Report on pH Predictive Models

DATA624 Spring 2022

Hector Santana, Christopher Ayre, Harris Dupre

Table of Contents

[Introduction 1](#_Toc103531381)

[Data Exploration 2](#_Toc103531382)

[Variable Correlations 2](#_Toc103531383)

[Missing Data 3](#_Toc103531384)

[Data Distributions, Density, and Boxplots. 4](#_Toc103531385)

[Data Transformation 5](#_Toc103531386)

[Model Building and Evaluation 6](#_Toc103531387)

[Variable Importance 8](#_Toc103531388)

[Conclusions 8](#_Toc103531389)

# Introduction

Our task was to develop a predictive model to better understand how a set of variables influence the pH of beverages. Our team acquired two sets of data, one whose purpose was training the models and the other for evaluating the selected predictive methodology.

The training data had 33 variables, nearly all of which are numerical including the target variable **PH**.

**Brand.Code** was the sole non-numeric variable consisting of single letter codes (A, B, etc.).

# Data Exploration

## Variable Correlations

Our team first analyzed the correlations between variables as highly correlated variables would reduce the predictive reliability of the model

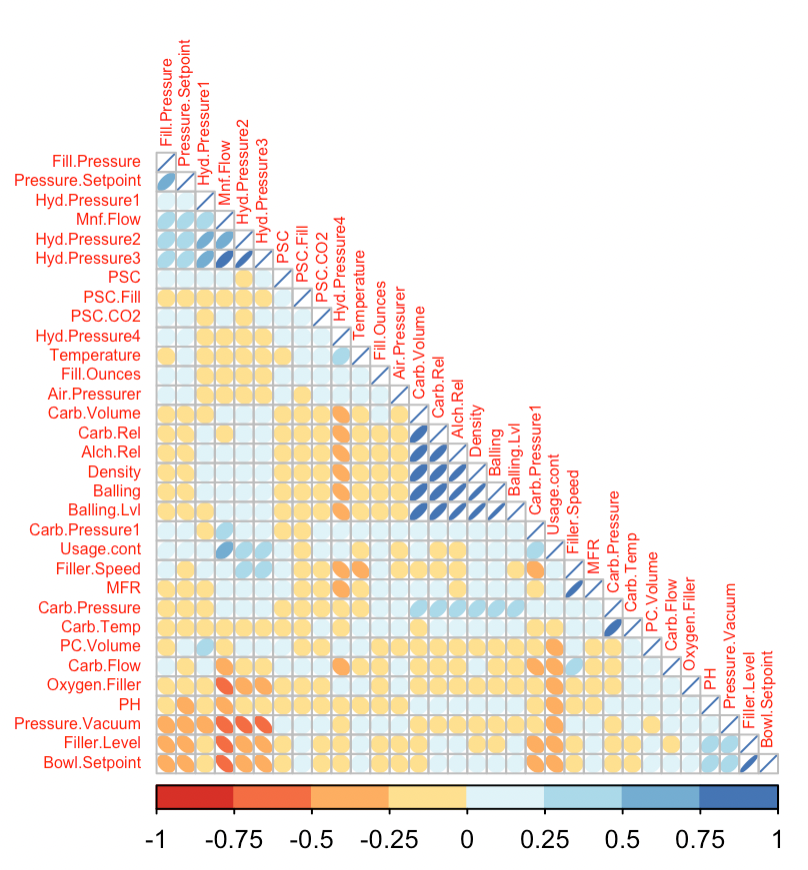


Figure 1 Correlation plot of the training variables.

Some observations of note:

* **Carb.Volume** has a strong positive correlations with variables **Carb.Rel**, **Alch.Rel**, **Density**, **Balling**, and **Balling.Level**. There are also significant positive correlations between these variables.
* Variable **Mnf.Flow** has strong negative correlation with **Oxygen.Filler**, **Pressure.Vacuum**, **Filler.Level**, and **Bowl.Setpoint**.
* **Pressure.Vacuum** also has significant negative correlations with **Hyd.Pressure2** and **Hyd.Pressure3**

## Missing Data

The data were also analyzed to identify variables with a significant percentage of missing data (potentially indicating that the variable should be dropped) and any patterns.

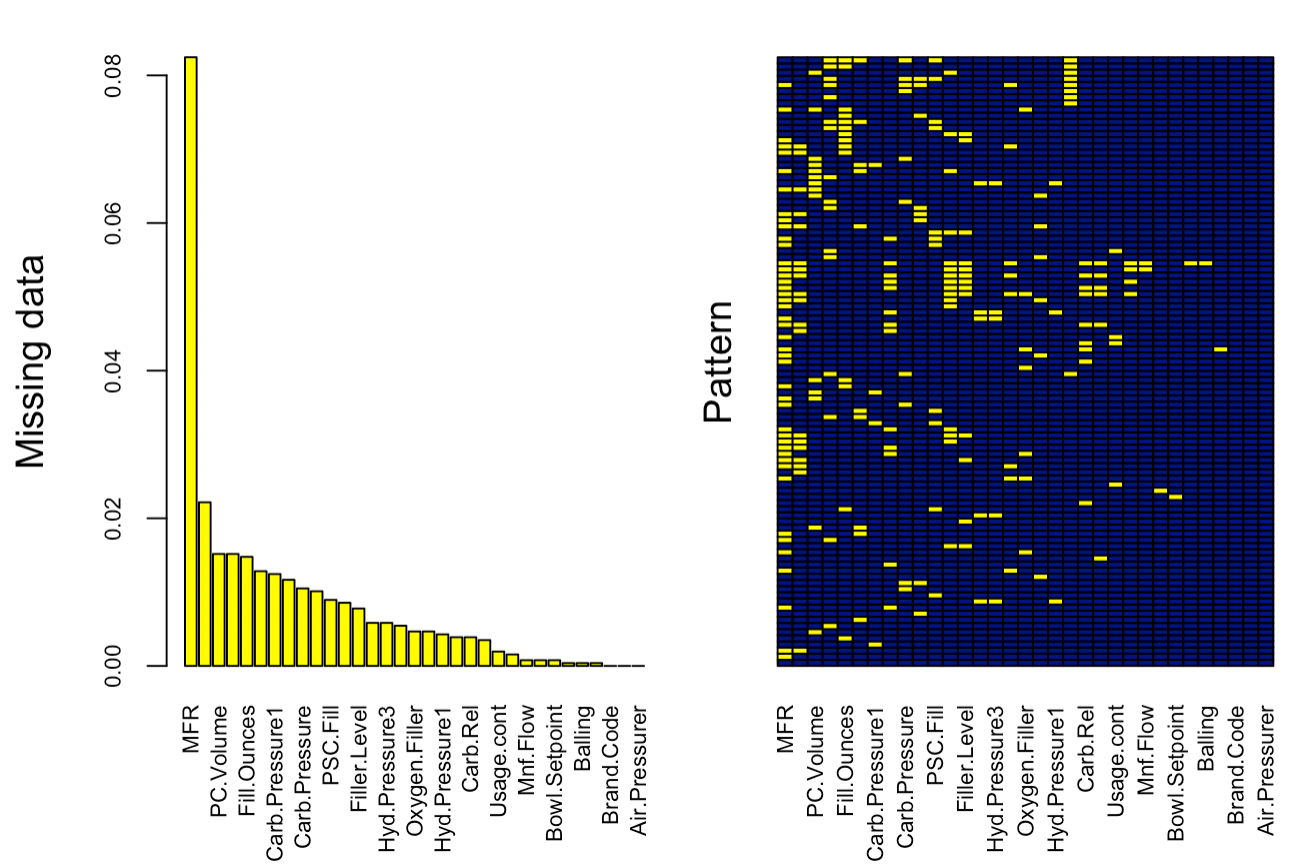


Figure 2 An analysis of missing data by % and pattern

The variable **MFR** had the highest percentage of missing data at approximately 8% of rows, which is still low enough that the field can be used in predictive models after imputation. There was no discernible pattern in the rows with missing data that would indicate, for example, that when one field is empty another would likely also be empty.

## Data Distributions, Density, and Boxplots.

Distribution and density plots are useful for visualizing the normality of each variable. Some predictive methodologies assume normality, and so accuracy will suffer with skewed or otherwise non-normal variables.

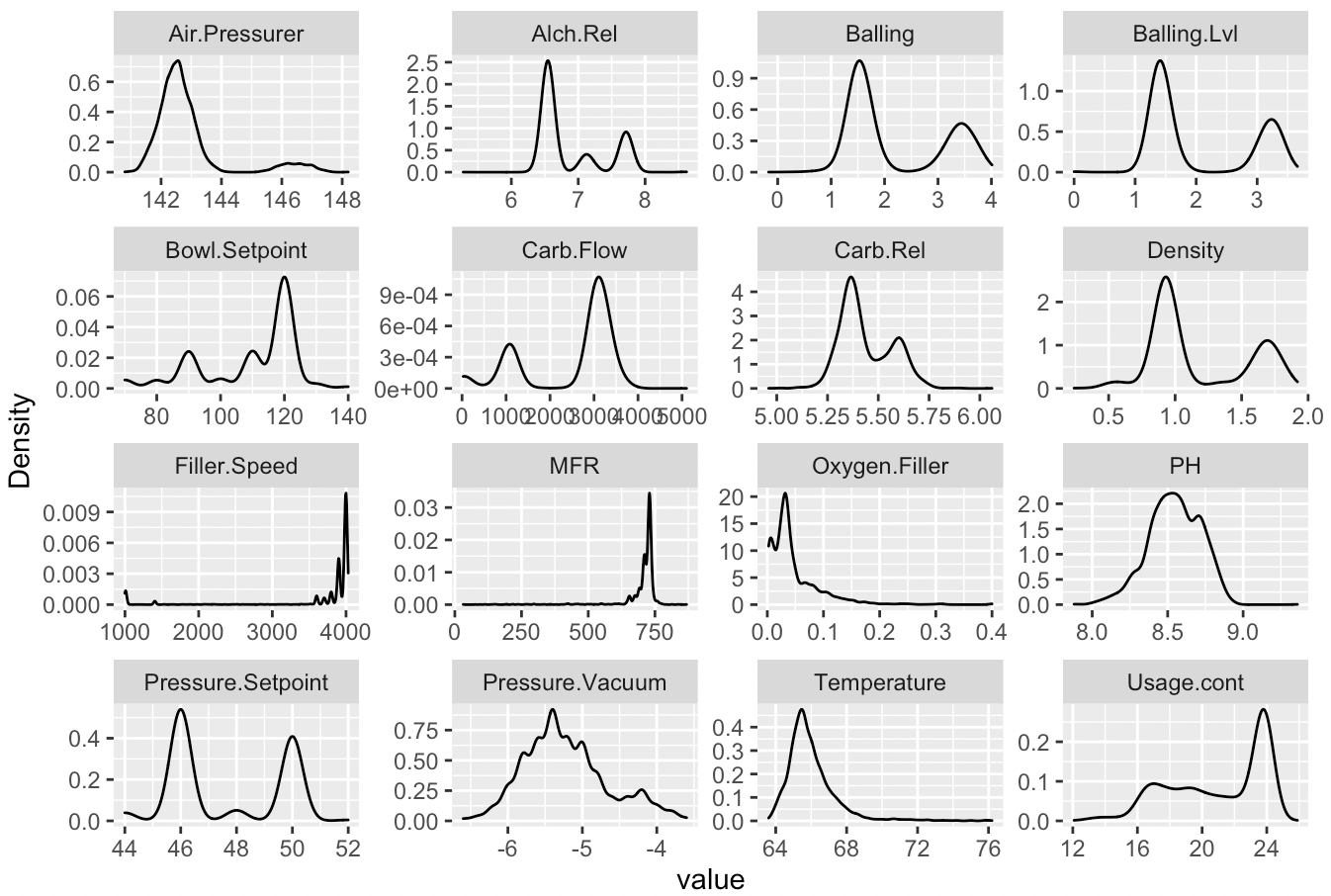


Figure 3 Density distributions of some variables, including the target variable PH.

Boxplot analysis allows us to visualize the quartile distribution of the data based on the target variable, as well as potentially identify outliers.

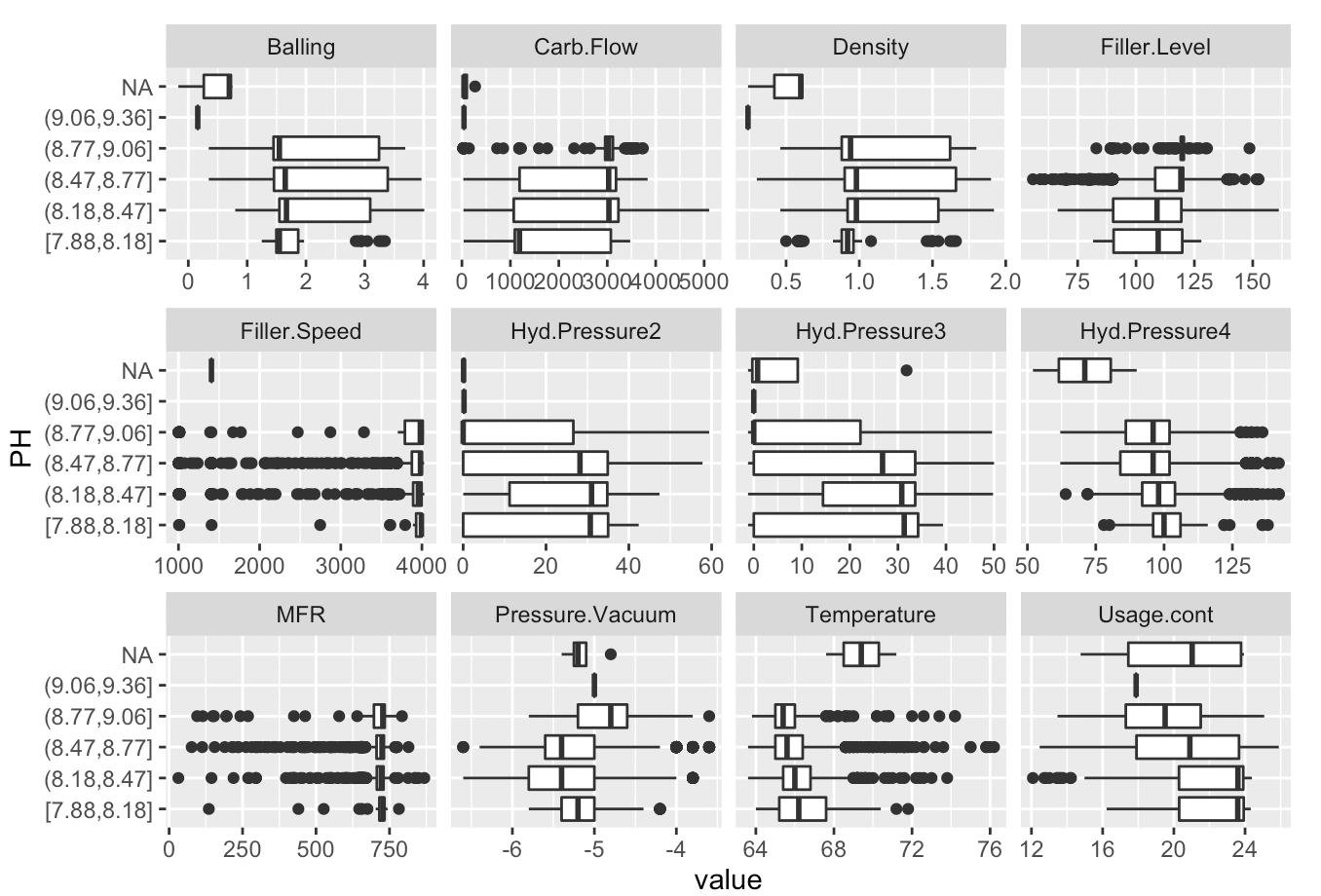


Figure 4 Box-plot analysis based on PH variable.

# Data Transformation

The data was preprocessed prior to model training. This included centering, scaling, imputing missing values, removing highly correlated variables, and removing variables with near zero variance.

The imputation method selected was KNN impute for its processing efficiency and proven accuracy.

Rows where there was no value for pH would have no predictive worth, so they were removed before the preprocessing step.

The training data was split into two sets, split into two partitions, one for training and one for validation as the models ran.

# Model Building and Evaluation

We attempted linear models, non-linear models, decision trees and neural networks. An exhaustive list can be found in the results table below.

The neural network model was not viable, it produced warnings throughout its run and had some of the longest processing times.

Overall, the Cubist model had the best performance, with the highest R squared and the lowest RMSE. MAE was also the lowest.

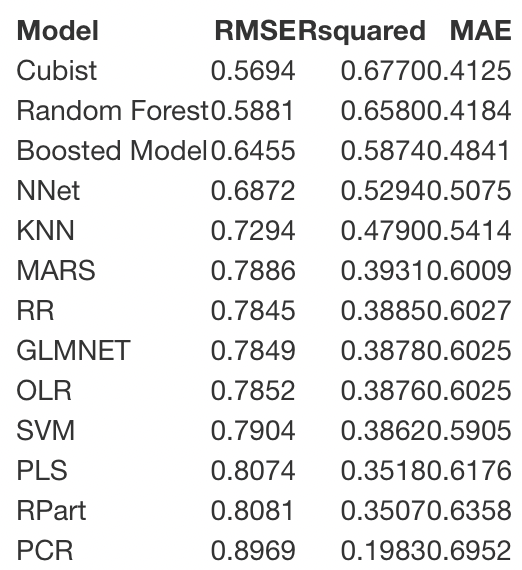


Figure 5 Cubist model showed best performance

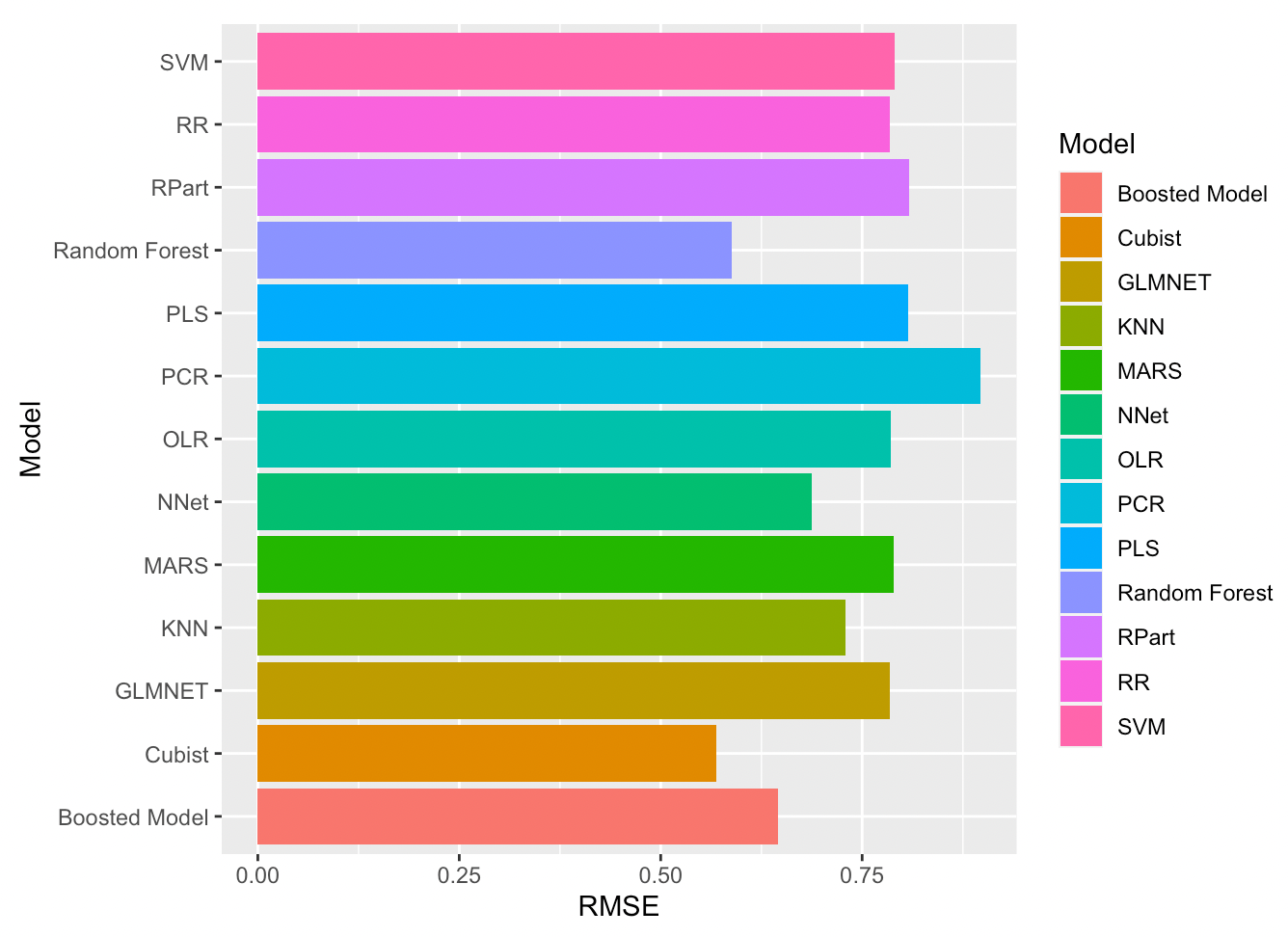


Figure 6 Cubist model had the lowest RMSE, with Random Forest a close second.

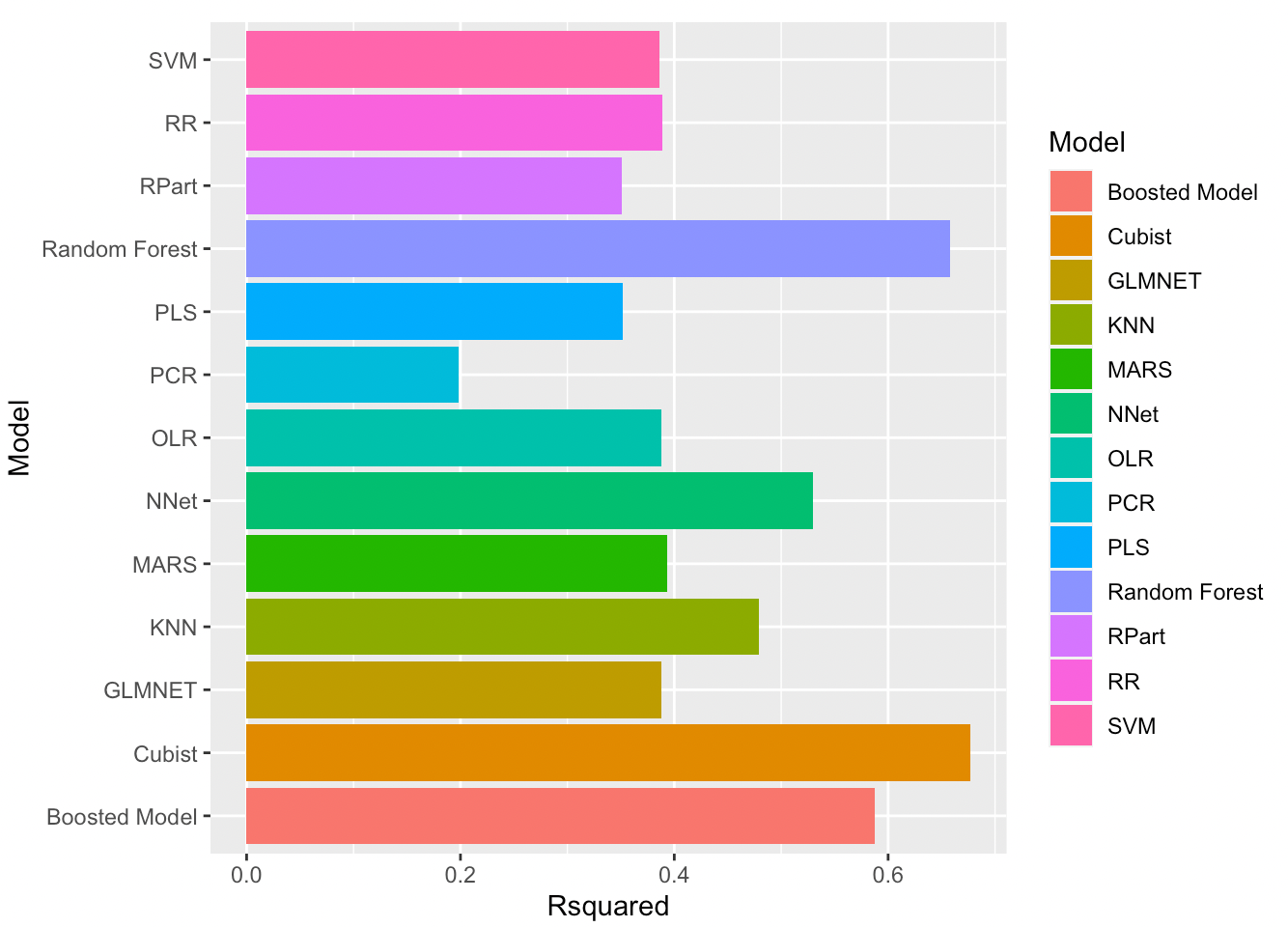


Figure 7 Cubest model also had the highest R-squared value.

# Variable Importance

An analysis of variables importance on the best performing model (the Cubist model) showed that the most important variable in terms of predictive capability for the target variable **PH** is **Mnf.Flow**. **Pressure.Vacuum**, **Density**, **Air.Pressurer** also had high importance in this model.

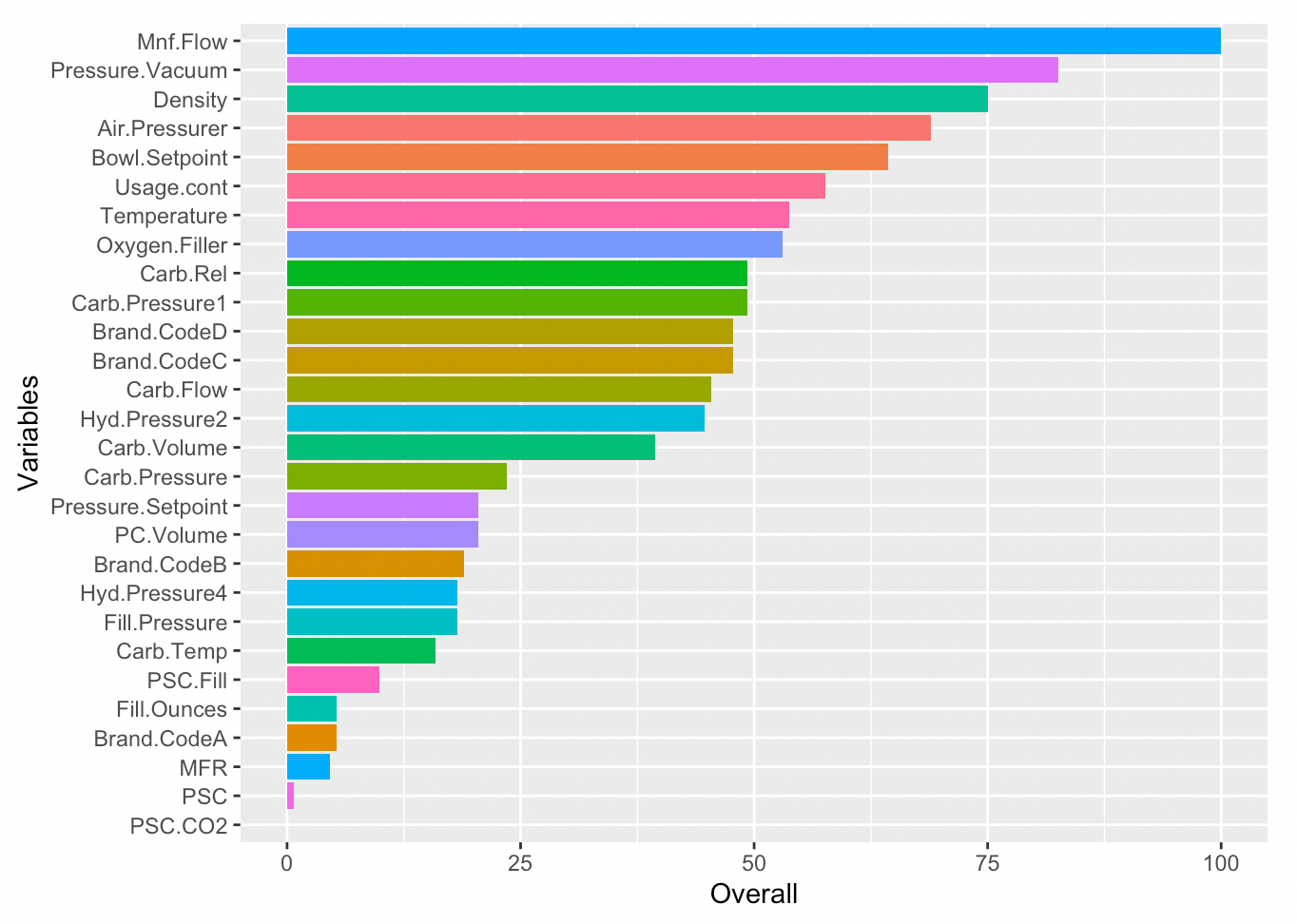


Figure 8 Variable importance.

# Conclusions

Predictive accuracy with these data will likely be highest with the Cubist model, as it had the best performance of all thirteen evaluated models.

**Mnf.Flow** is highly important in the prediction of pH, with other variables such a **Pressure.Vacuum** and **Density** achieving importance as well.