Analyzing Wine Reviews Using Text Analytics and Machine Learning

Katie Cao

BACKGROUND

- Wine is an alcoholic beverage made with the fermented juice of grapes.
- Wine tasting is the sensory observation and appraisal of wine. Wine reviews usually describe the flavors, aroma, comparing to other non-grape fruit or aromas and flavors such as citrus, honey, leather.
- We want to perform exploratory analysis on a relatively large dataset on wine reviews.
- Our analysis will help us determine the main criteria through which a quality in wine is determined by testing various features such as price, rating, review, region, etc.

Proposal

- The primary goal is to **rank the wines based on sentiment score** derived from reviews and the points given to each wine by wine enthusiasts.
- Secondly, based on that sentiment rank, **identify the top three countries**, regions, and provinces from which the most popular wines are produced and perform text analysis specifically on each of those top countries.
- Identify **the most popular wine varieties** based on the reviews given to each wine.
- Build **a machine learning model** that can **predict the variety of a wine based on the review description**.

HYPOTHESES

- Wine price is positively correlated with the higher rank given to review.
- Popular wine producing countries (such as France, Italy) are likely to produce higher ranked wines.
- Wines with the perfect score i.e., 100 also have the best sentiment score.
- Better ranked wines also have a positive correlation with longer description.

DATA DESCRIPTION

Our dataset is a list of 129,971 reviews of different wines around the world by professional wine tasters at Wine Enthusiast, one of the preeminent wine magazines. The data was scraped from winemag.com during the week of June 15th, 2017.

Variables	Meaning
country	Country where the wine is from
description	Review of the qualities of the wine
designation	Vineyard within the winery where the grapes that made the wine are from
points	Number of points Wine Enthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
price	Cost for a bottle of the wine
province	Province or state that the wine is from
region_1	Wine growing area in a province or state
region_2	Sometimes there are more specific regions specified within a wine growing area
taster_name	Name of wine taster
taster_twitter_handle	Twitter handle of wine taster
title	Title of the wine review, which often contains the vintage
variety	Type of grapes used to make the wine
winery	Winery that made the wine

METHODOLOGY

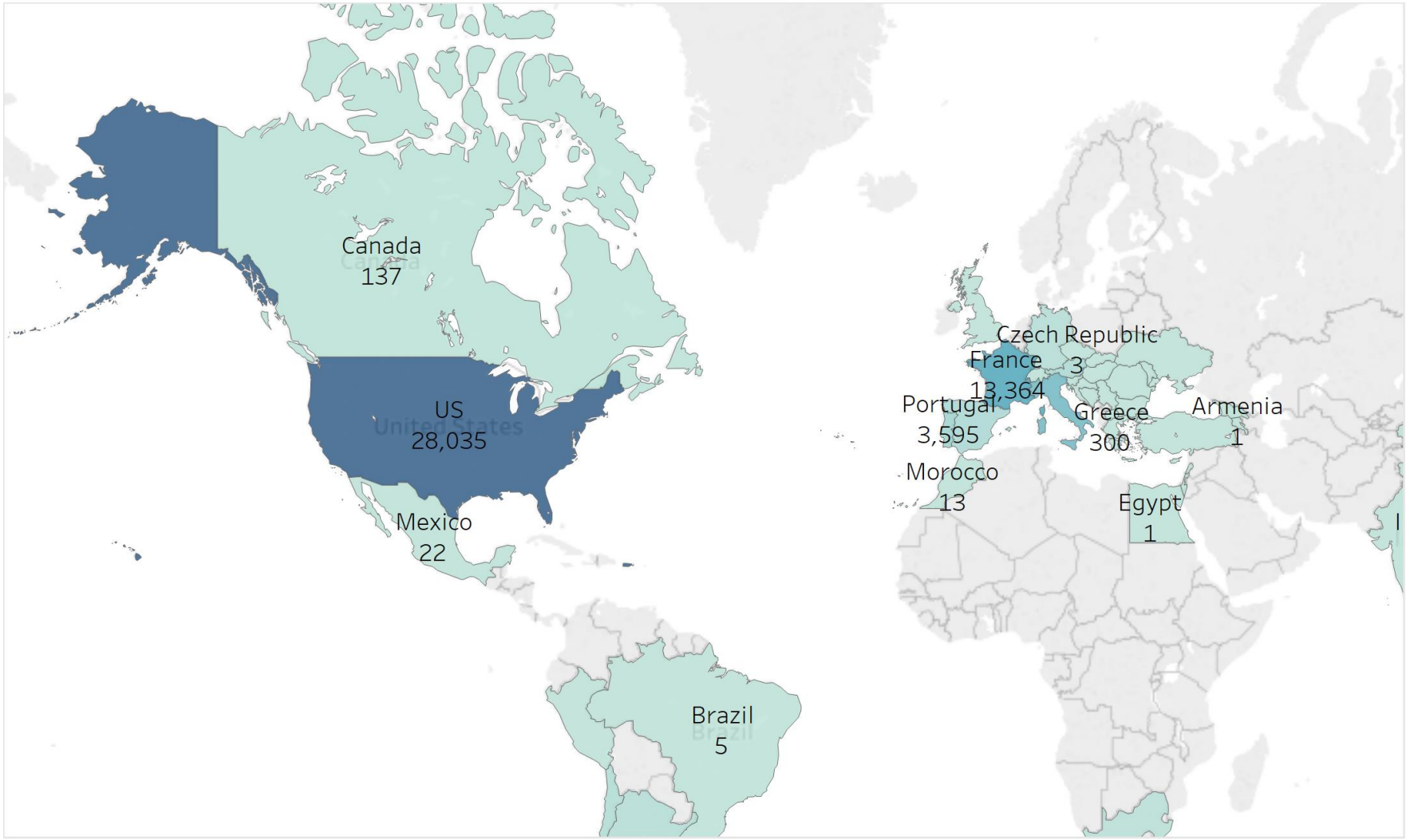
Data Collection/ Preprocessing → Sentiment Analysis Using Vader → Tokenization → Stopword Removal → - Stemming - POS tagging - N-grams → Analysis + Evaluation

Classified into Positive/Negative/Neutral
Assign rank based on sentiment score
Identify top three countries: France, Portugal, US

SPSS Modeler

Data Preprocessing → Variable Selection → Building ANN Model → Analysis + Evaluation

RESULTS & DISCUSSION – Top three countries

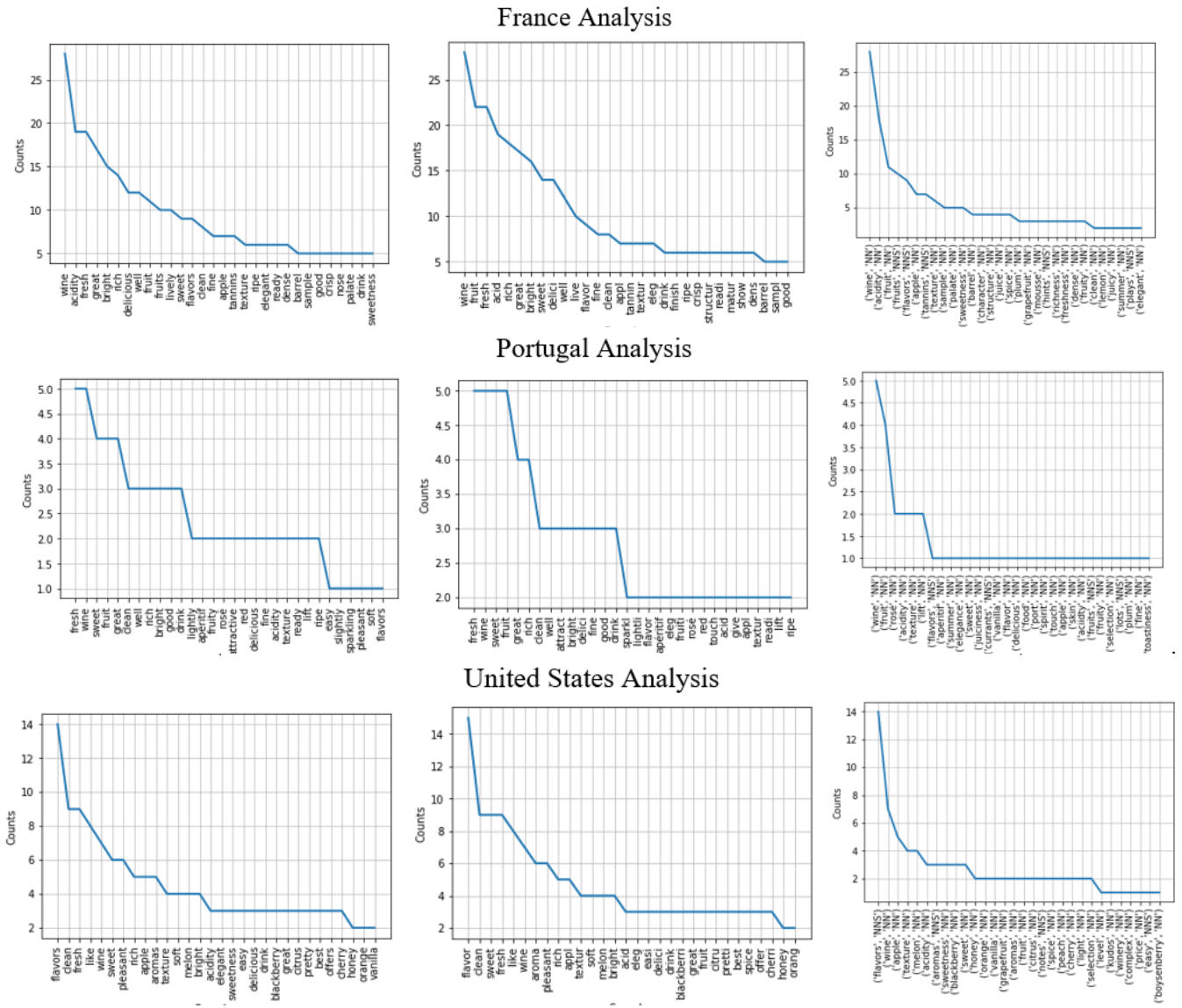


Vader Compound Score by Countries

RESULTS & DISCUSSION – Top three countries

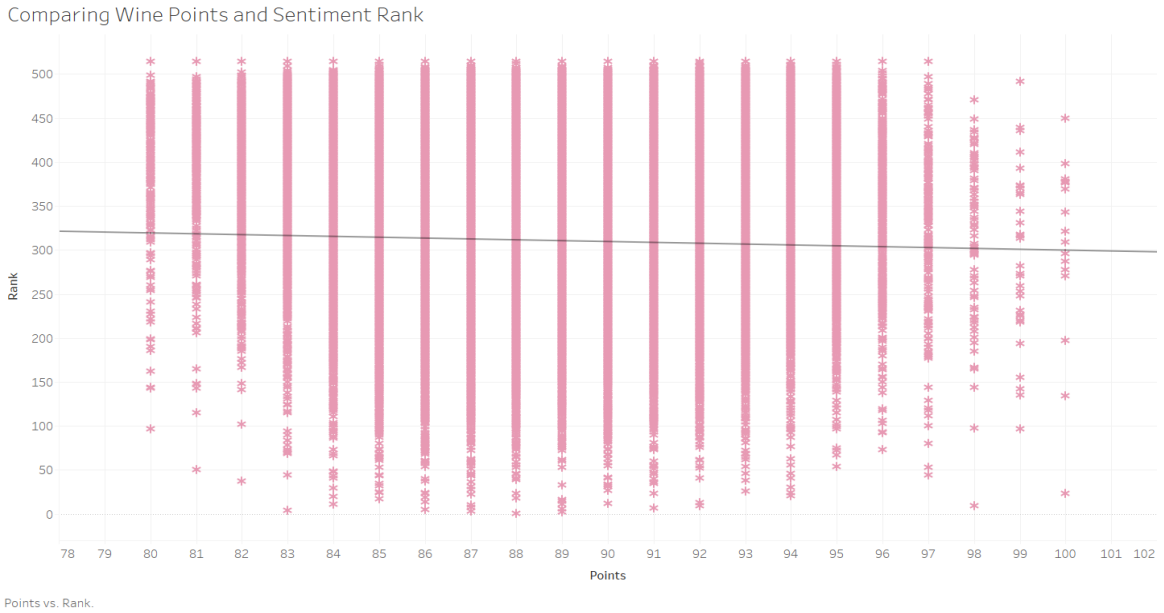
	France	Portugal	United States
After stopword removal	After stopword reduction we can see France wines have characteristics like fresh, acidity, bright, rich, lively, sweet.	After stopword reduction we can see Portugal wines have characteristics like sweet, rich, lightly, etc.	After stopword reduction we can see US wines have characteristics like pleasant, rich, aromatic etc.
Stemming	Helps in retaining root word like acid, sweet, rich, fine	Helps in retaining root word like light, rose. Rosé wine is preferred.	Helps in retaining root word like soft, clean, aroma, pleasant, spice etc.
POS tagging	Apple, plum, grapefruit, strawberry are among noticeable words.	Rose and vanilla are famous ones. Texture of wines is good.	Melon, blackberry, orange, grapefruit appear to be top descriptors.
n-gram	“fine” and clean” - most common “intense” and “sparkling” “apple” and “barrel” “sweet” and “fruits” “delicious” and “sweet”	“easy” and “fresh” – most common “slightly” and “sparkling” “pleasant” and “soft” “lightly” and “sweet”	“honey” and “orange” – most common “vanilla” and “flavors” “rich” and “balanced” “elegant” and “clean”

RESULTS & DISCUSSION – Top three countries

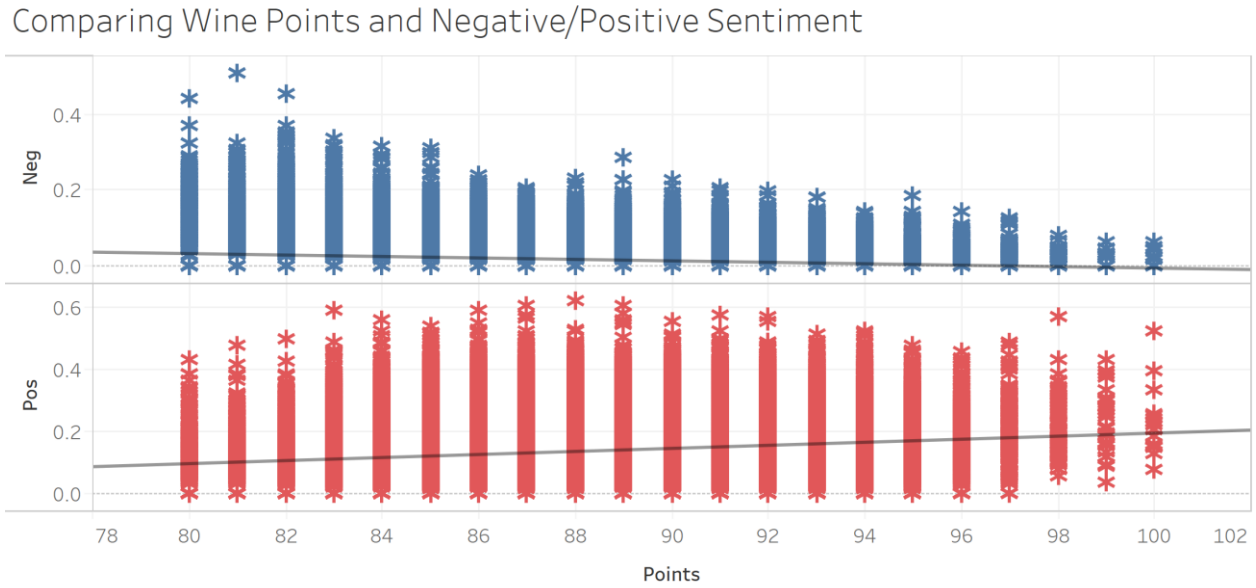


From left to right: Subplots showing word frequency of Tokenization after Stopword removal, Stemming, POS tagging

RESULTS & DISCUSSION – Visualization and Analytics

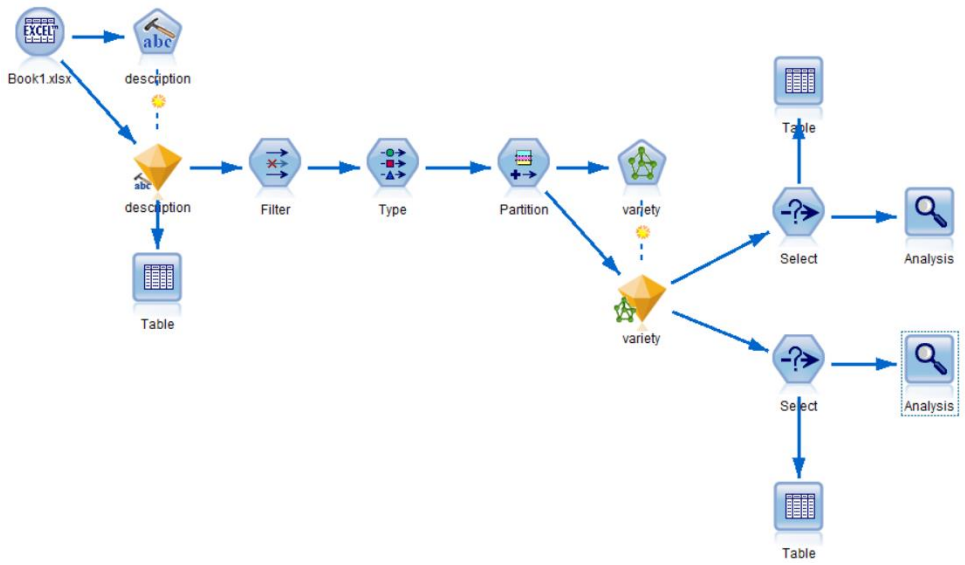


Comparing Wine Points and its Corresponding Sentiment Ranking



Comparing Wine Points and its Negative/Positive Sentiment

RESULTS & DISCUSSION – Machine Learning Model - ANN



Model Stream in SPSS Modeler

Observed	Predicted									
	Barbera	Chardonnay	Gamay	Merlot	Nebbiolo	Riesling	Syrah	Tempranillo	White Blend	Zinfandel
Barbera	0.0%	6.7%	2.1%	10.8%	36.1%	1.4%	19.7%	6.4%	1.6%	15.1%
Chardonnay	0.0%	89.9%	0.4%	0.2%	0.1%	5.6%	0.4%	0.4%	2.6%	0.5%
Gamay	0.0%	12.0%	71.4%	3.0%	1.3%	2.4%	3.9%	0.4%	0.6%	5.1%
Merlot	0.0%	2.6%	2.0%	54.6%	4.6%	0.5%	13.6%	7.4%	0.2%	14.4%
Nebbiolo	0.0%	1.5%	0.4%	2.5%	86.6%	0.6%	3.3%	3.0%	0.2%	1.9%
Riesling	0.0%	23.5%	0.5%	0.0%	0.1%	72.9%	0.5%	0.3%	1.9%	0.4%
Syrah	0.0%	2.4%	1.0%	6.5%	4.5%	1.1%	67.7%	6.0%	0.3%	10.6%
Tempranillo	0.0%	3.7%	0.6%	8.9%	9.8%	1.7%	18.0%	46.4%	0.8%	10.2%
White Blend	0.0%	30.9%	0.3%	0.4%	0.3%	8.0%	0.5%	0.7%	58.6%	0.3%
Zinfandel	0.0%	4.3%	1.9%	8.6%	3.9%	1.8%	15.4%	4.8%	0.6%	58.6%

Row Percent
100.00
80.00
60.00
40.00
20.00
0.00

Prediction Accuracy for top 10 Wine Varieties

RESULTS & DISCUSSION – Machine Learning Model - ANN

Analysis of [variety]

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field variety

Comparing $N-variety with variety

'Partition'	1_Training	
Correct	20,644	72.55%
Wrong	7,810	27.45%
Total	28,454	

Analysis of [variety] #1

File Edit

Analysis Annotations

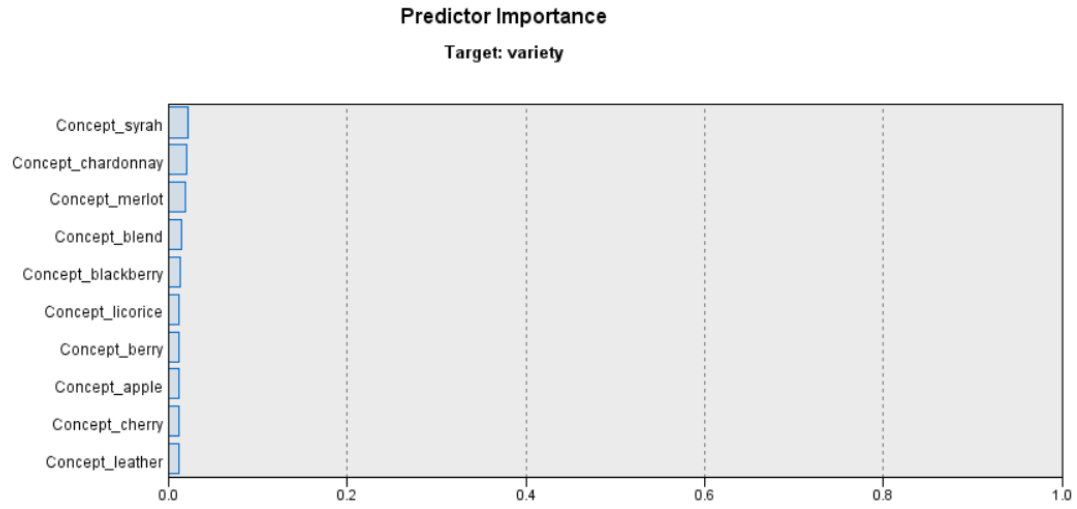
Collapse All Expand All

Results for output field variety

Comparing $N-variety with variety

'Partition'	2_Testing	
Correct	5,132	71.62%
Wrong	2,034	28.38%
Total	7,166	

Accuracy on Training and Testing Data



- If wine review contains the name of wine variety, it would increase the accuracy of the prediction
- If wine review contains common descriptors like fruits and material, it would also be more accurate to predict the wine variety.

CONCLUSION

Initial Hypotheses & Machine Learning Model

- Most of the top ranked wines are not considered expensive.
- Wines with high negative sentiment scores are also given low points.
- No significant correlation found between positive sentiment score/sentiment rank and wine points.
- Neural network model predicts wine variety well, with an accuracy of 71.62%
- Important predictors for wine variety are common sensory descriptors like blackberry, apple and cherry.

Interesting Findings About Wine

- The top 10 wine varieties are White Blend, Riesling, Chardonnay, Merlot, Gamay, Zinfandel, Syrah, Nebbiolo and Barbera.
- France, Portugal, and the United States frequently produce the top ranked wines. Our hypothesis related to France producing top wines has been satisfied, yet Italy does not rank high.
- Some of the most popular wines are Domaine Huët 2005 Pétillant (Vouvray), Château Haut-Simard 2009 Barrel sample (Saint), Straight Line 2011 Sauvignon Blanc.
- Provinces that produce highly-preceived and highly-ranked wines are Loire Valley, Bordeaux, California and Vinho Verde, to name a few.
- French wines are usually described as fresh, acidity, bright, rich, lively, sweet whereas American-produced wines as pleasant, rich, and aromatic.

FUTURE RECOMMENDATION AND FURTHER EXPLORATION

Recommendation

- Wine tasting- a subjective topic not backed by scientific research yet embedded in the culture and history all over the world
- What contributes to a popular wine variety and what drives the rating and the sentiment of wine tasters based on hundred thousand of reviews
- A professional sommelier: learn from other sommeliers with a large scale, fine-tune the art of food and wine pairing.
- A wine buyer: navigate through the delicacy of wine varieties and make a purchase without having to taste it.

Further Exploration

- Availability of data from other magazines beside WineMag and professional sommelier networks
- Further deep-dived sentiment analysis into:
  - The different vintages of one variety (vertical wine tasting)
  - Same vintage of different wine varieties (horizontal wine tasting)