Predicting Wine Ratings

Sections by:
Erick Sebastian Solis Martinez: exam number: 8
Literature review: paragraph 1,5
Data description and ethics: paragraph 3, 7, 11, 15, 19
Methods: paragraph 3, 7
Analyse results: paragraph 1, 5

Freja Damgård Høst-Madsen: exam number: 201
Literature review: paragraph 2, 6
Data description and ethics: paragraph 4 ,8, 12, 16, 20
Methods: paragraph 4, 8
Analyse results: paragraph 2, 6

Paul Alexander Kürstein-Jensen: exam number: 103
Literature review: paragraph 3
Data description and ethics: paragraph 1, 5, 9, 13, 17
Methods: paragraph 1, 5, 9
Analyse results: paragraph 3, 7

Victoria Hinley Deng: exam number: 121
Literature review: paragraph 4
Data description and ethics: paragraph 2 ,6, 10,14, 18
Methods: paragraph 2, 6, 10
Analyse results: paragraph 4

Sections shared: Introduction, discussion and conclusion.

Abstract

This paper analyses people's opinions online about wine consumption. We use data scraped through Vivinos website, the world largest wine marketplace. We compare the different characteristics of 13974 wines such as the country, grape type, type of wine, price and alcohol percentage. All in order to see if we can predict the wine rating based on these attributes. These ratings are from 1 to 5. We apply Ridge and Lasso, two different machine learning models with different polynomials degrees, to predict wine rating. Although our model was not able to predict the behaviour of the users correctly, based on these features, for wines rated above 3.5, the model performed better - but still was not sufficient to make a powerful assumption about the reviews it will describe.

Research question: Can wine ratings on Vivino.com be predicted?

Introduction

In recent years customer reviews have become all the more common online. Research shows that reviews affect what people choose to buy online and that the majority of all online shoppers use reviews to determine what products to buy and which services to purchase (Anderson). Therefore, predicting customer reviews based on the components of a product would be a powerful tool for manufacturers and businesses, as poorly reviewed items tend to sell less than products with excellent online reviews. This research paper investigates the wine reviews on Vivino.com. Vivino is the world's largest online wine marketplace (Vivino). Vivino is primarily used as a wine rating service where users can access scores, tasting notes, reviews and prices of wines from a simple scan of the label on their phone or by accessing the wines on Vivino's page online. The website and app also function as a marketplace, where customers can buy wines from Vivino's affiliated merchants. Users can rate wines on a scale from one to five stars, and the average score of the wine is clearly shown when customers/users scroll through the website looking for wine.

The average score thus has a prominent position on the website and is very noticeable when potential customers look for wine. This research paper intends to use machine learning to analyse online wine reviews on Vivino.com and which components and criteria affect the reviews of the wine. We aim to answer the question: How can a wine's components and other features such as region and taste be utilised to predict how well it will be received and reviewed?

This paper utilises supervised machine learning to create a model that aims to predict the rating of a wine on Vivino.dk. The rating of the wine is a numeric value. Ratings can be interpreted as a proxy for the quality and popularity of the wine. We investigate the relationship between rating and features we believe could affect this, using the machine learning methods LASSO and Ridge. We compare the performance of different learning algorithms to select the best model for predicting ratings. We will evaluate each model's ability to predict rating based on mean squared error, root mean squared error and mean absolute error.

Literature review

Big data has brought ample opportunity to analyse the behaviour of the consumers. So in the last few years, there has been significant research on taking data online with a focus on online reviews and comments from websites and apps like Twitter, Instagram and other review sites. This newfound access allows one to look deep into recreational eating on social networks. For example, authors like Guidry et al. have studied how hashtags have been used to express the approval or disapproval of fast-food chains. At the same time, Mejova et al. collect the data of Instagram showing that travel drives emotions associated by users to hashtags.

Companies have gained much knowledge about their customer’s needs. For example, using big data in Taiwan, Pizza Hut created a pizza with non-traditional topping features like a mash of pig’s blood cake and century eggs. In addition, Pizza Hut “scraped social media posts for the most talked-about foods and created word clouds of potential new ingredients” (Hou).

Scraping data for food and drink culture to study consumption habits has gained popularity, adding knowledge over the years. Machine learning and models that could predict people’s eating and drinking behaviour, based on several characteristics, is also another line of work.

There has not been much research in the field of machine learning and the consumption of alcoholic drinks behaviour. However, Chorley et al. has done some similar research, where they analysed data from Untappd. In this beer-oriented rating app, users can utilise similar services as the services provided by Vivino.

Kotonya et al. have also written a paper that aims at analysing the data from Vivino where they analyse the wine consumption. They compare “users’ perceptions of various wine types and regional styles across both New and Old World wines, examining them across price ranges, vintages, regions, varietals, and blends”. The paper’s primary finding is similar to what we are trying to accomplish; they found that the cost does not bias the ratings provided by Vivino. To achieve this, the authors use a model behaviour to develop a regression model for predicting wine ratings and a classifier for determining user review preferences.

Kotonya et al. focus their analysis on the social experiences around wine consumption. They use text analysis of reviews and focus on the social network aspects of Vivino’s platform. In this paper, the focus will primarily be on trying to predict how wine is received, purely based on the objective characteristics of the wine. This makes it possible for us to test if there is a clear pattern to how a wine is rated by users based on how and where it is produced and how it is priced. If we succeed in predicting wine reviews based on these characteristics, it could possibly be a powerful tool for wine producers who want to test how a new product will be received.

Data description and ethics

In this section, we will describe our data and how we obtained it.

Ethical concerns - getting permission.
Working with data online raises some ethical questions. When it comes to scraping data online, one has to be conscious of how one does it. Many sites restrict access through rate limits or directly state in their terms of service that scraping in any way, shape or form is not allowed.

Another issue arises with the ability to use the scraped data to gain an economic advantage over the places/sites that the data is scraped from. For example, in our case, we could potentially use the data from Vivino to sell wine in a manner where we could either reference the prices and ratings from Vivino and optimise our sales based on the information gained from Vivino.

We emailed Vivino to request permission to scrape and use scraped data for strictly academic/research purposes. Vivino replied, permitting us to scrape their site with the only restriction being rate-limiting our number of requests to 1000 requests every 10 minutes.

Dataset

As mentioned before, the data is scraped from vivino.com, where we receive the data in a JSON list with a total of 35637 wines, where we can export the variables that we wish to analyse. Unfortunately, not all relevant data is supplied in this list, such as the variables: distribution of reviews and grape type of wines. This motivates us to construct wine-specific URLs to map individual wines to scrape the distribution of ratings, alcohol percent and grape types on the wine-specific websites.

The URL's were constructed by looking for recognisable patterns in an existing URL of the wine pages we want data from. Here we found that the winery name, wine name, wine id and the year was used - example:

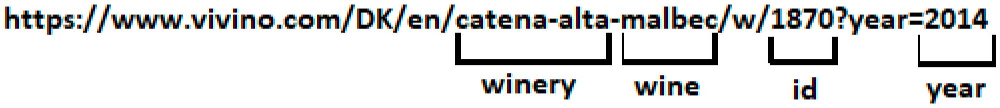

https://www.vivino.com/DK/en/catena-alta-malbec/w/1870?year=2014
winery
wine
id
year

Figure 1

For downloading the page, the request module was used. For parsing the HTML BeautifulSoup. We had to use precise regular expressions to find the information we wanted. Upon further inspection, we found that the information we wanted was in JSON format. Therefore, we had first to extract the JSON formatted HTML and convert it into a JSON file for easy extraction - by referring to the dictionary format of the JSON file.

An issue was that not all of the HTML was the same in terms of patterns and content. For example, a wine page on discount had a different HTML, and a vintage HTML page was different from an overall wine page. Furthermore, missing variables meant a lot of exception handling was needed to make the code run smoothly. To prevent losing all data in case of unexpected errors, we scraped in partitions of seven and later concatenated them into one dataframe.

The parameters we were looking for were the wine name, name of the winery, price (if discounted, discounted from price), vintage (year of harvest), country, region, grape type(s), number of reviews, distribution of star reviews and average review. The table utilised underneath shows the format of each parameter and what format it was changed to for ML. The cleaning process is described in the next section.

Parameter	**Format**	**Cleaned for ML**
Wine	string	Dropped
Winery	string	Dropped
price	float	float
discounted from	float	Dropped
vintage	integer	integer (2021-age)
country	string	dummy (integer)
region	string	Dropped
grape type	string	dummy (integer)
number of reviews	integer	integer
1-star reviews	integer	Dropped
2-star reviews	integer	Dropped
3-star reviews	integer	Dropped
4-star reviews	integer	Dropped
5-star reviews	integer	Dropped
average review	float	float

Figure 2

Data cleaning

We clean the data for duplicates which we check over multiple different variables. Wine, winery, vintage and type. Which dramatically reduced our data set of wines from 35650 wines down to 13974 wines. This reduction indicates that our initial mapping of wines was erred. Now having corrected the issue, we were now able to start. We apply descriptive analysis to our data as it is in this state.

We note that a “feature” of the Vivino website is that the average rating the website shows is typically “rounded” up. With a heavy emphasis on rounded. The average correction is +0.17 for all wines. Meaning that the average shown on the website is heavily skewed of what their actual rating average is. A wine with an accurate rating of 4.07 shows a 4.2 rating instead of 4.1, which would be the accurate number when rounding up.

Having noted this discrepancy, we will still focus on the rating shown on the website as this is the rating that customers will see.

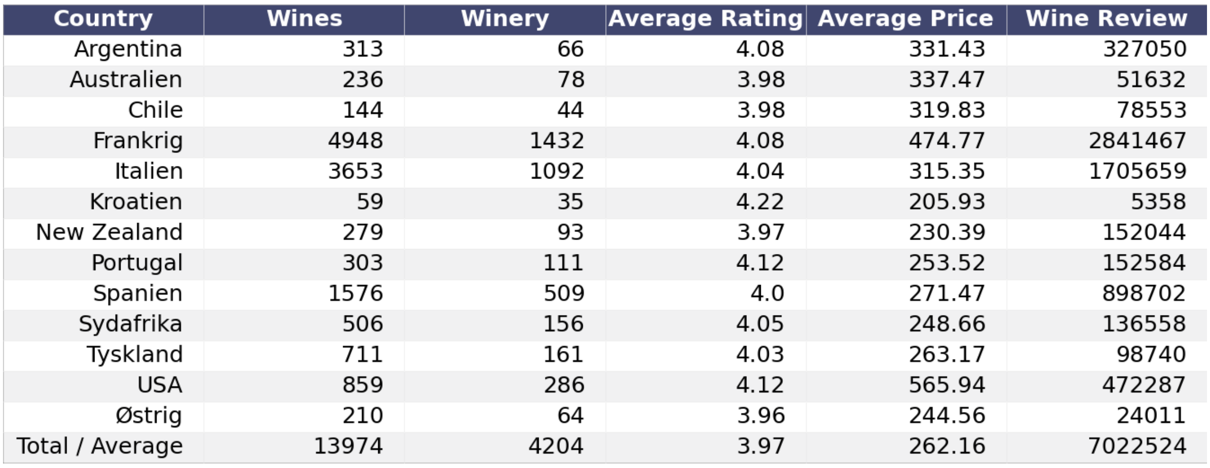
Not every wine had its alcohol percentage shown on their page. For example, 4239 wines did not show this value. Instead of removing these wines from the dataset or replacing missings with zero, we decided that the best way to handle these missing values was to replace them with the mean alcohol percent. It was clear that the missing values were not due to the wines not containing alcohol.

After analysing the scraped data, we decided to include the following features in our prediction of rating: wine type (there are four types: red, white, sparkling and rosé), country, grape type, number of reviews, alcohol percent and price.

To prepare the data for machine learning, we then transform the country and grape type into dummies, increasing the number of features to 432.

Descriptive Statistics

Figure 3 presents a summary of our dataset, with a total of 13,974 wines. The table shows the countries that have more than 50 different wines. We drop the rest from the table for illustration purposes (not from the analysis) because the number of countries was too large and did not have a significant weight.



Country	Wines	Winery	Average Rating	Average Price	Wine Review
Argentina	313	66	4.08	331.43	327050
Australien	236	78	3.98	337.47	51632
Chile	144	44	3.98	319.83	78553
Frankrig	4948	1432	4.08	474.77	2841467
Italien	3653	1092	4.04	315.35	1705659
Kroatien	59	35	4.22	205.93	5358
New Zealand	279	93	3.97	230.39	152044
Portugal	303	111	4.12	253.52	152584
Spanien	1576	509	4.0	271.47	898702
Sydafrika	506	156	4.05	248.66	136558
Tyskland	711	161	4.03	263.17	98740
USA	859	286	4.12	565.94	472287
Østrig	210	64	3.96	244.56	24011
Total / Average	13974	4204	3.97	262.16	7022524

Figure 3: Summary of the wines, wineries, average rating, average price and total of wines review per country datasets collected from the data scraping at Vivino website.

The total number of countries that produces wine that Vivino captures are 38, where Italy and France seem to have half of the share of the production, France with 35.4 and Italy with 26.1 percent while Spain (11.3 per cent) and surprisingly the United States (6.1 per cent) are on the top 4 of the producers offered through the site.

Regarding the rating, Croatian wines are well rated with a reasonable average price, but the number of reviews was meagre. As expected, France and Italy are also among the best ranked, followed by Argentina, Spain, Portugal, South Africa and the United States. The number of reviews per country is also similar, with France having around 2.5 million reviews, Italy around 1.6 million, and Spanish wines scratching a million reviews.

On the other hand, we have the Average price of each country, where french wines are as expensive as expected with a price of 474 DKK per bottle, but the United States wines are on the top of the most expensive ones with an average price of 566 DKK per bottle.

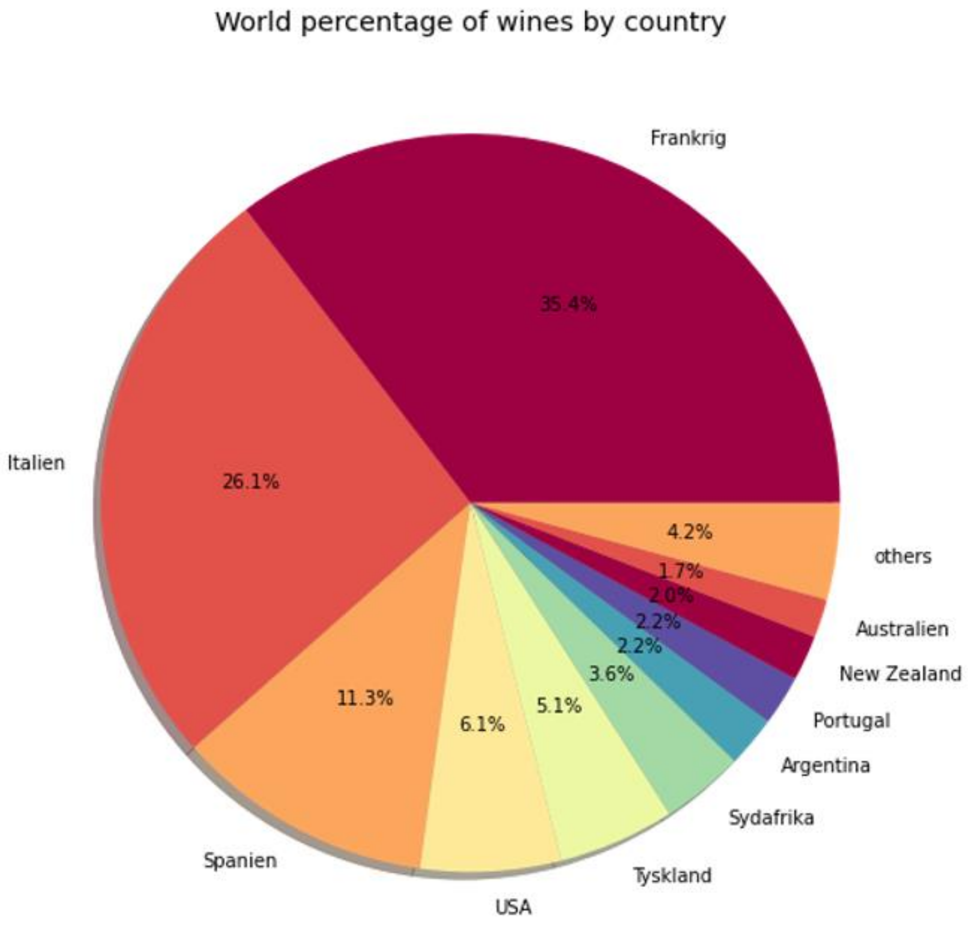

World percentage of wines by country

Figure 4

Another thing to point out is the apparent correlation between the price and the rating. In wines, as with many other products, the price is related to quality, taste, durability and other different types of characteristics. However, there is also a psychological reason to think that if it has a higher price, it will be better, and of course, in wines, it also adds social status.

For this reason, we decided to plot this relation and separate them by the type of wines. So the plot shows that there is a slight but not clear and nice positive correlation between these two variables, even though the scatter shows that there are many wines that are cheap and have a good review. However, it is rare to see an expensive wine with a bad review, so we can say this is still present.

The surprising one is the Rosé. From our data set, the most expensive Rosé rounds a price of 1,368.79 DKK. However, we can see a “higher” positive correlation than the other wines that are similar in line but seem like Rosé tends to have a heteroscedasticity form.

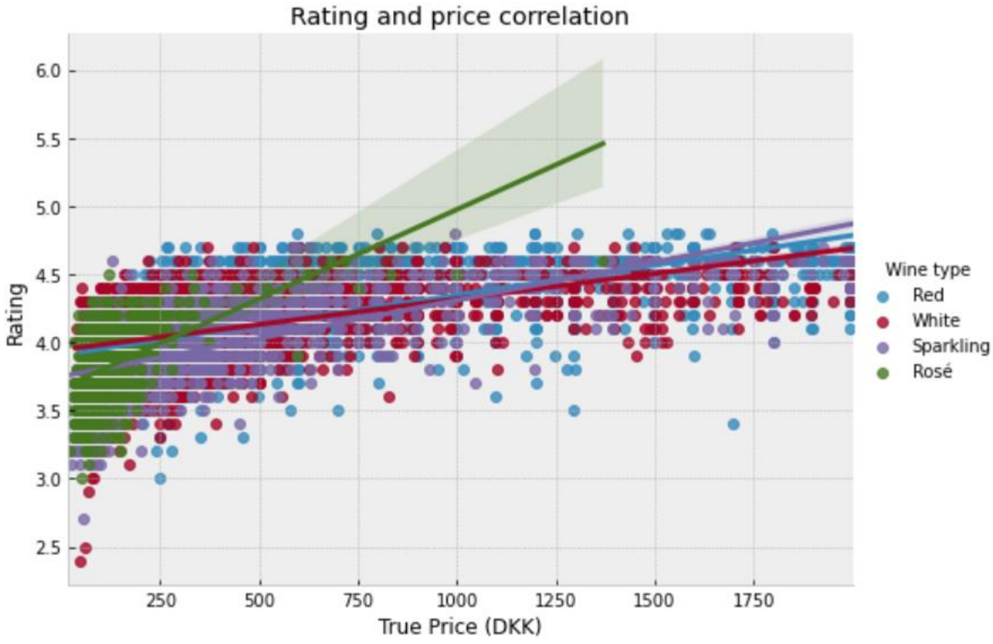


Figure 5

Methods

In the following supervised machine learning is utilised to create a model with the purpose of predicting the rating of wines on Vivino.dk. Supervised machine learning's primary goal is to train a model from labelled training data, which allows us to make predictions about data that is unseen for the model (Raschka and Marjalili 3). In order to test our results, we split our data into a development and a test dataset. The development data is used to develop the model.

The remaining data is kept out of the development process. This will ensure that there is no data leakage between the two processes. By keeping the testing data away from the development process, we ensure no bias in our results. Both training and test data are fitted to the training dataset distribution, as rescaling test observations relative to the test data set might cause overfitting and biased estimates of out-of-sample accuracy (Raschka and Marjalili 199). We use 1/3 of the data as test data and 2/3 thirds of the data as development data.

Regression Models

It is essential to mention overfitting and underfitting when it comes to the model's ability to make out of sample predictions. When the model is overfitted, it makes a more precise in-sample prediction, but if it is too overfitted, it captures all variance in the data, including the white noise. This will, in most cases, lead to out of sample prediction and, therefore, not desirable. On the other hand, underfitting would mean not taking enough variance into account and, therefore, not having a strong predictive power.

Furthermore, overfitting is often a problem in models with many parameters, which causes the model to perform well on training data but fail to generalise to test data. Given the dimensions of our data set, which have many features, we have decided to use the machine learning algorithms Lasso and Ridge. Lasso and Ridge regression are both regularised methods. The regularisation model is an approach for tackling overfitting by adding additional information to penalise extreme parameter weight variables (Raschka and Marjalili 74). We use the two most common penalty functions, L1 (LASSO) and L2 (Ridge) regularisation.

$$E[(Y_i - \widehat{Y_\iota})^2] + \lambda \cdot C(w)$$

Depending on the method used, the cost function C can look different. The penalty function for LASSO:

$$C_L(W) = \sum_{j=1}^{M} \left|w_j\right|$$

In the case of Ridge regression, the cost function is the sum of squared errors plus the sum of squared weights:

$$C_R(W) = \sum_{j=1}^{M} w_j^2$$

The LASSO method makes coefficients sparse. Depending on the size of the hyperparameter, lambda, certain weights can become zero, meaning that some variables are removed (Raschka and Marjalili 332).

In Ridge regression, increasing the regulation strength by increasing lambda shrinks the weights of our model. Weights can become close to but never equal to 0. Therefore Ridge is good if all variables are essential.

Hyperparameter

In both models, lambda is a hyperparameter. Therefore, we need to find the lambda that makes the best out of sample regression. For this purpose, the development data is split into training data and validation data. To decide on the best hyperparameter for each model, we use 10-fold cross-validation. This allows us to use all of our development data in the process. We use 66 equally spread lambda values between 10^{-4} and 10^{4}. Next, we split the development data into ten equally sized test bins. For every bin, we use the rest of the development data to train the model with a given lambda. We then use the mean squared error for every lambda and for every fold to decide which hyperparameter to use. After finding the optimal hyperparameter, we use the complete development data to fit the model with the chosen parameter (Raschka and Marjalili 193).

The models we use in this paper assume a linear relationship between features and target variables (Raschka and Marjalili 335). To account for the violation of this assumption, we use polynomial regression to model the nonlinear relationship. We fit LASSO and Ridge using second and third-degree polynomials to model the relationship between features and ratings.

Predictive performance

We will evaluate each model's ability to predict out of sample ratings based on the error measure mean squared error:

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$$

We also evaluate the root of mean squared error or RMSE:

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n}}}$$

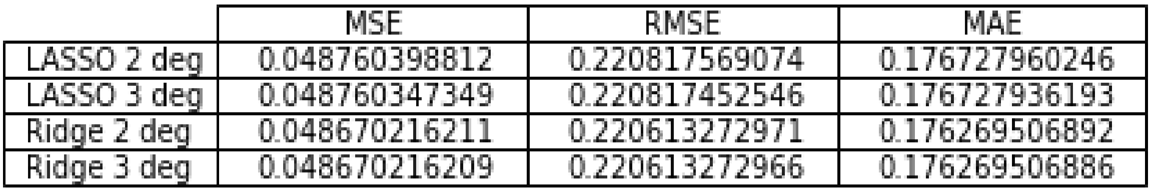
Lastly, we also calculate the mean absolute error for every model’s prediction:

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|}$$

By comparing models based on these measures, we will choose the model that performs best overall.

Analyse results

In this section, we will evaluate the results of the machine learning process described above. The features we decided to use for predicting the ratings were the number of reviews, country, grape types, alcohol percent. After fitting four models as described earlier and calculating the four error types, we get the following results:



	MSE	RMSE	MAE
LASSO 2 deg	0.048760398812	0.220817569074	0.176727960246
LASSO 3 deg	0.048760347349	0.220817452546	0.176727936193
Ridge 2 deg	0.048670216211	0.220613272971	0.176269506892
Ridge 3 deg	0.048670216209	0.220613272966	0.176269506886

Figure 6

From the table in figure 6 observe that Ridge regression performs best in all three error measures. Ridge regression with both 2 and 3 degrees of polynomial features perform pretty similarly. However, by looking at the 7th digit of the MSE, the 11th digit of the RMSE and the 11th digit of the MAE, we see that the Ridge regression 3 degrees of polynomial features perform better on all three error measures.

The optimal lambda chosen by the 10-fold cross-validation for the 3 degrees ridge regression is 11.12. In figure 7 below, the validation curve is shown for the k-fold process. It can be seen that the MSE for the training set clearly dips around this value. At the beginning of the curves, we see a high MSE for the validation set and a lower MSE for the training set - this indicates overfitting. On the other end, we see that MSE for training and validation is high and similar, indicating underfitting.

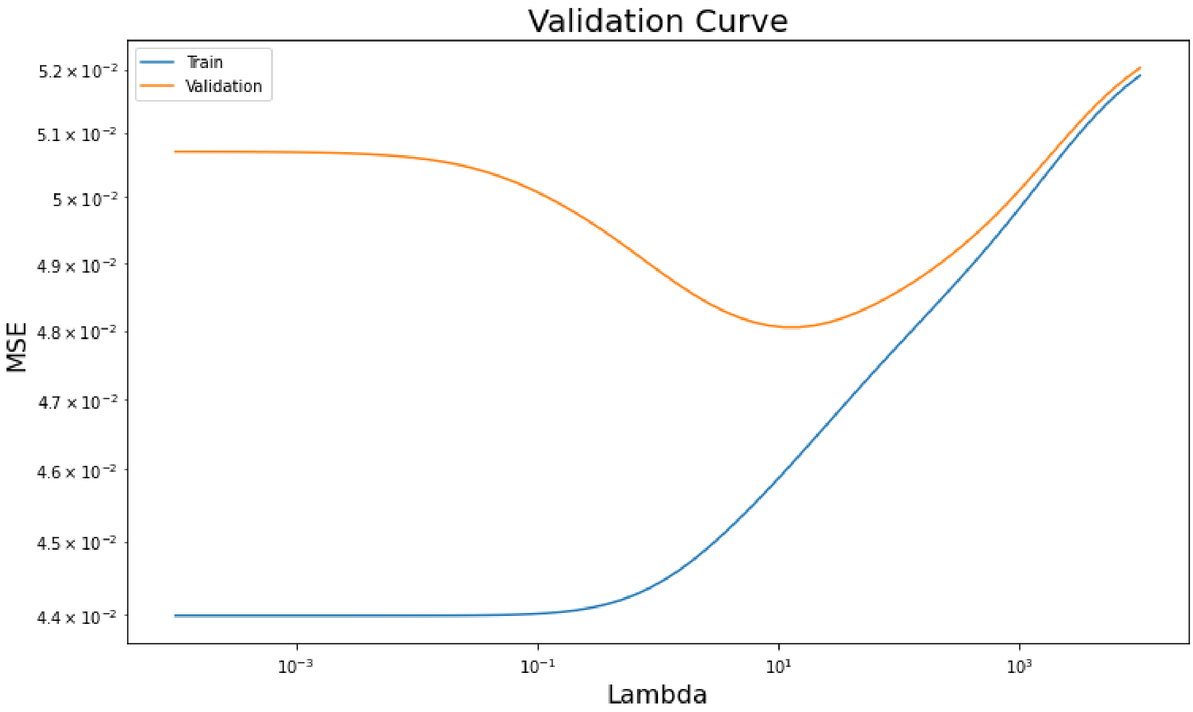


Figure 7

A way to test if our chosen model generalises well to unseen data is by plotting a learning curve. Learning curves can help detect over-or underfitting in a model (Raschka and Marjalili 196). Figure 8 shows the learning curve of the Ridge regression 3 degrees of polynomial features. The plot shows the accuracy as a function of train-set size. The lines represent the average performance on the validation- and training sets. The width of the coloured areas expresses the 95%-confidence interval of the performance.

Examining the plot (figure 8) allows us to determine whether the model suffers from high bias or high variance. From the plot, it is clear that the performance on the validation set is very inconsistent. When the sample size increases, it does not seem to improve the performance significantly, but, notably, the variance between the average performance on training and validation set seems to decrease as the sample sizes increase, as this suggests that the model might be improved by collecting more training data or reducing the model's complexity.

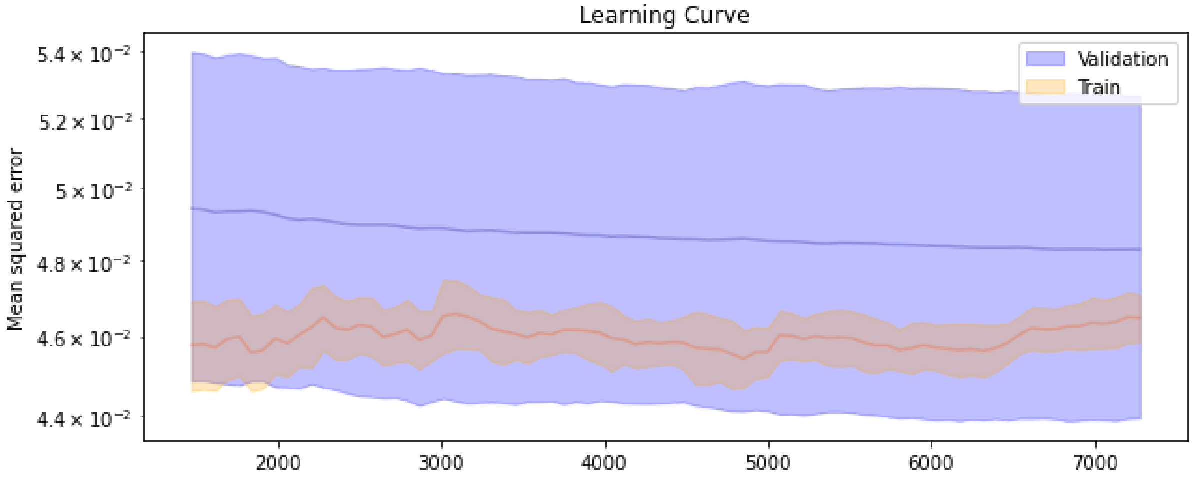


Figure 8

In figure 9 we are plotting the predicted ratings against the true ratings of the training set. It is clear that our model primarily predicts ratings over 3.5 stars. Therefore, the model fails to predict ratings for wines with a score under this value.

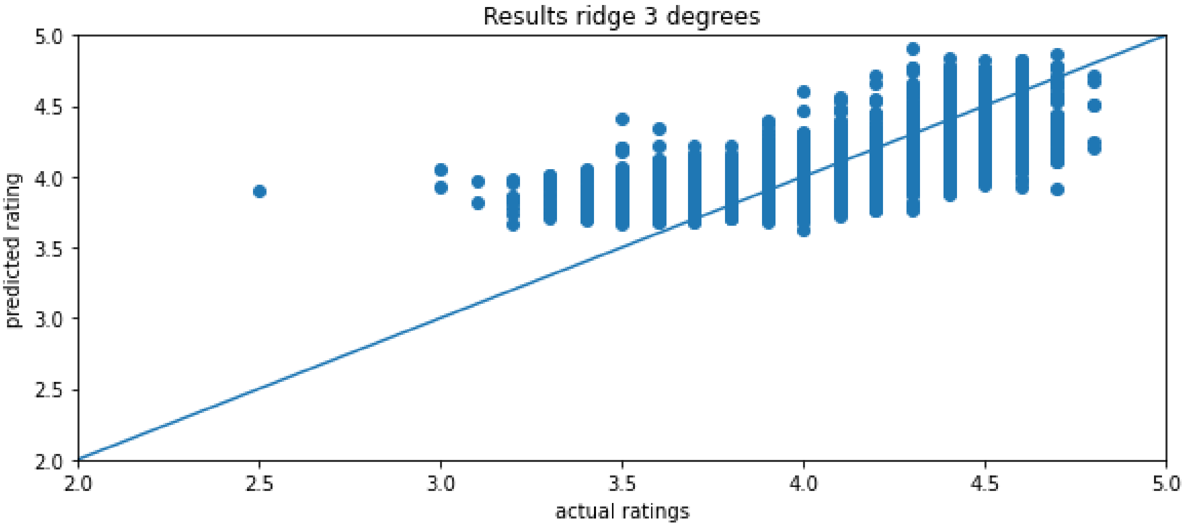


Figure 9

Based on the results, we can not conclude that wine reviews on Vivino.dk can be predicted based purely on objective factors, such as where wine is produced and how it is priced. Instead, this suggests that wine ratings depend more on subjective features and experiences, which are hard to capture in the data and models used in this paper.

Discussion

What could we have done differently?

First and foremost, as we noted in the data section, we could have used the actual rating as our target rating instead of the shown rating on the website. Though explained, the rating has little actual value as it is not shown to the users and therefore only found when the user actively calculates the rating with the number of ratings shown on the wine's webpage. We doubt that many day-to-day users do this and take the shown average rating for face value.

Another way to improve the model might be to increase the amount of data used. Increasing the 13000 wines is not a lot compared to the boasted 13 million wines on their website (https://www.vivino.com/about). Many of the wines on the website are not for sale and will not show up in the explore wine tab we used to map the wines we wanted. Maybe more wines could be included without including the price or by finding the price elsewhere.

When looking at the reviewers themselves, an option could be to scrape the top x-amount users from each country to see how they review wines. This would give an understanding of the general reviewing behaviour of the users. Perhaps the users generally only review great and bad wines and do not bother to rate a wine when the experience has been mediocre, which means there would be an over-representation of the good wines on Vivino. Vivino tends to show only the wines that are better reviewed.

Our analysis might also suggest that wine ratings depend more on subjective features and experiences, which are hard to capture with the data and models used in this paper. To better understand how people review wines, one could look into users' actual reviews. Reviews contain a great deal of information about how and what people review and look for in a wine. Using the text as data methods on these reviews would allow one to gain more insight into how the user's rate wines. The sentiment of the reviews would perhaps allow one to understand how and if the user rates cheap and expensive wines differently. Moreover, old expensive wine might be objectively better than cheap wine, but in the context of its price and age, the rhetoric and reviews of the expensive wine might be worse than the expensive wine. Understanding how people rate the way they do would enable one to integrate this knowledge into the ML algorithm and improve the predictability power of the algorithm. The main concern using text would be the ethical concern of the individual users' reviews being used. Although, of course, the user would be informed about their reviews being shown on the Vivino website. However, they would not be able to know of the scrapers that could potentially use their reviews. Therefore, raising the question about if and how one could use the reviews without notifying each individual reviewer and asking for permission to use the review for analysis. One could even look at the most influential Vivino users assuming that they are better at reviewing wine than regular users - making their reviews weigh more. Are the wine critics' reviews taken into account?

Conclusion

In this paper, we set out to analyse wine ratings and attempt at predicting wine ratings using a simple but hopefully effective machine learning method. Scraping Vivino.com, we were able to gather information on more than 13000 thousand wines obtaining information about the features of the wine such as age, grape type, price, number of reviews. We then analysed the data.

We fit LASSO and Ridge regressions using both second and third-degree polynomials to predict wine ratings based on the chosen features. Each model was evaluated and compared based on its ability to predict out of sample values. We used the error measures MSE, RMSE and MAE. Based on the result, it was clear that Ridge with third degree polynomials performed best in all three measures. However, by analysing this model further using a learning curve, it was clear that this model was a bit flawed, as the predictive performance did not seem to improve much when the sample size was increased, suggesting that the model might need improvements. Furthermore, the analysis of the predicted out of sample values also showed that the model systematically predicted low ratings wrong.

Our ML algorithm did not have the strong predictive power we hoped for. The model might have performed better if we had used more detailed information. We also had a lack of computing power to include regional features of the wines into our ML. Including more wines and more data could perhaps allow one to predict the rating more precisely. However, our findings might also suggest that wine ratings depend more on subjective features and experiences, which are hard to capture with the data and models used in this paper.

Bibliography

Anderson, Jill. “Why Are Customer Reviews So Important?” *https://medium.com/*, Revain, 2018, https://medium.com/revain/why-are-customer-reviews-so-important-185b915d4e5d. Accessed 16 08 2021.

Guidry, J. D., et al. “#mcdonaldsfail to #dominossucks: An analysis of Instagram images about the 10 largest fast food companies.” *Corporate Communications: An International Journal*, vol. 20, no. 3, 2015, p. 9.

Hou, Betty. “Big Data Turns Pig’s Blood and Century Eggs Into Hot Pizza Toppings.” *https://www.bloomberg.com/*, August 2021, https://www.bloomberg.com/news/articles/2021-08-12/big-data-turns-pig-s-blood-century-eggs-into-hot-pizza-options. Accessed August 2021.

Kotonya, Neema, et al. “Of Wines and Reviews: Measuring and Modeling the Vivino Wine Social Network.” *Tipicamente Wine Blog*, 2018, p. 9. *https://www.researchgate.net/publication/324859409_Of_Wines_and_Reviews_Measuring_and_Modeling_the_Vivino_Wine_Social_Network*.

Mejova, Y., et al. “Fetishizing Food in Digital Age: #Foodporn Around the World.” *ICWSM*, 2016.

Raschka, Sebastian, and Vahid Marjalili. *Python Machine Learning*. 2. ed., Packt Publishing, 2017.

Vivino. “About Vivino.” *Vivino*, https://www.vivino.com/about. Accessed 16 08 2021.