Ex12

Chathrua Gunasekara

November 29, 2014

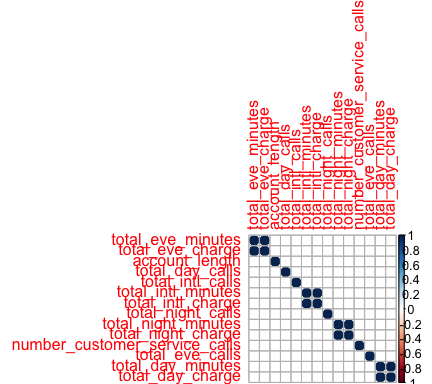
Q3. 1. Data Exploration

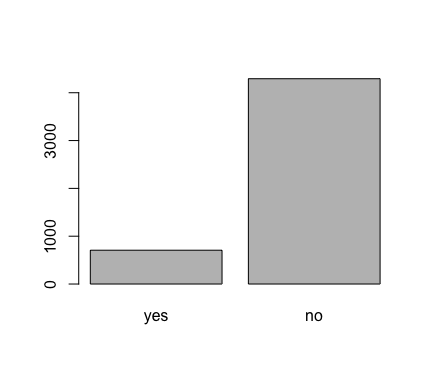
## Loading required package: lattice  
## Loading required package: ggplot2

Remove Near Zero Variance predictors

## [1] 6

Removed predictor V6 because the near zero variance.

 Removed the catagorical predictors and high correlated numerical predictors. Then the data set was pre processed with "centering" and "scaling"



The data set was splited to 75% training set and rest to testing set with stratified sampling beucase the class imbalance in the data set.

2.Fitting models and evaluating the model using ROC. Kappa statistic is used because th class imbalance.

1. Logistic Regression classification model

## Generalized Linear Model   
##   
## 3751 samples  
## 10 predictor  
## 2 classes: 'yes', 'no'   
##   
  
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)   
##   
## Summary of sample sizes: 2814, 2814, 2814, 2814, 2814, 2814, ...   
##   
## Resampling results  
##   
## ROC Sens Spec ROC SD Sens SD Spec SD   
## 0.7496665 0.07333333 0.9872795 0.01989304 0.01412562 0.004186292  
##   
##

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction yes no  
## yes 21 11  
## no 155 1062  
##   
## Accuracy : 0.8671   
## 95% CI : (0.847, 0.8854)  
## No Information Rate : 0.8591   
## P-Value [Acc > NIR] : 0.221   
##   
## Kappa : 0.1658   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.11932   
## Specificity : 0.98975   
## Pos Pred Value : 0.65625   
## Neg Pred Value : 0.87264   
## Prevalence : 0.14091   
## Detection Rate : 0.01681   
## Detection Prevalence : 0.02562   
## Balanced Accuracy : 0.55453   
##   
## 'Positive' Class : yes   
##

## Area under the curve: 0.7493

ii.Linear Discriminate Analysis

## Linear Discriminant Analysis   
##   
## 3751 samples  
## 10 predictor  
## 2 classes: 'yes', 'no'   
##   
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)   
##   
## Summary of sample sizes: 2814, 2814, 2814, 2814, 2814, 2814, ...   
##   
## Resampling results  
##   
## ROC Sens Spec ROC SD Sens SD Spec SD   
## 0.7580982 0.07060606 0.9839503 0.01841665 0.01839105 0.005519686  
##   
##

## Area under the curve: 0.7579

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction yes no  
## yes 17 14  
## no 159 1059  
##   
## Accuracy : 0.8615   
## 95% CI : (0.8411, 0.8802)  
## No Information Rate : 0.8591   
## P-Value [Acc > NIR] : 0.4231   
##   
## Kappa : 0.1274   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.09659   
## Specificity : 0.98695   
## Pos Pred Value : 0.54839   
## Neg Pred Value : 0.86946   
## Prevalence : 0.14091   
## Detection Rate : 0.01361   
## Detection Prevalence : 0.02482   
## Balanced Accuracy : 0.54177   
##   
## 'Positive' Class : yes   
##

1. Partial Least Squrare Discriminate Analysis

## Partial Least Squares   
##   
## 3751 samples  
## 10 predictor  
## 2 classes: 'yes', 'no'   
##   
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)   
##   
## Summary of sample sizes: 2814, 2814, 2814, 2814, 2814, 2814, ...   
##   
## Resampling results across tuning parameters:  
##   
## ncomp ROC Sens Spec ROC SD Sens SD   
## 1 0.7550104 0.01818182 0.9993540 0.01744653 0.009532619  
## 2 0.7552355 0.01818182 0.9993043 0.01769958 0.009532619  
## Spec SD   
## 0.00102186  
## 0.00101934  
##   
## ROC was used to select the optimal model using the largest value.  
## The final value used for the model was ncomp = 2.

## Area under the curve: 0.755

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction yes no  
## yes 3 1  
## no 173 1072  
##   
## Accuracy : 0.8607   
## 95% CI : (0.8402, 0.8794)  
## No Information Rate : 0.8591   
## P-Value [Acc > NIR] : 0.4553   
##   
## Kappa : 0.0272   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.017045   
## Specificity : 0.999068   
## Pos Pred Value : 0.750000   
## Neg Pred Value : 0.861044   
## Prevalence : 0.140913   
## Detection Rate : 0.002402   
## Detection Prevalence : 0.003203   
## Balanced Accuracy : 0.508057   
##   
## 'Positive' Class : yes   
##

## pls variable importance  
##   
## Overall  
## total\_day\_minutes 0.0747890  
## number\_customer\_service\_calls 0.0696121  
## total\_eve\_minutes 0.0335319  
## total\_intl\_minutes 0.0281295  
## total\_intl\_calls 0.0155282  
## total\_night\_charge 0.0135423  
## account\_length 0.0103439  
## total\_day\_calls 0.0091050  
## total\_night\_calls 0.0062863  
## total\_eve\_calls 0.0004964

1. Penalized Models

## 3751 samples  
## 10 predictor  
## 2 classes: 'yes', 'no'   
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)   
##   
## Summary of sample sizes: 2814, 2814, 2814, 2814, 2814, 2814, ...   
##

## Resampling results across tuning parameters:  
##   
## alpha lambda ROC Sens Spec ROC SD Sens SD   
## 0.0 0.100 0.7610631 0.003030303 1 0.02022662 0.004373866  
## 0.0 0.125 0.7610567 0.002424242 1 0.02019687 0.003606782  
## 0.0 0.150 0.7609814 0.002121212 1 0.02017541 0.003471648  
## 0.0 0.175 0.7609565 0.002121212 1 0.02018374 0.003471648  
## 0.0 0.200 0.7608858 0.001515152 1 0.02017503 0.003092790  
## 0.1 0.100 0.7575189 0.002424242 1 0.01834891 0.003606782  
## 0.1 0.125 0.7554726 0.001212121 1 0.01828654 0.002834589  
## 0.1 0.150 0.7533070 0.000000000 1 0.01859388 0.000000000  
## 0.1 0.175 0.7510710 0.000000000 1 0.01902832 0.000000000  
## 0.1 0.200 0.7485304 0.000000000 1 0.01946229 0.000000000  
  
##   
## ROC was used to select the optimal model using the largest value.  
## The final values used for the model were alpha = 0 and lambda = 0.1.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction yes no  
## yes 1 0  
## no 175 1073  
##   
## Accuracy : 0.8599   
## 95% CI : (0.8394, 0.8787)  
## No Information Rate : 0.8591   
## P-Value [Acc > NIR] : 0.4877   
##   
## Kappa : 0.0097   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0056818   
## Specificity : 1.0000000   
## Pos Pred Value : 1.0000000   
## Neg Pred Value : 0.8597756   
## Prevalence : 0.1409127   
## Detection Rate : 0.0008006   
## Detection Prevalence : 0.0008006   
## Balanced Accuracy : 0.5028409   
##   
## 'Positive' Class : yes   
##

1. Nearest Shrunken centroids Model

## 3751 samples  
## 10 predictor  
## 2 classes: 'yes', 'no'   
##   
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 0.75%)   
## Summary of sample sizes: 2814, 2814, 2814, 2814, 2814, 2814, ...   
  
## Resampling results across tuning parameters:   
## threshold ROC Sens Spec ROC SD Sens SD Spec SD  
## 0 0.7611647 0 1 0.01756043 0 0   
## 1 0.7561088 0 1 0.01751932 0 0   
## 2 0.7428150 0 1 0.01840524 0 0   
## 3 0.7334495 0 1 0.02019059 0 0   
## 4 0.7252872 0 1 0.02444128 0 0   
## 5 0.6727549 0 1 0.05473836 0 0   
## 6 0.5071470 0 1 0.02486087 0 0   
## 7 0.5000000 0 1 0.00000000 0 0   
## 8 0.5000000 0 1 0.00000000 0 0   
## 9 0.5000000 0 1 0.00000000 0 0   
## 10 0.5000000 0 1 0.00000000 0 0   
## 11 0.5000000 0 1 0.00000000 0 0   
## 12 0.5000000 0 1 0.00000000 0 0   
## 13 0.5000000 0 1 0.00000000 0 0   
## 14 0.5000000 0 1 0.00000000 0 0   
## 15 0.5000000 0 1 0.00000000 0 0   
## 16 0.5000000 0 1 0.00000000 0 0   
## 17 0.5000000 0 1 0.00000000 0 0   
## 18 0.5000000 0 1 0.00000000 0 0   
## 19 0.5000000 0 1 0.00000000 0 0   
## 20 0.5000000 0 1 0.00000000 0 0   
## 21 0.5000000 0 1 0.00000000 0 0   
## 22 0.5000000 0 1 0.00000000 0 0   
## 23 0.5000000 0 1 0.00000000 0 0   
## 24 0.5000000 0 1 0.00000000 0 0   
## 25 0.5000000 0 1 0.00000000 0 0   
##   
## ROC was used to select the optimal model using the largest value.  
## The final value used for the model was threshold = 0.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction yes no  
## yes 0 0  
## no 176 1073  
##   
## Accuracy : 0.8591   
## 95% CI : (0.8385, 0.8779)  
## No Information Rate : 0.8591   
## P-Value [Acc > NIR] : 0.5201   
##   
## Kappa : 0   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0000   
## Specificity : 1.0000   
## Pos Pred Value : NaN   
## Neg Pred Value : 0.8591   
## Prevalence : 0.1409   
## Detection Rate : 0.0000   
## Detection Prevalence : 0.0000   
## Balanced Accuracy : 0.5000

## pam variable importance  
##   
## Importance  
## total\_day\_minutes 0.262637  
## number\_customer\_service\_calls 0.244507  
## total\_eve\_minutes 0.116336  
## total\_intl\_minutes 0.097557  
## total\_intl\_calls -0.053727  
## total\_night\_charge 0.046986  
## account\_length 0.035843  
## total\_day\_calls 0.031577  
## total\_night\_calls -0.021738  
## total\_eve\_calls 0.001714

|  |  |  |
| --- | --- | --- |
| Model | ROC | Kappa for testing |
| Logistic Regression | 0.7496665 | 0.16 |
| Linear Discriminate Analysis | 0.7496665 | 0.12 |
| Partial Least Squared Discriminate Analysis | 0.7496665 | 0.02 |
| Penalized Model | 0.7610631 | 0.01 |
| Nearest Shrunken centroids | 0.7611647 | 0 |

Because the class imbalance in this data set kappa statistic is more important that the small difference in ROC between the models. So. Based on the table Logistic Regression is the best model to do the classification.