WINE BLIND TASTING PREDICTION

BT5153 Topics in Business Analytics
Final Project Report

GROUP 14
Hong Li(A0186004M)
Cun Xiaofei(A0186051J)
Liang Xinran(A0186708R）
Liang Shijie(A0186001U)
Sin Yu Fan(A0139284X)

Contents

1 Problem Description

The main purpose of our project is to solve an interesting game called “Blind Tasting”, which is also a test for every wine taster pass to get their professional certification. In the blind tasting test, the wine tasters are required to smell, taste and observe the state of each glass of wine, then give some comments about the wine based on their professional knowledge, like the production area(usually country) of certain wine, the vintage(production year), the grape type, winery, and ect. The more accurate and detailed the taster guesses, the higher the possibility of passing the test. However, the components of wine are complex and all the subjects should be completed by tasters’ experience, which leads to low accuracy rate of guessing. Therefore, our group decided to develop a prediction model to solve this blind tasting puzzle. Through aroma combination of certain wine abstracted from taster review text, we are able to predict its production year, production area as well as grape type, with excellent improved accuracy compared to professional wine tasters.

In addition, based on the prediction model developed above, we further use some document summarization method to build up a practical wine recommendation system for non-professional customers. As the threshold of wine knowledge is too high for customers to learn, those average consumers hardly know how to choose suitable kind of wine. In reality, most of wine purchasing behaviors are influenced by brand instead real demand. Thus, by using document summarization, we simplify and refine professional reviews into readable introduction for consumers. In terms of application, the customers only need to input their aroma and taste preference of wine in our wine recommendation system, then they will get a certain kind of wine contained those aroma and also a readable introduction of it, which greatly improves convenience of purchasing wine.

2 Data pre-processing and dataset description

2.1 Description of raw data

The raw dataset of our project is the Wine Reviews dataset in Kaggle(https://www.kaggle.com/zynicide/wine-reviews), which is crawled from a wine wine reviewing and ranking website *Wine Enthusiast*. The raw dataset contains 10 columns, recording basic information and professional descriptions from wine tasters of 129,970 bottles of wine. Basic information of wine includes title, grape variety, winery, price, country and region of production, and so on. One highlight of this dataset is the aroma and taste description of wine which can captured from tasters’ descriptions. This information seldom appear in other sources, but it is of great importance for consumers to correctly select wine by their preference of flavour.

2.2 Data pre-processing

Data pre-processing of our dataset includes the following steps:

- Delete rows with NA value in columns description, country, variety or title. Information in these columns will be extracted as independent and dependent variables in our model.
- Delete countries with less than 1,000 bottles of wine in the dataset. In raw data, distribution of wine between countries is unbalanced. Some countries only contain several bottles of wine, which may result in small training dataset for the prediction of these countries. Therefore, we only keep countries with more than 1,000 rows to guarantee sufficient training data.
- Filter the top 30 grape varieties in number, and delete the rest. There are more than 700 varieties in the raw data, but most of them contains only 1 or 2 bottles of wine, most of the wines are within a set of tens grape variety. So we selected top 30 most popular grape varieties and deleted other rows.
- Extract production year from wine's title. Title of a bottle of wine(eg. "Quinta dos Avidagos 2011 Avidagos Red (Douro)") contains its production year. Production year is extracted from this column and a new column "year" is created to store this information.
- Create "year generation" column. Since exact production year is too difficult to predict even by professional wine taster, "year generation" column is created from "year" column, which includes 4 generations: 2010s, 2000s, 1990s and earlier than 1990.
- Group by "country", "variety" and "year generation" to create wine categories. We define wines with same country, variety and year generation as one category, and this category is the label of prediction.

2.3 Dataset description and visualization

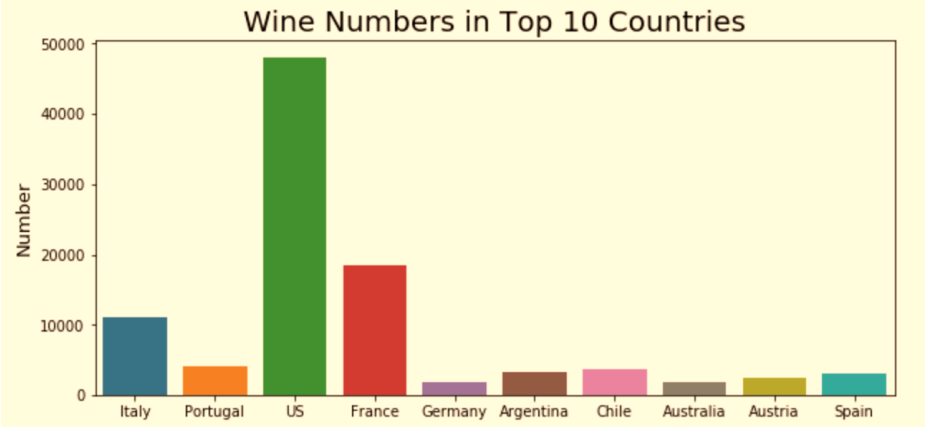
After data preprocessing, our dataset contains 2 tables: the first table has 100,318 rows, each row contains text description, production country, grape variety and year generation of one bottle of wine. The second table has 532 rows, each row shows one category of wine under our definition. The text description of each category is the sum of descriptions of all wines of the category.

description	country	variety	year_generation
Aromas inclu	Italy	White Blen	2010s
This is ripe ar	Portugal	Portugues	2010s
Tart and snap	US	Pinot Gris	2010s

index	description	count	year_generation	country	variety
('1990s', 'Argentina', 'Chardonnay')	Could the sc	1	1990s	'Argentina'	'Chardonnay'
('1990s', 'Australia', 'Cabernet Sauvig	Blackberry, h	8	1990s	'Australia'	'Cabernet Sauv
('1990s', 'Australia', 'Chardonnay')	This big wine	21	1990s	'Australia'	'Chardonnay'

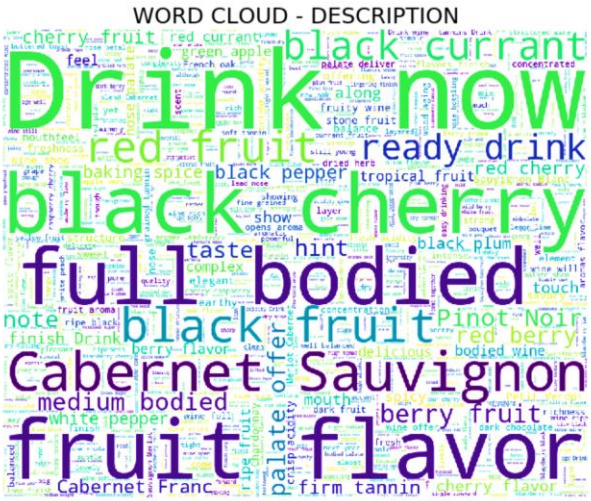
table 2.1 format of the dataset after data preprocessing

The country with the largest number of wines in the dataset is US, followed by France and Italy. In general, number of wines is fairly distributed between new world countries and old world countries.



Graph 2.1 wine numbers in top 10 countries

Word cloud of all text descriptions in the dataset is created. From the word cloud we can discover that most of the high-frequent words in descriptions are about the aroma, flavour and taste of wine.



Graph 2.2 word cloud of text descriptions

2.4 Dataset exploration: old world vs new world

A common sense of wine is that wines produced in old world countries and new world countries have different flavours and styles. We utilize our dataset to explore the grape variety distribution between old world and new world. The result is obvious to show that old world and new world countries are likely to product wine using different types of grape, which may because of their unique climate and soil. We also find that some countries have their own popular grape variety. For example, Shiraz is a popular grape type to product wine only in Australia.

New/Old World	Top 3 Grape Variety
Old World	Red Blend
	Bordeaux-style Red Blend
	Riesling
New World	Pinot Noir
	Cabernet Sauvignon
	Chardonnay

Table 2.2 top 3 grape varieties of old world and new world

Wine produced by different grape varieties have different aroma. To test whether it is true, we make use of the extracted words describing aroma using wine aroma dictionary(https://www.aromadictionary.com) from wine descriptions for feature forming of our model(which will showed in the following chapter) to explore how different aroma distribute between old world and new world countries. From the following table we observed that some aroma are only popular in either old world or new world countries. For example, oak and raspberry are in the top 10 list in new world countries, but they are not popular aromas in old world countries. Moreover, when we look at the most popular aroma of each country, we can also find some countries' most popular aroma is unique among all countries. The above discoveries show that aroma words are good predictors of wine's country, which encourage us to us aroma words as independent variables of our wine category prediction model.

Top 10 Aroma	
New World	Old World
Fruit	Fruit
Cherry	Cherry
Oak	Spice
Spice	Plum
Blackberry	Wood
Plum	Apple
Raspberry	Citrus
Vanilla	Pepper
Apple	Mineral
Chocolate	Blackberry

New/Old World	Country	Popular Aroma
New World	Argentina	Plum
	Australia	Vanilla
	Chile	Plum
	New Zealand	Cherry
	South Africa	Spice
	US	Cherry
Old World	Austria	Pear
	Franch	Wood
	Germany	Mineral
	Italy	Cherry
	Portugal	Wood
	Spain	Plum

Table 2.3 popular aroma in old world and new world countries

3 Methodology

In order to use the textual review descriptions to make predictions on "country", "year generation" and "grape variety", various techniques were used to capture data, as detailed in the following few subsections. A combination of these methodologies were included in the model and the accuracy of the model improved as captured in Section 4.

3.1 Aroma Dictionary Approach

As observed in section 2.4, the wine's aroma is a good indicator for the prediction of the "country" value. We wanted extract aroma words for each wine review description to help predict the three output variables. On www.aromadictionary.com, three documents were found containing keywords for white wine, red wine and wine making. After compiling the three list of words, a countvectorizer was fitted on this list of words.

In doing so, we built a countvectorizer with 295 single-word aroma descriptors, and the reviews of the individual wines can then be transformed by the countvectorizer to count the occurences of each aroma descriptors in each review. The reviews now have an additional 295 features. Although sparse, these features are able to improve the accuracy of predictions.

3.2 LDA

As the aroma dictionary is very limited in capturing information contained in the majority of the review description, we decided to use LDA topic modelling to process the review description. There may be underlying differences in the reviews along the lines of their grape varieties, and we believed that it would help in our predictions.

The probabilities given by the LDA model are additional variables that are added to the 295 features in section 3.1. After performing a grid search for the local optimal number of topics, we found that the best results was obtained with 10 topics.

3.3 Sentiment score

The next feature we wanted to capture was the sentiment score. We hypothesised that wines from the same group will receive similar sentiment scores from reviews, and therefore it would be a telling indicator for our predictions. Using text sentiment analysis, an additional feature ranging from -1 to 1 was created for each review.

3.4 Bag of Words

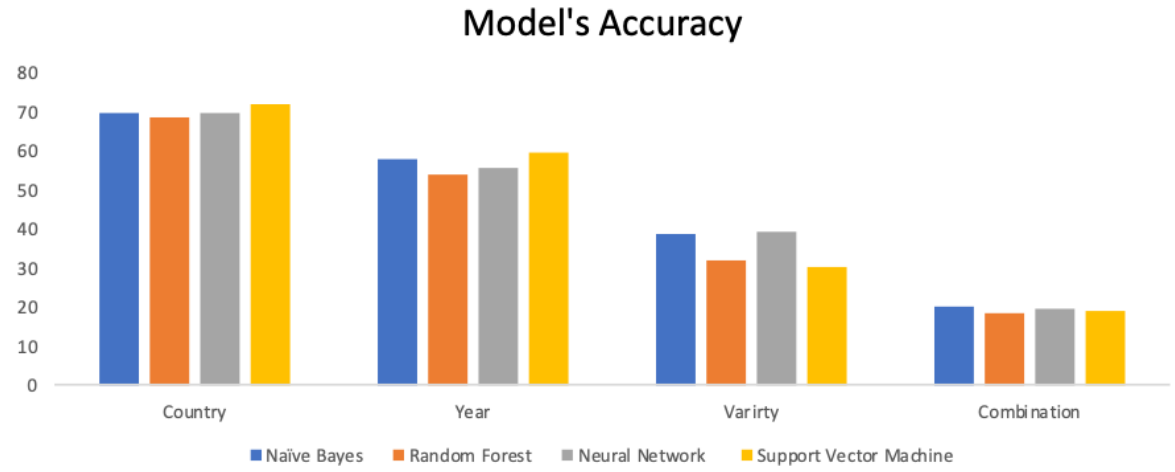
Lastly, we decided to process the entire review using countvectorizer. A countvectorizer with vocabulary consisting of monogram and bigram words from the reviews was constructed. As the signals from the aroma dictionary and LDA modelling will be captured by this Bag of Words approach, the model is built with only the countvectorizer and sentiment score inputs.

The results of the various predictive models built using a combination of the features stated above are shown in section 4.

4 Model Results

To solve this multi class prediction model, we tried to use the description of different tasters' reviews to predict the year, country and grape category of wines in three different models and combine the three models' result together to get the final result. Our final goal is to find the most optimal model to predict the combination result of wines as accurate as possible.

We trained models such as naïve bayes, random forest, neural network and support vector machine (SVM). We use only use aroma related words extracted from description as features to do prediction and the result of the models is shown in the following graph.The graph shows although SVM performs better in country and year prediction, on the whole, Naïve Bayes performs better in the combination, so we use Naïve Bayes in this case.



Graph 4.1 Accuracy of different models

By using text mining methods with a step-by-step procedure, we want to find which combination of text mining methods can lead to a better result.

	predict_grape_type (30 types)	predict_country (12 countries)	predict_year (4 time interval)	combination 532-kind wines
Benchmark(use the highest proportion)	12.98%	47.85%	68.75%	7.34%
Aroma Dictionary(295 features)	39.07%	58.13%	70.08%	20.42%
Aroma(295) + LDA(10 topics)	41.07%	60.04%	71.83%	22.08%
Aroma + LDA + Sentiment	42.67%	62.43%	73.06%	24.04%
countvector 1-gram + sentiment	59.35%	82.58%	78.23%	40.69%
1-gram & 2-gram + sentiment	60.58%	84.31%	78.72%	42.63%

Table 4.2 Accuracy of different NB models

The final result reaches 42.63% in our final model, which is better than we predicted. In this case, we use the highest label's proportion as model's benchmark. Since it is a multi-class, so we suppose the most naive way is that the machine will only output the most voted label as the result whatever the input dataset is. We find that when using Aroma dictionary alone and using countvectorizer with sentiment, the performance increases dramatically compared to the former model. The reason behind is that we increased a great number of features in and the features will contain some information about year, country and variety. Model 4's prediction result is better than Model 1's prediction result means that our aroma dictionary doesn't have enough scale to contain all the information that we need for the prediction.

The difference of accuracy between the last two rows shows combination of single word and double words can lead to better prediction on the final result than just using single word, because we assume that the combination adjective and noun can more clearly direct to a specific label.

5 Business Application 2

5.1 Procedure

For our second application, the recommendation system for wine buyers who have only the basic or even little knowledge about wine, what we expected is when a buyer inputs the basic wine descriptive words, our system will automatically predict a category of wine to him. Thus, the model's inputs are wine descriptive words, such as aroma, taste and color, while the output is one wine category. This model is actually the same as one of models we built in previous step when realizing the first application. Refer to part 4 for the model's details and performance result.

Apart from the type of wine recommended, we also want to attach a brief introduction to help our buyer understand the wine better. However, the wine descriptions given by professional reviewers are usually quite long. As a result, we did text summarization on each type of wine's description.

5.2 Text Summarization

Text summarization is most applied in media industry to give readers a brief summary, which help people decide whether to jump in and read the whole article under busy schedule. The key of text summarization is concise and fluent while preserving the key information.

5.2.1 The algorithms of text summarization chosen

In general, there are two types of text summarization, abstractive and extractive. The former aims at generating the most representative words, which can be compared to the way human read and summarize the text using our own words, so the words generated may not even exist

in original text. Extractive method focuses on sentence level, which weights the sentences and uses the same to form the summary. In order to make our attached summary as professional as possible, we chose to use the extractive method.

The weight of a sentences can be decided by either the similarity or the importance. Because we have already combined the different reviewers' description of the same type of wine, which make the first definition of sentence weight biased, we chose to use the term frequency to weight the sentence.

Stemmer is of great importance when calculating term frequency. If using Lancaster stemmer, some short words will be totally obfuscated and thus not as intuitive as Porter, which can be verified in the example below. Thus, we decided to use Porter stemmer to remove the suffix and prefix of a term before assigning it to a term index.

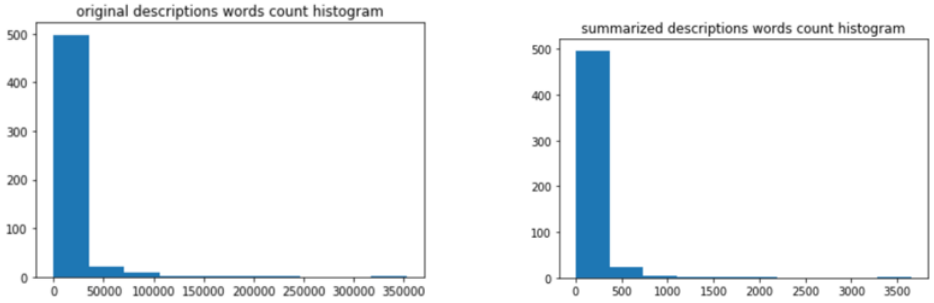
Graph 5.2.1 Example of result using "Porter stemmer" and "Lancaster stemmer"

5.2.2 The step of text summarization

a. Created the word frequency table.
   Created dictionary for the word frequency table from descriptions of each category. When tokenized word, we also removed the stopWords and English punctuations.
b. Tokenized the sentences.
c. Scored the sentences
   We scored one sentence by adding the frequency of every word in it. Because a potential issue is that long sentences have an advantage over short sentences, we divided every sentence score by the number of words in the sentence.
d. Found the threshold
   After ranked sentences, we chose the score of the sentence that situated at the first 5% to be threshold. Besides, also set the maximum number of sentences to be 10.
e. Generated the summary
   Used the threshold and sentence scores to generate the summary.

5.2.3 The results

Below are two distribution diagrams of the number of words in wine descriptions. It is obviously that the overall descriptions changed a lot after text summarization. The average number of words in one type of wine's description decreased from about 8571.29 to 147.38, while the longest description, which contained 362639 words originally, now only include 3649 words, shortened about 100 times, but still preserved the key meaning of the original wine review.



Graph 5.2.3 Distribution of word count beore and after document summarization

5.3 Example of application two

To illustrate, give one example of application two. For instance, if a customer inputs wine descriptive words: “black”, “dusty”, “pepper”, “tobacco”, “vanilla”, the output will be the “Red Blend” produced in “US” in “2010s”. And a brief summary: “This blend is made of Zinfandel, Syrah, Petite Sirah, Cabernet Sauvignon and Mourvèdre. It's a versatile, easy-to-like wine..."

6 Conclusion

- The improvement of prediction result is incredible and also meet our expectations. The prediction accuracy of grape type and production country skyrocket by using aroma dictionary and 1-gram countvector, while the prediction accuracy for vintage(production year) hardly improves.
- The aroma dictionary input in model contributes a lot for grape type and country accuracy, we believe that this is because the flavor and aroma of wine mainly depended on grape type. Secondly, the flavor style of different countries varies a lot. In reality, it can make sense as people always emphasize the production country of wine.
- The improvement contributed by 1-gram countvector might due to the insufficiency of aroma dictionary. In other words, the aroma category in dictionary is unable to cover all the flavor in tasters’ review. Therefore, as we develop countvector, some new aroma can be caught by model and further improve the accuracy.
- When we use 2-gram countvector instead of 1-gram, the model improves a little, which might due to the combination of 2 aroma phase, like the flavor of “black berry”, “blue berry” are different from “berry”. The 2-gram countvector further accurate the flavor combination and the model improves reasonably.
- The prediction accuracy of vintage(production year) did not improve a lot compared with others for mainly 2 reasons. Firstly, the benchmark accuracy is 68.75% and leave small promotion space. Secondly, as vintage year increases, a chemical substance called tannin content increases as well. However, the taste of tannin is hard to be described by aroma phase, which makes text mining of review invalid.