**Logistic Regression Approach to Predict the Success of Bank Telemarketing**

**Student Number:** D20123764

**Name:** Shakti Sampad Dash

**Programme Code:** TU60

**Note:** I have used the direct marketing campaigns of a Portuguese banking institution [dataset](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) to complete the independent project.

Contents

[Abstract 3](#_Toc60848097)

[Section 1 - Research Question(s) 5](#_Toc60848098)

[Question-1 5](#_Toc60848099)

[Question-2 5](#_Toc60848100)

[Section 2 - Dataset 5](#_Toc60848101)

[How to handle missing data 5](#_Toc60848102)

[Findings & Actions 6](#_Toc60848103)

[Representativeness 7](#_Toc60848104)

[Variables of interest and their statistical measurement type 7](#_Toc60848105)

[Euribor3m 7](#_Toc60848106)

[nr.employed 8](#_Toc60848107)

[emp.var.rate 8](#_Toc60848108)

[cons.price.idx 9](#_Toc60848109)

[cons.conf.idx 9](#_Toc60848110)

[Section 3 - Results 10](#_Toc60848111)

[Section 3.1 - Statistical Evidence 10](#_Toc60848112)

[Forward stepwise selection 11](#_Toc60848113)

[Backward stepwise selection 14](#_Toc60848114)

[Section 3.2 - Model1| Using predictors from Stepwise forward selection 16](#_Toc60848115)

[Report 16](#_Toc60848116)

[Fit and Usefulness of the model 16](#_Toc60848117)

[Predictors and their significance 16](#_Toc60848118)

[Assess the model against assumptions 16](#_Toc60848119)

[Model Equation and Illustration 16](#_Toc60848120)

[Section 3.3 - Model 2| Using predictors from Stepwise backward selection 16](#_Toc60848121)

[Report 16](#_Toc60848122)

[Fit and Usefulness of the model 17](#_Toc60848123)

[Predictors and their significance 17](#_Toc60848124)

[Assess the model against assumptions 17](#_Toc60848125)

[Model Equation and Illustration 17](#_Toc60848126)

[Section 4 – Discussion/Conclusion 17](#_Toc60848127)

[PLEASE NOTE: 17](#_Toc60848128)

# Abstract

We propose to do a logistic regression approach to understand the attributes which influences the decision for subscribing to a bank long term deposit and predict the success of telemarketing calls for selling bank long term deposit. We are given the data of direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (target variable y).

The dataset consists of direct marketing campaigns data of a banking institution. The dataset was picked from UCI Machine Learning Repository which is an amazing source for publicly available datasets. There were four variants of the datasets out of which we chose “bank-additional-full.csv” which consists of 41188 data points with 20 independent variables out of which 10 are numeric features and 10 are categorical features. The data is collected during the period of May 2008 and November 2010. The list of features available to us is given below:

**Bank client data:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Feature type** | **Description** |
| age | numeric | Age of the customer |
| job | categorical | Type of job (admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown) |
| marital | categorical | Marital status |
| education | categorical | Education details (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown) |
| default | categorical | Has credit in default? (no, yes, unknown) |
| housing | categorical | Has housing loan? (no, yes, unknown) |
| loan | categorical | Has personal loan? (no, yes, unknown) |

**Related with the last contact of the current campaign:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Feature type** | **Description** |
| contact | categorical | Contact communication type (cellular, telephone) |
| month | categorical | Last contact month of year (jan, feb, mar etc.) |
| day\_of\_week | categorical | Last contact day of the week (mon, tue, wed, thu, fri) |
| duration | numeric | Last contact duration, in seconds |

**Previous campaign attributes:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Feature type** | **Description** |
| campaign | numeric | Number of contacts performed during this campaign and for this client |
| pdays | numeric | Number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted) |
| previous | categorical | Number of contacts performed before this campaign and for this client |
| poutcome | categorical | Outcome of the previous marketing campaign ( failure, nonexistent, success) |

**Social and Economic context attributes:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Feature type** | **Description** |
| emp.var.rate | numeric | employment variation rate — quarterly indicator |
| cons.price.idx | numeric | consumer price index — monthly indicator |
| cons.conf.idx | numeric | consumer confidence index — monthly indicator |
| euribor3m | numeric | euribor 3 month rate — daily indicator |
| nr.employed | numeric | number of employees — quarterly indicator |

# Section 1 - Research Question(s)

## Question-1

Do the social and economic indicators, client data and previous campaign attributes have an influence on a customer subscribing to a term deposit subscription?

## Question-2

What is the probability of a customer subscribing to a term deposit subscription depends on social and economic indicators, client data, and previous campaign attributes?

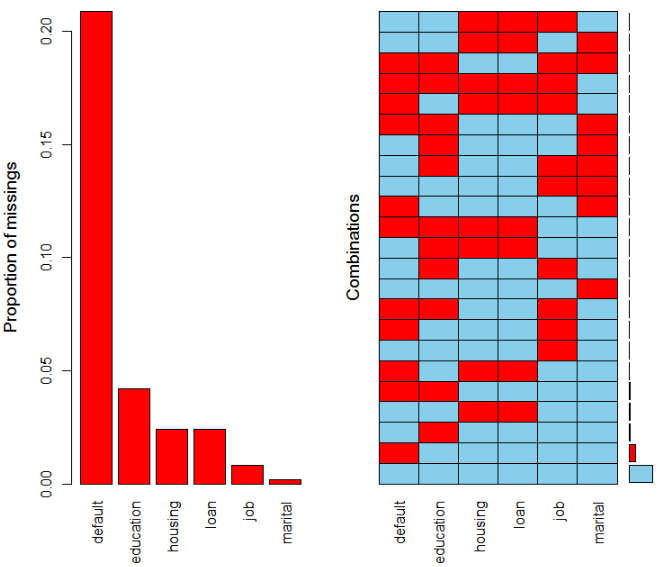
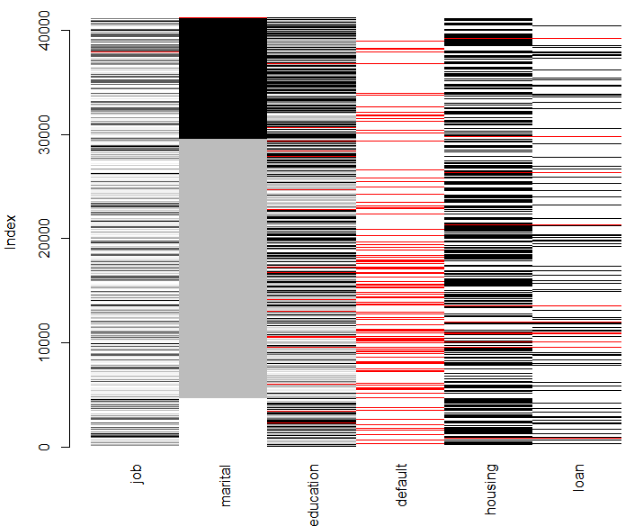
# Section 2 - Dataset

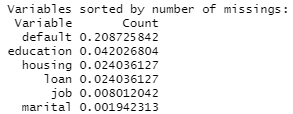
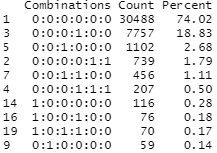
## How to handle missing data

The missing data are present in then below columns and marked as ‘unknown’:

1. job
2. marital
3. education
4. default
5. housing
6. loan

We have replaced the ‘unknown’ with NA and find out the pattern in missing data. Below are the test results:



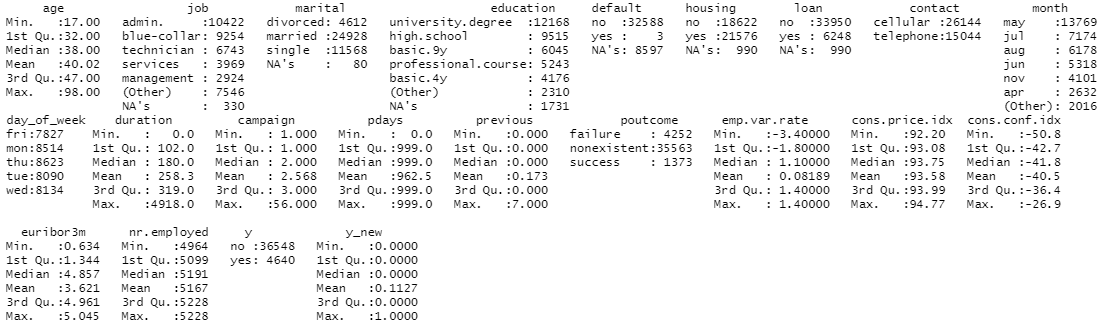


### Findings & Actions

* We can see the most common pattern is where all variables are having data - no missing data (74.02%). The next most common is where ‘default’ alone is missing - 7757 cases. Then when the ‘education’ alone is missing – 1102.
* The ‘default’ column is having a much skewed data and more than 20% value missing. So this variable is dropped/not used in the upcoming models.



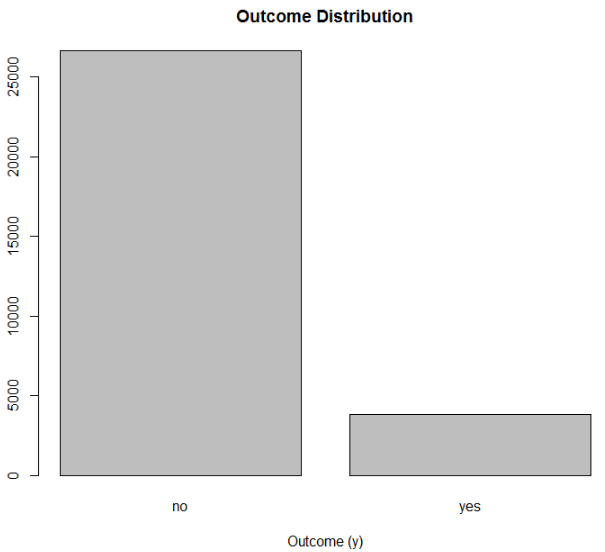
* When ‘housing’ is missing the ‘loan’ column is missing and vice versa. There are total 990 such records present in the data. We took the summary of the whole data set when housing and loan is missing and found the missingness is not dependent on any other variable.



* We will ignore all these missing data as it is missing completely at random and delete them from the dataset.

## Representativeness

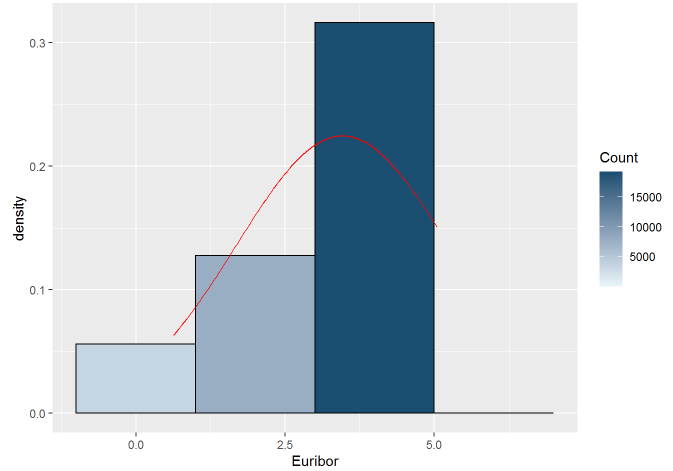
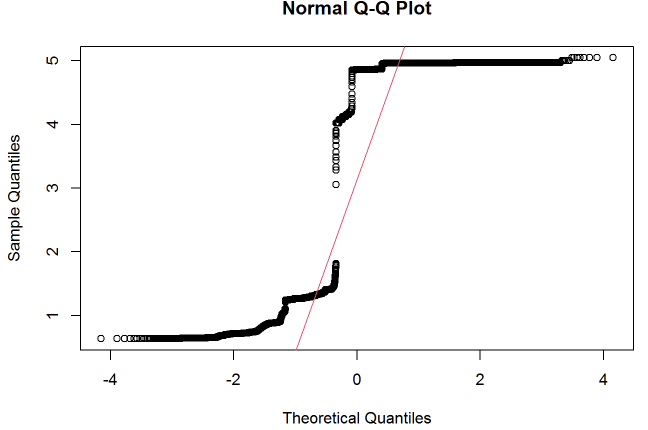
The study considers real data collected from a Portuguese retail bank from May 2008 to November 2010. As there was a global recession going on it is quite obvious that most of the people need money in hand and as a result most of the customers did not opt for the term deposit. So, the dataset is unbalanced, after removing the missing data only 3859 (12.65%) records are related to success.



## Variables of interest and their statistical measurement type

### Euribor3m

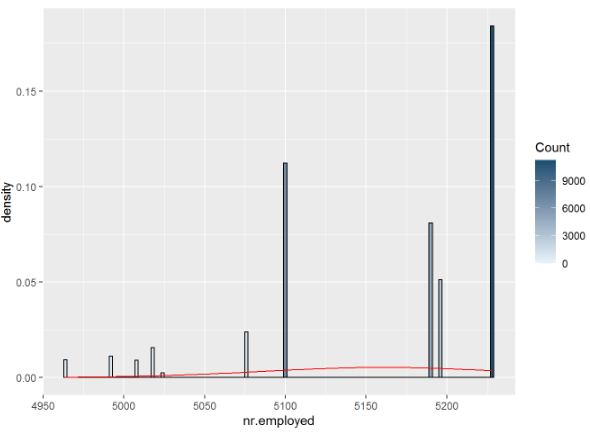
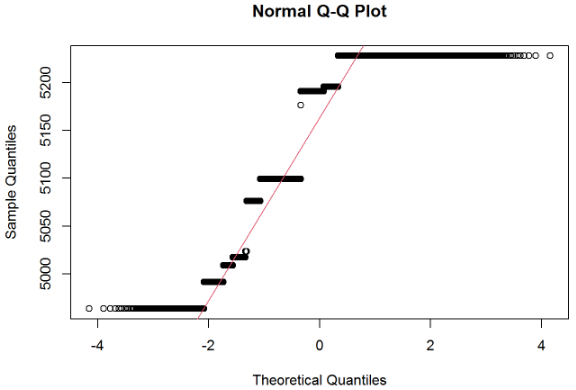
This is a social economic numeric variable. We will check the normality of the variable.

Euribor3m score data was assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 1 and Figure 2) identified some issues with skewness and kurtosis. The standardised score for kurtosis (-58.17) and the standardised score for skewness (-37.43) were outside the acceptable range using the criteria proposed by West, Finch and Curran (1996). However 100% of standardised Euribor3m fall within the bounds of +/- 3.29, using the guidance of Field, Miles and Field (2013) the data can be considered to approximate a normal distribution (m=3.46, sd=1.78, n=30488).

### nr.employed

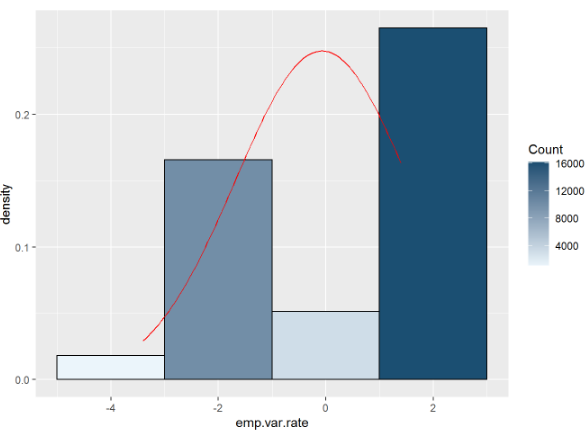
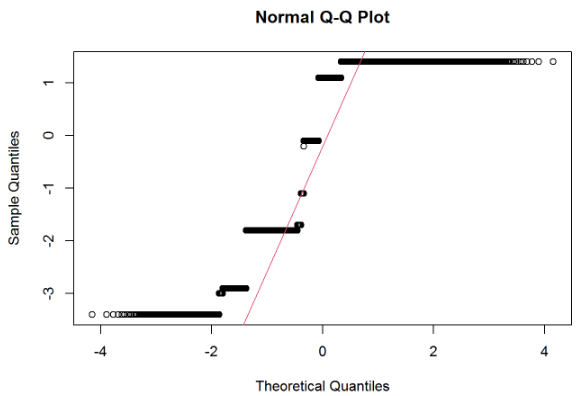
This is a social economic numeric variable. We will check the normality of the variable.

nr.employed score data was assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 1 and Figure 2) identified some issues with skewness and kurtosis. The standardised score for kurtosis (-12.53) and the standardised score for skewness (-63.71) were outside the acceptable range using the criteria proposed by West, Finch and Curran (1996). However 100% of standardised nr.employed fall within the bounds of +/- 3.29, using the guidance of Field, Miles and Field (2013) the data can be considered to approximate a normal distribution (m=5160.81, sd=75.16, n=30488).

### emp.var.rate

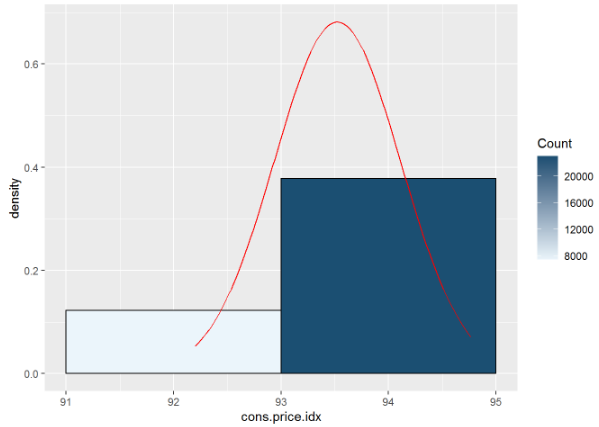
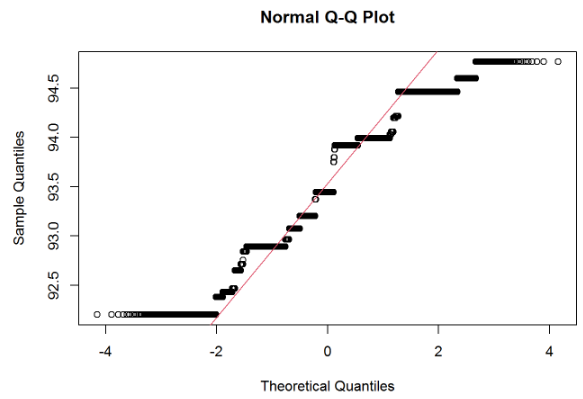
This is a social economic numeric variable. We will check the normality of the variable.

emp.var.rate score data was assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 1 and Figure 2) identified some issues with skewness and kurtosis. The standardised score for kurtosis (-45.32) and the standardised score for skewness (-39.13) were outside the acceptable range using the criteria proposed by West, Finch and Curran (1996). However 100% of standardised emp.var.rate fall within the bounds of +/- 3.29, using the guidance of Field, Miles and Field (2013) the data can be considered to approximate a normal distribution (m=-0.07, sd=1.61, n=30488).

### cons.price.idx

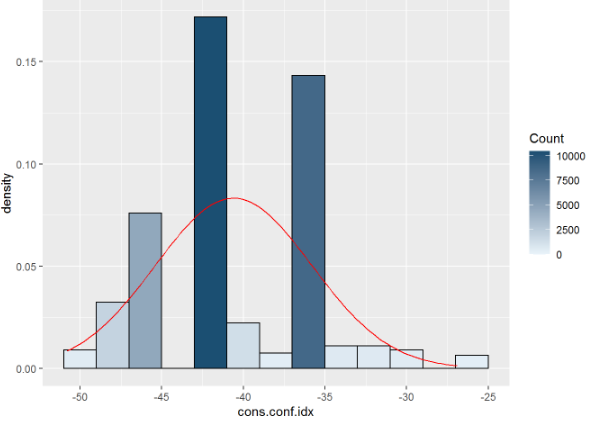
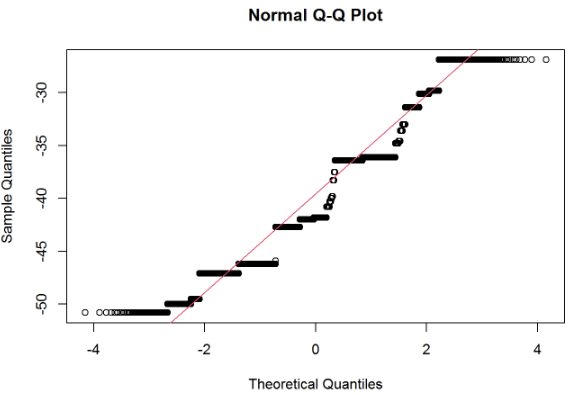
This is a social economic numeric variable. We will check the normality of the variable.

cons.price.idx score data was assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 1 and Figure 2) identified some issues with skewness and kurtosis. The standardised score for kurtosis (-30.77) and the standardised score for skewness (-8.46) were outside the acceptable range using the criteria proposed by West, Finch and Curran (1996). However 100% of standardised cons.price.idx fall within the bounds of +/- 3.29, using the guidance of Field, Miles and Field (2013) the data can be considered to approximate a normal distribution (m=93.52, sd=0.59, n=30488).

### cons.conf.idx

This is a social economic numeric variable. We will check the normality of the variable.

cons.conf.idx score data was assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 1 and Figure 2) identified some issues with skewness and kurtosis. The standardised score for kurtosis (-11.53) and the standardised score for skewness (26.66) were outside the acceptable range using the criteria proposed by West, Finch and Curran (1996). However 100% of standardised cons.conf.idx fall within the bounds of +/- 3.29, using the guidance of Field, Miles and Field (2013) the data can be considered to approximate a normal distribution (m=-40.6, sd=4.79, n=30488).

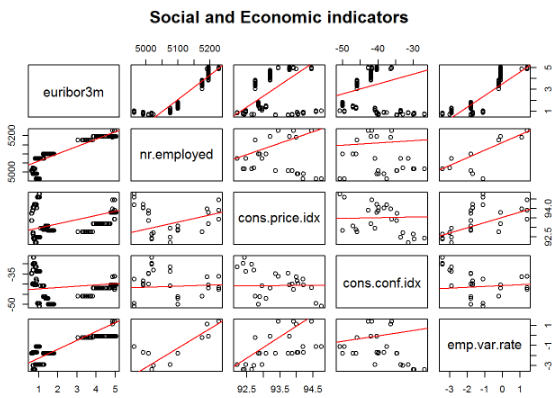
# Section 3 - Results

## Section 3.1 - Statistical Evidence

* The last call duration should be removed while creating the models as this data can only be collected once the call is over and by then the outcome is already known. So, to predict the outcome call duration can’t be used.
* As discussed in the missing data analysis section, ‘default’ (Has a credit in default?) column is having a much skewed data and more than 20% value missing. So this variable is dropped/not used in the upcoming models.



* All the social economic attributes are highly correlated as all of them are trying to narrate the current economic situation. As per one of the assumption of the logistic regression, the independent variables should not be too highly correlated to each other. We conducted Pearson correlation test to examine the correlation between the social economic attributes. Pearson’s test is chosen as in the previous section we have already found out the entire social and economic variables are normally distributed.



* First we created the correlation matrix for all the social economic attributes and also created the scatter plots. Then correlation tests are performed between the number of employees per quarter and all other social economic attributes.
* The relationship between number of employees per quarter and euribor 3 months rate was investigated using a Pearson correlation. A strong positive correlation was found (r =0.945, n=30486, p<.001).
* The relationship between number of employees per quarter and consumer pricing index monthly rate was investigated using a Pearson correlation. A strong positive correlation was found (r =0.489, n=30486, p<.001).
* The relationship between number of employees per quarter and consumer confidence index monthly rate was investigated using a Pearson correlation. A weak positive correlation was found (r =0.075, n=30486, p<.001).
* The relationship between number of employees per quarter and employment quarterly variation rate was investigated using a Pearson correlation. A strong positive correlation was found (r =0.900, n=30486, p<.001).
* Only number of employees per quarter and consumer confidence index monthly rate social economic attributes will be used for the model creation as they are weakly correlated. All the other social economic attributes are strongly correlated with number of employees per quarter, so those attributes will not be used in model building.
* All the below attributes will be used to create the model:
  1. Number of employees per quarter
  2. Consumer confidence index monthly rate
  3. Age of the customer
  4. Job of the customer
  5. Marital status of the customer
  6. Education of the customer
  7. Housing loan status
  8. Personal Loan status
  9. Customer communication type
  10. Last contact month
  11. Last contact day of the week
  12. Number of contacts performed during this campaign and for this client
  13. Number of days that passed by after the client was last contacted from a previous campaign
  14. Number of contacts performed before this campaign and for this client
  15. Outcome of the previous marketing campaign
* Below two different variable section methods are used two narrow down from the 15 attributes mentioned above to the most statistically significant attributes which will provide the better fit and efficient model.
  1. Stepwise forward selection
  2. Stepwise backward selection

### Forward stepwise selection

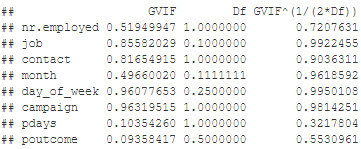
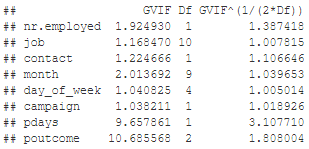
In this selection method, we pick one attribute from the above list of 15 variables and create a logistic regression mode. We check the statistical significance of the attribute for the prediction and performance of the model to decide whether the variable is helping the prediction or should be removed from the model. By applying this technique we create below 16 models and compared them by Predictor probability, AIC and AUC and choose the final list of variables and optimum model to predict the outcome variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model #** | **Predictor added** | **Pr(>|z|)** | **AIC** | **AUC** | **Conclusion** |
| 1 | Number of employees per quarter | <0.001 | 19532 | 0.758 | Predictor selected. |
| 2 | Consumer confidence index monthly rate | <0.001 | 19470 | 0.748 | Predictor rejected as AUC dropped significantly (0.01). |
| 3 | Age of the customer | <0.05 | 19526 | 0.756 | Predictor rejected as AUC dropped and very small improvement in AIC. |
| 4 | Job of the customer | <0.05 for 50% of categories | 19421 | 0.763 | Predictor selected. Rise in AUC and decrease in AIC. |
| 5 | Marital status of the customer | 0.44 for married  0.12 for single | 19422 | 0.764 | Predictor rejected as Probability >0.05 and increase in AIC. |
| 6 | Education of the customer | 0.59 for basic.6y  0.41 for basic.9y  0.79 for high.school  0.07 for illiterate  0.51 for professional.course  <0.05 for university.degree | 19410 | 0.766 | Predictor rejected as more than 80% of the categories are having Probability >0.05 |
| 7 | Housing loan status | 0.38 | 19422 | 0.763 | Predictor rejected as Probability >0.05 and increase in AIC. |
| 8 | Personal loan status | 0.52 | 19422 | 0.763 | Predictor rejected as Probability >0.05 and increase in AIC. |
| 9 | Customer communication type | <0.001 | 19292 | 0.774 | Predictor selected. Rise in AUC and decrease in AIC. |
| 10 | Last contact month | <0.05 for 77% of categories | 18835 | 0.786 | Predictor selected. Rise in AUC and decrease in AIC. |
| 11 | Last contact day of the week | <0.05 for 75% of category | 18789 | 0.788 | Predictor selected. Rise in AUC and decrease in AIC. |
| 12 | Number of contacts performed during this campaign and for this client | <0.001 | 18767 | 0.789 | Predictor selected. Rise in AUC and decrease in AIC. |
| 13 | Number of days that passed by after the client was last contacted from a previous campaign | <0.001 | 18351 | 0.795 | Predictor selected. Rise in AUC and decrease in AIC. |
| 14 | Number of contacts performed before this campaign and for this client | <0.001 | 18285 | 0.797 | Predictor selected. Rise in AUC and decrease in AIC. |
| 15 (a) | Outcome of the previous marketing campaign | <0.001 | 18258 | 0.797 | Predictor selected. Decrease in AIC. |
| 15 (b) | Number of contacts performed before this campaign and for this client | 0.205 | 18258 | 0.797 | Predictor rejected as Probability >0.05 |
| 16 | Removed: Number of contacts performed before this campaign and for this client | NA | 18258 | 0.797 | Predictor removed as per the conclusion drawn from 15 (b). AIC and AUC remain unchanged. |

Following are the list of attributes which are selected in the forward stepwise selection:

1. Number of employees per quarter
2. Job of the customer
3. Customer communication type
4. Last contact month
5. Last contact day of the week
6. Number of contacts performed during this campaign and for this client
7. Number of days that passed by after the client was last contacted from a previous campaign
8. Outcome of the previous marketing campaign

We created a model with above attributes. Examination for multicollinearity showed that the tolerance and variance influence factor measures for ‘Number of days that passed by after the client was last contacted from a previous campaign’ (pdays) predictor were not within acceptable levels (tolerance >0.4, VIF <2.5 ) as outlined in Tarling (2008). So, ‘Number of days that passed by after the client was last contacted from a previous campaign’ (pdays) is removed from the final model.



### Backward stepwise selection

In this selection method, we pick all 15 attributes and create a logistic regression mode. We remove all the attributes which are statistically not significant in the model (Probability >0.05). By applying this technique we create below 2 models and compared them by Predictor probability, AIC and AUC and choose the final list of variables and optimum model to predict the outcome variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model #** | **Predictors with Pr(>|z|) >0.05** | **Pr(>|z|)** | **AIC** | **AUC** | **Conclusion** |
| 1 | Age of the customer | 0.869 | 18256 | 0.797 | Predictor rejected as Probability >0.05 |
| Job of the customer | blue-collar <0.05  entrepreneur 0.45  housemaid 0.92  management 0.60  retired <0.05  self-employed 0.73  services 0.06  student <0.05  technician 0.78  unemployed 0.82 | Predictor rejected as Probability >0.05 for 70% categories. |
| Marital status of the customer | Married 0.62  Single 0.54 | Predictor rejected as Probability >0.05 |
| Education of the customer | basic.6y 0.28  basic.9y 0.84  high.school 0.33  illiterate 0.11  professional.course 0.43  university.degree 0.08 | Predictor rejected as Probability >0.05 |
| Housing loan status | 0.26 | Predictor rejected as Probability >0.05 |
| Personal loan status | 0.66 | Predictor rejected as Probability >0.05 |
| Number of contacts performed before this campaign and for this client | 0.33 | Predictor rejected as Probability >0.05 |
| 2 | All attributes except the once rejected in the above model | <0.05 | 18260 | 0.796 | Predictors selected as probability <0.05 and very small change in AIC and AUC. |

Following are the list of attributes which are selected in the backward stepwise selection:

1. Number of employees per quarter
2. Consumer confidence index monthly rate
3. Customer communication type
4. Last contact month
5. Last contact day of the week
6. Number of contacts performed during this campaign and for this client
7. Number of days that passed by after the client was last contacted from a previous campaign
8. Outcome of the previous marketing campaign

We created a model with above attributes. Examination for multicollinearity showed that the tolerance and variance influence factor measures for ‘Number of days that passed by after the client was last contacted from a previous campaign’ (pdays) predictor were not within acceptable levels (tolerance >0.4, VIF <2.5 ) as outlined in Tarling (2008). So, ‘Number of days that passed by after the client was last contacted from a previous campaign’ (pdays) is removed from the final model.

## Section 3.2 - Model1| Using predictors from Stepwise forward selection

### Report

A binomial logistic regression analysis was conducted to know the outcome of a bank telemarketing campaign using Number of employees per quarter, Job of the customer, Customer communication type, Last contact month, Last contact day of the week, Number of contacts performed during this campaign and for this client and Outcome of the previous marketing campaign as predictors.

The data met the assumption for independent observations. Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (tolerance >0.4, VIF <2.5) as outlined in Tarling (2008). The Hosmer Lemeshow goodness of fit statistic did not indicate any issues with the assumption of linearity between the independent variables and the log odds of the model (χ2(n=8)= =48.538, p <0.001).

### Fit and Usefulness of the model

* AIC, AUC, ROC Curve
* The model is significant Chi sq= 4962.4, p<0.01. We have calculated the pseudo R Square statistics using Cox and Snell and Nagelkerke formula. The pseudo R square indicate that between 15.02% and 28.22% of the variability of responses to whether the customer opts for the term deposit or not.

### Predictors and their significance

* Log odds
* Predictors and their probability

### Assess the model against assumptions

* Requires the dependent variable to be binary/nominal with multiple categories
* ordinal logistic regression requires the dependent variable to be ordinal.
* Requires the observations to be independent of each other.
* Observations should not come from repeated measurements or matched data.
* Requires there to be little or no multicollinearity among the independent variables.
* Independent variables should not be too highly correlated with each other.
* Requires linear relationship between the log odds and the predictors.

### Model Equation and Illustration

Equation and one illustration

## Section 3.3 - Model 2| Using predictors from Stepwise backward selection

### Report

A binomial logistic regression analysis was conducted to know the outcome of a bank telemarketing campaign using Number of employees per quarter, Job of the customer, Customer communication type, Last contact month, Last contact day of the week, Number of contacts performed during this campaign and for this client and Outcome of the previous marketing campaign as predictors.

The data met the assumption for independent observations. Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (tolerance >0.4, VIF <2.5) as outlined in Tarling (2008). The Hosmer Lemeshow goodness of fit statistic did not indicate any issues with the assumption of linearity between the independent variables and the log odds of the model (χ2(n=8)= =48.538, p <0.001).

### Fit and Usefulness of the model

* AIC, AUC, ROC Curve
* The model is significant Chi sq= 4962.4, p<0.01. We have calculated the pseudo R Square statistics using Cox and Snell and Nagelkerke formula. The pseudo R square indicate that between 15.02% and 28.22% of the variability of responses to whether the customer opts for the term deposit or not.

### Predictors and their significance

* Log odds
* Predictors and their probability

### Assess the model against assumptions

* Requires the dependent variable to be binary/nominal with multiple categories
* ordinal logistic regression requires the dependent variable to be ordinal.
* Requires the observations to be independent of each other.
* Observations should not come from repeated measurements or matched data.
* Requires there to be little or no multicollinearity among the independent variables.
* Independent variables should not be too highly correlated with each other.
* Requires linear relationship between the log odds and the predictors.

### Model Equation and Illustration

Equation and one illustration

# Section 4 – Discussion/Conclusion

In this section you should reflect on your results from the perspective of your research question(s). You should also suggest some changes/additional research that could be conducted to try to better answer this research question.

# PLEASE NOTE:

You are required to provide R Code (either .R or .RMD) which will generate the statistics/visuals to which you refer in this report as part of your submission. This should be appropriately commented so that it is easy to locate the code that generates the statistics/visuals for each section. It would be really helpful if you included the output generated by this R code with your submission.