学生就业影响因素的研究————学术与专业能力方面

孙文强 181830166　　姜玉骅 181870080　　党子宸 181830037

摘　要

每年的学生就业问题都会引起社会广泛的重视，而影响学生就业的因素有很多，本篇文章着重在学术与专业能力方面研究，从三个角度层层递进对一个就业数据集进行挖掘，首先利用皮尔逊假设检验和相关系数矩阵初探各个特征与就业状况的相关性，接着利用 lasso 回归过滤掉对就业状况影响甚微的特征，最后利用支持向量机、随机森林、lasso 回归对学生就业状况及薪资进行分类与回归的研究。分析发现，根据学术水平能够判断一个学生是否就业，但难以估计其薪资。

关键词: 皮尔逊卡方检验，随机森林，lasso 回归，支持向量机，学生就业

1　介绍

利用统计、机器学习的方法探讨学生就业相关的应用已有许多先前的案例。如基于模糊决策树挖掘高校就业数据，分析影响学生就业的因素；如利用分析数据库挖掘学生就业影响因素的关联规则；影响学生就业的因素也有许多，涉及政府、高校、用人单位、学生自身因素等多个方面，本文着重分析学术水平对学生就业的影响。

本文的所有分析与结论都是基于由 Jain University 的 Dhimant Ganatara 博士提供的数据集[1]，该数据集是关于 Jain University 学生的就业数据。它包括中等和高等中等学校的百分比和专业。还包括学位专业，类型，工作经验和已获得工作的薪资。数据集特征的详细解释见表1。

我们从三个角度层层递进对该数据集进行挖掘，首先利用皮尔逊假设检验和相关系数矩阵初探各个特征与就业状况的相关性，接着利用 lasso 回归过滤掉对就业状况影响甚微的特征，最后利用支持向量机、随机森林、lasso 回归对学生就业状况及薪资进行分类与回归的研究。

本篇论文的结构如下：第二部分，将我们简要阐释我们采取的方法的理论原理，如皮尔逊卡方检验、支持向量机、lasso 回归、随机森林。第三部分，应用我们的方法，从三个角度详细的阐释我们的分析过程。第四部分，对我们的方法进行讨论、补充说明。第五部分，总结本文得到的结论。

[1]数据集地址https://www.kaggle.com/benroshan/factors-affecting-campus-placement

数据标签	标签解释
sl_no(序列号)	指学生编号
gender(性别)	学生性别
ssc_p(中学教育程度-十年级)	中学课程均分
ssc_b(中学教育委员会)	中学是中央直属或者地方归属
hsc_p(高中教育百分比-十二年级)	指高中课程学习状况
hsc_b(高中教育委员会)	高中为中央直属或者地方归属
hsc_s(高中专业化)	高中的专业方向
degree_p(学位百分比)	本科课程平均成绩
degree_t(学士学位领域)	本科毕业学位类别
workex(工作经验)	有无工作经验
etest_p(就业能力测试)	由大学进行的就业能力测评
specialisation(MBA 毕业专业)	获得 MBA 学位的专业方向
mba_p(工商管理硕士百分比)	读 MBA 课程平均成绩
status(就业状况)	获取 MBA 学位后就业状况
salary(就职薪水)	就业薪水

表 1: Dataset

2 基本理论

2.1 皮尔逊卡方检验

皮尔逊卡方检验可用于两种情境：适配度检验和独立性检验。适配度检验是验证一组观察值的次数分配是否异于理论上的分配，独立性检验是验证从两个变量抽出的配对观察值是否相互独立。本文主要利用独立性检验。

令一个个体有两个变量分别是 r 元和 c 元变量，研究两个变量的相关性。原假设是两个变量相互独立。首先，将该个体的两个变量编排到双向表或者说列联表中 (contingency table)。则列联表共有 r 行，c 列，在两个变量相互独立的假设下，每个字段的理论次数为：

$$E_{i,j} = \frac{\left(\sum_{n_c=1}^{c} O_{i,n_e}\right) \cdot \left(\sum_{n_r=1}^{r} O_{n_r,j}\right)}{N} \tag{1}$$

其中 N 是样本大小，O_{ij} 是各个字段的实际次数。皮尔逊卡方统计值的公式是：

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{2}$$

卡方的自由度是 $df = (r-1)(c-1)$。根据卡方的自由度可得到卡方检验的拒绝域，进一步根据皮尔逊卡方统计值即可判断是否拒绝原假设。

2.2 lasso 回归

lasso 回归是最早为应用最小二乘法而定义的算法，它在基本的最小二乘法的基础上加上了 lasso 正则化，即 l_1 范数，它的目标方程可表示为以下形式：

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{N}\|y-X\beta\|_2^2+\lambda\|\beta\|_1\right\} \tag{3}$$

其中 X 为输入变量，y 为预测变量，β 为回归系数，N 为样本大小。目标方程还可以写为：

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{N}\|y-X\beta\|_2^2\right\} \text{ subject to } \|\beta\|_1\leq t \tag{4}$$

lasso 回归通过强制让回归系数的绝对值之和小于某固定值，即强制一些回归系数为 0，有效的选择了与预测变量最相关的输入变量，使模型变得更简单。因此 lasso 回归在线性拟合的用途之外，还拥有过滤特征的性质。

2.3 支持向量机

支持向量机最初是用于二元、线性、可分的分类问题，它的目标是找到一个" 最大间隔" 的划分超平面，所谓" 间隔"，即样本点到划分超平面的距离。支持向量机的优化目标即：最大化离划分超平面最近的样本点到超平面的距离。其目标函数的原始型表示为：

$$\begin{aligned}&\min_{w,b}\quad \tfrac{1}{2}\|w\|^2\\&\text{s.t.}\qquad y_i(w\cdot x_i+b)-1\geqslant 0,\quad i=1,2,\cdots,N\end{aligned} \tag{5}$$

这是一个凸二次规划问题。利用拉格朗日乘子法可得到目标函数的对偶型：

$$\begin{aligned}&\max_{\alpha}-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)+\sum_{i=1}^{N}\alpha_i\\&\text{s.t. }\sum_{i=1}^{N}\alpha_i y_i=0\\&\alpha_i\geqslant 0,\quad i=1,2,\cdots,N\end{aligned} \tag{6}$$

其中 α_i 为拉格朗日乘子。目前已有许多高效算法求解上述问题，如最典型的 SMO(Sequential Minimal Optimization) 算法。

支持向量机可以进一步推广，解决各类各样的分类问题。如针对多元分类问题，可以将多分类任务拆分二分类任务，最经典的拆分策略 One vs.One、One vs.Rest、Many vs.Many。针对非线性问题，可以利用核函数将原始样本空间映射到高维空间，使其在高维空间中是一个线性问题。针对不可分问题，引入软间隔的思想，允许一些样本不符合划分超平面的约束。

2.4 随机森林

随机森林是并行式集成学习方法的代表，在 Bagging 方法的思想基础上加以改进。先来简要说明一下 Bagging 方法。

Bagging 利用了 bootstrap 的思想，给定包含 m 个样本的数据集，我们先随机取出一个样本放入采样集，再把该样本放回初始数据集，使得下次采样样本仍然可能被选中。经过 m 次这样的随机采样操作，就得到了含 m 个样本的采样集。照这样，可采样出 T 个含 m 个样本的采样集，基于每个采样集训练处一个基学习器，再将这些基学习器结合起来，分类任务使用简单投票法结合，回归任务使用简单平均法结合，这就是 Bagging 的基本流程。

随机森林在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入了随机属性的选择。具体来说，传统决策树在选择划分属性时是在当前结点的属性集合中选择一个最优属性，而在随机森林的每个基决策树中，先从该结点的属性集合随机选择一个子集，然后再从子集中选择一个最优属性用于划分。

随机森林简单、容易实现、计算开销小，在许多现实任务中展现出强大的性能，被誉为" 代表集成学习技术水平的方法"。

3 方法

3.1 特征相关性初探

本文使用的数据集中，status(就业状况)、salary(就职薪水) 是我们的目标特征，除此之外共有 12 种特征，其中离散特征 7 种，连续特征 5 种。我们将分别探究这 7 种离散特征、5 种连续特征与 status(就业状况) 的相关性。

3.1.1 离散特征

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。

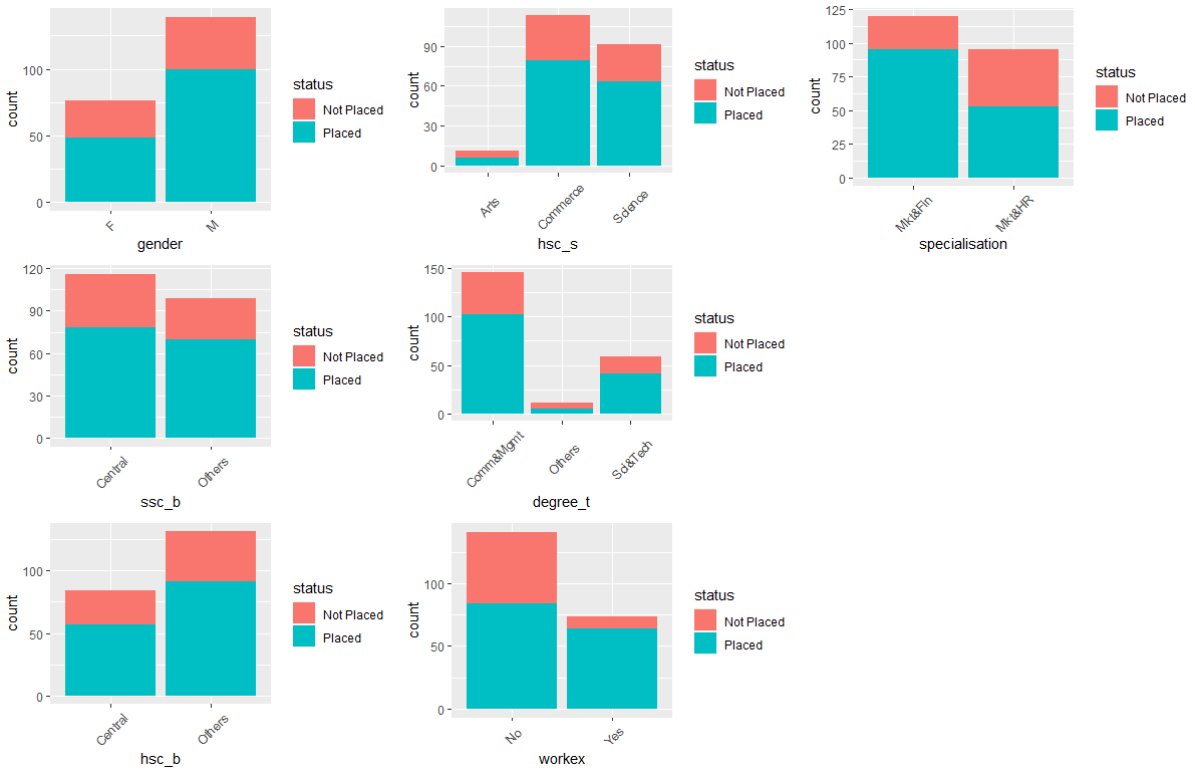

图 1: 各个离散特征与就业状况分布图

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

3.1.2 连续特征

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。

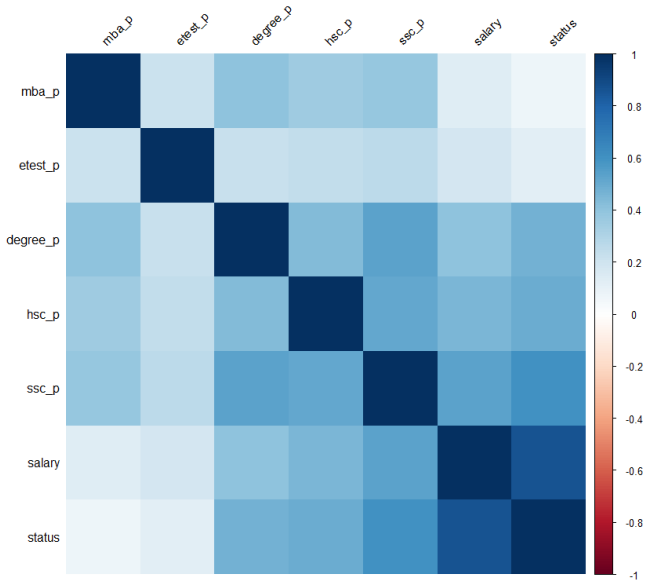

图 2: 连续特征与就业状况的热力图

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。

3.2 特征选择

数据集中可能影响就业状况的特征共有 12 种，用数值编码所有的离散特征，即用从 0 开始的数字表示每个单独的类别。根据前面假设检验与相关性的探究发现有若干特征对就业状况几乎没有影响，因此在分类、回归之前，我们以 status 为因变量，利用 lasso 回归去除无关的特征，以提高模型的准确度。

使用 lasso 回归需要选取适当的 λ 值，我们使用 10 折-交叉验证法，并以 *AUC* 值作为评估指标。不同的 λ 取值情况下，lasso 回归的 AUC 值变化见图3，其中上轴坐标值代表选取的特征数。可

以看出在 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5种特征。更细致的分析，发现选取的8种特征为："gender","ssc_p","hsc_p","degree_p","mba_p","workex","specialisation",反过来说，即中学教育委员会、高中教育委员会、高中专业方向、学士学位、就业能力测试5种因素对就业状况的影响甚微。因此将这5种特征去除，仅用剩余的8种特征用于就业状况的分类与回归。

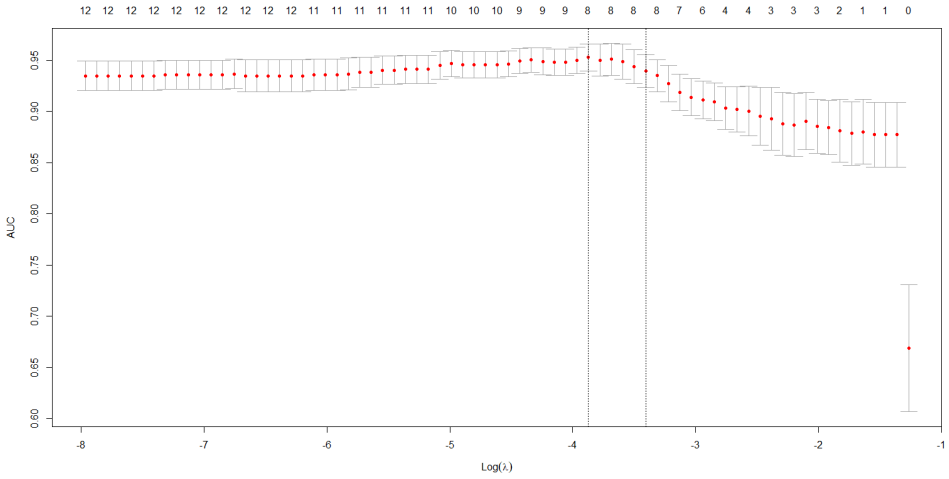

图 3: AUC-λ

3.3 就业状况分类与回归

首先，status(就业状况) 是一个二值变量，已经就业或未就业，我们以此作为目标变量，以上节筛选出的 8 种特征为特征变量，构建一个分类任务。为解决该分类任务，我们采取支持向量机和随机森林两种方法进行效率、准确率上的比较。我们将数据集以 3:1 的比例随机划分为训练集和测试集。算法在训练集上训练、测试集上评估。算法方面，支持向量机使用高斯核函数，超参数 γ 为数据集特征个数的倒数即 1/9；随机森林使用 100 棵子树，不设置树高度限制。以准确度 (Accuracy)、查全率 (Recall)、查准率 (Precision) 三个指标评估算法在测试集上的表现，结果见表10。可以看出两种分类算法性能都非常不错，但随机森林整体上好于支持向量机。另一

	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

方面，也说明了选取的 8 个特征构建的模型用于判断学生是否能够就业已经足够。

其次，salary(就业薪水) 是一个连续变量，我们同样以此作为目标变量，同样采取上节筛选的 8 种特征，将数据集以 3:1 比例随机划分为训练集和测试集，构建一个回归任务。我们采取 lasso 回归与随机森林两种方法进行执行该任务。根据上节抽取特征的经验，lasso 回归使用的 λ 值为 1e-4；随机森林使用 100 棵子树，不设置树高度限制。以均方误差 (MSE) 评估算法在测试

集上的表现，结果见表11。可以看出无论是 lasso 回归还是随机森林在测试集上的表现不符理想，合理的解释是这 8 种特征不足以预测学生的就业薪水，若要缩小预测的误差，可行的方法是增加新的有关学生就业特征，以提高模型的复杂度。

	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 讨论

我们主要从三个方面分析了整个数据集。

第一个特征相关性研究方面，通过皮尔逊卡方检验分析某个特征是否对就业状况有影响。7 对假设检验中，只有两对拒绝了原假设，并且显著性相当高，可以断定这两个特征对就业状况有显著影响。反过来由于假设检验对原假设偏好性，并不能断定着其它 5 个特征对就业状况无影响。这也是后续我们继续采用 lasso 回归抽取特征，探究相关性的原因。

第二个特征选择方面，除了采用 lasso 回归以外，我们也尝试了递归特征消除 (Recursive Feature Elimination) 等抽取特征的方法，结果及其相似，因此论文展示方面以 lasso 回归为主。

第三个分类与回归方面，首先分类任务上，支持向量机的表现不如随机森林，这似乎也是大势所趋。目前各大数据挖掘竞赛表现出色队伍所采取的方法基本以集成方法和神经网络为主，支持向量机的表现还是差强人意，但支持向量机所应用的统计方法和思维非常值得我们学习；其次回归任务中，无论是 lasso 回归还是随机森林的表现与预期相差甚多，研究过程中也试图通过其他方法缩小误差，如使用全部特征、特征归一化。根据已有特征构建新的特征提高模型复杂度、使用其它集成算法等，但收效甚微，唯一可行方法可能就是搜集新的与就业相关的特征，同时说明了仅仅依靠学生学术和专业能力预测其薪资是不切实际的。

5 结论

本篇文章根据该学生就业相关的数据集，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、MBA 毕业专业对就业状况有显著影响。(ii) : 利用 lasso 回归具有过滤特征的性质，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、就业能力测试 5 种因素对就业状况的影响甚微。(iii) : 根据学术和专业能力我们可以判断学生能不能找到工作，但无法准确估计其薪资，若要减小估计的误差还需将其它因素考虑进来。

6 参考文献

- 周志华. 机器学习 [M]. Qing hua da xue chu ban she, 2016.
- 盛骤, 谢式千, 潘承毅. 概率论与数理统计 [J]. 教育出版社, 2008.