Building Predictive Model for Banking Marketing

**# Problem Description:**

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The business goal is to find a model that can explain success of a contact, i.e.if the client subscribes the deposit. Such model can increase campaign efficiency by identifying the main characteristics that affect success, helping in a better management of the available resources (e.g. human effort,phone calls, time) and selection of a high quality and affordable set of potential buying customers.

Model goal is come up with "Better communication with customers and prospects – Recipient should feel that we understand him or her as an individual – “Send the right message to the right person at the right time”

**Data Description:**

On the input data:

Predictor Variables are considered:

Customer Demographics (Age, Marital)

Call Statistics (Duration, Pdays, Previous, day\_week)

Employment (Job, nremployed)

Credit and Loan History (Default, housing, loan)

social and economic context attributes (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed)

Output variable (desired target):

y - has the client subscribed a term deposit? (binary: 'yes','no')

**Partitioning the Data:**

In any Predictive Model work, the data set has to be partitioned appropriately so as to avoid overfitting/under fitting issues amongst other things.

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | % of Data | Used for | Remarks |
| Training | 80 | Training a Model | 80% of data was used to train a model |
| Validation | 20 | Testing | 20% of data was used for evaluating effectiveness. The model producing the least amount of error usually gets to the Test stage |

**Prediction Accuracy & Model Selection:**

Models built on TRAINING data set are validated using the VALIDATION data set. It is common to build multiple models including ensembles and compare their performance. The model that eventually gets deployed is the one that benefits the business the most, while keeping the error rate within acceptable limits.

Algorithm:  Decision Tree

precision recall f1-score support

0 0.95 0.96 0.95 7319

1 0.65 0.56 0.60 919

avg / total 0.91 0.92 0.91 8238

Prediction Accuracy for DT = 91.69%

### Technology

* Python/Scikit Learn was used to build, validate, and test the models with the 40 thousand transaction data set.

**Results**

* Model predicts the likelihood of Customer Subscriptions with high accuracy.
* Key variables that were impacting Customer Subscribe or causing significant impact on the "Y" were:
  + Duration of the call
  + Month (In which campaign is conducted)
  + nremployed

Relevant Papers:

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014  
  
S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]