# Analytics for Engineers

## Case Study: Detecting Pneumoconiosis

**10/3/2013**

# Case Study: Detecting Pneumoconiosis

## 1 Introduction

A leading hospital wishes to develop a screening program for coal miners, in order to facilitate early detection of Pneumoconiosis. The standard detection procedure involves taking a chest x-ray and examining it for abnormalities that indicate the onset of Pneumoconiosis. A typical doctor's report divides each lung into three zones (upper, middle and lower) and labels them as normal/abnormal (0=Normal, 1=Abnormal).

This report outlines an automated screening program that will identify whether a particular patient is diagnosed as normal or not. The goal is to identify as many of the abnormal cases as possible, while keeping the false positives low.

The recommended approach "leave-one-patient-out cross-validation" is used to determine the correct model and report the results in terms of average performance across cross-validation samples. Alternative training/validation/testing approaches have also been utilized to validate the results from the recommended approach and improve up on it.

## 2 Feature description

A team of image analysts have already developed algorithms to segment the lung and divide it into three zones. They have done 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.

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 in order to describe each region of interest. These are described below:

1. **Intensity based** We extract a set of 6 features based on the histogram of intensity values – mean, standard deviation, skewness, kurtosis, energy and entropy. Apart from calculating these on the original ROI, we also extract these features after applying a difference filter on the image for the purpose of local enhancement. If $I(x, y)$ denotes the image gray value at $(x, y)$, the first and second order filters are defined as:

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

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

where $d$ is the difference scale and $\theta$ is the orientation at which the difference is computed. $f_x$ and $f_y$ represent the first order difference while $f_{xx}, f_{yy}, f_{xy}$ 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. **Co-occurrence matrix based**: Set of 5 features extracted 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 have been provided for this study. There are also columns on Patient ID and Label.

# 3 Methods:

## 3.1 Tools used for analysis:

The tools used for analysis are JMP [SAS JMP Project, 2013] for high level exploratory analysis and R [R Project, 2013] for detail analytical/statistical analysis and automation of final solution.

## 3.2 Abbreviations:

In this report the following abbreviations are used:

- LL: Left Lower
- LM: Left Middle
- LU: Left Upper
- RL: Right Lower
- RM: Right Middle
- RU: Right Upper
- SVM [SVM]: Support Vector Machine

## 3.3 Exploratory Analysis:

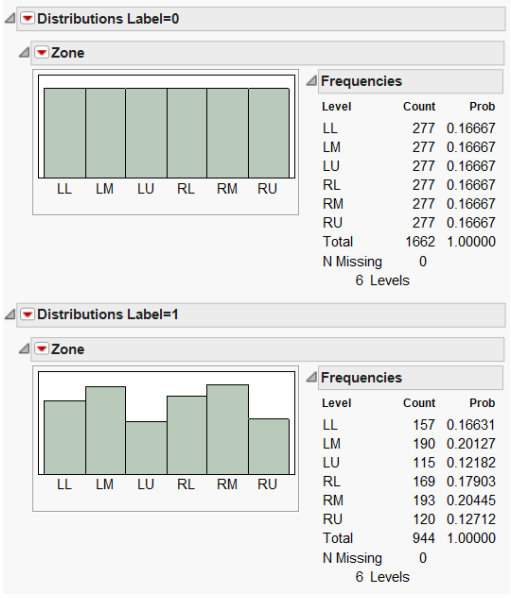
### 3.3.1 Data Summary

Data for this study downloaded from [AE. (2013)], and went through some exploratory analysis to

- Identify missing values (if any)
- Verify the quantity and quality of the data
- Determine the features to be used in the model.

There are 39 features in this data set, each represented as one column. Feature description section of this report has more detail on these features. No missing data (empty or NAN fields) was observed. There are total of 2606 records for all zones from which 434 are in LL, 467 in LM, 392 in LU, 446 in RL, 470 in RM, and 397 in RU.

As can be seen from Figure 1, number of reports on patients in each category is not the same. This means that some patients are missing data for some zones. On the other hand, there is only one report on each patient on any zone category. Number of patients reported as normal in each zone area is identical. But number of reported as abnormal is different in each zone. One reason for this is some patients missing information in some zones. Another reason is that same patients can show up as normal for some zones and abnormal for others. This will have direct effect on the statistical analysis results. The Histograms for label 1 indicates number of patients identified as abnormal are different for zones.




| Zone | Label | N |
|---|---|---|
| LL | 0 | 277 |
| | 1 | 157 |
| LM | 0 | 277 |
| | 1 | 190 |
| LU | 0 | 277 |
| | 1 | 115 |
| RL | 0 | 277 |
| | 1 | 169 |
| RM | 0 | 277 |
| | 1 | 193 |
| RU | 0 | 277 |
| | 1 | 120 |

| Zone | N |
|---|---|
| LL | 434 |
| LM | 467 |
| LU | 392 |
| RL | 446 |
| RM | 470 |
| RU | 397 |

*Figure1: Statistics on the Patient reports on each zone area. Number of patients with Abnormal patients is not the same in each data zone. This is due to Abnormal patient data missing from some zones. It can be seen that the number of patients diagnosed as normal is the same. But number of patients diagnosed as abnormal differ among the zones.*

### 3.3.2 Data Distributions - PCA analysis

To understand the distributions of data sets, principle component analysis is performed (PCA) on vectors $\vec{x}_{ij}, j \in \{1 \dots 6\}$ where 39 features in each data set represented as top 2 principle components that represent majority of data. As shown in Figure 2, data sets in group1 :{RL,LM,LL} and group2:{RU,RM,LU} each have similar trends within the group. Distributions are so similar that the colors for RU, RM, LU are so much overlapped that the color black for RU is covered by RM and LU. This information is critical when distributing the data consistently into training/validation/testing to reach maximum statistical analysis precision.

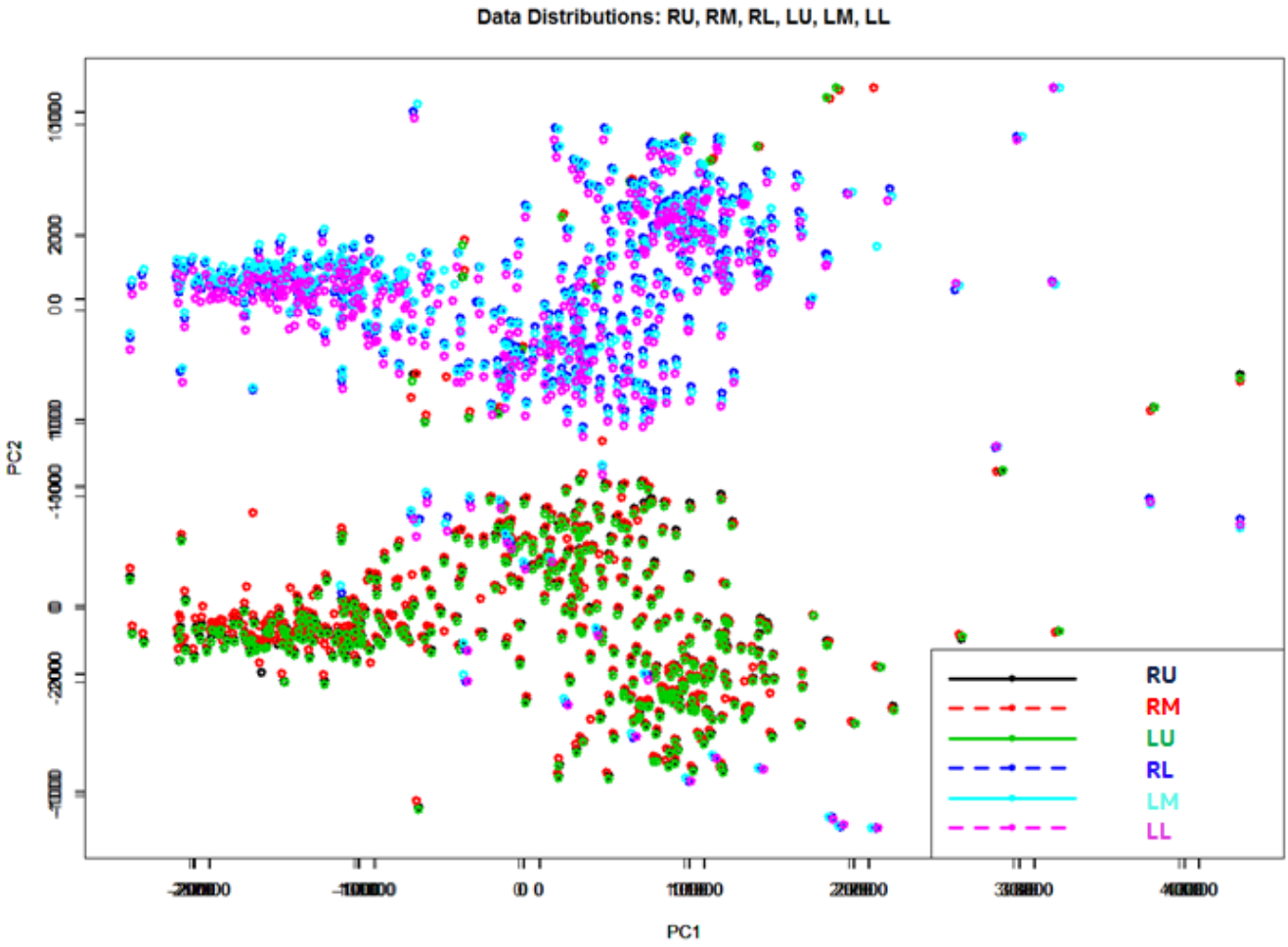



*Figure 2: Principle Component analysis on data (with all zones included) shows two major distributions that can be categorized as group1:{RU,RM,LU} and group2:{RL,LM,LL}. The top 2 most important components PC1 and PC2 are charted in this figure.*

### 3.3.3 Data Outliers

Analyzed features for possible outliers. For improvements in accuracy and precision, developed statistical modeling on data with outliers removed. Figure 3 visualizes some of the identified outliers in

the study. These outliers could be the result of the abnormal readings from scanners, or post processing of the data.

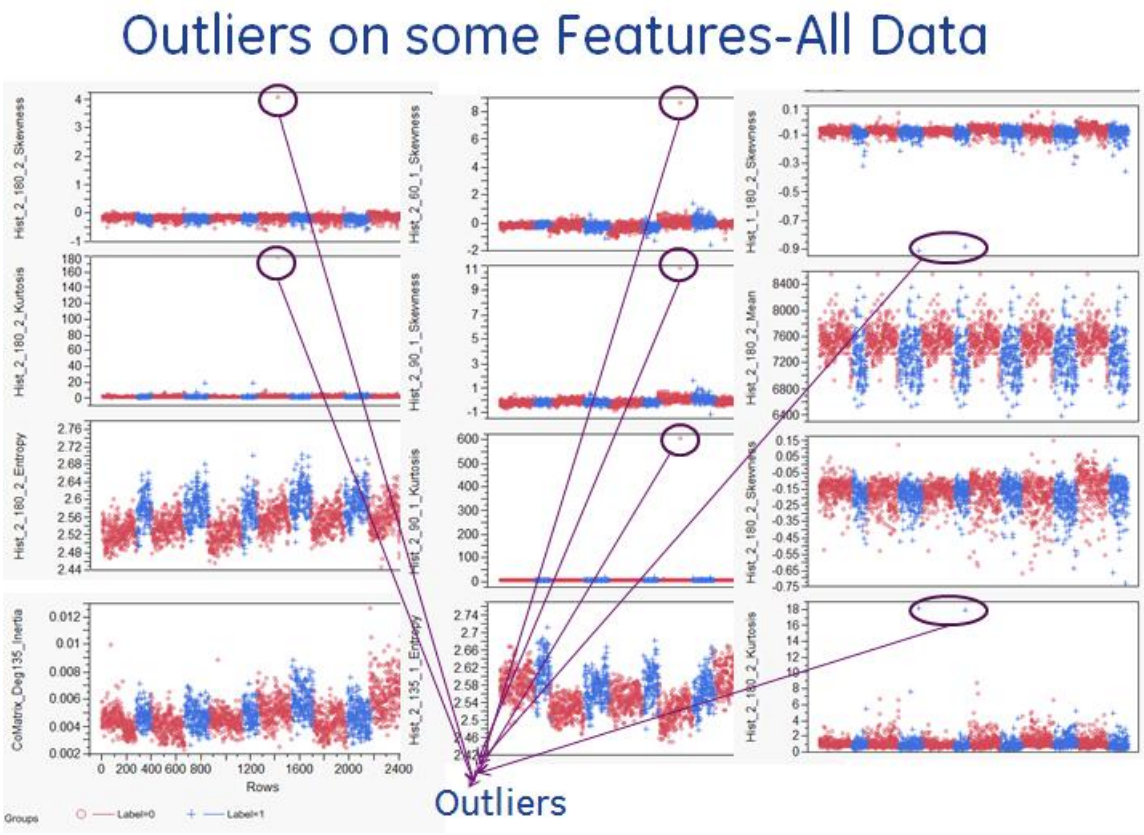



*Figure 3: Identifying outliers in features within circles. Each color band interval signifies a different zone data*

The indicated outliers were minimal and due to the patient data as listed in Table1.

| PatientNumMasked | Label | Zone |
|---|---|---|
| 168 | 1 | RU |
| 168 | 1 | RM |
| 375 | 0 | RL |
| 203 | 0 | LL |

*Table 1: Data related to the outlier points*

### 3.3.4 Data Combination Scenarios

Experiments are performed in two different scenarios. In one scenario models are run on each zone data and average outcome on all zones is reported. To have a better understanding of the data and explore alternative improved analytics models, in another scenario combined data from all zones into one. An additional feature column added as Zone indicating zones each data belongs to.

### 3.3.5 Modeling and Validation Approach

Two methods are selected for modeling and validation:

- Method1: Leave one-patient-out cross-validation (As suggested)

- Method2: Training/Validation/Testing (Seeking improvements on Method1 results and validating its results)

Both of these models are following a framework listed below:

1. Add additional column to each Zone data to signify the zone it belongs to.
2. Concatenate data from all zones.
   Experiments were run on each zone separately averaging the results. Both approaches resulted in the same precision/accuracy. As described in this report, concatenating all zone data approach was selected due to its convenience and the fact that features reported higher weight in importance when combining zone data.
3. Check for quality of data, identify outliers (explore/analyze)
4. Break data into Training, Validation and Testing.
5. Create a model based on the Train/Validation data.
6. Predict Test data per generated model.

When dealing with Method1, the following is used for modeling

1. Create a list of all unique Patient IDs.
2. For each unique patient ID, create list of data with only the patient IDs (as Test data), and the list not including the patient IDs (as Train/Validation data).

When dealing with Method2

1. Distributions of data are identified.
2. Data is divided into Training and Testing while making sure each set contains data from every distribution. 2/3 of data is assigned to Training, and 1/3 to testing.
3. Perform random sampling across the training data ensuring every zone in the training data had equal number of observations.
4. Perform another random sampling across training data to subset ¼ of the training data as validation data.

### 3.3.6 Feature Selection

After some high level analysis using JMP tool on the relationships among the features, used R Statistical Random Forest model to report the top important features with reports on the percent importance on each feature. Table 2 includes lists of features sorted based on their importance according to their MeanDecreaseGini parameter results. Figure 2.1 shows a graphical view of the Mean Decrease Accuracy and Mean Decrease Gini parameter on most features.

Experiments show that order of importance in features differs from zone to zone and is different for all data combined. The percent importance of top features is high when data is combined. This prompts the idea of combining data for analysis rather than running individual analysis on each zone data and then combines them. Experiments have been run for both scenarios.

All features report positive importance. Therefore **to maximize the accuracy**, all 39 features are selected for analysis. Number of features is reasonable and can be handled by the Classifying models in reasonable turnaround time.

All Data

| Number | Feature | % Importance |
|---|---|---|
| 1 | Hist_2_150_2_Entropy | 104.0581931 |
| 2 | Hist_2_30_2_Entropy | 79.66662042 |
| 3 | Hist_2_180_2_Entropy | 68.31438203 |
| 4 | Hist_1_180_2_StdDev | 66.52360527 |
| 5 | Hist_1_120_2_Mean | 65.05469307 |
| 6 | Hist_2_150_2_Mean | 61.35340388 |
| 7 | Hist_2_30_2_Mean | 50.02504848 |
| 8 | Hist_1_135_2_Entropy | 47.08513594 |
| 9 | Hist_2_180_2_Mean | 43.29273483 |
| 10 | Hist_1_180_2_Mean | 39.92341606 |
| 11 | CoMatrix_Deg45_Local_Homogeneity | 36.97519982 |
| 12 | CoMatrix_Deg135_Local_Homogeneity | 36.79859885 |
| 13 | Hist_1_135_2_Mean | 34.83287339 |
| 14 | Hist_1_90_2_Skewness | 34.59803835 |
| 15 | Hist_1_30_2_Mean | 33.84484011 |
| 16 | Hist_2_90_1_Kurtosis | 30.2238713 |
| 17 | CoMatrix_Deg90_Local_Homogeneity | 29.13501768 |
| 18 | Hist_2_135_1_Entropy | 25.87528638 |
| 19 | Hist_2_45_1_Entropy | 24.79261482 |
| 20 | Hist_2_90_2_Mean | 23.91219156 |
| 21 | Hist_2_180_2_Kurtosis | 21.7791701 |
| 22 | Hist_0_0_0_Mean | 20.52867433 |
| 23 | Hist_2_150_2_Kurtosis | 18.75272206 |
| 24 | Hist_2_90_2_Skewness | 15.80308056 |
| 25 | CoMatrix_Deg135_Inertia | 15.75467472 |

Right Lower Zone Data

| Number | Feature | % Importance |
|---|---|---|
| 1 | Hist_2_150_2_Entropy | 22.68178536 |
| 2 | Hist_2_30_2_Entropy | 13.08701061 |
| 3 | Hist_1_180_2_StdDev | 12.60906196 |
| 4 | Hist_2_180_2_Entropy | 12.30568497 |
| 5 | Hist_2_135_1_Entropy | 11.5470686 |
| 6 | Hist_1_120_2_Mean | 10.06704679 |
| 7 | Hist_2_30_2_Mean | 8.183499782 |
| 8 | CoMatrix_Deg45_Local_Homogeneity | 8.151436663 |
| 9 | Hist_1_135_2_Entropy | 7.880486202 |
| 10 | Hist_2_150_2_Mean | 7.857142073 |
| 11 | CoMatrix_Deg135_Local_Homogeneity | 7.379662252 |
| 12 | Hist_2_180_2_Mean | 6.724880922 |
| 13 | Hist_1_135_2_Mean | 5.453438394 |
| 14 | Hist_1_30_2_Mean | 4.991921992 |
| 15 | Hist_1_180_2_Mean | 4.954702403 |
| 16 | Hist_2_45_1_Entropy | 4.927597091 |
| 17 | Hist_2_150_2_Kurtosis | 4.405496003 |
| 18 | CoMatrix_Deg135_Inertia | 4.256999566 |
| 19 | Hist_2_90_1_Kurtosis | 4.069932692 |
| 20 | Hist_1_90_2_Skewness | 3.939604174 |
| 21 | Hist_2_180_2_Kurtosis | 3.604036504 |
| 22 | Hist_0_0_0_Mean | 3.417006416 |
| 23 | CoMatrix_Deg90_Local_Homogeneity | 3.199732459 |
| 24 | Hist_2_90_2_Mean | 2.764344087 |
| 25 | Hist_2_180_1_Skewness | 2.550974322 |

Right Upper Zone Data

| Number | Feature | % Importance |
|---|---|---|
| 1 | Hist_2_150_2_Entropy | 19.66192281 |
| 2 | Hist_1_180_2_StdDev | 11.32295773 |
| 3 | Hist_1_135_2_Entropy | 10.46200492 |
| 4 | Hist_2_180_2_Entropy | 9.097017311 |
| 5 | Hist_2_135_1_Entropy | 8.699080683 |
| 6 | Hist_2_30_2_Entropy | 8.372749968 |
| 7 | Hist_2_90_1_Kurtosis | 6.838102947 |
| 8 | CoMatrix_Deg135_Local_Homogeneity | 6.429578915 |
| 9 | Hist_1_120_2_Mean | 6.062403968 |
| 10 | Hist_2_180_2_Mean | 6.030150679 |
| 11 | Hist_1_180_2_Mean | 5.364031293 |
| 12 | Hist_2_150_2_Mean | 5.288417986 |
| 13 | Hist_1_90_2_Skewness | 4.536164603 |
| 14 | CoMatrix_Deg90_Local_Homogeneity | 4.507967806 |
| 15 | Hist_2_30_2_Mean | 4.338447351 |
| 16 | Hist_1_135_2_Mean | 4.201447775 |
| 17 | Hist_1_30_2_Mean | 3.746829873 |
| 18 | CoMatrix_Deg45_Local_Homogeneity | 3.533577983 |
| 19 | CoMatrix_Deg135_Inertia | 2.969242097 |
| 20 | Hist_2_180_2_Kurtosis | 2.794740903 |
| 21 | Hist_1_180_2_Skewness | 2.318658079 |
| 22 | Hist_2_60_2_Skewness | 2.288404054 |
| 23 | Hist_2_45_1_Entropy | 2.271962924 |
| 24 | Hist_2_90_2_Mean | 2.042358743 |
| 25 | Hist_2_150_2_Kurtosis | 2.025928993 |

*Table 2: Data indicating the importance of the featues reported by the Random Forest statistical model for All Data(Table on Left), Data From Right Lower Zone (Middle Table), Data from Right Upper Zone (Table on Right). Reported importance for All Data is the highest. That is one of the motivations for combing all zone data into one set for analysis.*

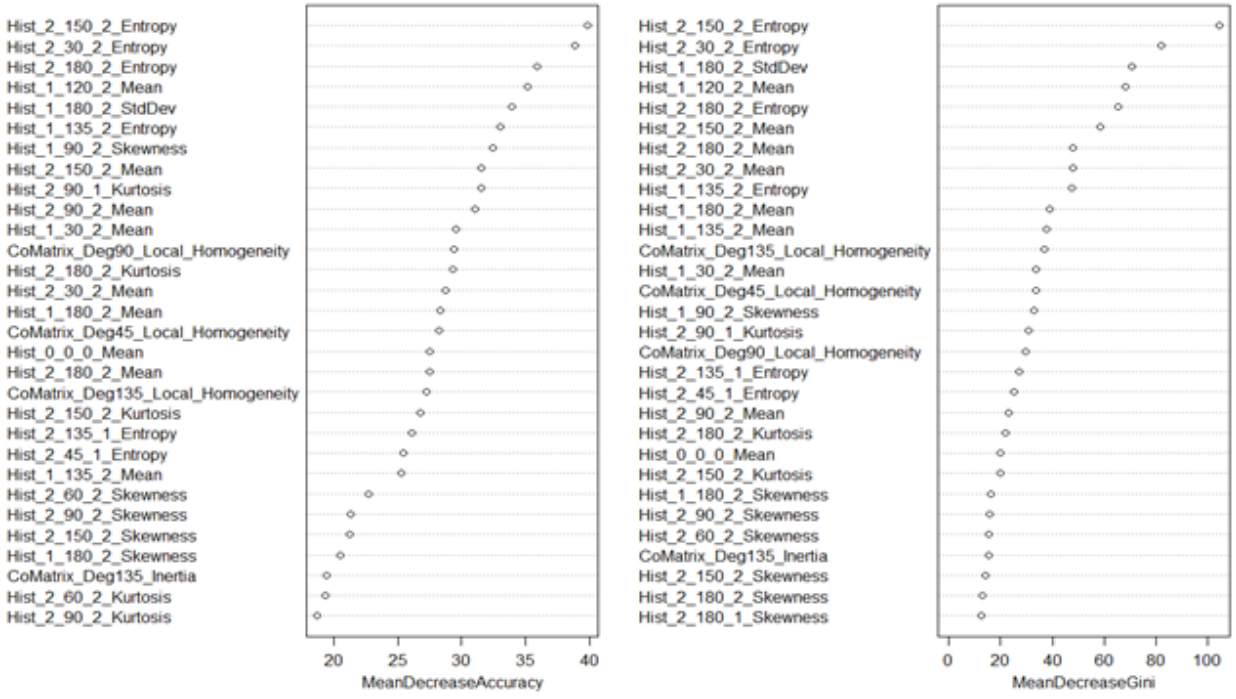



*Figure 2.1: Mean Decrease Accuracy and Mean Decrease Gini parameter graphical reports on most features.The higher the importance of a feature, the higher its MeanDecreaseAccuracy and higher its MeanDecreaseGini parameters.*

### 3.3.7 Correlations

Figure 5 and 6 highlight the positive and negative correlations among the top 6 high important features. As can be seen from the analysis most of these features (Especially the ones with highest importance) are highly correlated with each other. The correlations are used to study the relationships and dependence among the features. Studying such relationships can lead identifying confounders.

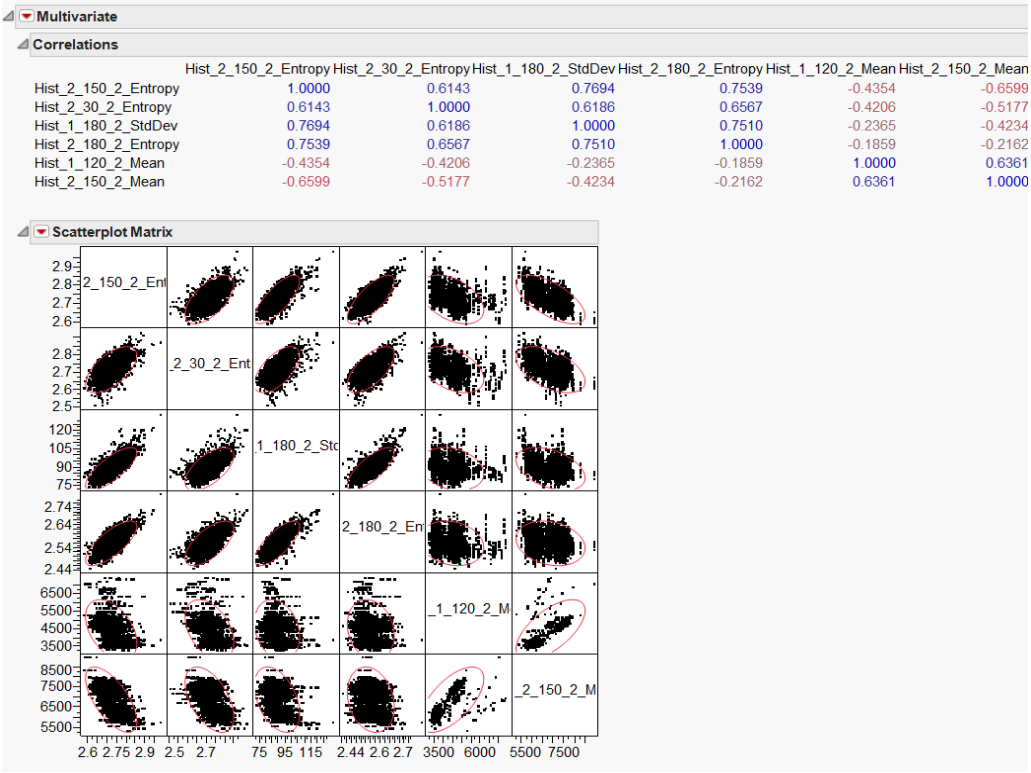



*Figure 5: Correlation matrix for the top 6 important features reported by the analysi*

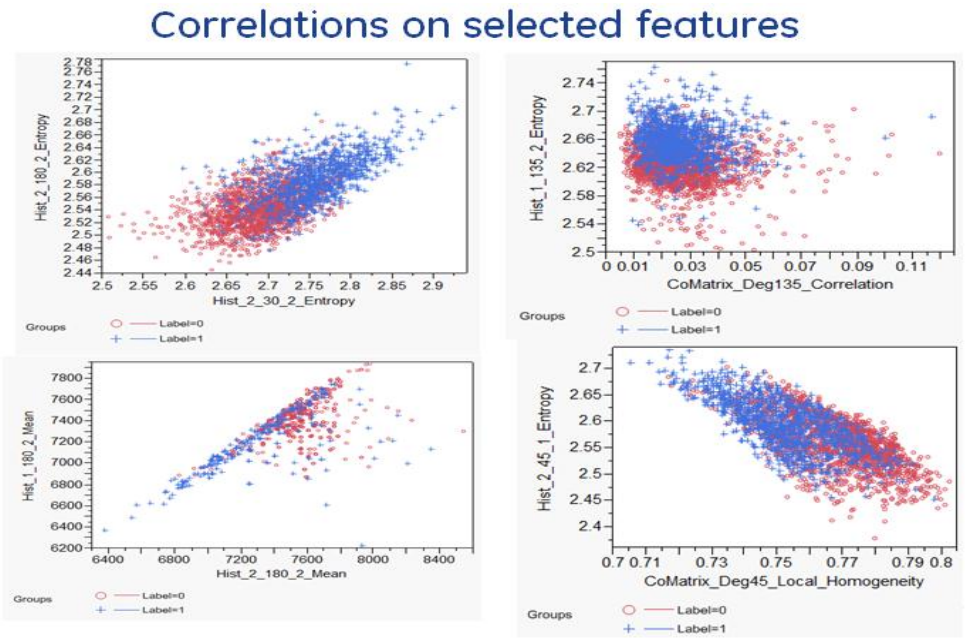



*Figure 6: Correlations on selected top important features color categorized by the Labels*

Figure 6.1 lists partial dependence plots on some of the selected features utilizing the random forest model [RandomForest Package]. Partial dependence plot gives a graphical depiction of the marginal effect of a variable on the class probability. The top important feature Hist_2_150_2_Entropy reaches a high dependency of 100% in this figure, whileone of the low important features plotted in this figure (Hist_0_0_0_Skewness with 10% importance) barely reaches 50% range in the dependency plot.

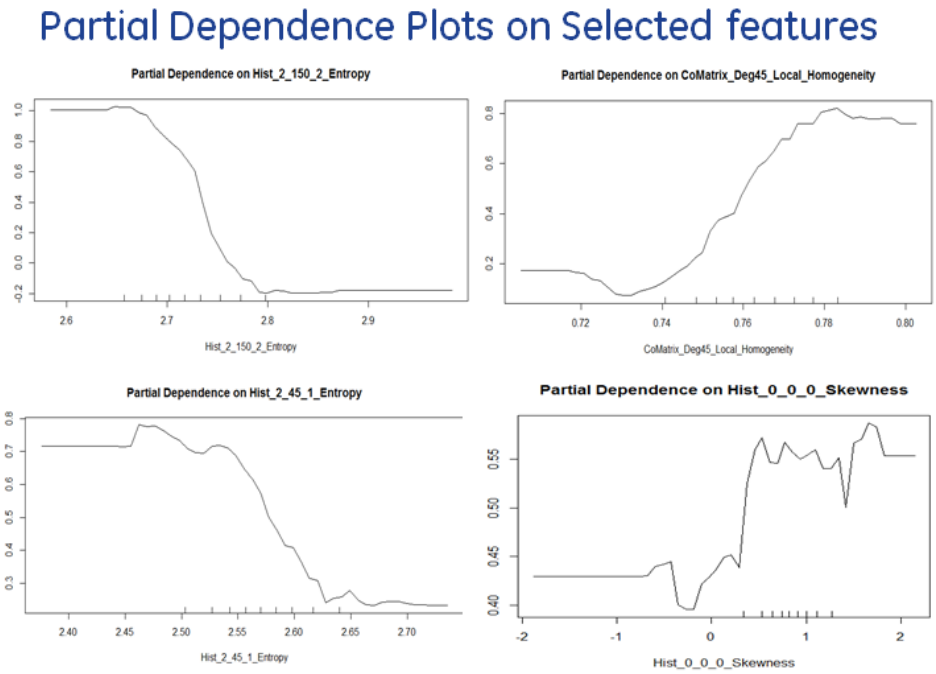



*Figure 6.1: Partial Dependence plots on selected features shows the dependence between the diagnostics and the selected features, marginalizing over the values of all other features (the complement features).*

### 3.3.8 Normalizing Data

Normalized the data using R function from [Normalize 2009]. This did not improve on the accuracy, indicating that the data is already in its normal form. Exploring the data with JMP tool also indicated that most features are already in normal form.

```
#https://stat.ethz.ch/pipermail/r-help/2009-September/210808.html
std=function(x){if(length(which(is.na(x)))==0) (x-mean(x))/sd(x) else

  (x-mean(x,na.rm=T))/sd(x,na.rm=T)
}

cols=cols-1
for(i in 1:cols)
{
  GE_Data_1[,i] = std(GE_Data_1[,i])
}
```

# 4 Analysis

## 4.1 Model Selection

**Random Forest, SVM and Neural Networks** were examined as models for classification.

According to [RF_Ecology (2011)]: "The Advantages of RF compared to other statistical classifiers include (1) very high classification accuracy; (2) a novel method of determining variable importance; (3) ability to model complex interactions among predictor variables; (4) flexibility to perform several types of statistical data analysis, including regression, classification, survival analysis, and unsupervised learning; and (5) an algorithm for imputing missing values."

We decided to use Random Forest over Trees for reasons also listed under [RandomForestTrees (2007)]:

- No need for pruning trees
- Accuracy and variable importance generated automatically
- Over fitting is not a problem
- Not very sensitive to outliers in training data
- Easy to set parameters

Random Forest works as collection of decision trees and final decision is based on the voting on final classification of data. A random forest with 500 trees will divide subsets of the data into each tree and the final result is based on the most popular classification across all the 500 trees.

For the Training/Validation/Testing method of our studies, we used Random Forest for the modeling since we had the option to choose data with similar distributions for each category of Training, Validation and Testing.

The Label column type changed to type 'Factor' in order for the R statistical models function properly.

## 4.2 Handling of Zone Data

It is important to find the best approach for handling of the data. We had two scenarios as following:

- To model per each zone then combine all the results by and reporting average percent on accuracy/precision/recall.
- To combine all the data first, then perform modeling on all data.

To come up with the right strategy, we did run experiments on both scenarios.

- Leaving-one-patient-out cross-validation method used for modeling.
- SVM used for this particular modeling (since during prototyping did show promising performance).

As can be seen from Figure 7, the results of accuracy, precision, recall for both scenarios are compatible and do not make much difference. Combined data is indicated by Zone type of 'All'. In addition, the outliers are removed from combined data and a report on SVM model results show that only slight

improvements gain due to the outlier removals. This must be due to small number of outliers in the data.

| SVM | Zone | Accuracy | Precision | Recall |
|---|---|---|---|---|
| 1 | RU | 0.921914 | 0.958763 | 0.775 |
| 2 | RM | 0.919149 | 0.932961 | 0.865285 |
| 3 | LU | 0.936224 | 0.978723 | 0.8 |
| 4 | RL | 0.894619 | 0.858824 | 0.863905 |
| 5 | LM | 0.895075 | 0.902857 | 0.831579 |
| 6 | LL | 0.882488 | 0.863014 | 0.802548 |
| 7 | ALL | 0.89985 | 0.895713 | 0.818856 |
| | ALL (Outliers Removed) | 0.905037 | 0.897936 | 0.832094 |
| | | | | |
| SVM | Avg 6 Zones | 0.908245 | 0.915857 | 0.823053 |

*Figure 7: Accuracy, Precision, Recall results when running model (SVM) on indvidual zones then taking their averages (Blue color rows) compared to the results when All data first combined then the SVM modeling applied. Results show both approaches are compatible and do not differ much in results. We did remove the outliers from the combined data and did run the modeling. The results did not show only a very slight improvement.*

As a result of this study, we continued for the rest of the studies with All Data combined for its convinience and ease of handling/automation.

## 4.3 Modeling and Validation

We used two approaches for modeling and validation.

- Leave one-patient-out cross-validation
- Training/Validation/Testing

### 4.3.1 Method1: Leave one-patient-out cross-validation

#### *4.3.1.1 Method Detail*

1. Let $P = \{P_1 \dots P_N\}$ be the patients for whom data is provided, where $P^+ \subseteq P$ are the patients who have been diagnosed with Pneumoconiosis and $P^- \subseteq P$ are those haven't been diagnosed with the disease.
2. For each patient $P_i$, let the vectors $\vec{x}_{ij}, j \in \{1 \dots 6\}$ represent the feature vectors extracted for the 6 lung zones (Right Upper, Right Middle, Right Lower, Left Upper, Left Middle, Left Lower), and let $y_{ij} \in \{0,1\}$ represent the zone-level labels (1=Pneumoconiosis, 0=healthy). Note that, even if one of the zones show evidence of Pneumoconiosis, the patient is diagnosed as having the disease. In other words, the patient label $y_i = max_j\ (y_{ij})$.
3. For each patient $P_i$:
    a. Build a model (or a set of models) on the data for $P - P_i$. This can be a single model that uses all the data across all 6 zones, or individual zone-level models whose predictions are then combined.

b. 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 in some fashion, or to use some method that predicts a single patient-level label on the basis of all the $\vec{x}_{ij}$ feature vectors in one go. Let the predicted label be $\hat{y}_i$.

c. Steps a and b are to be repeated for each patient; therefore, the predicted labels $\hat{y}_1 \ldots \hat{y}_N$ will each be a prediction from a different model (with respect to which that patient is an unseen sample), and any aggregate performance metric that compares them with $y_1 \ldots y_N$ can be seen as a measure of generalization ability (i.e., performance on unseen examples).

#### *4.3.1.2 Required Measurements*

Following are the required measurements reported for this study:

$$Accuracy = \left(\frac{1}{N}\right)|\{i \mid y_i = \hat{y}_i\}|$$

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

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

#### *4.3.1.3 Statistical Modeling*

This approach involves separating data for each patient into a test group and put the rest of the data into the training group. After combining data for all zones, the test group could have up to 6 records for each patient (1 from each zone category). As mentioned in data exploratory analysis segment of this report, some patients are missing data from some zones. Therefore the test data set can vary between 1 to 6 records.

Three classification modeling used: SVM, Random Forest and Neural Networks.

- **Neural Network**: (Default options selected, plus number of units in the hidden layer (size=10))

```
nnet1 = nnet(Label ~ ., data = train, size=10)
predTest <- predict(nnet1, newdata=test, type='class')
```

- **SVM**: (Default options selected)

```
svm1 <- svm(Label ~ ., data = train)
predTest <- predict(svm1, newdata=test, type='class')
```

- **Random Forest**: The options used for the Random Forest are as follows:

```
rfModel <- randomForest(Label ~ ., data = train,importance = TRUE,prox=TRUE,keep.forest=TRUE,
                        ntree=1001, do.trace=100,mtry=round(length(train)^(1/3)))
predTest <- predict(rfModel, newdata=test, type='class')
```

- ntree: number of trees to grow. We have selected this number to be 1001 (large enough to get a good accuracy/precision in results). Increasing number of trees requires more processing time. Random Forest model library has parallel processing capability that can be utilized for speed. We did not use this capability and as a result it took around 4 hours to complete this analysis with 1001 trees.
- mtry: number of variables randomly sampled as candidates at each split. Through some experimentations, this number has been selected to be round ($\sqrt[3]{number\ of\ features}$).
- Importance and prox are set to TRUE to assess the importance of the predictors.
- Keep.forest is set to TRUE to retain the forest in the output and allow for the predictions to be possible. Also, this parameter is needed to get the importance and partial dependency plots.

#### *4.3.1.4 Results*

The summary results on accuracy, precision and recall on each category can be seen in figure 8. The results for Neural Network are not reliable due to its low accuracy and precision reports.

| Modeling1: leave-one-patient-out cross-validation | | | |
|---|---|---|---|
| Model | accuracy | precision | recall |
| SVM | 0.899847 | 0.895713 | 0.818856 |
| Random Forest | 0.898312 | 0.896149 | 0.813559 |
| Neural Network | 0.659632 | 0.652406 | 0.129237 |

*Figure 8: Summary statistical report on accuracy, precision and recall for Leave-one-patient-out cross-validation modeling approach. The results from SVM and Random Forest are almost identical. The results from Neural Networks are not reliable.*

### 4.3.2 Method2: Training/Validation/Testing

#### *4.3.2.1 Method Detail*

Method2 breaks the data into Training, Validation, Testing groups. Generates model based on Training and Validation groups and tests it on the Testing group. This method can be used to validate the results captured from Method1 and improve up on them.

- To ensure similarity in distribution of training and test data, zones selected for each group from each distributions.
- Training/validation data is comprised of "RU","RM","RL","LM" zones with total of 1780 records
- Testing data is comprised of "LU" and "LL" with total of 826 records.

A random sampling across the training data was performed ensuring every zone in the training data had equal counts, 397. Then ¼ of the training data is assigned as validation data via additional random sampling. The final data distribution is as follows:

- Training: 1191 observations, 39 features
- Validation: 397 observations, 39 features
- Test: 826 observations, 39 features

#### 4.3.2.2 Required Measurements

The required measurements are the same as Method1.

#### 4.3.2.3 Statistical Modeling

Random Forest is used for this modeling for reasons highlighted in Model Selection section of this report. In particular Random Forest automates the training and validation steps as in one call. Figure 9 includes the prediction results from modeling in this methodology. Results show an improvement in accuracy and recall compare to the Method1 modeling.

The following code snapshot includes the call with its parameters to Random Forest along with the calls to creation of the confusion matrix from the actual and predicted values. By increasing number of trees large enough (1001), and the random sampling approach taken, cross validation should not be necessary here.

```
rfModel  <-randomForest(training2[,1:length(training2)-1], training2$Label,
                        xtest=validation2,ytest=validationValues,keep.forest=TRUE,
                        ntree=1001, do.trace=100,mtry=round(length(training2)^(1/3)))
predictedValues <- predict(rfModel, testing, type="response")
conTable0<-confusionMatrix(table(predictedValues,actualValues))
conTable0
conTable1<-confusionMatrix(table(predictedValues,actualValues),positive='1')
conTable1
```

#### 4.3.2.4 Results

The results from this modeling are: accuracy of 0.92131, precision of 0.88192, and recall of 0.87868. There is a slight (2%) improvement in results compare to the Method1 modeling.

| Modeling2: Training/Validation/Testing | | | |
|---|---|---|---|
| Model | accuracy | precision | recall |
| RandomForest | 0.92131 | 0.88192 | 0.87868 |

*Figure 9: Summary statistical report on accuracy, precision and recall for Training/Validation/Testing modeling approach.*

The result of the confusion matrix for normal and abnormal classes is reported in figure 10. It can be seen that out of 554 normal individuals, the prediction reports 32 of them as abnormal. And out of 272 normal individuals, the prediction reports 34 of them as normal. The formulas for all the statistics reported by the confusion matrix can be found in Appendix A. As listed in Appendix A, the

```
Confusion Matrix and Statistics

                actualValues
predictedValues   0   1
              0 522  34
              1  32 238

               Accuracy : 0.9201
                 95% CI : (0.8995, 0.9377)
    No Information Rate : 0.6707
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8188
 Mcnemar's Test P-Value : 0.902

            Sensitivity : 0.9422
            Specificity : 0.8750
         Pos Pred Value : 0.9388
         Neg Pred Value : 0.8815
             Prevalence : 0.6707
         Detection Rate : 0.6320
   Detection Prevalence : 0.6731

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

                actualValues
predictedValues   0   1
              0 522  34
              1  32 238

               Accuracy : 0.9201
                 95% CI : (0.8995, 0.9377)
    No Information Rate : 0.6707
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8188
 Mcnemar's Test P-Value : 0.902

            Sensitivity : 0.8750
            Specificity : 0.9422
         Pos Pred Value : 0.8815
         Neg Pred Value : 0.9388
             Prevalence : 0.3293
         Detection Rate : 0.2881
   Detection Prevalence : 0.3269

       'Positive' Class : 1
```

*Figure 10: Confusion Matrix report for both Normal and Abnormal classes*
*Accuracy: 0.9201, 95% CI: (0.8995, 0.9377), p-Value [ACC > NIR] : <2e-16*

The actual error rates are reported in Figure 10. It can be seen that the sensitivity value for abnormal cases (0.8750) is lower compare to the normal cases (0.9422). This indicates that the weakness of our model is in predicting patients who are actually have the disease (Label=1, Abnormal). Accuracy reported in the confusion matrix (0.9201) is consistent with the accuracy we have calculated from running the model on the test data (0.92131).

We can claim that this is an accurate model since the Prevalence (as the rate of actual occurrence of each class) is very close to the value of detection rate (as the rate of prediction of occurrence of each class).

## 4.4 Reproducibility

All analyses performed in this report are automated and could be reproduced using R programs written for this analysis [RStudio, 2013] (available on request). To reproduce the exact results presented in this manuscript, the analysis must be performed on the same data set (available on request).

## 4.5 Confounders

The 12% margin error margin of diagnosing the abnormal cases as normal might be due to confounders. Since not all the features are provided to this case study, it is difficult to identify the exact confounders. Among the available features, correlations and dependencies are examined to identify if any of the low important features had any high correlation or dependency with the dependent variable (Label). We could not find any strong correlations between the features with the least importance and the features with the most importance reported. Finding such correlations

could have helped us identifying some confounders that did not have direct impact on the outcome, but their importance on the most impactful features could be shown. Having more features could definitely help identifying confounders.

# 5 Conclusions:

We proposed different methods on handling of the zone data along with different statistical modeling and validation approaches to predict whether a particular patient is diagnosed as normal or abnormal. We conclude that the random forest model (with all zones data combined, then divided into training, validation and testing; with accuracy between 89.95% and 93.77% at 95% confidence interval with small P value) can claim to be an accurate model in predicting the disease for any selected patient (Although the model had a relatively significant error margin of ~12% in misclassifying an abnormal patient case as normal).

Improvements on the Method2 results in this study are partly due to test data and training data chosen to have the same distributions. While Method2 resulted in slightly higher accuracy and precision, Method1 approach that is validated by Mehtod2 can also be used as a reliable approach for modeling this case either using SVM or Random Forest.

We used Random Forest for Method2 and part of Method1 studies, for reasons listed under Model Selection section. Considering the results of this particular study, SVM can be used to get results with the same accuracy as Random Forest does in fraction of time Random Forest is using.

- **Next Steps:**
  To further increase the accuracy in the future, more data with more features should be studied. Having all features could have given the SVM/Random Forest models the opportunity to identify the most related/important features resulting in better prediction results. Also having more features could help identifying the specific confounders and ways to controls them in the experimental environments.

- **Possible Problems:**
  For studying cases with large number in features and observations in the future, a model scaling approach should be taken to ensure accurate and reliable results. Then feature selection and data sampling approaches should be examined. A method should be in place identifying and controlling the confounders while experimenting and collecting data.

# 6 References


- SVM : Introduction to Support Vector Machines (SVMS) http://www.louisaslett.com/Courses/Data_Mining/ST4003-Lab7-Introduction_to_Support_Vector_Machines.pdf
- AE. (2013). Analytics Engineer Case Studies: http://libraries.ge.com/foldersIndex.do?entity_id=36905398101&sid=101&SF=1#36970369101
- RF_Ecology (2011). RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY https://www.nescent.org/wg/cart/images/1/1a/HighResRandomForestsandAppendices.pdf Ecology, 88(11), 2007, pp. 2783–2792 2007 by the Ecological Society of America
- RandomForestTrees (2007). Introduction to decision trees and random forests http://www.whrc.org/education/indonesia/pdf/DecisionTrees_RandomForest_v2.pdf http://en.wikipedia.org/wiki/Random_forest
- RandomForest Package, The randomForest Package (for R) description
- Caret Package, http://cran.r-project.org/web/packages/caret/caret.pdf
- R Project (2013). *R Project*. Retrieved from http://www.R-project.org
- RStudio. (2013). *Using R Markdown with RStudio*. Retrieved from http://www.rstudio.com/ide/docs/authoring/using_markdown
- SAS JMP. (2013) SAS JMP Project. http://www.jmp.com/ads/adwords.shtml?gclid=CPC2zr6B07kCFVGi4AodjFkABQ
- Normalize (2009): https://stat.ethz.ch/pipermail/r-help/2009-September/210808.html


# 7 Appendix A

The following definitions on the confusion matrix can be found in the [Caret Package] Reference Document.

Suppose a 2x2 table with notation

| | Reference | |
|---|---|---|
| Predicted | Event | No Event |
| Event | A | B |
| No Event | C | D |

The formulas used here are:

$$Sensitivity = A/(A+C)$$

$$Specificity = D/(B+D)$$

$$Prevalence = (A+C)/(A+B+C+D)$$

$$PPV = (sensitivity*Prevalence)/((sensitivity*Prevalence)+((1-specificity)*(1-Prevalence)))$$

$$NPV = (specificity*(1-Prevalence))/(((1-sensitivity)*Prevalence)+((specificity)*(1-Prevalence)))$$

$$DetectionRate = A/(A+B+C+D)$$

$$DetectionPrevalence = (A+B)/(A+B+C+D)$$

See the references for discusions of the first five formulas.

For more than two classes, these results are calculated comparing each factor level to the remaining levels (i.e. a "one versus all" approach).