# Predicting and Recommendation Using Basic Linear Regression Models

Christopher Almajose
A10463176
calmajos@ucsd.edu

## ABSTRACT

With regards of a large data set and with the linear regression model, we are able to create predictors to determine the outcome of rating based on various labels (or features) into the model and evaluating the results as to whether or not certain labels (and the fitting of more labels) can in fact effect the results of the predictor.

### Categories and Subject Descriptors

[**Programming Languages**]: *Python*

**[Libraries]:** *sklearn, numpy, plotpy*

### General Terms

Algorithms, Data Analytics, Data Mining, Linear Regression model, (Data Compression

## 1. INTRODUCTION

Many of the webs recommendation systems and predictors used to be revolved on trying to implement faster algorithms to solve problems. Today, much of the algorithm design to perform these tasks are now able to be done when we are exposed to larger amounts of labeled data. Although there are very complex algorithms and models that can be done to complete those tasks, the most basic yet fundamental models to grasp are the simple ones such as the linear regression model.

As a fan of beer, the data set chosen for this demonstration is using the BeerAdvocate.com data set which contains various labels to help explain the how the linear regression model can help create a predictor on a specific label.

## 2. EXPLORATORY ANALYSIS

### 2.1 Data Set Information

The data set that was chosen, as previously stated, was taken from a website called BeerAdvocate.com taken from Stanford Large Network Dataset Collection and contains data with regards to labeled data, graphs and networks. The type of are data is just beer reviews that contain the following labels and sub labels:

1. Beer Labels:
    1. beer/name :
    2. beer/beerId
    3. beer/brewerId
    4. beer/ABV
    5. beer/style
2. Review Labels:
    1. review/appearance
    2. review/aroma
    3. review/palate
    4. review/taste
    5. review/overall
    6. review/time
    7. review/profileName
    8. review/text

Statistics in regards of the dataset include 1,586,259 reviews, 33,387 users, 66,051 beers, and times span from Jan. 1998 – Nov 2011. In regards to the specific fields we will be using being used for the assignment, the main labels being used in the linear regression model will be the review labels. Specifically, our truth vector for our labels will be focused around the 'review/overall' as it can sum up the overall impression of the beer based on the reviewer. However, the 'text' field will only be used in regards to the length of the label, not the contents itself.

### 2.2 Chosen Data for Analysis

Because the data is particularly larges, the data is restricted to a limited number of reviews and labels do due the memory constraints of the programming language. Thus, we will only be choosing at random

100,000 reviews. We choose at random because the data set given is mildly organized and choosing at random best represents real world data. This means that we only choose 6% of the data to operate on with 75% of the chosen data is used for a training set while the other 25% of the chosen data is used as a test set.

## 2.3 Statistics of Chosen Data

From the 100,000 random data points chosen, the following are some statistics of the data of the test set (75,000 points of data):

| Label/Rate | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 |
|---|---|---|---|---|---|
| Overall | 364 | 345 | 1448 | 2237 | 6443 |
| Appearance | 60 | 162 | 907 | 1515 | 6432 |
| Aroma | 0 | 398 | 1498 | 2291 | 7632 |
| Taste | 0 | 506 | 1516 | 2376 | 6114 |
| **Label/Rate** | **3.5** | **4.0** | **4.5** | **5.0** | |
| Overall | 12921 | 28702 | 17547 | 4891 | |
| Appearance | 13416 | 32791 | 16008 | 3707 | |
| Aroma | 15969 | 28258 | 14947 | 3801 | |
| Taste | 13544 | 26983 | 18624 | 5055 | |

**Table 1: Chosen Data Statistics**

The following is simple data revolving around the averages of the labels as well as its min/max ranges the data will be working with:

1. Avg Overall: 3.89729 (Min: 0.0 | Max: 5.0)
2. Avg Appearance: 3.9185 (Min: 0.0 | Max 5.0)
3. Avg Aroma: 3.834833 (Min: 1.0 | Max: 5.0)
4. Avg Taste: 3.90336 (Min: 1.0 | Max: 5.0)
5. Avg Text Length: 127 (Min: 3 | Max: 892)
6. Avg ABV: 7.2961 (Min: 0.1 | Max: 57.7)

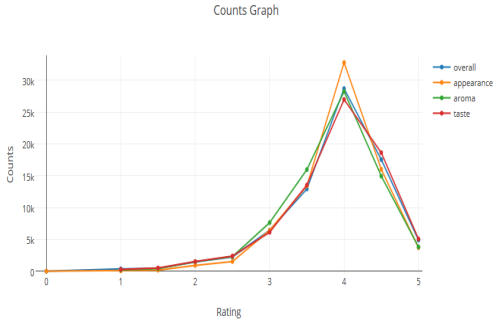


**Fig.1 Overall values of dataset**

# 3. Predictive Task

## 3.1 Overview

The overall predicting task is to first determine which features are the most important of all the labels that help correlate to taste. Then we use that information to help predict if the user will rate high on taste based on the new information as well as use the model to create a simple classifier for recommendation.

## 3.2 Value Choices for Prediction

### *3.2.1 Average as feature for prediction*

Taking the average of one label and using it as a predictor to predict taste is not very useful and doesn't help with gaining information on recommending to the user. We can join features and compress features (remove features that do not seem relevant) and see if it creates any differences in predicting, in particular, the taste rating. We will accomplish this u

We are able to calculate the Mean Squared Error (MSE) to calculate the error of our test set using the equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - Xi * \Theta)^2$$

Where x_i and theta represents the features and multipliers. The selected MSE against the test values of the average of the feature to predict itself with futre data:

| Labels | MSE (TRAIN) | MSE (TEST) |
|---|---|---|
| Overall | 0.475044 | 0.4767966 |
| Appearance | 0.47522 | 0.3451 |
| Aroma | 0.4797 | 0.4484 |
| Taste | 0.475003 | 0.486857 |
| Text (Length) | 15154.77 | 4922.487 |
| ABV | 11.9841 | 19419.509025 |

**Table 2: Avg MSE on feature to predict**

As the data show, not very exciting or accurate from the training and testing data to conclude that the average to predict even itself is enough to make a solid predictor.

### 3.2.2 *Bundled Features for prediction*

Because we really don't have any information of the beer, we can add the ABV of the beer as an automatic feature. Also, Overall seems to be 'unspecific' as it assumes what is the overall impression, not necessary the specific details. We can again run linear regression as well as calculate MSE upon using the features of ABV and Appearance/ Aroma/ Text Length as the bundle of features to determine if somehow correlate well to taste, as for beer drinkers, the taste is the most important part of the beer.

| Labels | MSE (TRAIN) | MSE (TEST) |
|---|---|---|
| Appearance | 0.33085 | 0.3336 |
| Aroma | 0.24432 | 0.24259 |
| Palate | 0.22789 | 0.227895 |
| Text (Length) | 0.43178 | 0.441124 |

**Table 3: Beer ABV and Additional label MSE to predict**

Upon evaluating table 2, we can see a significant difference of both the training and test sets when we bundle features together to predict the most important factors of a beer's: 'taste'. However, based on the scores of the test set of 25,000 data points against the training set calculated by using the coefficient of determination (a R^2 statistic) :

$$r^2 = 1 - FVU(f) = \frac{MSE(f)}{Var(y)}$$

FVU(f) = franction of variance unexplained. We can determine how accurate the predictors guessed as shown in table[4] where scores closer to 0 represent bad (trivial) predictors when values closer to 1 represent perfect predictors. By evaluating the results, we see that the labels 'ABV', 'Palate' and 'Aroma' corrolate the best when determining the taste of a beer.

| Labels | Appearance | Aroma | Palate | Text (Length) |
|---|---|---|---|---|
| Score | 0.31485 | 0.5017 | 0.5231 | 0.09405 |

**Table 4: Scores of predictors based on ABV and Label**

### 3.2.3 *Accumulation of features for prediction*

Finally, we can then bundle all the features and calculate how well all the features together can predict the taste. Upon the results of Table[5], we conclude that all the features together create a solid predictor with low MSE and score to predict somewhat accurately.

| MSE(train) | MSE(test) | Score |
|---|---|---|
| 0.178014 | 0.177228 | 0.636021 |

**Table 5: Results of predictors based on ABV, Palate, Aroma**

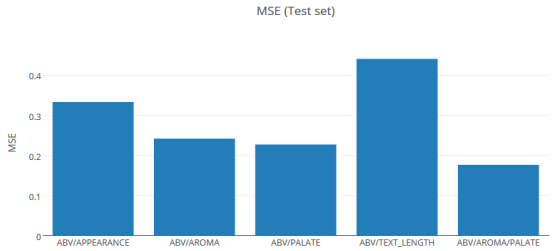


**Fig. 2: MSE of Mixed Labels**

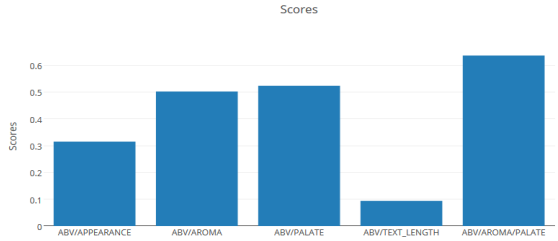


**Fig 3: Scores (Accuracy) of Mixed Labels**

### 3.2.4 *Simple Recommendation Predictor*

Now that the predictor some sort of label basis for predicting user taste to beer, we can then use our predictor and score value, to help decide if to recommend based on users previous ratings. We are, however stranded with users who have never rated anything before or have little rating information.

The following classification model determines if a certain beer will be recommended to a user:

$$r(u,i) = \begin{cases} 1 \text{ if } X_i \cdot \Theta > t \\ 0 \text{ otherwise} \end{cases}$$

where r(u,i) is recommendation of an item to a user, X_i is the item features, and t is a threshold value. The threshold value is used such the if the users previous reviews and features match closely to the items features that are also based on user reviews, then you can recommend the item. For calculating

this model, we first choose a user from our test data and determine its bias based on features related to taste. Then, any beers outside the training model, we can then obtain its feature matrix and truth values, and run them against the user. If the features scores pass the threshold, we guarantee that the MSE of the between the features are low, and then recommending something that may relate to the user.

As earlier stated, due to lack of information of new users input, our model becomes irrelevant. We can throw random feature to determine the recommending of a item.

## 4. Literature

### *4.1.1 Overview*

In regards to any literature to the study, there has not been very much studies due to the data only revolves around how others review beer. However, there have been ongoing studies revolving around how other senses of the body can react to other parts of the nervous system. An example would be of when one finds stimulus in the smell of a particular item, one may relate it the taste of the stimulus, whether or not it they are pleasant or horrid.

### *4.1.2 Dataset background*

This particular dataset was given to us via SNAP, that contains various large datasets involving networks, graphs, etc. There have been similar dataset used in the past that relate to this assignment and was used during the course. It was used to determine if ABV was related to the taste as well as time. However, the features were limited to the dimensionality of the features, and not bundled as this one suggests.

### *4.1.3 Models*

The main methods of the study uses the basic linear regression to help predict values of specific labels. The simple model involves around:

$$X\Theta = y$$

where X is the bundle of features, theta is our unknowns, which include an intercept and coefficients that alter our prediction of y, our truth table.

Also, to user the linear regression model, we also needed to other functions regarding the calculation of variance of data Var(y), mean squared error function MSE(f), and FVU(f) which helps explain the unknown of the variances.

The overall method we will be using to solve the study is the following:

$$taste = \Theta_0 + (\Theta_1 * ABV) + (\Theta_2 * palate) + (\Theta_2 * aroma)$$

The dimensionality is large as it contains three features, however, the assumption is that these features, based on previous calculations, does not effect the over fitting problem due to having multiple features, as many labels were left out if it had no effect on the taste of beer.

There are three different models that can be performed on the linear regression model.

1. Linear Least Squares
2. Ridge Regression
3. Random Forest Regression

Linear Least Squares is the most basic as it of linear regression as it only tries to find the best fit for the dimensions and result value. However, there is no regularization parameter to help smooth the results of this model

Ridge Regression, or Tikhonov regularization, is another method of linear regression the does have a Regularization value. With the regularization value, it helps smooth the noise of the data and fits the data better.

Finally was Random Forest Regression uses an ensemble learning method of machine learning. It applies the decision tree approach, determining predictions based on its neighbors. This algorithm is extremely slower, but possibly better if the data does not show any linear continuity.

## 5. Results and Conclusions

Upon reviewing the results of testing the training set against the test set, the predictor was able to predict a marginally successful rate. Due to this being a simple linear regression model, there were not any other opportunities to try against other various models.

| | MSE(train) | MSE(test) | Score |
|---|---|---|---|
| LST | 0.178014 | 0.177228 | 0.636021 |
| RIDGE | 0.178014 | 0.177228 | 0.636021 |
| RFR | 0.15852 | 0.1817246 | 0.626743 |

**Table 6: Results of predictors based on ABV, Palate, Aroma**

Based on the results of table 6, there weren't that many differences in the different models used. In regarding to the Ridge regression model, constraining the regularization constant or fitting the penalties to the targets did alter it, it varies based on each weight you give to the features.

In regards to the recommendation system, users who previously reviewed items get better recommendations due to previous input. Although it can limit the selection of the items recommended, it fits the needs of the users features. However, when users have never reviewed anything, they are usually at the fray of random features or, with an optimization, of mean values.

References

[1] Jure Leskovec and Andrej Krevl, SNAP Datasets: Stanford Large Net- work Dataset Collection . http://snap.stanford.edu/data, Jun, 2014

[2] J. McAuley and J. Leskovec, From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.

[3] BeerAdvocate.com, Provided datasets

[4] NIH News, Research uncovers Little-Known Impacts of smell and Taste on Health

http://www.nih.gov/news/health/jul2008/nidcd-15.htm

[5] Ratebeer.com - Developing Your Beer Palate for Beginners

http://www.ratebeer.com/Story.asp?StoryID=292