STATS 101B - Spring 2022

General Guidelines for the Final Project

Summary of the Project

The final project involves identifying the most important tuning the parameters of **Random Forest**, an important algorithm in data science for tackling classification problems. The complete details of the task are in the file "Final_Project_Spring_2022.pdf" which is available on the section 'Modules' on Bruin Learn under **Final Project**.

Project Outline

- You will carry out this project in teams. The maximum number of students in a team is **3**. Please designate one of you as the team leader and keep in mind that, after the final project has been released, you will not be allowed to change teams. **The team members must be from the same lecture.**

- You have to submit a comprehensive report in a pdf file. Your report accounts for 75% of your grade for the final project. Your report should demonstrate your understanding of all concepts and methods seen throughout the course.

- At the end of the report, you must include a **statement of contribution** in which you describe the specific contributions of each team member to the project.

- Additionally, you have to submit a short, high-quality presentation summarizing your whole analysis. You presentation must be pre-recorded and available on YouTube. Your presentation accounts for 25% of your grade for the final project. Your presentation should be aimed at a general audience and demonstrate the benefits of design of experiments. All team members must participate in the presentation.

- You must use R and RStudio to conduct your analysis.

- The due date for submitting both the report and presentation of the final project is **June 5 at 5 pm (PDT)**.

More information about the final project is on the syllabus on Bruin Learn.

Report Guidelines

Here, I provide guidelines on the components your report should *at least* have.

1. **Introduction.** Brief description of the problem. My advice is that you do not write a lengthy introduction, but just give the necessary components for your report. For instance, the goal of the problem. Another advice is that you avoid a report summary, since the grader will read the entire report anyway.

2. **Methodology**

   a) *Experimental Design*. Your answers to the first four questions of the project.

   b) *Data Analysis*. Your answers to the last two questions of the project.

   Do not include any R code here, just your answer to each question. For instance, you can include figures and tables in the format that matches your document.

3. **Conclusions.** An explanation of why you think your design and models work well or failed to work well. In other words, it should include a critical assessment of all the decisions made. You should also include recommendations for future experimentation.

4. **Appendix.** An Appendix showing your R code. The code should be such that I can run it and get the exact same output as you. Moreover, it should be well-documented and organized.

5. **Statement of Contribution.** Clearly state the contribution of each team member to this project and report.

6. ***Optional*: References.** Please make a correct use of the references and the citations within the main text. For instance, you can follow an APA format. This is in case you would like to refer to specific sections in the required and optional textbooks.

Once again, the components mentioned above are the least a report must have. In your report, every decision and method you use should be well motivated and clearly explained. For instance, you might want to transform your response or remove outliers. All these decisions should be documented clearly and with a good level of detail.

The length of the report should be no more than **5 one-sided pages** (not counting the Appendix, References, nor the Statement of Contribution.) The pages should have a single-column layout and be single space with 11 pt. The report should include informative graphs which are legible and of a good size. The correct way to include graphs is to label them (e.g. Figure 1), and discuss them within the text. For instance, "Figure 1 shows the residuals of our model." That is the proper way to include figures in a report. **If you are concerned about the page limit, you can take advantage of the Appendix.** For instance, you can include explanations of complicated procedures there, and simply refer to the specific line or section of the Appendix in the main text.

From your report, the grader should be able to follow your entire statistical reasoning smoothly and without too much trouble. Please also take into account that your grader will be tired at some point in time, which means that if your report is evaluated at the end, the grader will probably not have the same energy as when he started grading. So, it is very important that your report is well-written and comprehensive.

The report will be graded on a scale of 0 to 100. Below, I provide a *rough* guide of the scoring of the report.

Question(s)	Points
1	10
2	30
3	10
4	15
5	5
6	20
Introduction and Conclusions	10

When grading your answers, the four main evaluation rubrics are: completeness, comprehensiveness, well-motivated, and making a correct use of statistical techniques. Therefore, you will get the full points for a question if your answer is entirely comprehensive, well-motivated and complete. Additionally, it demonstrates a correct use of statistical techniques if needed.

Presentation Guidelines

The maximum allowed time for the presentation is **5 minutes** and all team members must participate in it. The presentation should be as professional as possible, meaning that you should use a dedicated presentation software such as Microsoft PowerPoint, Apple Keynote, OpenOffice Impress, or Prezi. You should also have slides with an appealing format. You can use any recording software for your presentation. I personally recommend Zoom since we are familiar with this software. It would be good if you also appear in the video, but this is not a requirement. The most important aspect is that you clearly convey your ideas.

To watch the presentations, we will use YouTube. This implies that you must upload your presentation to that platform. Please make sure that your YouTube video is accessible from the USA. You must include the YouTube link in the corresponding row in the Google spreadsheet used to register your team.

Below, I provide guidelines on the components your presentation should *at least* have.

- **Introduction.** Engaging introduction. Something that sparks interest in the problem.
- **Methodology and Results.** Explain your design and analysis process. At minimum, you should motivate your experimental design in economical (number of runs) and statistical terms (multicollinearity), show the initial model and how discard effects from it, and final model and evaluation (residual analysis). The key here is to demonstrate how you reached

the final model that explains the cross-validation accuracy in terms of the tuning parameters. The challenge is to convey all this technical information in an easy-to-understand manner.
- **Conclusions.** Provide recommendations and highlight the strengths of your design and model.

Once again, the components mentioned above are the least a presentation must have. Your presentation should convince us that your model provides significant insights into the data and the problem.

After watching your presentation, the TAs, the readers and I will have **2 minutes** to ask one question. To this end, we will select one team member at random. So, all team members must be fully aware of the experimental design and analysis process. Example of questions are:

- Why do you think the experimental design works well?
- What is the performance of your design for studying main effects or interactions?
- What are the most important tuning parameters?
- How confident are you that the final model is working well?
- How would you explain this specific component to your boss or little brother?

The presentation will be graded on a scale of 0 to 100. Below, I provide a *rough* guide of the scoring of the presentation.

100 If the presentation is excellent: engaging, entirely clear, sticks to the time limits and rules. Additionally, your team satisfactorily answers the question.
75 If the presentation is excellent: engaging, entirely clear, sticks to the time limits and rules. However, your team does not satisfactorily answer the question.
50 If your presentation lacks one of the following aspects: engaging, entirely clear, sticks to the time limits and rules, your team satisfactorily answers the question.
25 If your presentation lacks two of the following aspects: engaging, entirely clear, sticks to the time limits and rules, your team satisfactorily answers the question.
0 If the presentation is not submitted.

We will watch the presentations and ask questions during the scheduled time for the final exam. **We will carry out all these activities through Zoom.** To this end, use the Zoom link for the lectures which is available in the Syllabus on Bruin Learn.

For Lecture 3, this is June 6 from 8 am to 11 am (PDT). For Lecture 4 this is June 9 from 3 pm to 6 pm (PDT).

Material

The designs and statistical methods seen in the course should be enough for tackling this project. If you wish, you can also use methods and concepts in the required and optional textbooks, even if we did not discuss them in class. If you use a method not seen in class, you can refer to its specific section in the textbook in the main text and include a reference in the **References** section.

Project Updates

All clarifications and updates to this project will be posted on **Campuswire**.

Report Submission

The team's contact person must submit the report to **Gradescope**.

On Gradescope, create a group submission and include all team members. Submit one report per team.