2123515_ec349_rmd

2023-12-05

Academic integrity 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)

This is the end of the statement to be included.

Link to Github project: https://github.com/shauryagiri5/EC349_assignment

Word Count: 1151 words

Introduction:

This project aims to use data from Yelp to robustly and accurately predict the amount of "stars" given by user i to business j in a review on the website. In order to do so, data into the users themselves and the characteristics of the reviews they post would intuitively prove to be valuable. Through the process of building this model, this intuition will be tested.

The Data Science Methodology used to engage with this project is the CRISP-DM (IBM, 2017) Framework (Cross-Industry Standard Process for Data Mining). The phased nature of this methodology provides a framework to turn this potentially convoluted project into an intuitive step-by-step procedure. Moreover, its cyclical nature allows for potential improvements to be iterated and evaluated rapidly, especially if it is applied loosely alongside principles of Agile. Given the unique nature of this project in its academic context, features such as the time-consuming and redundant documentation can be discarded to best fit the project's requirements. This ease and adaptability makes it effective for this project.

Data Understanding:

Due to computational limits on the device used to complete this project, the large database with approximately 7 million reviews could not be used. Instead two smaller files entitled "yelp_review_small" and "yelp_user_small" were used in order to complete this project. As the names of these datasets imply, they contained data detailing reviews as well as the users.

The dataset "yelp_review_small" contains details of 1398056 reviews, including details such as the text of the review, the amount of stars it received, the opinions of other Yelpers etc. A comprehensive list of the variables and their explanations can be found in Table 1 of the appendix.

The dataset "yelp_user_small" contains details of 397579 users. Its key features include the number of reviews by a user, other users' opinions of their reviews, and their average star rating of their reviews, as Table 2 details.

The majority of the data included in these datasets is numeric. However, two variables are not directly numerical in nature. The first is the "text" variable. Since it is the text of the review, processes would have to quantify it. The second variable to consider is the target variable "stars" itself, as it can be a considered a categorical variable. However, unlike most categorical variables, each "category", or stars ranking, can be assigned a numerical value (for example, values of 0-5 for zero stars to 5 stars), thereby providing a logical reason to consider it a numeric variable.

Data Exploration and Preparation:

Given that the majority of the variables available are numerical, in order to facilitate analysis, it may be worthwhile to consider converting the 'text' variable into a numeric measure. A simple numeric measure can be created simply by counting the number of words in the review text. However, this does not provide much insight into the character of the review. This can be achieved using the Bing Sentiment Lexicon, a dataset that categorises English words as either positive or negative (CRAN, 2004). This allows for a quantification of the nature of a text review, transforming the characters to a numeric variable, allowing for a numerical exploration of the relevant variables.

From Figure 1, it can be seen that "stars" does not follow a normal or a uniform distribution. There are significantly more 4 and 5 star reviews than 1, 2 or 3 star reviews.

From the correlation matrix in Table 3, it can be seen that there are variables across both datasets that have significant correlation coefficients with the target "stars" variable. These variables are "average_stars", "sentiment_score" and "text_count". This indicates that these variables may be useful in an attempt to model "stars".

Modelling:

From the data exploration, it can be seen that "stars" is a a variable that may not be normally or uniformly distributed, and has correlation with a plethora of variables. Thus, in order to model for "stars", a modeling method that is versatile, adaptable and computationally exhaustive is a necessity.

Given the nature of the data and the computational power available, a tree regression emerges as a strong option. Given that the data is numerical, i.e. continuous, using a tree regression would allow for an intuitive solution. In order to mitigate potential overfitting, a pruning mechanism can be implemented, which can be tuned using a cross-validation function. After filtering a merged and processed dataset (removing null values, quantifying text etc), a sample of 100000 observations is taken. This sample is further split into a training dataset of 80000 and a testing dataset of 20000 observations. An implementation of this method can be seen in the decision tree depicted in Figure 2. In order to understand model validity and accuracy, the measures of r^2 and Root Mean Standard Error (RMSE) are used. Through iteration and consideration of the correlation matrix, the optimum choice of variables is achieved.

Model Evaluation:

The most efficient version of this model derived has an RMSE of 0.957 and an r^2 of 0.578. These numbers imply that the model explains 57.8% of the variance in the true star ratings, based on the features extracted from the review and the user. These values also suggest that the model has avoided the typical tendency of tree regression over-fitting. The resultant decision tree also has significant insights. Whilst multiple variables such as 'useful.x' were included in the initial tree model code, their omission from the resultant decision tree demonstrates their limited "informativeness". This is in line with the weaker coefficient demonstrated by this variable. This points to a limit to the explanatory power of a model built only using user and review data.

In comparison, a traditional linear regression using the same variables has a RMSE of 1.03 and r^2 of 0.513, thereby indicating a lower explanatory power with higher error for the linear regression. This does serve as a demonstration of the comparative merit of the tree regression model

Conclusion:

Thus, it can be said that a tree regression can predict the majority of the amount of "stars" user i would award to business j on Yelp, with an error of less than a star using data on the review and characteristics of the review.

However, the fact that not all of the available variables in the Yelp dataset are suitably incorporated indicates that the explanatory power and robustness of this model could be improved. Potential improvements could include looking at the other associated datasets such as the business data.

The biggest challenge of the project was incorporating sentiment analysis from the review text. As the report demonstrates, the resultant variables have some of the most significant correlation coefficients. However, the process of integrating them was not simple as it required syntax and libraries from unfamiliar libraries in R. Other papers such as Rafay, Suleman and Alim (2020) create thier own dictionaries. This was unfeasible for this project. Thus, it involved a learning curve in order to create efficient code.

References:

- CRAN (n.d.). R: Bing sentiment lexicon. [online] search.r-project.org. Available at: https://search.r-project.org/CRAN/refmans/textdata/html/lexicon_bing.html.
- Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04, [online] pp.168–177. doi:https://doi.org/10.1145/1014052.1014073.
- IBM (2017). IBM SPSS Modeler CRISP-DM Guide. [online] www.ibm.com. Available at: https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=spss-modeler-crisp-dm-guide.
- Rafay, A., Suleman, M. and Alim, A. (2020). Robust Review Rating Prediction Model based on Machine and Deep Learning: Yelp Dataset. [online] ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9080713.
- Therneau, T.M., Atkinson, E.J. and Foundation, M. (2015). An Introduction to Recursive Partitioning Using the RPART Routines. [online] Available at: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Appendix:

Table 1: dataset overview of "yelp_review_small"

Note: all variables considered in tree regression in bold

Name of Variable	Variable description	Variable type
review_id	Unique identity code for each review	Characters
user_id	Unique identity code for user writing the review	Characters
business_id	Unique identiy code for the business the review pertains to	Characters
stars	The amount of stars given by the user to the business	Categoric/Numeric
useful.x	Amount of other users who clicked that found the content of the review useful	Numeric
funny.x	Amount of other users who clicked that found the content of the review funny	Numeric
cool.x	Amount of other users who clicked that found the content of the review cool	Numeric
text	The actual text of the review	Characters
date	The date when the review was published	Date type

Table 2: dataset overview of "yelp_user_small"

Note: all variables considered in tree regression in bold

Name of Variable	Variable description	Variable type
user_id	Unique identity code for user writing the review	Characters
name	First name of user	Characters
review_count	Number of reviews written by user	Numeric
yelping_since	Date user joined Yelp	Date type
useful.y	Amount of other users who have found the content of the reviews of this user useful	Numeric

Name of Variable	Variable description	Variable type
funny.y	Amount of other users who have found the content of the reviews of this user funny	Numeric
cool.y	Amount of other users who have found the content of the reviews of this user cool	Numeric
elite	List of years for which the 'Elite' review status was given to the user	List
friends	List of user IDs of "friends" of the given user on Yelp	List
fans	Amount of 'fans' of the user's reviews	Numeric
average_stars	The average stars given by the user across their reviews	Numeric
Compliment variables	Amount of people who provide compliments detailing adjectives such as 'hot', 'more reviews' etc	Numeric

Table 3: Correlation matrix of key user and review variables with "stars"

	useful.x	funny.x	cool.x	useful.y	funny.y	cool.y	review_count	average_stars	fans	stars	text_count	sentiment_score
useful.x	1.00000000	0.58310058	0.72463156	0.44884116	0.42616700	0.43808102	0.27643639	-0.01572350	0.35803940	-0.07918287	0.29076387	0.08235807
funny.x	0.58310058	1.00000000	0.67411135	0.48653788	0.51567631	0.48578640	0.25033576	-0.00165345	0.35527900	-0.04623929	0.18255315	0.04583128
cool.x	0.72463156	0.67411135	1.00000000	0.60591645	0.58515794	0.60743210	0.33346488	0.06722497	0.44440865	0.08282198	0.17739373	0.16375630
useful.y	0.44884116	0.48653788	0.60591645	1.00000000	0.95338581	0.98953443	0.61163249	0.03121432	0.74387533	0.02126865	0.11609577	0.08231428
funny.y	0.42616700	0.51567631	0.58515794	0.95338581	1.00000000	0.95290391	0.50548918	0.02346491	0.71375053	0.01659563	0.09013079	0.05655262
cool.y	0.43808102	0.48578640	0.60743210	0.98953443	0.95290391	1.00000000	0.52008805	0.03053375	0.67148781	0.02092980	0.09333864	0.06617922
review_count	0.27643639	0.25033576	0.33346488	0.61163249	0.50548918	0.52008805	1.00000000	0.04924428	0.66128135	0.03216652	0.17583606	0.15361285
average_stars	-0.01572350	-0.00165345	0.06722497	0.03121432	0.02346491	0.03053375	0.04924428	1.00000000	0.04356634	0.58371310	-0.10486868	0.31679600
fans	0.35803940	0.35527900	0.44440865	0.74387533	0.71375053	0.67148781	0.66128135	0.04356634	1.00000000	0.02724294	0.14205835	0.10990991
stars	-0.07918287	-0.04623929	0.08282198	0.02126865	0.01659563	0.02092980	0.03216652	0.58371310	0.02724294	1.00000000	-0.19911001	0.47927911
text_count	0.29076387	0.18255315	0.17739373	0.11609577	0.09013079	0.09333864	0.17583606	-0.10486868	0.14205835	-0.19911001	1.00000000	0.30134055
sentiment_score	0.08235807	0.04583128	0.16375630	0.08231428	0.05655262	0.06617922	0.15361285	0.31679600	0.10990991	0.47927911	0.30134055	1.00000000

Figure 1 : Distribution of "stars" in sample_data

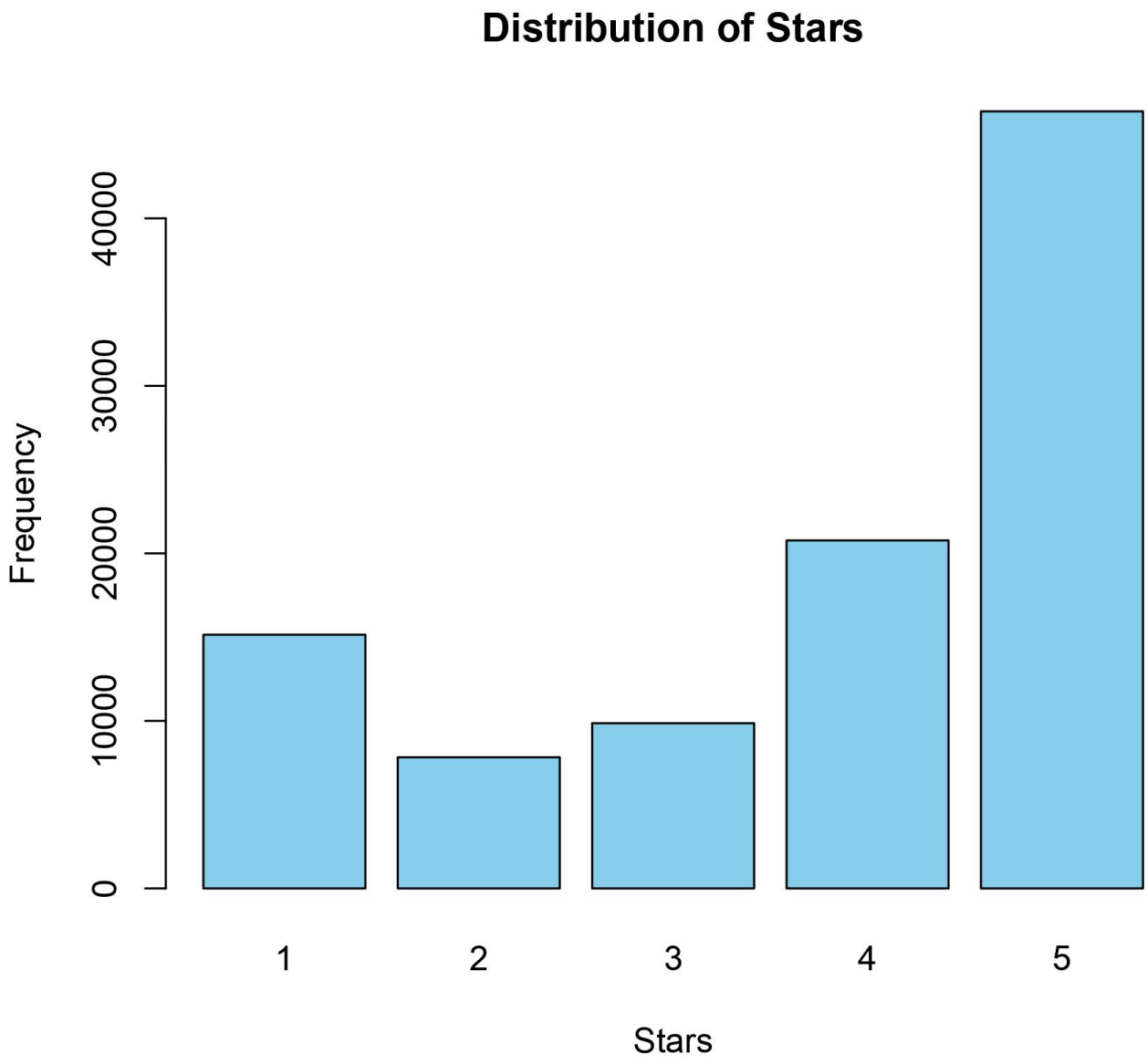
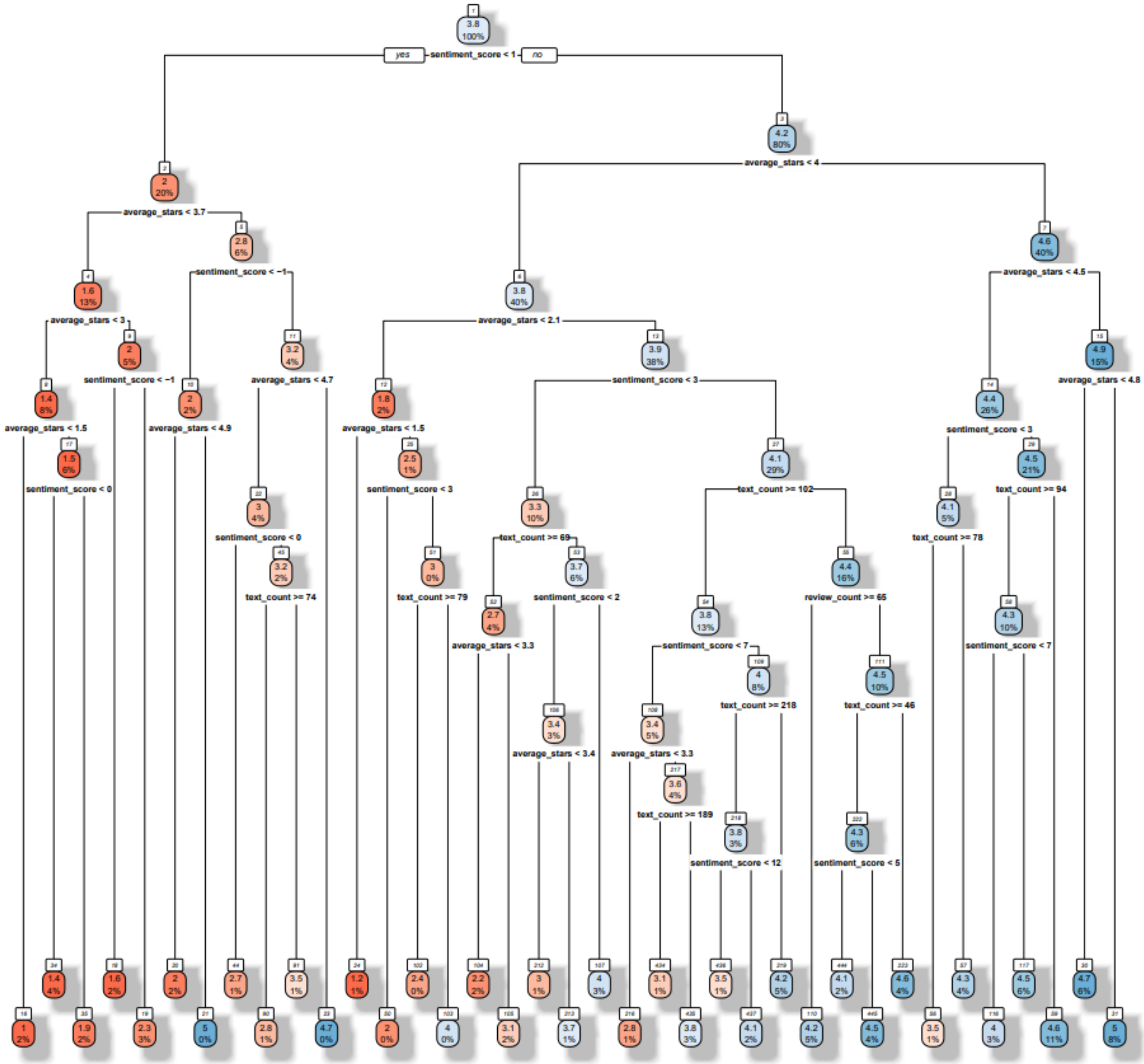


Figure 2: Tuned Decision Tree


yes
sentiment_score < 1
no
average_stars < 4
average_stars < 3.7
sentiment_score < -1
average_stars < 4.5
average_stars < 3
average_stars < 2.1
sentiment_score < -1
average_stars < 4.7
sentiment_score < 3
average_stars < 4.8
average_stars < 1.5
average_stars < 4.9
average_stars < 1.5
sentiment_score < 3
sentiment_score < 0
sentiment_score < 3
text_count >= 102
text_count >= 94
sentiment_score < 0
text_count >= 69
text_count >= 78
text_count >= 74
text_count >= 79
sentiment_score < 2
review_count >= 65
average_stars < 3.3
sentiment_score < 7
sentiment_score < 7
text_count >= 218
text_count >= 46
average_stars < 3.4
average_stars < 3.3
text_count >= 189
sentiment_score < 12
sentiment_score < 5