# Feature Engineering

Feature selection being done based on information value. Following R code was used to generate the information value for each of the variable.

Variables with information between 0.1 and o.5 are generally the good predictors.

Libraries-“devtools”,”woe”,”tomasgreif”

iv.mult(data,”response,”TRUE”)

iv.plot.summary(iv.mult(data,”response,”TRUE”))

Weight of Evidence (WoE) was calculated for each of the independent variables at all levels.

options(digits=2)  
iv.mult(data,"gb",vars=c("var1","var2",……,”varn”))

# Data Partitioning

Data was divided as training and validation sets in the ratio of 70:30. The following R code was used for data partitioning.

sub<-sample(nrow(data),floor(nrow(data)\*.70))

training<-data[sub,]

validation<-data[-sub,]

# Model Development

A full model (all variables included) was built to start with. Post full model development, stepwise algorithm was used to arrive at the most parsimonious model. Akaike Information Criteria (AIC) was used to select the model and the model with the lowest AIC value will qualify as the final model. The R codes for the full model and the stepwise model are given below.

logit\_fullmodel<-glm(Churn~gender+SeniorCitizen+Partner+Dependents+tenure\_cat+PhoneService+MultipleLines+InternetService+OnlineSecurity+OnlineBackup+DeviceProtection+TechSupport+StreamingTV+StreamingMovies+Contract+PaperlessBilling+PaymentMethod+MonthlyCharges\_cat+TotalCharges\_cat,data=train,family="binomial")

logit\_reducedmodel<-stepAIC(logit\_fullmodel,direction="backward")

using “MASS” library

The final model code is as below:

Logit\_finalmodel<-glm(formula = Churn ~ SeniorCitizen + Dependents + tenure\_cat +MultipleLines + InternetService + OnlineSecurity + OnlineBackup + TechSupport + StreamingTV + StreamingMovies + Contract + PaperlessBilling + PaymentMethod + TotalCharges\_cat, family = "binomial", data = train)

# Validate Model

The model thus developed was validated and the predictions were doe on the test data. Post predictions, the diagnostic measures like Area Under the Curve (AUC) and KS measures were computed. The details are as below. The AUC was 83.74% which indicated the goodness of fit of the model.

The area under the curve was computed using pROC and ROCR packages.

rocobj<-roc(probabilities\_predicted\_test$Churn,probabilities\_predicted\_test$predict\_testdata)

confusionMatrix(predicted,class)