

School of Computer Science

Data Mining in Fulfilment of

DATA9910

Maksymilian Drzezdzon

C15311966

Degree: TU060/1

Module Coordinator: Brendan Tierney

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: 🗹

**Date: 2020/11/09**

To-Dos

The purpose of this data mining project is to identify customers who are most likely to subscribe to a term deposit account based on previous marketing campaigns. (check course notes, yesterdays)

You are required to produce a report detailing your work investigating the data, building classification models, analysing the results, and comparing your results with the original findings.

The first task you should complete is a data investigation exercise, where you will document the characteristics and other information that you can determine about each Feature.

Identify any data insights discovered and detail all data preparation tasks and any decisions made.

In this tool you can have a number of different classification techniques and within each of these you can modify the various parameter settings.

You will need to evaluate the results from each of the models to determine which of the models gives the best results for you.

You can then compare your results with the original research and discuss the outcomes.

**Deliverables**

You will be required to document your approach to solving and evaluating this classification problem, based on the CRISP-DM process and documentation template guide.

Your report will probably be between 16-20 pages long. The maximum length 20 pages.

The report should clearly show your work in the following areas (similar to CRISP-DM):

- Definition of problem

- Data Exploration and Descriptive Analytics

- Identification of data insights from previous step

- Details of any additional data preparation (cleaning, transformations, etc), data enrichment, feature engineering, feature reduction, etc

- Details of each data mining algorithm used, the configuration settings used, etc

- Details of the evaluation and performance measures from your data mining models. Examine which one performed best, why this might have been the case and how the results compare across all the models

- Discussion of how your results compared to the results from the original research and any conclusions that you can draw from this comparison

**Marking Scheme**

The marking scheme for this assignment is:

- 25% Problem Definition, Descriptive Analytics, Data Insights, etc & summary of initial findings/insights

- 15% Details of any additional data preparation, data enrichment, feature engineering, feature reduction, etc

- 15% Details of each data mining algorithm used, the configuration settings used, etc

- 20% Details & Discussion of the evaluation and performance measures from your data mining models.

- 25% Discussion of how your results compared to the results from the original research and any conclusions that

you can draw from this comparison

Data Preparations

First data was explored in R and SaS enterprise miner. Values that can be dropped are **poutcome** because it’s all unknown, **previous** all values are 0, **pdays** are all -1, **months** is always may, **contact** is mostly unknown, this was done with the code below.

df = read.table("bank/bank-full.csv", header = T, sep = ";")

drop = c("poutcome", "previous", "pdays", "months", "contact")

df = df[,!(names(df) %in% drop)]

unique(df$job , incomparables = FALSE)

unique(df$education , incomparables = FALSE)

names(df)[names(df) == "y"] = "subscribed\_term\_deposit"

df$subscribed\_term\_deposit[df$subscribed\_term\_deposit == "no"] = 0

df$subscribed\_term\_deposit[df$subscribed\_term\_deposit == "yes"] = 1

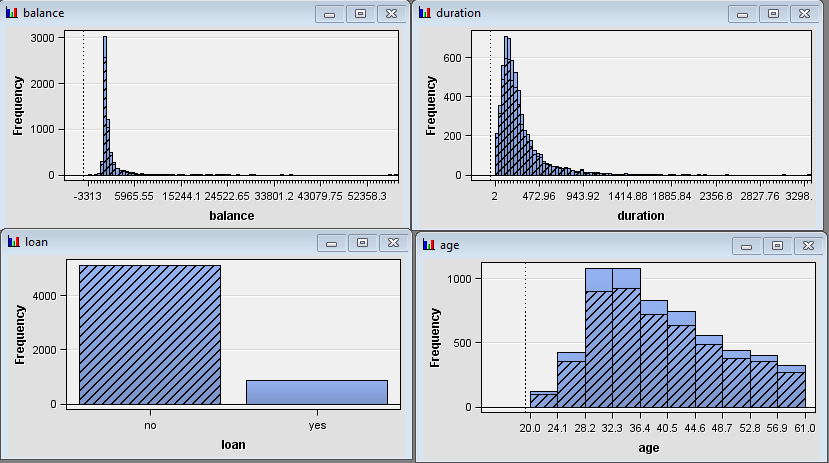
df$job[df$job == "unknown"] = NA

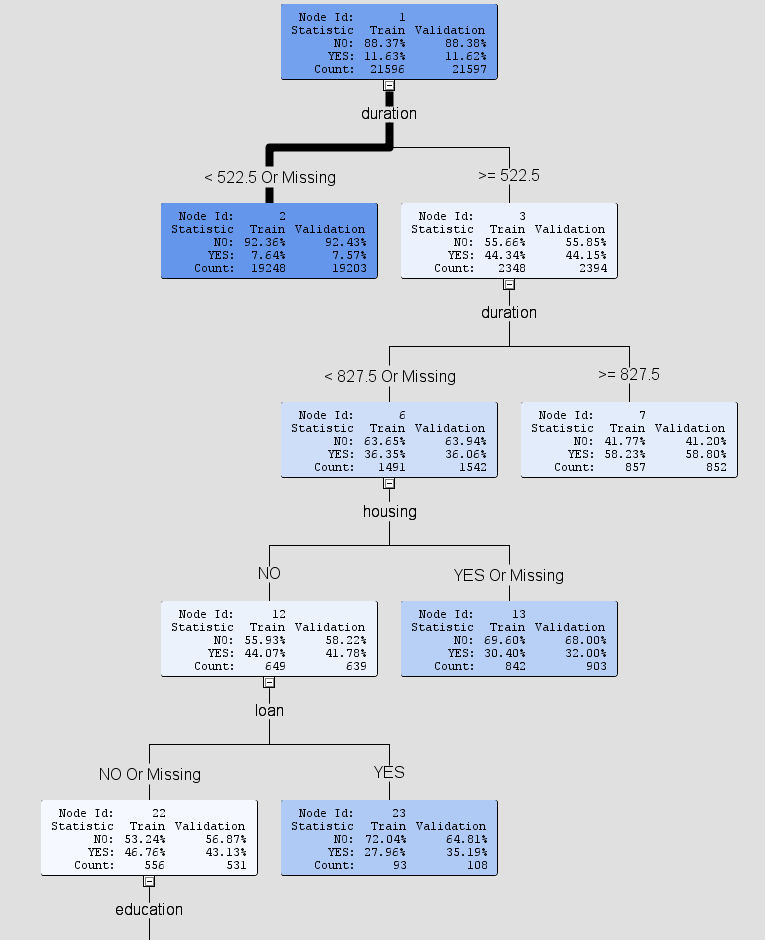
df$education[df$education == "unknown"] = NA

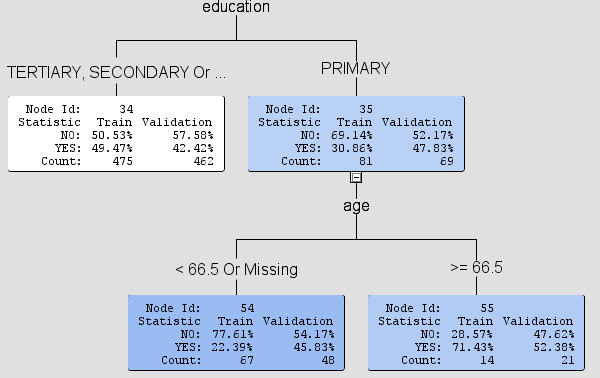
df = na.omit(df)

write.csv(df, "bank-full-updated.csv", row.names = F)

 This decision which variables to keep was made by selecting all variables in SaS and clicking explore which generates visualisations, when you select a column all other graphs reflect this action and show the representation of that column in other graphs, **no** in the loan bar chart was selected.



You can see the representation of all other participants that don’t have a loan across the age, balance and duration bar charts. Another augmentation that was made to what SaS already provides was the target variable **Y** was renamed to “subscribed\_term\_deposit” as its more descriptive in the R code and all yes or no values were converted to 1 or 0 as neural networks perform better on binary data. Bin sizes were increased for duration and balance to get a better read for the representation of clumped up values. After that the data was passed through a decision tree results are as follows 



The thicker the line the more data has been passed in that direction/leaf. The lighter the shade the more impure/bad the data is at predicting the target variable, in this case it being **Y** which is whether the customer accepted the thing. According to the decision tree, the best predicators are a person doesn’t have a house, doesn’t have a loan has a higher education status of above primary school, with longer conversations not materialising into good costumers as one would expect.

Appendix

References