Stat 333: Applied Linear Regression

Spring 2019, UW-Madison

Project 2 Description

Motivation

Consider the following reviews from Yelpers reviewing the restaurant 1847 At the Stamm House in Madison.

“*We've been waiting for 1847 Stamm House to open. Glad to miss the opening kinks being worked out some reviews mentioned. Went last night, and it was fabulous!! The fish fry was a new take on a Wisconsin fav, a whole piece of whitefish with delicious cole slaw and picked sides. All amazing! Can't wait to return to try the rest of the menu. Maybe on a night that's quiet, if there is one. Glad to see the place was hopping!”*

*“I was visiting from out of state and had this place recommended for a Friday night fish fry. What a mistake. I was told over the phone that the Cod fish fry was $15.I person we were told the Cod cost $16 and the Bluegill was $18. When the bill came I was informed that the Cod was $18. For one lousy piece of Cod. Outrageously overpriced. I also ordered a Moscow Mule. That was served in a glass, instead of the customary copper mug.”*

*“I too am hoping that this unique and beautiful restaurant works out the kinks and begins to perform up to the high expectations. My wife and I experienced the same problems during week #1 that the other reviewers have described (e.g. the relatively expensive burger and not-so-good white fish). We loved the atmosphere, albeit a bit too noisy, but the food and the service didn't measure up to the competition that day. I suggest the chef and servers visit Brasserie V to experience another way of doing things! Their best burger is actually worth $12! We are looking forward to our next meal at the Stamm House and are sure we will experience a marked improvement in all aspects of the dining experience!”*

It’s easy for humans to understand the *sentiment* behind the reviews, specifically whether a review is positive or negative, even without having any context about the business. In fact, it’s pretty easy for a human to guess the Yelp ratings based on the text alone. In contrast, for machines, understanding the sentiment of a text is difficult because it requires understanding not only the underlying language structure, but also the meaning it conveys. In fact, this is an active area of research in a field known as “sentiment analysis.”

Goals:

The two goals for this project are:

1. Find out **what** makes a review positive or negative based on the review and a small set of attributes by using a variety of interpretable models.
2. Propose **a prediction model** to predict the ratings of reviews based on the text and related attributes and submit the predictions to Kaggle (see grading description document).

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. As of Q4 2016, there are more than 121 million reviews and 89 million unique (average) monthly visitors have visited its website. Yelp utilizes automated software (<https://www.yelp-support.com/article/What-is-Yelp-s-recommendation-software?l=en_US>) to filter out bad reviews.

Recently, Yelp has released some of its 171 million reviews to the public as a part of the “Yelp Dataset Challenge.” This data contains ~6.7 million reviews from ~193K businesses from the following cities: Edinburgh (U.K.), Karlsruhe (Germany), Montreal (Canada), Waterloo (Canada), Pittsburgh (U.S.), Charlotte (U.S.), Urbana-Champaign (U.S.), Phoenix (U.S.), Las Vegas (U.S.), Madison (U.S.), Cleveland (U.S.). Yelp “challenges students to use this data in an innovative way and break ground in research” and they provide cash rewards for the winners.

To make the data more manageable for Stat 333, we have narrowed the original Yelp challenge dataset to restaurant or food businesses that are open and located in the state of Wisconsin and reviews that are written in English. Also, the businesses are included in the data must have at least 3 reviews older than 14 days. Also, 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, you will be analyzing 85,543 reviews, of which 51,326 reviews are for training and the rest are for testing/validation.

Training, Testing, and Validation Data

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.

Variable Names and Predictors

We include the following 41 predictors for you to use in your data analysis.

|  |  |
| --- | --- |
| Name | Description |
| Id | Id number of the review. |
| star | 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. |
| name | Name of the business |
| city | City in which the business is located |
| postal\_code | Postal zip code of the business |
| text | Review text |
| nchar | Number of characters in the review |
| nword | Number of words in the review. |
| gem | Number of times the word “gem” appears in the review text. |
| incredible | Number of times the word “incredible” appears in the review text. |
| die | Number of times the word “die” appears in the review text. |
| BLANK | Number of times the word “BLANK” appears in the review text.  There are a total of 500 words. |

As you can see, we provided you with 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.

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/).