```
/* loading dataset */
proc import out=bank
    datafile="/home/u60815483/ST662/bank-additional-full.csv"
    dbms=csv replace;
    getnames=yes;
run;

/* loading the test dataset */
proc import out=bank2
    datafile="/home/u60815483/ST662/bank-additional.csv"
    dbms=csv replace;
    getnames=yes;
run;

proc print data = bank2;
run;

proc contents data = bank;
run;

/*check the number of rows in the dataset */
proc sql;
 select count(*) as N from bank;
quit;

/* data exploration */
proc sgplot data = bank;
vbar y;
run;

proc sgplot data = bank;
vbar marital / group = y groupdisplay=cluster;
run;

/* Model 1*/
/* original model */

/* Split the dataset into training and test data */
proc surveyselect data=bank rate=0.8
```

```
out= bank_select outall seed=123
method=srs;
run;

data bank_train bank_test;
set bank_select;
if selected =1 then output bank_train;
else output bank_test;
run;

/* fit logistic regression model */
proc logistic data = bank_train outmodel= bank_train_logistic descending;
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;

/* confusion matrix for the splitted training and test dataset */
proc logistic inmodel=bank_train_logistic;
score data=bank_test(drop= selected) out=test_y outroc=test_roc;
run;


PROC FREQ DATA=test_y;
TABLE f_y*i_y/ nopercent norow nocol;
RUN;

/* ROC curve*/
proc logistic data = bank_test descending plots(only)=roc plots(maxpoints=none);
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;

/* confusion matrix for the test dataset */
proc logistic inmodel=bank_train_logistic;
score data=bank2 out=test_y2 outroc=test_roc2;
run;

PROC FREQ DATA=test_y2;
TABLE f_y*i_y/ nopercent norow nocol;
RUN;
```

```
/* ROC curve for test data*/
proc logistic data = bank2 descending plots(only)=roc;
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;

/* ############################################################################### */
/* Model 2 (Imputation and Scaling) */

/* check number of unknowns in all the columns */
proc freq data = bank order=freq;
tables job marital education default housing loan pdays;
run;

/* Imputation */
Data bank;
SET bank;
If Job = "unknown" Then Job = "admin."; else Job = Job;
run;

Data bank;
SET bank;
If marital = "unknown" Then marital = "married"; else marital = marital;
run;

Data bank;
SET bank;
If education = "unknown" Then education = "university.degree"; else education = education;
run;

Data bank;
SET bank;
If housing = "unk" Then housing = "yes"; else housing = housing;
run;

Data bank;
SET bank;
If loan = "unk" Then loan = "no"; else loan = loan;
run;
```

```
Data bank;
SET bank;
If pdays = 999 Then pdays = -1; else pdays = pdays;
run;
```

```
proc sql;
delete from bank
where default = "yes";
run;
```

```
proc sql;
delete from bank
where education = "illiterate";
run;
```

```
/* Scaling */
proc stdize data=bank out=bank_scaled method=std;
   var age duration campaign pdays previous 'emp.var.rate'n 'cons.price.idx'n 'cons.conf.idx'n euribor3m 'nr.employed'n;
run;
```

```
/* splitting the dataset into train and test */
proc surveyselect data=bank_scaled rat=0.8
out= bank_select2 outall seed = 123
method=srs;
run;
```

```
data bank2_train bank2_test;
set bank_select2;
if selected =1 then output bank2_train;
else output bank2_test;
run;
```

```
proc logistic data = bank2_train outmodel = bank2_train_logistic descending;
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;
```

```
/* for confusion matrix */
proc logistic inmodel=bank2_train_logistic;
score data=bank2_test(drop= selected) out=test2_y outroc=test2_roc;
run;

PROC FREQ DATA=test2_y;
TABLE f_y*i_y/ nopercent norow nocol;
RUN;

/* for roc curve */
proc logistic data = bank2_test descending plots(only)=roc plots(maxpoints=none);
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;

/* for test data */

proc logistic inmodel=bank2_train_logistic;
score data=bank2 out=test3_y outroc=test3_roc;
run;

PROC FREQ DATA=test3_y;
TABLE f_y*i_y/ nopercent norow nocol;
RUN;

/* for roc curve */
proc logistic data = bank2 descending plots(only)=roc plots(maxpoints=none);
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;

/* ############################################################################### */
/*Model 3 (Imputation, Scaling and Undersampling) */

/* Split the dataset into training and test data */
proc surveyselect data=bank_scaled rat=0.8
out= bank4_select outall seed = 123
method=srs;
run;
```

```
data bank4_train bank4_test;
set bank4_select;
if selected =1 then output bank4_train;
else output bank4_test;
run;

DATA bank4_yes;
set bank4_train;
if (y = "ye") then output;
run;

DATA bank4_no;
set bank4_train;
if (y = "no") then output;
run;

/* Sample 3705 obs from no dataset */
proc surveyselect data = bank4_no
out = bank4_no_sample seed = 123
method = srs
sampsize= 3705;
run;

data bank4_sampled;
set bank4_yes bank4_no_sample;
run;


/* after undersampling fit on the logistic regression */
proc logistic data = bank4_sampled outmodel= bank4_train_logistic descending;
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;

proc logistic inmodel=bank4_train_logistic;
score data=bank4_test(drop= selected) out=test4_y outroc=test4_roc;
run;

PROC FREQ DATA=test4_y;
```

```
TABLE f_y*i_y/ nopercent norow nocol;
RUN;

proc logistic data = bank4_test descending plots(only)=roc plots(maxpoints=none);
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;

/* for test dataset */

proc logistic inmodel=bank4_train_logistic;
score data=bank2 out=test4_y outroc=test4_roc;
run;

PROC FREQ DATA=test4_y;
TABLE f_y*i_y/ nopercent norow nocol;
RUN;

proc logistic data = bank2 descending plots(only)=roc plots(maxpoints=none);
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;


/* ############################################################################### */
/*Model 4 (Imputation, Scaling and Oversampling) */

/* splitting the dataset into train and test */
proc surveyselect data=bank_scaled rat=0.8
out= bank3_select outall seed = 123
method=srs;
run;

data bank3_train bank3_test;
set bank3_select;
if selected =1 then output bank3_train;
else output bank3_test;
run;
```

```
/* over sampling */

PROC SQL;
   SELECT y,COUNT(*) FROM bank3_train GROUP BY y;
QUIT;


DATA bank3_yes_train;
set bank3_train;
if (y = "ye") then output;
run;

DATA bank3_no_train;
set bank3_train;
if (y = "no") then output;
run;

proc surveyselect data=bank3_yes_train  method = urs sampsize = 29229
   rep=1 seed=123 out=bank3_yes_samples_train outhits;
run;


data bank3_sampled_final_train;
set bank3_yes_samples_train bank3_no_train;
run;

PROC SQL;
   SELECT y,COUNT(*) FROM bank3_sampled_final_train GROUP BY y;
QUIT;

/* Fit logistic regression model after imputation and scaling */
proc logistic data = bank3_sampled_final_train outmodel = bank3_train_logistic descending;
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;


/* for confusion matrix */
proc logistic inmodel=bank3_train_logistic;
score data=bank3_test(drop= selected) out=test7_y outroc=test7_roc;
```

```
run;

PROC FREQ DATA=test7_y;
TABLE f_y*i_y/ nopercent norow nocol;
RUN;

/* for roc curve */
proc logistic data = bank3_test descending plots(only)=roc plots(maxpoints=none);
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;



/* for test dataset */

proc logistic inmodel=bank3_train_logistic;
score data=bank2 out=test8_y outroc=test8_roc;
run;

PROC FREQ DATA=test8_y;
TABLE f_y*i_y/ nopercent norow nocol;
RUN;

proc logistic data = bank2 descending plots(only)=roc plots(maxpoints=none);
class y job marital education default housing loan contact month day_of_week poutcome / param = ref;
model y = age job marital education default housing loan contact month day_of_week duration campaign pdays previous poutcome;
run;
```