统计分析与建模-期末项目

期末项目

- 总分占比：40%
- 团队人数：4人
- 提交内容：代码(github)/报告(PDF)/答辩(PPT)
- 考核内容:
  a. 使用R进行数据分析和建模的能力
  b. 提出问题、分析问题、解决问题、并获得有效结论的能力
  c. 文献查阅、撰写报告的能力
  d. 表达和思辨能力
  e. 团队合作能力，强调共同参与, 分工协作

1. 回归模型

- **Mission:**

该数据集提供了一个全球范围内的教育数据，包括29列，内容涵盖失学率、学业完成率、熟练程度、识字率、出生率以及小学和高等教育入学统计等信息。希望你能够通过数据分析对不同国家和地区的教育动态提供深刻的见解，希望你的工作能够给全球教育工作者一个评估、加强和重塑全球教育系统的机会。

- **Candidate Tasks:**

1. 数据分析&结论
2. 数据预处理
3. 数据建模及模型质量评估
   a. 建立高等教育总入学人数预测模型（Y：Gross_Tertiary_Education_Enrollment）
   b. 建立失业率预测模型（Y：Unemployment_Rate）
4. 模型解读（影响因素分析）
5. 根据以上数据分析工作得到的教育改进启示和建议

- **Data Introduction:**

Canvas文件栏目：项目/Global_Education.csv

1. **Countries and Areas**: Name of the countries and areas.
2. **Latitude**: Latitude coordinates of the geographical location.
3. **Longitude**: Longitude coordinates of the geographical location.
4. **OOSR_Pre0Primary_Age_Male**: Out-of-school rate for pre-primary age males.
5. **OOSR_Pre0Primary_Age_Female**: Out-of-school rate for pre-primary age females.
6. **OOSR_Primary_Age_Male**: Out-of-school rate for primary age males.
7. **OOSR_Primary_Age_Female**: Out-of-school rate for primary age females.
8. **OOSR_Lower_Secondary_Age_Male**: Out-of-school rate for lower secondary age males.
9. **OOSR_Lower_Secondary_Age_Female**: Out-of-school rate for lower secondary age females.
10. **OOSR_Upper_Secondary_Age_Male**: Out-of-school rate for upper secondary age males.
11. **OOSR_Upper_Secondary_Age_Female**: Out-of-school rate for upper secondary age females.
12. **Completion_Rate_Primary_Male**: Completion rate for primary education among males.
13. **Completion_Rate_Primary_Female**: Completion rate for primary education among females.
14. **Completion_Rate_Lower_Secondary_Male**: Completion rate for lower secondary education among males.
15. **Completion_Rate_Lower_Secondary_Female**: Completion rate for lower secondary education among females.
16. **Completion_Rate_Upper_Secondary_Male**: Completion rate for upper secondary education among males.
17. **Completion_Rate_Upper_Secondary_Female**: Completion rate for upper secondary education among females.
18. **Grade_2_3_Proficiency_Reading**: Proficiency in reading for grade 2-3 students.
19. **Grade_2_3_Proficiency_Math**: Proficiency in math for grade 2-3 students.
20. **Primary_End_Proficiency_Reading**: Proficiency in reading at the end of primary education.
21. **Primary_End_Proficiency_Math**: Proficiency in math at the end of primary education.
22. **Lower_Secondary_End_Proficiency_Reading**: Proficiency in reading at the end of lower secondary education.
23. **Lower_Secondary_End_Proficiency_Math**: Proficiency in math at the end of lower secondary education.
24. **Youth_15_24_Literacy_Rate_Male**: Literacy rate among male youths aged 15-24.

25. **Youth_15_24_Literacy_Rate_Female**: Literacy rate among female youths aged 15-24.
26. **Birth_Rate**: Birth rate in the respective countries/areas.
27. **Gross_Primary_Education_Enrollment**: Gross enrollment in primary education.
28. **Gross_Tertiary_Education_Enrollment**: Gross enrollment in tertiary education.
29. **Unemployment_Rate**: Unemployment rate in the respective countries/areas.

2. 分类模型

- **Mission:**

泰坦尼克号的沉没是历史上最著名的沉船事故之一。1912年4月15日，在处女航中，被认为“永不沉没”的皇家邮轮泰坦尼克号与冰山相撞后沉没。不幸的是，船上没有足够的救生艇容纳所有人，导致2224名乘客和船员中的1502人死亡。虽然生存有一些运气因素，但似乎有些群体比其他群体更有可能生存下来。请使用乘客数据（包括姓名、年龄、性别、社会经济阶层等）建立一个预测模型，尝试思考和回答以下问题：“什么样的人更有可能存活下来？”。

- **Candidate Tasks:**

1. 数据分析&结论
2. 数据预处理
3. 数据建模及模型质量评估
    a. 建立是否存活(Survived)的预测模型（Y：Survived）
4. 模型解读（影响因素分析）
5. 根据以上数据分析工作回答问题：什么样的人更有可能存活下来?

- **Data Introduction:**

Canvas文件栏目：项目/titanic.csv

1. PassengerId
2. Survived: 0=No, 1=Yes
3. Pclass: Ticket class, 1=1st, 2=2nd, 3=3rd
4. Name: Name of passenger
5. Sex: Gender
6. Age: Age in Years
7. SibSp: No. of siblings/spouses aboard the Titanic
8. Ticket: Ticket number
9. Fare: Passenger fare
10. Cabin: Cabin number

11. Embarked: Port of Embarkation:C = Cherbourg, Q = Queenstown, S = Southampton

3. 时序模型

- **Mission:**

该数据集共包含25161行，每行代表特定公司在给定日期的股市数据。数据是从www.nasdaq.com通过网络抓取收集的信息包括上市公司的股价和交易量，如苹果、星巴克、微软、思科系统、高通、Meta、亚马逊、特斯拉、Advanced Micro Devices和Netflix。请对这些公司的数据进行统计分析，希望您能够为投资者提出可靠并宝贵的建议。

- **Candidate Tasks:**

1. 数据分析&结论（以下仅为参考）
    a. 可视化：每个公司的股票价格随时间的分布，可以使用使用折线图、条形图(barplot)和热图(heatmap)等来可视化股市数据中的趋势、季节性和模式
    b. 相关性分析：调查不同公司收盘价之间的相关性，以确定潜在的关系，计算相关系数，生成相关系数矩阵。
    c. 股票不动性分析
    d. 推荐最佳潜力股
2. 数据预处理
3. 数据建模及模型质量评估
    a. 建立股价预测模型（Y：Survived）
4. 模型解读（影响因素分析）
5. 根据以上数据分析工作给投资者提出的几点建议

- **Data Introduction:**

Canvas文件栏目：项目/stock.csv

1. Company: The stock ticker symbol of the company, used to uniquely identify it in the stock market. For example, "AAPL" represents Apple, Inc.
2. Date: The date on which the stock market data was recorded or reported. It represents the trading day for which the stock data is relevant.
3. Close/Last: TThe closing price or the last traded price of the company's stock on the given date. It represents the final price at which the stock was traded for the day.
4. Volume: The total number of shares of the company's stock traded on the given date. It indicates the level of investor interest and liquidity for the stock on that day.
5. Open: The opening price of the company's stock on the given date. It is the price at which the first trade occurred during the trading session.

6. High: The highest price at which the company's stock traded on the given date. It indicates the highest price reached during the trading session.
7. Low: The lowest price at which the company's stock traded on the given date. It indicates the lowest price reached during the trading session.