A picture containing bottle, indoor, food, vegetable

Description automatically generated

PREDICTIVE MODELING of PH

ABC BEVERAGE COMPANY MANUFACTURING PROCESS

December 12, 2021

*Data Scientists*

*Diego Correa*

Lisa Szydziak

**Project Goals**

It is powerful tool from a manufacturer’s standpoint to understand and identify the key drivers in their business. New regulations require an understanding of the manufacturing process, the predictive factors in predicting PH and a predictive model of PH. The results of a machine learning model could mean huge profits or gains if implemented in a plan.

Using the available data, we built several predictive models that could shed light on the key drivers of “PH”.

The strength and power of the model is evidenced by comparing all models and choosing the best performing model.

The model will give us the variables in order of importance when predicting PH.

In addition, we are charged with providing predicted values PH of an evaluation dataset. We will use our best model to provide these predictions.

# **Data Collection and Design**

The data file contained 2, 571 records with 33 variables.

The variables were used in the model:

|  |
| --- |
| Brand Code |
| Carb Volume |
| Fill Ounces |
| PC Volume |
| Carb Pressure |
| Carb Temp |
| PSC |
| PSC Fill |
| PSC CO2 |
| Mnf Flow |
| Carb Pressure1 |
| Fill Pressure |
| Hyd Pressure1 |
| Hyd Pressure2 |
| Hyd Pressure3 |
| Hyd Pressure4 |
| Filler Level |
| Filler Speed |
| Temperature |
| Usage cont |
| Carb Flow |
| Density |
| MFR |
| Balling |
| Pressure Vacuum |
| Oxygen Filler |
| Bowl Setpoint |
| Pressure Setpoint |
| Air Pressurer |
| Alch Rel |
| Carb Rel |
| Balling Lvl |

All variables were quantitative except for Brand, which took on values of A, B, C, D.

The dataset contained several missing values.

We removed degenerate variables, imputed missing variables, and considered removing highly correlated variables. In addition, we excluded records that were missing Brand.

# **Models**

The dataset was fit with several models and assessed accordingly for the best fitting and performing model.

The first model considered was a Multiple Linear Regression Model, a traditional statistical model.

The model is in the form:

i+…

We want to fit a dependent variable, *Y* and predictor variables, *x1, x2, x3, etc,* that are linear in the parameters. In other words, the response variable Y will be predicted by the explanatory variables, x1, x2, x3, etc. In the same way, for instance, height can be modeled to predict weight.

Once we run the analysis, we want to make sure we are not missing anything. Fit diagnostics such as residuals, Studentized residuals, leverage points, quantile plots, CooksD points are assessed for model validity.

Once the analysis is complete, we will have a model that predicts the “Y” variable. Inherent in the model, the predictor variables can be sorted in order of importance, and it will be evident what are the key drivers of the response variable.

Data Science methodologies offer several machine learning models which we will incorporate into our analysis.

The next model considered is Elastic net model that incorporates a penalty to the bias as a trade-off to reduce the variability.

We consider a Neural network model, a nonlinear technique inspired by biological neural networks to produce predictions.

The Support Vector Machine model is a supervised learning algorithm that provides classification or regression for data groups.

The Cubist model will be assessed as well. It is a rule-based predictive model.

The Random Forest model combines multiple decision trees to create a forest.

# **Data Analysis and Model Interpretation**

We seek to predict PH. The figure below is a histogram of the values of PH in the dataset.

Chart, histogram

Description automatically generated

Here are the descriptive statistics for PH:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 7.880 | 8.440 | 8.540 | 8.548 | 8.680 | 9.360 |

Our best performing model was the **Random Forest model**, which we recommend for ABC Beverage to predict PH.

What are the most important predictor variables sorted in order of importance?

|  |  |
| --- | --- |
| 1 | MnfFlow |
| 2 | BrandCodeC |
| 3 | OxygenFiller |
| 4 | AlchRel |
| 5 | AirPressurer |
| 6 | PressureVacuum |
| 7 | BallingLvl |
| 8 | CarbRel |
| 9 | CarbPressure1 |
| 10 | Temperature |
| 11 | Usagecont |
| 12 | CarbFlow |
| 13 | HydPressure3 |
| 14 | FillerSpeed |
| 15 | Density |
| 16 | Balling |
| 17 | FillerLevel |
| 18 | PCVolume |
| 19 | FillOunces |
| 20 | CarbVolume |
| 21 | FillPressure |
| 22 | MFR |
| 23 | HydPressure2 |
| 24 | BowlSetpoint |
| 25 | PSC |
| 26 | PSCFill |
| 27 | HydPressure4 |
| 28 | CarbPressure |
| 29 | CarbTemp |
| 30 | PSCCO2 |
| 31 | PressureSetpoint |
| 32 | BrandCodeA |
| 33 | BrandCodeB |
| 34 | BrandCodeD |

It would be advisable to focus on the top 5 parameters to drive predicting PH.

All models indicated that the most important variable used is the **MnfFlow** in predicting the **PH** value. In most models, this variable was followed by **UsageCont**, **BowlSetpoint**, and **FillerLevel**. However, this is not the case in the **Random Forest** model, which is our best performing model.

We recommend the top performing model, the random forest with top 5 predictor variables: **MnfFlow**, **BrandCodeC**, **OxygenFiller**, **AlchRel**, **AirPressurer**.

**Final Discussion**

Predictive Modeling of PH very powerful, it could help manufacturers with quality control and perhaps profitability.

The Random Forest Model performed best and is recommended to predict PH. The model has variables in order of importance which gives information about the manufacturing process.

As requested, attached is an excel file of the evaluation dataset with predicted PH.

Furthermore, the process of ongoing modeling can help manufacturers keep apprised of the key drivers in the dynamically changing environment.