Booz | Allen | Hamilton

PREDICTING WINE POINT REVIEWS

Data Science 5k

Evan Rosa

December 2018

PRESENTATION OVERVIEW

- Problem Statement and Data Review
- Data Cleaning & Dictionary Review
- EDA and Trends
- Modeling the Data
  - Linear Regression
  - NLP Random Forest Regression
  - Logistic Regression
- Final Thoughts and Next Steps

PROBLEM STATEMENT AND DATA REVIEW

Problem Statement

Using a wine review dataset, we want to see if we can develop a model that can predict wine point reviews.

Data Review

- Data for our models were scraped from WineEnthusiast
- The original dataset has about 13 columns and 130k rows.
  - Each row represents a single wine review for one bottle of wine.
- Contains mostly descriptive data
  - Country, Designation, Province, Region 1, Region 2, Variety, Winery, Description etc.
- Only 2 numeric data columns.
  - Price and Points

DATA TAKEN FROM KAGGLE

DATA CLEANING & DICTIONARY REVIEW

Data Cleaning

- Changed points data from an object to an integer
- Changed price data from object to float
- Changed accent marks on variety of wine (i.e. Rosé vs Rose)
- Dropped all null values
- Created dummy data points and concatenated them to our original dataset.
  - Dummy data includes: Country, designation, province, region 1, region 2, and variety

Data Dictionary

Please see below for all of the columns within the dataset. To review the data dictionary please follow the link below.

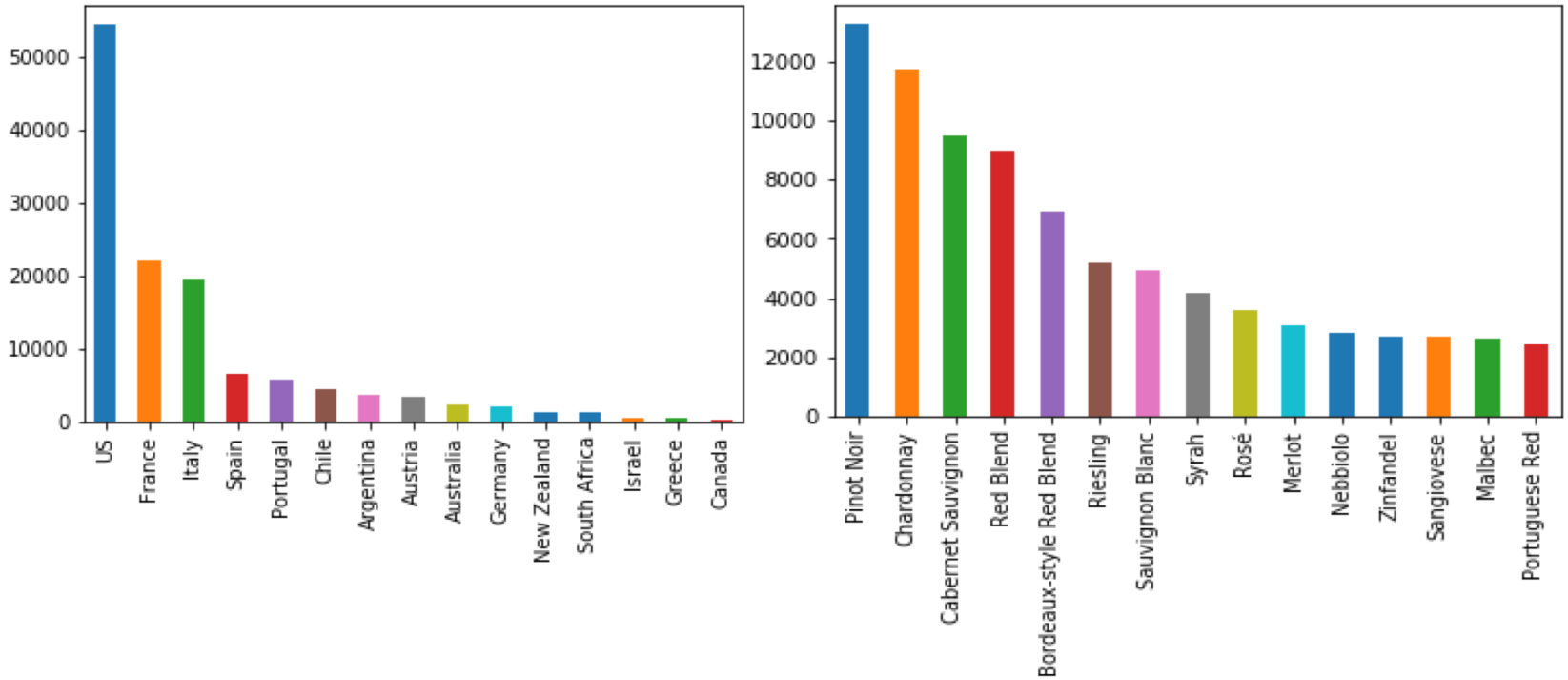
- Country, Points, Price, Province, Region 1, Region 2, Taster Name, Variety, Winery

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

EXPLORATORY DATA ANALYSIS

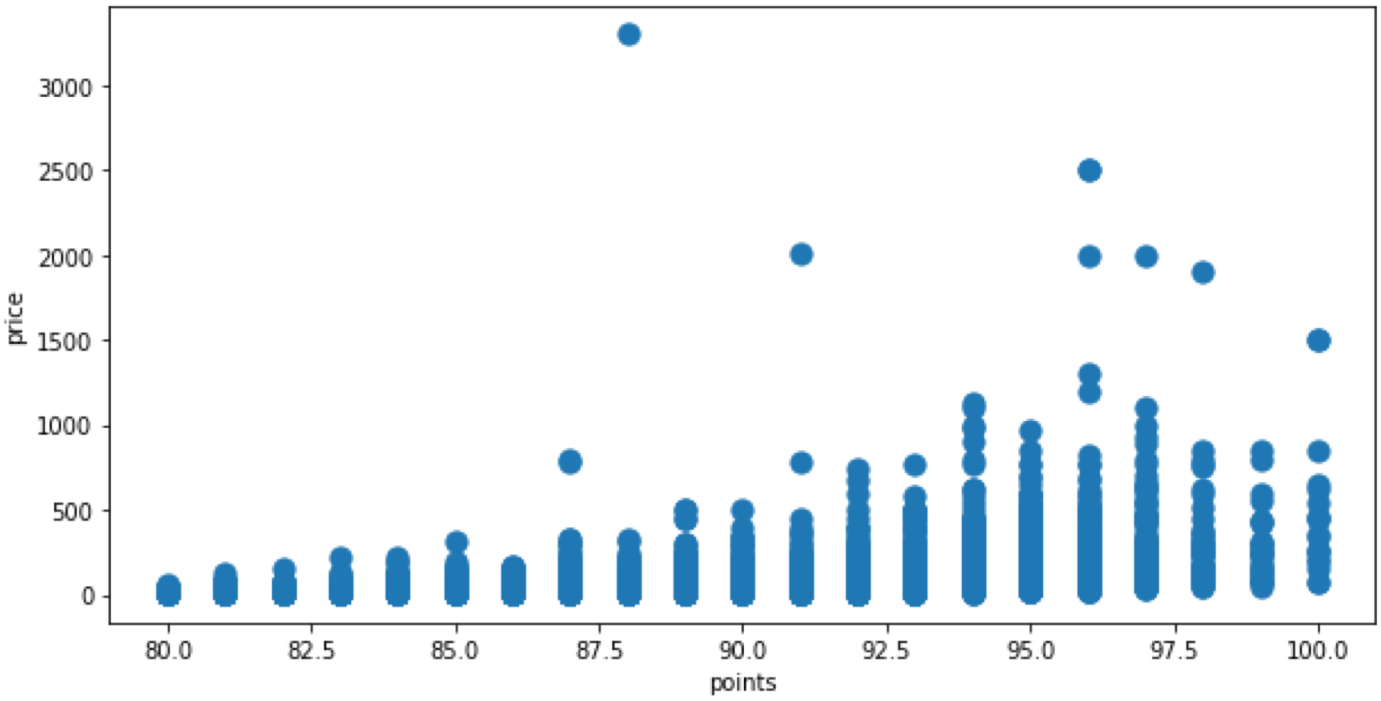
EXPLORATORY DATA ANALYSIS & TRENDS

In our initial EDA we take a quick look at the distribution of wine varieties and countries they were made in.



EXPLORATORY DATA ANALYSIS & TRENDS

The information below includes price and points data and the it shows a pretty good distribution. That said, it looks like as points get closer to a perfect 100 the more expensive wine becomes.



EXPLORATORY DATA ANALYSIS & TRENDS

The average points review for all wines is 88 out of 100 while the average price is $35. The cheapest wine is $4 while the lowest review is 80 out of 100. Conversely, the most expensive wine is $3300 while the highest review is 100.

	Points	Price
Count	129,971	120,975
Mean	88.45	35.36
Std	3.04	41.02
Min	80	4
25%	86	17
50%	88	25
75%	91	42
Max	100	3300

MODELING THE DATA

MODELING THE DATA: LINEAR REGRESSION

The overall idea is to use linear regression, to predict points review based on categorical or numeric columns. Based on correlation data, price and variety of wine have the highest correlation to points review.

Linear Model 1:

Using Price as my only feature.

- R-squared: .17
- RMSE: 2.74
- Y-intercept: 87.37

Linear Model 2:

Using Price and All Varieties of wine.

- R-squared: .24
- RMSE: 2.66
- Y-intercept: 84.77

Linear Model 3:

Using all categorical and numeric columns.

- R-squared: .24
- RMSE: 2.66
- Y-intercept: 84.77

Conclusion: When trying to predict wine point reviews, when compared to the null RMSE of 3.04, our best linear regression model was model 2 with a RMSE of 2.66. That said, linear models are not a reliable model due to our low R-squared value of 0.24.

Note: Linear Model 3 is not a typo. The reason it's not the best model is because it takes up too many computing resources. In this case, adding more data then needed isn't good practice.

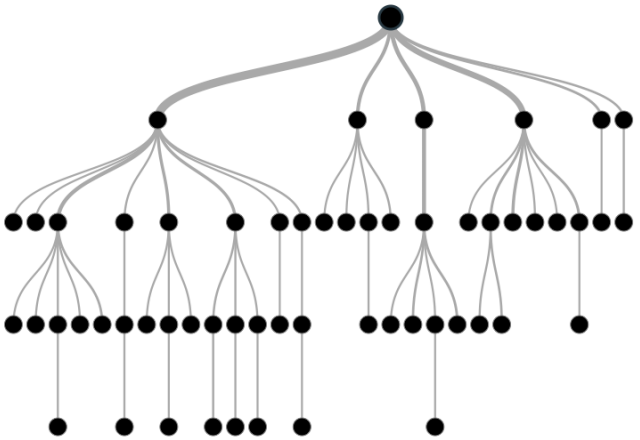
MODELING THE DATA:
NLP RANDOM FOREST REGRESSION

A random forest model predicts the outcome based on a series of decision trees. And a decision tree is basically a flow chart that uses a branching method to illustrate possible outcomes of a decision. In this model we use description to try and predict wine points review.

Random Forest Model:

Using description of wine.

- RMSE: 2.29
- Out of Bag (similar to R2): 0.5



Conclusion: The Random Forest model performed better than our linear regression model with a RMSE of 2.29. That said, our out of bag was 0.5 which doesn't make this model ideal for real-world applications. The argument can be made that adjusting the tuning parameters could help the model perform better, however, I wasn't able to do so due to the amount of computing resources it required.

MODELING THE DATA: LOGISTIC REGRESSION

Unlike the previous regression models, logistic regression doesn't try to predict the value of a numeric variable (in our case points review). Instead, we get a probability that the given input point belongs to a certain class. So, sense we can't predict the value of points let's try to predict good and bad wine.

A good wine is defined as any wine rated above 92, which is 1 rating point above the top 75% projectile.

Using price and variety of wine as our main features.

- Accuracy Score: 0.91

Conclusion: When trying to predict good or bad wine, when compared to the null accuracy of 90%, we get an accuracy score of 91 which is the percentage of correct predictions. That said, if you compare this to our null value you can see that it only does slightly better with only a 1% increase in accuracy. Moreover, the assumption can be made that if we were to increase our threshold for a good wine, to let's say 95, we would expect a more accurate prediction.

FINAL THOUGHTS AND NEXT STEPS

Final Thoughts

- Feature importance for each model didn't changed too much from model to model. Each model listed the most important characteristics of wine as price and the various varieties of wine.
- The point review is not easily predicted

Next Steps

- Research other regression models to try and predict the points review.
- Gather more data to include in the wine dataset, maybe a full wine review that describes if the wine tastes good or not.

THANK YOU!
ANY QUESTIONS?