EC349 Yelp User Review Prediction

Chin Howe Tsai | u2105215

2023-12-04 | wordcount: 1023/1250

GitHub Link (https://github.com/Gyro007/Yelp-Review-Prediction)

Tabula Statement

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements. In submitting my work I confirm that:

1. I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
2. I declare that the work is all my own, except where I have stated otherwise.
3. No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
4. Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
5. I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
6. Where a proof-reader, paid or unpaid was used, I confirm that the proofreader was made aware of and has complied with the University's proofreading policy
7. I consent that my work may be submitted to Turnitin or other analytical technology. I understand the use of this service (or similar), along with other methods of maintaining the integrity of the academic process, will help the University uphold academic standards and assessment fairness.

Privacy statement

The data on this form relates to your submission of coursework. The date and time of your submission, your identity, and the work you have submitted will be stored. We will only use this data to administer and record your coursework submission.

- Related articles
- Reg. 11 Academic Integrity (from 4 Oct 2021)
- Guidance on Regulation 11
- Proofreading Policy
- Education Policy and Quality Team
- Academic Integrity (warwick.ac.uk)

Introduction

Yelp is an American company headquartered in San Francisco which publishes user reviews about businesses in the United States of America. The objective of this project is to predict the number of stars user i gives to business j based on 5 Yelp datasets. The next section covers the DS Methodology, followed by challenges faced, model and justification, results and conclusion.

DS Methodology (150 Max)

John Rollin's General DS Methodology was used as it bears advantages such as incremental learning. This is due to its strong focus on iterating through its DS cycle. Rollin's methodology thus increases business value delivery to Yelp as the use of data science to create recommender systems is often time critical against competitors who seek to achieve similar outcomes. In phase 1, the problem of maximising prediction accuracy of the amount of "stars" was identified and it was ascertained that predictive analysis should be used. In phase 2, it is ascertained that check_in_data and tip_data can be forgone as only business_data, user_data_small and review_data_small contained variables of interest. As many variables were string responses, data cleaning is needed to turn them numeric. In phase 3, models of sentiment analysis and random forests were selected and the variables combinations were iterated until the best accuracy was found.

Most Difficult Challenge (200 Max)

While running the code, two challenges emerged.

First, due to the additional computations required for the random forest model, the time taken for the code to run was significantly longer than a simple linear regression model. However, while iterating through the process of finding the best combination of variables, I realised the accuracy of both models seem correlated, although the random forest model outperformed the other. I capitalised on the short running time of the simple regression model at each iteration until the accuracy made significant improvements and used it as a signal to warrant prediction using random forest.

Second, in view of the poor computing power of my laptop and being unable to access library computers as I am at home undergoing medical intervention for an injured knee. I greatly reduced the training and test sample sizes to 2000 and 1000 (2:1 train-test-split) respectively and only raised them 10-fold when significant improvements were obtained in the model's accuracy.

Model and Justification

Data Cleaning

Following Rollin's DS Methodology, data cleaning was conducted to obtain a set of usable data. From user_data_small, the variables, "useful", "funny", "cool", "fans", "average_stars" and "review_count" were extracted and renamed to reflect their association to each user. While the average rating of users had accorded to multiple businesses mirrors the strictness of users in determining the quality of their experience, variables like "review_count", "fans", "useful", "funny" and "cool" provided some insight on whether the reviews were reliable.

From business_data, the variables "stars", "review_count", "attributes.RestaurantsPriceRange2" and "attributes.NoiseLevel" were extracted and renamed to reflect their association to each business. While "stars" and "review_count" demonstrated the reputation of a business, "attributes.RestaurantsPriceRange2" and "attributes.NoiseLevel" displayed characteristics that had prominent economic effects on user reviews. Price served as a form of signal to consumers and affected their demand for a restaurant while noise pollution are externalities that can ruin consumer experience. Furthermore, as some variables were recorded in Booleans, each Boolean was accorded a particular value to convert the data into numeric form for analysis.

From review_data_small, the variables "text", "useful", "funny" and "cool" were kept while "stars" were used as the dependent variable for prediction. These "useful", "funny" and "cool" variables in this dataset reflected the reliability of a particular review, which has starkly different meaning from those in the user_data_small dataset. Sentiment

analysis was subsequently conducted on “text”.

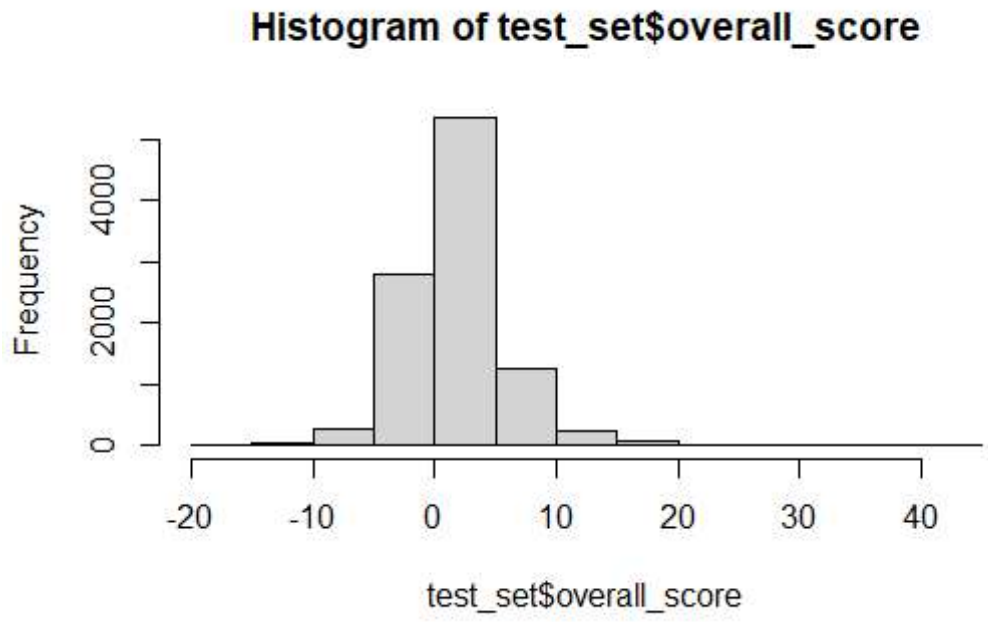
Train-Test-Split

To prepare the training and test data, I combined the 3 modified datasets by matching the business_id and user_id. Empty values and “None” responses were removed from the merged_data. With a total of slightly over 30000 observations in the merged_data, a train-test-split of 2:1 was conducted with 20000 observations for training and 10000 observations for test.

Sentiment Analysis

In view that there are large chunks of texts, sentiment analysis was conducted. The texts separated by words and augmented through techniques such as stop words removal and filtered through the bing lexicon to determine whether every word conveyed positive or negative emotions. The cumulative sentiments for each review_id was computed into an overall_score which represented the extent of positivity/negativity of user i’s review of business j.

A histogram of the sentiment distribution is shown below:



The final set of variables used are as follows:

S/N	Variable
1.	overall_score
2.	useful
3.	funny
4.	cool
5.	fans
6.	user_review_count
7.	user_useful
8.	user_funny

S/N	Variable
9.	user_cool
10.	user_leniency
11.	business_rating
12.	business_review_count
13.	attributes.RestaurantsPriceRange2
14.	attributes.NoiseLevel

$$Y_{train} = \sum_{i=1}^{p} \text{stars}_i$$

$$X_{train} = \sum_{i=1}^{p} (\text{overall\_score}_i + \text{useful}_i + \ldots + \text{attributes.NoiseLevel}_i)$$

Modelling

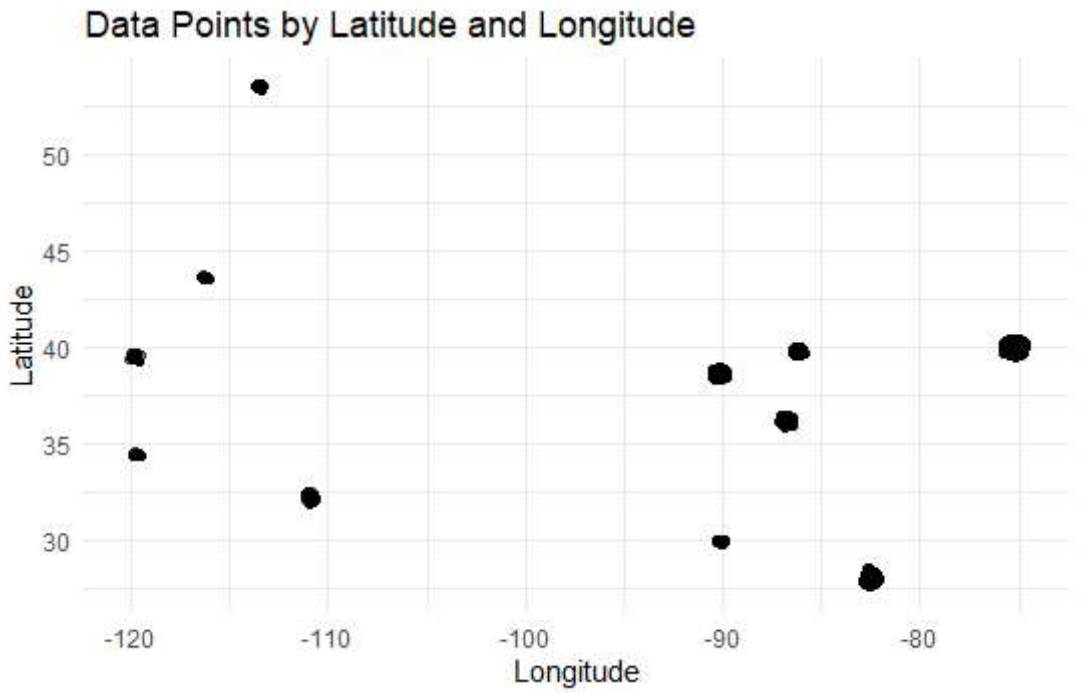
Random Forests is a special type of decision tree. It selects random subsets of covariates from the training set X_train to split the tree. It then attempts to find a particular variable j from the subset drawn that has the best splitting point. The splitting continues until a stopping criteria is reached. This process minimizes the prediction error while reducing variance in predictions as it makes predictions uncorrelated by randomly selecting subsets. The Random Forest process is represented by the equation below:

$$min_{j,s} \sum_{i:x_1 \in R_1(j,s)} (Y_i - \hat{Y}_{R_1(j,s)})^2 + \sum_{i:x_2 \in R_2(j,s)} (Y_i - \hat{Y}_{R_2(j,s)})^2 + \ldots + \sum_{i:x_m \in R_m(j,s)} (Y_i - \hat{Y}_{R_m(j,s)})^2$$

where the subset M contains m number of covariates $<= p$ which is the total number of variables in the cleaned dataset.

Results and Conclusion

From the results, accuracies of ~ 48% were observed. While iterating the process with different combinations of variables, it was found that sentiment scoring, restaurant's price and a user's strictness in scoring had strong effects on the accuracy of the model.



Originally, K-means clustering was used to conduct spatial analysis of the restaurants where a new feature “cluster” was created and each business_id being accorded a value based on their proximity to centroids. Given the latitude and longitude of the restaurants, 11 clusters were clearly observed in a spatial plot (as shown above). However, not only did clustering not improve the accuracy of the model, it reduced it by ~ 3% when tested with random forests. As such, the latitudes, longitudes and cluster labels were ultimately excluded from the model. However, further studies may wish to expand on this concept, by first grouping the business_data by “city” and conducting city-specific K-means clustering. This might create more meaningful interpretations and influence the accuracy of prediction models.