Predicting Term Deposit Subscription

Using Logistic Regression in SAS, R and Python

Introduction

A number of campaigns are being run by banks now-a-days aiming to get their customers to subscribe for different products like term deposits, credit cards, insurance and many more. In order to convince the customers to subscribe for the banking products multiple phone calls are made and a large amount of data is gathered by banks. Analysing such data can assist in identifying patterns which helps banking institutions to understand their customers to improve their business.

This project is oriented to develop a predictive model that analyses the customer behaviour for subscribing to one such banking products "term deposits". Our group was provided with the Portuguese banking institution marketing campaign dataset that aims to access whether term deposit would be subscribed or not based on several predictors. The data consists of 20 predictors (10 categorical and 10 numerical) and one response variable (categorical; whether the client subscribed to term deposit or not). The predictive model is trained and tested using the machine learning technique of logistic regression that classifies the response. In order to achieve the objective, logistic regression models were built in three different softwares; SAS, R and Python.

The analysis shows that the model on original data without any data pre-processing step does better.

Data Wrangling and Descriptive Analysis:

Imputation

No missing values were found in the dataset. However, there were few categorical variables containing "unknown" as a category.

```
> sapply(bank, function(x) length(which(x == "unknown")))
          age           job       marital     education       default
            0             0            80          1731          8597
      housing          loan       contact         month   day_of_week
          990           990             0             0             0
     duration      campaign         pdays      previous      poutcome
            0             0             0             0             0
 emp.var.rate cons.price.idx cons.conf.idx    euribor3m   nr.employed
            0             0             0             0             0
            y
            0
```

Figure 1: Number of "Unknown" values in all the variables

Imputations were performed in order to deal with "unknown" values in each of the variables. Further, some of the categories with rare occurrences were removed.

Variable	**Methods applied**
Job	"Unknown" imputed by the largest category "admin."
Marital	"Unknown" imputed by the largest category "married"
Education	• "Unknown" imputed by the largest category "university.degree" • 18 cases of category "Illiterate" removed
Default	3 cases of category "yes" removed
Housing	"Unknown" imputed by the largest category "yes"
Loan	"Unknown" imputed by the largest category "no"
pdays	Replaced "999" with "-1" to avoid skewness in the data

Table 1: Data Imputation and Cleaning

The encoding was performed for categorical variables for simplicity in model fitting.

Scaling

The distributions for a few feature variables are shown in Figure 2.

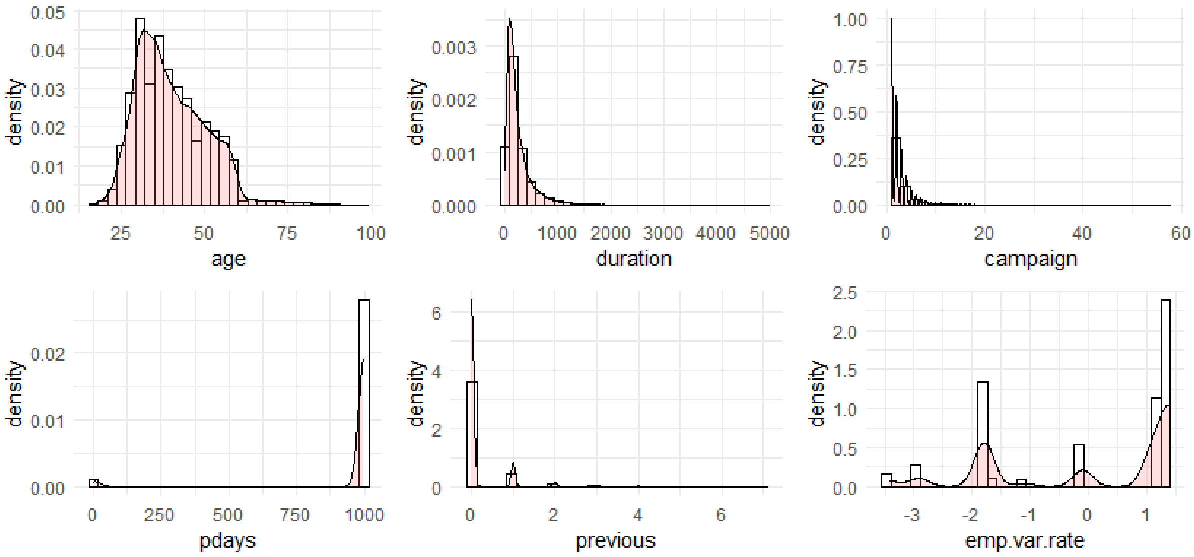

Figure 2: Distributions of numeric predictors

It can be seen that most of the distributions are highly skewed and the range of the features varies a lot. Thus, the numeric attributes were standardized to be centred at 0 with standard deviation 1.

Correlation

Figure 3 shows the correlation between the numerical variables.

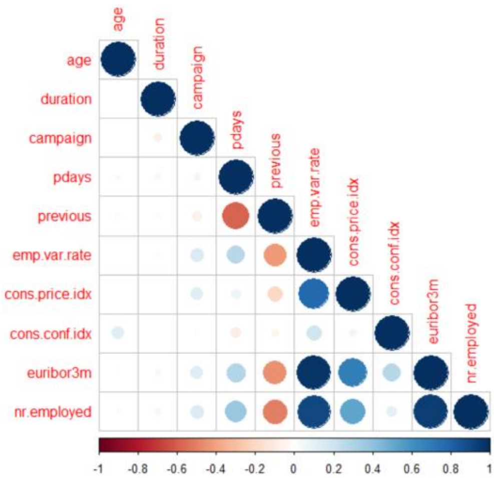

Figure 3: Correlation

It can be observed from the graph that the variables euribor3m, emp.var.rate and nr.employed are highly correlated.

Outcome Imbalanced

The predicted outcome (y) is highly imbalanced with 11.26% of "yes" responses and 88.74% of "no" responses.

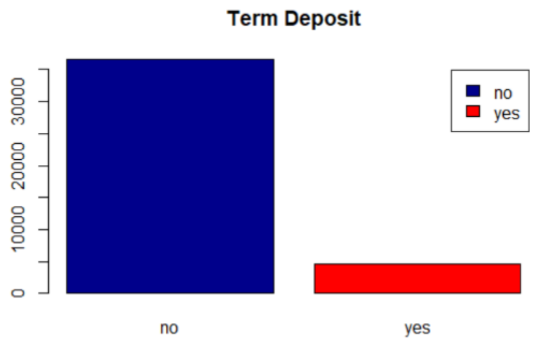

Figure 4: Proportion of Responses: Yes or No

In order to overcome the imbalanced proportion of responses, different sampling techniques (undersampling and oversampling) were performed.

Undersampling: A proportion of random samples were extracted from majority class ("no" responses) which were equal to the number of "yes" responses.

Oversampling: Random duplicates were generated from minority class ("yes" responses) which were equal to the number of "no" responses.

Model Fitting, Interpretation and Results

The dataset was split into training and test data with 80:20 ratio. Four different models were fitted and the scores obtained are listed in the below table.

Model	Accuracy	AIC	AUC
Model 1: Original Model (No imputation/scaling/sampling techniques)	91.0%	14163.5	0.9265
Model 2: Model with Imputation and Scaling	91.1%	14439.4	0.9318
Model 3: Model with Imputation, Scaling and Undersampling	85.1%	5025.0	0.9325
Model 4: Model with Imputation, Scaling and Oversampling	85.3%	39940.2	0.9337

Table 2: Model performance comparison

The ROC curves and AUC scores for all the models are shown figure 5.

Figure 5: ROC Curve

Since the data is imbalanced and the aim is to successfully identify people who subscribed to term deposit, the performance measures ROC (Receiver Operating Character) curve and AUC (Area Under Curve) were used to select the final model. It can be observed that the difference in AUC scores is not significant and ROC curves seems to overlap for all the models. Therefore, the final model is the original model since the model is performing equally well without applying any imputation, scaling or sampling technique.

Interpretation of coefficient estimates:

For the positive coefficient estimates, the predictor has a positive impact on subscription of term deposit, which implies that these predictors increase the chance of subscribing to term deposit. However, if the coefficient estimate is negative, then the chances of subscribing to term deposit are reduced.

For instance, it is estimated that the odds of subscribing to a term deposit drop by a factor of 0.2($e^{-1.597}$) for 1 unit increase in employment variation rate keeping all other predictors fixed.

Similarly, the estimated odds of subscribing to a term deposit for customers with university degree is approximately 1.2($e^{0.1924}$) times than customers with 4-year basic education.

Marginal Effects:

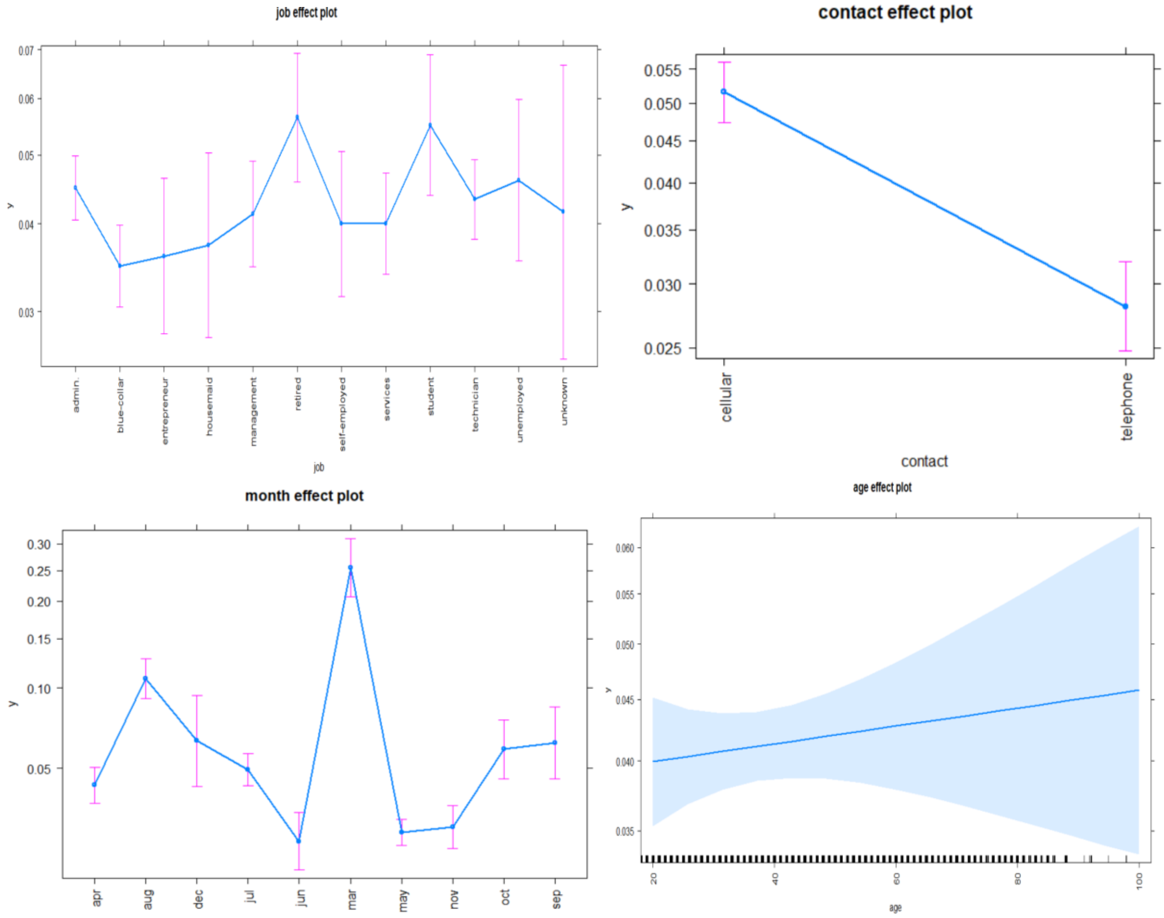

Figure 6: Marginal Effects for job, contact, month and age (left to right)

The average effect of changes in variables job, contact, month and age on the change in probability of y is shown in the marginal effects plots.

The *job effect plot* shows that the difference in the predicted probabilities for retired is approximately 0.055 times more than the admin (reference level), that is the retired persons are more likely to subscribe to term deposit as compared to admin.
The *contact effect plot* shows that the difference in the predicted probabilities for telephone is approximately 0.029 times less than the cellular (reference level), that is the customers who have telephone are less likely to subscribe to term deposit as compared to the customers who have cellular phones.

The *month effect plot* depicts that the customers contacted during the month of March were more likely to subscribe to the term deposit.
The *age effect plot* represents the average change in probability of subscribing to term deposits (y = “yes”) when age increases by 1 year. The effect of age on response; subscribing to a term deposit throughout is positive.

Comparison between the software

All the four models were built in SAS, R and Python. The AUC score, AIC and estimates were found to be approximately equal in all the three software. Interestingly, SAS returns a complete summary of the model, while certain coding needs to be done in R and Python to achieve those summaries.

The formula for logistic regression is quite similar in every software. In R, the base line functions glm(), summary() and predict() were used to fit models, evaluate performances and make predictions. Its output underlines the AIC and the deviance distribution. Whereas in python, sklearn.linear_model was imported to access LogisticRegression function in which the output also comprises of pseudo R-adjusted value and log likelihood value unlike in R. Further, it can be observed that the estimates are more precise in SAS and Python, however, they are in exponential terms in R.

References:
[1] https://core.ac.uk/download/pdf/55616194.pdf
[2] https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1143&context=math_theses
[3] https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/
[4] https://www.scribbr.com/statistics/akaike-information-criterion/#:~:text=The%20Akaike%20information%20criterion%20(AIC,best%20fit%20for%20the%20data.
[5]https://www3.nd.edu/~rwilliam/stats3/Margins02.pdf#:~:text=For%20categorical%20variables%20with%20more%20than%20two%20possible,Jews%20were%20to%20succeed%20than%20were%20Catholics%2C%20etc.

Final Model Outputs:

SAS Output:

The LOGISTIC Procedure

Model Information	
Data Set	WORK.BANK_TRAIN
Response Variable	y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	32951
Number of Observations Used	32951

Response Profile

Ordered Value	y	Total Frequency
1	ye	3657
2	no	29294

Probability modeled is y='ye'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	22973.168	13519.609
SC	22981.570	13964.956
-2 Log L	22971.168	13413.609

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9557.5586	52	<.0001
Score	11921.0866	52	<.0001
Wald	5261.0290	52	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-241.9	197.9	1.4936	0.2217
age		1	-0.00092	0.00274	0.1133	0.7364
job	admin.	1	-0.0331	0.2579	0.0165	0.8979
job	blue-collar	1	-0.1966	0.2615	0.5653	0.4521
job	entrepreneu	1	-0.1507	0.2871	0.2754	0.5998
job	housemaid	1	0.1011	0.2955	0.1171	0.7322
job	management	1	-0.0899	0.2675	0.1130	0.7368
job	retired	1	0.3317	0.2710	1.4974	0.2211
job	self-employ	1	-0.1865	0.2842	0.4306	0.5117
job	services	1	-0.1579	0.2678	0.3475	0.5555
job	student	1	0.1211	0.2781	0.1895	0.6633
job	technician	1	-0.0944	0.2620	0.1300	0.7185
job	unemployed	1	0.0576	0.2866	0.0404	0.8406
marital	divorced	1	-0.1307	0.4774	0.0749	0.7843
marital	married	1	-0.1394	0.4729	0.0869	0.7681

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	0.999	0.994	1.004
job admin. vs unknown	0.967	0.584	1.604
job blue-collar vs unknown	0.822	0.492	1.371
job entrepreneu vs unknown	0.860	0.490	1.510
job housemaid vs unknown	1.106	0.620	1.975
job management vs unknown	0.914	0.541	1.544

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.7	Somers' D	0.873
Percent Discordant	6.3	Gamma	0.873
Percent Tied	0.0	Tau-a	0.172
Pairs	107128158	c	0.937

R output:

```
Call:
glm(formula = y ~ ., family = "binomial", data = dftrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7076  -0.3046  -0.1878  -0.1351   3.3134

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.089e+02  4.206e+01  -4.968 6.77e-07 ***
age                          1.248e-03  2.686e-03   0.465 0.642267
jobblue-collar              -2.184e-01  8.871e-02  -2.462 0.013828 *
jobentrepreneur             -1.040e-01  1.368e-01  -0.760 0.447200
jobhousemaid                -5.869e-02  1.662e-01  -0.353 0.723938
jobmanagement                1.702e-02  9.395e-02   0.181 0.856231
jobretired                   3.205e-01  1.182e-01   2.710 0.006727 **
jobself-employed            -1.549e-01  1.320e-01  -1.173 0.240738
jobservices                 -1.285e-01  9.547e-02  -1.346 0.178257
jobstudent                   2.855e-01  1.235e-01   2.312 0.020779 *
jobtechnician                2.926e-02  7.947e-02   0.368 0.712717
jobunemployed                9.696e-02  1.397e-01   0.694 0.487505
jobunknown                  -1.121e-02  2.558e-01  -0.044 0.965055
maritalmarried              -7.462e-02  7.473e-02  -0.999 0.318000
maritalsingle                4.904e-03  8.550e-02   0.057 0.954262
maritalunknown               1.184e-01  4.504e-01   0.263 0.792553
educationbasic.6y            8.135e-02  1.350e-01   0.603 0.546657
educationbasic.9y            1.200e-03  1.062e-01   0.011 0.990986
educationhigh.school         4.865e-02  1.022e-01   0.476 0.633968
educationilliterate          1.006e+00  8.648e-01   1.163 0.244856
educationprofessional.course 5.650e-02  1.128e-01   0.501 0.616306
educationuniversity.degree   1.924e-01  1.023e-01   1.881 0.060013 .
educationunknown             1.723e-01  1.319e-01   1.307 0.191375
defaultunknown              -3.545e-01  7.486e-02  -4.735 2.19e-06 ***
defaultyes                  -7.250e+00  1.134e+02  -0.064 0.949043
housingunknown               1.378e-02  1.534e-01   0.090 0.928433
housingyes                   1.789e-02  4.587e-02   0.390 0.696430
loanunknown                         NA         NA      NA       NA
loanyes                     -2.432e-02  6.327e-02  -0.384 0.700634
contacttelephone            -5.581e-01  8.425e-02  -6.624 3.49e-11 ***
monthapr                    -1.948e+00  1.598e-01 -12.189  < 2e-16 ***
monthmay                    -2.465e+00  1.353e-01 -18.218  < 2e-16 ***
monthjun                    -2.383e+00  2.308e-01 -10.327  < 2e-16 ***
monthjul                    -1.751e+00  1.683e-01 -10.404  < 2e-16 ***
monthaug                    -1.213e+00  1.423e-01  -8.526  < 2e-16 ***
monthsep                    -1.669e+00  1.727e-01  -9.667  < 2e-16 ***
monthoct                    -1.876e+00  1.682e-01 -11.155  < 2e-16 ***
monthnov                    -2.406e+00  1.604e-01 -14.999  < 2e-16 ***
monthdec                    -1.744e+00  2.382e-01  -7.321 2.46e-13 ***
day_of_weektue               1.926e-01  7.212e-02   2.670 0.007575 **
day_of_weekwed               2.976e-01  7.186e-02   4.142 3.44e-05 ***
day_of_weekthu               1.736e-01  7.020e-02   2.473 0.013403 *
day_of_weekfri               1.275e-01  7.311e-02   1.744 0.081224 .
duration                     4.689e-03  8.292e-05  56.544  < 2e-16 ***
campaign                    -4.553e-02  1.297e-02  -3.511 0.000447 ***
pdays                       -8.781e-04  2.452e-04  -3.581 0.000342 ***
previous                     1.702e-02  6.610e-02   0.257 0.796819
poutcomenonexistent          5.373e-01  1.049e-01   5.122 3.02e-07 ***
poutcomesuccess              1.022e+00  2.382e-01   4.292 1.77e-05 ***
emp.var.rate                -1.597e+00  1.569e-01 -10.180  < 2e-16 ***
cons.price.idx               1.957e+00  2.776e-01   7.049 1.80e-12 ***
cons.conf.idx                1.876e-02  8.609e-03   2.179 0.029350 *
euribor3m                    2.648e-01  1.438e-01   1.842 0.065487 .
nr.employed                  4.643e-03  3.430e-03   1.354 0.175861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23417  on 32949  degrees of freedom
Residual deviance: 13872  on 32897  degrees of freedom
AIC: 13978

Number of Fisher Scoring iterations: 10
```

Python Output:

```
         Current function value: 0.214317
         Iterations 8
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.390
Dependent Variable: y                AIC:              14163.5039
Date:               2022-04-23 19:33 BIC:              14331.5589
No. Observations:   32950            Log-Likelihood:   -7061.8
Df Model:           19               LL-Null:          -11575.
Df Residuals:       32930            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     8.0000
-----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
-----------------------------------------------------------------
age               0.0058   0.0021   2.8031 0.0051  0.0018  0.0099
job               0.0105   0.0062   1.7043 0.0883 -0.0016  0.0226
marital           0.1415   0.0402   3.5239 0.0004  0.0628  0.2203
education         0.0516   0.0110   4.6767 0.0000  0.0300  0.0732
default          -0.3627   0.0728  -4.9832 0.0000 -0.5053 -0.2200
housing          -0.0001   0.0227  -0.0045 0.9964 -0.0446  0.0444
loan             -0.0015   0.0310  -0.0491 0.9608 -0.0622  0.0592
contact          -0.6465   0.0663  -9.7494 0.0000 -0.7765 -0.5165
month            -0.1163   0.0093 -12.5016 0.0000 -0.1345 -0.0981
day_of_week       0.0533   0.0162   3.2909 0.0010  0.0216  0.0851
duration          0.0045   0.0001  56.3180 0.0000  0.0044  0.0047
campaign         -0.0367   0.0128  -2.8670 0.0041 -0.0618 -0.0116
pdays            -0.0010   0.0002  -5.7841 0.0000 -0.0013 -0.0007
previous         -0.0440   0.0612  -0.7180 0.4727 -0.1639  0.0760
poutcome          0.4349   0.0841   5.1706 0.0000  0.2701  0.5998
emp.var.rate     -0.9352   0.0708 -13.2085 0.0000 -1.0739 -0.7964
cons.price.idx    0.6862   0.0342  20.0915 0.0000  0.6193  0.7532
cons.conf.idx     0.0166   0.0050   3.3497 0.0008  0.0069  0.0263
euribor3m         0.6812   0.0820   8.3055 0.0000  0.5205  0.8420
nr.employed      -0.0134   0.0007 -20.0376 0.0000 -0.0148 -0.0121
=================================================================
```