**ABC Beverage Company: Our pH Predictive Model in Light of the New Beverage Control Regulations**

As you all know, new beverage control regulations are requiring us to understand our manufacturing process, the predictive factors and our predictive model of pH. The purpose of this document is to explain in layman’s terms how our data team created our pH predictive model and what it can tell us.

1. **What is a predictive model?**

When we construct a predictive model, we identify predictive factors and find the best mathematical algorithm (a process or set of rules to be followed in making a calculation) to ensure our predictions are as reliable as possible. Think of it this way: imagine a new restaurant has opened down the block and you want to predict whether it’s worth your while to give it a look. You will determine the factors that go into your decision (reviews, type of food, what it looks like through the window, etc.). Then you will need a way to put those factors together - you might give each factor a score and average them, or you might think “if it’s Chinese food I’ll give it a try, but if it’s Italian food I’ll only give it a try if there’s pizza on the menu”). The former exercise is determining predictors and the latter is determining an algorithm.

1. **How did we construct our predictive model?**

In order to predict pH, we identified 32 potential predictors, including Brand, Mnf Flow, Usage cont, Alch rel, temperature, and more. We knew going in that it was likely that not all of the predictors would be useful. We also knew that some of the predictors were redundant, in that they had very similar effects on pH to other predictors, and so they might need to be removed. We also had some predictors that had missing values, and so we needed mathematical techniques to fill in those values.

Having identified our predictors, we processed them through 11 of the most common machine learning algorithms, including regression, neural net, support vector machines, random forests, and others. These algorithms each define a different set of rules and processes for evaluating the relationship between the predictors and Ph in order to come up with a prediction. This gave us a set of 11 different predictions for pH based on the 32 predictors.

1. **How did we evaluate our model in order to choose the best one?**

When we create a predictive model, we only use 80% of the data in the creation (or training) stage. This data includes all of our predictors as well as the pH for instances of beverages we have already tested in the field. In this case, we had a total of 2571 beverages we had tested, and so we trained our model on 2057 randomly chosen beverages. In the training stage, we develop a chosen algorithm to model the relationship between the predictors and pH.

Then we move into the testing stage: Using the algorithm we constructed, we make predictions on the remaining (in this case 514) beverage records. We then compare the predictions we made to the actual pH for those records and calculate a metric that we can use to compare to other algorithms. There are a number of different such metrics, including Root Mean Square Error and R Squared.

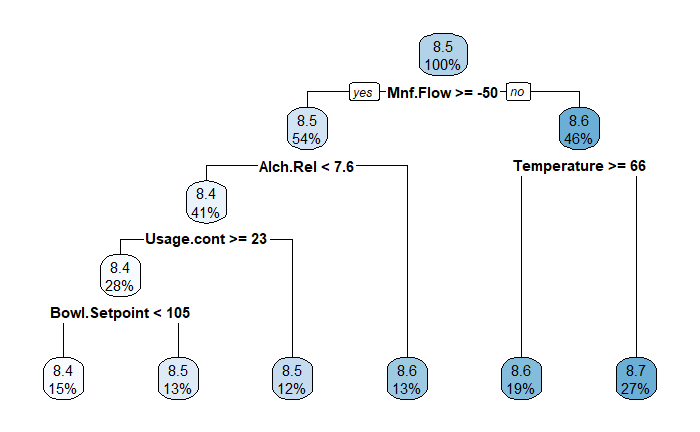
1. **How did the models perform?**

On the whole, we did not see stellar performance from any of the models, but our best model (a Random Forest model) performed adequately with an R-squared of .72. In layman’s terms, this means that our model explains about three quarters of the variation in pH for our beverages. This suggests we will usually be in the ballpark with our predictions but that there will still be significant room for small errors up or down.

1. **What is a Random Forest Model?**

In the restaurant example above, one of our algorithms was a decision tree (if it’s Italian, I’ll only give it a try if there’s pizza). Decision trees map a set of sequential decisions. Random Forest algorithms run a number of different decision trees on the same data and average out the results. This gives us a more reliable picture than one decision tree.

The following chart shows one decision tree generated by our Random Forest algorithm. According to the tree, when Mnf.Flow is not greater than or equal to -50, we can use temperature to determine pH. Otherwise, we can evaluate whether Alch.Rel is less than 7.6, and follow the tree accordingly.



1. **What does our Random Forest model tell us about how to predict pH?**

When we examine the most important predictors, not only in our Random Forest model but in some of the other better-performing models as well, we find that Mnf.Flow is the most important predictor, followed by Usage.cont and Bowl.Setpoint. Alch.Rel, Filler.level and Brand (especially Brand C) are also factors. We can see in the tree above, in this example all of our decisions begin with an evaluation of Mnf.Flow. While our Random Forest model aggregates many different trees, we can presume that many of them begin with Mnf.Flow.

1. **How does this information help our company?**

The new beverage control regulations require us to communicate our predictive model to regulators, but beyond that they also require us to have better control over the pH of our beverages. The most critical predictors for pH are in fact a small subset of the 32 potential predictors. This gives us an easy and quick way to predict whether our pH levels will fall into compliance or not.

It is important to understand that a predictive model does not necessarily map causality. A wagging tail might be predictive of a happy dog, but you can’t make the dog happy by wagging its tail. Similarly, we would not want to suggest here that pH is best regulated through manipulating Mnf.Flow – rather, our predictive model is important because it allows us to anticipate and predict compliance issues so they may be addressed before beverage audits occur.