CONTENT

1. ABSTRACT

I have implemented two supervised classification problem (Heart disease prediction and Springleaf Marketing Response) using machine learning algorithms. In heart disease prediction, we need to predict the status of heart disease which can be 0, 1, 2, 3 and 4. These 5 classes are not cleared by data dictionary but 0 means diameter narrowing is less than 50%.

In Springleaf marketing response, we have a large anonymous dataset (with 1933 features) to predict whether the customers likely to respond and be good candidates for their services.

2. Introduction

Machine learning is the science of getting computers to learn from data without being explicitly programmed. It is a study of pattern recognition, prediction algorithms and computational learning theory in artificial intelligence.

Here I have explained classified supervised learning.

Classified supervised learning is used to predict which class data point is part of (discrete value). Output is discrete, i.e.

True or false, yes or no, male or female, orange or apple, red or blue, 1 or 0, orange, apple, banana or none. The dependent feature i.e. the feature that we want to predict should be factor and level of factor can be more than two.

There are several machine learning algorithms to perform this tasks are:-

KNN(kth Nearest Neighbour), Naïve bayes, logistic regression, Neural networking, SVM (Support Vector Machine), GBM (Generalized Boosted Regression Model), LDA, etc. But when it comes to large data-set, we need an approach to solve our problem.

To explain both cases I have done 2 projects:-

1] **Heart Disease Diagnosis Prediction**:-

Heart disease dataset contains 4 dataset concerning heart disease diagnosis. Dataset contains 74 attributed from which only 14 are used and this was further reduced to 10. In this we need to predict whether a patient's heart status which have 5 possible values and it's unclear from the data dictionary what these 5 status means.

I have implemented lot of machine learning algorithms (Generalized boosted regression, linear discriminant analysis, random forest, logistic regression, support vector machine, etc.) and cross validate all results. The data is collected from 4 different location:-

- " Cleveland Clinic foundation"

- "Hungarian Institute of cardiology"

- "V.A Medical Centre, Long Beach "

- "University Hospital"

I have used Cleveland Clinic foundation dataset. We need to predict the diagnosis of heart disease which is either 0, 1, 2, 3 or 4(5 Classes)

0: <50% diameter narrowing

1: >50% diameter narrowing

But it contains value 0, 1, 2, 3, 4 which is 5 factors. It is unclear what these status means.

Number of attributes are 76 but I have used 14 main attributes that are:

-Age, sex(0:female & 1:male),

cp(Chest pain type):

1: for typical angina

2: for atypical angina

3: for non-Anginal pain

4: for Asymptomatic

Trestbps(Resting blood pressure), chol, fbs, restecg, thalach(max heart rate achieved), exang, oldpeak, slop, ca, thal(3 = normal; 6 = fixed defect; 7 = reversible defect) and num (the predicted attribute)

2] Springleaf Marketing Response:-

Springleaf offers their customers personal and auto loans that helps their customers to take control of their lives and finances. Springleaf's team connect with their customers through direct mail whom may be in need of a loan.

Direct mails provides huge value to customers who is in need of loan and it's a fundamental part of Springleaf's marketing strategy. Now here in order to improve their targeted efforts, our job is to predict the customers who are likely to respond and a good candidate for their services. Data provided is anonymous and large.

Using a large anonymised data-set, I predicted which customer will respond to a direct mail. A different approach is required to work with such a large dauntingly dataset.

3.1 Technical Overview

Before applying any machine learning algorithms, we need to do few things which is (my approach):-

- Dealing with NULL values by imputing values or omitting them.
- Data cleaning i.e. Cleaning/removing impurities from dataset for good result.
- Determining important features.
- Data Analysis
- Extracting and defining new useful features and eliminating less important features.
- Avoiding Overfitting (Smoothing and regularization is a solution for overfitting)

We need to take care of all these above approach.

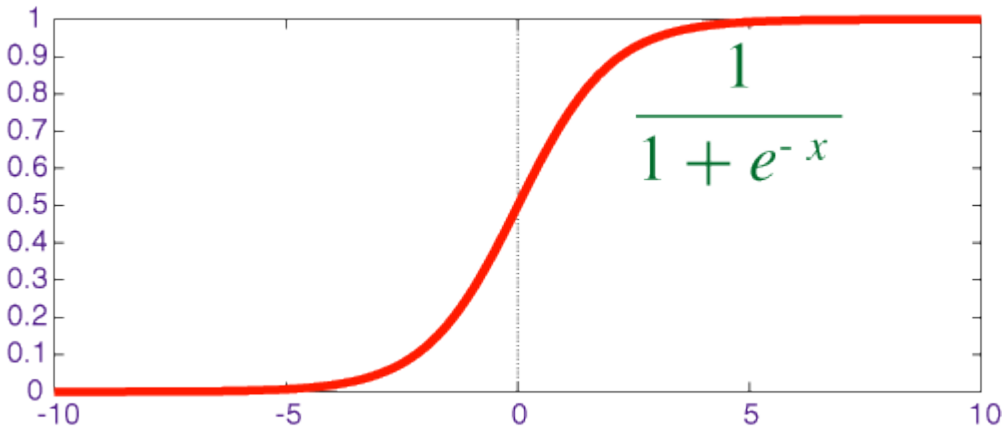
3.2 Algorithms: -

I will be using following machine learning algorithms to implement my projects

3.2.1 Multinomial Regression:-

Multinomial regression is a machine learning algorithm (generalizes logistic regression) that is used to predict classified/discrete feature like 1 or 0. Multinomial regression means multiple logistic regression. In logistic regression, dependent variable have only to factor which is 0 or 1. But in multinomial regression we can have more than two classes.

Unlike linear regression, here we don't fit a linear line, instead of using linear regression we use sigmoid function.



Where x = h(x)

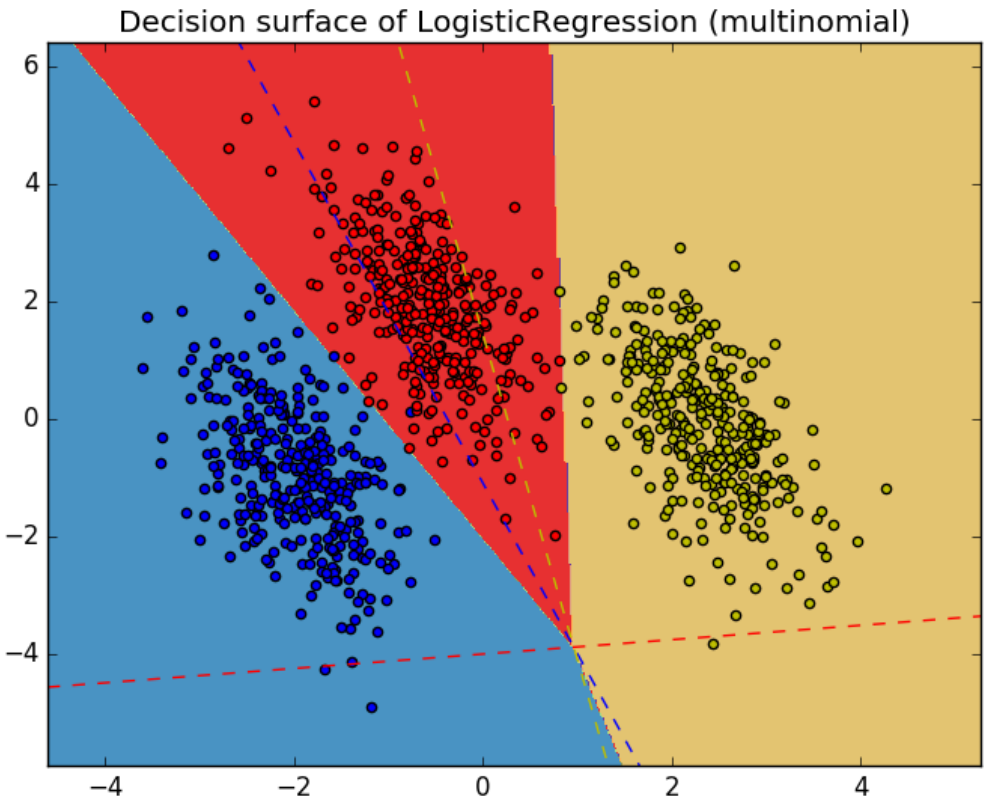
h(x) will be the equation containing constants and variable x(feature in data).

Example: h(x) =c0+ c1*x1 + c2*x2 + c3*x3 + ……. + cn^2 * xn^2 + ……

Where x1, x2 ,……xn are features in data i.e. thal, sex, chest pain, etc. and c0,c1,….cn are constant.

Constants will be determined by optimising cost function to its minimum value.

Cost function, g(x) = - (1/m)* ∑ [ylog(h(x)) +(1-y)log(1-h(x))]



Now here we have three classes for which we will find probability for each classes. The class with high probability will be consider as final result.

3.2.3 Random Forest :-

Random Forest is a machine learning ensemble classifier which uses many decision tree model to take decision.

What is ensemble model?

Ensemble model combines the result from different models and then take final decision by voting by very individual model. The result from ensemble model is usually better than the result from one individual model.

In random forest we take sample of n rows (n=no. of rows in training data-set) and m features (m<no. of features in training dataset) with replacement, selected at random from the original data.

The best split on these m is used to split the feature or we can say node. The value of 'm' is no of features used for sampling from main data. The value of m is held constant during the forest growing.

Each tree is grown to the largest extend possible.

3.2.4 Naïve Bayes:-

The naïve Bayes classifier is based on Bayesian theorem which is suited when the dimensionality of the input dataset is high i.e. no. of features are high. In Naïve Bayes, we assume that all features are independent of each other

In this algorithm we calculate the probability for example in titanic survival, we will calculate probability of passenger being survived if passenger's class is 1, probability of passenger being survived if passenger's class is 2, probability of passenger being survived if passenger's gender's is male and so on for every features provided in training dataset.

Naïve Bayes is widely used like in spam classifier, probability of email being spam if it contains word 'free' or '$' or 'reward'.

$P(A|B) = [P(A) * P(B|A)] / P(B)$

Example:-

$P(Status=2|cp=1)=[P(Status=2) * P(cp=1|status=2)] / P (cp=1)$

Naive Bayes is easy to implement it is predict class of test data set fast and perform well when input dataset with high dimensions.

Limitation of Naïve Bayes is that we assume that all features are independent which is almost impossible in real world that we get a set of predictor which are completely independent.

3.2.5 Gradient Boosting Modelling (GBM) :-

It is based on ensemble learning which is weighting combination of predictors.

Focus on new learners on example that other get wrong. It trains learners sequentially. In this algorithm, we first do a simple fitting or we can say prediction and then we calculate error residual. Then we fit and learn from error residual and then combine that with previous result. This gives a better predicting model. We keep doing this until out cost function reaches it's minimum value.

This is called gradient boosting because it uses gradient descent optimization techniques to optimize cost function.

Cost function $c(x) = 1/2m [\sum (y - y(i))^2]$

Uses simple regression model to start. Subsequent models predict the error residual of the previous prediction to minimize cost/error residual for next prediction. Overall prediction is given by weighted sum of collections and votes.

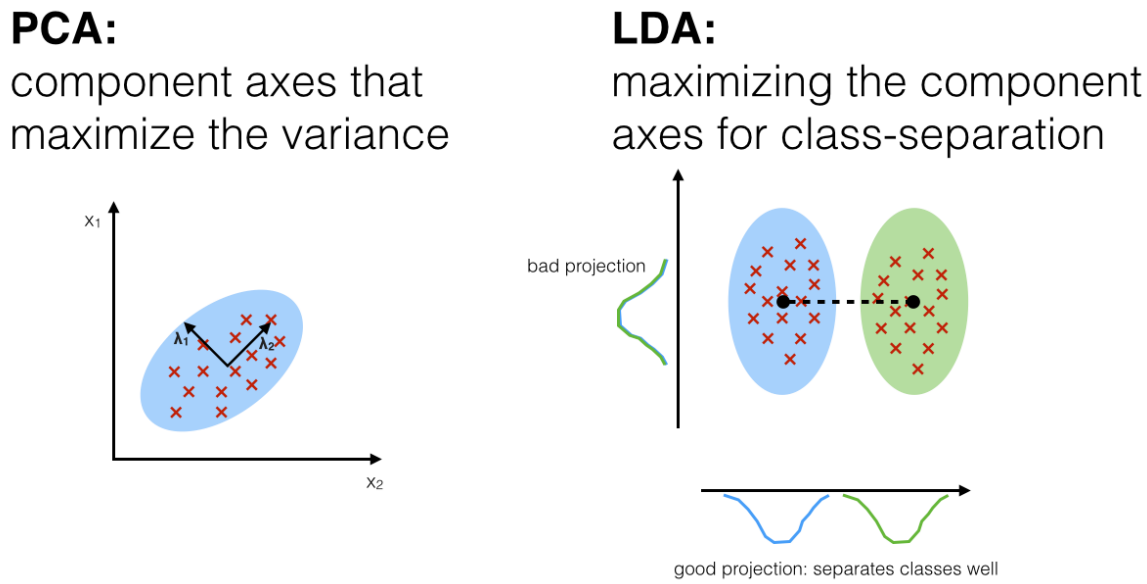
3.2.6 Linear Discriminant Analysis (LDA) :-

Linear Discriminant analysis is most commonly used to reduce dimension of the input data while at the same time preserving as much as of the class discrimination information as possible. It's helpful when we have lot of features in our dataset. It provides a lower dimensional space with good class separation.

It is similar to principle component analysis. It finds the component axes to minimize distances of every feature from the projection line (it's different then linear regression) same as PCA but additionally it also finds the axes that maximizes the separation between classes.

- It will first reduce the dimension of data i.e. chest pain type (cp), sex, , age, …., n to kth dimension where k<n.

- Then it will makes the cluster and will find the axes that maximizes the separation between these two clusters i.e. inter-cluster distance must be maximum.

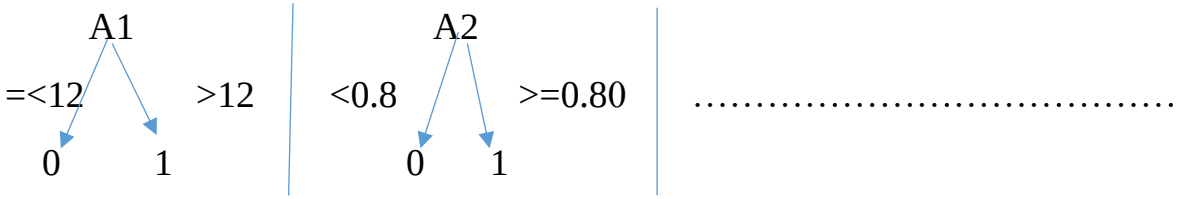


3.2.7 XGBOOST (eXtreame Gradient Boosting) :-

It is a scalable and efficient implementation of gradient boosting framework. The reason behind that it is faster is that it uses the concept of parallelism. The algorithm do boosting parallel unlike gradient boosting.

It has both linear model solver and tree learning model. It's capacity to do parallel computation in single machine makes this algorithm much faster i.e. at least 10 times faster than existing gradient boosting and compared to other machine learning algorithms.

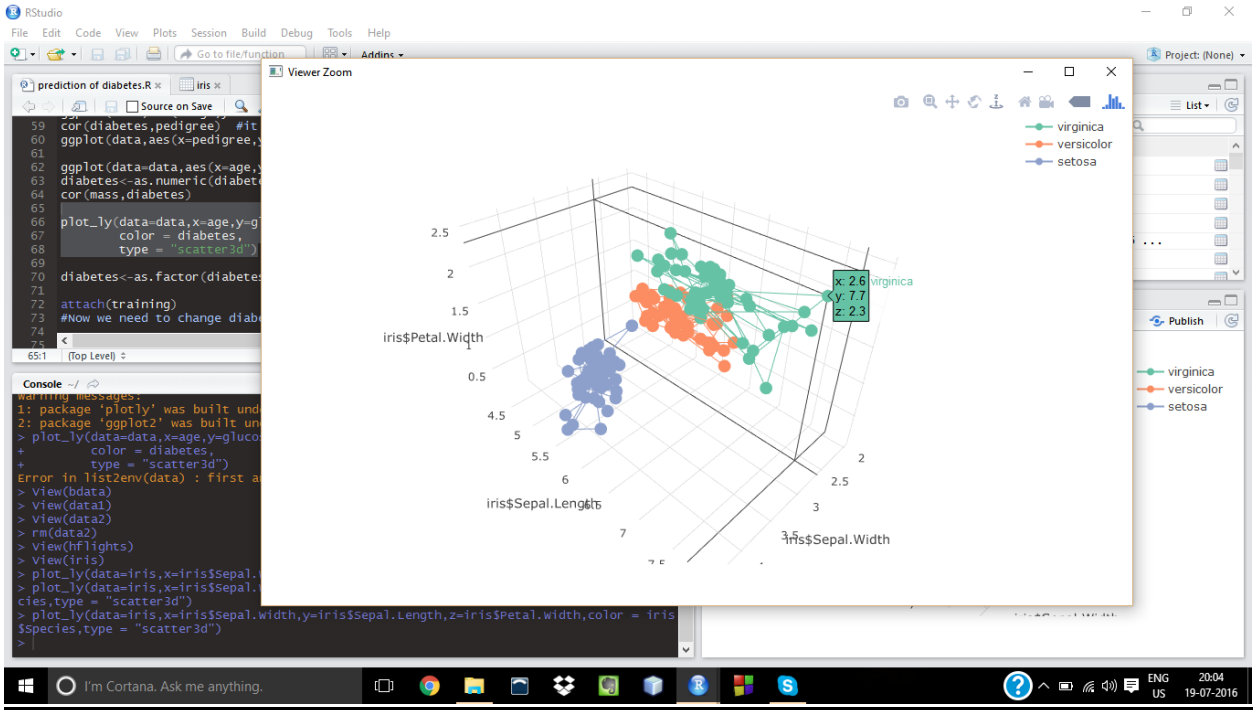
Many number of weak learners/predictors gives a better prediction after combining that a one complex predictor.



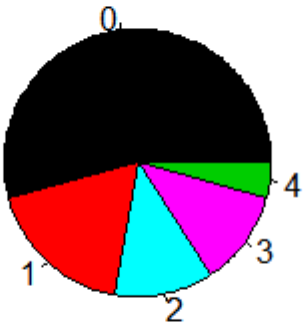
All these weak/ simple predictor votes and the for the answer and the maximum voted answer is said to be the final prediction.

4. Tool Used :-

R Studio

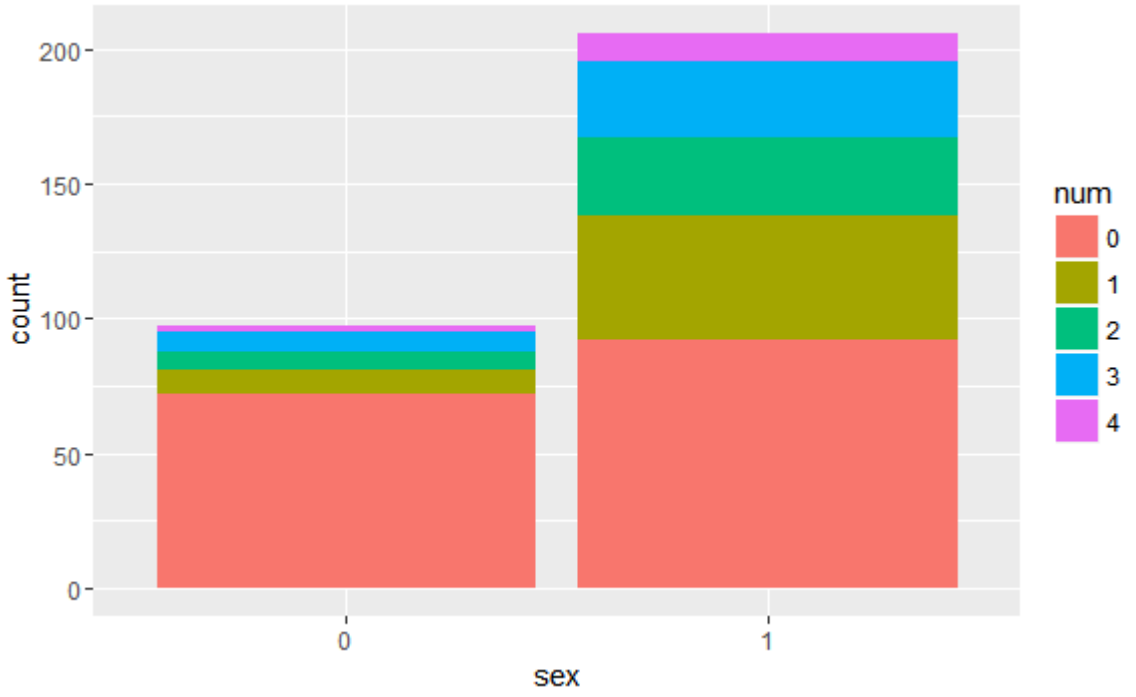


5. Results and Reports :-

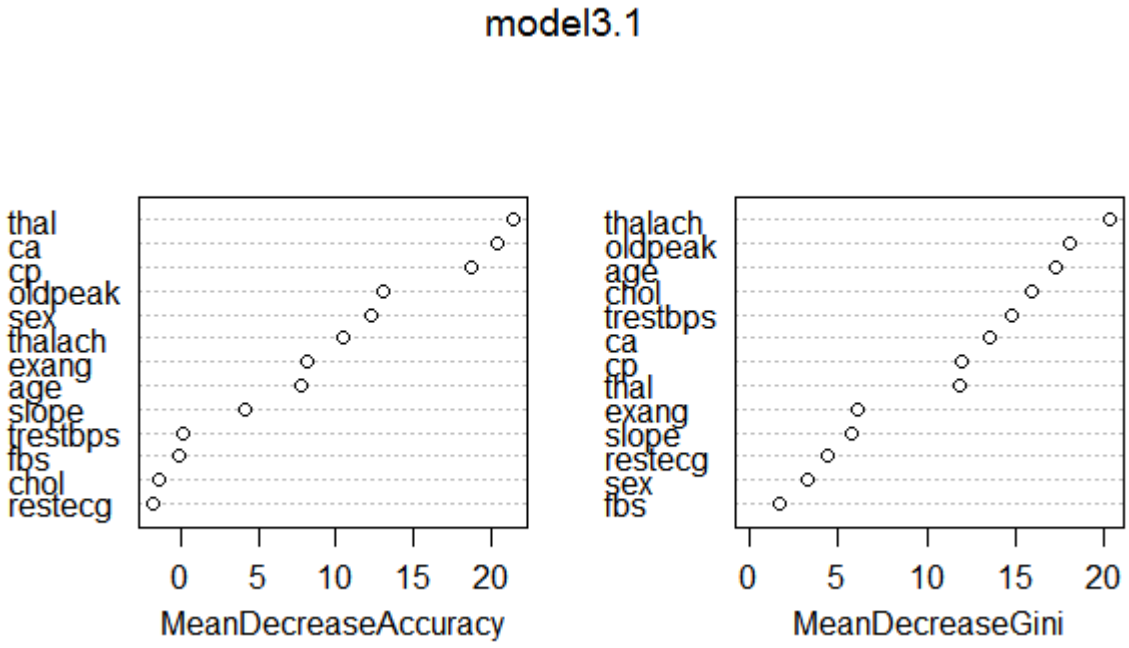


Over more than 54 % of patient's heart status is 0 which means their heart is less than 50 % diameter narrowing.

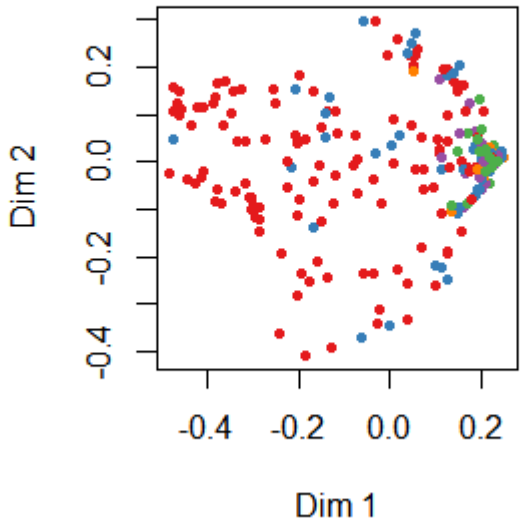
0	1	2	3	4
0.54125413	0.18151815	0.11881188	0.11551155	0.04290429



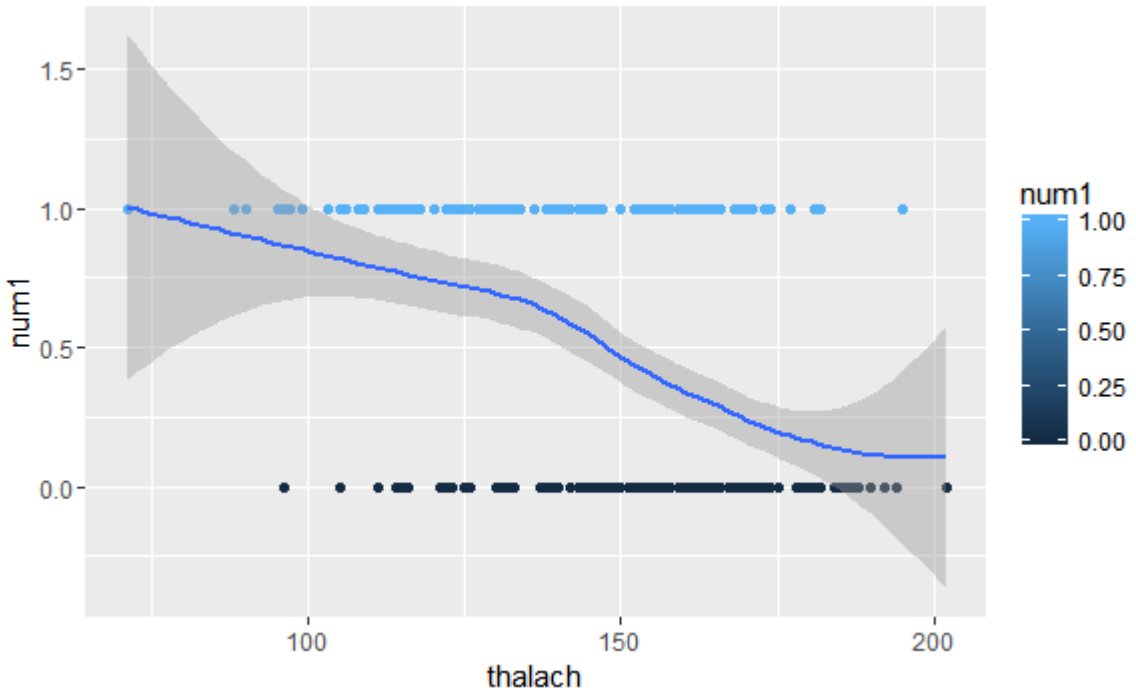
This bar chart shows that probability of having heart status equal to 0 is more in female as compared to male.



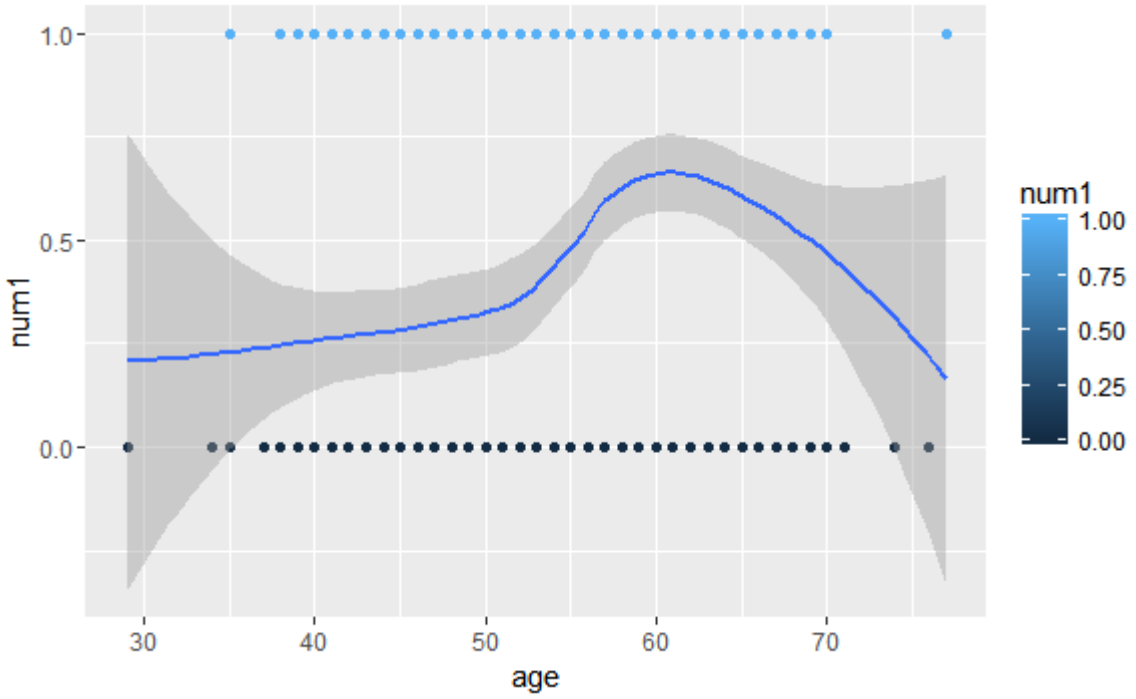
This is a variable importance plot that shows how important a variable is for predictive analysis. MeanDecreaseAccuracy shows the importance of variable in predictive analysis. In this case, thal(3 = normal; 6 = fixed defect; 7 = reversible defect) is most important variable and then ca serum (cholesterol in mg/dl)



This is called multi-dimensional scaling plot of proximity matrix and it is generated by random forest's model. A matrix of proximity measures among the input (based on the frequency that pairs of data points are in the same terminal node). It has reduced the dimension of data to two dimension to understand data. As we can see it doesn't look good as clusters are not separated good enough therefore will effect accuracy of prediction.

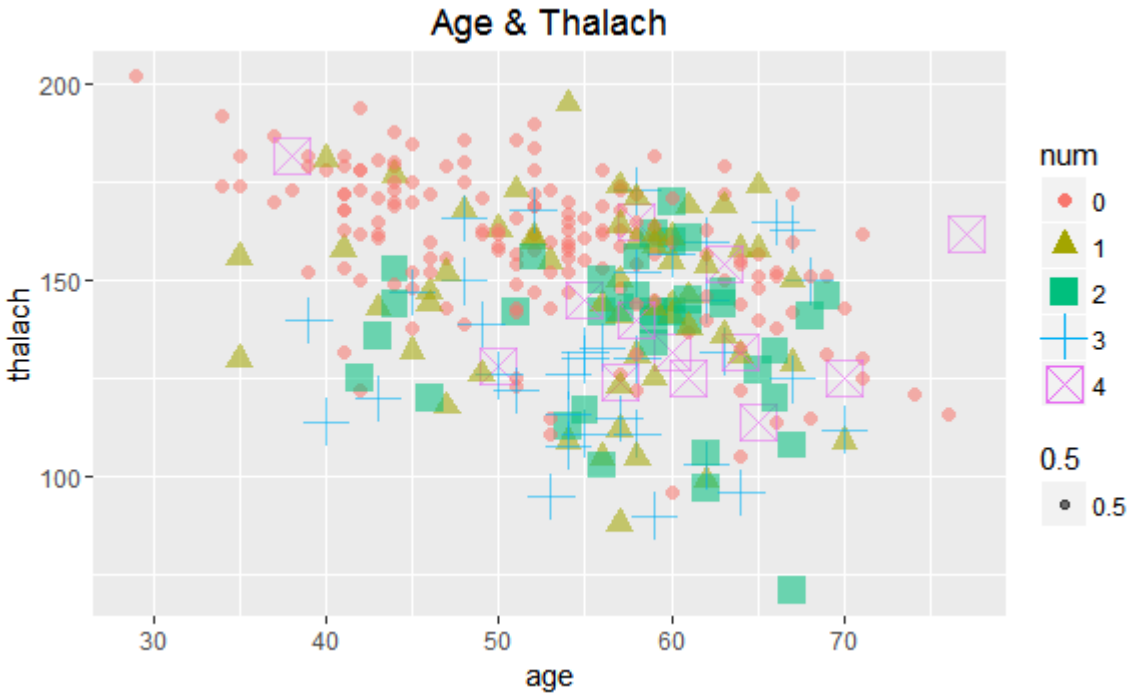


This is a correlation curve (I have modified num(dependent) features as 0 or 1 as num1) which shows how heart disease (num1) depends on thalach. This shows negative correlation which means probability of status being 0 decreases with increase of thalach in patient.

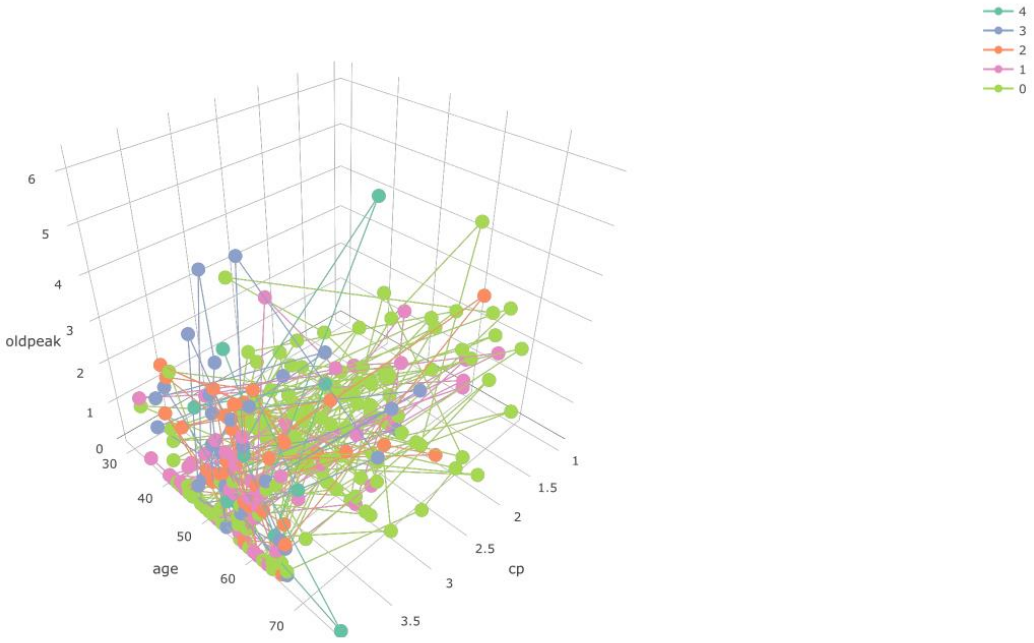


This is a correlation curve (I have modified num(dependent) features as 0 or 1 as num1) which shows how heart disease (num1) depends on age.

Here we can't say if it's a negative or a positive correlation but age between 50 to 60, it's a positive correlation and from 60 to 60 above it's a negative correlation.

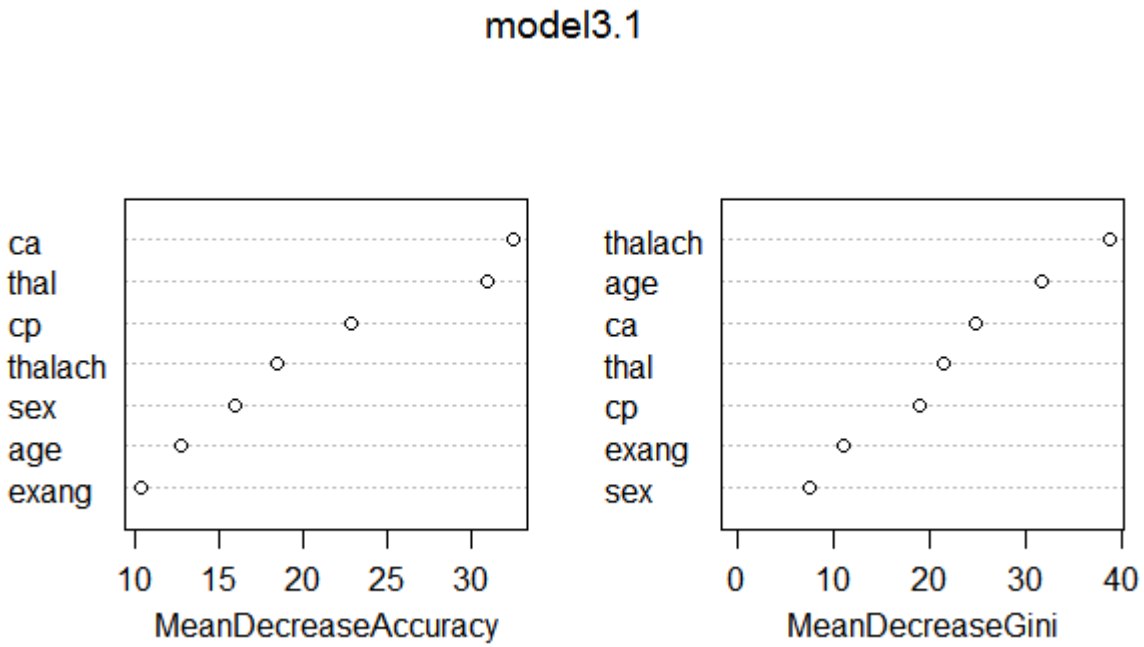


This shows how heart disease status depends on thalach and age of the patient. The status of heart disease can be distinguish by color and shape. As we can see red dots are denser when thalach (y-axis) increases.



This 3-D plot shows how status of heart disease depends on age, oldpeak and chest pain type. Green dots are more denser when chest pain type is 2 and 3 and is also dense when cp is 1 but not much denser when chest pain type is 4. So we

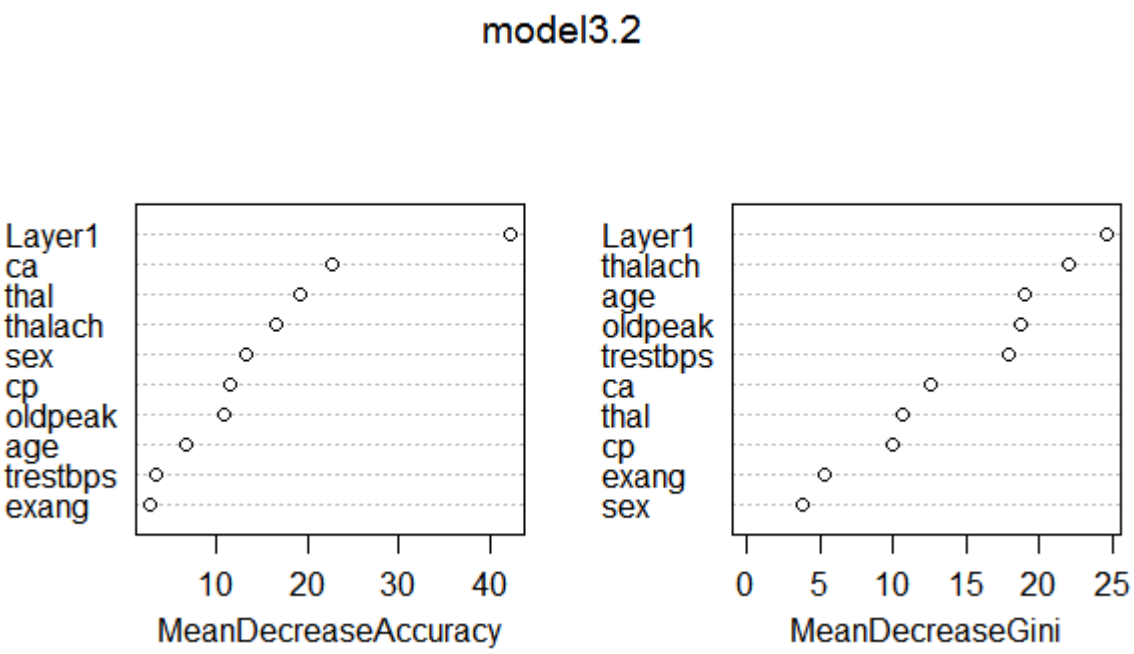
can say that when chest pain is 2 and 3 than probability of heart disease status being 0 is more



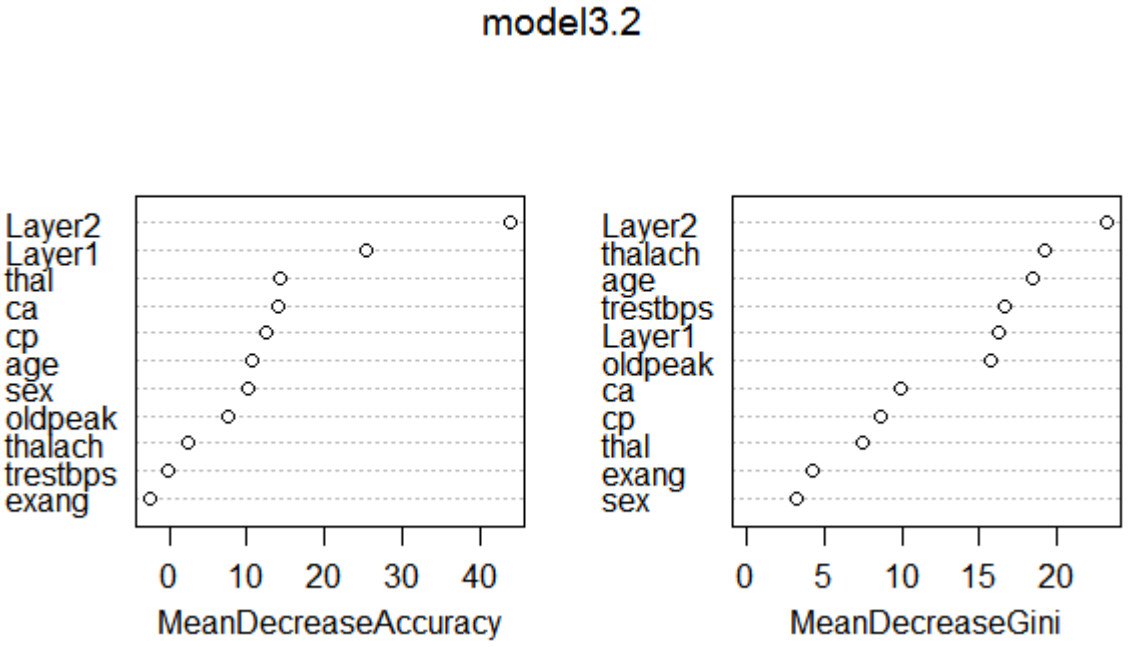
This is a variable importance plot of features that I have eliminated. These are top important variable among which ca(no. of major vessels(0-3)) is top in variable importance plot followed by thal, cp(chest pain type), thalach, sex, age and exang. As we have 5 classes here(0, 1, 2, 3, 4), getting high accuracy is a challenge.

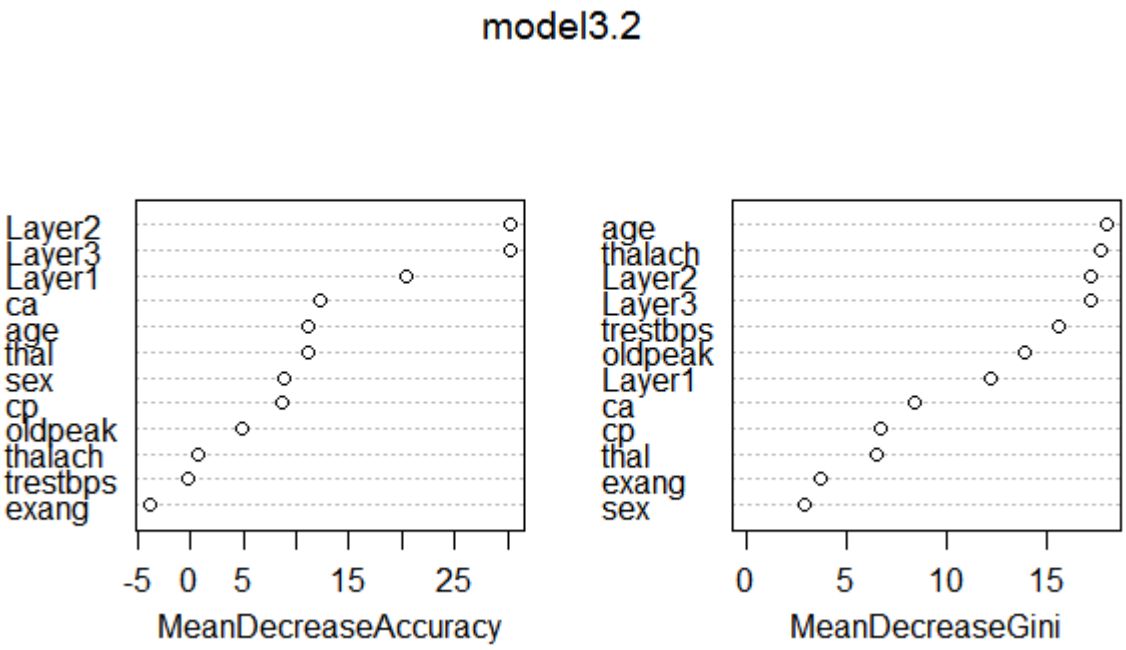
But these variable are not enough to to get accuracy above 70 %. So here I added my way of increasing accuracy which is called making layers.

I have compared results of accuracy of prediction without layer, with one layer and with two layers.



After adding layer, we can see Layer1 is at the top of the important variable plot which will for sure increase our accuracy. I have also added Layer2 and Layer3 and the variable importance plot are:-





I haven't performed data analysis for Springleaf marketing response because data set are anonymous and it has lot of features (1933 features/columns)

Confusion matrix with lda (Maximum accuracy obtained)

	References				
Prediction	**0**	**1**	**2**	**3**	**4**
0	39	2	1	1	0
1	1	6	0	1	1
2	0	0	6	1	0
3	0	5	2	5	0
4	1	0	0	0	2

Heart Disease Prediction Accuracy (By training 75% data)

Algorithm	**Accuracy (Without Layer)**	**Accuracy (1 Layer)**	**Accuracy (2 Layers)**	**Accuracy (3 Layers)**
Random Forest	63.51 %	67.567 %	70.27 %	74.324 %
Gradient Boosting	63.51 %	68.92 %	68.92 %	71.62 %
Multinomial Regression	58.11 %	67.56%	67.6%	68.92 %
Naïve Bayes	55.41 %	63.51 %	73 %	74.324 %
Linear Discriminant Analysis	63.51 %	68.91%	**78.38 %**	**78.38 %**

Springleaf Marketing Response

Algorithm	Accuracy
eXtream Gradient Boosting	75.32%
Naïve Bayes	74.24%

6. CONCLUSION

In Heart Disease diagnosis we have achieved the highest accuracy of 78.38% from linear discriminant analysis using two layers and achieved same accuracy using three layers.

In Springleaf marketing response, I performed only two algorithms as data provided is very large with lot of features. As xgboost is very faster (10 times than existing gradient boosting) and Naïve Bayes is recommended when we have lot of features, therefore I have implemented these two algorithms. Here xgboost just won by getting 75.32 % accuracy where Naïve Bayes scored 74.24 %.

7. Implications for Future Research

I will try to get more accuracy and will work on my layers concept and will implement more projects and will see how this layers concept is working out and how can I get more better predictive accuracy. Also working with big data Hadoop and will take this to the higher level by implementing machine learning algorithms on big data using parallelism by map-reduce or spark (I'm already done with data analysis on big data).