Regression Model on Dummy data

Nikhil Muthukrishnan

2 November 2018

Proof of Concept for Feature Selection and Regression Modelling

Objective: Show effectiveness of Regression Models on dummy Machine Data

Sprint: PO_SPRINT_12_OCT_2018

Contents: An excel file of 3 batches,3 sensors and 2 dependent classes viz: ABD pass or fail.

```
##   BatchNumber     GlobalTimeStamp s1 s2 s3 abd
## 1           1 2018-11-02 18:02:15  7 81  3   1
## 2           1 2018-11-02 18:03:15  1 40  2   1
## 3           1 2018-11-02 18:04:15 97 68 96   1
## 4           1 2018-11-02 18:05:15 74 56 29   1
## 5           1 2018-11-02 18:06:15  5 43 31   1
## 6           1 2018-11-02 18:07:15 53 12 55   1
```

The data seen above contains 5 coloumns Batch Number limited to 3 batches of 72 hours each GlobalTimeStamp ranges from 02-11-2018 18:02 to 11-11-2018 18:01 s1/2/3 are sensors 1/2/3

The data has been rigged to show sensor 1 as the cause for failure This can be seen by comparing the central limits of sensor 1 where is 0 and is 1.

```
##        s1               s2               s3              abd
##  Min.   :  1.00   Min.   :  1.00   Min.   :  1.00   Min.   :1
##  1st Qu.: 26.00   1st Qu.: 25.00   1st Qu.: 25.00   1st Qu.:1
##  Median : 51.00   Median : 51.00   Median : 50.00   Median :1
##  Mean   : 50.71   Mean   : 50.33   Mean   : 50.46   Mean   :1
##  3rd Qu.: 76.00   3rd Qu.: 76.00   3rd Qu.: 76.00   3rd Qu.:1
##  Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :1
```

```
##        s1               s2               s3            abd
##  Min.   : 90.00   Min.   :  1.00   Min.   :  1.0   Min.   :0
##  1st Qu.: 92.00   1st Qu.: 26.00   1st Qu.: 26.0   1st Qu.:0
##  Median : 95.00   Median : 51.00   Median : 50.0   Median :0
##  Mean   : 95.05   Mean   : 50.76   Mean   : 50.6   Mean   :0
##  3rd Qu.: 98.00   3rd Qu.: 76.00   3rd Qu.: 75.0   3rd Qu.:0
##  Max.   :100.00   Max.   :100.00   Max.   :100.0   Max.   :0
```

From the summary of the data we can see that mean and median of s1 for ABD pass and ABD fail are very DIFFERENT. This is intended to show the ability of the following models(2) to capture the effect of s1 on batch quality.

Model 1

The data has been made to simulate real data as much as possible and shows the format the data needs to be in to apply statistical techniques. The first model is a simple linear regression model with no specification. The coeffiecents of the model is below

```
##
## Call:
## lm(formula = abd ~ s1 + s2 + s3, data = dam)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42784 -0.34637 -0.05779  0.30704  0.67743
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.318e+00  1.046e-02  125.948   <2e-16 ***
## s1          -9.883e-03  9.832e-05 -100.517   <2e-16 ***
## s2          -1.609e-05  1.072e-04   -0.150    0.881
## s3          -6.039e-05  1.072e-04   -0.563    0.573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3534 on 12956 degrees of freedom
## Multiple R-squared:  0.4382, Adjusted R-squared:  0.4381
## F-statistic:  3368 on 3 and 12956 DF,  p-value: < 2.2e-16
```

As expected the linear model returns the highest coeffiecent for sensor 1 and lesser for sensor 2 and sensor 3. However the coeffiecents of sensor 3 are too close to sensor 2. Thus this model is inconclusive and a more sophisticated model will have to be used.

Model 2

The second model that will be fit on the simulated batch data is a logistic regression model. The details of the model are out of scope in this document and the conclusions will be made by comparing coeffiecents between model 1 and 2.

```
##
## Call:  glm(formula = abd ~ s1 + s2 + s3, family = "binomial", data = dam)
##
## Coefficients:
## (Intercept)           s1           s2           s3
##  23.8203376   -0.2657278    0.0001135    0.0003971
##
## Degrees of Freedom: 12959 Total (i.e. Null);  12956 Residual
## Null Deviance:       16500
## Residual Deviance: 6168  AIC: 6176
```

Here we see absolute numbers of the coeffiecents are lower than model 1 however difference across coeffiecents is much larger. Hence glm is slightly better than lm in identifying important features. However even this model does not take into consideration that the data is actually in a panel format meaning the cause of ABD failure are across several rows not just the where the dependent variable is 0.

Such a dataset requires pooled linear models which have several specifications that can be made to capture the effect of sensor readings on ABD failure. Pooled Linear Models will be described in detail if the results are more conclusive than glm.