

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.

You will need to work through/develop a number of classification models.

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

Values that can be dropped are poutcome is all unknown, previous all values are 0, pdays are all -1, months is always may, contact is mostly unknown, This was solved by

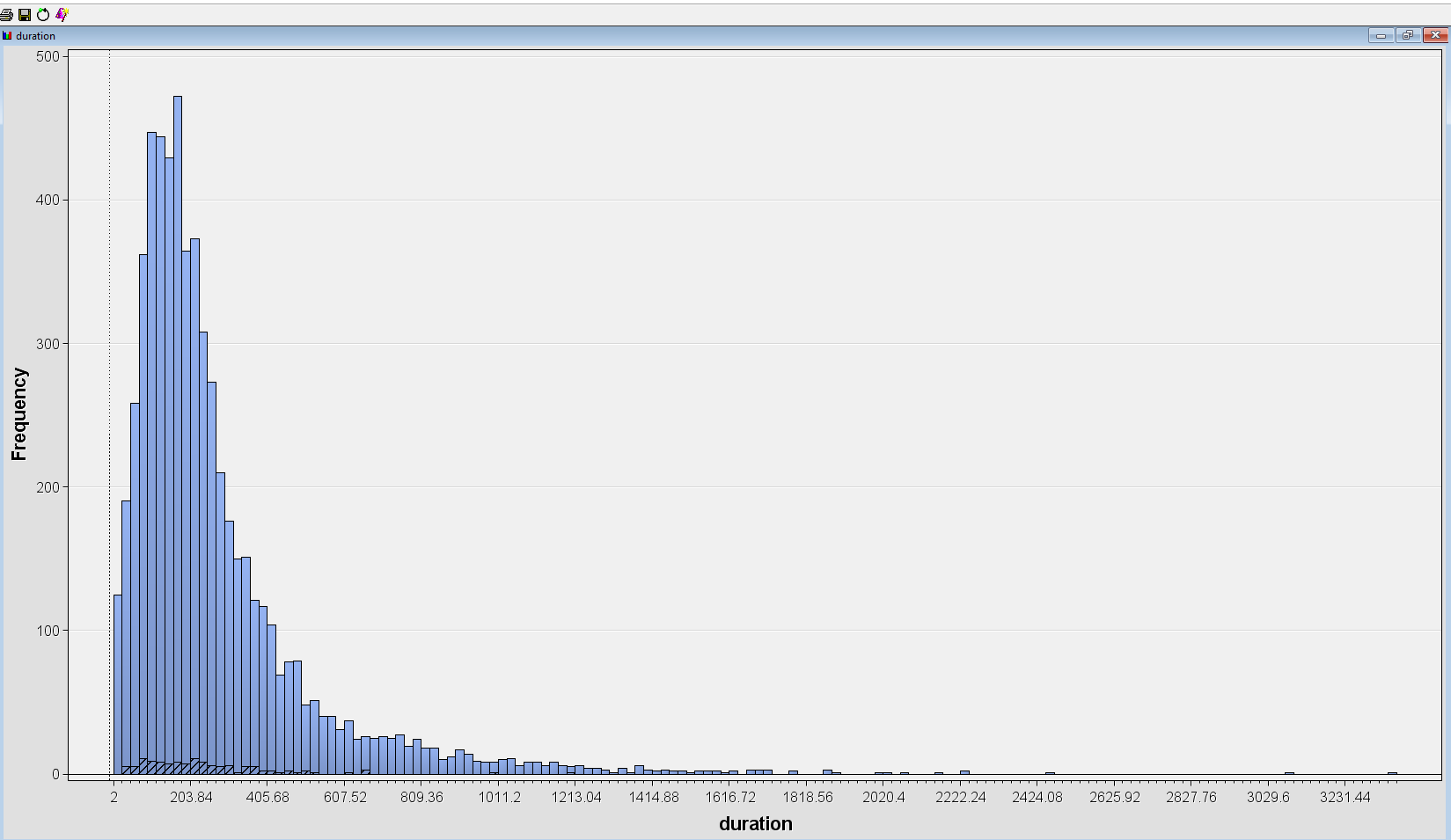
df = read.table("bank/bank-full.csv", header = T, sep = ";")

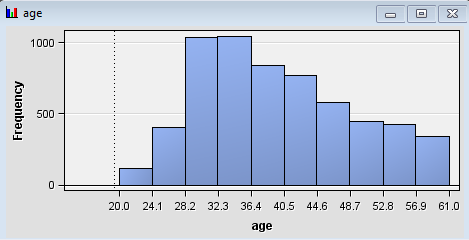
drop = c("poutcome", "previous", "pdays", "months", "contact")

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

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

 y is the target variable as that is what you are looking for.





Appendix

References