Individual Project

Fall 2019, by Regan Chan (ttchan2)

Due: Monday, Nov 25 by 11:59 PM Pacific Time

Contents

Project description and summary

Wine score analysis and prediction

Given a large amount of wine tasting reviews, is it possible to build a statistical model to estimate an expert review score of a bottle given its source winery, country and other parameters?

The problem is not straight forward however. While we are provided with some ready to use parameters like price, region, etc. which are great for running regression on, we cannot ignore that the description field may provide the most significant info, yet it's not cleansed or analyzed for us.

In order to extract some useful information out from the description field, without going through the complexities of NLP, I will use the `quanteda` library to extract words as tokens, then transform each description into a word vector, merge that with the other features provided and feed the resulting vectors to regression algorithms

Data processing

- There are numerical values such as price and points. They are readily fed to the regression algorithms

Non-numeric (categorical) values

- There are values like region, winery, etc. Which are more appropriately stored as factors. In order to run regression on them, they are converted to a model matrix in advance
- Except for "Napa-Sonoma", all region_1s are in a many-to-one cardinality to region_2. Hence region_2 is less precise/descriptive than region_1 so I can be safely discard it; however, region_1 values are often missing from the data, reducing its usefulness.
- Winery name could not alone identify the winery since it could be in different countries. But winery+country can uniquely identify a winery

- NA values are either removed or replaced with empty strings because regression algorithms do not work well with NAs.
- There seems to be a weak correlation between winery, variety and designation. I could not safely remove them without sacrificing accuracy.

Textual values

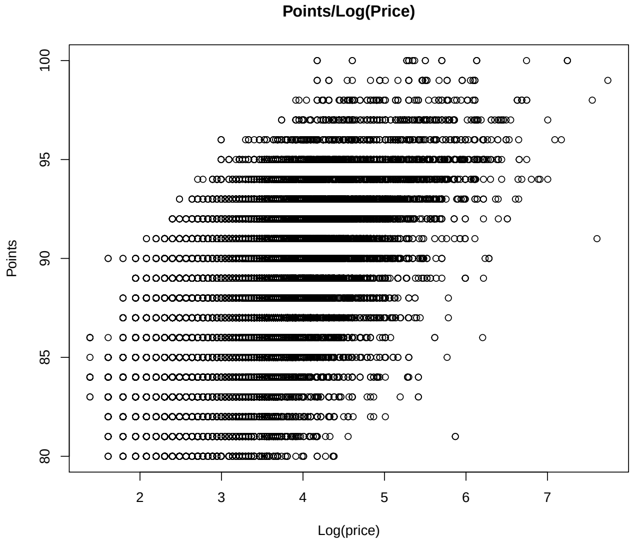
- Quanteda provides a complete suite to extract tokens and documents from input data
- Additional filtering like striping punctuation marks, deleting low frequency terms, and deleting stop-words are also performed

Duplicate rows

- There are many duplicate rows in the data. They should however not change the outcome of the regression

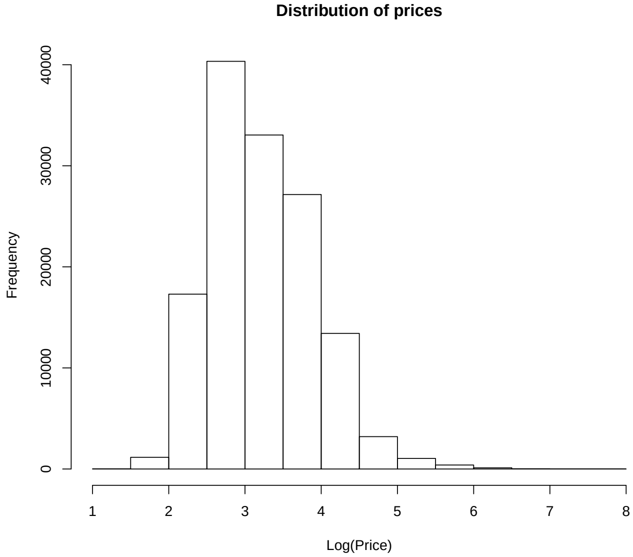
Descriptive statistics

Here I am hoping that there are some obvious correlations in the data that can help with regression

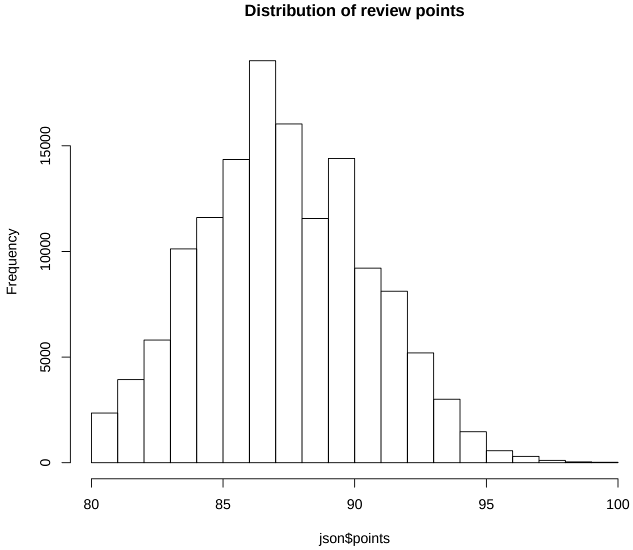


The first thing that came to mind is, does price dictate the quality of the wine? While it is true from this plot that higher price does roughly translate to higher scores, most of the wine is centered around a low price point and they still get very good scores.

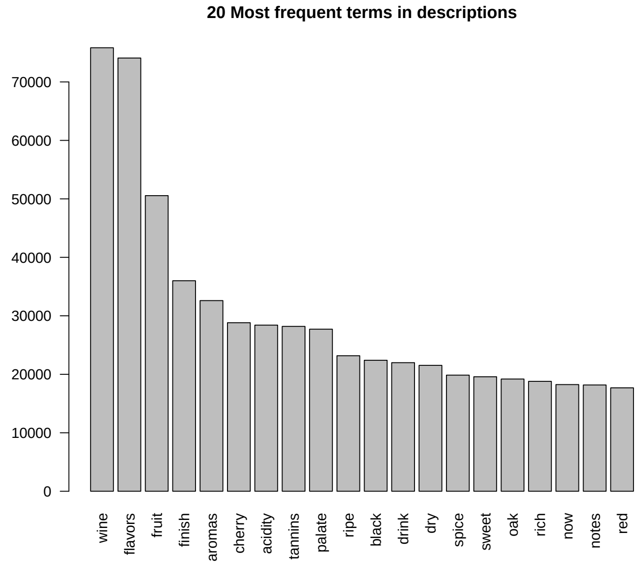
Another observation is that, a linear regression on this relationship should be very doable



Wines can be cheap or the sky is the limit! A significant portion of wines are at very low price point, though. Having taken a logarithm on the price may help model it better. The logarithm of the prices seem to be normally distributed



The points appears to be normally distributed around 87. The wine reviewers tend to give pretty good scores to all wines. Or maybe wine is wine, they all taste the same. Because of this, a gaussian model should give a good estimate.

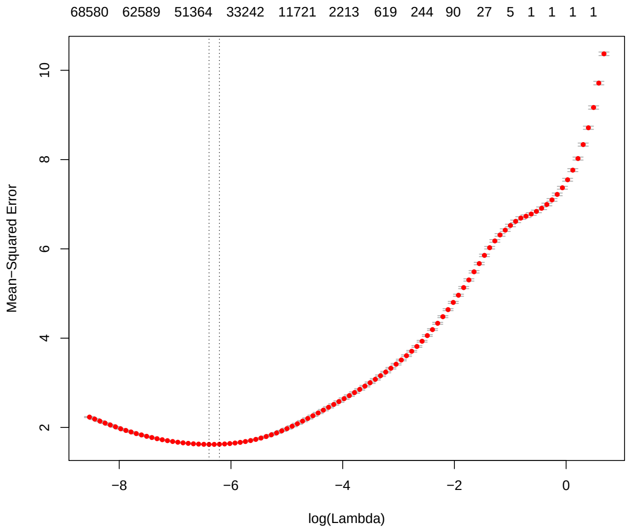


The most frequently occurring terms like "wine", "flavours" seems to be very generic and hence meaningless in our context. I filtered them away with max_termfreq.

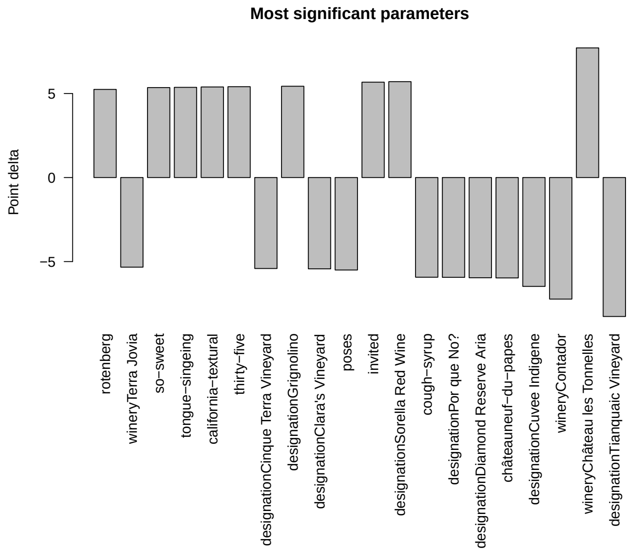
Regression model analysis

Generalized linear model with Lasso penalty

The generated training matrix has 87953 features. Lasso regularization with CV will reduce that number to something more manageable.



This 1se model will consider a staggering 44014 words/parameters, roughly half of the total 87953 parameters!



The winery “Château les Tonnelles” produces great wines, with a score 7.7 above average. Similarly, A designation of “Sorella Red Wine” or a description containing “tongue-singeing” also raises the score by 5.7 and 5.36 respectively. On the contrary, a designation of “Tianquaic Vineyard” lowers the score by 8; winery “Contador” and “Terra Jovia” are pretty bad, they got minus average scores of -7 and -5 respectively. Having “cough-syrup” in the description also is a very bad sign, subtracting -6 from the score. A designation of “Colheita White” also mean a -15 score.

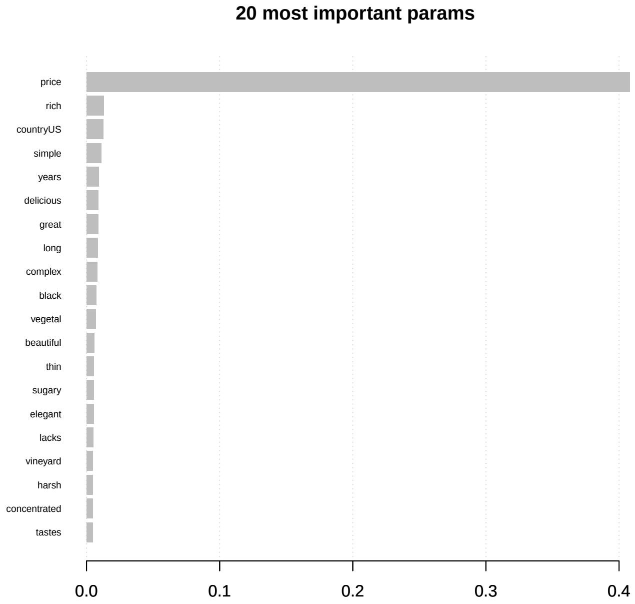
Model accuracy:

```
## [1] "Training RMSE:  0.821028322481652"
```

```
## [1] "Test RMSE:  1.25277753308355"
```

Random forest model with xgboost

A decision tree based model will give more insights as to what determines the wine scores.



Now the random forest has spoken. Price most clearly declares winners and losers. That’s not a surprise as we saw in a previous plot: While a good wine isn’t necessary expensive, expensive wines are definitely good!

Other than that, whether or not the word "rich" was mentioned in the description also matters. Obviously, whether it came from the US also was a major decision point.

Model accuracy:

```
## [1] "Training RMSE:  0.979274473270963"
```

```
## [1] "Test RMSE:  1.34622036650338"
```

The random forest model does slightly worse than the glmnet model but still rougly equivalent. Surprisingly, their decisions are based on totally different parameters!

Recommend wineries for fruity pinot noir

Since we know in advance the customer is interested in just one variety of wine (Pinot Noir), we can disregard any wineries that: 1) do not sell this type of wine, or 2) their price does not fit customer's budget

Now let's use our regression models to predict the scores of a "Pinot Noir" with "Fruity taste" in the description among those wineries with matches:

winery	glmnet	tree
Clos de Tart	91.44983	90.32743
Domaine Bruno Clair	90.35005	90.32743
Domaine Perrot-Minot	89.79542	90.67559
Domaine Méo-Camuzet	89.30035	90.32743
Domaine Jean Grivot	89.22133	90.32743
Semper	89.33144	89.45432
Domaine Henri Rebourseau	88.69450	89.92066
Barden	89.68834	88.70171
Lioco	90.26898	87.99841
Louis Jadot	88.26436	89.83159

According to our prediction models, the top choices for a "fruity taste" Pinot Noir are in the above table, along with their expected scores.

If we look at the underlying data, however, Shingle Peak, Stadlmann, Tangley Oaks, Alta Maria and Fiddlehead should have been our pick for the best wineries. We can get that if we don't replace the description in our test model.

However, if we assume that the customer is mostly interested in a fruity taste, basically ignoring all the other text in the description, then we get Clos de Tart, Domaine Bruno Clair, Domaine Perrot-Minot, Domaine Méo-Camuzet and Domaine Jean Grivot as our top picks for wineries.