PNEUMOCONIOSIS CASE STUDY

GE User

GE [Company address]

Table of Contents

1. Introduction

A leading hospital wishes to develop a screening program for coal miners, to facilitate early detection of Pneumoconiosis.

A team of image analysts have already developed algorithms to segment the lung and divide it into three zones. This for a set of images where the doctor's labeling for the lung zones is known, and characterized each lung zone in terms of a set of features. Each patient is identified by a unique patient number in the attached Excel spreadsheet. We need to develop a predictor model of the abnormality either at each zone and segment level or at the combined all the six observations for each patient as single row label as normal or abnormal.

2. Methods:

Build a model (or a set of models) on the data for $P - P_i$. We build model zone-level models whose predictions are then combined. Since

Predict the label for P_i. Part of the challenge will be to figure out whether to predict individual labels for each $\vec{x}_{ij}$ and then combine them. Let the predicted label be $\hat{y}_i$.As I = value in each of the zones is not uniform and Px is not represented in all the regions so if we merge in the single matrix we may lose data clarity also as each zone data has different behavior.

With this reason, we have taken approach of prediction byour problem is divided into parts.

1. Chest level classification

The weighted voting is a useful measure widely used. Herein, the weighting factors are calculated for the six regions. Each region votes for the region to be normal or abnormal. The final probability of an image being abnormal based on the six regions is given by

$$Prob = \sum_{i=1}^{N} Wi\ Pi$$

where Pr ob denotes the final probability, Pi, which ranges from 0 to 1, denotes the prediction result of the ith region, N is the number of regions, herein N = 6, and Wi is the weighting factor for region i determined by the AUC value of each region obtained in the training phase given by

$$Wi = \frac{Avi}{\sum Avi}$$

2. Zone Level Classification

As the image data is given each zone level with the data of patient is not same at each level as depicted in fig-1 down we must arrive at the zone level model accuracy and aggregate it into chest level. So at the each zone level we need to find the better classifier to predict zone level. More details for taking this approach has been explained in the Data understanding section of the data mining lifecycle.

To refine, analyze, model and to develop the model we would use CRoss Industry Standard Process for Data Mining (**CRISP -DM**). As part of this any data analysis project can be divided into following distinctive Steps:

1. Business Understanding:

A typical doctor's report divides each lung into three zones (upper, middle and lower) and labels them as normal/abnormal. However, due to the lack of trained doctors with expertise and the large number of patients to be screened, they have requested GE to develop a computer-aided detection system. So, if there is predictable system which can give the decision on presence of the disease large number of cases can be screened in short span of time. So, we needed to develop a model which can predict higher degree of accuracy classifier model

2. Data Understanding

The feature data for the six lung zones for each patient, along with the zone label (0=Normal, 1=Abnormal). Thus, a total of 40 features for each lung zone has been provided. The first column in each worksheet (one sheet per zone) gives the patient number, while the last column gives the label.

There are two categories of the features on which this decision is based on:

Intensity based We have a set of 6 features based on the histogram of intensity values – mean, standard deviation, skewness, kurtosis, energy and entropy. 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 $Hist\_d\_\theta$.

Co-occurrence matrix based: We have 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. 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.

Data has been given with six different samples with Right and Left segment divided into upper, middle and lower zones.

Fig-1 showing number of data at each of the zone level.

Explorative data Analysis:

1. If we look at the data of each zone we find that all the parameters are same across zone number of data points are different in each of the zone. On further exploration, we find out that same patient's data may not be available each zone. (as in Fig-1)

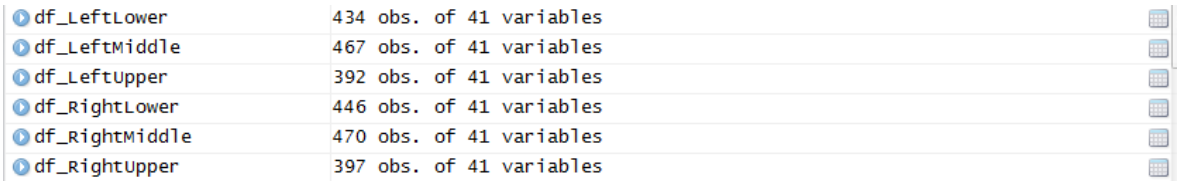

Figure 1 Data ranges in each Zone

We can see all the columns are having numerical variables so this makes our model easy to implement.

2. If we explore the distribution of data for Intensity based and co-occurrence feature

distribution we find out that data is not normalized for many features of intensity based features as shown in fig-2.

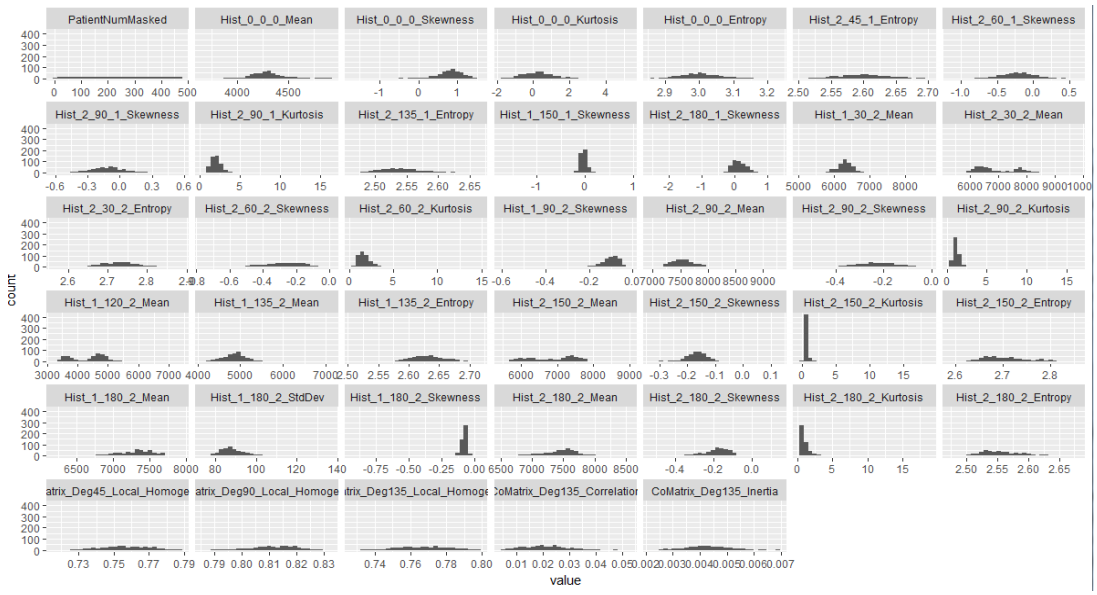

Figure 2 on Right Upper zone various parameters data distribution with intensity based and co-occurrence based feature (similar distribution is observed across all the zones)

Fig-2 on Right Upper zone various parameters data distribution with intensity based and co-occurrence based feature (similar distribution is observed across all the zones)

3. We also notice that scaling is not same for co-occurrence and intensity nature of the data as shown fig-2 above.
4. We check if data is consistent or any of the missing data for feature and found that data is consistent across all the six zones.
5. If we want to check the correlation interaction with response variable (in our case "Label") and explanatory variable by drawing correlation is very high between few of the explanatory variable which may bias the result as shown in the correlation fig-3 below. We can see there are many blue dots depicting high correlation.

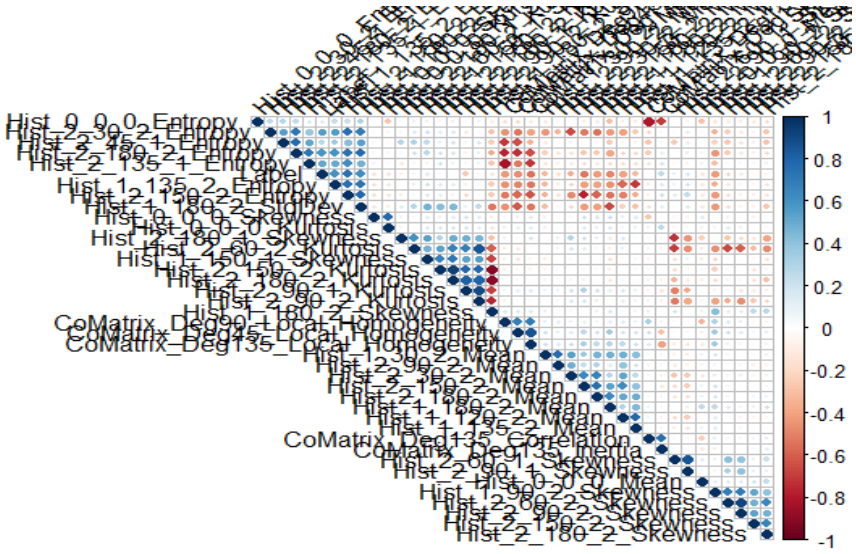

Figure 3-Correlation of the variables

6. One more observation on the response variable is of the binary value (if patient is having disease or not).
7. Since data points are different in each region and for deciding the response variable by combining all the zones is not viable option. If we combine there are two issues

   Since for the missing patient data in one zone we are required impute or remove that data for available zone. In the first case, there is not a solid a replacement which would impact nature of the data and behavior and in the second case, we are left with very less data to take the decision.

So, best option left with us is to predict individual labels for each $\vec{x}_{ij}$ and then combine them in some fashion each zone separately predicts and combine prediction as one single result.

Data Preparation:

1. Normalize the data for the skewed variables for 1st Quartile and third quartile + 1.5 times the inter quartile range.
2. Scale the variables so they are on the single uniform scale.
3. Patient id is insignificant variable for the predicted variable so we can remove that column.
4. We have 38 variables predicting the response which is way too high. As we have seen in the Explorative Analysis all variables may not be significantly impacting response, and we are interested in the prediction as per the problem definition and all the independent variables are numeric, so we can use Principal Component Analysis (PCA) for dimension variability.
   This analysis helps us to separate significant vector components on their variability from insignificant one.

We can do the PCA for all the regions for e.g. for right upper zone as shown in the Figure.4 first Principal component explains 20% of the variance and next one 18% so we can see as we go till 30th component variance explained is less than 1%.

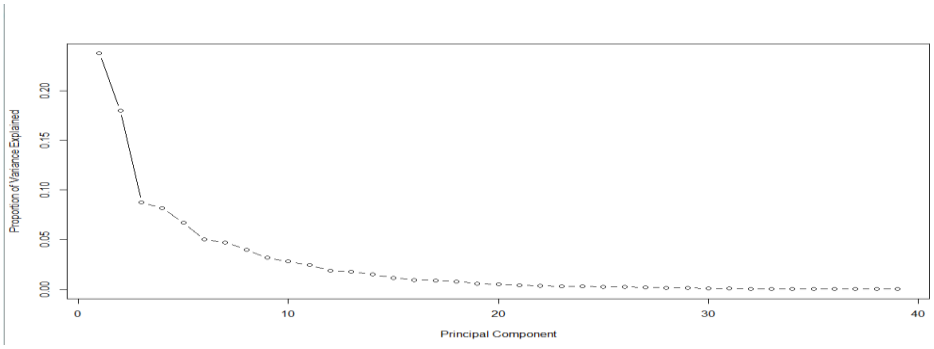

Figure 4 Variance explained vs principal component

If we cumulate the variance explained as shown in fig-6 98% of the variance is explained by first 32 component so rest 6 components can be less significant for results so in the right upper zone. we can build model with 32 principal components as shown in the Figure.5 . Similarly 28 an32, 33,28,26 qand 25 are applicable for the right middle , right lower , left upper , left middle and lower zones respectively.

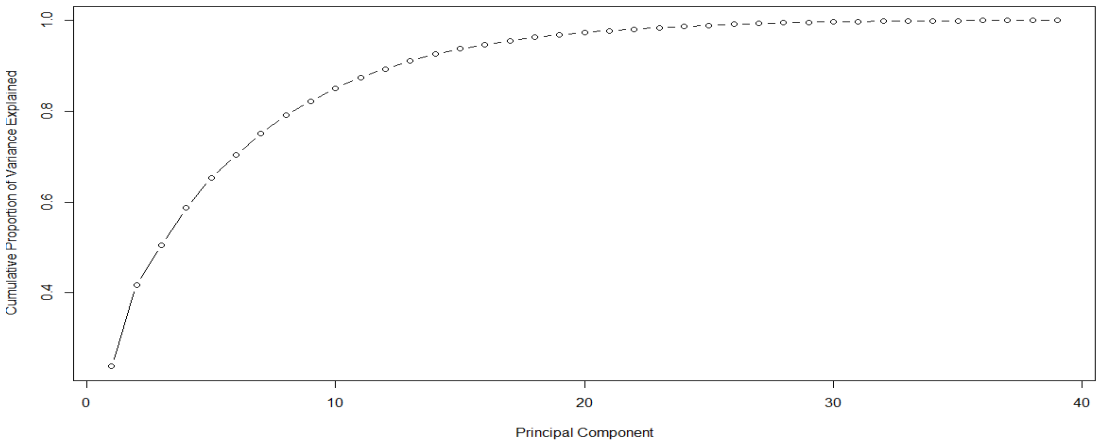

Figure 5 principal component vs cumulative proportion of variance explained.

Modeling:

Now we can see that data is ready for the modelling and data in each of regions is modeled using the first example of the model. If we look at the problem at hand we are needed to classify each patient zonal data into either having disease (i.e. equal to 1) or not having disease (i.e. equal to 0). Since it is guided by the supporting data for this classification, our modelling can be categorized as Supervised Classification type.

For the supervised classification type of problem following models are more suitable:

- Naïve Bayes
- Support Vector Machines(SVM)
- Random Forests
- Artificial Neural Network(ANN)

Now we would take right upper zone as an example and go through the modelling steps and check the suitability of each model. We divide the principal component data of the first 32 components into train and test. By doing so it allows us to train our model on train data and validate our model on the small set of data to check the more near real time efficiency and qualify.
We may use the validation data in SVM and ANN by diving into samples while training data itself explained in each of the model down.

3. Analysis

Naïve Bayes:

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. So, if I need differentiate the data between disease presence or not present we can use receiver operating characteristic curve, i.e. ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. So basically

We divide the data into area under AUC curve is "0.74" and the accuracy of 83.3%. and Sensitivity: 0.9647 and Specificity: 0.5143. we are looking for the equal specificity and sensitivity we need to check if we can have better differentiator model. If we draw 50% accuracy model we can see curve is very near to it at the high false rate, which is depicted in lower specificity as shown in the fig -6

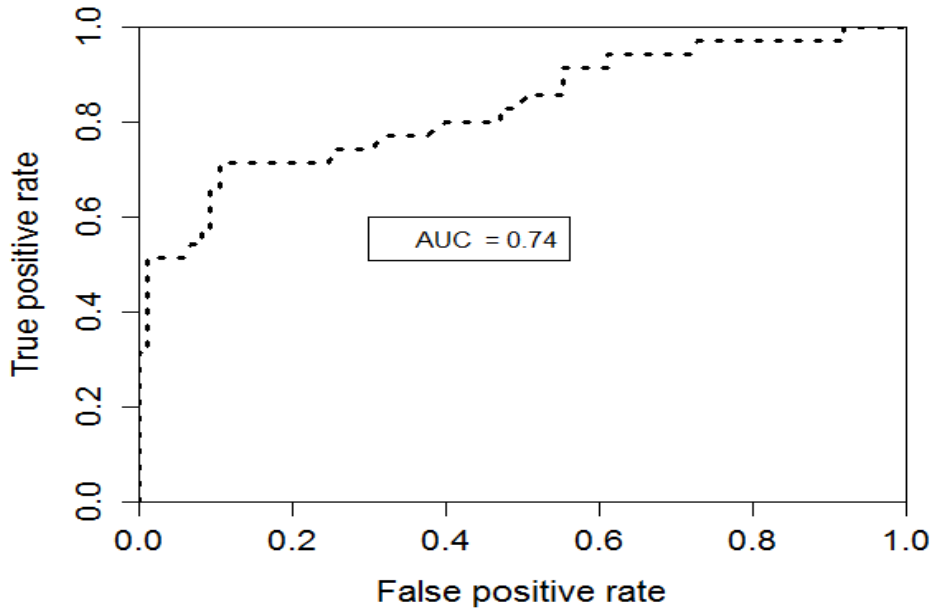

Figure 6 ROC Curve for Naive Bayes

Support Vector Machine (SVM):

Support Vector Machine is one more elegant classifies which helps to classify by drawing a separator Support plan for the multivariate classifier, this come with two types of classifiers mainly linear and Radial which are different from how do they calculate the mean distance and go on iterative mode of classification, we can look at which method would be more suitable for 5 fold cross validation.

If we look at the value distribution box plot for the line at and radial classifier we can observe that radial cv=5 is more reliable as shown in the fig-7

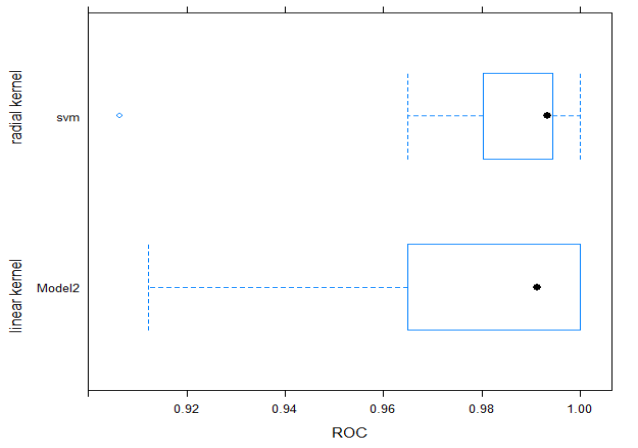

Figure 7 classification of the radial and linear classified value box plot

Fig-7 classification of the radial and linear classified value box plot

So for the cross validation of 5 and radial mode if we look for the tuning different parameters of this model as degree=3, sigma = 0.015, gamma=0.0025, C = 5, we get the accuracy of Accuracy- .85 , Sensitivity : 0.9647, Specificity : 0.5714 so corresponding AUC (Area Under Curve) which is measure of model's capacity to differentiate has improved marginally to 0.77 as shown in the Figure.8

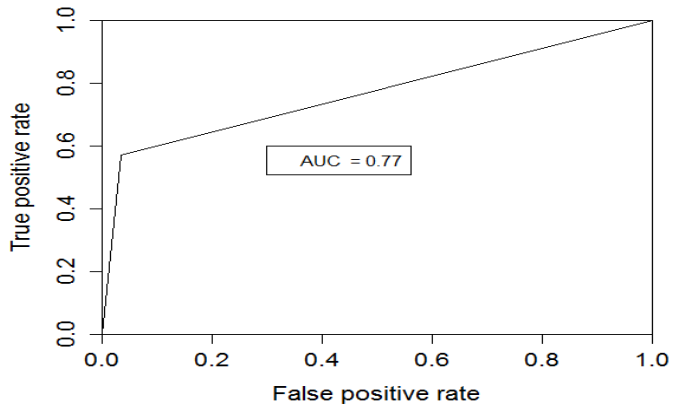

Figure 8 Curve for SVM which is True Positive Rate vs False Positive Rate

So, this is quite evidential that though there is increase in the sensitivity there is not much improvement in the specificity and AUC lesser than 90% this model is not a good differentiator.

Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification). On this model if use the tuning or training of the parameters, we arrive at No. of variables tried at each split: 17 among Number of trees: 500.

So at this optimized performance we get an Accuracy: 84.1%, Sensitivity: 0.9765 and specificity : 0.5143 and an AUC (Area Under Curve which is diagnostic differentiator parameter for the test as 0.83(as shown in the Fig-9) which is little below the industrial practice of the good classifier AUC of 0.9. This can be even seen that specificity 51% making curve leaning towards to middle line for higher value of False Positive rate of just more than 0.6.

So, this points to the fact that our better model search is not ended so far.

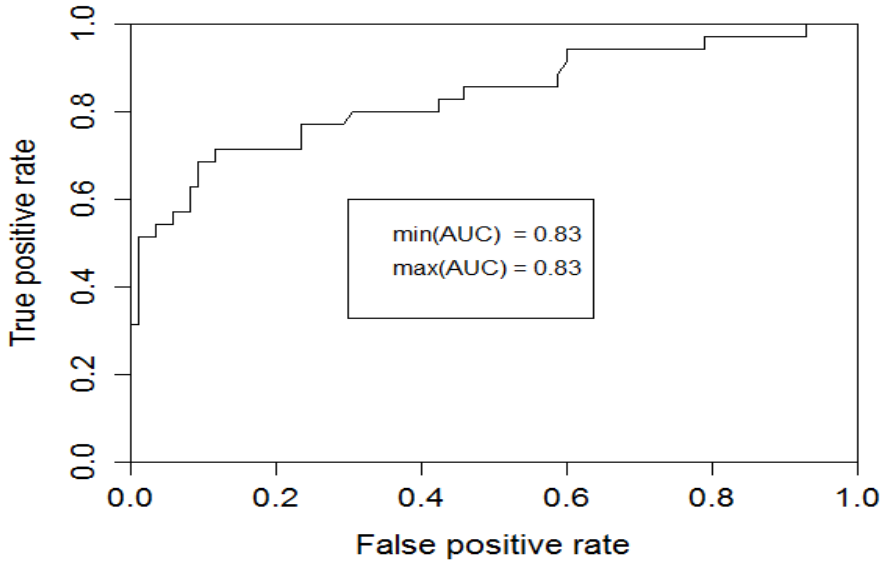

Figure 9Curve for Random Forest which is True Positive Rate vs False Positive Rate

Neural Network:

Search of the model of higher accuracy brought us to the Neural Network which gives higher efficiency. A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally, the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the pneumococcus classification problem at hand.

For tuning the model for maximum AUC (area under cover) above 90% in our model we have used activation function mainly Maxout, rectifier, RectifierWithDropout and maxoutWithDropout as this is binomial classification problem, with the hidden layers 50,20 and 10 epochs= 100 and cross validation of 5.

We arrived at the AUC of 0.999816 and Accuracy- 94.58% specificity and Sensitivity as high as 0.99 so this one looks to be better model for all the six-region's data set as shown in the Fig -10.

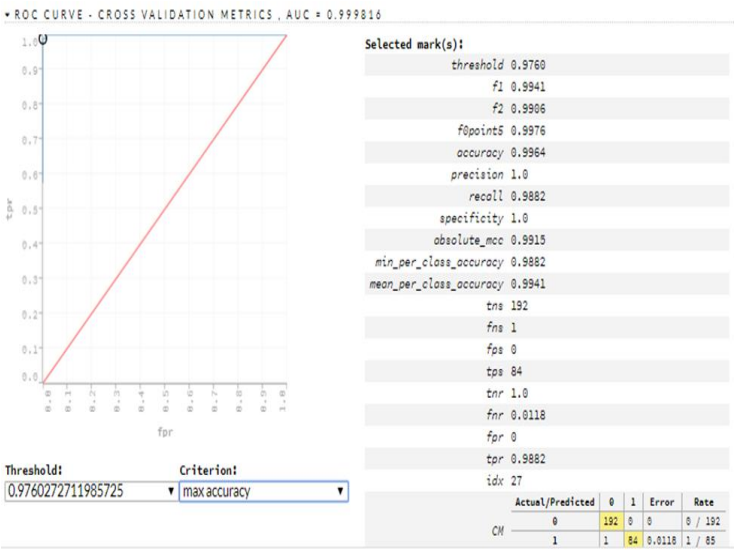

Figure 10 Right Upper zone ROC and statistics

Now if we look at the cross validation quality by looking at the lift and gain as shown in the graph of cumulative lift and gain we can see inflectuation point is between 0.2 and 0.3 which indicates well differentiated model at the crossvalidation of 5 groups as shown in the Fig-11.

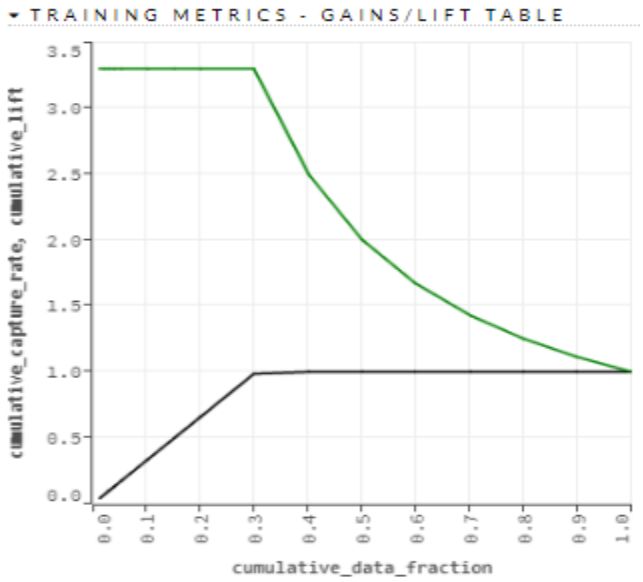

Figure 11- Gains/Lift graph for the Right Upper zone

So across 6 zones we can sif we look for tuning of highest Area Under Curve weare getting AUC of 0.97- 0.99 and proabality of prediction within the range of 0.94-0.98 which is quite good accuracy for the model as shown in the fig -12 and fig -13. If we look for all the companents instead of focussing we can increase of accuracy to the tune of 0.001 as shownin the table 1 and table-2.Also we notice that there

is good specificity and sensitivity across all the six regions in the range of 0.97 – 0.99 . so we have arrived at the good optimised model.

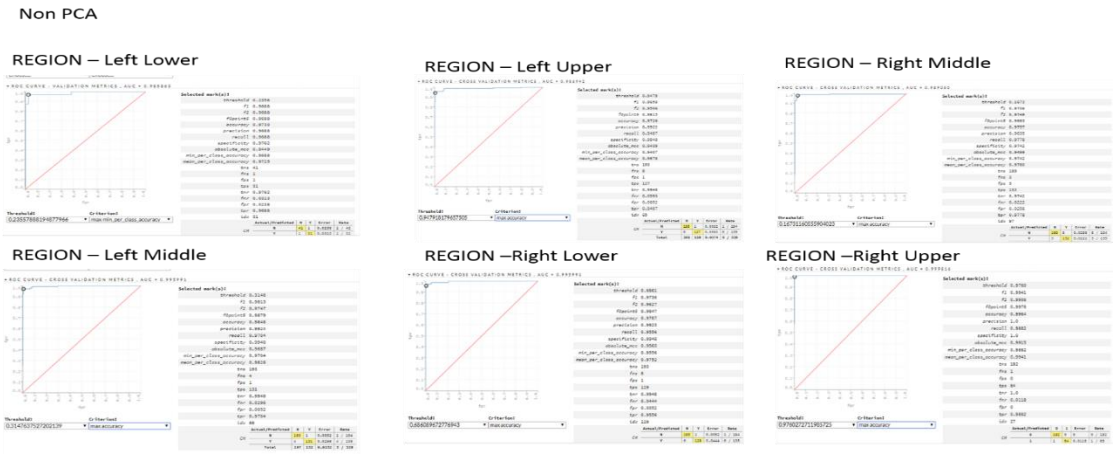

Figure 12 All the six zones ROC and statistics for test data for Non - PCA data

Figure 13 PCA data ROC curve and statistics

	PCA - Model					
	Right Upper	Right Middle	Right Lower	Left Upper	Left Middle	Left Lower
Probability	0.9718	0.9736	0.984	0.9964	0.9877	0.9701
Specificity	0.999	0.9845	0.9894	0.99902	0.9947	0.9773
Sensitivity	0.999	0.9722	0.9836	0.99905	0.9927	0.9565

Table 1 PCA Model Probability, Specificity and Sensitivity

	NON PCA - Model					
	Right Upper	Right Middle	Right Lower	Left Upper	Left Middle	Left Lower
Probability	0.9964	0.9757	0.9787	0.9726	0.9848	0.9726
Specificity	0.999	0.9742	0.9948	0.9948	0.9948	0.9922

Table 2 NON- PCA Model Probability, Specificity and Sensitivity

Using the AUC value and finding the cumulative AUC value across all the six zones we can arrive weightage factor for probability predictability on each zone using the equation-1 as shown in the Table-3 below.

Equation 1- weightage calculation at each zone level

$$Wi = \frac{Avi}{\sum Avi}$$

Weightage		
Right Upper	W_{RU}	0.1625436
Right Middle	W_{RM}	0.1678164
Right Lower	W_{RL}	0.1670742
Left Upper	W_{LU}	0.1677065
Left Middle	W_{LM}	0.1677967
Left Lower	W_{LL}	0.1670627

Table 3 Weightage calculated all the six zones

So, overall probability for the chest level classification by using the equation -2 using weighted average method. We get the accuracy of 98 % as shown in the table – 4

Equation 2 Overall proabality for chest level classification

$$Prob = \sum_{i=1}^{N} Wi\ Pi$$

Weightage (w)			Probability (p)	WXP
Right Upper	W_{RU}	0.1625	0.9718	0.1580
Right Middle	W_{RM}	0.1678	0.9736	0.1634
Right Lower	W_{RL}	0.1671	0.984	0.1644
Left Upper	W_{LU}	0.1677	0.9964	0.1671
Left Middle	W_{LM}	0.1678	0.9877	0.1657
Left Lower	W_{LL}	0.1671	0.9701	0.1621
			Total Proabability	0.98065

Table 4 Calculation of the total probability at the chest region for training data

4. Conclusion:

Weightage(w)			Probability(p)	W X P
Right Upper	W_{RU}	0.1625	0.9061	0.1473
Right Middle	W_{RM}	0.1678	0.9787	0.1642
Right Lower	W_{RL}	0.1671	0.9806	0.1638
Left Upper	W_{LU}	0.1677	0.9818	0.1646
Left Middle	W_{LM}	0.1678	0.9755	0.1637
Left Lower	W_{LL}	0.1671	0.9958	0.1664
			Total Proabability	0.970057

Table 5 Total probability Chest reason for test data

After going through all classifier models, we got the high accuracy in the Artificial Neural Network in the range of 98 % and better sensitivity and specificity in the range 0.97- 0.99. This model also high AUC which can differentiate with higher accuracy in the test and validation data range. So, accuracy is obtained is not an overfitting model.

By default, H2o uses F1-optimal threshold but as per the problem definition we needed maximum accuracy because, downside of guessing a person having a disease when in actual he is not having one is and for guessing not having disease when he is having disease are equal as in earlier case he would go through many invasive treatments, and in later case he may end up losing his life. So, when evaluating on the test data, threshold

would be calculated as the average of the threshold that has given maximum accuracy on train and validation data.

If you look at the Table5 an error of 0.03% for test data shows our data has low bias, and if you compare with the Table4 training set error of 0.02% we are having low variance model too. So, we arrived at low bias and low variance optimal model.

Further Steps

So, any of the further real-time has got changed behavior way different from test model and accuracy drops below 90% following are the further area for enhancing model:

1. We can use the ensemble models with Neural Network basically looking at the data if there is an interaction between group of independent models which can contribute to the model behavior.
2. If we can capture more amount of the data we can eliminate if there is model overfitted with multiple cross fits.
3. Our assumption of weightage average method to calculate the AUC weightage can be experimented with multiplication of weightage if there is inter play between region wise data, that means probability of I region identifying is exclusive and we want to categorize each zone differently.
4. We can look at the dimension reducing techniques like t-SNE is the very popular algorithm to extremely reduce the dimensionality of your data in order to visually present it. It is capable of mapping hundreds of dimensions to just 2 while preserving important data relationships, that is, when closer samples in the original space are closer in the reduced space. t-SNE works quite well for small and moderately sized real-world datasets like ours and does not require much tuning of its hyperparameters.