Planning your California Wine Vacation

Macie Hanneken

**Abstract**

In this analysis, I use factors from the wine-mag-130k data set from Kaggle in order to predict what gives a particular wine a higher ranking score. After data clean up, there were only four predictors: region\_1, region\_2, price, and variety. I used these predictors to predict the number of points a wine would receive from the wine tasting. I used simple multiple regression, principal component analysis, principal component regression, random forest, support vector machine, gradient boosting machine, and a neural network. I found that the random forest model predicted my data the best. That being said, it still had an MSE of 3.874.

**Introduction**

The data set used for this project was the Wine Reviews data set from Kaggle (found here: <https://www.kaggle.com/zynicide/wine-reviews>). This data has 129,971 observations and 14 different variables. The variables are: country, description, designation, points, price, province, region\_1, region\_2, taster\_name, taster\_twitter\_handle, title, variety, and winery. The country is what country the wine is from; the description is a description of the wine from the wine taster about its look, feel, and smell; the designation is the vineyard within the winery where the grapes that made the wine are from; the points are the amount of points given to the wine by the wine taster; the price is the price of the wine; the province is the province or state that the wine is from; region\_1 is the wine growing area in the state or province; region\_2 is a more specific area within region\_1; taster\_name is the name of the person who tasted the wine, gave the description, and gave the wine a point value; taster\_twitter\_handle is the twitter handle of the person who tasted the wine; title is the title of the wine review, the variety is the types of grapes used to make the wine, and the winery is the winery that made the wine. In this data set, the taster tasted the wine, gave a description, and then ranked the wine on a scale of 1-100, but the tasters only taste wine that is 80 or above, so it was really on a scale of 80-100.

Wine tasting is the sensory examination and evaluation of wine (<https://en.wikipedia.org/wiki/Wine_tasting>). Wine tasting has been around since the making of wine, but since the 18th century, the terminology and methodology for the practice has become more evolved, due to Linnaeus and Poncelet bringing and updated understanding of taste (<https://en.wikipedia.org/wiki/Wine_tasting>). Wine is judged based on its bouquet and taste. The bouquet is an aromatic experience. In order to get the most out of the bouquet, it is best to swirl the wine in a wine glass. This allows for the aromatic components of the wine to be exposed to more oxygen and release their aromas. This allows for the taster to get an idea of what the wine is composed of before they taste it, making them more prepared to appropriately taste the wine (<https://en.wikipedia.org/wiki/Wine_tasting>). Important factors in the tasting are: varietal character, integration, complexity, expressiveness, and connectedness. The varietal character describes how much a wine presents its grape aromas, the integration is how well balanced the components of the wine are, the complexity is affected by the multiplicity of the wines flavors, the expressiveness is the quality the wine possess when the aromas and flavors are well-defined and clearly projected, the connectedness of the wine is the bond between the wine and its land of origin (<https://en.wikipedia.org/wiki/Wine_tasting>). After the wine has been tasted, it is scored. The scoring process is based on the appearance of the wine, the smell, the taste, and overall. There are different scoring systems, where different weights are given to each of the listed categories (https://en.wikipedia.org/wiki/Wine\_tasting).

**Methods**

The objective of this study was to predict the number of points each wine received from a particular taster, based on variables from the Wine Reviews data set described above. I have recently been told that I need to cut gluten out of my diet. So, the best gluten free replacement for beer is most definitely wine. Since I have been drinking the same wine for so long, I figured now would be the time to branch out, but I didn’t really know much about wine. So, I figured this project would give me an insight into what factors play into making a wine well perceived.

Data Cleaning

In order to reduce bias, the first step I took for the data cleaning was to choose a singular wine taster. I did this because even though these tastings and rankings are supposed to be impartial, everyone has the wine that they themselves prefer. I wanted someone from America simply because American wines are a bit easier to classify since they don’t have the regions the wine is from within the name of the wine. I chose Matt Kettman as my taster. He exclusively tasted wines in California. Since the country, the state, and the taster were all the same, I could exclude those columns from my data frame. Then I was left with: points, price, description, designation, region\_1, region\_2, title, variety, and winery. I then got rid of description because it was a paragraph and the description was too long to be able to appropriately shorten it to a few descriptive words. After that was gone, I got rid of all of the rows that had missing values. Next, I looked to see how many levels each factor variable had, and if it was over 300, I eliminated it for simplicity. Designation had 2074 levels, title had 4054 levels, and winery had 643. I was then left with points, price, region\_1, region\_2, and variety. Points was my value to be predicted. Price was a numeric value. Region\_1 had 56 levels, region\_2 had 8 predictors, and variety had 88 predictors. One of the goals of this project was to determine what factors made a wine good, so I wanted to be able to see what levels of each factor were important in predicting the wine points. In order to do this and be computationally efficient, I made dummy variables for each factor level. Dummy variables take each level in each of your factor variables and makes a new column for them. They have a one if that particular observation had that level, and a zero if they don’t. This left me with 158 columns and around 6000 observations to do this analysis with.

Data Analysis

For the analysis, I first split the data into testing and training data. I did this using the caret package to randomly pick 75% of the data for training, and the resulting 25% for testing. I then used 6 different models to predict the number of points each wine would receive: simple linear regression, principle components regression, random forest, neural network, support vector machine (linear, radial, polynomial, and sigmoid kernel), and a gradient boosting machine. For each of these methods, I trained the model using the training data set and then tested the model on the test data set. I computed the mean square error (MSE) for each model in order to compare the error rates and to determine which model had the best prediction ability.

Simple Linear Regression:

For the simple linear regression, there were no parameters to tune, so I simply ran the model with all of the predictors.

Principle Components Regression:

This analysis was done using the pls package in R. I tuned the principle components regression model using cross-fold validation. I then used the summary from the cross validation on the training set to pick the number of components to include in predicting the test set. I did this by looking at the amount of variance explained with the addition of each new variable and set the number of components to be equal to the number at which 100% of the variance was explained. I then played around with the numbers and found the optimal number of components.

Random Forest:

I used the randomforest package in R for this model. I tuned the random forest model by first running the random forest in the caret package, so it would tune the model for me. Then using the output from the tuned model, I ran a random forest model without the caret package in order to see the importance of the variables used in the model. I ran this model with the number of variables randomly sampled as candidates at each split (mtry) set equal to 40, based off the results of the caret output (See Figure 1). The number of trees was set at 500.

Neural Network:

I originally ran a neural network without tuning using the nnet package in R and I got a very large MSE value, I then rand the neural network within the caret package to see if it could have been a tuning issue. I received the same MSE, but with a lot of errors due to their being too many weights.

Support Vector Machine (SVM):

I ran the svm model, within the e1071 package in R, with the four different kernels: linear, radial, polynomial, and sigmoid. I compared the MSE of each of these models and then I tuned the model that had the best MSE, because I assumed it would have the most room for improvement. I tuned the radial kernel svm using a range of costs: 001, .1, 1, 10, and 100. I also used a range of gammas: .1, .5, 1, 2, and 10. Using the tune function, I found the optimal cost and gamma and used them to predict my test data.

Gradient Boosting Machine:

In order to complete this analysis, I used the gbm package in R with the gaussian distribution and 5000 trees. In order to obtain the optimal fit, I tuned this model by first adjusting the shrinkage value until I got the lowest MSE, and then going back and adjusting the interaction depth with the new shrinkage value

**Results**

|  |  |
| --- | --- |
| **Model** | **MSE** |
| Simple Linear Regression | 5.05 |
| Principal Component Regression | 4.98 |
| Random Forest | 3.87 |
| Neural Network | 7932.00 |
| SVM: Linear Kernel | 4.41 |
| SVM: Radial Kernel | 4.13 |
| SVM: Polynomial Kernel | 7.78 |
| SVM: Sigmoid Kernel | 1086.60 |
| Gradient Boosting Machine | 3.97 |

The table above represents the results given from each model. If the model was tuned, the tuned MSE is shown in the results.

Simple Linear Regression:

This model showed that 54 out of the 158 modeled were significant at the 95% confidence level. This model had an MSE of 5.05 though, so not very good predictive ability.

Principal Component Regression:

This model used 66 components and had an MSE of 4.98, which was just barely better than the simple linear regression.

Neural Network:

The neural network model had the highest MSE at 7932, which was due to their being too many weights to get an accurate prediction.

SVM (linear, polynomial, and sigmoid):

The svm model with the linear kernel gave an MSE of 4.41, which was better than pcr and the simple linear regression. The svm with the polynomial model, gave an MSE of 7.78, which was significantly worse than the previous tests, except for the neural network. The sigmoid kernel svm did the worst by far of the svm models with an MSE of 1086.60.

SVM (radial kernel):

The radial kernel svm had a cost of 1 and a gamma value equal to .1 and an MSE of 4.13, which is an improvement from the untrained svm with a radial kernel that had an MSE of 4.24.

Gradient Boosting Machine:

The gradient boosting machine predicted very close to the random forest which was the best predictor for this analysis. The tuned model had an interaction depth of 6 and a shrinkage value of .01 with 5000 trees, and 53 variables tried at each split.

Random Forest:

This model used 500 trees which is the default of the random forest and 40 different variables tried at each split. The model had the lowest MSE at 3.87. This is still a large MSE, but since it was the lowest MSE of all of the models used to predict the number of points, I will discuss a little bit about what variables were important in this model’s predictions. Figure 2 is a plot of the importance of the top 30 most important variables in predicting the wine points. It looks like price and Pinot Noir are the most important in predicting the wine points for the random forest model.

**Conclusions**

In summary, the random forest model predicted the best out of all of the models. The MSE for this model was still quite high though, at 3.87. It showed that price, the variety Pinot Nior, the regions Paso Robles and Santa Maria Valley, and the variety Syrah were some of the most important factors in predicting wine scores. This is to be taken lightly though, since the prediction error was still so high. In order to better predict this data set in the future, I have a few suggestions. I think that it would be helpful to classify each country into the continent they belong to. This wasn’t a problem that I had to address since I dealt solely with the United States, but it was an original issue before I decided that that was my best course of action. I also think that it would be nice to be able to reduce the descriptions of the wines that the tasters give into a few short words, such as: bitter, sweet, fruity, etc. That would obviously be time consuming and one would have to know what words come up frequently in wine reviews. I think it would also be helpful to reduce the number of varieties by classifying each as: sweet red, bitter red, sweet white, bitter white, fruity, etc. This way there wouldn’t be 88 different varieties, just in the United States. All of these recommendations would reduce the number of levels of each factor making things a little bit simpler, which I think would also increase the predictive ability.

Something that has come up while reading about wines and wine tasting is the bias that comes with tasters knowing the price, origin of the wine, and even the color! There have been several experiments done in each of these categories where someone submits the same wine, but dresses one bottle up to make it look quite nice and fancy and makes the other bottle seem cheap. The tasters then taste and review the wine and give the cheap bottle a worse review than the nice bottle, even though they are the same wine ((<https://en.wikipedia.org/wiki/Wine_tasting>). A similar experiment was done with the color of wine, where the same wine was served, but one was dyed red. The description of the wine for the red (even though it was technically a white, just dyed) was descriptive of a very typical red wine. And when they tasted the original white wine, they gave very typical white wine reviews. A blind taste test was given to wine tasters with wine from California and wine from France. The California wine outcompeted the French wine, which has not served true in other non-blind tastings (<https://en.wikipedia.org/wiki/Wine_tasting>). There is clearly a lot of bias in wine tastings. This being said, I’m not sure if the data that I used was from a blind tasting or a non-blind tasting which would induce more bias. This is also something to keep in mind though for future studies. Having data collected from blinded tasters would lead to the most truthful, unbiased data, which will give the best predictability.

The largest complications that came with this data were the large number of levels for each factor. I think by using the methods mentioned previously to somewhat reduce these would help tremendously. I also think that the small prediction scale, only from 80-100 made the data more challenging to predict. That is not something that can be rectified though. You could potentially make this a classification problem by assigning bad wines 80-85, decent wines 86-90, good wines 91-95, and great wines 96-100. Then work on classifying each wine into those categories versus predicting the number of points they would receive.

I was really hoping to get a more straight-forward answer when I started this project. Like, Napa Valley is the best region and super important in predicting good wine, also the best wine is a Pinot Noir. I obviously also wanted better predictive ability so that the results that I did achieve would be a good indicator of what makes a wine good. This just goes to show that data in real life is not always clean and easy to handle. I do think that with enough time to have made the changes as stated above to the factor levels that I could have increased by predictive power.

This data set could stand to be cleaned up a bit, which may help with better predictive ability. At this time, the best model to predict this data is a random forest regression model that used price, the Pinot Noir variety, the regions Paso Robles and Santa Maria Valley, and the variety Syrah were some of the most important factors in predicting wine scores. That being said, a highly ranked Pinot Noir is: Meiomi Pinot Noir made in the great state of California (according to <https://www.totalwine.com/shop/best-tasting-pinot-noir>).

**References**

Wine tasting. (2018, November 24). Retrieved from [https://en.wikipedia.org/wiki/Wine\_tasting](https://en.wikipedia.org/wiki/Wine_tasting#Connoisseur_wine_tasting).

**Appendix**

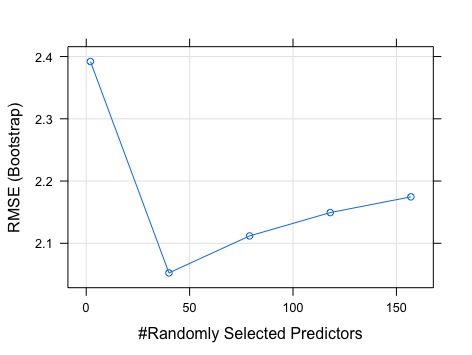
****

Figure 1. Showing RMSE values plotted against the randomly selected predictors for the random forest model ran in the caret package. This shows that the RMSE is at its lowest value somewhere around 40 randomly selected predictors.

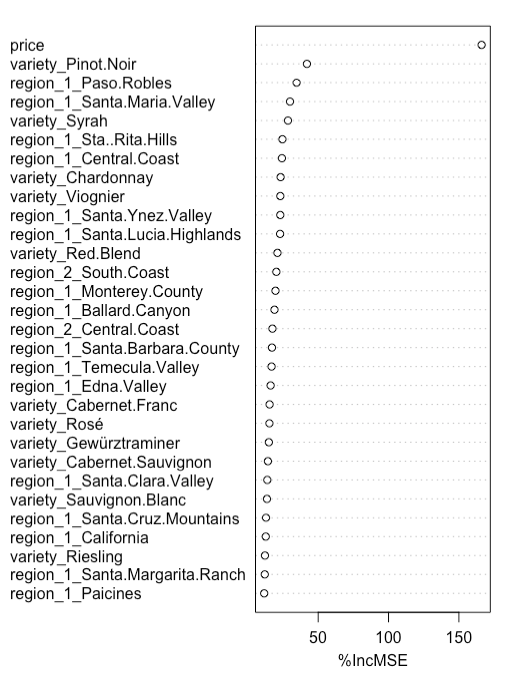


Figure 2. Plot of importance from the random forest model. The more to the right the dots are, the more important they were to the analysis.