A Data-Driven Approach to Wine Quality Ratings and Consumer Preferences

Team Members: Charles Lazaroni, Lucas Attias, and Mark Cappiello

IST 707 – Applied Machine Learning

Project Report

Introduction

The wine industry is simultaneously complex and expansive. Every winery is unique and produces its own distinctive wines and flavors. In addition, there is an immense number of individuals who rate and review these wines. These factors propagate a vast assemblage of choices and decisions that can disconcert not only the consumer, but also the winemakers in the industry. Resources like Wine Enthusiast Magazine can help mitigate the confusion by providing insights into what bottle (or bottles) might be of interest to the consumer, as well as the factors that contribute to the production of the desired wines.

However, this data is difficult to make sense of as it is presented – a block of 130,000 wine reviews sourced from Wine Enthusiast Magazine 2017-2018. This is raw data. To gain insights from it, someone will need to process it. Noticing this inaccessibility- we decided to undertake the effort to extract useful information. Our goal is simple- to help wineries make better decisions about the wine they grow, and to help consumers and distributors make better decisions about the wine they purchase.

We achieve this through the application of machine learning methods and analysis. We will uncover patterns and relationships within the data to develop classification models that estimate wine quality ratings and generate wine recommendations based on features such as geographic origin, grape variety, and taster's review. By integrating both structured data and natural language processing techniques for text-based descriptions, this project illustrates how machine learning can be utilized to address complex pragmatic problems.

Literature Review

The relationship between a wine's country of origin and its perceived value has been the focus of some research, particularly for high-quality wines[2]. However, there remains limited investigation into which wines are consistently rated higher and why. Furthermore, there is a distinct lack of accessible, open-source tools that allow consumers to retrieve personalized wine recommendations. This research aims to address these underexplored areas by providing data-driven insights into wine ratings and characteristics.

Wine ratings and reviews, though highly subjective, play a significant role in consumer decision-making. While personal preferences influence these reviews, sommeliers possess extensive and trained palates capable of identifying objective truths about wine flavors, bodies, and hints. This duality creates a challenge for consumers who rely on subjective reviews to choose wines and for wineries that use these ratings to understand what contributes to highly ranked wines.

Despite the importance of wine reviews, there is a lack of data-driven perspectives examining the patterns and relationships between wine characteristics, descriptions, and ratings. Existing

research rarely explores the influence of factors such as grape variety and taster profiles on wine ratings. By classifying wine ratings and constructing taste profiles based on sommelier reviews, we aim to bridge this research gap and deliver actionable insights for both consumers and wineries.

For consumers, wine ratings are essential for discovering new wines that align with their preferences. Our goal for this study is to identify similar taste preferences across reviewers and provide tailored wine recommendations. This feature not only enhances the user experience but also empowers consumers to make informed purchasing decisions.

For wineries, understanding the factors that contribute to higher quality scores is equally valuable. Insights derived from this research can help wineries identify optimal grape varieties, production methods, and regional growing strategies. By classifying quality scores and exploring correlations among features, wineries can refine their production goals and strategies to meet consumer expectations and improve their market position.

Data

The data originates from reviews in Wine Enthusiast magazine -- A 50+ year old company that specializes in creating high quality wine reviews and ratings from sommeliers. The dataset used in this study can be downloaded from the following link:
www.kaggle.com/datasets/zynicide/wine-reviews

After inspecting the dataset, fourteen columns and 130,000 rows were displayed. An exploratory data analysis was performed on the dataset to aid in further visualization of patterns and trends. A comprehensive explanation of the features is described below. Most of the features are a string datatype except where otherwise specified:

- Index (int)
- country: The country of origin of the wine.
- description: A textual description of the wine's appearance, aroma, and taste.
- designation: The vineyard within the winery where the grapes that made the wine are from
- Points (int): The wine's rating, on a scale of 0 to 100.
- Price (float): The price for one bottle of the specified wine.
- province: The province or state that the wine is from.
- region_1: The primary wine-growing region within the province.
- region_2: A more specific region within region_1.
- taster name: Name of the taster providing the text in the description field.
- taster twitter handle:
- title: The title of the wine review, which often contains the vintage.

- variety: The grape variety or blend used to make the wine.
- winery: The name of the winery that produced the wine.

Methods

The first section of this program contains an exploratory data analysis. The purpose of this section is to analyze patterns and trends in the data. The methods used in this section involve plotting/graphing the numerical data with histogram and scatterplot, clustering the data using unsupervised techniques such as k-means clustering, and analyzing descriptive text using word cloud. Graphing the data provides insight on how to clean and process the data before building a model for machine learning. Pair plots and correlation matrices were utilized to determine if any patterns exist in relationships between features. Clustering is a useful technique for visualizing the groupings of objects in a dataset. By using unsupervised clustering methods on the cleaned dataset, we grouped these datapoints into clusters based on similarity. The method of k-means clustering facilitated pattern recognition during the exploratory phase of the analysis.

The data was then pre-processed and cleaned for machine learning. Missing values were handled with imputation and removal methods, categorical features were encoded using OneHotEncoding, text data from the wine reviews was standardized using tokenization and stop-word removal, and numeric columns were normalized using a standard scaler to be compared with one another. Outliers were identified using the IQR method and imputed with the mean for each numeric column. The data was split into a consumer dataframe and a winery dataframe based on the features necessary for the stakeholders' needs. A 10% sample of the data was also extracted from winery dataframe to prevent crashing the program.

For consumers, two methods were used. The first was a series of in-depth EDA extractions to better understand Sommeliers. We started off by grouping the reviews by Sommelier. Then we calculated the mean review for each of them. We also noted the total number of reviews. After this, for each Sommelier we plotted a histogram. Then we took note of which wines were most popular regardless of score, and which wines were present in top 25% quartile reviews vs bottom 25% quartile reviews.

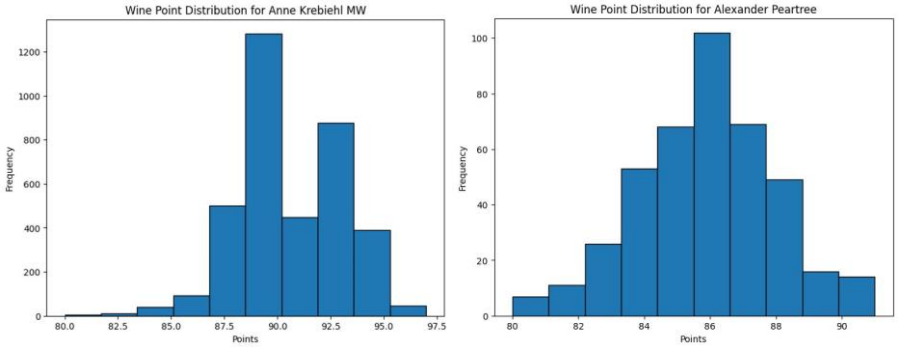
The second method we used for consumers was for the recommendation system. A series of data was prepared using the following features: price, region_1, variety, and winery. After this, the data was PCA reduced, and Euclidean distances to ID similar wines were calculated at inference time. Originally, we tried more complex clustering methods, but they yielded poorer performance and results compared to simply calculating distance. We hypothesize that this is a result of the complexity of wine and its varieties.

For wineries, supervised learning techniques were leveraged, and regression models were used to build a classifier that predicts wine scores based on the necessary features. Wines were placed into categories (low, medium, or high quality) based on their rating. Wines that received over 90 points were classified as high quality. Wines that were within the range of 87 to 89 points were classified as medium quality, and wines rated lower than 87 points were classified as low quality. These quality ratings were then encoded for classification by the regression models. Decision trees were bagged and gradient boosted to improve error reduction. Random forest and Support Vector Machine (SVM) models were also initialized and trained. The models were each evaluated using K-fold cross-validation and compared by analyzing F1-score precision and accuracy. The best performing model was chosen for the winery stakeholders.

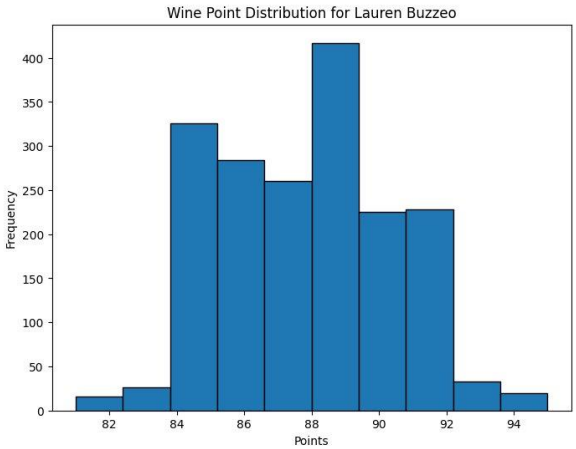
Results

Consumer Stakeholders

The EDA on sommeliers uncovered some interesting points about how sommeliers review wine over time. We expected wine to be evenly distributed in a normal curve. While that is close to what we found, there were distinct characteristics that reviewers showed over time.



To illustrate this, let's look at the easiest and harshest wine reviewers. As you can see, Ms. Krebiehl frequently gives scores above 90 - her average score, whereas Mr. Peartree rarely gives anything above a 90, an outlier score.



Another interesting pattern was Ms. Buzzeo, whose reviews almost appeared uniform in distribution from range 84-92. These results indicate a wide variety of choices in terms of quantifying the quality of wine when we begin to look from a bird's eye view.

	variety	review_count
0	Pinot Noir	12785
1	Chardonnay	11077
2	Cabernet Sauvignon	9384
3	Red Blend	8466
4	Bordeaux-style Red Blend	5340
5	Riesling	4971
6	Sauvignon Blanc	4780
7	Syrah	4086
8	Rosé	3261
9	Merlot	3061
10	Zinfandel	2708
11	Malbec	2593
12	Sangiovese	2377
13	Nebbiolo	2331
14	Portuguese Red	2196
15	White Blend	2167
16	Sparkling Blend	2027
17	Tempranillo	1788
18	Rhône-style Red Blend	1404
19	Pinot Gris	1388

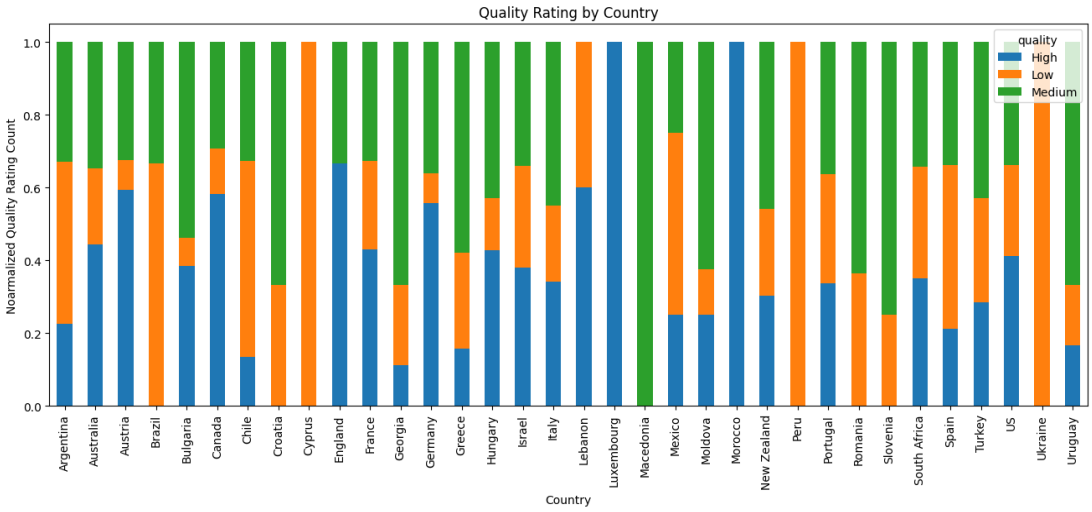
The second aspect of our tool to understand Sommeliers attempted to uncover which wines were most popular in positive reviews and negative reviews. We found that Rose and Portuguese Whites were over-represented in negative reviews, and Bordeaux, Ports, and Champagne were prominent in highly reviewed wine. We found Rosé to be particularly interesting in this scenario- it could represent a bias in reviewers or in lower quality vineyards producing Rosé because it is so popular.

The other tool we built for Consumers & Distributors was the Wine Recommendation System. We were pleased with these results- although they were difficult to quantify. Deploying this system to an API endpoint would be quite easy and be a large step towards our goal of providing accessible information on wine to consumers.

	country	price	province	variety	winery
60406	Chile	11.0	Maule Valley	Cabernet Sauvignon	Francisco Gillmore
85449	Chile	10.0	Central Valley	Cabernet Sauvignon	Misiones de Rengo
83292	Chile	9.0	Central Valley	Cabernet Sauvignon	Santa Rita
46219	Chile	10.0	Maipo Valley	Cabernet Sauvignon	Haras de Pirque
124828	Chile	10.0	Curicó Valley	Cabernet Sauvignon	Montes
61987	Chile	9.0	Central Valley	Cabernet Sauvignon	Espiritu de Chile
15393	Chile	12.0	Maule Valley	Cabernet Sauvignon	Bossy Boots
90611	Chile	10.0	Curicó Valley	Cabernet Sauvignon	San Nicolas
22026	Chile	10.0	Maipo Valley	Cabernet Sauvignon	Tres Palacios
67550	Chile	10.0	Maipo Valley	Cabernet Sauvignon	Vistamar

Winery Stakeholders

The exploratory data analysis for the winery stakeholders displayed interesting findings. The 3 countries producing the most wine found in this dataset are The United States, France, and Italy by a large margin. The wine quality classifications were also grouped by country, normalized, and graphed. The 3 countries producing the most wine have a balanced distribution between low, medium, and high quality. Italy produces slightly more higher quality wines than the other 2 dominant countries. The plot shows that although countries like Luxembourg and Morocco do not produce many wines, they are of the highest quality. The plot also shows that countries like Cyprus, Peru, and Ukraine tend to produce wines of lower quality. The graph of wine quality by country is displayed below:



The first regression model used to build a classifier to make predictions on wine qualities based on the features in the wineries dataframe was a decision tree classifier with gradient boosting. Gradient boosting works particularly well on regression problems but can also be adapted for classification in this case. This model yielded good results in the 5-fold cross-validation:

```
Cross-Validation Accuracy Scores: [0.57142857 0.5437788  0.52995392 0.51152074 0.55092593]
Mean Accuracy: 54.15%
```

F1-score precision and accuracy were also obtained from a classification report to validate the cross-validation results:

```
Decision Tree with Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.52      0.60        56
           1       0.47      0.44      0.45        79
           2       0.58      0.72      0.64        82

    accuracy                           0.57       217
   macro avg       0.59      0.56      0.57       217
weighted avg       0.58      0.57      0.56       217
```

The same process was repeated for a decision tree classifier with bagging. The results were compared with the decision tree classifier with gradient boosting to decide the best method for reducing error in the decision tree model. The results of the bagging method are shown below:

```
Cross-Validation Accuracy Scores: [0.51612903 0.49769585 0.53456221 0.48847926 0.52777778]
Mean Accuracy: 51.29%

Decision Tree with Bagging Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.54      0.57        56
           1       0.48      0.41      0.44        79
           2       0.58      0.72      0.64        82

    accuracy                           0.56       217
   macro avg       0.56      0.55      0.55       217
weighted avg       0.55      0.56      0.55       217
```

After evaluating the results of the two decision tree classifiers, gradient boosting will be used with the decision tree model to reduce error. Gradient boosting performed slightly better than bagging with a higher F1-score accuracy, and slightly higher macro average and weighted average precision scores. The 5-fold cross-validation scores were all higher for the decision tree with gradient boosting as well.

The random forest and SVM models were then initialized, trained, and a classification report was generated for the two models:

```
Random Forest Model Report:
              precision    recall  f1-score   support

           0       0.72      0.50      0.59        56
           1       0.43      0.42      0.43        79
           2       0.54      0.67      0.60        82

    accuracy                           0.53       217
   macro avg       0.56      0.53      0.54       217
weighted avg       0.55      0.53      0.53       217

SVM Model Report:
              precision    recall  f1-score   support

           0       0.59      0.52      0.55        56
           1       0.45      0.47      0.46        79
           2       0.60      0.62      0.61        82

    accuracy                           0.54       217
   macro avg       0.55      0.54      0.54       217
weighted avg       0.54      0.54      0.54       217
```

These two models scored well in F1-score accuracy and precision and performed similarly to the decision tree classifiers. However, the decision tree classifier with gradient boosting out-performed the other models. the winery stakeholders will utilize the decision tree classifier with gradient boosting to make classifications on wine quality rating based on the features necessary for the wineries.

Discussion

Overall, the consumer and winery stakeholders will be pleased with the results of this program. The taste profiles for different reviewers were successfully created and the recommendation system effectively matches consumers with wines of similar preference. This program drastically improves consumers' ability to filter the broad range of wines available to them to a narrower selection for purchasing and tasting. Wineries are also able to make predictions on wine quality based on the features necessary for production with solid accuracy and precision. Wineries can base their production in optimal locations and harvest the most advantageous grapes to bottle wines that return the best ratings.

Limitations

The main limitations of this program include the data quality, the complexity and bias of text analysis, and overfitting of the predictive models.

There are many rows with missing data such as missing price or wine review score. This limitation was minimized by preprocessing the data carefully and imputing or, where appropriate, removing data from analysis.

Because reviews are subjective, this made it difficult to understand the correlation between reviews and wine ratings. Bias also plays an impactful role based on the wine reviewer. This limitation was handled by engineering a taster profile feature based on a combination of reviewer and review description.

As with all data projects, there is a risk of overfitting the training data which could lead to poor performance. The use of cross-validation helped to mitigate overfitting. Due to the size of the dataset, processing capability for the regression models is limited. Only a sample of the data can be fully analyzed at once to avoid crashing the program. A more powerful server is needed to handle the processing overhead.

Future Work

To verify the reliability of Wine Enthusiasts Magazine, it might be possible to compare some reviews with alternative sources. This could involve scraping review data from other wine review sites.

More features could also be collected and included in the dataset. For example, if the numeric year value that the wine was bottled was included in the dataset, then winery stakeholders could be better informed about the optimal timing for bottling their wines. Although the "title" feature does contain the year bottled for some of the wines, it is difficult to use this value because it is part of a string datatype including the name of the wine.

Another design for this program that could be implemented in the future involves making predictions on competitive prices for the bottled wines based on their rating and geographic features. This would give wineries a vying advantage when pricing their wines. Consumers would also be provided with insight into how much they should be spending when purchasing new bottles.

Citations

1. Zackthoutt. "Wine Reviews." *Kaggle*, 27 Nov. 2017, www.kaggle.com/datasets/zynicide/wine-reviews
2. Bowe et al "Old dogs, new tricks- rethinking country image studies" https://www.researchgate.net/publication/259536227_Old_dogs_new_tricks_-_Rethinking_country-image_studies