Wine Rating Predictor Client Summary Report

Data Revenue

Ceren Iyim

02.04.2020

Introduction

In this report, an online wine seller's sample inventory data set is analyzed consisting of 13 variables and 10.000 rows. The primary goal is to seek the possibility of building a good wine predictor as a proof of concept of a full-production solution.The rating is available as "points" in the sample dataset and it is a measure of quality of a wine between 80 and 100.

Rest of the variables are as follows: country, description, designation, price, province, region1, region 2, taster name, taster twitter handle, title, variety and winery. The variable that we are trying to predict, points, will be referred to as ***target.*** The rest of the variables, potential predictors will be referred to as ***features*** in the report.

Another important terminology that will be used throughout the report is, training and test sets:

Training set will be used to map patterns between the features and target and to build the wine rating predictor. The process of building is also called ***training the model*** or ***building the model***.

Test set will be used to generate ratings/points of a wine using the patterns found between the features and the target in the training set. Feeding the test set into the model and generating predictions process is called ***evaluating the model.***

Objectives

- Identify variables to built a good wine predictors
- Build a machine learning model that can predict points of a wine given the sample dataset
- Interpret model results, identify the model's errors and further improvement areas

The detailed analysis can be found in the Wine Rating Predictor 1, 2 and 3 notebooks. This report will focus on the most important findings of the overall process.

Data Cleaning

Sample dataset had some missing data and the majority of the (61.5%) of the region 2 values are missing. Since this feature isn't likely to provide significant information to the machine learning model, it was eliminated. In addition to that, designation, winery and taster twitter handle features and duplicate rows are removed. Finally, if there were any missing values in the target, also they would be removed.

After the data cleaning, the sample dataset is left with 9948 rows, 8 features and a target in the total:

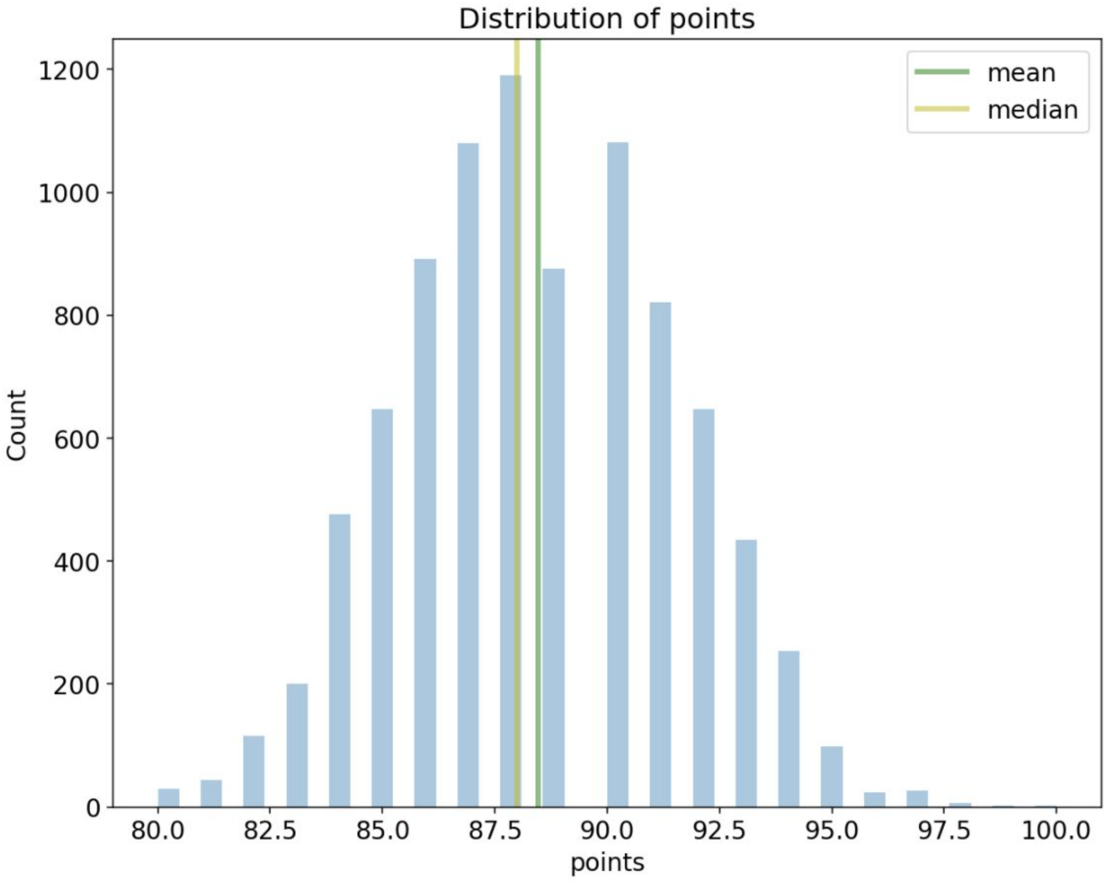
- 8948 rows in training set
- 1000 rows in the test set

Data Exploration

As we are mainly concerned with predicting the points, training dataset's features will be explored and visualized with their relationships to the target.

Distribution of Points

Distribution is a description of a variable's range and how data is spread in that range. Following is a ***histogram*** of showing the distribution of points. Each bar is a count of data points that exist within the bar range.



Points show a normal distribution (a bell shaped curve obvious) as expected from any random variable. Range of wine points are distributed between 80 and 100 with the ***mean*** (average) of 88.45 and ***median*** (the middle value of the sorted data) of 88.

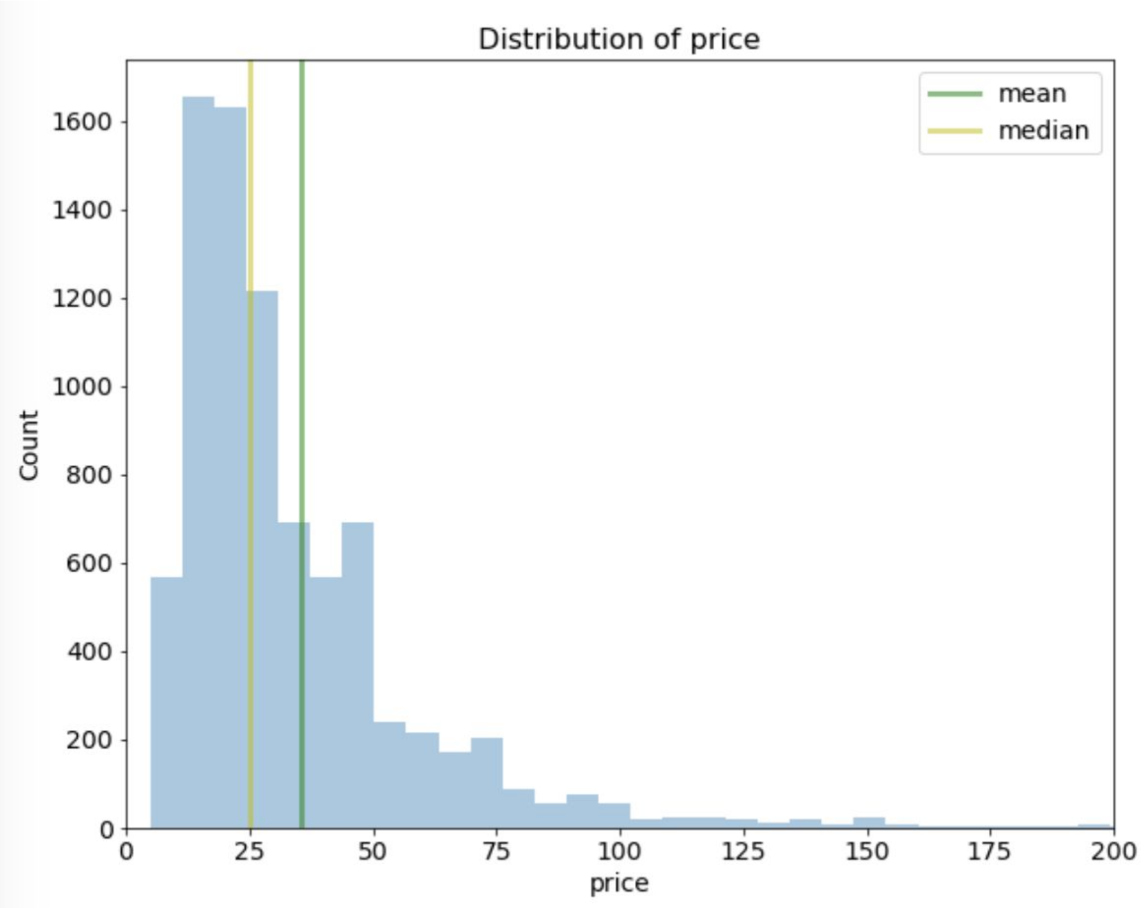
Moreover, ***standard deviation***, a measure of spread of a data range, is 3.03, resulting in a ***variance*** of 9.1, as the squared standard deviation.

Based on this statistics, we can confidently say that:

- 68% of the points lie in the 85.5 - 91.5 range
- 95% of the points lie in the 82.5 - 94.5 range
- 99% of the points lie in the 79.5 - 97.5 range

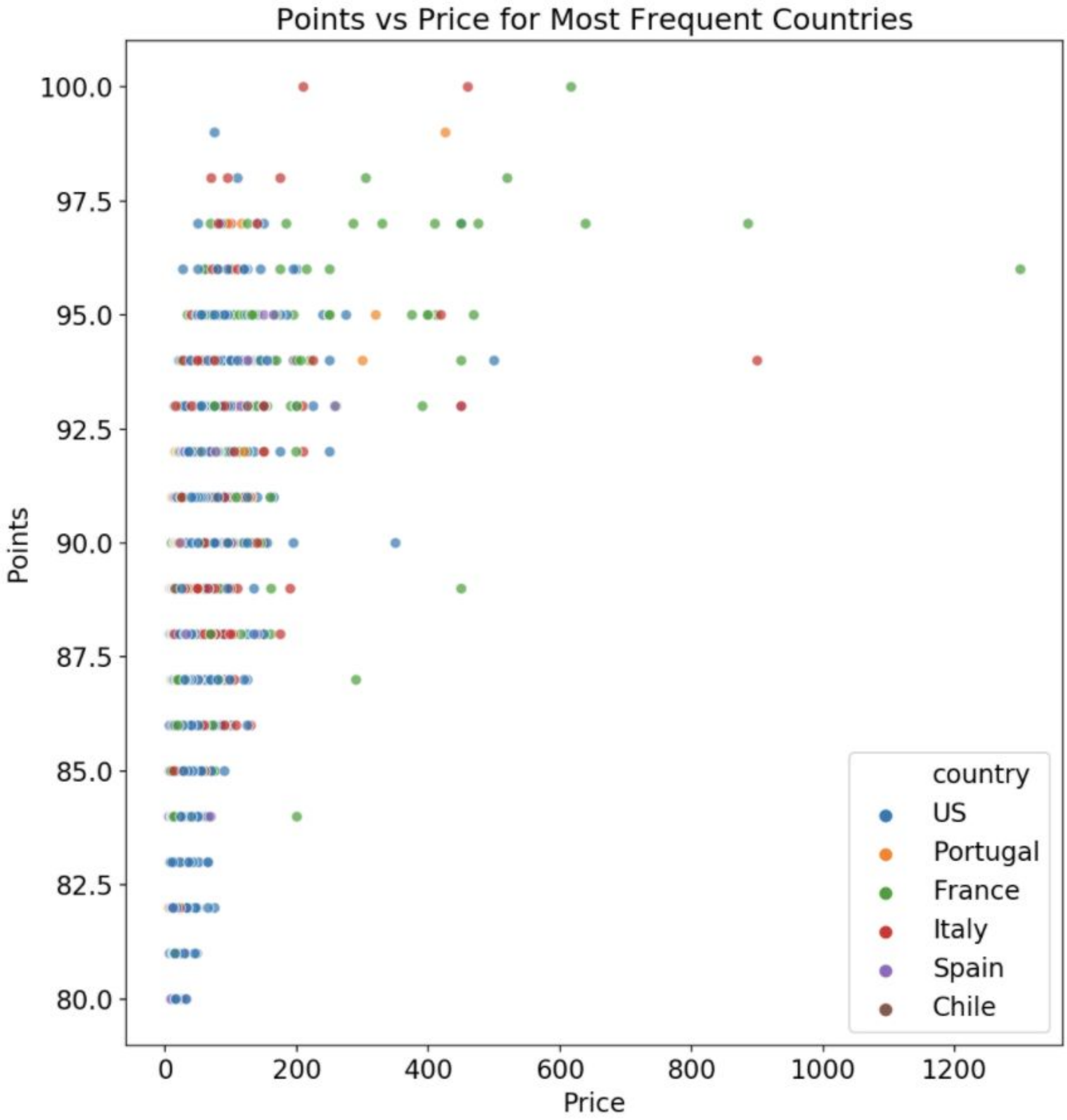
Distribution of Price

Price is concentrated between the range 0-200 and looking at it a full range which is 5-1300, did not convey much information. So, it is visualized within the 0 and 200 range with a histogram:



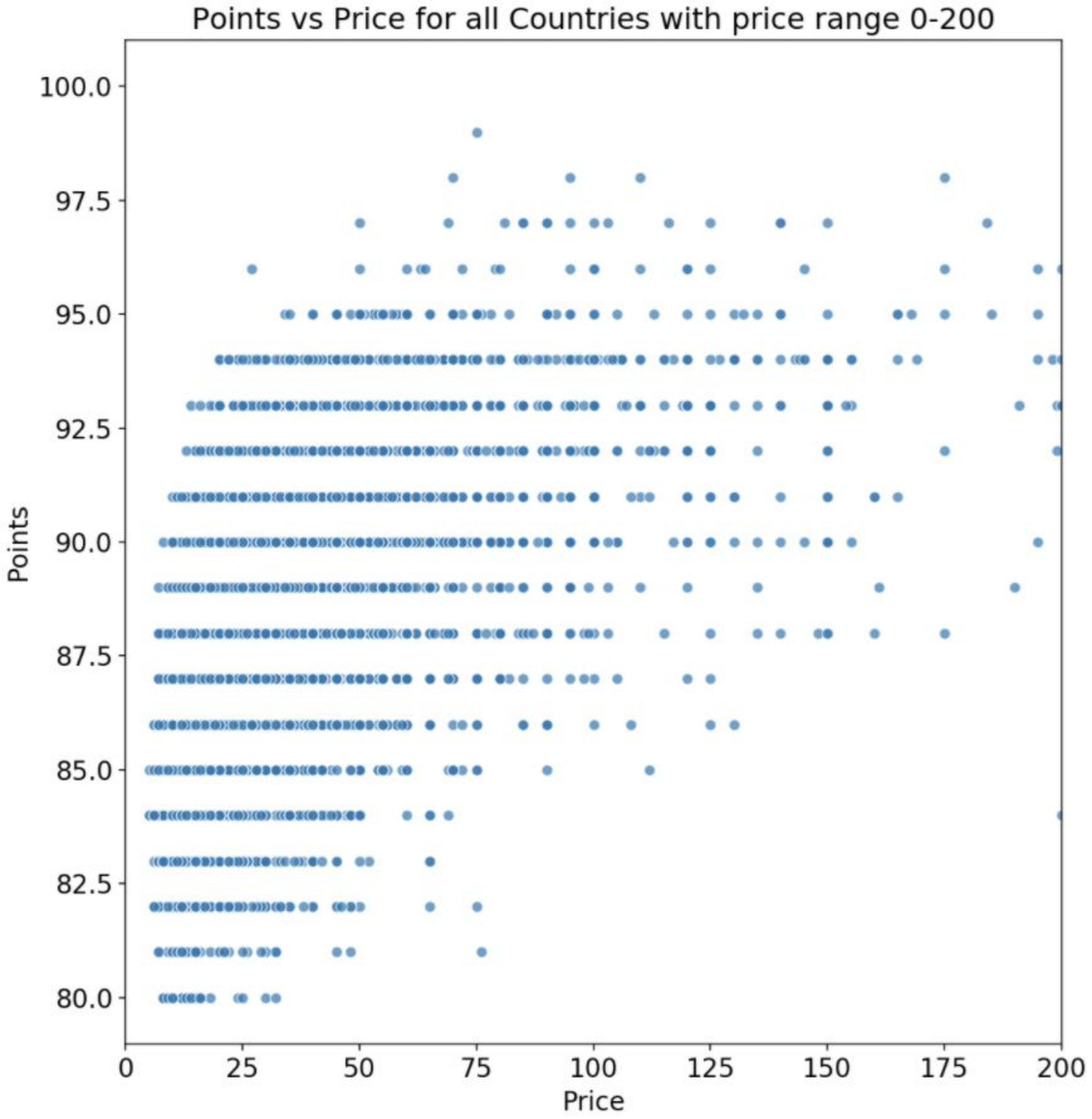
Majority of the wine prices are less than 50. Median of the price is 25 and the mean price is 35.5. Price is an indicator of wine's quality and age which is expected to be an important predictor for the model.

Points and Price



This is a plot of points and price relationship for the wines produced by the US, Portugal, France, Italy, Spain and Chile. There are two wines from Italy and one from France that received full points from tasters, all of their prices are greater than 200.

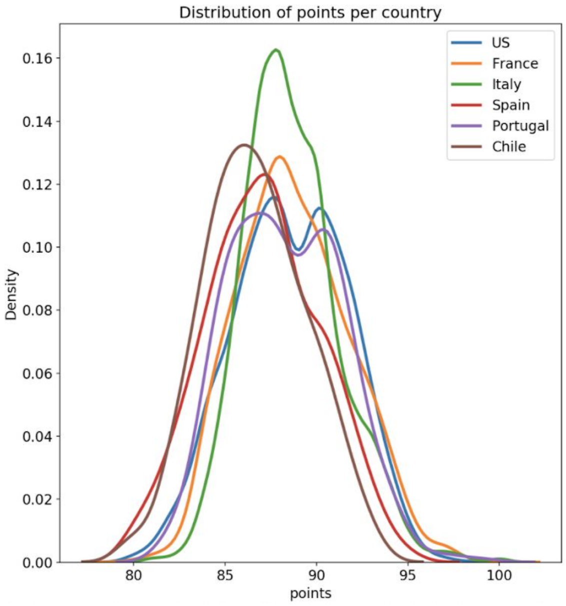
If we zoom in to the 0-200 price range for all countries' data:

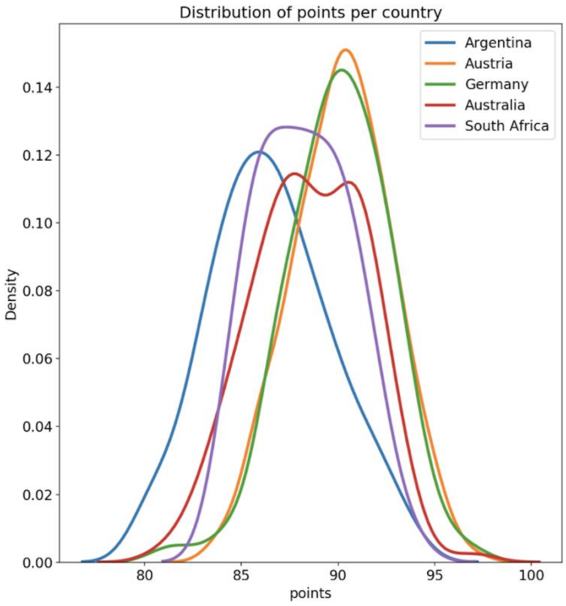


Both plots show that there is a positive trend between points and price, which serves as a proof that price will be an important predictor for the model.

Points per Country

In order to determine if certain wines tend to score better or worse on the points, we can plot the distribution of points by country. Following is a ***density plot*** showing the distribution of points for the countries with more than 100 occurrences in the training dataset. The actual values in a density plot can be difficult to interpret, however, it is more instructive to focus on the distribution/shape of the figure.





Wine producing market is dominated by the US, France, Italy and Spain which is easily available from the plots. Looking at the shapes above, country is a differentiating feature to determine points of a wine. However, there are some exception country pairs that have same points distribution:

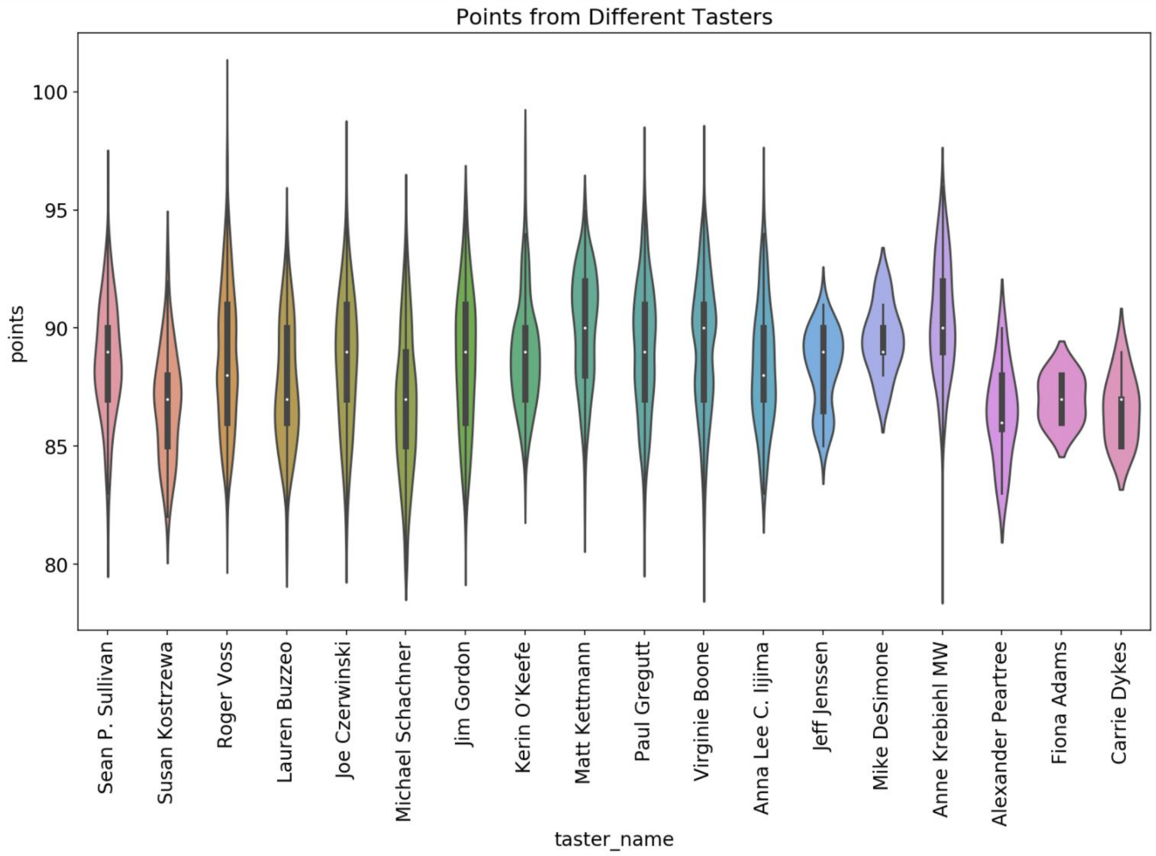
- USA - Portugal
- Austria - Germany
- Argentina - Chili

All in all, country provides location information in an accurate way when it is combined with a province, so both are going to be kept as predictors for the model.

Points per Taster

Target value, points, represents the score of a particular wine received from a taster. So, a different taster might imply different interpretations of a wine. Following is a violin plot of points for different tasters.

A ***violin plot*** conveys the distribution information with the shape outside and provides descriptive statistics (as summary of the statistics) with the slim rectangles inside. The median is represented as a white dot inside the rectangle.



Each taster's descriptive statistics and distribution of points are unique resulting in different violins above. Moreover, considering taster’s are describing a wine, it could be the second most important predictor of the model.

In addition to explored features above, variety and region_1 are a major element in differentiating grapes so as the wine's taste and quality, so they will remain in the final feature list.

Highlights from the Data Exploration:

1. 95% of the wine points are within the 82.5 - 94.5 range and points are normally distributed between 80 and 100.

2. Price is positively correlated with the points and it might be the most important predictor in determining points.

3. Taster name might be the second most important predictor in determining the points.

Feature Engineering & Pre-processing

Feature engineering and pre-processing is a process of extracting, transforming and removing features.

Machine learning models can only work with the numerical data and non-missing values. So, missing values and categorical columns were transformed using following steps:

1. Feature Extraction:

   - Description contains information about the wine's color, taste and notes (like citrus, tannins). Taste and color related words were searched and extracted from description as new features: is_red, is_white, is_rose, is_dry, is_sweet and is_sparkling. If a taste or color related word exists in the description, the corresponding feature to that word was assigned with value of 1, otherwise 0.

   - Title feature contains the production year of the wine. Year was searched and assigned to year feature, if year information is not available in the title year is assigned to 0.

   - Variety has the information about whether different types of grapes are blended. Is_blend feature is assigned to 1 if several grape varieties present, otherwise 0

   When the extraction was completed, description and title are removed from the feature set.

2. Categorical Feature Transformation:

   A ***categorical feature*** is a non-numeric feature. Categorical features (country, province, region_1, taster_name and variety) will be turned into numerical data, by using process of ordinal encoding. ***Ordinal encoding*** is the process of assigning positive integers consecutively starting from 1 to each unique value of a particular feature.

   For example, for the country feature US is assigned to 1, South Africa is assigned to 2 and Portugal is assigned to 3. Process is completed when all country values are assigned with an integer.

3. Missing Value Filling (Imputation):

   Missing values can be handled in several ways, for our model, the selected approach is imputation. ***Imputation*** is the process of filling missing values with several strategies. Following strategies are used for different categorical features:

- taster_name: imputed with constant value as 0, as stands for "Unknown taster"

- price and year: imputed with median value since this value is not affected by the extreme values.

- country, province, region_1 and variety: imputed with most_frequent value, since the most_frequent values of each feature aligns. Those are US, California, Napa Valley and Pinot Noir respectively.

One important consideration during imputing is, whatever imputing strategy is selected, both training and test set are imputed with the same imputed values. For instance, the median of the price in the training dataset was 25, 25 is assigned to missing values of the training dataset and the test dataset to prevent data leakage into the test set.

After feature pre-processing and engineering, the determined feature set is:

- Country
- Province
- Region_1
- Variety
- Price
- Year
- Taster name
- Is red
- Is white
- Is Rose
- Is dry
- Is sweet
- Is sparkling
- Is blend

Building Wine Rating Predictor

Wine rating predictor will be a ***supervised regression machine learning model*** built on the training set with the determined feature set.

- ***Supervised:*** predictions on the ratings will be generated using a defined feature set and target.

- ***Regression:*** the target is a continuous variable, within a range of 80 and 100.

Before building the predictor, an evaluation metric is set to measure the performance of the model. Besides, a baseline metric is determined to serve as a measure to reply to the initial question: “how good the predictor is in predicting points of a wine?”

Set Evaluation Metric

The evaluation metric is: ***Mean square error (MSE).*** It is the average of the sum of squared residuals where a ***residual*** is the difference between the actual and predicted value of a target variable. In other words, evaluation of the model is done by looking at the measure of how large the squared errors (residuals) are spread out.

MSE is selected because it is interpretable, analogous to variance and aligns with the selected model’s error minimization criteria.

Establish Baseline

A ***baseline*** can be explained as generating a naive guess of the target value by using expert knowledge or few lines of code. If the built-model (wine rating predictor) cannot beat this baseline, then the selected machine learning model may not be the best approach to solve this problem or whole pre-processing steps need re-consideration.

For the rating predictor, a simple baseline is to predict the variance of the mean of the training set to the test set. This approach aligns with our evaluation metric MSE as well.

Baseline MSE is calculated as 9.01. This shows that on average variance between the training points and testing points are 9.01. In other words, the sum of squared residuals of average training points to the validation points is 9.01.

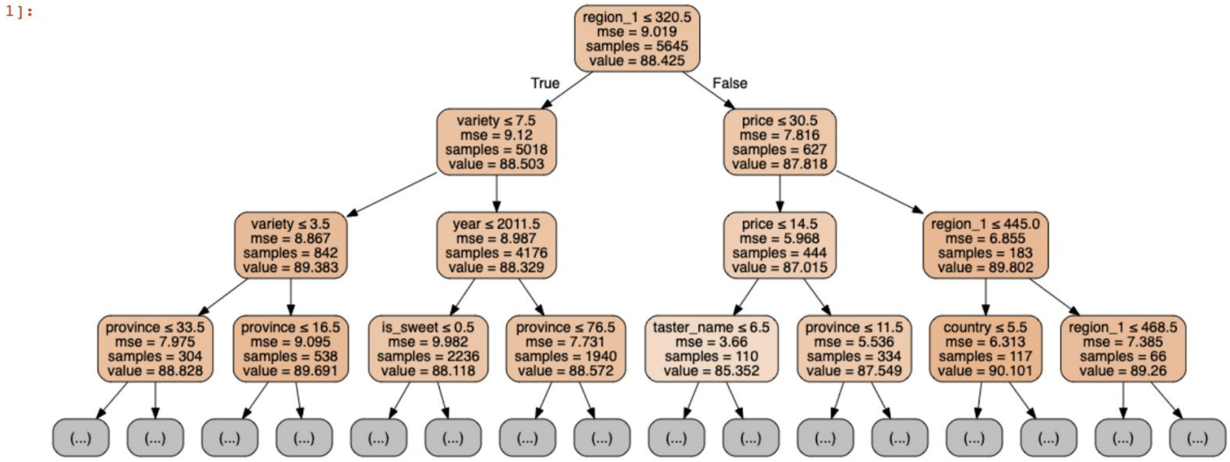
Predictor Model: Random Forest Regressor

Random forest regressor is an ensemble model built on multiple decision trees. In order to understand how the model works, the reasoning behind the decision trees should be understood.

Decision tree is a tree-like structure and uses this structure to make predictions. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the feature tested. Leaf node represents a final decision on the target.

Random forest regressor is an algorithm that builds multiple decision trees once and trains them on various sub-samples and various subsets of the features of the dataset. Its random selection of the dataset and features subsamples makes this algorithm more robust.

Visualization of a single decision tree from the wine rating predictor



Here is a series of yes/no questions asked and answered like a flowchart by a single decision tree visualized by the first 3 layers of nodes.

Root node asks if the region_1 is less than 320.5 and answers this question using a subset of the training dataset of size 5645. It can be easily realized that the initial MSE is the same as the baseline estimate of MSE. Predicted points is 88.425.

The false child node of the root node asks if the price is less than 30.5 and answers this question using a subset of the training dataset of size 627. The MSE improved slightly compared to the previous node, as an indicator of a more coherent node. Predicted points is 87.818.

This process continues iteratively until a leaf node is reached. When a node's MSE does not improve further or when the specified criteria for reaching a leaf node is met, the process stops.

Wine Rating Predictor (Model) Evaluation

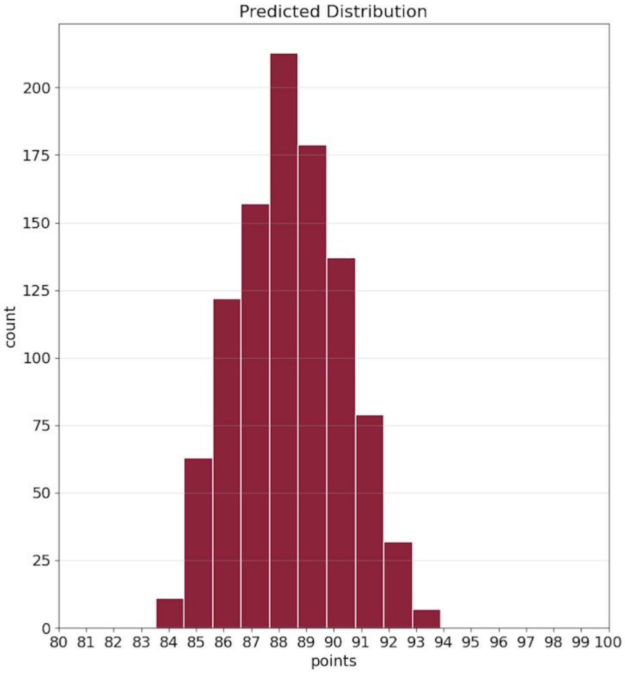
Before evaluating the model, the actual target values are held-out from the test set in order to utilize as a comparison to the predicted values.

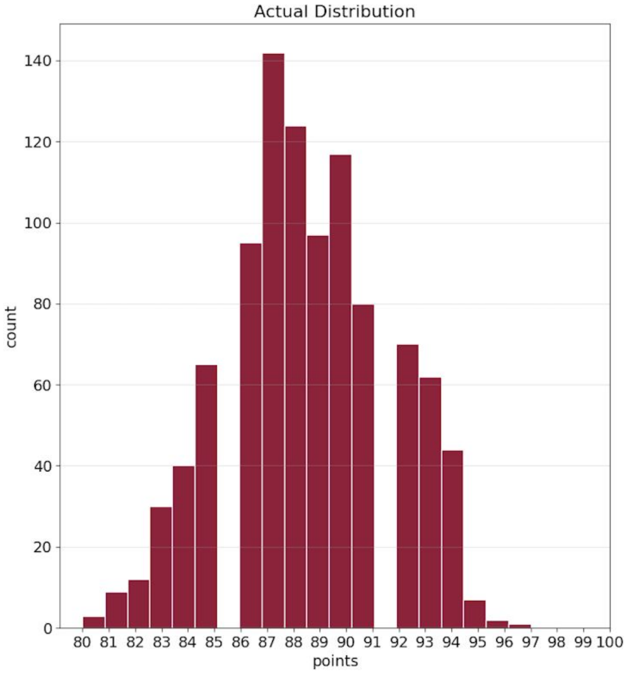
When the test set is inputted into the built-model, predictions are generated in return. When the MSE of the predictions are calculated, it has lowered from 9.01 to 4.9 showing 45% improvement.

This is a significant improvement from the baseline estimate. Moreover, it serves as a proof that the built-model (or wine rating predictor) is a good predictor of wine points.

Distribution of Predictions and Real Target Values

Following is a histogram of predictions and actuals:



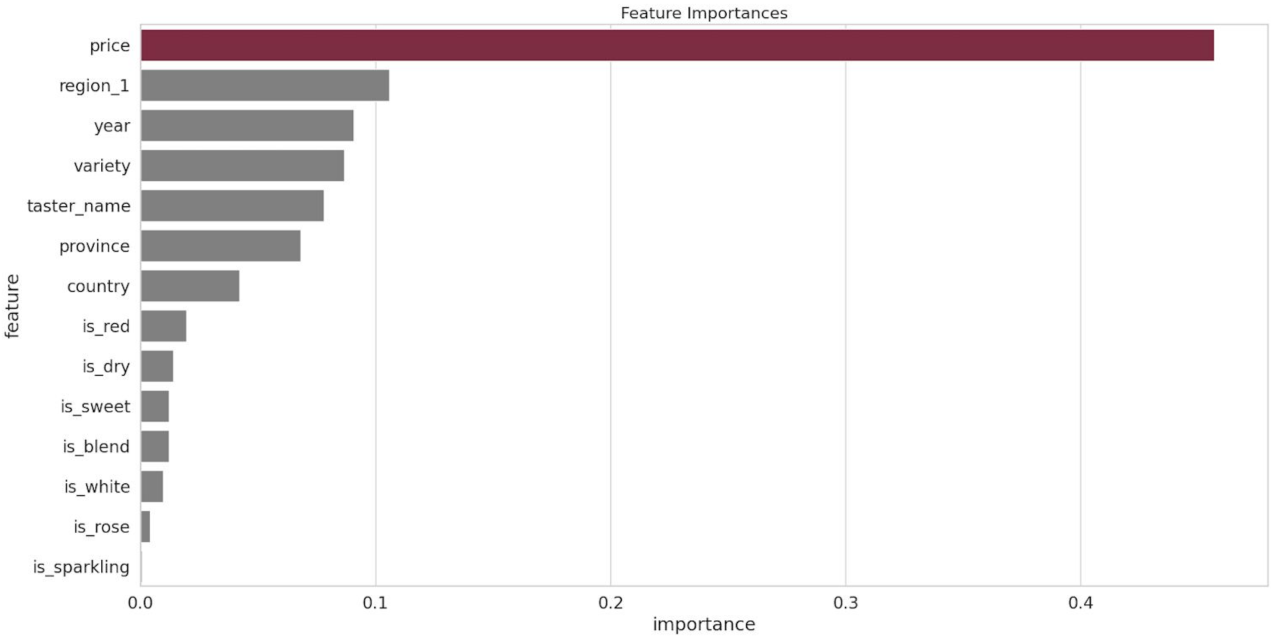


The model can predict the points between 84 and 94, but is not able to predict the points below 84 and points above 94 which are less-frequent values of the distribution. This is also the reason for the narrower distribution of the predictions.

This is a further improvement area of the wine rating predictor, but for now it has been shown with the 45% improvement that the model built with the current set of features is a good predictor.

Feature Importances

Another way to evaluate the model is to look at which features that the model considers most important. Following is the a number for the relative contribution of each feature in determining a point.



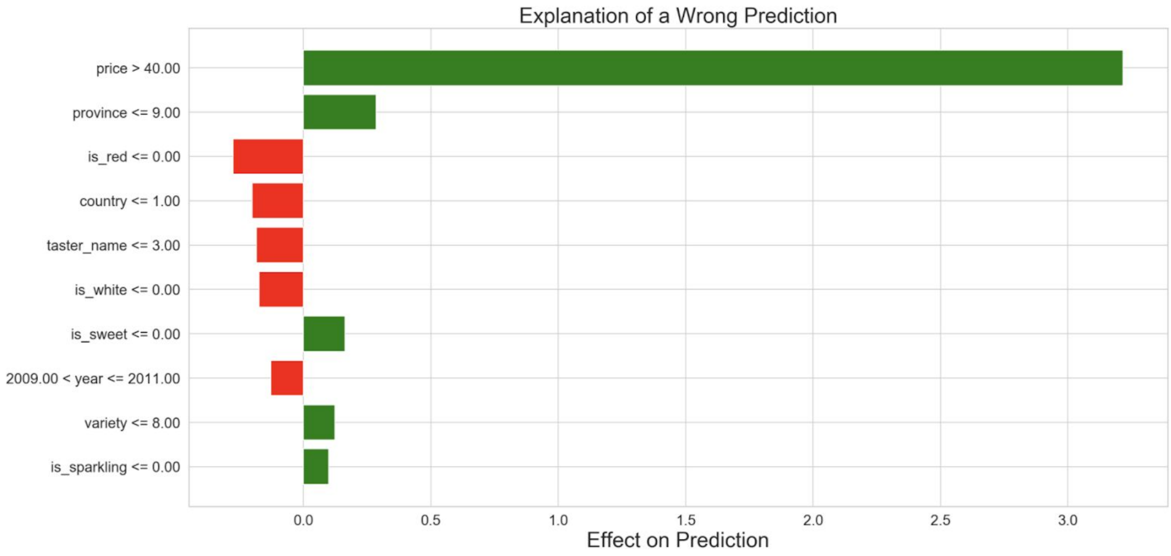
According to the model, the strongest predictor is price. The following most important features for predicting the points of a wine are: region_1, year, variety and the taster name.

These are inline with our preliminary observations developed during the Data Exploration section. From the features extracted, year is the most important among them and whether a wine is sparkling or not is the least important predictor of points.

Another significant observation is, note and taste related features extracted from variety and description features become moderately-important features for our predictor. A comprehensive text-mining and sentiment analysis may help to find more valuable features from description and variety features.

Detail Look at the Wrong Prediction

Following plot explains how the model concluded a wrong prediction. It predicted 90 points whereas the actual point was 82. This is the wrongest prediction of the model:

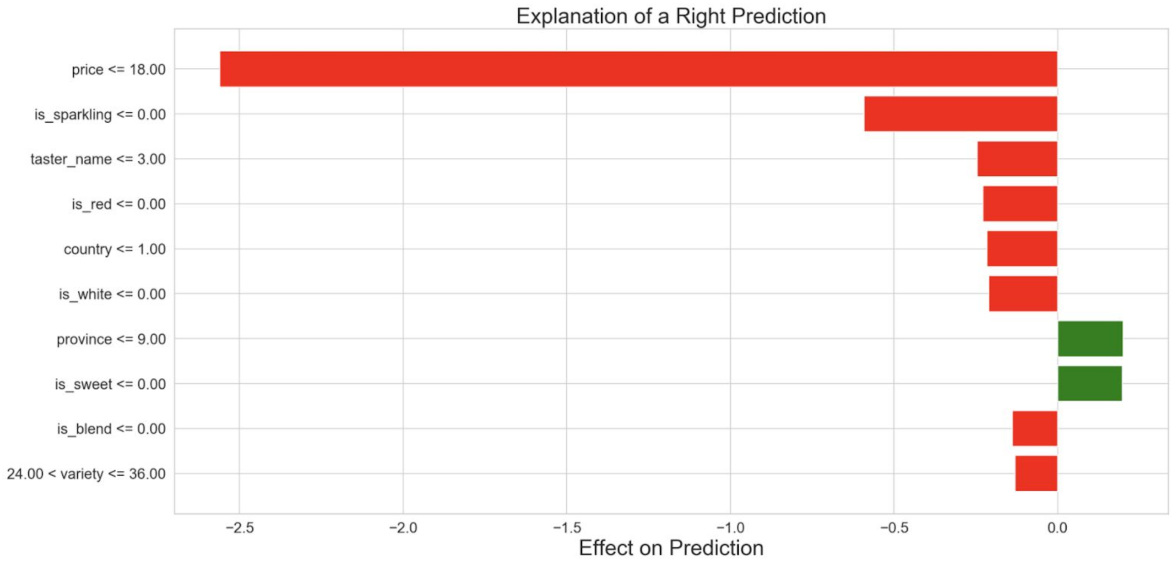


This shows the contribution of the feature to the wrong prediction example. Positive contributions are colored in green and negative contributions are colored in red.

As the price being the strongest predictor as a feature and having a price value above 40, significantly increased the model's prediction. On the other hand, country, is_red and is_white balanced the predictions by decreasing it. This wrong prediction can be interpreted as we humans also fall into an error of “expensive wines are better”.

Detail Look at the Right Prediction

Following plot explains how the model concluded a right prediction. It predicted 87 points, whereas the actual points were 87. This is the most correct prediction of the model:



The plot again shows the contribution to the prediction of each feature for the right example. Price being lower than 18 decreased the predictions, so the other features except province and is_sweet.

Conclusion

We set out to answer the question: Can we build a good wine predictor using machine learning models to predict points of a wine and which set of features can be used for that

Given the sample dataset, determined set of features and the trained random forest regressor, the answer is **yes,** a full-production solution is applicable with those. The model significantly lowered (45%) the baseline estimate, serving as an evidence to the positive answer to the initial question.

The two most important findings of the wine rating predictor is:

1. Wine rating predictor can infer the points of a wine with a reasonable variance (MSE) of 4.9.

2. The most useful predictors are determined as price, region_1, year, variety and the taster name.

Further Improvement Areas:

1. Extending prediction range: so that the less-frequent target values can be predicted by the model. The current sample dataset is dominated by wines from the most dominant countries (US, Italy, France, Spain), feeding more data from non-present countries might help the model to learn better the determiners of the less-frequent points.

2. Using NLP to extract stronger predictors: currently, the strongest predictor is “price”. Extracting more features from the description related to wine taste and notes can help create strong predictors like price, decreasing prediction error of the model.