


## Project cover sheet

**Project title: Portuguese Banking Product**

**Group members: Adam Mills, Aveek Das, Niall Martin, Sujit Krishnankutty**

**All group members have read and agreed to the final version of all documents.**

### Signatures

1. \_\_\_\_\_
2. **AVEEK DAS**
3. **NIALL MARTIN**
4. **SUJIT KRISHNANKUTTY**

(If you cannot provide a digital or scanned signature, please type your name and send an e-mail to [rafael.deandrademoral@mu.ie](mailto:rafael.deandrademoral@mu.ie) confirming that you have read and agreed to the final version of all documents)

## Introduction

We have been provided with a Portuguese Banking Product and we have its marketing campaign information with many variables recorded along with its binary response “y” indicating whether or not the client subscribed to a term deposit. We are using logistic regression, a predictive modeling algorithm that is used when the Y variable is binary categorical (0 or 1). The goal is to determine a mathematical equation that can be used to predict the probability of event 1. The main objective of this project is to demonstrate the analysis, model building process, and capabilities for fitting logistic regression models using SAS, R, Python, and Minitab software.

## Methodology

In this project, we have tried to implement a simple Logistic Regression model and compared the fitting process using Python, R, SAS, and Minitab. In the following section, we will discuss each of the methodologies in brief.

### Python

The first step to start with was cleaning the dataset and removing unwanted variables that do not contribute to the response variable. Since there were no fields like IDs, we did not remove any columns. Then, we have checked for the missing values from the dataset as it affects the natural variation of the dataset and in turn, affects the model accuracy as well.

Moving forward, we started by exploring the categorical and the numerical variables and the distribution of the continuous and discrete numerical variables. The exploratory data analysis was continued by understanding the count of records per categorical variable. In the following section, we have encoded all the categorical variables to 0s and 1s based on the values and applied the same for the target variable as well. Finally, we have split the data into training and test in the ratio 70:30 which

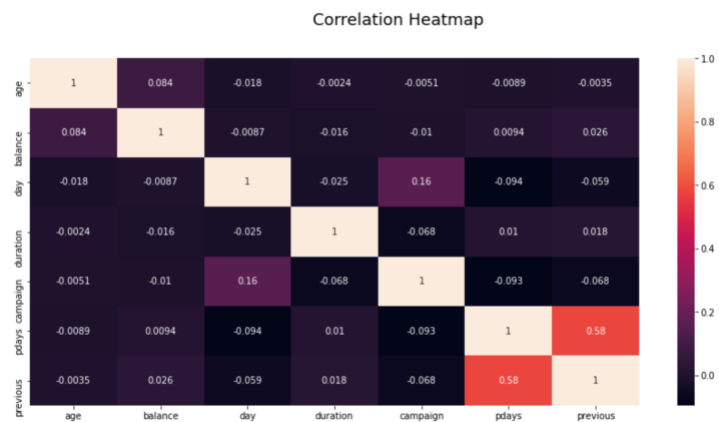


Figure 1 - Correlation Matrix of Categorical Variables

was then used to fit the model. For the model, we have used the LogisticRegression available under the sklearn.linear\_model library. Using the fit obtained by this model, we tried to predict the term deposits for the customers on the test dataset we obtained after splitting. In the later section of this report, we will discuss the results of the model and the accuracy obtained.

### R

For performing the logistic regression in R, different steps are executed before directly applying regression on the given dataset. Data cleaning is performed by checking if null values are present for which we don't have any values in our given dataset. We have removed the duration variable from the given dataset as it is not conducive to our model where our main objective is prediction. After data cleaning, categorical predictors are converted into factors. We have performed exploratory analysis for Age Distribution vs

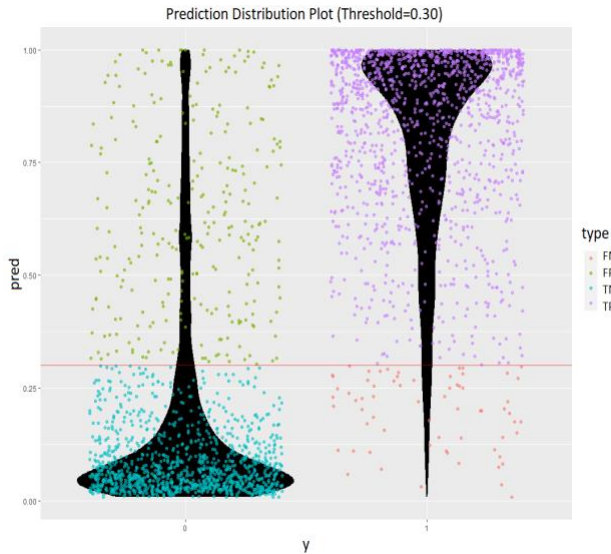


Figure 2 - Prediction Distribution Plot

fitted model with 10 iterations and measures like accuracy, specificity, sensitivity was calculated and discussed in the later sections.

## SAS

Next, we performed the logistic regression analysis in SAS. As mentioned before, it was important to ensure that the data was clean and of suitable structure to perform the analysis. It was not necessary to convert the response variable into 1 and 0 because it is possible to use a class statement in SAS. Having done this the next step was to perform exploratory analysis. We made plots using sgplot to compare the response variable with the explanatory variables. Some of the plots are displayed below. We can see from the first plot that the success of the previous marketing campaign clearly had an impact on whether the client subscribed to the loan. Marital status didn't have much of an impact and there were certain months that had much larger subscription rates than others. We then split the dataset into test and training datasets we used the 70:30 split as previously mentioned. This was done using `proc surveyselect`. The next part of the analysis was to perform the logistic regression. This was performed using `proc logistic`. We included `lackfit selection=backward slstay=0.05` to remove any insignificant variables from the model through backward elimination.

Marital Status that subscribes to the Term Deposit. We have found that the proportion of the response variable (Yes v/s No) is not balanced, hence we are observing skewness in the dataset which can lead to some bias. Data is reshaped into equal proportion response and it was split for training and test sets in the ratio: 70/30. This dataset is sufficiently large enough to justify the 70/30 split, as opposed to something like an 80/20 or more split. After performing logistic regression on the data, the model contains some insignificant variables, for which we have performed model fitting using Backward elimination due to which all the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation. Cross-validation is carried out for assessing the performance of the

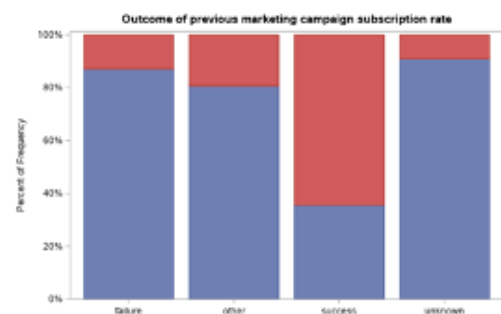


Figure 3 - SAS EDA 1

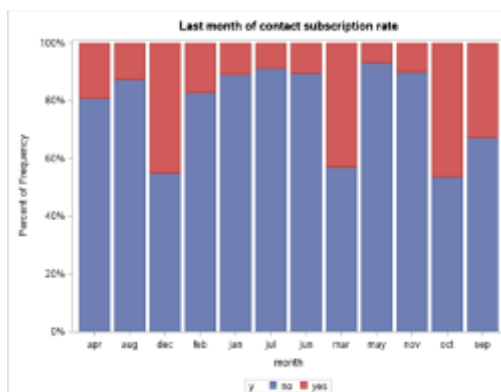


Figure 4 - SAS EDA 2

## Minitab

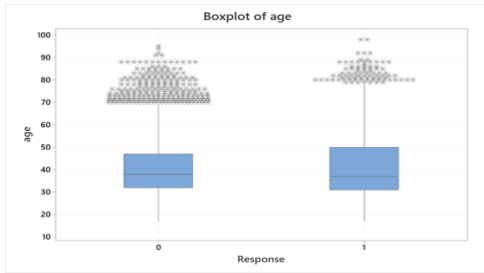


Figure 5 - Minitab EDA 1

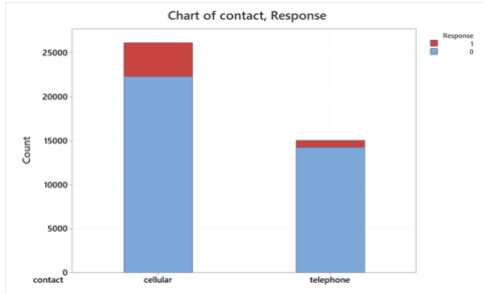


Figure 6 - Minitab EDA 2

As with the previous methods, duration is removed from the dataset as suggested by the brief. The response  $y$ , whether an individual subscribed to a term deposit, is converted to a binary “1” and “0” to allow for the modelling. We undertook some basic exploratory analysis using Minitab to examine the structure and content of the data. This included producing simple boxplots and stacked bar charts (see Fig. 5 and Fig. 6 for examples). The dataset was balanced by synthetically generating data in order to avoid problems with skewness. Prior to executing the regression, the data was split into training (70%) and test (30%) sets in order to judge the model’s performance. As we have quite a few predictors, it may be useful to only use a subset of these as some may be insignificant. Stepwise regression with alpha to enter and alpha to remove set to 0.15 was used in order to reduce the number of variables necessary in generating a well-performing logistic regression model.

## Results

After the performing the exploratory data analysis and fitting the logistic regression model, we have obtained results from all the four different tools which are discussed below. The output from Python has been displayed in Table 1 and the ROC Curve has been plotted in Figure 8. An accuracy of **87.77%** was obtained, which provided a sensitivity of **0.2036** and a specificity of **0.9723**. *Sensitivity* measures the proportion of positives that are correctly identified. *Specificity* measures the proportion of negatives that are correctly identified. *Accuracy* is the proportion of correctly identified cases from all the cases.

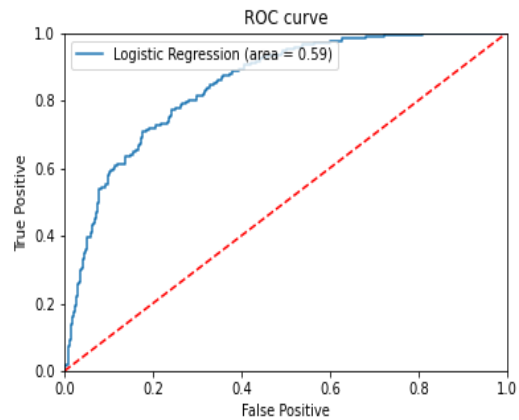


Figure 7 - ROC Curve from Python

## Discussion

In this section, we have performed a comparative study of predicting the term deposits using Logistic Regression using the four tools, Python, R, SAS and Minitab.

## SAS

SAS was a very useful software to utilize to perform a logistic regression analysis. In terms of cleaning the data, it proved efficient and non-time-consuming. There are some limitations with regards to exploratory analysis for example the aesthetics cannot be altered as much as other software. In relation to performing the logistic regression itself, the software had some useful features. For example, it was possible to remove

variables from the model that weren't significant by simply adding `lackfit selection=backward slstay=0.05` to the procedure. This then created an analysis of each step in the process of building a suitable model.

## Minitab

The use of the statistical software Minitab in producing this model had advantages and disadvantages. The point and click-based interface changes how the user interacts with the data slightly, compared with writing code in the likes of R or Python. Some operations, such as generating the logistic regression model, were found to be intuitive and fast. However, other operations such as data cleaning and manipulation proved to be slightly more difficult than they would have been in R. For example, the process of cleaning data or replacing particular variables in a dataset was found to be more time-consuming. Certain features appeared not to be available using Minitab, for instance, the balancing of the data was ultimately achieved in R via the caret package and not directly in the Minitab software.

## Python

Python is one of the most powerful tools when it comes to dealing with Machine Learning algorithms. The Exploratory Data Analysis was performed using the Pandas library and the graphs were plotted using the Seaborn and Matplotlib libraries. The model training and fitting was done using the LogisticRegression module from the Scikit Learn library. The results obtained from the model were used to fit using the training dataset and the prediction was performed on the test dataset. The performance measures such as accuracy, precision, sensitivity and specificity have been measured to understand how well the model has performed and an ROC curve has been obtained.

## R

R proved to be another useful software while performing logistic regression analysis. Data cleaning and transformation proved to be non-time consuming. As compared to other languages, R has much better options for plots for exploratory analysis. For data reshaping, caret package was used. R provides a descriptive summary of the logistic regression model. As compared to SAS, performing Model fitting using backward elimination proved to be more time-consuming. Cross-validation was also performed with 10 iterations. As we have performed down sampling, performance measures like sensitivity, accuracy, precision are much better as compared to other languages.

## References

UCI Machine Learning Repository, *Bank Marketing Data Set* viewed online on <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

James, G, Witten, D, Hastie, T, and Tibshirani, R 2013, "*An Introduction to Statistical Learning With Application in R*"

Lantz, B, 2015, "*Machine Learning With R*", second edition, published by Packt Publishing Ltd. Livery Place, 35 Livery Street, Birmingham B3 2PB, UK.