*This class assignment partially reproduces the analysis from the publication “A data-driven approach to predict the success of bank telemarketing” by Moro, Cortez, and Rita (2014).*

## Background & Summary

I consider a dataset extracted from ‘bank.csv’ which contains statistics for prospective customers of a Portugese banking institution. The data have been first analyzed in Moro, Cortez, and Rita (2014), but they used 150 variables, whereas this assignment provided only a subset of the data.

The dataset has statistics for 41,188 prospective customers, which include the values with regards to 21 variables (age, job, marital, education attainment, credit in default, housing loan, personal loan, contact communication type, month of last contact, day of last contact, duration of contact, number of contacts in this campaign, number of days after last contact, number of previous contacts, outcome of the previous marketing campaign, employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate, number of employees), all of which determine the outcome (Bernoulli) variable -- has the client subscribed to a term deposit? This is a binary dependent variable which takes on two values, 'yes’ or 'no', depending on the subscription to the term deposit.

I have computed a Logistic Regression (LR) model to estimate a model for the outcome variable.

## Exploratory analysis

### Data cleaning 1: Missing values

Missing values in the dataset were coded as ‘unknown’. There were 12,718 missing values, most of them in the default (8,597 missing values), education (1,731 missing values), housing (990 missing values), and loan (990 missing values). All observations with missing housing values were also missing loan values.

I did not delete the 8,597 missing observations from the default variable, but instead coded them as ‘unknown’ and treated them as potentially predictive. (Details in Appendix 1).

I deleted the 4121 observations with missing values in other variables, as rows with missing values did not differ with respect to the mean value of the outcome variable. Methods that would allow me to deal with the missing data in a more sophisticated way are beyond the scope of this course.

### Data cleaning 2: Categorical variables

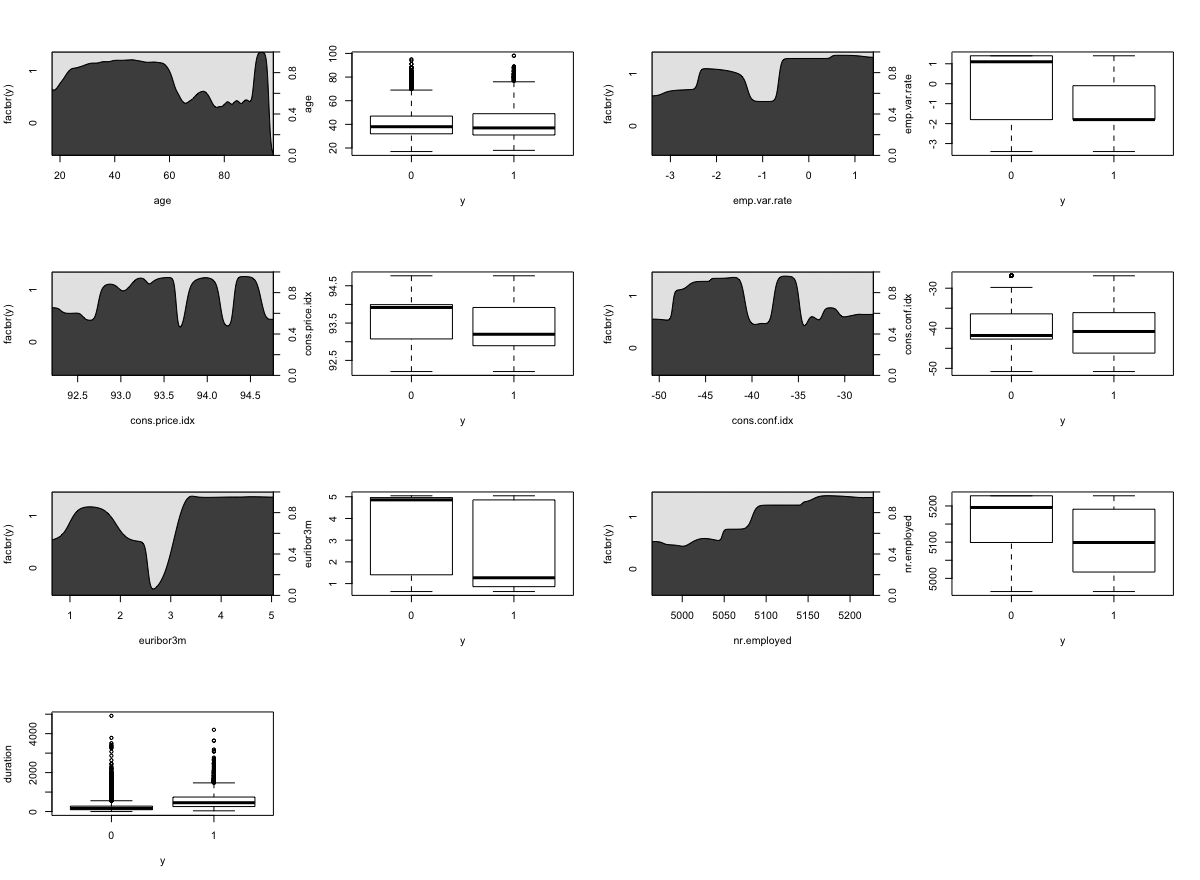
I explain how categorical variables were merged in (Appendix 1).

### Which predictors appear to affect the outcome variable?

During the exploratory stage of the analysis I compared the mean value of the outcome variable for each factor to decide which categorical variables were likely to be useful in the final model. All categorical variables except for day\_of\_week, housing, and loan were likely predictive.

For each continuous variable I made a boxplot and a conditional density plot (Figure 1) against the outcome variable. Nr. employed looks most promising. The plots showed that the relationship between euribor3m, cons.conf.idx, and cons.price.idx and the probability of success in outcome variable is likely not linear. The boxplot means of cons.conf.idx and age were close to each other. Boxplots of age and duration revealed multiple outliers.

**Figure 1:**

I also computed regressions of each variable and all were significant at 1%. I did not exclude any variables at this stage. 

## 

## From Initial to Final Model

### Null model vs. Full model

I chose variables and assess model adequacy using the -2Log Likelihood statistic, also known as the deviance.

I first found the deviance of the null model by fitting a regression with just an intercept, which was 26720 on 38244 degrees of freedom. I then fitted a full model, which had a lower deviance of 15,840 on 38,215 degrees of freedom. The difference in deviance is 10,874 and this reduction is in the critical region of an Chi-Square distribution (p = 2.2e-16); the full model is significantly better than the null model.

### Collinearity

Before exploring other models, I checked for collinearity using two methods: with a correlation table and variance inflation factors (VIF).

All five social and economic context attributes had a VIF above 5, likely due to correlations with one another. For this reason the VIF for month was also high. Predictably, was.contacted and past\_success were also suspect.

**VIFs for the full model:**

illiterate 1.002814

job\_merged 2.031890

confirmed.no.default 1.117213

campaign 1.046740

evermarried 1.366513

**was.contacted 11.093808**

**past\_success 9.490814**

**month 60.404659**

contact 2.377694

previous 2.062212

age 2.117878

duration 1.237489

**emp.var.rate 140.941668**

**cons.price.idx 67.000081**

cons.conf.idx 5.345445

**euribor3m 134.053807**

**nr.employed 167.994787**

The correlation matrix showed correlations of over 0.5 for past\_success and was.contacted (0.95), contact\_by\_phone and cons.price.idx (0.59), was.contacted and previous (0.58), past\_success and previous (0.51), cons.price.idx and nr.employed (.52)emp.var.rate and euribor3m (0.97), emp.var.rate and nr.employed (0.90), cons.price.idx and emp.var.rate (0.77), and euribor3m and nr.employed (0.95).

I dropped was.contacted (because past\_success had slightly lower correlations with other variables).

To choose which contextual variables I should keep, I used the correlation table, VIFs, and the conditional density plots (the plots informed me if the relationship with the outcome was linear): I dropped emp.var.rate and euribor3m, and I kept nr.employed, cons.conf.idx, and cons.price.idx.

For the dataset with fewer variables, the only high VIF was for month at 4.9. The highest correlations were between previous, contact\_by\_phone, cons.conf.idx, cons.price.idx, and nr.employed, but none exceeded 0.60.

I removed the variable campaign because its coefficient, although significant, was small.

### Model comparison based on deviance

After dropping nearly collinear predictors, fit was in practice identical to the previous model but with only 17 predictors. The marginal uncertainties for the remaining predictors are also smaller. The defiance was only slightly larger.

I then looked at multiple models sequentially using the Chisq test in the anova() function, adding one variable at a time and asking if each model with one additional variable significantly improves on the previous model.

Then, after finding that housing and loan variables do not significantly reduce deviance, which was expected based on the exploratory data analysis, I removed them and, once again, evaluated multiple models sequentially using the Chisq test on deviance reduction.

Some variables significantly reduced deviance but by a small amount. Those were age, previous, evermarried, cons.price.idx, and cons.conf.idx. A logit regression without them had higher deviance; the difference was significant at 1% but slight (magnitude 5). I dropped those variables to simplify the model and make it more interpretable.

I removed the variable duration because I did not know how to interpret it, and because it did not improve classification error, even though it reduced deviance by a large amount.

I called the resulting model glm.5.

### Model comparison based on classification error

As we can see below, the difference in prediction accuracy between the incumbent model (glm.5) and the null model is a slight 1.2%. When we only look at model accuracy in predicting positive outcomes, we notice a large improvement of the incumbent model over the null model.

88.8% is the proportion of observations correctly classified by the null model (this is the proportion of negative outcomes in the dataset). However, only 9.8% observations with positive outcomes (y=1) were correctly predicted. This improved markedly to 40.5% for the incumbent model.

The month variable significantly reduced deviance by a big amount but had a relatively high VIF score. The incumbent model without the month variable classified correctly 0.1% fewer overall observations and 2.4% fewer positive outcome observations. Including the month variable in the model improves classification accuracy, so I left it in the model.

**Table 1: Classification error**

| Model | repeated 10-fold CV prediction accuracy for all observations | repeated 10-fold CV prediction accuracy for y=1 | repeated 10-fold CV MSE | Leave-One-Out CV MSE |
| --- | --- | --- | --- | --- |
| glm.5 | 91.0% | 40.5% | 0.06334622 | 0.06332932 |
| glm.5 w/o month | 90.9% | 38.1% | 0.06510777 | 0.0650993 |
| null | 88.8% | 9.8% | 0.09894526 | did not compute |

## Results

The incumbent model has the following form:

**Model 1: ln (p/[1-p]) = 50.475 - 0.142 Low-wage-or-service + 0.322 Student + 0.293 Retired + 0.095 Has-university + 0.273 Confirmed-not-in-defaut + 1.474 Past-success - 0.408 Contacted-by-phone - 0.010 Nr.employed + 0.943 March - 0.617 May + 0.318 June + 0.265 July + 0.053 August - 0.235 September + 0.153 October - 0.263 November**

OR

**Model 2: P (y = 1) = inverse logit [50.475 - 0.142 Low-wage-or-service + 0.322 Student + 0.293 Retired + 0.095 Has-university + 0.273 Confirmed-not-in-defaut + 1.474 Past-success - 0.408 Contacted-by-phone - 0.010 Nr.employed + 0.943 March - 0.617 May + 0.318 June + 0.265 July + 0.053 August - 0.235 September + 0.153 October - 0.263 November]**

Based on the Wald statistic (Chi-square), all coefficients in the model above were significant at at least 1%, with the exceptions of August, September, and Has-university, which were significant at 5%. See Table 2 below:

**Table 2: Coefficients**

| Coefficients | Estimate | Std. Error | z value | Sig |
| --- | --- | --- | --- | --- |
| Intercept | 50.474 | 1.507 | 33.491 | < 2e-16 |
| Low-wage-or-service | -0.141 | 0.048 | -2.914 | 0.003 |
| Student | 0.321 | 0.100 | 3.196 | 0.001 |
| Retired | 0.293 | 0.074 | 3.920 | 8.85e-05 |
| Has-university | 0.095 | 0.042 | 2.258 | 0.023 |
| Confirmed-not-in-defaut | 0.273 | 0.058 | 4.634 | 3.59ee-06 |
| Past-success | 1.474 | 0.070 | 20.982 | < 2e-16 |
| Contacted-by-phone | -0.407 | 0.051 | -7.880 | 3.27e-15 |
| Nr.employed | -0.010 | 0.000 | -34.905 | < 2e-16 |
| March | 0.943 | 0.109 | 8.621 | < 2e-16 |
| May | -0.617 | 0.065 | -9.428 | < 2e-16 |
| June | 0.317 | 0.079 | 4.018 | 5.88e-05 |
| July | 0.265 | 0.074 | 3.548 | 0.000 |
| August | 0.05 | 0.007 | 0.2 | 0.046 |
| September | -0.235 | 0.113 | -2.077 | 0.037 |
| October | 0.153 | 0.103 | 1.482 | 0.138 |
| November | -0.235 | 0.113 | -2.077 | 0.037 |

The biggest limitation of this model and the data underlying it, is that they were all collected in the same year, making it impossible to ascertain the real significance of the contextual variables. It is recommended that comparable data be collected in future years, which will make it possible to interpret the month variable.

### Interpretation of coefficients

A difference of 1 in the nr.employed variable (continuous variable with interquartile range of 5099-5228) corresponds to a negative difference of -0.010 in the logit probability of positive outcome. The difference in logit probability (i.e log of odds) between a person in the 25th percentile and a person in the 75th is therefore negative 1.29.

Having information that a person is not in default increases their logit probability by 0.273 compared to not having that information, and having a university diploma increases logit probability by 0.095. If a person was previously contacted with success, their logit probability of subscribing increases by a relatively big 1.475. Being contacted by telephone, and not by cell, decreases logit probability by 0.408

Each month has fixed effects; being contacted in March increased logit probability of success by 0.943, while in May decreased by 0.617. It is difficult to know what this means for our inference without comparable data from other years. It is possible that those coefficients are seasonal effects, but it is also possible that they reflect differences in the data generating process.

Occupation matters for outcomes. Students and retirees have about 0.5 greater logit probability of positive outcome than low-wage and service workers, and 0.3 greater than other occupations.

This value is the logarithms of odds, which makes it more difficult to interpret than probability statements. We use Model 2 to arrive at probabilities in the lay section.

## Interpretation and conclusion for a lay audience

### Question of interest

We have extensive background information on the people who were contacted in a direct-marketing campaign; we also know who subscribed to a term deposit.

Based on those data, we are interested in finding out what worked and what is likely to work in the future. Who were the people who subscribed? What makes a person more likely to subscribe to a term deposit after being contacted by the banking institution?

### Results: what predictors made the outcome more/less likely?

First, look at Table 2. It contains the percentage who subscribed to the term deposit among the people who have a given attribute (e.g. 30.2% of students subscribed compared to 7.3% low-wage and service workers). We see that, for example, retirees and students subscribed more often than people employed in low-wage occupations, and that the rate of subscriptions was extremely high among those who have been successfully contacted in the past.

**Table 3: Subscription rate per attribute (biggest values in bold)**

| Variable name | Value | Percentage of people with this value who subscribed to the term deposit |
| --- | --- | --- |
| Job | Low-wage or service job | 7.3% |
|  | **Retiree** | **24.7%** |
|  | **Student** | **30.2%** |
|  | Other job | 11.7% |
|  |  |  |
| Has university degree | **True** | **13.7%** |
|  | False | 9.9% |
|  |  |  |
| Was successfully contacted in the past | **True** | **64.5%** |
|  | False | 9.3% |
|  |  |  |
| Contacted with a... | Telephone | 5.2% |
|  | **Cellphone** | **14.5%** |
|  |  |  |
| Is in default? | **No** | **12.7%** |
|  | Didn’t give information | 5.1% |
|  |  |  |
| Month (selected) | **March** | **50.7%** |
|  | May | 6.5% |
|  | June | 10.5% |
|  | July | 9.0% |
|  | August | 10.2% |
|  | **September** | **44.8%** |
|  | **October** | **45.1%** |
|  | November | 9.8% |

Second, look at the logistic regression model:

**Model 1: Probability of subscribing to term deposit = 63.12 - 0.206 Low-wage-or-service + 0.340 Student + 0.319 Retired + 0.149 Has-university + 0.273 Confirmed-not-in-defaut + 1.475 Past-success - 0.391 Contacted-by-phone - 0.13 Nr.employed + 1.415 March - 0.663 May + 0.532 June + 0.282 July + 0.368 August - 0.258 September + 0.222 October - 0.199 November**

The lay reader should consider the relative magnitude of the coefficients and whether they are negative or positive.

The following values make a successful outcome more likely: Having been successfully contacted in the past, being a student, being a retiree, having a university degree, confirmed information about not being in default, being contacted by cellphone, having been contacted in March, June, July, August, or October.

Alternative way of viewing the model is to see which values make subscribing to a term deposit less likely: working in a low-wage or service job, unknown default status, not having a university degree, being contacted by telephone, having been contacted in September, May, or November.

### What the results mean *in practice*

What does this mean *in practice* for four clients with different attributes?

Client 1: For a client working a low-wage or service job, who has no university diploma, did not provide us information about not being in default, who has not been successfully contacted in the past, was contacted by us by the telephone in May, when the nr.employed was high (5191), the probability of subscribing a term deposit is 1.98%.

Client 2: We can significantly increase the likelihood of subscribing by limiting our attention to clients, with whom we’ve interacted successfully in the past and making our calls at an opportune time. For a similar person to the one above, but who has been successfully contacted in the past and was contacted by us now when nr.employed is low (5099), the probability of subscribing a term deposit increases to 23.07%.

Client 3: Furthermore, we know that people who have confirmed not being in default, who are either retirees or students, and who have been contacted with a cellphone were more likely to subscribe. For example, a retiree with no university diploma, who confirmed not being in default, was contacted with a cellphone, had a probability of subscribing to a term deposit of 54.17%.

Client 4: Now let’s look at the person who has the highest probability of having subscribed to the term deposit: 89.83%. It is a student, who has confirmed that he is not in default, and was contacted with success during a previous campaign. Moreover, he was contacted with a cellphone and the contact happened at an opportune time: in March, and when the nr.employed was low (5008).

### Suggestions regarding the marketing strategy

Assuming a causal interpretation of the model, to maximize the returns on investment from a direct-marketing campaign one should:

* Target students and retirees; this product is best suited for them.
* Cultivate loyal customers, i.e., target persons with past success.
* Call on cell phones, not telephones.
* Spend more marketing money on people not in default, students, retirees, and people with degrees.
* When choosing a market, don’t choose low-wage or service clients.
* Contextual factors and time of contact matter.

## Appendix 1

The variable missing most values, default, had three values: ‘yes’, ‘no’, and ‘unknown’. There were only three ‘yes’ observations (n=3) to treat ‘yes’ as a separate level. Observations with missing default values had a smaller mean value of the outcome variable (0.05 compared to 0.12 for observations with default value of ‘no’), therefore I did not delete missing values from the default variable and treated them as potentially predictive.

Mean values of the outcome variable were low for low-wage blue-collar and service jobs, and high for students and retirees (see Table 1 in the lay part of the report). Therefore, I merged some levels of the job variable. New job\_merged variable had four levels: ‘low-wage’, ‘student’, ‘retiree’, and ‘other’.

I transformed the three-level marital variable into a binary evermarried variable.

Mean value of the outcome variable was unusually large for the illiterate, so I transformed the education variable into a binary illiterate variable. It was possible to leave more educational levels, but went with a binary variable for ease of interpretation.

I transformed the default variable into a binary variable called confirmed.no.default.

I transformed the numeric variable pdays into a binary variable was.contacted, which takes the value of 1 if there was contact in the past 27 days.

Lastly, based on the man values of the outcome variable, I transformed the categorical variable poutcome into a binary variable past\_success, which indicates the success of previous marketing campaigns.

References:

Moro, Sérgio, Paulo Cortez, and Paulo Rita. "A data-driven approach to predict the success of bank telemarketing." Decision Support Systems 62 (2014): 22-31.