Wine Review and Recommendation

Xinyi Song (xinyis8)

Project description and summary

I obtained wine review data with more than 120000 wine reviews written by at least 20 different toasters. The wines reviewed came from 42 different countries and ranged in price from $4 to $3300. Reviews included description and a rating of the wine on a 100-point scale.

In this project, I aimed to predict the points rated by toasters on a scale of 1-100 by performing machine learning models. Besides, based on results of analysis, I would also recommend five different wineries for customer who is interested in purchasing a fruity pinot noir, with a price less than 20 dollars.

I randomly divide whole data set into two parts: training data (75% of total data) and test data (25% of total data) and used the training data to train models and test data to calculate MSE (Mean Square Error) to evaluate model prediction accuracy. By applying text mining to choose features and machine learning models such linear regression with shrinkages and tree model, after tuning parameters, the xgboost model reduced the MSE to 2.6409 and improved prediction accuracy a lot. Also, based on my analysis, I gave several recommendation options of winery for customer with specific requirement for wines.

Data preprocessing and Feature Selection

The wine review dataset includes more than 120000 observations and 14 variables including points, description, designation, price, variety, winery, toasters and so on. And point is the response variable for analysis.

For feature selection in wine review description, I used package (‘tm’) in R to do the text mining in the following steps: Set all words to lowercase; remove stopwords and extremely rare words; remove punctuation and other symbols; remove unnecessary whitespace. Then, I established a documentary matrix to extract the features, for each key words, if it is mentioned in this review, the corresponding element in that matrix is 1, otherwise it is 0. Therefore, for the documentary matrix, each key word forms one column, each row stands for one observation, it is a sparse matrix. I set the parameter ‘sparse’ in ‘removeSparseTerms’ function as 0.997 to choose features considering both efficiency and model complexity.

For numeric variables such as price, I replaced the missing values with its mode. For categorical variables, I dealt with their missing values in the following steps:

(1) Designation: Winemakers typically label wines in one of two ways: by grape varietal or by style. And not all wines have designation, thus I label the wines with designation as 1 and otherwise as 0.

(2) For missing values in variable country, province, taster name, taster_twitter_handle, I label them as ‘others’.

(3) There is only one missing value in variable variety, I used the most frequent level to replace it.

And there are no missing values in points, title and winery. Finally, I choose features in description, designation and price into model building. I dropped the variety, region, province and country because there are too many levels in those variables, putting them into model building might lead to too much noise and lower efficiency. I consider this variable in the future recommendation part. To make it applicable in regression and xgboost model analysis, I used model.matrix to create the design matrix.

To properly evaluate model performance, I randomly divide the whole data set into train data (75% data) and test data (25%). I used the train data to train my model and calculated MSE with test data to evaluate model fitting and predication accuracy.

Descriptive statistics

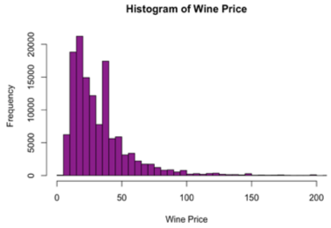
Part 1 Text Mining

For part of text mining, after feature selection, I chose approximate 1100 words to train my model.

Part 2 Numerical Variable

Table I Descriptive Statistics of Price

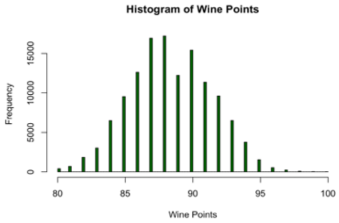
Min	1st Qu	Median	Mean	3rd Qu.	Max.
4.00	18.00	28.00	35.62	40.00	3300.00



Based on the results above, we can see that the range of price is extremely large, mean value is 35.62 and median is 28. But the maximum reaches 3300, I would track the wine with extremely large and low price and explore the influence of price on points of wine.

Table II Descriptive Statistics of Point

Min	1st Qu	Median	Mean	3rd Qu.	Max.
80.00	86.00	88.00	88.45	91.00	100.00



Based on the results above, we can find that the distribution of points approximates normal distribution, their mean and median approaches 88, and maximum of it is 100.

Part 3 Categorical Variable

Country: Except missing values which are labeled as ‘others’, there are 43 levels in that variable. And top three most frequent levels are US, France, Italy.
Province: There are 146 levels in variable province, and the most frequent level is California with 36247 wine reviews.
Variety: There are 708 levels in variable variety, and the top two most frequent levels are Pinot Noir and Chardonnay, both with more than 11700 wine reviews.
Winery: There are 16757 levels of winery, and the top four most frequent levels are Wines & Winemakers, Testarossa, DFJ Vinhos and Williams Selyem.

Finally, I choose features in description, designation and price into model building. I dropped the variety, region, province and country because there are too many levels in those variables, putting them into model building might lead to too much noise and lower efficiency. I will consider this variable in the future recommendation part.

Regression model analysis

In this part, I tried to use two kinds of models to predict points of each wine: Linear Model with shrinkage and XGboost model.

Linear Model

I tried to use linear regression model without shrinkage first with variable of features from description, price, designation. The R-square is 0.651 which indicates that it is a good fit. The variables acidity, adding, aftertaste, aged, alcohol, almond, American, appetizing, apple, barbecue, banana, berries, bitterness, blackberries, blackberry, blend, blue, blueberry, brightened, butter, cake, candied, cherry, chocolate, cigar, cinnamon, citrus, classic, cocoa, coconut, coffe, cola, cranberry, cream… have P-value less than 0.05, so do price and designation, which indicates they do have significant influence on point. And the MSE in linear model is 3.45.

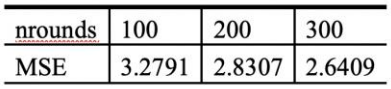
Then I tried both lasso regression with l1 norm shrinkage and ridge regression with l2 norm shrinkage. By using cross validation (cv.glmnet) I got lambda.min, and then used lamda.min to run my lasso and ridge model. The MSE shows that both ridge and lasso regression perform a little better than linear regression in terms of MSE but not reduce it a lot. The MSE of lasso regression is 3.3337 while that of ridge regression is 3.3319.

The advantage of linear model is that it is simple and has good interpretability. But since the relationship between points and predictors is not linear, it might not perform as well as I expect.

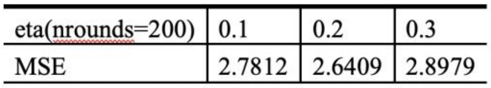
Tree Model

As for the tree model, I chose to use boosting tree to analyze our data and do prediction. For boosting tree, trees are grown sequentially: each tree is grown using information from previously grown trees and each tree is fitted on a modified version of the original data set. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. This algorithm uses a more regularized model formalization to control over-fitting. I tuned parameters by setting max number of iterations as 100, 200, 300 since too large nrounds will lead to low computation efficiency and overfitting and too small nrounds value will lead to lower accuracy. Also I set eta as 0.2 which is closed to default value at first, which controls the learning rate.

Here, nrounds controls the maximum number of iterations. For classification, it is similar to the number of trees to grow. And eta controls the learning rate, i.e., the rate at which our model learns patterns in data. After every round, it shrinks the feature weights to reach the best optimum. Lower eta leads to slower computation. It must be supported by increase in nrounds. And the MSE is the left table:



nrounds	100	200	300
MSE	3.2791	2.8307	2.6409



eta(nrounds=200)	0.1	0.2	0.3
MSE	2.7812	2.6409	2.8979

Then I tuned eta and set nrounds=300, the result is as above (right table). The result shows that for our data, eta=0.2 and nrounds =200 will give relatively smaller MSE. And the plot of the first 30 features in terms of feature importance is as following:

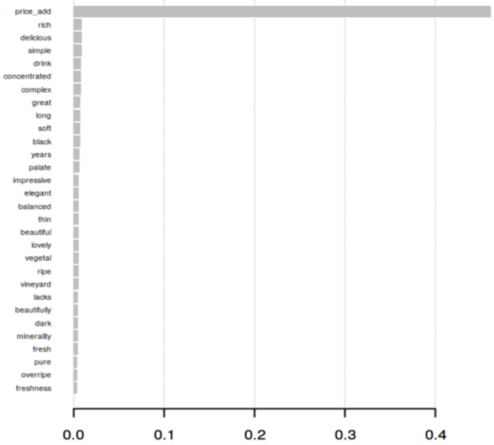


Figure II Top 20 Important Feautre of XGboost

Here, the price_add is the price of wine. The result of xgboost shows that variables such as price, delicious, soft, thin, vegetal, beautiful, elegant, balanced, ripe, vineyard, dark, minerality, pure, overripe, freshness could have great effect on points of wines.

Recommendation

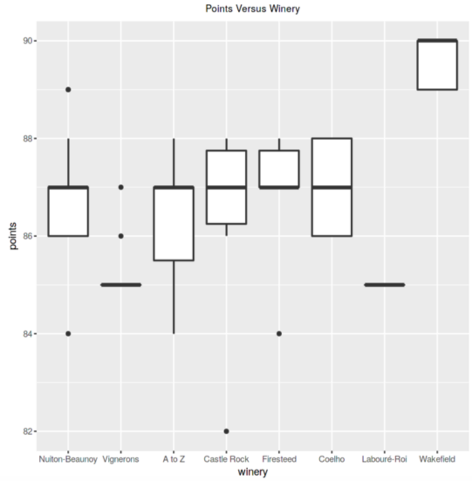
For customer who is interested in purchasing a pinot noir, with a price less than 20 dollars, and has a fruity taste, recommend five suitable wineries.

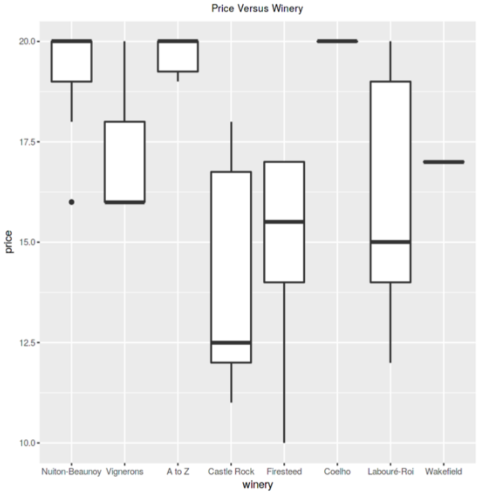
After screening the suitable values which satisfies the preferences above, I got 397 wineries. Then I sorted the wineries with more than five wines to narrow down the range since to pay more attention to different aspects with smaller screening range; Also, wineries with more targeting wines could have more chances to design suitable wines and thus be more likely to gain loyal customers. After screening part, I got the following wineries: Nuiton-Beaunoy, Vignerons de Buxy, A to Z, Castle Rock, Firesteed, Coelho, Labouré-Roi, Wakefield. For each wineries, I did analysis regarding their points, price, country, province and region.

Nuiton-Beaunoy located in Burgundy, France. And there are twelve kinds of targeting wines in this winery. The region of these wines are Bourgogne Hautes Côtes de Beaune with 2 wines, Hautes Côtes de Nuits with 4 wines, Bourgogne Hautes Côtes de Nuits with 4 wines, Bourgogne with 2 wines.

Vignerons de Buxy also located in Burgundy, France. And there are seven kinds of targeting wines in this winery. The region of these wines is Côte Chalonnaise. For points, they are all around 85, and its price is around 16.

A to Z located in Oregon, United States. And there are six kinds of targeting wines in this winery. The region of these wines is also Oregon. For points of its wines, they are around 87, and the price of them are around 19 and 20.





Castle Rock located in United States. And there are six kinds of targeting wines in this winery, three of them located in California while two of them located in Washington. The region of Castle Rock lies in Mendocino, Columbia Valley (WA) and Russian River Valley. Comparing with other wineries, the range of its points is relatively larger, its point is around 86.5. For price, its price is relatively cheaper than other wineries do. The mean and median of price is around 13, and the range of price is also relatively large.

Firesteed located in United States. And there are six kinds of targeting wines which are all located in Oregon. The region of Firesteed also all lies in Oregon. The median and mean of its point is around 87 and the its median and mean of its price is around 15. There is one kind of wine

Coelho located in Oregon, United States. And there are five kinds of targeting wines which are all located in Oregon. The region of Coelho is in Willamette Valley. The price of it is all 20, which could be more expensive than wines in other wineries. And also, the quality in terms of points of wine is also better and more stable, all of its wines suitable for this customer have points larger than 86. And the mean and median of it is 87.

Labouré-Roi located in France. And there are five kinds of targeting wines in this winery, two of them located in Bourgogne Hautes Côtes de Nuits while two of them located in Bourgogne. Only one of it located in Pays d'Oc. The region of Labouré-Roi lies in Mendocino, Columbia Valley (WA) and Russian River Valley. The province of them is Burgundy for four wines and Languedoc-Roussillon for one wine. The point of its observations is all 85 and the price is around 16.

Wakefield located in Australia. And there are five kinds of targeting wines in this winery: all of them are in province South Australia and region of Adelaide Hills. The point of it is relatively higher than wines in other wineries and all targeting price in this winery is 17. Based on the analysis above, I give the following recommendation options:

For people who cares price, I will recommend Wakefield, Firesteed, Castle Rock, Nuiton-Beauno and Vignerons de Buxy for their relatively high points. The reason is as following:

Wakefield has high cost-effective quality. Firesteed and Castle Rock will also be chosen for its relatively lower price and great performance in review points. And these three wineries are from Australia and USA. For diversity, in other words, to let my customer taste different styles of wines, I will also consider two wineries from other countries. I recommend Nuiton-Beaunoy since its wines have acceptable price and high points, also this winery has most targeting wines among all wineries, which indicates it could be more likely to produce wines that satisfies my customer' demand. Also, Vignerons de Buxy will be chosen, because it also produces many targeting wines for this customer and the min point value is 85, its wines' quality is most stable among all wines.

For people in USA who preferred local wines, I recommend Castle Rock, Coelho and Firesteed which located in USA. And also Wakefield in Australia and Nuiton-Beaunoy in France for both diversity and good taste.

For people in France who preferred local wines, I recommend Nuiton-Beaunoy, Vignerons de Buxy and Labouré-Roi which located in France; also, Wakefield and Firesteed for both diversity and great taste.

Table III Important Features of targeting wines in each winery

Winery	Features
Nuiton-Beaunoy	acidity, tannins, ripe, soft, bright, perfumed, attractive,
Vignerons de Buxy	red, light, juicy, tannins, friendly, balanced
A to Z	pinot, noir, lightly, cherry, candied, banana
Castle Rock	noir, tart, red, complexity, savory, oak, green, crisp, cherry, bay, aromas, add, acidity
Firesteed	pinot, cinnamon, cola, citrusy, bright
Coelho	tart, raspberry, sharp, gentle, earthy, cola, bitter, berry, aromas
Labouré-Roi	soft, cherry, red, light, tannins, texture, perfumed, fresh, flavor, aftertaste, adding
Wakefield	noir, aromas, raspberry, oaky, mouthwatering, mediumbodied, loaded, herbal, floral, fresh, cedar

For people who has special preferences, I have sorted the top features for each winery regarding its targeting wines. For people who has preference for tannins, I recommend Nuiton-Beaunoy, Labouré-Roi, Vignerons de Buxy. Since tannins usually combines with acidity and bitter, I will also recommend Castle Rock and Coelho.

For people who has preference for fruit berry and cherry, I will recommend A to Z, Castle Rock, Coelho, Labouré-Roi and Wakefield.

For people who has preference for light taste, I will recommend Vignerons de Buxy, A to Z, Labouré-Roi. Also, Nuiton-Beaunoy and Firesteed for its brightness.

Conclusion

Based on the analysis above, the XGBoost model with tuned parameters reduced MSE to 2.64 which reaches great score prediction accuracy. Besides, by text mining and analyze them from different dimensions, I gave different recommendations of wineries for different customer with specific requirement for wines.