GE Analytics Engineer Program

Case Study: Detecting Pneumoconiosis

Table of Contents

Detecting Pneumoconiosis 2

1. About the Data 2

2. Feature Description 3

2.1 Intensity Based Features 3

2.2 Co-occurrence Matrix Based Features 3

3. Exploratory Data Analysis 4

3.1 Data Gaps/Missing Data 4

3.2 Number of Normal and Abnormal Cases 5

3.3 Combining Data 6

3.4 Feature Related Inferences 6

3.4.1 Analysis Using Statistical Tools 6

3.4.2 Analysis by Visualization 9

3.4.3 Deeper Analysis on 'Difficult' Variables 13

3.5 Check for Correlated Variables 15

3.4 Check for Outliers 17

4. Feature Selection 18

5. Model Selection 20

6. Results and Conclusions 21

6.1 Results with one model for each Zone 23

7. Possible Problems and Next Steps 25

8. References 25

Detecting Pneumoconiosis

One of the occupational hazards for Coal Miners is Pneumoconiosis. A leading hospital wishes to develop a screening program for coal miners, to facilitate early detection of Pneumoconiosis.

Our aim is to build a computer program that can tell if a Patient has Pneumoconiosis based on the images of six different parts/zones of the Patient's Lungs.

This report details the approach, methods used and results obtained by writing such a program in Python 3.6 using Anaconda's Spyder.

1. About the Data

The data collected was based on a study conducted with Shanghai Pulmonary Hospital.

We're provided with an excel workbook with six tabs. Each tab has feature data for one of the six zones of the Patient's Lungs along with the Patient Identifier: PatientNumMasked and a Label (0=Normal, 1=Abnormal) indicating if that zone of the lung exhibits pneumoconiosis for a Patient.

The six zones a pair of human lungs are divided into are: Right Upper (Zone-1), Right Middle (Zone-2), Right Lower (Zone-3), Left Upper (Zone-4), Left Middle (Zone-5), Left Lower (Zone-6).

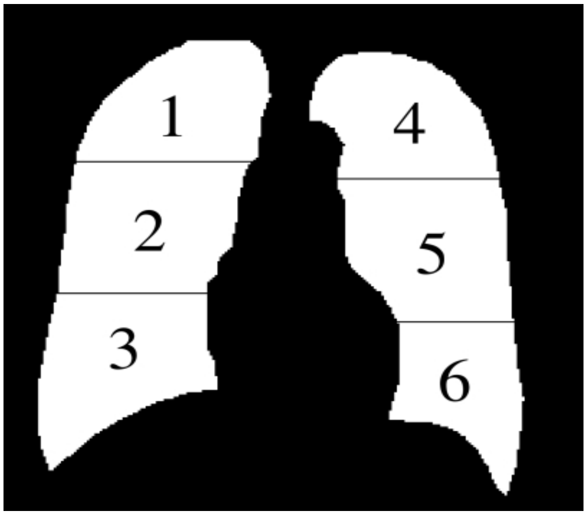


Fig-1

Even a single lung zone exhibiting abnormality amounts to concluding that the Patient has Pneumoconiosis.

2. Feature Description

Once a region of interest (lung zone) is segmented, it is characterized in terms of a set of features. We extract two types of features to describe each region of interest. These are described below:

2.1 Intensity based Features: We extract a set of 6 features based on the histogram of intensity values – mean, standard deviation, skewness, kurtosis, energy and entropy.

- **Mean**: Indicates the average intensity level
- **Variance**: Variation of Intensities around the Mean
- **Skewness**: Shows whether the histogram is symmetrical about the Mean
- **Kurtosis**: Shows whether the data is peaked or flat about the normal distribution
- **Entropy:** Measure of system disorder

- Apart from calculating these on the original ROI, we also extract these features after applying a difference filter on the image for local enhancement.
- If $I(x, y)$ denotes the image gray value at (x, y), the first and second order filters are defined as:

$$L_1^{\theta}(d) \quad = \quad f_x \cos\theta + f_y \sin\theta$$

$$L_2^{\theta}(d) \quad = \quad f_{xx} \cos^2\theta + f_{yy} \sin^2\theta + f_{xy} \cos\theta \sin\theta$$

Where,

d is the difference scale

θ is the orientation at which the difference is calculated

f_x and f_y represent the first order difference

f_{xx}, f_{yy}, f_{xy} represent the second order difference.

- We use the first and second order difference filter bank with given orientations $\theta \in \{0, 30, 35, 60, 90, 120, 135, 150, 180\}$ and given scale $d \in \{1, 2\}$.
- We can calculate 6 intensity-based features (mean, variance, skewness, kurtosis, energy, entropy) for each filtered image, along with the same features for the raw image without filtering, amounting to a total of 222 features.
- A subset of 34 features from this set has been provided in the attached data sheet. These features are labeled with the prefix $Hist\_d\_\theta$.

2.2 Co-occurrence Matrix based Features: We also extract a set of 5 features based on the gray level co-occurrence matrix computed for the ROI, namely energy, entropy, local homogeneity, correlation and inertia.

- The co-occurrence matrix allows us to capture the level of similarity and dissimilarity among adjacent pixels in an ROI. Thus, an ROI with an opacity will contain adjacent pixels with similarly high intensities, whereas a normal ROI will not contain such adjacent pixels.
- Computing these features for various orientations $\delta = \{0, 45, 90, 135\}$ captures this information for various types of adjacency.
- A subset of 5 of out of 25 such features has been provided in the attached data sheet. These features are labeled with the prefix $CoMatrix\_Deg\delta$.

Thus, a total of 39 features for each lung zone has been provided in the attached Excel spreadsheet. The first column in each worksheet (one sheet per zone) gives the patient number, while the last column gives the label.

For our analysis, ***'Label' is the target or dependent variable and all other features excluding 'PatientNumMasked' are the predictor variables and are 39 in number.***

3. Exploratory Analysis

3.1 Data Gaps/Missing Data

Each zone/tab has a different number of observations/samples.

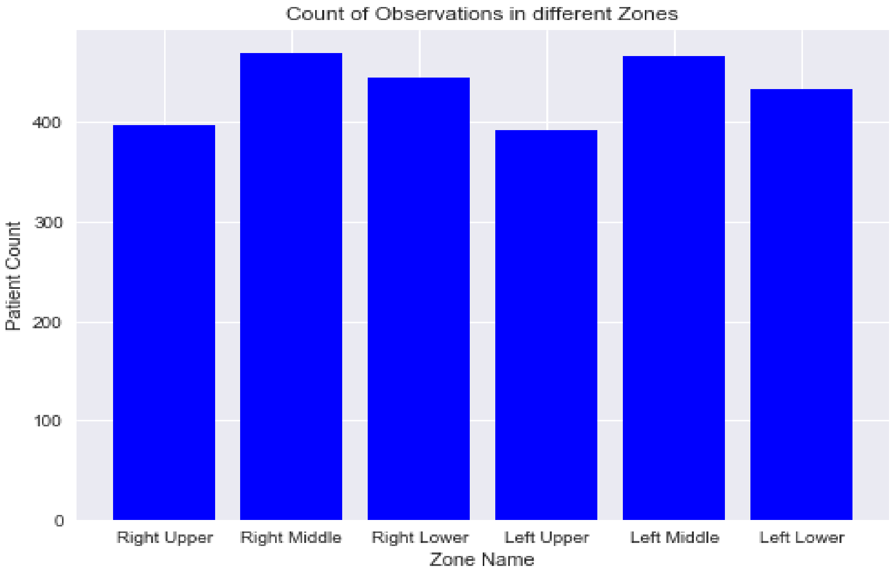


Fig-2

From Fig-2, we see that Left Upper zone has the least number of Observations/Patient count = 392 and the Right Middle zone has the highest with a count of 470.

There are 384 Patients with data in all the six zones and about 86 patients with data in only one or more zones. Also, we can find that there is **no duplicate Patient data in any of the zones.**

All six zones have the same features (39 in number). No samples with missing data for any features were found.

3.2 Number of Normal and Abnormal Cases

Fig-3, is a plot of the count of Patients marked as healthy and diseased in each Zone.

X axis - displays the possible values of 'Label' (0 and 1) column for each Zone. 0- Healthy, 1- Diseased

Y axis - displays the count of Patients for the distinct values of 'Label'

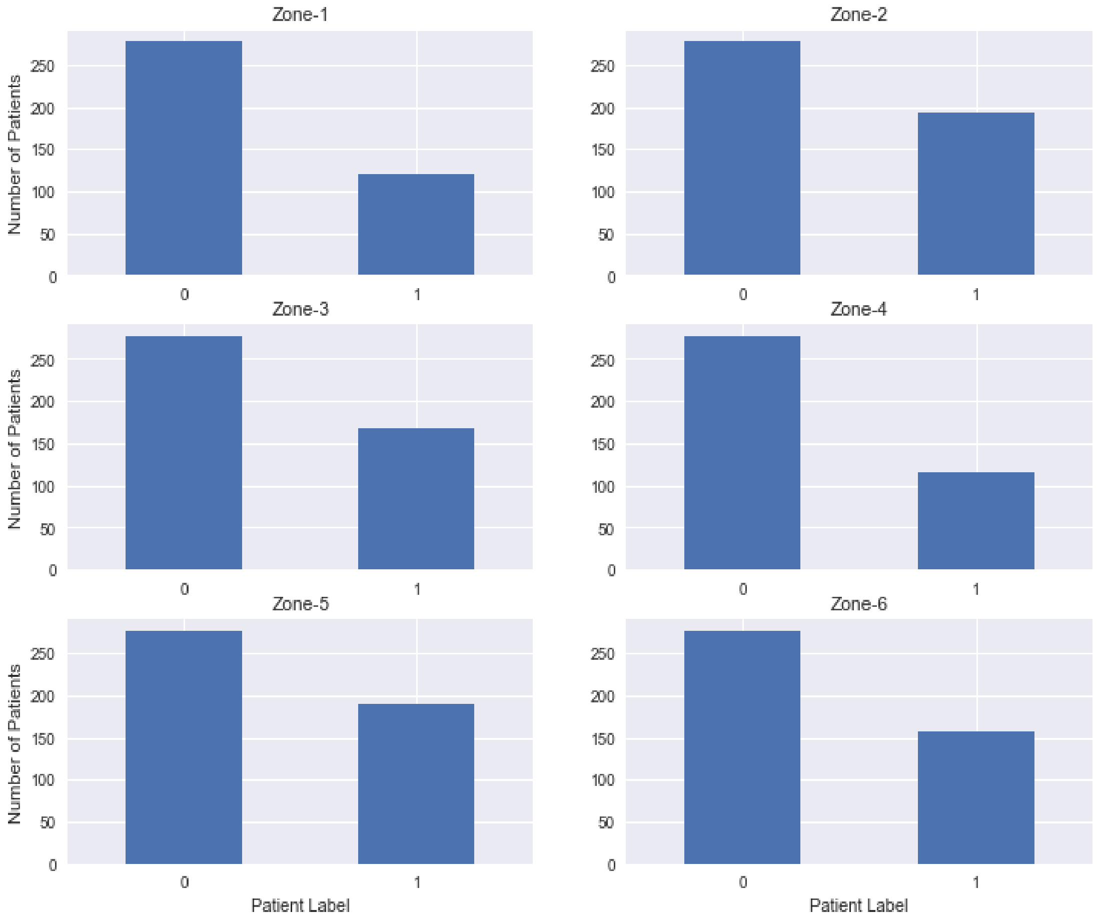


Fig-3

We can see that the **number of Normal/Healthy Patients is the same across all Zones**. But, the number of abnormal cases differ in each zone. Upon further examination, it is found that this difference is only due to the unequal number of Patients in each zone.

Also, **among all the Patients who have data for all six zones and have Pneumoconiosis, the Label is 1 for all zones. We may proceed to infer that anyone with Pneumoconiosis tend to exhibit abnormality in all six zones of the lungs.**

3.3 Combining Data

For analysis and model building, we can follow either of the two approaches

1. Combine data across zones into a single data set and predict results on test set
2. Perform analysis and modeling for each zone and then combine results to conclude as:

$$y_i = max_j\ (y_{ij}).$$

$y_{ij} \in \{0,1\}$ represent the zone-level labels (1=Pneumoconiosis, 0=healthy)

Here, *i* is the Patient identifier and $j \in \{1 \ldots 6\}$

Note that, even if one of the zones show evidence of Pneumoconiosis, the patient is diagnosed as having the disease.

We'll train our model on the consolidated dataset as well as the six different datasets and compare metrics to conclude on the best method/model.

3.4 Feature Related Inferences

As a first step in Analysis, I will **combine data of all zones as one data-set** and **check which of the 39 variables have a significant relationship with target variable.** This is to get a list of variables that influence Labels of all zones.

Statsmodels.formula.api provides us with logit package to perform **logistic regression**. We use this to **study the effect of all predictor variables, various combinations of predictor variables and individual predictor variables on 'Label' (target variable)**.

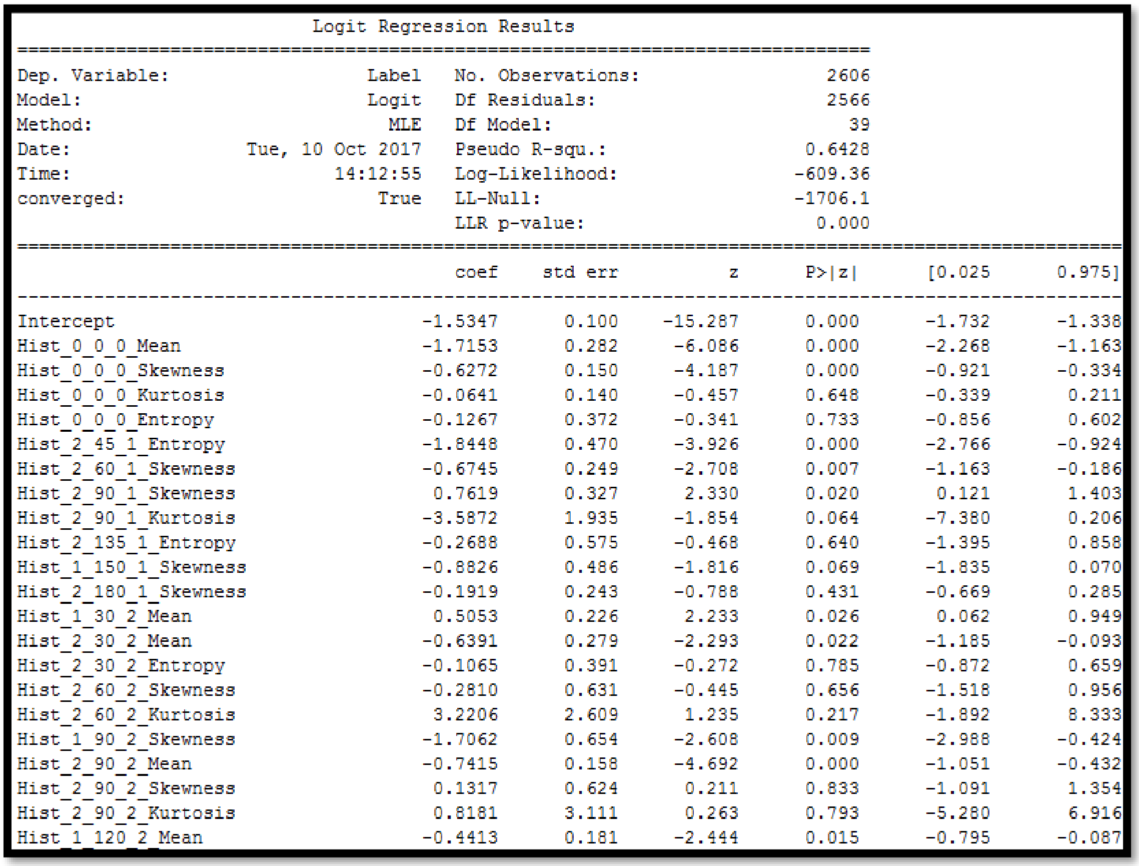
3.4.1. Analysis using Statistical Tools

Upon performing logistic regression on the consolidated data set, we see from the results shown below **only 22 variables of the 39 have a significant influence on the target** – Label across all six lung zones.

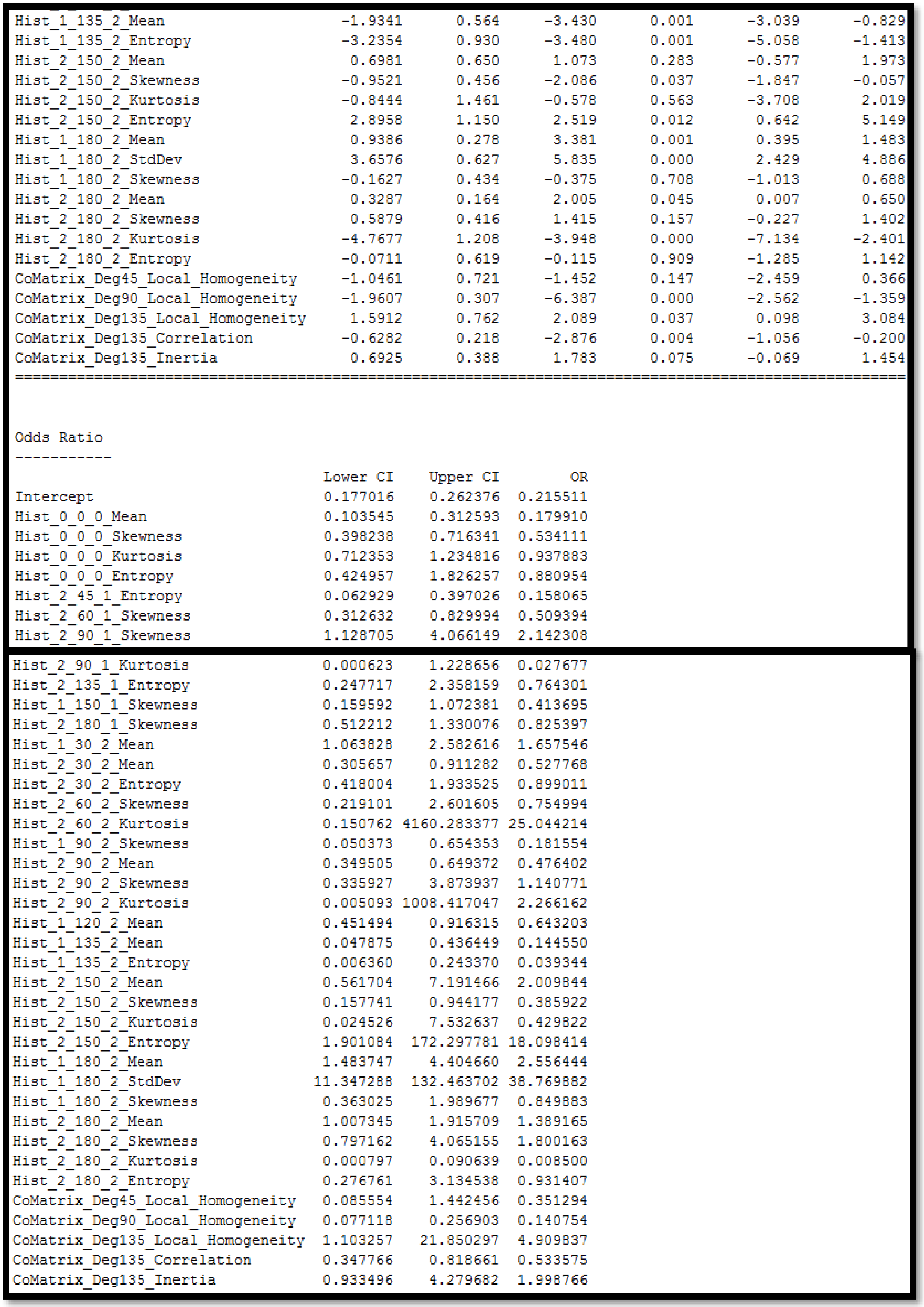
Note: This function is run after centering/normalizing all variables. All variables have different scales and drastically different average values. For example, the mean values such as Hist_0_0_0_Mean and others have very high average values in the range of a few thousands but some of the other variables have an average value of about 2.5.

Normalizing variables will prevent gross influence of prevents on the target and speeds up convergence.

From the Logit Regression Results, the variable coefficient(coef), p-value(P>\z\) are of importance to us for analysis and inference. The function results are displayed below along with the odds ratio and 95% confidence interval range values for the variables.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Label   No. Observations:                 2606
Model:                          Logit   Df Residuals:                     2566
Method:                           MLE   Df Model:                           39
Date:                Tue, 10 Oct 2017   Pseudo R-squ.:                  0.6428
Time:                        14:12:55   Log-Likelihood:                -609.36
converged:                       True   LL-Null:                       -1706.1
                                        LLR p-value:                     0.000
======================================================================================================
                                        coef    std err          z       P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------
Intercept                            -1.5347      0.100    -15.287       0.000      -1.732      -1.338
Hist_0_0_0_Mean                      -1.7153      0.282     -6.086       0.000      -2.268      -1.163
Hist_0_0_0_Skewness                  -0.6272      0.150     -4.187       0.000      -0.921      -0.334
Hist_0_0_0_Kurtosis                  -0.0641      0.140     -0.457       0.648      -0.339       0.211
Hist_0_0_0_Entropy                   -0.1267      0.372     -0.341       0.733      -0.856       0.602
Hist_2_45_1_Entropy                  -1.8448      0.470     -3.926       0.000      -2.766      -0.924
Hist_2_60_1_Skewness                 -0.6745      0.249     -2.708       0.007      -1.163      -0.186
Hist_2_90_1_Skewness                  0.7619      0.327      2.330       0.020       0.121       1.403
Hist_2_90_1_Kurtosis                 -3.5872      1.935     -1.854       0.064      -7.380       0.206
Hist_2_135_1_Entropy                 -0.2688      0.575     -0.468       0.640      -1.395       0.858
Hist_1_150_1_Skewness                -0.8826      0.486     -1.816       0.069      -1.835       0.070
Hist_2_180_1_Skewness                -0.1919      0.243     -0.788       0.431      -0.669       0.285
Hist_1_30_2_Mean                      0.5053      0.226      2.233       0.026       0.062       0.949
Hist_2_30_2_Mean                     -0.6391      0.279     -2.293       0.022      -1.185      -0.093
Hist_2_30_2_Entropy                  -0.1065      0.391     -0.272       0.785      -0.872       0.659
Hist_2_60_2_Skewness                 -0.2810      0.631     -0.445       0.656      -1.518       0.956
Hist_2_60_2_Kurtosis                  3.2206      2.609      1.235       0.217      -1.892       8.333
Hist_1_90_2_Skewness                 -1.7062      0.654     -2.608       0.009      -2.988      -0.424
Hist_2_90_2_Mean                     -0.7415      0.158     -4.692       0.000      -1.051      -0.432
Hist_2_90_2_Skewness                  0.1317      0.624      0.211       0.833      -1.091       1.354
Hist_2_90_2_Kurtosis                  0.8181      3.111      0.263       0.793      -5.280       6.916
Hist_1_120_2_Mean                    -0.4413      0.181     -2.444       0.015      -0.795      -0.087
```



```
Hist_1_135_2_Mean                     -1.9341      0.564     -3.430      0.001      -3.039      -0.829
Hist_1_135_2_Entropy                  -3.2354      0.930     -3.480      0.001      -5.058      -1.413
Hist_2_150_2_Mean                      0.6981      0.650      1.073      0.283      -0.577       1.973
Hist_2_150_2_Skewness                 -0.9521      0.456     -2.086      0.037      -1.847      -0.057
Hist_2_150_2_Kurtosis                 -0.8444      1.461     -0.578      0.563      -3.708       2.019
Hist_2_150_2_Entropy                   2.8958      1.150      2.519      0.012       0.642       5.149
Hist_1_180_2_Mean                      0.9386      0.278      3.381      0.001       0.395       1.483
Hist_1_180_2_StdDev                    3.6576      0.627      5.835      0.000       2.429       4.886
Hist_1_180_2_Skewness                 -0.1627      0.434     -0.375      0.708      -1.013       0.688
Hist_2_180_2_Mean                      0.3287      0.164      2.005      0.045       0.007       0.650
Hist_2_180_2_Skewness                  0.5879      0.416      1.415      0.157      -0.227       1.402
Hist_2_180_2_Kurtosis                 -4.7677      1.208     -3.948      0.000      -7.134      -2.401
Hist_2_180_2_Entropy                  -0.0711      0.619     -0.115      0.909      -1.285       1.142
CoMatrix_Deg45_Local_Homogeneity      -1.0461      0.721     -1.452      0.147      -2.459       0.366
CoMatrix_Deg90_Local_Homogeneity      -1.9607      0.307     -6.387      0.000      -2.562      -1.359
CoMatrix_Deg135_Local_Homogeneity      1.5912      0.762      2.089      0.037       0.098       3.084
CoMatrix_Deg135_Correlation           -0.6282      0.218     -2.876      0.004      -1.056      -0.200
CoMatrix_Deg135_Inertia                0.6925      0.388      1.783      0.075      -0.069       1.454
=====================================================================================================


Odds Ratio
-----------
                                    Lower CI     Upper CI         OR
Intercept                           0.177016     0.262376   0.215511
Hist_0_0_0_Mean                     0.103545     0.312593   0.179910
Hist_0_0_0_Skewness                 0.398238     0.716341   0.534111
Hist_0_0_0_Kurtosis                 0.712353     1.234816   0.937883
Hist_0_0_0_Entropy                  0.424957     1.826257   0.880954
Hist_2_45_1_Entropy                 0.062929     0.397026   0.158065
Hist_2_60_1_Skewness                0.312632     0.829994   0.509394
Hist_2_90_1_Skewness                1.128705     4.066149   2.142308
Hist_2_90_1_Kurtosis                0.000623     1.228656   0.027677
Hist_2_135_1_Entropy                0.247717     2.358159   0.764301
Hist_1_150_1_Skewness               0.159592     1.072381   0.413695
Hist_2_180_1_Skewness               0.512212     1.330076   0.825397
Hist_1_30_2_Mean                    1.063828     2.582616   1.657546
Hist_2_30_2_Mean                    0.305657     0.911282   0.527768
Hist_2_30_2_Entropy                 0.418004     1.933525   0.899011
Hist_2_60_2_Skewness                0.219101     2.601605   0.754994
Hist_2_60_2_Kurtosis                0.150762  4160.283377  25.044214
Hist_1_90_2_Skewness                0.050373     0.654353   0.181554
Hist_2_90_2_Mean                    0.349505     0.649372   0.476402
Hist_2_90_2_Skewness                0.335927     3.873937   1.140771
Hist_2_90_2_Kurtosis                0.005093  1008.417047   2.266162
Hist_1_120_2_Mean                   0.451494     0.916315   0.643203
Hist_1_135_2_Mean                   0.047875     0.436449   0.144550
Hist_1_135_2_Entropy                0.006360     0.243370   0.039344
Hist_2_150_2_Mean                   0.561704     7.191466   2.009844
Hist_2_150_2_Skewness               0.157741     0.944177   0.385922
Hist_2_150_2_Kurtosis               0.024526     7.532637   0.429822
Hist_2_150_2_Entropy                1.901084   172.297781  18.098414
Hist_1_180_2_Mean                   1.483747     4.404660   2.556444
Hist_1_180_2_StdDev                11.347288   132.463702  38.769882
Hist_1_180_2_Skewness               0.363025     1.989677   0.849883
Hist_2_180_2_Mean                   1.007345     1.915709   1.389165
Hist_2_180_2_Skewness               0.797162     4.065155   1.800163
Hist_2_180_2_Kurtosis               0.000797     0.090639   0.008500
Hist_2_180_2_Entropy                0.276761     3.134538   0.931407
CoMatrix_Deg45_Local_Homogeneity    0.085554     1.442456   0.351294
CoMatrix_Deg90_Local_Homogeneity    0.077118     0.256903   0.140754
CoMatrix_Deg135_Local_Homogeneity   1.103257    21.850297   4.909837
CoMatrix_Deg135_Correlation         0.347766     0.818661   0.533575
CoMatrix_Deg135_Inertia             0.933496     4.279682   1.998766
```

Fig-4

Inferences:

- All variables with a P>|z| or p-value>0.05 do not influence the Label significantly. Also, variables that have an odds ratio =1 results in a 50-50 chance of the Label being 0 or 1.
- **Variables with p-value <0.05 have a significant relationship with the Label**. List of variables with this characteristic are:
    - Hist_0_0_0_Mean
    - Hist_0_0_0_Skewness
    - Hist_2_45_1_Entropy
    - Hist_2_60_1_Skewness
    - Hist_2_90_1_Skewness
    - Hist_1_30_2_Mean
    - Hist_2_30_2_Mean
    - Hist_1_90_2_Skewness
    - Hist_2_90_2_Mean
    - Hist_1_120_2_Mean
    - Hist_1_135_2_Mean
    - Hist_1_135_2_Entropy
    - Hist_2_150_2_Skewness
    - Hist_2_150_2_Entropy
    - Hist_1_180_2_Mean
    - Hist_1_180_2_StdDev
    - Hist_2_180_2_Mean
    - Hist_2_180_2_Kurtosis
    - CoMatrix_Deg90_Local_Homogeneity
    - CoMatrix_Deg135_Local_Homogeneity
    - CoMatrix_Deg135_Correlation
    - CoMatrix_Deg135_Inertia
- **When the coef (variable Co-efficient) is or and the Odds ratio is >1, as the value of the variable increases, there is a higher probability of the Label being 1.**
- **Similarly, if the coef is negative or Odds ratio is <1, lower the variable value, higher is probability of the Label being 1.**

3.4.2. Analysis by Visualization

The 22 variables are divided into 4 bins labeled 'Very Low', 'Low', 'Average' and 'High'. Each bin has equal number of Patients. We visualize the relationship between these variables and the Label.

- **X axis represents the various bins of the predictor variable.**
- **Y axis represents the proportion of Patients diagnosed with Pneumoconiosis in each bin.**

For example, in Fig-5 the very first plot shows that:

- **More than 80% of the Patients with ‘High’ values of Hist_2_150_2_Entropy have Pneumoconiosis (Label=1)**
- **Less than 5% of the Patients with ‘Very Low’ values of Hist_2_150_2_Entropy have Pneumoconiosis (Label=1)**

We see a linear and **positive relationship between predictor(Hist_2_150_2_Entropy_Bins) and target** which reflects the Logit function results.

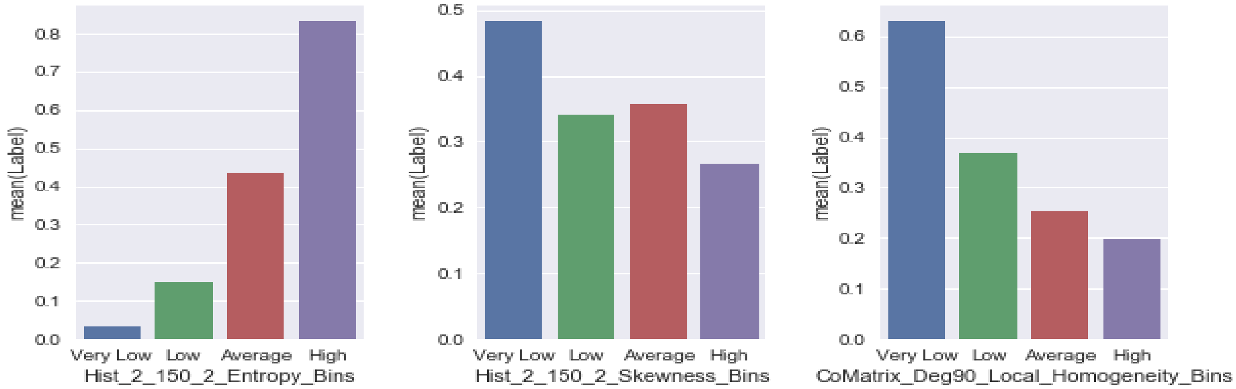


Fig-5

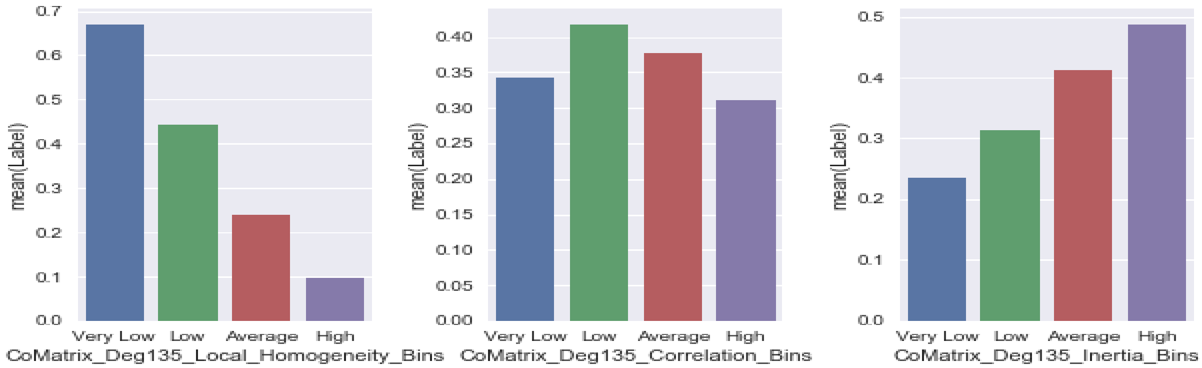


Fig-6

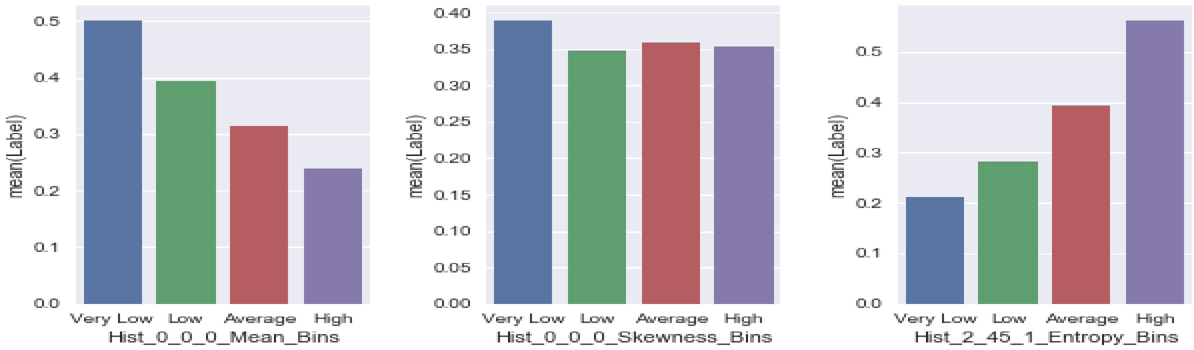


Fig-7

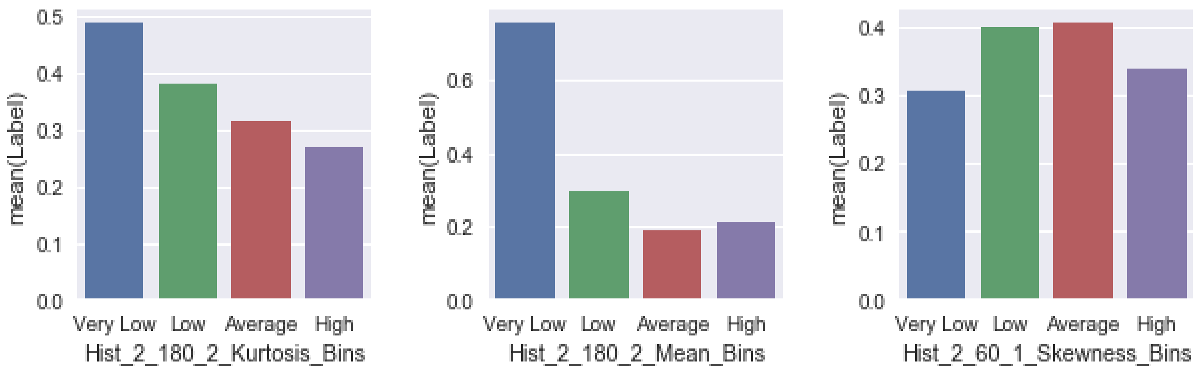


Fig-8

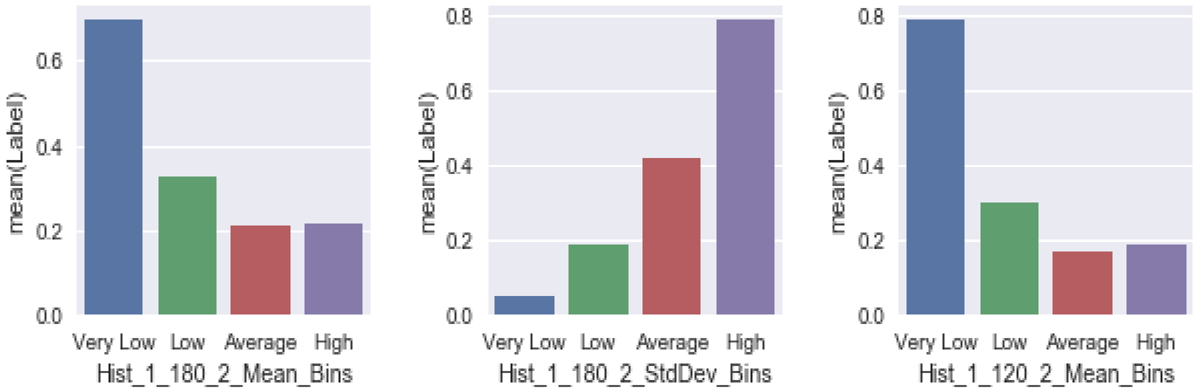


Fig-9

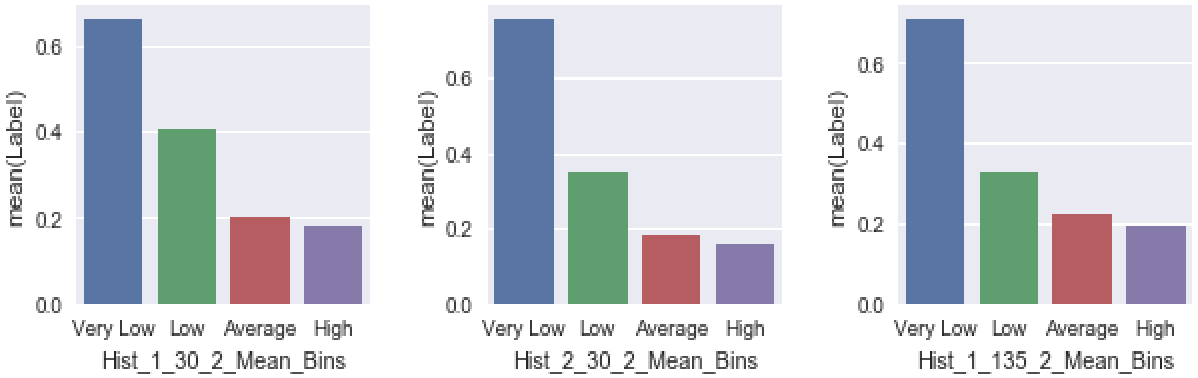


Fig-10

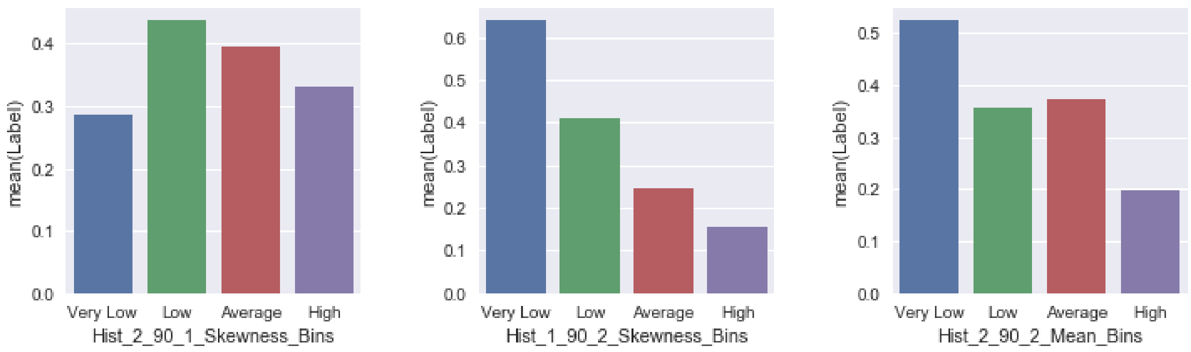


Fig-11

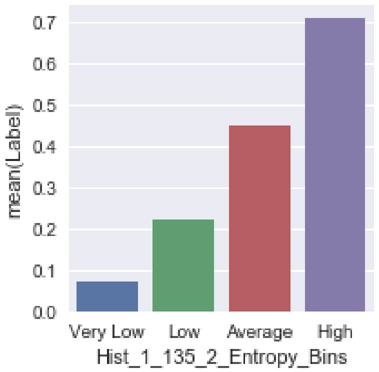


Fig-12

From the plots, we can infer that:

- **All 'Entropy' variables regardless of the filter** (for example: Hist_1_135_2_Entropy, Hist_2_45_1_Entropy etc.) **and CoMatrix_Deg135_Inertia, exhibit a positive relationship with target**. **Higher the entropy, larger are the chances that the Patient will have pneumoconiosis(Label=1)**

- Similarly, **all 'Kurtosis' and 'Mean' variables are negatively related to target. Lower the Mean/Kurtosis value, higher are the chances that the Patient will have Pneumoconiosis(Label=1).**

- **CoMatrix_Deg135_Local_Homogeneity and CoMatrix_Deg90_Local_Homogeneity also exhibit a negative relationship with target**

- However, there are some **variables whose relationship with target cannot be clearly determined.** They do not clearly show either a positive or negative relationship. They **also have Odds Ratios close to 1**. These are: **Hist_2_90_1_Skewness, Hist_0_0_0_Skewness, Hist_2_60_1_Skewness, CoMatrix_Deg135_Correlation.**
- **Seems like most 'Skewness' variable plots don't seem to clearly indicate the kind the relationship they have with 'Label'**

For the variables that haven't indicated a clear relationship with target, we perform further analysis to make more better inferences:

1. **Check if the relationship of the variable with the target is confounded** by another variable due to which our plots weren't indicative of any kind of influence on the 'Label'
2. **Visualize their relationship with 'Label' for each of the six zones**. This is to check if the variable's relationship with target cannot be generalized for all zones, maybe it can be clearly indicative at zonal level.

For brevity, not all plots are going to be displayed in the report; only the inferences/conclusions will be mentioned.

3.4.3. Deeper Analysis on 'Difficult' Variables

Hist_2_90_1_Skewness, Hist_0_0_0_Skewness, Hist_2_60_1_Skewness, CoMatrix_Deg135_Correlation are referred to as 'Difficult' variables as they did not exhibit a clear relationship with target during early stages of analysis.

We check if these variables show a distinct relationship with Zonal Labels i.e. target variable in each of the zones.

Hist_0_0_0_Skewness

We check if Hist_0_0_0_Skewness exhibits a distinctive relationship with the Label for any one or more zones individually rather than the combined data set by plotting it against the proportion of observations where the Label =1.

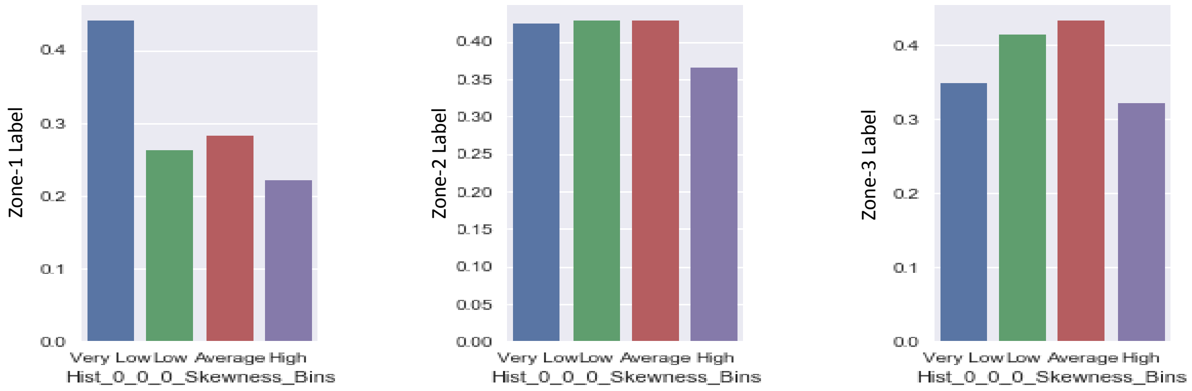


Fig-13

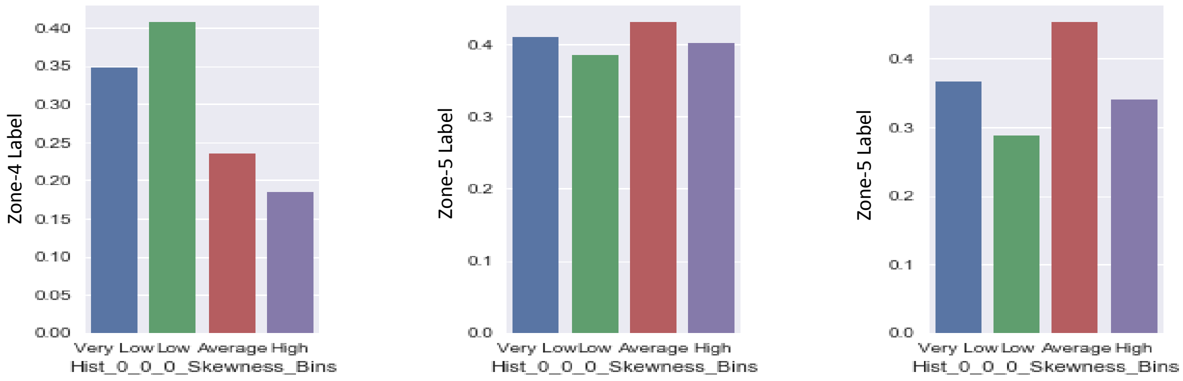


Fig-14

Hist_0_0_0_ Skewness variable exhibits a significant 'negative' relationship with the Label/target only in Zone-1. Hence, the relationship doesn't generalize well across all the 6 zones.

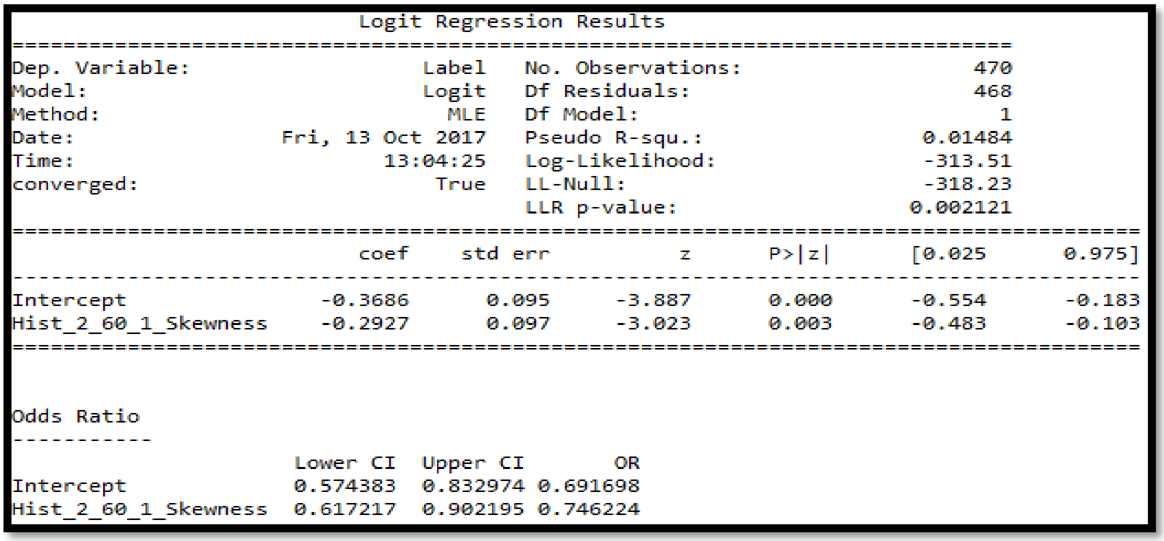
Hist_2_90_1_Skewness

On similar lines as Hist_0_0_0_Skewness, analysis on **relationship between Hist_2_90_1_Skewness and the Label also exhibits a positive relationship with the Label but only in the Right Upper and Left Upper Lung Zones.**

Hist_2_60_1_Skewness

We observe instances of the relationship between Hist_2_60_1_Skewness and 'Label' to be confounded. One such instance is shown below.

Logit results below show that Hist_2_60_1_Skewness has a significant relationship with the target/Label on the **Right Middle Zone**:



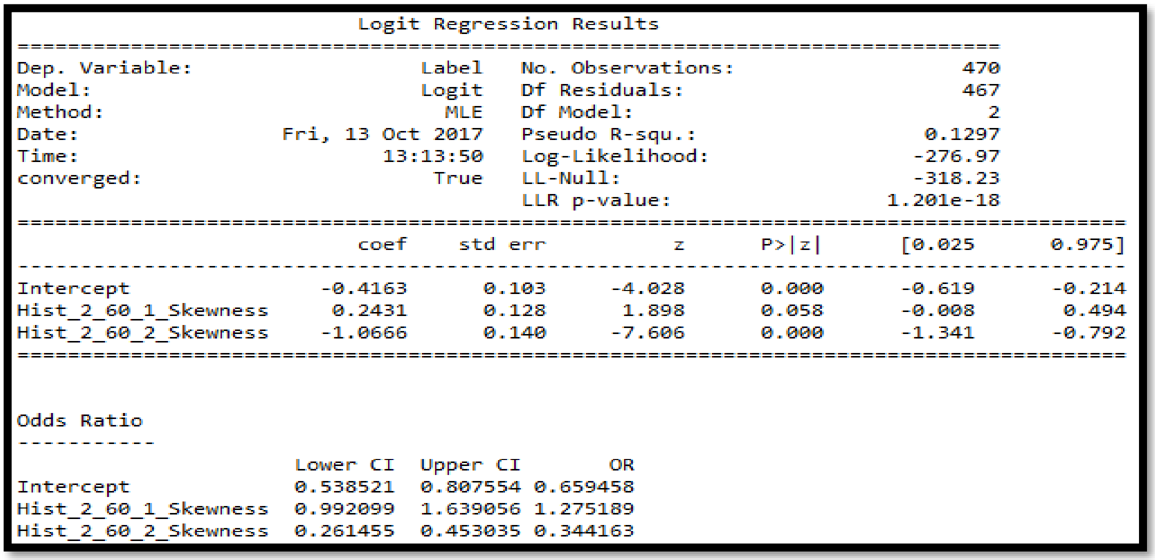
```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Label   No. Observations:                  470
Model:                          Logit   Df Residuals:                      468
Method:                           MLE   Df Model:                            1
Date:                Fri, 13 Oct 2017   Pseudo R-squ.:                 0.01484
Time:                        13:04:25   Log-Likelihood:                -313.51
converged:                       True   LL-Null:                       -318.23
                                        LLR p-value:                  0.002121
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept               -0.3686      0.095     -3.887      0.000      -0.554      -0.183
Hist_2_60_1_Skewness    -0.2927      0.097     -3.023      0.003      -0.483      -0.103
========================================================================================


Odds Ratio
-----------
                      Lower CI  Upper CI        OR
Intercept             0.574383  0.832974  0.691698
Hist_2_60_1_Skewness  0.617217  0.902195  0.746224
```

Fig-15

But when the effects of Hist_2_60_2_Skewness (skewness of the histogram of the image generated after applying 2nd filter at 60 degrees) on the Label are controlled for, the influence of Hist_2_60_1_Sknewness is no longer significant.

Fig-17 confirms our theory on a confounded relationship.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Label   No. Observations:                  470
Model:                          Logit   Df Residuals:                      467
Method:                           MLE   Df Model:                            2
Date:                Fri, 13 Oct 2017   Pseudo R-squ.:                  0.1297
Time:                        13:13:50   Log-Likelihood:                -276.97
converged:                       True   LL-Null:                       -318.23
                                        LLR p-value:                 1.201e-18
=======================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               -0.4163      0.103     -4.028      0.000      -0.619      -0.214
Hist_2_60_1_Skewness     0.2431      0.128      1.898      0.058      -0.008       0.494
Hist_2_60_2_Skewness    -1.0666      0.140     -7.606      0.000      -1.341      -0.792
=======================================================================================


Odds Ratio
-----------
                      Lower CI  Upper CI        OR
Intercept             0.538521  0.807554  0.659458
Hist_2_60_1_Skewness  0.992099  1.639056  1.275189
Hist_2_60_2_Skewness  0.261455  0.453035  0.344163
```

Fig-16

- After including Hist_2_60_2_Skewness, the p-value of Hist_2_60_1_Skewness is >0.05 indicating Hist_2_60_1_Skewness doesn't significantly influence the Target anymore.
- Hence, **Hist_2_60_2_Skewness here would be a possible confounder of the relationship between Hist_2_60_1_Skewness and Target as they are not correlated variables.**
- Again, using logit, while further examining influence of **CoMatrix_Deg135_Correlation on the Label on the combined data-set, it is found that CoMatrix_Deg90_Local_Homogeneity confounds that relationship.**
- Also, while **analyzing the left (3 zones) and right (3 zones) separately, it is found that the Co-occurrence matrix based features have a significant relationship with the Label of the Left-Upper, Left-middle and Left-Lower zones compared to the Right zones.**

3.5 Check for Correlated features

We've seen some cases of confounding relationships among the predictors. Hence, it is essential we check for correlated features as well.

Correlated features can be used to make new and more important features derived to be used in our model or some of them can be eliminated from model building as their effect on 'Label' may be redundant.

But, we do no eliminate correlated features manually before modeling. This is taken care of by the feature selection algorithm detailed in the next section.

Otherwise, using many correlated features deteriorates performance of the model.

The Heatmap in Fig-18 is generated using the Correlation Matrix of all features of the consolidated data-set (data of all six lung zones combined).

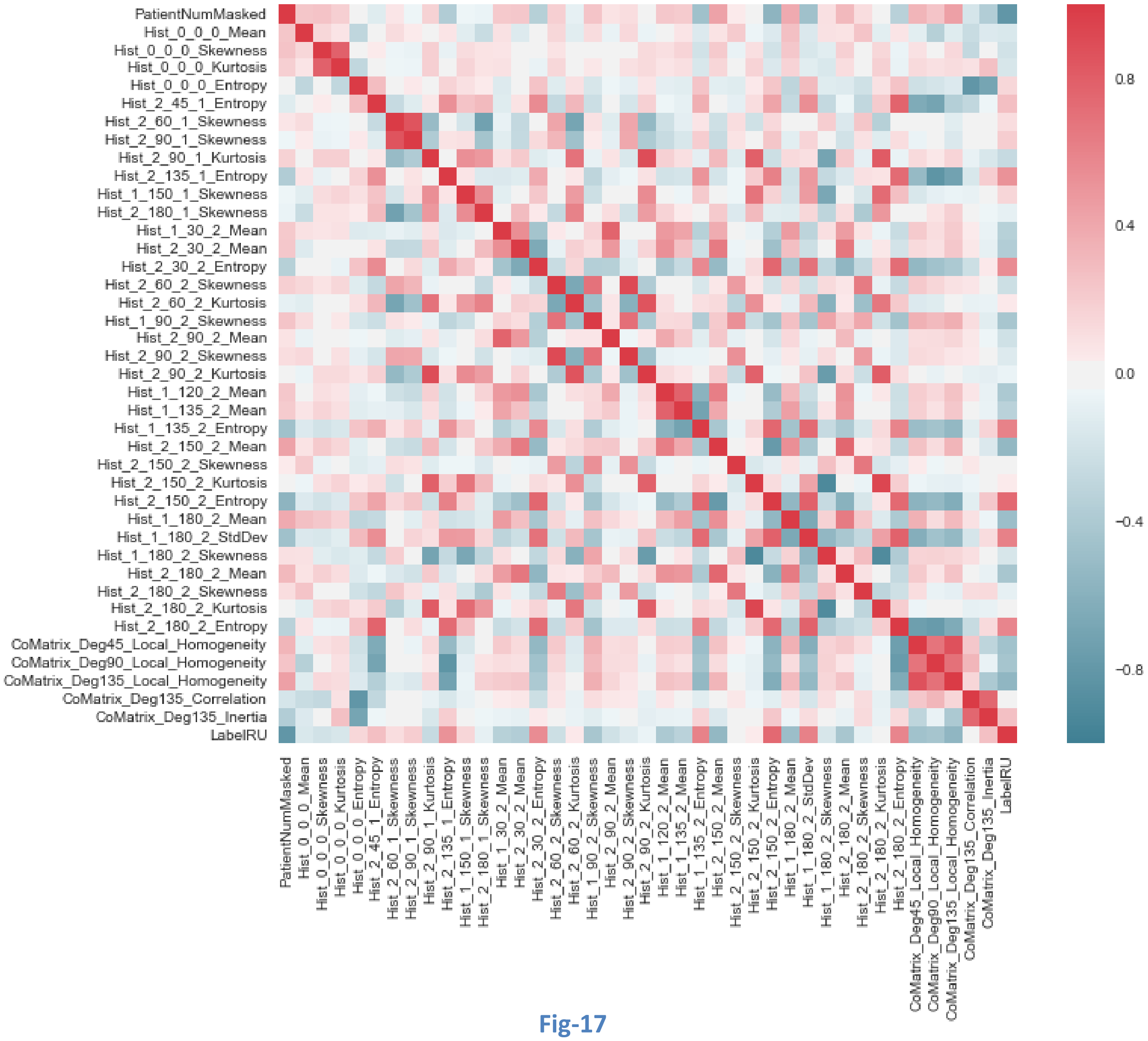


Fig-17

We can see that there are variables highly correlated with each other. The darkest red squares indicate high correlation.

Example, Hist_2_60_2_Skewness is highly correlated with Hist_2_90_2_Skewness and Hist_2_150_2_Kurtosis is highly correlated with Hist_2_180_2_Kurtosis.

On the other hand, CoMatrix_Deg135_Correlation is least correlated with Hist_0_0_0_Entropy.

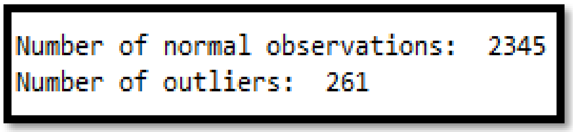
3.6 Check for Outliers

One of the many methods of **detecting outliers is by** using **IsolationForest.**

The IsolationForest 'isolates' observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

Since recursive partitioning can be represented by a tree structure, the number of splitting's required to isolate a sample is equivalent to the path length from the root node to the terminating node.

Output of the function is an array which has the value -1 for outliers and 1 for normal samples. In our consolidated data-set, we find that there are **261 samples which are considered outliers of the total of 2606.**



```
Number of normal observations:  2345
Number of outliers:  261
```

Fig -18

As the consolidated data-set has patient information for all six zones, we can even find out the number of outliers in each zone.

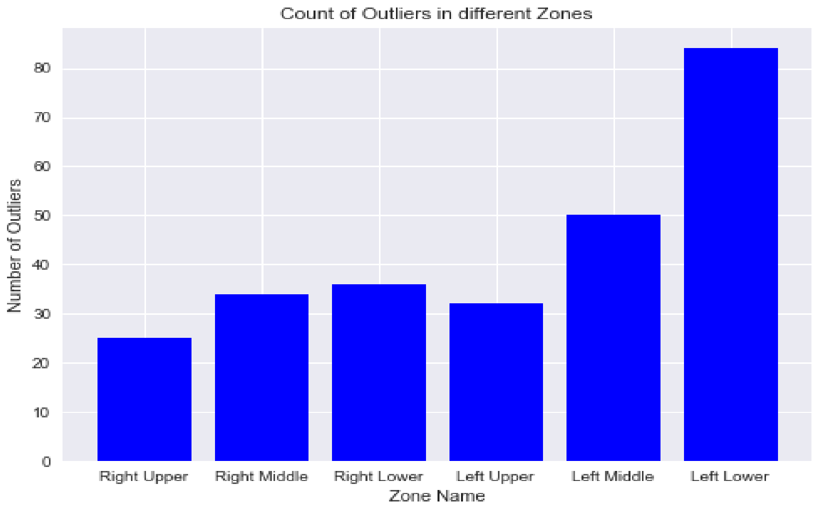


Fig -19

Left-Lower zone has 84, the most number of outliers; the Right Upper zone has 25, the least number of outliers.

Of the 261 outliers, there are 116 unique samples or Patients whose data is considered an outlier. Some **common causes for outliers are:**

- **Sensor/other equipment malfunction**
- **Wrong data entry**

Upon examination of the values of various outliers, it is found that the relationships we previously found/inferred do not change after the exclusion of outliers.

Model accuracy with and without outliers will be shown in section 5. We observe that there is no drastic improvement in accuracy when outliers are removed from our data set. Hence, we keep them in analysis.

4. Feature Selection

Redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model.

Feature selection helps in choosing some features of the available ones that will give us as good or **better accuracy whilst requiring less data.**

Having fewer attributes is desirable because it reduces the complexity of the model, and a simpler model is simpler to understand and explain.

One of the many techniques to perform feature selection is by using the **Extremely Randomized Trees method. This method will reduce the dimensionality of our data-set resulting in fewer features.**

Extremely Randomized Trees or Extra Trees Classifier is like Random Forests except that the splits it performs on the data are also randomized (instead of the best possible split as in Random Forests).

We must study the variation in Accuracy levels in 4 cases:

- Accuracy before feature selection on data with outliers
- Accuracy after feature selection on data with outliers
- Accuracy before feature selection on data without outliers
- Accuracy after feature selection on data without outliers

Below table captures the observations made.

	Accuracy Before Feature Selection	Number of Features Selected	Accuracy After Feature Selection
Data with Outliers	91.94 %	13	91.81 %
Data without Outliers	91.60 %	13	90.56 %

Fig - 20

We observe that after feature selection, with a simpler model we obtain a slight improvement in accuracy despite the data including outliers.

When outliers are dropped, there is only a slight dip in Accuracy levels before and after feature selection. Hence, we will retain outliers in our dataset.

Fig - 20 is the graphical representation of the importance's of all features in our data-set assigned by the Extra Trees classifier.

X axis – Feature Name
Y axis – Feature Importance

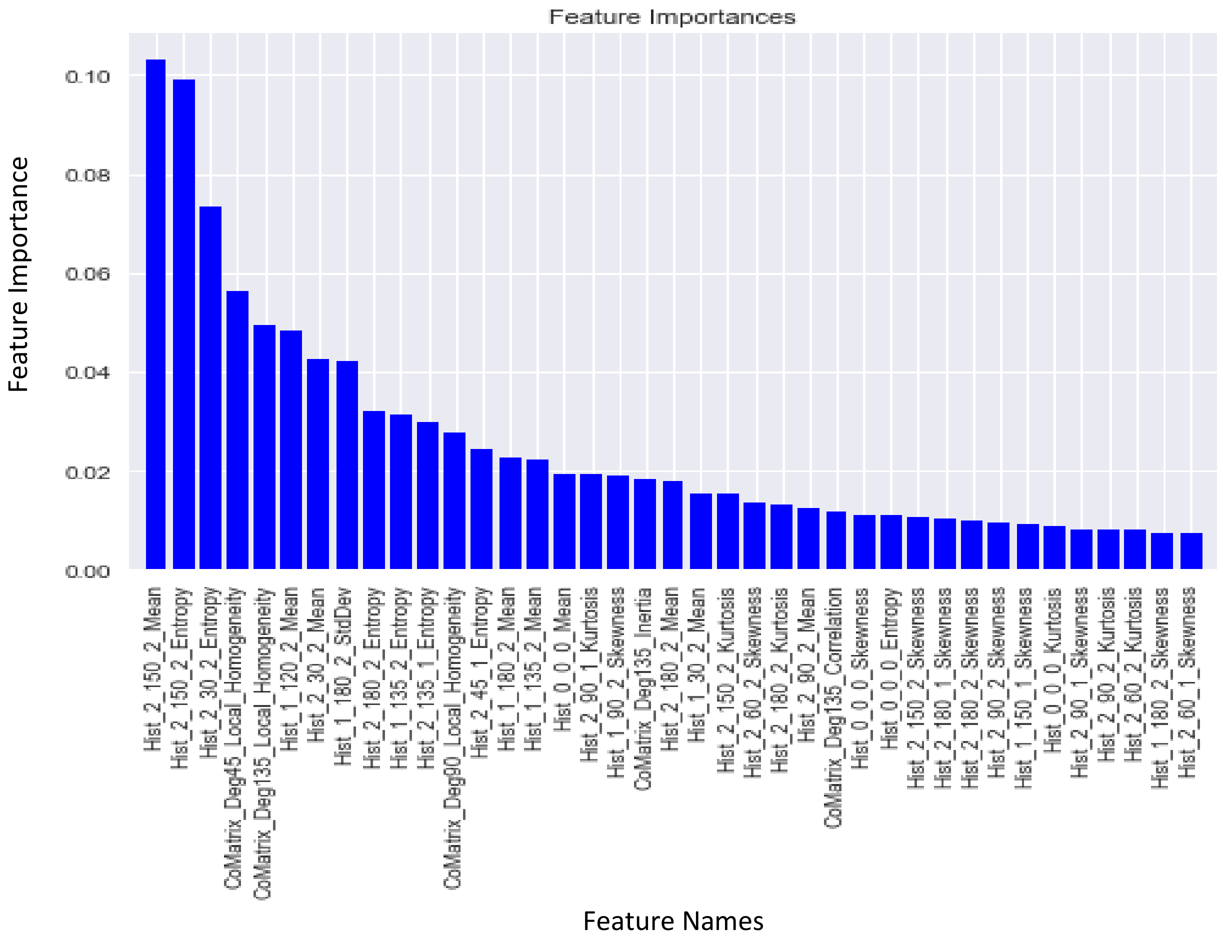


Fig - 21

Mean of the image histogram taken at 150degrees, after applying the 2nd filter: **Hist_2_150_2_Mean is of highest importance and Hist_2_60_1_Skewness is of lowest importance.**

An important observation that confirms our hypothesis is that **most of the skewness features seem to score low on importance and Mean and Entropy features are of highest importance for prediction.**

5. Model Selection

Classification algorithms **Logistic Regression, Extremely Randomized Trees** or Extra Trees Classifier, **K nearest neighbors and Support Vector Machine**s are a **mixture of linear (logistic regression) and non-linear algorithms (Extra Trees, K nearest neighbors and SVM)**

Among these classification methods, we'll choose the one that gives us better accuracy. The dataset is split randomly such that 30% of the data (781 observations of 2606) is marked as test data and remaining is the training data.

We **use leave one out cross validation method** and the mean of all cross-validation accuracy scores is taken to determine model accuracy over the validation set. This step is repeated for all four models using the 4 different algorithms mentioned earlier to get their mean accuracy scores.

Plot below depicts the accuracy scores of all 4 models.

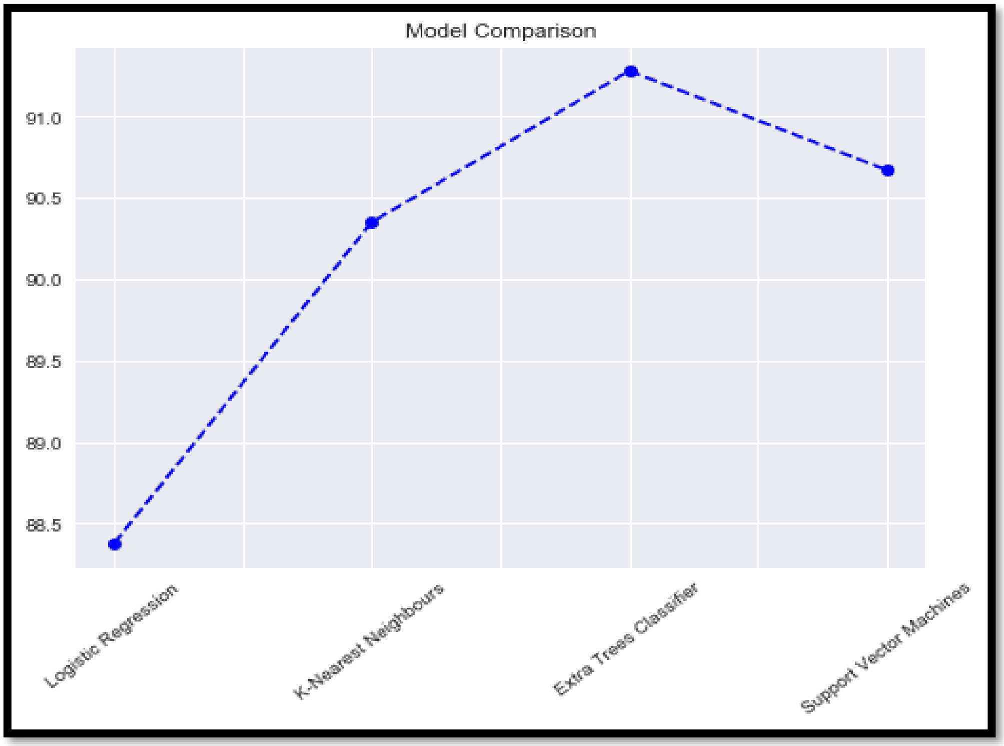


Fig - 22

We can see that **Logistic Regression performs poorly of all algorithms (88.37%) as the data or classes may not be entirely linearly separable.**

The Extra Trees Classifier gives us the highest accuracy of 91.28%. Hence, we'll use this algorithm to make predictions on our test set.

6. Results and Conclusions

The approach to be followed can be pictorially depicted as:

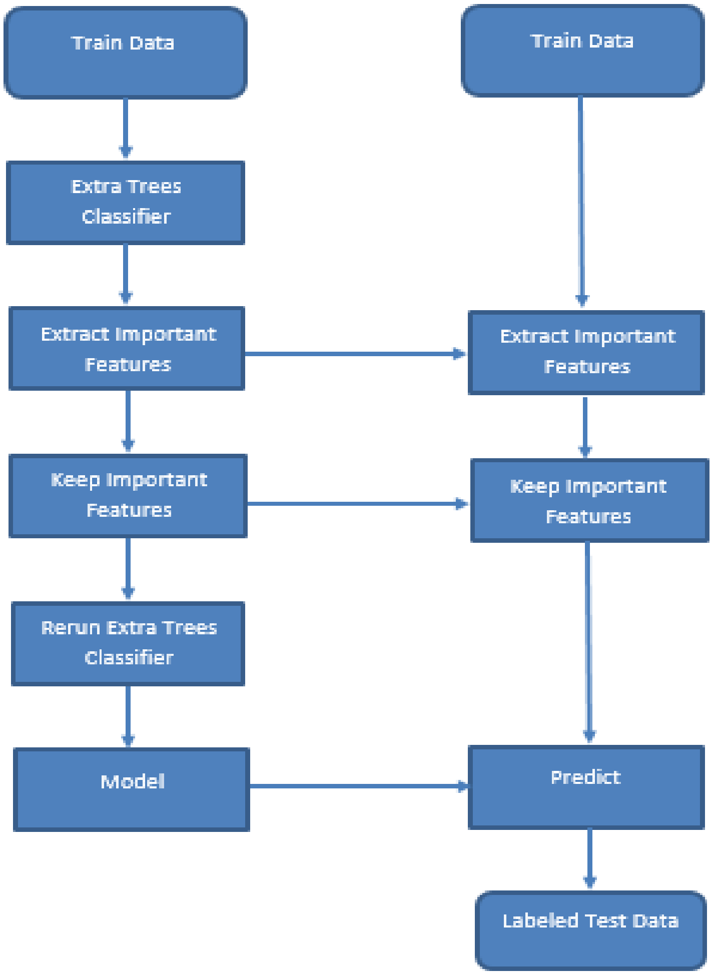
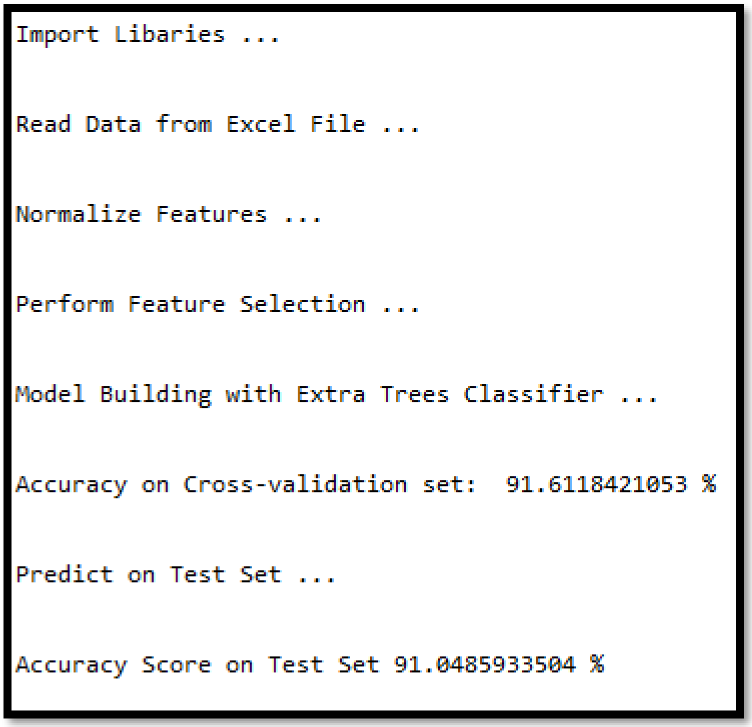


Fig - 23

We make use of the Extra Trees Classifier Algorithm to predict if a Patient has Pneumoconiosis on the consolidated dataset (dataset obtained by combining data from all six zones).

The results of our analytic are as shown in Fig-25.



```
Import Libaries ...

Read Data from Excel File ...

Normalize Features ...

Perform Feature Selection ...

Model Building with Extra Trees Classifier ...

Accuracy on Cross-validation set:  91.6118421053 %

Predict on Test Set ...

Accuracy Score on Test Set 91.0485933504 %
```

Fig -24

We see that the **model accuracy is 91.04%** on the Test set and 91.61% on the Training set.

The model is characterized by low bias and low variance as it generalizes well on both Train and Test sets. Hence no further regularization measures are taken.

Confusion matrix, precision and recall metrics for this model are given below.

Total = 782		Prediction	
		Positive	Negative
Actual	Positive	212	51
	Negative	19	500

Fig -25

Form the matrix:

True Positives = 212
True Negatives = 500
False Positives = 19
False Negatives = 51

A high precision value indicates a low rate of false positives. Precision of a model as calculated as follows:

$$Precision = \frac{|\{i \mid y_i = \hat{y}_i,\ \hat{y}_i = 1\}|}{|\{i \mid \hat{y}_i = 1\}|}$$

Calculated Precision= 0.92

Recall, indicates rate of false negatives. A high value of Recall indicates low rate of False positives. It is calculated as follows:

$$Recall = \frac{|\{i \mid y_i = \hat{y}_i,\ y_i = 1\}|}{|\{i \mid y_i = 1\}|}$$

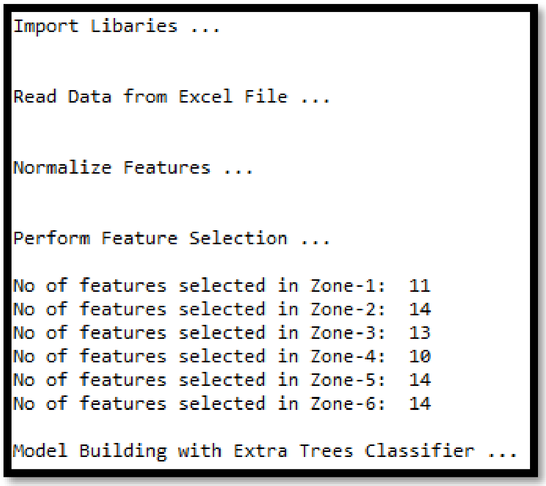
Calculated Recall = 0.81

From our confusion matrix, it is seen that the rate of False negatives is higher than number of False positives, therefore justifying the Recall value being slightly lower than Precision.

6.1 Results with one model for each Zone

Results/predictions seen so far were obtained using a Model that was trained on the consolidated dataset.

But, what happens when we **train 6 different Models (one model for each zone) and combine the zonal results/predictions based on PatientNumMasked to obtain a final Label**? Upon taking this approach, we observe the below results:



```
Import Libaries ...


Read Data from Excel File ...


Normalize Features ...


Perform Feature Selection ...

No of features selected in Zone-1:  11
No of features selected in Zone-2:  14
No of features selected in Zone-3:  13
No of features selected in Zone-4:  10
No of features selected in Zone-5:  14
No of features selected in Zone-6:  14

Model Building with Extra Trees Classifier ...
```

Fig - 26

Feature selection is performed separately, for each zone. This results in a **different number of features selected in each zone depending on the most important features required for optimal classification on each zone.**

We then move on to splitting data in each zone into training and test sets. Go on to perform Leave one out cross validation and prediction to observe the below results:

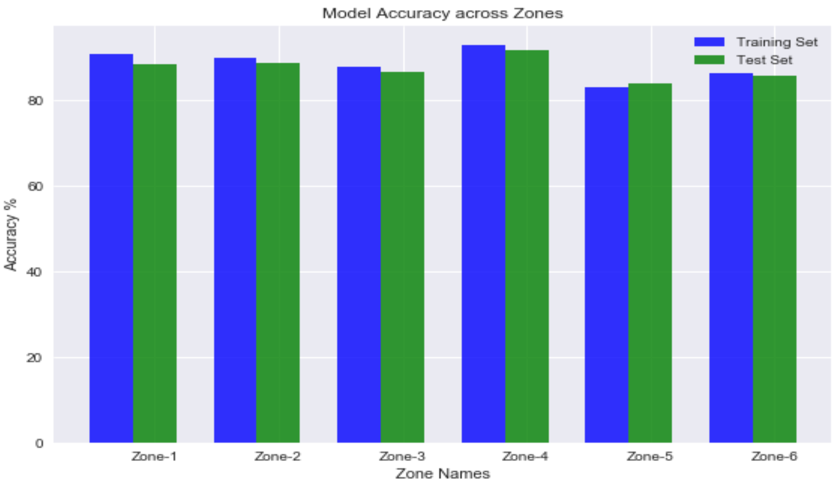


Fig -27

We see that the accuracy of the Zone-4 model is the highest and that of Zone-5 is the lowest. Zonal predictions: y1, y2, y3, y4, y5 and y6 are combined on PatientNumMasked (Patient Identifier) to calculate final prediction as:

$$y_i = max_j\left(y_{ij}\right)$$

Where $y_{ij} \in \{0,1\}$ represent the zone-level labels (1=Pneumoconiosis, 0=healthy)

The average accuracy obtained across all zones is 87.37%. Results obtained are shown in tabulated form for easier comparison with the results of model built on the consolidated data-set.

Metrics	Model on Consolidated Dataset	Zonal Model
Model Accuracy	91.04%	87.37%
Precision	0.917	0.867
Recall	0.806	0.8125

Fig -28

We see that the zonal models result in a lower accuracy than we obtained from the model built on the consolidated dataset. Also observed are comparatively lower values of Precision and Recall values indicate a higher ratio of false positives and false negatives respectively in comparison with the model built on the consolidated dataset.

7. Possible Problems and Next Steps

- A possible issue with our approach might be that we have used algorithms with default settings, we can look at tuning algorithm parameters to optimize model performance.
- Another classification algorithm we can experiment with is the Neural Network
- Further steps to find insights may be to perform multivariate analysis

8. References

1. ML What/How: http://scikit-learn.org/stable/tutorial/basic/tutorial.html
2. Confounding: https://aneeshaasc.tumblr.com/post/162939009733/confounders-thou-shall-not-escape
3. Logit: https://aneeshaasc.tumblr.com/post/163241507273/provedisprove-hypothesis-logistic-regression
4. Plotting feature importance: https://aneeshaasc.tumblr.com/post/164214738793/run-forest-run
5. Data management and Visualization: https://www.coursera.org/learn/data-visualization/home/welcome
6. The above are links to my technical blog that has articles written on my work with other datasets done as a part of MOOC: https://www.coursera.org/learn/regression-modeling-practice/home/welcome
7. For outlier detection: http://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html#sphx-glr-auto-examples-ensemble-plot-isolation-forest-py
8. For Feature Selection: http://scikit-learn.org/stable/modules/feature_selection.html
   https://chrisalbon.com/machine-learning/feature_selection_using_random_forest.html
9. Extra-Trees Classifier:
   http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
10. Confusion matrix:
    http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
11. Cross-Val score:
    http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html
12. Calculating Accuracy score:
    http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html