Developing an Application to Score Unreviewed Wines

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Jared Werner Hood

Spring, 2020.

Technical Project Team Members

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Charles Reiss, Department of Computer Science

Developing an Application to Score Unreviewed Wines

Jared Hood
Computer Science
University of Virginia
Charlottesville VA USA
jwh6ry@virginia.edu

ABSTRACT

Wine is an increasingly popular libation across the world and logically with that follows an increase in wine purchases. Inexperienced wine drinkers may not know how to determine the quality or value of a bottle of wine based on limited information available on the label. The aim of this project is to provide a score, from 1 to 100, for a bottle of wine in order for consumers to make educated purchasing decisions. The 100-point score is standard in the wine industry for critics to rate individual bottles and the score indicates the quality of the specific bottle. This project uses a dataset of ~150,000 wine reviews, split into training and testing sets, as input to a machine learning algorithm in order to train the model on predicting the wine score. The features of the dataset that are included are wine country of origin, region, price, province, vintage, variety, which will derive a score out of 100. The model was validated with the testing set for accuracy. Finally, the project implements a simple user interface to allow users to input the aforementioned data from the wine of interest and generate a score for that bottle. The significance of the project findings is that if wine can accurately be assessed then it will make selecting good bottles of wine easier for novice consumers.

1 Introduction

The wine industry has experienced significant growth in the past few years and with that growth has come a large number of inexperienced wine consumers. One of the obstacles that novice wine consumers face when first getting into wine is being able to identify quality wines on the shelves at the stores where they shop. With stores carrying over 100 different options [1] an inexperienced consumer may struggle to differentiate one bottle from another. One way that consumers can figure out what wine will be of higher quality is by utilizing the wine critic scores that stores sometimes display with the bottle. The score is created by a professional wine critic who judges the wine on a number of features and assigns it a score on a 100-point scale. This wine score can give the consumer a good indication of what a professional thought of the wine, which can give them assurance in the quality of their purchase [2]. However, with the mass number of wines available, not all of them can be scored and further if they have been scored not all of them are displayed with their score.

Currently if there is no wine score displayed at the store a consumer can attempt to find extrinsic information about the wines in two ways. First, the consumer can look up the bottle of wine online to see if there is an available score. Second if the bottle has not been scored yet by any critic, which is the case for most bottles, then the consumer can look up other information pertaining to the bottle such as winery reputation. The best-case solution where the score is available online would take the consumer around one minute to lookup. However, in the worst-case scenario, the consumer may need to do a fair amount of research depending on the obscurity of the wine in order to begin to determine its quality. This would also require the consumer to have some knowledge of wines in order to be able to know what to research in the first place. Thus, the problem with the current way to find information about wines for novices is that they must either do their own research, which has a steep learning curve, or must rely on the opinions of experts. Relying on expert opinion fails to provide the consumer with any usable information when a particular bottle of wine has not yet been reviewed. This leaves a large number of wines at the store that consumers will have no indication of the quality of the wine to base their purchases on.

2 Related Work

There are currently no other systems that take a wine bottles label information and price and then output a predicted wine score. One alternative way to find information about a particular wine bottle is to lookup the bottle in a wine database. One such database is www.wine-searcher.com [3], which allows users to search for specific bottles within their large database. Searching will yield the wine critic score if the bottle has already been scored as well as some other information about the bottle including community reviews. These community reviews can be used by a consumer to help aid their decisions, but community reviews are still not as reputable as an expert wine critics review.

Although there are no systems that do the exact same task as the one described in this paper, there are applications that will help users find wine that they will enjoy. Hello Vino, is a wine recommendation application that provides custom recommendations based on the user's individual preferences [4]. The application begins by asking the user about what kind of wines they enjoy and what their favorite flavor profiles are. It then searches through its wine database and recommends wines from the user's local area with the same features they selected. This application will give wine recommendations to the consumer but does not allow for the consumer to input specific wines and receive back a recommendation about whether they would enjoy the wine or not. This does not give the consumer a lot of flexibility when attempting to choose wines, which can be limiting when at a grocery store with a small selection of wines.

Another wine recommendation application on the market is Wine Ring. Wine Ring uses machine learning to help users identify wines that they will enjoy. The application works by first having the user rate several wines that they tried in the past in order to build a profile of that user's preferences [5]. Once the application has enough data on the user it can then begin to recommend wines to the user that will fit within their preferences. The application continues to fine tune the profile for each user as they rate each new wine they try. Users are also able to take a picture of a bottle of wine at the store and then get back whether the application believes the user will enjoy that particular bottle. The downside with this application is that it requires the user's preference data up front in order to recommend them wines. This can make the application difficult for novice consumers to use who do not have a lot of experience with wines and thus may not know their own preferences yet.

3 System Design

3.1 Overall System Architecture

The system was designed in a Jupyter Notebook hosted on Google Collaboratory. The system is composed of three parts, the underlying machine learning model, the user data input, and the output wine score. The underlying machine learning model is a random forest regression from the sklearn ensemble package. The model was trained on wine review data sourced from Kagle, which contained roughly 150,000 entries with 14 features. After data cleaning the features that the model was trained on included: country of origin, price, province of origin, region of origin, wine variety, winery name, and vintage. The model was trained in a separate Jupyter notebook and after training was completed the model was saved into an external file on Google Drive. This file was then loaded into the Jupyter notebook that contained the actual user application where it is used to quickly calculate the wine score. This saves the time that would be needed to retrain the model every time the application is run.

When the application is run, the user inputs the prompted features needed for the calculation. The inputted user data is then transformed through a pipeline to encode the data in such a way that it can be used by the model to predict the score. The pipeline one hot encodes the categorical features and uses a standard scalar on the numerical features. After the transformation the model takes the row of data that the user inputted and then predicts the wine score for that bottle of wine. The prediction is outputted along with the bottle's information that the user inputted.

3.2 Design Decisions

The first design decision in this application was to determine which of the features from the original dataset would be relevant to the generation of wine scores. The original dataset included 14 features, of which after dropping the irrelevant ones 10 remained including country, designation, points, price, province, region_1, region_2, variety, winery, and vintage. The wine designation is used by wineries to indicate unusual qualities of the wine, such as degree of sweetness or color. As specific designations are not regulated for consistency across wineries, of the 92,506 entries that included a designation, there were 37,979 unique designations. With this large number of unique designations, it would not have been possible to extrapolate a correlation between specific designations and wine score. Instead it was better to check if there was a correlation between having any designation versus none. The correlation coefficient was found to be 0.0528 between having designation and wine score indicating that designation has a very small positive influence on the wine score. However, this includes all the designations that have only a handful of occurrences, so it was also necessary to check whether the most frequent designations had an influence on the wine score. There were 11 unique designations with over 300 occurrences with "Reserve" being the most frequent with 2009 occurrences. When only keeping the designations that had more than 300 occurrences in the dataset, the correlation coefficient was found to be -0.0268 between having a designation and wine score. This indicates that the most frequent designations did not have any notable influence on the wine score. Therefore, designation was dropped from the feature set. The other feature that was dropped from the dataset was region_2. This feature was dropped because there were over 50,000 entries in the dataset that did not have a value for region_2.

To be used by the machine learning model the data needed to be transformed into a usable form. The first step in cleaning the data was to remove all of the rows containing missing values or to replace the missing values. For the price feature any value that was missing was changed to be the mean value of all the other price values. This allowed for more data points to be kept versus having to remove any of data points that did not have an included price. Additionally, for the price values it should be noted that the retail price for the same bottle of wine can change from store to store and day to day. This does not greatly affect the outcome of the prediction as the difference in price does not fluctuate enough to drastically alter the model's prediction. Therefore, for ease of use the customer can enter whatever the listed retail price is at the store the day they are there. For the categorical data points that had missing values, all of the missing values were changed to be "Unknown". This not only allowed for more data points to be kept for the model to train on, but also allowed the model to be able to account for instances where the wine bottle label information being inputted in by the user did not have all of the necessary features. As the wine label information is not required by the government to have all of the information that the model was trained on such as region and vintage [6], it is possible that the user may not be able to find some of the features on the label. Therefore, in order for users to be able to use the application on any wine bottle they may find at the grocery store it was important to have a way for the machine learning model to handle any data it hadn't seen before. For example, if a user were to input into the application a region that the model hadn't seen before, it would not be able to parse the data. Therefore, having at least one "Unknown" for each feature would ensure that if the model came across a label it hadn't seen before that it would be able to just treat it as an unknown value.

The final notable design decision in this project was the machine learning model used. Several different models were trained on the data and then were validated for accuracy. The model's accuracy was determined from the root mean square error (rmse) score for the data. This score indicates how far off on average the model was at predicting the wine score for a given bottle. Four different models were tested including a polynomial kernel regression model, gaussian kernel regression model, a random forest regression model, and a decision tree regression model all from the sklearn package. The random forest model performed the best out of all of them with a training error of 0.807 and a testing error of 2.007.

4 Procedure

To generate a wine, score the customer runs the application in the Jupyter notebook from the Google Colaboratory page. When run the application will prompt the user to input the information from the label of the wine bottle including the wine's country of origin, province, region, wine variety, winery, and vintage. An example of the user inputting the data can be seen below in Figure 1.

```
Country: US
Price: 9.00
Province: California
```

Figure 1: Application prompting user to input wine bottle label information.

The application will also ask the customer to input the price of the bottle of wine, for which the customer will input the listed retail price. If any of the features of the wine bottle are not available on the label, such as the wine region, the user can simply omit that piece of data and enter only what is available. After all of the information has been entered, the application will then calculate and display the predicted score for that particular wine. The application will also display the users inputted information in a table so that they can look back and make sure they entered everything correctly as seen in Figure 2.

```
Country: US
Price: 9.00
Province: California
Region:
Variety: Merlot
Winery: Kirkland Signature
Vintage: 2015


 Your Entered Information

  country  price    province region_1 variety             winery vintage
0      US    9.0  California     None  Merlot  Kirkland Signature    2015

Wine Score Prediction:  86.45
```

Figure 2: Application output including inputted information and wine score prediction.

The customer can then use this score to get a general idea of how critics would rate the wine, which can be used to aid purchasing decisions. The application can be run for as many bottles of wine as needed in order to give the consumer a fair idea of the qualities of the wines they are considering.

5 Results

The wine score generator allowed the user to generate a wine score in 30 seconds. This is a comparable speed to looking up the wine score for a given bottle of wine in a database. However, if the bottle had not yet been rated, the time it takes to generate the score with this application is much faster than the time it would take for a user to research the quality of a wine. Also, in order for a novice consumer to begin to research a wine they would likely have to learn about what makes a quality wine in the first place. Therefore, overall this application significantly decreases the amount of time necessary to get a general indication of the quality of a bottle of wine. This means that consumers will be able to use the application to efficiently get a general idea about the quality of a number of bottles at the grocery store, which they can then choose from when buying a wine.

In terms of accuracy the testing rmse of 2.007 indicates that on average a given wine score will be within ± 2 of the actual score. Being within 2 of the actual critic wine score is good enough for consumer purposes as it will be close enough to give a good indication of the quality of the wine. The model also revealed that the most significant factor affecting the wine score of a given bottle was the price. From the initial dataset the correlation coefficient between price and wine score was 0.416, indicating a semi-strong positive relationship between the two features. This was further exemplified in the final model's feature importance where the price was the most important feature with a feature importance score of 0.376 compared to the next most importance feature which was wine variety with a feature importance of 0.006. The next most important feature was country with an importance of .002 and the rest of the features were nearly negligible.

6 Conclusions

For this technical project I designed an application to meet the need of giving consumers access to wine bottle scores for unreviewed wines in order to help them make educated purchasing decisions. The finished product was able to calculate the wine score for a bottle of wine within ± 2 points. Since the purpose of the wine score is to give the consumer an indication of the quality of the wine that they are buying in order to reduce the chances they buy a bad wine, ± 2 points is good enough to give the consumer a usable indication of the wines' quality. Therefore, this application can be used by consumers to compare and contrast different wines at the store in order to select the best one. It is possible that the score predicted by the model will not match up with the preferences of the consumer, but this is an effect of individuals having different pallet preferences. As the data that the model was trained on was all from expert reviewers, the model is therefore a generalization of the tastes of the experts. Thus, the application may not be in line with every individual's preferences but can still give an indication of what wines the experts enjoy.

7 Future Work

The next step for this system would be to make it accessible to consumers on their phones. This could include either creating a mobile application or a mobile website where consumers can utilize the model to generate the wine scores. The addition of this feature would make it easier for consumers to use the application when at the grocery store where having a computer available is not always feasible. This is the most important step for the application to potentially gain traction in the consumer market as it is more in line with the use case of the wine score generator than the computer-based application.

Another feature to add to the application would be to allow users to take a picture of the wine bottle label and have the information automatically filled in. This feature would make using the application more streamlined and therefore would allow users to generate scores for more wine bottles in a smaller amount of time. The benefit to this is that it would enable users to compare more wines and thus have a larger selection of bottles to compare and contrast when deciding what to ultimately purchase.

A final step for this project would be to create a mapping between data that the users input to the data that the model was trained on. Since the dataset that the model was trained on is not fully comprehensive of the entire wine industry, the model is not able to understand all the possible inputs that the user may enter. Therefore, the model is not able to utilize all user inputs that may be important in determining the wine score. This was not implemented in this project due to time constraints, but in the future a database could be created to match user inputs to similar features in the model. For instance, if a user inputted a region that is not present in the model it could be checked if any of the regions present in the model are geographically nearby to that region, which would likely produce similar wine. This step would improve the model by enabling it to understand more user inputs and thus be able to provide more accurate predictions for a wider range of wine bottles.

REFERENCES

[1] Oana Bârbulescu (2018, December 1). Marketing Strategies Used By Retailers On The International Wine Market. Bulletin of the Transilvania University of Brasov. Series V : Economic Sciences, 11(2), 9 - 14. DOI:https://doaj.org/article/ecedc80d8584467b807a5e4205f93d1d

[2] Hopfer, H., & Heymann, H. (2014, March 1). Judging wine quality: Do we need experts, consumers or trained panelists?. Food Quality and Preference, 32(Part C), 221 - 233. DOI:https://doi.org/ 10.1016/j.foodqual.2013.10.004

[3] Searcher. Find and price wines, beers and spirits across online stores. (n.d.). Retrieved November 25, 2020, from http://www.wine-searcher.com/

[4] Hello Vino. (n.d.). Wine Assistant App. Retrieved November 25, 2020, from http://www.hellovino.com/

[5] Dillion, S., Dillion, P., & Sussman, A. (2014). U.S. Patent No. 8751429. Washington, DC: U.S. Patent and Trademark Office.

[6] Bugher, T. (2020, October 20). TTB: Wine Labeling: Home. Retrieved November 25, 2020, from https://www.ttb.gov/wine/labeling