Wine Reviews Prediction

Zhanming Xiao James Liu Kuhu Puranik Bhairavi Wadekar

Background, Motivation, & Business Question

We are interested in what features would affect the points of wines and seek to utilise machine learning to predict wine scores, addressing the challenge of the subjective and costly process of wine quality assessment. We will use the wine dataset from Kaggle, which contains 130k wine reviews with variety, location, winery, price, and description, to predict the points of the wine. Selecting the features that would affect the points of the wine is important for business because it can help the winery to improve the quality of the wine and increase the points of the wine, thereby enabling informed decisions on pricing, marketing, and inventory management. By transforming attributes (winery, location, types of grapes) into quantifiable data, our solution strives to offer strategic advantages such as optimised product design strategies and targeted marketing efforts. This innovation is crucial for gaining a competitive edge in the market, ensuring our client can better align their products with consumer expectations and market trends, thereby enhancing overall business performance in the competitive wine industry. We are particularly interested in the premium wine market and aim to empower winemakers and distributors with the ability to forecast wine quality, focusing on the features that would most significantly impact the points of the wine.

Data and Statistical Question

Statistical Question:

The primary statistical question we can explore is: How do various factors such as country, geographic region (points to soil quality), price, variety, and descriptive terms (tokens) relate to the wine score (points)? This question aims to identify the relationship between these variables and the wine score, which is crucial for predicting the quality of wine.
$X = production\ factors\ and\ flavour\ descriptions\ associated\ with\ wine\ scores$
$Y = wine\ scores$

Usefulness:

This question addresses the business need by identifying key predictors of wine quality, allowing for strategic decisions on pricing and marketing based on identifiable characteristics.
It enables the development of a predictive model that can forecast wine scores before they hit the market, offering a significant competitive advantage.
Understanding these relationships can help in tailoring the inventory to include wines that are likely to score higher, thus enhancing the product portfolio.

Shortcomings:

This question might not fully address external factors influencing wine scores, such as market trends, consumer preferences shifts, or the impact of specific tasters' ratings and the bias of a taster. It also does not account for the time-sensitive nature of wine scores, where the perception of quality might change over time or with ageing potential.

Feature Engineering:

We noticed that the points distribution is likely to be normal, with an average close to 88, thus we chose 88 as it marks the top 50% percent of wines, which we regard as premium wines.

In addition, in our quest to determine the characteristics that contribute to a wine being considered premium, we've opted to exclude certain columns from our dataset:

- 'Unnamed: 0': Essentially an index column.
- 'title' and 'description': The uniqueness of values in these columns closely matches the total row count, suggesting minimal impact on wine ratings. However, the description field is rich in information, detailing attributes of preferred wines (such as flavor). We plan to analyze it separately to identify impactful keywords on ratings.
- 'taster_name' and 'taster_twitter_handle': Given that the scoring process is a blind test, any personal biases are ostensibly eliminated, making these columns irrelevant.
- 'region_2': With nearly 60% of its values missing, this column fails to accurately represent the significance of subregions.
- 'price': Since price is considered an outcome rather than a determinant of quality, it will be excluded from our feature set.
- 'designation': With 30% missing data and half of the entries being unique, imputing values for this column could introduce significant bias. Therefore, it has been excluded.
- Following this rationale, we will discard the aforementioned columns, retaining the rest under the Major Features dataset. Additionally, we will isolate the description column in a separate dataset named Description Features.

By addressing the primary statistical question with the considerations mentioned above, including thoughtful feature engineering, the project can provide actionable insights into predicting wine scores, directly supporting the business question. This approach enables the client to make informed decisions that could positively impact their pricing strategy, marketing efforts, and inventory management, thus aligning with their goal of gaining a competitive edge in the wine industry.

Exploratory Data Analysis

The dataset comprises an array of features related to wine characteristics, including country of origin, designation, price, province, region, taster name, title, variety, winery, and a tokenized description capturing the essence of each wine's flavour profile and aroma ('tokens'). With a primary focus on predicting wine scores ('points'), the data presents a mixture of categorical variables (e.g., country, variety) and numerical variables (e.g., price), alongside textual descriptions. The presence of both global and specific wine attributes suggests a complex interplay of factors influencing wine quality as perceived by experts. Preliminary exploratory analysis indicates a skewed distribution of wines across countries and varieties, with a positive but variable correlation between price and wine scores, highlighting the dataset's potential for building predictive models that consider a multifaceted approach to understanding wine quality. This complexity, along with the richness of the descriptive text data, underscores the need for sophisticated feature engineering, particularly in processing textual data, to effectively capture and predict the nuanced attributes that contribute to a wine's score.

The exploratory data analysis (EDA) of the training dataset has revealed several key insights and potential considerations for building the predictive model (See figure 1):

- Distribution of Wine Scores: The distribution of wine scores appears to be approximately normal but slightly left-skewed, indicating that most wines have scores around the median, with fewer wines achieving very high scores. This distribution suggests that wine quality, as quantified by experts, tends to cluster around a central value, with exceptional qualities being rarer.

- Price vs. Wine Scores: The scatter plot of price against wine scores shows a positive relationship, indicating that higher-priced wines tend to have higher scores. However, there's considerable variability, especially in the lower to mid-price range, suggesting other factors also play significant roles in determining wine scores. The relationship is not strictly linear, and there's a wide spread of scores at different price points, highlighting the complexity of predicting wine quality based solely on price.

- Top 10 Countries by Wine Count: The distribution of wines by country shows a dominance of a few countries in the dataset, with countries like the US, France, and Italy having the highest number of wines. This distribution is important for modeling as it suggests that country-specific factors might significantly influence wine scores, but it also highlights potential biases towards wines from these countries.

- Top 10 Wine Varieties by Count: Similar to the country distribution, a few varieties dominate the dataset. This finding indicates the importance of variety in predicting wine scores. However, it also suggests that the model might perform better for these common varieties due to the availability of more data points.

Potential Issues and Considerations:

1. The wide variability in scores for wines across different price points suggests that while price is an important feature, the model will need to incorporate other features to accurately predict wine scores.
2. The dominance of certain countries and varieties in the dataset could introduce bias, meaning the model might perform better on wines from these countries or of these varieties. This will need to be accounted for, possibly through stratified sampling or by ensuring the model is tested on a diverse set of wines.
3. Feature engineering, especially for the 'tokens' column (which contains descriptive text of the wines), could uncover additional predictors of wine quality. Techniques such as

TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings could be useful in capturing the essence of wine descriptions.

4. The high multicollinearity among predictors can inflate R-squared values, making it challenging to discern the true contributions of individual predictors to the target variable. This can occur when predictors are highly correlated with each other, leading to unstable estimates of regression coefficients.

Conclusions for the wineries

Quality control: The models used in our analyses focus on providing insights into identifying key factors that contribute to the perceived quality of wine, allowing producers to adjust their processes to enhance quality.
Pricing Strategy: On feature engineering we can derive from the model the pricing strategies by correlating wine characteristics and quality with market prices, ensuring competitiveness and profitability, prioritising the wines with higher scores.
Supply Chain Optimization: By predicting quality more accurately, the wineries can better manage their supply chain processes, from selecting the right grape suppliers to optimising distribution channels based on quality tiers.

Method & Results

Our approach to addressing our inquiries is divided into two segments. The first segment concentrates on production-related factors, including country, province, region_1, variety, and winery, and is referred to as the Major Features dataset. The second segment utilizes the description column from the original dataset, aiming to leverage keywords that contribute to high scores to design products aligned with customer preferences. Since the description involves text mining, we might employ different classification techniques, such as Naive Bayes, distinct from those used for classifying the major features. We label this second segment as the Description Features Models. Both models will be developed from the same data split to prevent any training biases, yet they are intended to answer the same business question.

Major Features Models

- Dummy Classifier (Base Model): This model serves as a baseline to evaluate the performance of more sophisticated models. By providing a point of comparison, it helps to establish the minimum performance benchmark, indicating whether other models are capturing meaningful patterns or merely performing better than random guessing.
- Decision Trees: Decision trees are useful for their interpretability and ability to handle non-linear relationships. They can model complex interactions between features related to the wine's production, such as country, province, and variety.
- Random Forest: An ensemble of decision trees, Random Forest, improves model accuracy and robustness by averaging multiple trees' predictions.
- Logistic Regression: This model offers a statistical approach to predicting binary outcomes, providing a straightforward interpretation of how each feature (e.g., variety, winery) influences the likelihood of a wine being classified as premium.

Description Features Models

- Dummy Classifier (Base Model): As with the Major Features Dataset, the Dummy Classifier here establishes a baseline performance for models analyzing text data from wine descriptions, ensuring that subsequent models are genuinely insightful.
- Multinomial Naive Bayes: This model is particularly suited for text classification problems due to its assumption about the independence of features in a document. It's effective for analyzing the frequency of keywords in wine descriptions that correlate with premium quality, making it a natural choice for text-based insights.
- Random Forest: This method can capture the importance of various words and phrases in predicting wine quality, benefiting from Random Forest's ability to manage high-dimensional data.
- Logistic Regression: When applied to text data, Logistic Regression can quantify the impact of specific keywords on the classification as premium or non-premium wines. This model is useful for identifying which terms are most strongly associated with premium wines, providing clear guidance for marketing and product description strategies.

Feature pre-processing

In our analysis, we will employ specific data preprocessing techniques and analytical methods for handling categorical features and text data:
For categorical features in the Major Features(e.g., country, province, variety, winery), we will use imputation to fill in missing values and one-hot encoding to convert categories into a numerical format that can be efficiently processed by machine learning models.
For text analytics of the wine descriptions, we will utilize the TF-IDF vectorizer to transform the text into a numerical representation that highlights the importance of words in relation to their frequency in the document and across all documents.
These methods are integral to our approach, ensuring that our models can accurately interpret and analyze the dataset's diverse information types.
Please refer to the Feature Engineering section for feature selections and other details.

Hyperparameter tuning

In tuning our models, we employed the RandomSearchCV method to systematically explore a wide range of hyperparameters and identify the optimal settings for each model. Please refer to the code for more details.

Reflection on selected performance metrics and Model selection

Please refer to the appendices for the performance metrics output.

Major Features Classification Models

- Accuracy: Logistic Regression achieves the highest accuracy of 0.7521 among all models tested, indicating it correctly predicts wine quality more often than the other models.
- Balance between Precision and Recall: The Logistic Regression model shows a strong balance between precision and recall for both classes (0 and 1), with precision values of

0.76 and 0.74, and recall values of 0.77 and 0.73, respectively. This balance is crucial in scenarios where both false positives and false negatives are of concern.

- F1-Score: The f1-scores for Logistic Regression are the highest among the models for both classes, indicating a robust balance between precision and recall. This suggests that Logistic Regression not only makes accurate predictions but does so in a balanced manner across both classes.
- Comparative Performance: When compared to the other models, Logistic Regression consistently outperforms the Decision Tree, Random Forest, and certainly the Dummy Classifier across all key metrics (accuracy, precision, recall, f1-score). The Decision Tree and Random Forest models show decent performance but are not as balanced or as high performing as Logistic Regression.

Given these points, the Logistic Regression model is the best choice for predicting wine quality based on the dataset provided. It offers the highest accuracy and the best balance between identifying positive cases and minimizing false positives, making it the most reliable model among those tested for this specific task.

Description Classification Models

- Highest Accuracy: Logistic Regression has the highest accuracy score, indicating it makes the most correct predictions for wine quality out of all models tested.
- Balanced Precision and Recall: It offers a balanced performance between precision and recall for both classes, with precision at 0.83 for class 0 and 0.82 for class 1, alongside recall at 0.84 for class 0 and 0.81 for class 1. This balance is crucial for making reliable predictions in both classes without bias toward predicting one class more accurately than the other.
- Strong F1-Score: The F1-scores, which are the harmonic mean of precision and recall, are also highest for the Logistic Regression model, indicating a strong balance between the precision and recall across both classes. This suggests that Logistic Regression not only accurately identifies positive cases but does so with a minimal number of false positives and negatives.
- Comparatively, while Naive Bayes shows commendable performance with an accuracy of 0.8050, it slightly lags behind Logistic Regression in both accuracy and the balance of precision and recall across classes. Random Forest has a lower accuracy of 0.7458 and demonstrates a significant imbalance between precision and recall, especially for class 1, indicating a tendency to misclassify a higher proportion of actual positives compared to Logistic Regression and Naive Bayes.
- The Dummy Classifier is intended as a baseline with an accuracy close to random guessing, as expected, and is significantly outperformed by the other models.

Given these insights, Logistic Regression is the most suitable model for predicting wine quality in this context due to its superior accuracy, balanced precision and recall, and strong F1-scores, making it the most reliable and balanced model among those evaluated.

Business Insights

Major Features

Utilizing the output from the permutation importance analysis in the context of designing premium wines involves focusing on the features that have the most significant impact on wine quality as identified by the logistic regression model.

1. Leverage the Winery's Reputation

- Focus on Branding: With "winery" having the highest importance score, the reputation and branding of the winery are crucial. Develop a strong brand story that emphasizes quality, craftsmanship, and the unique aspects of your winery that contribute to producing premium wines.
- Quality Consistency: Ensure that the winery maintains high standards in wine production to strengthen and uphold its reputation as a marker of premium quality.

2. Highlight the Geography

- Province and Region: The importance of "province" and "region_1" suggests that the geographic origin of the wine significantly affects its perceived quality. Highlight the unique characteristics of the province and region in your marketing, focusing on terroir, climate, and how these factors contribute to the distinctive qualities of your wines.
- Country of Origin: With "country" also being a significant factor, leverage the country's image in the wine market. For example, certain countries are renowned for their wine quality and specific varieties. Use this to your advantage in storytelling and branding.

Description Keywords

The output description mining model identifies keywords highly associated with the score of wines, provides valuable insights into the characteristics that are often mentioned in descriptions of higher-quality (premium) wines. To apply these keywords in designing premium wines, we may use the following strategies:

1. Focus on Quality and Complexity

- Complexity and Depth: Terms like "complex," "layered," "depth," and "complexity" suggest that premium wines are appreciated for their multifaceted flavors and aromas. Aim to develop wines that offer a rich array of sensory experiences.
- Balance and Structure: Keywords such as "balanced" and "structured" indicate the importance of a harmonious blend of acidity, tannins, sweetness, and alcohol. Ensure your winemaking process focuses on achieving a well-rounded wine.

2. Aim for Desirable Taste Profiles and Aromas

- Richness and Opulence: Descriptors like "rich," "opulent," "lush," and "velvety" point to a preference for wines with a full-bodied feel and luxurious taste. Consider grape varieties and fermentation processes that enhance these attributes.

- Elegance and Refinement: Words such as "elegant," "refined," "polished," and "lovely" suggest a market appreciation for wines that are sophisticated and gracefully composed. Strive for subtlety and finesse in flavor profiles.

3. Highlight Specific Vintage and Aging Potential

- Vintage and Drinkability: The presence of specific years (e.g., 2020, 2021, 2022) and phrases like "drink 2020" imply the importance of vintage quality and optimal drinking windows. Highlight the best vintages and advise on aging potential to guide consumers.
- Aging and Cellaring: "Years," "cellar," "aged," and future years (e.g., 2025) suggest consumers value wines that age well. Focus on producing wines that will develop positively over time, offering guidance on cellaring practices.

4. Emphasize Terroir and Winemaking Excellence

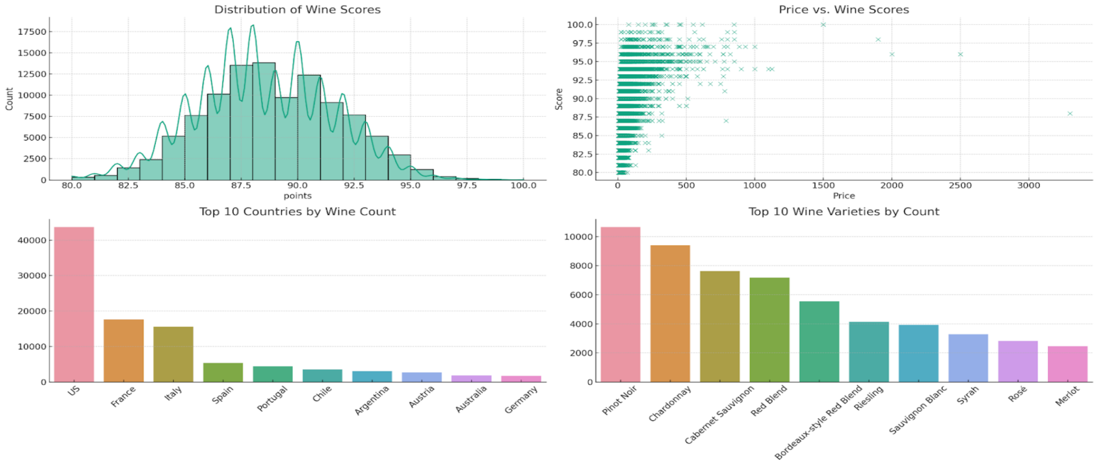
- Vineyard and Terroir: "Vineyard," "minerality," and "terroir-focused" indicate the value placed on the origin and unique characteristics imparted by specific locations. Promote the unique aspects of your vineyard's terroir in your wine's branding.
- Winemaking Techniques: Use of "barrel sample" and specific tasting notes (e.g., "minerality," "elegance") can reflect the winemaking process's influence on the wine's character. Adopt and highlight winemaking techniques that contribute to the desired profiles.

5. Marketing and Labeling

- Descriptive Marketing: Incorporate these keywords into your wine labels, marketing materials, and tasting notes. Describing your wine using terms found to be associated with higher scores can attract consumers looking for premium wine experiences.
- Storytelling: Build narratives around your wines that include these descriptors, focusing on the craftsmanship, vineyard heritage, and the sensory journey the wine offers.

Appendix 1

Exploratory data analysis (distribution of Wine Scores, Price vs. Wine scores, Top 10 Countries by Wine Count, and Top 10 Wine Varieties by count).



Appendix 2: Performance Metrics

Major Feature Models:

```
Dummy Classifier Score: 0.5031736872475476
Decision Tree Score: 0.6889401808040008
Random Forest Score: 0.6722831313714176
Logistic Regression Score: 0.7520677053279476
Dummy Classifier Report:
              precision    recall  f1-score   support

           0       0.53      0.53      0.53     13697
           1       0.47      0.47      0.47     12298

    accuracy                           0.50     25995
   macro avg       0.50      0.50      0.50     25995
weighted avg       0.50      0.50      0.50     25995

Decision Tree Report:
              precision    recall  f1-score   support

           0       0.68      0.78      0.72     13697
           1       0.70      0.59      0.64     12298

    accuracy                           0.69     25995
   macro avg       0.69      0.68      0.68     25995
weighted avg       0.69      0.69      0.69     25995

Random Forest Report:
              precision    recall  f1-score   support

           0       0.66      0.78      0.71     13697
           1       0.69      0.55      0.62     12298

    accuracy                           0.67     25995
   macro avg       0.68      0.67      0.66     25995
weighted avg       0.68      0.67      0.67     25995

Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.76      0.77      0.77     13697
           1       0.74      0.73      0.74     12298

    accuracy                           0.75     25995
   macro avg       0.75      0.75      0.75     25995
weighted avg       0.75      0.75      0.75     25995
```

Description Models:

```
Scores:
Dummy Classifier: 0.49474899019042123
Naive Bayes: 0.8050009617234083
Logistic Regression: 0.82488940180804
Random Forest: 0.7457587997691864

Reports:
Dummy Classifier:
              precision    recall  f1-score   support
```

```
           0       0.53      0.53      0.53     13697
           1       0.48      0.47      0.48     12298

    accuracy                           0.51     25995
   macro avg       0.50      0.50      0.50     25995
weighted avg       0.51      0.51      0.51     25995

Naive Bayes:
              precision    recall  f1-score   support

           0       0.81      0.82      0.82     13697
           1       0.80      0.79      0.79     12298

    accuracy                           0.81     25995
   macro avg       0.80      0.80      0.80     25995
weighted avg       0.80      0.81      0.80     25995

Logistic Regression:
              precision    recall  f1-score   support

           0       0.83      0.84      0.84     13697
           1       0.82      0.81      0.81     12298

    accuracy                           0.82     25995
   macro avg       0.82      0.82      0.82     25995
weighted avg       0.82      0.82      0.82     25995

Random Forest:
              precision    recall  f1-score   support

           0       0.70      0.92      0.79     13697
           1       0.86      0.55      0.67     12298

    accuracy                           0.75     25995
   macro avg       0.78      0.74      0.73     25995
weighted avg       0.77      0.75      0.74     25995
```

Appenix 3 Top 50 Keywords

complex, long, years, delicious, 2020, elegant, rich, impressive, vineyard, lovely, beautiful, 2022, beautifully, concentrated, 2025, great, 2021, excellent, velvety, powerful, pure, drink 2020, fine, 2018, 2019, opulent, 2023, lush, deep, balanced, minerality, focused, cellar, layered, gorgeous, richness, drink 2022, elegance, structured, refined, drink 2021, depth, intense, age, polished, 2024, best, mineral, winemaker, balance.