Assignment 1 EC349

Sören Twietmeyer

2023-12-04

EC349 Assignment 1: Analysis of Yelp Data

Note: The code can also be found on GitHub (https://github.com/TwietmeyerSoeren/EC349)

Outline of the goal

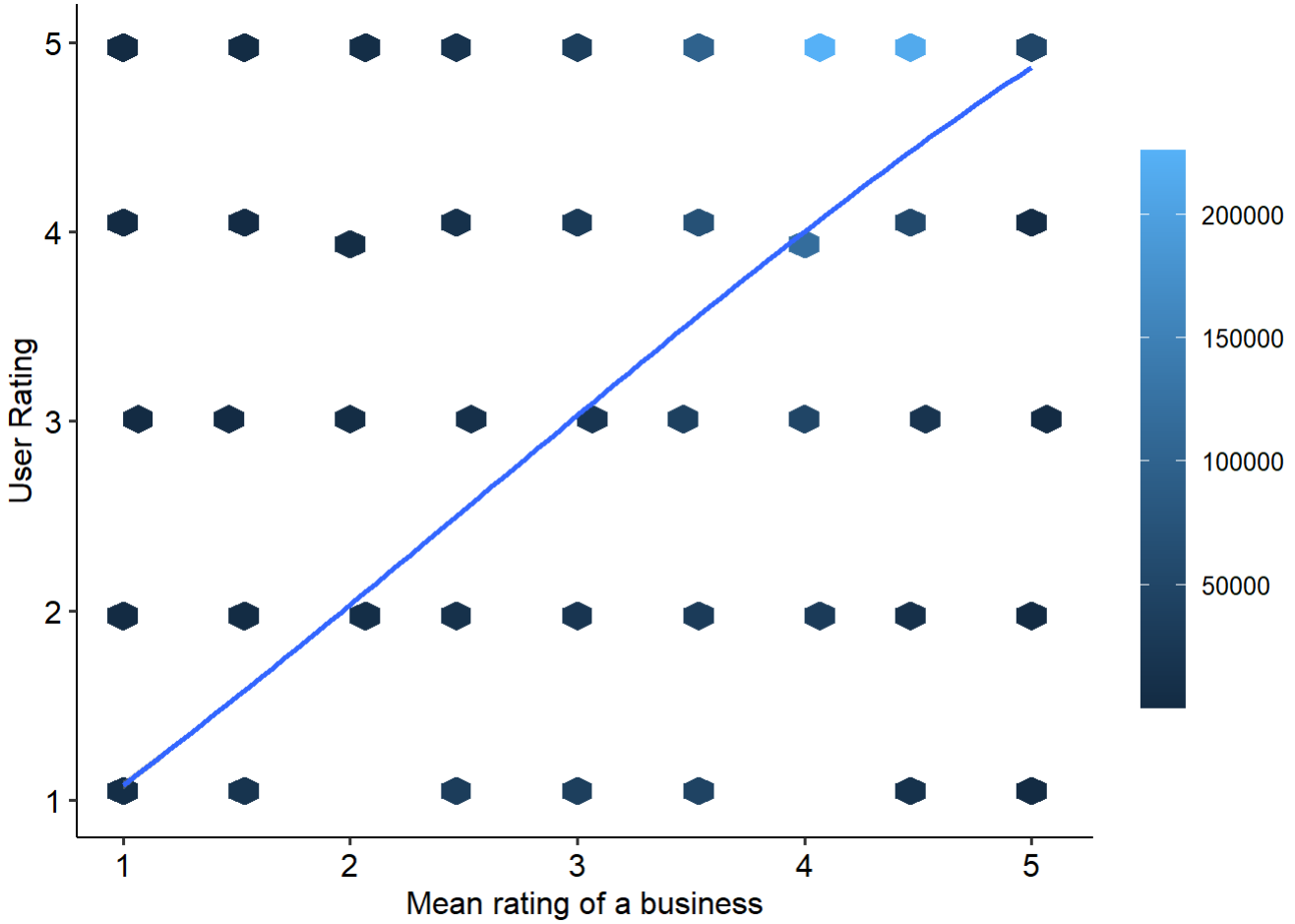
This project is aiming to predict user ratings on Yelp using data on the reviews submitted, the user, the business as well as data on additional comments made about a particular business. The goal is to achieve a precise prediction given the information available and does not require an evaluation of the reviews themselves, i.e., why users gave a specific review. This will allow us to trade off interpretability of our models for gains in performance.

Methodology

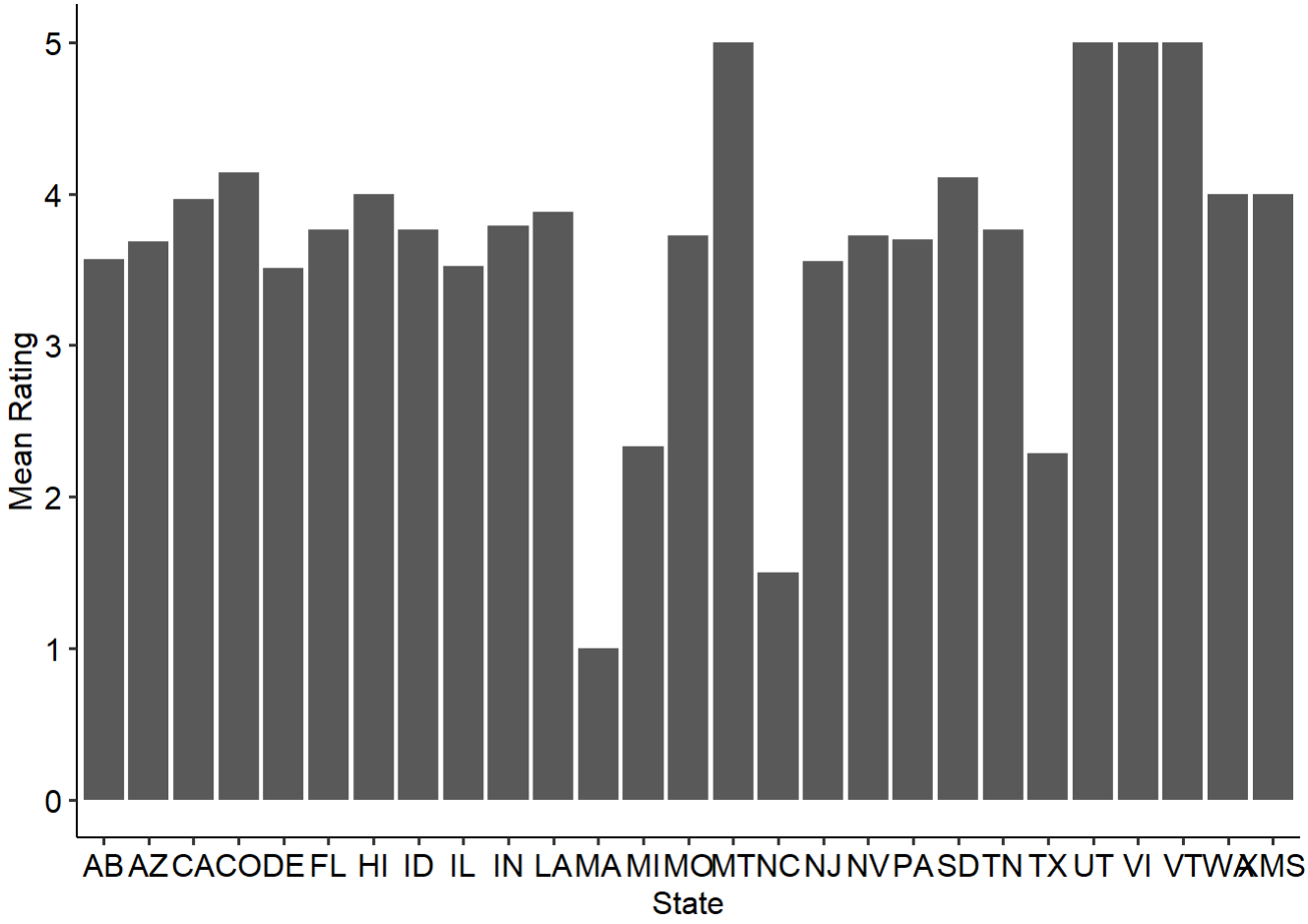
This data science project requires a methodology that is iterative, intuitive and suitable for an individual project. Given the absence of a business case, steps included in methods such as CRISP-DM relating to the understanding of the business are only marginally relevant. Additionally, CRISP-DM is documentation heavy, which can be helpful in longer and larger projects but is less required for an individual project. The John Rallin DS methodology, TDSP and the KDD methodology are very similar in structure and contain irrelevant parts regarding the business understanding and stakeholder involvement. The OSEMN methodology is the most appropriate since it only covers the core parts of the aforementioned methodologies and is well suited for individual projects which require less documentation. Nevertheless, I have adapted the structure to include a phase for a clear problem definition. This is crucial despite the absence of other parties with interest in the project.

Data Exploration

The data exploration phase showed that the business data set includes useful information, however, some columns suffer from large shares of missing data, which negatively impacts its usefulness. Having a lot of missing values means that, if we want to include the variables, we need to drop a lot of reviews, leaving us with too few reviews, thus negatively impacting performance. Nevertheless, as shown in the graph below, the average business rating is highly positively correlated with the rating of a review and can thus serve as an important predictor. Similarly, businesses in different states tend to get different ratings, wherefore it is sensible to include state or city dummies in our models. However, given that there are more than 1400 cities and the linked risk of overfitting, the city variable will not be used. The characteristics useful, funny, cool show a non-linear relationship, which is why a squared and cubic term is added for each.

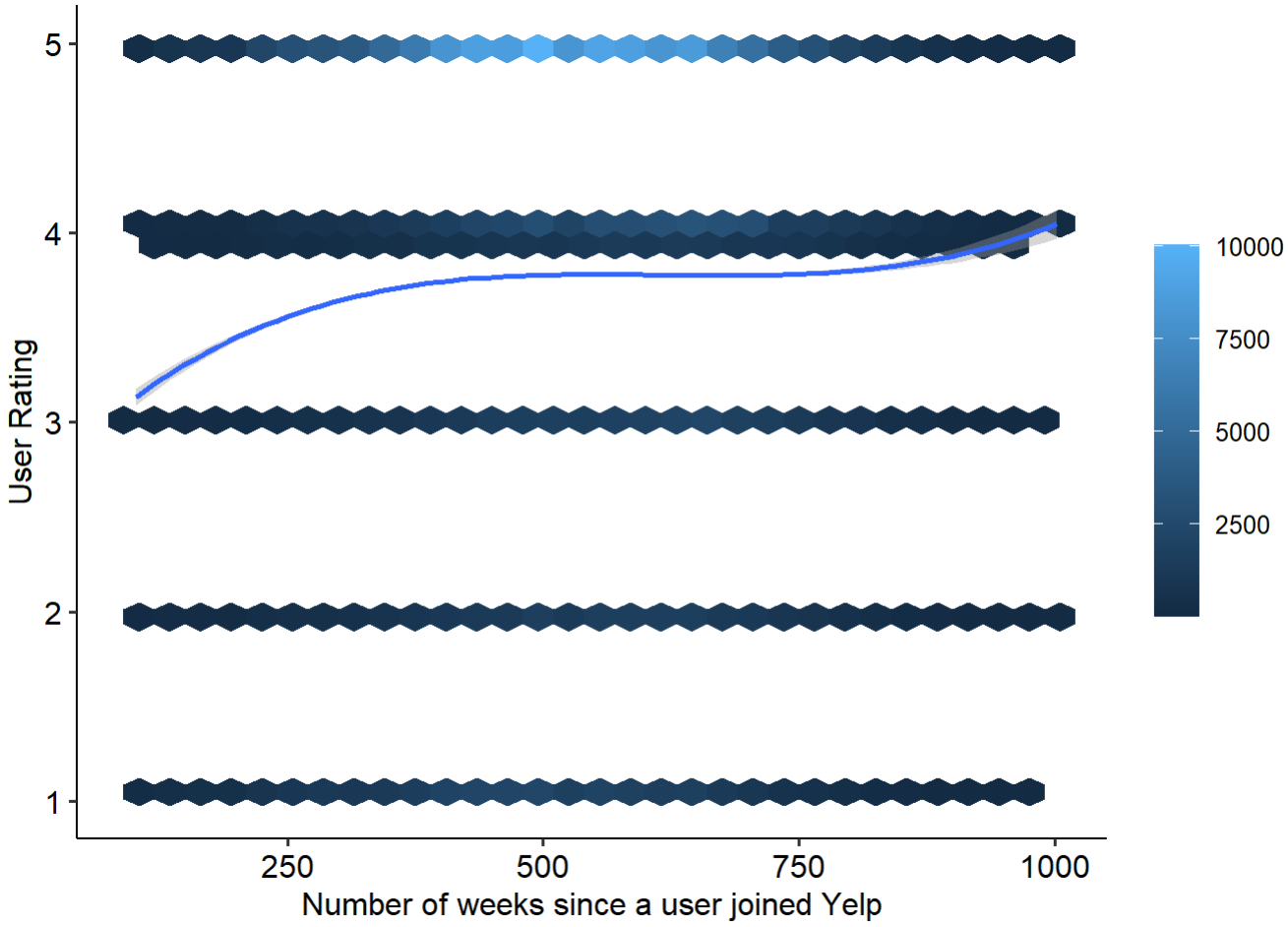


Relationship between user ratings and average business ratings.



Average rating of businesses in each state.

The tip data only includes additional comments made from a user to a business and can be omitted given the inclusion of the comments of the reviews from the review data set. The check-in data contains the times a user looked at a particular business. After aggregating the data, a slight negative non-linear relationship can be seen in the data, qualifying it as a feature. Whilst the user data set contains valuable information that show significant correlation with our variable of interest such as the number of weeks since a user joined yelp (see graph below), there are only 20% of the users in the review data set present in the user data set. Therefore, we are unable to construct meaningful averages to fill the data or ignore the missing data points forcing us to omit the data entirely.



Relationship length of Yelp usership and ratings.

Modelling

Model choice

Whilst the reviews are given in integers, it is an ordered variable, which nevertheless has an underlying continuous meaning. Therefore, we can use regression methods to predict the outcome instead of classification methods. Our data set contains a small number of parameters relative to a large number of observations. However, only few variables such as the average number of stars given to a business are likely to be highly correlated with our variable of interest. In these cases, models performing variable selection/ shrinking will do better since they are able to largely reduce the variance which emerges due to the presence of many parameters. Ridge and LASSO models both benefit from the ability to shrink coefficients to pick the most predictive ones hence reducing the negative impact of variables of low explanatory power. Additionally, the text is unlikely to be related in a highly non-linear way, wherefore the linear models perform well. This is amplified by the inclusion of non-linear terms that account for some of the non-linear relationships and by the fact that the strongest predictors from the review and business characteristics are linearly related to the outcome.

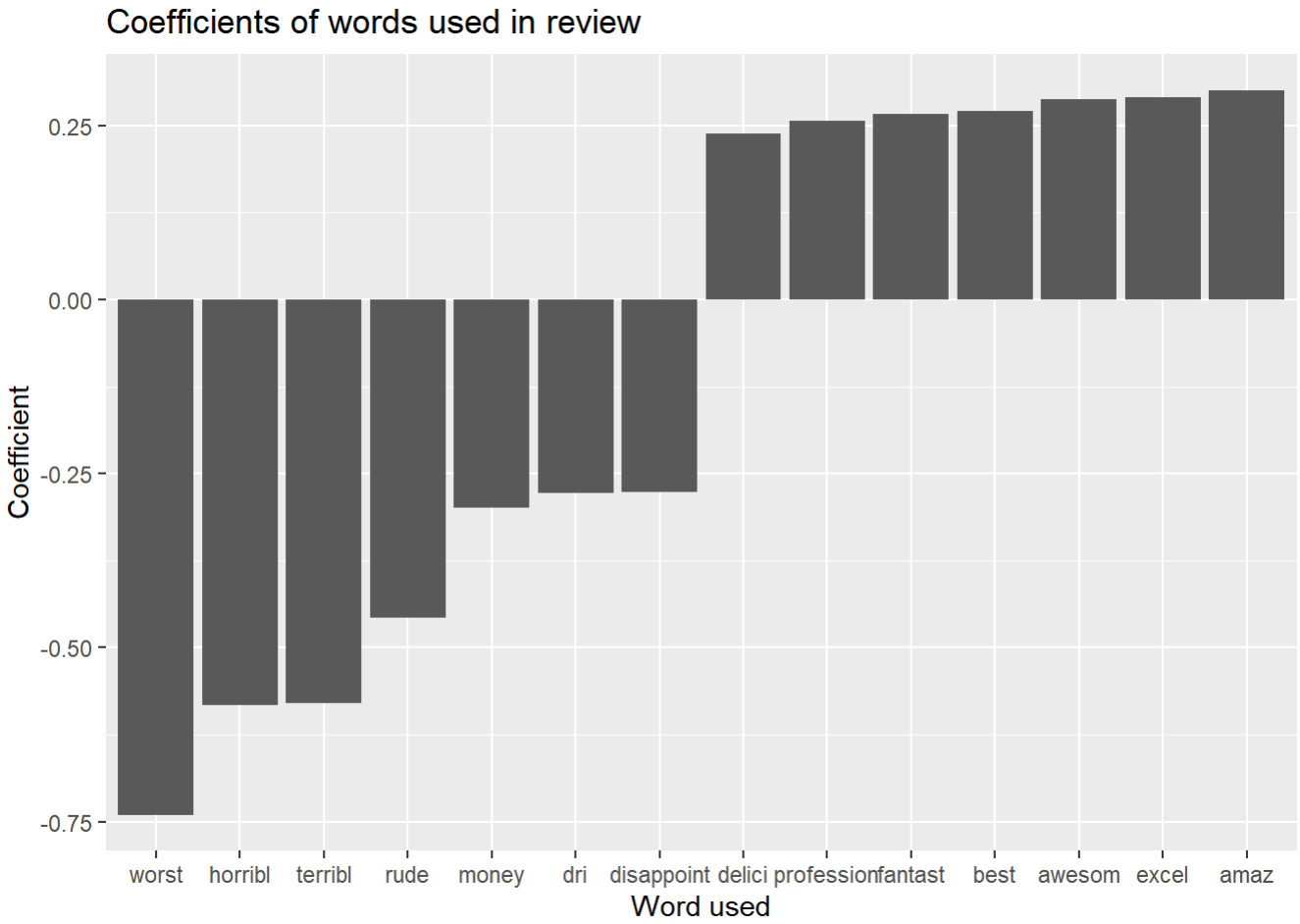
Performance evaluation

Linear models

Performance/Model	LRM	LASSO	Ridge	LASSO binary
MSE on test data	0.96	0.96	0.97	0.91
MSE on training data	0.97	0.95	0.96	0.88
R^2 on test data	0.54	0.53	0.54	0.56
R^2 on training data	0.54	0.54	0.54	0.57

Ridge and LASSO can shrink the estimators to reduce the variance of the estimators. However, given the large number of observations and relatively small number of parameters, the variance of the OLS estimator is already relatively low. In combination with the cross-validated lambda being very close to zero for both Ridge and LASSO, this explains why they don't outperform OLS. The small decrease in variance that can be obtained through shrinking is made redundant by the increase in bias generated by the shrinkage methods. A better performance can be achieved by changing the data format slightly to only indicate the usage of a word without specifying the frequency. This is due to the fact that often used words are prevented from getting too much weight.

Including the text of reviews leads to a significant increase in performance given the relatively large informativeness of some words as shown in the graph of the coefficients for shrinkage estimators.



The seven highest and seven lowest LASSO coefficients on the text used in reviews.

Non-linear models

Simple regression trees are unlikely to perform well given that there are over 500 parameters, most of which have low levels of informativeness. Under these circumstances, a large tree would be required, leading to a high variance and poor performance. Using Random Forests to decrease the variance will lead to poor performance due to the large number of parameters and few good predictors, as evidenced by the fact that only around 12% of features have a an absolute value of the coefficient larger than 0.1. This would lead to strongly biased trees in cases when only poor predictors are chosen at each node and given the high probability of this happening, the bias is likely to remain. Nevertheless, boosting is able to achieve a performance comparable to the shrinkage estimators. This is due to the combination of bagging, learning and the ability to capture non-linear relationships between the variables. Bagging reduces the variance of the estimates by using averages and by letting the estimator explain the previously unexplained data points, enabling us to achieve a more precise estimate. By reducing the learning rate, we can achieve a lower MSE due to the prevention of overfitting.

Performance/Model	Unpruned Tree	Boosted Tree
MSE on test data	1.21	0.86
R^2 on test data	0.30	0.64

The merely marginal difference between the linear and non-linear methods can be explained by the likely linear relationship between the explanatory and outcome variables. It is unlikely that the variables such as the words are related in a highly non-linear way in which case boosted trees would perform better. Additionally, the strong predictors such as business characteristics are related in a linear way further highlighting why more flexible models don't significantly outperform linear models.

Project Challenges

A major challenge during the project was to determine what data structure is required, how the data can be transformed to be made useful and how the ultimate goal of prediction can best be reached. Due to the lack of prior experience in text analysis, understanding how to prepare text and which models perform well with text-based data was challenging. However, using an iterative approach of exploring the data, assessing its usefulness whilst referring to the initial goal has proven itself particularly instructive. Applying the same strategy to the modelling phase was equally effective.

Statement of Academic Integrity

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) (https://warwick.ac.uk/services/gov/calendar/section2/regulations/academic_integrity/)

Guidance on Regulation 11 (https://warwick.ac.uk/services/aro/dar/quality/az/acintegrity/framework/guidancereg11/)

Proofreading Policy (https://warwick.ac.uk/services/aro/dar/quality/categories/examinations/policies/v_proofreading/)

Education Policy and Quality Team (https://warwick.ac.uk/services/aro/dar/quality/az/acintegrity/framework/guidancereg11/)

Academic Integrity (https://warwick.ac.uk/students/learning-experience/academic_integrity)