ABC Beverage

PH Predictors

December 2024

Chafiaa Nadour
Darwhin Gomez
John Ledesma
NFN TENZIN DAKAR
Puja Roy
Will Jasmine

Introduction

In response to new regulations requiring ABC Beverage to better understand our manufacturing process and predictive factors for PH levels, our data science team analyzed historical production data to build a reliable forecasting model. The goal was to identify the key factors influencing PH and create accurate predictions to support compliance and process optimization.

Methodology

We started by cleaning the data, addressing missing values, removing duplicates, and encoding categorical variables. Numerical features were standardized to ensure consistency. For exploratory data analysis, we visualized the data, checked for outliers, and analyzed correlations between features.

Next, we trained and tested several models including linear regression, decision trees, neural networks, random forests, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). Model performance was evaluated using RMSE and R-squared metrics.

Final Model and Key Findings

The random forest model delivered the best performance with an R-squared of 0.65 and an RMSE of 0.10. The top five features driving PH predictions were usage_cont, filler_level, temperature, brand_codeC, and carb_flow. SHAP analysis confirmed their importance and provided deeper insights into their impact on PH levels.

PH forecasts have been generated using the final model and added to the StudentEvaluation.xlsx file for further review and action.

The table below highlights the models tested during the study. We evaluated our models using R-squared (R^2) and Root Mean Squared Error (RMSE).

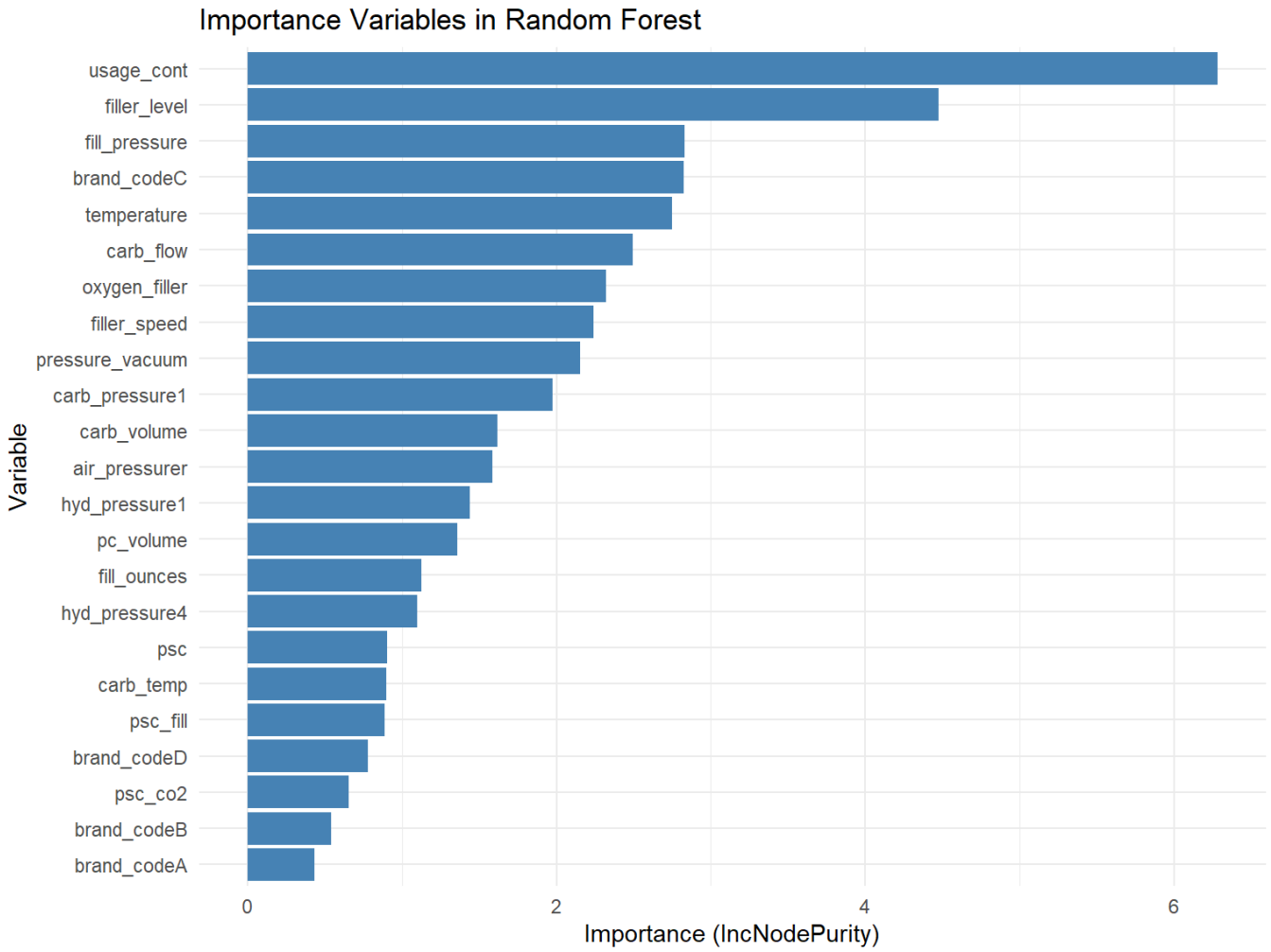
R^2 measures the model's effectiveness in explaining the variance in the target variable, PH. A higher R^2 value indicates a better-performing model.

RMSE represents the magnitude of errors in the model's predictions. For RMSE, smaller values are preferred as they indicate higher accuracy.

```
##            Model      RMSE R_Squared
## 1  Random Forest 0.1023506 0.6513140
## 2            SVM 0.1113341 0.5731233
## 3            KNN 0.1187727 0.5122407
## 4 Neural Network 0.1250552 0.4540662
## 5     Tree-Based 0.1349601 0.3675322
## 6   Linear Model 0.1351587 0.3599708
```

After analyzing the results from our model testing, we selected the Random Forest model as the best option. It achieved an R^2 of 0.65, meaning it explains 65% of the variance in PH values. Additionally, it has the lowest RMSE, indicating it provides the most accurate predictions among all the models we tested.

Below are the variables of importance in predicting PH



usage_cont

- Importance: Ranked the most important variable in the Random Forest model.
- SHAP Insight: SHAP values show both positive and negative impacts on PH, suggesting variability in its effect.
- Correlation Plot: Shows a negative correlation with PH. Higher usage_cont values lead to lower PH levels.
- Conclusion: usage_cont is a strong predictor, likely capturing flow rate and system efficiency, which impact acidity levels.

filler_level

- Importance: The second most important variable.
- SHAP Insight: SHAP values suggest it has a consistent positive influence on PH.
- Correlation Plot: Indicates a slight positive correlation. Higher filler levels tend to increase PH.
- Conclusion: Variations in filler levels affect ingredient concentration, influencing PH balance.

temperature

- Importance: The third most important variable.
- SHAP Insight: SHAP values indicate a negative influence on PH.
- Correlation Plot: Shows a clear negative correlation. Higher temperature leads to lower PH.
- Conclusion: Temperature impacts carbonation and chemical reactions, affecting the acidity and lowering PH.

brand_codeC

- Importance: The fourth most important variable.

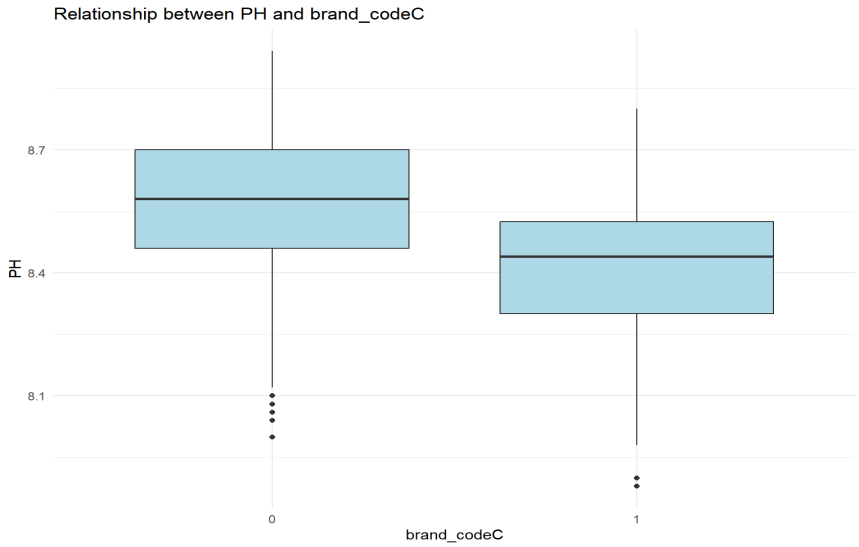
- SHAP Insight: SHAP values suggest a strong negative impact when Brand C is present.
- Boxplot: PH is consistently lower when brand_codeC = 1 (Brand C is present).
- Conclusion: The formulation or process used by Brand C leads to lower PH levels compared to other brands.

fill_pressure

- Importance: The fifth most important variable.
- SHAP Insight: SHAP values show a negative impact on PH.
- Correlation Plot: Displays a negative correlation. Higher fill pressure tends to decrease PH levels.
- Conclusion: Fill pressure affects the pressure conditions during filling, which can influence ingredient concentration and acidity, thereby lowering PH.

carb_flow

- Importance: The sixth most important variable.
- SHAP Insight: SHAP values show a positive impact on PH.
- Correlation Plot: Displays a slight positive correlation. Higher carb_flow increases PH.
- Conclusion: Carb flow affects the amount of dissolved CO_2, altering acidity and thereby increasing PH.

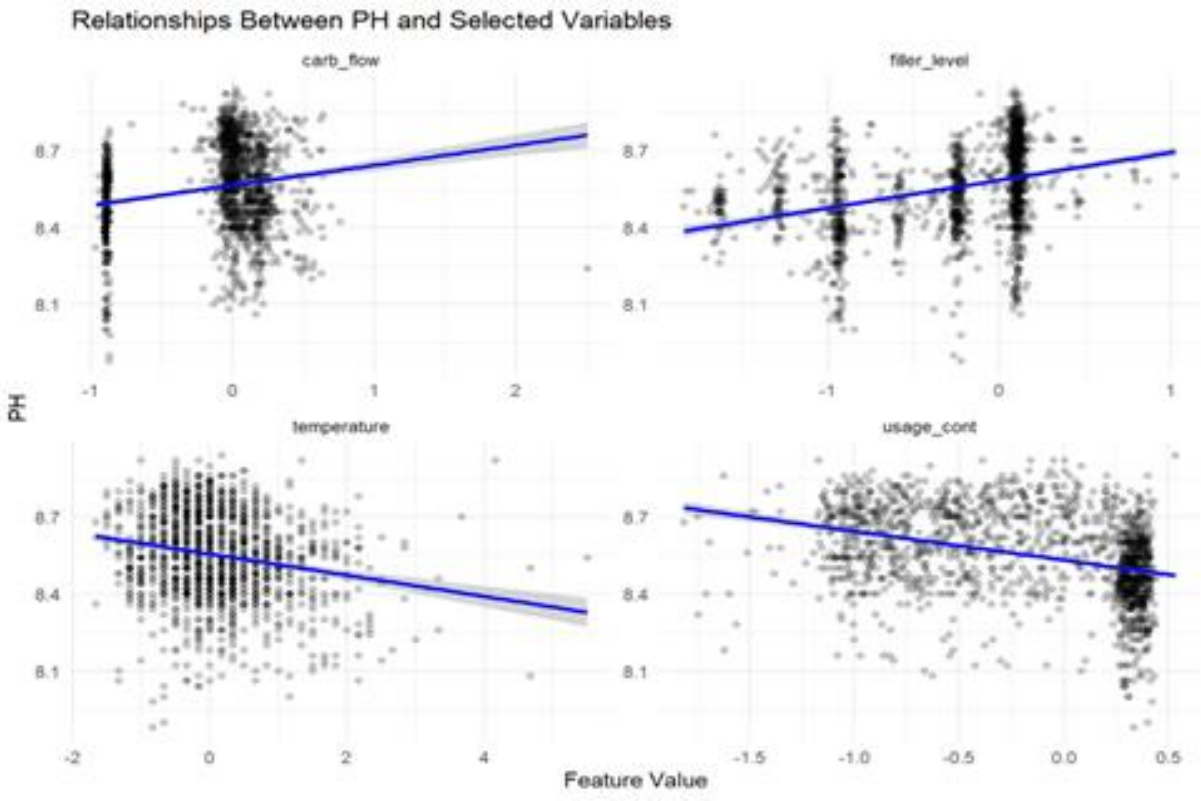


usage_cont and temperature have strong negative relationships with **PH**.

filler_level and **carb_flow** exhibit positive influences on **PH**.

fill_pressure negatively influences **PH**.

brand_codeC distinctly lowers **PH**.



Conclusion

The Random Forest model has proven to be the most effective for predicting PH levels, with an R-squared of 0.65 and an RMSE of 0.10. This means the model explains 65% of the variation in PH values and offers the most accurate predictions among the models we tested. However, 35% of the variation remains unexplained, indicating there is still room for improvement.

Moving forward, we recommend addressing gaps in the data or using imputation methods to handle missing values. Additionally, incorporating more relevant features or refining the existing ones could help improve the model's performance. It's important to note that the current forecasts are based on data without imputations, which may affect accuracy.

Our team is confident in the insights we've provided so far. With more complete data and some refinements, we can achieve even more reliable and accurate predictions in the future.

Please find our enclosed predictions of PH in the file:

StudentEvaluation_WithForecasts.csv