Wine prices prediction from wine reviews dataset

Bakirov Aslan

12/11/2020

0.1 Introduction.

The wine reviews dataset from Kaggle provides users with a myriad of opportunities to explore, visualize, and construct models from using different techniques and is also great for wine lovers to explore. This paper will talk over the methods used to explore the famous dataset and the construction of models, stemming from data exploration, which predict the wine prices.
The goal of the project is to create a price-predicting model that uses the variables extracted from the data set to predict the price of the wine. So, a potential user of the model can compare the wine characteristics (e.g variety and vintage) and obtain the price of the bottle similar to what he or she possess. This is supposed to be advantageous for relative valuation of wine bottles in terms of price, similar to house pricing models.

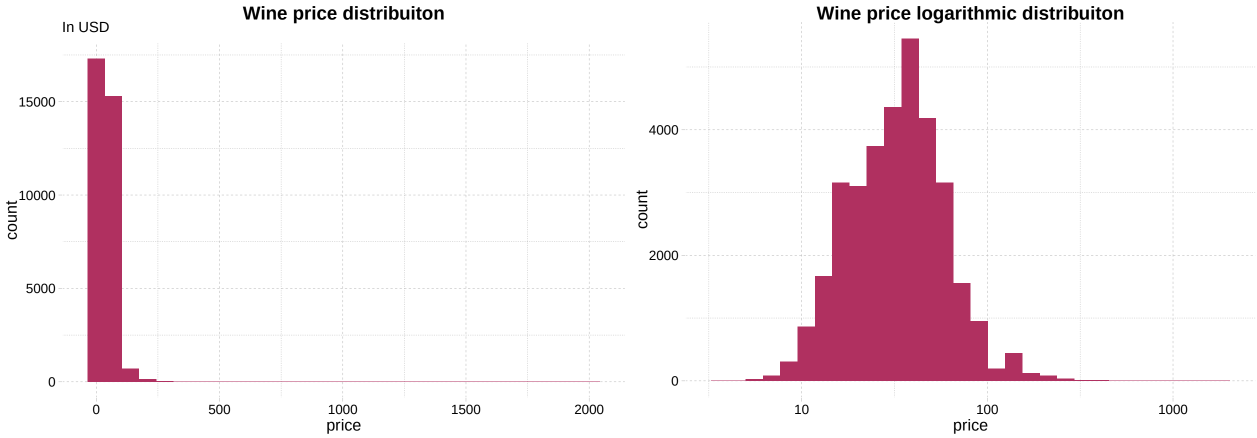
0.2 Data exploration.

To begin with, let us fix U.S.A and California as the origin of the wine. Such a decision has two reasons: the chosen location has the highest number of observations (about 36000), and there are too many dimensions, thus it would intricate and enormously expand the model. Also, the dataset provides more information about province, region1 and 2 and designation for U.S.A, while for some countries these are rarely available. The year when the data was scraped last time was 2017, hence we take it as base year where needed.

0.2.1 Pricing in reality:

Wine World has it's own pricing segments and specific characteristics such as region, type of grapes, pure one sort or mashup, how a bottle of wine tastes "varietally" correct (typicity), environmental factors in which wine is produced including the soil, climate and topography (terroir) form a wine segment. "Extreme value" wine is the cheapest segment and it usually costs under 4 USD, consumers can expect bulk wine with no distinction. The most expensive segment is Icon, and its price starts from 200 USD," Icon" is unique wines, wineries, and microsites. If a wine contains terroir, typicity, and hand craft is presented in its production, then most probably it falls into "Super-premium" segment. It is a median segment, price varies between 20-30$.

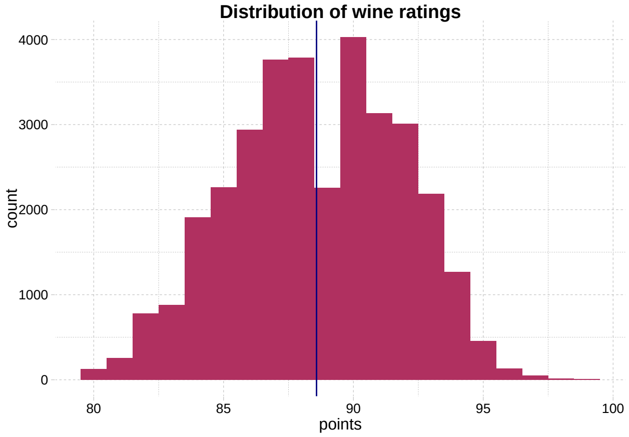
Additionally, wine prices are mostly defined by the cost of the land on which the winery is located. The twisty part is the fact that those land prices are also affected by the prices of wine produced on them, in previous years.Hence in reality the prices of wine are most probably serially autocorrelated.

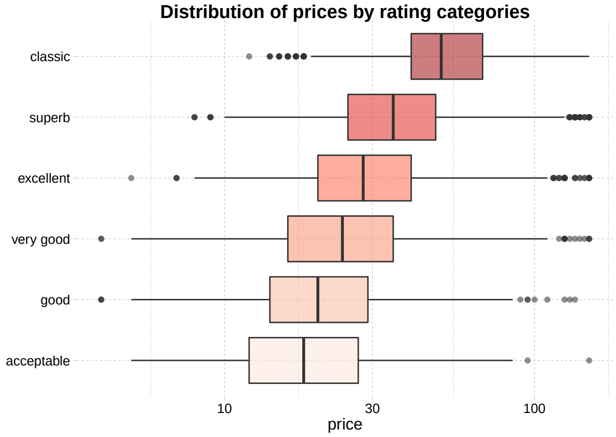


The price variable, being a numeric, has a skewed distribution and contains huge outliers, as we can see from the histogram with absolute prices. It is better to first convert it to the logarithmic scale and then broom the outliers. This way we save more observations, instead of cutting price under 72USD, we cut it under 150$.

Luxurious (Icon) wines for skyrocketing prices usually are traded on auctions and it is unreasonable to include them in the model and expect rational pricing there only based on the variables we have, since in auction case value for individual is essential to know, which we do not know. Additionally, the right-side skewed distribution pinpoints the fact that majority of wines are under 200 dollars. The boxplot confirms this observation.

0.3 Points distribution

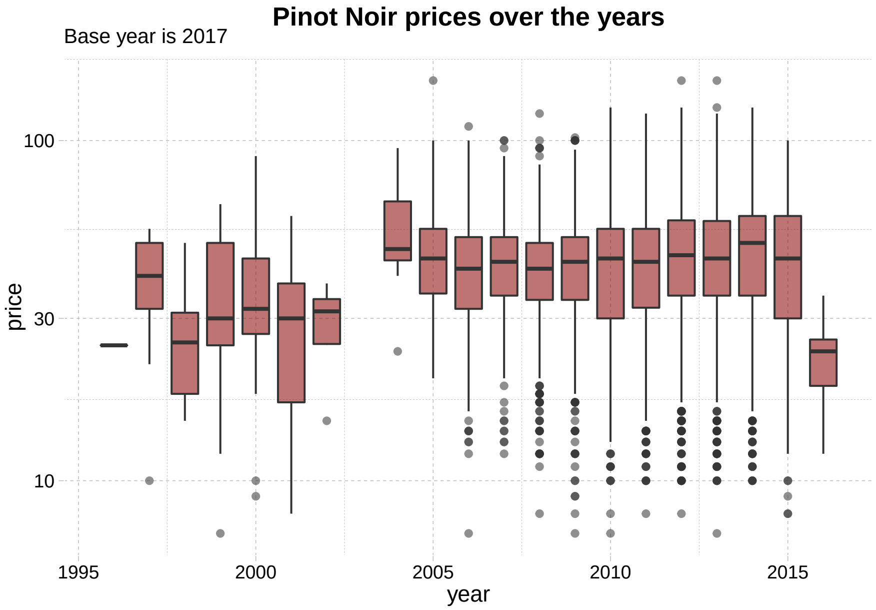




The **winemag** journal, from which the dataset has been scraped, provides reviews and concludes points as their conclusion. These points range from 80 to 100, and the wines are classified as Acceptable,Good,Very Good,Excellent,Superb, and Classic. The wines deemed unacceptable (under 80) are not included. The above histogram provides an insight that the wine points are fairly normally distributed. We can not only use points as is, but also create aforementioned categories and include them as factor variables.

0.4 Feature exploration from the description/title.

The title column is extremely useful for the analysis, since it contains one of the most essential traits of wine - its year. The relationship between wine price and its year of production is of a major interest in the real world. Simply including it to our model would not give the whole picture, due to this convoluted relationship. In reality, wine becomes pricier when it ages, yet after hitting a certain extent it expires and starts losing its value. Except for other factors which affect the price through the year, such as the harvest quality during a given year, or possible occurrence of a cosmic event, and even the performance of the economy in a year, this assumption should perfectly hold. To replicate this kind of relationship, I decided to express price as quadratic function of year.



0.4.1 Oaked and non-oaked wines.

Among other features extracted from description/title, are oak, showing whether a wine was contained in oak barrels. The oaked wines have a tendency to be more expensive, since the containers themselves are costly (a couple of thousand dollars for a new one). Apart from this straightforward reason, another advantage is the oaky notes which take the taste of wine to a whole new level of richness. The oak feature has been examined during our bootstrapping homework where we checked whether the oaked wines are pricier than the rest.

0.4.2 Reserved wines

Another feature capturing our interest is whether the wine is reserve or not. The idea is that winemakers reserve a portion of a rare quality vintage wine, thus ensuring that wine is also aged. This quality is represented by the label on a bottle, and the validity of the label is regulated in some countries. For instance, in Spain it is a must for a wine to be preserved at least 3 years, 6 months of which should be in oak barrels to put the label "Riserva". Yet, some countries have no back-up for the whole reserve story and thus it is of greater importance, especially in US, where no regulation binds winemakers to reserve a wine for a couple of years before putting "reserve" on the bottle. The hypotheses that prices are the same have been rejected at extremely significant levels, stating that at least the Californian wines we have in the dataset, follow the tradition.

reserved FALSE TRUE 30598 1962

```
Welch Two Sample t-test
```

data: wine_revs[reserved == 1, price] and wine_revs[reserved == 0, price] t = 11.404, df = 2126.3, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 6.086727 8.614944 sample estimates: mean of x mean of y 44.84149 37.49065 ### Vintage wines.

The vintage is the reference to the year of the grapes from which the wine was made. In other words, vintage wines are the ones in production of which goes the harvest of one particular year. On the other hand, wines that are made from a mix of grapes from distinct years, are marked as non-vintage (NV). I decided to parse it from the description to get some insight from the taster, since almost all wines at this step can be considered vintage. It happened when I ditched the NA values from year column. Thus, this vintage column represents wines which are recognized by the taster and hence the description of the wine contains the word vintage.

0.5 Text analysis

0.5.1 The number of words and number of charachters.

Next, I decided to make use of one most important column and applied text analysis with *tidytext* package to unnest words from the descriptions. The aforementioned step is also used to add two new variables – number of words and number of characters in description for each wine, since it might be the case that these are strongly correlated.

0.5.2 Sentiment analysis.

We have established above that the description is the most important column in our dataset, yet it is to blunt to measure the effect of the vast description on the price. A taster can say a myriad of words to express a neutral position on the particular wine, as well as positive. In an attempt to understand the attitude of the taster, I took the *AFINN lexicon* for a light sentiment analysis, which sorts words by a score from -5 to +5 according to their attitude (negative-positive). Then, I added another variable as a sum of all those sentiments in the description of the wine. Now I estimate how price changes when the taster use a slightly more positive word/words.

```
# quantitative representation of sentiment with afinn:
desc_text<-
anti_join(desc_text,stop_words,"word") %>%
  filter(word!="wine") %>%
           filter(word!="drink") %>%
  mutate(sent_num=0) %>%
  left_join(get_sentiments("afinn"),by="word") %>% .[is.na(value)==T,value:=0]
```

```r
# excluding oaked words:
desc_text<-desc_text[!str_detect(desc_text$word,
                                 "oak"%R%optional(ANY_CHAR)%R%optional(ANY_CHAR))&
                       !str_detect(desc_text$word,DGT%R%optional(DGT))&
                       !str_detect(desc_text$word,
                                 "reserve"%R%optional(ANY_CHAR)%R%optional(ANY_CHAR))&
                       !str_detect(desc_text$word,
                                 "vintage"%R%optional(ANY_CHAR)%R%optional(ANY_CHAR))]
## the variety and region names to be excluded:
wine_names<-unnest_tokens(wine_revs,wine_names,variety) %>% .[,unique(wine_names)]
desc_text<-desc_text[!word%in%wine_names]
reg_names<-unnest_tokens(wine_revs,wine_names,region_1) %>% .[,unique(wine_names)]
desc_text<-desc_text[!word%in%reg_names]
wine_revs<-left_join(wine_revs,desc_text[,.(sentiment=sum(value)),V1],"V1")
glimpse(wine_revs)
```

0.6 Grouping and adding all variables together.

Some of the most frequently appearing varieties (e.g., Pinot Noir) and regions such as Napa Valley are distinguished, and the rest of these columns were grouped as factor stating that they represent Other variety and region. This is an easing step to include some of the important variables to the model without over-complicating it. Afterwards, all these factor levels are casted into dummy new columns and added to the set of predictors.

```
##           subregion   N.x             region_1  N.y
## 1:          Sonoma 18974        Sonoma Coast 1367
## 2:          Sonoma 18974       Sonoma County 1120
## 3:          Sonoma 18974 Russian River Valley 2834
## 4:   Central Coast 18974         Paso Robles 2099
## 5:            Napa 18974         Napa Valley 3991
## 6: California Other 18974          California 2175
```

After grouping, we should pay attention to the sub-regions within regions, since they may overlap. In US that may happen since the division does not follow clear geographical pattern, instead the division is made in accordance with the set of geopolitical factors (from soil quality to special legal aspects).

In particular, notice an overlap of sub-regions in Russian River Valley, Sonoma Coast, Sonoma County and Other regions, all these have sub-regions named Sonoma. Besides, this issue occurs with Napa Valley, Paso Roble regions too. One solution to this is to add these sub-regions occurring several times as dummy variables, and in the model include as interaction variable with clearly defined regions. Thus, we would have price effect of wine's location in Sonoma-Russian River Valley apart from effect of Sonoma-Other region. This way we fully establish grounds for leaving the "Other" factor level as the base to compare the rest with.

0.6.1 Adding most frequent words and casting.

Finally, I use a thousand of the most commonly used words in the dataset, to add to our models. The words we have cleaned above are supposed to present a greater help in contrast with such methods as tf-idf, which if applied for description, would only give us exclusive words for one particular wine, not helping much with predictions.

Afterwards, let's finalize the data exploration with modifying our dataset from long to wider form, adding binary variables for each of factor levels in columns. I leave out grouped categories as base levels, and well defined factors as predictors which should compare to the base category.

0.7 Model building.

Firstly, let's subset our data into the train and test sets in the 75/25 ratio. I am training models and then using them to predict prices in the test set. While building models I tried to include all the features which have been revealed in the data exploration part. The basis for predicting the price of a Californian wine is comprised of its region within the state, its sub-region (for a couple of regions), variety, year, taster's name, 1000 most frequent words used by the tasters for various wines, rating category (base-Acceptable), number words and characters in the description, sentiment score for the description, whether the wine is oaked, reserved and vintage.

Two of these models are linear regressions, one only including all the predictors, and the other containing additional interactive terms and quadratic form of year. Furthermore, I used the regularised regressions, with 4-fold cross validation to penalize for large number of variables.

```
## glmnet
##
## 22992 samples
##  1037 predictor
##
## No pre-processing
## Resampling: Cross-Validated (4 fold)
## Summary of sample sizes: 17245, 17244, 17244, 17243
## Resampling results across tuning parameters:
##
##   lambda       RMSE       Rsquared   MAE
##   0.001000000  0.3762670  0.5722593  0.2968221
##   0.001438450  0.3762670  0.5722593  0.2968221
##   0.002069138  0.3762670  0.5722593  0.2968221
##   0.002976351  0.3762670  0.5722593  0.2968221
##   0.004281332  0.3762670  0.5722593  0.2968221
##   0.006158482  0.3762670  0.5722593  0.2968221
##   0.008858668  0.3762670  0.5722593  0.2968221
##   0.012742750  0.3762670  0.5722593  0.2968221
##   0.018329807  0.3762670  0.5722593  0.2968221
##   0.026366509  0.3762597  0.5722545  0.2968213
##   0.037926902  0.3762550  0.5721314  0.2968615
##   0.054555948  0.3763445  0.5718509  0.2969975
##   0.078475997  0.3765875  0.5713496  0.2973007
##   0.112883789  0.3770839  0.5705178  0.2978397
##   0.162377674  0.3779633  0.5692460  0.2987311
##   0.233572147  0.3794221  0.5673798  0.3001225
##   0.335981829  0.3817314  0.5647290  0.3022572
##   0.483293024  0.3852398  0.5611219  0.3054651
##   0.695192796  0.3903637  0.5564452  0.3100932
##   1.000000000  0.3975633  0.5506745  0.3165519
##
```

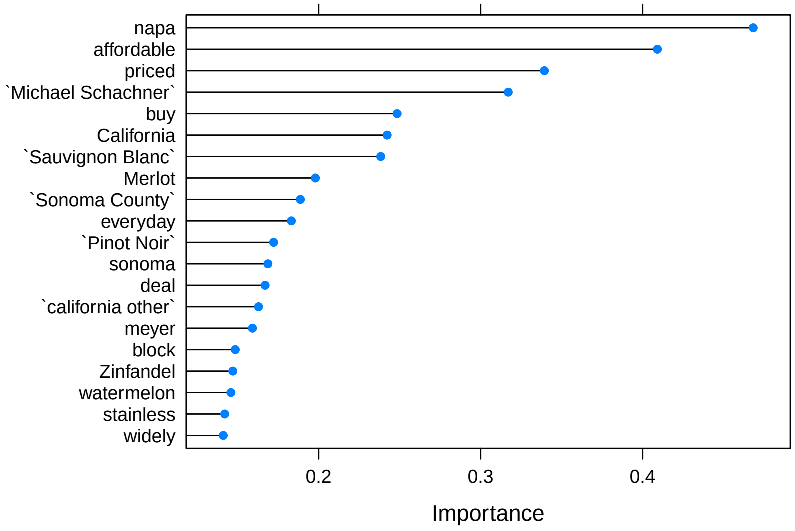
```
## Tuning parameter 'alpha' was held constant at a value of 0
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 0 and lambda = 0.0379269.

## glmnet
##
## 22992 samples
##  1037 predictor
##
## No pre-processing
## Resampling: Cross-Validated (4 fold)
## Summary of sample sizes: 17244, 17244, 17244, 17244
## Resampling results across tuning parameters:
##
##   lambda        RMSE       Rsquared   MAE
##   0.0001000000  0.3762270  0.5730432  0.2965896
##   0.0001373824  0.3760361  0.5734182  0.2964349
##   0.0001887392  0.3757905  0.5739005  0.2962339
##   0.0002592944  0.3754838  0.5745031  0.2959797
##   0.0003562248  0.3751208  0.5752156  0.2956811
##   0.0004893901  0.3747032  0.5760352  0.2953320
##   0.0006723358  0.3742323  0.5769658  0.2949482
##   0.0009236709  0.3737636  0.5778984  0.2945878
##   0.0012689610  0.3733365  0.5787738  0.2942990
##   0.0017433288  0.3729489  0.5796343  0.2940417
##   0.0023950266  0.3728418  0.5800002  0.2940557
##   0.0032903446  0.3733290  0.5792255  0.2945801
##   0.0045203537  0.3747424  0.5766153  0.2958592
##   0.0062101694  0.3773922  0.5714933  0.2982794
##   0.0085316785  0.3815615  0.5632407  0.3020175
##   0.0117210230  0.3873534  0.5515975  0.3070616
##   0.0161026203  0.3944921  0.5373918  0.3132301
##   0.0221221629  0.4030896  0.5206307  0.3209357
##   0.0303919538  0.4139062  0.4996788  0.3307735
##   0.0417531894  0.4277398  0.4722024  0.3430755
##   0.0573615251  0.4418078  0.4495010  0.3561089
##   0.0788046282  0.4607193  0.4191362  0.3734845
##   0.1082636734  0.4882702  0.3598219  0.3976992
##   0.1487352107  0.5137749  0.3374912  0.4177176
##   0.2043359718  0.5506807  0.2769656  0.4459418
##   0.2807216204  0.5751888        NaN  0.4651009
##   0.3856620421  0.5751888        NaN  0.4651009
##   0.5298316906  0.5751888        NaN  0.4651009
##   0.7278953844  0.5751888        NaN  0.4651009
##   1.0000000000  0.5751888        NaN  0.4651009
##
## Tuning parameter 'alpha' was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 1 and lambda = 0.002395027.
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Wed, Sep 28, 2022 - 1:04:36 PM

Table 1: RMSE

	model1	model2	lasso_cv	ridge_cv
1	0.358	0.357	0.363	0.360
2	0.377	0.376	0.375	0.377

In the table we can observe how each model performed in practice. The first row presents the RMSE statistics obtained from training data, and the second is for test data. As it could have been expected, the regularized LASSO model performed better than others, most probably due to the fact that we included more than a thousand of variables and those are mostly correlated and hence-penalized.



0.8 Conclusion.

To sum up, the wine reviews dataset proves to be rich in data exploration, which is fantastic for getting features out of it, yet the models we can build on it are only one-sided. The price prediction model does not take into account either costs to winemaker, or the fact that the price of wine depends also on the price of the land, which in turn depends of the previous years' prices of wines. Nonetheless, the model is good to predict and reveals that, for instance, if *Michael Schachner* tasted the *Pinot* from *Sonoma County*, and in his verdict used the word *priced*, it would cost almost 100% more than another *acceptable* wine from *Other region*, and *Other variety*. Ultimately, these aspects can be applied to conduct relative valuation of wines with respect to base levels.