MA578 Finalproject Proposal

Bingtian Ye

2023-10-30

Problem Background:

Each student will have a different performance in the Probability and statistics course, which is specifically reflected in their grade. In this final project, I want to explore which of the usual exams (The dataset has three usual exams), attendance, and homework indicators have the greatest impact on students' final exams.

The Significance of The Question:

Since this course is so relevant to us, I thought it would be interesting to study this topic. At the same time, I think all students care about their scores.

In addition, studying this issue has certain pedagogical implications. For example, eventually, we can build a Bayesian regression model or other models. To determine which indicators, have a greater impact on students' final grades. This way, students can draw attention when they find themselves with low values on these indicators.

Dataset:

The dataset is the Probability and Statistics course from Kaggle(https://www.kaggle.com/datasets/indikawickramasinghe/probability-and-statistics-course-performance), which includes students' performance in a college-level course in Probability and Statistics from 2016 to 2022. Also, the dataset has nine features, which are Year, Semester, Exam.1, Exam.2, Exam.3, Homework, Attendance, FinalExam, FinalGrade.

Work Plan:

I did a regression and the results are as follows, and show the p-value as follow:

```
##  (Intercept)       Exam.1       Exam.2       Exam.3     Homework   Attendance
## 6.992565e-02 1.588418e-06 6.848755e-12 1.628468e-15 8.853613e-09 4.145772e-09
##    FinalExam
## 5.462972e-19
```

The p-values in the results are all very low, which means that FinalGrade can be considered a weighted sum of several other indicators. Then it makes no sense to use FinalGrade as the target variable. So use FinalExam as target variable.In terms of data, I will first perform a series of processing and EDA on the data. Then select the prior and build the model. For example, whether establishing Exam.1 will affect Exam.2 and Exam.3. The dataset includes the data from 2016 to 2022, so I want to use the data of the previous year is used as prior information, and then the data of the current year is sampling data. At the same time, I hope to use the current year's data as the prior for the next year, which means constantly updating the prior to see if the final regression results will be better.

Also, if possible, I would like to cite methods for time series analysis. Considering that the situation of each semester is different, the impact of grades further back will be smaller and smaller. So I hope to add a coefficient so that the further the time is, the smaller the influence index of the prior information given.

Excepted Result:

If everything goes according to plan, I expect to end up with a sense of the impact of different factors on the final exam, which would be reflected in the coefficients of each indicator if I were using Bayesian regression. In addition, I will build different models and use different prior distributions to get different results, and observe which prior distribution has better results.