高等教育学生表现分析

学号：202058508227 作者：代晓强

目录

1 引言

1.1 问题的研究背景

- 问题产生的背景是什么？(R Core Team 2019)

  高等教育（即 HigherEducation）指在完成中等教育的基础上进行的专业教育和职业教育，是培养高级专门人才和职业人员的主要社会活动。包括我们现在所受到的大学教育也是其中一种，而与其他教育不同的是，在接受高等教育过程中，学生在相对开放的环境下会受到各种各样因素的影响，因此学生的成绩表现也不尽相同。

1.2 问题的研究目的和研究意义

- 为什么要做这个问题，有什么意义？

  此次的分析通过对某大学采集的高等教育学生的所处环境、父母教育、初高中学校等不同因素的共同作用下的学生表现进分析，不同的环境因子通过对样例数据集中高等教育学生表现的分析，可以得到学生受不同因素的影响下成绩表现的好坏，从而进一步了解不同因素对学生最终表现（Performances）的影响，进而在一定程度上预测哪种学生在高等教育中表现更好。

2 方法

2.1 问题的描述和统计模型的选取

- 对问题进行数学描述，并分析选用何种统计学模型进行建模。

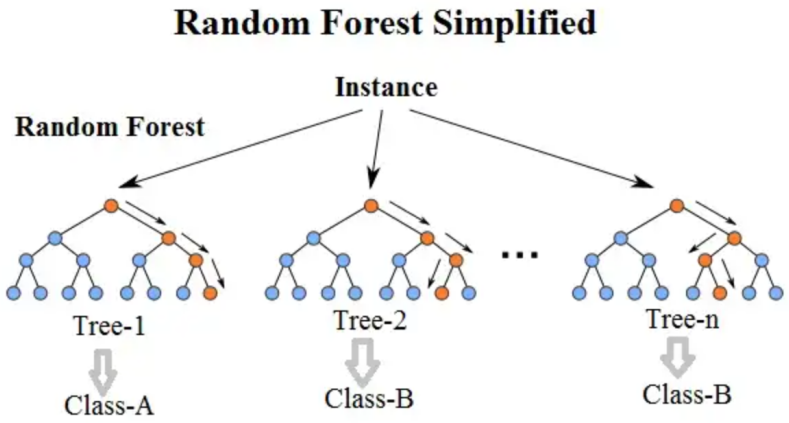
  该问题主要通过对样例数据中的学生在不同影响因子下共同作用下的表现进行分析，进而通过其在共同影响下的最终表现成绩（GRADE 列），得出对于学生的表现（Performances）影响大小的影响因子排序。这里使用的是随机森林模型，对其进行建模分析。

2.2 统计模型的介绍与应用方法

- 对统计学模型进行详细介绍。

  随机森林是一种由决策树构成的（并行）集成算法，属于 Bagging 类型，通过组合多个弱分类器，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度和泛化性能，同时也有很好的稳定性，广泛应用在各种业务场景中

  随机森林核心点是「随机」和「森林」，也是给它带来良好性能的最大支撑。随机森林，顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。



- 介绍如何应用该模型进行问题分析。

  在此次对高等教育学生表现分析的实验中，便是根据不同的影响银子设立决策树，然后根据各个组合的弱类型选择器所得出的成绩（Grade 列）去进行影响因子的分类，最终确定各特征列中的各个类别中对高等教育学生表现影响较重的项提取出来进行排序和可视化，从而得到对应的结构，即对高等教育学生表现影响较大的情况。当然由此分类出的结果有一定误差，但也可供我们参考。

3 实验与分析

3.1 数据来源

- 数据来源和实验环境

  选自 kaggle 的数据集———“Higher Education Students Performance Evaluation”，数据主要收集于某国外高校工程和教育科学学院 2019 年的学生。当然，其中的数据类型基本都是数字因子类型，后期会对其进行数据的重构，

- 实验环境介绍

  – R 4.2 + RStudio

– R 中的 `tidyverse` 库和 `caret` 库（主要）

3.2 数据介绍

- 数据相关说明和对应字典

  数据中共有 30 余列，除去首列学生编号（STUDENTID）与末列的成绩表现（GRADE）以外，即有 30 种对学生表现的影响因子，而每个影响因子对应的值通过"1~N" 的数字代表其相应的含义，如下表格所示（部分）：

Variables(列名)	Description(介绍)
STUDENTID	学生编号（不作为影响因子）
AGE	学生年龄（1：18-21，2：22-25，3：26 以上）
GENDER	性别（1：女性，2：男性）
SCHOLARSHIP	奖学金类型：（1：无，2：25%，3：50%，4：75%，5：全额）
WORK	额外工作：（1：有，2：没有）
ACTIVITY	定期的艺术或体育活动: (1: 是, 2: 否)
PARTNER	是否有伴侣：（1：是，2：否）
MOTHER_EDU	母亲的教育：（1：小学，2：中学，3：高中，4：大学，5：硕士，6：博士）
FATHER_EDU	父亲的教育：（1：小学，2：中学，3：高中，4：大学，5：硕士，6：博士）
LIVING	住宿类型：（1：租房，2：宿舍，3：与家人一起，4：其他）
#_SIBLINGS	姐妹/兄弟的数量 (1：1，2:，2，3：3，4：4，5：5 或以上）
KIDS	父母状况：（1：已婚，2：离异，3：死亡–其中一人或两人）

Variables(列名)	Description(介绍)
STUDY_HRS	每周学习时间：（1：无，2：<5 小时，3：6-10 小时，4：11-20 小时，5：20 小时以上）
MOTHER_JOB	母亲的职业：（1：退休，2：家庭主妇，3：政府官员，4：私人部门雇员，5：自营职业，6：其他）
FATHER_JOB	父亲的职业：（1：退休，2：政府官员，3：私人部门雇员，4：自营职业，5：其他）
READ_FREQ_SCI	阅读频率（科学书籍/期刊）(1: 没有, 2: 有时, 3: 经常)
ATTEND_DEPT	参加与本系有关的研讨会/会议：（1：是，2：否）
GRADE	最终表现 Performance（0：不及格，1：DD，2：DC，3：CC，4：CB，5：BB，6：BA，7：AA）
……	……

- 按照以下过程进行数据分析，要求有代码和**文字描述**，文字描述需要简单介绍代码的基本思路和代码运行后的结果。

3.3 数据导入

- 3.3.1 利用 `library()` 导入此次数据分析中需要用到的 R 包

```
library(tidyverse)
library(rattle)
library(caret)
```

- 3.3.2 利用 `read_csv()` 导入原生数据集

```
dset.0 <- read_csv("./data_csv/student_prediction.csv")
# show 数据集
dset.0
```

```
## # A tibble: 145 x 33
##    STUDENTID   AGE GENDER HS_TYPE SCHOLAR~1  WORK ACTIV~2 PARTNER SALARY TRANS~3
##    <chr>     <dbl>  <dbl>   <dbl>     <dbl> <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
##  1 STUDENT1      2      2       3         3     1       2       2      1       1
##  2 STUDENT2      2      2       3         3     1       2       2      1       1
##  3 STUDENT3      2      2       2         3     2       2       2      2       4
##  4 STUDENT4      1      1       1         3     1       2       1      2       1
##  5 STUDENT5      2      2       1         3     2       2       1      3       1
##  6 STUDENT6      2      2       2         3     2       2       2      2       1
##  7 STUDENT7      1      2       2         4     2       2       2      1       1
##  8 STUDENT8      1      1       2         3     1       1       1      2       2
##  9 STUDENT9      2      1       3         3     2       1       1      1       1
## 10 STUDENT10     2      1       2         3     2       2       1      3       4
## # ... with 135 more rows, 23 more variables: LIVING <dbl>, MOTHER_EDU <dbl>,
## #   FATHER_EDU <dbl>, `#_SIBLINGS` <dbl>, KIDS <dbl>, MOTHER_JOB <dbl>,
## #   FATHER_JOB <dbl>, STUDY_HRS <dbl>, READ_FREQ <dbl>, READ_FREQ_SCI <dbl>,
## #   ATTEND_DEPT <dbl>, IMPACT <dbl>, ATTEND <dbl>, PREP_STUDY <dbl>,
## #   PREP_EXAM <dbl>, NOTES <dbl>, LISTENS <dbl>, LIKES_DISCUSS <dbl>,
## #   CLASSROOM <dbl>, CUML_GPA <dbl>, EXP_GPA <dbl>, `COURSE ID` <dbl>,
## #   GRADE <dbl>, and abbreviated variable names 1: SCHOLARSHIP, ...
```

3.4 数据整理

- 3.4.1 原生数据遵循以列为变量、行为观察值的原则，且无明显缺失或重复值
- 3.4.2 为方便理解和后续操作，将数据集的列名由大写转为小写，且将原本表示学生父母的婚姻状况中”kids” 列重新命名列名为”parents_status”

```
dset_im.0 <- dset.0
names(dset_im.0) <- tolower(names(dset.0))
dset_im.0 <- dplyr::rename(dset_im.0, parents_status = kids)
dset_im.0
```

```
## # A tibble: 145 x 33
##    studentid   age gender hs_type scholar~1  work activ~2 partner salary trans~3
##    <chr>     <dbl>  <dbl>   <dbl>     <dbl> <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
##  1 STUDENT1      2      2       3         3     1       2       2      1       1
##  2 STUDENT2      2      2       3         3     1       2       2      1       1
##  3 STUDENT3      2      2       2         3     2       2       2      2       4
##  4 STUDENT4      1      1       1         3     1       2       1      2       1
##  5 STUDENT5      2      2       1         3     2       2       1      3       1
##  6 STUDENT6      2      2       2         3     2       2       2      2       1
##  7 STUDENT7      1      2       2         4     2       2       2      1       1
##  8 STUDENT8      1      1       2         3     1       1       1      2       2
##  9 STUDENT9      2      1       3         3     2       1       1      1       1
## 10 STUDENT10     2      1       2         3     2       2       1      3       4
## # ... with 135 more rows, 23 more variables: living <dbl>, mother_edu <dbl>,
## #   father_edu <dbl>, `#_siblings` <dbl>, parents_status <dbl>,
## #   mother_job <dbl>, father_job <dbl>, study_hrs <dbl>, read_freq <dbl>,
## #   read_freq_sci <dbl>, attend_dept <dbl>, impact <dbl>, attend <dbl>,
## #   prep_study <dbl>, prep_exam <dbl>, notes <dbl>, listens <dbl>,
## #   likes_discuss <dbl>, classroom <dbl>, cuml_gpa <dbl>, exp_gpa <dbl>,
## #   `course id` <dbl>, grade <dbl>, and abbreviated variable names ...
```

3.5 数据操作

- 3.5.1 利用 `c()` 构建数据筛选的筛选向量 `vars_im`

```
## 除去无用的首列 studentid，只保留了要用到的影响因素和最终成绩表现 (grade) 列
vars_im <- c("gender", "scholarship", "work", "activity", "partner",
```

```
             "parents_status", "study_hrs", "attend_dept", "grade")
vars_im
```

```
## [1] "gender"         "scholarship"    "work"           "activity"
## [5] "partner"        "parents_status" "study_hrs"      "attend_dept"
## [9] "grade"
```

- 3.5.2 利用 `select()` 结合上述的 `vaes_im` 对数据列筛选

```
dset_im.1 <- dset_im.0 |>
  select(one_of(vars_im))
dset_im.1
```

```
## # A tibble: 145 x 9
##    gender scholarship  work activity partner parents_sta~1 study~2 atten~3 grade
##     <dbl>       <dbl> <dbl>    <dbl>   <dbl>         <dbl>   <dbl>   <dbl> <dbl>
##  1      2           3     1        2       2             1       3       1     1
##  2      2           3     1        2       2             1       2       1     1
##  3      2           3     2        2       2             1       2       1     1
##  4      1           3     1        2       1             1       3       1     1
##  5      2           3     2        2       1             1       2       1     1
##  6      2           3     2        2       2             1       1       1     2
##  7      2           4     2        2       2             1       2       2     5
##  8      1           3     1        1       1             1       1       1     2
##  9      1           3     2        1       1             1       1       1     5
## 10      1           3     2        2       1             1       2       1     0
## # ... with 135 more rows, and abbreviated variable names 1: parents_status,
## #   2: study_hrs, 3: attend_dept
```

- 3.5.3 利用 `recode()` 对数据进行重新编码
  注:数据的重新编码可参考上述 “2.3 介绍数据” 中的 `Variables-Description` 表格

```
dset_im.2 <- dset_im.1
dset_im.2$gender <- recode(dset_im.1$gender, '1' = 'Female', '2' ='Male')
dset_im.2$scholarship <- recode(dset_im.1$scholarship, '1' = '0%', '2' = '25%',
                                '3' = '50%', '4'= '75%', '5'= '100%')
dset_im.2$work <- recode(dset_im.1$work, '1' = 'Yes', '2' = 'No')
dset_im.2$activity <- recode(dset_im.1$activity, '1' = 'Yes', '2' = 'No')
dset_im.2$partner <- recode(dset_im.1$partner, '1' = 'Yes', '2' = 'No')
dset_im.2$parents_status <- recode(dset_im.1$parents_status, '1' = 'Married',
                           '2' = 'Divorced', '3' = 'Died - one of them or both')
dset_im.2$study_hrs <- recode(dset_im.1$study_hrs, '1' = 'None',
                    '2' = '<5 hours', '3' = '6-10 hours', '4' = '11-20 hours',
                    '5' = 'More than 20 hours')
dset_im.2$attend_dept <- recode(dset_im.1$attend_dept, '1' = 'Yes', '2' = 'No')
dset_im.2
```

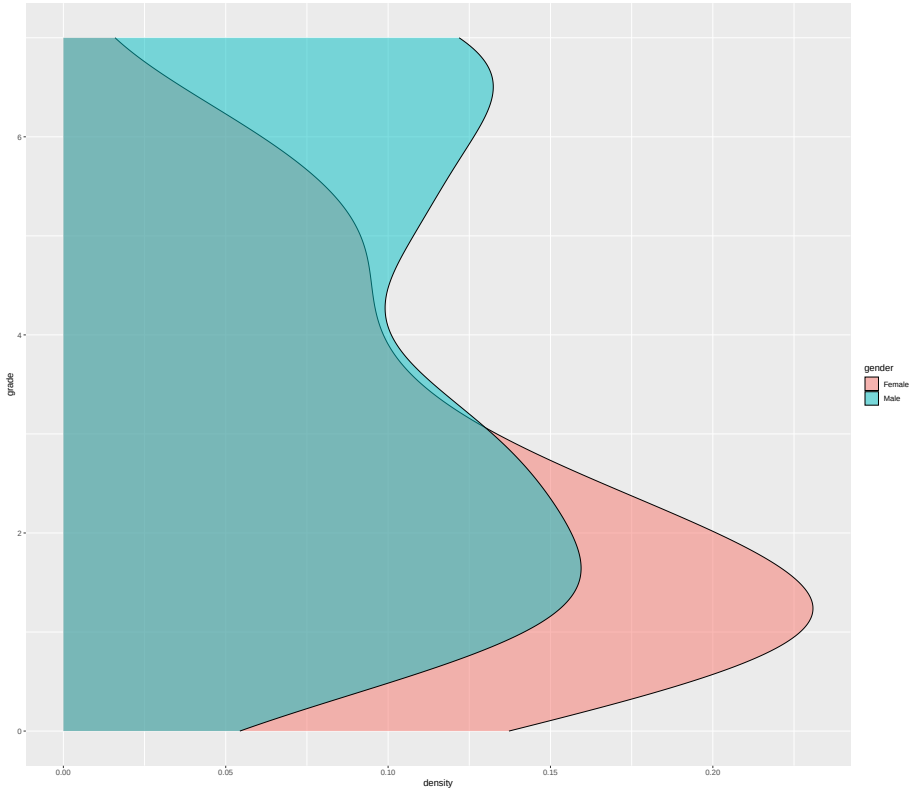
```
## # A tibble: 145 x 9
##    gender scholarship work  activity partner parents_sta~1 study~2 atten~3 grade
##    <chr>  <chr>       <chr> <chr>    <chr>   <chr>         <chr>   <chr>   <dbl>
##  1 Male   50%         Yes   No       No      Married       6-10 h~ Yes         1
##  2 Male   50%         Yes   No       No      Married       <5 hou~ Yes         1
##  3 Male   50%         No    No       No      Married       <5 hou~ Yes         1
##  4 Female 50%         Yes   No       Yes     Married       6-10 h~ Yes         1
##  5 Male   50%         No    No       Yes     Married       <5 hou~ Yes         1
##  6 Male   50%         No    No       No      Married       None    Yes         2
##  7 Male   75%         No    No       No      Married       <5 hou~ No          5
##  8 Female 50%         Yes   Yes      Yes     Married       None    Yes         2
##  9 Female 50%         No    Yes      Yes     Married       None    Yes         5
## 10 Female 50%         No    No       Yes     Married       <5 hou~ Yes         0
## # ... with 135 more rows, and abbreviated variable names 1: parents_status,
## #   2: study_hrs, 3: attend_dept
```

3.6 数据可视化

- **注：**可视化中主要绘制了两种图像—密度分布图（对于”YES/NO” 双观测值变量）和条形图（观测值在两个以上的变量）

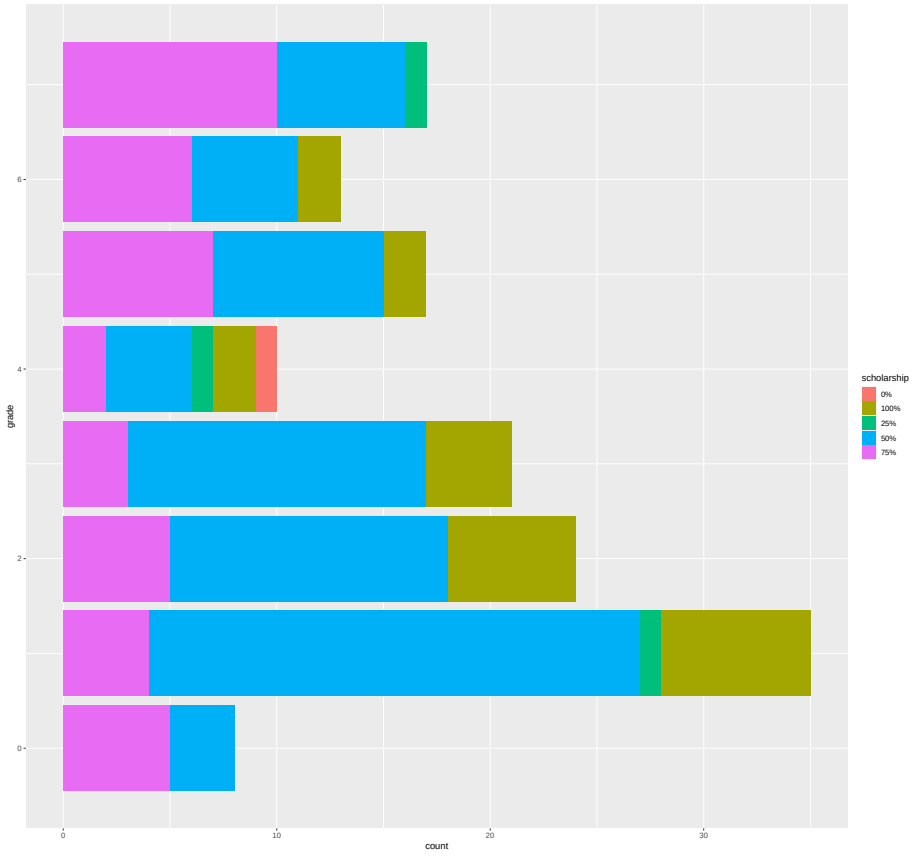
- 3.6.1 绘制 “成绩表现（grade）” 关联 “学生性别（gender）” 的密度分布图

```
# 选取工程学院的采集数据，男生会比较多、且成绩表现男生相对好一些
dset_im.2 |>
ggplot(aes(grade)) +
  geom_density(aes(fill = gender), alpha = 0.5) +
  coord_flip()
```



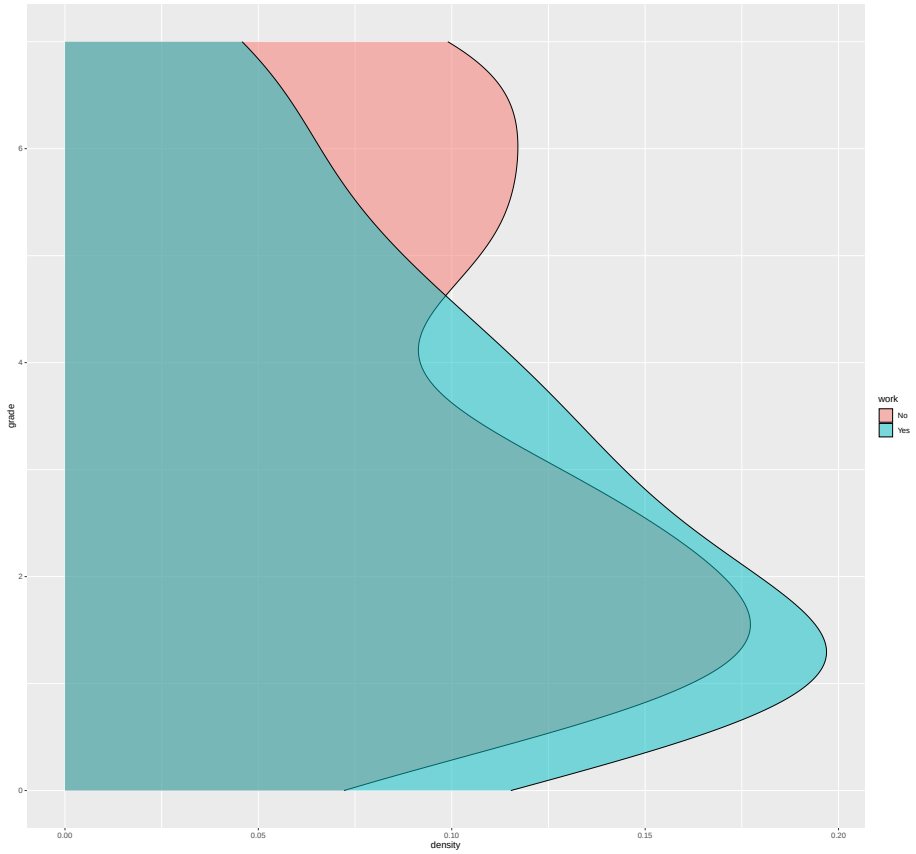
- 3.6.2 绘制获得 “奖学金级别” 与 “成绩表现（grade）” 的条形图

```
# 获得最高奖学金的未来成绩表现不一定最好
dset_im.2 |>
ggplot(aes(grade, group = scholarship)) +
  geom_bar(aes(fill = scholarship)) +
  coord_flip()
```



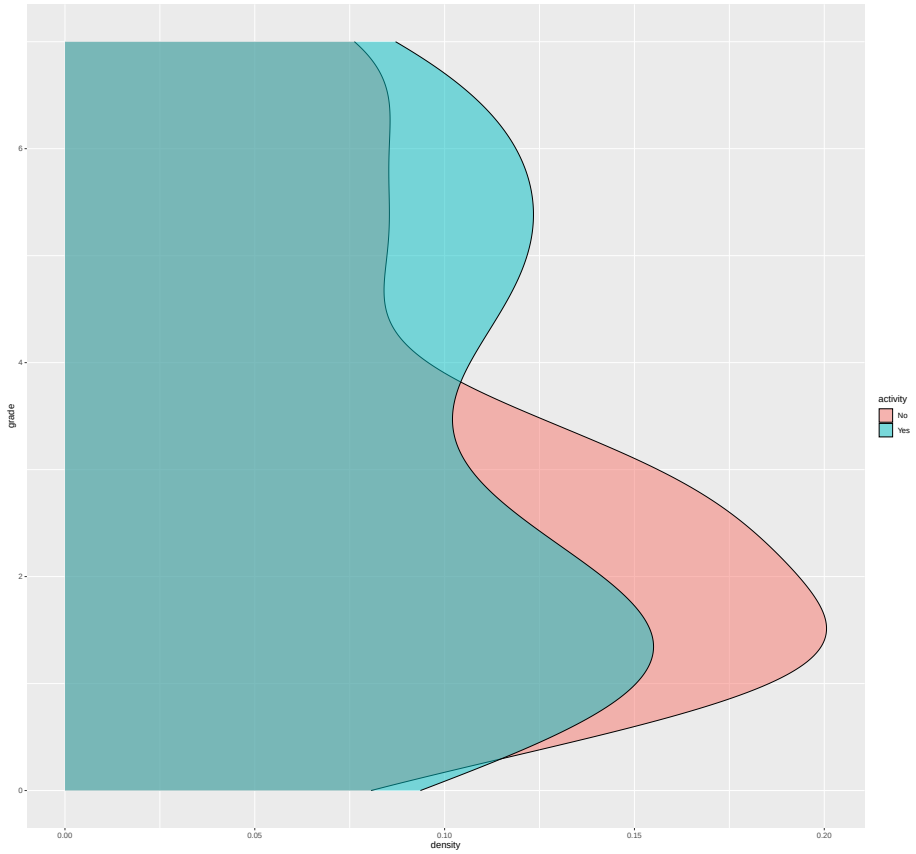
- 3.6.3 绘制“有无额外工作（work）”与“成绩表现（grade）”的密度分布图

```
# 无额外工作在高成绩占比较大
dset_im.2 |>
ggplot(aes(grade)) +
  geom_density(aes(fill = work), alpha = 0.5) +
  coord_flip()
```



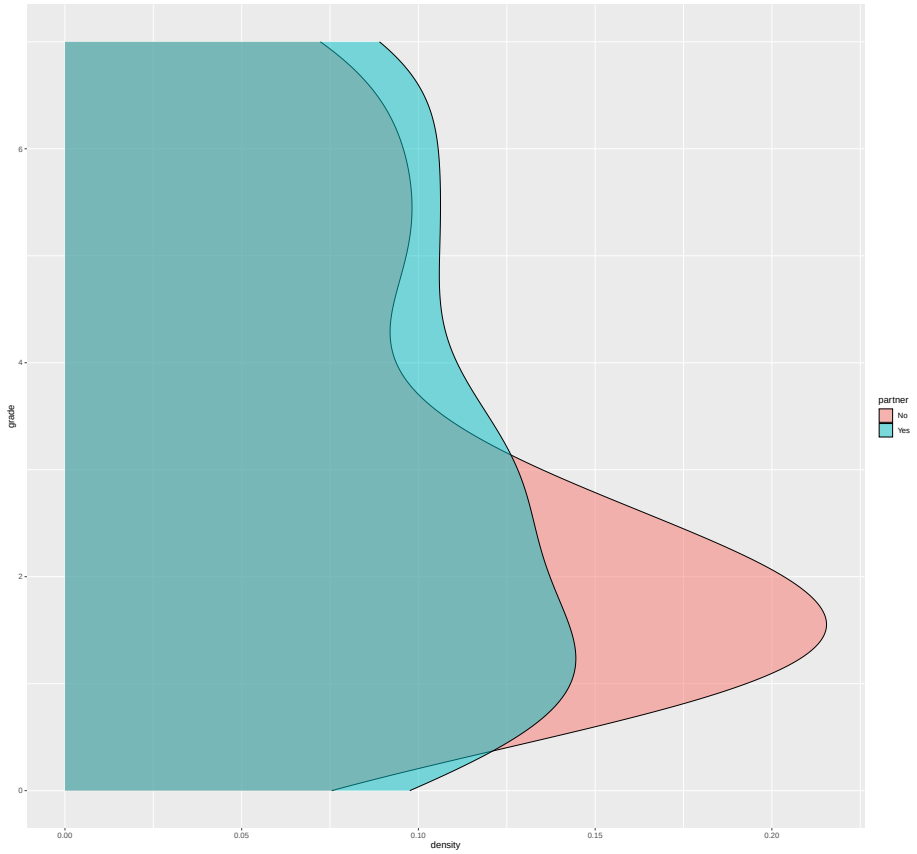
- 3.6.4 绘制 “是否喜欢参加体育/艺术活动（activity）” 与 “成绩表现（grade）” 的密度分布图

```
# 经常参加体育/艺术活动的学生反而成绩表现好一些
dset_im.2 |>
ggplot(aes(grade)) +
  geom_density(aes(fill = activity), alpha = 0.5) +
  coord_flip()
```



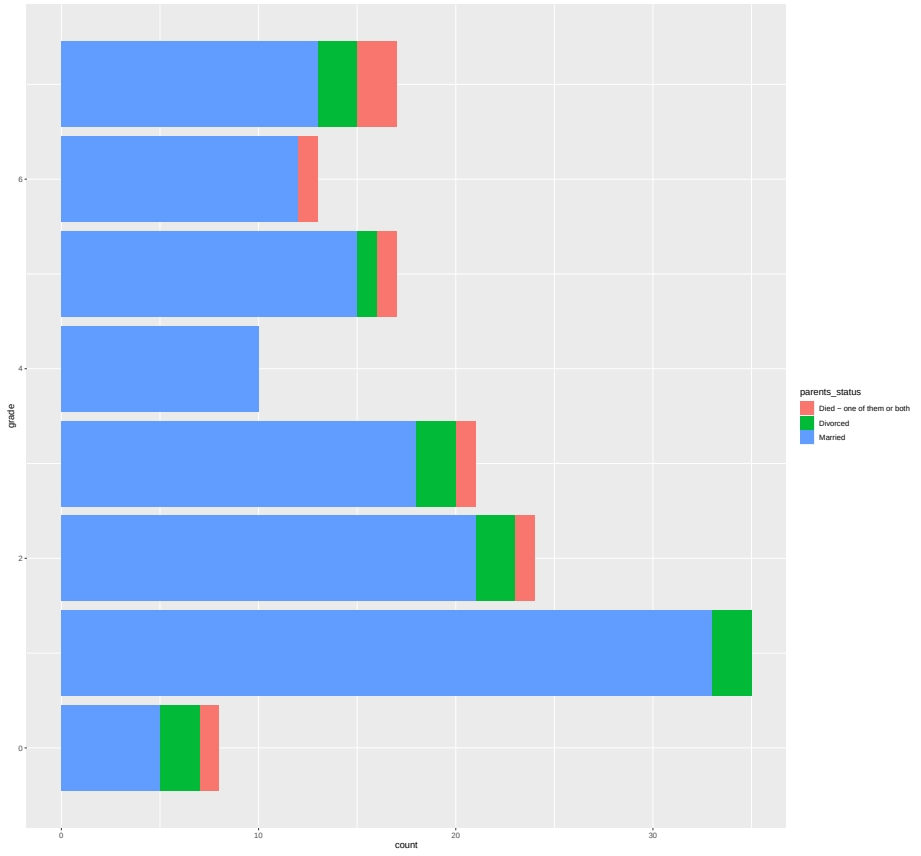
- 3.6.5 绘制 “是否有伴侣（partner）” 与 “成绩表现（grade）” 的密度分布图

```
# 有伴侣的学生成绩表现要好一些
dset_im.2 |>
ggplot(aes(grade)) +
  geom_density(aes(fill = partner), alpha = 0.5) +
  coord_flip()
```



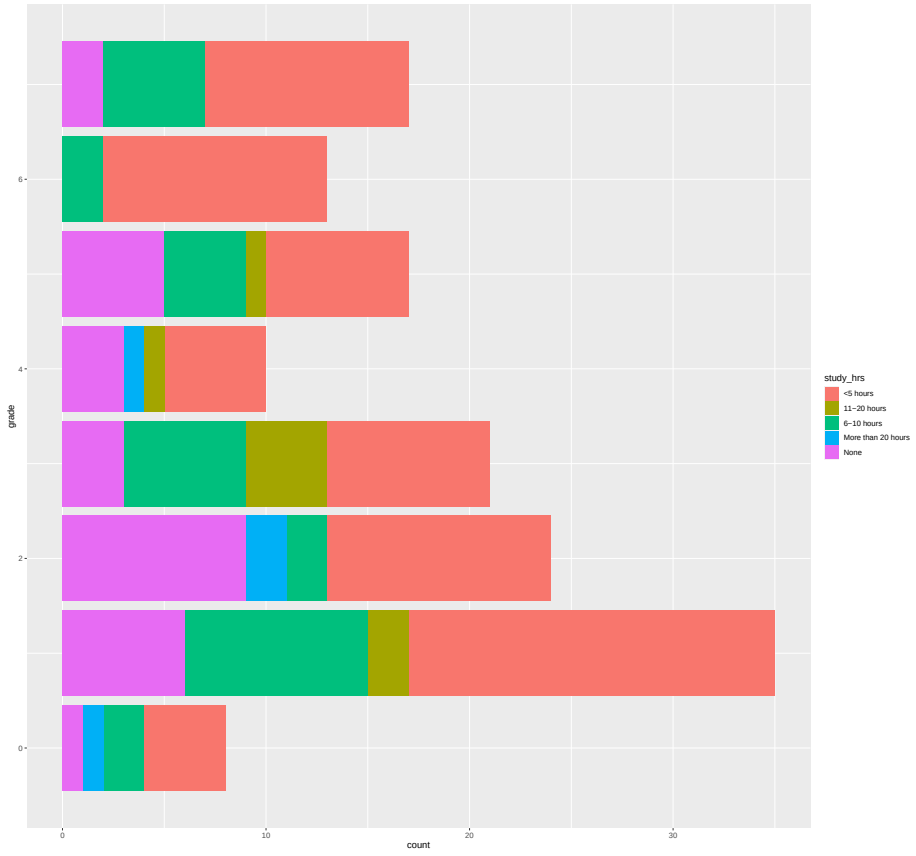
- 3.6.6 绘制 “父母的婚姻状况（parents_status）” 与 “成绩表现（grade）” 的条形图

```
# 分布比较均匀，但由于数据集规模太小参考性不高
dset_im.2 |>
ggplot(aes(grade, group = parents_status)) +
  geom_bar(aes(fill = parents_status)) +
  coord_flip()
```



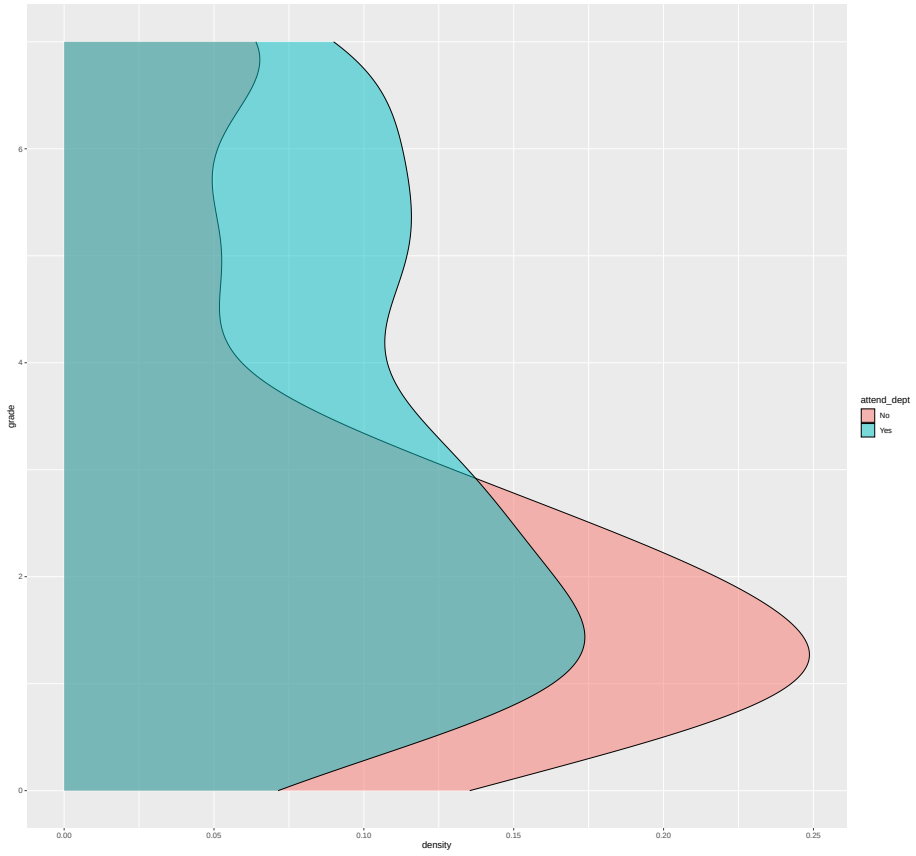
- 3.6.7 绘制 “每周学习时间（study_hrs）” 与 “成绩表现（grade）” 的条形图图

```
# 成绩表现较好的反而每周的学习时间没有那么长，每周适当学习时间表现较好
dset_im.2 |>
ggplot(aes(grade, group = study_hrs)) +
  geom_bar(aes(fill = study_hrs)) +
  coord_flip()
```



- 3.6.8 绘制“是否有参加系里研讨会（attend_dept）”与“成绩表现（grade）”的密度分布图

```
# 显然有参加研讨会的同学在高成绩占比比较大
dset_im.2 |>
ggplot(aes(grade)) +
  geom_density(aes(fill = attend_dept), alpha = 0.5) +
  coord_flip()
```



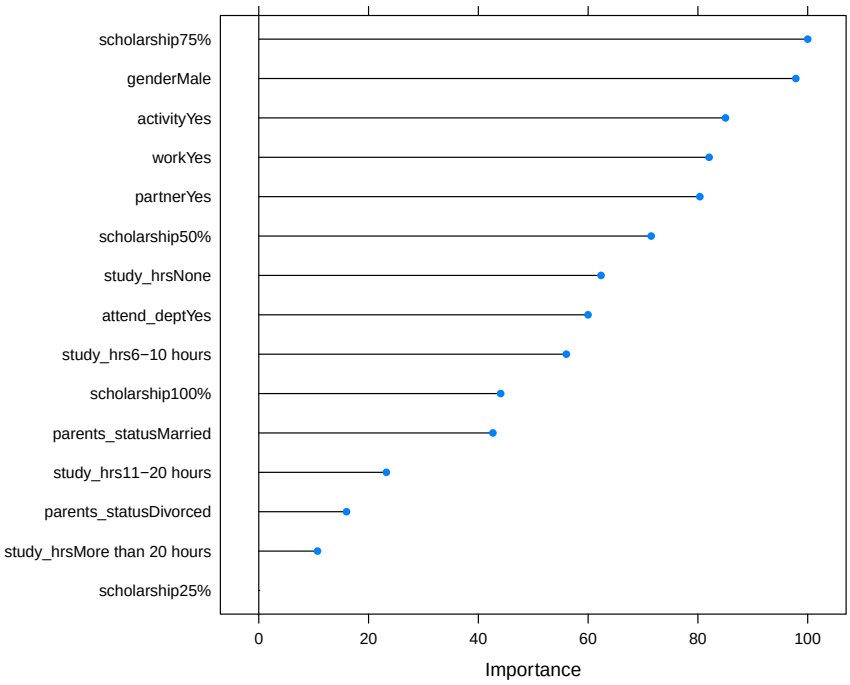
3.7 数据建模

- 3.7.1 对数据进行划分处理，将数据按一定的比例划分成训练集和测试集

```
# 分层抽样
set.seed(777)  # 设置随机种子
# 按照目标进行 8: 2 的分层抽样
train_list <- createDataPartition(dset_im.2$grade, p = 0.8, list = FALSE)
train <- dset_im.2[train_list,]
test <- dset_im.2[-train_list,]
# 将因变量转为因子类型
train$grade <- as.factor(train$grade)
test$grade <- as.factor(test$grade)
```

- 3.7.2 构建简单的随机森林模型并可视化各变量（观察值）对成绩表现（grade）的”importance”

```
# 构建模型的控制对象（采用 5 折交叉验证）
trControl<-trainControl(method = 'cv',number = 5)
set.seed(777)
# 随机森林模型构建
model_rf<-train(grade~.,data = train, method='rf',trControl=trControl)
# 对测试集进行预测
pred_rf<-predict(model_rf,test)
# 单变量直接用 plot 函数画出即可
plot(varImp(model_rf))
```



3.8 数据分析结果

- 对数据可视化分析结果汇总

Variables(列名)	Effects(对"grade" 影响大小)
gender	选取工程学院的采集数据，男生较多且成绩表现相对更好一些
scholarship	获得最高奖学金学生的成绩表现不一定是最好，而 75% 等奖学金学生成绩表现比较突出
work	无额外工作学生高成绩表现占密度相对大一些
activity	参加体育/艺术活动的学生成绩表现相对更好一些

Variables(列名)	Effects(对"grade" 影响大小)
partner	有伴侣的学生成绩反而比没有的成绩表现更好一些
parents_status	父母状况相对的成绩分布比较均匀，数据集规模太小参考性不高
study_hrs	每周学习时长最长的学生反而可能成绩表现最差，每周适当学习时间的学生成绩表现比较好
attend_dept	有参加系研讨会的学生在高成绩表现的密度明显更大

4 结论和讨论

4.1 根据实验结果分析得到哪些结论？

综合随机森林模型给出的各变量观测值"importance" 的分析结果以及各变量关联学生成绩表现（grade）的数据可视化分析结果我们可以得出：

- 该工程和教育科学系学生中奖学金类别、参加系研讨会、性别、参加体育/艺术活动、有无伴侣几个因素对学生在高等教育中成绩表现（grade）影响较大
- 奖学金类别在 75% 的学生成绩表现（grade）相对较好
- 参加系研讨会的学生成绩表现（grade）明显要好很多
- 工程系学生中男学生相对女学生成绩好一些（男学生人数要多很多）
- 参加体育/艺术活动的学生成绩表现（grade）相对较好
- 有伴侣（男女朋友）的学生成绩表现（grade）要比没有的好一些

- 也可以介绍该工作的不足以及未来可以改进的地方。

5 参考文献

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Hadley Wickham,Garrett Grolemund.R for Data Science[M].O'Reilly Media.2016

本文贡献

- 代晓强负责此次课程设计实现和课程论文书写。

6 附录

- **源代码已附在前文对应章节中**