Predicting pH for ABC Beverage

DATA 624 | Glen Dale Davis & Tora Mullings

Introduction:

Our analysis of the relationship between pH and the attributes ABC Beverage measures for its products has revealed:

- The five most relevant features when accurately predicting pH are:
  - ALCH_REL | DENSITY | MNF_FLOW | BOWL_SETPOINT | BALLING
- The best of 15 models for predicting pH is a Support Vector Machines (SVM) Model that utilizes the above features and others and relies on K Nearest Neighbors (KNN) imputed data for missing values in observations
- Other models using the same data performed at least 10% worse

Model Development Process

- We built models in three broad categories:
  - linear: Multiple Linear Regression
  - nonlinear: Multivariate Adaptive Regression Splines (MARS), K Nearest Neighbors (KNN), and Support Vector Machines (SVMs)
  - tree-based: Single Regression Trees
- These models were all trained and tested on three different sets of data:
  - primary: missing values in observations were imputed
  - secondary: observations with missing values were deleted
  - tertiary: missing values in observations were imputed and skewed predictors were transformed

Model Selection Process

- Our primary concerns in model selection were:
  - performance accuracy, so we only considered the models with the lowest Root Mean Squared Error (RMSE)/highest Predictive R-Squared on each dataset as final contenders
- Our secondary concern was:
  - applicability, so we chose a model that could handle observations with missing values over a model that performed better, but could not handle missing values

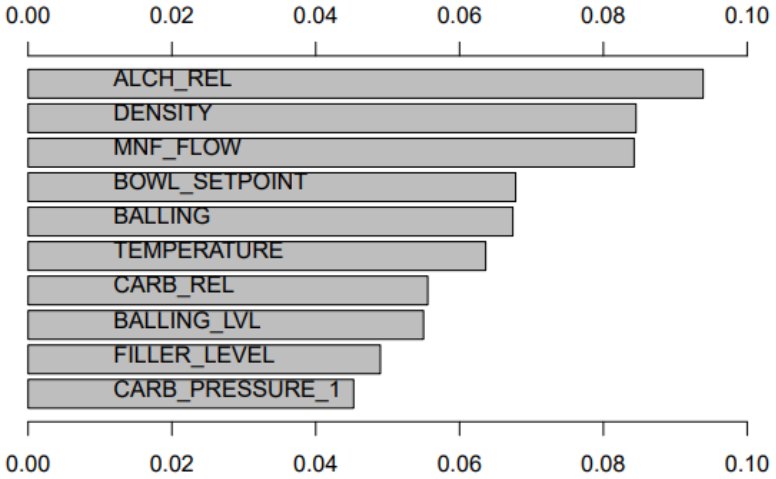
Final Model Metrics

- Model SVM:RB:1 was selected as the final model for predicting pH. It had the lowest RMSE and the highest Predictive R-Squared among its peers:

Model	Predictive R-Squared	RMSE
Model LM:1	0.3984	0.1357
Model MARS:1	0.4532	0.1294
Model KNN:1	0.4651	0.128
Model SVM:RB:1	0.567	0.1151
Model Tree:1	0.4432	0.1306

Final Model Feature Importance

- This nonlinear model relies on features for predicting pH that some of the linear models could not utilize effectively. Their relative importance is depicted below:



Conclusions

- Nonlinear regression methods outperformed linear and tree-based regression methods in this application
- Transformations were not useful in improving model performance
- The search for the best model was by no means exhaustive, and research should continue using SVM:RB:1 as the new baseline