MS Business Intelligence & Analytics
Fall 2016

Optimization and Process Analytics
BIA 650-A

December 15, 2016

A DATA DRIVEN APPROACH TO OPTIMIZE THE SUCCESS RATE OF BANK TELEMARKETING

Instructor: Prof. Edward Stohr

Members:
Abhinav Panwar
Arun Krishnamurthy
Piyush Bhattad
Puneet Manocha
Vaibhav Desai

ABSTRACT

Credit on international markets became more restricted for banks because of global financial crisis, putting emphasis on internal clients and their deposits to gather funds. This fact led to a demand for knowledge about client's response to telemarketing campaigns.

This work describes a data mining approach to extract valuable knowledge from recent Portuguese bank telemarketing campaign data. Data Analysis, Data Visualization, Regression Methods, Clustering and, Segmentation are conducted using the SAS tool. Optimization and Sensitivity analysis are carried out using Excel's Solver Table Functionality.

OBJECTIVE

The objective of the model is to increase the success rate of bank telemarketing by optimizing the combination of attributes that contribute to successful term deposit subscription.

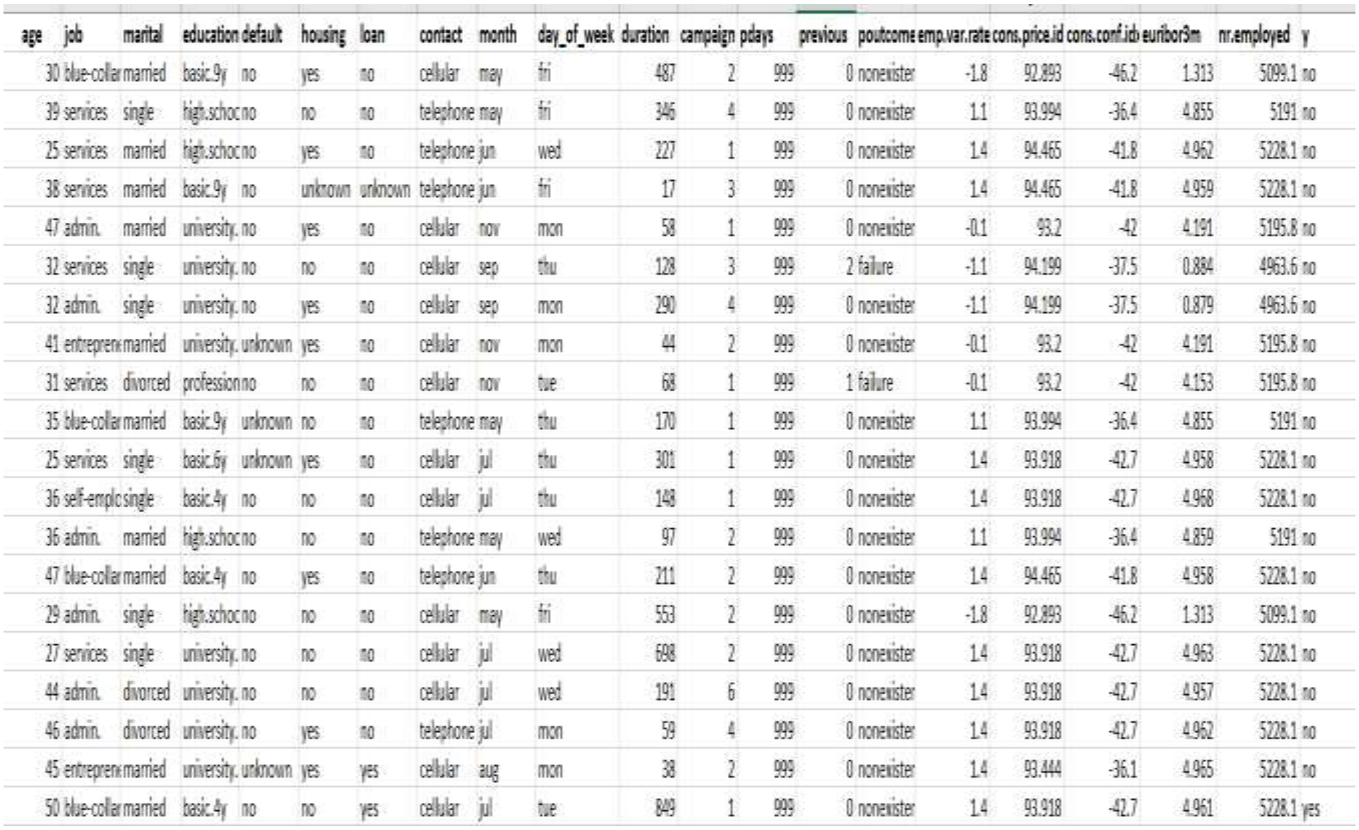
DATASET INFORMATION

1. **Age:** Numeric
2. **Job:** type of job (Categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. **Marital:** marital status (categorical: 'divorced', 'married', 'single', 'unknown')
4. **Education**: (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. **Default**: has credit in default? (categorical: 'no', 'yes', 'unknown')
6. **Housing**: has housing loan? (categorical: 'no', 'yes', 'unknown')
7. **Loan**: has personal loan? (categorical: 'no', 'yes', 'unknown')
8. **Contact**: contact communication type (categorical: 'cellular', 'telephone')
9. **Month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. **Day_of_Week**: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
11. **Duration**: last contact duration, in seconds (numeric).
12. **Campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. **Pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric)
14. **Previous**: number of contacts performed before this campaign and for this client (numeric)
15. **Poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
16. **Employment Variation Rate**: employment variation rate - quarterly indicator (numeric)
17. **Consumer Price Index**: consumer price index - monthly indicator (numeric)
18. **Consumer Confidence Index**: consumer confidence index - monthly indicator (numeric)
19. **Euribor3m**: euribor 3 month rate - daily indicator (numeric)
20. **Number of Employees**: quarterly indicator (numeric)

Output variable (desired target):

21. **Y**: has the client subscribed a term deposit? (binary: 'yes','no')

Below is the snapshot of the dataset we used for processing:



age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.id	cons.conf.idx	euribor3m	nr.employed	y
30	blue-collar	married	basic.9y	no	yes	no	cellular	may	fri	487	2	999	0	nonexister	-1.8	92.893	-46.2	1.313	5099.1	no
39	services	single	high.schoc	no	no	no	telephone	may	fri	346	4	999	0	nonexister	1.1	93.994	-36.4	4.855	5191	no
25	services	married	high.schoc	no	yes	no	telephone	jun	wed	227	1	999	0	nonexister	1.4	94.465	-41.8	4.962	5228.1	no
38	services	married	basic.9y	no	unknown	unknown	telephone	jun	fri	17	3	999	0	nonexister	1.4	94.465	-41.8	4.959	5228.1	no
47	admin.	married	university.	no	yes	no	cellular	nov	mon	58	1	999	0	nonexister	-0.1	93.2	-42	4.191	5195.8	no
32	services	single	university.	no	no	no	cellular	sep	thu	128	3	999	2	failure	-1.1	94.199	-37.5	0.884	4963.6	no
32	admin.	single	university.	no	yes	no	cellular	sep	mon	290	4	999	0	nonexister	-1.1	94.199	-37.5	0.879	4963.6	no
41	entreprene	married	university.	unknown	yes	no	cellular	nov	mon	44	2	999	0	nonexister	-0.1	93.2	-42	4.191	5195.8	no
31	services	divorced	profession	no	no	no	cellular	nov	tue	68	1	999	1	failure	-0.1	93.2	-42	4.153	5195.8	no
35	blue-collar	married	basic.9y	unknown	no	no	telephone	may	thu	170	1	999	0	nonexister	1.1	93.994	-36.4	4.855	5191	no
25	services	single	basic.6y	unknown	yes	no	cellular	jul	thu	301	1	999	0	nonexister	1.4	93.918	-42.7	4.958	5228.1	no
36	self-emplo	single	basic.4y	no	no	no	cellular	jul	thu	148	1	999	0	nonexister	1.4	93.918	-42.7	4.968	5228.1	no
36	admin.	married	high.schoc	no	no	no	telephone	may	wed	97	2	999	0	nonexister	1.1	93.994	-36.4	4.859	5191	no
47	blue-collar	married	basic.4y	no	yes	no	telephone	jun	thu	211	2	999	0	nonexister	1.4	94.465	-41.8	4.958	5228.1	no
29	admin.	single	high.schoc	no	no	no	cellular	may	fri	553	2	999	0	nonexister	-1.8	92.893	-46.2	1.313	5099.1	no
27	services	single	university.	no	no	no	cellular	jul	wed	698	2	999	0	nonexister	1.4	93.918	-42.7	4.963	5228.1	no
44	admin.	divorced	university.	no	no	no	cellular	jul	wed	191	6	999	0	nonexister	1.4	93.918	-42.7	4.957	5228.1	no
46	admin.	divorced	university.	no	yes	no	telephone	jul	mon	59	4	999	0	nonexister	1.4	93.918	-42.7	4.962	5228.1	no
45	entreprene	married	university.	unknown	yes	yes	cellular	aug	mon	38	2	999	0	nonexister	1.4	93.444	-36.1	4.965	5228.1	no
50	blue-collar	married	basic.4y	no	no	yes	cellular	jul	tue	849	1	999	0	nonexister	1.4	93.918	-42.7	4.961	5228.1	yes

DATA CLEANSING

Merging the datasets

Merging Telemarketing data to the corporate database to get variables for modeling

Creating new variables

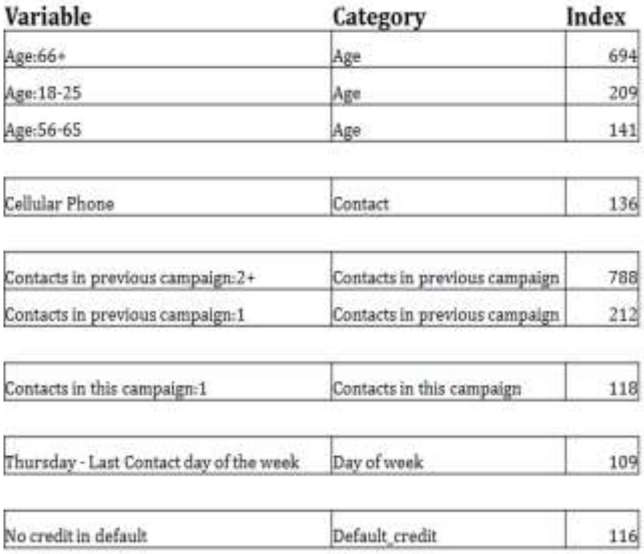
Add Age brackets to make more meaning of the data.

Restructuring the data

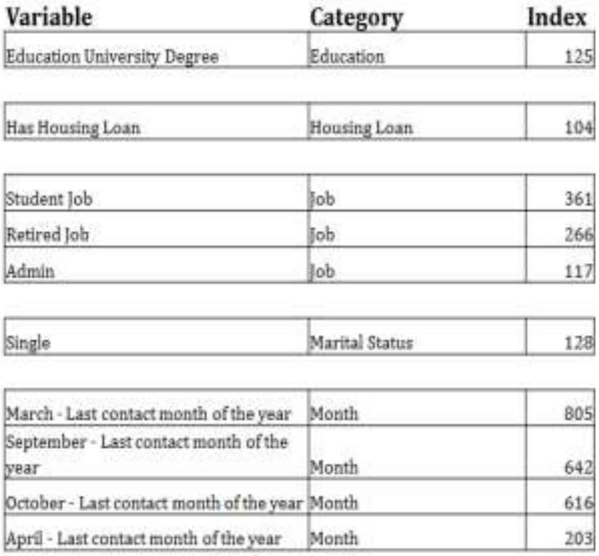
Convert all variables to 0/1 for Logistic regression.
Remove Outliers

Profile : Responders vs Nonresponders

Term-deposit subscribers have the following characteristics



Variable	Category	Index
Age:66+	Age	694
Age:18-25	Age	209
Age:56-65	Age	141
Cellular Phone	Contact	136
Contacts in previous campaign:2+	Contacts in previous campaign	788
Contacts in previous campaign:1	Contacts in previous campaign	212
Contacts in this campaign:1	Contacts in this campaign	118
Thursday - Last Contact day of the week	Day of week	109
No credit in default	Default_credit	116



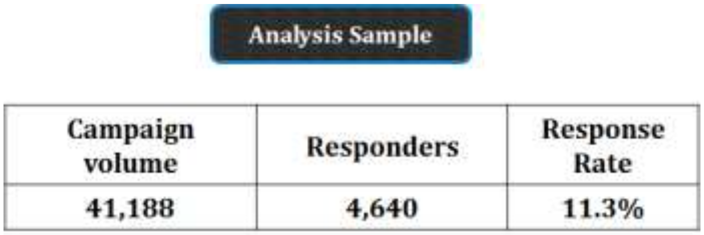
Variable	Category	Index
Education University Degree	Education	125
Has Housing Loan	Housing Loan	104
Student Job	Job	361
Retired Job	Job	266
Admin	Job	117
Single	Marital Status	128
March - Last contact month of the year	Month	805
September - Last contact month of the year	Month	642
October - Last contact month of the year	Month	616
April - Last contact month of the year	Month	203

The marketers can use straight-selects to target customers.
Variables with indices can be selected.

Before building a model, the marketers can just pick customers who have are above the age of 66, have a cellular phone, Single and have a Education university degree. These categories have indices and could give good results in the next campaign.

METHODOLOGY

The objective of the model is to segment customers into “good”, “fair” and “poor” responders in order to allow a more strategic targeting based on a customer’s propensity to subscribe to a term deposit.



Analysis Sample

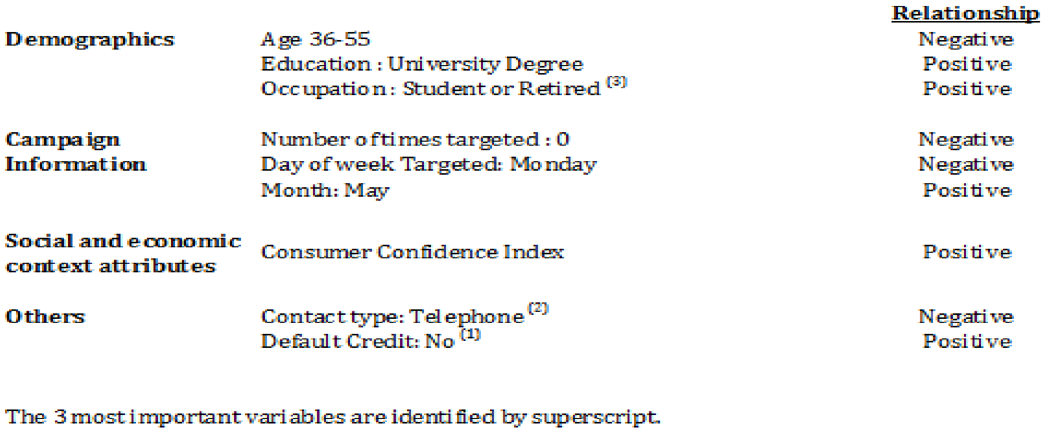
Campaign volume	Responders	Response Rate
41,188	4,640	11.3%

The telemarketing data - customers who have said Yes/No to the term-deposit offer is merged with the customer database. The database has more than 20+ variables of Demographic, Geographic, Socio economic attributes and previous campaign information.

A logistic regression model was then built to rank subscribers by their likelihood of responding to the offer. The model was built on a random 50% sample of the customers. The remaining 50% of the customers were set aside for model validation.

The model identifies good and poor responders for term-deposits.

The model identifies some unique characteristics which differentiates the responders from Nonresponders. The responders have the following characteristics:



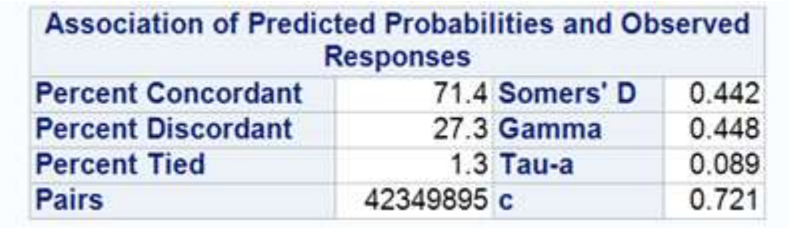
		Relationship
Demographics	Age 36-55	Negative
	Education : University Degree	Positive
	Occupation : Student or Retired [3]	Positive
Campaign Information	Number of times targeted : 0	Negative
	Day of week Targeted: Monday	Negative
	Month: May	Positive
Social and economic context attributes	Consumer Confidence Index	Positive
Others	Contact type: Telephone [2]	Negative
	Default Credit: No [1]	Positive

The 3 most important variables are identified by superscript.

The top three statistically significant variables - highest Wald-chi square values are

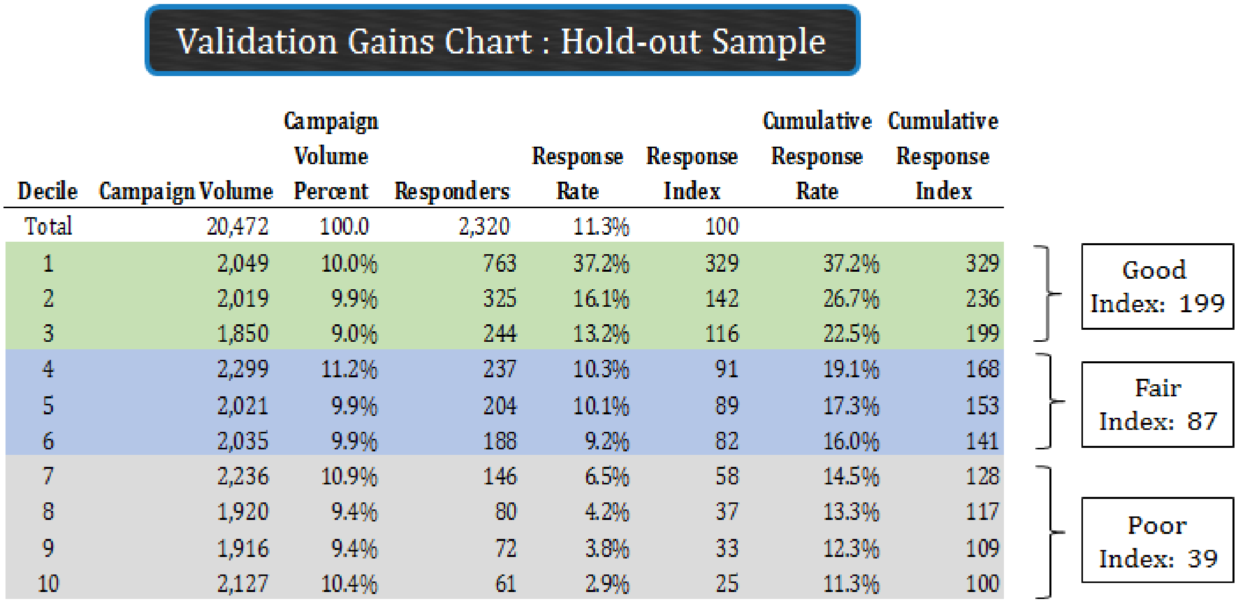
1. Subscribers who don't have a default credit
2. Contact type: Telephone
3. Student/Retired occupation

Relationship positive signifies a higher score from that variable in the model and negative signifies a lower score from that variable.



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	71.4	Somers' D	0.442
Percent Discordant	27.3	Gamma	0.448
Percent Tied	1.3	Tau-a	0.089
Pairs	42349895	c	0.721

The model had a percent concordance of 71.4. It basically explains explains the model reliability.
A c-value of 0.721 shows the there's good association between the predictor variables and dependent variable.



Validation Gains Chart : Hold-out Sample

Decile	Campaign Volume	Campaign Volume Percent	Responders	Response Rate	Response Index	Cumulative Response Rate	Cumulative Response Index	
Total	20,472	100.0	2,320	11.3%	100			
1	2,049	10.0%	763	37.2%	329	37.2%	329	Good Index: 199
2	2,019	9.9%	325	16.1%	142	26.7%	236	
3	1,850	9.0%	244	13.2%	116	22.5%	199	
4	2,299	11.2%	237	10.3%	91	19.1%	168	Fair Index: 87
5	2,021	9.9%	204	10.1%	89	17.3%	153	
6	2,035	9.9%	188	9.2%	82	16.0%	141	
7	2,236	10.9%	146	6.5%	58	14.5%	128	Poor Index: 39
8	1,920	9.4%	80	4.2%	37	13.3%	117	
9	1,916	9.4%	72	3.8%	33	12.3%	109	
10	2,127	10.4%	61	2.9%	25	11.3%	100	

The model is used to score, rank and divide the customers into 10 deciles based on the their propensity to respond. The top decile will have the maximum number of responders and the bottom decile will have the minimum number of responders.

The index in the first decile is 329 which indicates by targeting the first decile, the response rate is improving by a factor of 3.29

Cumulative index of 236 in decile indicates when you target the first two deciles, you improve the campaign by a factor of 2.36.

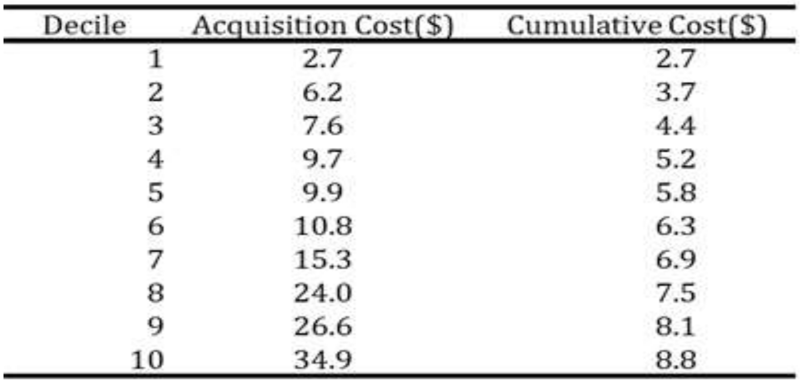
Deciles 1-3 each with an index above 100 can be segmented as good responders. They have a cumulative index of 199.
Deciles 4-6 can be segmented as fair responders. They have a cumulative index of 87.
Deciles 4-6 can be segmented as poor responders. They have a cumulative index of just 39.

The marketer can use different strategies across these segments to obtain a good response rate.

Another way to interpret the gains chart:

Let us assume it costs $1 to target each customer. The bank has spend $41,888 to acquire 4,640 customers. This indicates that the bank spends approximately $8.8 to acquire one customer.

The bank can use different cost allocation strategies across these deciles to optimize campaign budget.



Decile	Acquisition Cost($)	Cumulative Cost($)
1	2.7	2.7
2	6.2	3.7
3	7.6	4.4
4	9.7	5.2
5	9.9	5.8
6	10.8	6.3
7	15.3	6.9
8	24.0	7.5
9	26.6	8.1
10	34.9	8.8

From the above table, we can see that it costs $2.7 to target a customer in the top decile. By targeting the top 5 deciles, the bank can save close to 50% of the campaign budget.

Optimization Model

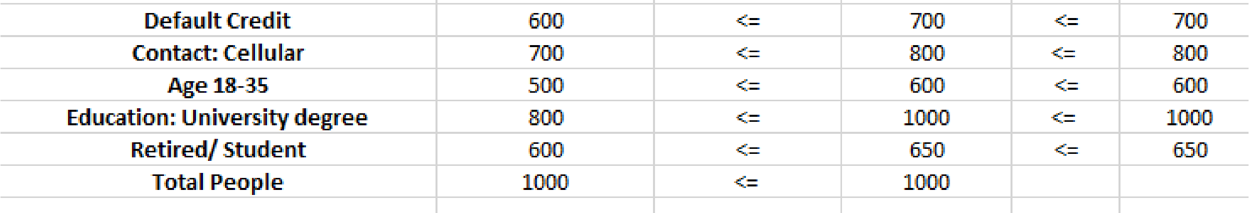
Once we get the significant variables from logistic regression which predict the chances of a person subscribing to term deposit , next step is to optimize the model according to Marketer. Following are the significant variables which predict the outcome:

a. Default Credit
b. Contact type-Cellular
c. Age- 18 to 35
d. Education- University degree
e. Retired or Student

Out of the attributes we get above, a Marketer may want to focus on specific attributes only. A Marketers may have a strategy according to the interest of their organisation, So that they would be attracting people with certain characteristics.

Now, from the analysis we did in previous step, people with characteristics mentioned above predicts that the person will become customer or not. So, a Marketer will want to focus on these characteristics only, as the rest attributes cannot predict the outcome.

Below is the snapshot of strategy which a Marketer can have:



Default Credit	600	<=	700	<=	700
Contact: Cellular	700	<=	800	<=	800
Age 18-35	500	<=	600	<=	600
Education: University degree	800	<=	1000	<=	1000
Retired/ Student	600	<=	650	<=	650
Total People	1000	<=	1000		

Following is the description of what above strategy means:

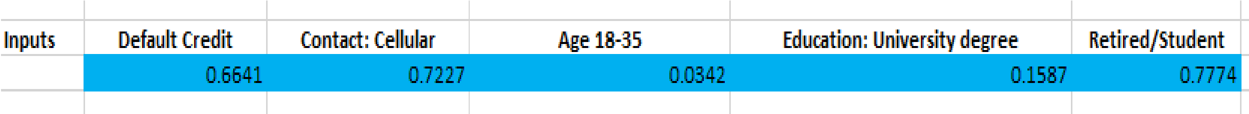
1. Total number of people to target is 1000
2. People having default credit should be at least 600 and not more than 700
3. People having cellular contact type should be at least 700 and not more than 800
4. People having age between 18 and 35 should be at least 500 and not more than 600
5. People having university degree should be at least 800 and not more than 1000
6. People who are either retired or student should be at least 600 and not more than 650

We created a spreadsheet model where a marketer can enter the details of his/her strategy, and a solution will be proposed which can be used to generate maximum positive response subjected to constraints provided by the marketer. The strategy of a marketer is nothing but a constraint which should be satisfied always.

Following is the description of all the components in the spreadsheet model:

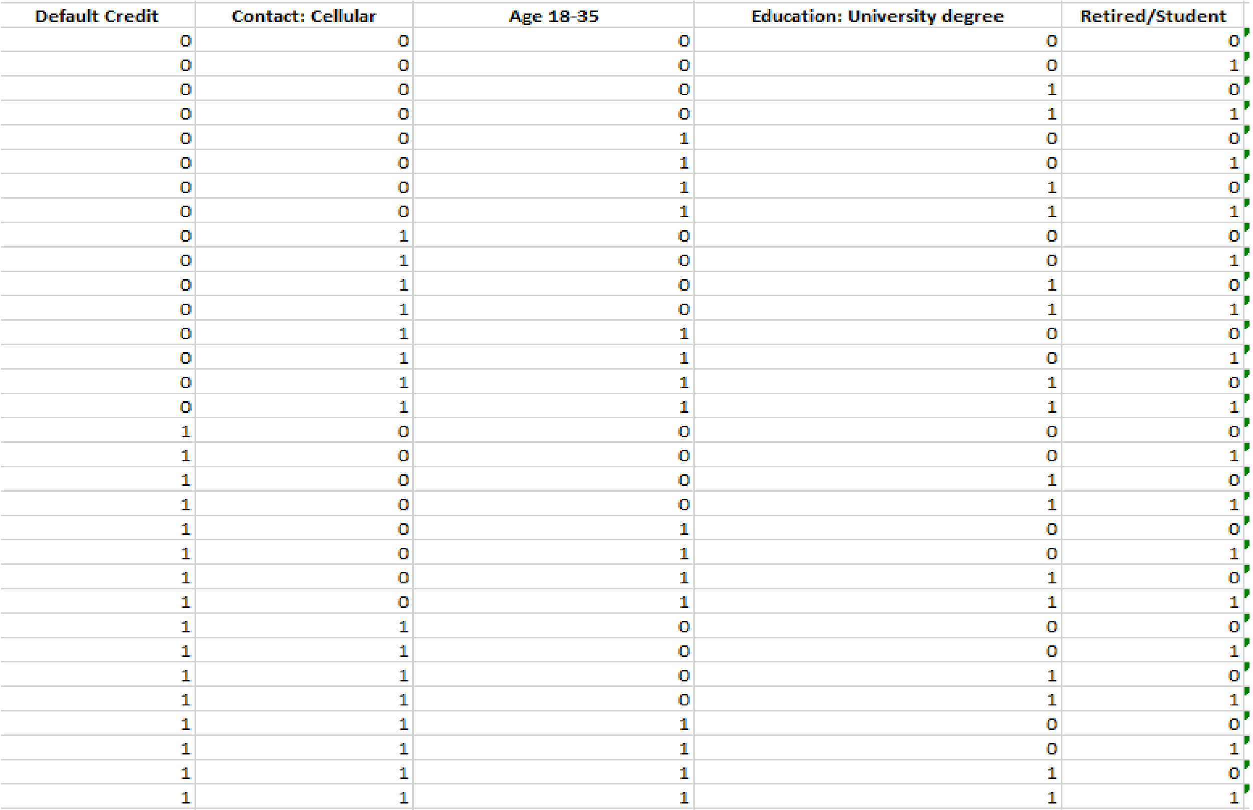
1. Inputs:

a. Coefficients of all the parameters we get from regression equation.



Inputs	Default Credit	Contact: Cellular	Age 18-35	Education: University degree	Retired/Student
	0.6641	0.7227	0.0342	0.1587	0.7774

b. All possible combinations of all the predictor variables. Since there are 5 variables, there can be 32 (2^5) combinations in total.



Default Credit	Contact: Cellular	Age 18-35	Education: University degree	Retired/Student
0	0	0	0	0
0	0	0	0	1
0	0	0	1	0
0	0	0	1	1
0	0	1	0	0
0	0	1	0	1
0	0	1	1	0
0	0	1	1	1
0	1	0	0	0
0	1	0	0	1
0	1	0	1	0
0	1	0	1	1
0	1	1	0	0
0	1	1	0	1
0	1	1	1	0
0	1	1	1	1
1	0	0	0	0
1	0	0	0	1
1	0	0	1	0
1	0	0	1	1
1	0	1	0	0
1	0	1	0	1
1	0	1	1	0
1	0	1	1	1
1	1	0	0	0
1	1	0	0	1
1	1	0	1	0
1	1	0	1	1
1	1	1	0	0
1	1	1	0	1
1	1	1	1	0
1	1	1	1	1

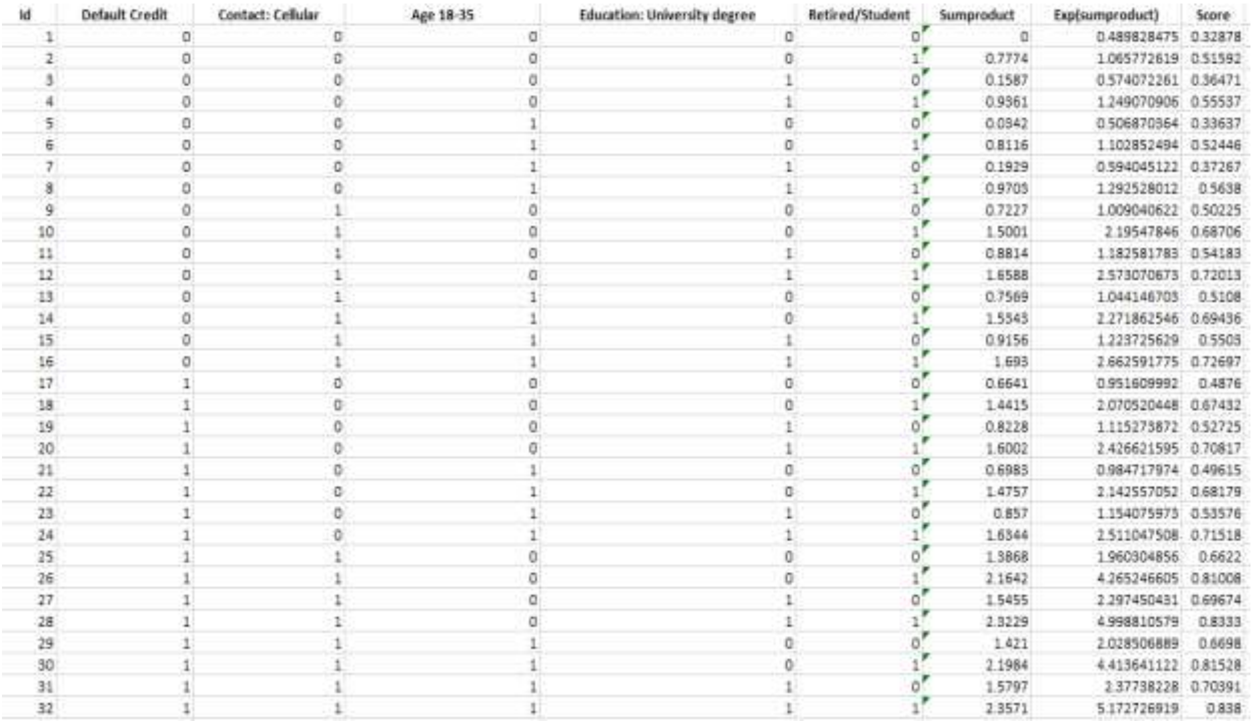
C. For each of the combination above, we calculated score using logistic regression equation. Below is the logistic regression equation we got from SAS processing:

Sumproduct = -0.7137 + 0.6641*Default_credit +0.7227*Cellular Contact type + 0.0342*Age 18-35 + 0.1587*University degree + 0.7774*Retired or Student

Score = EXP(Sumproduct)/(1+ EXP(Sumproduct))

This means higher the score, more is the probability that the person will respond positively to term subscription.

Below is the snapshot of score computed for each of the combination:

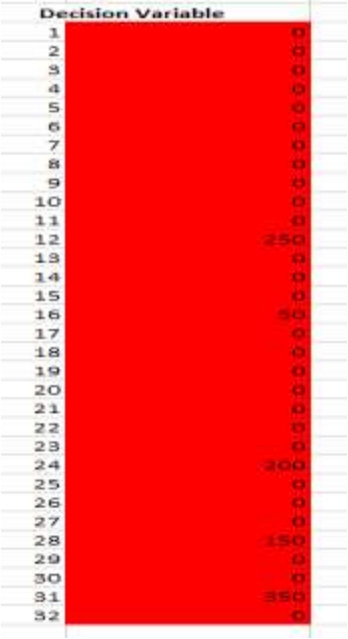


Id	Default Credit	Contact: Cellular	Age 18-35	Education: University degree	Retired/Student	Sumproduct	Exp(sumproduct)	Score
1	0	0	0	0	0	0	0.489828475	0.32878
2	0	0	0	0	1	0.7774	1.065772619	0.51592
3	0	0	0	1	0	0.1587	0.574072261	0.36471
4	0	0	0	1	1	0.9361	1.249070906	0.55537
5	0	0	1	0	0	0.0342	0.506870364	0.33637
6	0	0	1	0	1	0.8116	1.102852494	0.52446
7	0	0	1	1	0	0.1929	0.594045122	0.37267
8	0	0	1	1	1	0.9703	1.292528012	0.5638
9	0	1	0	0	0	0.7227	1.009040622	0.50225
10	0	1	0	0	1	1.5001	2.19547846	0.68706
11	0	1	0	1	0	0.8814	1.182581783	0.54183
12	0	1	0	1	1	1.6588	2.573070673	0.72013
13	0	1	1	0	0	0.7569	1.044146703	0.5108
14	0	1	1	0	1	1.5545	2.271862546	0.69436
15	0	1	1	1	0	0.9156	1.223725629	0.5503
16	0	1	1	1	1	1.693	2.662591775	0.72697
17	1	0	0	0	0	0.6641	0.951609992	0.4876
18	1	0	0	0	1	1.4415	2.070520448	0.67432
19	1	0	0	1	0	0.8228	1.115273872	0.52725
20	1	0	0	1	1	1.6002	2.426621595	0.70817
21	1	0	1	0	0	0.6983	0.984717974	0.49615
22	1	0	1	0	1	1.4757	2.142557052	0.68179
23	1	0	1	1	0	0.857	1.154075973	0.53576
24	1	0	1	1	1	1.6344	2.511047508	0.71518
25	1	1	0	0	0	1.3868	1.960304856	0.6622
26	1	1	0	0	1	2.1642	4.265246605	0.81008
27	1	1	0	1	0	1.5455	2.297450431	0.69674
28	1	1	0	1	1	2.3229	4.998810579	0.8333
29	1	1	1	0	0	1.421	2.028506889	0.6698
30	1	1	1	0	1	2.1984	4.413641122	0.81528
31	1	1	1	1	0	1.5797	2.37738228	0.70391
32	1	1	1	1	1	2.3571	5.172726919	0.838

We have given ID to each possible combination.

2. **Decision Variable:**

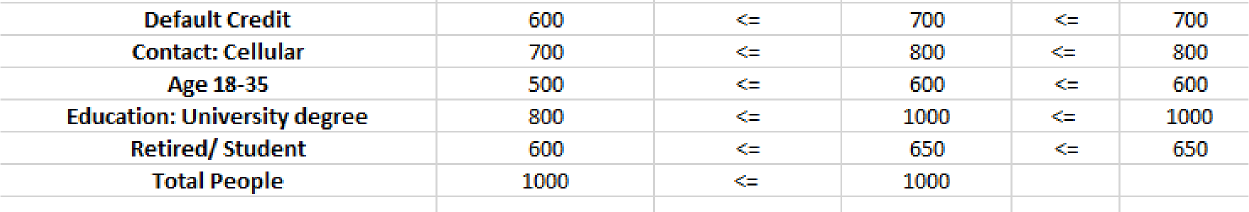
This is the variable which the marketer wants to find. It is equal to number of people to be targeted for each combination. We have given ID to each possible combination. Following is the snapshot of the decision variable from spreadsheet model we created:



Decision Variable	
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	250
13	0
14	0
15	0
16	50
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	200
25	0
26	0
27	0
28	150
29	0
30	0
31	350
32	0

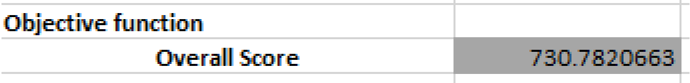
From above result we can see, that 150 people out of 1000 people to be selected, are of ID 28, which means their characteristics are Default Credit-Yes, Cellular phone-Yes, University degree-Yes, Age 18-35-No and Retired/Student-Yes.

3. **Constraints:** Constraints are basically the strategy of a marketer which must be satisfied always. As mentioned above, for our model this is strategy we implemented:



Default Credit	600	<=	700	<=	700
Contact: Cellular	700	<=	800	<=	800
Age 18-35	500	<=	600	<=	600
Education: University degree	800	<=	1000	<=	1000
Retired/ Student	600	<=	650	<=	650
Total People	1000	<=	1000		

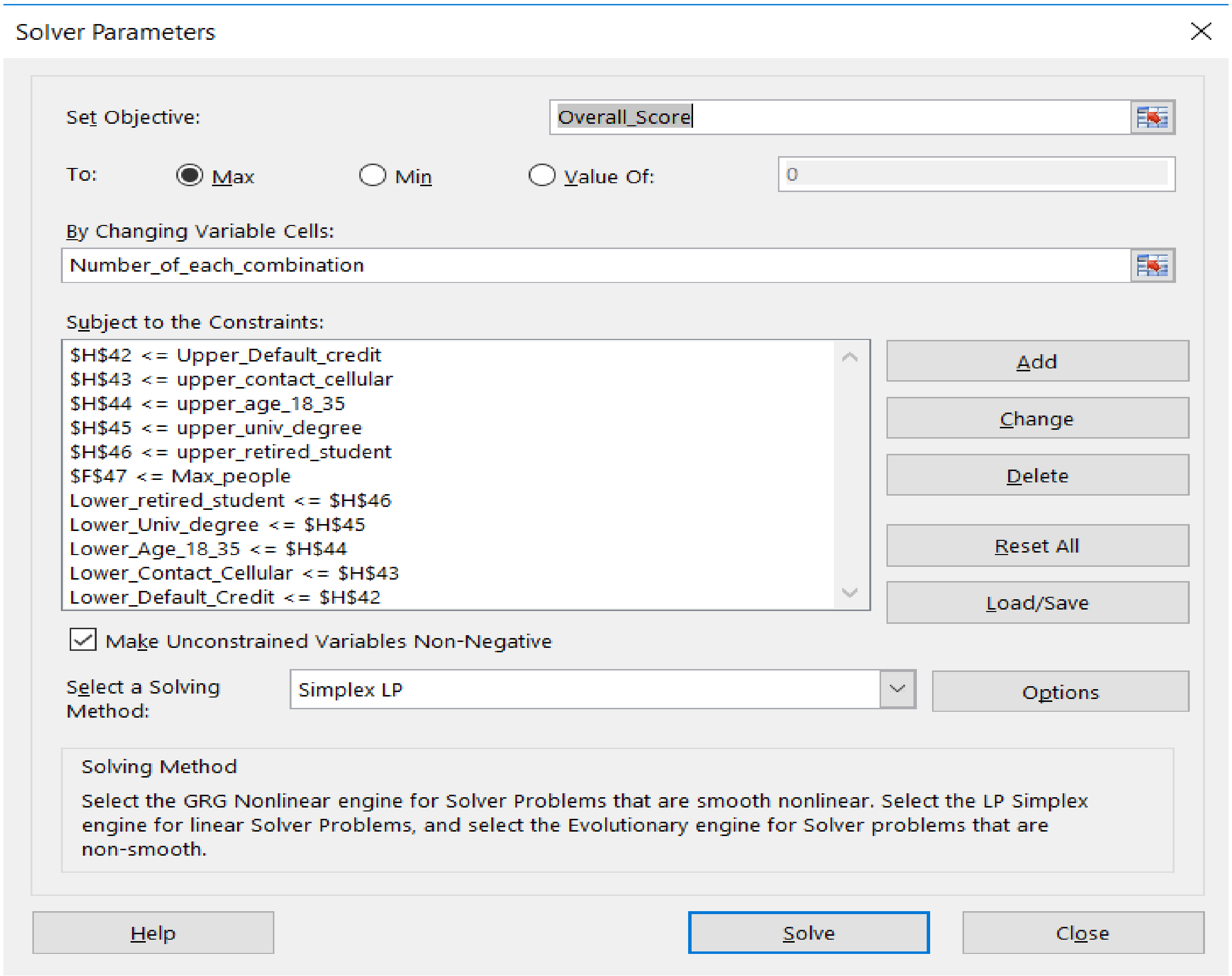
4. **Objective Function:** This is the total score which needs to be maximized as large as possible such that all the constraints are satisfied. This total score is calculated by multiplying number of combination of each ID by their corresponding score , and then adding them. Following is the overall computed score:



Objective function	
Overall Score	730.7820663

The above model we created is a linear model. We used Excel's plug-in, Solver with Simplex algorithm, to compute the solution.

All the above information (objective function cell, decision variable cells, and all constraints) are added to Solver dialog box. Below is the snapshot of the solver dialog box, where all information is fed:



Solver Parameters

Set Objective: Overall_Score

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells:
Number_of_each_combination

Subject to the Constraints:

```
$H$42 <= Upper_Default_credit
$H$43 <= upper_contact_cellular
$H$44 <= upper_age_18_35
$H$45 <= upper_univ_degree
$H$46 <= upper_retired_student
$F$47 <= Max_people
Lower_retired_student <= $H$46
Lower_Univ_degree <= $H$45
Lower_Age_18_35 <= $H$44
Lower_Contact_Cellular <= $H$43
Lower_Default_Credit <= $H$42
```

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Simplex LP

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

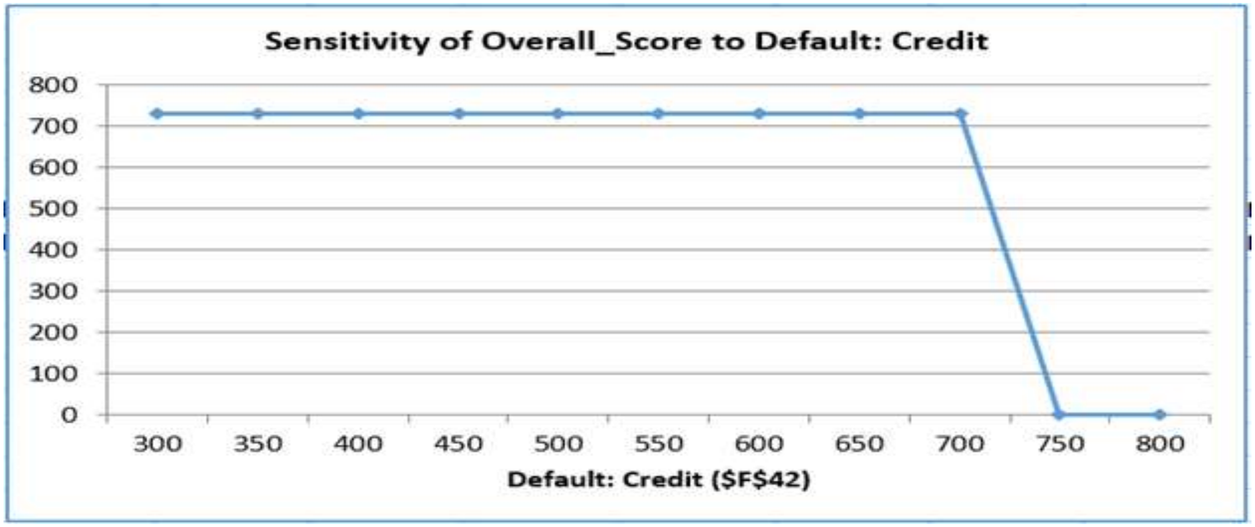
Once we run the solver, we get the solution. The solution computed has the highest probability to respond positively to term subscription subject to constraints provided by the marketer.

OUTPUT ANALYSIS

After running the solver, we get the number of people of each possible combination, which the marketer should target.
For our solution, out of the 1000 people we targeted following is the solution:

- 250 people having cellular contact type, university degree and Retired/Student
- 50 people having cellular contact type, Age between 18 to 35, university degree and Retired/Student
- 200 people having default credit, Age between 18 to 35, university degree and Retired/Student
- 150 people having default credit, cellular contact type, university degree and Retired/Student
- 350 people having default credit, cellular contact type, Age between 18 to 35 and university degree

We also did sensitivity analysis using solver-table. We varied the lower limit of number of people having Default Credit-Yes from 300 to 800 to see its effect on the overall score. Below is the graph depicting the effect:



As we can see, from 300 to 700, there is no effect on the overall score. But after 700, the score starts decreasing. And once the number reaches 750, it has no contribution to the overall score.

A marketer can use sensitivity analysis to see the effect of change in input on output, and thereby forming appropriate marketing strategy.

RECOMMENDATIONS

- A straight-select of customers can be used from the Profile report

- Deciles 1-3 can be categorized as good responders, and they have a response index of 199. Deciles 4-6 can be categorized as fair responders, and they have a response index of 87. The bottom 4 deciles can be categorized as poor responders and they have a response index of just 39.

- Cost allocation and different marketing techniques should be used across deciles in order to achieve optimum response rate.

- Always target people with following attributes:
  Default Credit: Yes
  Contact Cellular: Yes
  Age 18-35: Yes
  University Degree: Yes
  Retired/Student: Yes

- If a Marketer wants to target some specific group, use spreadsheet model with Solver to identify the people from the database who can respond positively

REFERENCES

- Portuguese Bank Telemarketing Data Set-[DB/OL] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

- S. May & V. A. Clark Practical Multivariate Analysis Fifth Edition [M].NW: CRC Press, 2012

- Wayne L. Winston & S. Christian Albright Practical Management Science Fourth Edition,2011