Logistic Regression (Module -9)

1.) I have a dataset containing family information of married couples, which have around 10 variables & 600+ observations.

Independent variables are ~ gender, age, years married, children, religion etc.

I have one response variable which is number of extra marital affairs.

Now, I want to know what all factor influence the chances of extra marital affair.

Since extra marital affair is a binary variable (either a person will have or not),

so we can fit logistic regression model here to predict the probability of extra marital affair.

install.packages('AER')

data(Affairs,package="AER")

	X	naffairs	kids	vryunhap	unhap	avgmarr	hapavg	vryhap	antirel	notrel	slghtrel	smerel	vryrel	yrsmarr1	yrsmarr2	yrsmarr3	yrsmarr4	yrsmarr5
1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
2	2	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
3	3	3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
4	4	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
5	5	3	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
6	6	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
7	7	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
8	8	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0
9	9	7	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
10	10	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
11	11	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
12	12	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
13	13	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
14	14	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
15	15	12	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
16	16	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0
17	17	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
18	18	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0
19	19	1	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0
20	20	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0

2.) Output variable -> y
y -> Whether the client has subscribed a term deposit or not
Binomial ("yes" or "no")

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
7	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
11	41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
12	29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
13	53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
14	58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
15	57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
16	51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
17	45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no

3.) Suppose we are interested in the factors that influence whether a political candidate wins an election.

The outcome (response) variable is binary (0/1); win or lose.

The predictor variables of interest are the amount of money spent on the campaign, the amount of time spent campaigning negatively and whether or not the candidate is an incumbent.

	Election.id	Result	Year	Amount.Spent	Popularity.Rank
1	*NA*	*NA*	*NA*	*NA*	*NA*
2	122	0	32	3.81	3
3	315	1	48	6.32	2
4	201	1	51	3.67	1
5	965	0	40	2.93	4
6	410	1	52	3.60	1
7	150	0	35	4.20	4
8	743	1	39	5.66	2
9	612	1	42	4.32	3
10	206	1	44	3.26	3
11	792	0	50	4.52	4

Hints:

1. Business Problem

    1.1. Objective
    1.2. Constraints (if any)

2. Data Pre-processing

    2.1 Data cleaning, Feature Engineering, EDA etc.

3. Model Building

    3.1 Partition the dataset
    3.2 Model(s) - Reasons to choose any algorithm
    3.3 Model(s) Improvement steps
    3.4 Model Evaluation
    3.5 Python and R codes

4. Deployment

    4.1 Deploy solutions using R shiny and Python Flask.

5. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

Note:

1. For each assignment the solution should be submitted in the format

2. Research and Perform all possible steps for improving the model(s) accuracy.

    Ex: Transformations, Feature Engineering, Hyper Parameter tuning, Outlier treatment, etc.

3. All the codes (executable programs) are running without errors

4. Documentation of the module should be submitted along with R & Python codes, elaborating on every steps mentioned here.