Gautam Kumar

Breakdown classification at reflow ovens

# **Business Goal of the Project**

Analysis of data and building a classifier for predicting failure of reflow ovens in manufacturing line of mother boards

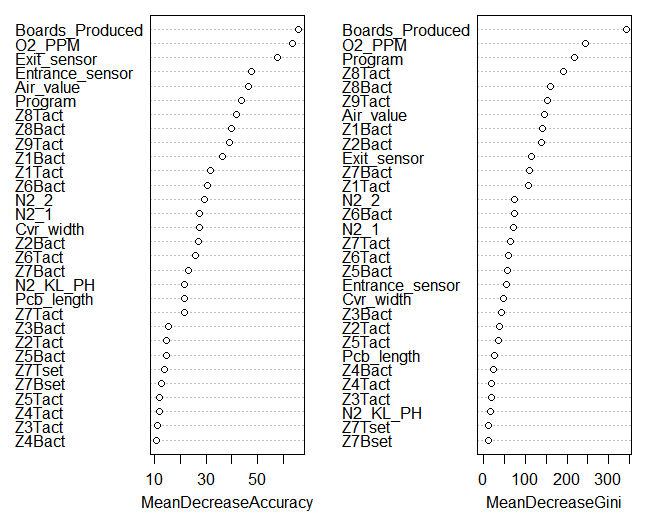
# **Introduction**

1. There are 132840 observations in oven 1 data, where red light=1 but numPcbs == 0, I am considering these observations as normal operation without any failure.
2. Valid Failure percentage is very less 0.08%. I am using Under Sampling over sampling and SMOTE to create test and train set for this skewed problem.

# **Over Sampling:**

## **Random Forest in case of over sampling**

### **Importance of variables plot for random forest using all explanatory variables**



|  |
| --- |
|  |
|  |
| |  | | --- | | Plot of variable importance using random forest by keeping mtry=8. We are seeing an elbow near “Pcb\_length”. For classification problems we see the Decrease in the Gini. So we will consider all the variables above and including Z6Tact in these plots as statistically important for our Modelling.  Variables chosen in this case.  **{ Boards\_Produced, O2\_PPM, Exit\_sensor, Air\_Value, Z8\_Tact, Program, Z9Tact, Z8Bact, Z1Bact,Air\_value, Z2\_Bact,**  **Z6Bact, Z7Bact, Z6Tact, N2\_2,Z7Tact,Z6Tact,Z5Bact,Entrace\_sensor,Cvr\_width,Z3Bact,Z2Tact,Z5Tact,Pcb\_length}** | |

### **Confusion Matrix after creating this model using Random Forest.**

Confusion Matrix and Statistics

Reference

Prediction FALSE TRUE

FALSE 1795 9

TRUE 103 1786

Accuracy : 0.9697

95% CI : (0.9636, 0.975)

No Information Rate : 0.5139

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.9394

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9457

Specificity : 0.9950

Pos Pred Value : 0.9950

Neg Pred Value : 0.9455

Prevalence : 0.5139

Detection Rate : 0.4861

Detection Prevalence : 0.4885

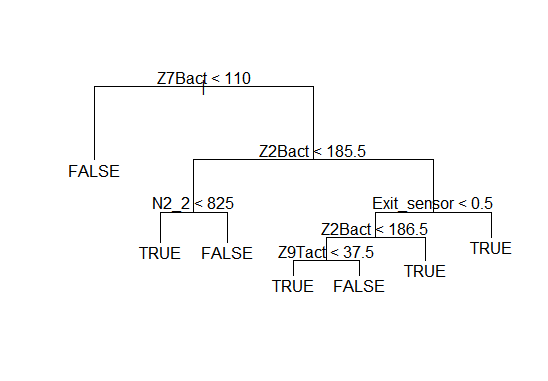
Balanced Accuracy : 0.9704

'Positive' Class : FALSE

### **Confusion Matrix after creating this model using SVM.**

|  |
| --- |
| Confusion Matrix and Statistics  Reference  Prediction FALSE TRUE  FALSE 1269 491  TRUE 629 1304    Accuracy : 0.6967  95% CI : (0.6816, 0.7115)  No Information Rate : 0.5139  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.3942  Mcnemar's Test P-Value : 4.246e-05    Sensitivity : 0.6686  Specificity : 0.7265  Pos Pred Value : 0.7210  Neg Pred Value : 0.6746  Prevalence : 0.5139  Detection Rate : 0.3436  Detection Prevalence : 0.4766  Balanced Accuracy : 0.6975    'Positive' Class : FALSE |
|  |

### **Decision Tree plot and Confusion Matrix with Decision Tree with Over Sample.**



### **Confusion Matrix and Statistics using Decision Tree.**

Reference

Prediction FALSE TRUE

FALSE 1656 858

TRUE 242 937

Accuracy : 0.7021

95% CI : (0.6871, 0.7169)

No Information Rate : 0.5139

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.3982

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.8725

Specificity : 0.5220

Pos Pred Value : 0.6587

Neg Pred Value : 0.7947

Prevalence : 0.5139

Detection Rate : 0.4484

Detection Prevalence : 0.6807

Balanced Accuracy : 0.6973

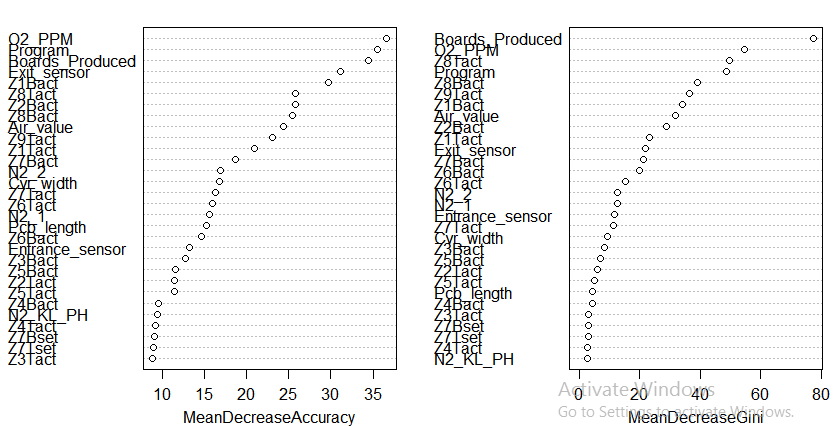
'Positive' Class : FALSE

### **Logistic Regression using over sampling**

Logistic regression is not converging. And we are getting huge coefficients which do not look normal for modelling.

# **Under Sampling.**

## **Variable Importance plot using random forest.**



Plot of variable importance using random forest by keeping mtry=8. We are seeing an elbow near, Pcb Length. For classification problems we see the Decrease in the Gini. So we will consider all the variables above and including Z6Tact in these plots as statistically important for our Modelling.

Variables chosen in this case.

**{Boards\_Produced, O2\_PPM, Exit\_sensor, Air\_Value, Z8\_Tact, Program, Z9Tact, Z8Bact, Z1Bact, Air\_Value, Z2\_Bact, Z6Bact, Z7Bact, Z6Tact, N2\_2,Z7Tact,Z6Tact,Z5Bact,Entrace\_sensor,Cvr\_width,Z3Bact,Z2Tact,Z5Tact,Pcb\_length**

**}**

### **Confusion Matrix after creating this model Random Forest under Sampling.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Confusion Matrix and Statistics  Reference  Prediction FALSE TRUE  FALSE 301 70  TRUE 62 260    Accuracy : 0.8095  95% CI : (0.7783, 0.8381)  No Information Rate : 0.5238  P-Value [Acc > NIR] : <2e-16    Kappa : 0.6178  Mcnemar's Test P-Value : 0.5423    Sensitivity : 0.8292  Specificity : 0.7879  Pos Pred Value : 0.8113  Neg Pred Value : 0.8075  Prevalence : 0.5238  Detection Rate : 0.4343  Detection Prevalence : 0.5354  Balanced Accuracy : 0.8085    'Positive' Class : FALSE      **Confusion Matrix after creating this model Random SVM under Sampling.**  Confusion Matrix and Statistics  Reference  Prediction FALSE TRUE  FALSE 239 84  TRUE 124 246    Accuracy : 0.6999  95% CI : (0.6642, 0.7338)  No Information Rate : 0.5238  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.4016  Mcnemar's Test P-Value : 0.006848    Sensitivity : 0.6584  Specificity : 0.7455  Pos Pred Value : 0.7399  Neg Pred Value : 0.6649  Prevalence : 0.5238  Detection Rate : 0.3449  Detection Prevalence : 0.4661  Balanced Accuracy : 0.7019    'Positive' Class : FALSE  **Confusion Matrix after creating this model Decision Tree using Under Sampling**  Confusion Matrix and Statistics  Reference  Prediction FALSE TRUE  FALSE 294 38  TRUE 61 300    Accuracy : 0.8571  95% CI : (0.8288, 0.8824)  No Information Rate : 0.5123  P-Value [Acc > NIR] : < 2e-16    Kappa : 0.7146  Mcnemar's Test P-Value : 0.02703    Sensitivity : 0.8282  Specificity : 0.8876  Pos Pred Value : 0.8855  Neg Pred Value : 0.8310  Prevalence : 0.5123  **Plot of Decision Tree with under sampling.**      Detection Rate : 0.4242  Detection Prevalence : 0.4791  Balanced Accuracy : 0.8579    'Positive' Class : FALSE **Confusion Matrix and Logistic regression Model for under sampling.**   |  | | --- | | Confusion Matrix and Statistics  Reference  Prediction FALSE TRUE  FALSE 350 188  TRUE 17 138    Accuracy : 0.7042  95% CI : (0.6687, 0.738)  No Information Rate : 0.5296  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.3884  Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.9537  Specificity : 0.4233  Pos Pred Value : 0.6506  Neg Pred Value : 0.8903  Prevalence : 0.5296  Detection Rate : 0.5051  Detection Prevalence : 0.7763  Balanced Accuracy : 0.6885    'Positive' Class : FALSE | | **Logistic Regression Model Summary.**  |  | | --- | | Call:  glm(formula = valid\_failure ~ Boards\_Produced + O2\_PPM + Exit\_sensor +  Air\_value + Z8Tact + Program + Z9Tact + Z8Bact + Z1Bact +  Z2Bact + Z6Bact + Z7Bact + Z6Tact + N2\_2 + Z7Tact + Z6Tact +  Z5Bact + Entrance\_sensor + Cvr\_width + Z3Bact + Z2Tact +  Z5Tact + Pcb\_length, family = binomial, data = train\_oven\_1\_df)  Deviance Residuals:  Min 1Q Median 3Q Max  -2.1382 -1.0126 0.0000 0.9899 3.1123  Coefficients: (1 not defined because of singularities)  Estimate Std. Error z value Pr(>|z|)  (Intercept) -3.037e+01 1.125e+04 -0.003 0.99785  Boards\_Produced -7.542e-04 1.431e-04 -5.271 1.36e-07 \*\*\*  O2\_PPM 3.652e-04 1.440e-04 2.536 0.01121 \*  Exit\_sensor 1.267e+00 2.071e-01 6.119 9.43e-10 \*\*\*  Air\_value 3.664e-05 1.041e-03 0.035 0.97193  Z8Tact -6.127e-02 2.272e-02 -2.697 0.00700 \*\*  ProgramENG158R1 1.082e+01 5.338e+02 0.020 0.98382  ProgramENG280R1 -1.168e+04 4.170e+05 -0.028 0.97765  ProgramENG316R1 -1.127e+04 4.017e+05 -0.028 0.97761  ProgramENG329R1 -4.157e+00 1.063e+00 -3.912 9.15e-05 \*\*\*  ProgramENG356R1 -8.312e+03 2.958e+05 -0.028 0.97758  ProgramENG513R1 -8.308e+03 2.960e+05 -0.028 0.97761  ProgramENG592R1 -8.293e+03 2.956e+05 -0.028 0.97762  ProgramENG611R1 8.843e+00 5.338e+02 0.017 0.98678  ProgramENG846R1 -1.108e-01 1.432e+00 -0.077 0.93830  ProgramENG849R1 -4.702e+00 1.031e+00 -4.560 5.11e-06 \*\*\*  ProgramENG857R1 1.081e+01 5.338e+02 0.020 0.98384  ProgramMAINT-154-7\_DAYCHECKP\004 -3.296e+03 1.161e+05 -0.028 0.97736  ProgramPROD\_CLEAN\_MAX\_WIDTHR\004 1.121e+04 4.013e+05 0.028 0.97773  Z9Tact 3.118e-02 2.759e-02 1.130 0.25843  Z8Bact NA NA NA NA  Z1Bact -2.323e-03 2.056e-03 -1.130 0.25868  Z2Bact -1.934e-01 2.136e-01 -0.906 0.36509  Z6Bact -3.114e-02 1.489e-01 -0.209 0.83434  Z7Bact 1.370e+00 5.213e-01 2.628 0.00859 \*\*  Z6Tact 1.974e-01 1.858e-01 1.063 0.28794  N2\_2 7.760e-04 1.063e-03 0.730 0.46555  Z7Tact 6.331e-01 4.719e-01 1.341 0.17976  Z5Bact -3.725e+00 1.531e+00 -2.432 0.01500 \*  Entrance\_sensor -1.681e-01 1.108e-01 -1.517 0.12917  Cvr\_width -7.472e+00 2.669e+02 -0.028 0.97767  Z3Bact -1.283e+00 5.051e-01 -2.541 0.01105 \*  Z2Tact -1.466e+00 6.016e-01 -2.436 0.01485 \*  Z5Tact 3.836e+00 1.381e+00 2.778 0.00547 \*\*  Pcb\_length 7.399e+01 2.639e+03 0.028 0.97763  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 2243.0 on 1618 degrees of freedom  Residual deviance: 1684.6 on 1585 degrees of freedom  AIC: 1752.6  Number of Fisher Scoring iterations: 20 | |  | | |  | | --- | |  | | | |  | |
|  |
| **SMOTE** Variable importance using Random Forest when sampling is done using SMOTE.     |  | | --- | |  | |

Plot of variable importance using random forest by keeping mtry=8. We are seeing an elbow near, Pcb\_lenght. For classification problems we see the Decrease in the Gini. So we will consider all the variables above and including Z6Tact in these plots as statistically important for our Modelling.

Variables chosen in this case.

**{Boards\_Produced, O2\_PPM, Exit\_sensor, Air\_Value, Z8\_Tact, Program, Z9Tact, Z8Bact, Z1Bact, Z2\_Bact, Z6Bact, Z7Bact, Z6Tact, N2\_2,N2\_1,Z7Tact,Z6Bact,Z5Bact,Entrace\_sensor,Cvr\_width,Z3Bact,Z2Tact,Z5Tact,Pcb\_length,Z1Tact**

**}**

### **Confusion Matrix using SMOTE and Random Forest Model.**

Confusion Matrix and Statistics

Reference

Prediction FALSE TRUE

FALSE 302 53

TRUE 63 275

Accuracy : 0.8326

95% CI : (0.8027, 0.8597)

No Information Rate : 0.5267

P-Value [Acc > NIR] : <2e-16

Kappa : 0.6648

Mcnemar's Test P-Value : 0.4034

Sensitivity : 0.8274

Specificity : 0.8384

Pos Pred Value : 0.8507

Neg Pred Value : 0.8136

Prevalence : 0.5267

Detection Rate : 0.4358

Detection Prevalence : 0.5123

Balanced Accuracy : 0.8329

'Positive' Class : FALSE

### **Confusion Matrix using SMOTE and SVM Model**

Confusion Matrix and Statistics

Reference

Prediction FALSE TRUE

FALSE 242 91

TRUE 123 237

Accuracy : 0.6912

95% CI : (0.6553, 0.7254)

No Information Rate : 0.5267

P-Value [Acc > NIR] : < 2e-16

Kappa : 0.3837

Mcnemar's Test P-Value : 0.03408

Sensitivity : 0.6630

Specificity : 0.7226

Pos Pred Value : 0.7267

Neg Pred Value : 0.6583

Prevalence : 0.5267

Detection Rate : 0.3492

Detection Prevalence : 0.4805

Balanced Accuracy : 0.6928

'Positive' Class : FALSE

**Confusion Matrix with SMOTE and Decision Tree Model.**

Confusion Matrix and Statistics

Reference

Prediction FALSE TRUE

FALSE 353 217

TRUE 12 111

Accuracy : 0.6696

95% CI : (0.6331, 0.7045)

No Information Rate : 0.5267

P-Value [Acc > NIR] : 1.664e-14

Kappa : 0.3155

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9671

Specificity : 0.3384

Pos Pred Value : 0.6193

Neg Pred Value : 0.9024

Prevalence : 0.5267

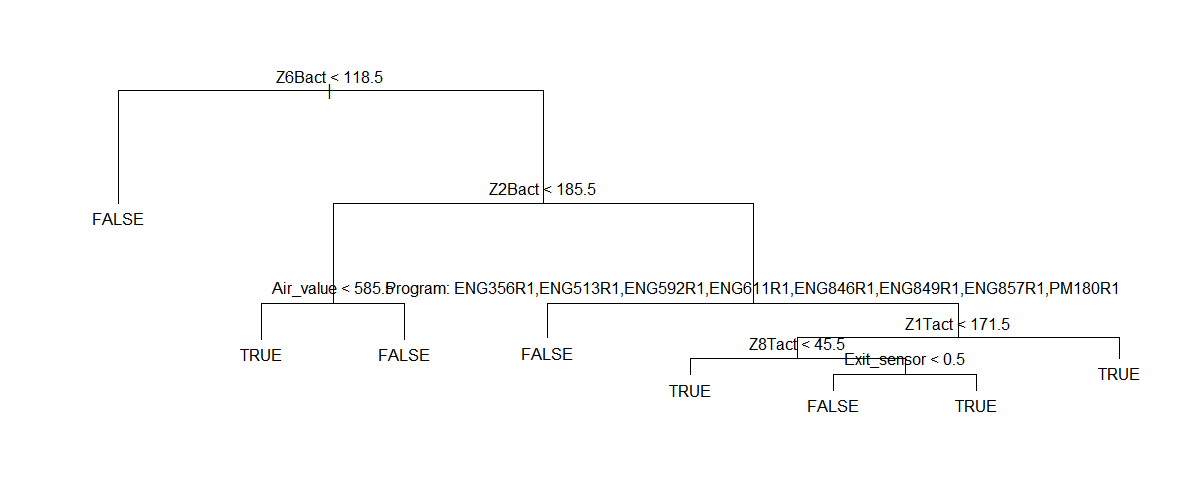
Detection Rate : 0.5094

Detection Prevalence : 0.8225

Balanced Accuracy : 0.6528

'Positive' Class : FALSE

**Plot of decision tree with SMOTE sampling.**



**Confusion matrix with SMOTE and Logistic Regression model**

Confusion Matrix.

|  |
| --- |
| Confusion Matrix and Statistics  Reference  Prediction FALSE TRUE  FALSE 341 190  TRUE 24 138    Accuracy : 0.6912  95% CI : (0.6553, 0.7254)  No Information Rate : 0.5267  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.3643  Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.9342  Specificity : 0.4207  Pos Pred Value : 0.6422  Neg Pred Value : 0.8519  Prevalence : 0.5267  Detection Rate : 0.4921  Detection Prevalence : 0.7662  Balanced Accuracy : 0.6775    'Positive' Class : FALSE |
| **Model Summary with SMOTE and Logistic Regression** |
| |  | | --- | |  | |

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9875 -1.0172 0.0000 0.9955 3.3483

Coefficients: (2 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) 6.275e+03 2.077e+04 0.302 0.762574

Boards\_Produced -6.968e-04 1.420e-04 -4.905 9.33e-07 \*\*\*

O2\_PPM 3.530e-04 1.625e-04 2.173 0.029797 \*

Exit\_sensor 1.252e+00 2.124e-01 5.892 3.81e-09 \*\*\*

Air\_value 1.066e-03 1.066e-03 1.000 0.317546

Z8Tact -5.872e-02 2.272e-02 -2.584 0.009771 \*\*

ProgramENG158R1 4.729e+00 2.686e+01 0.176 0.860249

ProgramENG280R1 1.150e+01 7.176e+01 0.160 0.872657

ProgramENG329R1 -2.774e+00 1.219e+00 -2.276 0.022835 \*

ProgramENG356R1 -2.700e+01 7.005e+03 -0.004 0.996924

ProgramENG513R1 -2.239e+01 6.289e+03 -0.004 0.997159

ProgramENG592R1 -5.474e+00 5.086e+01 -0.108 0.914277

ProgramENG611R1 3.798e+00 2.687e+01 0.141 0.887593

ProgramENG846R1 -1.317e+00 1.183e+00 -1.113 0.265501

ProgramENG849R1 -3.931e+00 1.184e+00 -3.319 0.000902 \*\*\*

ProgramENG857R1 4.664e+00 2.686e+01 0.174 0.862164

ProgramMAINT-154-7\_DAYCHECKP\004 -3.853e+01 8.178e+02 -0.047 0.962421

ProgramPM180R1 -2.461e+01 1.773e+04 -0.001 0.998892

ProgramPROD\_CLEAN\_MAX\_WIDTHR\004 3.849e+03 1.278e+04 0.301 0.763164

Z9Tact 3.782e-02 2.710e-02 1.396 0.162841

Z8Bact NA NA NA NA

Z1Bact -1.564e-02 5.401e-02 -0.290 0.772173

Z2Bact -4.644e-01 2.778e-01 -1.672 0.094599 .

Z6Bact -5.121e-02 1.629e-01 -0.314 0.753192

Z7Bact 1.266e+00 4.818e-01 2.628 0.008580 \*\*

N2\_2 2.322e-03 1.126e-03 2.062 0.039174 \*

N2\_1 NA NA NA NA

Z7Tact 3.022e-01 4.632e-01 0.652 0.514192

Z6Tact 4.556e-01 1.822e-01 2.501 0.012382 \*

Z5Bact -4.449e+00 1.756e+00 -2.534 0.011283 \*

Entrance\_sensor -1.146e-01 1.069e-01 -1.072 0.283765

Cvr\_width -4.081e+00 1.346e+01 -0.303 0.761638

Z3Bact -6.266e-01 6.445e-01 -0.972 0.330914

Z2Tact -9.479e-01 7.271e-01 -1.304 0.192335

Z5Tact 3.261e+00 1.571e+00 2.075 0.037989 \*

Pcb\_length 1.558e-03 4.539e-01 0.003 0.997261

Z1Tact 8.632e-01 1.978e-01 4.365 1.27e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2243.2 on 1618 degrees of freedom

Residual deviance: 1666.7 on 1584 degrees of freedom

AIC: 1736.7

Number of Fisher Scoring iterations: 19

Appendix: