**Bank Marketing**

**MIS 436**

**Group 1**

**Yen Vo, Zhan Xu, Keegan Wolfe, Yu Pei**

**Table of Contents**

[**Background**](#_mp09nlsygntd) **3**

[**Variables and Data**](#_n89224z7wfe) **4**

[**Model and Interpretation**](#_l17vweak4rut) **6**

[Data Mining techniques](#_eaye69zdox9r) 6

[Data partitioning:](#_ek3mil53gi81) 6

[Decision Tree](#_rmpdy4w9zhkk) 6

[Maximal Tree:](#_rxitqgoh6cu) 7

[Probability Tree:](#_ad17j74cdacm) 8

[Optimal Tree:](#_lgigrtao7qob) 9

[Regression Model:](#_rwgmf15kogwl) 10

[Model Comparison](#_7y6ku0j1tnfv) 17

[**Conclusion and Recommendations**](#_a8ca5gnz2yh9) **18**

[**References**](#_hs0kj229ja2) **19**

# Background

In 2008, there was an economic crisis, which affected not only American banking but also European countries, especially Portugal. The Portuguese economy did not grow that year and fell nearly 3% in 2009. After the financial crisis, bankers found that two Portuguese banks had been accumulating losses for years because of bad investments, embezzlement, and accounting fraud(1).

In order to improve their financial situation, Portuguese banking institutions launched a direct-to-direct marketing campaign, which was mainly telephone-based. Bankers need to contact multiple contacts of the same customer in order to access if the product (bank fixed deposit) would be subscribed. The banker recorded all the information and created a data set.

The general purpose of the project is to find out which groups are more likely to subscribe to a term deposit. Four statistical methods have been used to analyze the trend of consumers, including decision trees, logistic regression, neural networks, and clustering.We aim to give precise results by comparing the predictive accuracy of probability among different statistical methods to address the following specific issues:

1. Which clients are more likely to make deposits based on factors: Age, Job, Marital, education, credit in default, having a loan, and house?
2. Does the previous marketing campaign outcome exist and target the right customers? Can we learn anything from it?
3. Is telemarketing a good strategy to approach clients?
4. What is the future marketing strategy for the bank?

# Variables and Data

The dataset includes 41,188 observations and 20 inputs, ordered by data (from May 2008 to November 2010).

| Group | Variable | Type | Missing / Unknown |
| --- | --- | --- | --- |
| Client | Age | Interval | no |
| Job | Nominal | 330 |
| Marital | Nominal | 80 |
| Education | Ordinal | 1,731 |
| Default | Binary | 8,597 |
| Housing | Binary | 990 |
| Loan | Binary | 990 |
| Last contact of the current campaign | Contact | Binary | no |
| Month | Nominal | no |
| Day\_of\_week | Nominal | no |
| Duration | Ratio | no |
| Other Attributes | Campaign | Interval | no |
| Pday | Ratio | 39,673 |
| Previous | Ratio | no |
| Poutcome | Nominal | no |
| Social and economic factors | Emp.var.rate | Ratio | no |
| Cons.price.idx | Interval | no |
| Cons.conf.idx | Interval | no |
| Euribor3m | Interval | no |
| Nr.employed | Interval | no |

Based on the target question we deal with, we decide to not use these five variables for future analysis. Because pdays shows that over 96% clients were not previously contacted, we will exclude pdays in our dataset. Because poutcome shows that over 86% of the observations are nonexistent, that means we can’t get any useful information from the previous campaign. We will exclude outcomes in our dataset. Because previous shows that over 86% of the observations are 0, we can assume most of the clients are new. We will exclude previous in our dataset.

Variables we will include in out dataset:

1. Housing
2. Job
3. Duration
4. Education
5. Marital
6. Month
7. Loan
8. Campaign
9. Contact
10. Age
11. Day\_of\_week
12. Default
13. Target : y

# Model and Interpretation

## Data Mining techniques

In the era of information technology, big data plays a key role in predicting consumer trends and the information that we extract can help the bank gain improved insight about the clients and make a better move to improve the business situation. In this particular project, we used two important data mining techniques to explore the data set, including supervised learning and unsupervised learning.

Supervised learning predicts a target variable and to evaluate predictions, we need to “validate” our models on a holdout sample ( partitioning). Three methods that we used for this technique are decision tree, regression, and neural network.

Unsupervised learning detects patterns in data. We used cluster analysis to measure the similarity and dissimilarity in different segmentations in the data set.

The results of the two mining techniques employed in our project are summarized as follows:

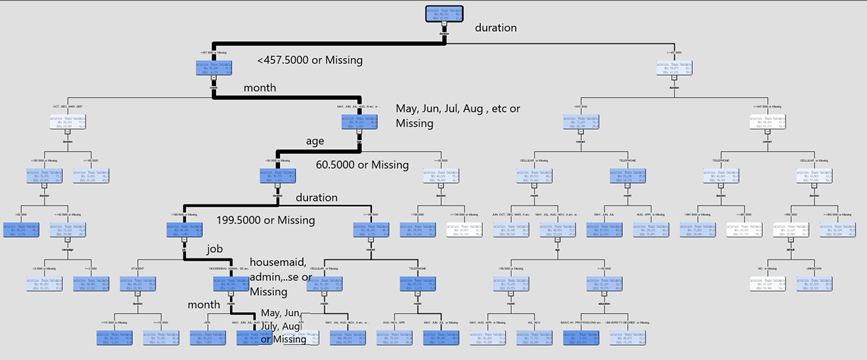
## Data partitioning:

Before training data, we added data partition node to the diagram and assign 50% of the data for training and 50% for validation

**Decision Tree**

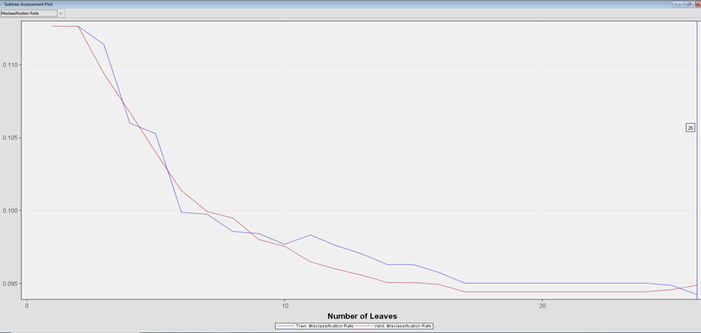
**The decision tree belongs to categorical prediction. It predicts categorical labels of the dependent variable using other available attributes. In this model, we used three different techniques to analyze the data, including maximal tree, probability tree, and optimal tree.**

**Maximal Tree:**

****

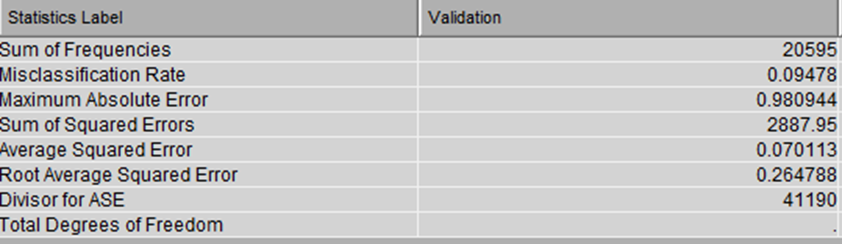
**Fig. 1. Maximal tree.**

**The tree generated 26 leaf nodes**

****

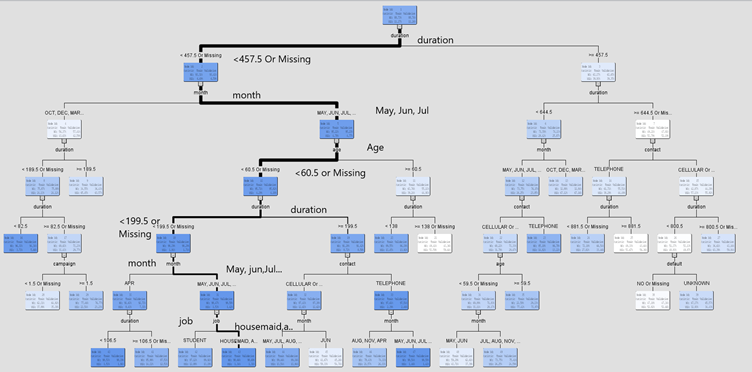
**Fig. 2. Subtree assessment plot of maximal tree.**

**The average squared error is 0.069143 and the misclassification rate is 0.094877 which means we predict 6.9 % of the mistake on validation data based on misclassification rate and 9.48 % based on average squared error.**

****

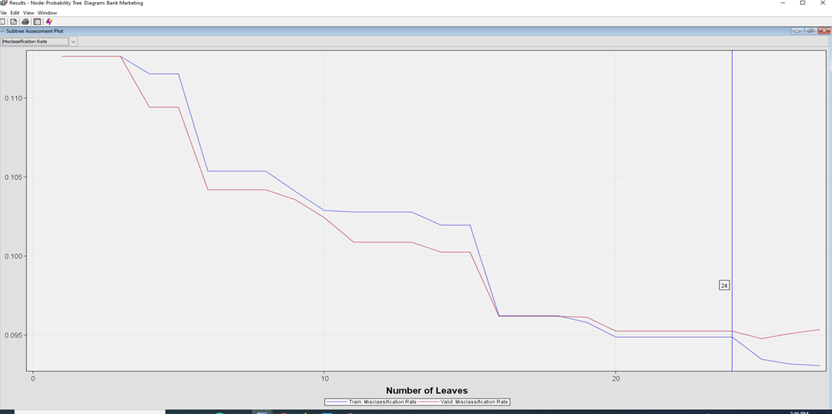
**Fig. 3. Fit Statistics of the maximal tree.**

**In Probability Tree, the branch selection is the same as the maximal tree. However, it is much simpler and has fewer leaf nodes.**

****

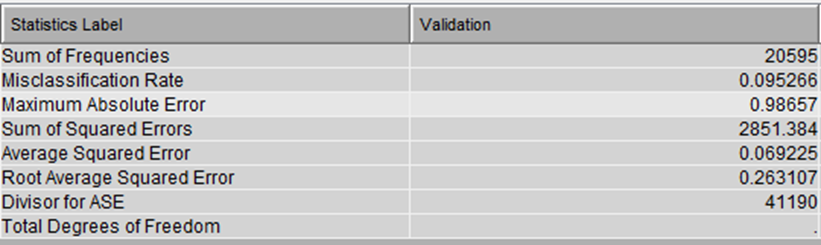
**Fig. 4. Probability Tree**

**There is 24 leaf nodes in the tree, and the best performance was based on the misclassification rate at 24 leaf nodes.**

****

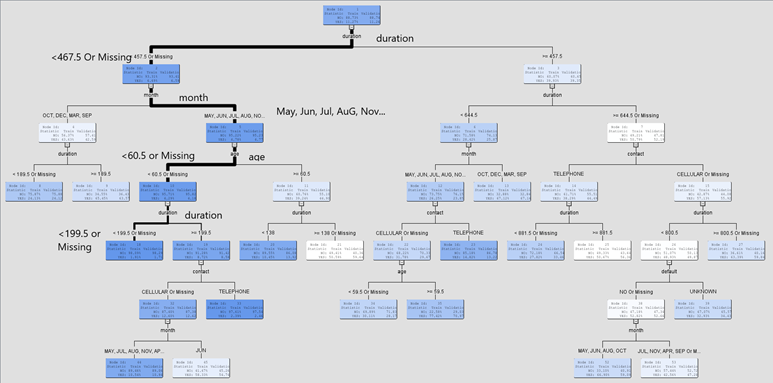
**Fig. 5. Probability tree’s subtree assessment plot.**

**We are wrong 9.53 % on validation data of the time based on Misclassification rate and 6.92% based on ASE. The result is slightly higher than the Maximal Tree.**

****

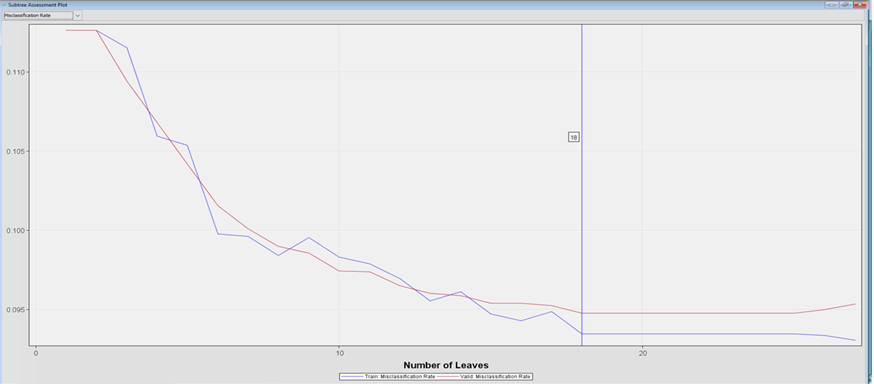
**Fig. 6. Fit Statistics of probability tree with the highlight of important numbers.**

**Optimal Tree: The selection branch has a duration of fewer than 467.5 seconds, and it looks like customers were more willing to answer in the summer months: May, June, July, and August. People less than 60.5 years old were most likely to respond with a duration less than 199.5 seconds. This is the simplest version of the decision tree model.**

****

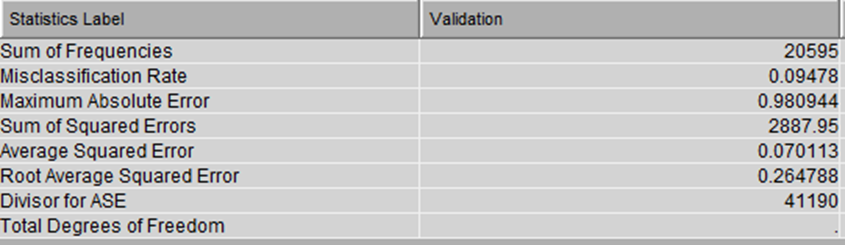
**Fig. 7. Optimal Tree.**

**There is 18 leaf nodes in the optimal tree. It looks like the optimal tree is the best model for the decision tree.**

****

**Fig. 8. Optimal Tree’s subtree assessment plot.**

**We are wrong 9.4 % on validation data of the time based on Misclassification rate and 7 % based on ASE. The misclassification rate is slightly lower than both Maximal and Probability Tree, ASE is higher than others.**

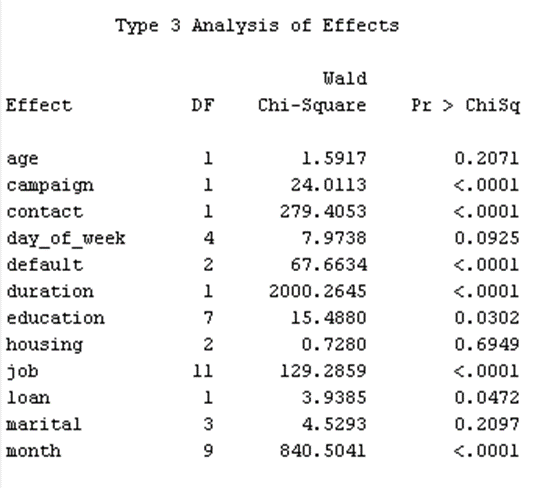
****

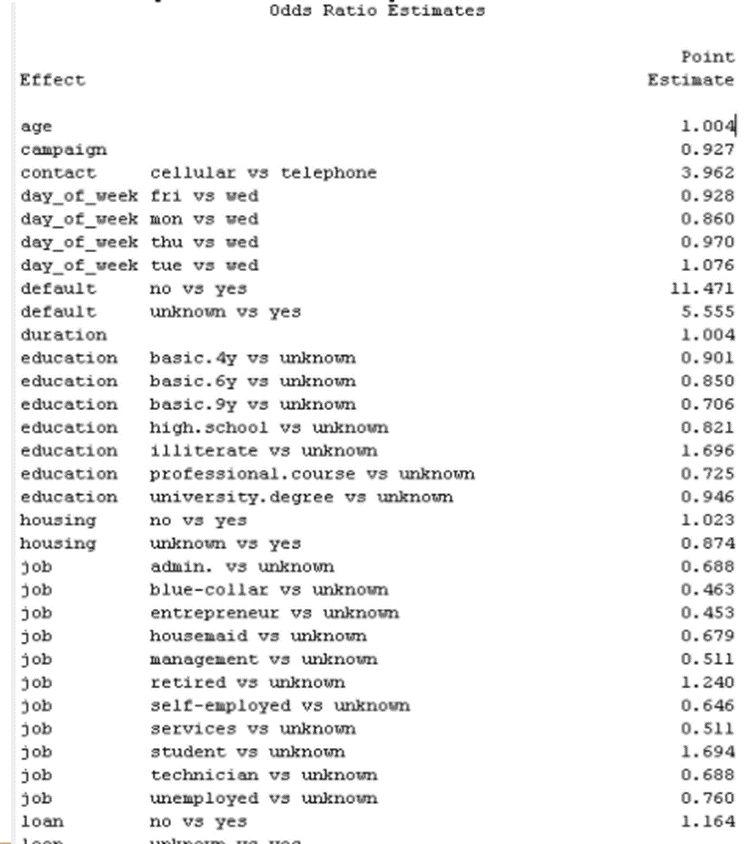
**Fig. 9. Fit Statistics of the optimal tree with the highlight of important numbers.**

**In general, the Optimal tree is the best model in the Decision Tree with the lowest misclassification rate value. However, there is an issue with the Decision Tree with a biased result toward the target variable. The percentage of saying yes and no is unbalanced. As a result, we cannot use the Decision Tree model for our final interpretation.**

**Regression Models are a prolific and useful way to create predictions. New cases are scores using a prediction formula. Inputs are selected via a sequential selection process. Model complexity is controlled by fit statistics calculated on validation data. Three models were implemented for this model: default regression, default stepwise and stepwise misclassification.**

**Default Regression is determined by the target’s measurement level. The type 3 Analysis tests the statistical significance of adding indicated input to a model that already contains other listed input. The important values we can use to answer four questions are campaign, contact, default, duration, education, job, loan, and month.**

****

****

**In this model, age, contact by cellular, not having credit card in default, duration of contact, Tuesday (compared to Wednesday), not a homeowner increases the chance of a customer making deposit.**

**Default Stepwise combines elements from both the forward and backward selection procedures. The default stepwise selects the model from step 7 as the model with optimal complexity. The variables are in the selected model including campaign, contact, default, duration, education, job, and month. These variables are also important in order predict the clients who more likely making deposit as they have the p value less than 0.05.**

**Unknown value means undetermined or nonexistent outcome. It does not make sense to compare a value with an undetermined value. As a result, it can’t be used to predict the target variable in this scenario. The value in odds ratio estimates can help the bank compare the relationship between variables and the target outcome. Here is the important information the bank can learn from the model. If the bank representative targeted customer using cellular, the odd ratio of making deposit is 295.4%. If a client doesn’t have a credit in default, the odd of making deposit is very large. Approach the clients during December, March, and October versus September also has a large impact in making deposit to the bank.**

**Stepwise Misclassification**

**For stepwise misclassification, duration plays an important role along with contact and month in order to increase the change a client decided to make a deposit because the three variables have p value < 0.05.**

**Using cellular as contact mode clearly plays a key role in this model. Increase the duration of call is also increase the term of subscribe a deposit. The stepwise misclassification got the same month’s result as default stepwise.**

**Neural Network**

**Default Neural Network is a default option in SAS enterprise miner. A massive divergence in training and validation misclassification rate occurs near iteration 29. The rapid divergence of the training and validation fit statistics is cause for concern.**

**Fig. 9. Default Neural Network Iteration Plot.**

**Neural Network Misclassification improved the best model from 29 iterations to 17 iterations based on the misclassification rate.**

**Fig. 10. Misclassification Neural Network Iteration Plot.**

**Auto Neural Network use an automatic way to explore alternative network architectures and hidden unit counts.**

**The final model shows the hidden units added at each step and the corresponding value of the objective function. There are 2 hidden units were added at each step.**

**Model Comparison**

**The fit statistics from model of comparison are shown below (). Optimal Tree is the best model based on misclassification rate and Maximal Tree (Decision Tree 4) is the best model based on average squared error. However, if using ROC value to measure the sensitive of the model, default Neural Network is the best model.**

**Fig. 11. Model of Comparison Table**

**Unsupervised technique**

**Clustering attempts to group training data set cases based on similarities in input variables. There is 6 segments in the model, and the 6th segment largely dominated the cases.**

**Fig. 12. Clustering Pie Chart**

**Fig. 12. Clustering Fit Statistics.**

**As seen in the Mean Statistics window, segment frequency counts vary from 3 cases to more than 34,000 cases. The segment 6 is the most important segment as it has the most cases. The range of age in this segment is around 40 years old, the customer has been contacted twice during the campaign and with the duration around 187 seconds.**

**Fig. 13. Segment Profile Plot.**

**The sixth segment has the duration of call and the number of contacts performed for a particular client more than average, whereas the second segment has the duration of call lesson than average, and the third segment has the number of contact is less than average. Clearly, the bank representative should focus on the duration and the number of contact performed in segment 6th as it has the most cases.**

**Fig. 14. Worth Variables Plot.**

**The window shows the relative worth of each variable in characterizing each segment. For example, segment 6 is largely characterized by the duration input, but campaign input still plays a role.**

**Decision Tree**

**The decision tree belongs to categorical prediction. It predicts categorical labels of the dependent variable using other available attributes. In this model, we used three different techniques to analyze the data, including maximal tree, probability tree, and optimal tree.**

**Maximal Tree:**

**Fig. 1. Maximal tree.**

**The tree generated 26 leaf nodes**

**Fig. 2. Subtree assessment plot of maximal tree.**

**The average squared error is 0.069143 and the misclassification rate is 0.094877 which means we predict 6.9 % of the mistake on validation data based on misclassification rate and 9.48 % based on average squared error.**

**Fig. 3. Fit Statistics of the maximal tree.**

**In Probability Tree, the branch selection is the same as the maximal tree. However, it is much simpler and has fewer leaf nodes.**

**Fig. 4. Probability Tree**

**There is 24 leaf nodes in the tree, and the best performance was based on the misclassification rate at 24 leaf nodes.**

**Fig. 5. Probability tree’s subtree assessment plot.**

**We are wrong 9.53 % on validation data of the time based on Misclassification rate and 6.92% based on ASE. The result is slightly higher than the Maximal Tree.**

**Fig. 6. Fit Statistics of probability tree with the highlight of important numbers.**

**Optimal Tree: The selection branch has a duration of fewer than 467.5 seconds, and it looks like customers were more willing to answer in the summer months: May, June, July, and August. People less than 60.5 years old were most likely to respond with a duration less than 199.5 seconds. This is the simplest version of the decision tree model.**

**Fig. 7. Optimal Tree.**

**There is 18 leaf nodes in the optimal tree. It looks like the optimal tree is the best model for the decision tree.**

**Fig. 8. Optimal Tree’s subtree assessment plot.**

**We are wrong 9.4 % on validation data of the time based on Misclassification rate and 7 % based on ASE. The misclassification rate is slightly lower than both Maximal and Probability Tree, ASE is higher than others.**

**Fig. 9. Fit Statistics of the optimal tree with the highlight of important numbers.**

**In general, the Optimal tree is the best model in the Decision Tree with the lowest misclassification rate value. However, there is an issue with the Decision Tree with a biased result toward the target variable. The percentage of saying yes and no is unbalanced. As a result, we cannot use the Decision Tree model for our final interpretation.**

**Regression Models are a prolific and useful way to create predictions. New cases are scores using a prediction formula. Inputs are selected via a sequential selection process. Model complexity is controlled by fit statistics calculated on validation data. Three models were implemented for this model: default regression, default stepwise and stepwise misclassification.**

**Default Regression is determined by the target’s measurement level. The type 3 Analysis tests the statistical significance of adding indicated input to a model that already contains other listed input. The important values we can use to answer four questions are campaign, contact, default, duration, education, job, loan, and month.**

**In this model, age, contact by cellular, not having credit card in default, duration of contact, Tuesday (compared to Wednesday), not a homeowner increases the chance of a customer making deposit.**

**Default Stepwise combines elements from both the forward and backward selection procedures. The default stepwise selects the model from step 7 as the model with optimal complexity. The variables are in the selected model including campaign, contact, default, duration, education, job, and month. These variables are also important in order predict the clients who more likely making deposit as they have the p value less than 0.05.**

**Unknown value means undetermined or nonexistent outcome. It does not make sense to compare a value with an undetermined value. As a result, it can’t be used to predict the target variable in this scenario. The value in odds ratio estimates can help the bank compare the relationship between variables and the target outcome. Here is the important information the bank can learn from the model. If the bank representative targeted customer using cellular, the odd ratio of making deposit is 295.4%. If a client doesn’t have a credit in default, the odd of making deposit is very large. Approach the clients during December, March, and October versus September also has a large impact in making deposit to the bank.**

**Stepwise Misclassification**

**For stepwise misclassification, duration plays an important role along with contact and month in order to increase the change a client decided to make a deposit because the three variables have p value < 0.05.**

**Using cellular as contact mode clearly plays a key role in this model. Increase the duration of call is also increase the term of subscribe a deposit. The stepwise misclassification got the same month’s result as default stepwise.**

**Neural Network**

**Default Neural Network is a default option in SAS enterprise miner. A massive divergence in training and validation misclassification rate occurs near iteration 29. The rapid divergence of the training and validation fit statistics is cause for concern.**

**Fig. 9. Default Neural Network Iteration Plot.**

**Neural Network Misclassification improved the best model from 29 iterations to 17 iterations based on the misclassification rate.**

**Fig. 10. Misclassification Neural Network Iteration Plot.**

**Auto Neural Network use an automatic way to explore alternative network architectures and hidden unit counts.**

**The final model shows the hidden units added at each step and the corresponding value of the objective function. There are 2 hidden units were added at each step.**

**Model Comparison**

**The fit statistics from model of comparison are shown below (). Optimal Tree is the best model based on misclassification rate and Maximal Tree (Decision Tree 4) is the best model based on average squared error. However, if using ROC value to measure the sensitive of the model, default Neural Network is the best model.**

**Fig. 11. Model of Comparison Table**

**Unsupervised technique**

**Clustering attempts to group training data set cases based on similarities in input variables. There is 6 segments in the model, and the 6th segment largely dominated the cases.**

**Fig. 12. Clustering Pie Chart**

**Fig. 12. Clustering Fit Statistics.**

**As seen in the Mean Statistics window, segment frequency counts vary from 3 cases to more than 34,000 cases. The segment 6 is the most important segment as it has the most cases. The range of age in this segment is around 40 years old, the customer has been contacted twice during the campaign and with the duration around 187 seconds.**

**Fig. 13. Segment Profile Plot.**

**The sixth segment has the duration of call and the number of contacts performed for a particular client more than average, whereas the second segment has the duration of call lesson than average, and the third segment has the number of contact is less than average. Clearly, the bank representative should focus on the duration and the number of contact performed in segment 6th as it has the most cases.**

**Fig. 14. Worth Variables Plot.**

**The window shows the relative worth of each variable in characterizing each segment. For example, segment 6 is largely characterized by the duration input, but campaign input still plays a role.**

# **Conclusion and Recommendations**

The bank should approach segmentation of clients around 39 - 40 years old, had been contacted twice during the campaign, and the length of phone call should be around 187.4 seconds, or just over 3 minutes. Anyone who is single was slightly more likely to make a deposit compared to married and divorced couples. Anyone who has credit in default is very unlikely to make a deposit. Not having any loans or not having a house also had a slight effect on deposits. Because the results from our regression model compared age, job, marital, and education to unknown (missing values), the results are not meaningful to interpret. The previous campaign was unhelpful, and the data was excluded from the models. Overall, telemarketing proved to be effective, with the months of December, October and March ( compared to September) being the most effective months, with Mondays and Tuesdays being the most effective days of the week. Our recommendation for the bank is to focus on those key areas where the data showed positive results. Targeting any combination of the following traits will prove helpful: single, non-homeowners, no pre-existing loans, credit not in default, and around 39-40. Focusing on contacting customers during the months of March, October, and December will yield the best results. Additionally, contacting them on Monday or Tuesday, and making sure the calls last around 3 minutes will also be effective.

# References

1. “2010–2014 Portuguese Financial Crisis.” Wikipedia, Wikimedia Foundation, 24 Oct. 2021, https://en.wikipedia.org/wiki/2010%E2%80%932014\_Portuguese\_financial\_crisis.