# Logistic Regression Case Study

About the Case

The dataset comes from the [UCI Machine Learning repository](https://archive.ics.uci.edu/ml/index.php), and it is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict whether the client will subscribe (1/0) to a term deposit (variable y).

Data Description

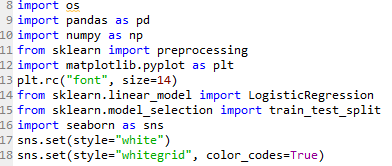
1. age (numeric)  
2. job : type of job (categorical: “admin”, “blue-collar”, “entrepreneur”, “housemaid”, “management”, “retired”, “self-employed”, “services”, “student”, “technician”, “unemployed”, “unknown”)  
3. marital : marital status (categorical: “divorced”, “married”, “single”, “unknown”)  
4. education (categorical: “basic.4y”, “basic.6y”, “basic.9y”, “high.school”, “illiterate”, “professional.course”, “university.degree”, “unknown”)  
5. default: has credit in default? (categorical: “no”, “yes”, “unknown”)  
6. housing: has housing loan? (categorical: “no”, “yes”, “unknown”)  
7. loan: has personal loan? (categorical: “no”, “yes”, “unknown”)  
8. contact: contact communication type (categorical: “cellular”, “telephone”)  
9. month: last contact month of year (categorical: “jan”, “feb”, “mar”, …, “nov”, “dec”)  
10. day\_of\_week: last contact day of the week (categorical: “mon”, “tue”, “wed”, “thu”, “fri”)  
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y=’no’).

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14. previous: number of contacts performed before this campaign and for this client (numeric)  
15. poutcome: outcome of the previous marketing campaign (categorical: “failure”, “nonexistent”, “success”)  
16. emp.var.rate: employment variation rate — (numeric)  
17. cons.price.idx: consumer price index — (numeric)  
18. cons.conf.idx: consumer confidence index — (numeric)  
19. euribor3m: euribor 3 month rate — (numeric)  
20. nr.employed: number of employees — (numeric)

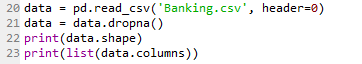
**Predict variable (desired target)**

y — has the client subscribed a term deposit? (binary: “1”, means “Yes”, “0” means “No”)

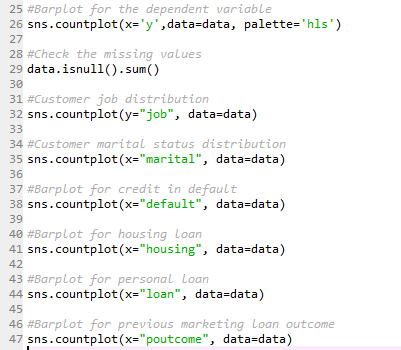
**Importing packages:**



**Data**:



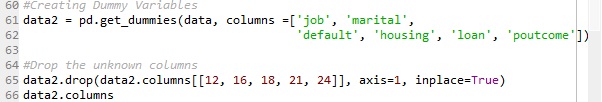
Checking for missing variables and barplot of the data:



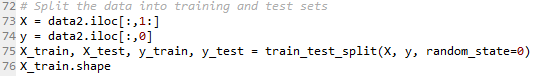
Now we drop the variables which we will not take for fitting the model.



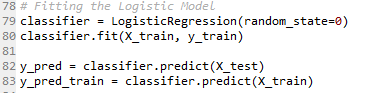
**Creating the dummy variables:**



**Splitting the data into two, test and train.**

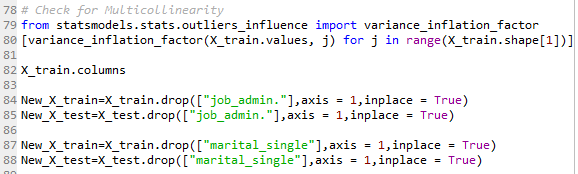


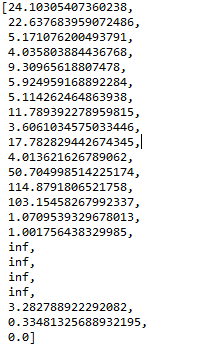
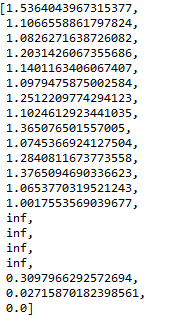
**Fitting the model:**



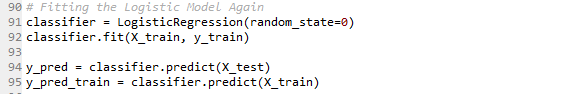
**Checking** **multicollinearity and removing them:**

multicollinearity states that there should be no perfect linear relationship between two or more of the predictors or independent variables. This is tested with the vif function and any variable with a value of GVIF significantly more than 4 will have to be removed the model. In our case multicollinearity was present



**Fitting the model again:**

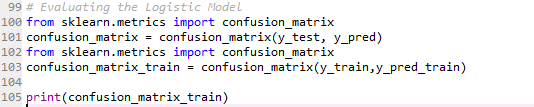




.

So here we see the accuracy of our model is 90%.

**Confusion Matrix:**

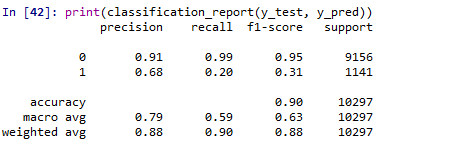




So it means we have predicted 27049+647 correctly and 2852+ 343 predictions were wrong.

**Classification report:**

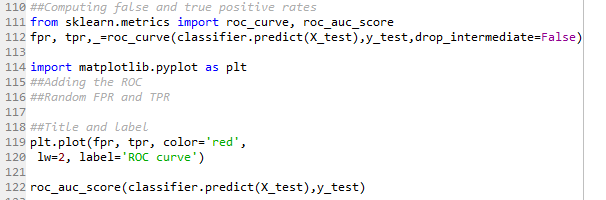


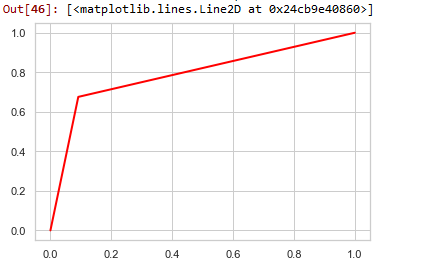


**Interpretation**: Of the entire test set, 90% of the promoted term deposit were the term deposit that the customers liked. Of the entire test set, 90% of the customer’s preferred term deposits that were promoted.

**ROC Curve:**

Receiver Operating Characteristic(ROC) summarizes the model’s performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model. Below is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals 0. This curve will touch the top left corner of the graph.







AUC / C-Statistic ranges from 0.5-1. It explains the trade-off between true positive rate (Sensitivity) and false positive rate (1-Specificity). The Higher the better. Here we got 0.79 which is good enough. Hence the model is reasonably good.