## Linear Discriminant Analysis:

In order to predict/classify Foundation Type, We would build a LDA model using the training data set. Since LDA does not support inclusion of categorical variables in the model, We will analyze continuous variables as independent variables for this model.

**Assumptions:**

- Normality criteria for LDA has been taken care in our earlier proc **glmselect** model. We have transformed the required variables which we continue to use.

- Homogenous variance-covariance:

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 169.127118 | 950 | 1.0000 |

Since the Chi-Square test fails for homogenous variance, we will address this by using pool covariance.

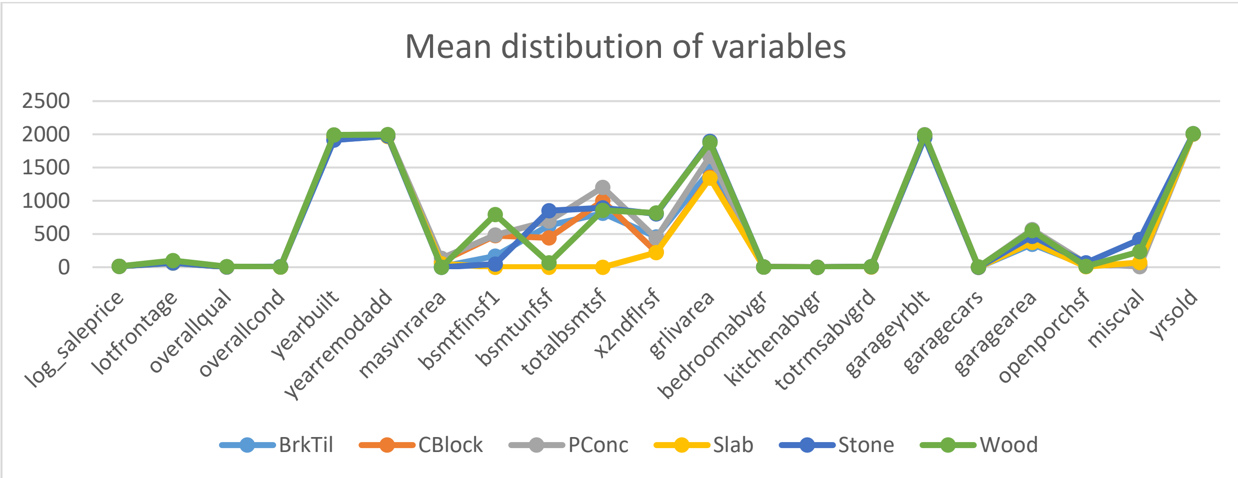
**Analysis:**

In our model, we start off by looking at difference in mean value for each independent variable against foundation factor. Table of mean and frequency across foundation type.

| Variable | BrkTil | CBlock | PConc | Slab | Stone | Wood |
|---|---|---|---|---|---|---|
| Frequency | 146 | 634 | 647 | 24 | 6 | 3 |
| | Mean | Mean | Mean | Mean | Mean | Mean |
| log_saleprice | 11.7225277 | 11.8700797 | 12.2616651 | 11.5329593 | 11.9331562 | 12.1024793 |
| lotfrontage | 61.4394358 | 70.1638853 | 70.994341 | 66.1343133 | 66.6666667 | 102.2149013 |
| overallqual | 5.4452055 | 5.4211356 | 6.9799073 | 4.2916667 | 5.6666667 | 6.6666667 |
| overallcond | 6.1986301 | 5.829653 | 5.202473 | 4.75 | 7 | 5.6666667 |
| yearbuilt | 1921.02 | 1961.25 | 1993.31 | 1959.58 | 1912.67 | 1990.33 |
| yearremodadd | 1971.62 | 1975.22 | 1998.05 | 1965.17 | 1978.33 | 1997 |
| masvnrarea | 7 | 86.4600188 | 133.6680639 | 51.4583333 | 0 | 0 |
| bsmtfinsf1 | 165.8424658 | 477.1246057 | 484.0032077 | 0 | 45.8333333 | 791.6666667 |
| bsmtunfsf | 629.0479452 | 443.4716088 | 695.3292117 | 0 | 849.1666667 | 65.3333333 |
| totalbsmtsf | 814.6232877 | 1001.49 | 1200.88 | 0 | 895 | 857 |
| x2ndflrsf | 455.0068493 | 228.714511 | 436.8809892 | 218.8333333 | 800.8333333 | 818 |
| grlivarea | 1452.08 | 1355.5 | 1667.63 | 1339.46 | 1894.67 | 1876 |
| bedroomabvgr | 2.9178082 | 2.8609825 | 2.8438949 | 2.9166667 | 3.5 | 3 |
| kitchenabvgr | 1.0619494 | 1.05843 | 1.0093425 | 1.4583333 | 1.3333333 | 1 |
| totrmsabvgrd | 6.5547945 | 6.1340694 | 6.8686244 | 6.5 | 8.1666667 | 7 |
| garageyrblt | 1951.18 | 1967.66 | 1995.86 | 1969.08 | 1950.5 | 1990.33 |
| garagecars | 1.3082192 | 1.4952681 | 2.1468315 | 1.5 | 1.6666667 | 2 |
| garagearea | 344.6575342 | 410.8533123 | 566.1483771 | 375.0416667 | 464.3333333 | 555 |
| openporchsf | 26.9133409 | 32.4453435 | 63.0543646 | 8.7083333 | 67.8333333 | 14 |
| miscval | 27.5342466 | 33.5073863 | 9.3972179 | 71.868394 | 416.6666667 | 233.3333333 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Yrsold | 2007.73 | 2007.87 | 2007.77 | 2008.04 | 2008.67 | 2008.33 |

Lets, Plot the above mean distribution.




We can note that number of Cblock and PConc is very high and consumes 88% of the dataset. While slab and wood have negligible amount of data. This might cause misclassifying slab or wood type.

Variables masvnarea, bsmtfinsf1, bsmtunfsf, totalbsmtsf and x2ndflrsf differ noticeably between each foundation type. These predictor will have higher impact of classifying on foundation type to another.

Since chi-square test for within covariance failed, we will use pooled covariance in our model.

**<u>Model building:</u>**

We have run variaous models, thrown away variables with pvalue > 0.05. The below summary is of our final model.

Statistics:

| | | | |
|---|---|---|---|
| Total Sample Size | 1460 | DF Total | 1459 |
| Variables | 12 | DF Within Classes | 1454 |
| Classes | 6 | DF Between Classes | 5 |

| Pooled Covariance Matrix Information | |
|---|---|
| Covariance Matrix Rank | Natural Log of the Determinant of the Covariance Matrix |
| 19 | 79.03045 |

| Multivariate Statistics and F Approximations | | | | | |
|---|---|---|---|---|---|
| S=5 M=6.5 N=717 | | | | | |
| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
| Wilks' Lambda | 0.17625037 | 31.55 | 95 | 6992.1 | <.0001 |
| Pillai's Trace | 1.26646257 | 25.71 | 95 | 7200 | <.0001 |
| Hotelling-Lawley Trace | 2.56341658 | 38.71 | 95 | 5549 | <.0001 |
| Roy's Greatest Root | 1.65300566 | 125.28 | 19 | 1440 | <.0001 |
| NOTE: F Statistic for Roy's Greatest Root is an upper bound. | | | | | |

Since the mean values of each response variable differ from the factor levels we move on to univariate analysis.

## <u>Frequency and Priors</u>

| Class Level Information | | | | | |
|---|---|---|---|---|---|
| **foundation** | **Variable Name** | **Frequency** | **Weight** | **Proportion** | **Prior Probability** |
| **BrkTil** | BrkTil | 146 | 146.0000 | 0.100000 | 0.422535 |
| **CBlock** | CBlock | 634 | 634.0000 | 0.434247 | 0.140845 |
| **PConc** | PConc | 647 | 647.0000 | 0.443151 | 0.415962 |
| **Slab** | Slab | 24 | 24.0000 | 0.016438 | 0.015023 |
| **Stone** | Stone | 6 | 6.0000 | 0.004110 | 0.003756 |
| **Wood** | Wood | 3 | 3.0000 | 0.002055 | 0.001878 |

From the information we can see that priors are arbitrary. The decision to set is based on frequency and probability of them happening, good accuracy model.

PConc has a prior of 0.42 because of its domination on the dataset. Although CBlock is frequent it has been reduced and BrkTile has been increased to reduce the misclassifications of BrkTile into CBlock.

Slab, Stone and wood have been kept low because of its frequency and probability of happening.

## <u>Comparing Mahalanobis  distance within foundation types</u>

| Squared Distance to foundation | | | | | | |
|---|---|---|---|---|---|---|
| **From foundation** | **BrkTil** | **CBlock** | **PConc** | **Slab** | **Stone** | **Wood** |
| **BrkTil** | 0 | 6.64709 | 14.33441 | 32.45816 | 4.95205 | 29.90824 |
| **CBlock** | 6.64709 | 0 | 5.52319 | 24.34768 | 13.74802 | 18.33124 |
| **PConc** | 14.33441 | 5.52319 | 0 | 31.72738 | 22.02268 | 16.20705 |
| **Slab** | 32.45816 | 24.34768 | 31.72738 | 0 | 34.04687 | 39.00079 |
| **Stone** | 4.95205 | 13.74802 | 22.02268 | 34.04687 | 0 | 36.27761 |
| **Wood** | 29.90824 | 18.33124 | 16.20705 | 39.00079 | 36.27761 | 0 |

| Prob > Mahalanobis Distance for Squared Distance to foundation | | | | | | |
|---|---|---|---|---|---|---|
| **From foundation** | **BrkTil** | **CBlock** | **PConc** | **Slab** | **Stone** | **Wood** |
| **BrkTil** | 1.0000 | <.0001 | <.0001 | <.0001 | 0.0819 | <.0001 |
| **CBlock** | <.0001 | 1.0000 | <.0001 | <.0001 | <.0001 | <.0001 |
| **PConc** | <.0001 | <.0001 | 1.0000 | <.0001 | <.0001 | 0.0003 |
| **Slab** | <.0001 | <.0001 | <.0001 | 1.0000 | <.0001 | <.0001 |
| **Stone** | 0.0819 | <.0001 | <.0001 | <.0001 | 1.0000 | <.0001 |
| **Wood** | <.0001 | <.0001 | 0.0003 | <.0001 | <.0001 | 1.0000 |

From the figure above we can note that the distances between Stone and BrkTil disnce is low, We can see that these fail the <u>**Mahalanobis**</u> significance test. classification for these cannot be justified.

Among all the types Slab, Stone and wood are reasonably different from each other and easier to classify.

## Univariate analysis

| Univariate Test Statistics | | | | | | | |
|---|---|---|---|---|---|---|---|
| F Statistics, Num DF=5, Den DF=1454 | | | | | | | |
| Variable | Total Standard Deviation | Pooled Standard Deviation | Between Standard Deviation | R-Square | R-Square / (1-RSq) | F Value | Pr > F |
| log_saleprice | 0.3995 | 0.3339 | 0.2410 | 0.3037 | 0.4361 | 126.81 | <.0001 |
| lotfrontage | 20.1000 | 19.8817 | 3.4772 | 0.0250 | 0.0256 | 7.44 | <.0001 |
| overallqual | 1.3830 | 1.1293 | 0.8772 | 0.3355 | 0.5049 | 146.83 | <.0001 |
| overallcond | 1.1128 | 1.0459 | 0.4216 | 0.1197 | 0.1360 | 39.53 | <.0001 |
| yearbuilt | 30.2029 | 19.6159 | 25.1807 | 0.5796 | 1.3789 | 400.98 | <.0001 |
| yearremodadd | 20.6454 | 16.8902 | 13.0461 | 0.3330 | 0.4992 | 145.18 | <.0001 |
| log_masvnrarea | 2.6135 | 2.4929 | 0.8739 | 0.0932 | 0.1028 | 29.90 | <.0001 |
| bsmtfinsf1 | 435.2895 | 420.7455 | 125.1297 | 0.0689 | 0.0740 | 21.52 | <.0001 |
| bsmtunfsf | 441.8670 | 418.8021 | 156.6076 | 0.1048 | 0.1170 | 34.03 | <.0001 |
| totalbsmtsf | 418.2740 | 374.8027 | 204.7443 | 0.1998 | 0.2497 | 72.61 | <.0001 |
| log_x2ndflrsf | 3.2933 | 3.2004 | 0.8749 | 0.0589 | 0.0625 | 18.18 | <.0001 |
| grlivarea | 496.1339 | 473.1157 | 166.3551 | 0.0938 | 0.1035 | 30.08 | <.0001 |
| kitchenabvgr | 0.2063 | 0.1972 | 0.0677 | 0.0899 | 0.0988 | 28.73 | <.0001 |
| totrmsabvgrd | 1.6254 | 1.5875 | 0.3954 | 0.0493 | 0.0519 | 15.09 | <.0001 |
| garageyrblt | 23.9946 | 17.6747 | 17.8067 | 0.4593 | 0.8493 | 246.98 | <.0001 |
| garagecars | 0.7473 | 0.6645 | 0.3769 | 0.2121 | 0.2692 | 78.28 | <.0001 |
| garagearea | 213.8048 | 196.2095 | 93.8589 | 0.1607 | 0.1915 | 55.68 | <.0001 |
| openporchsf | 61.1182 | 58.9561 | 18.0448 | 0.0727 | 0.0784 | 22.80 | <.0001 |
| log_miscval | 1.1683 | 1.1576 | 0.1885 | 0.0217 | 0.0222 | 6.45 | <.0001 |

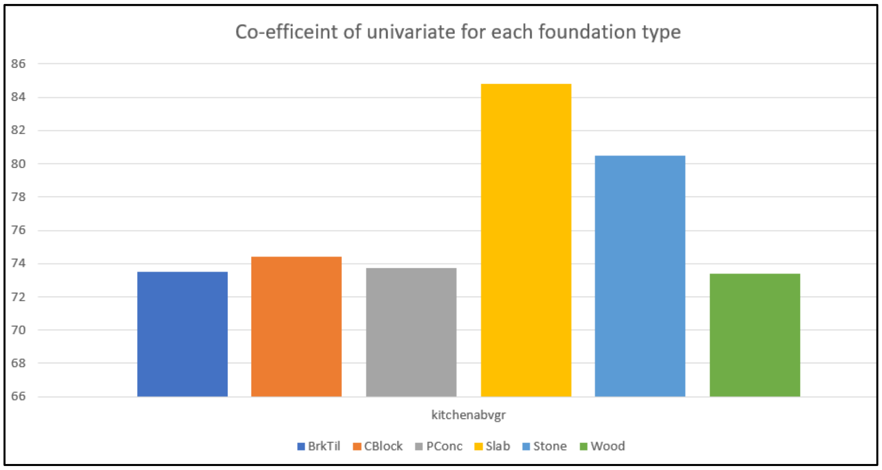
We can note the dependent variables included in the model have pvalues < 0.01, making them all signification in classifying the foundation type.

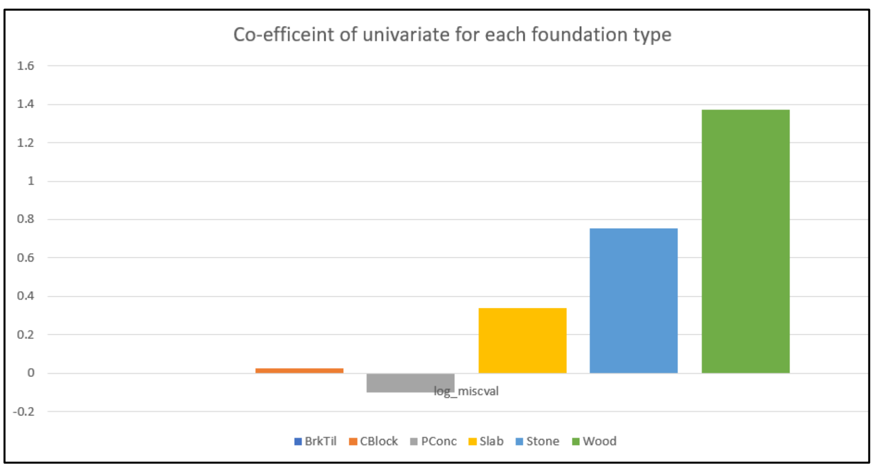
## LDA score for each foundation type:

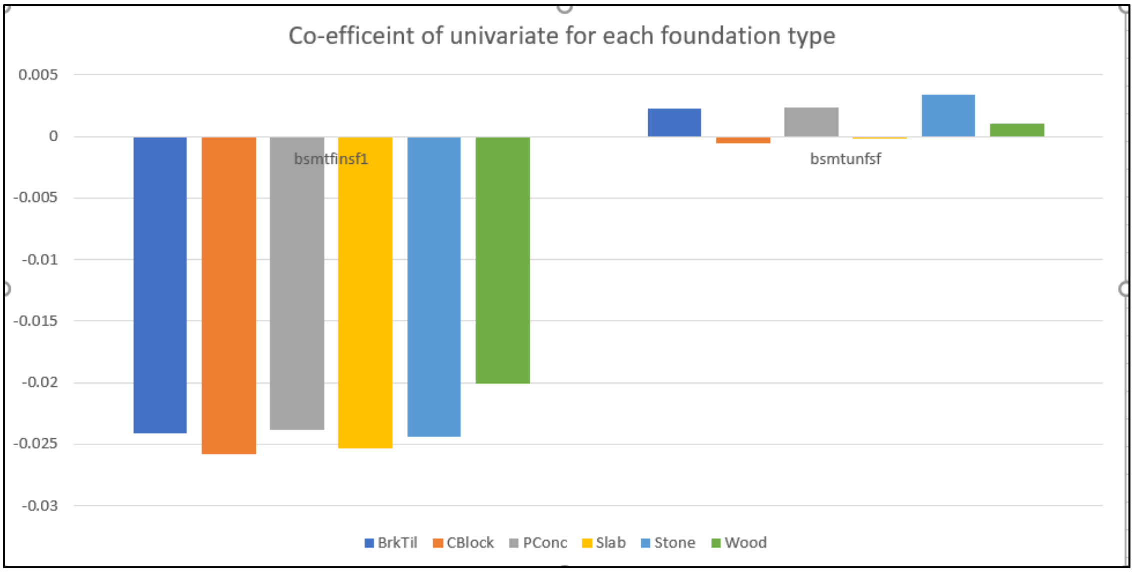
| Linear Discriminant Function for foundation | | | | | | |
|---|---|---|---|---|---|---|
| Variable | BrkTil | CBlock | PConc | Slab | Stone | Wood |
| Constant | -12738 | -12935 | -13174 | -12969 | -12758 | -13205 |
| log_saleprice | 218.23800 | 217.92699 | 220.49971 | 221.18633 | 220.20438 | 207.08187 |
| lotfrontage | 0.67947 | 0.70881 | 0.69563 | 0.71612 | 0.67222 | 0.82314 |
| overallqual | -68.13145 | -69.17323 | -68.59344 | -68.84250 | -68.50602 | -67.79816 |
| overallcond | -7.62658 | -6.97092 | -7.87450 | -8.25044 | -7.00043 | -6.83491 |
| yearbuilt | 3.00656 | 3.15990 | 3.17716 | 3.18482 | 2.98795 | 3.26491 |
| yearremodadd | 5.21780 | 5.17542 | 5.22318 | 5.15534 | 5.23066 | 5.23261 |
| log_masvnrarea | -6.16668 | -5.99265 | -6.21283 | -6.07605 | -6.22185 | -6.72188 |
| bsmtfinsf1 | -0.02415 | -0.02578 | -0.02388 | -0.02538 | -0.02437 | -0.02005 |
| bsmtunfsf | 0.00222 | -0.0006017 | 0.00234 | -0.0002185 | 0.00335 | 0.00105 |
| totalbsmtsf | -0.00691 | -0.00491 | -0.00722 | -0.02505 | -0.00820 | -0.00870 |
| log_x2ndflrsf | 6.81118 | 6.68527 | 6.86850 | 5.04112 | 6.72852 | 7.09369 |
| grlivarea | -0.03490 | -0.03332 | -0.03447 | -0.02501 | -0.03416 | -0.02715 |
| kitchenabvgr | 73.48621 | 74.43396 | 73.75574 | 84.77637 | 80.50236 | 73.38054 |
| totrmsabvgrd | 1.02602 | 0.90734 | 0.85179 | 0.97275 | 1.09025 | 0.09840 |
| garageyrblt | 3.73369 | 3.72606 | 3.76881 | 3.71416 | 3.72852 | 3.75051 |
| garagecars | -60.44115 | -61.49126 | -60.84745 | -60.82226 | -60.38015 | -61.29631 |
| garagearea | -0.06953 | -0.06862 | -0.07057 | -0.06863 | -0.06697 | -0.06782 |
| openporchsf | -0.27703 | -0.27611 | -0.27406 | -0.27539 | -0.26729 | -0.29041 |
| log_miscval | -0.00523 | 0.02361 | -0.10301 | 0.33745 | 0.75381 | 1.37211 |

Looking at the co-efficients, Bsntunsf and log_misCval are good classifier for foundation. Totrmsabvfrg,  garagearea helps to differenciate the foundation types.

To understand the co-efficent better, Below are the plots that visually show the difference in co-efficient for a particular variable.






## Classification summary

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.TRAIN
Resubstitution Summary using Linear Discriminant Function

| Number of Observations and Percent Classified into foundation | | | | | | | |
|---|---|---|---|---|---|---|---|
| From foundation | BrkTil | CBlock | PConc | Slab | Stone | Wood | Total |
| BrkTil | 133<br>91.10 | 8<br>5.48 | 0<br>0.00 | 0<br>0.00 | 5<br>3.42 | 0<br>0.00 | 146<br>100.00 |
| CBlock | 86<br>13.56 | 438<br>69.09 | 99<br>15.62 | 9<br>1.42 | 1<br>0.16 | 1<br>0.16 | 634<br>100.00 |
| PConc | 44<br>6.80 | 39<br>6.03 | 557<br>86.09 | 3<br>0.46 | 2<br>0.31 | 2<br>0.31 | 647<br>100.00 |
| Slab | 0<br>0.00 | 2<br>8.33 | 1<br>4.17 | 21<br>87.50 | 0<br>0.00 | 0<br>0.00 | 24<br>100.00 |
| Stone | 2<br>33.33 | 1<br>16.67 | 0<br>0.00 | 0<br>0.00 | 3<br>50.00 | 0<br>0.00 | 6<br>100.00 |
| Wood | 0<br>0.00 | 0<br>0.00 | 1<br>33.33 | 0<br>0.00 | 0<br>0.00 | 2<br>66.67 | 3<br>100.00 |
| Total | 265<br>18.15 | 488<br>33.42 | 658<br>45.07 | 33<br>2.26 | 11<br>0.75 | 5<br>0.34 | 1460<br>100.00 |
| Priors | 0.42254 | 0.14085 | 0.41596 | 0.01502 | 0.00376 | 0.00188 | |

| Error Count Estimates for foundation | | | | | | | |
|---|---|---|---|---|---|---|---|
| | BrkTil | CBlock | PConc | Slab | Stone | Wood | Total |
| Rate | 0.0890 | 0.3091 | 0.1391 | 0.1250 | 0.5000 | 0.3333 | 0.1434 |
| Priors | 0.4225 | 0.1408 | 0.4160 | 0.0150 | 0.0038 | 0.0019 | |

- Out of 146 BrkTil Type only 133 got classified correctly.
- 5% of BrkTile ended up as CBlock.
- 86% of CBlock was accurately classified.
- A high % of misclassification of CBlock is PConc with 15%.
- Overall error rate is 0.14% making an accuracy of this model to 86%

**Test data results:**

The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using Linear Discriminant Function

| Observation Profile for Test Data | |
|---|---|
| Number of Observations Read | 1459 |
| Number of Observations Used | 1459 |

| Number of Observations and Percent Classified into foundation | | | | | | | |
|---|---|---|---|---|---|---|---|
| From foundation | BrkTil | CBlock | PConc | Slab | Stone | Wood | Total |
| BrkTil | 150<br>90.91 | 10<br>6.06 | 1<br>0.61 | 3<br>1.82 | 1<br>0.61 | 0<br>0.00 | 165<br>100.00 |
| CBlock | 70<br>11.65 | 422<br>70.22 | 90<br>14.98 | 15<br>2.50 | 2<br>0.33 | 2<br>0.33 | 601<br>100.00 |
| PConc | 46<br>6.96 | 45<br>6.81 | 561<br>84.87 | 2<br>0.30 | 2<br>0.30 | 5<br>0.76 | 661<br>100.00 |
| Slab | 0<br>0.00 | 2<br>8.00 | 0<br>0.00 | 23<br>92.00 | 0<br>0.00 | 0<br>0.00 | 25<br>100.00 |
| Stone | 4<br>80.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 1<br>20.00 | 0<br>0.00 | 5<br>100.00 |
| Wood | 0<br>0.00 | 0<br>0.00 | 2<br>100.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 2<br>100.00 |
| Total | 270<br>18.51 | 479<br>32.83 | 654<br>44.83 | 43<br>2.95 | 6<br>0.41 | 7<br>0.48 | 1459<br>100.00 |
| Priors | 0.42254 | 0.14085 | 0.41596 | 0.01502 | 0.00376 | 0.00188 | |

| Error Count Estimates for foundation | | | | | | | |
|---|---|---|---|---|---|---|---|
| | BrkTil | CBlock | PConc | Slab | Stone | Wood | Total |
| Rate | 0.0909 | 0.2978 | 0.1513 | 0.0800 | 0.8000 | 1.0000 | 0.1494 |
| Priors | 0.4225 | 0.1408 | 0.4160 | 0.0150 | 0.0038 | 0.0019 | |

Pattern of classification of test is very similar to train data.

- BrkTil has accuretly classified to 90%.
- A large % of misclassification on BrkTil is CBlock.
- A large % of misclassification on CBlock is Pconc type, 14%.
- Overall error rate is 0.15 making accuracy of the result to 85%.

**Conclusion:**

The goal of this project was to the best of knowledge apply LDA as a classification technique (and hence impute the missing variables). The model built accurately classifies 85% of the cases into correct foundation type.