Predictive modelling

George, John & William

11/17/2021

Data Preparation and cleaning

Getting the missing values percentages

```
##   Number_Missing Missing_Rate        Variable
## 1           7087     56.53318      enrollment
## 2           7087     56.53318      employment
## 3           9448     75.36694 employment_type
## 4           9448     75.36694 weekly_work_hrs
## 5           7361     58.71889       ethnicity
## 6           7383     58.89438          gender
```

```
## [1] 5095   18
```

missing value treatment

Method I

Treating imbalance classification

```
library(ROSE)
 data_pad_balance<-ovun.sample(permanent_address ~ ., data = data_pad, method = "both", p=0.5,N=NROW(da
 dim(data_pad_balance)
```

```
## [1] 5095   18
```

Converting predictors to category

Partitioning data set

```
## [1] 3415   18
```

```
## [1] 1680   18
```

The training data has 76 observations with 1887 now (old =1057 when compared) variables. The testing data has 32 observation with 1887 now (old= 1057 when compared) variables.

Model fitting

```r
#------ Model building -----------
# Create a wrapper function to abstract away the common aspects of model fitting
formula<- permanent_address~.
fit.model <- function(method, tunegrid="", data=NULL, formula=NULL) {

  data <- training
  if(is.null(formula)) formula<- permanent_address~.

  # Train the model
   train(
           formula,
           data = data,
           method = method,
           trControl = trainControl(method = "cv", 5),
           preProcess = c("center","scale"),
           tuneGrid = tunegrid)

}
```

```r
# Logistic Regression
log <-train(formula,
                 data=training,
                 method="glm",
                 family = binomial(link = "logit"),
                trControl = trainControl(method = "cv", 5),
                preProcess = c("center", "scale"))
```

```r
# LDA
lda <- train(formula,
                 data=training,
                 method="lda",
                trControl = trainControl(method = "cv", 5),
                preProcess = c("center", "scale"))
```

```r
#-------------- Elastic Net Models -------------------
# fit a LASSO model
lasso <- fit.model("glmnet", expand.grid(.alpha=1, .lambda=seq(0,0.1,0.01)))

# Fit a Ridge regression model
ridge <- fit.model("glmnet", expand.grid(.alpha=0, .lambda=seq(0,0.1,0.01)))
```

```r
# Bagging
# bag <- fit.model("rf", data.frame(mtry=11))
bag <- train(formula,
                 data=training,
                 method="rf",
                trControl = trainControl(method = "cv", 5),
                preProcess = c("center", "scale"),
                tuneGrid = data.frame(mtry=11),
                ntree = 1000)
```

```r
# Random Forest
# rf <- fit.model("rf", data.frame(mtry=1:10))
# rf <- train(formula,
#                   data=training,
#                   method="rf",
#                  trControl = trainControl(method = "cv", 5),
#                  preProcess = c("center", "scale"),
#                  tuneGrid = data.frame(mtry=1:10),
#                  ntree = 1000)
```

```r
#--------------

# Support Vector Machine with linear kernel
set.seed(125)
trctrl <- trainControl(method = "cv", number=5)
svc <- train(formula, data = training, method = "svmLinear",
                    trControl=trctrl, prob.model=T,
                    tuneLength = 10)
```

```r
# Support Vector Machine with radial kernel
set.seed(125)
trctrl <- trainControl(method = "cv", number=5)
svmR <- train(formula , data = training, method = "svmRadial",
                    trControl=trctrl, prob.model=T,
                    tuneLength = 10)
```

Making predictions

Metrics

```r
library(mlr3measures)
```

```
## In order to avoid name clashes, do not attach 'mlr3measures'. Instead, only load the namespace with

##
## Attaching package: 'mlr3measures'

## The following object is masked from 'package:pROC':
##
##     auc

## The following objects are masked from 'package:caret':
##
##     precision, recall, sensitivity, specificity

## The following object is masked from 'package:MASS':
##
##     fbeta
```

```r
# Create a custom confusion matrix with performance metrics
metrics <- function(model_object, response="", test_data=NULL) {
  # response = "permanent_address"
  # model_object <- log
  #
  if(is.null(test_data)) test_data <- testing

  # make predictions
  prediction <- predict(model_object, test_data)
  prediction_A <- predict(model_object, test_data, type="prob")

  target <- test_data[, response]

  cmat <- confusionMatrix(prediction, target, mode = "prec_recall")
  ROC <- roc(target, predictor = prediction_A[,2])
  AUC_m<-round(ROC$auc, digits=4)

  # Plotting the ROC_auc curves
  plot <- ggroc(ROC, colour = 'blue', size = 2) +
  ggtitle(paste0('(AUC = ', AUC_m, ')')) +
    theme(plot.title = element_text(hjust = 1))+
  theme_minimal()

  misscal<- round(mean(prediction != target),digits = 2)

 # Returned outputs
 return(list(
   accuracy = (1-misscal),
   mcr = misscal,
   sens = round(cmat$byClass[1],2),
   spec = round(cmat$byClass[2],2),
   fbeta = round(cmat$byClass[7],2),
   auc = AUC_m,
   plot = plot
 ))
}
```

```r
metric_log <- metrics(lasso, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2

## Setting direction: controls < cases
```

```r
#------- Compute performance metrics for the full models ---------------
log.metric <- metrics(log, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```r
lda.metric <- metrics(lda, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```r
# knn.metric <- metrics(knn)
lasso.metric <- metrics(lasso, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```r
ridge.metric <- metrics(ridge, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```r
bag.metric <- metrics(bag, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```r
# rf.metric <- metrics(rf)
svc.metric <- metrics(svc, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```r
# svmP.metric <- metrics(svmP)
svmR.metric <- metrics(svmR, response = "permanent_address")
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

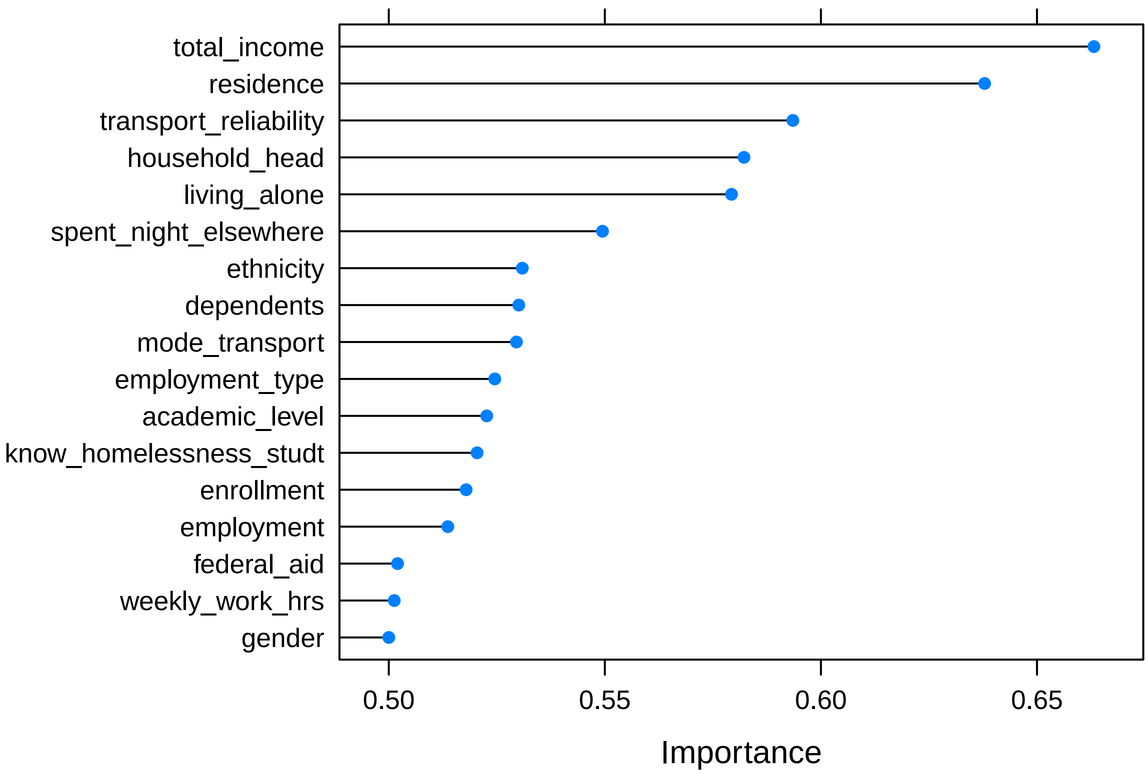
```r
mod.sum <- data.frame(rbind(
                          c("Logistic", log.metric$mcr, log.metric$accuracy, log.metric$sens, log.metri
                          c("LDA",  lda.metric$mcr, lda.metric$accuracy, lda.metric$sens, lda.metric$spe
                          c("LASSO", lasso.metric$mcr, lasso.metric$accuracy, lasso.metric$sens, lasso.
                          c("Ridge", ridge.metric$mcr, ridge.metric$accuracy, ridge.metric$sens, ridge.
                          c("Bagging", bag.metric$mcr, bag.metric$accuracy, bag.metric$sens, bag.metric
                          c("SVC",  svc.metric$mcr, svc.metric$accuracy, svc.metric$sens, svc.metric$spe
                          c("SVM (Radial Kernel)", svmR.metric$mcr, svmR.metric$accuracy, svmR.metric$s

names(mod.sum) <- c("Model", "Misclassification Rate", "Accuracy", "Sensitivity", "Specificity", "fbeta

kable(mod.sum, align = "lcccccc", caption = "Table : Evaluation metrics for Housing Insecurity with Per
  kable_paper("hover", full_width = F)%>%
       kable_styling(font_size = 12)
```

Base on our table of results SVM with radial basis function is the best

```r
Var <- varImp(svmR, scale = FALSE)
plot(Var)
```

Table 1: Table : Evaluation metrics for Housing Insecurity with Permanent Address as a response

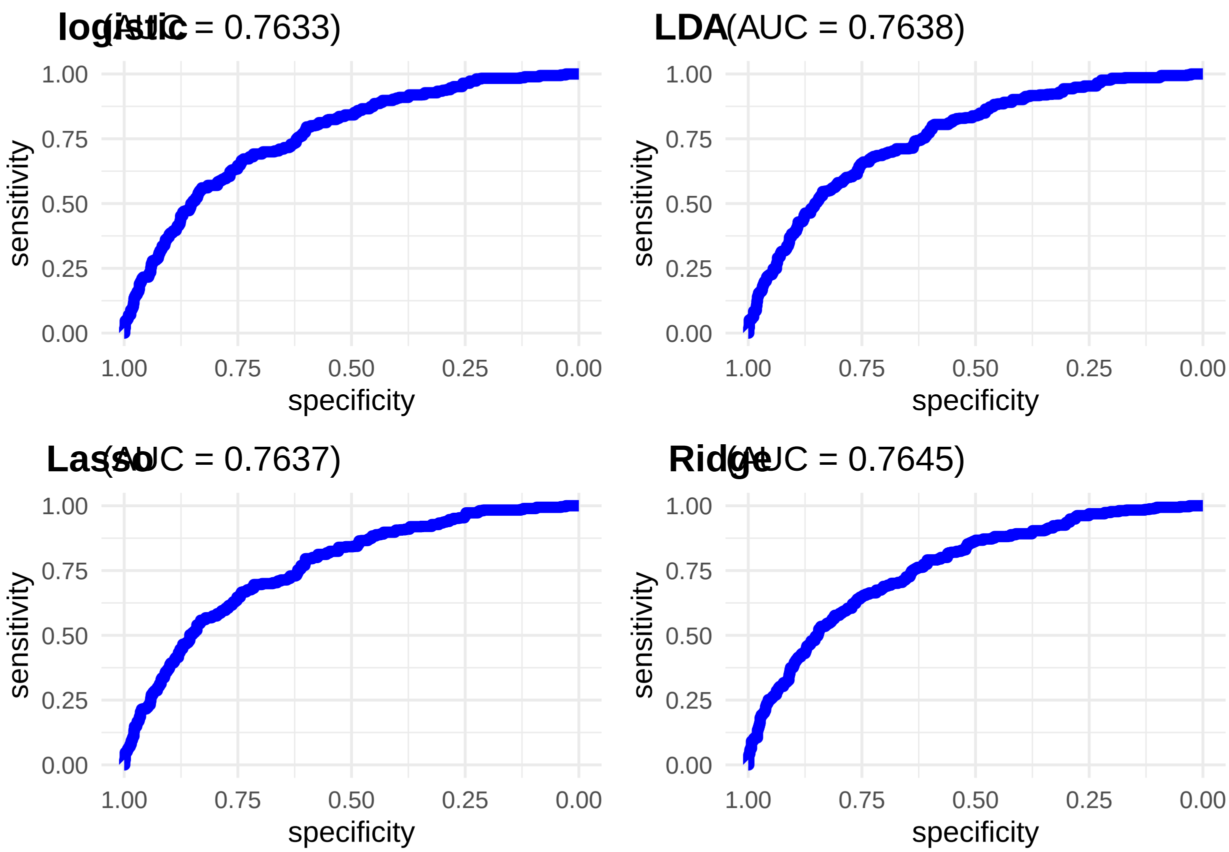
Model	Misclassification Rate	Accuracy	Sensitivity	Specificity	fbeta	AUC
Logistic	0.3	0.7	0.74	0.66	0.71	0.7633
LDA	0.3	0.7	0.75	0.66	0.72	0.7638
LASSO	0.3	0.7	0.74	0.65	0.71	0.7637
Ridge	0.3	0.7	0.75	0.65	0.71	0.7645
Bagging	0.03	0.97	0.95	0.99	0.97	0.9987
SVC	0.33	0.67	0.81	0.54	0.71	0.7532
SVM (Radial Kernel)	0.04	0.96	0.93	1	0.96	0.9668



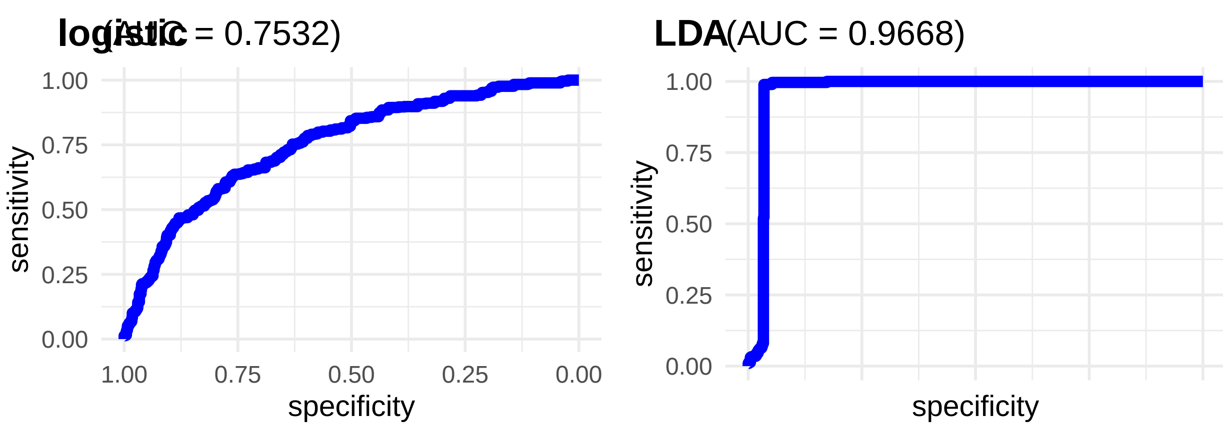
Graphing

```
library(ggpubr)
ggarrange(log.metric$plot, lda.metric$plot, lasso.metric$plot, ridge.metric$plot, svc.metric$plot, svmR
          labels = c("logistic", "LDA", "Lasso", "Ridge","SVM Linear","SVM Radial"),
          ncol = 2, nrow = 2)
```

```
## $`1`
```




```
##
## $‘2‘
```




```
##
## attr(,"class")
## [1] "list"      "ggarrange"
```