Wine Reviews: Predicting Price and Points Using Characteristics of Wine

Department of Biostatistics & Bioinformatics
Duke University School of Medicine

Department of Biostatistics and Bioinformatics, School of Medicine, Duke University
Yiwen Liu, Benji Wagner, Wendi Xiao

Background and Objectives

- The data was scraped from WineEnthusiast during the week of November 22nd, 2017 and uploaded to Kaggle[1]
  - Contains 119,988 entries of wine with 14 variables (**Table I**)
- **Objectives**:
  - Predict price/rating points using only points/price via GAM univariate models
  - Predict price/points using different machine learning models such as Lasso/Ridge, Random Forest, Gradient Boosting Machine (GBM) and Support Vector Machine (SVM)
  - Compare different models

Table I: Description of variables in the Wine Review dataset

Variable	Type	Levels/Range with Missingness (%)
country	Categorical	43 levels with 59 NAs (0.05%)
description	String	0 NAs (0.0%)
designation	Categorical	37979 levels with 34545 NAs (28.8%)
points	Numeric	80 - 100 with 0 NAs (0.0%)
price	Numeric	$4 - $3300 with 8395 NAs (7.0%)
province	Categorical	425 levels with 59 NAs (0.05%)
region_1	Categorical	1229 levels with 19560 NAs (16.3%)
region_2	Categorical	17 levels with 73219 NAs (61.0%)
taster_name	Categorical	19 levels with 24917 NAs (20.8%)
taster_twitter_handle	Categorical	15 levels with 29446 NAs (24.5%)
title	Categorical	118840 levels with 0 NAs (0.0%)
variety	Categorical	707 levels with 1 NA (<0.01%)
winery	Categorical	16757 levels with 0 NAs (0.0%)

Exploratory Analysis

- **Variable Selection and Transformation**
  - Removed description, designation, region_1, region_2, taster_twitter_handle and winery based on missingness (**Figure 1**)
  - Merged country and province into location
  - Extracted production year from title
  - Converted points to Percentile Points
- **Missing Data Manipulation**
  - Converted NA values in taster_name into "Missing" level
  - Imputed price based on variety

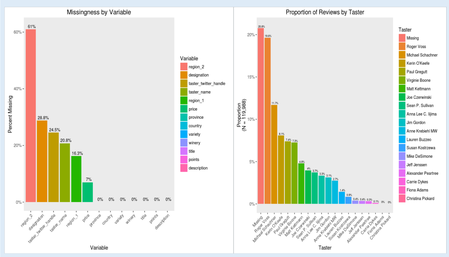

Figure 1. Bar chart of Missingness by Variable (**left**) and Bar chart of Proportion of Reviews by Taster (**right**).

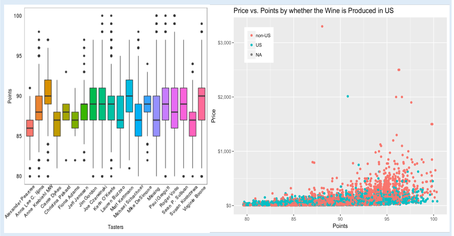

Figure 2. Boxplot of wine points given by each taster (**left**); and price vs. points by whether the wine is produced in the US (**right**).

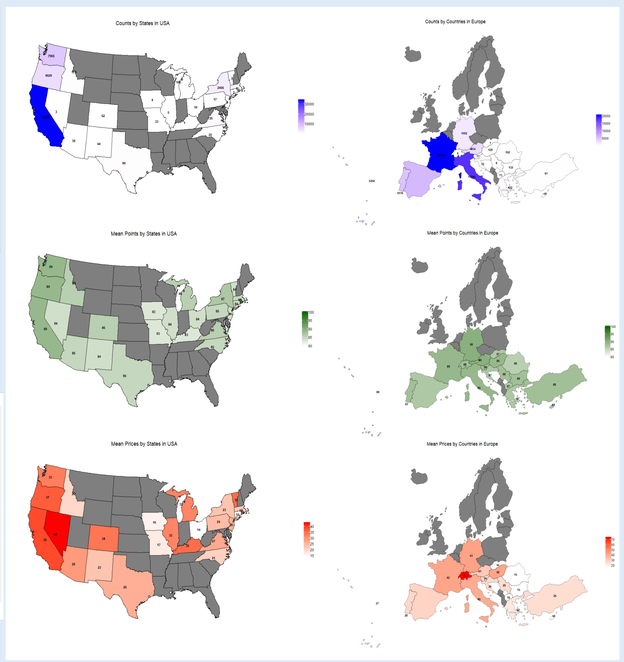

Figure 3. The counts(**top**), mean price (**middle**), and mean points (**bottom**) of states in USA (**left**) and countries in Europe (**right**).

Methods

- **Split** dataset into training and test set, pre-process and clean train and test set separately
- **Univariate models** were fitted for predicting price using only Percentile Points and vice versa and f**our multivariate models** each were fitted on the training set in our analysis for predicting price or points:
  - LASSO/RIDGE – allows for coefficient regularization, mitigate overfitting problem, useful in prediction models
  - Random Forest – effective at modelling non-linear and heterogenous effects, good with categorical and continuous variables
  - GBM – builds additive model in a forward fashion, allows for optimization of given loss function
  - SVM – effective in high dimensional space, use subset data in decision function, memory efficient

Results

- Univariate models for predicting price/percentile points performed worse than multivariate models
- Among all models, SVM models performed the best for predicting both price and points

MAE_price	MSE_points
16.44	589.35
17.23	588.39
16.23	593.43
15.96	589.34
15.93	601.68
16.14	575.25
16.12	586.48
16.03	592.96
15.83	580.97
16.39	575.75

Lasso	Ridge	Random Forest	GBM	SVM
14.78	14.00	24.08	15.64	**12.66**
14.79	14.00	24.08	15.57	**12.72**
14.80	14.00	24.08	15.73	**12.63**
14.78	14.04	24.08	15.71	**12.87**
14.78	14.00	24.08	15.61	**12.54**
14.78	14.00	24.08	15.79	**12.70**
14.77	14.04	24.08	15.65	**12.81**
14.81	13.98	24.08	15.68	**12.77**
14.79	14.00	24.08	15.71	**12.82**
14.79	14.00	24.08	15.64	**12.66**

Lasso	Ridge	Random Forest	GBM	SVM
6.66	6.69	1254.20	7.09	**5.07**
6.61	6.69	1254.20	7.10	**5.06**
6.58	6.75	1254.22	7.18	**5.04**
6.69	6.72	1254.20	7.28	**5.05**
6.64	6.69	1254.23	7.27	**5.03**
6.60	6.67	1254.26	7.40	**5.05**
6.60	6.69	1254.20	7.23	**5.05**
6.63	6.72	1254.20	7.17	**5.06**
6.58	6.69	1254.21	7.24	**5.05**
6.58	6.69	1254.26	7.13	**5.04**

Table II. (**left**) Mean absolute error of price (left column) and mean squared error of points (right column) in the 10 test folds via univariate models; (**middle**) Mean absolute error of price and (**right**) Mean squared error of points in the test folds via multivariate models

Interpretation and Conclusions

- The MAE for price is 13.76 and the MSE for points is 5.54 on the test set
- **Price Prediction**
  - Predicted prices were lower than observed prices (**Figure 4 left**)
  - Predictions for higher price were worse than predictions for lower price
  - Price absolute error and price were correlated for a given production year
- **Points Prediction**
  - Points were under-estimated across all ranges of price (**Figure 4 middle**)
  - No predictions for points above 97 and no observations for points below 80
  - Positive correlation between points given by tasters and points squared errors (**Figure 4 right**)
  - The number of observations per state greatly influences the performance of model (**Figure 5**)
- **Conclusions**
  - The SVM model is the best model for predicting both price and points
  - Both price & points were under-estimated

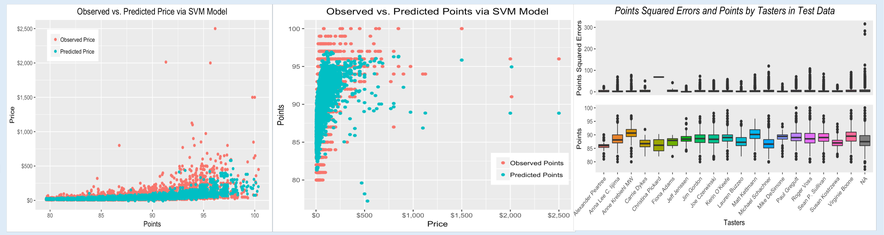

Figure 4. (**left)** Observed vs. Predicted Price by Points via SVM Model; **(middle)** Observed vs. Predicted Points by Price via SVM Model; **(right)** Points Squared Errors (top) and Points (bottom) by Taster in Test Data

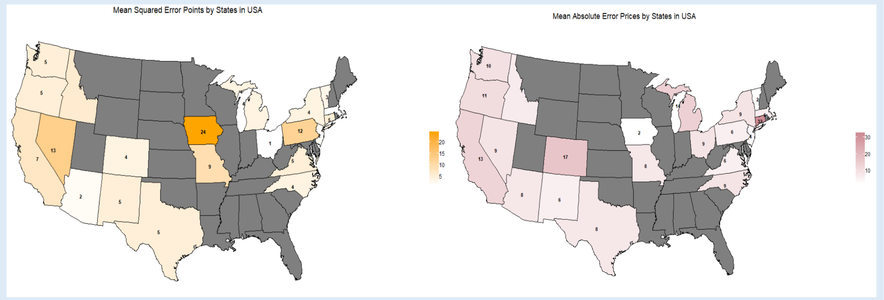

Figure 5. The mean squared error counts (**left**) and mean absolute error price (**right**) of states in the USA.

Future Work

- Include variables such as description (for natural language processing) and winery in the analysis
- Use domain knowledge to further group variables such as variety to reduce the dimension of the model

References

1. Dataset obtained from https://www.kaggle.com/zynicide/wine-reviews