Neil Byers and [Name Removed]
Dr. Jaime Arguello
INLS 613
December 6, 2019

Predicting Wine Scores Based on Review Content
Final Report

Introduction

Product reviews have come to play a large role in modern consumer culture. Nowadays, a large corpus of written opinions and numeric ratings can be found for virtually any product. This is certainly true in the alcoholic beverage industry, where entire magazines and online communities exist specifically for this purpose. Whether or not 'experts' rate a certain beer, wine, or liquor highly is a factor that influences both producers and consumers of a given beverage. For this project, we attempt to pick apart this phenomenon using a large set of wine reviews scraped from Wine Enthusiast's website, *winemag.com*, a site where experts rate and describe hundreds of thousands of different wines from around the world.

Specifically, we hope to use text mining techniques to determine how well we can predict the score given to a specific wine by its reviewer based on the contents of the review itself. Doing so relates to real-world concerns in a number of ways. First, consumers use wine reviews to help decide which wines they should purchase or seek out. Assessing whether or not certain characteristics of a review can help predict a wine's score can inform consumers as to how seriously they should take these reviews when making purchasing decisions. Secondly, text-feature analysis can help wine producers understand what flavors and characteristics are associated with or predictive of higher scores. If we were able to determine, for example, that certain flavor descriptors are associated with higher scores, it would help wine producers to tailor their product to achieve higher scores and, presumably, more business.

Finally, this project serves to assess the credibility of the reviewers themselves. Can we really take seriously what reviewers have to say about their wines? Do the text features of their

reviews have any real bearing on the scores assigned to the wines? We assume that more expensive wines are rated more highly than cheaper wines. Similarly, it might be possible that certain varieties of wine or wines from certain regions are viewed more favorably than others. Do these factors play any role in determining a wine's score, as opposed to the words a reviewer uses to describe that wine? Should consumers and producers of wine even read these reviews? Or can they safely guess a wine's probable score based solely on its price and origins? This project seeks to address, at least in part, all of these questions.

To do so we have reduced the problem to a number of specific research questions. Generally, we ask: Can we use the features of a wine review to predict the score given to that wine by its reviewer? In order to better answer this question, we split our project into two approaches. First, we use different feature representations to answer the following question: Do features external to a wine's description (price, variety, province, etc.) affect our ability to predict a wine's score when compared to a purely description-based approach? This approach will allow us to examine how useful a wine's actual description is when paired with or tested against other features present in the review. We initially feared that the reviewers might be influenced factors like price and province when reviewing a wine, meaning that the contents of the description would not be independent from the other attributes of the review. If this were the case, it would significantly hinder our ability to draw conclusions about the predictive performance of the descriptions as opposed to other attributes. However, because winemag.com follows a strictly blind tasting and review policy, we can safely assume independence.[1]

Secondly, we address the issue of credibility by examining the text of descriptions between different varieties of wine. Presumably, different varieties of wine lend themselves to noticeably different descriptions. This should be true especially when comparing say, a dry red

[1] https://www.winemag.com/2019/10/24/blind-wine-tasting-rate-fairly/

wine like a Cabernet Sauvignon and a sweeter white wine like a Riesling. If a model built solely using the text content of descriptions for a single wine variety can prove just as predictive for wines in general as a model built using a wide variety of wine varieties, it will imply that reviewers do not clearly identify 'good' and 'bad' descriptors specific to certain varieties of wine. In order to ascertain whether or not this is the case, we ask: "Are models built using single varieties of wine better or worse at predicting scores for wines in general than models built using a random sample of wines from many different varieties?"

Our hope is that this project can use text-mining and machine learning to shed light on real-world concerns. Not only do we hope to help consumers and producers of wine make better-informed decisions, but also to shed light on flavor descriptions of wine in general. Can we use machines to identify counter-intuitive flavor characteristics? Do reviewers focus on or appreciate certain characteristics that don't agree with common perceptions? In the food and beverage industry, people's perceptions of products are just as important as the processes and ingredients used to produce them. Using these technologies will help us gain a greater understanding of said perceptions.

Related Work

Previous studies have explored various approaches for using machine learning in applications related to wine. Several of the scholarly articles (Gupta 2018; Aich, et al 2018; Aich, et al 2019) we examined attempt to predict the quality rating of wine using a feature set of physio-chemical attributes, such as alcohol content, total phenol count, alkalinity, color intensity, etc. These studies rely on the publicly available UCI dataset which includes such features for two different varieties of Italian Wine. These articles use a wide range of techniques and algorithms (J48 Trees, SVMs, LDA, Random Forest, etc.) to predict the numeric score of wines and achieve accuracy ratings as high as 98%. Another source (Yeo, et al 2015) attempts utilize

historical price data along with clustering and regression algorithms to predict wine price trends. Though they can give us some perspective, the predictive tasks of these studies differ substantially from that associated with our project. While physio-chemical properties and price trends are important aspects of the wine domain, they do not address the human aspect of said domain. People's perception of wine can be just as important of a factor as these other characteristics, and our project attempts to address this. As such, our dataset presents an opportunity to fill a significant gap in our understanding of the wine as a consumer product.

In order to gain a stronger foundation for our particular task, we started with several more general studies from outside of our domain. Fazzolari 2017 evaluates the performance of models built using several algorithms (SVMs, Naive Bayes, PART, and J48) for text-score disagreement in hotel reviews. This study not only points to SVMs as being the most effective learners, but also indicates the difficulty of distinguishing scores from reviews that fall in the middle of the possible score range. Given our heavy concentration of instances around our median score, this problem could particularly affect our dataset. Finally, while many studies indicate that more sophisticated algorithms can be particularly successful (Fazzolari 2017; Vinodhini and Chandresekaran 2012), we elected to perform our analysis using the Naive Bayes algorithm. The simplicity, computational efficiency, and proven baseline effectiveness of Naive Bayes was deemed adequate for our purposes. We admit, however, that different algorithms might achieve better performance should they be applied in our context.

While we could not locate any academic sources that utilized our specific dataset, we did find several student projects and blogs that attempt analyses using the same dataset provided by Kaggle. These different individuals all attempted to solve a problem similar to our own: implementing machine learning techniques to predict the a wine's score. These other projects used as a target variable the wine's score, which can theoretically range from 0 to 100. Just as we hoped to do with our models, they attempt to predict a wine's score using various feature combinations provided in the dataset. Rather than binarizing the target variable into two

classification groups ('high' and 'low') as we did for this project, these individuals grouped the instances into multiple classification groups based on point ranges. Having done so, they work to build models that predict which point range a particular instance might fall into.

These sources used various machine learning techniques, including Recurrent Neural Network (Gu 2019) and Random Forest (Bolstad 2018), and evaluating their models using simple accuracy measures. The neural network implemented (Gu 2019) combed through the content of the reviews, treating each as a sequence of vectors, then trained and tested the model on the vocabulary from the reviews (accuracy – 64%). Bolstad implemented Random Forest Decision Trees by taking into account non-text features like country, province, region, variety, year, and price and achieved an accuracy of 59% (Bolstad 2018). Another student group utilized linear regression to predict score classification ranges (Xu, Wang 2017) and calculate mean error rates for evaluation purposes.

Looking at how other students have applied machine learning techniques on the same dataset provided us with concerning about how we might go about designing our models to address our problem definition. While we decided to use a different classification algorithm from those used in the projects mentioned above, our classification task remained broadly similar.

Data and Data Preparation

Our data, collected from Kaggle, was scraped from winemag.com in June 2017.[2] The original set consists of 130,000 instances of wine reviews, each of which included attributes such as description, title, country of origin, points, province of origin, variety, price, etc. The data also included information about the individual reviewers, such as name and Twitter handle. To prepare our data, we first removed all reviewer-specific information, as this was not relevant for our purposes. We then split our data into two binary class values based on the score ("points")

2 https://www.kaggle.com/zynicide/wine-reviews

achieved by each wine. We split the data at the median score (88 points), with the “low” category including the median. In order to reduce computational requirements, we randomly sampled a 10,000 instance training set and a 500 instance test set. For distribution graphs by attribute, please see the Appendix.

Several considerations might affect the conclusions we can draw from this data. For one, winemag.com may not be representative of all wine reviews. There may be linguistic or cultural phenomena associated with this particular organization which affect our ability to generalize beyond our dataset. Furthermore, the reviews are not representative of all wines. Almost half of our dataset consists of wines from the US (~54,000 instances), while the vast majority of these wines were made in California (~36,000 instances). Other attributes have similarly skewed distributions. The wine reviews in our set also rate wines on a scale of 80-100. We were unable to determine whether this meant that the website only reviews the best wines, or that 80 points is the minimum score for all wines reviews. As such, it is possible that our dataset might only be representative of highly-rated wines as opposed to all wines.

Approaches

In order to go about answering our research question we followed two general approaches. The first of which was to determine how well we could predict a wine’s score (high or low) using just the text features of a wine’s description in comparison with using other features like price, variety, region, and country of origin. To do so we first extracted a random sample of 10,000 wine reviews from the original dataset to use as training data, as well as a separate random sample of 500 reviews to use as a test set[3] We then used the training data to extract a basic feature representation using the textual features of the descriptions alone. We

[3] We sampled using Python’s built-in random.sample method. A random sample of 10,500 instances was compiled using this method. We then separated the data into training and test sets using Microsoft Excel.

used a simple Naive Bayes algorithm to build a model and evaluated using our test data as well as 10-fold cross validation.

Having built and evaluated our baseline text-based model, we performed a similar operation using the non-description features mentioned above (price, variety, country, province). We first built and evaluated a model for each individual feature. In other words, we used feature representations in which price or variety, for example, was the only feature. We then did so using a feature representation that included all four non-description features. This allowed us to see how well a model might do using each feature in isolation as well as all of these features in unison. We could then use our results from this process to compare models built using description-derived text features and models built using non-description features. Could we achieve similar or better performance using one or the other?

Having established how well description-based features and non-description features performed separately from one another, we then used feature ablation to determine whether or not the addition of certain non-description features to our text-derived representations would improve model performance. To do so we built six further feature representations: one each with the text features plus price, variety, country, and region; one with the text features plus price and region,[4] and one with the text features plus all four non-description features. Models were then built and evaluated using each of the six feature representations.

Here we ran into an important obstacle that somewhat hindered our ability to draw meaningful conclusions from these tests. When using the test dataset to evaluate models built using feature representations which included non-text-based features ('Column Features' in LightSIDE), we consistently generated an error in LightSIDE. It became clear that LightSIDE would produce an error every time it came across values for the non-description features in the test set that were not included in the training set. For example, when LightSIDE encountered a

[4] Having achieved more favorable results with these two features in WEKA and LightSIDE than with the others, we thought it useful to build a separate model with these two features and not variety or country.

certain value for the ‘variety’ column in the test set that was not present in the training set (i.e. ‘Manzoni’), it immediately stopped the evaluation procedure. For a screenshot of the error message, please see the Appendix. We could not find a way to avoid this error except to evaluate our models using 10-fold cross validation instead of using the test set. While this established a useful baseline for each feature representation, it barred our ability to compare performance results using never-before-seen instances.

In any case, through this approach we built models using only the text features as well as models using only the non-text features. We then combined these features to determine how well these features might perform when combined. Doing so allowed us to address the first half of our research question: Do features external to a wine’s description (price, variety, province, etc.) affect our ability to predict a wine’s score when compared to a purely description-based approach?

The second approach used was to build, evaluate, and compare models using only instances of specific wine varieties. Here we wanted to test how well models built using description-only text features from single-variety training sets would perform when evaluated using a test set that includes instances from a wide range of wine varieties. Intuitively it seems that certain text features would be specific to certain varieties and the flavors associated with them. As a result, we wanted to determine if a model’s ability to generalize would suffer when built using only a single variety.

To do so, we used WEKA to determine the most heavily-represented varieties in our raw dataset. These were:

1. Pinot Noir (13272 instances)
2. Chardonnay (11753 instances)
3. Cabernet Sauvignon (9472 instances)
4. Red Blend (8946 instances)
5. Bordeaux-style Red Blend (6915 instances)
6. Riesling (5189 instances)
7. Sauvignon Blanc (4967 instances)
8. Syrah (4142 instances)
9. Rosé (3564 instances)

10. Merlot (3102 instances)

Out of these, we chose to include 1, 2, 3, 5, 6, and 9 in our tests because these represented an adequately wide flavor range between styles. We built models using the description-derived text features as above. We then evaluated each of these models using 10-fold cross validation as well as our 500-instance sample randomly collected across all varieties. For comparison's sake we included a similar model built using the 10,000-instance randomly selected training set used in the previous section. This model was also evaluated using cross validation and the separate training set. While no shared instances existed between the randomly selected training and test sets, it is likely that instances present in the latter were also present in some of our variety-specific training sets. As a result, the performance evaluation of these models on the test set is probably somewhat inflated.

Using these two approaches allowed us to evaluate the credibility of our reviewers from two perspectives. First, we looked at whether or not reviewers' score assignments could be predicted using just a wine's description, just the non-description features, or both in unison. Here, high performance with just the text features would indicate a higher degree of credibility. Secondly, we tried to determine how well models built using specific varieties would generalize to non-variety-specific test instances. Here, lower performance likely would indicate higher credibility because it would imply that reviewers describe drastically different wine varieties using vocabularies specific to those varieties. In other words, it would be harder to take the words of reviewers seriously if the good and bad words used to describe a Riesling highly similar to those used to describe a Cabernet Sauvignon.

Evaluation

We chose accuracy as our main evaluation metric. This decision resulted from several considerations. First of all, our data, having been split on the median into separate classes, was more or less evenly split between ‘high’ and ‘low’ scores (~47% high vs. ~53% low). As a result, we could not achieve a high accuracy rating simply by choosing one class value over another as a rule. Another consideration was that neither class was more important to us than the other. Building models that can accurately predict low scores was deemed to be just as important from a domain perspective as building those that can predict high scores. Finally, we chose accuracy for simplicity’s sake. Our analysis attempts to compare a wide range of models and feature representations, meaning that having a single number for each iteration can be very helpful. We admit that deeper analysis of precision and recall in regards to highs or lows for each of our models would be useful. Knowing that certain non-description features or certain varieties lead to a large number of false ‘lows’ or ‘highs’ would be helpful for determining how these attributes relate to a wine’s score. While we deemed such an analysis to be beyond the scope of our project, it would certainly prove to be an interesting avenue for future studies.

Results - Approach 1

As part of our first approach we examined the performance of description-derived text feature representations compared against as well as paired with the reviews’ non-description features (price, country, province, and variety). These results were produced using a 10,000-instance randomly sampled training set and, when possible, a 500-instance randomly sampled test set. The first two tables helped establish our ‘baseline’ accuracy ratings for each category of feature. The third table resulted from iterative feature ablation of our best description-only feature set along with several non-description features.

Description Only: Accuracy ratings for different description-derived text feature representations

	Unigrams	**Previous + Skip Stopwords, Track Hit Location**	Previous + Word/POS pairs	Previous + Bigrams	Previous + Stemming
10-Fold Cross Validation	79.45%	**79.6%**	79.76%	79.97%	79.97%
Test Set Accuracy	79.76%	**81.36%**	80.96%	81.16%	80.76%

Non-Description Attributes: Accuracy ratings and information gain for non-description-attribute feature representations

	Price	Country	Province	Variety	All
10-Fold Cross Validation Accuracy	64.75%	55.99%	59.19%	59.60%	**68.35%**
Information Gain	**.1856**	.0263	.0631	.0618	------

Feature Ablation: Accuracy ratings for description-derived text features with their non-description counterparts

	Just Text	+Price	+Country	+Province	+Variety	+Province & Price	**+All**
10-Fold Cross Validation	79.6%	79.76%	79.66%	80.07%	79.42%	80.29%	**80.34%**

Results - Approach 2

As part of our second approach we compared models built using the descriptions of single varieties of wine. The first two tables show the top seven most positively and negatively correlated features for each variety. The third table compares the performance of models built using single varieties with our baseline model built using instances randomly sampled from all varieties. We evaluated each model using the 500-instance randomly selected test set (“Test”)

as well as with 10-Fold cross validation (CV). The number of instances for each variety has been provided as well.

Variety-Specific feature representations: positively correlated features

General	Bordeaux-Style Red Blend	Cabernet Sauvignon	Chardonnay	Pinot Noir	Riesling	Rose
Years	Barrel	Through	Years	Vineyard	Long	Rich
Through	Sample	Years	Vineyard	Black	Freshness	Age
Rich	Cabernet	Cellar	Minerality	Years	Yet	Concentrated
Concentrated	Dark	Black	Wine	Complex	Ripe	Complex
Black	Rich	2020	Age	Through	Concentrated	Estate
Complex	Great	Vineyard	Rich	Rich	Through	Impressive
2020	Sauvignon	Wine	Complex	2020	Rich	Mineral

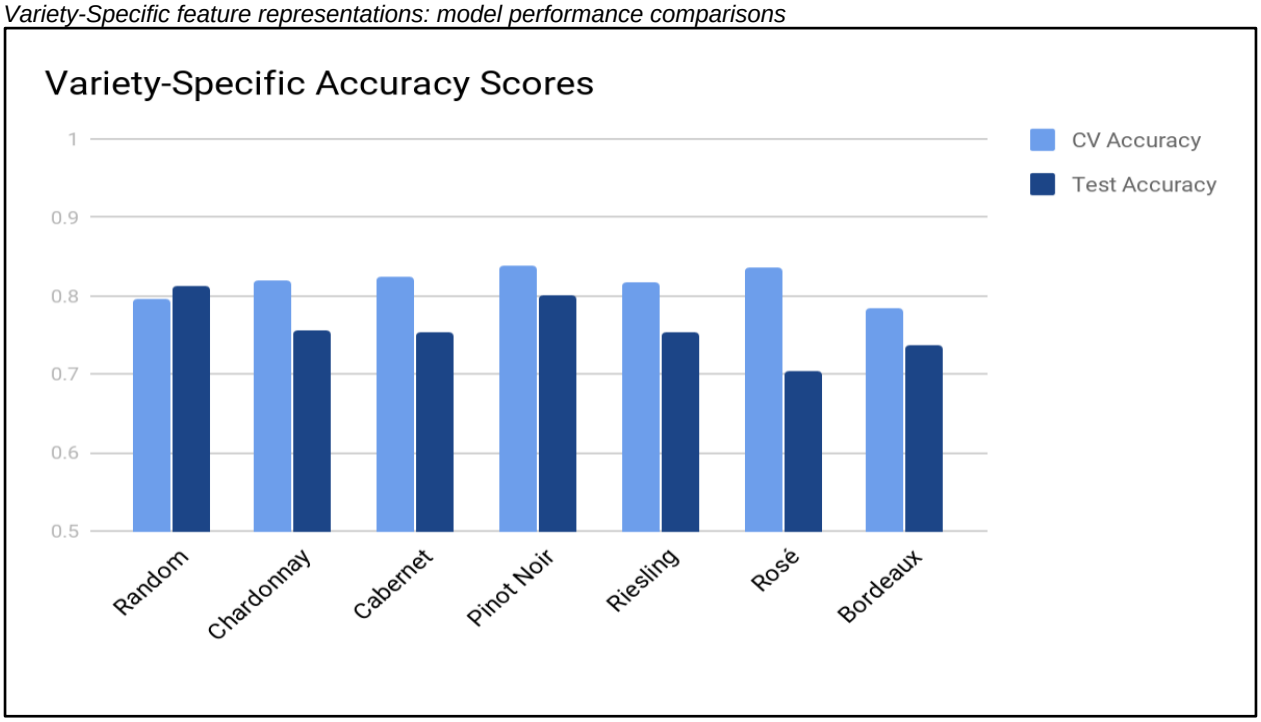
Variety-Specific feature representations: positively correlated features

General	Bordeaux-Style Red Blend	Cabernet Sauvignon	Chardonnay	Pinot Noir	Riesling	Rose
Simple	Light	Simple	Simple	Simple	Off-dry	Soft
Flavors	Fruity	Berry	Sweet	Pinot	Pressed	Sweet
Easy	Stalky	Herbal	Flavors	Flavors	Straightforward	Caramel
Soft	Dry	Green	Soft	Thin	Apple	Easy
Light	Attractive	Aromas	Tastes	Little	Bit	Now
Fruity	Lean	Bit	Feels	Pleasant	Simple	Fruity
Bit	2016	Flavors	Oaky	Bit	Slightly	Attractive

Variety-Specific feature representations: model performance comparisons

	Accuracy, Kappa (CV)	Accuracy, Kappa (Test)	Instances
Random Training Sample	.7960, .5899	.8136, .6277	10000
Bordeaux-Style Red Blend	.7837, .5654	.7375, .475	6915
Cabernet Sauvignon	.8237, .6473	.7535, .5084	9472
Chardonnay	.8202, .6397	.7555, .5199	11753
Pinot Noir	.8394, .6669	.7996, .5994	13272
Riesling	.8160, .5629	.7535, .5082	5189
Rosé	.8367, .5899	.7034, .4088	3564

Variety-Specific feature representations: model performance comparisons



Discussion

Based on the results from the models that we built, we determined that the non-text features of the wines did little to make our models more predictive. On their own, the non-description features performed poorly when compared with the baseline description features

(68.35% Cross-fold accuracy vs. 79.6% Cross-fold accuracy). While this seems to imply that descriptions are more predictive than all of the other features working together, the next step was to evaluate how well these separate features performed together. Did the non-description features significantly improve the performance of our baseline model? The answer was no. Having built models that incorporated additional features such as country of origin and price, we found that there was little improvement in the accuracy we achieved with our baseline model built using only description features (79.6% baseline vs. 80.34% with added features). Even with different combinations of these different non-text features, it was difficult to improve our baseline model. It is impossible for us to say how well these models would perform with previously unseen test instances due to the error we encountered in LightSIDE. That being said, cross-validation still allows to establish ball-park performance levels using each of these variables. These results seem to indicate that a reviewer's impressions of a wine are more indicative of a wine's score than just the price, variety, or area of origin alone. Indeed, these added features do little to make the description features any more predictive.

Our second approach led to a number of interesting results. First, looking at the most positively and negatively correlated unigrams across varieties, we can see that many of the features which best distinguish 'high' wines from 'low' wines are common to many varieties (high: "rich," "complex"; low: "simple," "light," "soft"). Such features were also found to be strongly correlated with "high" or "low" wines in the random sample training set. That being said, there are also many unigrams in these feature representations that are unique to one or a few varieties, such as "minerality" (high) for Chardonnay and "herbal" (low) for Cabernet Sauvignon. What can we take from these results? It seems that while reviewers are discerning enough to assign associate specific words with "good" Chardonnays and "bad" Cabernet Sauvignons, for example, they still look for relatively vague and non-descriptive characteristics across all wines. In other words, reviewers seem to have some idea of what they like and don't like about

different varieties of wine, but much of their make-or-break characteristics seem to generalize across all varieties.

This observation is supported by the performance of the single variety models in comparison with the random training sample model. In general, the models performed more or less in line with our expectations. The random sample model did not perform as well as the single variety models when evaluating using cross-validation. This seems intuitive, because a model built using Pinot Noir-specific features would presumably perform better at predicting good or bad Pinot Noirs (83.94%) than a model built using a random sample would perform predicting good or bad wines in general (79.60%). The variety-specific model would benefit from the added specificity of the features described above. Similarly, the random-sample model would presumably exhibit better generalization performance across all varieties than models built using specific varieties. This expectation was borne out by the fact that the random sample model performed much better when evaluated using the random test set (81.36%) than did models built using specific varieties (ranging from 70.34% to 79.96%). These results seem in line with the lessons taken from the positively and negatively correlated features. The improved (though not drastically) cross-validation performance of the variety-specific models and the improved (though not always drastically) test-set performance of the random sample set support a situation in which reviewers can agree on certain variety-specific flavor characteristics but still use much of the same vocabulary to distinguish good wines from bad across varieties. In other words, our variety-specific models do better when predicting instances of their own variety, but not that much better. They do worse when predicting instances sampled across all varieties, but not always that much worse.

We also hoped to discuss possible limitations to our project, as these affect our ability to interpret our results. We must consider the possibility that the original source of our data, reviews from *winemag.com*, may not be representative of all wines. We noticed when looking at the distributions of our dataset that there seemed to be an overrepresentation of wines from the

United States as well and an overrepresentation of certain varieties of wines. If our data cannot be said to representative all wines, then our models may not be as effective performing classification tasks on a corpus that is more representative. We also considered the idea that the reviewers from our dataset may not be representative of all wine reviewers. These reviewers could use different terms and vocabulary than other reviewers, meaning that the features that we extracted here may differ substantially from those extracted from other reviews. Such differences would likely also affect our ability to make predictions outside of our own dataset.

Another possible limitation are certain flaws in the data itself. For one, we found many missing values amongst our instances. This is to be expected with large datasets, as not all instances will come with values for all available attributes. Other limitations arose from our data preparation process. For one, we used random sampling to select our non-variety-specific training and test sets. Doing so always comes with a risk that the subset may not have the same distribution as the original dataset. We recognize the possibility of a difference in distribution between our random sample and the overall dataset and admit that such a difference would mean different feature representations and potentially different model performance. Another consideration is that many of the instances in our dataset are clustered around the median score value of 88 (see Appendix for distribution graph). As such, most of the instances in the "high" class were not very far removed from most of the instances in the "low" class. We might have achieved better results had we only sampled instances from the extreme ends of our score range.

Conclusions

From our models we conclude that the actual written descriptions or wines provided in our dataset do matter when trying to predict wine scores. We were able to achieve higher accuracies for models whose features included textual elements of the written description. How does this relate to our original research question? Combined with the knowledge that all of

these reviews were made using blind taste tests, our results indicate that we should indeed strongly consider the written impressions of a reviewer over the other features present in a review. This can mean that when customers go to purchase wine, they can comfortably take into account a wine's description as opposed to just focusing on price or place of origin. Customers can also determine higher quality wines by looking at wine reviews with terms found in our feature selection that are found to correlate with higher scoring wines. Wine producers can also take this into account when they are making and improving their wines. One use, for example, would be examine which flavor description terms reviewers used in high-scoring wines and attempt to emulate or recreate these flavor profiles in their own product. Customers and producers alike can also take this conclusion to mean that, for simplicity's sake, wine reviewers do write descriptions that align with their overall impressions. We also determined that the addition of the other non-text features did improve our accuracy, though only slightly. While it is more useful to actually read a reviewer's description, consumers and producers might still benefit from doing so along with consideration of other features. For those hoping to determine which non-text features might be most informative, we also determined through information gain calculations which attributes would be more helpful in making accurate predictions. This might give individuals some direction as to which characteristics they should consider over others.

In regards to our second approach, we established that, though differences separate feature representations built using single varieties, these differences are not as salient as one might expect. Reviewers used specific words to describe certain wines as good or bad, but much of their vocabulary translated across different varieties (rich, complex, soft, light, simple, etc.). In other words, what they like and don't like in different wines seems to be balanced by what they like and don't like across all wines. This observation was supported by the cross validation and test-set performance of single-variety models vs. the random sample model. What are the ramifications of this? Wine producers should certainly look to the variety-specific features when deciding what flavors to aim for, but should keep in mind the more generally

praised characteristics as well. Similarly, consumers looking for quality examples of specific varieties will have certain words to look for while those with less specific tastes can benefit from features that here seem more or less “good” or “bad” across all varieties.

In regards to credibility, our study paints a favorable picture for wine reviewers. Their descriptions performed better as features than did features such as wine’s price or place of origin. Similarly, we detected noticeable differences in the performances of models built using their descriptions of single varieties over the performance of general sample. This seems to indicate a high degree of credibility. These reviewers can blindly taste wines and write descriptions that do better at predicting whether or not these wines will have high or low scores than do a wine’s other characteristics. They also exhibit reasonable degrees of consistency describing the favorable and unfavorable characteristics of specific varieties of wine. While these distinctions are not as drastic as one might expect, they still indicate a high amount of expertise on the part of the reviewers. Overall, we conclude that, at least for our specific and narrow dataset, it benefits both consumers and producers of wine to pay attention to the written aspects of wine reviews. Their writers seem to “know what they’re talking about” in a way that captures quality more fully than the raw attributes of a given wine.

References

Aich, S., Al-Absi, A.A., Hui, K.L., Lee, J.T., Sain, M. (2018) A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques, in *International Conference on Advanced Communication Technology*, pp. 139–143. doi: https://doi.org/10.23919/ICACT.2018.8323674

Aich, S., Al-Absi, A.A., Hui, K.L., Sain, M. (2019) Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques, in *International Conference on Advanced Communication Technology,* pp. 1122-1127. doi: https://doi.org/10.23919/ICACT.2019.8702017

Bolstad, L. E. (2018). Predicting Wine Ratings Using Machine Learning. Retrieved from Medium website: https://medium.com/@larserikbolstad/predicting-wine-ratings-using-machine-learning-aa64167d25b9

Fazolari, M., et. al. (2017). "A study on text-score disagreement in online reviews." *Cognitive Computation* 9(5), 689-701. https://doi.org/10.1007/s12559-017-9496-y

Gu, M. (2019). Predicting Wine Quality using Text Reviews. Retrieved from Towards Data Science website: https://towardsdatascience.com/predicting-wine-quality-using-text-reviews-8bddaeb5285d

Gupta, Y. (2018). "Selection of important features and predicting wine quality using machine learning techniques." *Procedia Computer Science* 125, 305-312. https://doi.org/10.1016/j.procs.2017.12.041

Sullivan, S. P. (2019, October 24). "Blind Tasting is the Only Way to Rate Wine Fairly." Retrieved from https://www.winemag.com/2019/10/24/blind-wine-tasting-rate-fairly/

Vinodhini G., Chandrasekaran, R. M. (2012). "Sentiment Analysis and Opinion Mining: A Survey." *International Journal of Advanced Research in Computer Science and Software Engineering* 2(6), 282-292. https://pdfs.semanticscholar.org/261e/26ae134b8f63270dbcacf2d07fa700fdf593.pdf

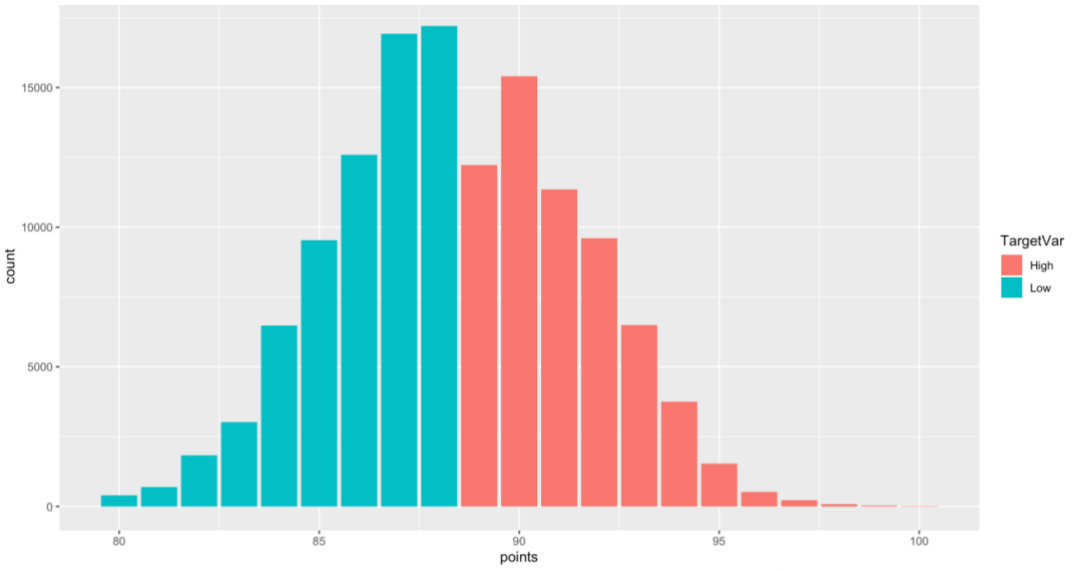
Xu, K., & Wang, X. (2017). Wine Rating Prediction. Retrieved from: https://pdfs.semanticscholar.org/599d/3935682044ef1304b5f6d180b0172027610f.pdf

Yeo, M., Fletcher, T., & Shawe-Taylor, J. (2015). Machine Learning in Fine Wine Price Prediction. *Journal of Wine Economics*, 10(2), 151-172. doi: https://doi.org/10.1017/jwe.2015.17

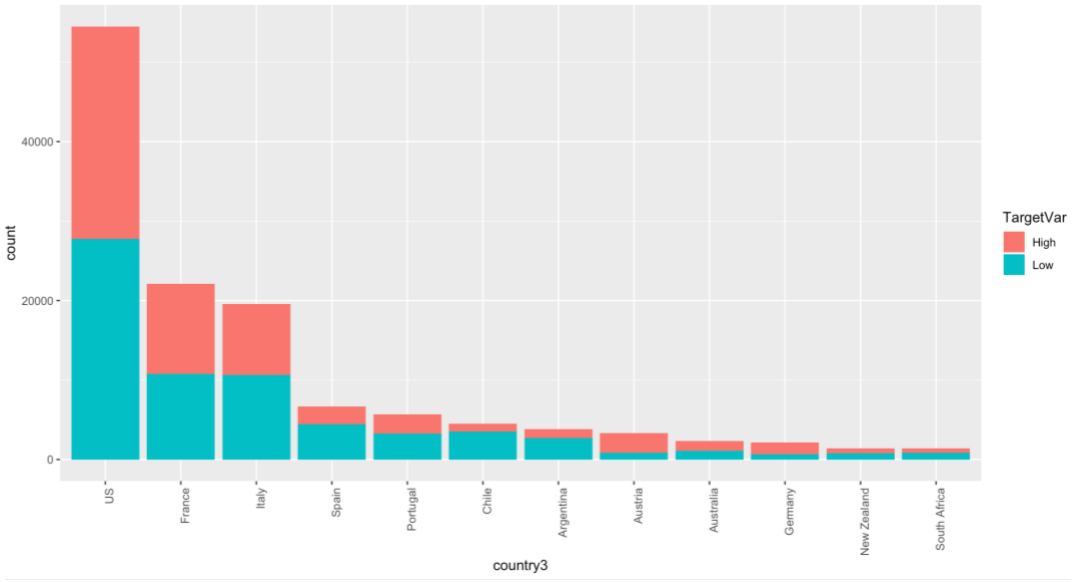
Appendix - Graphs and Confusion Matrices

Distribution Graphs by Attribute

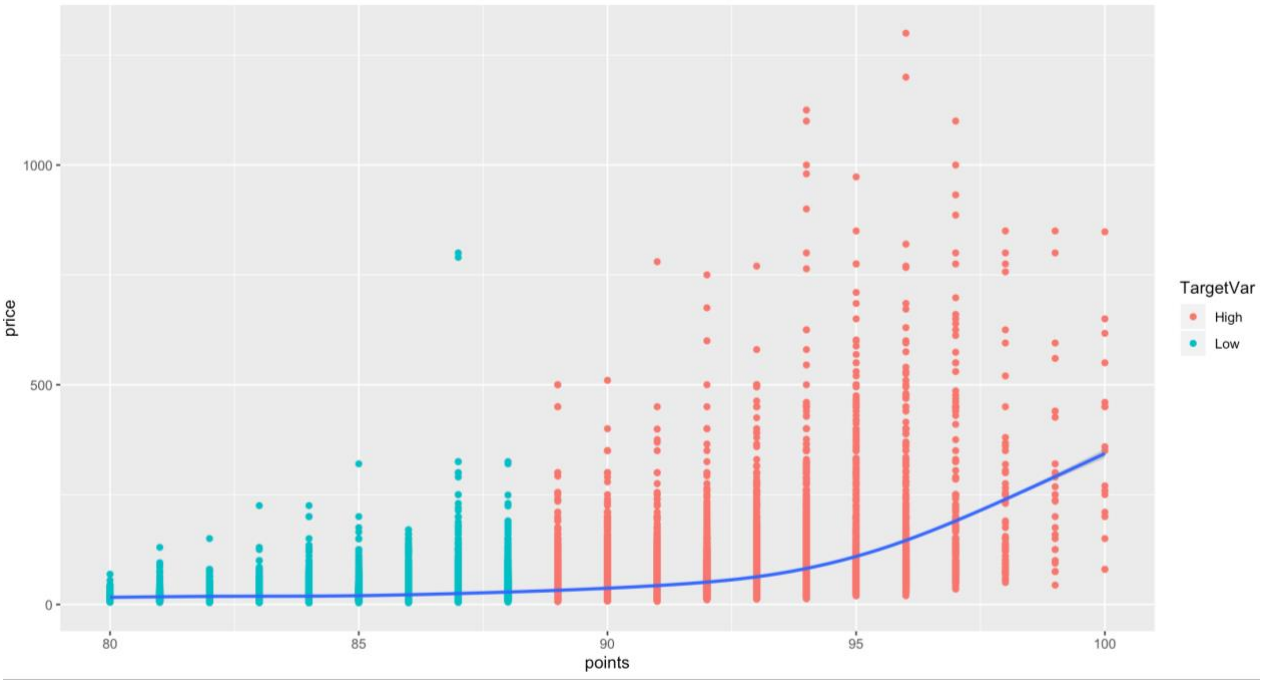
Score (Target Variable)



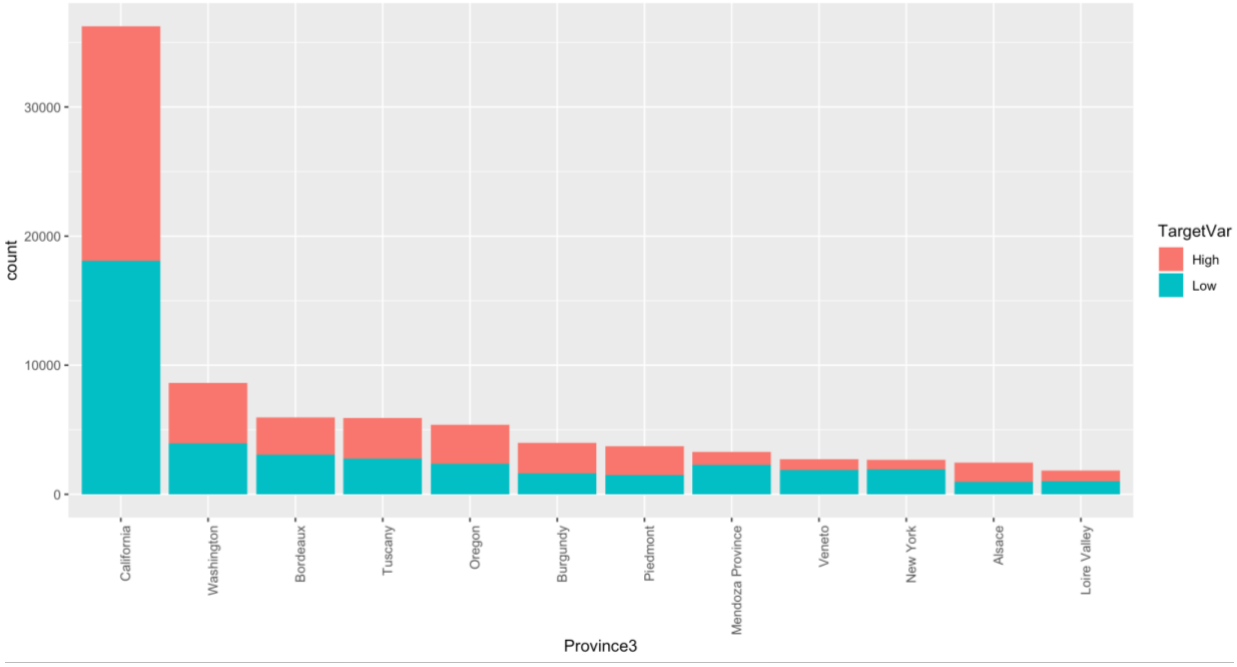
Country



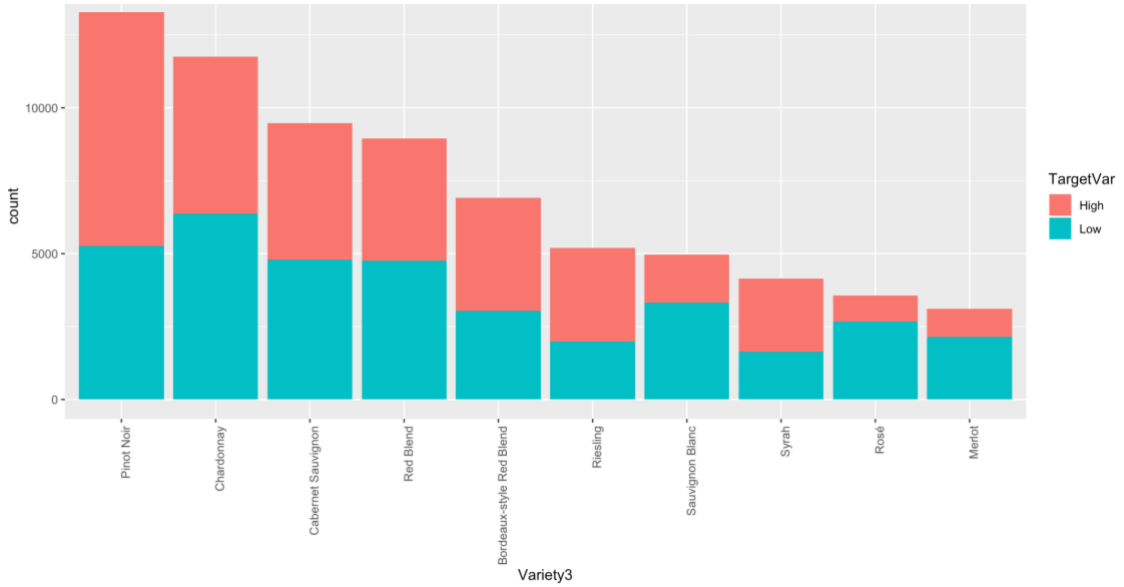
Price



Province



Variety



Feature Ablation Confusion Matrices

Naive Bayes Model - Features: Description Alone - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.796
Kappa	0.5899

Model Confusion Matrix:

Act \ Pred	High	Low
High	3614	1083
Low	957	4347

Naive Bayes Model - Features: Description and Price - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7996
Kappa	0.5958

Model Confusion Matrix:

Act \ Pred	High	Low
High	3494	1203
Low	801	4503

Naive Bayes Model - Features: Description and Country - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7966
Kappa	0.5913

Model Confusion Matrix:

Act \ Pred	High	Low
High	3634	1063
Low	971	4333

Naive Bayes Mode - Features: Description and Province - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8007
Kappa	0.5995

Model Confusion Matrix:

Act \ Pred	High	Low
High	3649	1048
Low	945	4359

Naive Bayes Model - Features: Description and Variety - Evaluation: 10-Fold Cross Validation

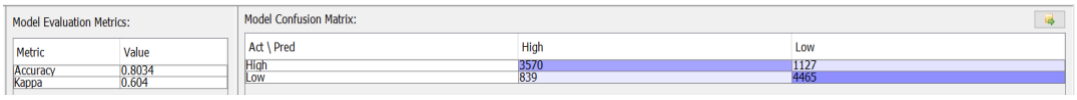
Model Evaluation Metrics:

Metric	Value
Accuracy	0.7942
Kappa	0.5865

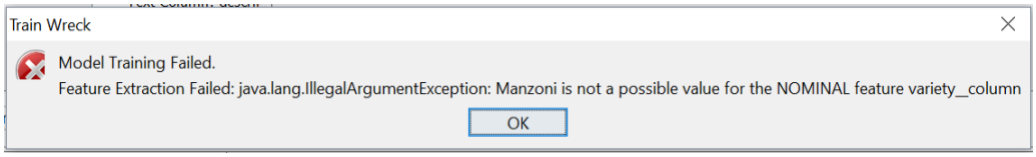
Model Confusion Matrix:

Act \ Pred	High	Low
High	3626	1071
Low	987	4317

Naive Bayes Model - Features: Description with Price, Country, Province, and Variety - Evaluation: 10-Fold Cross Validation



Error Encountered when Applying Test Set for Evaluation



Confusion Matrices for Variety-Specific Models

Variety: Random Sample (many varieties) - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.796
Kappa	0.5899

Model Confusion Matrix:

Act \ Pred	High	Low
High	3614	1083
Low	957	4347

Variety: Random Sample Training Set (many varieties) - Evaluation: Random Sample Test Set

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8136
Kappa	0.6277

Model Confusion Matrix:

Act \ Pred	High	Low
High	195	58
Low	35	211

Variety: Chardonnay - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8202
Kappa	0.6379

Model Confusion Matrix:

Act \ Pred	High	Low
High	4332	1045
Low	1068	5308

Variety: Chardonnay - Evaluation: Random Sample Test Set

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7555
Kappa	0.5119

Model Confusion Matrix:

Act \ Pred	High	Low
High	175	78
Low	44	202

Variety: Cabernet Sauvignon - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8237
Kappa	0.6473

Model Confusion Matrix:

Act \ Pred	High	Low
High	3811	871
Low	799	3991

Variety: Cabernet Sauvignon - Evaluation: Random Sample Test Set

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7535
Kappa	0.5084

Model Confusion Matrix:

Act \ Pred	High	Low
High	164	89
Low	34	212

Variety: Riesling - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.816
Kappa	0.6243

Model Confusion Matrix:

Act \ Pred	High	Low
High	2533	670
Low	285	1701

Variety: Riesling - Evaluation: Random Sample Test Set

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7535
Kappa	0.5082

Model Confusion Matrix:

Act \ Pred	High	Low
High	168	85
Low	38	208

Variety: Rosé - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8367
Kappa	0.5629

Model Confusion Matrix:

Act \ Pred	High	Low
High	595	295
Low	287	2387

Variety: Rosé - Evaluation: Random Sample Test Set

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7034
Kappa	0.4088

Model Confusion Matrix:

Act \ Pred	High	Low
High	148	105
Low	43	203

Variety: Pinot Noir - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8394
Kappa	0.6669

Model Confusion Matrix:

Act \ Pred	High	Low
High	6832	1170
Low	961	4309

Variety: Pinot Noir - Evaluation: Random Sample Test Set

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7996
Kappa	0.5994

Model Confusion Matrix:

Act \ Pred	High	Low
High	197	56
Low	44	202

Variety: Bordeaux-Style Red Blend - Evaluation: 10-Fold Cross Validation

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7837
Kappa	0.5654

Model Confusion Matrix:

Act \ Pred	High	Low
High	2980	895
Low	601	2439

Variety: Bordeaux-Style Red Blend - Evaluation: Random Sample Test Set

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7375
Kappa	0.475

Model Confusion Matrix:

Act \ Pred	High	Low
High	185	68
Low	63	183