Acquisition Analytics Assignment: Bank Marketing

We will understand the business objective and understand the data using univariate and multivariate analysis. Then we'll build supervised learning models, choose the best among them and analyze the financial benefits of the analytics project.

- Data Understanding
- Data Cleaning
- Data Preparation
- Data Modelling
- Model Evaluation

The objective is to reduce the customer acquisition cost by targeting the ones who are likely to buy and to improve the response rate, i.e. the fraction of prospects who respond to the campaign.

Now there are two problems with including the variable ‘duration’ in the model:

- The prospect data procured by the marketing team does not contain ‘duration’, since the call has not been made yet.
- In your analysis of marketing cost and response, you assumed that the cost of a phone call is independent of its duration (₹1 per call), which is not true.

To solve these problems, you should resolve to build another model without including the variable ‘duration’. This will help you understand the relationship of the other variables with the response.

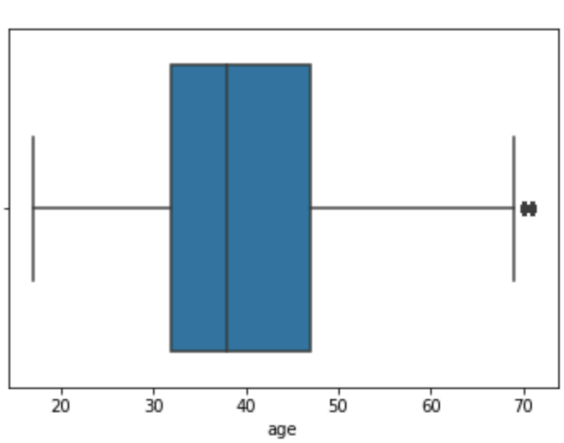
Also, set the business objective to achieving 80% of total responders at the minimum possible cost. The total number of responders is the total number of prospects who responded, from the available data of about 45,000 prospects.

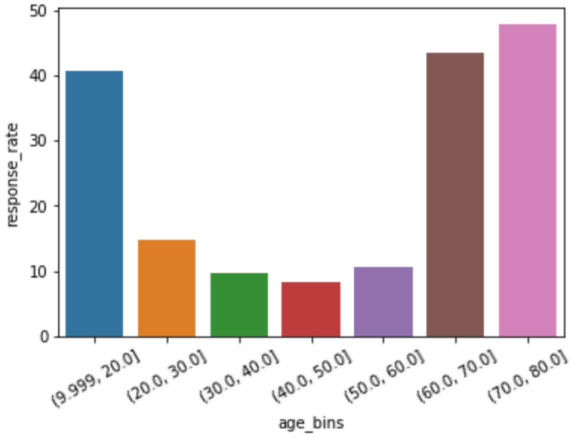
Based on this information, calculate the X in the top X%, i.e., how many prospects should be called to meet the business objective. Further, create a presentation for the CMO highlighting the results and the methodology employed.

Exploratory Data Analysis

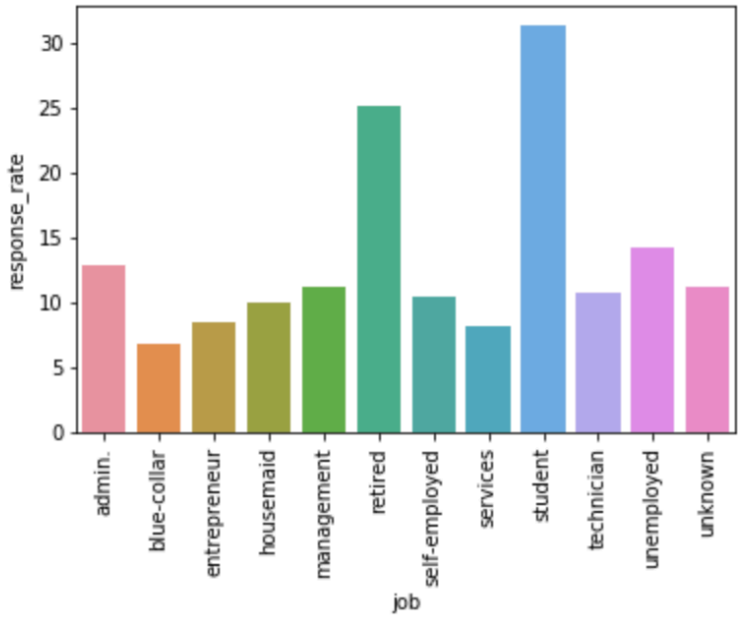
Now, we will do Exploratory Data Analysis on individual features and look for patterns.
First, we will look at the client data.

- Age
- Job : type of job
- Marital : marital status
- Education
- Default: has credit in default?
- Housing: has housing loan?
- Loan: has personal loan?

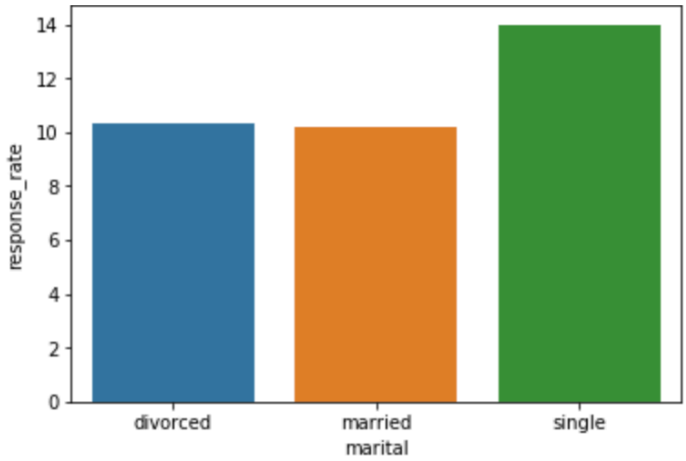




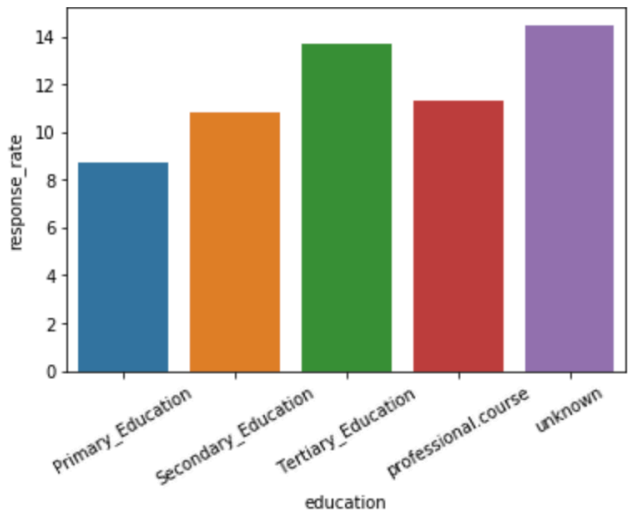
As we can see, the youngest and the eldest are the most likely to buy the investment product being offered by the bank.



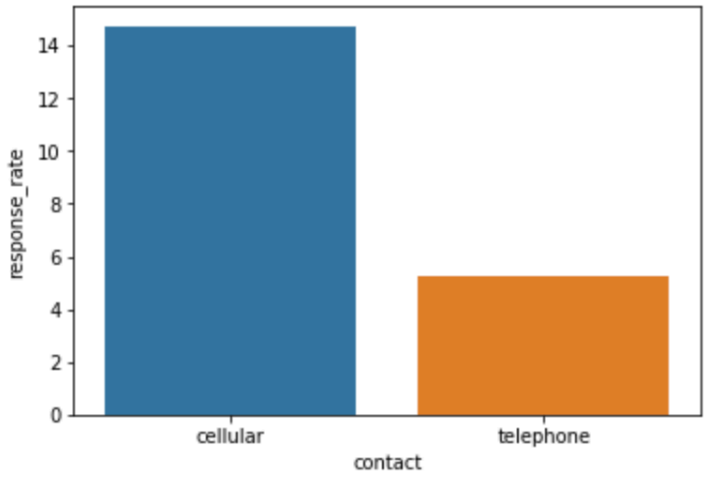
We can draw similarities from the age response analysis where we found that the youngest and eldest were most likely to respond in a positive manner. It is reiterated by the above analysis, where we notice that student and retired have the highest response rates.



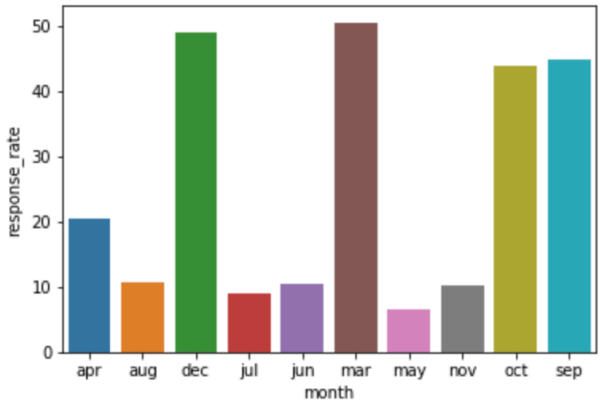
You may note that there is a slight uptick in response rate for single customers, this can be attributed in part to the fact that people under the age of 20 have a very high response rate.



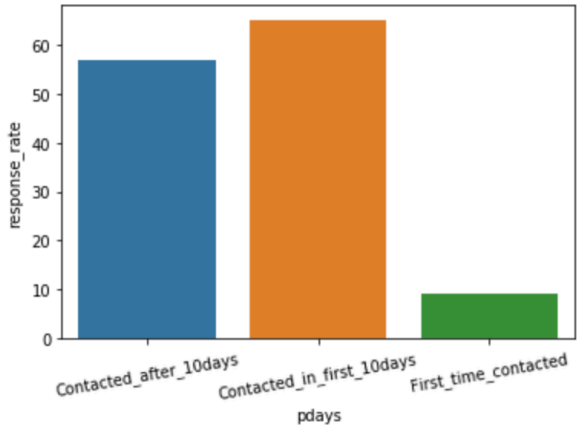
You can note that the response rate is lowest for people with highest education qualification as primary education, which may be attributed, in part, to lack of awareness and a lack of trust between the bank and the customer. It is the highest for people with tertiary education.



You can see that the prospective customers contacted through cellular medium are significantly ore likely to opt in to the investment scheme compared to customers contacted through telephonic medium.



We can note that the four months of december, march, october and september appear to be the best to contact the potential customers. However, please note that these our months have the fewest data entries as well, so it is not certain, how well it would behave when calls are made at a high volume.



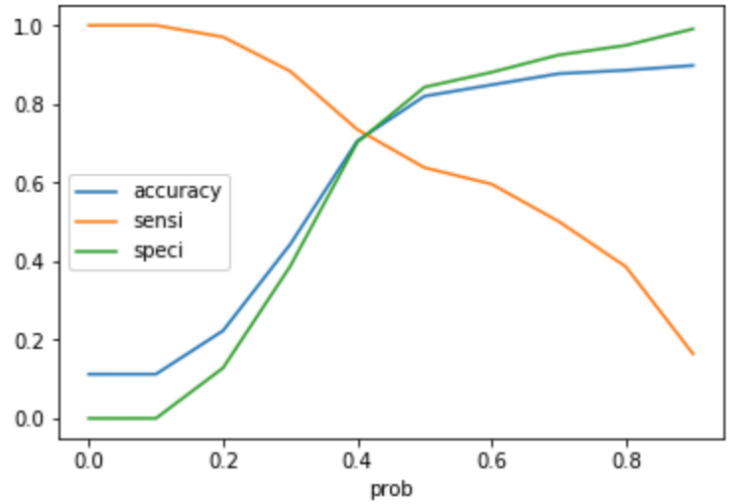
Though there is a huge difference in the response rates of the three categories, it is important to note that there is also a huge difference in the volume of data entries in these categories

Data Modelling

We drop the duration column and proceed with the data modelling.
Also we add a unique id to each prospect
We use RFE to select the most significant 15 features, then perform further fine tuning by checking and eliminating features with high VIF or p value
After the fine tuning we get the below features to be most important for prospect selection

- contact_telephone
- cons.price.idx
- euribor3m
- previous_Never contacted
- month_may
- month_aug
- poutcome_success
- month_mar
- job_student
- month_dec

Then we further create our model using Logistic regression and use the cutoff as 0.4



Sensitivity: 0.74
Specificity: 0.7

Thus a good model that can predict prospects.

Then we create a new dataframe with each prospect along with the duration of call and call cost (considering 1$ per 60 sec)

	actual	pred_prob	predicted	prospect_id	duration	call_cost
0	0	0.193235	0	1	261	4.35
1	0	0.166248	0	2	149	2.48
2	0	0.178931	0	3	226	3.77
3	0	0.188811	0	4	151	2.52
4	0	0.187114	0	5	307	5.12

Next we find the number of top X% prospects you should target to meet the business objective
Report the average call duration for targeting the top X% prospects .
Business objective to achieving 80% of total responders at the minimum possible cost

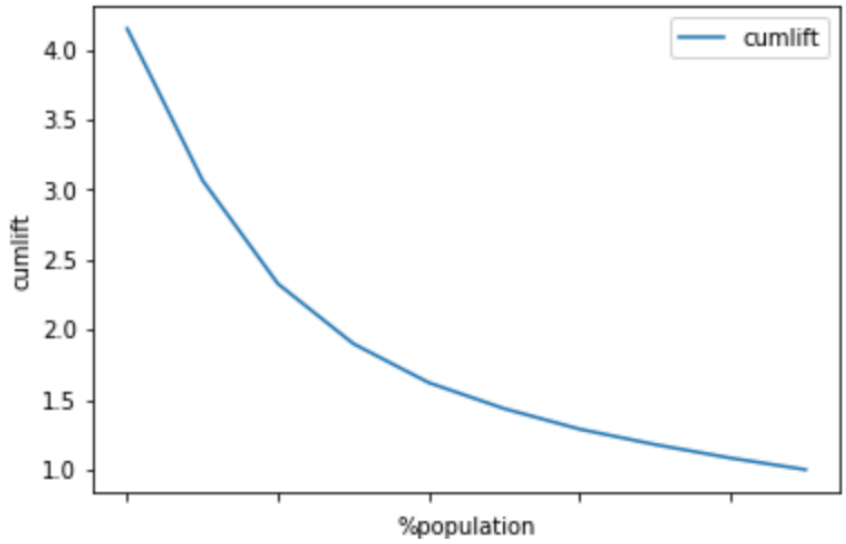
We first create 10 ranking or decile using qcut()

	decile	total	actual	duration	call_cost	cumresp	gain	cumlift
9	1	4118	1925	366.832208	11769.01	1925	41.496012	4.149601
8	2	4117	920	414.689130	6358.55	2845	61.327872	3.066394
7	3	4117	394	661.796954	4345.70	3239	69.821082	2.327369
6	4	4117	286	751.818182	3583.65	3525	75.986204	1.899655
5	5	4117	236	813.161017	3198.31	3761	81.073507	1.621470
4	6	4118	236	780.309322	3069.08	3997	86.160811	1.436014
3	7	4117	191	774.492147	2465.39	4188	90.278077	1.289687
2	8	4117	190	826.868421	2618.35	4378	94.373787	1.179672
1	9	4117	145	876.503448	2118.14	4523	97.499461	1.083327
0	10	4118	116	852.163793	1647.46	4639	100.000000	1.000000

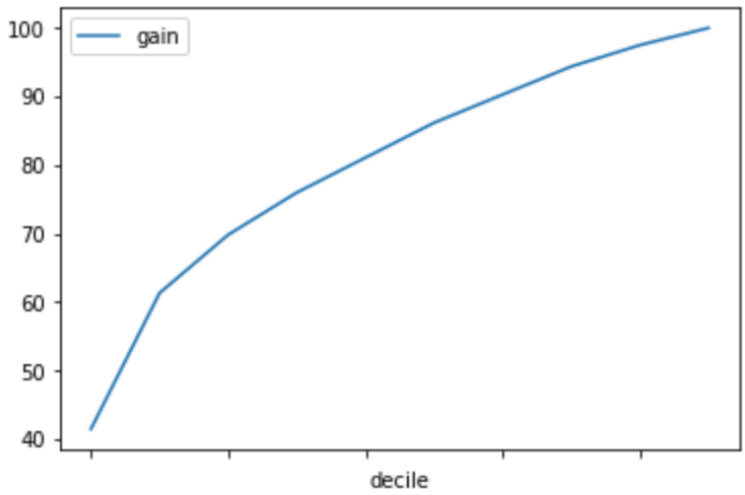
We can attain 80% of total conversions by targeting only 50% of the total client base. This can be used in cost optimization and depending on the cost per call, we can determine, how many people should be part of the tele marketing campaign

Average call duration for targeting the top 50% prospects is 600 seconds (~10 mins)

Next we plot the gain charts and lift charts



LIFT CHART



GAIN CHART

Now we calculate the Total cost of campaign is 174562.17
Cost incurred for acquiring 80% of customer using predictive model is : 29255.22
Percentage of cost of acquisition saved is 83.24
We are able to save 83.24% of fund using the predictive model

Therefore, it is advisable to use a predictive model to identify suitable prospects for the loan an d use a cost effective method during the campaign.