ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

MSC BUSINESS ANALYTICS

Course: Statistics For Business Analytics II

Tutor: D.Karlis

Assignment: Bank Loan Deposits Data Set

Dataset Code: 2

Full Name: Mallios Charalampos

Student ID: P2821912

Table of Contents

0. Introduction

The given dataset relates to telemarketing phone calls to sell long-term deposits. Within a campaign, the agents make phone calls to a list of clients to sell the product (outbound) or, if meanwhile the client calls the contact-center for any other reason, he is asked to subscribe the product (inbound). Thus, the result is a binary unsuccessful or successful contact.

This study considers real data collected from one of the retail bank, from May 2008 to June 2010, in a total of 39883 phone contacts. 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. We are going to use the subset of data having Code = 2 in the particular column.

1. Data Preparation and Data Cleaning

From the given dataset containing 39884 observations and 22 variables a we will use the 3958 observations tagged with code 2 out of the main dataset. Variable code will be dropped out of the dataset since we are not going to use it further since no further information is provided. The second step is to transform the variables to numeric and factors respectively based on the description of dataset. Our dataset does not contain any missing values but there are categories for some variables that are categorized as unknowns. We suppose based on the certain topic (bank loans) that is category of values provide useful information and for that reason, we are going to use them as a level of factor variables. Moreover, we observe that some variables could be better represented as factors since they have some discrete levels. Based on this, we transform the variable pdays to factor since ‘999’ is far away from the other values of the variable. On this dataset we are going to examine our analysis in the next sections.

2. Data Exploration and Full Model

Firstly, we examined the full model investigation analysis taking all the variables as selected to out generalized linear regression model. The median is close to zero since the median deviance residual is close to zero, this means that our model is not biased in one direction (i.e. the outcome is neither over- nor underestimated).

```
glm(formula = new1$SUBSCRIBED ~ ., family = "binomial", data = new1)
```

Figure 1.1: Full model formula

Deviance Residuals				
Min	1Q	Median	3Q	Max
-5.6865	-0.3092	-0.2033	-0.1362	2.9898

Figure 1.2: Full model deviance Residuals

Null deviance: 2439.9 on 3957 degrees of freedom
Residual deviance: 1566.1 on 3906 degrees of freedom
AIC: 1670.1

Figure 1.3: Full model summary

Fit = Residual deviance/ degrees of freedom = 1532.5 /3892 = 0.40

Figure 1.4: Full model fit performance

Some of the variables are statistically significant at 0.05% level. Most influential variable at this point seems to be duration, but let's see find it out based on the boxplot with the response SUBSCRIBED variable below:

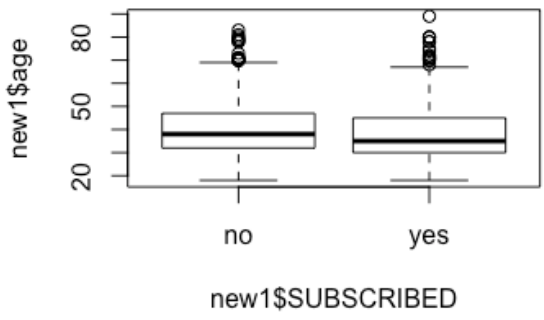


Figure 1.5: Impact of Age on Subscription to loan

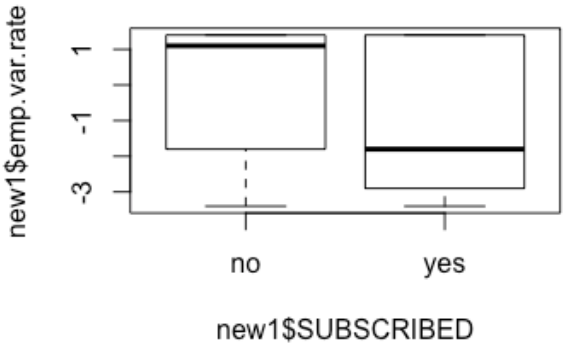


Figure 1.6: Impact of EmpVar on Subscription to loan

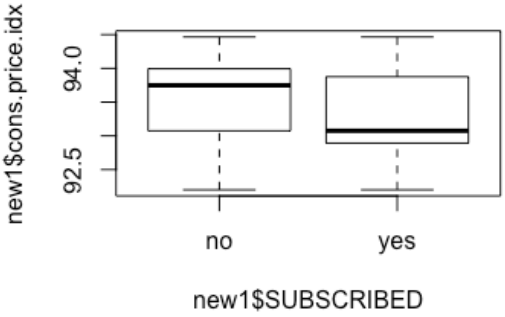


Figure 1.7: Impact of Cons.Price on Subscription to loan

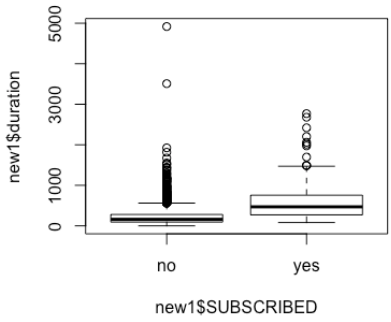


Figure 1.8.1: Impact of Duration on Subscription to loan

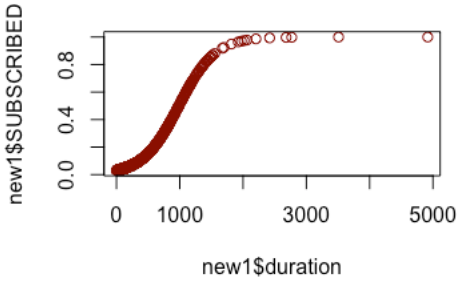


Figure 1.8.2:Logit probabilities if we had the model only with duration

The boxplot confirms that duration is highly influential on subscription prediction probability.

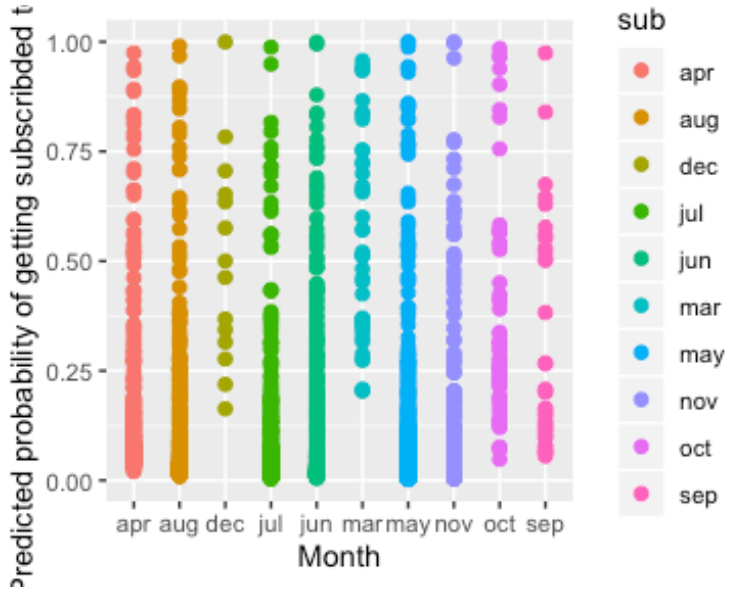


Figure 1.9: Spread on Probability of subscription to loan on month level

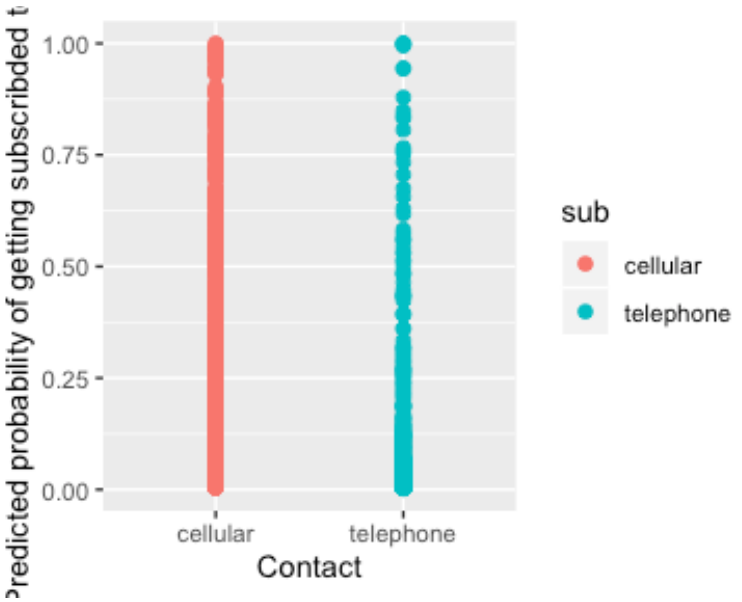


Figure 2.0: Spread on Probability of subscription to loan on contact way levels

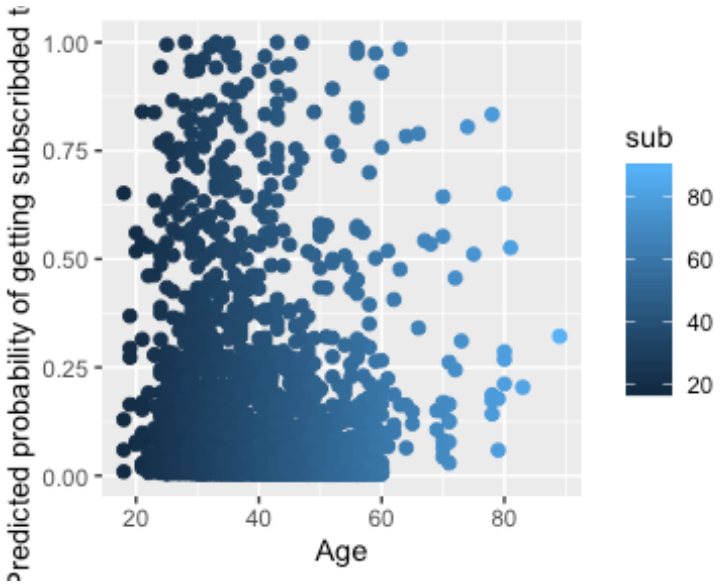


Figure 2.1: Spread on Probability of Subscription to loan on Age

3. Model Selection with Stepwise Procedure

In order to reduce variables from our model and keep only the significant ones we run Stepwise method with both directions for variable selection (adding and deleting each variable in each step based on AIC score). After we have run the stepwise method we finally end with this model:

```
new1$SUBSCRIBED ~ age + default + contact + month + duration +
    pdays + previous + nr.employed


AIC= 1633.3
```

Figure 2.2: Model after stepwise method

Null deviance: 2439.9 on 3957 degrees of freedom
Residual deviance: 1599.3 on 3941 degrees of freedom
AIC: 1633.3

Figure 2.3: Step model summary

Fit = Residual deviance/ degrees of freedom = 1599.3 /3941 = 0.405

Figure 2.4: Step model fit performance

As we observe many of the variables from the initial dataset were removed, keeping only the variables in Figure 2.2. Indeed, the model selection seems to be properly because the AIC was reduced (1670.1 ->1633.3) and the Fit slightly increased (+0.05). This means that we missed less information beside the fact we removed some variables from the full model.

Coefficients of Stepwise Model				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1107670	0.3927353	-0.282	0.77791
age	-0.0144897	0.0062416	-2.321	0.02026 *
defaultunknown	-0.4883296	0.2182633	-2.237	0.02526 *
contacttelephone	-0.5583119	0.1982952	-2.816	0.00487 **
monthaug	0.7595573	0.2823602	2.690	0.00714 **
monthdec	0.8909845	0.6307628	1.413	0.15779
monthjul	0.5643234	0.2958050	1.908	0.05642 .
monthjun	0.7185385	0.2790595	2.575	0.01003 *

monthmar	1.3621006	0.4225396	3.224	0.00127 **
monthmay	-0.6487220	0.2411365	-2.690	0.00714 **
monthnov	-0.2917488	0.3092114	-0.944	0.34541
monthoct	0.4475247	0.3974283	1.126	0.26014
monthsep	-0.4743738	0.4840198	-0.980	0.32705
duration	0.0042803	0.0002322	18.431	**< 2e-16** ***
pdays1	1.7326346	0.3094856	5.598	**2.16e-08** ***
previous	-0.3280596	0.1592296	-2.060	0.03937 *
nr.employed	-0.4105454	0.0406595	-10.097	**< 2e-16** ***

Figure 2.5: Model with stepwise method coefficients

Most of the variables are statistically significant. The **duration**, **pdays1**(not previously contacted) and ***nr.employed*** are statistically more significant(***) but the constant one seems to be not statistically significant(p-value : *0.77791*).Moreover some months are also not statically significant but since some of the month levels variable are statistically significant we are not going to remove it. If we remove the constant one we get the following results:

Null deviance: 5487 on 3958 degrees of freedom
Residual deviance: 1599.3 on 3941 degrees of freedom
AIC: 1633.3

Figure 2.6: Step model summary without constant

Fit = Residual deviance/ degrees of freedom = 1599.3 /3941 = 0.405

Figure 2.7: Full model fit performance without constant

The results are the same so the removal of the constant didn't have positive impact on our model. If we check our variables for multicollinearity we expect none of them to have high GVIF. Indeed, the results are acceptable as we see below :

Variable	GVIF
age	1.063876
default	1.088317
contact	1.445333
month	2.796297
duration	1.175231
pdays	1.932747
previous	2.084716
nr.employed	2.626589

Figure 2.8: Collinearity of model coefficients

For our model we can see the deviance for each variable, and see how much value each variable adds.

Variable	Deviance
age	4.73
default	27.02
contact	49.77
month	152.68
duration	**431.68**
pdays	77.27
previous	0.02
nr.employed	94.35

Figure 2.9: Model with Deviance for each Variable

It is clear that duration offers information to our model in a very influential way following by month and previous seems to be less influential than that other variables.

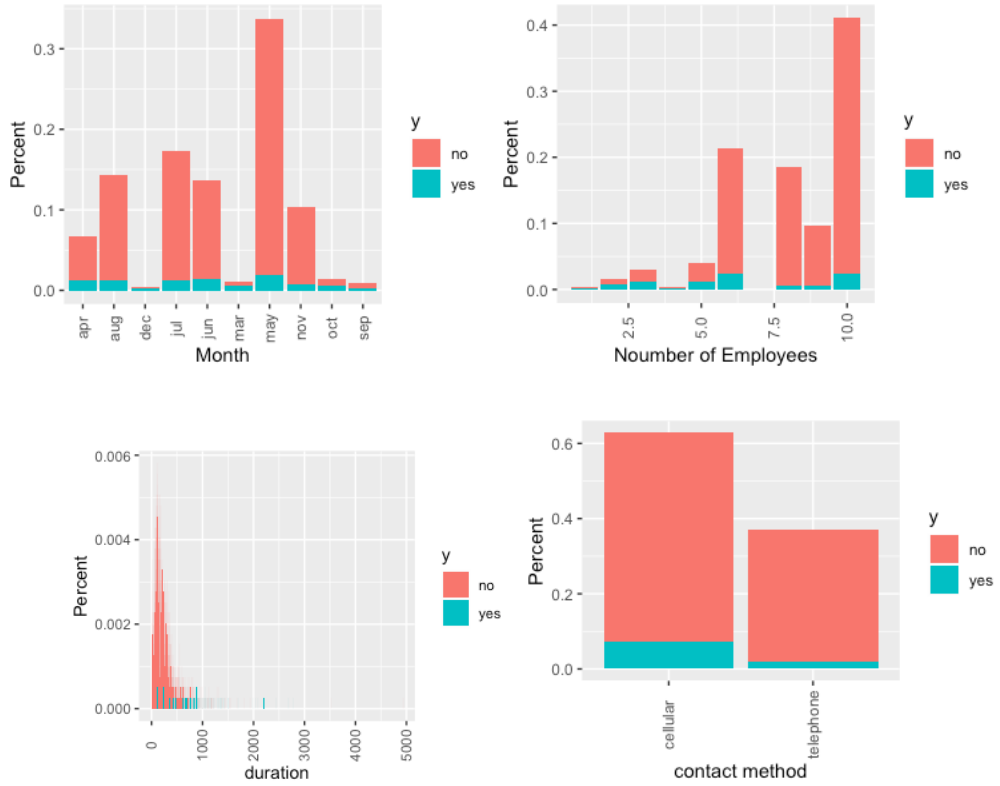


Figure 2.10: Barplots of high deviance variables

4. Interpretation

Subscribed ~ -0.01449 + -0.48833age - 0.48833defaultunknown 0.55831contacttelephone + 0.75956monthaug + 0.89098monthdec + 0.56432 monthjul + 0.71854 monthjun + 1.36210monthmar - 0.64872monthmay - 0.29175monthnov + 0.44752monthoct -0.47437monthsep + 0.00428 duration +1.73263pdays1-0.32806previous -0.41055 nr.employed

Figure 3.1: final model Interpretation

1 unit increase in age *decreases* the log odds of subscription to loan by 0.014, assuming that all other variables are constant. 1 unit increase in defaultunknown *decreases* the log odds of subscription to loan by 0.48833, assuming that all other variables are constant For each month we can keep one month as baseline and see how it going. If we keep August as baseline. $\beta 0$ = -0.1107670: log odds of subscription with the other variables = 0. $\beta 0 - 0.5583119$= -0.1107670− 0.5583119 =-0.6690789: log odds of subscription with the other variables = 0 for a customer contacted by telephone , when the other variables = 0.1 unit increase in has credit in default unknown decreases the log odds of subscription to loan by 0.48, assuming that all other variables are constant.1 unit increase in contact with telephone decreases the log odds of subscription to loan by 0.55, assuming that all other variables are constant.1 unit increase in duration increases the log odds of subscription to loan by 0.00428, assuming that all other variables are constant.1 unit increase in pdays1 (have contacted customer) increases the log odds of subscription to loan by 1.73263, assuming that all other variables are constant.1 unit increase in previous decreases the log odds of subscription to loan by 0.32806, assuming that all other variables are constant.1 unit increase in nr.employed decreases the log odds of subscription to loan by 0.41055, assuming that all other variables are constant.

5. Hypothesis testing

We can test for each covariate if it has effect on response variable SUBSCRIBED.
We can test if β_1 (age covariate) has effect on Subscription for the user with the wald test below :

H_0:β_1 =0 vs H_1: not H_0

Wald test:

Chi-squared test:
X2 = 33.5, df = 1, P(> X2) = 7e-09

But in general, we can test if out model is better than the model with only the constant variable.

```
Wald test

Model 1: new1$SUBSCRIBED ~ age + default + contact + month + duration +
    pdays + previous + nr.employed
Model 2: new1$SUBSCRIBED ~ 1
  Res.Df  Df  Chisq Pr(>Chisq)
1   3941
2   3957 -16 534.21  < 2.2e-16 ***
```

Figure 4.1: Wald Test for Hypothesis Testing

As we can see our model, based on p-value(<0.05), our final model is better than the constant one. Likelihood ratio test also confirms the results.

```
Model 1: new1$SUBSCRIBED ~ age + default + contact + month + duration +
    pdays + previous + nr.employed
Model 2: new1$SUBSCRIBED ~ 1
  #Df   LogLik  Df  Chisq Pr(>Chisq)
1  17  -799.65
2   1 -1219.93 -16 840.55  < 2.2e-16 ***
```

Figure 4.2: Walt Test for Hypothesis Testing

To see how good the model fits our data we can conduct X^2 test. The value of p is 1 which means that fits perfectly the data, but seems to good to be true . Furthermore, we can compute McFaden pseudo-R^2 to conduct investigation.

```
McFaden :  R^2 : 0.344
```

Figure 4.3: McFaden R^2

Which indicates how good the model explains our data. The value is not the so high but indicates that our model tries to explain the data.

6. Predicting Performance

Our model seems to explain the data in satisfactory way, but we would like to investigate furthermore its predicting ability. For this reason, we create a partition to our data, a training dataset and a testing dataset(80/20 ratio). We want to explore if our model provides an appropriate classification to SUBSCRIBED (YES or NO) to loan.

The training classification :

No Yes
3010 157

The testing classification :

No Yes
718 73

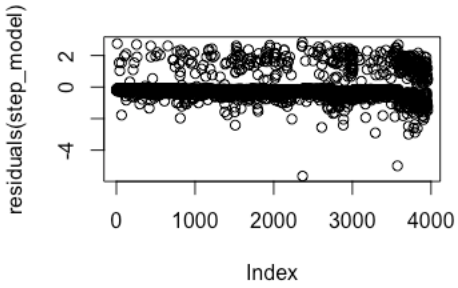
Figure 5.1: Classification of our model

Training Model Ratio : 3010 No / 157 Yes = 19.17

Testting Model Ratio : 718 No / 73 Yes = 9.83

The model tries to appropriately classify the values giving more classified as No but in other hand in the testing dataset add more subscriptions to loan Yes and this way we understand that treats the data differently that it should do and so the accuracy is not good. For the testing dataset the Yes classifies are two times more than it should.

We can also check for the residuals of the model to see if there any outliers. Indeed sum(residuals(step_model))^2 should be close to 1(values between 0-1) but is huge : 315609.3 and below is the graph with multiple values outliers.



7. Conclusion

In a nutshell, having transformed our data and running a model keeping the best in terms of Akaike we ended up with a model better than the constant or the full model with all the variables included. The model tries to fit the data in an appropriate way but the high variance of values doesn't make it appropriate for prediction purpose of use.