Wine score prediction

Stefan Natu

Objective

1) Train a model to predict quality score for wines from training data.
2) Design a wine with the highest possible score
3) Design a wine that would be super expensive but still score >95.

Dataset -- WineEnthusiast from zackthoutt

Data contains numerical as well as categorical variables which need to be dealt with separately.

Remove duplicates and Nan’s from description, points column to get final dataframe for analysis.

Data Imputation: we replace missing values from region_1 with corresponding provinces

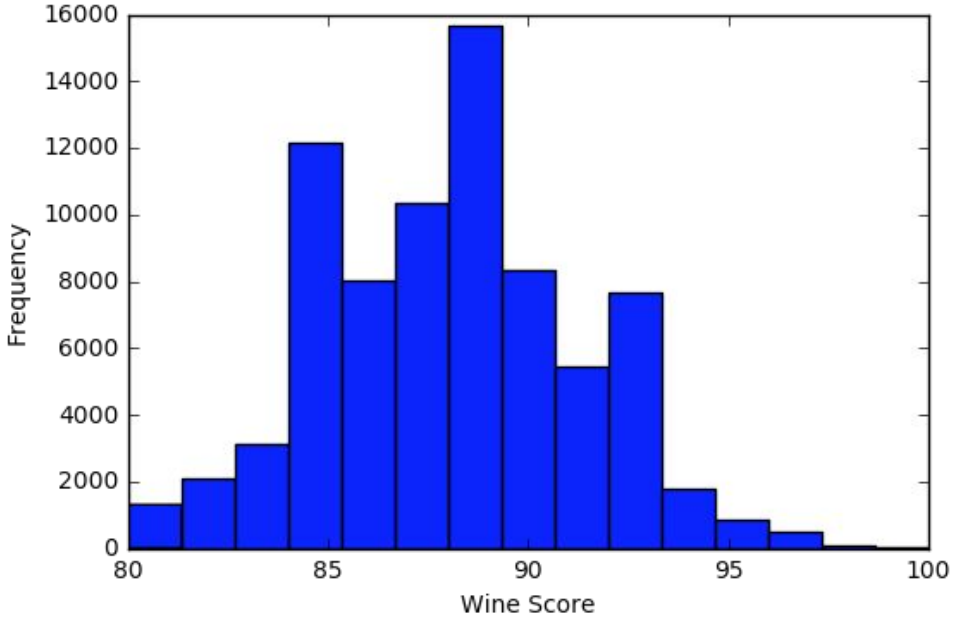
Data Frame

	country	description	designation	price	province	region_1	region_2	variety	winery
0	France	Smells leathery and a bit overripe, with scent...	NaN	45.0	Rhône Valley	Châteauneuf-du-Pape	NaN	Rhône-style Red Blend	E. Guigal
1	Chile	Earthy and staunch on the nose, with gentle bu...	Reserve	18.0	Colchagua Valley	NaN	NaN	Cabernet Sauvignon	El Huique
2	US	Soliloquy was a hit from the very start and th...	Soliloquy	25.0	California	Oakville	Napa	Sauvignon Blanc	Flora Springs
3	US	Drinks tighter and steelier than PR Chards of ...	Private Reserve	35.0	California	Napa Valley	Napa	Chardonnay	Beringer
4	US	Lush enough to drink now, with decanting, but ...	Quail Hill Vineyard	40.0	California	Russian River Valley	Sonoma	Chardonnay	Lynmar

Exploratory Data Analysis

Histogram of Points reveals points continueously scattered from 80-100

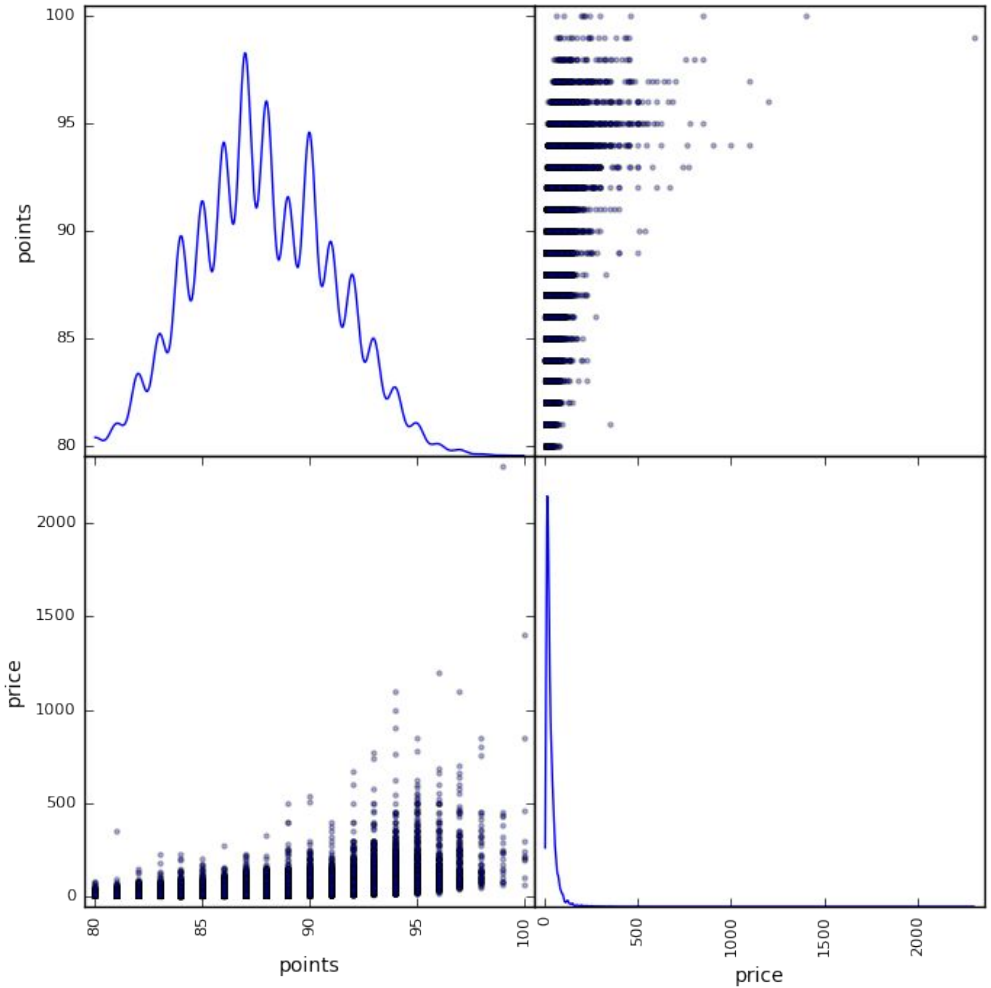
Use Regression model. Can also bin the scores and use a classifier (future work)



Exploratory Analysis

Plot the points versus price

Higher priced wines have a higher mean score but also higher variance



Categorical Variables - country, region, province, variety

After some EDA we find that -- good wines come from a lot of countries, and a lot of varieties.

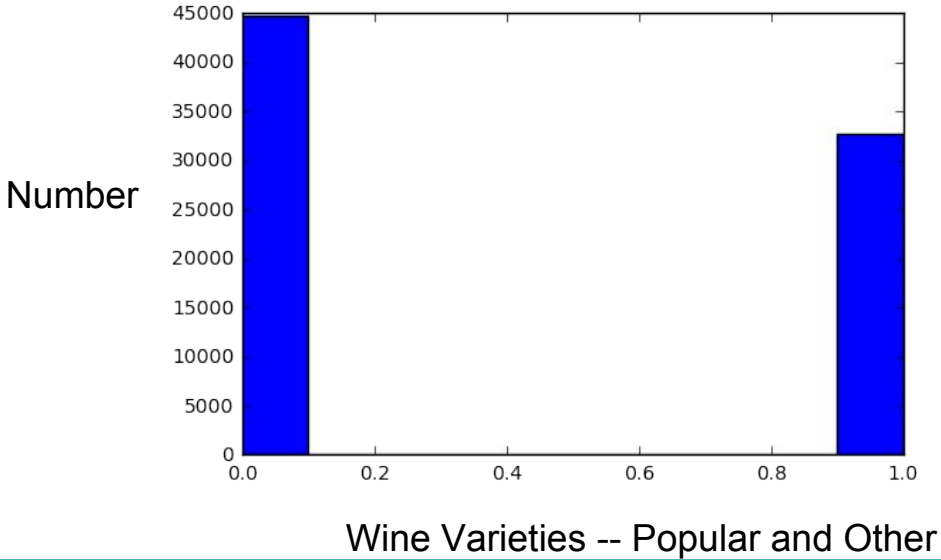
Since we don’t have enough data on all countries/wineries/varieties, we bin the categorical data into categories based on the amount of data.

For example: USA, France, Italy, Argentina, Chile and Australia have a lot of data, but other countries don’t -- so we put all other wines into a single category called “Other”.

Categorical Variables contd.

Break the varieties up into 2 labels -- popular (top 10) labeled 0 and other (labeled 1). (Choose 10 to minimize class imbalance)

Break the world up into 30 primary wine producing regions and provinces, putting the rest in Other.



Corr Plot comparing categorical variables to points

Correlation plot of categorical variables

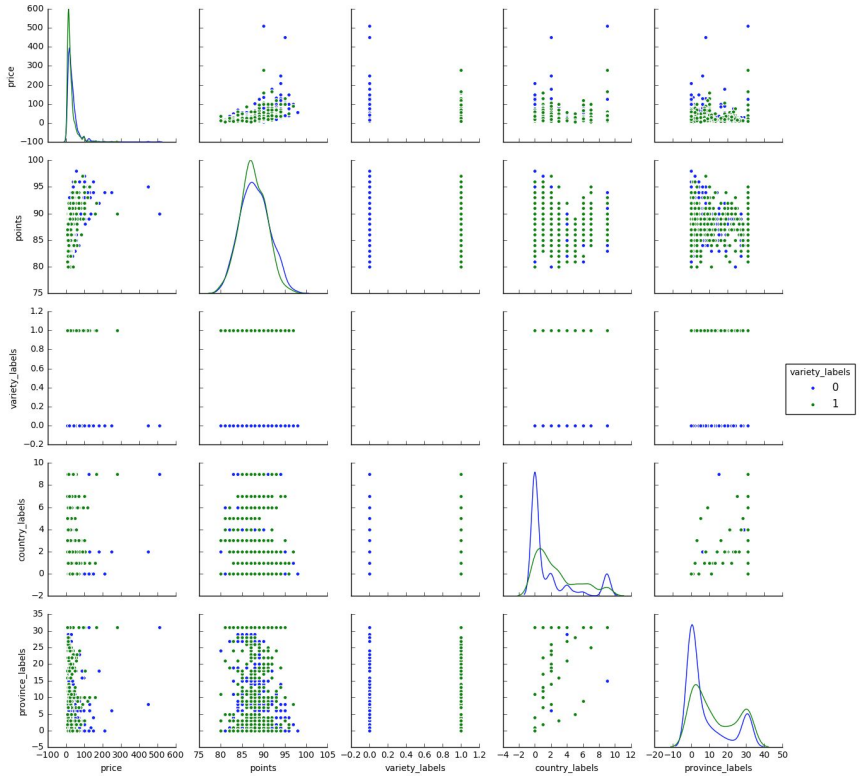
With points and prices color coded with

the 2 variety labels (popular and other)

High point wines come in both varieties.

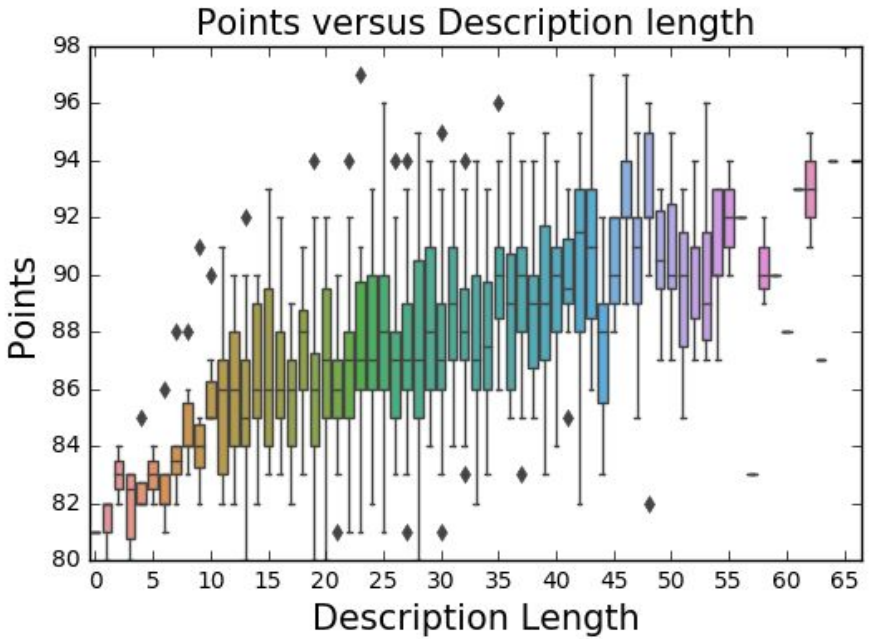
More popular wines tend to be

more expensive on average.



Compare the Description length to score

This is a new feature we add -- how long is the description



NLP: tfidf on descriptions! -- this is cool

To convert the descriptions to word vectors, we use a term-frequency/inverse document frequency model.

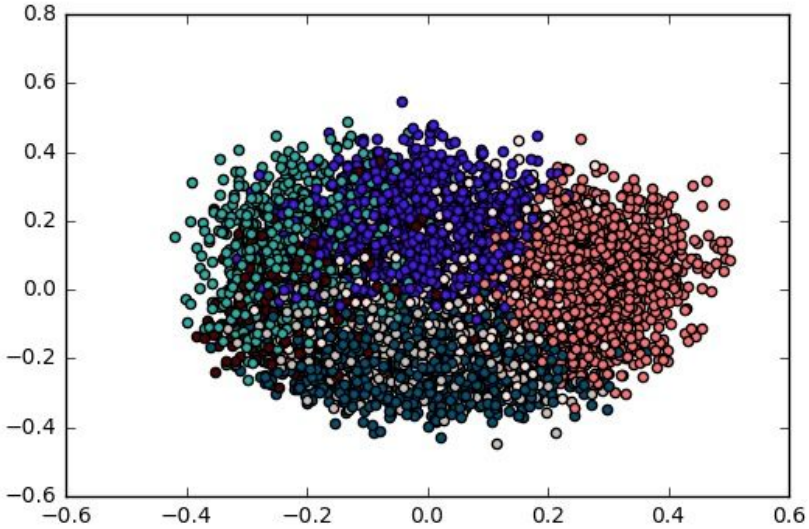
We extract a 100 word feature vectors. After some simple topic modeling with NMF, here are the categories that emerge. -- dry, spicy, fruity, sweet, blends, acidity, berries etc.

```
Topic 0:
wine fruit spice character vineyard wood great mouth shows mineral
Topic 1:
apple citrus fresh white crisp green acidity peach lemon pear
Topic 2:
palate finish aromas nose notes fruit plum long herbal medium
Topic 3:
sweet oak vanilla flavors like chardonnay little toast smoky raspberry
Topic 4:
dry flavors pinot good cherries cola tannic shows little cherry
Topic 5:
black cherry tannins blackberry dark chocolate pepper plum currant licorice
Topic 6:
cabernet blend sauvignon merlot syrah cassis tobacco style blackberry oak
Topic 7:
soft drink easy fruity flavors chocolate texture light cherries blackberry
Topic 8:
ripe rich acidity fruits drink tannins years age texture wood
Topic 9:
red berry cherry bright raspberry spice light plum fresh flavors
```

T-Sne/PCA to plot word features

We can use T-SNE or PCA to plot the word feature vectors in 2d space to get a sense for how clustered or distinct they are.

Plot shows that there are distinct categories that pop out.



Modeling -- Construct 3 models 2 sets of features

We construct 3 models -- simple Linear Regression, Random Forest Regression and Gradient Boosted Forest (XGBoost).

For each model we consider 2 sets of features -- one set including the NLP features and one without.

Perform Standard Scaling prior to fitting the model.

Split the training dataset into 80 training and 20 for holdout.

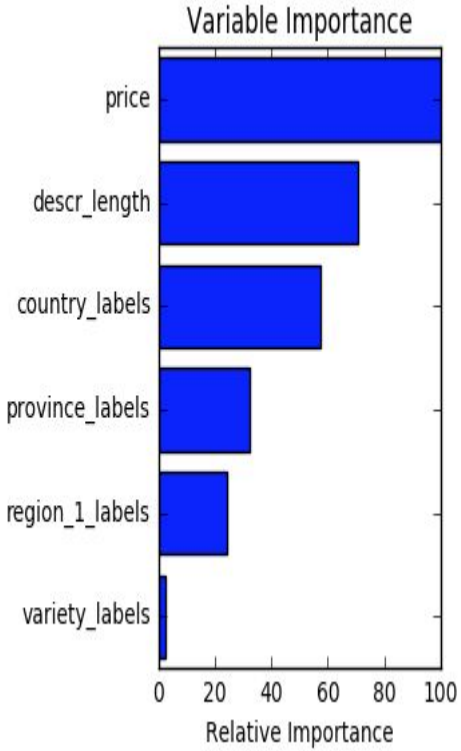
Mean Squared Error results without NLP

Linear Regression: 6.26

Random Forest: 5.4

Gradient Boosted Regression: 4.63

Top features ranked by importance for XGBoost. -- price and description length are Highest.



MSE with the NLP parameters

Linear Regression: 6.26

Random Forest: 5.2

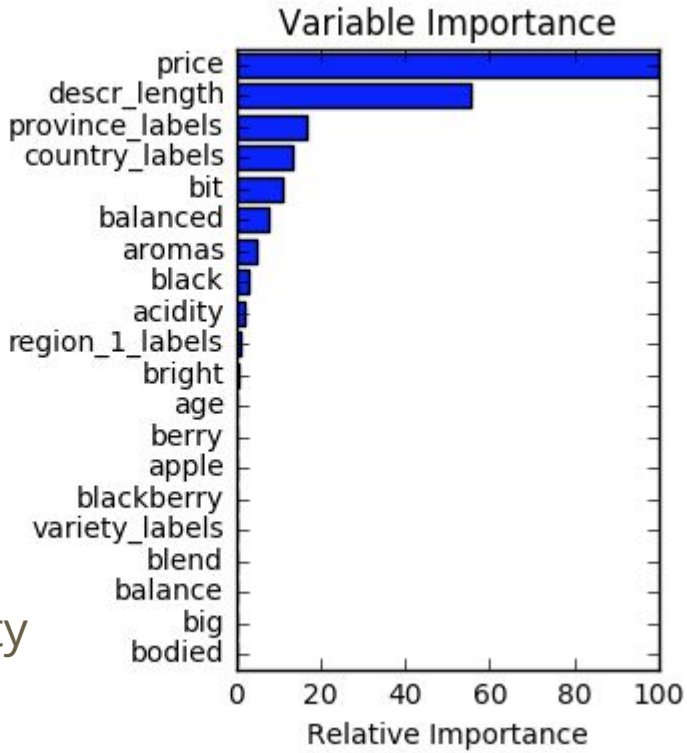
Gradient Boosted Regression: 4.23

Top features ranked by importance for

XGBoost. -- price and description length are

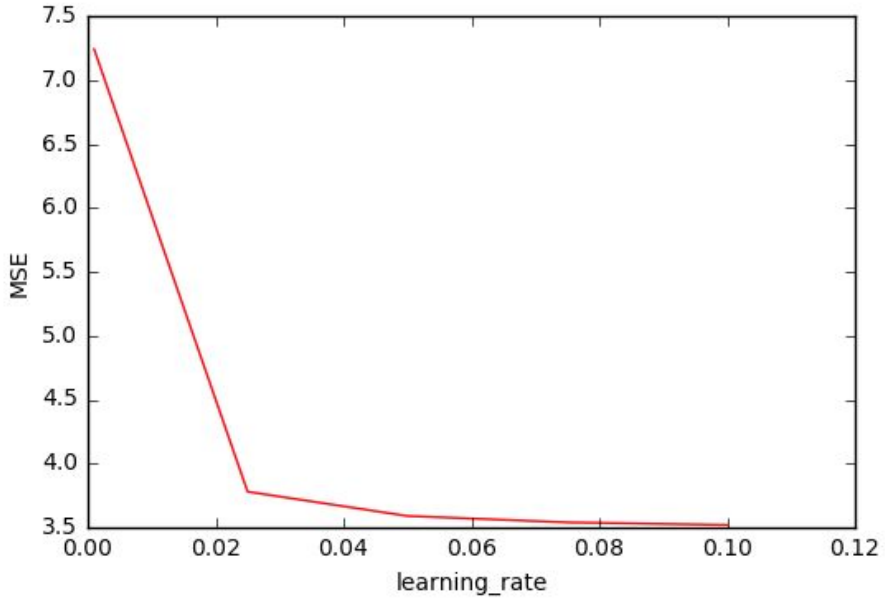
Highest, but bit, balance and aroma and acidity

Also matter!



With a little bit of hyperparameter tuning....

Tune the learning rate for XGBoost. Lowest MSE occurs at learning_rate = 0.1



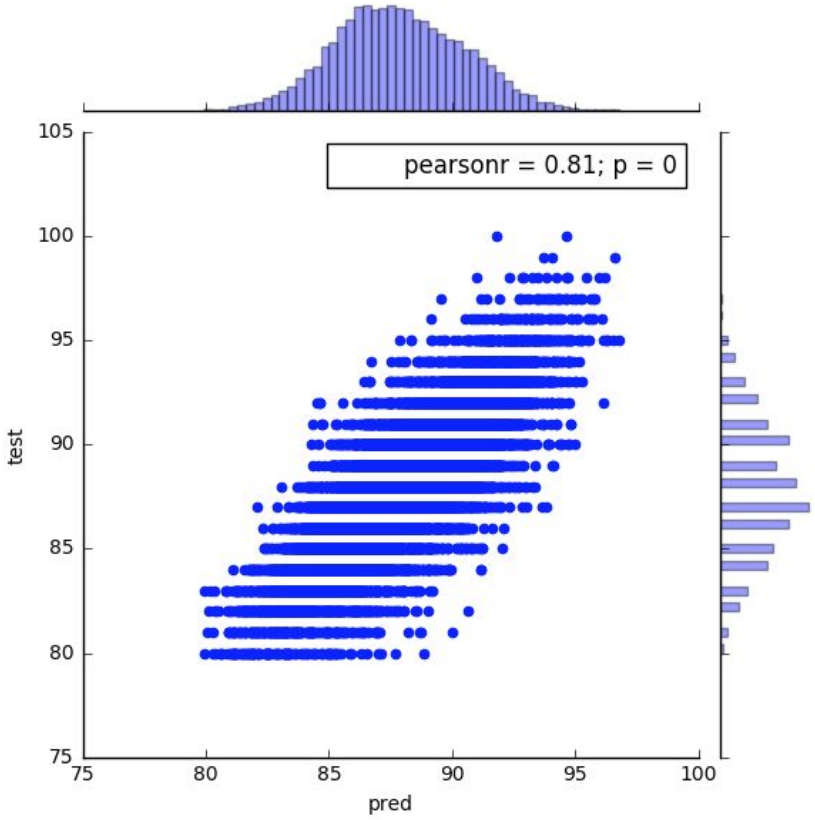
Validation prediction comparison

The model does pretty well on the

Validation set.

Ofcourse having more training data will increase the accuracy of the model.

One can also change the number of labels used for the categorical variables to play with the bias-variance tradeoff



Run the XGBoost model on the test data

Train the final XGBoost model on the entire training dataset and run it on the test data.

Training MSE = 3.05

Max wine score

Max Score: 97.3

Wine: pd.loc(pd['score'] == max_score)

	country	description	designation	price	province	region_1	variety	winery
24628	France	The purest Cabernet Sauvignon fruit, with dark...	NaN	1300.0	Bordeaux	Pauillac	Bordeaux-style Red Blend	Château Mouton Rothschild

Max Price yet score over 95

Use pd.sort_values([price, points], ascending = [False, False])

country	France
description	A wine that has created its own universe. It h...
designation	Clos du Mesnil
price	1400
province	Champagne
region_1	Champagne
variety	Chardonnay
winery	Krug
country_labels	2
province_labels	23
region_1_labels	23
variety_labels	0
descr_length	58
pred_score	97.0056