Project 2

Total points: 100 points (+ extra points)

Presentation Time & Location:

Presentations will be on Apr 21, Apr 23, Apr 28 during lecture time at the lecture hall.

Timeline:

**11:59pm Wednesday Mar 11, 2020:** self chosen group assignment submitted to the instructor by email.

**In class Thursday Mar 12, 2020:** random group assignments ready for students without self chosen groups. Group presentation order will be randomly generated in class and posted on Canvas.

**11:59pm Monday, Apr 20, 2020:** Kaggle submission deadline. Slides and R codes submission deadline (email to TA).

**Apr 21 Tuesday, Apr 23 Thursday, Apr 28 Tuesday, 2020:** In class presentation.

**5pm Apr 28 Tuesday, 2020:** Write-up submission deadline (to Canvas).

Data Background:

Yelp is an Internet company founded in 2004 to “help people find great local businesses” by providing a platform for users to write reviews of businesses. It holds a data challenge regularly (<https://www.yelp.com/dataset/challenge>) to “challenges students to use this data in an innovative way and break ground in research” and they provide cash rewards for the winners.

The raw data can be downloaded from <https://www.yelp.com/dataset/download>. This data contains 6.7 million reviews of 193K businesses from more than one thousand cities. Our class project will focus on (i) finding out what makes a review positive or negative and (ii) predicting a review’s rating based on its text and a small set of relevant attributes. To make the data more manageable, we have narrowed the original Yelp challenge dataset to restaurants of the Madison area. Specifically, our subset consists of almost-all English Yelp reviews of Madison area businesses with the tag “Restaurants” as part of its category values. Only reviews that were recommended at the time of the data collection are included (see link here for how Yelp recommends reviews to its users <https://www.yelp-support.com/Recommended_Reviews>).

In total, the dataset contains 92,236 reviews of 1,361 Madison businesses.

We have randomly split the data into three pieces whereby 60% of the subset is training data, 20% of the subset is reserved for testing, and the last 20% is validation data.

We include the following predictors for you to use in your data analysis. It is some starting set of predictors based on the review text. However, you are **strongly encouraged** to come up with new predictors based on the predictors we have given you, especially the review text. However, **the total number of predictors you use in your final model cannot exceed 3000**.

|  |  |
| --- | --- |
| Name | Description |
| stars | Stars of the review (1=worst, 5=best). The value only takes integers between 1 and 5. This is only available in the training data. |
| Id | Id number of the review. |
| text | Review text |
| date | Date of review |
| useful | Number of Yelp users who thought the review was useful. This value takes whole numbers (i.e. 0,1,2,…) |
| funny | Number of Yelp users who thought the review was funny. This value takes whole numbers (i.e. 0,1,2,…) |
| cool | Number of Yelp users who thought the review was cool. This value takes whole numbers (i.e. 0,1,2,…) |
| city | City in which the business is located |
| categories | An un-parsed string that categorizes the businesses into Yelp’s categories. See the full list here: <https://www.yelp.com/developers/documentation/v3/category_list> |
| nchar | Number of characters in the review |
| nword | Number of words in the review. |
| sentiment | A score of the text’s sentiment using the AFINN lexicon, which ranges from -5 (very negative) to 5 (very positive). See details here: <https://www.tidytextmining.com/sentiment.html> |
| gem | Number of times the word “gem” appears in the review text. |
| incredible | Number of times the word “incredible” appears in the review text. |
| perfection | Number of times the word “perfection” appears in the review text. |
| BLANK | Number of times the word “BLANK” appears in the review text. |

Groups:

You will work in **groups of two students (preferred) or individually**. Each group will be responsible for (1) the presentation (2) the four-page write-up, and (3) the code that your group used to analyze the data and make rating predictions.

For this project, you may choose your own groups. You must report by email to the instructor the groups you have chosen by **11:59pm on** **Mar 11, 2020**. If you are without a group by this time, the instructor will randomly assign you to a group in class. The presentation order will be randomly assigned and available in class on Mar 12, 2020.

You are strongly encouraged to form your own groups due to the grading scheme for the course. It is also in your interest to form a group with diverse backgrounds, say majors from CS, English, and/or linguistics. You should use Piazza’s discussion forum to find your group.

Deliverables:

For the presentations, each group must **e-mail to the TA** the presentation slides (in .ppt,, .pptx, .pdf, or html format) before the first day of presentations. It is your responsibility to check with the TA in advance that your slides can be displayed properly on the classroom projector before your presentation time. Once submitted, the slides **cannot be changed** during the presentation days.

For the R codes, your group must submit **a single .R or .Rmd file** to the **TA’s e-mail.** It is your responsibility to make sure that the code is well-documented, runs without any error, is not copied/plagiarized/fabricated from unauthorized sources, and produces the identical set of results as reported in your other deliverables when it comes time for grading. **No other programming languages are allowed** in the entire data analysis including data preprocessing (consult with the instructor if you have confusion on this).

For the four-page write-up, **each individual** student must submit **one electronic copy in .pdf format** to Canvas by the due time. The electronic copy should **include your rating on your peer group members**.

Finally, your group must submit your final predictions to Kaggle. It is your responsibility to make sure you can submit to Kaggle before the deadline.

Presentations:

Your group will prepare a **five-minute**, 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 five-minute time limit will be *strictly enforced*. To encourage this behavior, every additional 30 seconds after five minutes will incur a penalty of 1 point. While we will provide time cards and warnings for you during the presentation, it is ultimately **your responsibility** to rehearse your presentation so that it stays **under five minutes**.

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

Four-Page 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, equations, and references. The summary must not exceed four pages. All figures and tables must be legible with proper labels and fonts.

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

Kaggle

The class Kaggle website is the following:

<https://www.kaggle.com/c/uwmadison-2020spring-stat333/>

**Each group should only use one ID to submit to Kaggle. The ID should be your group ID number.**

The ranking on Kaggle is purely based on prediction performance measured using Root Mean Squared Error (RMSE).

Your group must submit your predicted Yelp ratings for the testing and the validation data before the first day of the presentation. 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 leader board only presents your standing based on the validation data set and is a good proxy for your performance in the testing data. The private leader board, which will be revealed to everyone at the end of the presentations, reveals your standings based on the validation data set.

Rules and 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/).

Grading Rubric:

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

|  |
| --- |
| Presentation (35 points) |
| 1. Clear, takeaway messages (5 points) 2. Relevant, concise, clear, and understandable summary of statistical analysis and plots (10 points) 3. Statistically correct and interpretable model(s) with an understanding of its strengths and weaknesses? (8 points) 4. Overall, did the group present convincing evidence for their finding? (6 points)   Overall, was the delivery clear and easy to understand? (6 points) |
| Write-up and R Code (50 points) |
| 1. Introduction, background information, and thesis statement (5 points) 2. Motivation for the model(s)used and statement of the model(s) (5 points) 3. 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” (12 points) 4. Clear, laymen’s interpretation of the estimates and inferential quantities (8 points) 5. Correct and interpretable model with an understanding of its strengths and weaknesses by checking model assumptions and using model diagnostics (8 points) 6. Conclusion (2 points) 7. R Code runs without error and produces the results stated in the write-up and the presentation (10 points) |
| Peer rating (5 points) |
| You will rate each of your group members with a score from 0 to 5 based on their participation in the group project. 5=full participation. 0=no participation. This rating should be included in the electronic copy of your write-up submitted to Canvas. |
| Kaggle (10 points + extra) |
| 1. Exceeds the benchmark on both the testing and validation data (5 points) 2. Achieve RMSE of <0.85 on both the testing and validation data (5 points) 3. 5 extra credit points will be awarded based on group’s ranking on the validation data. 5 extra points for the 1st group, 4.5 extra points for the 2nd group, 4 extra points for the 3rd group, and so on to 0.5 extra point for the 10th group. No extra points for groups ranking below the 10th.   You will not receive any of these “ranking-based” points unless (i) the benchmarks (a)-(b) are met, (ii) your R codes can run without errors, and your R codes can yield the same results as your write-up, slides, and the submission that leads to your ranking on Kaggle, (iii) the total number of predictors you use in your final model is below 3000.  The extra points will be added to the final grade of the course before curving. |