SAS® GLOBAL FORUM 2018

USERS PROGRAM

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF

The Ultimate Data-Driven Guide to U.S. Canned Craft Beers Using SAS® Enterprise Miner™ 14.2

Author: Linyi Tang
Master of Science in Business Analytics, Oklahoma State University, 74075

Abstract

The explosive growth in the number of available online reviews has provided important guidance for shoppers who are considering the purchase of a product. However, the number of reviews and product choices can be overwhelming. In order to alleviate the problem of information overload, the ability to filter, emphasize, and efficiently deliver relevant information to the customer becomes crucial. Furthermore, product rating prediction based on reviews can be beneficial for online shopping portals to shape their recommendation system and for marketers to generate marketing strategies. Beer is one of the most popular drinks worldwide. In recent years, with the success of microbreweries, the breadth of beer options available is massive. In this study, we provide a data-driven guide to U.S canned craft beers and conduct rating prediction based on online beer reviews. Text mining was implemented to extract key words of interest.. Decision tree, linear regression, and k-means clustering were used and evaluated for rating prediction. Linear regression model was selected based on the least mean squared error.

Methodology

Figure1. Project Flow

The first data set 'craft canned beer" was obtained from Kaggle website. It contains 2410 US craft beers, 510 US breweries and was collected in January 2017 on CraftCans.com. The second data set 'Beer Advocate' was obtained from snap.standford.edu. Beer Advocate is a membership-based reviews website that beers are ranked based on various categories including overall, taste, aroma, appearance, and palate.

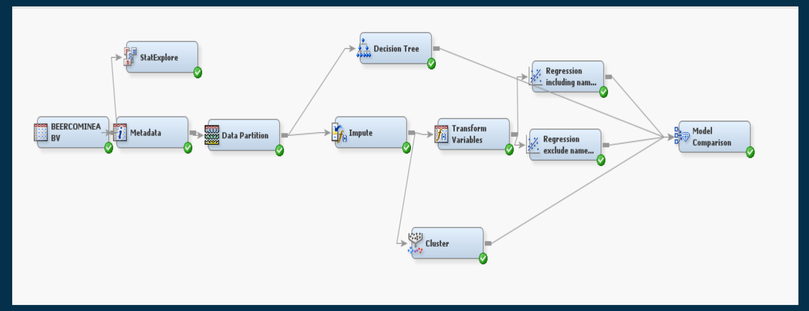

Figure 2. SAS Enterprise Miner 14.2 ® Project Diagram

SAS Enterprise Miner 14.2 ® is employed for data partition, data imputation, variable transformation and predictive modeling including decision tree, regression model and k-means clustering.

Results

I. Exploratory Data Analysis

The combined data set was used for exploratory data analysis. There are total 63,833 observations in the dataset. Top 20 beers were identified and sorted by average overall ratings.

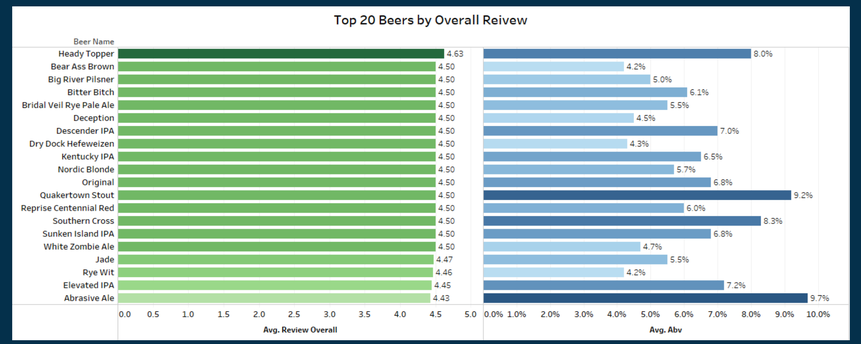

Figure 3. Top 20 Canned Craft Beer by Overall Rating

II. Predictive Modeling and Model Comparison

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F
1	review_taste	1	1	59912.7	<.0001
2	review_palate	1	2	4888.13	<.0001
3	style	68	3	14.11	<.0001
4	review_aroma	1	4	553.68	<.0001
5	review_appearance	1	5	322.14	<.0001
6	SQRT_IMP_abv	1	6	97.16	<.0001
7	PWR_IMP_ibu	1	7	25.03	<.0001

The selected model is the model trained in the last step (Step 7). It consists of the following effects:

Intercept PWR_IMP_ibu SQRT_IMP_abv review_appearance review_aroma review_palate review_taste style

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	74	11784	159.241412	1032.30	<.0001
Error	44628	6884.268285	0.154259		
Corrected Total	44702	18668			

Model Fit Statistics

R-Square	0.6312	Adj R-Sq	0.6306
AIC	-83480.4512	BIC	-83478.1961
SBC	-82827.3665	C(p)	74.0849

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
PWR_IMP_ibu	1	3.8607	25.03	<.0001
SQRT_IMP_abv	1	14.5537	94.35	<.0001
review_appearance	1	51.7606	335.54	<.0001
review_aroma	1	62.2831	403.76	<.0001
review_palate	1	602.6948	3907.03	<.0001
review_taste	1	2182.7827	14150.1	<.0001
style	68	103.6993	9.89	<.0001

References

•J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. ICDM, 2012.
•J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.


#SASGF
SAS®
GLOBAL
FORUM
2018
April 8 – 11 | Denver, CO
Colorado Convention Center