**2023-2024学年第2学期**

**期末大作业**

**学生成绩因素分析**

**课程名称： 商业数据分析**

**任课教师： 戴杨 丁月**

**姓 名： 彭榆丰**

**学 号： 2021116286**

评审日期\_\_\_\_\_\_\_\_\_\_ 成绩\_\_\_\_\_\_\_ \_\_ 评审教师(签名)\_\_\_\_\_\_\_\_ \_\_

教师评语：

目录

[一 数据集介绍 1](#_Toc170570968)

[二 数据探索与数据可视化 2](#_Toc170570969)

[三 回归分析 6](#_Toc170570970)

[1线性回归模型 6](#_Toc170570971)

[2逻辑回归模型 8](#_Toc170570972)

[四 聚类分析 10](#_Toc170570973)

[五 结论以及建议 11](#_Toc170570974)

**学生成绩因素分析**

学生的考试成绩一直都被重点关注，除去先天的智力因素，人们往往将考试成绩与平时的努力画上等号，如果孩子成绩不理想，很多父母更趋向于认为孩子不够努力，而孩子总是抱怨父母给的压力太重，感觉得不到父母的关心与支持，从而孩子与父母的关系变得渐渐疏远起来，这是唯分数论带来的坏处。那么事实是怎样呢，本报告将通过机器学习模型对学生成绩进行预测，得出影响学生分数的各种各种因素。

# 一 数据集介绍

本报告所选数据集下载自kaggle。数据集包含美国 2,392 名高中生的全面信息，详细介绍了他们的人口信息、学习习惯、父母参与、课外活动和学业成绩。目标变量成绩等级将学生的成绩分为不同的类别。具体变量说明见表1-1所示。

表 1‑1：数据字段说明表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 字段类型 | 字段描述 | 取值范围 |
| Student id | 连续性数据 | 学号 | [1001,3392] |
| Age | 分类型数据 | 学生年龄 | [15,18] |
| Gender | 分类型数据 | 学生性别 | 0:男 1:女 |
| Ethinicity | 分类型数据 | 学生种族 | 0:白种人1:非裔美籍 2:亚洲人  3:其他 |
| ParentalEducation | 分类型数据 | 父母受教育程度 | 0:无 1:高中 2:专科院校 3:本科  4:更高 |
| StudyTimeWeekly | 连续型数据 | 每周学习时长 | [0,20] |
| Absences | 连续型数据 | 每学期缺席次数 | [0,30] |
| Tutoring | 0-1数据 | 是否有家教 | 0:无 1:有 |
| ParentalSupport | 0-1数据 | 父母支持程度 | [0,4] |
| Extracurricular | 0-1数据 | 是否参与课外活动 | 0:no 1:yes |
| sports | 0-1数据 | 是否参与体育活动 | 0:no 1:yes |
| Music | 0-1数据 | 是否参与音乐活动 | 0:no 1:yes |
| Volunteering | 0-1数据 | 是否参与志愿活动 | 0:no 1:yes |
| GPA | 连续型数据 | 成绩 | [0,4]数值越高成绩越好 |
| Gradeclass | 分类型数据 | 成绩等级 | [0,4]数值越低成绩越好 |

# 二 数据探索与数据可视化

对数据集先进行描述性的统计，见表2-1,从数据的统计结果可以看出美国高中生每周学习的平均时长不足10小时，最大值也只有20小时，大大小于我们国家的高中生，而缺课次数的均值达到了14次，从这两个方面可以看出，与我国高中生相比，美国高中对成绩的重视程度较低。在课外活动方面，美国高中生更喜欢参加体育活动。

表2-1：数据集描述性统计

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | Gender | Ethnicity | ParentalEducation | StudyTimeWeekly | Absences | Tutoring |
| count | 2392 | 2392 | 2392 | 2392 | 2392 | 2392 | 2392 |
| mean | 16.46865 | 0.51086957 | 0.877508361 | 1.746237458 | 9.771991919 | 14.541388 | 0.3014214 |
| std | 1.123798 | 0.49998636 | 1.028475776 | 1.000411069 | 5.652774236 | 8.46741738 | 0.45897125 |
| min | 15 | 0 | 0 | 0 | 0.001056539 | 0 | 0 |
| 25% | 15 | 0 | 0 | 1 | 5.043079162 | 7 | 0 |
| 50% | 16 | 1 | 0 | 2 | 9.705363213 | 15 | 0 |
| 75% | 17 | 1 | 2 | 2 | 14.40840955 | 22 | 1 |
| max | 18 | 1 | 3 | 4 | 19.978094 | 29 | 1 |
|  | ParentalSupport | Extracurricular | Sports | Music | Volunteering | GPA | GradeClass |
| count | 2392 | 2392 | 2392 | 2392 | 2392 | 2392 | 2392 |
| mean | 2.122073579 | 0.3833612 | 0.303511706 | 0.196906355 | 0.157190635 | 1.9061863 | 2.98369565 |
| std | 1.122812854 | 0.48630676 | 0.459870375 | 0.397744113 | 0.364056518 | 0.91515582 | 1.23390756 |
| min | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25% | 1 | 0 | 0 | 0 | 0 | 1.17480345 | 2 |
| 50% | 2 | 0 | 0 | 0 | 0 | 1.89339269 | 4 |
| 75% | 3 | 1 | 1 | 0 | 0 | 2.62221617 | 4 |
| max | 4 | 1 | 1 | 1 | 1 | 4 | 4 |

图2-1是以人种分类的GPA箱线图，可以看出，白种人（0）的成绩相对最低，而非裔美籍学生（1）和其他人种（3）的成绩相对更高 ，但各人种成绩分布基本相同，无显著差异。图2-2是以父母受教育程度为分类的GPA箱线图，数值越高代表父母受教育程度越高，从图中可以看出，学生GPA并未与父母受教育程度呈线正相关性，受教育程度为最高的那一批父母的孩子在GPA上的表现甚至是最差的。

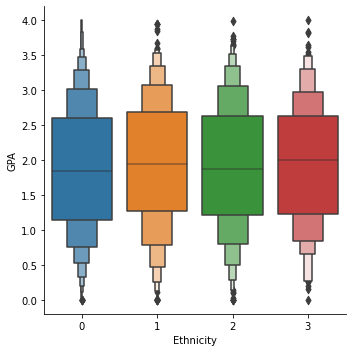
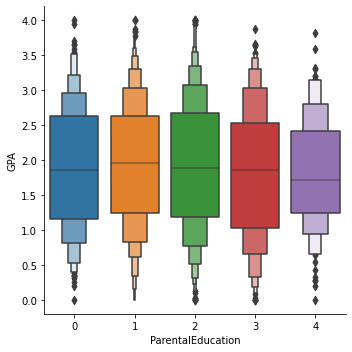
 

图2-1 图2-2

图2-3为父母对孩子的支持程度与孩子为分类的GPA箱线图，从图中可以明显的看出，父母对孩子的支持程度与孩子的GPA呈显著正相关，父母对孩子支持程度越高，孩子的成绩越好。图2-4中更是明显的看出在GPA最高的那部分学生，父母的支持是必不可少的。而图2-5则是研究父母的受教育程度与父母对孩子的支持程度是否有关系，从图中可以看出，父母的受教育程度与对孩子的支持程度无显著的关系，受教育程度低的父母也可以给孩子足够的支持。

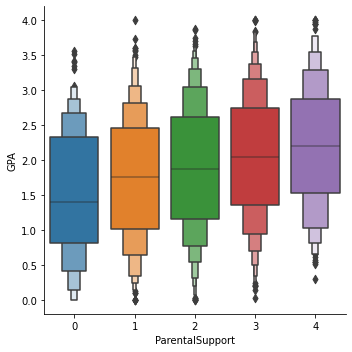


图2-3

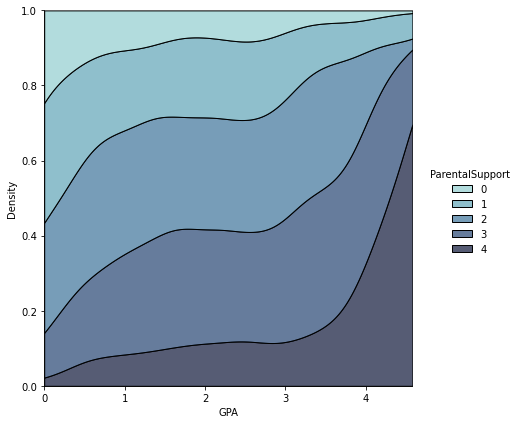


图2-4

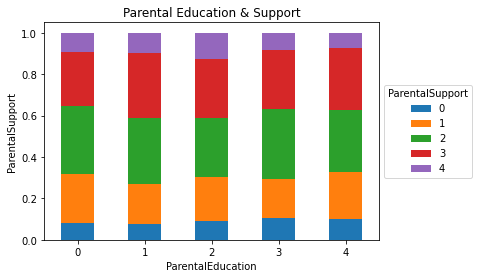


图2-5

图2-6中，性别对GPA的影响也是微乎其微的，而请家教则对孩子的GPA有明显积极的提高作用。图2-8为学习时间与GPA的热力图，可以看出，学习时间与GPA存在着相关性，但相关性不是特别强。增加学习时间有利于成绩的提高。图2-9为缺课次数与GPA的散点图，图中呈现了强烈的负相关性，可以看出缺课对学生成绩的影响是非常大的。

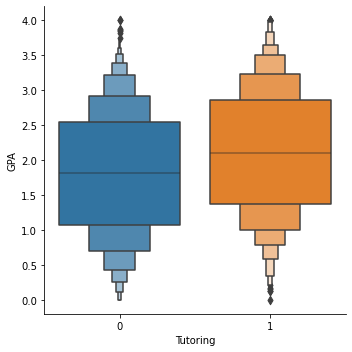
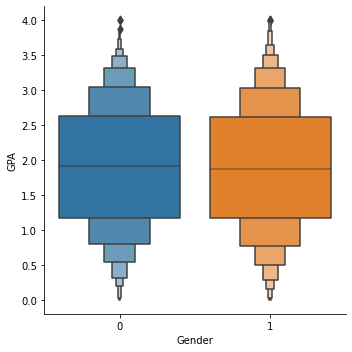


图2-6 图2-7

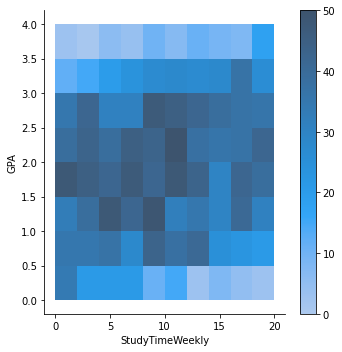


图2-8

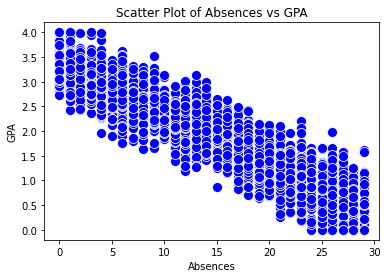


图2-9

# 三 回归分析

## 1线性回归模型

根据图3-1各变量之间的相关性以及第二部分的可视化分析结果，去除部分与GPA相关性较低的列，包括Age,Ethnicity,Gender列，因为GPA与GradeClass高度相关，为避免多重共线性，进行线性回归分析时，删除GradeClass列，留下连续型数据GPA进行建模，再对Absences列和StudyTimeWeekly列进行标准化。

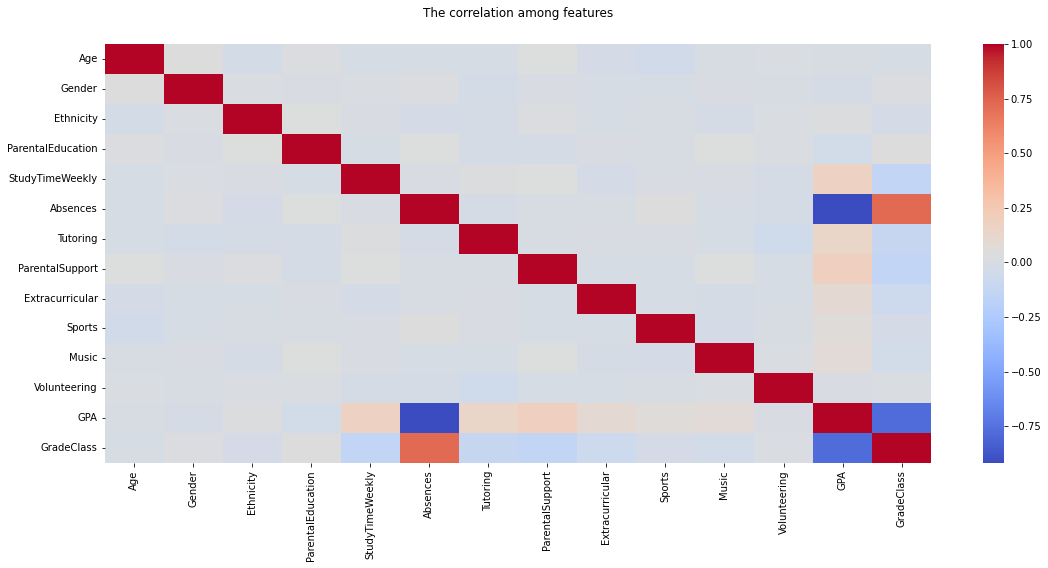


图3-1

进行分析前，先考察特征列是否可以降维，对特征列数据进行kmo值检验

kmo值：0.16726782058757675

kom值较低，特征列之间相关性不高，不适合进行降维分析。以GPA为目标列，ParentalEducation,Tutoring,ParentalSupport,Extracurricular,Sports,Music,Volunteering列划为分类数据，StudyTimeWeekly,Absences划为连续数据继进行建模。

lm1 = ols("GPA~C(ParentalEducation)+StudyTimeWeekly+Absences+C(Tutoring)+C(ParentalSupport)+C(Extracurricular)+C(Sports)+C(Music)+C(Volunteering)", data=df\_1).fit()

对数据进行线性回归建模后结果如表3-1,R-squared与Adj.R-squared相差较小，证明模型中没有过多的解释变量，而F检验量值为3290,且Prob(F-stastic)为0，证明模型回归效果显著。根据图3-2,3-3,3-4可以看出，模型不存在异方差现象，模型的残差符合正态分布。

表3-1:线性回归报告

|  |  |
| --- | --- |
| **OLS Regression Results** | |
|  | |
| **Dep. Variable:** | **GPA** |
| **Model** | **OLS** |
| **R-squared** | **0.954** |
| **Adj. R-squared** | **0.901** |
| **F-statistic** | **3290** |
| **Prob(F-statistic)** | **0** |
| **AIC:** | **-973** |
| **BIC:** | **-880** |

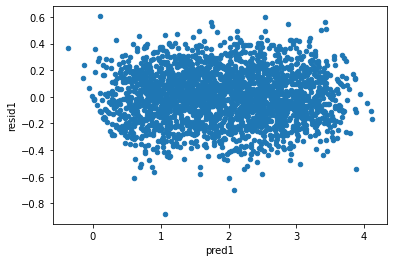


图3-2：残差图

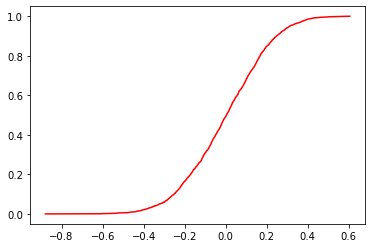


图3-3

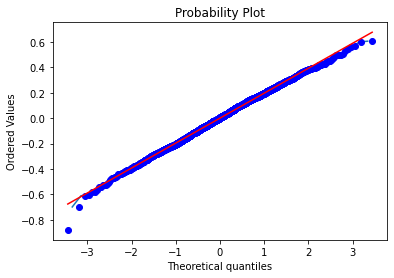


图3-4：QQ图

## 2 逻辑回归模型

该数据分类数据过多，考虑逻辑回归分析是否更加适合。以df\_1数据为基础，先对分类变量进行哑变量处理，并将GradeClass小于3的划分为成绩较好学生，赋值为1，将大于3的学生赋值为0，GradeClass作为标签列后划分数据集进行建模。

根据模型对测试机的预测结果画中混淆矩阵表3-2，并根据混淆矩阵画出ROC图（图3-5），Roc面积达到0.89，且模型准确率为85.7%，拟合效果良好。

表3-2：混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| **confusion matrix** | |  |  |
|  | **pred\_0** | **pred\_1** |  |
| **real\_0** | **303** | **22** |  |
| **real\_1** | **47** | **107** |  |

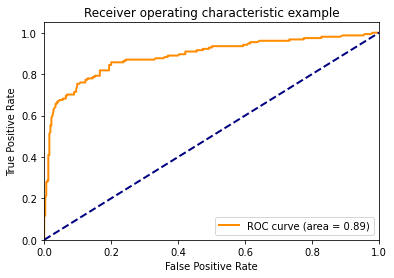


图3-5：ROC图

# 四 聚类分析

采用K-means方法对数据集进行聚类分析，聚类数据为剔除了Gradeclass和Gpa的无标签数据，经过滚石图以及手肘法确定k值为5 。

km = KMeans(n\_clusters=5,random\_state=1234).fit(X)

聚类结果中各标签占比如表4-1。各聚类结果下GradeClass占比如图4-1，可以看出聚类标签为0的类别成绩最好，而聚类标签为2的类别成绩最差。

表4-1：各标签占比及成绩平均值

|  |  |  |
| --- | --- | --- |
| **标签占比及标签下成绩均值** | | |
| **label** | **percent** | **Gpa\_mean** |
| **0** | **0.19** | **2.88** |
| **1** | **0.19** | **1.49** |
| **2** | **0.18** | **1.21** |
| **3** | **0.21** | **1.55** |
| **4** | **0.2** | **2.38** |

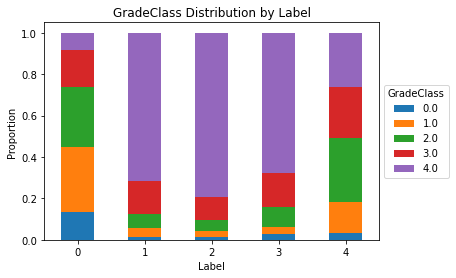


图4-1:各标签下成绩等级占比

# 五 结论以及建议

表6-1分别为线性回归分析与逻辑回归分析报告中摘选的部分自变量，回归系数，以及p值，表6-2为聚类分析结果以标签为分类，选取的部分特征列的均值。由表6-1中我们可以看到缺课次数的线性回归系数为-0.8446，逻辑回归中系数为-3.2457，在两个模型中均为所有自变量中最高，可以得出缺课次数在模型的所有自变量中对因变量的影响最大，其次分别为父母对孩子的高支持度，家教，课外活动参与，学习时间，而父母的受教育程度以及参加音乐，志愿等活动不能被认为与成绩有显著性关联。再根据聚类分析的结果，图4-1中得出的标签为0的类别成绩最好，再由表6-2可知，标签为0的学生父母支持度，学习时间，都为所有标签中的较高水平，而缺席次数则是显著最低，而成绩最差的标签为2的学生则是父母的支持度最低，仅为0.84，缺课次数也是所有标签中的第二。聚类分析的结果与回归分析结果相互印证，证明本报告模型拟合效果好。

综上所诉，如果学生想取得好成绩，首先需要尽量降低缺课次数，如果家里有条件的情况下，可以请家教来帮助学习，学历低的父母也不必担心，你们的受教育程度对孩子的成绩并无显著负面影响，而对孩子提供尽可能高的支持才是孩子成绩好的关键。令人感到不解的一点是，学习时间对成绩起的作用并没有想象中的那么大，本报告对此有两方面的推测，一是这个结果从侧面印证了学习效率的重要性，二是本数据集中学生学习时间的范围跨度太小，并未考虑长期的学习时间的积累，量太小还不足以引起质变。最后，孩子的成绩与父母的支持程度哪一个为自变量哪一个为因变量还需要进一步考虑。

表6-1：回归分析结论表

|  |  |  |
| --- | --- | --- |
| **线性回归特征列（部分）** | **coef** | **p>|t|** |
| **C(Tutoring\_1)** | **0.2496** | **0** |
| **C(ParentalSupport\_4)** | **0.6138** | **0** |
| **StudyTimeWeekly** | **0.1635** | **0** |
| **Absences** | **-0.8446** | **0** |
| **C(Music\_1)** | **0.1423** | **0** |
| **C(Sports\_1)** | **0.194** | **0** |
| **C(Volunteering\_1)** | **-0.008** | **0.471** |
| **C(Extracurricular\_1)** | **0.1913** | **0** |
| **C(ParentalEducation\_4)** | **0.0171** | **0.436** |
| **C(ParentalEducation\_3)** | **-0.0169** | **0.3** |
|  |  |  |
| **逻辑回归特征列（部分）** | **coef** | **p>|z|** |
| **Tutoring\_1** | **0.5851** | **0** |
| **ParentalSupport** | **0.118** | **0.66** |
| **StudyTimeWeekly** | **0.5242** | **0** |
| **Absences** | **-3.2457** | **0** |
| **Music** | **0.1619** | **0.318** |
| **Sports** | **0.2078** | **0.124** |
| **Volunteering** | **-0.2464** | **0.182** |
| **Extracurricular** | **0.3784** | **0.004** |
| **ParentalEducation\_4** | **-0.2614** | **0.03** |

表6-2：聚类分析结论表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **label** | **ParentalEducation** | **StudyTimeWeekly** | **Absence** | **ParentalSupport** |
|  |  |  |  |  |
| **0** | **2.1** | **13.91** | **7.05** | **2.82** |
| **1** | **1.05** | **13.95** | **20.84** | **2.71** |
| **2** | **2.34** | **10.77** | **19.89** | **0.84** |
| **3** | **2.11** | **4.26** | **17.36** | **2.83** |
| **4** | **1.11** | **6.83** | **7.73** | **1.28** |