# Analysis of Glass Identification Data

**Final Project Report - Surya Suresh(V00998151)**

## Introduction

Multiclass classification problems are those where a label must be predicted, but there are more than two labels that may be predicted. These are challenging predictive modeling problems because a sufficiently representative number of examples of each class is required for a model to learn the problem. Problems of this type are referred to as imbalanced multiclass classification. The main purpose of the classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence, if correctly identified.

## Objective

The glass identification dataset is a standard dataset for exploring the challenge of imbalanced multiclass classification. The main objective of this project is to fit a regression model for the glass identification dataset to correctly classify the types of glasses.

## Dataset Description

The glass identification dataset was obtained from the UCI machine learning repository and credited to Vina Spiehler in 1987. The dataset describes the chemical properties of glass and involves classifying samples of glass using their chemical properties. The dataset consists of 214 instances and 11, all attributes are continuously valued.
Below listed 10 attributes are the input variables that summarize the properties of the glass dataset.

1. Id number – Describes the 1-214 instances
2. RI – Refractive index of the glass
3. Na – Sodium
4. Mg – Magnesium
5. Al – Aluminum
6. Si – Silicon
7. K – Potassium
8. Ca – Calcium
9. Ba – Barium
10. Fe – Iron
11. Type of glass – Class attribute

Among all these variables, Types of glass are the **response variable**, and the other 10 variables are predictors. Based on their oxide content the predicted variable can be classified into 7 classes

- **Class 1**: Building windows (float processed)
- **Class 2**: Building windows (non-float processed)
- **Class 3**: Vehicle windows (float processed)
- **Class 4**: Vehicle windows (non-float processed)
- **Class 5**: Containers
- **Class 6**: Tableware
- **Class 7**: Headlamps

Here the data is imbalanced, and I observed that the response variable with class 4(non-float processed) has 0 examples. Although there are minority classes, all classes are equally important in this prediction problem. Here float glass refers to the process used to make glass.

The dataset can be divided into window glass (classes 1-4) and non-window glass (classes 5-7). There are 163 examples of window glass and 51 examples of non-window glass

- **Window Glass**: 163 examples
- **Non-Window Glass**: 51 examples

Another division of the observations would be between float processed glass and non-float processed glass.

- **Float Glass**: 87 examples
- **Non-Float Glass**: 76 examples

However, I found that in the case of window glass only the division is more balanced.

In this project, I would like to consider type of glass as windowglass which has class range from 1-4 & non-window glass which has class range from 5-7and fit a logistic regression model; and then predict the most important variables in the model. Logistic regression model is chosen because it is simple to implement, understand, and can be easily extended to multi-class classification.

The summary of the response variable (glasstype) in the dataset is as shown below:

```
> summary(mydata$glasstype)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    2.00    2.78    3.00    7.00
> table(mydata$glasstype)

 1  2  3  5  6  7
70 76 17 13  9 29
```

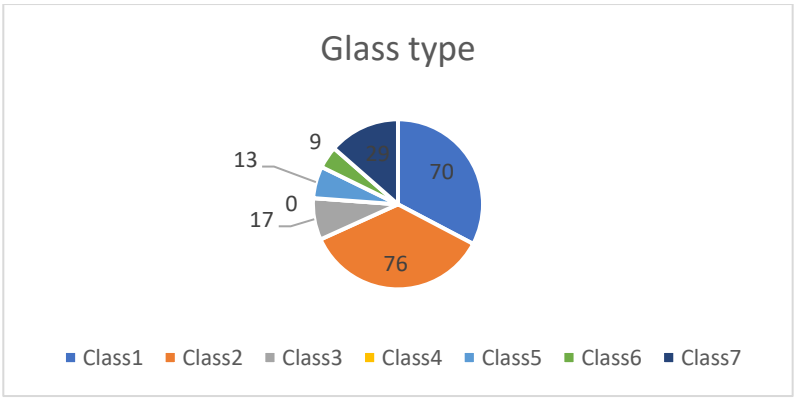
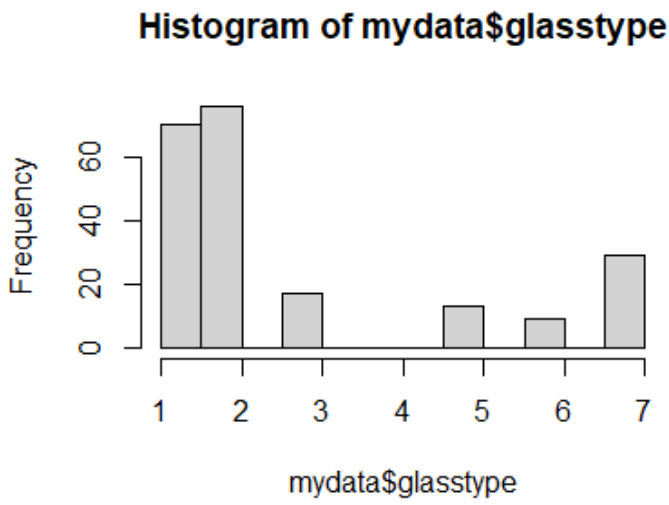



**Figure 1 : Classification of Data**




Here from the above histogram, we can see class 1-3 have more values, so we will be fitting the regression model on those classes under windowglass and classes from 5-7 refer in non-window glass.

## Empirical Analysis of Data (Part 1):

The main objective of this analysis is to find out what all chemical properties ( or variables) that used to classify the different type of glass

As we discussed above, we have 10 predictors and 1 response variable (Type of glass). Let us fit the linear regression model for the overall data,

Type of glass = intercept +b1(id)+ b2(RI)+ b3(Na)+ b4(Mg)+ b5(Al)+ b6(Si)+ b7(K)+ b8(Ca)+ b9(Ba)+ b10(Fe)

**Summary(Glass types) :**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.666e+02  8.228e+01  -3.241 0.001393 **
id           1.905e-02  1.072e-03  17.776  < 2e-16 ***
RI           9.290e+01  4.720e+01   1.968 0.050411 .
Na           1.349e+00  4.478e-01   3.012 0.002923 **
Mg           6.169e-01  4.677e-01   1.319 0.188630
Al           1.665e+00  4.719e-01   3.528 0.000517 ***
Si           1.331e+00  4.589e-01   2.901 0.004125 **
K            9.842e-01  4.719e-01   2.085 0.038275 *
Ca           7.587e-01  4.760e-01   1.594 0.112503
Ba           1.220e+00  4.786e-01   2.550 0.011506 *
Fe          -9.648e-01  5.267e-01  -1.832 0.068463 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6785 on 203 degrees of freedom
Multiple R-squared:  0.9009,    Adjusted R-squared:  0.896
F-statistic: 184.4 on 10 and 203 DF,  p-value: < 2.2e-16
```

From the above data, we can clearly see that there are only 2 important significant predictors

1. Id
2. Al

Considering only the important predictors we have linear regression model fit as

Glasstypes =(1.905)* Id +(1.665) *Al

Now let us focus on prediction the Class1- 7 glass types in 2 different categories.

- mydata$windowglass <- mydata$glasstype <=4
- mydata$nonwindowglass <- mydata$glasstype >=5

Using the same predictors, we have logistic regression model fit for windowglass type glasses as:

```
Call:
glm(formula = windowglass ~ id + Al, family = binomial(link = "logit"),
    data = mydata)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.335e-04   2.000e-08   2.000e-08   2.000e-08   3.250e-04

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1127.564 135919.112   0.008    0.993
id             -6.543    751.454  -0.009    0.993
Al            -37.233  65755.175  -0.001    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.3503e+02  on 213  degrees of freedom
Residual deviance: 2.2247e-07  on 211  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25
```

Windowglass = 1127.564+ (-6.543)*id +(-37.233) *AI

## Empirical Evaluation

The evaluation Metrics that I used for the classification task are:

- **Sensitivity** = TP/(TP + FN) = (Number of true positive assessment)/(Number of all positive assessment)
- **Specificity** = TN/(TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)
- **Accuracy** = (TN + TP)/(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)
- **Precision** = (TP) / (TP + FP) = quantifies the number of positive class predictions that belongs to the positive class.
- **Recall** = TP/(TP +FN) = quantifies the number of positive class predictions made from all positive examples in the dataset

## 10-Fold Cross Validation:

We use 10 – fold Cross validate to empirically evaluate the model.

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. Here the data is randomly partitioned into 10 equal sized sub-samples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. This will be repeated 10 times till test data covers all the subsamples.

### CONFUSION MATRIX

The output that we receive after using 10-fold cross validation is as follows:

```
> totalConfusion
       myrpartPredict
        FALSE TRUE
  FALSE   162    1
  TRUE      0   51
```

| | Predicted False | Predicted True |
|---|---|---|
| Actual False | 162 (TN) | 1(FN) |
| Actual True | 0 (FP) | 51(TP) |

**Table 1: Confusion Matrix of whole data**

From the above table 1, we can observe that the data has more negatives compared to positives. This is due the imbalanced dataset. In multiclass classifications on imbalanced datasets, accuracy is not a valid measure and I tried to calculate precision, recall measures as well.
Evaluation results for the Glass identification data

Sensitivity = TP/(TP + FN= 51/(51+1) = 98%

Specificity = TN/(TN + FP) = 162/(162+0) = 100%

Accuracy = (TN + TP)/(TN+TP+FN+FP) = (162+51)/(162+51+1+0) = 99.53%

Precision = (TP) / (TP + FP) = (51)/(51+0) =100%

Recall = TP / (TP + FN) = 51/(51+1) =98%

| Metric | Value |
|---|---|
| Sensitivity | 98% |
| Specificity | 100% |
| Accuracy | 99.53% |
| Precision | 100% |
| Recall | 98% |

**Table 2: Evaluation result of whole data**

From the above results, we can say that the model tried to classify the type of glasses most accurately. Compared to Specificity sensitivity values are less because, the whole dataset has very few positive values .

# Part 2: PCA (Principal Component Analysis)

## PCA:

PCA, is also called as dimensionality reduction1 method is often used to reduce the features of large data sets, by transforming a large set of variables into a smaller set of variables that contains most of the information in the large set.

In our data, the response variable ‘windowglass’ is removed and PCA is done on the 10 variables.

```
> summary(mypca)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     61.9571 1.60495 1.31336 0.82660 0.79683 0.54418 0.44075 0.31379 0.09273
Proportion of Variance  0.9984 0.00067 0.00045 0.00018 0.00017 0.00008 0.00005 0.00003 0.00000
Cumulative Proportion   0.9984 0.99905 0.99950 0.99968 0.99984 0.99992 0.99997 1.00000 1.00000
                          PC10      PC11
Standard deviation     0.03775 0.0009758
Proportion of Variance 0.00000 0.0000000
Cumulative Proportion  1.00000 1.0000000
```

From the above results, we can see that PC1 alone has variance of 99.84% of the total variance. Here we can clearly see that PC1 can explain 99.84% of the variances of all the predictors. PC2-PC5 are significantly low in explaining the predictors.

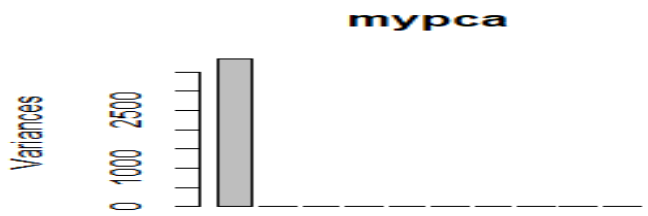



**Figure 2: The variance level among the PCs**

Now the first 5 PCs are taken into consideration, and model is fit against 5 PCs and response variable 'Windowglass'. This is how my new data set with 5 PCs and response variable looks like with 214 observations:

```
> mydfpca
    myresponse          PC1          PC2          PC3          PC4          PC5
1        FALSE -106.5197307 -0.108600255  0.983764145  1.333378006  0.2067157443
2        FALSE -105.5057471 -0.637196912  1.919685870  0.448837229  0.0104872717
3        FALSE -104.5065582 -0.630590048  1.907631350  0.050476161 -0.1082910907
4        FALSE -103.5112080 -0.265135114  1.509522681  0.110506070  0.2073411713
5        FALSE -102.5105891 -0.399371042  1.645447893 -0.143718524 -0.1365641155
6        FALSE -101.5120138 -0.360388519  1.526518948 -0.487944773  0.1775750541
7        FALSE -100.5114914 -0.321492157  1.565525816 -0.124825182 -0.1649465814
8        FALSE  -99.5130690 -0.255526385  1.487848939 -0.286024177 -0.2804617184
9        FALSE  -98.5082931 -0.269567559  1.525295622  0.925660142  0.4402459116
10       FALSE  -97.5134825 -0.110987234  1.354688361 -0.270855579 -0.0132407137
11       FALSE  -96.5130444 -0.314374465  1.461008031 -0.714742527  0.0386201343
12       FALSE  -95.5165515  0.010375029  1.163248965 -0.384354597 -0.0215889301
13       FALSE  -94.5136270 -0.369372138  1.455280350 -0.659535719 -0.0736298331
14       FALSE  -93.5161591 -0.124561565  1.243562383 -0.491517224 -0.1884158753
15       FALSE  -92.5179498 -0.021527556  1.128196207 -0.701069974 -0.1796342476
16       FALSE  -91.5173735 -0.119850104  1.178417472 -0.558746378 -0.1913454068
17       FALSE  -90.5203035  0.103547847  0.939503870 -0.505328548 -0.1168263182
18       FALSE  -89.5167792  0.335490207  0.781861784  1.903598533  0.3879964943
19       FALSE  -88.5166453  0.172613184  0.938888311  1.095214061 -0.0068710371
20       FALSE  -87.5171639 -0.128889435  1.071010689 -0.180879388  0.2055190842
21       FALSE  -86.5195162 -0.033531837  0.933774923 -0.289698268  0.1647891026
```

Now calculating the Logistic fit to the model,

Myresponse = Intercept + (b1) * PC1 + (b2) * PC2 + (b3) * PC3 + (b4) * PC4 + (b5) * PC5

```
Call:
glm(formula = myresponse ~ PC1 + PC2 + PC3 + PC4 + PC5, family = binomial(link = "logit"),
    data = mydfpca)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.538e-05  -2.100e-08  -2.100e-08  -2.100e-08   5.607e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -38.0892 66854.1702  -0.001    1.000
PC1             0.9958   504.0409   0.002    0.998
PC2             1.2640 25092.7609   0.000    1.000
PC3            18.2856 51255.3177   0.000    1.000
PC4           -10.8162 43737.9959   0.000    1.000
PC5             6.8071 35512.4462   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.3503e+02  on 213  degrees of freedom
Residual deviance: 8.3524e-09  on 208  degrees of freedom
AIC: 12

Number of Fisher Scoring iterations: 25
```

Here I am unable to see any significant predictors. So, I guess, the PC1 will only explain most of the predictions.

## Empirical evaluation on PCA: Confusion Matrix:

The output for model with PCs that we receive after using 10-fold cross validation is as follows

```
> totalConfusion
       myrpartPredict
        FALSE TRUE
  FALSE   163    0
  TRUE      0   51
```

Myresponse = -2.45 + (0.016) * PC1 + (0.03) * PC2 + (-0.199) * PC3 + (-0.205) * PC4 + (-1.103) * PC5

**Evaluation results for the Glass identification data after PCA**

Sensitivity = TP/(TP + FN= 51/(51+1) = 98%

Specificity = TN/(TN + FP) = 163/(163+0) = 100%

Accuracy = (TN + TP)/(TN+TP+FN+FP) = (163+51)/(163+51+1+0) = 99.53%

Precision = (TP) / (TP + FP) = (51)/(51+0) =100%

Recall = TP / (TP + FN) = 51/(51+1) =98%

| Metric | Value |
|---|---|
| Sensitivity | 98% |
| Specificity | 100% |
| Accuracy | 99.53% |
| Precision | 100% |
| Recall | 98% |

**Table 3: Metrics of data with 5 principal components as variables**

The results of PC reduced data have same accuracy and metric values compared to the metrics of whole data. This is because the proportion of variance in the PC1 will explain almost 99.84% of all predictors. Therefor we can't say that the model is better or worse if it uses 5 principal components.

## Conclusion

In this Project I tried to fit and evaluate a model for the imbalanced multiclass glass identification dataset. And it helps me to understand how to load the data and generate some idea for the model selection. Based on the classification I tried to fit the logistic regression model and evaluated the model with robust test harness which helps the model to predict the class labels with high accuracy.

## References:


1. Jolliffe Ian T. and Cadima Jorge 2016Principal component analysis: a review and recent developmentsPhil. Trans. R. Soc. A.3742015020220150202
2. https://machinelearningmastery.com/imbalanced-multiclass-classification-with-the-glass-identification-dataset/