Sommeliers and BERT:

Grape Variety Classification

Frank Novak
Capstone Project
DSI 1010

The Process

01 Problem

02 Collection

03 Exploration

04 Theory

05 Modeling

06 Conclusions

01

Problem

Text to Tool

Problem Statement

Can Sommelier reviews be used as predictive features for a multiclass classification model to distinguish specific grape varieties as a means to educate those new to wine and facilitate growth as a wine consumer?

02

Collection

Scraping winemag.com

Reviews

There's a mineral-laced start to the nose of this bottling, with hints of asphalt and graphite, followed by darker stewed fruit on the back. The palate is layered in creamy, lavish tannins, as flavors of cassis and caramel aim to impress through the clean acidity. ***—MATT KETTMANN***

RATING	**92** POINTS
PRICE	$45, BUY NOW
DESIGNATION	Reserve
VARIETY	Cabernet Sauvignon
APPELLATION	Paso Robles, Central Coast, California, US
WINERY	Smith & Hook

03

Exploration

Overall and Text

Exploration

Overview

- Price
- Points
- Countries
- Regions
- Wineries

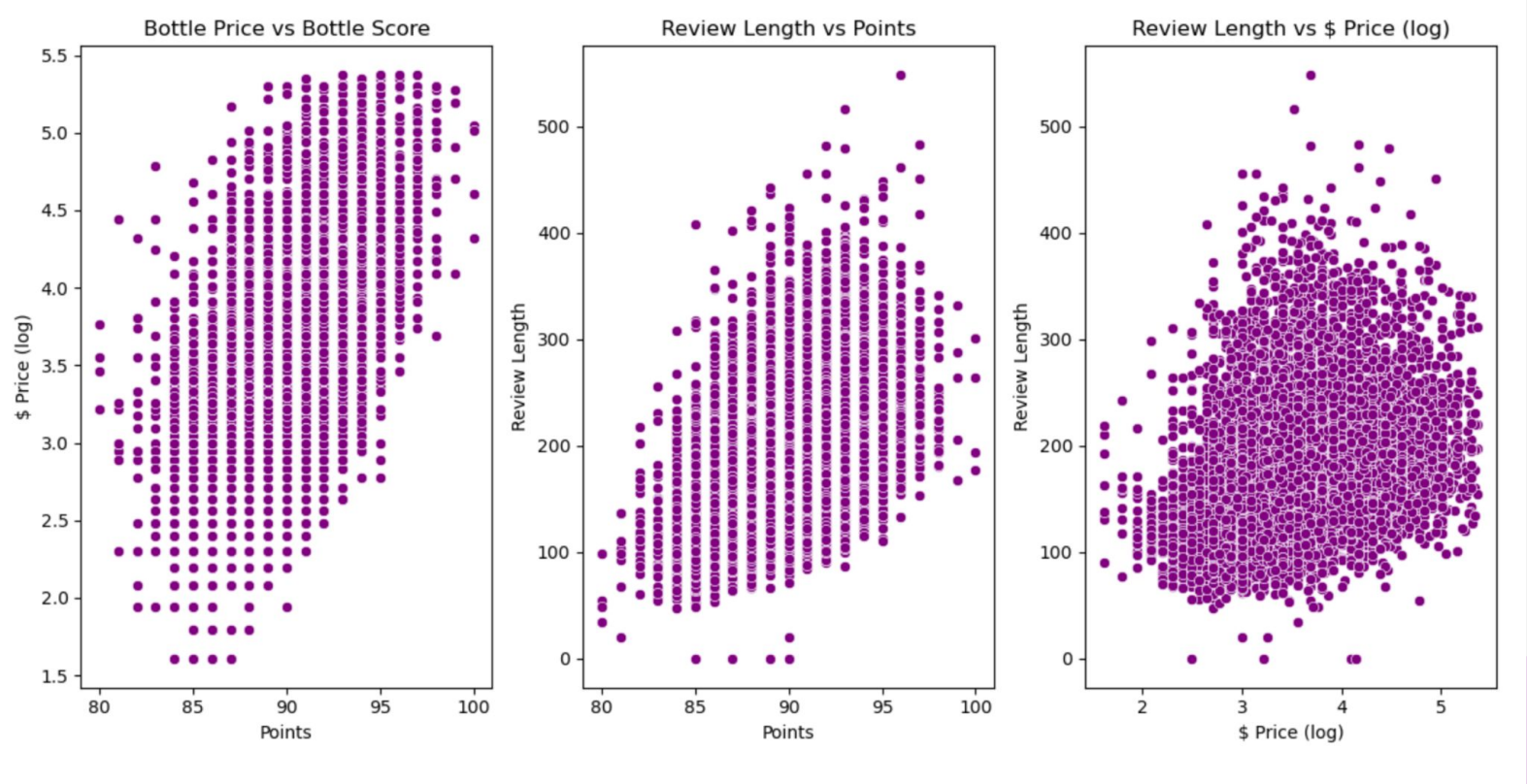
Focus

- Reviews
- Varieties

Overview

- Unique Sommeliers: 18
- Unique countries: 23
- Unique varieties: 328
- Unique wineries: 4496
- Unique bottles: 13782

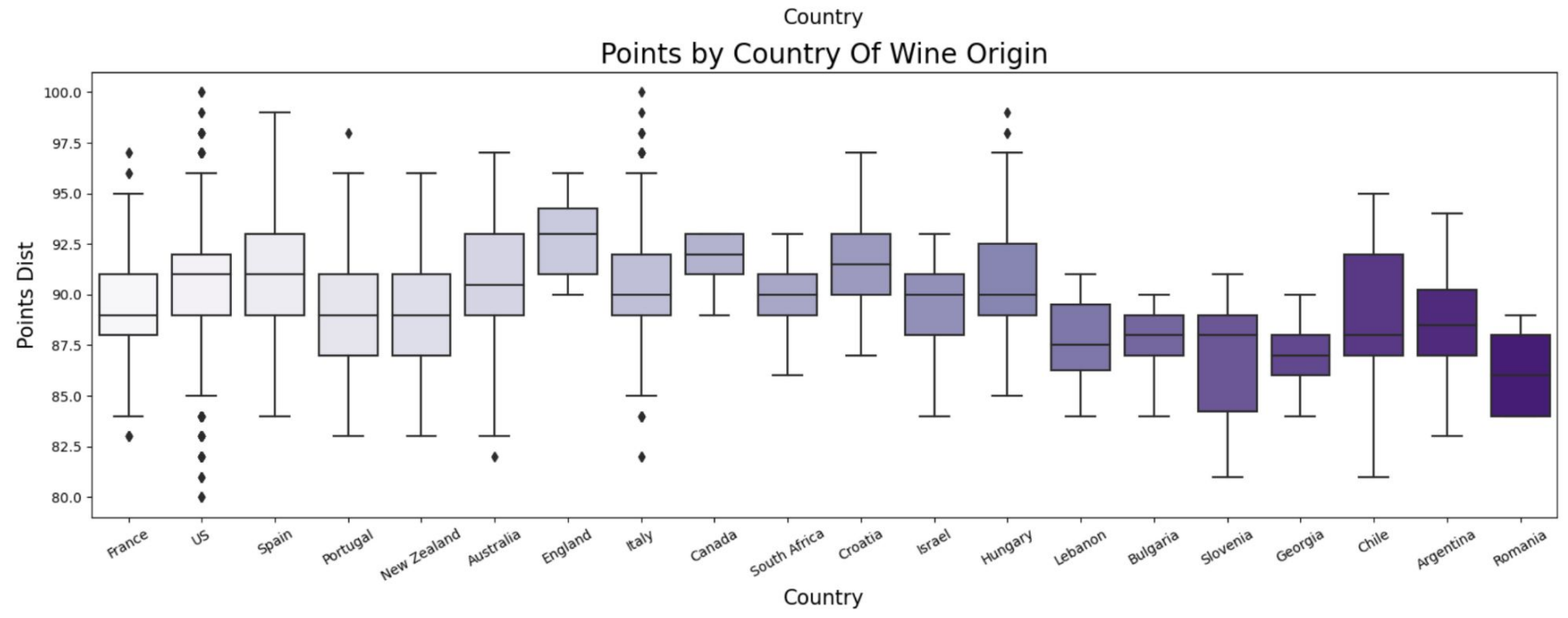
Overview


Bottle Price vs Bottle Score
$ Price (log)
Points
Review Length vs Points
Review Length
Points
Review Length vs $ Price (log)
Review Length
$ Price (log)

Overview

Points	Country	Bottle	Variety
95	Spain	Rolland & Galarreta 2014 Tempranillo (Rioja)	Tempranillo
95	Australia	Chambers Rosewood Vineyards NV Muscat (Rutherglen)	Muscat Blanc à Petits Grains
95	Italy	Paltrinieri 2020 Radice (Lambrusco di Sorbara)	Lambrusco di Sorbara
94	Australia	El Vinculo 2018 Crianza Tempranillo (La Mancha)	Tempranillo

Overview


Country
Points by Country Of Wine Origin
Points Dist
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
France
US
Spain
Portugal
New Zealand
Australia
England
Italy
Canada
South Africa
Croatia
Israel
Hungary
Lebanon
Bulgaria
Slovenia
Georgia
Chile
Argentina
Romania
Country

Overview

Points	Country	Bottle	Variety
92	Hungary	Dúzsi Tamás 2020 Kékfrankos Rosé (Szekszárd)	Rosé
91	Hungary	Royal Tokaji 2019 The Oddity Dry (Tokaji)	Furmint
91	Hungary	Gál Tibor 2020 Egri Csillag White (Eger)	White Blend

Focus

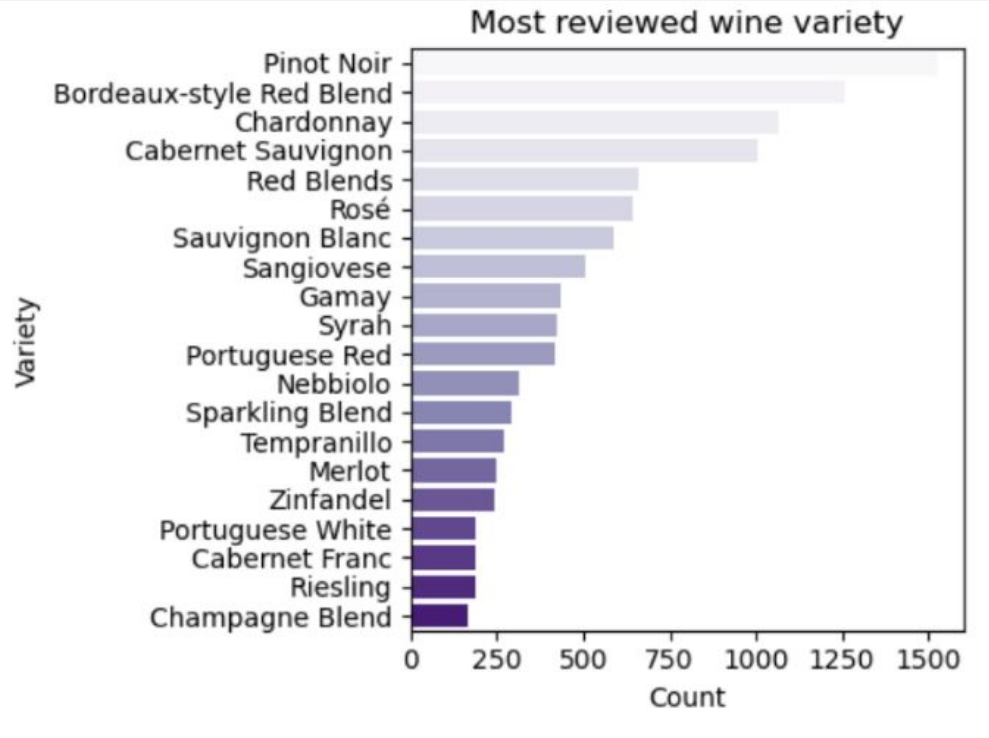
Varieties

- Unique wine varieties: 328
- Limit to those with most reviews for data
- Remove “grouped” varieties (ie: blends)

Text

- Is there information to suggest that wine varieties can be distinguished from one another?

Focus


Most reviewed wine variety
Pinot Noir
Bordeaux-style Red Blend
Chardonnay
Cabernet Sauvignon
Red Blends
Rosé
Sauvignon Blanc
Sangiovese
Gamay
Syrah
Portuguese Red
Nebbiolo
Sparkling Blend
Tempranillo
Merlot
Zinfandel
Portuguese White
Cabernet Franc
Riesling
Champagne Blend
Variety
0
250
500
750
1000
1250
1500
Count

04

Theory

Vectorization and BERT

Text Vectorization

Conversion of text to numerical values for model integration

Statistics based

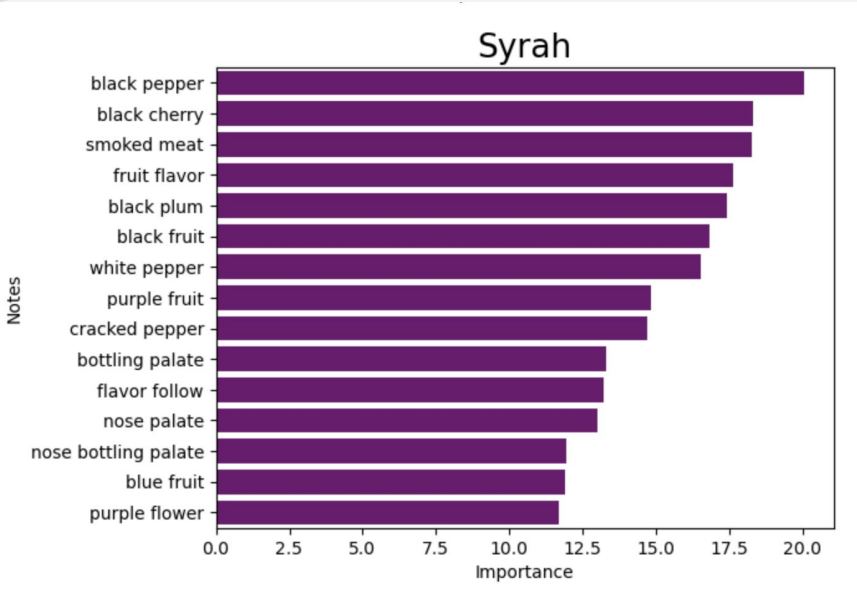
- OHE:
- Count Based:

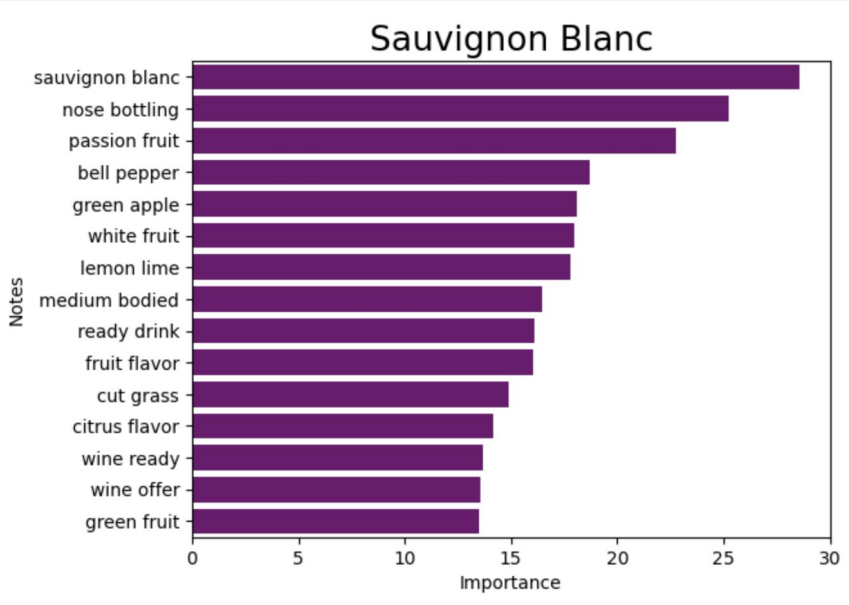
Predictive Based

- Sequential RNN:
- Transformers:

Focus

TF-IDF Vectorization





BERT

"Bidirectional Encoder Representations from Transformers"

- Stacked Transformers
- Better Contextual Relations
- Pretrained
- Fine-Tuning

05

Models

Statistical, Untrained and Pre-trained

Models

Model	Train Accuracy	Test Accuracy
Multinomial Naive Bayes	0.908	0.819
Sequential RNN	0.824	0.7026
BERT	0.987	0.8952

DEMO

Welcome to the Bottle-O-Wine Selector

The app to help you choose a wine based on specific tasting notes.

What are tasting notes are you looking for?

pink starbursts and smoke

What is the maximum price you would like to spend?

40

Submit

Pinot Noir

Here are some recs:

	points	title	↓ price
1389	94	Donnachadh 2021 Pinot Noir (Sta. Rita Hills)	40.0000
13233	96	Au Bon Climat 2019 Bien Nacido Vineyard Pinot Noir (Santa Maria Valley)	40.0000
6571	96	Williams Selyem 2020 Pinot Noir (San Benito)	39.0000
14034	94	Scar of the Sea 2020 Bassi Vineyard Pinot Noir (San Luis Obispo County)	36.0000

06

Conclusions

Improvements and Further Research

Conclusions

- BERT model increased accuracy by almost 10% from Naive Bayes
- Exponential increase of computing power and memory
- Some data leakage occurred since names of the grapes were mentioned within the review.
- Good use case established for continued development of an app

Further Research

- Improve model by adding more data to expand classifying to 300+ types
- Try out different pre-trained models
- Use clustering grapes before variety classification
- Add additional features like grape color (red, white, orange)

Thanks

Questions?