**Helping Brewers Predict Alcohol by Volume More Accurately**

**By Jerry Mischel, DS 740**

**Goal of Analysis and Practical Purpose for Target Audience**

In my final project for DS740, I analyzed a dataset of approximately [74,000 homebrewed beer recipes](https://www.kaggle.com/jtrofe/beer-recipes/home) from Brewer’s Friend to predict alcohol by volume (ABV) from various predictor variables. Currently, there is a standard formula for estimating ABV, which is the difference between the original and final gravities multiplied by 131.25. These gravity readings refer to the total amount of dissolved sugars in the beer before (original gravity) and after (final gravity) fermentation, the process which creates alcohol. My goal for this project was to build a more accurate predictive model (than this traditional method) to estimate ABV using appropriate data mining techniques. I also wanted to confirm whether original gravity and final gravity are the most influential factors to estimate ABV or if there are other variables contributing to the accuracy of my model(s).

If you’re a homebrewer, you have a passion for beer and want to invest the time to create a beer you like. Homebrewers also prefer a certain amount of alcohol in a beer, which reveals how much punch each bottle will pack and has an impact on the overall flavor and body. This predictive model could be valuable to homebrewers, who want to know if their investment will produce the beer with the ABV they want and what factors most influence ABV, as well as the American Homebrewers Association, which could arm homebrewers with this tool to help with members’ brewing process. Additionally, potential craft brewers would be another target audience, as they need to ensure a consistent ABV in their beers sold to the public.

**Background on the Dataset**

This dataset included 24 variables—a mix of both factor and quantitative. These variables included the wort density (or original gravity, OG, which contains the sugars that will be fermented by the brewing yeast to produce alcohol), final gravity (FG, the density of the fermented beer), boil time, boil gravity, color, brew method, efficiency (a measurement of potential fermentables converted into sugar in your wort) and measurements about the size of the batch. All variables from this dataset are described in the Variable Guide at the bottom of this document.

**Initial Data Analysis and Data Exploration**

An initial examination of the data revealed 7 variables which could not be used as predictors, including BeerID, Name, URL, Style, StyleID, PrimingMethod, and PrimingAmount. Style, PrimingMethod, and PrimingAmount would’ve been nice to include but there were too many styles of beer (176 in total) and the other two variables each had more than 69k missing values and inconsistent entries/scales. Out of the remaining variables, a significant number of missing values existed within BoilGravity, MashThickness, PitchRate and PrimaryTemp on the upwards of more than 25,000 each. Removing all missing values from these variables would’ve resulted in losing more than 50,000 observations in total and forcing only one BrewMethod to remain. As a result, I decided to remove the MashThickness variable which had the most impact on dwindling the four BrewMethods.

There were also inconsistencies with the SugarScale variable, as the values were based on two different scales (Plato and Specific Gravity). I used a conversion formula, found on [this website](https://www.brewersfriend.com/plato-to-sg-conversion-chart/), to adjust all Plato measurements into Specific Gravity to ensure ABV estimates weren’t incorrectly calculated. As a result, all observations were moved into the Specific Gravity category, so the SugarScale variable became obsolete as a predictor variable.

Finally, more than 200 observations had an ABV less than 2%. While there are folks who want to produce beer without the “kick” of alcohol, further investigation into the dataset and website (from which the data was pulled) revealed some incorrect entries for ABV, some entered as 0. As such, I decided to remove all observations with ABV less than 2%.

By the end of the cleaning stage, I was left with more than 31,000 observations, 12 predictor variables (11 quantitative and 1 factor) and 1 response variable (ABV) which are all highlighted in the Variable Guide. While I lost more than half of my dataset, I finally had a clean dataset with no missing values, proper scales for OG and FG, and plenty of observations left to create my predictive models.

**Benchmark ABV MSE**

Since the goal of my project was to provide a more accurate estimation of ABV through various data mining techniques than the traditional calculation formula described above, I needed to create a benchmark value from which to compare my models. Using my clean dataset of 31,000 beer recipes, I leveraged the ABV formula to estimate all ABVs and calculated the mean-squared error (MSE) rate between those estimations and the actual ABVs. This benchmark ABV MSE value was 0.01, which is already quite accurate.

**Determining Appropriate Data Mining Techniques**

To improve upon the accuracy of the traditional method to estimate ABV, I chose three different supervised learning data mining techniques for regression in my project – multiple linear regression, decision trees and artificial neural networks (ANNs).

*Multiple Linear Regression:* A multiple linear regression is used to explain the relationship between one continuous response variable (i.e., ABV) and two or more independent variables (i.e., OG and FG).  Since the data didn’t show much collinearity (or strong linear relationships) between predictor variables (see Figure 1 below), multiple linear regression was a good technique for this project. In Figure 1, you can also see a pretty solid linear relationship between OG and ABV, which lent well to this technique. Within this multiple linear technique, I fit 12 models using forward selection (i.e., starting with one predictor variable and ending with all 12 predictor variables), one model using the regsubsets function, which selects more influential variables based on predictive accuracy, and one model using just OG and FG, which are the two main variables used in the traditional method of calculating ABV. I hypothesized that the model just using OG and FG would be the most accurate out of these models.

As for the other two data mining techniques, the dataset showed heavy skewness and outliers within distributions for some predictor variables as well as non-linear relationships between almost all variables (except OG) and the response (see Figure 1 below). Both decision tree and ANN techniques allow for complex relationships, including non-linear relationships between the predictor variables and the response.

*Decision Trees:* Decision trees are a supervised learning technique which uses a tree-like framework to show potential consequences of decisions. For example, if OG is more than a certain value and FG is less than a certain value, then the model computes a decision pathway which outputs a certain value for ABV. Decision trees don’t require much data preparation (i.e., aren’t sensitive to outliers and skewness) and can help uncover the most influential variables contributing to the accuracy of the model. Within the decision tree technique, I chose three different methods – pruning, boosting and bagging – all of which have more sophisticated ways of building decision pathways and predicting values.

*Artificial Neural Networks:* ANNs are a supervised learning technique, inspired by biological neural networks, which take inputs (predictor variables) and produce outputs (predictions for a response variable like ABV) by leveraging a nonlinear model created by the hidden nodes (or layers) of the network which conduct the computations and weight the influence of each predictor variable. ANN models aren’t sensitive to outliers and skewness if data is normalized or standardized. Like decision trees, ANNs also help to identify influential variables, which contribute the most to predicting ABV. Within the ANN technique, I used a method called cross validation to select optimal parameters for both the number of hidden nodes (which help compute the weights for each predictor variable) and the weight decay, which penalizes models with large values of weights and helps reduce the risk of overfitting.

|  |
| --- |
| A close up of a device  Description generated with high confidence |
| *Figure 1: Pairs plot for clean dataset (plotting relationships between all variables)* |

**Model Selection and Assessment**

Determining the best model out of these data mining techniques and methods required finding a model which predicts ABV from the 12 predictor variables with the least amount of error, using a training dataset of 25,000 (or approximately 80%) and optimal parameters (like for boosting and ANNs). Predictions were made for each of these models and measured against the original ABV values using the MSE calculation to see how accurate they fit. Essentially, the model with the lowest MSE wins!

Out of these data mining techniques, the regsubsets function had the lowest MSE at 0.0097 (see Table 1 below), which is a slight reduction in error from the traditional ABV calculation method. In fact, 11 out of the 13 multiple linear regression models tested had all very comparable MSE values and only differed by a maximum of 0.00002. This multiple linear regression function, chosen by regsubsets, included five variables to predict ABV – OG, FG, BoilTime, PitchRate and PrimaryTemp. The first two variables make sense, as they’re included in the traditional ABV formula. BoilTime (how long the wort is boiled), PitchRate (the yeast added during fermentation), and the PrimaryTemp (temperature during fermentation stage) also made sense given all three control the amount of sugars fermented, which aids with creating alcohol.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1: Lowest MSE Values for Data Mining Techniques** | | | | | |
|  | **Multiple Linear Regression (regsubsets)** | **Artificial Neural Networks** | **Bagging/ Random Forests** | **Boosting** | **Pruning** |
| MSE | 0.0097 | 0.0098 | 0.0306 | 0.1150 | 0.3680 |

While all aspects of the brewing process are important to create quality product, OG and FG are the most significant variables contributing to ABV (as anticipated and demonstrated by large coefficients in formula below) with smaller but important contributions from the other three variables. Homebrewers might consider watching these variables closely when predicting ABV. For any future predictions of ABV, here’s the suggested formula for homebrewers to apply:

*ABV-estimate = -1.678 + 131.5\*OG - 129.9\*FG + 0.000235\*BoilTime + 0.007222\*PitchRate + 0.0004571\*PrimaryTemp*

The ANN data mining technique revealed the second lowest MSE value at 0.0098 with decision trees not faring quite as well. Further analysis of the best ANN model showed that OG was the most influential variable but this model didn’t reveal a similar strong influence by FG – rather an equal influence from FG, IBU and BrewMethod. FG and BrewMethod make sense, but the IBU relationship is interesting as it’s usually seen as an outcome from the brewing process not necessarily a predictor of ABV.

A screenshot of a cell phone

Description generated with very high confidence**Impact on Validation Set**

Following the analysis to choose the best model from the training data set, I ran the regsubsets model against the test data (6,644 beer recipes), which was approximately 20% of the clean data set. This model showed similar accuracies predicting new data, or ABV, as the MSE value was slightly lower (0.0096) with an R-squared value of 0.9972, which highlights how much of the variance was explained by the model. This is a highly predictive model. You can see this strong correlation between actual ABV values in the test data versus those ABV estimates predicted from the regsubsets model in Figure 2.

Figure 2: Plot correlation between the test data values for ABV (y-axis)

**Final Conclusions**

My final multiple linear regression model, using OG, FG, BoilTime, PitchRate and PrimaryTemp, produced a slightly lower MSE value (0.0097) than the traditional ABV calculation method (0.01). While this change may seem insignificant, any small changes in the brewing process can change various results including ABV, flavor and complexity. This means any additional accuracy with ABV predictions would probably still be interesting to the more dedicated homebrewers and craft brewers, who must have accurate estimations and calculations for ABV since they sell to the public. Further research into predicting ABV with additional variables (like PrimingMethod, PrimingAmount, MashThickness, etc.) and additional data mining techniques might reveal increased accuracy.

**Variable Guide**

* Response Variable
  + ABV – Alcohol By Volume (percentage)
* Predictor Variables
  + Size (L) – Amount brewed for recipe listed (in liters)
  + OG – Specific gravity of wort before fermentation
  + FG – Specific gravity of wort after fermentation
  + IBU – International Bittering Units
  + Color – Standard Reference Method - light to dark ex. 40 = black
  + BoilSize – Fluid at beginning of boil
  + BoilTime – Time wort is boiled (in minutes)
  + BoilGravity – Specific gravity of wort before the boil
  + Efficiency – Beer mash extraction efficiency - extracting sugars from the grain during mash
  + BrewMethod – Various techniques for brewing (all grain, brew in a bag, extract, partial mash)
  + PitchRate – Yeast added to the fermentor per gravity unit - M cells/ml/deg P
  + PrimaryTemp -- Temperature at the fermenting stage (in Celsius)
* Removed Variables
  + BeerID – Record ID
  + Name – Name of the brewer
  + URL – Location of recipe webpage at <https://www.brewersfriend.com>
  + Style – Type of brew
  + StyleID – Numeric ID for type of brew
  + MashThickness – Amount of water per pound of grain
  + SugarScale – Scale to determine the concentration of dissolved solids in wort (specific gravity or plato)
  + PrimingMethod – The type of priming sugar used
  + PrimingAmount – Amount of priming sugar used
  + UserID