Wine!

How Quantitative and Qualitative Variables Differ in Predicting Wine Quality.

About the Datasets

Wine Quality Dataset:

contains about 6500 rows of data comparing the chemical breakdown of red and white wines with their rating.

Wine Reviews Dataset:
Contains 130k different wines describing their characteristics such as points, brand, reviewer, country of origin, etc. This data was gathered from a WineEnthusiest website.

Quality vs. Points

Wine Quality

- Median of at least 3 evaluations made by wine experts. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Points

Wine Spectator tasters review wines on the following 100-point scale:

- **95-100** Classic: a great wine
- **90-94** Outstanding: a wine of superior character and style
- **85-89** Very good: a wine with special qualities
- **80-84** Good: a solid, well-made wine
- **75-79** Mediocre: a drinkable wine that may have minor flaws
- **50-74** Not recommended

Thoughts Behind This Analysis

- **Goal:** Is there bias in the Wine Reviews data?
  - Chemical makeup and variety are both indicators to how the wine is going to taste

- Wine Quality data doesn't disclose brand information
  - More scientific in collecting data

- Wine Reviews is actively trying to sell to customers
  - The reviewers are named

Wine Reviews

Building a model

Wine Reviews Dataset

Correlation between variable and points:

description_length	0.5282095837360421
price	0.45111736218256504
taster_name_Michael Schachner	-0.24683378570475292
country_Chile	-0.2018660991979389
title_length	0.16559143968374196

Created a separate dataframe only containing variables where the correlation > .1 and < -.1.

['price', 'title_length', 'description_length',
'country_Argentina', 'country_Chile',
'designation_none',
'province_Mendoza Province',
'region_1_California', 'region_1_Mendoza',
'region_1_None', 'region_2_California Other',
'taster_name_Anne Krebiehl\xa0MW',
'taster_name_Matt Kettmann',
'taster_name_Michael Schachner',
'variety_Pinot Noir',
'variety_Sauvignon Blanc']

Wine Reviews Dataset Methods:	Results Using All of the columns	Results Using Correlated Columns
SVM	Accuracy: < .2 for all c --------- Evaluating model: C= =10------ Accuracy: 0.19542572463768115 Avg. F1 (Micro): 0.19542572463768113 Avg. F1 (Macro): 0.07555018046321006 Avg. F1 (Weighted): 0.26239090996739955	=~.2 For all c --------- Evaluating model: C= =6.0---- Accuracy: 0.21460436722235265 Avg. F1 (Micro): 0.21460436722235265 Avg. F1 (Macro): 0.12962588206615633 Avg. F1 (Weighted): 0.22544413943357228
Naive Bayes	Accuracy = ~0.186	Accuracy = ~0.137
Random Forest	Accuracy = ~0.28 n_estimators =50, max_depth = 300	Accuracy = ~0.243 n_estimators =80, max_depth = 16
Quadratic Regression (Degree =2)	R^2= -8166956925598678.0	R^2 = 0.5158

Wine Quality

Building a model

Wine Quality Methods:	Results
Random Forest	Accuracy = 0.65
KNN K = 3	Accuracy 0.548
Naive Bayes	Accuracy = 0.467
Linear Regression	R^2= 0.337

Random Forest

```
Evaluating model: n_estimators =50,
 max_depth = 12
Accuracy: 0.6548076923076923
Confusion Matric:
[[  0   0   1   1   0   0   0]
 [  0   2  28  11   0   0   0]
 [  0   0 243  91   2   0   0]
 [  0   0  84 340  25   0   0]
 [  0   0   2  85  90   0   0]
 [  0   0   0  15  13   6   0]
 [  0   0   0   0   1   0   0]]
```

```
Evaluating model: n_estimators =80,
 max_depth = 12
Accuracy: 0.6596153846153846
Confusion Matric:
[[  0   0   1   1   0   0   0]
 [  0   3  28  10   0   0   0]
 [  0   1 240  94   1   0   0]
 [  0   0  78 348  22   1   0]
 [  0   0   4  84  89   0   0]
 [  0   0   0  17  11   6   0]
 [  0   0   0   1   0   0   0]]
```

Knn

```
---------- EVALUATING MODEL: k = 3
Accuracy: 0.5480769230769231
Confusion Matric:
[[  0   0   1   1   0   0   0]
 [  1  11  22   7   0   0   0]
 [  3  13 213  99   8   0   0]
 [  3   9 132 254  48   3   0]
 [  0   1  22  65  83   6   0]
 [  1   1   0  12  11   9   0]
 [  0   0   0   1   0   0   0]]
```

Bayes

```
Accuracy: 0.4673076923076923
Confusion Matric:
[[  0   0   2   0   0   0   0]
 [  0   0  20  17   4   0   0]
 [  0   0 214  93  29   0   0]
 [  0   0 170 200  79   0   0]
 [  0   0  24  81  72   0   0]
 [  0   0   5  15  14   0   0]
 [  0   0   0   0   1   0   0]]
```

Conclusion

Which Model performed the best?

Dataset:	Wine Reviews	Wine Quality
Best Model on Training Data	Quadratic regression (degree = 2) With Degree= 2 Mean Squared Error, 4.3398024242813955 R^2 0.5158349218104092 Median Absolute Error 1.4536091386854082 Expected Variance Explained: 0.5161060460423037 Actual Predicted 107813 88 88.208834 2172 90 89.797836 65831 86 86.463853 115828 87 86.035168 102028 86 87.153470	Random Forest ------ Evaluating model: n_estimators =50, max_depth = 12 ------ Accuracy: 0.6413461538461539 Confusion Matric: [[0 0 1 1 0 0 0] [0 2 29 10 0 0 0] [0 0 243 91 2 0 0] [0 0 89 324 34 2 0] [0 0 5 80 92 0 0] [0 0 0 16 12 6 0] [0 0 0 0 1 0 0]]
Best Model on Testing Data	R^2 stayed around 0.51 Actual Predicted 124861 88 87.996376 12566 87 88.389977 122629 83 83.452636 129224 86 86.619013 92659 88 87.129701 Mean Squared Error, 4.28397870193224 R^2 0.5151282648791193 Median Absolute Error 1.3827326230444612 Expected Variance Explained: 0.5153940228563683	Accuracy stayed around 0.64 ---------- Evaluating Random Forest Model ---------- using 50 estimators and a max depth of 12 Accuracy: 0.6461538461538462 Confusion Matric: [[0 0 3 2 0 0 0] [0 3 20 17 2 0 0] [0 0 306 129 5 0 0] [0 0 113 424 29 0 0] [0 0 5 101 99 0 0] [0 0 0 22 11 8 0] [0 0 0 1 0 0 0]]