Stat 333: Applied Linear Regression

Spring 2019, UW-Madison

Project 2 Grading Guidelines

Total points: 100 points (+10 extra credit points)

Participation:

We will only grade your project 2 if (i) you work in a group of size three or four, and (ii) you turn in all the deliverables below by the deadline. If you do not meet these two conditions, your grade for project 2 will be a zero.

You are also encouraged to “unofficially” participate in the project, say by submitting Kaggle predictions, without submitting the other deliverables.

Regardless of how you decide to participate, please remember the academic rules of integrity specified below.

Deliverables and Deadlines:

|  |  |
| --- | --- |
| Deliverables | Due Date |
| Form groups on Canvas | Tuesday, Apr. 9, 2019, 11:59pm CST |
| Presentation Slides (.ppt, .pptx,.pdf) | Monday, Apr. 22, 2019, 11:59pm CST |
| R/Python Code (.Rmd or .ipynb) | Monday, Apr. 22, 2019, 11:59pm CST |
| Four-page Executive Summary (.docx,.doc,.pdf) | Monday, Apr. 22. 2019, 11:59pm CST |
| Kaggle Submission | Monday, Apr. 22, 2019, 11:59pm CST |

You need to submit the following files to Canvas:

1. Presentation Slides: Each group must submit presentation slides (in .ppt,, .pptx, or .pdf format) to Canvas by Apr. 22nd (Monday), 11:59pm CST.
2. R/Python Code: Each group must submit code used in your analysis in a **single markdown file in .Rmd or .ipynb format** to Canvas by Apr. 22nd (Monday), 11:59pm CST. You are \*\**required\*\** to use R-markdown or Jupyter Notebook to document your code.

Utilizing *Github* for collaboration and version control is \*\**strongly encouraged*\*\*

1. Executive Summary: Each group must submit an electronic copy of the executive summary to Canvas by Apr. 22nd (Monday), 11:59pm CST.

You need to submit the following to Kaggle’s website:

1. Kaggle predictions. Use the following link to participate in the Kaggle competition: <https://www.kaggle.com/t/781c5bd72b8e442fbd8aa21a86e971aa>. The class Kaggle website is here: <https://www.kaggle.com/c/uw-madison-sp19-stat333/>. Follow the instructions on the website to submit your predictions.

Please do not share the participation link to people outside of the class.

For Canvas submission, you must submit **only one** **file** for each deliverable, i.e. one file containing all your slides, one file containing all your code (no multiple code files), and one file for the executive summary.

Also, the slides, the R markdown, the executive summary, and the Kaggle predictions **cannot be changed after the due dates specified above**. It is **your responsibility** to make sure the three files are submitted on time to Canvas and Kaggle.

No late submissions will be accepted.

Groups:

You will work in groups of three or four. Each group will be responsible for (1) the presentation, (2) the two-page write-up, (3) the code that your group used to analyze the data and produce the plots in your slides and the write-up, and (4) the Kaggle predictions.

For this project, you may choose your own groups. You must indicate your group on Canvas by **Monday April 9th, 2019, 11:59pm CST**. Groups cannot be changed after this point. Also, *if you are without a group by this time, your project will not be graded.*

We encourage you to use Canvas’ discussion forum to find your group as well as form groups so that all of your members are in the same discussion section.

All members of the group must mutually agree to participate in the project. In particular, if you join an existing group without explicitly notifying AND obtaining permission from its members, you will be removed from the group.

Presentation Time & Location:

Presentations will be on Apr. 23th, 2019 (Tuesday) and Apr. 25th, 2019 (Thursday) during lecture time, 2:30pm-3:45pm, at the lecture hall, Chamberlain 2241

Presentations:

Your group will prepare a 5 minute 30 seconds, in-class presentation of your data analysis, followed by questions from the audience. The goal of the presentation is to practice presenting your statistical findings in a concise and clear manner. The presentation should include key evidence (e.g. plots, tables, inferential methods, etc.) that support your findings. Your presentation must be clear and precise enough that **any Stat 333 student** should be able to understand what statistical analysis you used and how you have reached your conclusion. The exact grading rubric for the presentation is outlined below.

Due to time constraints, the 5-minute 30 seconds time limit will be *strictly enforced*. To encourage this behavior, every additional 30 seconds after the time limit will incur a penalty of 2 points. It is **your responsibility** to rehearse your presentation so that it stays under 5 minutes and 30 seconds. It is also **your responsibility** to check with the TA that your slides can be displayed properly on the classroom projector before your presentation time (e.g. during the discussion sections the week before).

Each member of your group must speak for at least 1 minute during the presentation. All members of the group must work on the presentation and be prepared to answer questions from the teaching staff or the students.

The exact time of your group’s presentation will be determined randomly on the first day of the presentation.

R/Python Code

Each group must provide **a single .Rmd or .ipynb file** that can **replicate** your analysis from start (i.e. reading in the data) to finish (your presentation slides and two-page summary). Your code must be able to replicate every step of the data analysis, including, but not limited to: data cleaning, outlier removal, model building, evaluation of different models, statistical testing, prediction, and any and all intermediary plots, tables, and analysis. In particular, your code must reproduce the exact tables, plots, and other analysis in your summary and the presentation (i.e. exact labels for axis, color shading plots, etc.). Please refer to my lecture notes as an example of this.

All members of the group must contribute to the code. On the code, the group must clearly indicate each member’s contribution to the code, including (i) who created the code, (ii) who revised/reviewed/maintained the code, and (iii) who is/are ultimately responsible for different portions (or all portions) of the code.

Your R code also must be **well-documented** so that **any Stat 333/327 or CS 300** student can read and understand your code. This is important for reproducibility and to track down potential bugs/errors in your analysis

Finally, your code must run, from start to finish, **without producing errors** given the dataset on Canvas. In particular, if we copy/paste your code under the directory where the dataset is stored, it should not throw an error in the latest R or Python console.

We remind all students that it is **your** r**esponsibility** to make sure that the code is not copied/plagiarized/fabricated from unauthorized sources, and your code produces the identical set of results as reported in your other deliverables when it comes time for grading (see reminder below).

Four-Page Executive Summary

Your group must submit a four-page summary of the data analysis. The goal of the four-page summary is to provide a summary of your data analysis. In particular, the summary must include (i) your overall findings, (ii) relevant and important evidence for your findings (e.g. plots, tables), and (iii) important details of your statistical analysis (e.g. type of model used, inferential quantities, outliers, leverage points, modeling assumptions, etc.). Your summary should be detailed enough that any data scientist can read your summary and replicate your analysis. The exact grading rubric for the summary is outlined below.

All members of the group must contribute to the summary. On the summary, the group **must clearly indicate** each member’s contribution to the project, **including each member’s contribution to the presentation and the code.**

The summary must be typed in 12-inch Times New Roman or Sans Serif font, single-spaced, with 1-inch margins and must include all relevant figures/tables, and equations. The summary cannot exceed four pages; the list of references can exceed the four-page limit, but must be at most one page. Each group must submit **a single electronic copy** to Canvas in .pdf, .doc, or .docx format.

All figures and tables must be legible. If figures are in color, you must print the summary in color.

You may follow any reasonable guidelines for formatting the references (e.g. MLA, APA, Chicago Manual of Style, etc.)

It is your responsibility to make sure that your electronic submission adheres to these guidelines.

Kaggle

The class Kaggle website is the following:

<https://www.kaggle.com/c/uw-madison-sp19-stat333/>.

Your group must submit your predicted Yelp ratings by the due date. On Kaggle, **YOU MUST** use the team name “Group(GroupNumber)” when you submit your Kaggle predictions. Remember to replace the GroupNumber with your group number. Only one “team” is allowed per group; you may not “join” multiple teams or create additional teams on Kaggle.

The maximum number of submissions per day is 20. Please keep this in mind, especially when you are reaching the deadline.

It is your responsibility to learn how to use Kaggle and make sure the submission is properly formatted.

Note that you will be graded based on both testing AND the validation data. The public leaderboard only presents your standing based on the testing data set and is a good proxy for your performance in the validation data. The private leaderboard, which will be revealed to everyone at the end of the presentations, reveals your standings based on the validation data set.

Grading Rubric:

We will use the following grading rubric to grade your deliverables.

|  |  |
| --- | --- |
|  | Possible Points |
| Presentation | 30 |
| 1. Clear, takeaway message with a “rule-of-thumb” that is easy to use and accurate 2. Relevant, concise, and clear summary of statistical analysis 3. Relevant (no extraneous plots!) and visually accurate plots 4. Strengths and weaknesses of the analysis 5. A simple, illustrative example to demonstrate the rule-of-thumb? 6. Overall, did the group present convincing evidence for their finding? 7. Overall, was the delivery clear and easy to understand? |  |
| Executive Summary | 30 |
| 1. Introduction with clear motivation and thesis statement 2. Background information about the data 3. Motivation for the model used and statement of the model 4. Concise and relevant summary about estimation and inference of relevant parameters, which may include estimated coefficients, R^2, standard errors, confidence intervals, p-values, hypothesis testing statements, and etc. No “data/printout dump” 5. Clear, laymen’s interpretation of the estimates and inferential quantities 6. Model diagnostics and checking modeling assumptions with plots 7. Strengths and weakness of the group’s data analysis 8. Conclusion |  |
| R Code | 30 |
| 1. Replicates the tables/plots **exactly** in the presentation? 2. Replicates the tables/plots **exactly** in the summary? 3. Replicates data cleaning, model building, model evaluation, estimation, testing, and any intermediary analysis. 4. Well-documented R code for each “portion” of the data analysis? Sound variable names and concise functions for analyzing data? 5. Error free code? Does it run, without any errors, from start to finish? |  |
| Kaggle | 10 points  (10 extra credit points) |
| 1. Exceeded the benchmark on the testing data: 5 points 2. Exceeded the benchmark on the validation data: 5 points. 3. Extra credit: A group will get extra credit points based on the following scoring system.    1. The top 10 scoring groups extra credit points in descending order (i.e. 1st place🡪 10 points, 2nd place 🡪 9 points,…,10th place 🡪 1 point)   The maximum possible extra credit is 10 points. |  |

Rules and Academic Integrity

Each year in Stat 333, we have a few students who violate the academic integrity standards laid out in class. We take this opportunity to remind students of the policies regarding academic integrity.

Each student assumes the responsibilities of an active participant in UW-Madison’s community of scholars in which everyone’s academic work and behavior are held to the highest academic integrity standards. Academic misconduct compromises the integrity of the university. Cheating, fabrication, plagiarism, sabotaging other groups’ work, unauthorized collaboration, and helping others commit these acts are examples of academic misconduct. Specific examples include, but are not limited to,

1. Copying, plagiarizing, stealing, fabricating any of the deliverables, especially the code or the predictions on Kaggle, from other groups, students outside of the class, or the Internet. In particular, while you may ask other groups for general ideas and questions, you cannot ask for help cleaning the data set, analyzing the dataset, and doing other activities that would be inconsistent with the academic integrity at UW-Madison. If you are unsure, you are always welcome to ask the TA or the professor.
2. Using unauthorized sources, including the original Yelp dataset on Yelp’s website or the original ratings (or summaries of ratings of businesses) which can be derived from Yelp’s website. You are also not allowed to directly copy, steal, plagiarize, paraphrase, or use any analysis that was already conducted on the Yelp data by others (e.g. data science courses online, someone’s blog post or R markdown, Google Cloud’s API platform for sentiment analysis, any pre-written software/code that does sentiment analysis automatically, etc.).

However, you are **strongly encouraged** to browse through Yelp, resources on natural language processing (NLP), sentiment analysis, and other researchers’ analysis of the Yelp data and gather **background information**. You are strongly encouraged to use the information from your background research **to complement** your own analysis and **provide proper attributions**. In short, your analysis of the data must be **original** and **must be your own work**. Or, in industry-lingo, you should not be stealing others’ intellectual property.

If you have any questions about this, please come talk to the TA or the professor.

1. Attempting to gain an unfair advantage by recreating the original Yelp data and using predictors that are not part of the data set. You must only work with the data set you were provided with.

You are strongly encouraged to create your own predictors based on the data set you were given. Again, please come talk to the TA or the professor if you have any questions about this.

1. You may not ask someone to do any part of the analysis on your behalf.

Committing said acts can result in disciplinary action, which includes, but is not limited to failure on the assignment/course, disciplinary probation, or suspension. Substantial or repeated cases of misconduct will be forwarded to the Office of Student Conduct & Community Standards for additional review. For more information, refer to [students.wisc.edu/student-conduct/academic-integrity/](https://students.wisc.edu/student-conduct/academic-integrity/).