# 对外经济贸易大学

# 课 程 报 告

项 目 名 称 _____

课 程 名 称 _____

班　　级 _____

学　　号 _____

姓　　名 _____

二级学院 _____

专业名称 _____

指导教师 _____

2023 年 6 月

**摘要：**每年大家都会格外关注学生就业问题，影响学生就业的因素确实很多。

在本文中，我主要关注**学术**与**专业能力**这个方面，通过从三个不同的角度逐步深入挖掘一个就业数据集。首先，我使用了**皮尔逊假设检验**和**相关系数矩阵**，初步研究了各个特征与就业状况的关联性。接着，通过使用 lasso **回归**，过滤掉对就业状况影响微乎其微的特征。最后，我运用**支持向量机**、**随机森林**和 lasso 回归来进行对学生就业状况及薪资的分类与回归研究。

**关键词：皮尔逊假设检验，lasso 回归，支持向量机，随机森林**

# 目录

# 引言

许多之前的研究已经使用统计和机器学习的方法来探讨学生就业相关的应用。例如，有基于模糊决策树挖掘高校就业数据的案例，分析了影响学生就业的因素。而这次报告专注于分析学术与专业能力对学生就业的影响。

所有的分析和结论都基于由 Jain University 的 Dhimant Ganatara 博士提供的数据集。本文从**三个角度两个方向**层层递进地对该数据集进行挖掘。第一部分，通过使用**皮尔逊假设检验**对**离散变量**分析和**相关系数矩阵**对**连续变量**分析，初步研究了各个特征与就业状况的相关性。接着，通过使用 lasso 回归，过滤掉对就业状况影响微乎其微的特征。最后，通过使用**支持向量机、随机森林**对于**离散变量，lasso 回归和随机森林**对于**连续变量**，对学生就业状况及薪资进行分类与回归的研究。

| 数据标签 | 标签解释 |
|---|---|
| sl_no | 学生编号 |
| gener | 学生性别 |
| ssc_p | 中学课程均分 |
| ssc_b | 中学是中央直属 |
| hsc_p | 高中课程学习情况 |
| hsc_b | 高中为中央直属 |
| hsc_s | 高中的专业方向 |
| degree_p | 本科课程评价成绩 |
| workex | 有无工作经验 |
| etest_p | 大学能力测试 |
| specialisation | MBA 专业方向 |
| mba_p | MBA 平均成绩 |
| status | MBA 后就业情况 |
| salary | 薪水 |

表 1

# 1. 方法

## 1.1 数据集分析

在我们使用的数据集中，就业状况（status）和就职薪水（salary）是我们的目标特征，除此之外还有 12 种特征，其中包括 7 种离散特征和 5 种连续特征。我们将分别研究这 7 种离散特征和 5 种连续特征与就业状况（status）之间的相关性。

### 1.1.1 离散数据分析

图 1 分别展示了 7 种离散特征与就业状况的分布图，从图中难以直观看出各个离散特征是否与 status(就业状况)相关。因此，我们通过列出 7 种离散特征分别与 status(就业状况)的双向表，见表 2 至表 8,并采取皮尔逊卡方检验的方法，检验各个离散特征是否与 status(就业状况)相独立。

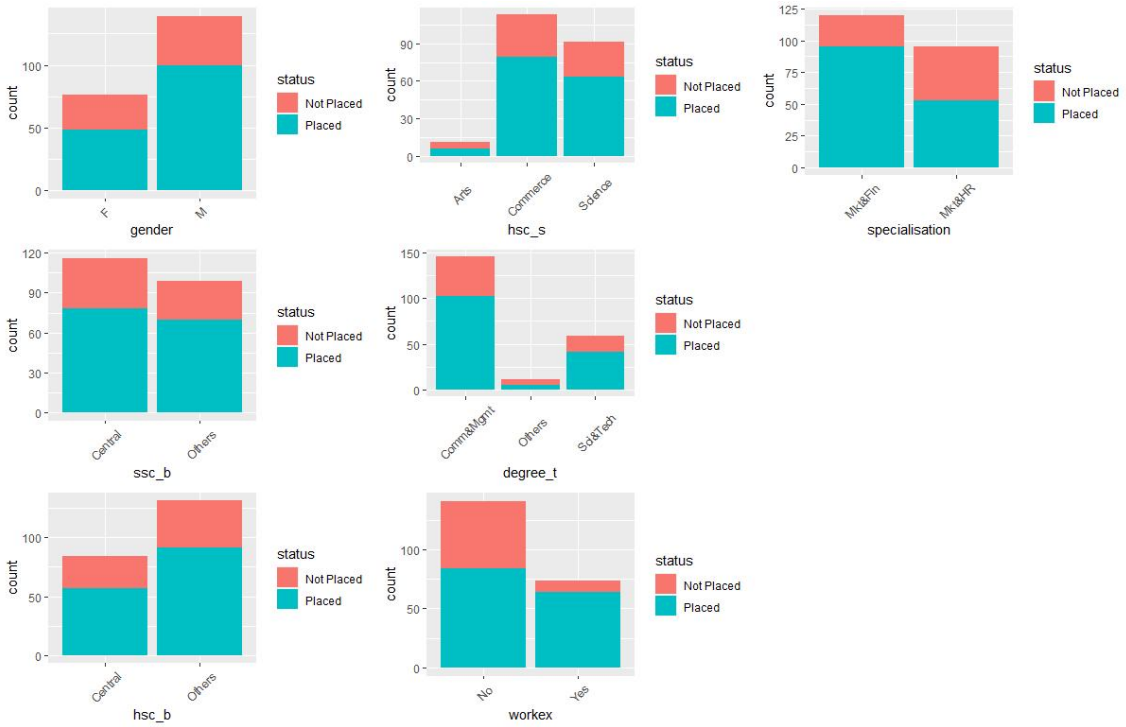


图 1：各个离散特征与就业状况分布图

表2: gender-status

| | Not Placed | Placed |
|---|---|---|
| F | 28 | 48 |
| M | 39 | 100 |

表3: ssc_b-status

| | Not Placed | Placed |
|---|---|---|
| Others | 29 | 70 |
| Central | 38 | 78 |

表4: hsc_b-status

| | Not Placed | Placed |
|---|---|---|
| Others | 40 | 91 |
| Central | 27 | 57 |

表5: workex-status

| | Not Placed | Placed |
|---|---|---|
| No | 57 | 84 |
| Yes | 10 | 64 |

表6: degree_t-status

| | Not Placed | Placed |
|---|---|---|
| Sci&Tech | 18 | 41 |
| Comm&Mgmt | 43 | 102 |
| Others | 6 | 5 |

表7: hsc_s-status

| | Not Placed | Placed |
|---|---|---|
| Commerce | 34 | 79 |
| Science | 28 | 63 |
| Others | 5 | 6 |

表8: specialisation-status

| | Not Placed | Placed |
|---|---|---|
| Mkt&Fin | 25 | 95 |
| Mkt&HR | 42 | 53 |

一共7 对假设检验，依次为

- $H_0$：性别(gender) 对就业状况(status) 无影响 vs $H_1$：性别对就业状况有影响
- $H_0$：中学教育委员会(ssc_b) 对就业状况无影响 vs $H_1$：中学教育委员会对就业状况有影响
- $H_0$：高中教育委员会 (hsc_b) 对就业状况无影响 vs $H_1$：高中教育委员会对就业状况有影响
- $H_0$：高中专业方向(hsc_s) 对就业状况无影响 vs $H_1$：高中专业方向对就业状况有影响
- $H_0$：学士学位领域(degree_t) 对就业状况无影响 vs $H_1$：学士学位领域对就业状况有影响
- $H_0$：有无工作经验(workex) 对就业状况无影响 vs $H_1$：有无工作经验对就业状况有影响
- $H_0$：MBA 毕业专业(specialisation) 对就业状况无影响 vs

  $H_1$：MBA 毕业专业对就业状况有影响

| **feature** | gender | ssc_b | hsc_b | hsc_s | degree_t | workex | specialisation |
|---|---|---|---|---|---|---|---|
| ***p*-value** | 0.2398 | 0.6898 | 0.9223 | 0.5727 | 0.2266 | 9.907e-05 | 4.202e-04 |

表9: 皮尔逊卡方检验结果

选用 p 值来表示假设检验的显著性，皮尔逊卡方检验的结果如表 9。设置信度为 95%，可得出以下结论：

•性别对就业状况无影响
•中学教育委员会对就业状况无影响
•高中教育委员会对就业状况无影响
•高中专业方向对就业状况无影响
•学士学位领域对就业状况无影响
•有无工作经验对就业状况有显著影响
•MBA 毕业专业对就业状况有显著影响

### 1.1.2 连续数据分析

讨论连续特征与就业状况的相关性相对容易一些，直接通过相关系数即可得出其相关性。图 2 给出了 5 个连续特征与 status、salary 的相关性。

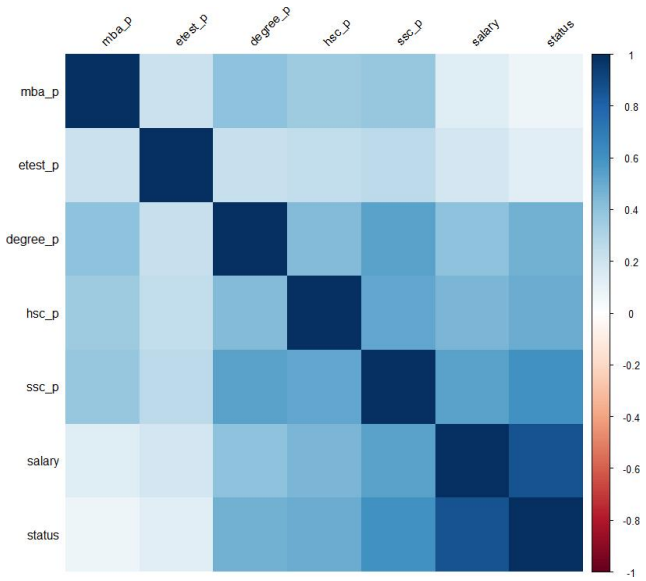


图 2：连续特征与就业状况的热力图

根据热力图可以得出：status、salary 与 5 个连续特征之间全部呈现正相关，但相关性不是非常强，其中 hsc_p(高中课程学习情况)，ssc_p(中学课程均分) 与就业状况的相关性相对较强。

# 2. 特征选择

在数据集中，有 12 种可能影响就业状况的特征。为了进行分类和回归分析，在对离散特征进行数值编码的基础上，我们使用 lasso 回归以提高模型的准确度。通过前面的假设检验和相关性探究，我们发现有几个特征对就业状况几乎没有影响。

在进行 lasso 回归之前，需要选择适当的 λ 值。我们采用了 10 折交叉验证法，并以 AUC 值作为评估指标。在不同的 λ 取值情况下，lasso 回归的 AUC 值变化如图 3 所示，其中上轴坐标值代表选取的特征数。可以观察到在 λ 取第一条虚线处的值时，模型的 AUC 值最大。此时选取的特征数为 8，过滤掉了 5 种特征。

更详细的分析表明，选取的 8 种特征为："gender"，"ssc_p"，"hsc_p"，"degree_p"，"mba_p"，"workex"，"specialisation"。反过来说，中学教育委员会、高中教育委员会、高中专业方向、学士学位、MBA 成绩、工作经验、专业方向等 8 种因素对就业状况的影响较为显著。因此，我们将这 5 种特征去除，仅使用剩余的 8 种特征进行就业状况的分类与回归分析。这有助于提高模型的精确度并更好地理解特征对学生就业的影响

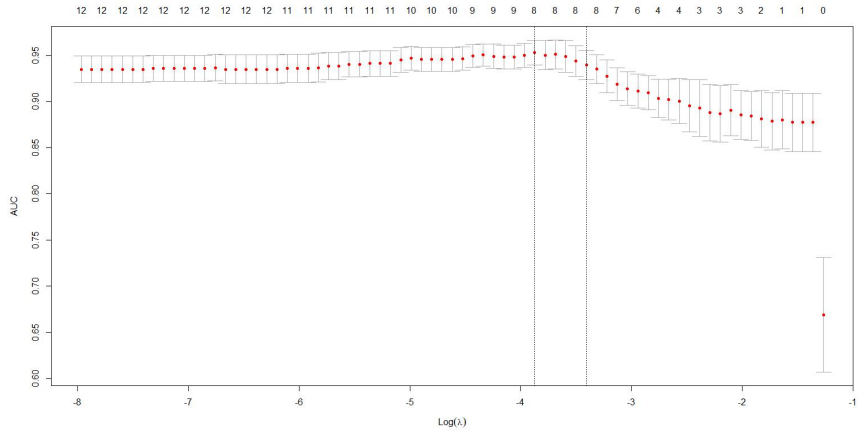


图3: AUC-λ

# 3. 就业情况分类和回归

在分类和回归任务方面，首先在分类任务上，支持向量机的表现似乎不如随机森林，这也符合当前数据挖掘竞赛中出色队伍的趋势。目前，许多表现优异的团队主要采用集成方法和神经网络，而支持向量机的表现相对较差。尽管如此，支持向量机所采用的统计方法和思维仍然具有很大的学习价值。

其次，在回归任务中，无论是 lasso 回归还是随机森林的表现都与预期相去甚远。在研究过程中，尝试了一些方法来缩小误差，如使用全部特征、特征归一化、根据已有特征构建新的特征以提高模型复杂度，以及尝试使用其他集成算法等。然而，这些尝试的效果并不明显，唯一可行的方法似乎是搜集新的与就业相关的特征。

这也强调了仅仅依靠学生的学术和专业能力来预测其薪资是不切实际的。预测学生的薪资需要更多的维度和多样化的特征，可能包括一些与实际工作经验、行业需求以及其他非学术因素相关的信息。这一发现为未来改进模型提供了启示，需要更广泛地考虑各种与就业相关的因素。

## 3.1 离散数据对就业的回归分析

首先，我们要解决的问题是预测学生的就业状况，即已就业或未就业，这是一个二元分类任务。我们选择了上一部分筛选出的 8 种特征作为模型的输入，而就业状况则是我们的目标变量。

为了解决这个分类问题，我们采用了支持向量机和随机森林两种不同的方法，并在效率和准确率上进行了比较。我们将整个数据集按 3:1 的比例随机划分为训练集和测试集。然后，我们分别在训练集上进行算法训练，再在测试集上进行算法评估。

在算法方面，支持向量机使用了高斯核函数，其中超参数 $\gamma$ 被设定为数据集特征个数的倒数，即 1/9；而随机森林使用了 100 棵子树，而且没有设置树的高度限制。我们以准确度（Accuracy）、查全率（Recall）和查准率（Precision）这三个指标来评估算法在测试集上的性能，具体结果见表 10。

总体来说，两种分类算法在测试集上都表现得非常不错，但随机森林整体上要优于支持向量机。这也意味着通过选取的 8 个特征构建的模型足以用于判断学生是否能够就业。

| | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| 支持向量机 | 0.811 | 0.846 | 0.892 |
| 随机森林 | 0.926 | 0.902 | 1.000 |

表10: 分类任务：随机森林与支持向量机结果对比

## 3.2 连续数据对就业的回归分析

其次，我们将目光转向了学生的就业薪水（salary），这是一个连续变量。同样地，我们以就业薪水作为目标变量，选取了上一部分筛选出的 8 种特征，然后将数据集按 3:1 的比例随机划分为训练集和测试集，构建一个回归任务。

在这个回归任务中，我们选择了使用 lasso 回归和随机森林两种方法。根据之前的特征选择经验，lasso 回归使用的λ值为 1e-4；而随机森林使用了 100 棵子树，没有设置树的高度限制。我们以均方误差（MSE）来评估算法在测试集上的表现，具体结果见表 11。

然而，从结果可以看出，无论是 lasso 回归还是随机森林在测试集上的表现都不够理想。一个合理的解释可能是这 8 种特征不足以有效地预测学生的就业薪水。要降低预测误差，一个可行的方法是增加新的与学生就业相关的特征，以提高模型的复杂度。这可以为未来的研究提供一个方向，以改进对学生就业薪水的准确预测。

| | MSE |
|---|---|
| lasso 回归 | 6094313921 |
| 随机森林 | 6129612493 |

表11: 回归任务：lasso 回归与随机森林结果对比

# 4. 总结

在第一个特征相关性研究方面，我们通过皮尔逊卡方检验来分析每个特征是否对学生的就业状况产生影响。在进行了 7 组假设检验后，发现只有两组拒绝了原假设，而且显著性水平相当高，因此我们可以断定这两个特征对学生的就业状况有显著影响。然而，由于假设检验有对原假设的偏好，不能确定其他 5 个特征是否对就业状况无影响。因此，我们后续采用 lasso 回归进行特征抽取，以进一步探究相关性。

在第二个特征选择方面，除了使用 lasso 回归之外，我们还尝试了递归特征消除等其他特征抽取的方法，但结果相似。因此，在论文的展示方面，我们主要以 lasso 回归为主。通过对学生就业相关的数据集的研究，我们得出了以下三个结论：

(i) 利用皮尔逊卡方检验发现，有无工作经验和 MBA 毕业专业对学生的就业状况有显著影响。

(ii) 利用 lasso 回归的特征过滤性质，我们发现中学教育委员会、高中教育委员会、高中专业方向、学士学位以及就业能力测试这五个因素对学生的就业状况影响微乎其微。

(iii) 根据学术和专业能力，我们能够判断学生是否能找到工作，但却无法准确估计其薪资水平。要减小估计的误差，需要考虑其他因素的影响。