task3_galaxy_nzv

Code

Hide

```
samsung <- read.csv("galaxy_smallmatrix_labeled_9d.csv")
```

Near Zero Variance Variables

nzv

Hide

```
#nearZeroVar() with saveMetrics = TRUE returns an object containing a table including: f
requency ratio, percentage unique, zero variance and near zero variance

nzvMetrics <- nearZeroVar(samsung, saveMetrics = TRUE)
nzvMetrics
```

	freqRatio <dbl>	percentUnique <dbl>	zeroVar <lgl>	nzv <lgl>
iphone	5.039313	0.20912400	FALSE	FALSE
samsunggalaxy	14.090164	0.05421733	FALSE	FALSE
sonyxperia	44.111888	0.03872667	FALSE	TRUE
nokialumina	495.500000	0.02323600	FALSE	TRUE
htcphone	11.427740	0.06970800	FALSE	FALSE
ios	27.662132	0.04647200	FALSE	TRUE
googleandroid	61.248780	0.04647200	FALSE	TRUE
iphonecampos	10.526217	0.23236000	FALSE	FALSE
samsungcampos	93.176471	0.08519867	FALSE	TRUE
sonycampos	347.081081	0.05421733	FALSE	TRUE

1-10 of 59 rows Previous **1** 2 3 4 5 6 Next

Hide

```
# returns column 2, iphonecamunc, same as nvzMetrics
# nearZeroVar() with saveMetrics = FALSE returns an vector
nzv <- nearZeroVar(samsung, saveMetrics = FALSE)
nzv
```

```
 [1]  3  4  6  7  9 10 11 12 13 14 15 16 17 19 20 21 22 24 25 26 27 29 30 31 32 34 35 36
37 39 40
[32] 41 42 44 45 46 47 49 50 51 52 53 54 55 56 57 58
```

Remove Near Zero Variance Variables

Hide

```
# create a new data set and remove near zero variance features
samsungNZV <- samsung[,-nzv]
str(samsungNZV)
```

```
'data.frame':   12911 obs. of  12 variables:
 $ iphone         : int  1 1 1 0 1 2 1 1 4 1 ...
 $ samsunggalaxy  : int  0 0 1 0 0 0 0 0 0 0 ...
 $ htcphone       : int  0 0 0 1 0 0 0 0 0 0 ...
 $ iphonecampos   : int  0 0 1 0 0 1 0 0 0 0 ...
 $ iphonecamunc   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ iphonedispos   : int  0 1 0 0 0 0 2 0 0 0 ...
 $ iphonedisneg   : int  0 1 0 0 0 0 0 0 0 0 ...
 $ iphonedisunc   : int  0 1 0 0 0 0 0 0 0 0 ...
 $ iphoneperpos   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ iphoneperneg   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ iphoneperunc   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ galaxysentiment: int  5 3 3 0 1 0 3 5 5 5 ...
```

Train Model

Train and Test Set

Hide

```
# convert variable types, categorical
samsungNZV$galaxysentiment <- as.factor(samsungNZV$galaxysentiment)
```

Train and Test Set:

Hide

```
# Create Train and Test Set for samsungNZV
# create 75% sample of row indices
in_training <-createDataPartition(samsungNZV$galaxysentiment, p = .7, list = FALSE)
# create 75% sample of data and save it to trainData
trainData_samsungNZV <- samsungNZV[in_training, ]
 # create 25% sample of data and save it to test_data
testData_samsungNZV <- samsungNZV[-in_training, ]
# verify split percentages
nrow(trainData_samsungNZV) / nrow(samsungNZV)
```

```
[1] 0.7001781
```

Models

Hide

```
#c5
c5_samsungNZV <- train(galaxysentiment ~., data = trainData_samsungNZV, method = "C5.0",
                trControl = fitControl)
```

Hide

```
# randomforest
rf_samsungNZV <- train(galaxysentiment ~., data = trainData_samsungNZV, method = "rf",
                trControl = fitControl)
```

Hide

```
# svm   (kernlab)
#svm_samsungNZV <- train(galaxysentiment ~., data = trainData_samsungNZV, method = "svmL
inear",
#                 trControl = fitControl)
```

Hide

```
# kknn
kknn_samsungNZV <- train(galaxysentiment ~., data = trainData_samsungNZV, method = "kkn
n",
                trControl = fitControl)
```

Hide

```
# gbm
#gbm_samsungNZV <- train(galaxysentiment ~., data = trainData_samsungNZV, method = "gb
m",
#                 trControl = fitControl)
```

Model Summaries

Compare Accuracy on Prediction Results:

Hide

```
#c5
prediction_c5_samsungNZV <- predict(c5_samsungNZV, testData_samsungNZV)
postResample(prediction_c5_samsungNZV, testData_samsungNZV$galaxysentiment)
```

```
 Accuracy     Kappa
0.7561354 0.5061385
```

Hide

```
#randomforest
prediction_rf_samsungNZV  <- predict(rf_samsungNZV, testData_samsungNZV)
postResample(prediction_rf_samsungNZV, testData_samsungNZV$galaxysentiment)
```

```
Accuracy    Kappa
0.757427 0.504549
```

Hide

```
#svm
#prediction_svm_samsungNZV  <- predict(svm_samsungNZV, testData_samsungNZV)
#postResample(prediction_svm_samsungNZV, testData_samsungNZV$galaxysentiment)
# kknn
prediction_kknn_samsungNZV  <- predict(kknn_samsungNZV, testData_samsungNZV)
postResample(prediction_kknn_samsungNZV, testData_samsungNZV$galaxysentiment)
```

```
 Accuracy     Kappa
0.7473521 0.4905414
```

Hide

```
modelData_samsungNZV <- resamples(list(C50 = c5_samsungNZV, randomForest = rf_samsungNZ
V,
                                       #svMLinear = svm_samsungNZV,
                                       kknn = kknn_samsungNZV))
```

Hide

```
summary(modelData_samsungNZV)
```

```
Call:
summary.resamples(object = modelData_samsungNZV)

Models: C50, randomForest, kknn
Number of resamples: 10

Accuracy
                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
C50          0.7411504 0.7435130 0.7477830 0.7507724 0.7555310 0.7721239    0
randomForest 0.7436464 0.7488238 0.7533186 0.7538713 0.7571982 0.7701657    0
kknn         0.7215470 0.7389503 0.7421139 0.7390501 0.7430786 0.7447514    0

Kappa
                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
C50          0.4686177 0.4748606 0.4816509 0.4921860 0.5024918 0.5429144    0
randomForest 0.4727955 0.4813335 0.4963733 0.4964004 0.5043570 0.5425005    0
kknn         0.4261212 0.4707905 0.4809050 0.4738417 0.4844310 0.4892646    0
```