Data 624 Final Business Report

By: Matthew Tillmawitz, Heleine Fouda, Marjete Vucinaj, Lewris Mota, and Kim Koon

Introduction

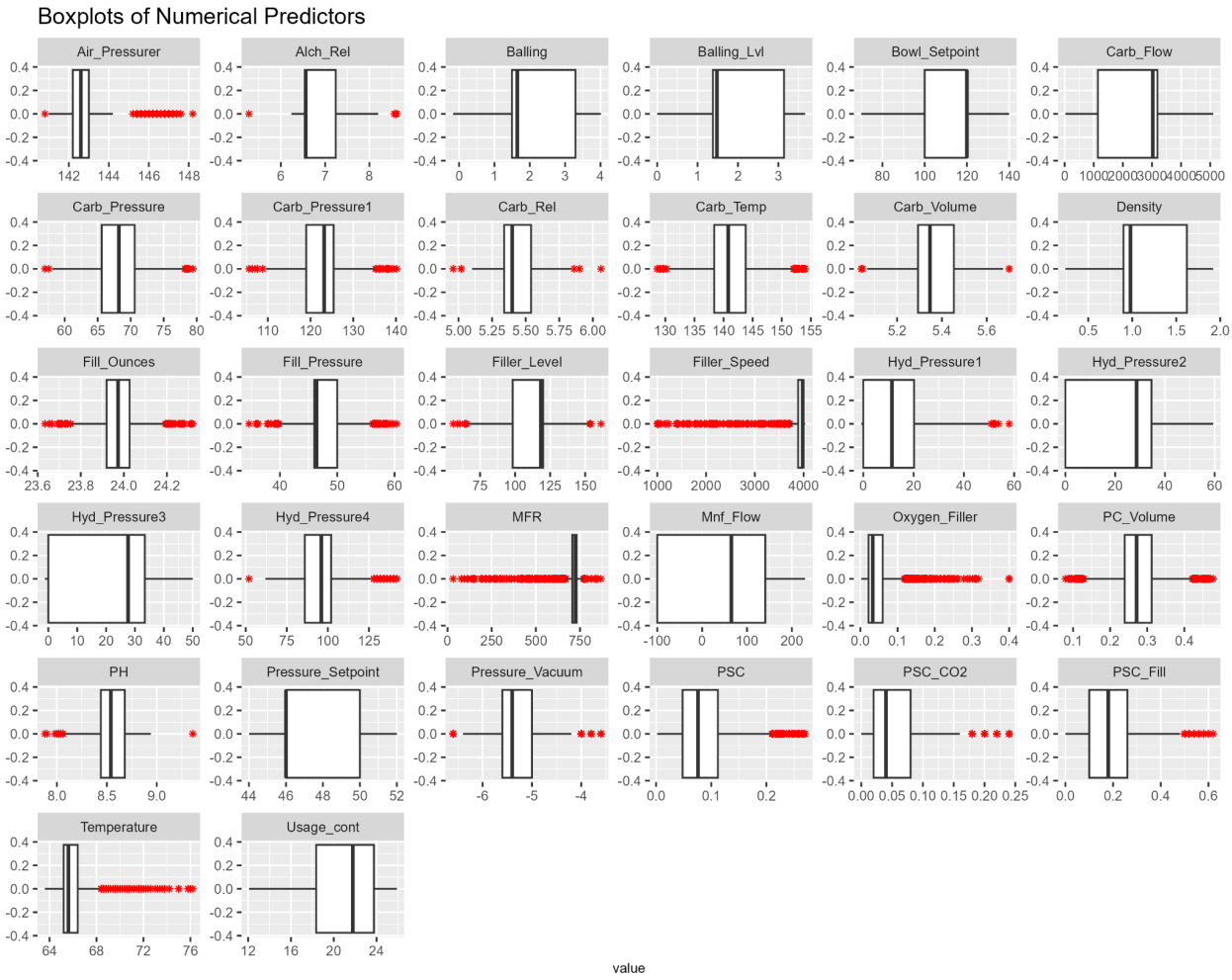
Recent regulatory changes have highlighted the need for a deeper understanding of manufacturing processes and their impact on product quality. At ABC Beverage, pH levels are a critical factor in ensuring consistency and maintaining high product standards. This analysis focuses on identifying the key drivers of pH variability and developing a predictive model that aligns with regulatory requirements. By leveraging advanced analytical and predictive modeling techniques, the study provides actionable insights to enhance process understanding and ensure compliance.

The scope includes data preprocessing, identifying predictive factors, and building a robust pH prediction model. This structured approach uncovers meaningful relationships within manufacturing data, supporting better decision-making. Accurate pH predictions will not only meet regulatory standards but also improve quality control, reduce variability, and ensure ABC Beverage consistently delivers exceptional products. These findings will enable the company to address compliance needs while optimizing its manufacturing processes through data-driven strategies.

High Level EDA

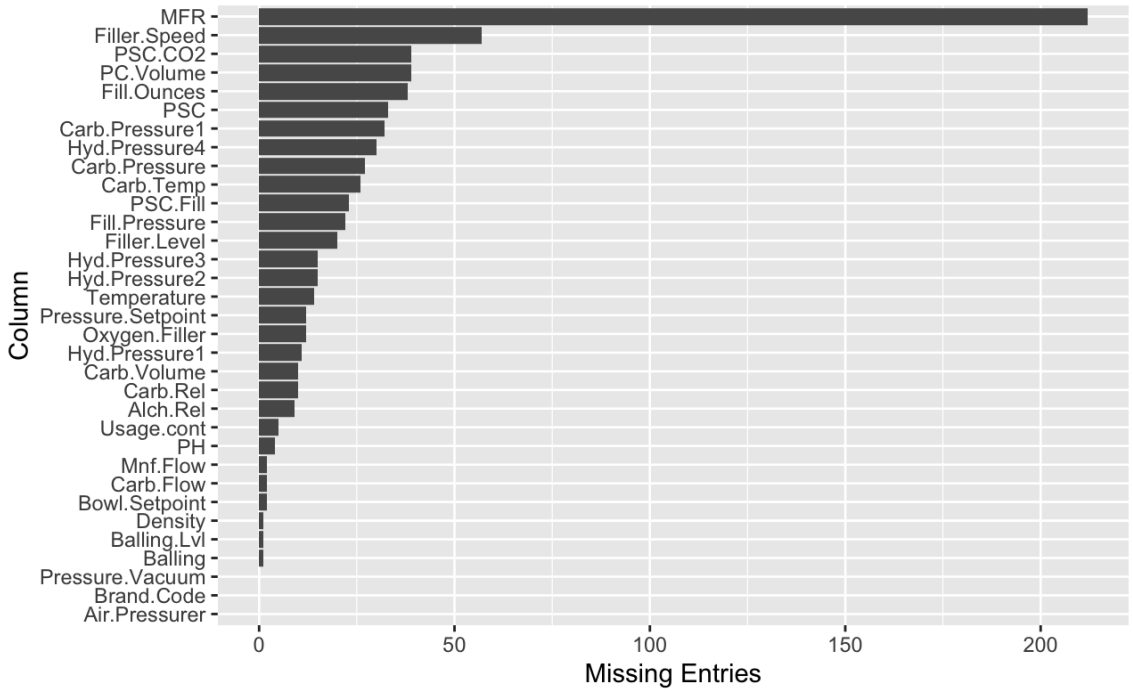
The dataset provides a detailed snapshot of variables that are predominantly numeric, along with a single character variable, Brand_Code. Most variables exhibit high levels of completeness, with many exceeding 98% complete. The target variable, pH, is highly complete (99.8%) with only four missing values. See the Imputation section of this report for further discussion on missing values.

The numeric variables in the dataset reveal diverse distributions and variability. Some, such as Carb_Volume, Fill_Ounces, PSC, and Temperature, have tightly clustered values and small standard deviations, indicating consistent measurements. On the other hand, variables like Filler_Speed, MFR, and Carb_Flow show wide ranges and high variability, suggesting potential outliers. Distributions are also varied, with variables like PSC_Fill and Alch_Rel showing skewness, while Filler_Speed exhibits a peaked distribution with clustering near the upper bounds. This variability underscores the need for preprocessing steps such as normalization or transformations to handle skewness and outliers effectively.



Imputation

Missing data is a reality that needs to be dealt with in order to get the best performance from our models. Instead of ignoring or dropping all the observations with missing data, we can impute the values using a method known as Boosted Aggregation or “Bagging”. This method performs better than other methods such as KNN when there are many variables with missing values as it can use the missingness as a factor when imputing. The figure below demonstrates why this is desirable, as only three of our variables are not missing any observations. As we can see in the same figure, the MFR variable is missing the most observations by far and at first glance could be worrying. The proportion of missing entries for this variable is not large enough at slightly over 8% for us to justify ignoring it, and it should be evident that discarding over 200 observations would have an adverse effect on our models. Using Bagging we are able to impute values for all the missing data in a manner that will not skew or unduly influence model performance the way other replacement methods, such as averaging, would. It should be noted that there are a few observations in our training data missing pH readings. These observations are being dropped from our training data, as there are only a few and pH is the variable we are training our model to predict.


MFR
Filler.Speed
PSC.CO2
PC.Volume
Fill.Ounces
PSC
Carb.Pressure1
Hyd.Pressure4
Carb.Pressure
Carb.Temp
PSC.Fill
Fill.Pressure
Filler.Level
Hyd.Pressure3
Hyd.Pressure2
Temperature
Pressure.Setpoint
Oxygen.Filler
Hyd.Pressure1
Carb.Volume
Carb.Rel
Alch.Rel
Usage.cont
PH
Mnf.Flow
Carb.Flow
Bowl.Setpoint
Density
Balling.Lvl
Balling
Pressure.Vacuum
Brand.Code
Air.Pressurer
Column
0
50
100
150
200
Missing Entries

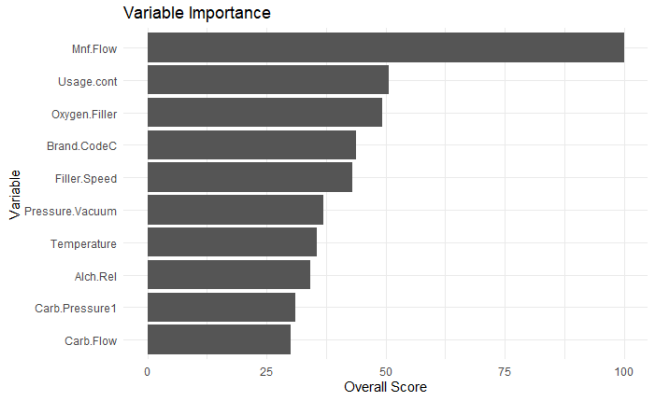
Model Selection

Model	R^2	RMSE	Advantages	Limitations
Random Forest	0.73	0.09	• Simple configuration with strong performance • More robust tuning, reduces risk of overfitting	• Requires more time and computational power • Resource Intensive with Cross-Validation
XGBoost	0.72	0.09	• High accuracy and performance • Optimized through tuning	• Requires more time and computational power • Lack of interpretability
Cubist	0.71	0.09	• Rules used are easy to access and understand • Ensemble method resilient to outliers and missing data	• Rules give a false sense of interpretability • Multiple ways to calculate predictor importance
Support Vector Machine with RBF	0.61	0.11	• Often achieves high predictive accuracy for classification and regression tasks, especially with well-tuned hyperparameters.	• Performance can degrade in the presence of high noise or overlapping classes.
Averaged Neural Network	0.60	0.11	• Automatically distinguishes important variables • Less prone to overfitting compared with a single neural network	• Less interpretability • Long computation time
Neural Network	0.55	0.12	• Automatically distinguishes important variables	• Less interpretability • Long computation time
CART	0.48	0.12	• Interpretability • Models non-linear relationships well	• Overfitting risk • Instability
LASSO	0.43	0.13	• Linear model with high interpretability • Conducts feature selection • Able to handle highly dimensional and highly correlated data	• Poor performance compared to non-linear models • Less regularization than Ridge
Elastic Net	0.43	0.13	• Generalization of LASSO with the benefits of Ridge	• Poor performance compared to non-linear models
Ridge Regression	0.41	0.13	• Linear model with high interpretability • Resilient to overfitting • Able to handle highly dimensional and highly correlated data	• Poor performance compared to non-linear models • No feature selection
PLS	0.37	0.14	• Handles multicollinearity • High dimensional data	• Not interpretable • Assumes linear relationships

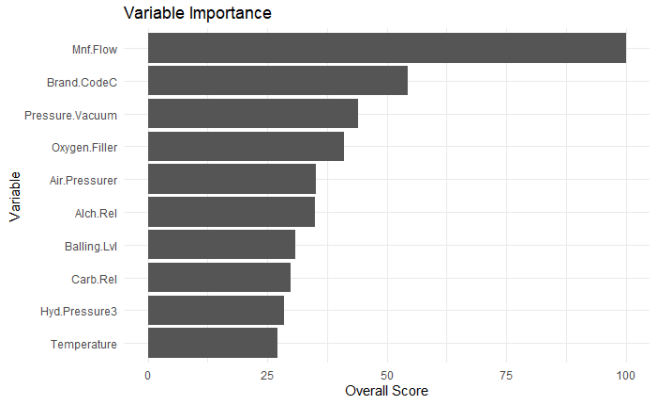
Selection of the Random Forest Model

From the model summary table above, the best performing models were the XGBoost model and the Cross-Validated Random Forest model. The Random Forest model and the XGBoost model exhibited similar performance metrics, both with an R^2 of approximately 0.72 and both with an RMSE of 0.89 on the test set, and similar variable importance distributions. While XGBoost models are less prone to overfitting due to its built in regularization, comparing the in-sample R2 and RMSE to the out-of-sample R2 and RMSE show that neither models are overfitted. As a result, the Random Forest model was chosen for its better interpretability.

XGBoost



Random forest



Conclusion

This project evaluated multiple predictive models to meet the new regulatory requirement of understanding and predicting pH levels in ABC Beverage's manufacturing process. The analysis began with data preprocessing, including bagging imputation for missing values, removing rows with missing pH values, and one-hot encoding of the Brand.Code variable.

A range of models were tested, including linear (PLS, Ridge, LASSO), non-linear (SVM, Neural Networks), tree-based (CART, Random Forest, XGBoost), and hybrid models (Cubist). While Cubist (R^2 =0.66) and XGBoost (R^2 =0.71) demonstrated strong performance, they were not selected due to limited interpretability and weaker training performance.

The Random Forest model was chosen as the best-performing model. It achieved an R^2 of 0.73 on the test set, with an RMSE of 0.089 and an MAE of 0.065, demonstrating excellent predictive accuracy and generalization. Key predictors identified by Random Forest include Manufacturing Flow Rate (Mnf.Flow), Continuous Usage (Usage.cont), and Oxygen in Filler (Oxygen.Filler), all critical for regulating pH levels and ensuring product consistency.

Future improvements include incorporating additional predictors, enhancing feature engineering, and increasing the dataset size for deeper insights. Exploring more interpretable models or refining current models through advanced tuning could further enhance performance while maintaining transparency. Accurate pH predictions are essential for regulatory compliance, improving quality control, and reducing variability, ensuring ABC Beverage delivers consistent, high-quality products.

References:

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://link.springer.com/book/10.1007/978-1-4614-6849-3

CUNY School of Professional Studies – The Graduate Center. (2024). *DATA 622: Machine learning* [Course]. CUNY School of Professional Studies – The Graduate Center, Fall 2024

Näf, J. (n.d.). *What is a good imputation for missing values?* Towards Data Science. Retrieved [Dec 13 2024], from https://towardsdatascience.com/what-is-a-good-imputation-for-missing-values-e9256d45851b