# Acquisition Analytics

By: Pramodini V. Nayak

# Abstract


Objective:

To build a response model for the bank marketing data set.  The business objective is to achieving 80% of total responders for a loan product at the minimum possible cost.

Problem statement:

We have to build the model without ‘duration’ as variable. Because, the prospect data procured by the marketing team will not contain ‘duration’, since the call has not been made yet.

# Analysis methodology

**Data collection and cleaning**

- Import the data
- Identifying the data quality issues and clean the data

**Visualizing the data**

- Visualizing original data variables to look for any pattern or correlation.
- Analyzing the outlier and treating accordingly.

**Data preparation**

- Dropping duration column
- Creating dummy variables.
- Test and train split of the data.
- Standardizing all the continuous variables.

**Modelling and Evaluation**

- Feature selection using RFE for selecting top 20 variables.
- Using GLM statsmodel to select significant variables.
- Hyperparameters tuning our model.
- Finding optimal probability cut-off.
- Prediction on test data.
- Evaluating the model accuracy , sensitivity and specificity.

**Decision Making**

- Based on business objective to achieve 80% of total responders selecting suitable number of decile.
- Calculating average call duration for targeting 80% responders.
- Calculating cost of acquisition
- Calculating the amount saved after the influence of the model.

# Feature selection using RFE and stats-model:

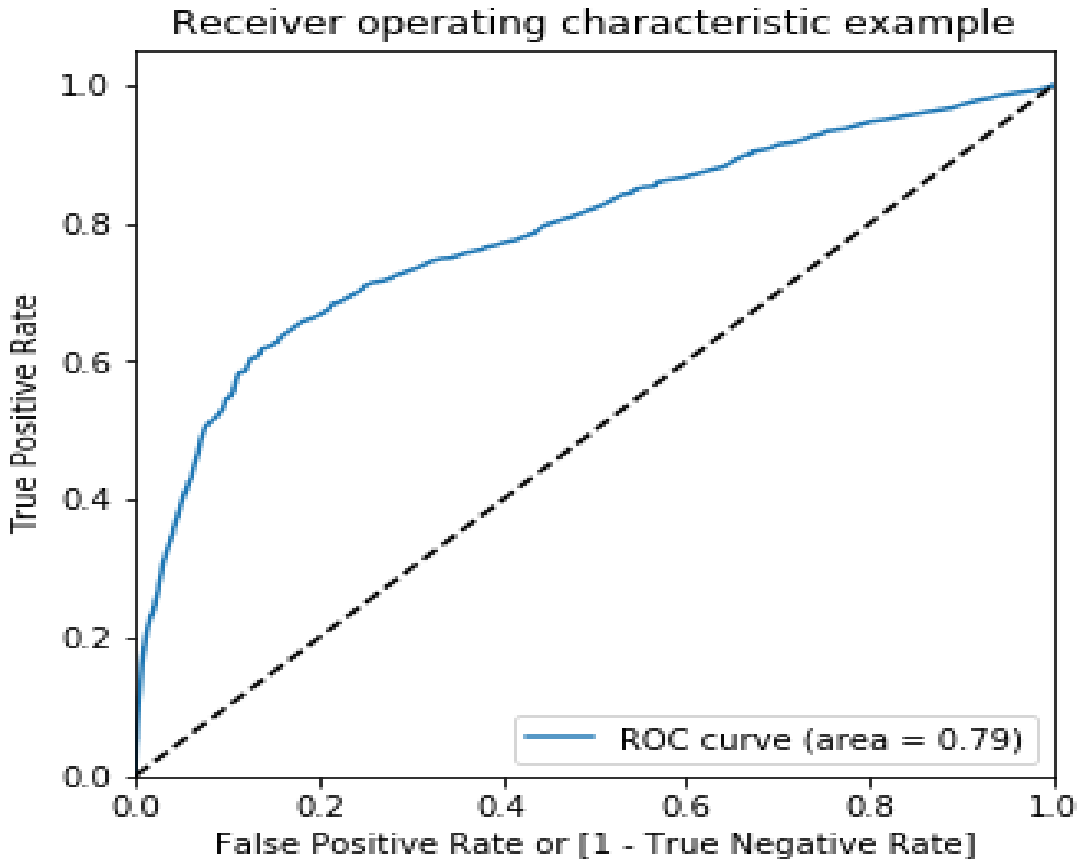
Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | response | **No. Observations:** | 28831 |
| **Model:** | GLM | **Df Residuals:** | 28818 |
| **Model Family:** | Binomial | **Df Model:** | 12 |
| **Link Function:** | logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -8120.1 |
| **Date:** | Mon, 04 Nov 2019 | **Deviance:** | 16240. |
| **Time:** | 16:13:38 | **Pearson chi2:** | 3.08e+04 |
| **No. Iterations:** | 6 | **Covariance Type:** | nonrobust |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -2.5298 | 0.065 | -39.104 | 0.000 | -2.657 | -2.403 |
| **cons.price.idx** | 0.1815 | 0.025 | 7.402 | 0.000 | 0.133 | 0.230 |
| **euribor3m** | -0.9292 | 0.028 | -33.095 | 0.000 | -0.984 | -0.874 |
| **job_retired** | 0.4338 | 0.081 | 5.333 | 0.000 | 0.274 | 0.593 |
| **job_student** | 0.4666 | 0.102 | 4.553 | 0.000 | 0.266 | 0.667 |
| **default_unknown** | -0.3635 | 0.067 | -5.424 | 0.000 | -0.495 | -0.232 |
| **contact_telephone** | -0.1933 | 0.059 | -3.263 | 0.001 | -0.309 | -0.077 |
| **month_dec** | 0.4441 | 0.196 | 2.269 | 0.023 | 0.060 | 0.828 |
| **month_mar** | 0.8193 | 0.113 | 7.243 | 0.000 | 0.598 | 1.041 |
| **month_may** | -0.8794 | 0.052 | -16.809 | 0.000 | -0.982 | -0.777 |
| **month_sep** | 0.2711 | 0.116 | 2.344 | 0.019 | 0.044 | 0.498 |
| **previous_Never contacted** | 0.4076 | 0.063 | 6.484 | 0.000 | 0.284 | 0.531 |
| **poutcome_success** | 1.8859 | 0.090 | 21.017 | 0.000 | 1.710 | 2.062 |

- Our final model has 12 features.
- The variables with positive co-efficient effect the output positively. Variables such as: retired, students etc.
- The variables with negative co-efficient effect the output negatively. Variables such as: contacting via telephone, euribor3m etc.

# ROC curve




The area under the curve of the ROC is 0.79 which is quite good.

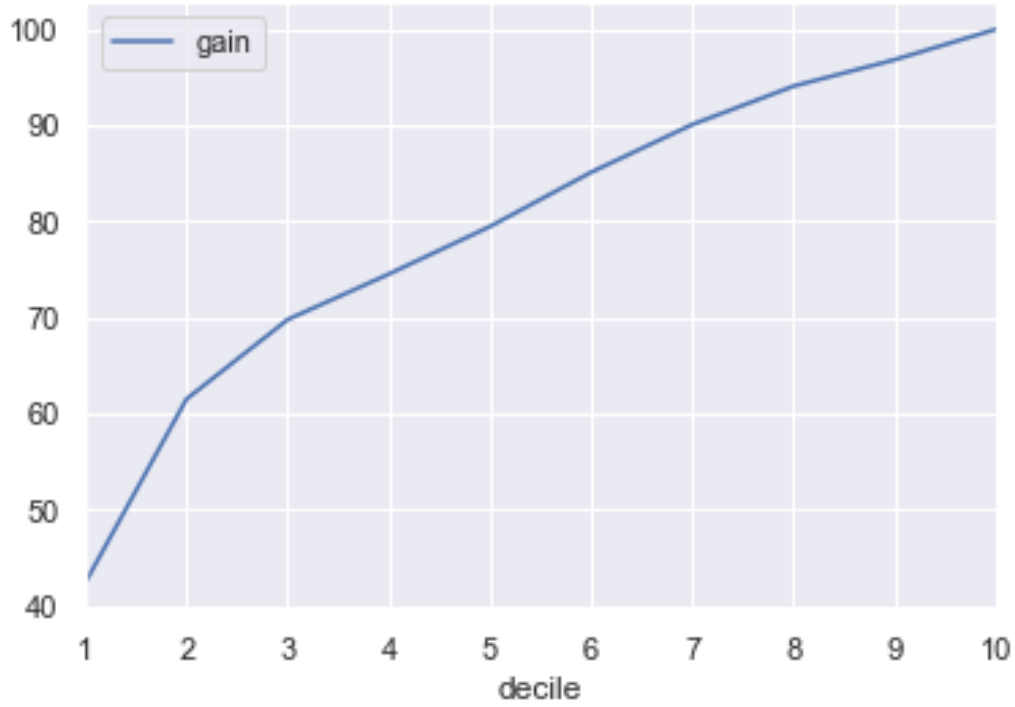
# Optimal Probability Cut off


We see that the optimal probability cut-off is 0.405.

With 0.405 as cut-off, our model evaluation are as follow:
Accuracy is 72.02 %
Sensitivity is 70.26%
Specificity is 72.24%

# X % target analysis

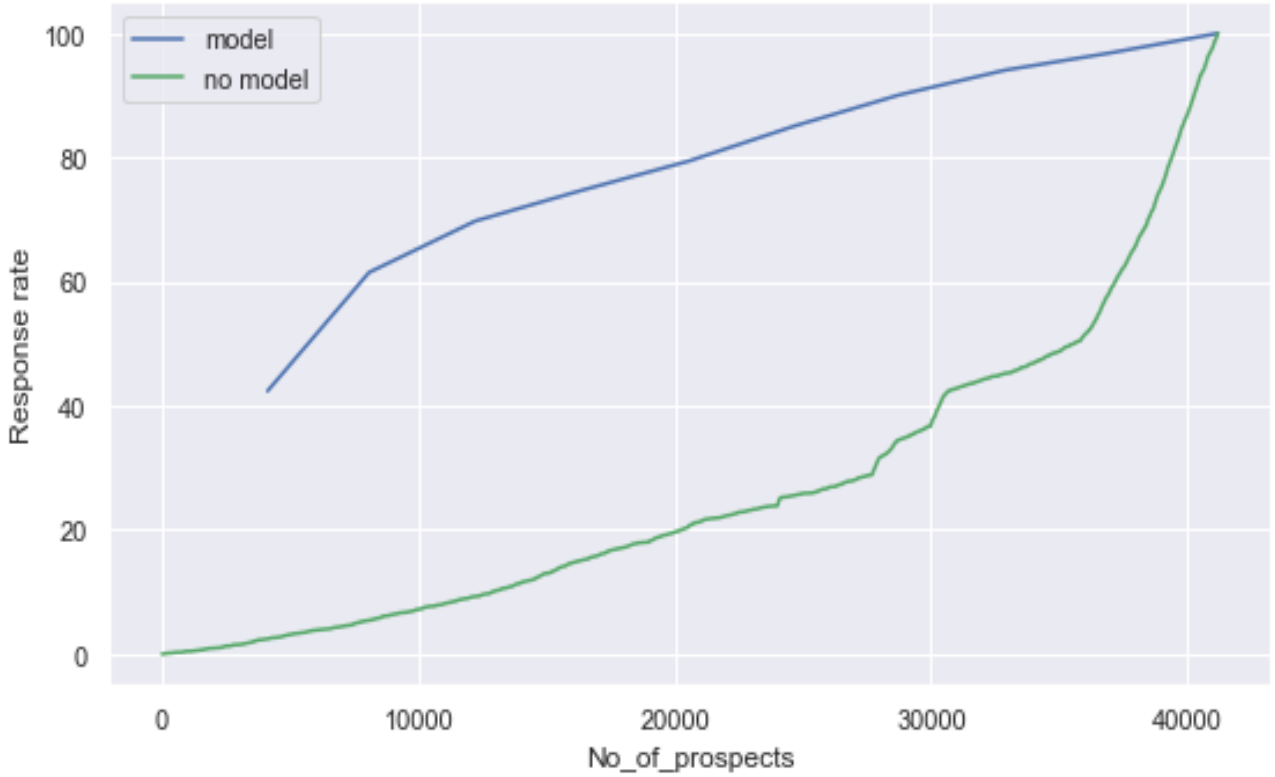
| | decile | total | response | cumresp | gain | cumlift | duration |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 4117 | 1966 | 1966 | 42.370690 | 4.237069 | 1183069 |
| 1 | 2 | 3963 | 887 | 2853 | 61.487069 | 3.074353 | 1074082 |
| 2 | 3 | 4139 | 384 | 3237 | 69.762931 | 2.325431 | 1065006 |
| 3 | 4 | 3976 | 220 | 3457 | 74.504310 | 1.862608 | 959090 |
| 4 | 5 | 4382 | 230 | 3687 | 79.461207 | 1.589224 | 1087228 |
| 5 | 6 | 4118 | 262 | 3949 | 85.107759 | 1.418463 | 1073691 |
| 6 | 7 | 4063 | 231 | 4180 | 90.086207 | 1.286946 | 917071 |
| 7 | 8 | 4126 | 185 | 4365 | 94.073276 | 1.175916 | 972874 |
| 8 | 9 | 4086 | 128 | 4493 | 96.831897 | 1.075910 | 1021963 |
| 9 | 10 | 4218 | 147 | 4640 | 100.000000 | 1.000000 | 1122062 |




We see from the table that from top 5 decile we get 79.46% (Approx. 80%) of response. This can be confirmed again from gain graph.

But if we want exact 80% or above 80%(85.10%) we can reach out to top 6 decile. However, it will cost us additional 4118 calls.

# Lift chart




By looking at the lift chart we clearly see that, model has huge influence on response rate.
By contacting 20000 prospects we had only 20% response rate. But by using logistic regression model we see a response rate of 80% for the same number of contacts.

# Summary

By using logistic regression model for analyzing bank marketing data without the 'duration' variable, we found the following observation:

- The average call duration for targeting the top 80% prospects is 261 seconds, which is about 4.35 minutes.
- Cost of acquisition:

A. If we assume that Cost for each call will be cost per **minute*duration**:

Cost for targeting 80% prospects = Rs. 44, 737.29
Cost for targeting all the customer = Rs. 87,301.13.
So we save Rs. 42,563.84 which is approx. 49% savings.

B. If we assume that Cost = **1*number of contacts** made in the current campaign:

Cost for targeting 80% prospects = Rs. 20, 577
Cost for targeting all the customer = Rs. 41,188.
So we save Rs. 20,611 which is approx. 50% savings.

**Note**: Assumed the cost of call is 50 paisa per minute.