Development of Pattern Discovery and Analysis across a Clinical Program

Prem Narasimhan

Abstract: Development of pattern discovery and analysis across a clinical program (Drug A and Drug A combination) programs for analysis of Immune Mediated and Myocarditis Events Propose of this project is to probe a machine learning approach to fast detection of risk signals in ongoing clinical trials. In this project want to show by employing a linear regression model to predict, the safety signal can be discovered earlier compared to other techniques used in the past. This project demonstrates the potential value of machine learning algorithms in improving near real-time clinical trial surveillance.

Background

- During the clinical development of Drug A and within the post-marketing experience, Immune mediated adverse events and Myocarditis events have been reported, consistent with checkpoint inhibitor therapy
- As per FDA request, additional data is being collected for these events across Drug A and Drug A combination programs to further characterize these events.
- Aggregate and patient level analysis of th**is** additional information will help in better medical understanding the nature of these clinical event, better addressing case management and identifying patient safety considerations. In addition, exploratory analysis will also accelerate responses to Health Authorities.

Problem Statement

- Current clinical analysis templates does not includes information that allow analysis of the data being collected for Immune Mediated adverse events and Myocarditis events
- Additionally, the integration of studies is currently done across protocols and not by programs, which does not allow aggregate analysis across Drug A, and Drug A with combination programs

Anticipated Benefits

Integrated analysis will allow for aggregate and patient level analysis of Immune Mediated adverse events and Myocarditis events in a uniform manner across different groups

- Allow for exploratory analysis co-relating different information without development of complex reports (i.e. conmed vs AEs, labs vs. AEs)
- Increased efficiency by
  - Development of integrated sets, Identify patterns, clusters, elimination redundant efforts
  - Reducing errors inherent in manual analysis/tabulations
  - Assist in determining safety and tolerability responses to Health Authorities
  - Automate the process as close to real-time

Primary Objectives and Workflow Diagram

- Discovering patterns in the data where a set of subjects shares some features (such as AEs, diagnosis, lab test, dose, demographics etc.) that occur frequently together or strongly correlated in the data set**.**
- The pattern discovery helps in determining the safety and tolerability of the Anti-XXX-Monoclonal (A Drug) administered alone and in combination with the Anti-PD-1 Monoclonal Antibody (B with A drug) in subjects with advanced solid tumors.

Flow Diagram

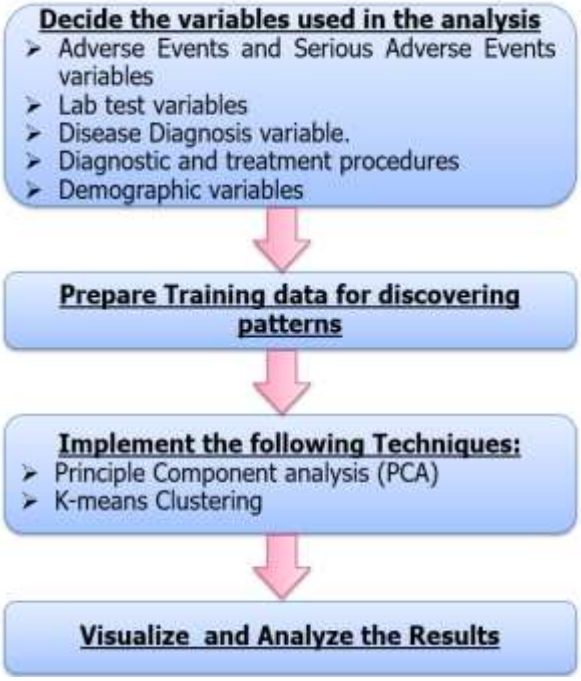

Decide the variables used in the analysis
➢ Adverse Events and Serious Adverse Events variables
➢ Lab test variables
➢ Disease Diagnosis variable.
➢ Diagnostic and treatment procedures
➢ Demographic variables
Prepare Training data for discovering patterns
Implement the following Techniques:
➢ Principle Component analysis (PCA)
➢ K-means Clustering
Visualize and Analyze the Results

Figure 1

K-means clustering with PCA

- Started with 41 features including AEs, lab test, disease diagnosis, diagnostic and treatment procedures and the demographic variables.
- Implement principle component analysis (PCA) to reduce the dimensionality.
- Then, applied k-means clustering to group samples into five different clusters.

Methods

This project contains data from 3 different clinical trials have a population of 1600 subject and from the following domain AEs, diagnosis, lab test, dose, demographics and biomarkers and Tumor Type. Some of the subject were on Drug A and some of them on combination.

Supervised Learning

One way to get the post-drug treatment distribution is to apply a classification algorithm to predict the binary post-drug treatment event (1=yes, 0=no) for each member in the test set. Where age, gender, baseline weight, baseline height, AE's and Bio-markers are used as input features. As a next step, split them into 5 clusters.

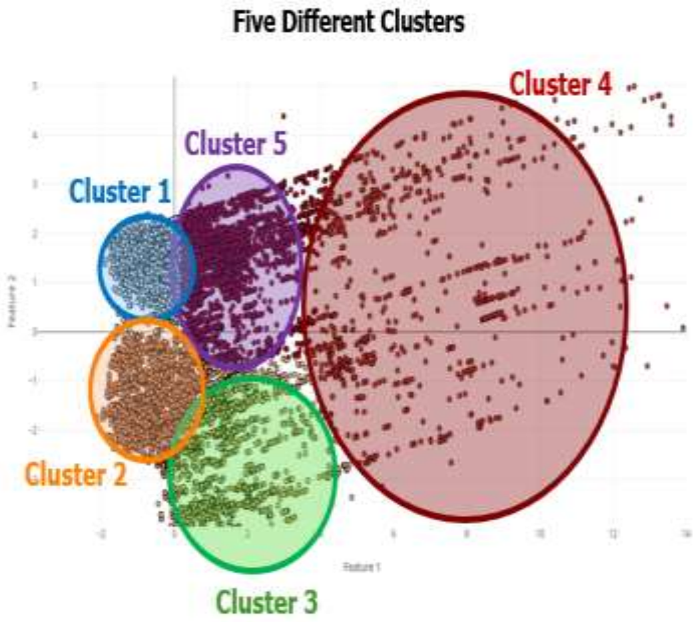


Figure 2

The important feature in the GMB modelling is the variable importance. The table below ranks the individual variables based on their relative influence, which is measure indicating the relative importance of each variable in training the model.

Gradient Boosting Machine model

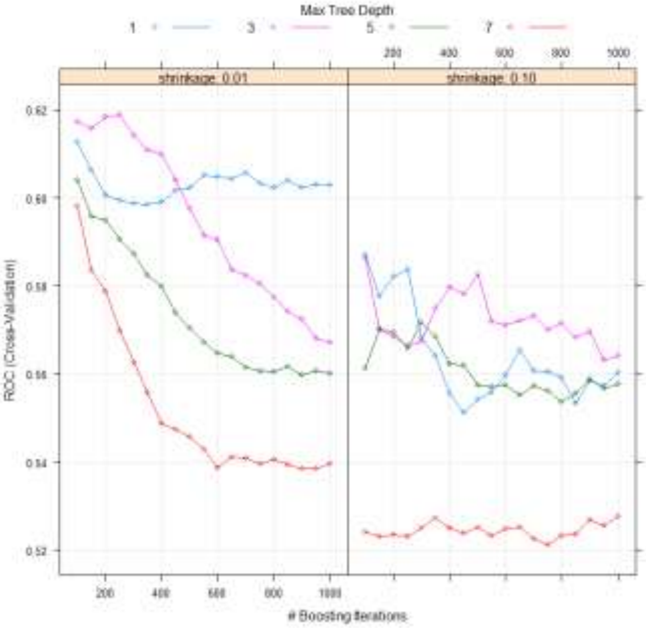


Figure 3

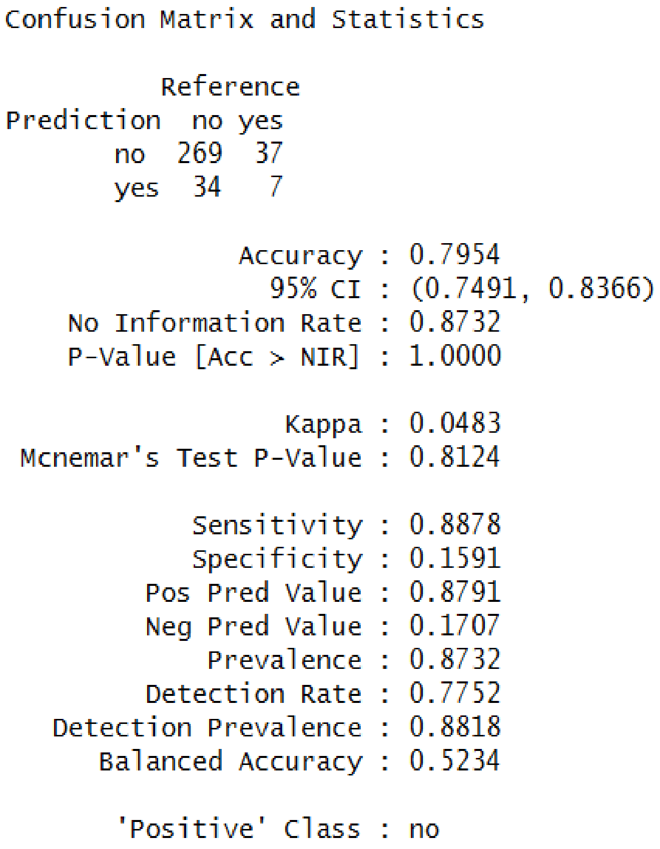
Variable Importance

A234:- Dermatitis Contact

A355:- Conjunctival Haemorrhage

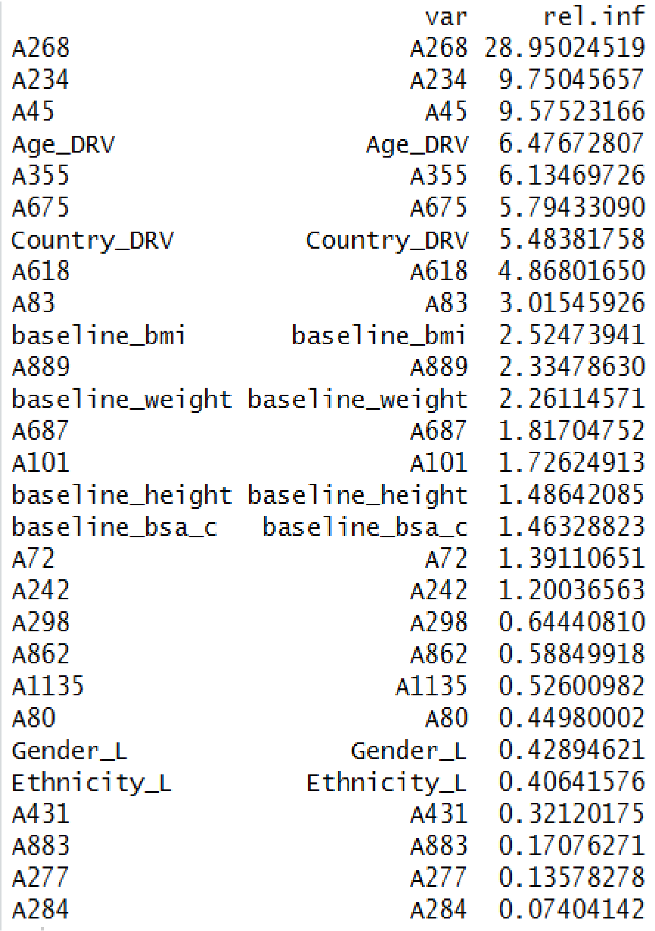
A618:- Alanine Aminotransferase Increase

Accuracy:-0.7954



```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  269  37
       yes  34   7

               Accuracy : 0.7954
                 95% CI : (0.7491, 0.8366)
    No Information Rate : 0.8732
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0483
 Mcnemar's Test P-Value : 0.8124

            Sensitivity : 0.8878
            Specificity : 0.1591
         Pos Pred Value : 0.8791
         Neg Pred Value : 0.1707
             Prevalence : 0.8732
         Detection Rate : 0.7752
   Detection Prevalence : 0.8818
      Balanced Accuracy : 0.5234

       'Positive' Class : no
```

Table 1



```
                                   var     rel.inf
A268                              A268 28.95024519
A234                              A234  9.75045657
A45                                A45  9.57523166
Age_DRV                        Age_DRV  6.47672807
A355                              A355  6.13469726
A675                              A675  5.79433090
Country_DRV                Country_DRV  5.48381758
A618                              A618  4.86801650
A83                                A83  3.01545926
baseline_bmi              baseline_bmi  2.52473941
A889                              A889  2.33478630
baseline_weight baseline_weight  2.26114571
A687                              A687  1.81704752
A101                              A101  1.72624913
baseline_height baseline_height  1.48642085
baseline_bsa_c          baseline_bsa_c  1.46328823
A72                                A72  1.39110651
A242                              A242  1.20036563
A298                              A298  0.64440810
A862                              A862  0.58849918
A1135                            A1135  0.52600982
A80                                A80  0.44980002
Gender_L                      Gender_L  0.42894621
Ethnicity_L                Ethnicity_L  0.40641576
A431                              A431  0.32120175
A883                              A883  0.17076271
A277                              A277  0.13578278
A284                              A284  0.07404142
```

Table 1.1

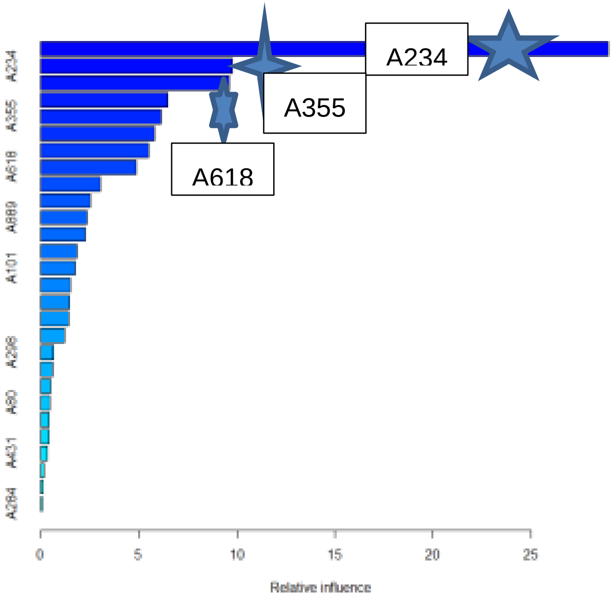


Figure 4

Testing set

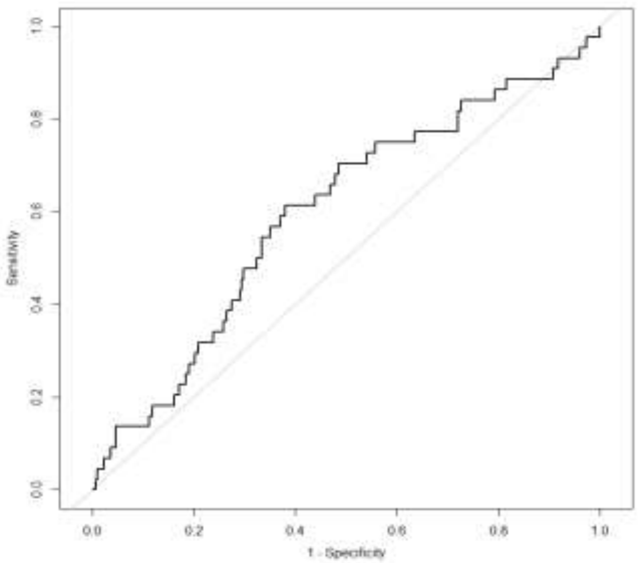
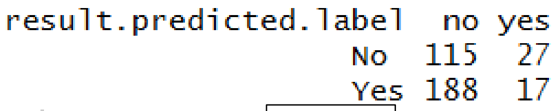


Figure 5

Use 50% probability as cut-off, the testing set prediction accuracy was ~79.5%.

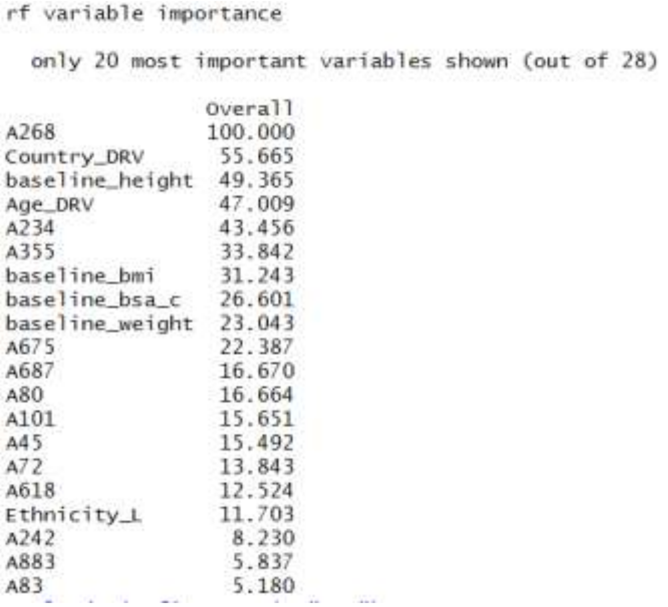


```
result.predicted.label  no yes
                   No  115  27
                   Yes 188  17
```

Table 2

Random Forest Model

This is an imbalance classification problem, so accuracy is not an appropriate metric. Instead we’ll measure Receiver Operaing Characteristic Area Under the Curve (ROC AUC), a measure fomr 1= yes and 0=no with a radom guess scoring 0.5

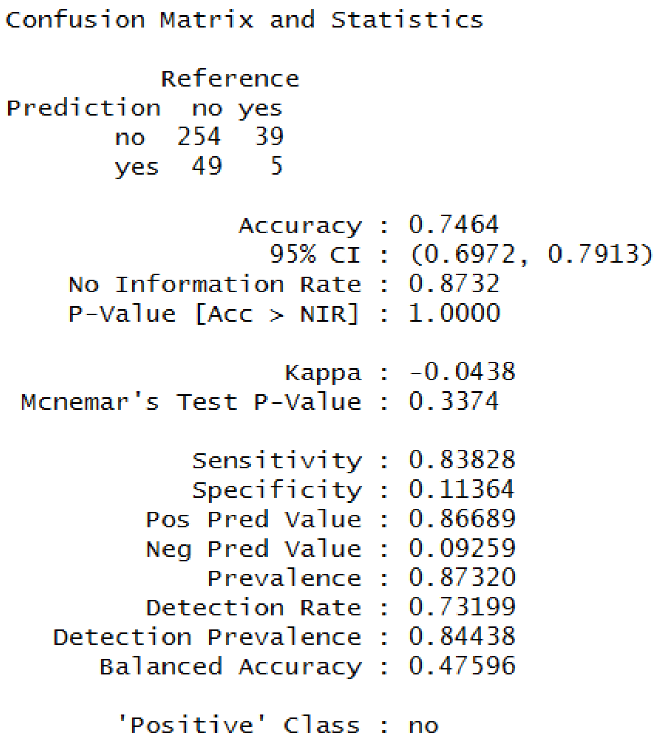


```
rf variable importance

  only 20 most important variables shown (out of 28)

                Overall
A268            100.000
Country_DRV      55.665
baseline_height  49.365
Age_DRV          47.009
A234             43.456
A355             33.842
baseline_bmi     31.243
baseline_bsa_c   26.601
baseline_weight  23.043
A675             22.387
A687             16.670
A80              16.664
A101             15.651
A45              15.492
A72              13.843
A618             12.524
Ethnicity_L      11.703
A242              8.230
A883              5.837
A83               5.180
```

g

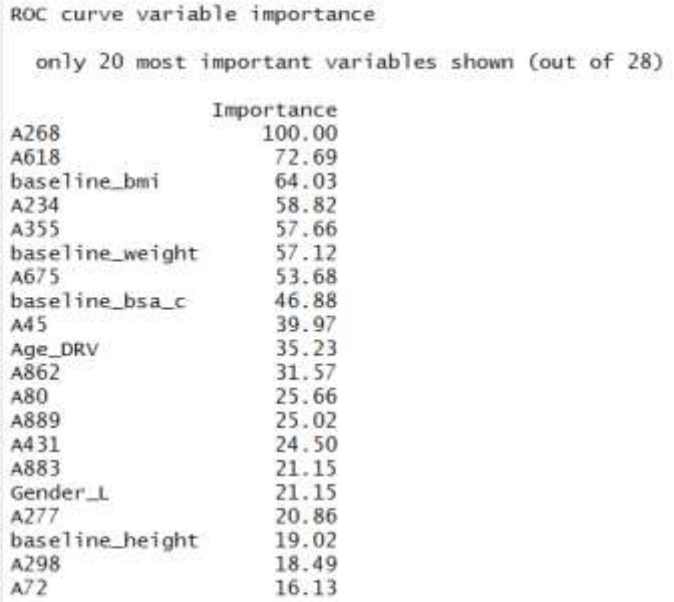
Project

Table 1

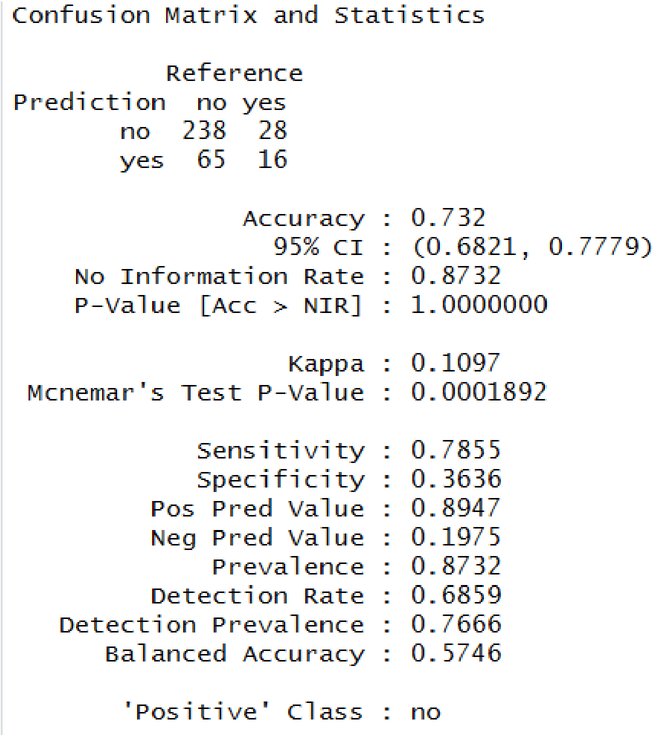


```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  254  39
       yes  49   5

               Accuracy : 0.7464
                 95% CI : (0.6972, 0.7913)
    No Information Rate : 0.8732
    P-Value [Acc > NIR] : 1.0000

                  Kappa : -0.0438
 Mcnemar's Test P-Value : 0.3374

            Sensitivity : 0.83828
            Specificity : 0.11364
         Pos Pred Value : 0.86689
         Neg Pred Value : 0.09259
             Prevalence : 0.87320
         Detection Rate : 0.73199
   Detection Prevalence : 0.84438
      Balanced Accuracy : 0.47596

       'Positive' Class : no
```

Support Vector Machine Model



```
ROC curve variable importance

  only 20 most important variables shown (out of 28)

                Importance
A268                100.00
A618                 72.69
baseline_bmi         64.03
A234                 58.82
A355                 57.66
baseline_weight      57.12
A675                 53.68
baseline_bsa_c       46.88
A45                  39.97
Age_DRV              35.23
A862                 31.57
A80                  25.66
A889                 25.02
A431                 24.50
A883                 21.15
Gender_L             21.15
A277                 20.86
baseline_height      19.02
A298                 18.49
A72                  16.13
```



```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  238  28
       yes  65  16

               Accuracy : 0.732
                 95% CI : (0.6821, 0.7779)
    No Information Rate : 0.8732
    P-Value [Acc > NIR] : 1.0000000

                  Kappa : 0.1097
 Mcnemar's Test P-Value : 0.0001892

            Sensitivity : 0.7855
            Specificity : 0.3636
         Pos Pred Value : 0.8947
         Neg Pred Value : 0.1975
             Prevalence : 0.8732
         Detection Rate : 0.6859
   Detection Prevalence : 0.7666
      Balanced Accuracy : 0.5746

       'Positive' Class : no
```

ROC

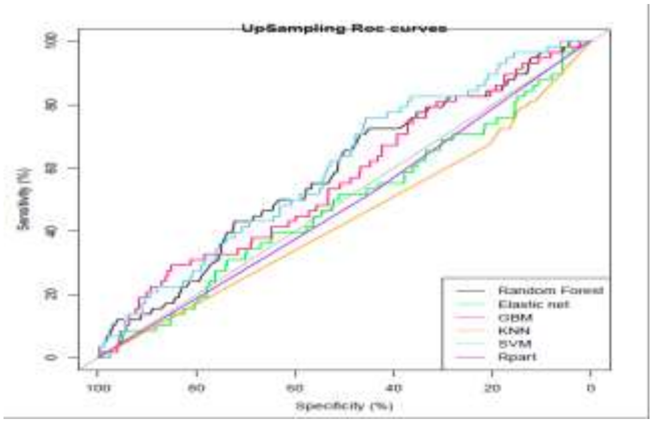


Figure 5

Conclusion:

- The project demonstrates the potential values of machine learning algorithms in improving near real-time Immune mediated AE and Myocarditis surveillance. Rapid signals can be detected by the methods can trigger timely investigation for underlying reasons.
- Liver metastasis was more likely to occur in patients with stage IV cancer and most commonly seen in patients aged 60 - 90.
- There is a strong correlation between the metastasis and resistance to treatment.
- We can implement the same analysis on any new program to assess the most frequent AE and its relationship with other factor that might help assessing the safety of the program and across programs
- Several limitation with the datasets, lack of complete data table, gaps in the data, not all attributes had sufficient information