Predicting Scores and Prices for Wines

Xinyang Liu: xl9qw
Boh Young Suh: bs6ea
Ni Zhuang: nz4gg

I. Introduction to the Problem

With the improvement of people's living standard, an increasing number of people begin to develop an interest in wine tasting. However, not everyone can identify top quality wines through blind tasting like a master sommelier. Most wine review websites only have basic descriptions and ratings for relatively well-known types of wines, which are either very difficult to find in ordinary stores or beyond the financial capabilities of the majority of consumers. Therefore, it would be desirable to have an alternative way to recommend wines according to a person's demand and financial capacity. By developing a predictive model, we will be able to identify the quality of wine more easily based on the description and some other important features such as its place of origin or the possible price. We believe that wine lovers who are not professional in wine tasting but want to make better purchase decisions would show high interest in our research and deep learning to predict wine ratings. Successful implementation of a model from the wine dataset to predict wine quality will enable wine tasters to better predict variety of wines, select an ideal wine, and also enrich their wine tasting experience.

II. Objectives and Metrics

Our three main objectives are

- What seems to be associated with high review scores and high prices?
- Can the descriptions on wine bottles affect their scores or prices?
- Is there a positive correlation between review scores and prices?

The main objective of our project is to optimize buyers' wine selection process, so that non-professionals can also find the ideal wine that would match their taste. Our obtained model should demonstrate the relations between variables and enable people to predict the score and price of wine based on given information. We will measure how well our model predicts the wine ratings by performing cross validations.

III. Data

Dataset	Source	Attributes
winemag-data_first150k.csv	https://www.kaggle.com/zynicide/wine-reviews/data	Details about each wine (Country, Description, Point, Price, Variety, Winery, etc.)

The dataset from Kaggle contains information about over 150K observations of various wines and our team plans to use this dataset to build our model. Previous researches have investigated in the general summary of the dataset -- different levels for categorical variables and the range of values for numeric variables. The only data mining method that has been applied to the data was the basic linear model, and the results it achieved, like Residual Standard Error (squared root of mean squared error) and Adjusted R-squared, were not as desired. As a result, we still have not been able to find an appropriate way to predict the rating and price of certain wine, which is necessary in order to make recommendations for non-professional wine lovers. Also, one possible problem with our dataset is that several predictors are related to the origin of certain wines, which might indicate some degrees of multicollinearity between variables.

IV. Hypothesis

Our preliminary hypotheses are

- The descriptions on wine bottles affect the wine's score and price (Whether or not certain words are included)
- There is a positive correlation between review scores and prices
- The place of origin will affect the wine's prices and ratings

We plan to use both parametric and nonparametric, linear and nonlinear statistical method to analyze our data, applying basic Linear Regression, K-Nearest Neighbor (KNN), and Random Forest to test our hypothesis mentioned above. In addition, with a descriptive column in our data, we plan to make use of text mining to uncover the underlying relationship among wine description, the score and its price. Text mining will focus on the description of each wine to figure out how the description of wine with certain words affect the evaluation of wine. With text mining, if descriptions of top rated wines and of low rating wines shows a big difference, then it supports our assumption that the descriptions on wine bottles affect the wine's score and price. Also if there is a sign of significance between the review score and price as well as a positive

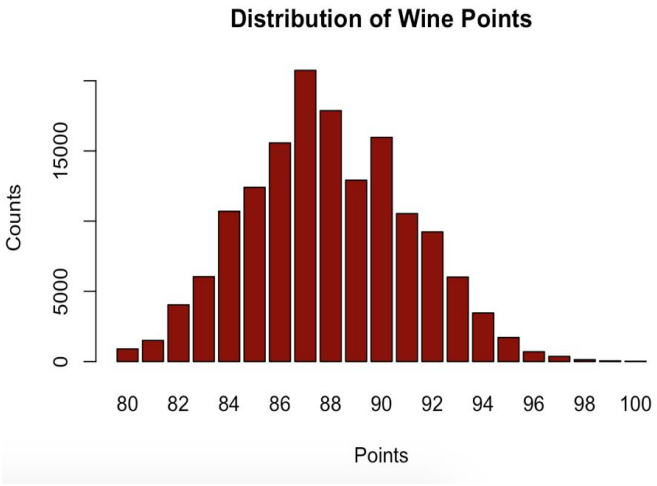
coefficient seen in the model obtained for each other, we can conclude that a positive correlation exists between the two variables.

V. Approach & Result

- Data Preprocessing

Our original dataset contain 150K rows of wine data. There are two region information, one for the wine growing area in a province or a state, the other for a more specific region specified within a wine growing area. However, we found out that there were a lot of missing data on information for the second region. Therefore, we replaced those missing values by matching them with general area information from the first region.

We also converted points for each wine to a 10-based rating system to categorize points for more effective analysis. In order to have a similar number of observations for each rating to avoid imbalanced dataset, we made some adjustments to the thresholds that split the data. The below graph showed the distribution of our wine data points. Most of the wine points ranged from high 80s to low 90s. In addition, some of the categorical variables in our dataset had many different levels, such as designation which has more than 23,000 different levels, and many of them only had one observation. Thus, we decided to reclassify the ones that have fewer observations in each level and renamed it to “Others”. This decreased the number of levels for each variable to less than 53 and as a result made our further analysis much smoother.

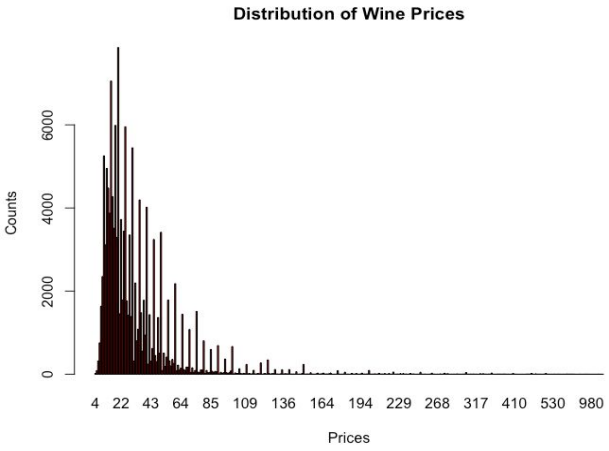


→

Wine Points	Rank
80 =< points < 84	1
84 =< points < 86	2
86 =< points < 87	3
87 =< points < 88	4
88 =< points < 89	5
89 =< points < 90	6
90 =< points < 91	7
91=< points < 92	8
92 =< points < 93	9
93 =< points < 100	10

- Data Exploration

After appropriate data preprocessing, exploratory analysis was done to have a better understanding of the variables within our dataset. Distribution of wine prices showed a highly skewed result with the lowest price wine being $ 4 and the highest price $ 2,300. The majority of wine prices were ranged from $10 to $50 as shown in the right graph. Also, as the distribution of prices was extremely imbalanced, we attempted to do log transformation for prices during our analysis and prediction.



In addition, while trying to figure out the relationship between wine points and their prices, we plotted the points against prices. From figure 1, we can see that there was not a clear correlation between the two variables, 0.4598634 to be exact; although lower scored wines tend to be cheaper, not all expensive wines received high ratings. After log transformation on price, there appeared to be an obvious upward trend in the plot of points versus log of price, and the correlation became much more explicit, with a value of 0.6111787.

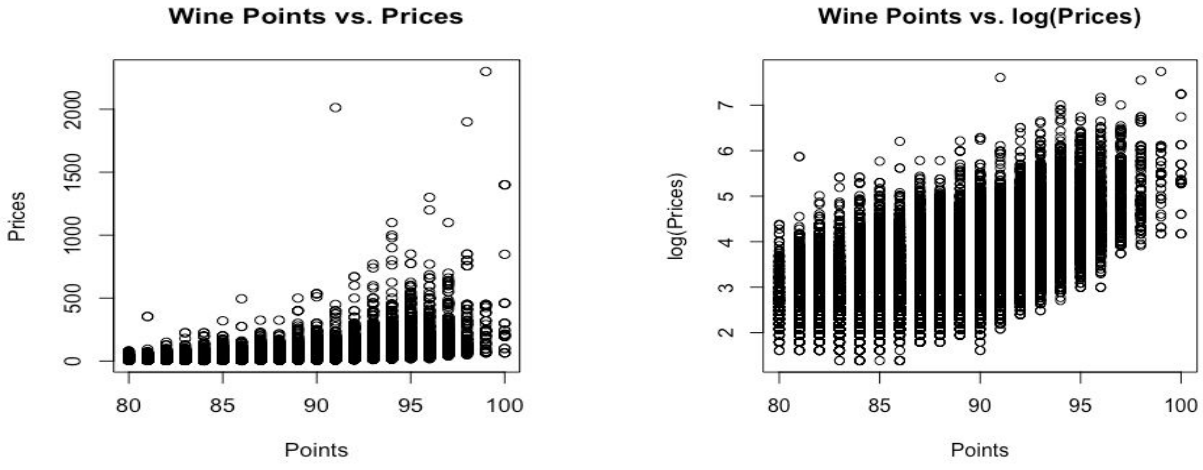


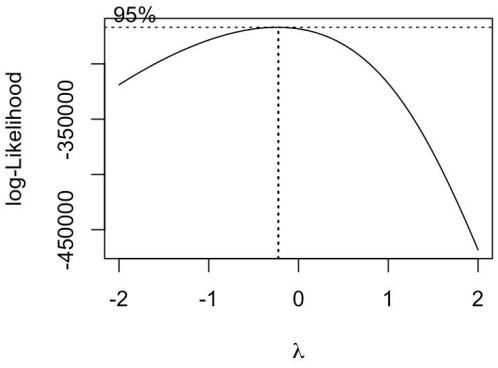
Figure 1

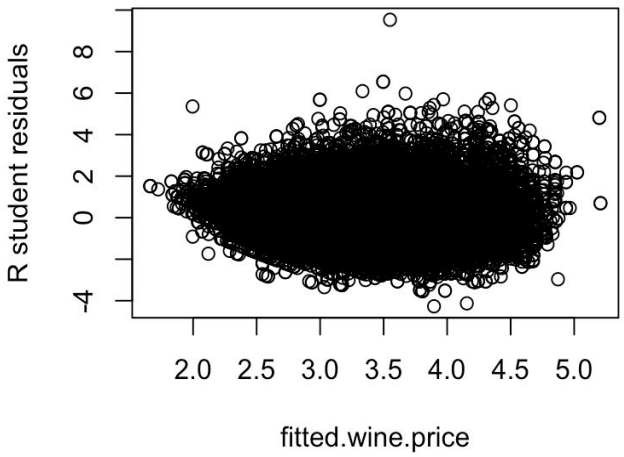
- Model Implementation

Our team applied various data mining approaches in order to analyze the wine dataset, such as basic Linear Regression, K-Nearest Neighbor (KNN), Random Forest, and Text Mining. The predicting powers of our models were not so good in general, but KNN gave a relatively better results compared to other methods.

Linear Model

We first decided to begin our analysis executing linear regression and followed a parametric approach to predict wine prices and points. Our outcomes for predicting points had a very low explanatory power, which was about 36.29%. This could be due to the large number of categorical variables in the dataset. After performing a forward, backward and stepwise model selection using penalty criteria such as AIC, our final model kept every variable we had without removing any. “Country” variable had very high multicollinearity with region and province, but we decided to keep it in the model because the R squared score and sum of residual squares were even worse if we removed the “country” variable.





The linear model used for predicting wine price performed better than predicting points. We performed a log transformation on the response variable “price” based on the lamda score from the box cox plot. Transformation improved the R squared score from 34.83% to 58.26%, and the training MSE dropped from 869 to 0.2 after transformation.

The normal probability plot showed relatively normal distribution but were not perfect in tails, and the residual plot had relatively constant variance. However, we can observe that as wine price went up, the variance of residual was a little unstable. Our explanation for this relative

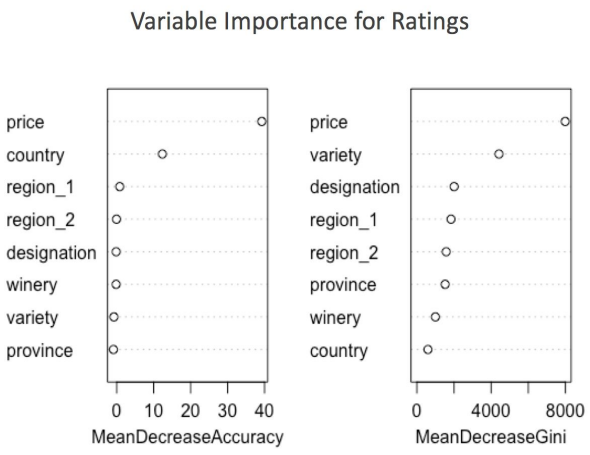
inconsistency was that high priced wine had a wider range of ratings and relatively inaccurate predictions, which resulted in the larger variance in residuals. But the predictions for the lower priced wine are likely to be more accurate.

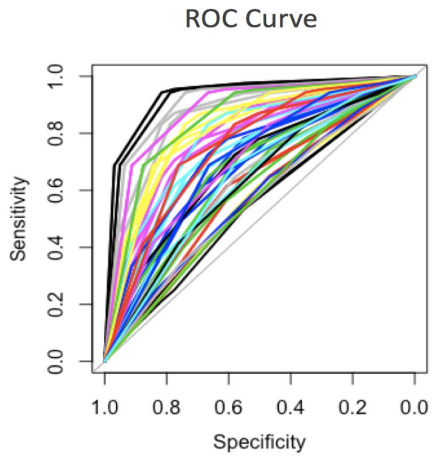
K-Nearest Neighbor (KNN)

KNN model was built for both wine prices and wine ratings. In this process we had to remove columns that had low variance. NearZeroVar function was used setting frequency cut 19 and unique cut 10. As a result, the best model was selected using the smallest value of RMSE which was 2.112333. The final value used for k was 5 and the R squared was 43.54%. On the other hand, the model built to predict price had a slightly better results. RMSE was also selected to identify the best model. The smallest RMSE was 25.75544 and that was when k = 5. R squared of the model was around 50.16%, explaining half of the data.

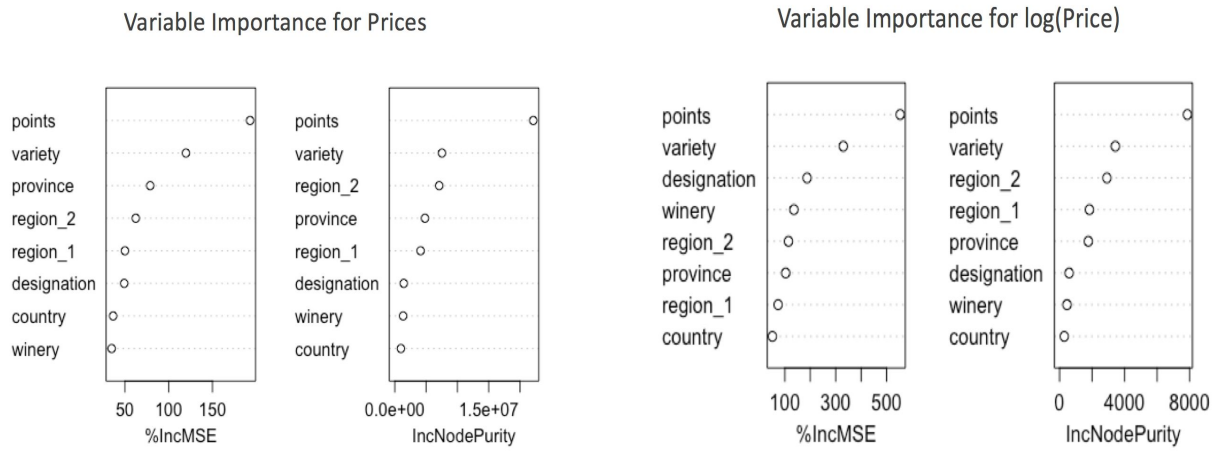
Random Forest

We also ran random forest to analyze the data using nonlinear statistical learning method. According to the variable importance plot for ratings, we can see that based on mean decrease in accuracy, price of wine and its country of origin affected the rating the most, and in terms of mean decrease in the Gini index, price and wine variety were the most crucial factors. Prediction on rating exhibited a very high misclassification rate, but an Area Under the Curve (AUC) of 0.7095 indicated that much of the data were explained despite of the high misclassification rate.





On the other hand, applying random forest on price resulted in a pretty high MSE of 489.0404 at first. However, after log transformation on the values for wine price, the MSE dropped dramatically to only 0.1355413. In both cases, review points of wines and their varieties were the two most important variables contributing to the prices.



Text Mining

In addition to building regression models on the wine dataset, our team also wanted to focus on the description on the wine to see if any specific words in the description were correlated to high, low scores and prices. We subset top scored wines and low scored wines as well as high price and low price wines based on the overall distribution of the variables, and hoped to see that there would be some differences in the wording of their wine descriptions. When we first implemented word cloud, we found out that the top words shown in each model were the same, with “wine” and “flavor” being the top two words, so we decided to remove those two words. Below are the four different word clouds resulted from text mining. The description of top rated wine and high priced wine were focused more on the age of the wine and specific ingredients that were used to make wine. Words such as “rich”, “age”, “tannin”, and “complex” were founded in common. On the other hand, the descriptions on low scored wine and low priced wine had more words describing the taste, such as “sweet”, “aroma”, “acid”, and “dry”, showing a clear difference from the high scored and high priced wines.

Next, we built a nonparametric model to make use of information from description to predict wine ratings and prices. The goal here was to allow people to assess the quality of certain wine based on its descriptions or past reviews of that wine. We implemented K-Nearest Neighbor

(KNN) to make such predictions. As for wine price, the prediction accuracy was about 98.24% when setting K to 5 and 98.25% when K equals to 10. As log transformed price tend to give us better results in our previous analysis, we tried to transform the price to see if that would make any difference in predicting. Surprisingly, the accuracy dropped to about 38.75% using 5-folds and 40.65% using 10-folds. Thus, using the original variable to implement KNN approach was better. The accuracy of wine rating predictions from KNN was even higher, with the accuracy rate as perfect as 100% in either 5 or 10-folds.

Word Cloud for Top Scored Wine

Word Cloud for Low Scored Wine

Word Cloud for High Pricing Wine

Word Cloud for Low Pricing Wine

VI. Conclusion

On account of all our previous data process and model analysis, we can conclude that our preliminary hypotheses are mostly supported. Our analysis indicates that the description of wine is closely related to wine rating and price, and differ in content based on their values. Moreover, according to the correlation values produced before and after log transforming price, we can see that there is a positive relationship between rating score and price. Also, all factors are significant in predicting wine score and price, especially the place of origins for wines, which is determinative in the wine qualities. With the results of our analysis, consumers would be able to estimate the value and quality of certain wine based on the information about the wine's origin, variety, and its description. We are positive that any amateur wine lovers can also appreciate wine like connoisseur while employing our model to make a better choice in wine selection in the future.