2/12/2018

Predicting whether customer will subscribe to term deposit

A Data-Driven Approach using Logistic Regression

SUBMITTED TO

DR. RITA CHAKRAVARTI

INSITITUTE OF SYSTEM SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

PREPARED BY

Team Abraca**DATA**

Bhabesh Senapati	(A0178349M)
Dibyajyoti Panda	(A0178271Y)
Diksha Kumari Jha	(A0178275R)
Gopesh Dwivedi	(A0178338R)
Sravani Satpathy	(A0178226Y)

Table of Contents

1. Executive Summary

The objective for any successful financial marketer is to focus on improving the customer experience across channels. Exploitations of the financial data, domain knowledge and statistical tools to solve the real-world problems have been to the fore. Improving the marketing communications process from the consumer's perspective will drive growth, loyalty and profitability. Managing the marketing process is highly inclusive of various technological tools available, one of them being **Direct Telemarketing**. Retention of the customers can be achieved by probably selling the liability products of the bank to the existing customers. In this regard, banks are planning to campaign more on their term deposits and hence give a locking period for the deposits. This will help build a long-term relationship with the customers.

The increasing number of marketing campaigns over time has reduced their effects on the public. First, due to competition, positive response rate to mass campaigns are typically very low, according to a recent study, less than 1% of the contacts will subscribe a term deposit. Second, direct marketing has drawbacks, such as causing negative attitude towards banks due to intrusion of privacy. To save costs and time, it is important to filter the contacts but keep a certain success rate. The aim of this activity is to **increase campaign efficiency by identifying the main factors that affect the success of a marketing campaign and target clients effectively with the help of statistical tools.**

2. Business Objective

In this project, we aim **to increase campaign efficiency** by identifying the main factors that affect the success of a campaign and predicting whether the campaign will be successful to a certain client, or in other words, **predict whether the client will subscribe a term deposit.**

3. Assumptions

- A customer will either buy or does not buy the term deposit i.e. outcome to be predicted is **dichotomous**
- Only the **meaningful and contributing** factors should be included in the decision making
- All the predicators have **little or no dependence** among other variables
- The data should **not be biased** towards a class. If so, proper measures should be taken before proceeding with modelling

4. Preliminary Exploratory Analysis

4.1 Data Source

- The dataset, which is publicly available for research, is related to direct marketing campaigns (phone calls) of a Portuguese Banking Institution
- The data is taken from: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

4.2 Data Description

- The dataset contains **21** variables and **41,188** observations

Name	Type	Description
age	Continuous	Age of Customer, various
job	nominal	type of Job, 12 Levels
marital	nominal	Marital Status of Customer, 4 Levels
education	nominal	Education Status, 12 Levels
default	nominal	Has credit default?
housing	nominal	Has housing loan?
loan	nominal	Has Personal Loan?
contact	nominal	communication type, cellular or telephone
month	nominal	last contact month of year
day_of_week	nominal	last contact day of week
duration	continuous	last contact duration in seconds
campaign	continuous	number of contacts performed during this campaign
pdays	continuous	number of days that passed by after the client was last contacted
previous	continuous	number of contacts performed before this campaign
poutcome	nominal	outcome of previous marketing campaign
emp.var.rate	continuous	employment variation rate
cons.price.idx	continuous	consumer price index
cons.conf.idx	continuous	consumer confidence index
euribor.3m	continuous	Euribor 3 month rate
nr.employed	continuous	number of employees

Table 1. Data Description

4.3 Cleaning & Treatment of Missing Values

- In direct marketing campaigns, the data is collected over telephone. This creates a lot of missing information since responders sometimes give incomplete response or tend to hide sensitive information
- The existence of missing data may blur the real pattern hidden in the data thus making it more difficult to extract information. Therefore, we chose different methods to deal with those missing data for different attributes
- We first dropped rows where **Job, Marital, Housing and Loan** where unknown because percentage of missing records is quite low
- **Default** has quite large no. of unknowns (~8K), therefore we cannot delete the records. However, since the default corresponds to “Yes” only for **3** records, we cannot apply any

imputation methods as well. Therefore, we treated "unknown" as a valid factor in the category

- For **Education**, we could have imputed missing values based on the proportion of various segments. However, for simplicity we considered "unknown" as valid category
- After these steps, the total observations were reduced to ***39,803***

Factors	# of "Unknowns"	Treatment
Job	330	Drop observations where Job is unknown
Education	1731	Treat "unknown" as separate valid factor
Marital	80	Drop observations where Marital is unknown
Default	8597	Special case, Treat "unknown" as valid factor
Housing	990	Drop observations where Housing is unknown
Loan	990	Drop observations where Loan is unknown

Table. 2: Treatment of Missing Values

- **Dropping "Duration"** from our analysis: This attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this predictor should only be included for benchmark purposes and therefore discarded from realistic predictive analytics

4.4 Biased Data

- The data is biased as the number of "yes" i.e. customer subscribed to the term deposit is **only ~11%** of the total records. This problem can be resolved by following methods
  - **Oversampling**: This amount to sampling the minority class with replacement until we have the same number of observations as the majority class.
  - **Under sampling**: We can take a subsample of the majority class so that it equals to minority class
  - **Reweight the classes**: For logistic regression, we can reduce the loss function to penalize a misclassified minority case much more heavily than a misclassified majority class
  - **Simulate data:** SMOTE to generate new data to increase the minority class, in this case subscribing to term deposit
  - **NOTE:** We implemented the model by Under sampling i.e. taking 2000 "yes" and randomly selecting 2000 "no". However, since the predictive gain was not sufficient we decided to continue with the original biased dataset

4.5 Inferences from Visual Analysis

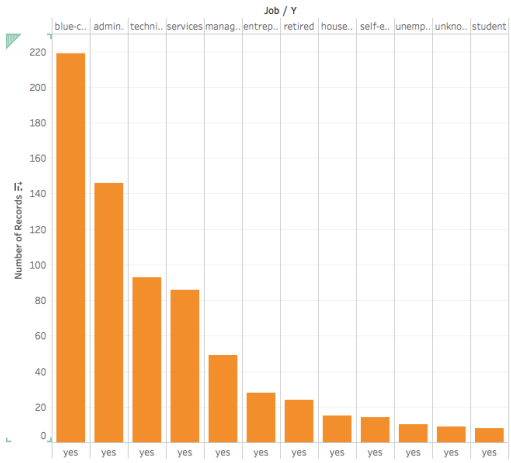


Fig 1. Job vs Subscriptions

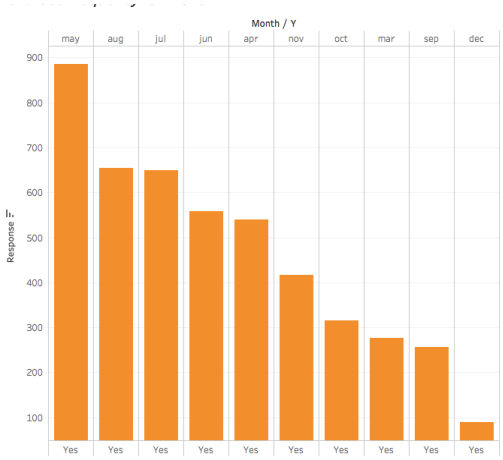


Fig 2. Month vs Subscriptions

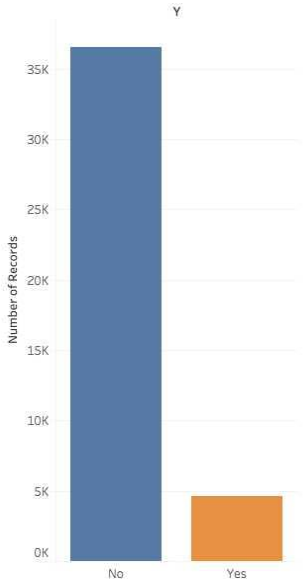


Fig 3. Subscribers Distribution

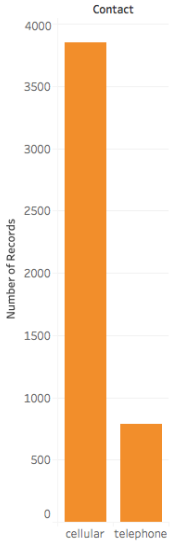


Fig 4. Contact vs Subscriptions

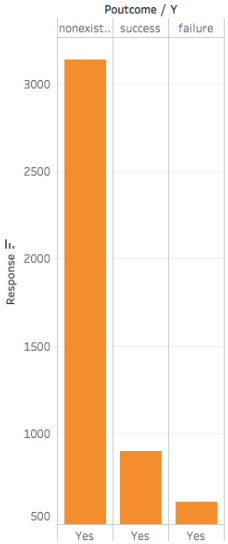


Fig 5 Poutcome vs Subscriptions

- People in different job segments tend to subscribe differently i.e. Blue-collared Customers subscribes the most.
- Subscriptions also varies according to month when the customer was contacted
- Cellular customers tend to subscribe more as compared to Telephone Customers
- Campaign success also depends on previous campaign performance for the customer which indicates the purchasing behavior

5. Modelling using Logistic Regression

» We used R for data cleaning and preparation. After the cleaning, the dataset was split into training and testing set using **70:30** as the split ratio. 70% of the total data was randomly selected for training the model and 30% was kept for testing model performance on unseen data. The training dataset contains **27,862** records whereas testing consist of **11,941**

Why 70-30 as the split ratio?

- If we take 50:50 as the split ratio, there would be less training data which might create under-fitting and model performance might degrade.
- We did not take 90:10 as the split ratio since the test or validation data will have very few positive cases i.e. yes. This might create misleading results
- Hence, the best solution would be found when data is split into 75-25 or 70-30. We chose 70-30 as it is widely used for regression modelling

» From visual inspection, we noticed few factors have strong relationship with the dependent variable (propensity to subscribe) like month, job, contact etc.

» We start by building the base model which uses all variables as predictors for the dependent variable in JMP

Source	LogWorth	PValue
month	55.501	0.00000
emp.var.rate	23.850	0.00000
contact	22.985	0.00000
cons.price.idx	15.889	0.00000
day_of_week	6.948	0.00000
poutcome	6.220	0.00000
pdays	5.754	0.00000
cons.conf.idx	4.407	0.00004
campaign	4.227	0.00006
default	3.689	0.00020
nr.employed	1.442	0.03615
job	1.128	0.07441
euribor3m	1.003	0.09921
previous	0.824	0.15012
education	0.337	0.46044
age	0.222	0.60011
marital	0.165	0.68398
loan	0.054	0.88385
housing	0.018	0.95958

Fit Details

Measure	Training
Entropy RSquare	0.2196
Generalized RSquare	0.2835
Mean -Log p	0.2749
RMSE	0.2787
Mean Abs Dev	0.1554
Misclassification Rate	0.0986
N	27862

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
age	1	1	0.27464857	0.6002
job	10	10	16.8386893	0.0780
marital	2	2	0.76561223	0.6819
education	5	5	4.64648597	0.4605
default	2	2	17.8600444	0.0001*
housing	1	1	0.00256872	0.9596
loan	1	1	0.02136661	0.8838
contact	1	1	112.619453	<.0001*
month	9	9	275.057098	<.0001*
day_of_week	4	4	38.6469064	<.0001*
campaign	1	1	18.2125356	<.0001*
pdays	1	1	22.3916139	<.0001*
previous	1	1	2.05331241	0.1519
poutcome	2	2	29.4476572	<.0001*
emp.var.rate	1	1	104.122186	<.0001*
cons.price.idx	1	1	68.4635307	<.0001*
cons.conf.idx	1	1	16.9615809	<.0001*
euribor3m	1	1	2.72035936	0.0991
nr.employed	1	1	4.38149029	0.0363*

	Predicted Values	
Actual Values	Not subscribe	Will Subscribe
Not Subscribe	24,363	358
Subscribe	2,389	752

Table 3: Confusion Matrix for Base Model

Parameters	Base Model
Accuracy	*90.14 %*
AUC	*0.795*
AIC	*15409.9*
Sensitivity	*23.94 %*
Specificity	*98.55 %*

- We observe that though the accuracy of our model is **quite good i.e. 90%** but the type II error (**Type II = 1-sensitivity**) is also **high ~76%.** It means the model is misclassifying subscribing customers as not subscribers. This may defeat our purpose to maximize the term depositors
- From confusion matrix, **2,389 instances** were found where the model predicted the subscribing customers as not a subscriber. We should try to reduce this and optimize our model by minimizing predictors while maintaining accuracy. Overall misclassification rate is around 9%
- From Variable Effect summary, we can notice that month, emp.var.rate, contact, cons.conf.idx, day_of_week, pdays, poutcome, campaign have low p-value and are significant
- Also, **Housing and Loan** are least significant with very high p-value. Next, we try to improve the model by feature selection and varying the cut-off threshold for probability

6. Detailed Analysis and Model Improvement

6.1 Feature Selection

- Feature selection can be done in various ways, namely, forward, backward and bi-directional. Here, we used the bi-directional for our modelling.
- From the base model, since **Housing** has the least signification and highest p-value, we removed it from our analysis
- Next, we removed **Loan, marital, education, previous, euribor3m** due to low significance
- We checked the accuracy, AUC and Type 1 and Type 2 error at every step and reverted if the model performance degraded
- Here is the summary and reasoning behind our predictors selection

Variable	Name	Status	Reason
1	age	removed	Most of the subscribers comes from 30-40 age group, therefore no strong relation with dependent variable for rest. Also, p-value was quite high i.e. insignificant
2	job	included	For different job segments, the subscribers varied. Hence, making it significant for business purpose. Also, p-value in the final model was quite low
3	marital	removed	very high p-value, therefore low significance
4	education	removed	very high p-value, therefore low significance
5	default	removed	though it was significant on “unknown” but Prob>ChiSq is very high (~0.99) and variable is unstable. Also, after removing Type I and Type II error decreased.
6	housing	removed	very high p-value, therefore low significance

7	loan	removed	very high p-value, therefore low significance
8	contact	included	very low p-value, highly significant
9	month	included	Subscribers varied according to months, which makes sense as People tend to buy term deposits when financial cycle is near
10	day_of_week	removed	removing this did not degrade the model significantly as from the plot, no. of subscribers remains almost constant throughout the week
11	duration	removed	Due to business logic, if duration=0, y is also 0, highly influential
12	campaign	included	Both Campaign and pdays are logically important as greater the memory of the last call, higher the chances of customer subscribing the term deposit. Hence, they are significant to the business problem
13	pdays	included	
14	previous	removed	high p-value, hence low significance. From business point of view, customers contacted long before the campaign tend to forget information
15	poutcome	included	Low p-value, highly significant. Also, customers behavior can be predicted from past purchase
16	emp.var.rate	removed	Initially significant but does not affect model performance. Also, Prob>ChiSq is high making it redundant
17	cons.price.idx	removed	
18	cons.conf.idx	included	low p-value i.e. high significance. If customers are more optimistic, there are higher chance of subscriptions
19	euribor.3m	removed	Not significant due to high p-value
20	nr.employed	included	Initially not significant but turns out to be most significant after removal of other redundant predictors

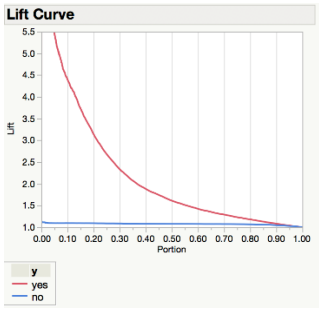
6.2 Predicators for Final Model

» After many iterations and combinations, we come up with the final list of predictors as below:

Source	LogWorth	PValue
nr.employed	185.548	0.00000
month	60.945	0.00000
poutcome	21.600	0.00000
contact	18.086	0.00000
cons.conf.idx	6.441	0.00000
campaign	5.950	0.00000
pdays	5.605	0.00000
job	3.537	0.00029

7. Performance Metrics

- The Final Model was implemented on training and testing set separately and key performance indices were noted
- We optimized the model using **8 Predictors** which are both significant technically and to our business problem



Effect Likelihood Ratio Tests

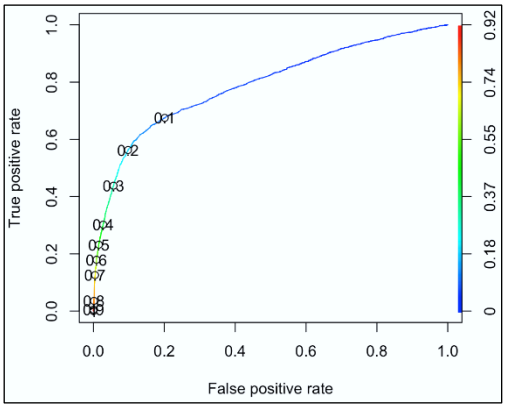
Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
month	9	9	311.131082	<.0001*
contact	1	1	78.4525902	<.0001*
poutcome	2	2	99.4733777	<.0001*
nr.employed	1	1	847.282771	<.0001*
pdays	1	1	22.1777789	<.0001*
campaign	1	1	23.7047129	<.0001*
cons.conf.idx	1	1	25.8872621	<.0001*
job	10	10	32.8325332	0.0003*

Fit Details

Measure	Training
Entropy RSquare	0.2110
Generalized RSquare	0.2732
Mean -Log p	0.2779
RMSE	0.2804
Mean Abs Dev	0.1572
Misclassification Rate	0.1000
N	27862

7.1 Receiver Operating Characteristics (ROC)

- Using ROC, we optimized the **threshold cut-off at 0.25**.
- This means if the probability of the customer being a subscriber **is greater than 0.25**, we would predict the customer will subscribe the term deposit
- The thresholds are skewed since the data is biased towards "no"
- The **AUC is 0.79** which is comparable to base model



7.2 Confusion Matrix

- We see the model is performing well on both training and testing data set. With 8 predictors, we have **88% accuracy** on training as well as testing data set.

Actual Values	Predicted Values (Train)		Predicted Values (Test)	
	Not subscribe	Subscribe	Not subscribe	Subscribe
Not Subscribe	23,083	1,698	9,838	757
Subscribe	1,620	1,521	687	658

Table 3: Confusion Matrix with Final Model for Training and Testing

Parameters	Training Set	Testing Set
Accuracy	88.12 %	87.90 %
AUC	0.787	0.784
AIC	15522	-
Sensitivity	48.42%	48.95%
Specificity	93.14%	92.85%

8. Conclusion and further scope

- A model with the accuracy of 88% shows that the mentioned problem focuses on the practical aspect of using the results efficiently placing lesser emphasis on model accuracy & improving Type II Error

- Table 3 above sets out the AUC and Accuracy values of the predicted model above, which have been trained on the training data and tested on the test set for both out base and final models. These AUC values measure the fit of these trained models in respect of the validation data, i.e. how well the trained model predicts if a customer in the validation data subscribes to the term deposit product

- A highly sensitivity in testing dataset means that there are few false negative results, and thus fewer cases of customers who are supposed to purchase the term deposit are missed. And specificity of a model is its ability to designate a customer who wouldn't purchase the term deposit as negative i.e. fewer False positive values. The current model is designed keeping the biased data in place. The sensitivity has **increased from 23% to 48%(doubled)** thereby balancing the specificity to 93%.

- The scope of the study above is constrained by the volume and availability of data – for example, the data does not provide an indication of the price and profitability of the product being marketed or the cost of making each call – it clearly demonstrates that predictive analytics can contribute **towards increasing efficiency; reducing the marketing cost** of the product and **improving profitability as a result**

- Thus, the **cost effectiveness of the model has resulted in the balancing of the type I error keeping in sight of the type II error** which would contribute to the fact of not missing out of the potential subscribers. More generally, it can be used widely and effectively to solve many real-world commercial problems

- The sensitivity can further be improved if we balance the biasness of the dataset using available techniques like SMOTE. However, we are not considering that into scope of the problem as of now.

9. Appendix

- Scatter plot matrix