Machine Learning

Machine Learning for Wine Reviews

The question to be answered by this anlaysis is ultimately a question of machine learning. We aim to recommend whether a wine should or should not be bought or suggest a wine that has a great value, review points for every dollar. It is easy to make a determination for a wine that is in the dataset of ~120,000 wine reviews that build the basis of our analysis, but of course, there are so many wines in the world that it is impossible to build a database of every single one. Using machine learning to build a predictive model will make it possible to make value predictions for a wine that is not in the database.

Machine Learning Methods

Fortunately, we have a large dataset of over 120,000 complete wine reviews that can be split into both a training and testing set that can be used to both create and evaluate our model.

The machine learning methods used in this analysis are supervised regression methods. We are attempting to predict value, a continuous variable.

Linear regression will serve as the main machine learning method, with others considered if linear regression does not serve to create a good predictive model.

Variables

The dependent variable in this analysis is value, created here from points and price and defined as points per USD. All other variables will be considered as independent variables in the regression. They include country, region, variety, points and price.

Model Evaluation

The linear regression model will be evaluated using R-squared and adjusted R-squared. Adjusted R-squared will be used to understand which independent variables should be used in the final model. R-squared will be used to determine the overall fit of the model, and whether other machine learning methods need to be considered.

In the case that the linear regression model does not explain a large percentage of the variance in the data, RMSE will be used to compare goodness of fit across various methods.