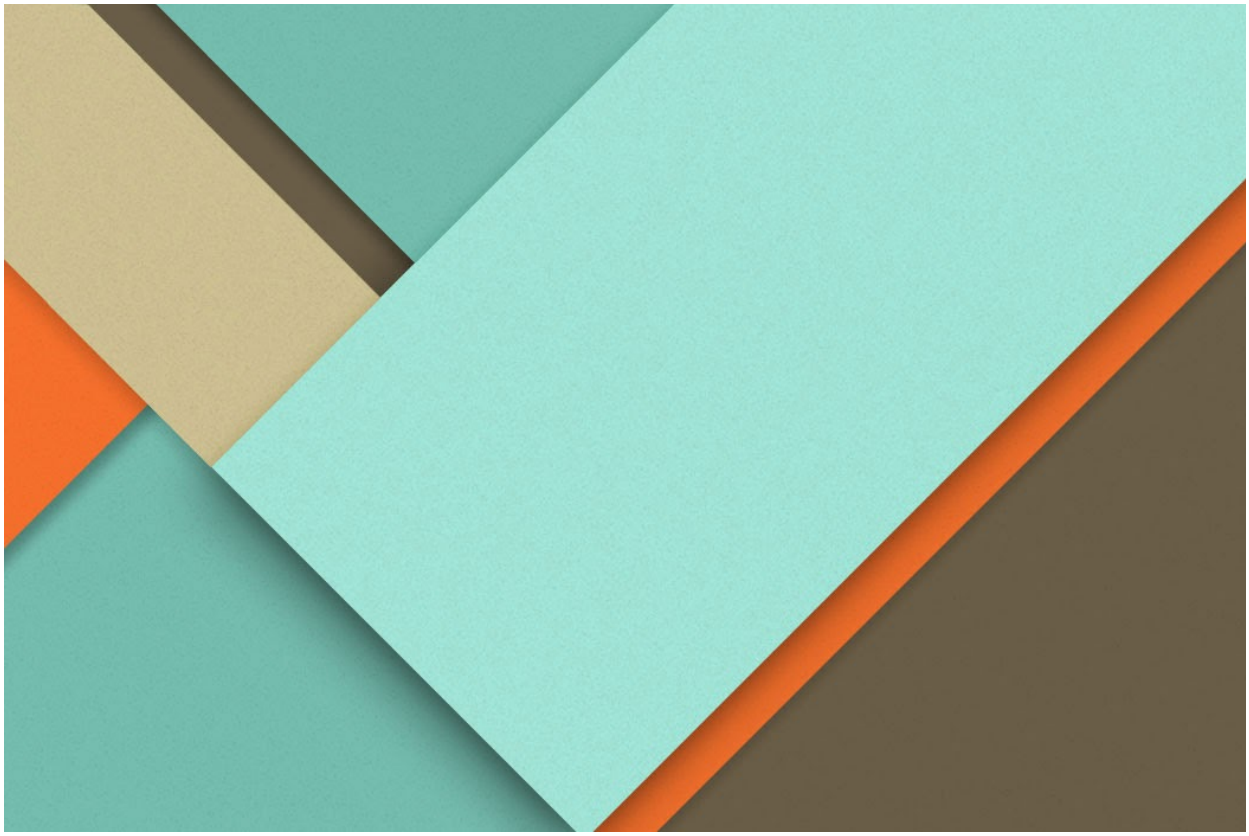


Group 4 - Project 2

2020-05-10

Paul Britton

Anthony Munoz

Luisa Velasco

Javern Wilson

Overview

This document serves as a non-technical summary of our analysis of the manufacturing process at ABC Beverages. As per incoming regulations, the intention of this document is to explain ABC's manufacturing process with specific purpose of highlighting both useful models for; and important factors in the prediction of beverage PH. This document is intended to be stand-alone, however, readers can also find the complete technical report here.

Data

A comprehensive set of factors encompassing all aspects of the manufacturing process was considered for this research. The general approach was to begin with as broad a dataset as possible and effectively de-emphasize or eliminate the less important predictors through our analysis and modelling.

Care was taken to ensure that the data were appropriately pre-processed using various techniques in an effort to maximize amenability to statistical modelling. As an example, missing values were filled in using a statistical sampling technique.

Modelling

Several "families" of models of varying complexity were explored in this work. Rather than try to anticipate which kinds of models would be most effective, or what types of relationships the data may hold, we elected to cast a broad net and run numerous models. This approach allows us to let the data instruct us as to what methods are most appropriate.

I. Procedures

The data was segmented in various ways in order to help guard from over-fitting, or building models which work well on the training data, but fail on new, unseen data. The general principle is that we can take a liberal and exploratory approach with our training data with where we try many ideas, while relying on our testing data to help

us validate (or invalidate) these techniques on fresh data. The data was segmented as follows:

A. Training Data

This dataset was used for model development and exploratory research. In the interest of creating robust results, the training set was randomly split on an 80/20 basis, meaning we randomly select 80% of the data points for development and hold out the remaining 20% for validation.

B. Testing Data

This data is entirely "out of sample" and predicting the target variable (PH) using this data is the ultimate objective.

II. Models

A. Linear Models

The simplest family of models tested, these models perform best when relationships are linear. They are generally easier to understand on an intuitive basis and were tested as there is obvious merit in keeping models as simple as possible.

B. Non-Linear Models

As the name implies, non-linear models are suited to problems where the relationships between variables may be less intuitive. They are more generally more complex than linear models - this complexity, however, may be required to produce an effective model depending on the relationships in the data.

C. Tree Models

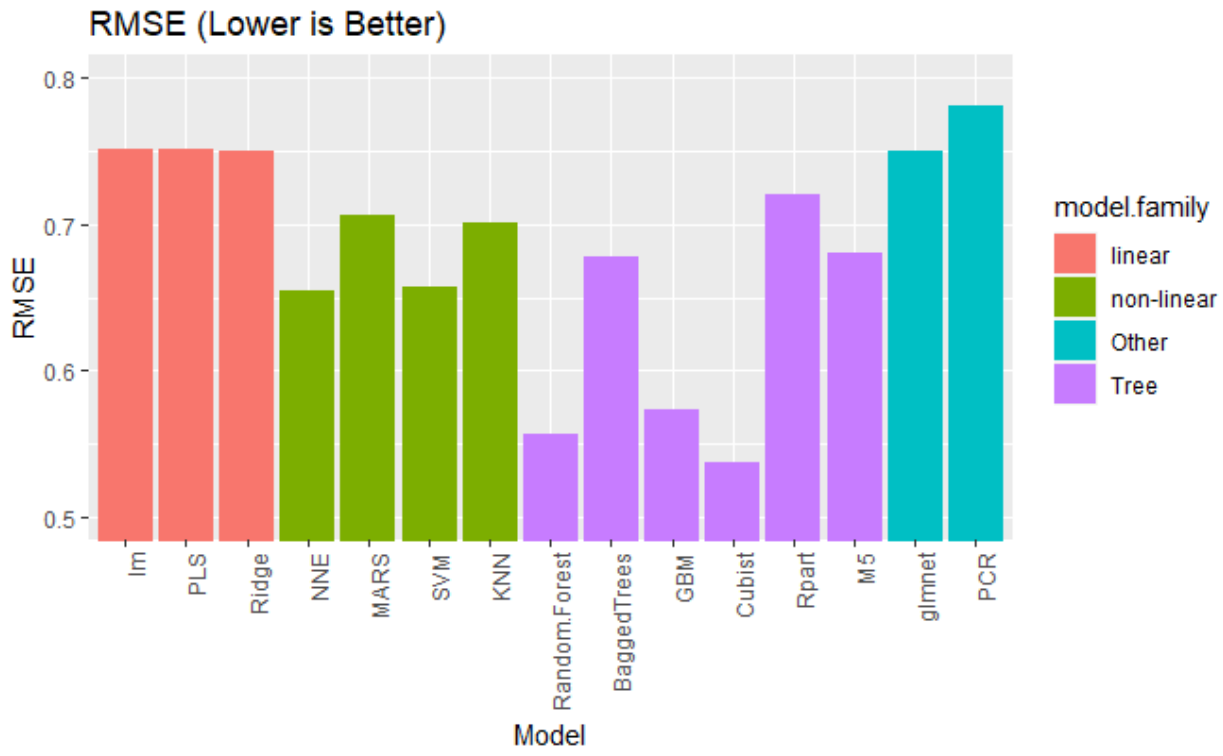
Tree models use a tree-like structure (similar to the structure of a file-system in a computer) to identify relationships in the data. Their merit lies in the fact that they are efficient, make no explicit assumptions about the data structure, are relatively insensitive to data quality, and most importantly, provide intuitive, non-technical output which is easily interpretable.

D. Other Models

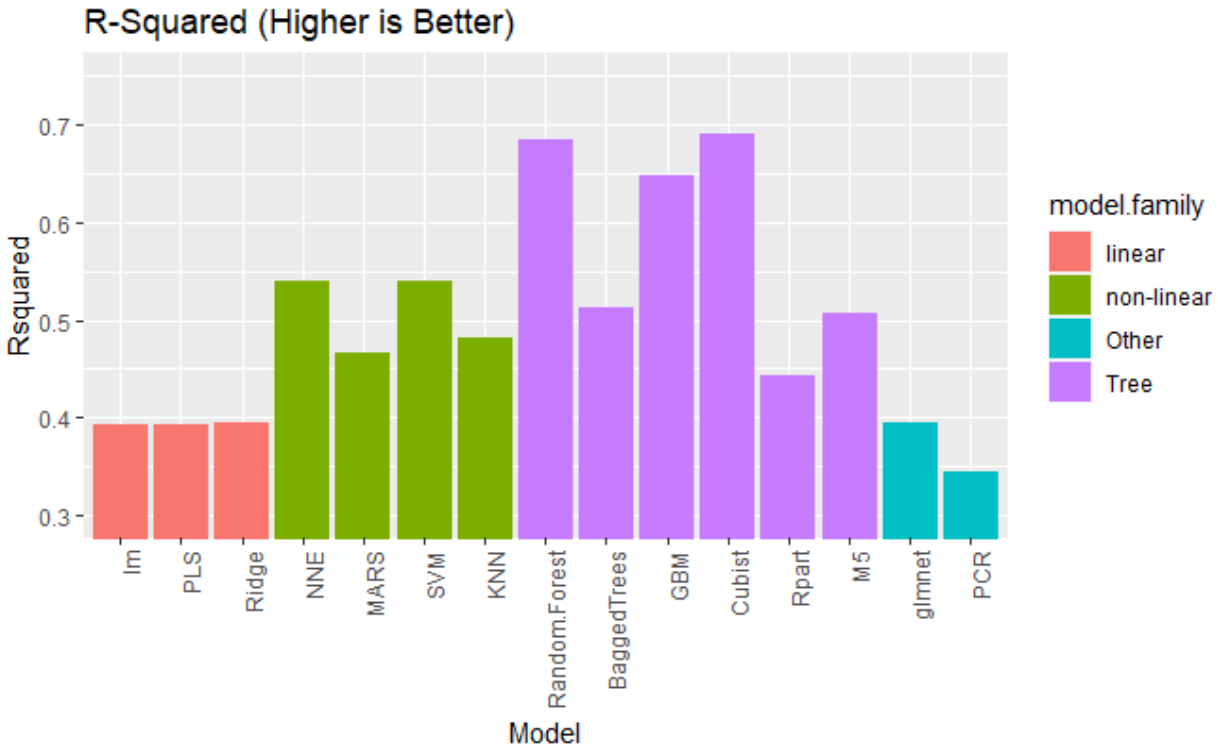
Several other models which do not fit cleanly into any of the 3 categories above were also analyzed, the details of which can be seen in the technical report.

Key Findings

Of the 15 models tested, the most favourable results were observed in the tree family, most specifically using the GBM and Cubist models. As can be seen from the plots below, the tree family (purple) of models generally produce the most favourable results followed by the non-linear (green) family. The linear and "other" models appear to be about equal. The finding that non-linear models seem to produce better performance may be evidence that there are some non-linear relationships present within the data. In the context of managing product PH, this may mean that small changes in process may lead to disproportionately large changes in output for some variables, which is noteworthy.



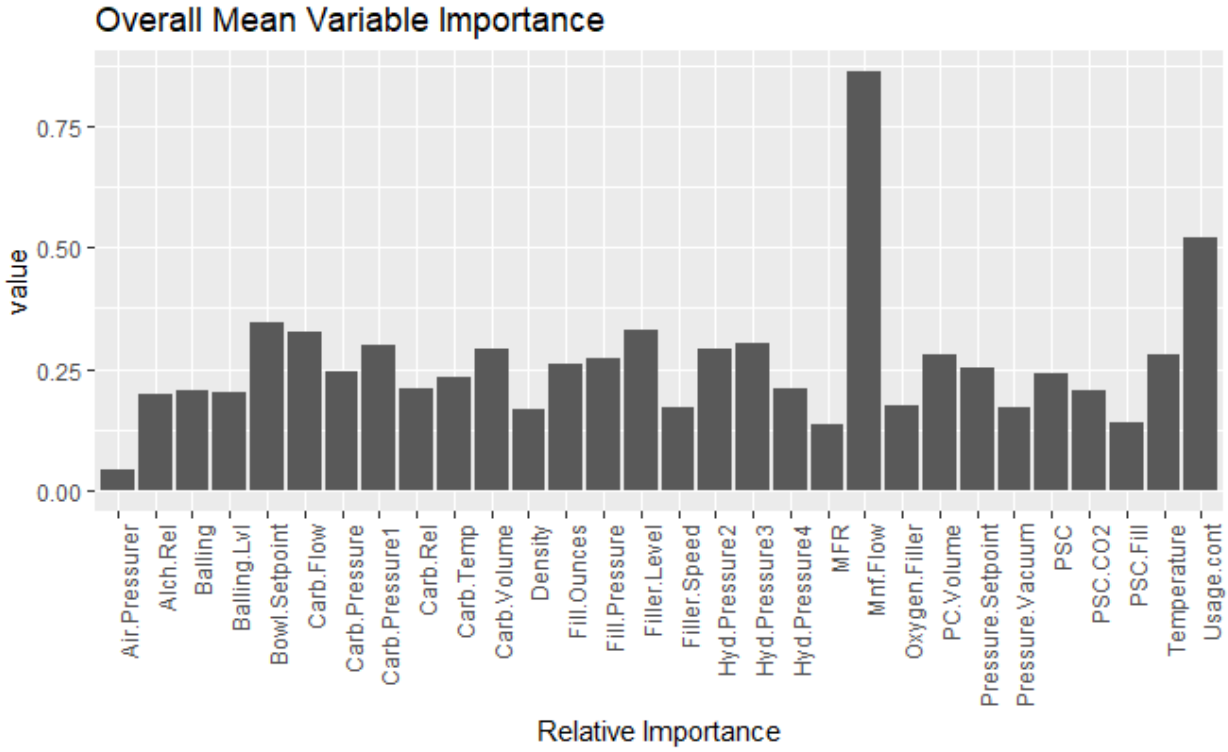
The RMSE, which is a measure of model error (differences between the actual values observed and the values predicted by the model) clearly show the Tree family, and particularly "Cubist", "Random Forest" and "GBM" as being the best models.



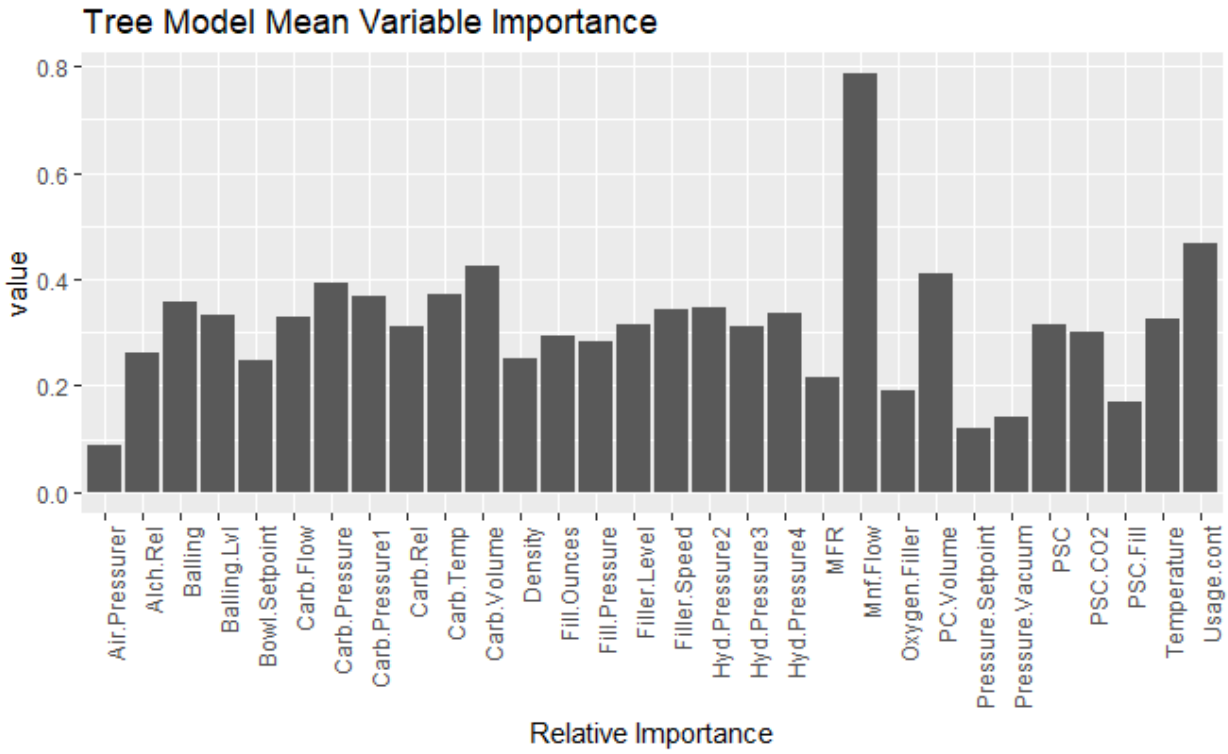
The R-Squares, which, in simple terms, tells us how much of the change in PH is explained by the model, once again shows that the tree family is superior. Whereas the linear and "other" models explain less than half of the variability in PH, "Cubist", "Random Forest" and "GBM" explain almost 70%.

III. Important Factors

When considering variable importance, we first look at overall mean variable importance, which shows the relative importance (1 = most important, 0 = least important) of each variable across all models. We can plainly see that 1 variable stands out above (Mnf.Flow) and that another (Usage.cont) is not far behind whereas most other variables are at 0.33 or below. Note that some of the models included in this average, however, were not particularly accurate. As such, we will also examine variable importance for just the models from the tree family.



If we consider only the tree family, we see a similar picture, in terms of top performing variables, however, we can see a generally similar picture with the differences between variable importance being found in the secondary relationships.



We can infer from these plots that most of the models are likely to capture the main, obvious relationships (Mnf.Flow, Usage.cont) whereas non-linear models have added value by detecting relationships that may be harder to see. The overall finding is that PH will likely be most sensitive to changes to "Mnf.Flow" and "Usage.cont" and as such, special care should be taken if adjusting these values.

IV. Recommended Model

The "Cubist" model shows superior performance to all other models examined and this is the recommended model. The Cubist model was selected using a combination of RMSE, RSquared and MAE. It was significantly better than the linear models and marginally better than the other tree-based approaches.

	RMSE	Rsquared	MAE
Cubist	0.5371	0.6908	0.4055
Random.Forest	0.5567	0.685	0.4278
GBM	0.5736	0.6473	0.436
NNE	0.6543	0.5392	0.5068
SVM	0.6571	0.5395	0.4915
BaggedTrees	0.6777	0.5123	0.5344

M5	0.6796	0.5073	0.5143
KNN	0.7012	0.4813	0.5455
MARS	0.7066	0.4669	0.5447
Rpart	0.7196	0.4423	0.5602
Ridge	0.7494	0.3945	0.5831
glmnet	0.7495	0.3943	0.5827
PLS	0.7507	0.3926	0.5832
lm	0.7508	0.3924	0.5827
PCR	0.7798	0.3444	0.6178

Conclusion

In summary

1. The variables of importance to executives at ABC Beverage are “Mnf.Flow” and “Usage.cont” and we advise that, independent of any model, care be taken when modifying these quantities.
2. Simple linear models do not fit the process as well as non-linear ones and as such, we advise that the Cubist model identified here is appropriate.