Team members: Calvin Wong, Omar Pineda, Murali Kunissery

Project Problem Statement

This is role playing. I am your new boss. I am in charge of production at ABC Beverage and you are a team of data scientists reporting to me. My leadership has told me that new regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH. Please use the historical data set I am providing. Build and report the factors in BOTH a technical and non-technical report. I like to use Word and Excel. Please provide your non-technical report in a business friendly readable document and your predictions in an Excel readable format. The technical report should show clearly the models you tested and how you selected your final approach. Please submit both Rpubs links and .rmd files or other readable formats for technical and non-technical reports. Also submit the excel file showing the prediction of your models for pH.

Executive Summary

New Regulations by ABC beverage company leadership requires the company’s production unit to better understand manufacturing processes, the predictive factors and its relationship to the PH values of the beverages.

Research Statement

The research is to develop a model to find the predictive variables related to the PH of beverages

Data Collection

The data set is historic data containing predictors associated to the PH and is provided in an excel format. We will utilize the training dataset to analyze the PH of beverages and the test dataset to predict PH. Two excel files are provided:

* The training data (StudentData.xlsx)
* The test data (StudentEvaluation.xlsx).

Model Selection

We will attempt multiple machine learning approaches and use RMSE and R2 to determine the best approach. The reason to use these measures is because they are widely used for evaluating. Both use SST (difference in distance between the actual value and the baseline model which is generated from the mean) and SSE (difference in distance between the actual value and predicted model which we consider to be the best fit model).

Linear Regression Models

Linear regression is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modeling the relationship between a scalar response (or [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable)) and one or more [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or [independent variables](https://en.wikipedia.org/wiki/Independent_variable)). In this instance, due to multiple dependent variables, we use a process called multiple linear regression. Three different approaches are selected, generalized linear model (GLM), glmnet and partial least squares (PLS).

The performance of each linear regression model is discussed within the technical document. The model chosen is the generalized linear model (GLM), although not the best performing, was very close in performance. The best model showed only a 0.001 improvement in the RMSE values. Due to computing burdens, we elect the simpler GLM model.

Non-linear Regression Models

Nonlinear regression models are those that are not linear in the parameters. Transformation is necessary to obtain variance homogeneity but transformation destroys linearity as well. In this exercise we have demonstrated a comparison model between linear and non linear regression and used Neural Networks to analyze and predict the density (pH) of the Beverage. The idea behind Neural Network is simple that it introduces a hidden layer between input and output layer. The addition of the hidden layer makes the neural network non linear. The selection of the number of hidden layers is usually a trial and error method and running the analysis through multiple iterations reveals the best selection model. The technical section reveals that two hidden layers followed by a 5 node neuron best fitted the dataset and the analysis was performed with that setting. RMSE on the neural network was determined to be 0.1366331 and was closer to other models performed but not the best.

Tree Models

Decision trees consist of one or more nested if-then statements for the predictors that partition the data. A model is used to predict the outcome within these partitions. The final regions are called leaves and the points where the splits occur are nodes. We considered a Basic Regression Tree, XGBoost Trees and Random Forest, and we chose Random Forest as the best model for this section. The Random Forest model is an ensemble method that incorporates several decision trees and reduces the correlation between trees. Random Forests are more computationally efficient on a tree-by-tree basis since the tree building process only needs to evaluate a fraction of the original predictors at each split. It is also possible to quantify the impact of predictors in the ensemble. The Random Forest that we trained had an RMSE of 0.096 and and R^2 of 0.739, meaning that about 74% of the variance in PH was explained by our predictors.

Conclusion

Three different type of models were analyzed to find the predictive variables related to the pH of beverages. RMSE was used as a comparative variable to determine the best fit for the dataset. The analysis revealed that the RMSE for the chosen Linear Model (glm) was 0.134, Non-Linear Model (Neural Net) was 0.136, and lastly, the RMSE for the chosen Tree model (Random Forest) was 0.096. So, we chose the Random Forest Model to predict the PH of our beverages given the predictors for beverages.

In the chosen Random Forest model, the top 5 predictors that we found to influence PH in the manufacturing process were Mnf.Flow, Brand.Code, Usage.cont, Oxygen.Filler, and Alch.Rel. We should pay close attention to these predictors when trying to control for the PH of our beverages given the new regulations that have been enacted.

We are also including PH predictions based off of this model for the beverages that you provided in the evaluation dataset. Please let us know if you have any questions.