Fabric Conditioner New Predictive System

Data Science Challange

Carlos Corro

Definition

Project Overview

Measuring the density of fabric conditioner can pose several challenges, particularly in terms of costs and labor involved. Obtaining accurate density measurements often requires specialized equipment like digital density meters or hydrometers designed for liquids. These devices can be expensive, and the cost might increase based on the precision and automation level required for the measurements.

I developed a predictive density model using machine learning algorithms that can accurately estimate the density of a substance based on various input parameters. The primary aim is to reduce the costs associated with traditional density measurements.

Problem Statement

The goal is to create model that can precisely predict density levels based on specific input parameters. The tasks involved are the following:

1. Data Cleaning and Organization.
2. Applied Supervised Regression Model for Predicting Density
3. Evaluate the model's performance using appropriate metrics.

If the model performs well, it will be used in the Globins Care Co.

Analysis

Data Exploration

The data was obtained using a Data Engineering tool developed by SMi that uses Machine Learning models to extract useful information from the produced batches based on cloud data. The batch manufacturing process is divided in phases, and the Tool can analyze each batch and get specific data from the phases during the production of each batch, for example: Temperature average in phase 2, pressure maximum value in phase 4, weight of the tank in phase 1, etc.

The dataset has the feature description of the dataset below:

1. SITE: Name of the site.
2. DIVISION AND CATEGORY: internal product classification in Globins Care Co.
3. LINE: Production line where the data was collected.
4. MODEL_NAME: Machine Learning model applied in the data extraction.
5. RECIPE_GROUP: Recipe group of the analyzed product.
6. UNIQUEID: Identification number of the batch.
7. PRODUCT_CODE: Identification number of the product.
8. MODEL_COMPLETE: Integrity of the batch data: "COMPLETE" means that the batch was fully analyzed by SMi Data Engineering tool. "INCOMPLETE" means that the model has failed, or the batch contains atypical characteristics.
9. CHAR_NAME: String containing the description of the characteristic measured like phase number, phase name, and metric evaluated.
10. CHAR_TIME: Timestamp of the measurement.
11. CHAR_VALUE: Value measured by SMi Data Engineering tool.

Data Cleaning and Organization

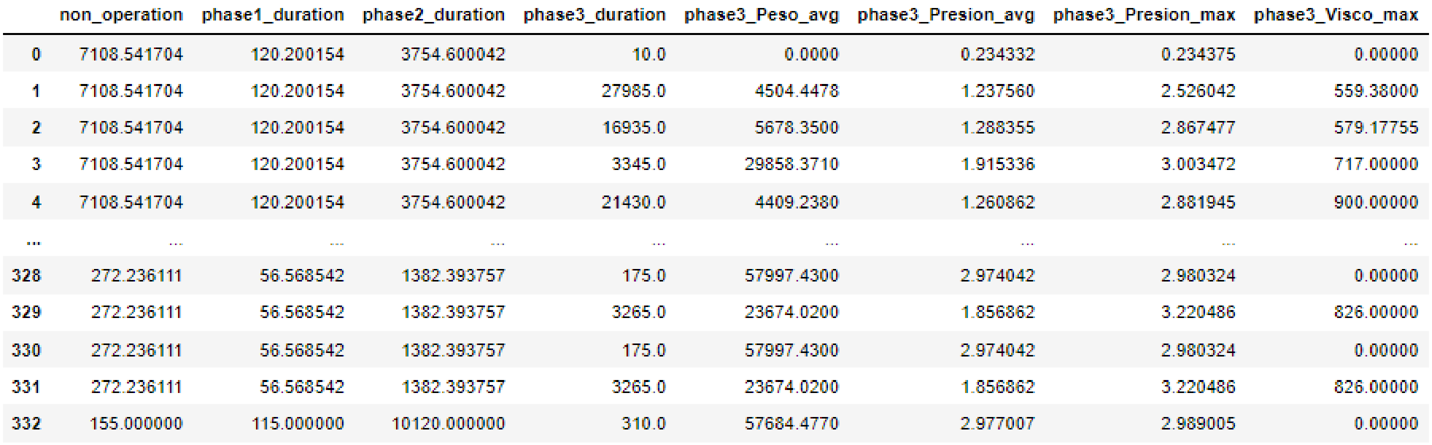
The Features that are purely descriptive will be dropped. We will only use the relevant columns to run the model, which are UniqueID, Model_Complete, Char_Name, Char_Time, and Char_Value.

As stated in the feature description, we will filter the dataset based on the Model_Complete column and use only the batches that have been completed. This is to prevent the model from attempting to predict incomplete or atypical information. Afterward, the 'Model_Complete' column will be dropped.

Next, we will pivot the 'Char_Name' column to examine the production phases. And we noticed that only the data from phase 3 is connected by UniqueId and Char_Time. The other phases, like Non-Operation Value, Phase 1 and Phase 2 are no connected. To use the data from these phases, we will employ standard deviation to classify them as time variability since the data in these phases consist of time measurements.

Inferring that it's not possible to have negative values in variables such as density, weight, pressure, and duration, we will drop the rows containing negative values.

Fig. 1 Part of Data Frame after the cleaning and data processing mentioned above.

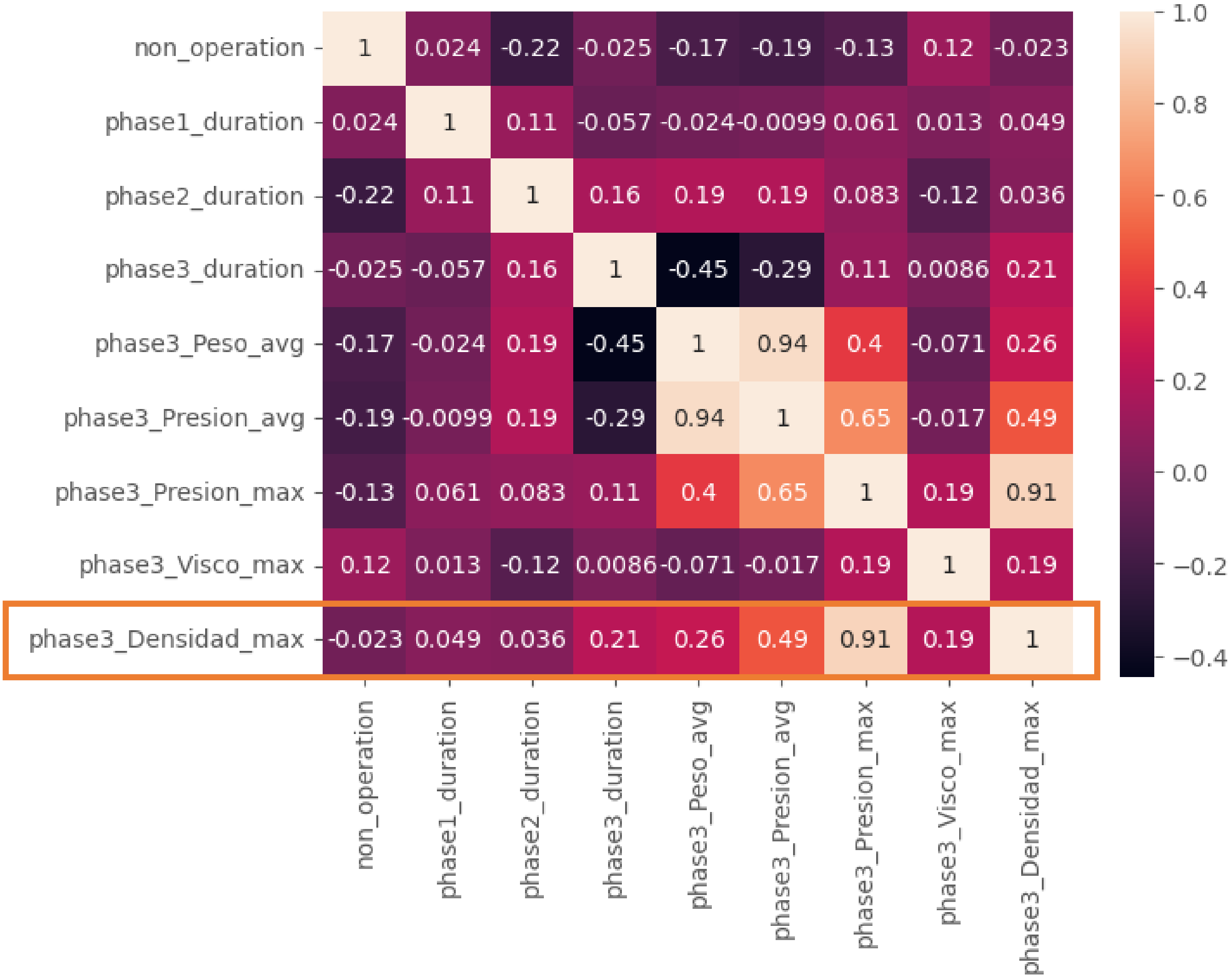


	non_operation	phase1_duration	phase2_duration	phase3_duration	phase3_Peso_avg	phase3_Presion_avg	phase3_Presion_max	phase3_Visco_max
0	7108.541704	120.200154	3754.600042	10.0	0.0000	0.234332	0.234375	0.00000
1	7108.541704	120.200154	3754.600042	27985.0	4504.4478	1.237560	2.526042	559.38000
2	7108.541704	120.200154	3754.600042	16935.0	5678.3500	1.288355	2.867477	579.17755
3	7108.541704	120.200154	3754.600042	3345.0	29858.3710	1.915336	3.003472	717.00000
4	7108.541704	120.200154	3754.600042	21430.0	4409.2380	1.260862	2.881945	900.00000
...	...	...	...	...	...	...	...	...
328	272.236111	56.568542	1382.393757	175.0	57997.4300	2.974042	2.980324	0.00000
329	272.236111	56.568542	1382.393757	3265.0	23674.0200	1.856862	3.220486	826.00000
330	272.236111	56.568542	1382.393757	175.0	57997.4300	2.974042	2.980324	0.00000
331	272.236111	56.568542	1382.393757	3265.0	23674.0200	1.856862	3.220486	826.00000
332	155.000000	115.000000	10120.000000	310.0	57684.4770	2.977007	2.989005	0.00000

Exploratory Visualization

The plot below shows the correlation matrix of the processed data.

Fig. 2 A plot showing the correlation matrix of the features. And we can observe that the dependent variable has a high correlation with pressure and none correlation with time variations in the non-operational, phases 1 and phase 2.



Due to a correlation very close to 0 among the variables **non-operation**, **phase 1** and **phase 2**, we will discard these features. We will only use the features from phase 3 to run the model.

Fig. 3 Analyzing the histograms of the features, it's possible to observe that the data of phase3_duration and phase3_Visco are mostly concentrated on the left side, while the features phase3_ Presion_max and phase3_Visco_max are more concentrated on the righ side, possibly indicating an asymmetric distribution, perhaps with predominance of smaller values. Meanwhile, the feature Presion_AVG is well distributed.

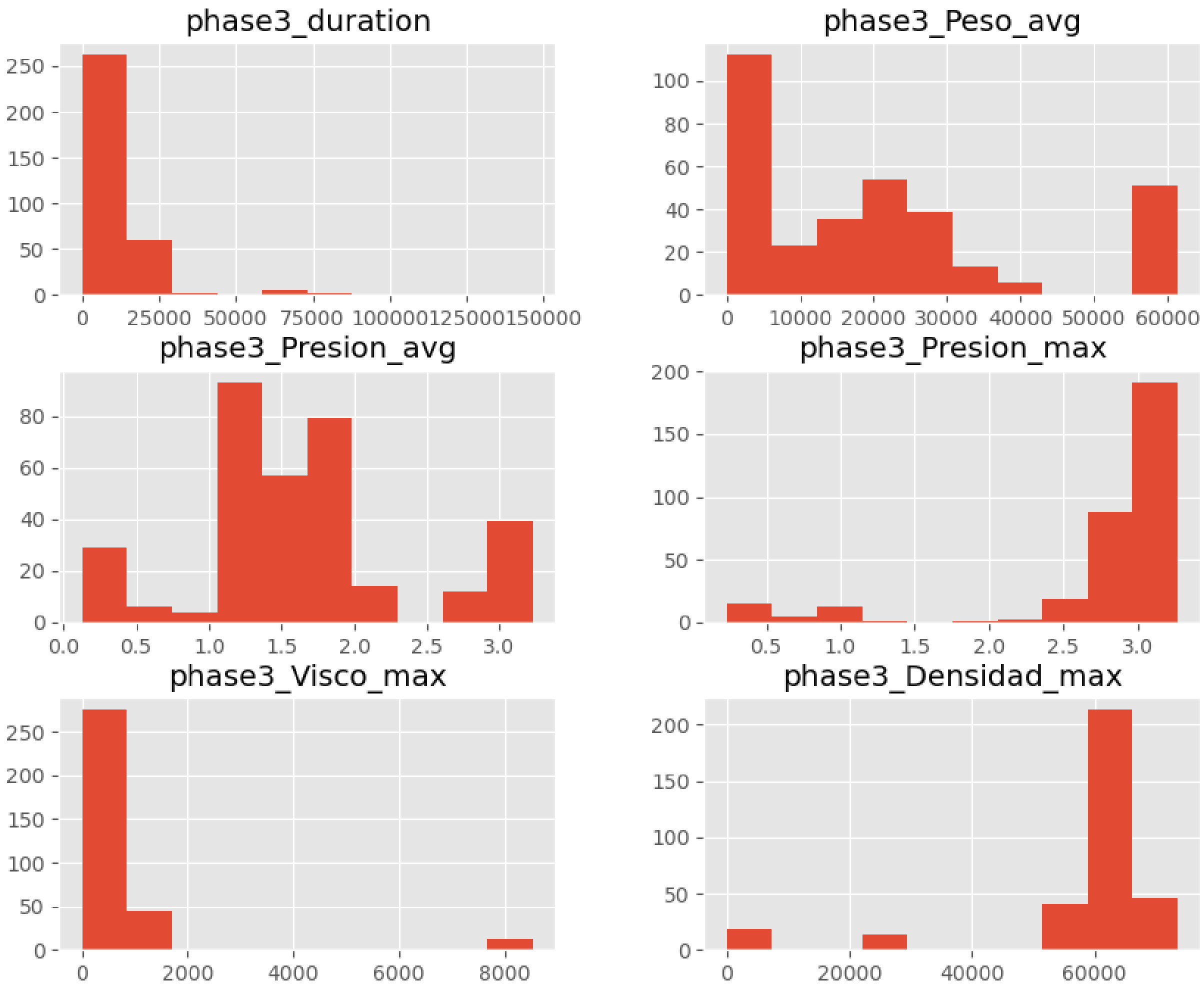
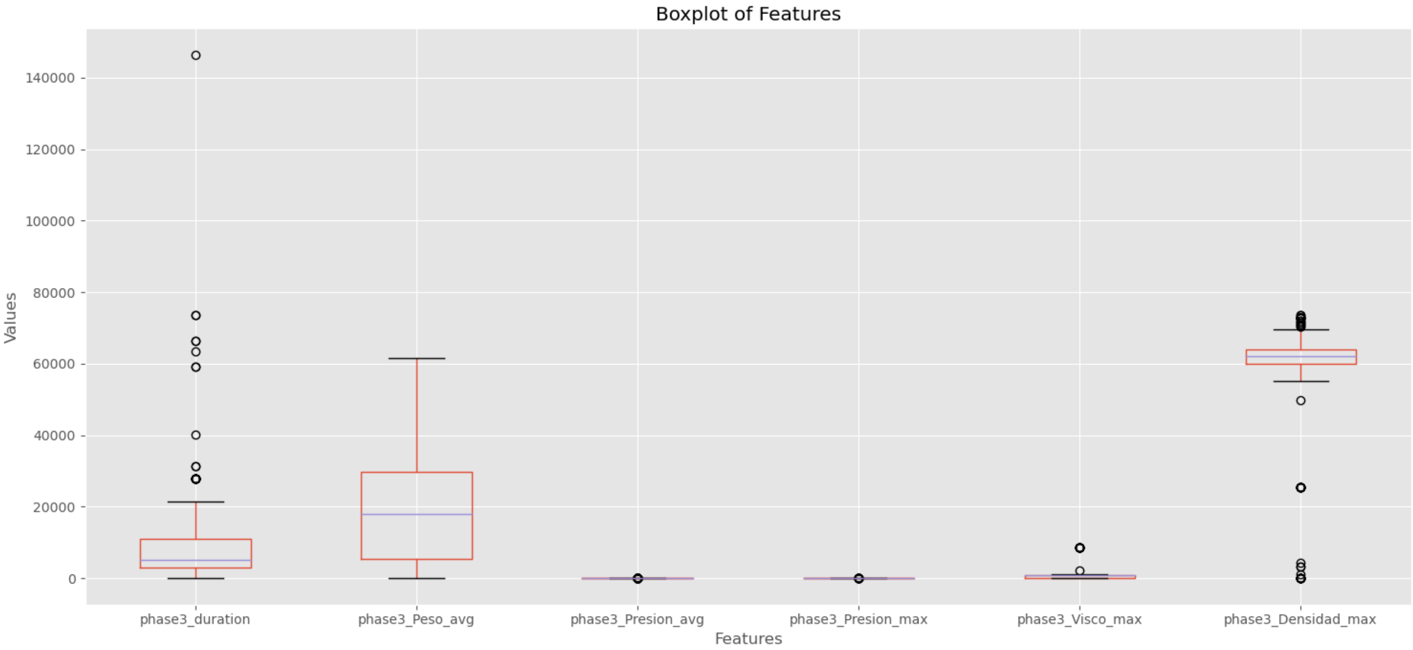
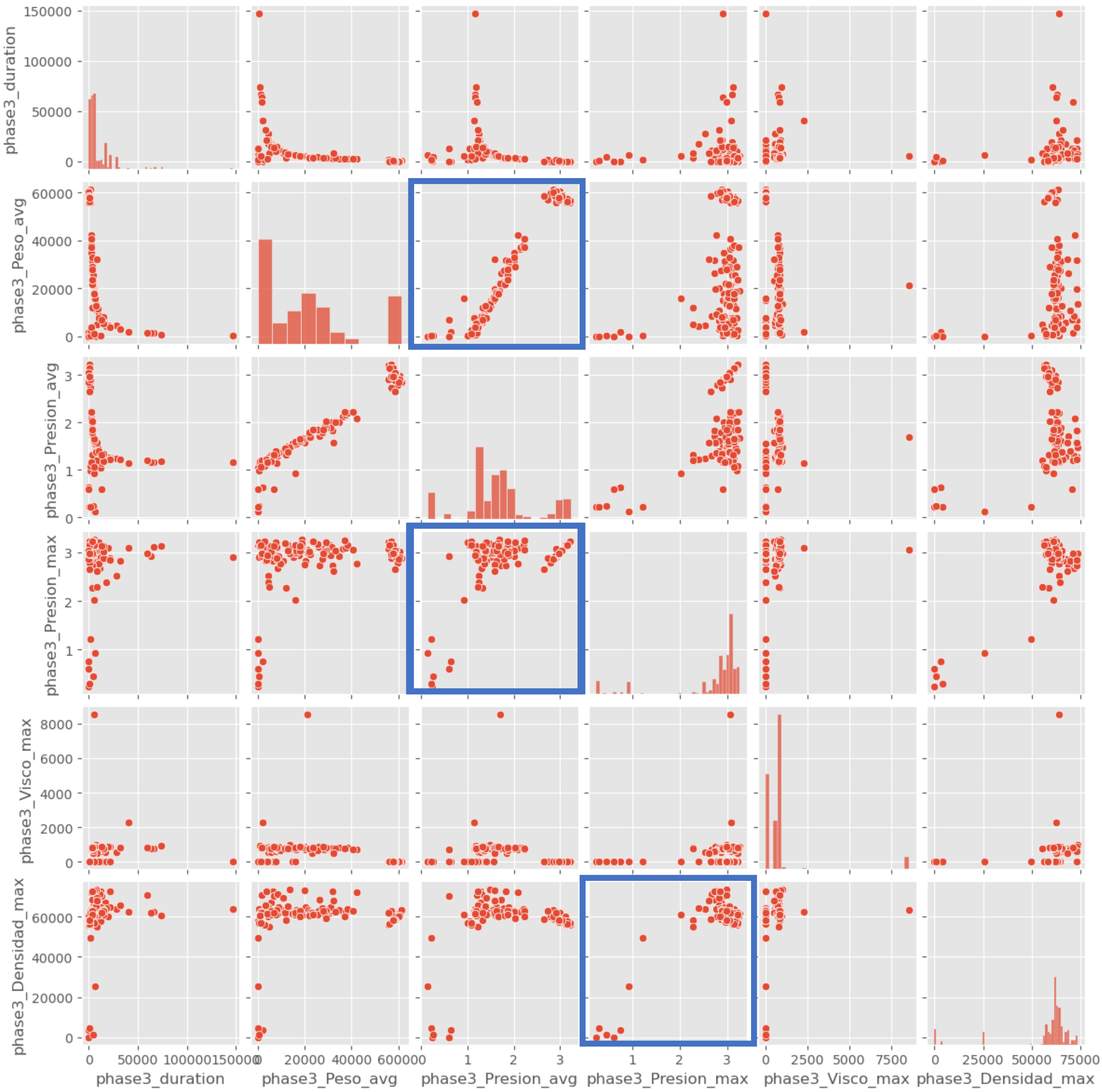


Fig. 4 From the boxplots below, it's possible to observe that there are many outlier data points.



However, after researching and analyzing, it's possible to infer that the outliers should not have been removed as they constitute a crucial part of the data. Since the data represents different products in different batches in a production, it's plausible to suggest that various types of products within the fabric conditioner can directly impact weight, density, pressure, and viscosity. Therefore, it was decided to retain the outliers to keep the model more faithful to the real-world scenarios.

Fig. 5 We can observe in the scatterplots a linear relationship between the features Peso_avg with Presion_avg, Presion_max with Presion_avg, and Densidad_max with Presion_max. We can see that the scatterplots of the rest of the variables are mostly random and show no relationship among them.



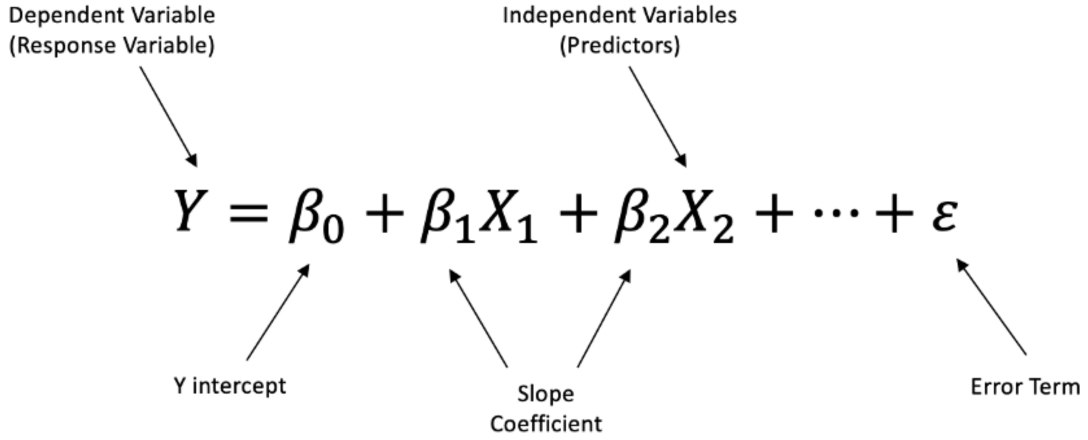
phase3_duration
phase3_Peso_avg
phase3_Presion_avg
phase3_Presion_max
phase3_Visco_max
phase3_Densidad_max
150000
100000
50000
0
60000
40000
20000
0
3
2
1
0
3
2
1
8000
6000
4000
2000
0
60000
40000
20000
0
0
50000
100000150000
0
20000
40000
60000
0
1
2
3
1
2
3
0
2500
5000
7500
0
25000
50000
75000
phase3_duration
phase3_Peso_avg
phase3_Presion_avg
phase3_Presion_max
phase3_Visco_max
phase3_Densidad_max

Machine Learning Model

Multiple Linear Regression Model

Due to a correlation very close to 0 among the variables non-operation, phase 1, and phase 2, we will discard these features. We will only use the features from phase 3 to run the model.

Fig. 6 Multiple Linear Regression Model Formula.



By using the sklearn multiple linear regression as the formula above illustrate. The variables are:

1. Dependent Variable: phase3_Densidad_max
2. Independent Variables: phase3_duration,
   phase3_.Peso_avg,
   phase3_Presion_avg,
   phase3_Presion_max,
   phase3_Visco_max.

After normalizing the data and splitting it into training and testing sets in an 80% to 20% ratio, randomly, the model reached the following Coefficients and Intercept numbers:

1. Coefficients: 0.33168688, 0.35572476, -0.6084219, 1.00683016, 0.00170674
2. Intercept: 0.10829946

Based on the coefficients, it's possible to infer some hypotheses. Like the relationship between Presion_Avg and Density, it suggests an inverse relationship. In this case, an increase in average pressure is associated with a decrease in the product's density. This relationship could be explored to understand how pressure affects density during the conditioner's production.

Another inference is that the remaining variables show a positive relationship with the dependent variable, except for the Visco_Max variable, which is close to zero.

Metrics

The selection of specific metrics like Explained Variance, Mean Squared Log Error, R-squared, MAE, MSE, and RMSE for evaluating multiple linear regression analysis is grounded in their unique abilities to assess different aspects of model performance.

- Explained Variance captures the proportion of variance explained by the model. It aids in understanding how well the model fits the data by quantifying the variance captured by the predictors.
- Mean Squared Log Error (MSLE) evaluates the accuracy of predictions on a logarithmic scale. It penalizes underestimation and overestimation equally and is particularly useful when the target variable has exponential growth patterns.
- R-squared (R2) signifies the goodness of fit, measuring the proportion of variance in the dependent variable explained by the independent variables. It provides an overall assessment of how well the model fits the observed data.
- Mean Absolute Error (MAE) calculates the average of the absolute differences between predicted and actual values. It provides insight into the average prediction error without considering the error's direction.
- Mean Squared Error (MSE) measures the average squared differences between predicted and actual values. Squaring errors emphasize larger errors, making it sensitive to outliers.

- Root Mean Squared Error (RMSE) represents the square root of MSE and provides an interpretable measure of the average magnitude of error. It helps in understanding the model's predictive performance in the original scale of the target variable.

The results of the model metrics were:

1. Explained_variance: 0.8967
2. Mean_squared_log_error: 0.0027
3. R2: 0.8967
4. MAE: 0.0632
5. MSE: 0.0073
6. RMSE: 0.0854

Fig. 7 As we can see in the graph, the prediction regression line follo ws the test data.

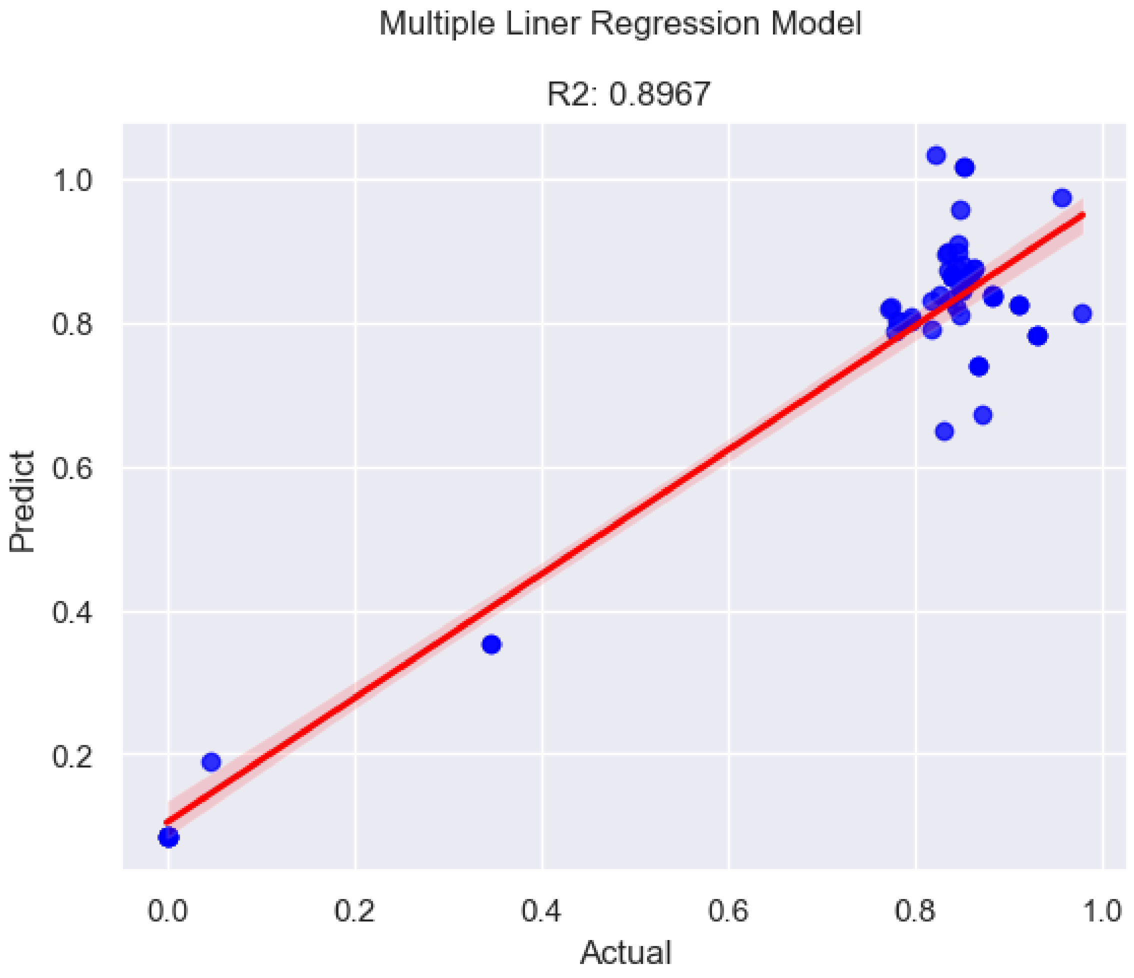
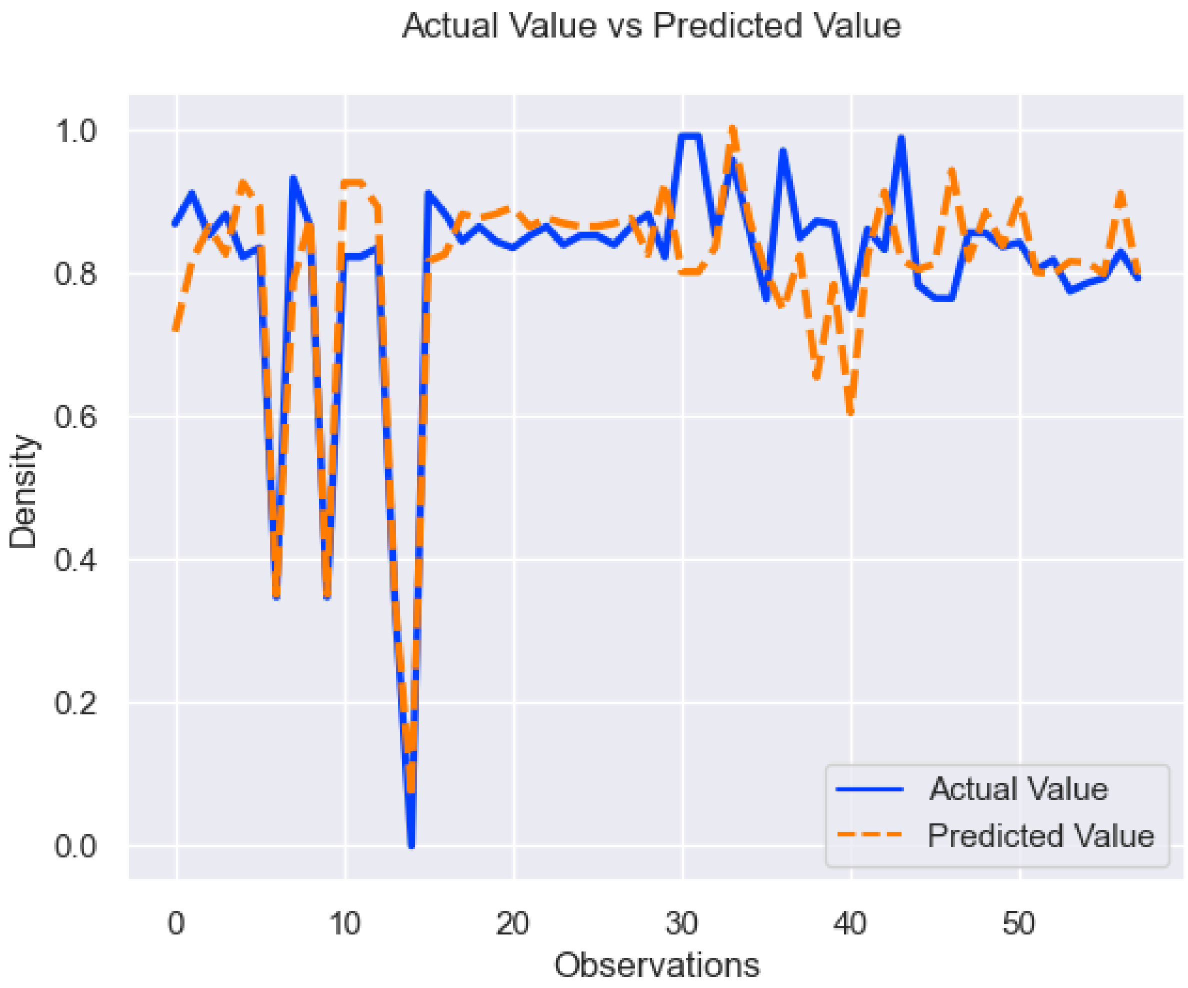


Fig. 8 The graph below compares the test data with the prediction. As evaluated by the R^2 metrics and errors, the image clearly illustrate s how well the model tracks the actual data and its ability to predict with minimal errors.



Conclusion

The metrics obtained from the multiple linear regression model showcase a compelling performance in predicting density. An explained variance of 0.8967 suggests that approximately 90% of the variance in the density of the Fabric Conditioner can be explained by the model, indicating a strong capacity to capture variability. The low mean squared log error (0.0027) signifies accurate predictions, particularly valuable when dealing with exponential growth patterns in density. Additionally, an R2 score matching the explained variance underlines the consistency of the model's fit.

The mean absolute error (MAE) of 0.0632 and mean squared error (MSE) of 0.073 demonstrate the average magnitude and squared differences between predicted and actual densities, respectively. Meanwhile, the root mean squared error (RMSE) of 0. 0854 provides an interpretable measure of prediction error in the original scale, aiding in understanding the model's predictive accuracy.

For an industry seeking to predict density, these metrics serve as vital indicators of the model's robustness. The high explained variance and R2, coupled with low error metrics like MSE, MAE, and RMSE, suggest that this multiple linear regression model reliably predicts Fabric Conditioner density. This reliability could profoundly benefit the industry by optimizing manufacturing processes, ensuring consistent product quality, and potentially reducing costs associated with manual density measurements. The model's accuracy and reliability in predicting density offer a valuable tool for decision-making and process optimization within the Globins Care Co.

Fabric Conditioner New Predictive System

Carlos Corro

LinkedIn

Github

Email