QUEEN'S UNIVERSITY BELFAST

DSA8023 – ANALYTICS IN ACTION

ANALYTATHON 1

PREDICTIVE ANALYTICS IN SOFT DRINK PRODUCTION

Safwan Sukeri

40307894

Msukeri01@qub.ac.uk

1. Introduction

Based on the task description and initial consultation, the assignment was to build predictive analytics to predict the quality of product at several stages of the process.

2. Data Preparation and Exploratory Data Analysis (EDA)

Initial step towards building a Predictive Model is the data preparation. A set of 12 preceding variables were provided to predict the targeted variable g4_var_2. EDA was used to filter the preceding variables (features) and the result are presented as follows:

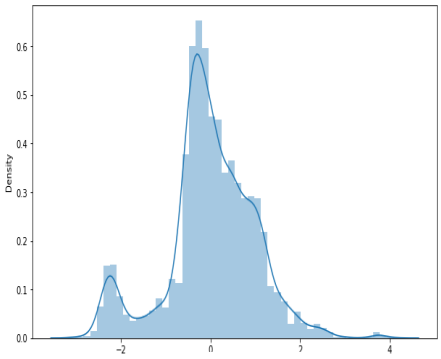

Figure 2.1

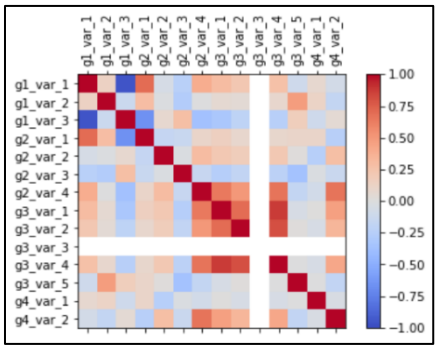

Figure 2.2

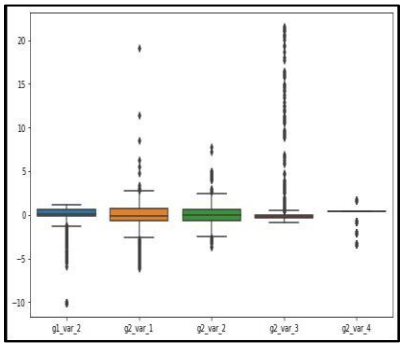

Figure 2.3

Figure 2.1 showed a distribution plot of the targeted variable g4_var_2. Due to the presence of missing values in the dataset, the symmetry shape distribution will imply that the missing value is best replaced with the nearest value to the period of missing data (as this is time series data). Next, g1_var_1 and g1_var_3 will be exempted from the analysis due to weak correlation with the targeted variable g4_var_2 which can visualized based on Figure 2.2. In addition, it was discovered that g3_var_3 had only a single value. Hence it will be dropped as well.

Figure 2.3 showcased the 5 out of 9 boxplots of the remaining features. Some of the key findings were, data have been scaled and normalised, outliers were presence in most of the features, g2_var_4 had only 5 values which might implied a categorical variable.

3. Methodology

Based on the findings from EDA, the method chose for this analysis will be Regression. The main justifications are explained as follows:

1. Regression models are among the best statistical models for studying relationships between independent and dependent variable for continuous data. This can help the manufacturing company identify key aspect affecting the targeted variable and quality
2. The presence of outliers can be overcome by using Random Forest Regressor Method
3. A good starting point to understand the landscape before initiating with more complex model such as LSTM or Deep learning which require proper data architecture.

Hence, the two model that will be used for this predictive analysis will be

- Linear Regression Method
- Random Forest Regressor

4. Results

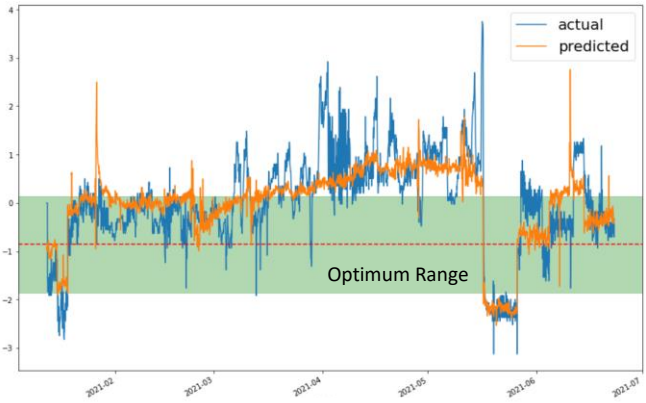

Figure 4.1 Linear Regression

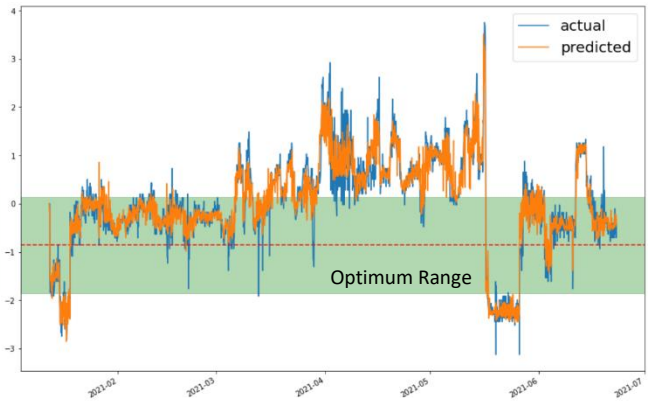

Figure 4.2 Random Forest Regressor

Based on figure 4.1 and 4.2, the Random Forest Regressor tend to perform better compared to Linear Regression. The R^2 Value are 61% and 89% respectively. The green region is the Optimum Range based on the given optimum value of -0.8572 ± 1 which is the standard deviation of the variable. The purpose of this Optimum Range is to determine whether the variable g4_var_2 tend to fall within optimum value or otherwise. This then relates to the objective of the predictive analysis which to predict for any spike (non-optimum) to happen beforehand. This then allows the company to anticipate any anomalies in the future. Thus, by using the train data set of which consist 20% of total data, the accuracy to predict such measures for both of the model are 87% and 95% respectively.

There has also been a finding about the preceding variable that is significantly affecting the value of the targeted variable. The variable g2_var_4 had a co-efficient of 0.682 for the Linear Regression Model and 0.425 for the Random Forest Regressor Model which is the highest among the variables in the model. This can help to anticipate the value of g4_var_2 if a significant change in the variable g2_var_4 is detected.

The final part of this task is to study the relationship between the targeted variable g4_var_2 against 3 quality variables, g6_var_2, g6_var_3 and g6_var_4. Based on the result of the correlation analysis between the respective variables, it was found that g4_var_2 has a moderate negative relationship with g6_var_2 of -0.33. Thus, it can be deduced that in the situation that model predicts high value of g4_var_2, it can be anticipated that the quality of g6_var_2 might slightly decay.

5. Conclusion

Predictive analysis can play a significant role in industry. With a proper solution architecture, the future is certainly there for the manufacturing industry to overcome the current limitation.