# 

# 

# **TERM DEPOSIT PREDICTION**

# By

Manaswini Rayappureddi

Table of Contents

[**TERM DEPOSIT PREDICTION** 1](#_Toc476769507)

[Executive Summary: 4](#_Toc476769509)

[Data Cleaning: 4](#_Toc476769510)

[Validating Data/Modeling types: 5](#_Toc476769511)

[Excluding Redundant variables: 5](#_Toc476769512)

[Resolving Inconsistencies: 5](#_Toc476769513)

[Treating Missing Values: 5](#_Toc476769514)

[Standardizing Variables: 6](#_Toc476769515)

[Partitioning the Dataset (Training, Validation and Testing): 6](#_Toc476769516)

[Training dataset: 6](#_Toc476769517)

[Validation dataset: 6](#_Toc476769518)

[Testing dataset: 6](#_Toc476769519)

[Modeling: 6](#_Toc476769520)

[Nominal Logistic Regression: 7](#_Toc476769521)

[Clustering 8](#_Toc476769522)

[Neural Networks: 8](#_Toc476769523)

[Recommendations 9](#_Toc476769524)

[Appendix 1: Calculation of Accuracy 10](#_Toc476769525)

[Logistic regression 11](#_Toc476769526)

[Neural networks 11](#_Toc476769527)

[Decision tree: 12](#_Toc476769528)

[Discriminant Analysis: 13](#_Toc476769529)

[Ensemble model 14](#_Toc476769530)

[Appendix 2: Modelling Technique 15](#_Toc476769531)

[Nominal Logistic: 15](#_Toc476769532)

[Neural Networks 18](#_Toc476769533)

[Discriminant Analysis 22](#_Toc476769534)

[Decision Tree 24](#_Toc476769535)

[Ensemble Model of DT, NN, LR, DA using LR: 27](#_Toc476769536)

[References: 29](#_Toc476769537)

# Executive Summary:

Direct marketing campaigns target clients/customers of specific segments by contacting them through various channels to meet a specific goal. The dataset used in this project is related to direct marketing campaigns of a Portuguese retail bank, where the campaigns were based on phone calls. The goal in this project is to predict whether the client will subscribe for the term deposit or not, so that the campaign team can target the right clients to maximize the number of term deposits. This prediction would also help the bank to reduce the cost that would incur for targeting wrong clients.

The dataset is chosen from University of California – Irvine (UCI), machine learning repository. The data set has 41,189 observations with 21 variables which indicate various attributes of a client such as age, job, loan details, previous campaign details and their socio economic status like employment variation rate, consumer price index etc. More information about the attributes can be found in the data dictionary.

Using the popular SEMMA approach, we have explored, cleansed the data, prepared a ready-to-model dataset to perform modelling and the results were analyzed. The cleansed dataset was split into three partitions for training, validating and finally testing the model. This is a binary classification problem, to predict the likelihood of customers’ subscription of term deposit. We chose Decision trees, Logistic Regression, Discriminant Analysis, Neural Networks, and Ensemble Modeling techniques. The performance metrics such as the accuracy of the model, benefit gained by targeting the good customers who will subscribe for term deposits and cost incurred when customer doesn’t subscribe for the term deposit (bad customer) are used as decisive parameters in selecting a model. With an accuracy of 81%, Logistic regression modeling performed well in predicting the customers who are likely to subscribe for a term deposit**.** It was followed by neural networks with an accuracy of 80%. However,selecting a model also depends on benefit and costs incurred for each model i.e. business cost of the misclassification.Performance metrics of various models are discussed in further sections of the report.

Further analysis of the results resulted in a few insights. The bank can actually target a particular age group, job positions to make the campaign more effective. For example, targeting students (term deposit acceptance rate: 31%) instead of customers with other professions, whose acceptance rate is mere 5%-8% would make the marketing campaign more target oriented thereby ensuring more term deposit rate.

# Data Cleaning:

Before performing data modeling techniques, the team checked for various types of data anomalies like incorrect datatypes, inconsistences in the data, missing observations, outliers and dimensionality. Team also standardized the dataset as performing modeling on inconsistent data might lead to wrong conclusions and undermine the analysis.

Following data cleansing steps have been performed to make sure that the data is ready for modeling:

### Validating Data/Modeling types:

Dataset has 21 variables (including target variable). Team considered 10 variables as continuous datatype, because these variables contain numeric data. Remaining 11 variables were considered as nominal datatype, because these variables had character data.

### Excluding Redundant variables:

The variable **“pdays”** (number of days that passed by after the client was last contacted from a previous campaign) was strongly correlated (correlation: 0.9) to **“previous”** variable (number of contacts performed before this campaign and for this client) so, we decided to exclude **“pdays”** variable from our analysis.

The attribute “**duration**” describes last call duration, in seconds. This attribute highly affects the output target (e.g., if duration=0 then target variable (y) = no i.e., the campaign team did not contact the client, therefore we don’t expect the client to subscribe for term deposit). The duration of the call is not known while predicting whether a client subscribes for a term deposit before the campaign team contacts that specific client. Thus, this variable should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. We excluded this variable and did not delete it from the dataset as it can be found useful in future.

### Resolving Inconsistencies:

The inconsistencies in the data were treated in order to bring the uniformity in the dataset. We found that the columns “**Education**” and “**Job**” had values with special characters and resolved these inconsistencies.

### Treating Missing Values:

Missing values occur when there is no data stored in a variable. There were a few missing values in this dataset. We decided to impute some of the missing values based on the patterns or relationships found between relevant variables and deleted some of the observations. However the observations can be chosen to be not excluded if the information in other variables is important.We found that for some observations, the variable “Education” is missing. We imputed the missing values by making use of the other variable “Job”. Education for an “Admin” was university degree for all the non-missing cases, hence we imputed university degree as the education for the missing cases.

### Standardizing Variables:

This is the process of bringing all the variables onto the same scale to avoid domination of variable with larger scale. The team standardized the variables **“campaign”** (number of contacts performed during this campaign and for this client), **“previous** (number of contacts performed before this campaign and for this client), **“emp var rate”** (employment variation rate), **“cons price idx”** (consumer price index), **“cons conf idx”** (consumer confidence), **“euribor 3m rate”** (euribor 3month rate), **“nr employed”** (number of employees) and created new variables with standard extension to the original variables. All continuous variables are scaled in the range of 0 to 1 Standardized variables are used in place of original variables for modeling.

# Partitioning the Dataset (Training, Validation and Testing):

After performing data cleansing steps, we retained 19 variables with 31,269 observations. About 14.8%(4,640) of the observations are with “Yes” response and remaining 85.2%(26,629) of the observations are with “No” response. Then we partitioned the main dataset into training, validation datasets (balanced) and testing dataset (unbalanced). Balancing a dataset is done with respect to the target variable.

Training dataset: Training dataset is used to build and train the models. A balanced training dataset with 3898 observations is taken to train the model.

Validation dataset:Validation dataset is used to assess the performance of the models built on training data. This gives us the option to go back and tweak our models, so as to make a model perform better with testing or unseen data. A balanced validation dataset with 3,900 observations has been used to validate the model.

Testing dataset:Testing dataset is used to find the performance of the models when they are used in real world. These data are generally unseen by the models and they closely resemble real world conditions. An unbalanced testing dataset with 5,000 random observations has been used to test the model.

# Modeling:

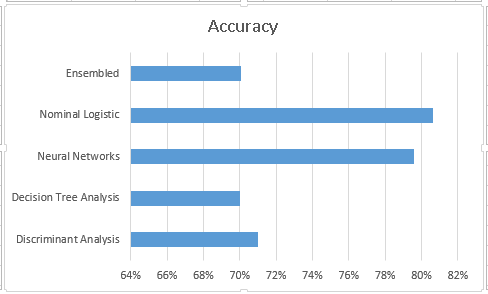
Classification models such as Logistic Regression (LR), Decision Trees (DT), Discriminative Analysis (DA), Neural Networks (NN) and Ensemble Modeling (EM) have been used in predicating whether a client subscribes for a term deposit or not, so that the campaign team can contact the client. Simple models such as Logistic Regression (LR) are less flexible but are easy to interpret the model, whereas models such as Neural Networks (NN) are more flexible when compared to other simple models but are difficult to understand and hard to draw inferences from them. And an ensemble model was built to improve the accuracy of the predictive classification.

During the modeling phase, along with the goal of predicting the target variable, the team has decided to use other parameters (confusion matrix) such as false negatives and false positives. False negatives indicate the cases where the model predicts that client will not subscribe for the term deposit, but in reality client will subscribe. Since the model has predicted that these clients will not subscribe, the bank will not contact such clients and may lose their term deposits. Minimizing false positives is necessary to improve the accuracy of the model.

On the other hand, false positives indicate the cases where the model predicts that client will subscribe for the term deposit, but in reality they will not subscribe. In this scenario, the bank bears the opportunity costs of wrongly investing its limited resources in inappropriate customers. It is also important to reduce this parameter to the save the cost.

After validating various modeling techniques, the team found out that logistic regression performed with an accuracy of 81% followed by neural networks with 80%. Accuracy calculations can be found in appendix 1.

Accuracy of various models are depicted in the below figure:



## Nominal Logistic Regression:

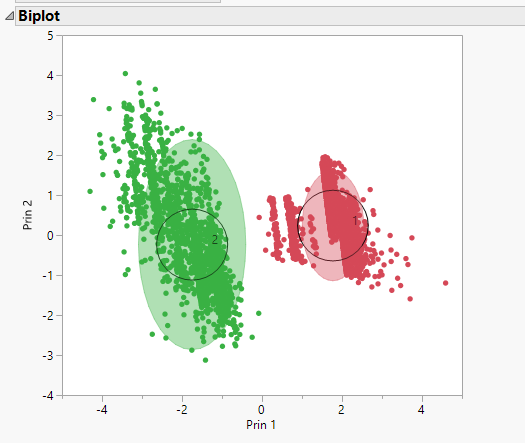
Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.

A Logistic Regression model with 19 variables resulted in a model with a AIC - 4398.95, BIC - 4686.16 values. Removing the insignificant variables like loan status, marital status and age resulted in a better model with AIC - 4390.58, BIC - 4603.08 values. AIC (Akakie Information criterion) BIC (Bayesian information criterion) are the measure of relative quality of statistical models. Complete list of insignificant variables and formulae for AIC and BIC can be found in the appendix.

By interpreting the results of logistic regression, the model shows 5.26% false negatives and 14.10% false positives. For the rest of 80.64% the model predicts exactly whether the client subscribes or not. The model classified 1183 clients would subscribe for term deposit out of which only 478 subscribe in reality, making 705 clients as bad clients.

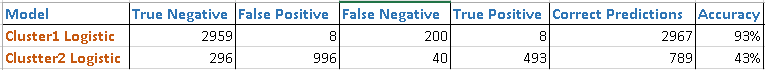
The bank has to check for: 478(benefit of each good client) and 705(cost of each bad client) to compare them with different models. Calculations of performance metrics can be found in appendix 1**.**

Clustering - Clustering is performed to find out if there are any significant groups/patterns in the data so that models can be performed for each cluster.



The Biplot shows that there are two different clusters, therefore two different logistic regression models are built for two clusters.

The following are the results when models were built for each cluster.



The model for cluster1 has a very high accuracy of 93%. However, the model for cluster 2 has an accuracy of 43%. The bank has to check for the benefit of good customers and cost of bad customers to compare with other models.

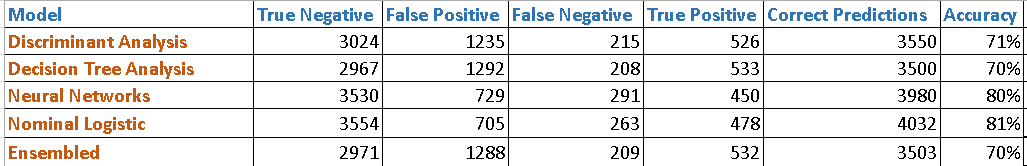
## Neural Networks:

Neural Networks is an information processing paradigm inspired by the way biological nervous systems, such as the brain, process the information. In this project, we used neural network with 2 layers, 3 nodes in the first layer and 2 nodes in the second layer. We found out Neural networks performed with an accuracy of 80%.

The model classified 1179 clients would subscribe for term deposit out of which only 450 subscribe in reality making 729 people as bad clients

The bank has to check for: 450 (benefit of each good client) and 729(cost of each bad client) to compare them with different models. Calculations of performance metrics can be found in appendix 1**.**

Comparision of various metrics of each model:

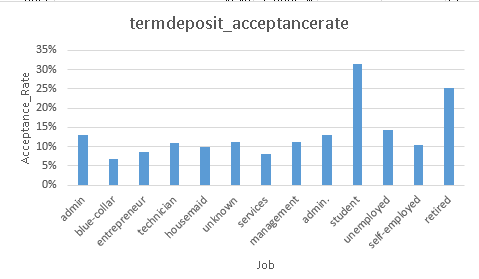


Recommendations:

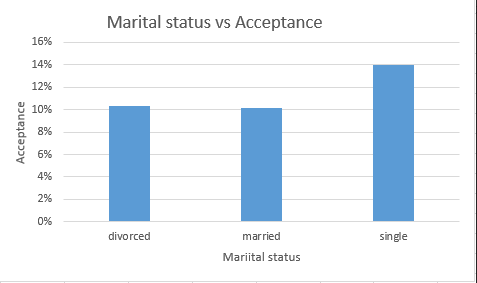
During the data cleansing and modeling phase, team found some key insights from the dataset which can be used for the target marketing strategies i.e. targeting the right clients. Analyzing the given datasets provided the below insights.

* Out of total students contacted, 31% of them subscribe for a term deposit. The bank can target more students as they have a high acceptance rate comparted to clients who belong to other job categories.
* Students are followed by retired clients in accepting the term deposits with an acceptance rate of 25%. The below figure shows the acceptance rates with respect to various job.

To ensure a more target oriented marketing, the bank can focus on students and retired employee job categories to achieve more success rate.



* The acceptance rate of term deposits for clients whose marital status is “single” is 14% in comparison with other categories which is 10% each. Bank can actually target more singles. Below figure displays the percentage of acceptance with respect to marital status of clients.



# 

# Appendix 1: Calculation of Accuracy

Formula for calculating accuracy:

## Logistic regression

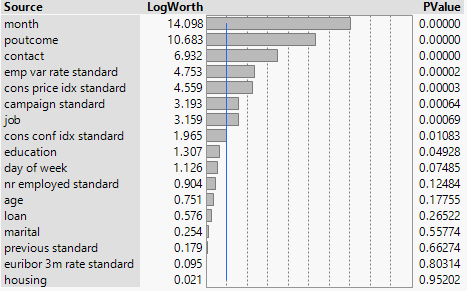


Fig: List of insignificant variables which were excluded during Logistic regression modeling.

**Testing data Calculations:**

Number of correct predictions: 3554+478 = 4032

Total number of observations = 5000

Number of clients predicted by model who will subscribe for term deposit = 705 + 478 = 1183

Accuracy of model in predicting clients who will subscribe for term deposit = 40.01 %

Overall accuracy of the model = 80.64%

Baseline accuracy: 14.82%

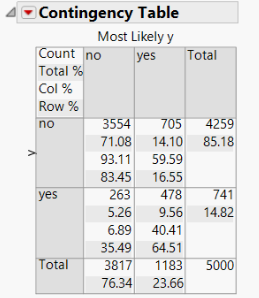


Fig. Contingency table for logistic regression.

## Neural networks

Number of correct predictions: 3530+450=3980

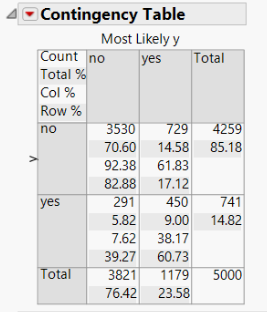
Total number of observations = 5000

Number of clients predicted by model who will subscribe for term deposit = 729 + 450 = 1179

Accuracy of model in predicting clients who will subscribe for term deposit = 38.16%

Overall accuracy of the model = 79.6%

Baseline accuracy: 14.82%

  
Fig. Contingency table for Neural networks.

## Decision tree:

Number of correct predictions : 2967 + 533= 3500

Total number of observations = 5000

Number of clients predicted by model who will subscribe for term deposit = 1292 + 533 = 1825

Accuracy of model in predicting clients who will subscribe for term deposit = 29.2%

Overall accuracy of the model = 70.0%

Baseline accuracy: 14.82%

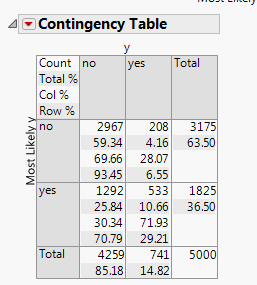


Fig. Contingency table for Decision tree.

## Discriminant Analysis:

Number of correct predictions : 3024+526 = 3550

Total number of observations = 5000

Number of clients predicted by model who will subscribe for term deposit = 1235+526=1761

Accuracy of model in predicting clients who will subscribe for term deposit = 29.87%.

Overall accuracy of the model = 71.0%

Baseline accuracy: 14.82%

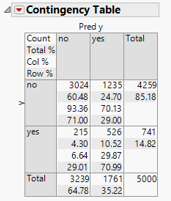
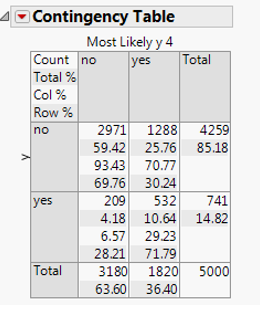


Fig. Contingency table for Discriminant Analysis.

Ensemble model:



Number of correct predictions: 2971+532 = 3503

Total number of observations = 5000

Number of clients predicted by model who will subscribe for term deposit = 1288+532=1820

Accuracy of model in predicting clients who will subscribe for term deposit = 29.23%.

Overall accuracy of the model = 70%

Baseline accuracy: 14.82%

|  |
| --- |
| Appendix 2: Modelling TechniqueNominal Logistic:   Fig: Confusion matrix of training Data in nominal logistic    Fig: Contingency table of validation Data in nominal logistic    Fig: Contingency table of testing Data in nominal logistic Neural Networks   Fig : Neural Networks diagram    Fig : Confusion Matrix of Neural network Training Data    Fig : Contingency table of Neural network Validation Data  Fig : Contingency table of Neural network Testing Data Discriminant Analysis   Fig: Discriminant Analysis on Training data    Fig: Contingency table of Validation data    Fig: Contingency table fTesting data Decision Tree A decision tree can be used as a model for a sequential decision problems under uncertainty. A decision tree describes graphically the decisions to be made, the events that may occur, and the outcomes associated with combinations of decisions and events. Probabilities are assigned to the events, and values are determined for each outcome.    Fig: Decision Tree Split    Fig: Fit Details of Training Data for Decision Tree    Fig: Contingency table of Validation Data for Decision Tree    Fig: Contingency table of Testing Data for Decision Tree Ensemble Model of DT, NN, LR, DA using LR: All 4 models yielded similar results and an ensemble model was built to improve the accuracy of the predictive classification. Since, Logistic regression gave better accuracy, the ensemble model was built using logistic regression.  **Result:**    Fig : Contingency table for testing Data |

# References:

1. <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing> – Source of dataset
2. <https://en.wikipedia.org/wiki/Logistic_regression>
3. <https://en.wikipedia.org/wiki/Artificial_neural_network>
4. <https://en.wikipedia.org/wiki/Consumer_confidence_index>
5. <http://www.bls.gov/cpi/>
6. <https://www.quandl.com/data/BOF/QS_D_IEUTIO3M-EURIBOR-3-Months-Daily>
7. <http://scu.edu.au/admin/hr/index.php/168/>