Models

Bank <- read.csv("~/Desktop/MSA\_6440/customer\_churn/data/processed/BankChurners\_filtered.csv")  
source('/Users/JonathanVoth/Desktop/R Codes/myfunctions.R')  
  
bank <- Bank[,c(-1,-2,-10)]  
  
RNGkind (sample.kind = "Rounding")

## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler  
## used

set.seed(0)  
p2 <- partition.2(bank, 0.7) # 70:30 partition  
training.data <- p2$data.train  
test.data <- p2$data.test

## Logistic Regression

# Logistic model with all variables  
log.model <- glm(Attrition\_Flag ~ ., family = binomial(link='logit'), data = training.data)  
summary(log.model)

##   
## Call:  
## glm(formula = Attrition\_Flag ~ ., family = binomial(link = "logit"),   
## data = training.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1799 0.0669 0.1772 0.3662 3.0069   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.052e+00 6.615e-01 -10.661 < 2e-16 \*\*\*  
## Customer\_Age 2.854e-02 1.112e-02 2.565 0.010310 \*   
## GenderM 7.311e-01 2.034e-01 3.594 0.000325 \*\*\*  
## Dependent\_count -9.162e-02 4.306e-02 -2.128 0.033371 \*   
## Education\_LevelDoctorate -7.316e-01 2.708e-01 -2.702 0.006898 \*\*   
## Education\_LevelGraduate -2.654e-02 1.858e-01 -0.143 0.886411   
## Education\_LevelHigh School -1.746e-01 1.970e-01 -0.886 0.375464   
## Education\_LevelPost-Graduate -3.537e-01 2.759e-01 -1.282 0.199922   
## Education\_LevelUneducated -1.899e-01 2.062e-01 -0.921 0.357074   
## Marital\_StatusMarried 6.989e-01 2.010e-01 3.478 0.000506 \*\*\*  
## Marital\_StatusSingle 5.768e-02 2.015e-01 0.286 0.774652   
## Income\_Category$40K - $60K 7.697e-01 2.756e-01 2.793 0.005224 \*\*   
## Income\_Category$60K - $80K 4.753e-01 2.415e-01 1.968 0.049065 \*   
## Income\_Category$80K - $120K 4.140e-02 2.229e-01 0.186 0.852693   
## Income\_CategoryLess than $40K 4.813e-01 2.959e-01 1.627 0.103813   
## Months\_on\_book -4.050e-03 1.093e-02 -0.371 0.710845   
## Total\_Relationship\_Count 4.604e-01 3.971e-02 11.594 < 2e-16 \*\*\*  
## Months\_Inactive\_12\_mon -5.484e-01 5.551e-02 -9.879 < 2e-16 \*\*\*  
## Contacts\_Count\_12\_mon -4.623e-01 5.184e-02 -8.919 < 2e-16 \*\*\*  
## Credit\_Limit 1.039e-05 8.311e-06 1.251 0.211055   
## Total\_Revolving\_Bal 9.706e-04 1.047e-04 9.272 < 2e-16 \*\*\*  
## Avg\_Open\_To\_Buy NA NA NA NA   
## Total\_Amt\_Chng\_Q4\_Q1 1.939e-01 2.661e-01 0.729 0.466179   
## Total\_Trans\_Amt -4.900e-04 3.209e-05 -15.267 < 2e-16 \*\*\*  
## Total\_Trans\_Ct 1.167e-01 5.235e-03 22.294 < 2e-16 \*\*\*  
## Total\_Ct\_Chng\_Q4\_Q1 2.656e+00 2.735e-01 9.710 < 2e-16 \*\*\*  
## Avg\_Utilization\_Ratio 4.159e-01 3.577e-01 1.163 0.245006   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4301.7 on 4956 degrees of freedom  
## Residual deviance: 2308.6 on 4931 degrees of freedom  
## AIC: 2360.6  
##   
## Number of Fisher Scoring iterations: 6

## CV Logistic Model

print(step\_cv)

## Generalized Linear Model with Stepwise Feature Selection   
##   
## 4957 samples  
## 18 predictor  
## 2 classes: 'Attrited Customer', 'Existing Customer'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 4461, 4461, 4461, 4461, 4461, 4462, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9031659 0.5904387

step\_cv$finalModel

##   
## Call: NULL  
##   
## Coefficients:  
## (Intercept) Customer\_Age   
## -6.8187405 0.0246414   
## GenderM Dependent\_count   
## 0.7297593 -0.0892104   
## Education\_LevelDoctorate Marital\_StatusMarried   
## -0.6181186 0.6396325   
## `Income\_Category$40K - $60K` `Income\_Category$60K - $80K`   
## 0.6513966 0.3824741   
## `Income\_CategoryLess than $40K` Total\_Relationship\_Count   
## 0.3819285 0.4600760   
## Months\_Inactive\_12\_mon Contacts\_Count\_12\_mon   
## -0.5457039 -0.4603949   
## Total\_Revolving\_Bal Total\_Trans\_Amt   
## 0.0010677 -0.0004878   
## Total\_Trans\_Ct Total\_Ct\_Chng\_Q4\_Q1   
## 0.1164674 2.7298641   
##   
## Degrees of Freedom: 4956 Total (i.e. Null); 4941 Residual  
## Null Deviance: 4302   
## Residual Deviance: 2315 AIC: 2347

# Confusion matrix for test data  
prob.test <- predict(step\_cv, newdata = test.data, type = "prob")  
y.test <- ifelse(prob.test[,2] > 0.8, "Existing Customer", "Attrited Customer")  
confusionMatrix(as.factor(y.test), as.factor(test.data$Attrition\_Flag), positive = "Existing Customer")

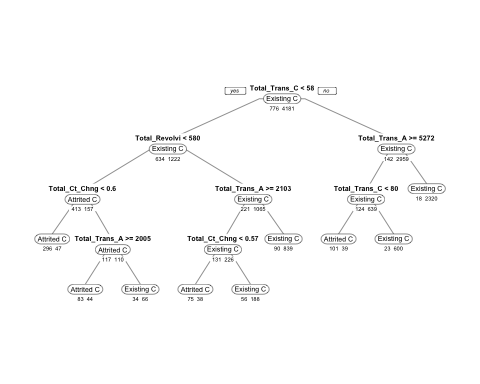
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Attrited Customer Existing Customer  
## Attrited Customer 267 218  
## Existing Customer 70 1569  
##   
## Accuracy : 0.8644   
## 95% CI : (0.8491, 0.8787)   
## No Information Rate : 0.8413   
## P-Value [Acc > NIR] : 0.001677   
##   
## Kappa : 0.5689   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.8780   
## Specificity : 0.7923   
## Pos Pred Value : 0.9573   
## Neg Pred Value : 0.5505   
## Prevalence : 0.8413   
## Detection Rate : 0.7387   
## Detection Prevalence : 0.7717   
## Balanced Accuracy : 0.8351   
##   
## 'Positive' Class : Existing Customer  
##

## Decision Tree Model

library(rpart)  
library(rpart.plot)  
library(caret)  
  
# Decision tree model with CV  
set.seed(0)  
train\_control <- trainControl(method = "cv", number = 10)   
tree\_cv <- train(Attrition\_Flag ~ ., data = training.data, method = "rpart", trControl = train\_control, tuneLength = 10, metric = "Kappa", control = rpart.control(minsplit = 150, minbucket = 100))  
# tuneGrid = data.frame(cp = seq(0, 0.01, 0.001)  
  
print(tree\_cv)

## CART   
##   
## 4957 samples  
## 18 predictor  
## 2 classes: 'Attrited Customer', 'Existing Customer'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 4461, 4461, 4461, 4461, 4461, 4462, ...   
## Resampling results across tuning parameters:  
##   
## cp Accuracy Kappa   
## 0.006443299 0.9049841 0.6335781  
## 0.007087629 0.9049841 0.6335781  
## 0.009020619 0.9041760 0.6269033  
## 0.012886598 0.9045793 0.6264613  
## 0.013745704 0.9045793 0.6222396  
## 0.018685567 0.9023615 0.6177340  
## 0.025773196 0.9025635 0.6205321  
## 0.039948454 0.8997430 0.5871590  
## 0.070876289 0.8944958 0.5467765  
## 0.164948454 0.8827900 0.4324316  
##   
## Kappa was used to select the optimal model using the largest value.  
## The final value used for the model was cp = 0.007087629.

# Decision tree rules plot  
prp(tree\_cv$finalModel, type = 1, extra = 1, under = TRUE, split.font = 2, varlen = -10)

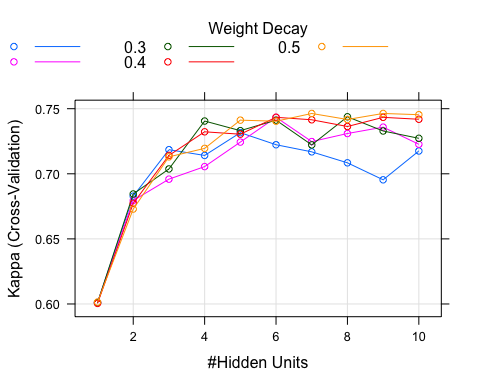


# Evaluation on the test data  
prob.test <- predict(tree\_cv, test.data, type = 'prob')  
y.test <- ifelse(prob.test[,2] > 0.8, "Existing Customer", "Attrited Customer")  
confusionMatrix(as.factor(y.test), as.factor(test.data$Attrition\_Flag), positive = 'Existing Customer')

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Attrited Customer Existing Customer  
## Attrited Customer 278 217  
## Existing Customer 59 1570  
##   
## Accuracy : 0.8701   
## 95% CI : (0.855, 0.8841)   
## No Information Rate : 0.8413   
## P-Value [Acc > NIR] : 0.0001161   
##   
## Kappa : 0.5911   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.8786   
## Specificity : 0.8249   
## Pos Pred Value : 0.9638   
## Neg Pred Value : 0.5616   
## Prevalence : 0.8413   
## Detection Rate : 0.7392   
## Detection Prevalence : 0.7669   
## Balanced Accuracy : 0.8517   
##   
## 'Positive' Class : Existing Customer  
##

## Neural Network

# Plot  
plot(cv.nn)



# Best parameters  
cv.nn$bestTune

## size decay  
## 45 9 0.5

# Evaluation on the test data  
prob.test <- predict(cv.nn, test.data, type = 'prob')  
y.test <- ifelse(prob.test[,2] > 0.8, "Existing Customer", "Attrited Customer")  
confusionMatrix(as.factor(y.test), as.factor(test.data$Attrition\_Flag), positive = 'Existing Customer')

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Attrited Customer Existing Customer  
## Attrited Customer 299 158  
## Existing Customer 38 1629  
##   
## Accuracy : 0.9077   
## 95% CI : (0.8946, 0.9197)   
## No Information Rate : 0.8413   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.698   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9116   
## Specificity : 0.8872   
## Pos Pred Value : 0.9772   
## Neg Pred Value : 0.6543   
## Prevalence : 0.8413   
## Detection Rate : 0.7669   
## Detection Prevalence : 0.7848   
## Balanced Accuracy : 0.8994   
##   
## 'Positive' Class : Existing Customer  
##