Predictive analytics to predict quality of product

DSA8023 Analytathon1 Yogesh Kashiram Bore

1. Proposed Challenge: To predict product quality at various stages of the production process.

2. Exploratory Data Analysis:

1. **Data Observations:** Dimensionality (columns) is 37 and size (rows) is 15499. A manufacturing process that occurs sequentially and divided into 6 stages. First 21 rows of null values after inspection for one variable " g4_var_2".Two variables "g3_var_3" and "g6_var_9" have of values 1's and 0's with no variance. Outliers identified with simple summary statistics (max and min is far off from median). Plotted the box plots to check the outlier *Fig (a).* Discover new patterns or associations between all the variables using correlation matrix. All data is divided into 6 groups and 5 stages. Checked the autoregressive nature of data by using Autocorrelation plot (ACF) *Fig (b).*

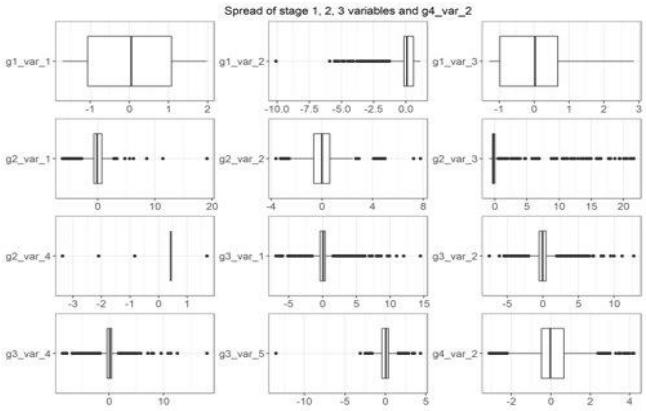


Fig. (a) Box plots

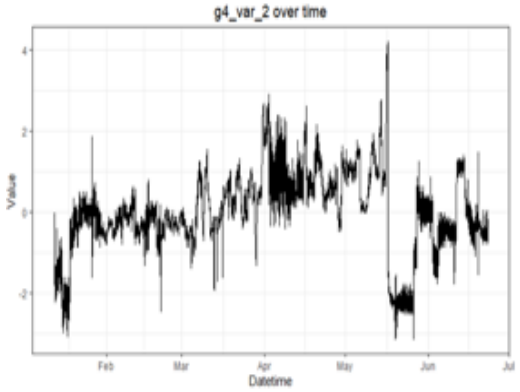


Fig. (b) Line graph

2. **Data Wrangling:** Imputation of missing values (Null values) with median as it is robust to outliers compared to mean. For g4-var2 column datatype changed to numeric from character.
3. **Model Selection:** Given we are keeping the outliers, we chose to use a model that was robust to outliers, random forest. The problem statement required that the analysis be explainable, so a black box model such a deep learning models would not be appropriate. We wanted to try both regression and classification to see how the predictive power of each approaches would differ.
4. **Techniques for model evaluation:** We have generated models having a different coefficient of determination R-squared (R2) statistical measure and Root Mean Square Error (RMSE) standard way to measure the error of a model in predicting quantitative data.

3. Model Implementation: As decided in the EDA phase we chose generalized additive models (GAM), generalized linear model (GLM) - Logistic regression, Random Forest for our prediction.

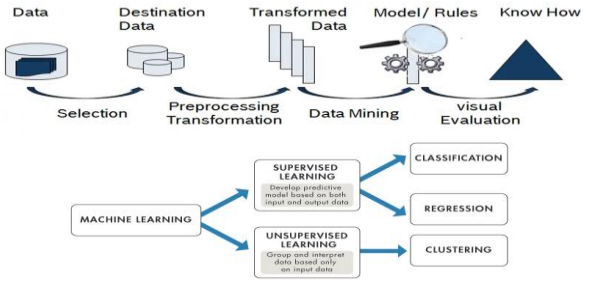


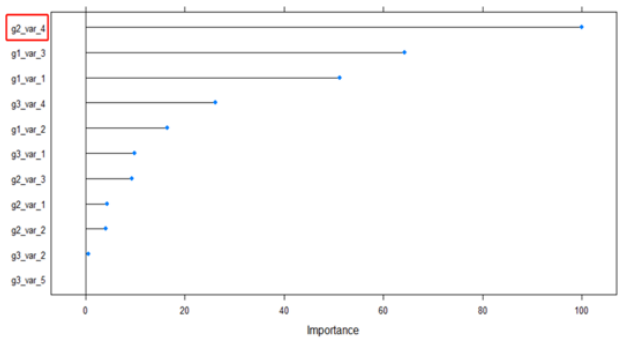
Fig. (c) Model Architecture

Model	R-Square	RMSE
GAM	0.706524	0.555097
GLM	0.2857014	4.153211
Random Forest	**0.9438298**	**0.238951**

Table (a): 3 model Results

Random Forest	Stage I	Stage II	Stage III
RMSE	0.2905181	0.2409913	0.2389513
R Squared	0.9185517	0.9428694	**0.9438298**

Best Model Results : Random Forest: Among the 3 models tried, we got best result for random forest.



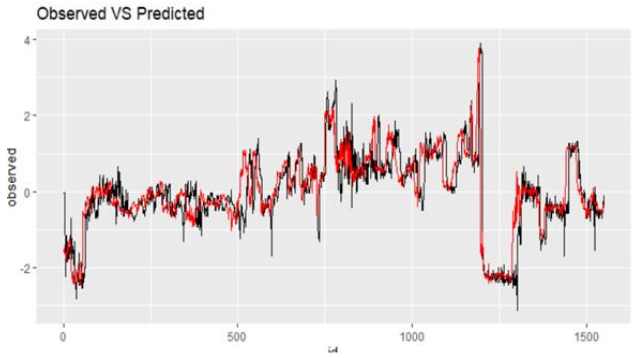
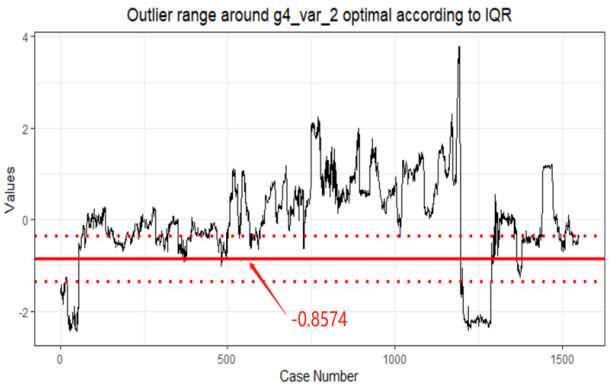


Fig. (d) Variable Importance

Fig. (e) Observed VS Predicted Line Graph

Variable importance *Fig. (d)* illustrates that a g2 variable and two g1 variables are more important. And we tried this model with all given data and same model with 3 different stages data using lag function and got a good accuracy of **94 %** as shown in the table a. Also the obeserved the predicted values are shown in *Fig. (e)* where red line shows the predicted value and black line observed values. Optimal value is - 0.8574.



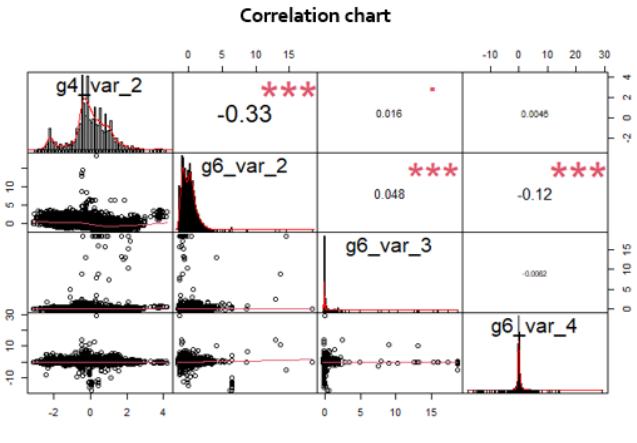


Fig. (f) Optimal Value IQR Line Graph

Fig. (g) Correlation Chart

In *Fig. (f)* predicted valued plotted along with inter quartile range of 0.25 to 0.75 and observed results.

4. Relationship Between Target variable:

Model	Group6 varible	Correlation Coefficient	Relationship
g4_var2	g6_var2	-0.33	Weak
g4_var2	g6_var3	-0.12	Very Weak Negative
g4_var2	g6_var4	0.05	Very Weak Positive

Table (b) Correlation

We find the relationship between the target variable g4_var2 and group 6 variables using correlation chart *Fig. (g)*. Histogram indicates distribution of all 4 variables, g4_var2 is normally distributed distribution. Below the diagonal all scatter plot is shown which gives the distribution of all 4 variables. Above the diagonal we can see the significant level of each variable and correlation coefficient of each variable. Check the table (b) for relationships for all 4 variables from group4 and group6.

5. Future Work:

Bootstrap Aggregation (Bagging) ensemble learning and implement a Random Forest Bagging model using a sklearn library. Boosting also improve the prediction power by training a sequence of weak models. The large standard and normalised data set can be used to train the model. For visualization web application can be developed using R shiny or Django for easy interaction with the user.

6. Conclusion:

We have successfully performed exploratory data analysis, data wrangling. Applied 3 different modelling techniques. Because of 15 minutes lag in data, applied random forest for each stage to increase accuracy. Evaluate the predicted values with the given optimal value -0.8574. Observed relationship between target variable and 3 variables from group 6. And finally checked the future scope and enhancement for proposed model.