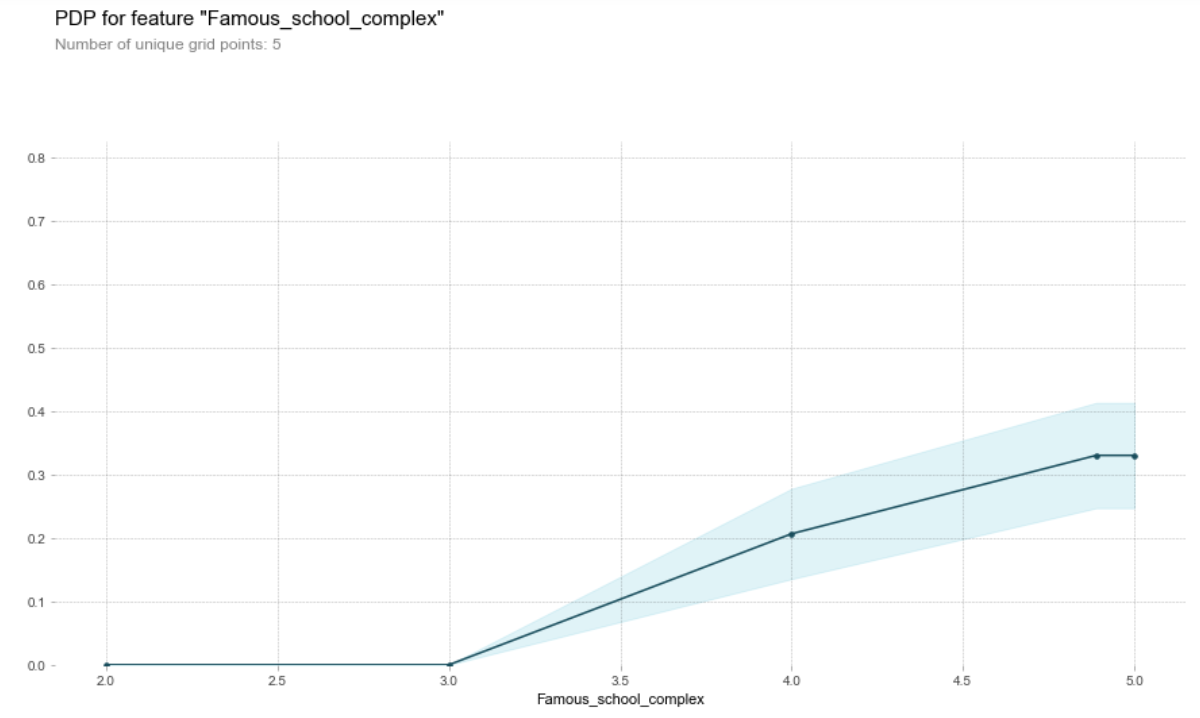


图 5‑8 名校情结PDP图

由图5-8可知，当“名校情结”的影响程度大于3的时候，开始对预测值产生影响，随者“名校情结”不断增加大，硕士对于“后悔考研”的概率也会逐渐加大，到接近5的时候放缓。

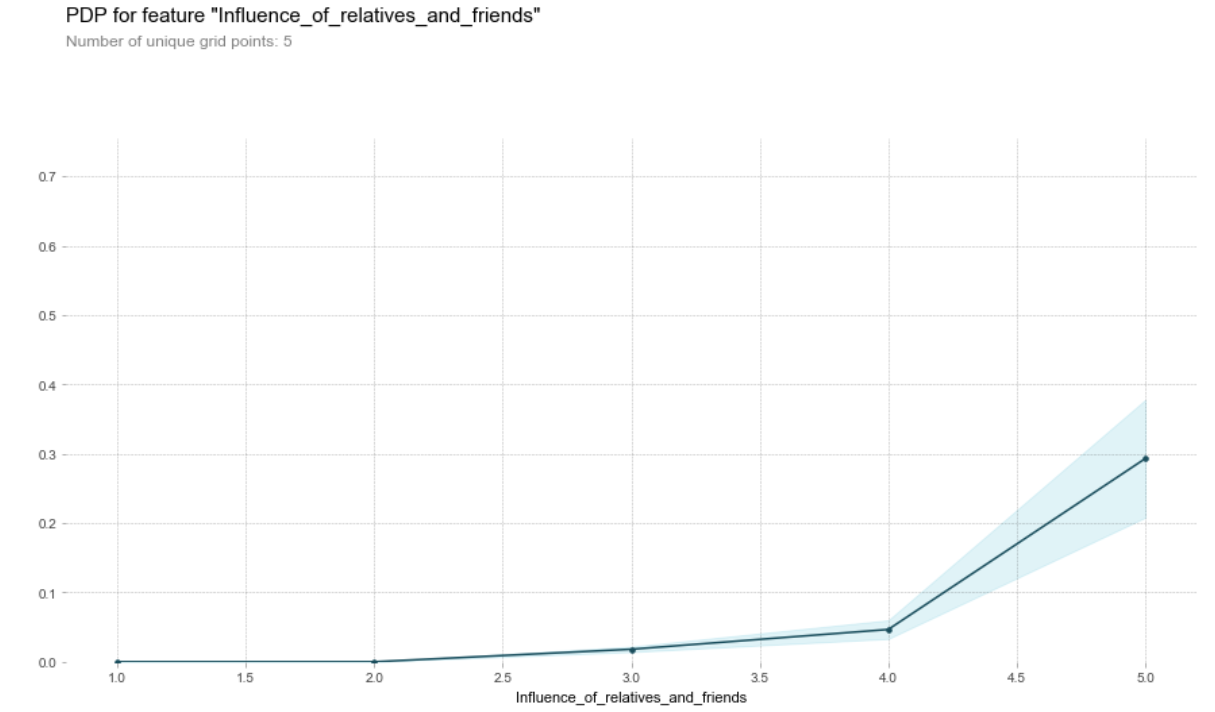


图 5‑9 父母和朋友的影响PDP图

由图5-9可知，当“父母和朋友的影响”大于2的时候，开始对预测值产生影响，从2到4之间以较缓慢的速度影响预测值，当大于4的时候，产生较大的影响，“后悔考研”的概率逐渐加大。

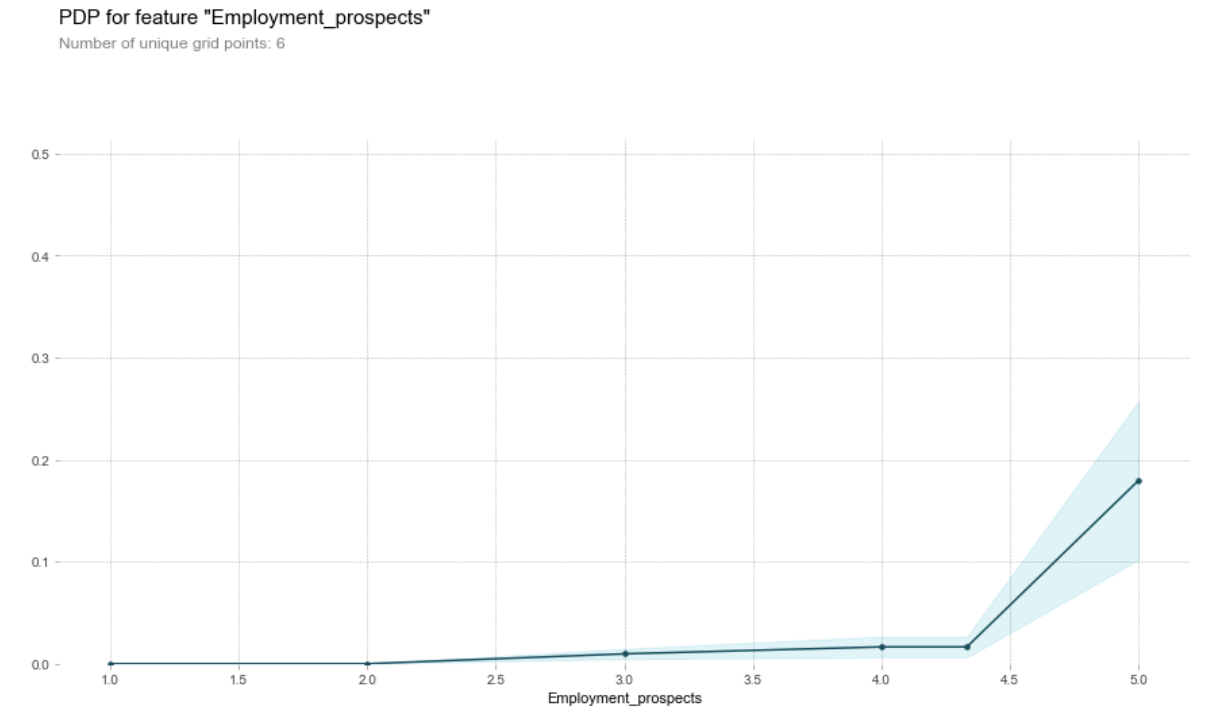


图 5‑10 看好专业未来就业发展前景PDP图

由图5-10可知，当“看好专业未来就业发展前景”大于2的时候，开始对预测值产生影响，从2到接近4.5之间以较缓慢的速度影响预测值，当大于4.5的时候，产生较大的影响，“后悔考研”的概率逐渐加大。

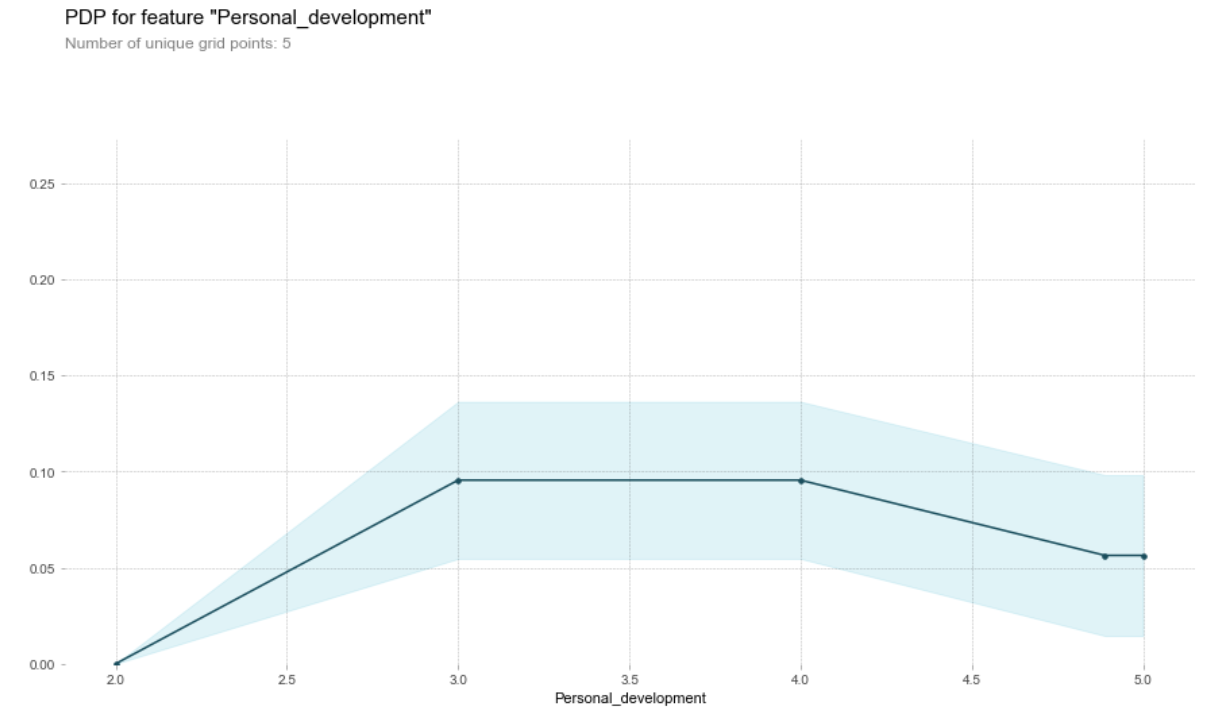


图 5‑11 考虑个人自身今后发展PDP图

由图5-11可知，当“考虑个人自身今后发展” 大于2的时候，直到等于3，开始对预测值产生较大影响，在3到4之间对预测值不产生影响，在4到5之间逐渐减少影响，降低了“后悔考研”的概率。

2．二维PDP

* x轴和y轴都表示两个变量影响程度
* 最右侧的颜色条表示的是模型预测相较于基线值或最左边的值的变化。

也就是说二维PDP图是可视化特征之间相互作用的结果。

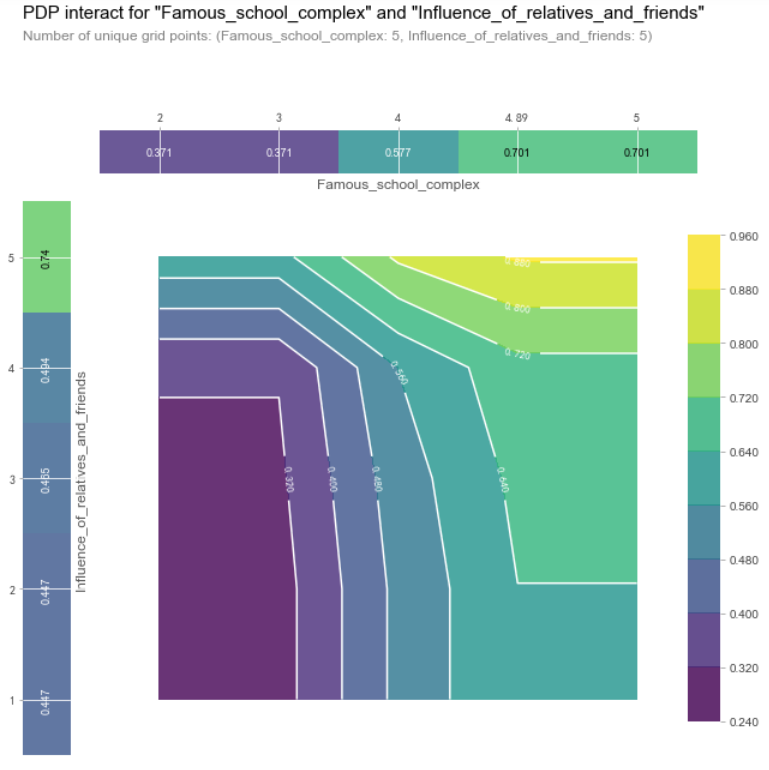


图 5‑12 名校情结和亲朋影响PDP图

由图5-12可知，“名校情结”和“父母和朋友的影响”都对“后悔考研”产生正向影响。其中“名校情结”影响占主导地位，当“名校情结”大于3的时候，对预测值的影响程度逐渐加大；而“父母和朋友的影响”在接近4的时候开始对预测值的影响程度缓慢加大；在两者的相互作用下，当“父母和朋友的影响”接近5，“名校情结”大于4.89之后，对预测值的影响程度达到最大，在[0.88，0.96]之间波动，即此时的硕士学历者更加“后悔读研”。

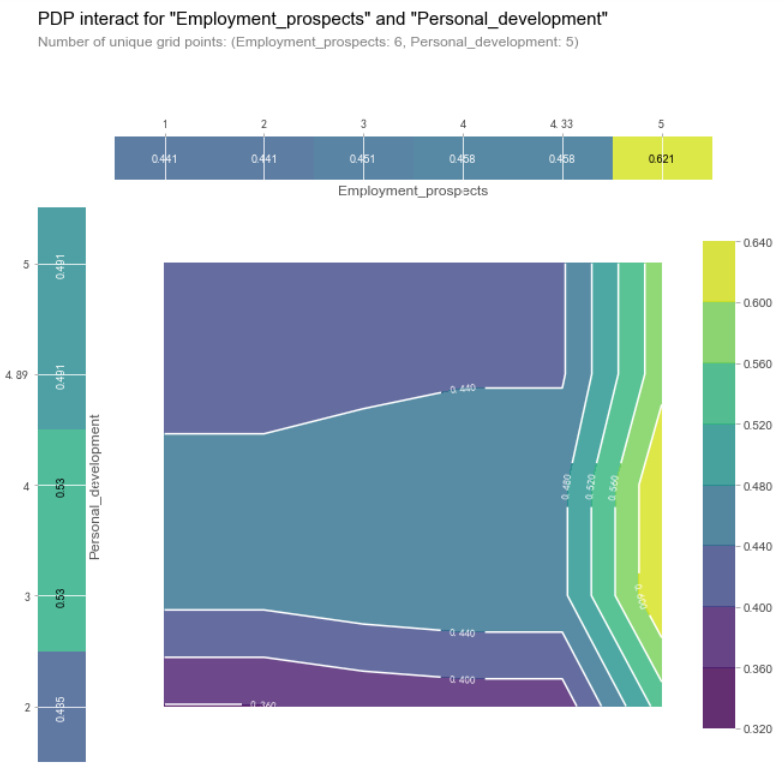


图 5‑13 就业前景和个人发展PDP图

由图5-13可知，“看好专业未来就业发展前景”对“后悔考研”产生正向影响，而“考虑个人自身今后发展”对“后悔考研”产生先增后减的影响。其中“看好专业未来就业发展前景”在4.33之前都趋于稳定的变化程度，当大于4.33的时候，对预测值的影响程度逐渐加大；而“考虑个人自身今后发展”在大于2的时候开始对预测值的影响程度缓慢加大，在[3，4.89]之间达到最大，并趋于稳定，随后在大于4.89的时候降低对预测值的影响程度；在两者的相互作用下，当“看好专业未来就业发展前景”接近5，“考虑个人自身今后发展”在[3，4]之间的时候，对预测值的影响程度达到最大，在[0.6，0.64]之间波动，即此时的硕士学历者更加“后悔读研”。

**（三）SHAP**

1．理论介绍

SHAP是一个用Python编写的“模型解释”工具包，其可解释任何机器学习模型的输出结果。SHAP的名称源自于SHapley Additive exPlanation。SHAP基于合作博弈理论的思想，构建了一个加性解释模型，将每个特征视为“贡献者”。对于每个预测样本，模型都会产生一个预测值，并将SHAP值分配给该样本中的每个特征，以反映特征对预测结果的贡献程度。因此，SHAP可以帮助我们理解机器学习模型如何进行决策，并揭示出每个特征的重要性。

假设第i个样本为，第i个样本的第j个特征为，模型对该样本的预测值为，整个模型的基线（通常是所有样本的目标变量的均值）为，那么SHAP value服从以下等式：

其中为的SHAP值。直观上看，就是第i个样本中第1个特征对最终预测值的贡献值，当>0，说明该特征提升了预测值，也正向作用；反之，说明该特征使得预测值降低，有反作用。

SHAP value最大的优势是SHAP能对于反映出每一个样本中的特征的影响力，而且还表现出影响的正负性。

2．Local Interper

Local可解释性提供了预测的详细信息，重点在于解释单个预测是如何生成的。它有助于让决策者信任模型，并解释每个特征如何影响模型做出单次决策。

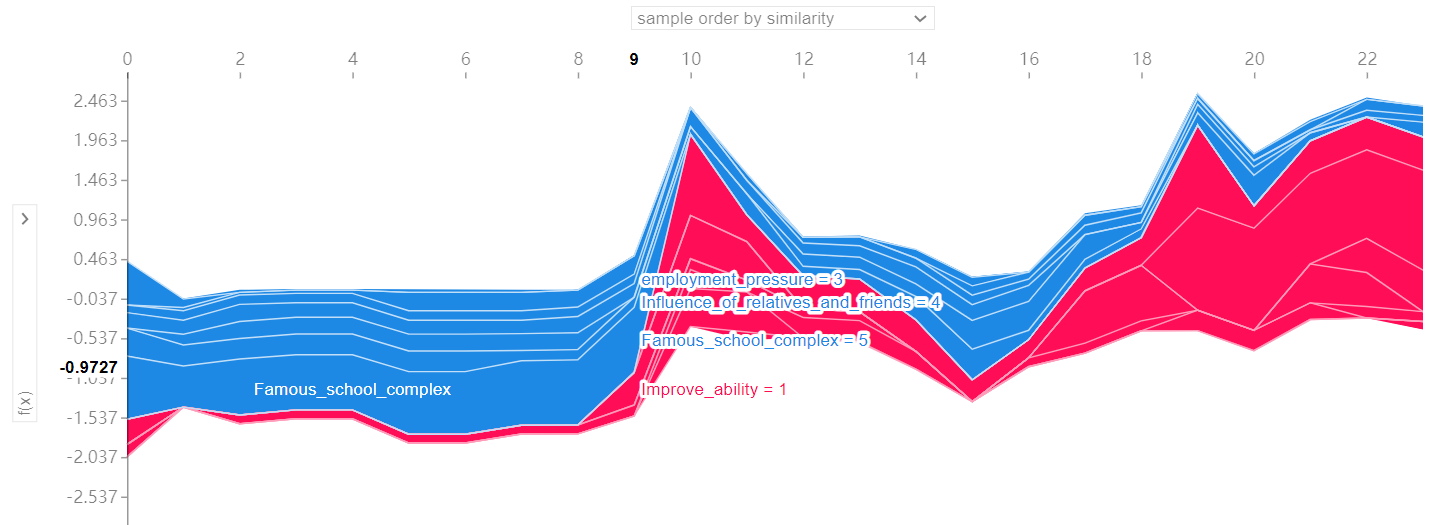


图 5‑14 多样本解释可视化

对于多个样本的解释，可以将单个样本的解释形式旋转90度并水平并排放置，得到力图的变体，从而能够看到整个数据集的解释结果。图5-14为交互图，是一种可交互的解释图形，通过颜色和长度的对比，可以清晰地看出每个样本对于“是否后悔考研”选择的影响以及其最大影响特征。这种方法可以更好地通过内在特征解释最终预测值。在上图中，可以通过上方和左方的选项卡任意选择单个变量的多个样本对模型输出结果的影响。

3．Global Interper

Global可解释性是指寻求理解模型的整体结构，而不仅仅是单个预测的解释。与局部可解释性不同，它涉及到对模型的一般工作原理作出说明，例如模型的决策边界、特征重要性排序、模型的复杂度等。全局可解释性可以帮助我们了解模型是如何使用特征来做出决策，从而更好地理解模型的预测结果和信任模型的决策。通常可以通过特征重要性分析、决策树可视化、模型复杂度分析等方式来实现全局可解释性。

summary plot（概要图）可以为每个特征绘制其SHAP值的总体分布，以帮助理解整体模式并发现异常值。它可以用来说明哪些特征对模型输出的影响最大，以及它们在数据集中的分布情况。

* 每一行代表一个特征，横坐标为SHAP值。
* 一个点代表一个样本，颜色表示特征值的高低(红色高，蓝色低)。

颜色使我们能够匹配特征值的变化如何影响风险的变化。重叠点在 y 轴方向抖动，因此我们可以了解每个特征的 Shapley value分布，并且这些特征是根据它们的重要性排序的。



图 5‑15 全局概要图

由图5-15可知，横向来看，以特征为依据，样本分布越为分散，那么代表该特征影响越大，故可以看到“名校情结”影响最大，这也与ELI5的特征重要性排序图相呼应；另外，发现特征“对所选专业感兴趣”，大多数的点弥漫在SHAP 为0附近，所以对大部分人都没啥影响，只对小部分人有影响，尽管此特征与ELI5的特征重要性排序图存在差异，但也给予我们从另外一个角度看待问题。总的来说，此全局概要图与ELI5的特征重要性排序图大体上一致。从各项特征来分析可以看出：

1. “名校情结”、“父母和朋友的影响”和“看好专业未来就业发展前景”的特征值与SHAP值呈现正相关。
2. “提高自身学历和能力”和“考虑个人自身今后发展” 的特征值与SHAP值呈现负相关。
3. “跟随读研大趋势”的特征值在很小的情况下对SHAP值的影响微乎其微，随着其特征值达到中等程度后，逐渐对SHAP值产生负向影响，而当特征值达到最大程度的时候，对SHAP值产生的影响最大，但是是正向影响，由此也可以看出如果盲目的跟随考研趋势，未来后悔读研的程度会越大。
4. “规避就业压力”的特征值在很小的情况下，对于SHAP值的影响是最大，正向影响，随着其特征值越来越大，对SHAP值的影响反而呈现负向影响，但是影响程度不大，从数据上看出，这与我们的的常规认知反而相反。

Interaction value（交互值）是指将SHAP值推广到更高阶交互的一种方法。在树模型中，可以通过快速而精确地计算两两交互来实现。计算将为每个预测返回一个矩阵，其中主要影响位于对角线上，交互影响位于对角线外。这些数值可以揭示有趣的隐藏关系。

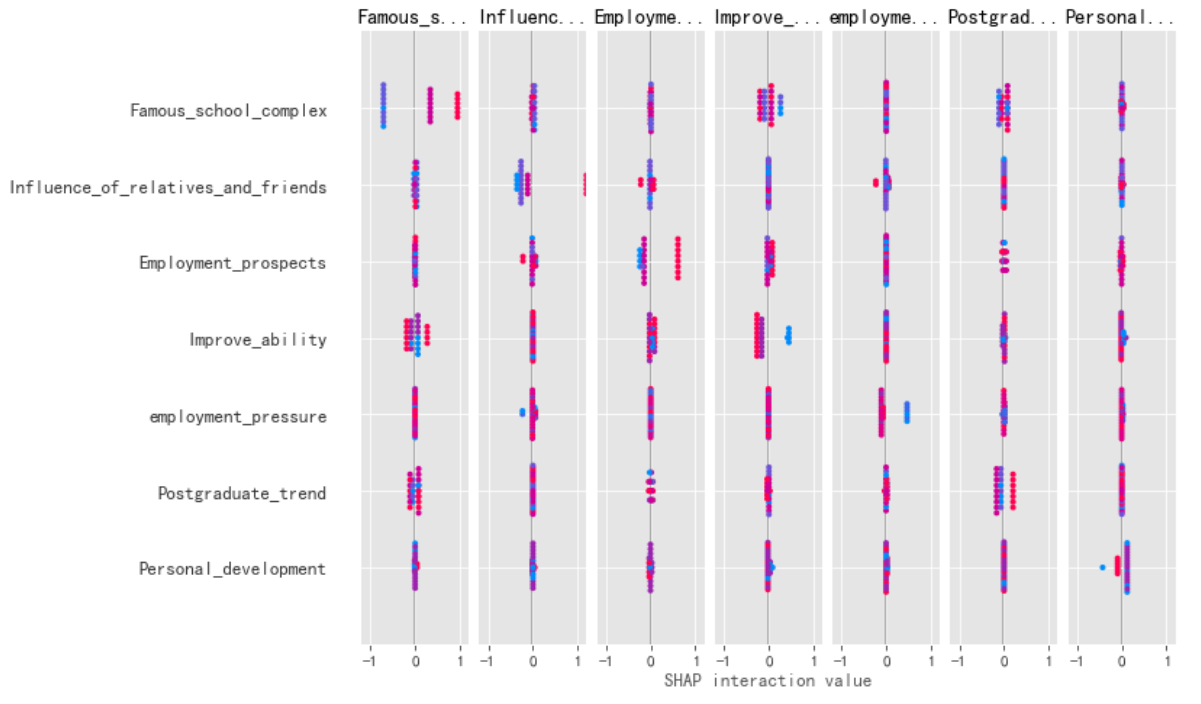


图 5‑16 特征交互图

由图5-16可知，主对角线可以看出各项特征对SHAP值的影响关系趋势，而对角线之外的，以上方横轴“名校情结”和左边纵轴“提升自身学历和能力”为例，在“名校情结”数值较小的情况下，不管“提升自身学历和能力”数值如何增加，对于SHAP值的影响始终是很小的，但随着“名校情结”逐渐增大，对SHAP值的影响也逐渐增大。