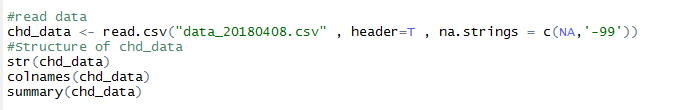
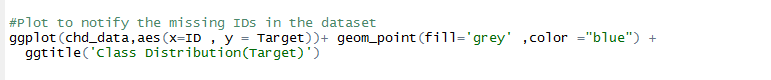
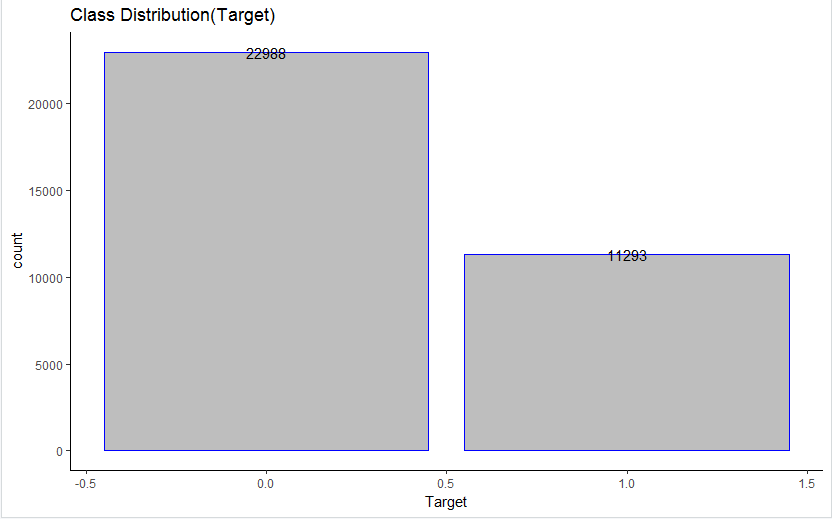
1. Read the CSV file accepting missing values as NA and –99 Values using na.strings and check the Structure, summary of Data.

Here na.strings replaces the missing values with NA and -99.



1. Checking the data distribution in the target variable i.e. how many id are available without heart disease and with disease.

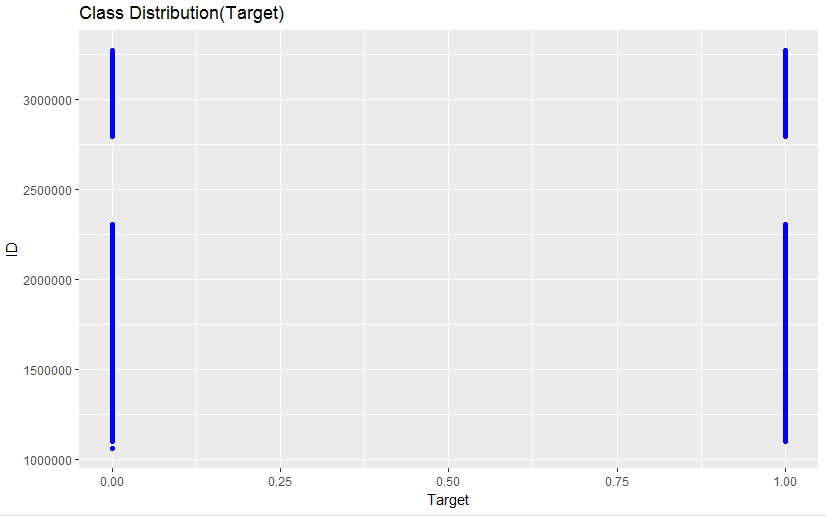




1. Checking is there any gap between the target variable with respect to ID’s

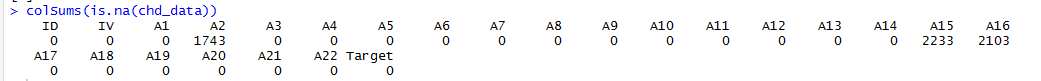
It can be observed that approximately there is gap in the ID’s from graph.





1. Checking the missing values in each column

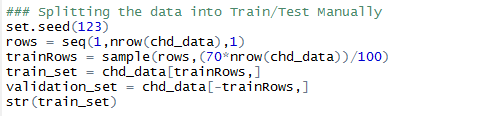
colSums(is.na(chd\_data))



1. ID column has no use in the model building and hence removing it along with column A11 as it contains constant value.

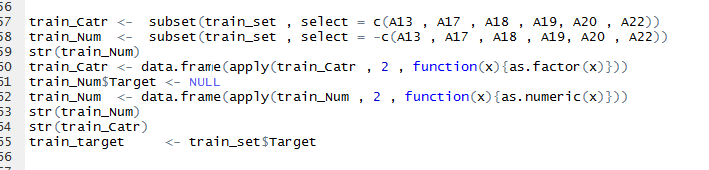


1. Splitting the data into train and test data in the ratio 70:30

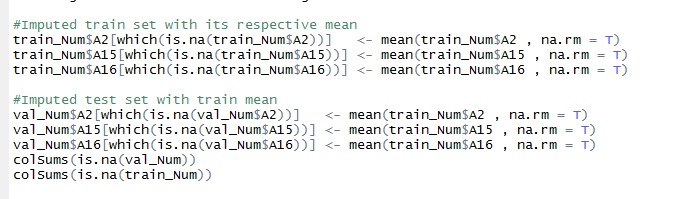


Here we use set.seed to get the reproducible random numbers i.e. when we sample it randomly picks the value but inorder to reproduce the same result we use set.seed

1. Divide train data into categorical and Numerical Similarly Test Data



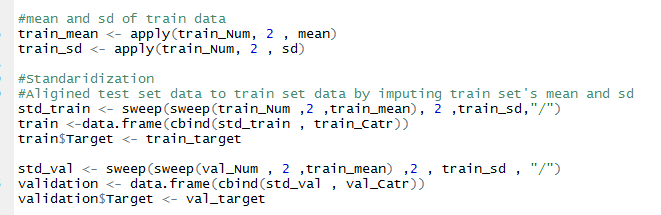
1. Imputing Missing values NA and –99 for train data and Test data separately



1. Checking correlation between independent variable

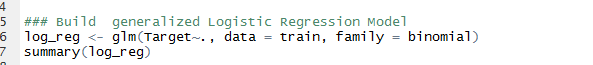
i.e. here correlation is more than one variable giving same information about target variable i.e. gives the strength and direction between 2 variable.

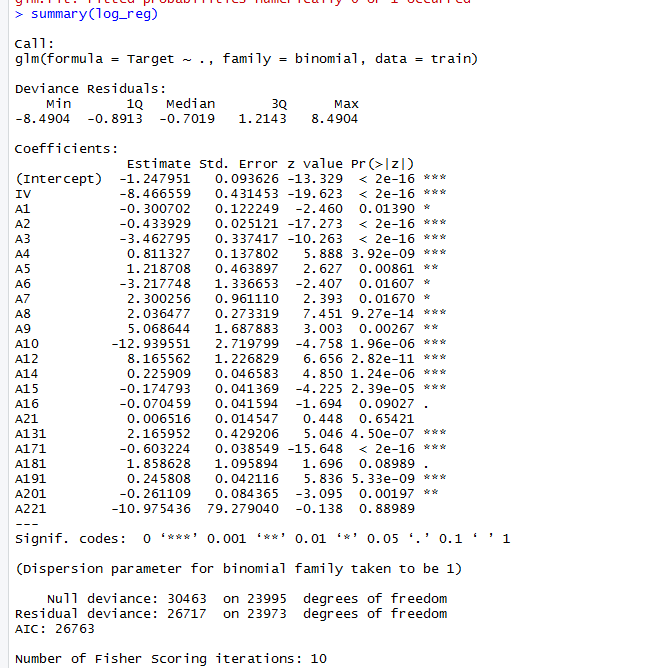
1. Standardization of data i.e. train and test data separately (test data will use mean and sd of train data)



**Building Model**

1. **Build a GLM**





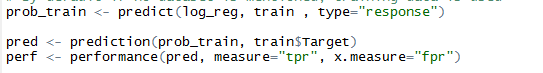
Number of Fisher Scoring Iteration: 10—it is the iteration take to find the minima(maximum likelihood).

Deviance Residuals: Unexplained variation between actual value and predicted value.

Family=Binomial-used for logistic regression as dependent variable will be TRUE or FALSE(YES or NO)(1 or 0)

AIC-Used to remove multicollinearity and finding the best model.

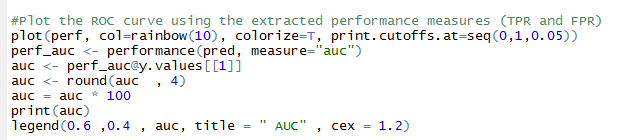
1. Predict values i.e. Y^

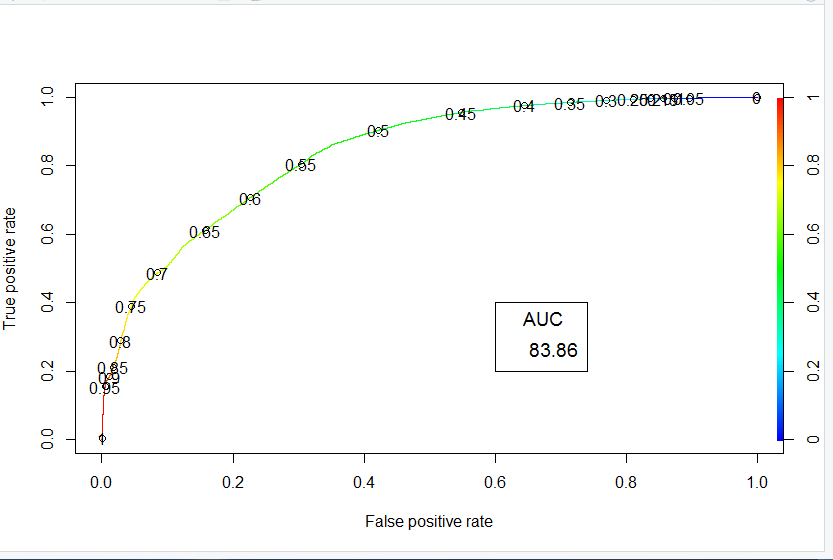


Type=calculate dependent variable

1. Plot The ROC Curve

What is ROC Curve

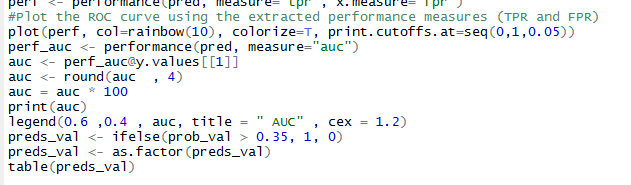


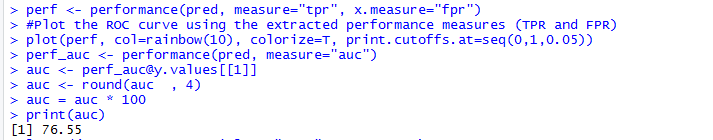


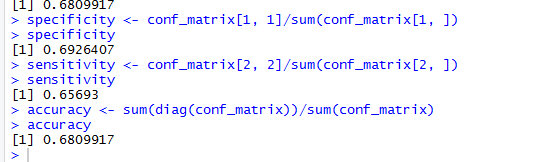
Meaning of below Code:

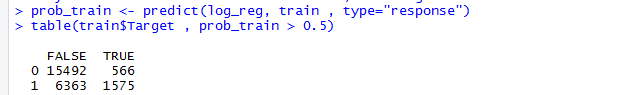


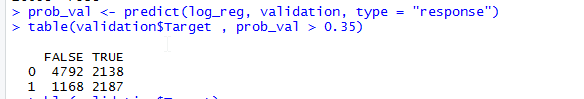
1. Perform the same on Test data and observe results



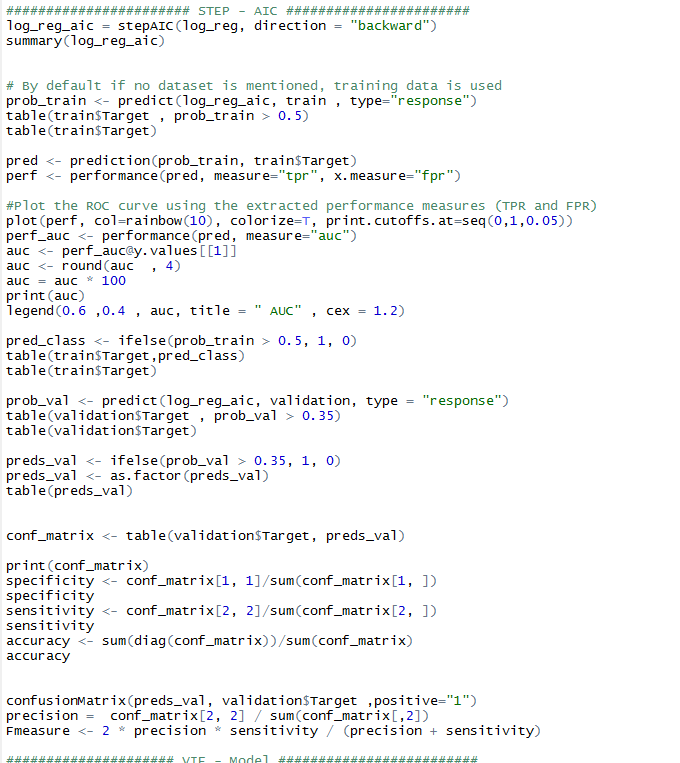




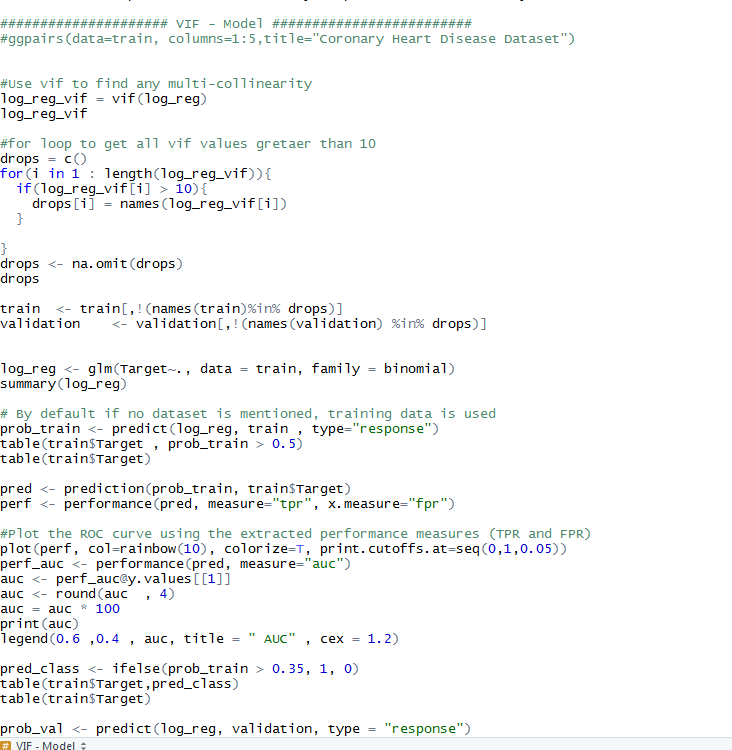


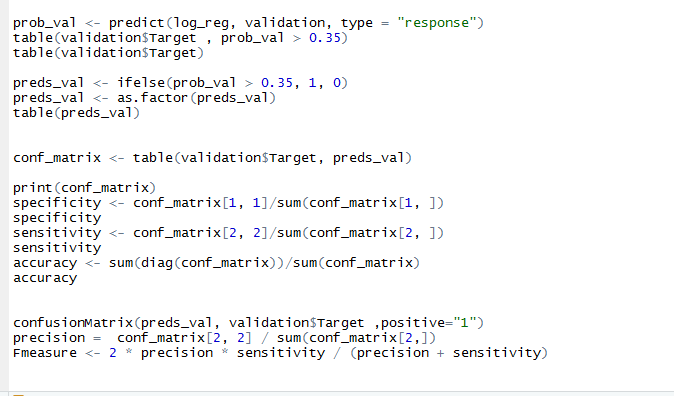


MODEL =2(AIC)



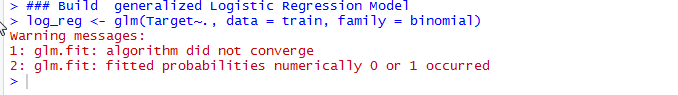
MODEL using VIF: still gives the same result as that of AIC





**Warning Message:**

If you have a variable which perfectly separates zeroes and ones in target variable, R will yield the following "perfect or quasi perfect separation" warning message:.



It can be eliminated by increasing the iteration using maxit=1000.

**SMOTE:**

With SMOTE-synthetic minority oversampling technique

In simple words, instead of replicating and adding the observations from the minority class, it overcome imbalances by generates artificial data. It is also a type of oversampling technique.

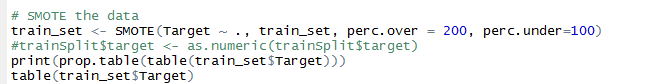
In regards to synthetic data generation, synthetic minority oversampling technique (SMOTE) is a powerful and widely used method. SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples. We can also say, it generates a random set of minority class observations to shift the classifier learning bias towards minority class.

To generate artificial data, it uses bootstrapping and k-nearest neighbors. Precisely, it works this way:

1. Take the difference between the feature vector (sample) under consideration and its nearest neighbor.
2. Multiply this difference by a random number between 0 and 1
3. Add it to the feature vector under consideration
4. This causes the selection of a random point along the line segment between two specific features

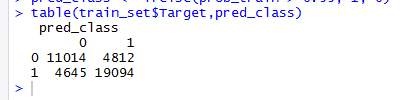
R has a very well defined package which incorporates this techniques. We’ll look at it in practical section below.





Confusion matrix:

Train Data:



Test data:

