Objective 1 - Checking Assumptions:

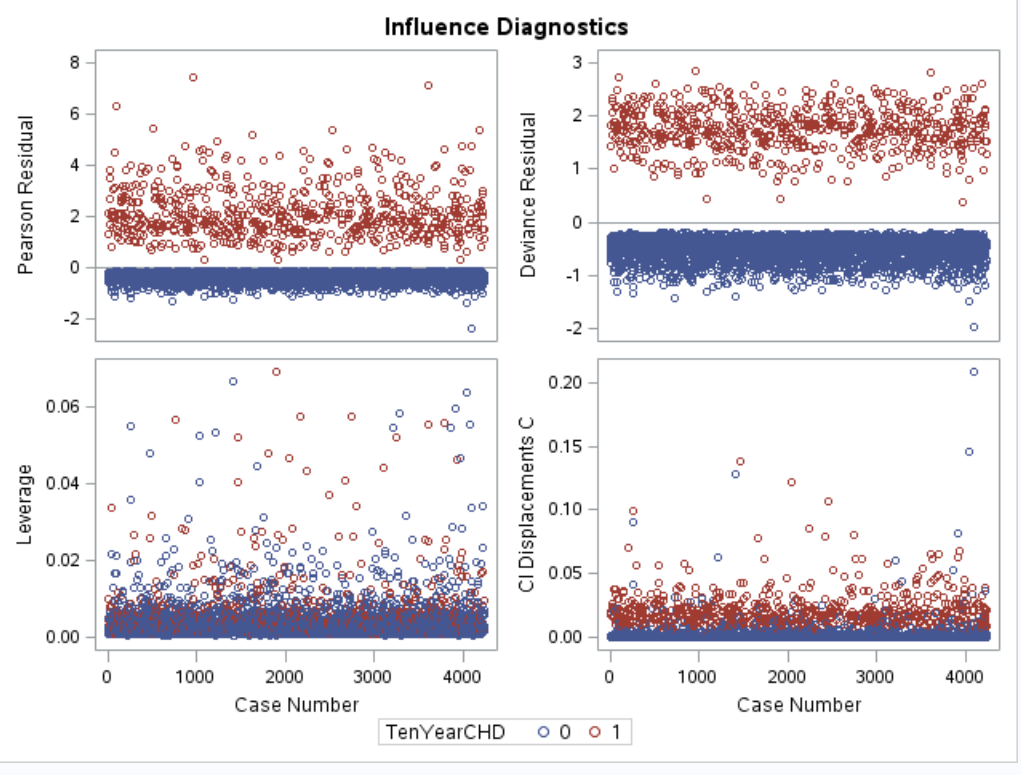
Using the following residual charts(fig xx) we determined that additional EDA is required to ensure usefulness for predictive model analysis.

* Pearson Residual
* Deviance Residual Observations with a deviance residual in excess of two may indicate lack of fit. Additional testing below confirm that there
* Leverage
* CI Displacement C -

Overall the residuals show that the data is

The Pearson and Deviance Residual plots show values greater than two and this may indicate lack of fit for the data population that was positive (value ‘1’) for Coronary Heart Disease (CHD). In charts below, the CHD = ‘1’ have approximately 30% of data population above two.

The Leverage Plot shows that the data is symmetrically distributed and there are no outliers of concern. The CI Displacement C chart does show a few potential influence data observations but reviewing the Leverage it does not warrant reviewing the observation from the predictive modeling analysis.

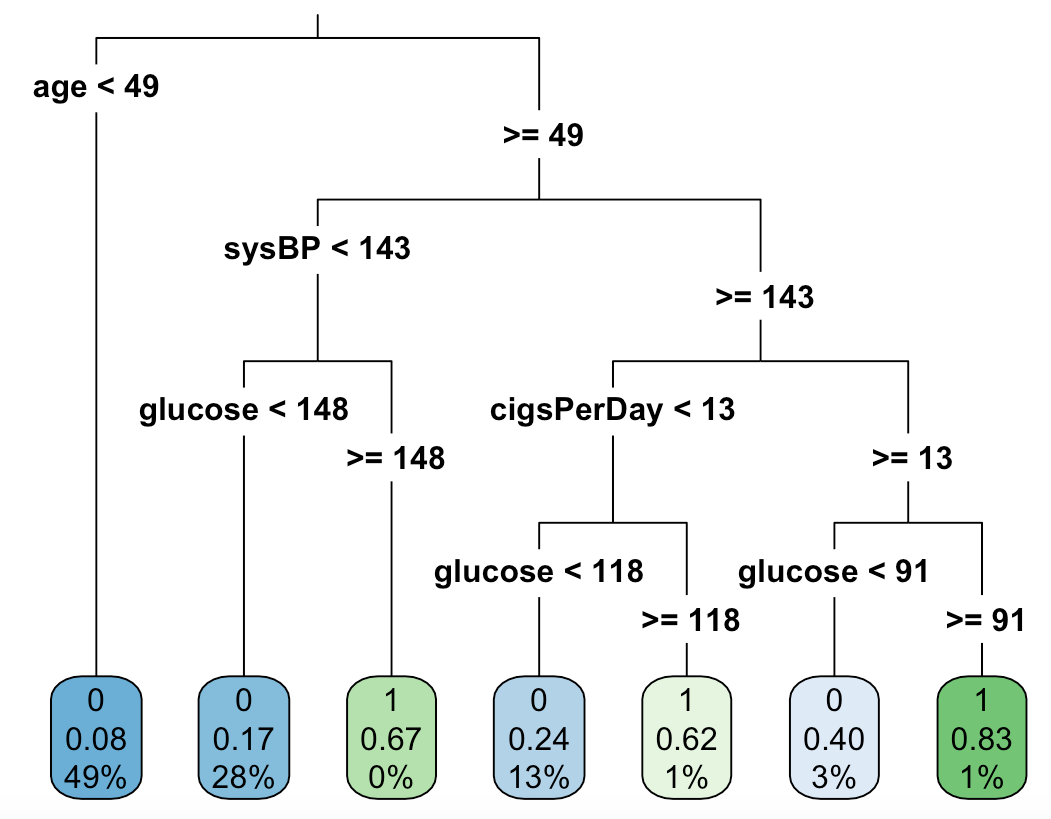


<<Mike and Spencer --- I believe we need to follow the above with the results from the Lack of Fit done in R. It will be helpful that we can confirm that Lack of Fit is not a concern or as part of conducting EDA we addressed the Lack of Fit issue in the data that is used in study.>>

Objective 2 – Decision Tree:

<<Mike & Spencer – I suggest we use the following at or near the beginning of the write-up for Model 2. I believe this helps flow from Objective 1 EDA and helps with outlining approach used for modeling in Objective 2 >>

As part of the predictive modeling analysis, in addition to conducting basic modeling analysis as show in EDA, we create a Decision Tree to help with identifying valid response variables to utilize in study. Below are the Decision Tree (code in appendix) results:

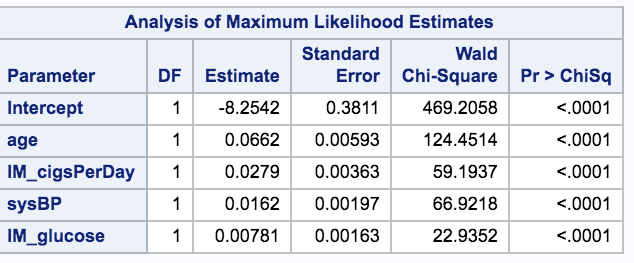


As shown in the Decision Tree we note that there are no categorical variables in results. The tree indicates that age, sysBP, cigsPerDay and glucose are highly predictive potentially to indicate the likelihood of CHD. The Tree’s response variables are used in LDA analysis below to attempt to produce the most optimal predictive model

<<Michael & Spencer -- If memory serves we were going to compare two top models. I believe one of the top models is the LDA with the Decision Tree variable. If you have two better modeling results please proceed with using the better models. If the below model is one of the better it will be helpful since it will help flow from the above analysis. I am good either way.>>

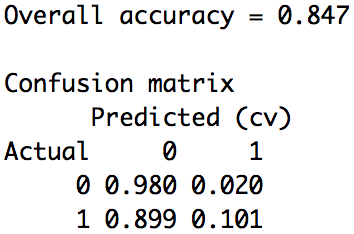
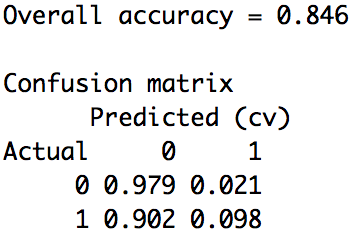
LDA Modeling:

Below is the LDA and Logistic modeling using Decision Tree results. As part of the analysis ran a quick logistic modeling test and the following regression model was created. All the response variables in regression show significant likelihood for predicting CHD. All p-values are < .0001.



The following confusion matrix shows a very high overall accuracy of 84.7 percent. That results compare very favorably against the other test models.

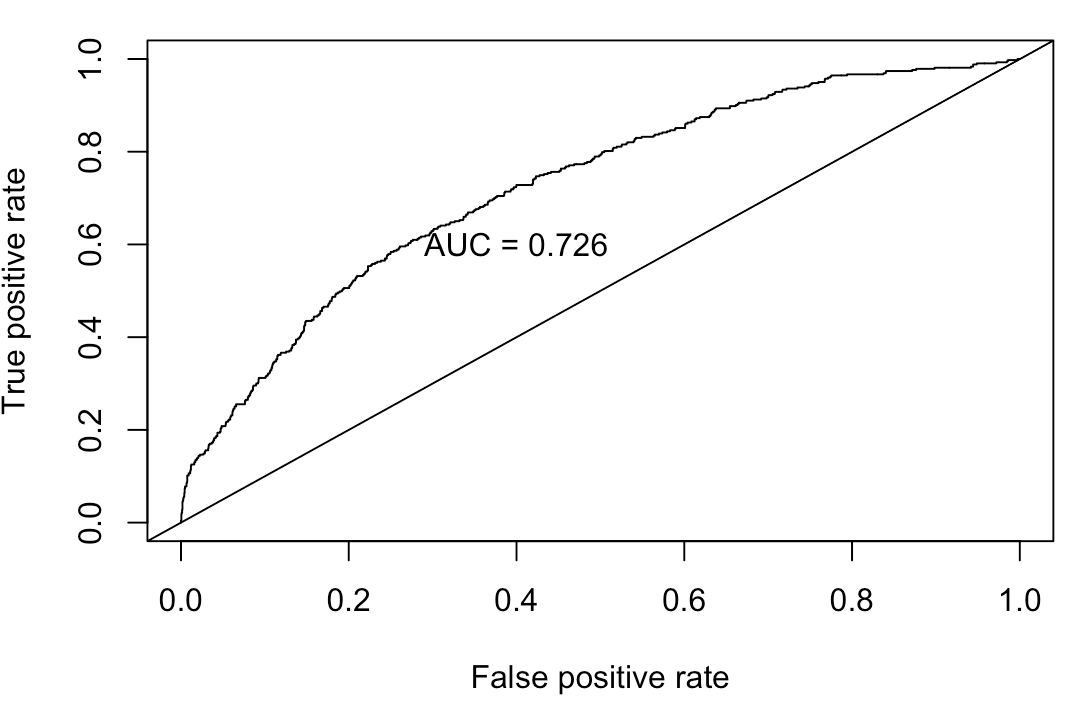
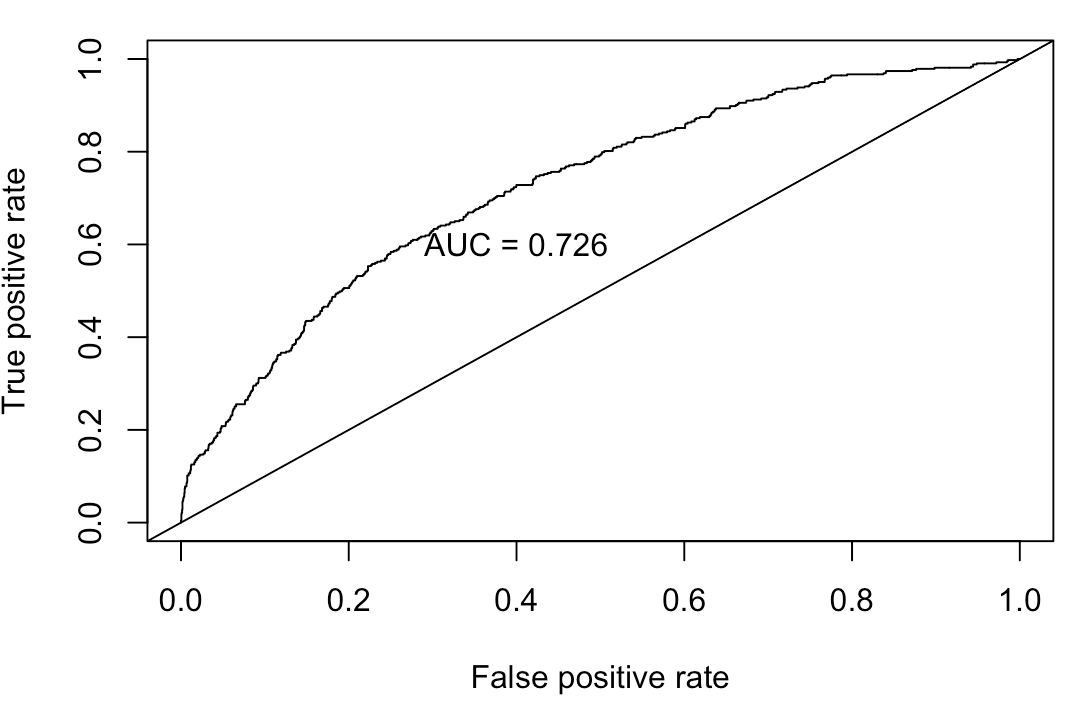
**LDA with Decision Tree variables LDA with All Variables**

As shown above, it is worth noting that even with LDA model with all the data population variables produced a high overall accuracy of 84.6%.

Below are the comparisons of AUC for both Decision Variables and All Variables. Both had the same AUC results:

**LDA with Decision Tree variables LDA with All Variables**

In addition, numerous tests with interactive variables on both models were conducted and all produced lower overall accuracy results. None of the results that included interaction variables were worth including in final analysis documentation.

Tom’s – Conclusion --- not need to include in completed write-up.

In conducting study of the response variables outlined in EDA, it was determined that for simplicity purposes utilizing the LDA model with Decision Tree variables would be most effective at predicting CHD with a test accuracy of 84.7% <<need CI>>.

The CHD prediction is assumed to be a random study of both cigarette and non-cigarette smokers and predictive model can be utilized for detecting likelihood of CHD for overall population which the study population originated.

Appendix:

## LDA , confusion matrix and ROC performance graph --- all the variables

#Training Set

dat.train <- heartData[1:3000,]

dat.train.x <- dat.train[,1:15]

dat.train.y <- dat.train$TenYearCHD

dat.train.y <- as.factor(as.character(dat.train.y))

dat.test <- heartData[3001:nrow(heartData),]

dat.test.x <- dat.train[,1:15]

dat.test.y <- dat.test$TenYearCHD

dat.test.y <- as.factor(as.character(dat.test.y))

fit.lda <- lda(dat.train.y ~ ., data = dat.train.x)

pred.lda <- predict(fit.lda, newdata = dat.test)$class

testhat <- predict(fit.lda, newdata = dat.test)

confusion(dat.test$TenYearCHD, testhat$class)

## LDA , confusion matrix and ROC performance graph --- with Decision Tree variables

#Training Set

dat.train <- heartData[1:3000,]

dat.train.x <- dat.train[,1:15]

dat.train.y <- dat.train$TenYearCHD

dat.train.y <- as.factor(as.character(dat.train.y))

dat.test <- heartData[3001:nrow(heartData),]

dat.test.x <- dat.train[,1:15]

dat.test.y <- dat.test$TenYearCHD

dat.test.y <- as.factor(as.character(dat.test.y))

fit.lda <- lda(dat.train.y ~ age+sysBP+cigsPerDay+glucose, data = dat.train.x, CV=True)

pred.lda <- predict(fit.lda, newdata = dat.test)$class

testhat <- predict(fit.lda, newdata = dat.test)

confusion(dat.test$TenYearCHD, testhat$class)

preds <- pred.lda$posterior

preds <- as.data.frame(preds)

pred <- prediction(preds[,2],dat.train.y)

roc.perf = performance(pred, measure = "tpr", x.measure = "fpr")

auc.train <- performance(pred, measure = "auc")

auc.train <- auc.train@y.values

plot(roc.perf)

abline(a=0, b= 1)

text(x = .40, y = .6,paste("AUC = ", round(auc.train[[1]],3), sep = ""))

# Decision tree model

term\_vars <- c("Gender","age", "education","currentSmoker", "cigsPerDay", "BPMeds", "prevalentStroke", "prevalentHyp", "diabetes", "totChol", "sysBP", "diaBP", "BMI", "heartRate", "glucose", "TenYearCHD")

CHD\_term\_train <- heartData[1:3000,]

CHD\_term\_test <- heartData[3001:nrow(heartData),]

set.seed(99) # set a pre-defined value for the random seed so that results are repeatable

rpart\_model <- rpart(TenYearCHD ~.,

data = CHD\_term\_train[term\_vars],

method = 'class',

parms = list(split='information'),

control = rpart.control(usesurrogate = 0,

maxsurrogate = 0))

pred.lda <- predict(rpart\_model, newdata = CHD\_term\_test)

# Plot the decision tree

rpart.plot(rpart\_model, roundint = FALSE, type = 3)

/\* simple logistic regression --- all variables \*/

proc logistic data=CHD\_clean;

class male IM\_education currentSmoker IM\_BPMeds prevalentStroke prevalentHyp diabetes/ param=ref;

model TenYearCHD(event='1')= male age IM\_education currentSmoker IM\_cigsPerDay IM\_BPMeds prevalentStroke prevalentHyp diabetes IM\_totChol sysBP diaBP IM\_BMI IM\_heartRate IM\_glucose/ scale=none lACkfit;

effectplot slicefit(sliceby=male plotby=IM\_BPMeds) / noobs;

run;

/\* simple logistic regression --- Decision Tree variables \*/

proc logistic data=CHD\_clean;

class male IM\_education currentSmoker IM\_BPMeds prevalentStroke prevalentHyp diabetes/ param=ref;

model TenYearCHD(event='1')= age IM\_cigsPerDay sysBP IM\_glucose/ scale=none lACkfit;

effectplot slicefit(sliceby=male plotby=IM\_BPMeds) / noobs;

run;