NYU Center for Data Science
Fall 2018 - Introduction to Data Science

Final Project:
Carpe Vinum: Predicting Wine Quality

Daniel Amaranto, Hee Seok Joo, Jaejin Kim

December 8, 2018

Contents

1 Business Understanding

The nature of vinification makes it hard for investors to make informed investment decision: one needs 10 years of runway to see significant returns. Also, there are overwhelming number of factors that can affect the price of a wine, from the market conditions and how, when and where it was produced to weather conditions in which the grapes were raised. There being more than 10,000 varieties of wine grapes and even more varieties of wine complicates the issue. It is extremely challenging for individual investors or small businesses to keep track of all the different combinations that make for success. Nonetheless, with the global wine market estimated to be around $302 billion and projected to grow by 7 percent in 2018[1], wine is becoming a more lucrative investment option. The London International Vintners Exchange (Liv-ex), which was founded in 2000, surpassed its previous exposure

– the total value of bids and offers on Liv-ex – record this year, reaching over £50 million.[2]

To tackle the discrepancy between wine being a potentially valuable investment and there being a number of obstacles, we wanted to build a model which predicts proxies of quality from a set of characteristics that describe a particular bottle of wine. The data instance in this project is a specific bottle of wine output from a maker, with a given year and title. The target variable would be points as rated by Wine Enthusiast on a scale of 80-100. For this, we used the wine data from Kaggle, which was scraped to feature unique reviews of various wine products, along with their country of origin, province, region, maker, grape variety, price, and points. We trained models to be able to predict points from this information.

This could benefit the wine sellers or investors greatly, since the model takes in certain baseline qualities that are predetermined before the products hit the market or mature and predict the quality as represented by points. The model will be able to advise not only investors, distributors, and sellers, but also those who want to create their own vineyard or winery.

2 Data Understanding

For this project, we used the wine reviews data from Kaggle, which has 280,901 reviews on commercially available wine.[3] The dataset has 13 other variables such as country, description, designation, points, price, etc. From the 14 variables, we initially removed 3 of them, which are 'taster_name', 'taster_twitter_handle', and 'title'. The three variables are all related to the reviews and reviewers themselves rather than the bottles of wine, hence are not good predictors to any proxies of quality. Next, NAs in column 'country' has been filled with 'Other', and those in column 'price' has been filled with the mean of the overall price. We then standardized values under 'price' column to allow for easier analysis.

There were 756 distinct grape varieties in the data, but 88 of those varieties were unique bottles and hundreds more had only a few instances. Therefore we encoded 'variety' so that the grape varietals with fewer than 2500 instances would fall into an 'Other' category. We created new variables for the analysis, namely 'over90' and 'continent', which are Boolean variable indicating whether the wine received points higher or equal to 90, and the continent that the country of origin is in. Being above 90 puts a bottle roughly into the top third of wines (34% of wines score at least that high). Finally, we created the variable 'textlen' which is the standardized length

of reviews. As will be specified below, we ultimately decided to not use this variable in our model. Initially, we thought it could be a feature of the model, but we feared data leakage as the review and points are assigned at the same time, so knowing one implies that the other is also known.

3 Data Preparation and Analysis

Our goal was to predict the points of the wine given various factors. More specifically, predict whether a wine received over 90 points in a review, so one of the variables created as mentioned above, 'over90', was used as the target variable. By using binary target variable instead of numeric target variable, we aimed to predict with few input features. The main features in our model consist of country of origin, price of wine, and grape variety.

As mentioned above, missing numeric data was filled in with the mean of the variable, and missing or ambiguous categorical data was replaced with 'Other'. The categorical data was transformed using OneHotEncoder from sklearn, which converts categorical data into binary values in order to fit them through the model. The textlen and the description of the wine was initially used because they seemed to have correlation with the points, but we suspected data leakage and seemed artificial since anyone can write a long negative review and it could erroneously lead to positive prediction. We also used countries as the geographic feature instead of specific regions because data on specific regions of the wineries were often missing even though large countries like US or Canada have drastically varying climate from region to region that could lead to some inaccuracy. Despite the wide variation in climates and wine quality within a given country, being marketed as "Canadian," or "Chilean," or some other nationality could have a particular impact on perceived quality that is useful to the model. The organized data set was then split into train, validation, and test data. The sets were split simply by taking 60% of the data set to be the train set, 20% to tune the parameters, and the remaining 20% to be the final holdout test set.

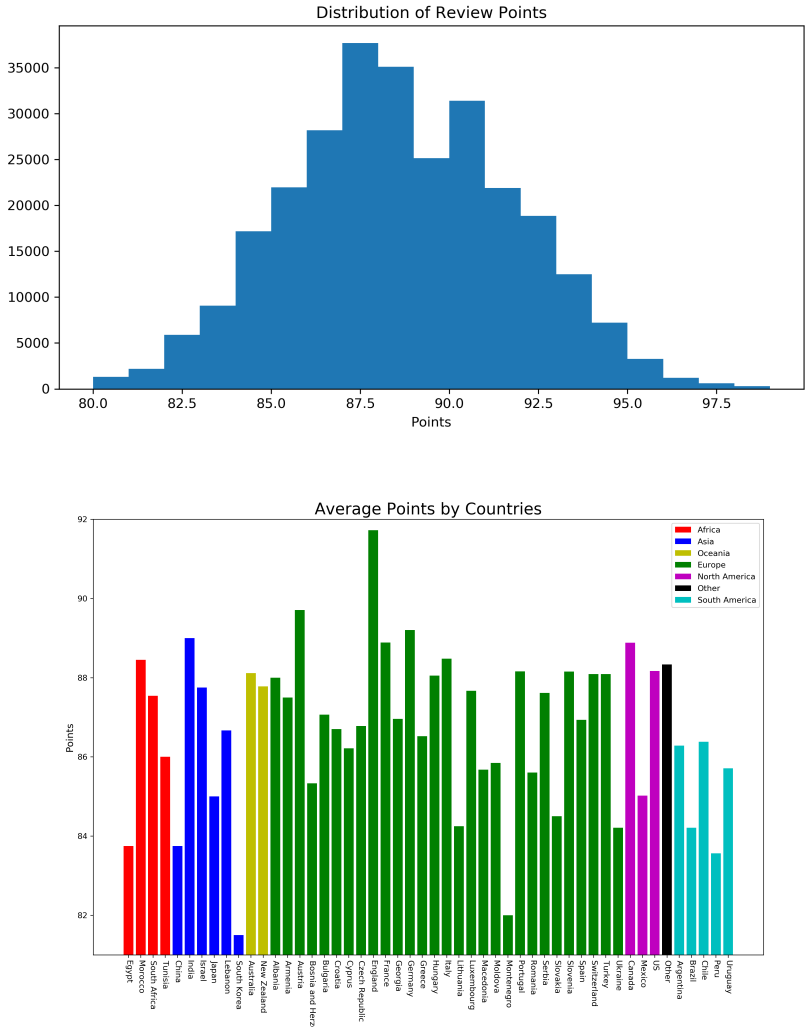

Figure 1: Some visualizations of the data are presented. The distribution of the points has almost a normal bell shape with the mean around 88.1 point. The average points by countries by each continent are presented to show wines from which country received highest points on average.

4 Modeling and Evaluation

For this classification task one of the main challenges is the scarcity of features. Another challenge is the large number of classes within the country

and variety groups. There are 50 countries and 24 varieties (after grouping scarce varieties into an 'Other' category). Our baseline was a Naive Bayes model that took price, country, and variety as input and had accuracy of 0.738 (see Figure 2). Several other models were also tried and compared, including some ensemble models. Unsurprisingly, the Naive Bayes and K-nearest neighbors models were the worst performers. One-hot encoding so many features results in many more columns in the dataset. KNN does particularly poorly in this setting because of the curse of dimensionality: instances are all quite far apart and the model struggles to determine important differences in distance. [4] Moreover, the model is very inefficient to train. Logistic regression was better than Naive Bayes or KNN, but the best models were the Random Forest and the Gradient Boosting Classifier.

From the data it originally seemed that tree based methods would be the best due to all the categorical features. Indeed this was so, but the one linear method did perform fairly well when compared to the tree based methods. One reason for this might be that quality tracks fairly consistently with price. The tree based methods are still superior because they are capable of picking up non-linear relationships in addition to capturing the importance of price as it relates to quality. The ROC curves and AUC/accuracy of each model are shown in Figure 2.

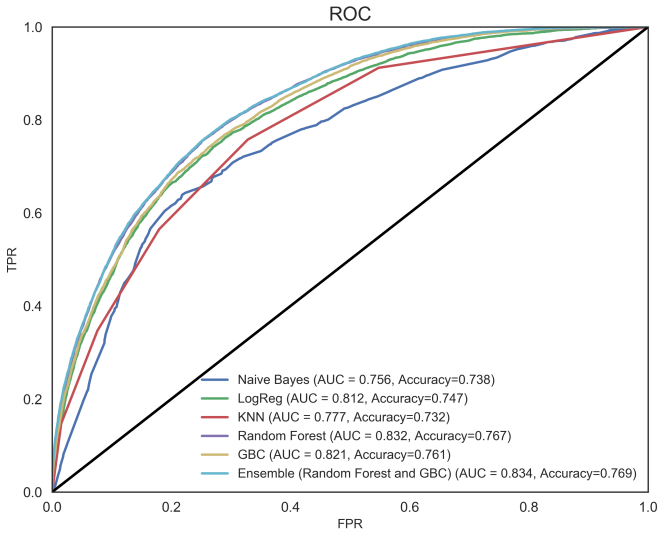

Figure 2: ROC Curves for Various Models

For each method we replaced country with continent to determine whether a more general indicator of origin would be sufficient to improve predictions. The motivation for this is that having fewer classes would make a more efficient model. Continent performed worse in every instance. After identifying that the ensemble methods were the best performers, we employed backward stepwise feature selection to see if each feature was necessary. Removal of any feature decreased model accuracy in each ensemble method, so we moved on to tune the models on the validation set with each of the original features. For each ensemble model we compared the results with 10,50,and 200 estimators, but the default values (10 for random forest and 100 for GBC) were optimal.

In both the random forest and the gradient boosting classifier, grid search was performed over 10 different values [2, 4, 8, 16, 32, 64, 128, 256, 512, 1024] of the minimum values required to split a node and the minimum number of samples required to exist in a leaf. These parameters affect the bias variance balance of each model. For each of these parameters, the lower the value the more you are allowing a model to over fit the training data. However, higher values will limit the model's flexibility and result in greater bias.[4] Performing grid search and checking accuracy on the validation set revealed that the optimal parameters for the random forest were a split size of 8 and a leaf size of 128, while for the gradient boosting classifier the optimal values were 256 and 16, respectively. It should be noted that the increase in accuracy was very small. We compared the optimized ensemble methods and an ensemble of each of those on the held out test set, revealing that the ensemble of the random forest and gradient boosting classifier was the best performer (see Table 1).

As constructed, this model takes very simple characteristics that describe a bottle of wine and finds out how they relate to the score that wines receive. It is remarkable that essentially 3 facts (price, nationality, and grape) can quite meaningfully improve a prediction about whether or not the bottle will score highly. Aside from the price that a wine businessperson pays for a bottle, there are no serious consequences that accompany these decisions. A false positive prediction may not sell well and a false negative could represent some lost revenue, but simple accuracy is a fair measure in this circumstance.

Table 1: Accuracy on Test Set

Model	Test Set Accuracy
Random Forest	0.7636
Gradient Boosting Classifier	0.7597
RF/GBC Ensemble	0.7656

4.1 Categorical Feature Embeddings

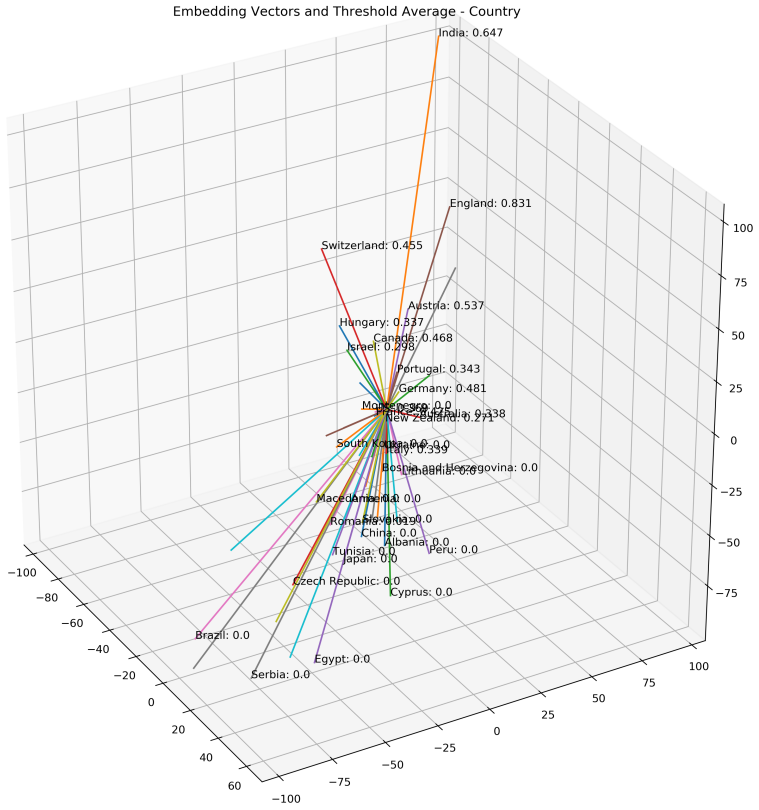

Figure 3: Embeddings for Country categories learned on Neural Net

In addition to one-hot encoding the categorical features, we also trained a neural net that treated each class within a categorical feature as an embedding vector of length=(number of classes/2).[5] So for example, the country category, which had 50 different classes, had randomized vectors of length=25 initiated before the the model was trained in a feed forward neural net. As the model trained, the embeddings were updated. The accu-

racy of this method on the test set (.7464) was not ultimately better than the traditional models we trained. However, analyzing the learned country embeddings in lower dimensions shows that they did distribute based on points. Figure 3 shows the embeddings of each country, as well as the proportion of wines from each country that were over 90. The graph clearly shows that countries with higher proportion of wines over 90 are located in the upper regions, whereas countries that had no wines over 90 spread out in the opposite direction. The embeddings of countries where wine performed in the middle are between those extremes. Despite the fact that the neural network itself didn't have superior accuracy, additional training or different deep learning configurations could clearly be explored with and they have the potential to outperform the traditional methods.

5 Deployment

Despite not having many features, our final model does a decent job at predicting whether or not a bottle of wine will score higher than 90. Some useful ways to imagine this model adding value would be to consider the decisions that must be made by someone who is setting out in the business of wine distribution or starting a high-end wine shop. Given the enormous size of the market, how do they choose which wines to buy? Even an expert sommelier can't sample every new wine that comes to market. The two lines in figure 3 represent two different buying strategies. The first is to simply buy all of the most expensive bottles of wine until you hit your budget (here the budget is arbitrarily chosen as $750,000). This would be one way to have a stock of premier bottles to sell. The other line represents buying all of the most expensive bottles of wine that have been predicted by the model to be have scores over 90. The distance between those lines is the benefit from using the model, and it can be interpreted in two ways.

Taking a look at the strategy without the model, the final composition of wines in stock would be 6,768 with scores above 90 and 2203 with scores below 90. With the model one would spend the same amount of money and end up with 6918 bottles above 90 and 2177 bottles below 90. Therefore you could measure the benefit from using the model as the additional high-quality bottles in stock.

Alternatively, you could describe the benefit from using the model in dollar terms by analyzing the vertical distance between the lines. If you had a shop with limited space and wanted to hit a target number of bottles, say 8000, choosing all of the most expensive bottles would cost $701,500. If

you applied the model to the most expensive bottles and only selected those predicted to be above 90, you would end up spending $695,600. The precise benefit that could come from using this model would vary depending on the budget and target quantities of a particular business. Continuous evaluation of the model's efficacy would be easy. As soon as new wines are judged the prediction can be verified and training data updated as well.

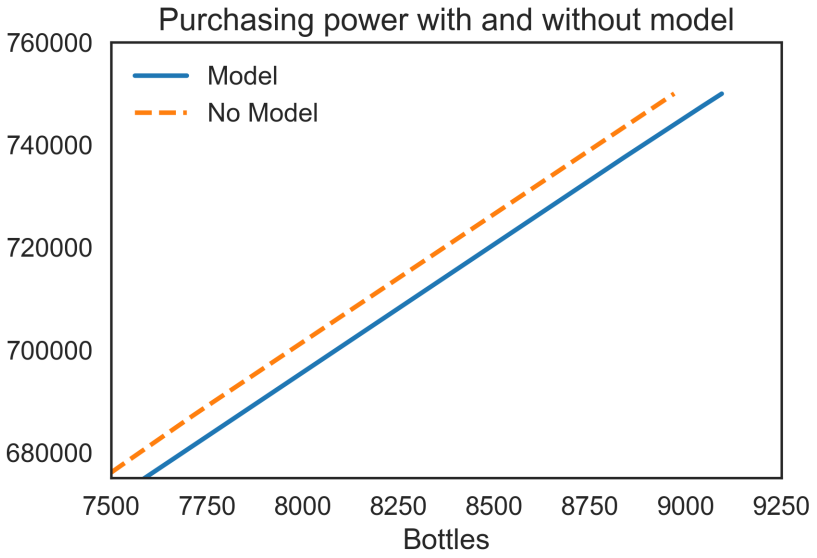

Figure 4: If you sort bottles by descending price and purchase all of them, at any point the basket assembled by the model is either cheaper than the basket assembeled without the model, or it costs the same amount but you end up with more bottles. This happens because the model rejects bottles that it identifies to be of lower quality.

6 Conclusion

We explored many models to predict a proxy of wine quality and were able to get reasonable results from a very limited set of inputs: price, country of origin, and grape variety. The resulting model would be useful for any buyers and sellers of wine. This model is promising and it could improve buying decisions even more with further tuning and feature development.

References

1. Euromonitor International Ltd.
2. https://www.liv-ex.com/2018/10/new-record-high-exposure-reaches-50million/
3. https://www.kaggle.com/zynicide/wine-reviews
4. James,Gareth, Witten,Daniela, Hastie,Trevor and Robert Tibshirani. *An Introduction to Statistical Learning.* Springer 2013.
5. https://yashuseth.blog/2018/07/22/pytorch-neural-network-for-tabular-data-with-categorical-embeddings/

Contributions

- Daniel Amaranto: Baseline models, final model outcome and deployment, training and visualizing categorical embedding vectors.
- Hee Seok Joo: Data preparation, data visualizations of unprocessed data.
- Jaejin Kim: Business understanding, data understanding and preparation, Random Forest model.