gbm(formula = bookings ~ ., distribution = "multinomial", data = t[train,

], n.trees = 300, interaction.depth = 2, shrinkage = 0.001,

cv.folds = 5)

A gradient boosted model with multinomial loss function.

300 iterations were performed.

The best cross-validation iteration was 300.

There were 17 predictors of which 3 had non-zero influence.

var rel.inf

lat\_destination lat\_destination 66.0221547

destination\_language destination\_language 33.5148812

sum\_secs\_elapsed sum\_secs\_elapsed 0.4629641

signup\_method signup\_method 0.0000000

signup\_flow signup\_flow 0.0000000

language language 0.0000000

affiliate\_channel affiliate\_channel 0.0000000

affiliate\_provider affiliate\_provider 0.0000000

first\_affiliate\_tracked first\_affiliate\_tracked 0.0000000

signup\_app signup\_app 0.0000000

first\_device\_type first\_device\_type 0.0000000

first\_browser first\_browser 0.0000000

counts counts 0.0000000

lng\_destination lng\_destination 0.0000000

distance\_km distance\_km 0.0000000

destination\_km2 destination\_km2 0.0000000

language\_levenshtein\_distance language\_levenshtein\_distance 0.0000000

gbm(formula = bookings ~ ., distribution = "multinomial", data = t[train,

], n.trees = 75, interaction.depth = 2, shrinkage = 0.001,

cv.folds = 5)

A gradient boosted model with multinomial loss function.

75 iterations were performed.

The best cross-validation iteration was 75.

There were 17 predictors of which 3 had non-zero influence.

var rel.inf

lat\_destination lat\_destination 62.7264098

destination\_language destination\_language 36.8311856

sum\_secs\_elapsed sum\_secs\_elapsed 0.4424046

signup\_method signup\_method 0.0000000

signup\_flow signup\_flow 0.0000000

language language 0.0000000

affiliate\_channel affiliate\_channel 0.0000000

affiliate\_provider affiliate\_provider 0.0000000

first\_affiliate\_tracked first\_affiliate\_tracked 0.0000000

signup\_app signup\_app 0.0000000

first\_device\_type first\_device\_type 0.0000000

first\_browser first\_browser 0.0000000

counts counts 0.0000000

lng\_destination lng\_destination 0.0000000

distance\_km distance\_km 0.0000000

destination\_km2 destination\_km2 0.0000000

language\_levenshtein\_distance language\_levenshtein\_distance 0.0000000

gbm(formula = bookings ~ ., distribution = "multinomial", data = t[train,

], n.trees = 50, interaction.depth = 2, shrinkage = 0.001,

cv.folds = 5)

A gradient boosted model with multinomial loss function.

50 iterations were performed.

The best cross-validation iteration was 50.

There were 17 predictors of which 3 had non-zero influence.

var rel.inf

lat\_destination lat\_destination 60.4125868

destination\_language destination\_language 39.1576868

sum\_secs\_elapsed sum\_secs\_elapsed 0.4297264

signup\_method signup\_method 0.0000000

signup\_flow signup\_flow 0.0000000

language language 0.0000000

affiliate\_channel affiliate\_channel 0.0000000

affiliate\_provider affiliate\_provider 0.0000000

first\_affiliate\_tracked first\_affiliate\_tracked 0.0000000

signup\_app signup\_app 0.0000000

first\_device\_type first\_device\_type 0.0000000

first\_browser first\_browser 0.0000000

counts counts 0.0000000

lng\_destination lng\_destination 0.0000000

distance\_km distance\_km 0.0000000

destination\_km2 destination\_km2 0.0000000

language\_levenshtein\_distance language\_levenshtein\_distance 0.0000000

gbm(formula = bookings ~ ., distribution = "multinomial", data = t[train,

], n.trees = 300, interaction.depth = 2, shrinkage = 1e-04,

cv.folds = 5)

A gradient boosted model with multinomial loss function.

300 iterations were performed.

The best cross-validation iteration was 300.

There were 17 predictors of which 3 had non-zero influence.

var rel.inf

lat\_destination lat\_destination 66.7410131

destination\_language destination\_language 32.7991807

sum\_secs\_elapsed sum\_secs\_elapsed 0.4598062

signup\_method signup\_method 0.0000000

signup\_flow signup\_flow 0.0000000

language language 0.0000000

affiliate\_channel affiliate\_channel 0.0000000

affiliate\_provider affiliate\_provider 0.0000000

first\_affiliate\_tracked first\_affiliate\_tracked 0.0000000

signup\_app signup\_app 0.0000000

first\_device\_type first\_device\_type 0.0000000

first\_browser first\_browser 0.0000000

counts counts 0.0000000

lng\_destination lng\_destination 0.0000000

distance\_km distance\_km 0.0000000

destination\_km2 destination\_km2 0.0000000

language\_levenshtein\_distance language\_levenshtein\_distance 0.0000000

CARET – GBM

ctrl = trainControl(method = "repeatedcv", number = 5)

gbmGrid <- expand.grid(interaction.depth = c(2,4,6,8), n.trees =(2:8)\*100,

shrinkage = c(0.01, 0.005, 0.001),

n.minobsinnode = c(100, 500, 2000))

m\_gbm = train(x=t, y=t$bookings, method="gbm", distribution = "multinomial", tuneGrid = gbmGrid,train.fraction = .8, verbose=TRUE, trControl=ctrl)

REDUCED GBM MODEL (JUST USING BOOST.T PARAMETERS)

gbm(formula = t.bookings ~ ., distribution = "multinomial", data = reduced[trainred,

], n.trees = 300, interaction.depth = 2, shrinkage = 0.001,

cv.folds = 5)

A gradient boosted model with multinomial loss function.

300 iterations were performed.

The best cross-validation iteration was 300.

There were 3 predictors of which 3 had non-zero influence.

var rel.inf

t.lat\_destination t.lat\_destination 63.8278527

t.destination\_language t.destination\_language 35.7077721

t.sum\_secs\_elapsed t.sum\_secs\_elapsed 0.4643752

predredtrain = predict(boost.tred, newdata = reduced[trainred,], n.trees = 300)

ff <- apply(predredtrain, 1, which.max)

cc <- as.factor(as.vector(reduced[trainred,]$t.bookings))

table(ff, cc)

Training accuracy = 0.830944

28868/170760

predredtest = predict(boost.tred, newdata = tr.test, n.trees = 300)

gg <- apply(predredtest, 1, which.max)

dd <- as.factor(as.vector(tr.test$t.bookings))

table(gg, dd)

Testing accuracy = 0.8287461

7311/42691

lat\_destination lat\_destination 66.7410131

destination\_language destination\_language 32.7991807

sum\_secs\_elapsed sum\_secs\_elapsed 0.4598062

signup\_method signup\_method 0.0000000

signup\_flow signup\_flow 0.0000000

language language 0.0000000

affiliate\_channel affiliate\_channel 0.0000000

affiliate\_provider affiliate\_provider 0.0000000

first\_affiliate\_tracked first\_affiliate\_tracked 0.0000000

signup\_app signup\_app 0.0000000

first\_device\_type first\_device\_type 0.0000000

first\_browser first\_browser 0.0000000

counts counts 0.0000000

lng\_destination lng\_destination 0.0000000

distance\_km distance\_km 0.0000000

destination\_km2 destination\_km2 0.0000000

language\_levenshtein\_distance language\_levenshtein\_distance 0.0000000

Pretty confident that we can drop all of these variables other than the first three:

lat\_destination lat\_destination 66.7410131

destination\_language destination\_language 32.7991807

sum\_secs\_elapsed sum\_secs\_elapsed 0.4598062

signup\_method signup\_method 0.0000000

signup\_flow signup\_flow 0.0000000

language language 0.0000000

affiliate\_channel affiliate\_channel 0.0000000

affiliate\_provider affiliate\_provider 0.0000000

first\_affiliate\_tracked first\_affiliate\_tracked 0.0000000

signup\_app signup\_app 0.0000000

first\_device\_type first\_device\_type 0.0000000

first\_browser first\_browser 0.0000000

counts counts 0.0000000

lng\_destination lng\_destination 0.0000000

distance\_km distance\_km 0.0000000

destination\_km2 destination\_km2 0.0000000

language\_levenshtein\_distance language\_levenshtein\_distance 0.0000000