Bayesian Networks Assignment 1

Daan Brugmans S1080742
Maarten Berenschot S1017215

I. Introduction

This paper contains a realization of the Assignment 1 project for the Radboud University (2023) course "Bayesian Networks and Causal Inference". This project focuses on applying techniques for constructing directed acyclic graphs (DAGs) and Bayesian networks on a real-world dataset. Students must construct a DAG to model the relationships between the variables in the dataset, then analyze their model statistically, draw conclusions, and make an updated version of their model.

This realization of the project uses a preprocessed dataset from a paper written by Moro et al. (2014) on the topic of bank marketing. Specifically, the paper in question is about using bank marketing data for constructing a prediction algorithm that could predict whether a customer is expected to subscribe to a long-term deposit at the bank or not. The results of this paper were aggregated in the dataset used for this project. This project focuses on using these results for *causal inference* purposes, not prediction purposes.

II. Problem Domain

Many banks wish to acquire more new long-term deposits from their customers. To that purpose, banks try out various methods for reaching out to customers that might be interested in starting a new long-term deposit. One such method is calling potential new depositors via phone.

Unfortunately, this method far from guarantees that the potential customers reached will actually subscribe to a term deposit. Preferably, we aim to contact only those potential depositors that have the highest likelihood of subscribing to a new term deposit. In an attempt to model the type of customer that would be most likely to subscribe to a new term deposit, multiple campaigns were undertaken. In these campaigns, current customers of the bank were contacted via phone and were asked about their interest in subscribing.

For these customers, this paper postulates the following: *which characteristics are the most relevant for determining the likelihood of a current bank customer subscribing to a new long-term deposit?*

III. Data

Table I describes the variables of the data used for modelling the initial DAG. For each variable, their name, data type and cardinality/range are given. The target feature for the model is shown at the end of the table.

Variable Name	Type	Cardinality	Range
Age	Continuous	-	[18, 95]
Annual Balance	Continuous	-	[-8.019, 102.127]
Call Duration	Continuous	-	[0, 4.918]
Current Campaign Calls	Continuous	-	[1, 63]
Education Level	Ordinal	3	-
Has Default	Binary	2	-
Has Housing Loan	Binary	2	-
Has Personal Loan	Binary	2	-
Job Category	Categorical	11	-
Marital Status	Categorical	3	-
Previous Campaigns Calls	Continuous	-	[0, 275]
Previous Campaign Outcome	Categorical	3	-
Has Subscribed to Deposit	Binary	2	-

Table I
Description of the preprocessed bank marketing dataset's variables.

IV. Network

All code written for the realization of the networks may be found in the following repository:

https://github.com/daanbrugmans/ru-bayesian-networks-and-causal-inference-23-24/tree/main/Assignments/Assignment%201

A. Graph

Figure 1 describes the manually constructed DAG based solely on domain knowledge, without the use of the dataset or the spread of the data. The direction of the edges was most often determined by chronology of the variables: one's age comes first, then their education, then their job, then their income, etc.

Figure 3 describes an altered version of the manually constructed DAG that shows altered nodes and edges. These changes were made by testing the (in)dependencies of the manually constructed DAG on the banking data and making alterations based on canonical correlations testing.

B. Preprocessing

In order to prevent incorrect statistical test results from influencing the result of the research, a set of data preprocessing steps were undertaken that should improve the quality and consistency of the dataset and make it more understandable and relevant for the purposes of this project.

The first preprocessing step was the filtering and renaming of variables. The original, unaltered dataset contained variables that, for the purposes of limiting the scope of this project, were not included in the DAG; features describing dates and times were excluded, for example. Furthermore, the remaining

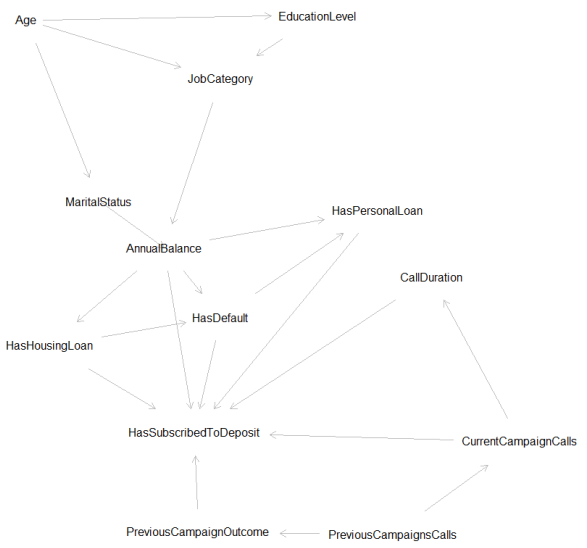

Figure 1. The manually constructed DAG based solely on domain knowledge.

variables were renamed to be more descriptive, as some of the original variable names did not capture the full meaning of the feature. For example, the variable ”previous” described the number of calls performed to a banking customer for a previous campaign aimed at acquiring new long-term deposit customers. This variable was renamed to ”Previous Campaign Calls”. The variable names shown in table I are the renamed versions of the variables.

The next step of the preprocessing step was to discard records with an ”unknown” value for the ”Education Level” and ”Job Category” variables. The ”unknown” value is considered to be a missing value, and records with that value for this variable are considered to have missing values. For the purposes of limiting the scope of this project, instead of applying an imputation algorithm, these records were discarded.

Next, all binary variables represented with strings were converted to be represented with integers. That is, ”yes” became 1 and ”no” became 0. Finally, all continuous variables were normalized to a mean of 0 and a standard deviation of 1. This is to prevent imbalanced variable scales from negatively influencing the fitting of the model.

It should be noted that, although no further data preprocessing was performed prior to testing and fitting, during the fitting process, canonical correlations are used to calculate coefficients of edges. In this process, features are dummy encoded, which may be perceived as a data preprocessing step.

C. Implementation

The preprocessing steps described earlier, alongside other code that had to be written for this project, was written in the R programming language (The R Foundation, n.d.). The main reason motivating this choice is that the R language has extensive support for building DAGs, fitting Bayesian networks and testing conditional independencies, some of which were used for this project.

To create the DAG, the DAGitty package and platform by Textor et al. (2016) were used. The DAGitty package allows the construction of DAGs and is capable of testing independencies in the graph using d-separation. For this project, DAGitty was used for these purposes, in addition to the calculation of canonical correlations, fitting the DAGs to the data, and comparing independencies in the DAG to independencies implied by the canonical correlations.

To perform the preprocessing of the data, packages from the package collection tidyverse (Wickham et al., 2019) were used, including ggplot2 (Wickham, 2016) for visualizing the dataset, dplyr (Wickham et al., 2023) for data wrangling, and stringr (Wickham, 2022) for processing string data.

D. Testing

Statistical tests were performed using DAGitty’s (Textor et al., 2016) $localTests$ function with the ”cis.pillai” test. Using DAGitty, canonical correlations between variables were calculated that were used to prove and disprove independencies both present and missing in the DAG. The result of this testing is a list of found independencies in the DAG and the statistic calculated on the data expressing the strength of the relationship between the variables. Table II shows the list of independencies found in the DAG in figure 1.

Some new observations were found from this analysis. For example, in the initial DAG from figure 1, there existed no direct dependency between age and annual balance. This dependency was only modeled indirectly, via job category. However, pre-fit testing showed that there is actually a noteworthy direct correlation between age and annual balance.

Additionally, the strength of some existing dependencies could be disproven. For example, it was found that the dependency between ”Marital Status” and ”Annual Balance” was relatively small. It was assumed prior to testing that the marital status of a customer would significantly influence their annual income, since customers in marriage might have two persons’ worth of incomes to their disposal. However, this does not reflect in the data. It may be concluded that the ”Annual Balance” variable was recorded per person, excluding total household income.

The result of the statistical testing, and the findings that could be drawn from it, was used to alter the initial DAG. Specifically, the estimates of the found independencies were used to construct new edges between existing nodes of the initial DAG and to remove edges of independencies that could not be proven. Only edges with an estimate outside of the range $(-0.06, 0.06)$ were chosen to be included in the altered DAG. This range was chosen as to balance the amount of edges in the DAG: excluding all coefficients in this range resulted in a DAG that showed near exclusively relevant relationships between variables. This process was repeated multiple times until the DAG contained all coefficients outside the range of $(-0.06, 0.06)$ that could be found using canonical correlations, while excluding all coefficients inside this range. Figure 3 shows this altered, final DAG, fitted to the data with canonical correlations. Figure 2 shows how the manually constructed DAG of figure 1 would fit to the data.

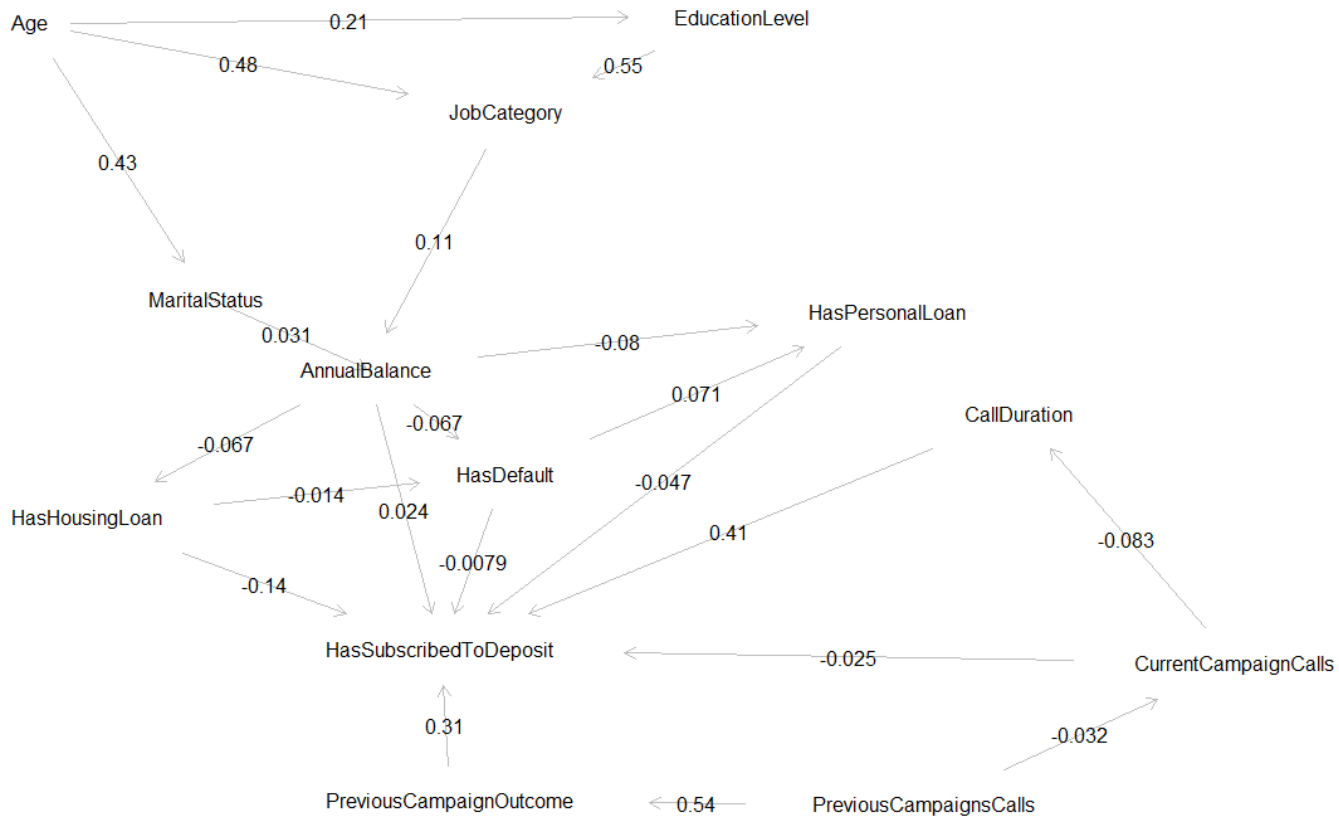

Figure 2. The initial DAG from figure 1 fitted to the dataset using canonical correlations.

E. Discoveries

One astounding discovery was that the average annual balance seemed to have little effect on the likelihood of a customer subscribing to a new long-term deposit. This goes against initial expectations, assuming that the opposite was to be the case. Two other variables that were expected to have a major influence on a customer making a deposit that proved not to have such an influence after testing are ”Has Default” and ”Has Personal Loan”. These variables, alongside ”Annual Balance”, seem not to be notably correlated with most other variables; only ”Education Level” and ”Age” have a noteworthy correlation with ”Annual Balance”. This discovery very much goes against the original expectations set by the initial DAG.

The DAG in figure 3 shows that four variables have a coefficient large enough to be considered relevant when determining whether a customer was likely to subscribe to a long-term deposit. These are ”Job Category”, ”Has Housing Loan”, ”Previous Campaign Outcome”, and ”Call Duration”. Here, ”Call Duration” has the highest correlation coefficient, closely followed by ”Previous Campaign Outcome”. These two variables, with coefficients of 0.41 and 0.31 respectively, have noticeably higher coefficients than the variables ”Has Housing Loan” (-0.12) and ”Job Category” (0.093). From these findings, we may conclude the following:

- The longer a customer is called for a campaign, the more likely it is they will open a long-term deposit.
- The more successful previous campaigns were, the more likely it is that a customer will open a long-term deposit.
- If a customer has a housing loan, it is less likely that they will open a long-term deposit.
- A customer’s job type is important to consider when estimating the likelihood of them subscribing to a long-term deposit.

A number of additional notable findings are as follows:

- Age is strongly correlated with what type of job a customer performs.
- The older a customer is, the more likely it is that they have participated in higher levels of education.
- The level of education is strongly correlated with a job type of a customer. In fact, it is the strongest correlation found in the DAG of figure 3.
- A customer’s characteristics most relevant in determining their annual balance are their age and highest finished education level.
- A customer’s job category is clearly correlated with whether they have a housing loan.
- The ”Previous Campaign Calls” and ”Previous Campaign Outcome” variables are strongly correlated, implying that performing campaign calls has a major effect on the success of a campaign.
- The more successful campaigns were held in the past,

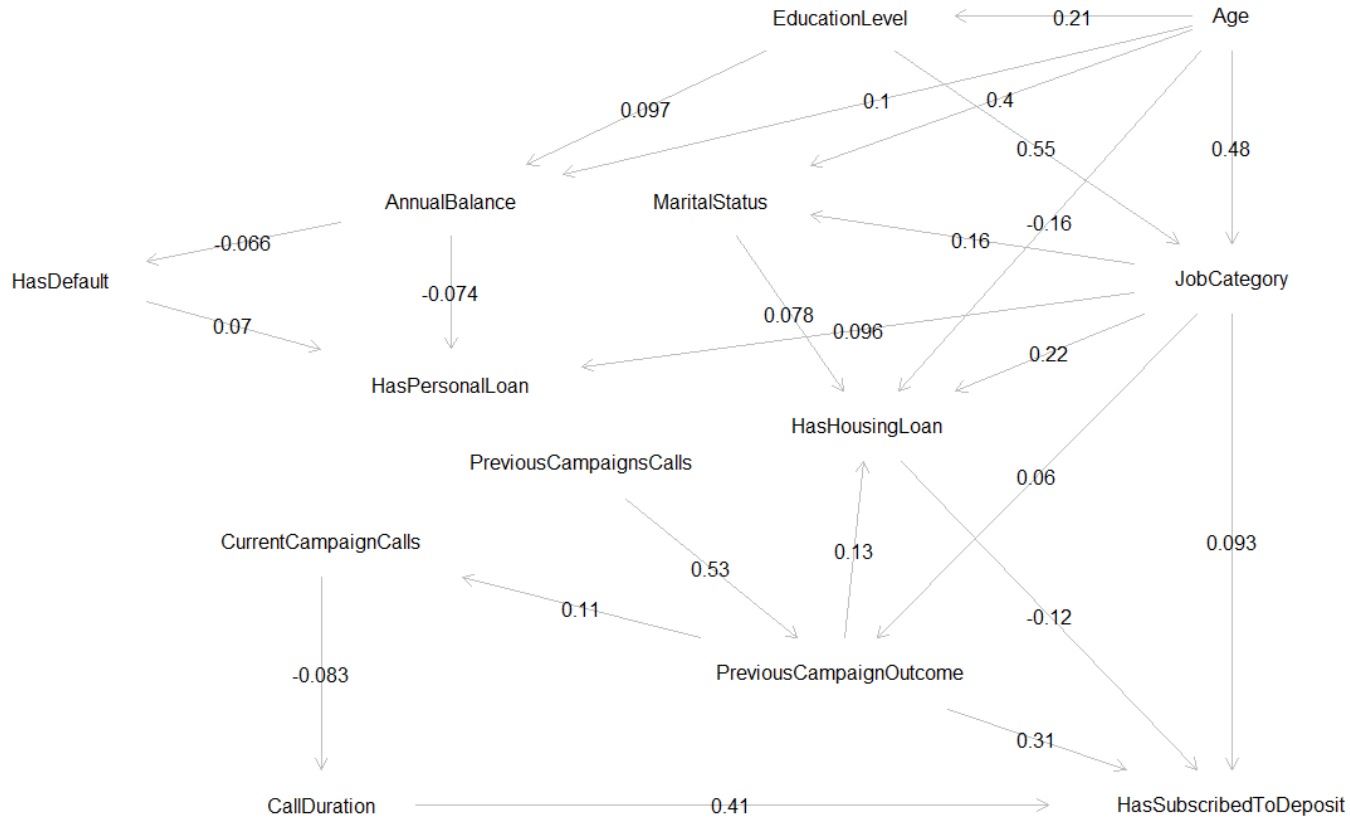

Figure 3. The tested DAG, whose nodes and edges were altered form the initial DAG from figure 1 using canonical correlations.

the more likely it is that a customer will subscribe to a deposit. This is likely customers being motivated by the successes of previous campaigns.

- The longer a campaign call to a customer takes, the more likely it is that the customer will subscribe to a deposit.

V. Research Question

The main research question posed for this project is as follows:

> *Which characteristics of the call and the callee influenced whether the callee would open a new term deposit the most?*

Alongside a supplementary research question:

> *To what extent did the average yearly household balance of the callee and the duration of the campaign call influence whether the callee would open a new term deposit or not?*

In order to answer these questions, this paper turns to the fitted DAG in figure 3. The fitted initial DAG in figure 2 shows that average annual balance has only a very small influence on whether a customer is likely to subscribe to a long-term bank deposit. In fact, this influence seems so small, that the relationship between the two was not included in the final DAG of figure 3. The duration of a campaign call, however, does have a major influence. In fact, it is the variable with the highest correlation to a customer subscribing. It is trivial to conclude that, when provided with the fitted DAG, the average yearly household balance of a banking customer has little influence on the likelihood of that customer subscribing to a long-term bank deposit, while the length of a campaign call has a strong influence on that likelihood.

Analyzing the fitted DAG, it may be concluded that the characteristics of the call have a greater influence on the likelihood of a customer subscribing to a long-term deposit than the characteristics of the callee: the variables ”Call Duration” and ”Previous Campaign Outcome” have the strongest correlations with that likelihood, while ”Has Housing Loan” and ”Job Category” are the only variables that are characteristics of the callee and have any notable correlation on the subscription likelihood. This means that campaigns of calling customers, encouraging them to subscribe to long-term deposits, are important in getting the customer to subscribe. If the banking firm wants to take into consideration what type of customers should be contacted, they should avoid customers with housing loans, and take the customer’s job category into consideration.

Further research should explore the differences between the individual job categories and explain which job categories have the customers that are most likely to subscribe to a long-term deposit.

VI. Discussion

An important point worth discussion is the fact that the constructed Bayesian network is not suited for prediction

Found Independency	Estimate	P-value
Age ⫫ AnnB | JbCt, MrtS	0.0881	3.9e-75
Age ⫫ CllD	-0.0050	3.0e-01
Age ⫫ CrCC	0.0040	4.0e-01
Age ⫫ HsDf | AnnB	-0.0098	4.1e-02
Age ⫫ HsDf | JbCt, MrtS	-0.0175	2.7e-04
Age ⫫ HsHL | AnnB	-0.1799	7.5e-311
Age ⫫ HsHL | JbCt, MrtS	-0.1570	1.9e-236
Age ⫫ HsPL | AnnB	-0.0018	7.1e-01
Age ⫫ HsPL | JbCt, MrtS	-0.0300	4.3e-10
Age ⫫ HSTD | AnnB	0.0199	3.6e-05
Age ⫫ HSTD | JbCt, MrtS	0.0256	1.1e-07
Age ⫫ PrCO	0.0382	1.3e-13
Age ⫫ PrCC	0.0011	8.2e-01
AnnB ⫫ CllD	0.0201	3.1e-05
AnnB ⫫ CrCC	-0.0163	7.3e-04
AnnB ⫫ EdcL | Age, JbCt	0.0534	0.0e+00
AnnB ⫫ EdcL | JbCt, MrtS	0.0452	0.0e+00
AnnB ⫫ PrCO	0.0391	2.8e-14
AnnB ⫫ PrCC	0.0166	5.8e-04
CllD ⫫ EdcL	0.0035	7.7e-01
CllD ⫫ HsDf	-0.0110	2.2e-02
CllD ⫫ HsHL	0.0040	4.0e-01
CllD ⫫ HsPL	-0.0132	6.0e-03
CllD ⫫ JbCt	0.0397	1.1e-10
CllD ⫫ MrtS	0.0229	1.2e-05
CllD ⫫ PrCO | PrCC	0.0448	0.0e+00
CllD ⫫ PrCO | CrCC	0.0445	0.0e+00
CllD ⫫ PrCC | CrCC	-0.0024	6.2e-01
CrCC ⫫ EdcL	0.0208	8.7e-05
CrCC ⫫ HsDf	0.0164	6.3e-04
CrCC ⫫ HsHL	-0.0256	1.1e-07
CrCC ⫫ HsPL	0.0104	3.0e-02
CrCC ⫫ JbCt	0.0555	0.0e+00
CrCC ⫫ MrtS	0.0305	2.0e-09
CrCC ⫫ PrCO | PrCC	0.1114	0.0e+00
EdcL ⫫ HsDf | AnnB	0.0103	1.0e-01
EdcL ⫫ HsDf | JbCt, MrtS	0.0238	4.8e-06
EdcL ⫫ HsDf | Age, JbCt	0.0244	2.5e-06
EdcL ⫫ HsHL | AnnB	0.1071	0.0e+00
EdcL ⫫ HsHL | JbCt, MrtS	0.0472	0.0e+00
EdcL ⫫ HsHL | Age, JbCt	0.0559	0.0e+00
EdcL ⫫ HsPL | AnnB	0.0572	0.0e+00
EdcL ⫫ HsPL | JbCt, MrtS	0.0395	2.1e-15
EdcL ⫫ HsPL | Age, JbCt	0.0409	2.2e-16
EdcL ⫫ HSTD | AnnB	0.0697	0.0e+00
EdcL ⫫ HSTD | JbCt, MrtS	0.0500	0.0e+00
EdcL ⫫ HSTD | Age, JbCt	0.0572	0.0e+00
EdcL ⫫ MrtS | Age	0.0841	0.0e+00
EdcL ⫫ PrCO	0.0432	0.0e+00
EdcL ⫫ PrCC	0.0250	1.3e-06
HsDf ⫫ JbCt | AnnB	0.0354	4.3e-08
HsDf ⫫ MrtS | AnnB	0.0170	2.0e-03
HsDf ⫫ PrCO	0.0401	5.8e-15
HsDf ⫫ PrCC	-0.0179	2.0e-04
HsHL ⫫ HsPL | AnnB, HsDf	0.0326	1.3e-11
HsHL ⫫ JbCt | AnnB	0.2638	0.0e+00
HsHL ⫫ MrtS | AnnB	0.0211	6.8e-05
HsHL ⫫ PrCO	0.1434	0.0e+00
HsHL ⫫ PrCC	0.0368	2.0e-14
HsPL ⫫ JbCt | AnnB	0.0964	0.0e+00
HsPL ⫫ MrtS | AnnB	0.0538	0.0e+00
HsPL ⫫ PrCO	0.0564	0.0e+00
HsPL ⫫ PrCC	-0.0114	1.8e-02
HSTD ⫫ JbCt | AnnB	0.1307	0.0e+00
HSTD ⫫ MrtS | AnnB	0.0677	0.0e+00
HSTD ⫫ PrCC | CrCC, PrCO	0.0084	7.9e-02
JbCt ⫫ MrtS | Age	0.1551	0.0e+00
JbCt ⫫ PrCO	0.0643	0.0e+00
JbCt ⫫ PrCC	0.0418	3.7e-12
MrtS ⫫ PrCO	0.0299	1.2e-14
MrtS ⫫ PrCC	0.0169	2.1e-03

Table II
AN OVERVIEW OF ALL INDEPENDENCIES FOUND BY TESTING THE INITIAL DAG TO THE DATA

purposes, only for causal inference. This was an intentional design decision and is relevant due to the variables used to construct the network. In the original paper by Moro et al. (2014), the data was used to predict the likelihood of a banking customer subscribing to a long-term deposit. This was used to determine which customers the banking company should contact, based on a set of characteristics describing the customer within a certain demographic. These contacts were part of a marketing campaign and were also stored as data. This data is part of the dataset used for this project. This means that some of the variables in the fitted Bayesian network cannot be used for prediction purposes. For example, the "Call Duration" variable describes the length of a call made to a potential customer in seconds. If the fitted Bayesian network would be used to predict whether a banking customer would subscribe to a new long-term deposit or not, it would be required that the length of a call to that customer would be known in advance, prior to actually calling the customer, since the prediction task is used to determine whether it would be worth to call the customer. However, this is impossible to know in advance. This makes the Bayesian networks in its current state unsuited for prediction purposes.

Worth mentioning are the limitations of the project and its data. Since the dataset used for this paper originates from research done by Moro et al. (2014), the data is limited in scope. For example, the research performed by Moro et al. (2014) was done using data from a Portuguese banking company. This means that the findings for this assignment should be interpreted with the following knowledge:

- Since all data comes from customers of a specific banking company, there may exist a bias in the data that is not found in more general populations. Customers of other banks may exert different behaviour and preferences, for example, which could make the findings of the assignment less generalizable to the population of all Portuguese banking customers.
- By extension, since this data only considers Portuguese banking customers, the findings may not be generalizable to customers of banks of other countries. Financial and economic behaviour can vastly differ between nations and their people: financial behaviour of Portuguese banking customers may not be representative of Norwegian, American, Chinese and Angolan customers' behaviour.
- Furthermore, the data used was collected in the period of 2008 to 2013 (Moro et al., 2014). Since financial behaviour of banking customers is influenced by the current state of banking, the economy and society as a whole, the behaviour of customers in this dataset may not be properly representative of Portuguese banking customers of the current age.

To wrap up this assignment, this paper ends with a brief discussion of the success of the fulfillment of the assignment. In short, the research questions could successfully be answered, although the answers partially went against initial expectations. It was expected that one's annual household balance would greatly influence whether they would open a long-term deposit, yet one's job category seemed more relevant than that.

That was an interesting discovery worth exploring further. One possibility is that there is something more abstract, related to one's own personality and decision-making preferences, that influence a multitude of variables, that in turn influence the subscription likelihood. Perchance that would be one's inherent personality or conditions growing up. This may be an interesting topic for latent variable analysis. It is clear to state that the conclusion of the research were intriguing and worth exploring further.

REFERENCES

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Radboud University. (2023). NWI-IMC012 Bayesian Networks and Causal Inference. https://www.ru.nl/courseguides/science/vm/osirislinks/imc/nwi-imc012/

Textor, J., van der Zander, B., Gilthorpe, M. S., Liskiewicz, M., & Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, *45*(6), 1887–1894. https://dagitty.net/

The R Foundation. (n.d.). What is R? https://www.r-project.org/about.html

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2022). *Stringr: Simple, consistent wrappers for common string operations* [https://stringr.tidyverse.org, https://github.com/tidyverse/stringr].

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation* [https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr].