What makes a good coffee?

Julia Jentsch
Matrikelnummer 4222855
julia.jentsch@student.uni-tuebingen.de

Max Grathwohl
Matrikelnummer 5866474
max.grathwohl@student.uni-tuebingen.de

Abstract

This study aims to predict the review ratings of different coffee blends using machine learning (ml) techniques. The dataset used for this research was obtained from Kaggle and underwent exploratory analysis to identify the five most correlated features for the ratings: aroma, acidity, body, flavor, and aftertaste. The relationship between the ratings and these features was modeled using multiple linear regression and random forests algorithms. The performance of both methods was evaluated and compared to determine which algorithm is better at predicting the ratings. Our assumption was that there is a significant correlation between some features and rating, therefore suggesting a correlation also between these features and the quality of a coffee. Further we assumed the random forest to be a better predictor than a regression, as it is generally seen as an advanced ml-algorithm. We could show that there is indeed a correlation between features and ratings. To our surprise, multiple-linear regression outperformed the random forest. This suggests a high linearity in the predicting function of our features.

1 Introdcution

In our project, we analyze the Coffee_Data_CoffeeReview dataset from Kaggle (https://www.kaggle.com/datasets/hanifalirsyad/coffee-scrap-coffeereview?select=coffee_fix.csv) which was scraped from https://www.coffeereview.com/. Coffeereview.com is one of the biggest professional coffee rating sites, publishing a monthly rating report. Reviews are performed by experts, the concept is described as: "We conduct blind, expert cuppings of coffees from all over the world, and report the findings in the form of 100-point reviews, (...)". The rating is based on a scale from 50 to 100. In our analysis, we thus assume that the price of coffees does not have an influence on the final rating. Neither should origin. These assumptions are confirmed during our analysis. The data consists of 2282 different coffees from over 2077 different brands, with 37 columns detailing characteristics including text-descriptions, binary encodings of origin countries and types, 0-10 ratings of the features aroma, acid, body, flavor and aftertaste and the 50-100 point ratings. Acid in the context of coffee refers to a pleasant crispness and has nothing to do with pH levels, making it a desirable characteristic. Body describes the texture and thickness of coffee.

2 Exploratory Analysis

In order to understand its underlying structure and patterns, we did an exploratory data analysis (EDA) of the complete data-set. The goal of EDA is to identify trends, patterns, outliers, and other features that may be of interest or that may need further investigation. Looking at the correlation matrix, we see clearly that a small subset of feature-columns are significantly more correlated with the overall rating (i.e. have a lighter color) than others.

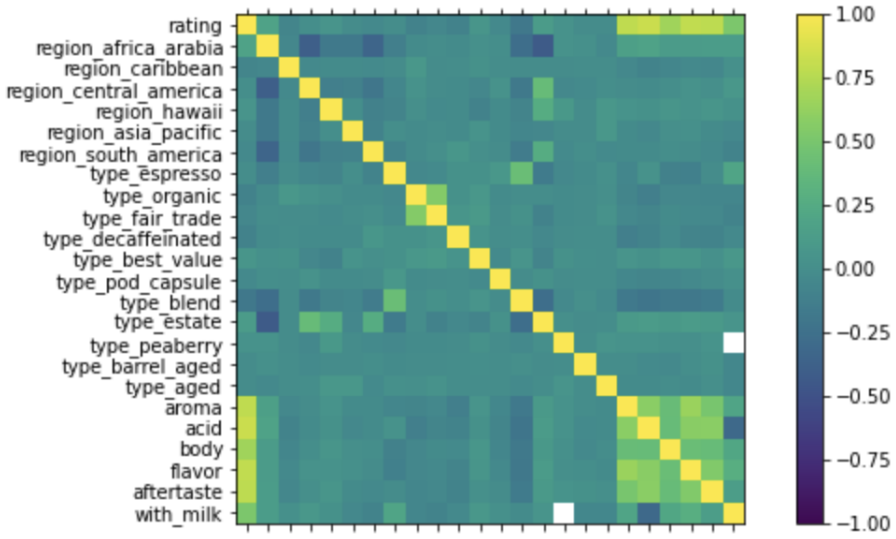

Figure 1: Correlation matrix of features and rating

We therefore focus on the features aroma, acid, body, flavor, aftertaste and with_milk.

2.1 Data Preprocessing

The data was already clearly structured into three date-frames, with one containing all columns and rows. We chose this one to further work with. Data was consistent in type wrt. to the columns (i.e. all integer for rating scores). As discussed in the paragraph "Exploratory Analysis", we found columns that have not shown a correlation with "rating" and therefore dropped these.
We found that the chosen columns still contained some "not a number" (NaN) values. Number of NaN per feature are: aroma: 26, acid: 327, body: 2, flavor: 2, aftertaste: 2, with_milk: 1933. We see that even though the feature with_milk seems to have some influence, it's too sparse to be a good feature. We therefore choose to drop this feature for our future analysis. In order to combat the NaN values, we choose to replace them with the respective mean of the features. Our reason is, that as acid is has the highest NaN count, but it is also assumed to have one of the biggest influences on the rating (according to coffeereview.com), we need to add a value that will not alter the prediction in a significant way. We decided against dropping rows with NaN values, because not all features for a given row were NaN at the same time, thus still offering potential information about significance of the other features.
After cleaning our data, we are now left with only numerical features. We notice that these features all tend to accumulate around high values, as one can see in Fig 2. This bias of the data to higher valued scores could limit our analysis and performance of prediction models. However it does make sense that reviewed coffees, generally have a high quality and thus high scores, as probably few coffee makers will send in a objectively bad coffee.

2.2 Linear Regression

We perform a linear regression on each feature and the rating separately, to further explore their relationship with the rating as dependent variable and the features as independent variable (Fig 2). We can see that there is a linear relationship between the features and the rating and that they are positively correlated, as indicated by the correlation matrix before.

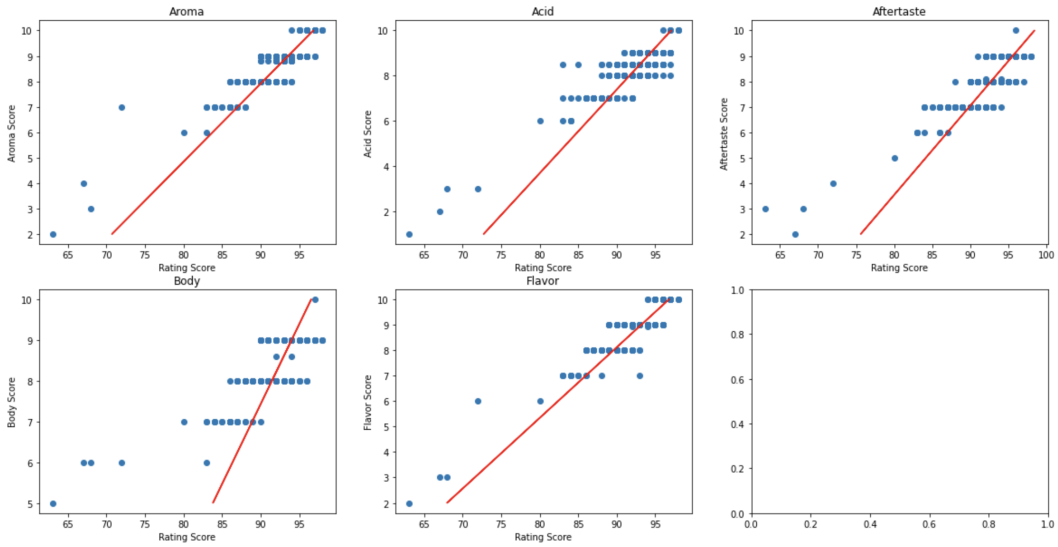

Figure 2: Linear regression of Rating on features

3 Prediction Methods and Results

Our goal was to build predictors and see how well they perform, i.e. how good they are modeling the data distribution. To evaluate the performance, we look at the accuracy and the mean squared error (Fig. 4), as well as plotting the distribution of the number of predicted ratings (Fig 3.), comparing them to the target rating distribution.
For the final experiments and reported results we used the same random train/test split, the latter with the size of 0.3, for all models of multiple linear regression (MLR) as well as for the random forest regressor (RFR).
We want to note that we did multiple iterations of these experiments as well as with newly randomized train/test splits for all different models. The results always overlapped, indicating strong predictive qualities which are not prone to outliers.

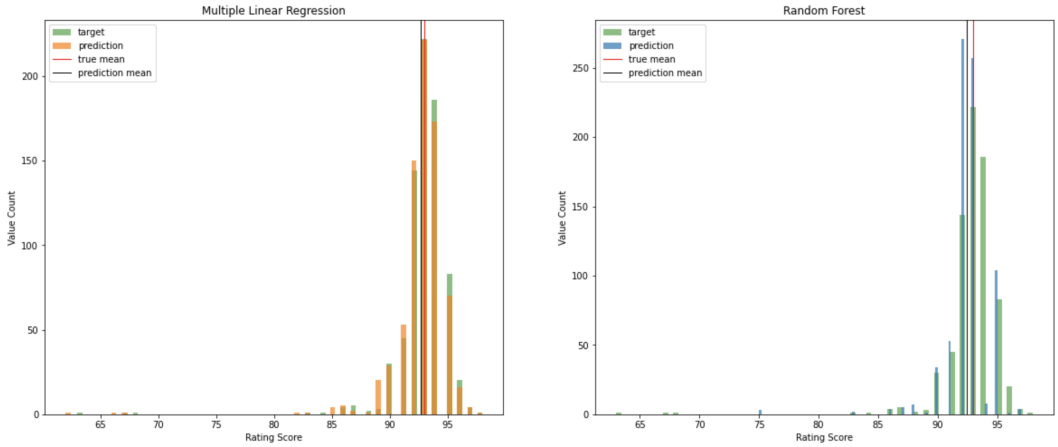

Figure 3: Comparison of predicted and true rating distributions.
MLR on the left, RF on the right

3.1 Multiple Linear Regression

Building on our results from linear regression, we performed a (MLR).

We see the prediction results in Fig. 3. The histogram shows a very similar distribution to the actual data, confirming the achieved accuracy of 99% and mean squared error of 0.066. The mean of predictions also almost overlaps with the actual mean. Performance always stayed in the same

confidence intervals. All of our coefficients amount to a value around 1, confirming again the positive correlation our features and the rating have.

3.2 Random Forest Regressor

We repeat the experiment from above using a (RFR). Fig. 3 shows the predictions. We can clearly see the difference to the actual targets. RFR almost completely misses predictions on rating score 94, while predicting almost double the actual number of targets with score 92. We also see the this drop in performance in accuracy: 93%, and mean squared error: 0.155.

4 Discussion

4.1 Comparison of Multiple Linear Regression and Random Forest Regressor

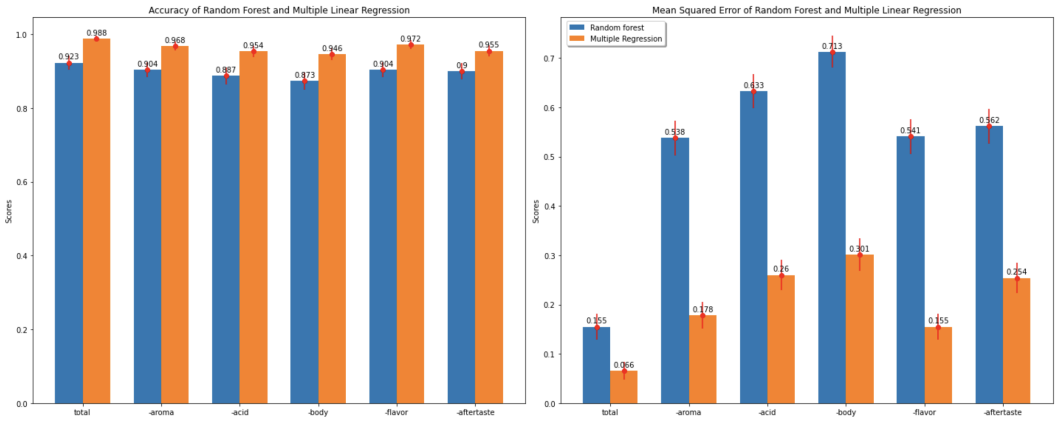

Figure 4: Comparison of accuracy and MSE of MLR and RF
95% Confidence Intervals in red

In order to compare the performance of the MLR and RFR, we use the mean squared error and accuracy. For each metric we calculated 95% confidence intervals, which support our findings. As reported, the MLR has a higher accuracy and a lower mean squared error, thus outperforming RFR. Since our features and the rating are positively correlated and have a linear relationship, as shown above, it is not surprising that a MLR performs well at predicting the ratings.
In the graphic above, accuracy and mse for the overall prediction can be seen on the very left in each plot, while the other bars indicate the values for the predictions being performed if the named feature is dropped out of the training, meaning only the other features are used to predict the rating. By doing this we can see that the body is the most important feature, since the error is the highest if we drop it from the data, followed by acid.

4.2 Limitations

We do want to note, that our results could be limited due to the high bias of our features. As we showed in the exploratory analysis at Fig. 2, feature scores accumulate around few values. This could lead our models to favourably predict high rating scores. However, in several experiment iterations the test data did contain occurrences of coffees which rating scores lay below the lowest score of the training set (e.g. 65, 67 and 68 in test vs 72 in training). The MLR succeeded in predicting these low ratings, even though it never saw such low scores during training. Futher iterations of the experiments confirmed these findings. We're therefore confident that MLR at least, would continue to make accurate predictions on higher variance data.

Concluding, we can confirm that our assumptions do hold and that it is possible to predict the experts rating of coffee, based on the given features. We showed that what makes a good coffee are: body and acid, followed by aftertaste, aroma and lastly the flavor.

References

[1] Cutler, A., Cutler, D.R., Stevens, J.R. (2012) Random Forests. In Zhang, C., Ma, Y. (eds), *Ensemble Machine Learning* , Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-9326-7_5

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[3] Osborne, Jason W. (2000) "Prediction in Multiple Regression", *Practical Assessment, Research, and Evaluation*, Vol. 7, Article 2.

[4] https://www.coffeereview.com