Wine Tasting with a Sip of Data

By: Brandon Jezek, Paula Koziol, Gautam Seenivasan & Jerome Bright

How Do You Pick a Wine?

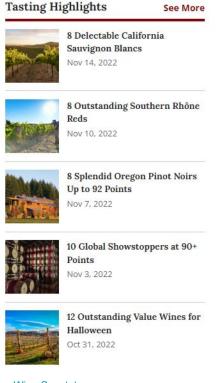## Sort through best sellers lists?

amazon prime
Best Sellers in Wine

Not a Wine!

#1

LOL Surprise Glitter Color Change Pearl Surprise (Purple) with 6 Surprises- Exclusive Collectible Doll & Lil Sister in Interactive…
★★★★★ 28
$24.99

Look for a pretty bottle? Or something on sale?

Juggernaut
Hillside Cabernet Sauvignon 2020

United States · California · Juggernaut · Red wine · Cabernet Sauvignon

4.1 ★★★★
617 ratings

☐ Add to Wishlist

ⓘ Latest vintage available

Look for Featured Wines?

Tasting Highlights   See More

8 Delectable California Sauvignon Blancs
Nov 14, 2022

8 Outstanding Southern Rhône Reds
Nov 10, 2022

8 Splendid Oregon Pinot Noirs Up to 92 Points
Nov 7, 2022

10 Global Showstoppers at 90+ Points
Nov 3, 2022

12 Outstanding Value Wines for Halloween
Oct 31, 2022

Wine Spectator

Compare Reviews?

"Ample aromas of blackberry, plum, cherry, licorice, **smoke** & vanilla. Palate shows flavors of ripe fruits with some earth notes of pepper & **cocoa** showing throughout. Finish is long with chocolate & **tobacco** spice…"

"**Mulberry**, spiced plum, cassis, cinnamon, clove, anise, **leather**, cream, oak, tobacco, vanilla."

"**Medium** plus body. Cherry leather vanilla on the nose. Dark fruit black fruit black cherry on the taste. Pretty balance acidity and tannin. **Pepper** finish"

Vivino

Welcome to Wine-ML!

Photo by Kym Ellis on Unsplash

What qualities are most important for you?

Streamlit

Welcome to Wine-ML!

What kind of wines do you like?

Use the input box to describe what kind of wine you're looking for - taste, aroma, pairing, etc and we'll find the best match based on our inventory.

Enter any text information that describes the kind of wine you're looking for, and we'll show you our best match!

> dessert riesling

Based on your preferences, we recommend this wine!

Title: BEDELL 2001 LATE HARVEST RIESLING RIESLING (NORTH FORK OF LONG ISLAND)

Variety: RIESLING

Designation: LATE HARVEST RIESLING

Country of origin: US

Province: NEW YORK
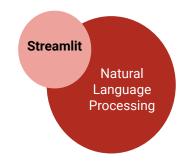
Region: NORTH FORK OF LONG ISLAND, LONG ISLAND

Winery: BEDELL

Taster name: JOE CZERWINSKI

Taster twitter handle: @JOECZ

Description: IT'S PICKED LATE, WHEN THE GRAPES ARE SUPERRIPE, THEN WINEMAKER KIP BEDELL PUTS THEM IN A FREEZER PRIOR TO CRUSH TO FURTHER CONCENTRATE THE RESULTING MUST. THE RESULT I A THICK, GOOEY, UNCTUOUS DESSERT WINE WITH FLAVORS OF TANGERINES AND HONEY. YUM.

Streamlit

Natural Language Processing

Welcome to Wine-ML!

Wine Quality Calculator

Use our models to help you pick great quality wines. Our models indicate that alcohol content, volatile acidity, and sulphates are the most important variables when it comes to dictating the quality of a wine.

Don't feel the need to stick to wines that have been reviewed when you can use our models to help you gauge its quality without having to taste it.

Enter as much information that you have about the wine below, and use our models to help you predict its quality!

Fixed acidity

8.32	−	+

Volatile acidity

0.53	−	+

Citric acid

0.27	−	+

Residual sugar

2.54	−	+

Chlorides

0.09	−	+

. . .
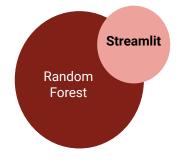
The parameters input correspond to a wine with LOW quality (rating of 6 or lower)

Random Forest

Streamlit

Red Wine Important Quality

White wine Importance Matrix

White wine highest correlation is Quality & Alcohol again! With .44

White Wine
important Quality

```python
random_forest = pd.DataFrame(zip(X_train.columns, rf_model.feature_importances_),columns = ["feature", "importance"])
random_forest.set_index("feature", inplace=True)
random_forest.sort_values(by = "importance", ascending = False, inplace = True)
random_forest.head()
```

Out[16]:

feature	importance
alcohol	0.173620
volatile acidity	0.120091
sulphates	0.106653
density	0.097363
total sulfur dioxide	0.090359

In [17]:
```python
#the top 3 important properties are alcohol, volatile acidity, and sulphates
```

In [18]:
```python
random_forest.plot.barh(figsize=(10,10))
plt.show()
```

WineOnTap/WinePrediction.ipynb ×   +

← C   https://github.com/bjezek/WineOnTap/blob/main/Brandon/WinePrediction.ipynb

Suggested Sites   Property Detail   Minority Workers...   Black Panther: The...   An End-to-End Gui...   SOLUTION: How m...   Downloadable Vide...   New Tab   Dashboard   Other favorites

On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [20]:
```python
y_pred_goodwine = rf_model.predict(X_test)
acc_rf = accuracy_score(Y_test,y_pred_goodwine)
print('Accuracy = ', acc_rf)
```

89% Accuracy in predicting rating of red-wine

Accuracy = 0.8895833333333333

In [21]:
```python
random_forest_report = classification_report(Y_test,y_pred_goodwine)
print(random_forest_report)
```

```
              precision    recall  f1-score   support

         no       0.93      0.95      0.94       417
        yes       0.60      0.49      0.54        63

   accuracy                           0.89       480
  macro avg       0.76      0.72      0.74       480
weighted avg       0.88      0.89      0.89       480
```

In [23]:
```python
#Classification report
```

```
[44]:  random_forest = pd.DataFrame(zip(X_train.columns, rf_model.feature_importances_),columns = ["feature", "importance"])
       random_forest.set_index("feature", inplace=True)
       random_forest.sort_values(by = "importance", ascending = False, inplace = True)
       random_forest.head()
```

[44]:

feature	importance
alcohol	0.155732
density	0.123412
residual sugar	0.092581
volatile acidity	0.087980
free sulfur dioxide	0.083631

```
[45]:  random_forest.plot.barh(figsize=(10,10))
       plt.show()
```

WineOnTap/WinePrediction.ipynb ×   +

← C   🔒 https://github.com/bjezek/WineOnTap/blob/main/Brandon/WinePrediction.ipynb   A⁰   ⭐

Q Suggested Sites   🗋 Property Detail   🗋   🗋   🗋 Minority Workers...   ✂ Black Panther: The...   🗋 An End-to-End Gui...   🗋 SOLUTION: How m...   📰 Downloadable Vide...   🗋 New Tab   🛡 Dashboard   >   📁 Other favorites

alcohol

```
     0.00      0.02      0.04      0.06      0.08      0.10      0.12      0.14      0.16
```

In [46]:

```python
y_pred_goodwine_1 = rf_model.predict(X_test)
acc_rf = accuracy_score(Y_test,y_pred_goodwine_1)
print('Accuracy = ', acc_rf)
```

87% Accuracy in predicting
rating of white-wine

Accuracy = 0.8727891156462585

In [47]:

```python
random_forest_report_1 = classification_report(Y_test,y_pred_goodwine_1)
print(random_forest_report_1)
```

	precision	recall	f1-score	support
no	0.89	0.95	0.92	1161
yes	0.76	0.58	0.66	309
accuracy			0.87	1470
macro avg	0.83	0.76	0.79	1470
weighted avg	0.87	0.87	0.87	1470

Dataframe Analysis

Top 10 Varieties

Top 10 Wineries
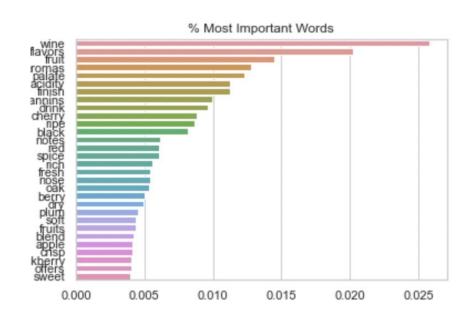
Best Bang for Your Buck

NLP

import seaborn as sns
import matplotlib.pyplot as plt
from nltk.tokenize.treebank import TreebankWordDetokenizer
from nltk.corpus import stopwords
from nltk import word_tokenize

% Most Important Words

Sentiment Analysis

import nltk

from nltk.sentiment.vader import **SentimentIntensityAnalyzer**

import nltk

nltk.download('vader_lexicon')

```python
[156]: df_sentiment = pd.read_csv('winemag-data-130k-v2.csv')

[ ]:   #Sentiment analysis of wine reviews

[157]: sid = SentimentIntensityAnalyzer()
       df_sentiment.reset_index(inplace=True, drop=True)
       df_sentiment[['neg', 'neu', 'pos', 'compound']] = df_sentiment['description'].apply(sid.polarity_scores).apply(pd.Series)
       df_sentiment

       df_sentiment["variety"] = df["variety"]

       df_sentiment.head()
```

designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery	neg	neu	pos	compound
Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia	0.000	0.935	0.065	0.1531
Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos	0.000	0.868	0.132	0.6486
NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm	0.053	0.947	0.000	-0.1280
Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest	Riesling	St. Julian	0.000	0.926	0.074	0.3400

Filtering out for High Sentiment

```python
filter5 = df_palate["Sentiment Score"] > .9

good_palate = df_palate.loc[filter5]

good_palate.dropna()
```

	description	Sentiment Score	Variety
119	Medium-gold in color. Complex and inviting nos...	0.9732	Riesling
128	Compelling minerality on the nose, Refined and...	0.9313	Pinot Blanc
133	Einaudi's wines have been improving lately, an...	0.9325	Nebbiolo
135	The color is just beginning to show signs of b...	0.9398	Nebbiolo
147	Black cherry, black plum and black currant are...	0.9386	Cabernet Sauvignon
...	...	...	...
129907	Toasty aroams of grilled apple, peach and hone...	0.9322	Chardonnay
129908	Pleasant scents of toasted apple and peach pit...	0.9528	Chardonnay
129917	This dark, meaty Malbec has some Agrelo fruit ...	0.9413	Malbec
129921	There is a select group of under-$20 Malbecs f...	0.9231	Malbec
129930	Arguably on the heavy and rich side, but if yo...	0.9387	Tempranillo Blend

Final Recommendation

```
good_palate["Variety"].describe()
```

```
count             3021
unique             190
top           Riesling
freq               363
Name: Variety, dtype: object
```

Streamlit Integration

Make a selection below to find the best wines associated with your preference:

| acidity | fruity | aromas | palate | tannins |

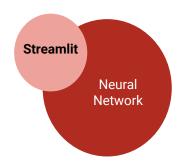Riesling

<u>Neural Network</u>

Neural networks are useful in predictive analysis because of the model's hidden layers.

The neural network uses the hidden layer to make predictions with higher precision because learns in similar way to humans. That's because it 'learns' the way a human does

Wine reviews are intended for consumers who want to make the best decisions providing them with the most descriptive information available but in many instances consumers gain more information on different wines and enable them to select wines to their liking more easily.

Using a Neural network model, the goal is to produce information that customers can use to make a good decision about wines before making their purchases.

Objective: Train Neural Network model predict to predict wine scores based on their reviews
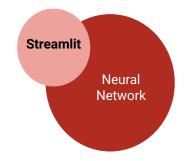
Neural Network

Streamlit

Welcome to Wine-ML!

Observe the overal data distribution

```
wine_data.describe()
```

|	Unnamed: 0	points	price
count	150930.000000	150930.000000	137235.000000
mean	75464.500000	87.888418	33.131482
std	43569.882402	3.222392	36.322536
min	0.000000	80.000000	4.000000
25%	37732.250000	86.000000	16.000000
50%	75464.500000	88.000000	24.000000
75%	113196.750000	90.000000	40.000000
max	150929.000000	100.000000	2300.000000

Streamlit

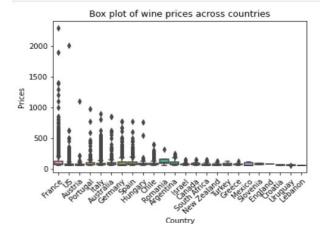Neural Network

Neural Network

Streamlit
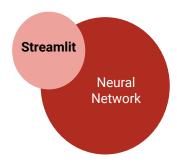
Prices vary by countries

```
]:   #Extract the most expensive wines
     priceVScountry = wine_data[["price", "country"]].dropna(how = "any").nlargest(20000, "price")


     ax = sns.boxplot(x="country", y = "price", data=priceVScountry)
     ax.set_title("Box plot of wine prices across countries")
     ax.set(xlabel = "Country", ylabel = "Prices")
     ax.set_xticklabels(ax.get_xticklabels(),rotation=45,ha="right",rotation_mode='anchor')

     plt.show()
```
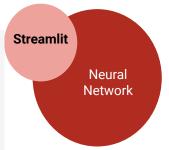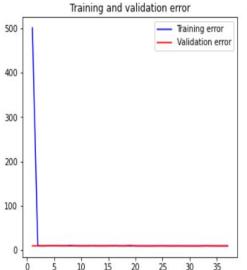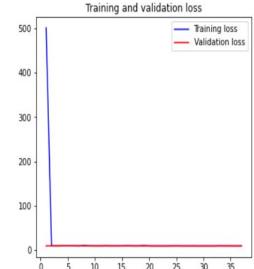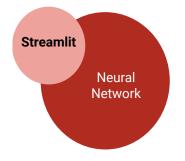

Box plot of wine prices across countries

Neural Network


```python
wordcloud = WordCloud(width = 800, height = 400,
                background_color ='white',
                stopwords = stopwords,
                min_font_size = 10).generate(all_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```

Neural Network

Expenses

Finalizing the development costs:
1) Analysis and raw estimation – Free

2) Prototype development – $5000

3) MVP development (prototype included) – $10000

4) Deployment and Release – $5000

How Much Does It Cost To Build A Recommendation System (azati.ai)

ROI and Pricing

- Charge a flat $10,000 a year fee to recommend wines
- Charge a maintenance fee of $250 a month
- Include a affiliate performance clause for a 3% royalty after a certain milestone is reached.

- ***50%+ ROI per client***

Next Steps

Our Deep Learning models can produce improved results with more information and stringent model tuning.

The benefits of these improved tests would allow us to explore how these models can assess consumer reviews and surveys. These studies could provide the customer with earlier consumer feedback as compared to the typical wine review of the professionals, like movie scores.

Additionally, we can apply these models to consumer tastes and expectations for vodka, tequila, or beer and supply similar benefits for customers as they try to gather the information they consider important for their next purchase. &

Run over & Undersampling check hedging parameters for accuracy results on a continued basis