Contents

[I. Introduction 2](#_Toc425085297)

[1. Purpose of the project 2](#_Toc425085298)

[2. Background information of the company or scenario 2](#_Toc425085299)

[3. General concepts and trends relevant to the chosen topic 3](#_Toc425085300)

[4. Project Objectives 6](#_Toc425085301)

[5. How the application of Business Analytics can contribute to meeting those objectives 6](#_Toc425085302)

[6. Business problem statement of the project 7](#_Toc425085303)

[7. Project Schedule 7](#_Toc425085304)

[8. Literature Review 8](#_Toc425085305)

[II. Data Preparation & Models Construction 10](#_Toc425085306)

[1. Data Understanding Stage 10](#_Toc425085307)

[A. Description of the Original Data Set 10](#_Toc425085308)

[B. Initial Data familiarization 10](#_Toc425085309)

[C. Initial Hypothesis 13](#_Toc425085310)

[2. Data Preparation Stage 15](#_Toc425085311)

[A. Unbalanced Data Set – Under Sampling of Non-Subscribers 15](#_Toc425085312)

[B. Activities Performed in SAS Enterprise Guide 16](#_Toc425085313)

[C. Activities Performed in SAS Enterprise Miner 18](#_Toc425085314)

[3. Modeling Stage 20](#_Toc425085315)

[A. Initial Models 20](#_Toc425085316)

[B. Re-run of Model(s) after Fine-Tuning 23](#_Toc425085317)

[C. Model Comparison 24](#_Toc425085318)

[III. Evaluation and Conclusion 25](#_Toc425085319)

[1. Evaluation Stage 25](#_Toc425085320)

[A. Evaluation of the Decision Tree Results 25](#_Toc425085321)

[B. Interpretation of the Decision Tree Results 28](#_Toc425085322)

[C. Findings & Implications 30](#_Toc425085323)

[D. Conclusion / Summary 34](#_Toc425085324)

[Reference 35](#_Toc425085325)

# Introduction

Telemarketing is an interactive marketing technique that companies can employ to solicit business from existing and new customers. The main aim of telemarketing is to get customers to buy the company’s product or services. Compared to other mass marketing media like television commercials and outdoor advertisements, the cost may be considered relatively low. However, since the nature of such telemarketing campaign involves making unsolicited calls to prospective customers; the chances of making a successful sale may not be very high.

# Purpose of the project

In this project, we attempt to help a bank to increase the new subscription for term deposits by maximizing the success rate from its new telemarketing campaign. This will help the bank to reduce its campaign costs and efforts and also improve its return on investment.

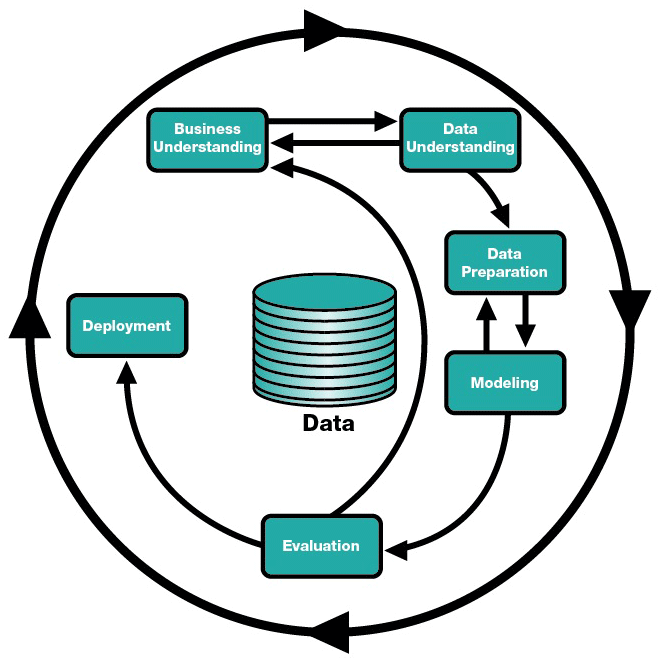
To achieve this, we will be using data-mining techniques to help the bank identify prospects that are more likely to respond positively to the bank’s unsolicited calls.

# Background information of the company or scenario

In the past the bank had conducted similar telemarketing campaigns. Results of each campaign (ie. whether a call resulted in a subscription), the detailed profile of the prospect called (ie. personal data like age, marital status and job nature etc) and the contact details (ie. number of current and previous contacts etc) had been recorded. In addition, macro-economic indicators prevailing at the time of a call was made were also recorded.

# General concepts and trends relevant to the chosen topic

* 1. **General Concept - Cross Industry Standard Process for Data Mining (CRISP-DM)**

The CRISP-DM model is the "de facto standard for developing data mining and knowledge discovery projects". [1] In this model, the whole process of data mining can be broken down into 6 consecutive major phases. However, the sequence of the phases is not strict and an analyst may need to move back to an earlier phase to adjust or fine tune certain variables or parameters if necessary.

* + 1. **Business Understanding**

The process starts with an initial understanding of the project objectives and requirements from a business perspective. After the business objectives has been clearly defined, only then we can proceed on to define what are the deliverables we want to achieve from the data mining exercise and plan for the necessary steps required. There is no value-add to business users if the data mining exercise produce accurate results that but do not help address the business problem at hand.

* + 1. **Data Understanding**

Data relevant to the business objectives are first collected. An initial familiarization with the collected data is needed to identify if there are any data quality problems and also to gain initial insights into the data. This phase is important as it may allow the analyst to identify interesting subsets within the data set and form hypotheses for hidden information that he/she can validate in the later stages.

* + 1. **Data Preparation**

In this phase, the analyst converts the initial raw data into the final data set. This may also be known as the ETL (Extraction, Transformation and Loading) process; where the analyst performs activities like record/attribute selection as well as transformation and cleaning of data to be fed into the modeling tool.

* + 1. **Modeling**

In this phase, one or multiple modeling techniques may be selected and applied to the data set. Usually several techniques can be applied to the same data mining problem type. Therefore, the analyst may need to run the different techniques to see which produce more suitable result for the task at hand. Also, each technique may be run more than one time as its parameters may be calibrated or fine-tuned to achieve optimal results.

* + 1. **Evaluation**

Having selected the optimal model, the analyst must now evaluate the model thoroughly and also review the steps executed to construct the model. He/she needs to consider it results from the selected model are able to sufficiently address the business objectives identified at the start of the project. A decision can then be made on whether to apply the data mining results to the business problem.

* + 1. **Deployment**

After the optimal model has been created, the analyst or business users need to come up a deployment plan to put the model to work. Also, there should be a monitoring and maintenance plan in place to ensure that the model is continually updated to remain valid and does not deteriorate in performance over time.

* 1. **Industrial Trends – Direct Marketing**

There are two main marketing methods that organizations today engage in their efforts to promote and sell their products, namely mass marketing and direct marketing.

Mass marketing is targeted at the general public through the use of traditional mass media. While it may reach a very large audience, most people exposed to the advertisement may not be keen or have no requirement at all for the product or service advertised. The changing lifestyles of today’s consumers also contribute to the decreased effectiveness of traditional mass media like television. People today are spending a greater portion of their time on new media types on the internet and mobile gadgets, therefore reducing their time spent on traditional media. Therefore, using media like newspapers, radios and television will result in high waste and low response rate from customers who will actually buy the product [2].

Direct marketing on the other hand focuses on specific audience group(s) within the general public. The selected audience is assumed to have keener interest in the marketer’s products or services. However, due to the ever increasing number of marketing campaign that consumers are exposed to; the effectiveness of such direct marketing campaigns has also diminished over time. To improve the response rate and investment returns from such campaigns, businesses are now tapping on Data Mining techniques to reveal a specific class of customers which are most likely to be interested in a particular product. This will allow the planning of a direct marketing campaign aimed toward a specific class of customers with the aim of achieving higher response [3].

Hence, direct marketing - aided by data mining techniques to pinpoint specific consumer group(s) to target at - will become the predominant marketing tool for most businesses in today’s competitive environment.

# Project Objectives

* 1. **Objectives of the project**

The main objective of this project is to help the bank in question to improve the customer response rate for its future telemarketing campaigns through the application of Business Analytics.

Based on the historical records of past campaigns that the bank had collected, we will apply the CRISP-DM model to help the bank identify which of its customers are most likely to subscribe to a new term deposit product. By helping the bank to identify customers with the highest potential to subscribe to its new product, we hope to help the bank to be able to better plan and improve the return on its marketing dollars.

# How the application of Business Analytics can contribute to meeting those objectives

Business Analytics can help meet above objectives by:

1. Identify the characteristics of customers who had purchased before
2. Identify the predominant characteristic(s) that have the greatest impact on the purchase decision
3. Allow testing of hypothesis formed on which are the predominant characteristic(s)
4. Provide a precise profile of the most likely customer
5. In addition to **Who** to contact, try to answer the question of **When** is the best time to contact a prospect (historical records contain contact details and general economic indicators at time of call also)
6. Help the bank to target customer with similar characteristics for its new product

# Business problem statement of the project

In this project, we will be using data-mining techniques to help the bank identify prospects that are most likely to respond positively to the bank’s unsolicited calls. In doing so, we can help the bank to execute a more cost-effective tele-marketing campaign and also improve the subscription rate for the new term deposit product.

The business problems can be stated as follow:

1. To understand the profile and predominant characteristics of a typical customer who will subscribe to term deposits
2. To predict which customers are most likely to purchase the new product
3. To determine when is the best time to call a potential customer

When translated into data mining problems, the following statements are applicable:

1. To use decision tree or clustering analysis to identify the important attributes of customers who will subscribe to the term deposit.
2. To construct a predictive model to identify customers who are most likely to subscribe to the new term deposit.
3. To estimate the probability of customers who will subscribe to the new term deposit.
4. To construct a predictive model to determine if there is any best time to launch the new product and telemarketing campaign (based on economic indicators) and/or when is the best time to contact a prospective customer.

# Project Schedule

Refer to attachment.

# Literature Review

Research into direct marketing strategy has shown that data mining is an effective tool for banks to improve their marketing campaigns. [4] In such campaigns, the quality of prospect data will very often determine successes. Organizations use data mining techniques to predict expected customers that have a higher probability to use their services.

In bank marketing campaigns that seek to use data mining to improve performance (Ling and Li, 1998)(Hu, 2005)(Li et al, 2010), a Classification DM approach was most commonly used. In this approach, the goal is to build a predictive model that can label a data item into one of several predefined classes (e.g. “yes”, “no”). Several DM algorithms can be used for classifying marketing contacts, each one with its own purposes and capabilities.

The prediction or classification is the most important task in the data mining that is usually applied to classify the group of data [5]. Especially for telemarketing campaigns, the Lift (the measure of the strength of association) is the most commonly used metric to evaluate prediction models (Coppock 2002). Based on the Lift, marketing managers can decide how many contacts to call (from the original set); and also check if there is an alternate model better suited for some other target responses.

In conjunction with classification, feature selection is another important method used in data mining for banks’ marketing campaigns. In feature selection ([6], [7]) the objective is to select the relevant features and discard the irrelevant or weak ones in the dataset; so that only a minimum set of features close enough to represent the original dataset will be selected.

Before proceeding with final deployment of the selected model, it is important to thoroughly evaluate the model and review the model’s construction to be certain it properly achieves the business objectives. While previous evaluation steps dealt with factors such as the accuracy and generality of the model, main objective of the evaluation step is to assess the degree to which the model meets the business objectives and determine if there is some business reason why this model is deficient. Another option here is to test the model(s) on real-world applications – if time and budget constraints permit. Moreover, evaluation also seeks to unveil additional challenges, information, or hints for future directions. [8]

# Data Preparation & Models Construction

# Data Understanding Stage

## Description of the Original Data Set

The data set we are using for this project is based on "Bank Marketing" UCI dataset ( <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>).

The data set records details of a bank's direct marketing campaigns based on phone calls to customers. It captures customers who had or who had not subscribed to the bank's term deposit in previous marketing campaigns.

The original data set contains a total of 41,188 records with 21 attributes. (Table 1)

## Initial Data familiarization

In this step, we try to ascertain if there are data quality problems with the collected data and also to gain initial insights into the data.

A data set that is not "clean" will produce misleading results. Therefore, it is important that we are able to identify the data quality problems; so that we can clean up the data before performing subsequent analyses. Areas we look at include: duplicate records, missing values, outliers and need for more appropriate classifications.

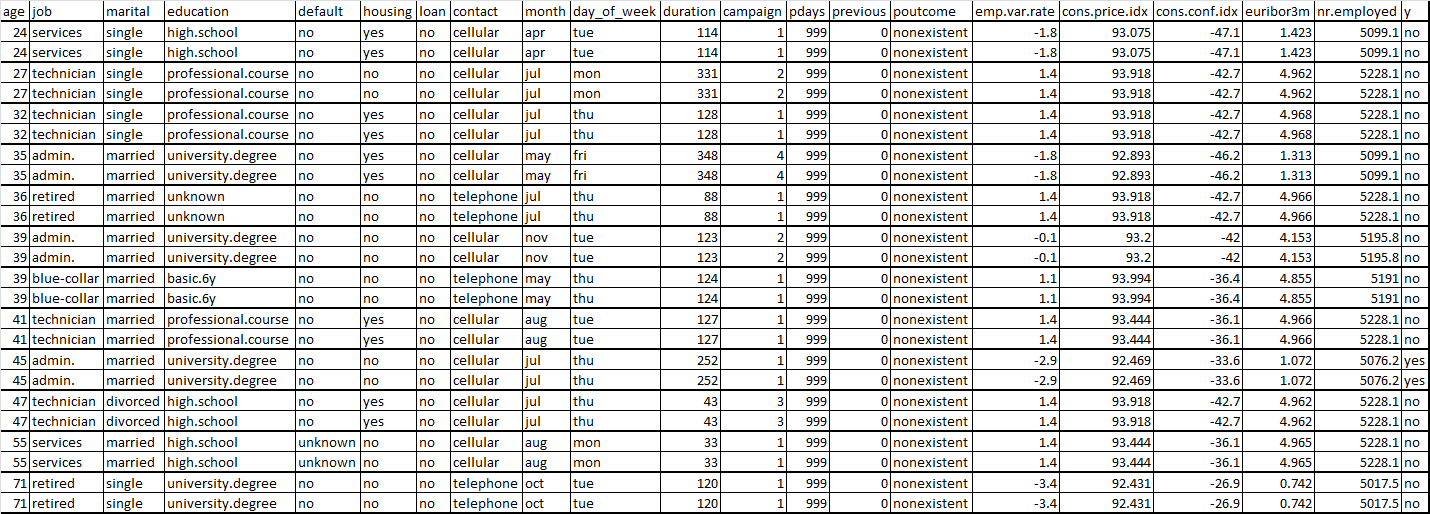
**Table 1 - Detail for the Bank Telemarketing Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| S/No | Attribute Name | Description | Type |
| 1 | Age | It is age of client. | Interval |
| 2 | Job | It is type of client’s job. | Nominal |
| 3 | Marital | It is client’s marital status. | Nominal |
| 4 | Education | What is the highest education of client? | Nominal |
| 5 | Default | Does client has credit? | Nominal |
| 6 | Housing | Does client has housing loan? | Nominal |
| 7 | Loan | Does client has personal loan? | Nominal |
| 8 | Contact | What is a contact communication type of client? | Nominal |
| 9 | Month | What is the last month of the year contracting to the client? | Nominal |
| 10 | Day of Week | What is the last day of the week contracting to the client? | Nominal |
| 11 | Duration | How long does it contact to the client? | Interval |
| 12 | Campaign | Number of contacts performed during this campaign and for this client | Interval |
| 13 | Pdays | Number of days that passed by after the client was last contacted from a previous campaign | Interval |
| 14 | Previous | Number of contacts performed before this campaign and for this client | Interval |
| 15 | Poutcome | Outcome of the previous marketing campaign | Nominal |
| 16 | Emp.var.rate | Employment variation rate | Interval |
| 17 | Cos.price.idx | Consumer price index | Interval |
| 18 | Cons.conf.idx | Consumer confidence index | Interval |
| 19 | Euribor3m | Euribor 3 month rate | Interval |
| 20 | Nr.employed | Number of employees | Interval |
| 21 | Y | Does the client has subscribed a term deposit? | Binary |

1. **Duplicate Records**

It is found that there are 12 duplicate records in the data set.

Action to be taken: Use SAS Enterprise Guide to remove duplicate records.



1. **Missing Values**
2. It is found that there are no missing values.
3. However, the following variables contain "Unknown" values:

|  |  |
| --- | --- |
| **Variable** | **Action to be taken** |
| Job | To reclassify later. |
| Marital | As these are categorical variables, we will leave the "Unknown" values as they are. |
| Education |
| Default |
| Housing |
| Loan |

1. The following variables contain "0" values:

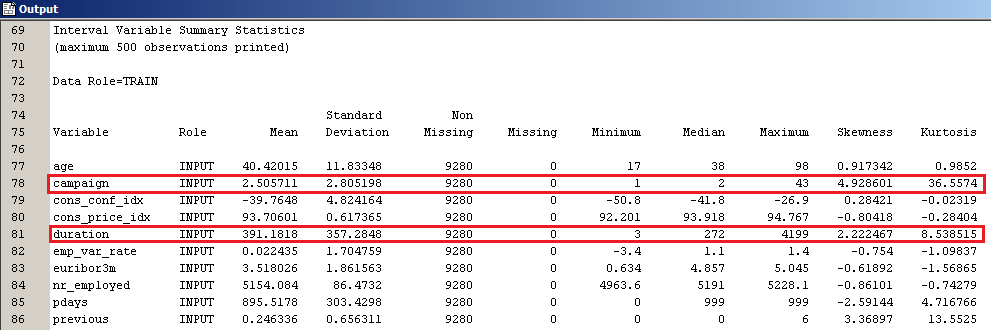
|  |  |
| --- | --- |
| **Variable** | **Action to be taken** |
| Duration | We assume "0" is legitimate value because customer did not pick up the call. |
| Previous | We assume "0" is legitimate value. |
| Pdays | To impute later as all records with "Pdays" = "0" has "Previous" value > 0; which does not make sense. |



1. **Outliers**

The following variables have extreme or outlying values:

|  |  |
| --- | --- |
| **Variable** | **Action to be taken** |
| Campaign | Apply transformation to these variables to normalize the distribution of values. |
| Duration |



1. **Need for More Appropriate Classification**

The following variables have values that are too widely spread to allow for effective management of target marketing campaigns based on customer demographics or calendar period. It will be easy to manage if we group them into smaller number of groups:

|  |  |
| --- | --- |
| **Variable** | **Action to be taken** |
| Age | Ranges from 17 to 98.  To reclassify into age groups. |
| Job | Currently have 12 values.  To reclassify into job groups. |
| Month | Currently have 10 values.  To reclassify into quarters – Q1 / Q2 / Q3 / Q4 |

## Initial Hypothesis

Based on initial exploration of the original data set, a customer that is more likely to respond positively to the bank's telemarketing campaign has the following characteristics:

|  |  |
| --- | --- |
| Variable | Description |
|  | Is contacted by cellular. |
|  | Was not contacted previously. |
|  | Is married. |
|  | Holding an admin, technician or blue-collar job. |
|  | Does not have personal loan. |
|  | Has university degree or high school certificate. |
|  | Is between 26 to 45 years old. |

# Data Preparation Stage

Having gotten ourselves familiarized with the data set and identified the data quality problem. We then proceed on to prepare the data for modelling. Some preparation activities are performed in the SAS Enterprise Guide and some activities are performed in the SAS Enterprise Miner.

## Unbalanced Data Set – Under Sampling of Non-Subscribers

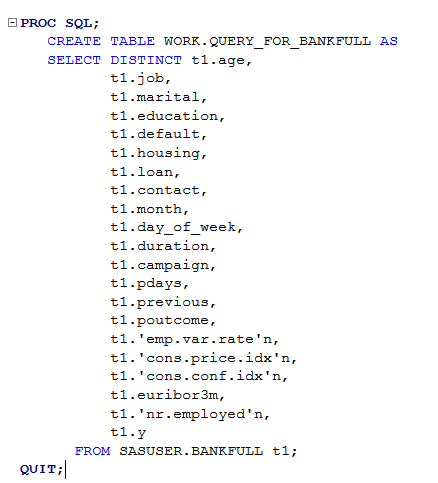
In the original dataset, there are 4,639 subscribers. However, there are 36,537 non-subscribers. As predictive modelling techniques like neural network and regression are sensitive to unbalanced data sets, we need to balance the data set used in this project.

In order to balance the number of subscribers and non-subscribers, we randomly selected 4,639 records from amongst the non-subscribers to match the number of subscribers. We then combine the 4,639 subscribers and 4,639 non-subscribers into a sample data set on which we later apply the different techniques to assess their effectiveness.

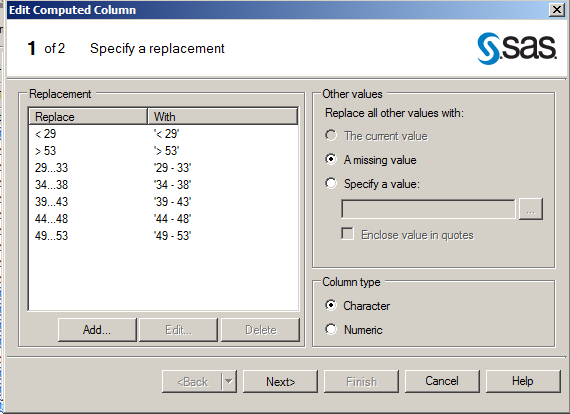
## Activities Performed in SAS Enterprise Guide

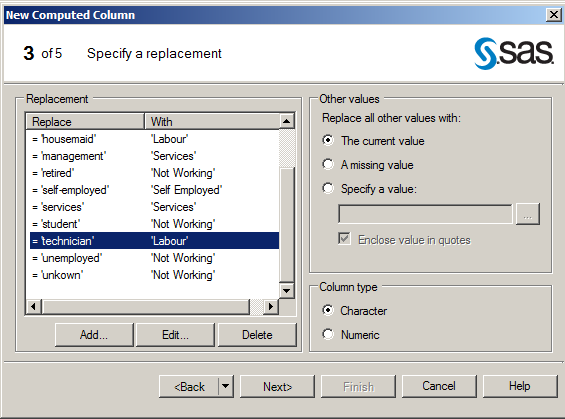
1. **Removal of Duplicate Records**

The following query was being run to remove duplicate records. A total of 12 duplicate records were removed. The total number of records in the data set was reduced to 41,176



1. **Reclassification/Recoding of Values**
2. Customer age which has values that range from 17 to 98 was regrouped into the 7 groupings and given a new variable name "AgeGroup":



1. Customer job which originally has 12 different values was regrouped into the 4 groupings and given a new variable name "JobGroup":

## Activities Performed in SAS Enterprise Miner

1. **Rejection of Macro-economic Variables**

The following variables are not directly related to the customers' demographics. They fall under macro-economic factors that the bank or customer has no direct control over. Nevertheless, they may provide indications on when are the best times to for the bank to conduct such telemarketing campaigns. We will revisit these macro-economic factors at the end of the report.

|  |  |  |  |
| --- | --- | --- | --- |
| S/No | Attribute Name | Description | Type |
| 1 | Emp.var.rate | Employment variation rate | Interval |
| 2 | Euribor3m | Euribor 3 month rate | Interval |
| 3 | Nr.employed | Number of employees | Interval |
| 4 | Cons.conf.idx | Consumer confidence index | Interval |
| 5 | Cos.price.idx | Consumer price index | Interval |

To make the models relevant to the business problem at hand, we will therefore leave out all macro-economic variables from the models by setting their roles to "Reject" for now.

1. **Imputation of "0"Values**

"0" values for Pdays were first replaced with missing value. The missing values were then replaced with the sample mean value.

|  |  |
| --- | --- |
| **Variable** | **Action taken** |
| Pdays | All records with "Pdays" = "0" has "Previous" value > 0; which does not make sense. To replace the missing value with missing " " first and impute later. |

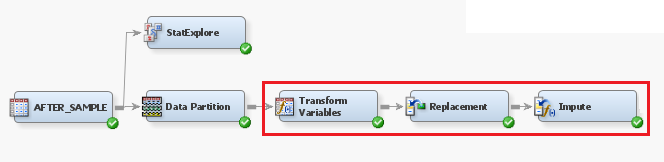
1. **Replacement of Values**

"Month" values were replaced by quarter values – Q1 / Q2 / Q3 / Q4.

1. **Transformation of** **extreme or outlying values:**

Log transformation was applied to the "Campaign" and "Duration" variables to normalize the distribution of the values.

|  |  |
| --- | --- |
| **Variable** | **Action taken** |
| Campaign | Apply transformation to these variables to normalize the distribution of values. |
| Duration |

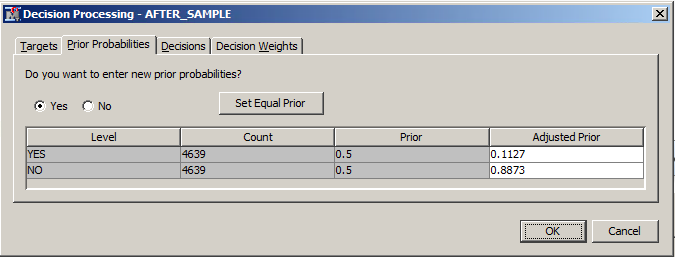


1. **Adjusting the Prior Probabilities**

Population probabilities:

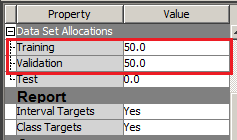
Yes = 0.112663

No = 0.887337



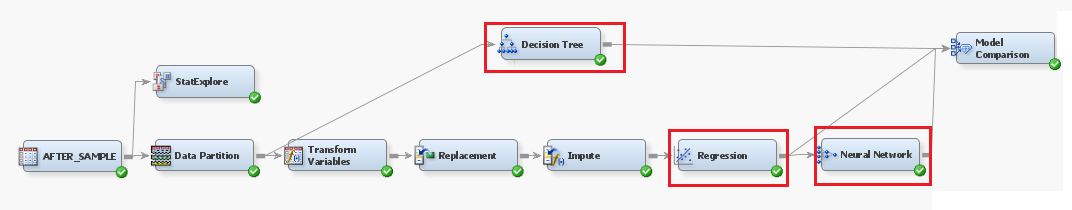
1. **Data Partitioning**

The data was split into 50% training and 50% validation.



# Modeling Stage

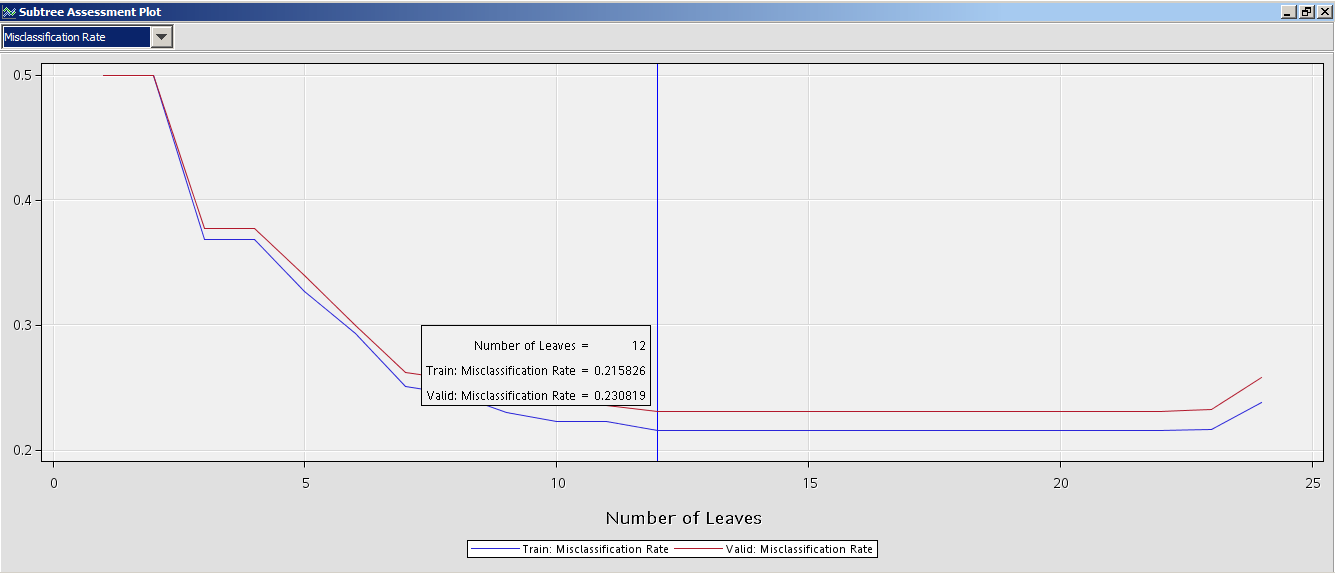
In this stage, we apply different data mining techniques to the data set to see the results that each technique produces. From here, we can then compare the results of each model and select the best technique most suitable for our project. Techniques used include the following: Decision Tree, Neural Network and Regression.

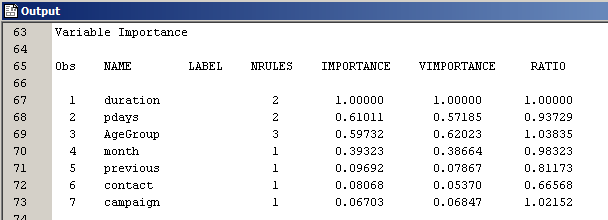


## Initial Models

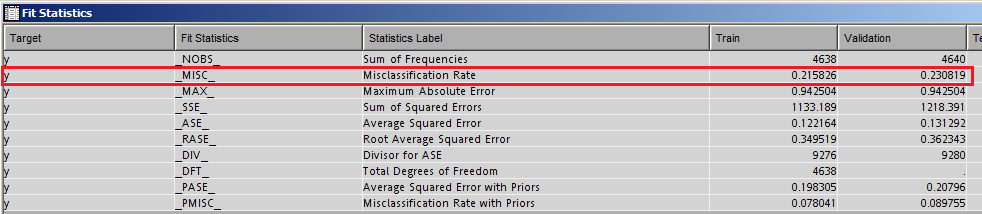
1. **Decision Tree**

The optimal tree has 12 leaves and there are 7 important variables.





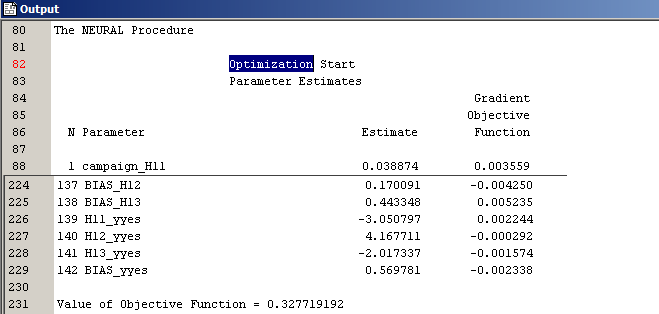
The miscalculation rates for the training and validation data are as follow:

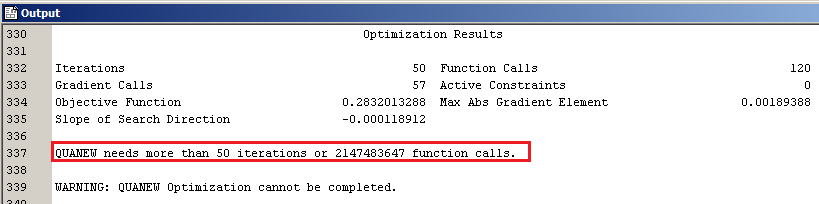


1. **Neural Network**

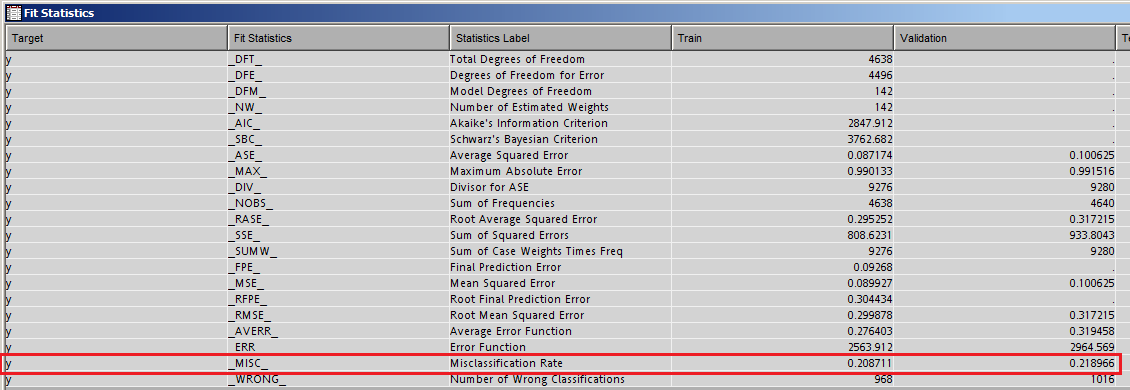
The model has 142 weights and needs more than 50 iterations to converge.

(We will try to limit the number of weights and iterations later.)



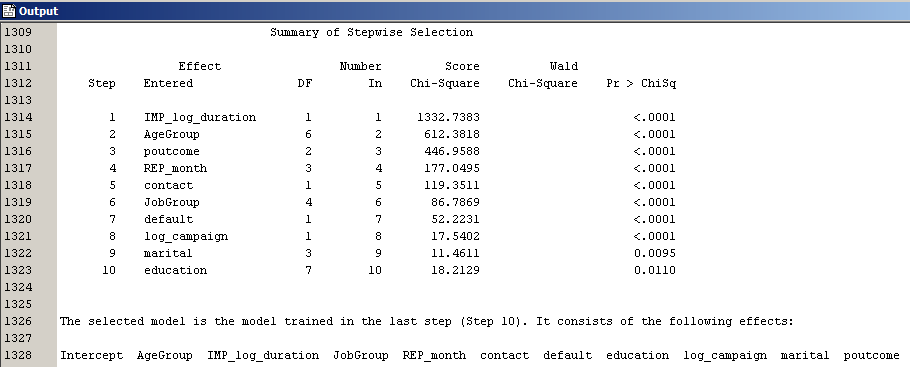


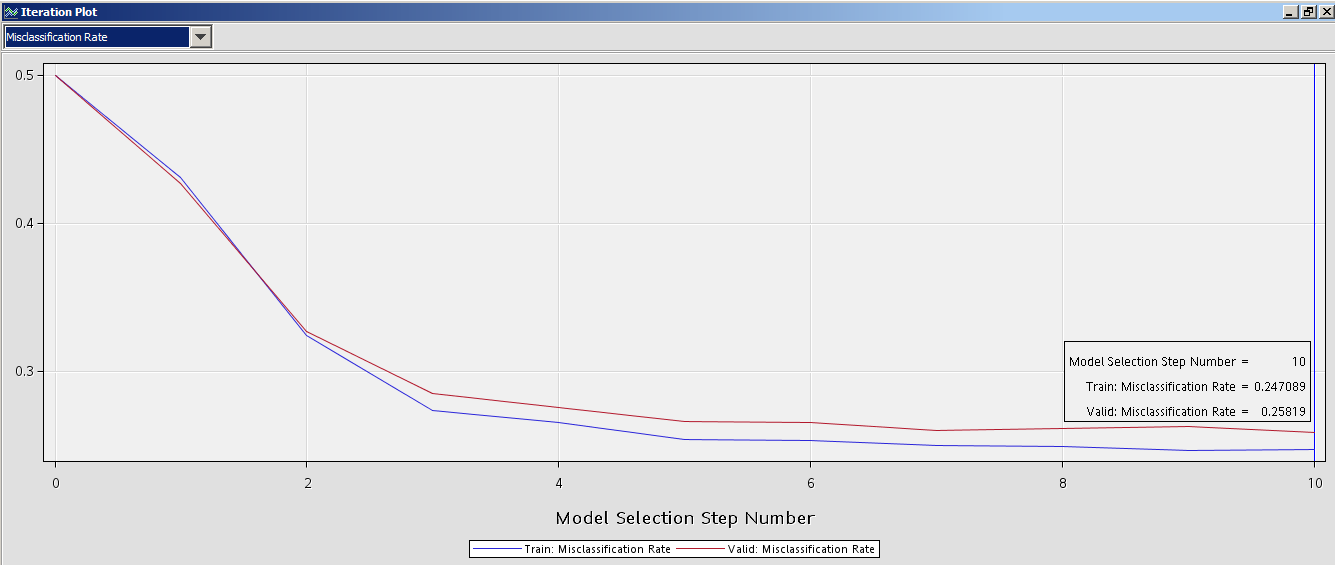
The miscalculation rates for the training and validation data are as follow:



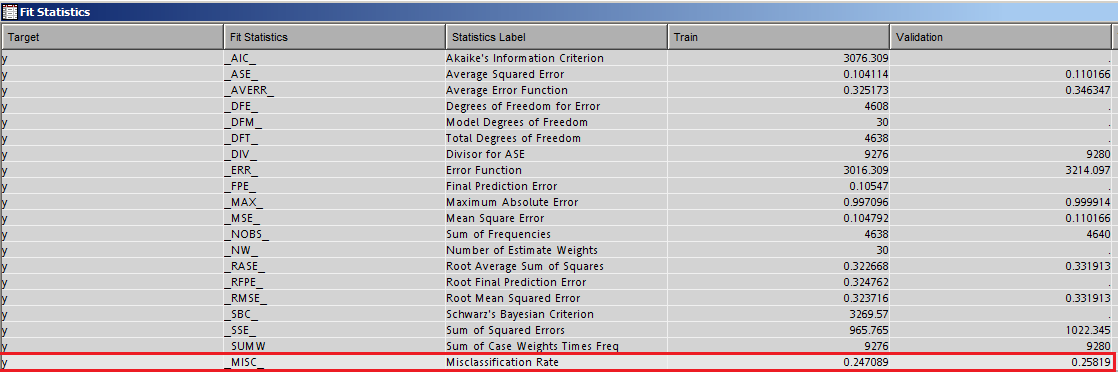
1. **Stepwise Regression**

The model has 10 effects and is optimized around step 10.





The miscalculation rates for the training and validation data are as follow:



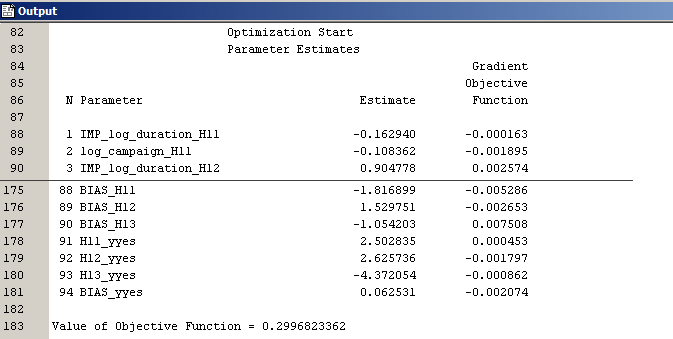
## Re-run of Model(s) after Fine-Tuning

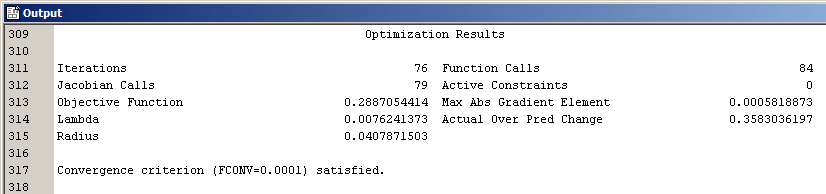
1. **Reducing Inputs and Weights for Neural Network**

The initial neural network has too many weights and takes a large number of iterations to converge. To overcome this problem, we used the regression model to select input for the neural network by connecting them together.

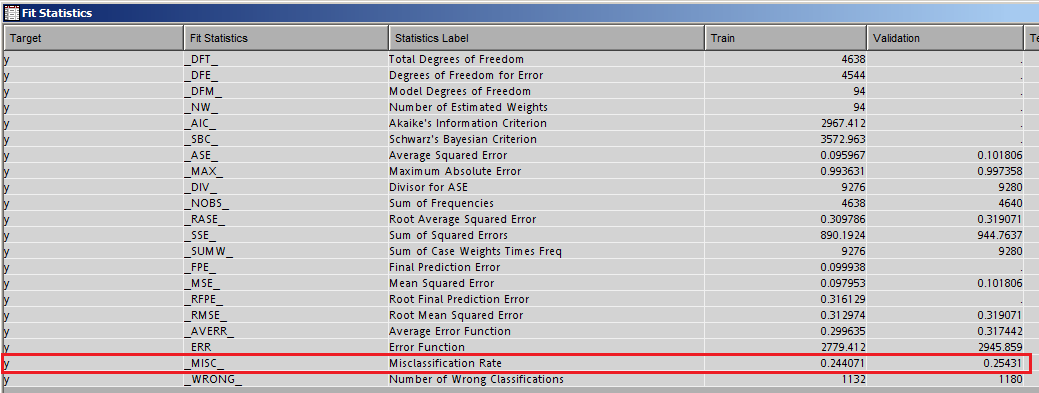
Re-run of the neural network yields the following results:

The model has 94 weights and needs more than 76 iterations to converge.



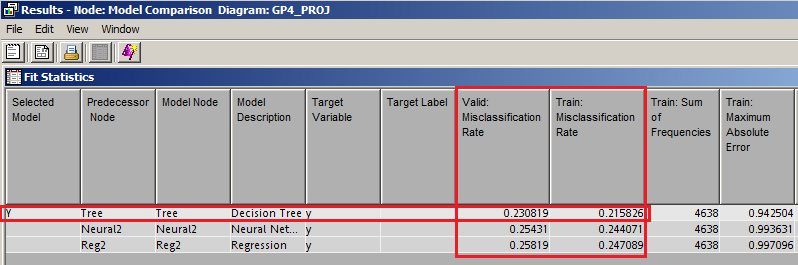


The miscalculation rates for the training and validation data are as follow:



## Model Comparison

We performed a model comparison of the decision tree, regression and neural network models. Based on the validation misclassification rate, the model with the least misclassification rate was selected as the best model to run our predictions. As shown below, the decision tree has the least misclassification rate.



# Evaluation and Conclusion

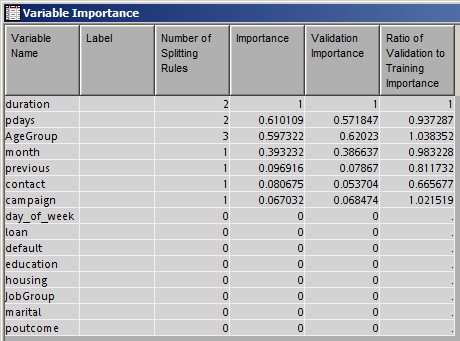
# Evaluation Stage

## Evaluation of the Decision Tree Results

The decision tree had the least validation misclassification rate and was therefore best suited to address our business problem. Based on the results of the decision tree, we can gain some valuable insights into the dataset.

1. **Variable Importance**

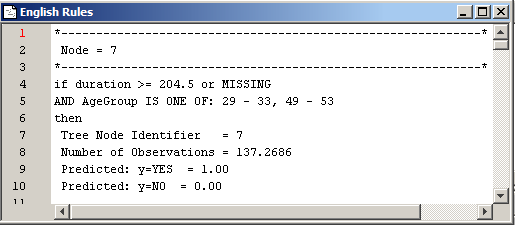
The Variable Importance window provides insight into the importance of inputs in the decision tree. Below are the variables that have the greatest influence on the predicted outcome of the decision tree.

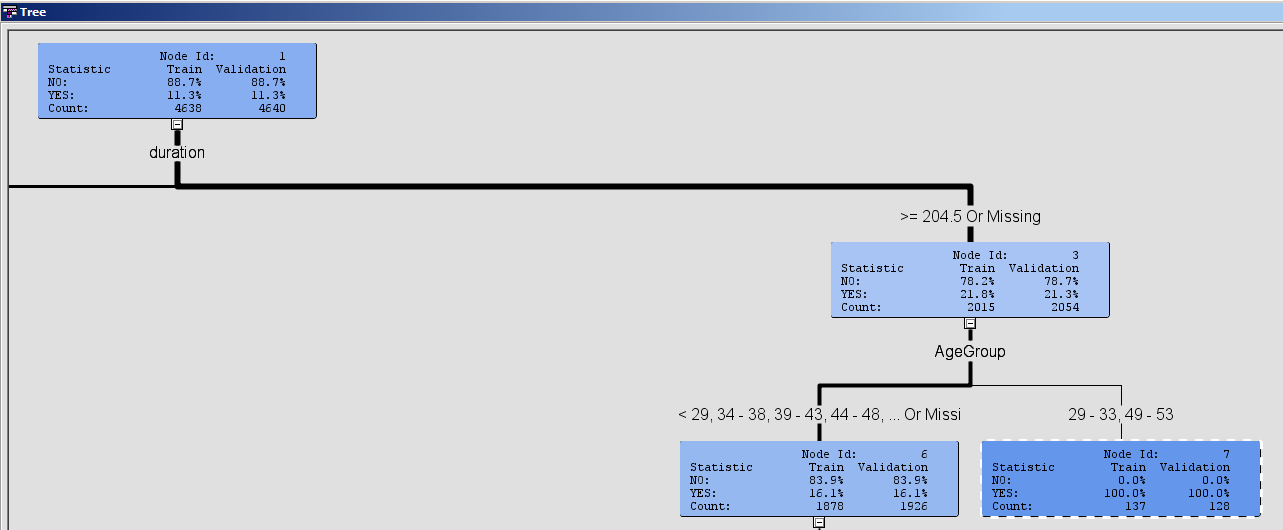


|  |  |  |
| --- | --- | --- |
| Order of Importance | Attribute Name | Description |
| 1 | Duration | How long does it contact to the client? |
| 2 | Pdays | Number of days that passed by after the client was last contacted from a previous campaign |
| 3 | Age Group | Age group of the client. |
| 4 | Month | What is the last month of the year contracting to the client? |
| 5 | Previous | Number of contacts performed before this campaign and for this client |
| 6 | Contact | What is a contact communication type of client? |
| 7 | Campaign | Number of contacts performed during this campaign and for this client |

1. **English Rules**

The English rule window displays the definition for the leaf nodes (ie. the combination of variable values) in a decision tree model. Also, it shows the number of observations included in a particular leaf and the proportion of predicted outcomes.

****

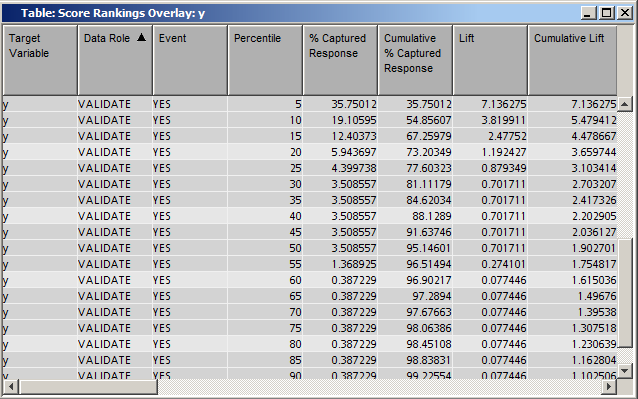


Summary of the leaf nodes:

|  |  |  |  |
| --- | --- | --- | --- |
| Node | Variable Values | Number of Observations | Predicted Ratio (Count) |
| 7 | if duration >= 204.5 or MISSING  AND  AgeGroup IS ONE OF: 29 - 33, 49 - 53 | 137.2686 | y=YES = 1.00 (137)  y=NO = 0.00 (0) |
| 8 | if pdays < 16  AND duration < 204.5  AND campaign < 3.5 or MISSING | 49.8638 | y=YES = 0.50 (25)  y=NO = 0.50 (25) |
| 9 | if pdays < 16  AND duration < 204.5  AND campaign >= 3.5 | 7.7746 | y=YES = 0.09 (1)  y=NO = 0.91 (7) |
| 10 | if pdays >= 16 or MISSING  AND month IS ONE OF: OCT, DEC, SEP, APR, MAR  AND duration < 204.5 | 241.635 | y=YES = 0.13 (31)  y=NO = 0.87 (210) |
| 16 | if pdays >= 16 or MISSING  AND month IS ONE OF: AUG, MAY, JUL, NOV, JUN or MISSING  AND duration < 204.5  AND AgeGroup IS ONE OF: < 29, 34 - 38, 39 - 43, 44 - 48, > 53 or MISSING | 2315.9136 | y=YES = 0.01 (23)  y=NO = 0.99 (2292) |
| 17 | if pdays >= 16 or MISSING  AND month IS ONE OF: AUG, MAY, JUL, NOV, JUN or MISSING  AND duration < 204.5  AND AgeGroup IS ONE OF: 29 - 33, 49 – 53 | 7.6636 | y=YES = 1.00 (7)  y=NO = 0.00 (0) |
| 18 | if pdays < 513  AND duration < 520.5 AND duration >= 204.5 or MISSING  AND AgeGroup IS ONE OF: < 29, 34 - 38, 39 - 43, 44 - 48, > 53 or MISSING | 81.953 | y=YES = 0.59 (48)  y=NO = 0.41 (33) |
| 19 | if pdays >= 513 or MISSING  AND duration < 520.5 AND duration >= 204.5 or MISSING  AND AgeGroup IS ONE OF: < 29, 34 - 38, 39 - 43, 44 - 48, > 53 or MISSING | 1354.523 | y=YES = 0.08 (108)  y=NO = 0.92 (1246) |
| 30 | if previous < 0.5 or MISSING  AND duration >= 520.5  AND AgeGroup IS ONE OF: > 53 | 134.9874 | y=YES = 0.16 (21)  y=NO = 0.84 (113) |
| 31 | if previous >= 0.5  AND duration >= 520.5  AND AgeGroup IS ONE OF: > 53 | 15.2128 | y=YES = 0.53 (8)  y=NO = 0.47 (7) |
| 32 | if duration >= 520.5  AND contact IS ONE OF: TELEPHONE  AND AgeGroup IS ONE OF: < 29, 34 - 38, 39 - 43, 44 - 48 or MISSING | 110.0892 | y=YES = 0.29 (32)  y=NO = 0.71 (78) |
| 33 | if duration >= 520.5  AND contact IS ONE OF: CELLULAR or MISSING  AND AgeGroup IS ONE OF: < 29, 34 - 38, 39 - 43, 44 - 48 or MISSING | 181.1154 | y=YES = 0.50 (91)  y=NO = 0.50 (90) |

1. **Score Rankings Overlay**

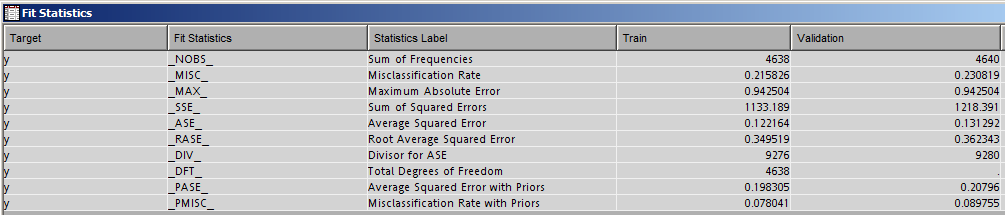
From the score rankings overlay, we are able to see the proportion of events captured by the model at different percentiles. The % captured response shows us the percentage of events captured by the model, and the model's lift tells us the ratio of the percentage of captured events as compared to a random model. For instance, at the 60th percentile the decision tree model captures almost 97% of all responses and 2.2 times as many responses compared to a random model.



|  |  |  |
| --- | --- | --- |
| Percentile | Cumulative % Captured Response | Cumulative Lift |
| 20 | 73.20 | 3.66 |
| 40 | 88.12 | 2.20 |
| 60 | 96.90 | 1.62 |
| 80 | 98.45 | 1.23 |

1. **Fit Statistics**

The fit statistics shows a validation misclassification rate of 23.08%. That means the decision tree model has a 76.92% accuracy rate in predicting the correct outcome.



## Interpretation of the Decision Tree Results

1. **Insights Gain from the Decision Tree Results**

To gain deeper insights into the characteristics of customers who are likely to respond positively to the bank's marketing campaign, we tabulate the results of the decision tree according to the important variables. The below tables show the total count of predictions for "y=YES" influenced by each variable and its corresponding values.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Nodes | Predicted total count for y=YES |
| Duration | >= 204.5 or MISSING | 7, 18, 19 | 293 |
| < 204.5 | 8, 9, 10, 16, 17 | 87 |
| >= 520.5 | 30, 31, 32, 33 | 152 |
| Interpretation: The likelihood of customer responding positively to the marketing campaign increases if the duration is longer. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Nodes | Predicted total count for y=YES |
| Pdays | < 16 | 8, 9 | 26 |
| >= 16 or MISSING | 10, 16, 17 | 61 |
| < 513 | 18 | 48 |
| >= 513 or MISSING | 19 | 108 |
| Interpretation: The longer the number of days that passed since a client was last contacted, the higher the chance of positive response. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Nodes | Predicted total count for y=YES |
| Age Group | 29 - 33, 49 – 53 | 7, 17 | 144 |
| < 29, 34 - 38, 39 - 43, 44 - 48, > 53 or MISSING | 16, 18, 19 | 179 |
| > 53 | 30, 31 | 29 |
| < 29, 34 - 38, 39 - 43, 44 - 48 or MISSING | 32, 33 | 123 |
| Interpretation: Not definitive age group, but it appears that the majority of customers most likely to respond positively are between 29 – 48 years old. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Nodes | Predicted total count for y=YES |
| Month | OCT, DEC, SEP, APR, MAR | 10 | 31 |
| AUG, MAY, JUL, NOV, JUN or MISSING | 16, 17 | 30 |
| Interpretation: The last month of the year customer was contacted does not substantially affect the response rate. | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Nodes | Predicted total count for y=YES |
| Previous | < 0.5 or MISSING | 30 | 21 |
| >= 0.5 | 31 | 8 |
| Interpretation: Customers who were not contacted before are almost 3 times more likely to respond positively. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Nodes | Predicted total count for y=YES |
| Contact | TELEPHONE | 32 | 32 |
| CELLULAR or MISSING | 33 | 91 |
| Interpretation: Customers who were contacted by cellular are 3 times more likely to response positively. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Nodes | Predicted total count for y=YES |
| Campaign | < 3.5 or MISSING | 8 | 25 |
| campaign >= 3.5 | 9 | 1 |
| Interpretation: Customers who has been contacted less frequent (1 to 3 times only) are much more likely to respond positively. | | | |

1. **Comparison with Initial Hypothesis**

During the data understanding stage, we formed the initial hypothesis that customers with certain characteristics are more likely to respond positively to the bank's telemarketing campaign. In below table we relook at the initial hypothesis to confirm if it was correct:

|  |  |
| --- | --- |
| Initial Hypothesis | Correctness |
| Is contacted by cellular. | True – 3 times more likely. |
| Was not contacted previously. | True – 3 times more likely. |
| Is married. | False. |
| Holding an admin, technician or blue-collar job. | False. |
| Does not have personal loan. | False. |
| Has university degree or high school certificate. | False. |
| Is between 26 to 45 years old. | Partially true – should be 29 – 48 years old. |

Additional characteristics not in the initial hypothesis but are important include:

|  |
| --- |
| Duration - the longer, the more likely to respond positively. |
| Pdays - the longer, the more likely to respond positively. |
| Campaign – the lesser, the more likely to respond positively. |

## Findings & Implications

The main objective of this project is to help the bank improve the customer response rate for its future telemarketing campaigns. To determine if we had achieved the objective, we need to look at whether we are able to address the business problems defined earlier, which are:

1. To understand the profile and predominant characteristics of a typical customer who will subscribe to term deposits
2. To predict which customers are most likely to purchase the new product
3. To determine when is the best time to call a potential customer
4. **Predominant Characteristics**

Problem i) has been addressed by identify the most important variables that affect the response outcome:

|  |  |  |
| --- | --- | --- |
| Order of Importance | Attribute Name | Description |
| 1 | Duration | How long does it contact to the client? |
| 2 | Pdays | Number of days that passed by after the client was last contacted from a previous campaign |
| 3 | Age Group | Age group of the client. |
| 4 | Month | What is the last month of the year contracting to the client? |
| 5 | Previous | Number of contacts performed before this campaign and for this client |
| 6 | Contact | What is a contact communication type of client? |
| 7 | Campaign | Number of contacts performed during this campaign and for this client |

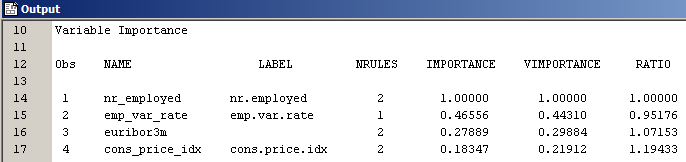
1. **Customer Profile**

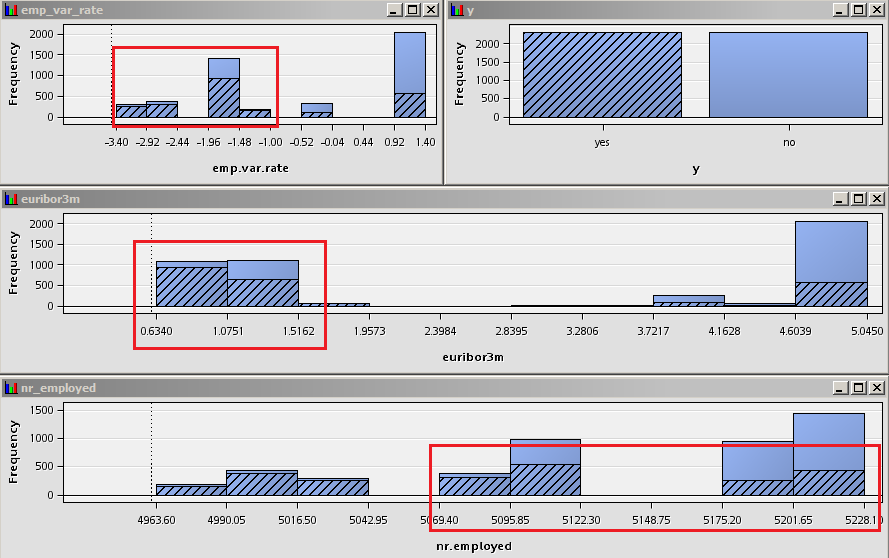
Problem ii) has been addressed by identify the characteristics of customers most likely to purchase:

|  |
| --- |
| Description |
| Is contacted by cellular (3 times more likely). |
| Was not contacted previously (3 times more likely). |
| Is between 29 – 48 years old. |
| Has been contacted (duration) longer. |
| Extended number of days had passed (Pdays) since last contacted from a previous campaign. |
| Was contacted only 1 – 3 times during current campaign (Campaign). |

1. **Best Time to Contact Potential Customers (or Launch Campaign)**

To address this problem, we rerun the decision tree model using only the macro-economic variables (which we rejected in the earlier models) as input to obtain the top 3 most important macro-economic variables that affect the outcome.





And the best time to contact potential customers (or run campaign) is when these macro-economic indicators are:

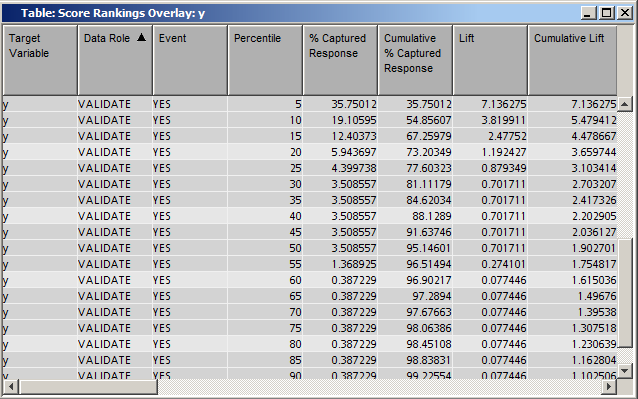
|  |  |  |  |
| --- | --- | --- | --- |
| Order of Importance | Attribute Name | Description | Increased likelihood of positive response when indicator is: |
| 1 | Nr.employed | Number of employees | High |
| 2 | Emp.var.rate | Employment variation rate | Low |
| 3 | Euribor3m | Euribor 3 month rate | Low |

Based on the above findings, we are able to help the bank address the business problems defined at the start of the project. We can therefore apply our predictive model created in this project (the decision tree model) to the bank's real-world scenario.

1. **Number of Potential Customers to Call**

The end result we are looking at is to help the bank improve returns on the its marketing dollars. Therefore, in addition to answering the questions of who to call and when is best time to call, we should also be able to advise the bank on the question of "how many to call?".

The Score Rankings Overlay table mentioned in earlier section helps us answer this question by showing the % captured response and the model's lift at different percentiles. For instance, at the top 20th percentile the decision tree model captures 73.2% of all responses and 3.66 times as many responses compared to a random model; and at the top 60th percentile the model captures almost 97% of all responses.



Using this table as a guide, the bank can decide on the total number of potential customers to call in order to achieve the target number of responses. This will help the bank avoid the problem of diminishing returns on its marketing dollars by not calling more customers than it need to.

## Conclusion / Summary

Predictive modelling allows us to answer the key question of: who to call, when to call and how many to call in a typical telemarketing campaign. In addition, it helps us to rank the importance of the characteristics (variables) of customers to allow us to understand which characteristics predominantly affect customer behaviours.

At the start of the project, we had formed certain hypotheses about the dataset. But only through the use of predictive models and evaluation of the model results, can we then conclude which of these hypotheses are valid and to what extend are they valid.

Therefore, the use of predictive modelling allows us to test and refine our assumptions made on the dataset. This will in turn allow marketers to fine-tune their marketing elements in the real world to target at the customer groups most likely to respond positively to the marketing campaign. The end result is improvement in returns from marketing investment.

## Reference

[1] Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009); [A Data Mining & Knowledge Discovery Process Model](http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf)

[2] P. Van der Putten, “Data Mining in Direct Marketing Databases”, Complexity and Management: A Collection of Essay, World Scientific. 1999.

[3] Y. Kim, “Toward a Successful CRM: Variable Selection, Sampling, and Ensemble”, Decision Support Systems, vol.41, no. 2, 2006, pp. 542-553.

[4] Sing L. et. al. “Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing”, IJCSI, vol.10, Issue 2, No.2, 2013.

[5] A. Suebsing, C. Vajiramedhin, Accuracy rate of predictive models in credit screening, Applied Mathematical Sciences, vol. 7 no.112, 2013, 5591-5597.

[6] G. John, R. Kohavi, and Pfleger. Irrelevant features and the subset selection problem. Int. Conf. on Machine Learning, Morgan Kaufman, San Francisco. 1994, 121-129.

[7] W. Duch, T. Winiarski, J. Biesiada, and A. Kachel, Feature Ranking Selection and Discretization. Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), Istanbul. 2003, 251-254.

[8] Colin Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining”, Journal of Data Warehousing, Volume 5, Number 4, Fall 2000