Dataset insights:

- The data set initially had 41189 observations including the missing values.
- The missing values are assumed to be unknown for variables.

Variable	Value
job	unknown
education	unknown
marital	unknown
loan	unknown
housing	unknown

Data Preparation:

- Missing values treatment
    1. Removed these variables to avoid noise in my final result.
    2. After removing the missing values we are left with a total of 38245 observations.

- Outliers Treatment
    1. proc univariate to find out data statistics
    2. Outliers are treated for the variables are treated individually
    3. Duration to 99th percentile

- Data Conversion
    1. Converted all the categorical variable to numerical equivalents
    2. Values are given like 0,1,2,3 depending on their business significance

Correlation:

- Made a correlation matrix and found out the variables with high covariance(positive or negative) index coefficient.
- Prediction looking at the correlation matrix:
    - Economic factors are strongly correlated among each other
    - Previous is negatively correlated to the economic factors
    - Contact is slightly related P to the economic variables
    - Response is correlated to strong positively duration and negatively correlated to “pdays” “previous” “poutcome”
    - Age is related to marital
    - Independent variables
        - Month
        - Day of week
        - Campaign
        - Job
        - Education
        - Housing Loan

Factor Analysis:

We observe multi co-linearity among different variables so we try to reduce the no of variables by discarding the not important ones which will otherwise create noise in our final result.

	Factor1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor10
emp_var_rate	0.965	-0.142	0.049	0.027	-0.065	-0.044	-0.015	-0.016	0.003	-0.015
euribor3m	0.942	-0.201	0.057	0.061	-0.081	-0.002	0.029	0.024	0.003	-0.015
nr_employed	0.866	-0.295	-0.027	0.020	-0.116	-0.177	-0.021	0.003	0.006	-0.006
cons_price_idx	0.820	0.111	0.079	-0.042	0.043	0.252	-0.124	-0.115	0.001	-0.040
previous	-0.310	0.882	-0.220	0.010	0.053	0.015	0.031	0.024	0.015	-0.012
pdays	0.208	-0.725	-0.584	0.008	-0.099	0.018	-0.018	0.001	-0.001	-0.006
poutcome	0.147	-0.074	0.940	-0.021	0.063	-0.028	-0.012	-0.033	-0.016	0.020
age	-0.033	0.042	0.042	0.826	0.014	-0.034	-0.110	-0.018	-0.023	0.024
marital	-0.046	0.039	0.049	-0.768	0.016	-0.002	0.078	0.009	-0.051	0.053
duration	0.041	-0.063	-0.067	-0.014	0.893	-0.009	-0.042	0.055	0.010	-0.007
response	-0.241	0.207	0.202	0.011	0.742	-0.033	0.075	0.016	0.001	-0.015
month	-0.197	0.025	-0.066	-0.045	-0.023	0.814	-0.080	0.072	0.016	-0.005
contact	0.483	-0.034	0.082	0.030	-0.023	0.708	-0.046	-0.076	-0.012	-0.053
education	-0.032	0.008	0.016	-0.164	-0.013	-0.145	0.728	-0.025	0.056	-0.021
job	0.062	0.104	-0.119	0.096	0.041	0.006	0.537	-0.086	-0.343	0.339
cons_conf_idx	0.227	0.004	0.334	0.312	-0.006	0.247	0.405	0.197	-0.025	-0.034
default	0.242	0.029	-0.015	0.212	-0.023	-0.003	-0.509	-0.039	-0.184	0.208
day_of_week	0.107	0.119	-0.009	-0.052	-0.050	-0.085	-0.050	0.841	0.023	-0.053
campaign	0.206	0.106	0.005	-0.033	-0.131	-0.122	-0.043	-0.521	0.055	-0.095
loan	0.027	0.025	-0.032	0.045	0.015	0.008	0.066	-0.041	0.913	0.113
housing	-0.070	-0.019	0.030	-0.053	-0.023	-0.037	-0.049	0.046	0.114	0.900

Analytics Initiatives:

- We can find key drivers for revenue or profit by store.
- Y (DV)= Response
- X's (IV's) = emp_var_rate previoous pdays poutcome age marital duration month job day_of_week campaign loan housing
- Technique = Linear Regression

Model Fitting:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	92.1	Somers' D	0.841
Percent Discordant	7.9	Gamma	0.841
Percent Tied	0	Tau-a	0.169
Pairs	71000235	c	0.921

Concordance of 92 %v approximation and correspondingly high Somer'd value

Partition for the Hosmer and Lemeshow Test					
		response_c = 1		response_c = 0	
Group	Total	Observed	Expected	Observed	Expected
1	2657	0	14.21	2657	2642.79
2	2658	1	23.93	2657	2634.07
3	2656	4	32.56	2652	2623.44
4	2656	13	43.75	2643	2612.25
5	2656	19	62.08	2637	2593.92
6	2656	49	99.74	2607	2556.26
7	2656	157	171.14	2499	2484.86
8	2656	403	300.71	2253	2355.29
9	2656	808	600.25	1848	2055.75
10	2657	1561	1666.63	1096	990.37

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
292.5712	8	<.0001

- p value is less tha 0.01% so we accept the H0
- H0 = Observed value for response equals our Estimated value
- H0=Ei-Oi=0

Conclusion: Dataset fit for Logistic Regression

Customer Targeting by Deciling:

We employ the method 2 as mentioned above of splitting the data into ratio of 70:30 named Development and Val. A proper logit function creates beta values for the individual multi-variate analysis. Now this model is used to verify the validation sample.

Development model 70% of data size				
Rank for variable newpred	cnt	response_cnt	min_p	max_p
10	2656	1561	0.345775	0.996356
9	2657	808	0.151467	0.345098
8	2656	403	0.083234	0.151366
7	2657	157	0.048447	0.083214
6	2656	49	0.028681	0.048438
5	2657	19	0.019333	0.028675
4	2656	13	0.014102	0.019332
3	2657	4	0.010573	0.014101
2	2656	1	0.007391	0.010572
1	2656	0	0.000619	0.00739
Total	**26564**	**3015**		

Validation Output on our predicted model				
Rank for Variable Prob	**cnt**	**response_cnt**	**min_p**	**max_p**
10	1168	593	0.00624	0.690195
9	1168	346	0.002533	0.006233
8	1169	186	0.00133	0.002531
7	1167	75	0.0007	0.001329
6	1169	34	0.000388	0.0007
5	1168	8	0.000245	0.000387
4	1168	0	0.000173	0.000245
3	1168	1	0.000127	0.000173
2	1168	0	0.000088	0.000127
1	1168	0	7.62E-06	0.000088
Total	**11681**	**1243**		

- Since our dependent variable response is in a yes or a no of customer being able to buy the product the best suited technique according to me would be to go for logistic regression.
- We are able to conclude our 91% customers who will give a response of yes in the top three deciles.
- So customer targeting becomes very easy for us.
- Thus there can be a huge reduction in cost to reach out profitable customers.
- The development and validation model are almost same.

Main Conclusion:

Taken a vif of 74%, the model is successfully build which can be used to predict the behaviour of a customer to buy the product in a yes or a no will depend on the 15 co-variant variables mentioned above in the report which were used in developing the model.

Submitted by:
Lokesh Sharma
Lokesharma92@gmail.com
9663653800