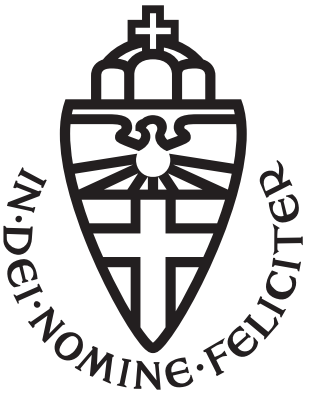
Bayesian networks



Radboud University Nijmegen

Using data to maximize profits for marketing campaigns

Authors:
Niek Janssen (s4297091)
Laurens Kuiper (s4467299)
Ward Theunisse (s4492765)

December 2018

Contents

1 Introduction

1.1 Problem domain

Companies engage in direct marketing campaigns by contacting potential customers to reach a specific goal (e.g. selling a product or service). Communicating is often done by telephone, initiated from the company's contact center. One of the biggest challenges the companies face is estimating what people are most likely to buy their product: many calls result in customers declining the offer. For a company to have a successful marketing campaign, a sufficient number of customers must follow up on their offer, otherwise it will cost more than it will bring in.

Companies typically collect data about their campaign to improve their success. In this project, we will use publicly available data from one such campaign to build a Bayesian network model about the factors involved in bank marketing.

1.2 Data

We use a publicly available bank marketing dataset retrieved from the UCI Machine Learning Repository [1]. This dataset contains data related to a direct marketing campaign by a Portuguese banking institution based on phone calls. Each row in the table corresponds one instance of telephonic contact with a potential customer in the latest marketing campaign. Our Bayesian network will contain the following variables related to a potential customer: Age, Eduction, Job, Marital Status, whether they currently have a Housing Loan, whether they have Credit in Default, and whether they have a Personal Loan. It will also contain the following variables related to the call: the Contact type (whether the potential customer used a mobile phone, landline), Contact Month, Contact Weekday, amount of Days Since Last Contact, Previous campaign outcome, Number of Contacts in this Campaign, Contact Duration, and the Outcome of the call (whether the customer subscribes or not). Lastly it will also contain these national economic variables: Consumer Price Index, Consumer Confidence Level, 3 Month Euribor (European banking loan) rate, Employment Rate, Employment Variation.

1.3 Research questions

We are mainly interested in which variables influence the customer's decision subscribing to a term deposit. We will be able to learn which variables affect the decision, and how large their effect size is. Using the final model, we can advise the bank about:

- Which people should the bank target? For example, people with certain types of jobs or education might be more susceptible to a term deposit.
- Given a certain person, what is the best time

to call this him/her? For example, students do not like being called early in the morning.
- How important is the current state of the economy for making this decision? Should the bank target different people as the economy evolves?

1.4 Implementation

All implementation was done in `R`, making use of the `dagitty`, `bnlearn` and `lavaan` packages, for model testing, structure learning, and structural equation models respectively.

2 Initial causal diagram

Our initial DAG as shown in Figure 1 was constructed with our prior beliefs about the data. A bigger version of this DAG is shown in the appendix (Appendix A Figure 10).

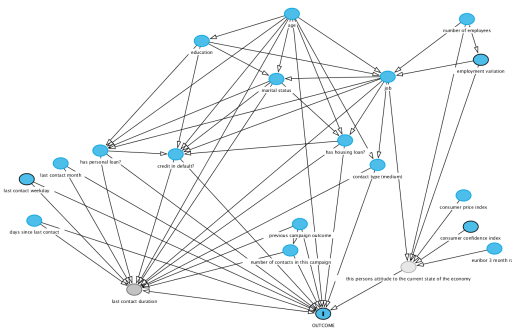


Figure 1: Initial DAG as proposed in exposee.

3 Testing the causal diagram

In this report we present the findings of testing and amending the initial model as proposed in our exposee. The model was iteratively tested using the χ^2-test, removing and adding variables/connections at each step.

Conditional independencies were automatically derived using the `ImpliedLocalDependencies` function from the `dagitty` package in `R`. The result of this can be found in Appendix D.

The function found a large amount of independencies, most of which are associated with either economic variables, contact information, or with the previous campaign. This is because we assumed these variables would mainly have an effect on the outcome, but not on personal attributes, with the exception of economic variables. In this report we put a big emphasis on economic variables, because we realized that this affects the bank heavily.

3.1 Approach

Tests were performed automatically using the `localTests` function from the `dagitty` package. However, the data has some attributes that can take on many different values. This would cause issues, which is why we performed some preprocessing. These are the attributes for which this was the case:

variable	**type**
age	Discrete values between 18-85
nr.employed	continuous values
emp.var.rate	continuous values
cons.price.idx	continuous values
cons.conf.idx	continuous values
euribor3m	continuous values

All other attributes were nominal values related to social and economical status of the person.

3.2 Data analysis

Because `dagitty` interprets our data as nominal, it is necessary to bin our data to allow for meaningful analysis. For each of these attributes, we inspect the histogram of value frequencies, and we apply domain knowledge to obtain acceptable bins.

Binning
We binned the data in following bins:

Variable	**Split on**
Age	18, 25, 30, 45, 60
Consumer confidence	35, 38, 45
Price index 1000	92.6, 93.3, 93.6, 94.3
Empl. var. rate	-0.5, 0.5, 2.1
Euribor 1000	3
Nr. employed	5050, 5150

For age, shown in section 3.2, we used our domain knowledge to choose stages of life ($< 18, 18-25, 26-30, ..., > 60$), whereas for the economic variables we chose bins that seemed fit the data. For euribor, shown in Figure 6, we only chose 2 bins, because there was a big gap in the middle of the range, resulting in 2 clusters.

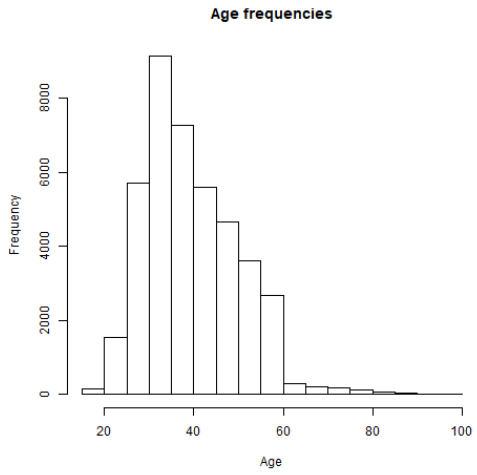


Figure 2: Distribution of age

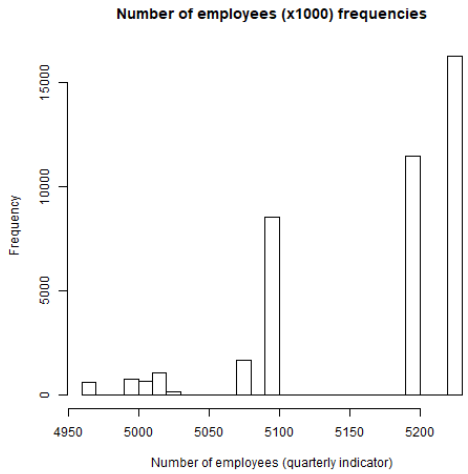


Figure 3: Distribution of number of employees

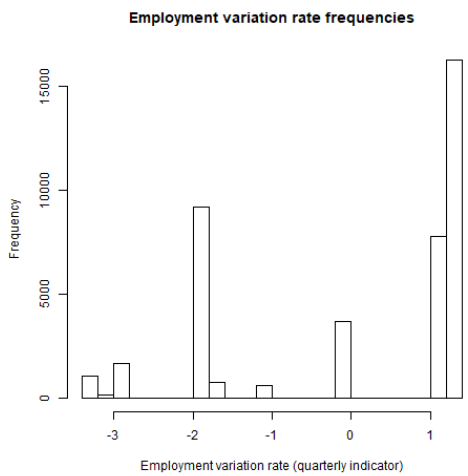


Figure 4: Distribution of employment variation

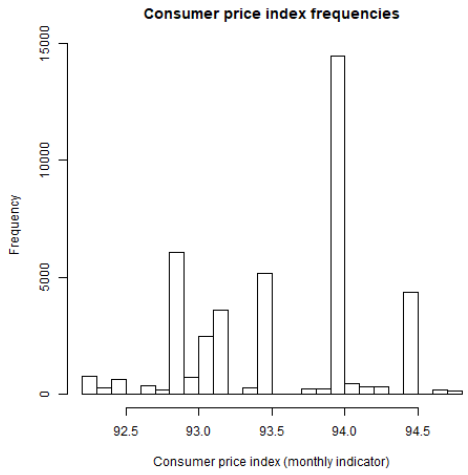


Figure 5: Distribution of consumer price index

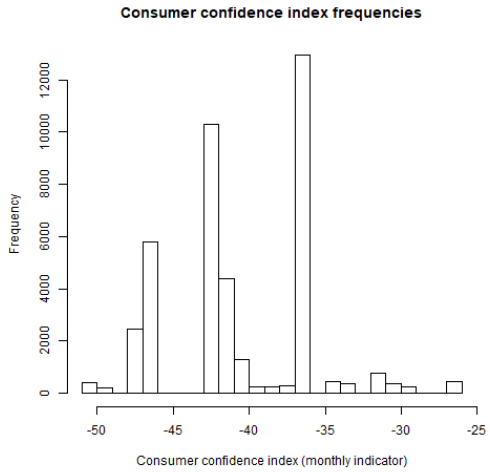


Figure 6: Distribution of consumer confidence index

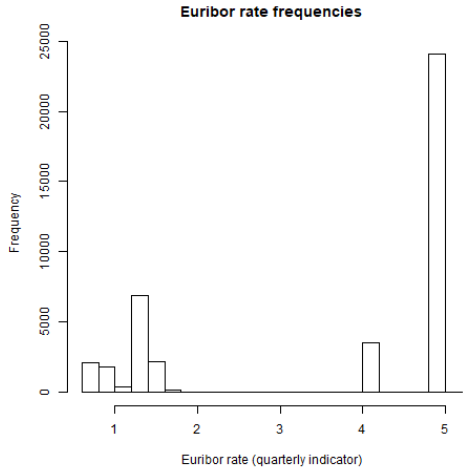


Figure 7: Distribution of three month euribor rate

Testing strategy
An iterative strategy was applied. In each iteration we ran `localTests`, starting with parameter `max.conditioning.variables` set to 0. We focused on the highest RMSEA value, applying a change to the model each time, and running the function again. This was repeated until we were satisfied (RMSEA values < 0.05, unless independency was deemed illogical), at which point we incremented the maximum amount of conditioning variables by 1, and repeated the process.

3.3 Changes

A changelog of the findings of our testing strategy is described in the subsections in this section.

Contact type
When we first tested the independencies of our DAG, we found a very strong dependency between the contact type variable (indicating whether the consumer reached on their cellphone or landline) and the month. After a quick look, this seemed to be corresponding to the data we have, so we added an arrow from month to contact type.

We also found a strong dependency between default and contact medium, conditioned on age and job. Because the month is strongly influential on the contact type, we deemed it likely that the month would have a connection with default. We thought it would also be logical that some months would be more expensive than others, so in these months people are more likely to have a default.

We also found a dependency from education to contact. We might be able to explain this because people with certain jobs need a mobile phone to work efficiently.

Later on in the process, we found that the contact type was continually the same over time, with a single point in time when it switched. The first two years of the data collection was done via landline, while the last three years all calls were made via a mobile phone. Because this is skewed so heavily, we decided to delete the contact type attribute altogether.

Latent variables
Education and duration were dependent on each other, conditioned on five other variables. Because of this, added a latent variable called patience. This latent variable disappeared again when we deleted duration. We removed duration because it highly affects the output target, but it is only known after a call is performed, thus not offering any predictive / actionable information. The authors of the paper that our data features in state that " this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model."

We also found a dependency between loan and housing, conditioned on age, job and marital status. Another dependency between education and y exists. Because of this, we add a latent variable "willingness to take a loan". This latent variable has connections from age, marital status and education, and has connections to housing, loan and y. All direct connections between these variables were removed. This latent variable later became wealth, where it has got connections to marital status and default.

At first, we thought we could remove patience because the latent variable wealth would suffice, but this wasn't the case. We could, however, replace it with a direct connection between education to duration.

More deletions
We read from the source of the dataset that duration is so closely relate to y (outcome), that it should not be included in predictive models. This is because, when people decide that they want a loan (y=true), the call becomes much longer because of the further administration that has to be done over the phone. This made us delete the duration altogether.

Additionally, we deleted the attributes day_of_week and campaign because they had no predictive value.

Economical variables
Because we are not economists, we did a little trial and error with the economical values. All economical variables are updated every three months. The connections we found are listed in the tables below.

Whether a person has a job depends partly on the global employment and on the employment variation rate

from	**to**
emp.var.rate	job
nr.employed	job

Whether a person can afford a house depends partly

on the global economy, so all four economic variables. This is also true for whether a person needs a loan, and/or is in default.

from	to
emp.var.rate	housing
nr.employed	housing
cons.conf.idx	housing
cons.price.idx	housing
emp.var.rate	loan
nr.employed	loan
cons.conf.idx	loan
cons.price.idx	loan
emp.var.rate	default
nr.employed	default
cons.conf.idx	default
cons.price.idx	default

The economical variables also influence one another. In short: Employment rate influences the consumer indexes. This is because the consumer economics do better when the job market performs better.

from	to
emp.var.rate	nr.employed
emp.var.rate	cons.price.idx
emp.var.rate	cons.conf.idx
nr.employed	cons.price.idx
nr.employed	cons.conf.idx
cons.price.idx	cons.conf.idx

Euribor3m is the 3 month average interest rate of a group of European banks. Banks also change their interest based on the current economical climate. Because of this, we created dependencies from euribor3m to all other economical variables. During testing we found that a lot of combinations between euribor3m values and other economical variables didn't exist in our data. This, and the fact that euribor3m only has two bins, made the `localTests` function return an error when such a combination occurred as a condition. Because the influence of the euribor3m attribute is captured strongly by the other economical variables, being dependent on it, we decided to delete the euribor3m variable.

We found high RMSEA values between marital and economical variables, conditioned on other economical variables. We think this is an illogical dependency, likely to be an artefact in our dataset because of some unobserved variable such as for instance company strategy, and decided not to change our DAG to address this.

Revised DAG
Our revised and final DAG is shown in section 3.3. A bigger version of this DAG is shown in the appendix (Appendix B Figure 11).

Figure 8: Final DAG after model testing

Final testing results
The final testing values can be found in Appendix E. Because we focused conditional independencies with an RMSEA value greater than 0.05, many independencies have disappeared. Mainly, economic variables have been connected with many other variables. We believe this was the right thing to do because the overall state of the economy is a good predictor of an individual's economical status and consequently, their willingness to take out loans.

Still, some independencies with an RMSEA value ≥ 0.5 remain. This is by choice. The RMSEA values that stand out the most are associated with the *marital* attribute in combination with economic variables. We are convinced that these are coincidental, an artifact of our binning, or an artifact of change in the marketing strategy of the bank, and that no new edges or latent variables should be introduced to explain them.

Conclusion
We are happy with our final DAG, not just because it deals with most of the high RMSEA values that we found, but also because we feel that the edges drawn can be explained with reason.

Finally, we intended to analyse our latent variable *wealth* using the `lavaan` package, but since it expects different input than `dagitty`, we did not have time to do so.

4 Structure learning

In this section we automatically construct a Bayesian network structure for our dataset using a structure learning algorithm. The PC algorithm

(named after its authors, Peter and Clark) is the state-of-the-art constraint based method for causal discovery. It is based on the Inductive Causation (IC) algorithm by Pearl and Verma, 1991. It uses many conditional independence tests, similar to the tests we performed in the previous section.

4.1 Software

The `bnlearn` package implements various structure learning algorithms. Their implementation of PC is called `pc.stable`, which takes the data, and many parameters, of which at least three must be considered before learning network structure, which are the following.

variable	type
`test`	the conditional independence test to use
`alpha`	target error rate
`max.sx`	maximum size of conditioning sets used in independence tests

4.2 Early observations

In the previous section we binned the data to nominal values to use the χ^2 test for conditional independencies. We set our RMSEA value threshold to 0.05 for these test, and a maximum size of the conditioning sets was not set. We aim to learn a network structure from the data that we can compare to the model that we created ourselves, so our first structure learning approach was to use `pc.stable` with `test = 'x2', alpha = 0.05, max.sx = NULL`. This gave us an extremely sparse graph with many variables not having an edge at all. Changing the test to other tests that are suitable for nominal values, as well as changing the other two parameters had little to no effect on the outcome. We assume the main reason for this to be that many combinations of variables have too few samples, which causes the conditional independence tests for nominal variables to be unreliable or even undefined, especially when the size of the conditioning set increases. This leads us to believe that we cannot work with nominal values if we wish to learn a network structure that can be compared with the network that we created in the previous section.

4.3 Approach

We aim to use the correlation test for independence, because it is considered to be more robust, and mainly, it does not suffer from the problem that we encountered. This test, however, needs continuous variables. To convert our data to continuous, we must again look into the data, as well as use our domain knowledge to convert all variables to `numeric` type in `R`.

Pre-processing
We start again from the unbinned data, with 41188 data points. The same variables that were dropped in the the previous section are dropped from the dataset again. That means that the variables *contact, month, day_of_week, duration, compaign, pdays, previous, poutcome* and *euribor3m* are out.

We will now discuss how the remaining variables were transformed to be continuous.

Age This variable was already an integer, so we simply cast it to numeric.

Job This variable has 12 levels: *admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown.* The levels were grouped based on salary, then mapped to a numeric value. The dataset is from Portugal, but data from us-news.com was used to estimate income, because it is publicly available, and gives a rough estimate. We came up with the following groups along with their values:

1. *unknown, unemployed, retired, student*
2. *housemaid, services, blue-collar*
3. *admin., technician*
4. *self-employed, entrepreneur*
5. *management*

We could not find data on *self-employed* and *entrepreneur*. Presumably becaues these incomes vary a lot. We chose to group these together, with an arbitrary position in the order.

Marital This variable has 4 levels: *divorced, married, single, unknown.* There were only 80 *unknown* entries, so these were added to *single.* Then, from this variable, the three binary variables *divorced, married, single* were created. This is known as one-of-K encoding.

Education This variable has 8 levels, which were simply ordered 1 through 9 from low to high, before being cast to numeric. The order chosen is: *unknown, illiterate, basic.4y, basic.6y, basic.9y, high.school, professional.course, university.degree.* This means that we assign a gap of 1 between each education level, which might not be realistic, but

the data does not give us enough information to do it otherwise.

Default This variable has 3 levels: *no, unknown, yes*. There were only 3 *yes* entries, so these were added to *unknown*. The variable was then made binary with 0 representing *no*, and 1 representing *unknown*.

Housing & Loan These variables both have 3 levels: *no, unknown, yes*. There were exactly 990 *unknown* entries in both, these were added to *no*. The variable was then made binary with 0 representing *no*, and 1 representing *yes*.

Economic variables The four economic variables *emp.var.rate, cons.price.idx, cons.conf.idx, nr.employed* are already numeric, and were not altered.

Outcome Outcome variable y has 2 levels: *no, yes*. These were simply mapped to 0 and 1, respectively.

Automatic network construction
Now that the variables are converted to `numeric`, `pc.stable` with correlation as independence test can be applied to the data, and the network structure can be automatically learned. Without tweaking the parameters, e.g. `test = 'cor', alpha = 0.05, max.sx = NULL` we get a DAG that has many edges, and resembles the final DAG from testing much more already.

However, there are some problems. Since the network is ignorant of the order in which variables happen, it learns that the outcome variable, y, is the predictor of some other variables. We do not want this to happen, therefore we add some edges to the `blacklist` parameter of `pc.stable`. Namely all outgoing edges from y, as well as all incoming edges to `age`, since age should not be influenced by other variables.

The learned network structure was inspected for many values of `alpha, max.sx`. We found that resulting DAGs with the maximum amount of conditioning variables set to 3 or greater did not differ. When it was set to lower than 3, many edges were introduced, resulting in an almost fully connected graph. We consider 3 conditioning variables to still be on the low side. Therefore we chose to set it to `NULL`, an unbounded amount of conditioning variables, because it did not change when greater than 3.

Tweaking the `alpha` parameter between 0.05 and 0.3 did not change the learned network. Setting it outside of this range either resulted in a network with too many edges, or with too few edges. We settled on `alpha = 0.1`.

4.4 Learned DAG

The DAG that was learned with our final parameter configuration is shown in Figure 9. A bigger version of this DAG is shown in the appendix (Appendix C Figure 12).

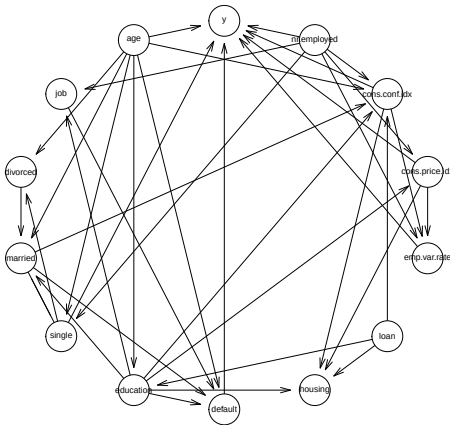


Figure 9: Learned DAG with PC algorithm.

The learned DAG is very similar to the DAG made by testing, save for some edges and edge directions. We are happy to see an edge from *education* to *job*, because it makes a lot of sense. However, we are surprised that neither of these two variables influence the outcome variable directly. In our final DAG these went into a latent variable *wealth*, which directly influence the outcome.

We also expected that *housing* and *loan* would have an edge to y, but the effect happens to be too small for the PC algorithm with our parameters. What is very interesting is that all four of the national economic indicator variables are connected to y. In our final DAG only two were connected to the outcome directly.

Finally, the algorithm finds a dependency between marital status and an economic variable, which we also found during model testing. Intuitively there should not be a dependency there. One of the reasons for this effect is that there could be a confounding variable.

5 Application

We apply a model to the data to find an answer our research questions.

5.1 Structural Equation Models

We apply our bayesian network by fitting it as a structural equation model (SEM) on the data. SEMs require continuous data, therefore we need to use a model that accepts it. Our tested model does not accept this, so instead we used the model that was learned with the PC algorithm.

Markov Blanket SEM
To answer our research questions, only variables that predict outcome variables y are important. Therefore we extract the Markov blanket of y from the network using the `bnlearn` package. The only includes the variables `age`, `default`, `single`, `nr.employed`, `cons.conf.idx`, `cons.price.idx`, and `emp.var.rate`.

The Markov blanket DAG is transformed into a format that the `lavaan` library accepts. Then, the function `fit()` is called with the DAG and the data as parameters to fit the SEM. The coefficients are then printed by calling the `summary()` function. We encountered some confusing values with most coefficients being extremely low, which caused us to wonder why PC had learned these edges. We realized that this could be solved by standardizing the values in the data. After standardization the coefficients were on the same scale, and therefore more interpretable.

The coefficients of the variables in the Markov blanket are shown in Appendix F, in the *Estimate* column. These are not all the coefficients in the SEM, just those that influence y. We can see that the variables that are related to the person (`age`, `default`, `single`) have a positive or negative coefficient of less than 0.05, while the variables that have to do with the national state of the economy have higher coefficients, ranging from 0.1 to 0.3.

Variables not in Markov Blanket
In the model that we made by testing we felt that the variables `education` and `job` would have a big influence on the outcome y. They were connected to each other, and to the outcome variable through the latent variable `wealth`. However, they did not appear in the Markov blanket of the learned structure.

We are still interested in their influence on the outcome, therefore we manually change the Markov blanket to include them. Both are connected directly to y, and `education` is connected to `job`. The SEM is fit again, and a summary is displayed. The coefficients are shown in Appendix G.

These variables turn out to be weak predictors of the outcome, with lower coefficients than the personal variables in the Markov blanket. As a bonus the coefficient between `education` and `job` was added, to show that the former predicts the latter rather well.

5.2 Conclusions about the Data

The SEM coefficients carry information about the relation between variables, that can help us answer our research questions. Our first research question is about which people the bank should target in their marketing campaign. The coefficients tell us that people are more likely to say yes when they get older, when they are single, and when they have a higher education. They are less inclined to say yes when they have credit in default, and when their job pays more. However, these effects are very small.

The second research question is about when to call people. During model testing we found that the moment of calling is one of the weakest predictors in the data. Therefore we chose to get rid of the time and date variables, as such, we do not think the bank should focus on on this.

The third and final question is about the state of the economy. The coefficients of the SEM tell us that these variables are the best predictors of the outcome. This means that the bank can gain much more by marketing at the right time, rather than targeting the right people. When employment rate goes up, outcomes are less positive. When the price and confidence indices go up, more positive outcomes occur. A sub-question to this one is whether the bank should target different people in different economic states. This is something we are not able to answer this question with the coefficients from the SEM. On top of that, we found biases in the data that relate to changes in the marketing campaign of the bank, that complicate answering this question. For instance, the contact medium changed from landline to strictly cellular at some point, and dependencies between a person's marital status and the national economic

state were found, which we assume is due to the bank's marketing strategy.

6 Discussion

We enjoyed working on this project a lot and feel like we learned a lot. Looking back, we are quite satisfied how it turned out. During the course of this project, we learned many things about working with Bayesian networks, from pre-processing to applying the model. Summing up, we learned how to:

- construct a Bayesian network based on domain knowledge, how to inspect and visualise data for bias and incorrect or missing data;
- bin and order bins, based on data distribution and domain knowledge;
- apply inference algorithms to reason probabilisticly in Bayesian networks;
- statiscally evaluate the fit of a Bayesian network model to a dataset using several independence tests;
- generate structural equation models using structure learning algorithms;
- apply a Bayesian network in order to answer causal questions about the original problem domain, based on data in this domain.

We feel that a strong point of our project is that the model made by testing has many similarities with the model that was automatically learned from the data, even though these were made with data that was pre-processed differently (nominal binning vs. ordered continuous), and had different independence tests. This indicates that our pre-processing choices were valid, and contributed to the validity of the model.

Areas where we feel our current work could be improved are as follows:

- We did not bin certain numerical attributes in our data before starting the model testing, only during. Over the course of the course we learned of the benefit of binning our data with respect to the tests, but we did not start over with the model testing on this binned version of the dataset. This likely does not have had a substantial impact on the end result, because we also evaluated the value of adding edges to the network, as opposed to only removing them.
- The ordering we decided on for the attribute "job" should be scrutinized. It is currently based on our intuition with respect to the problem domain combined with income data for these jobs we retrieved online, but it is still not a given whether to group "entrepeneur" and "self-employed" together and where to place this group in relation to other groups. Ideally this grouping should be verified by checking the coëfficient between "education" and "job".

If we were to redo this project, we would bin the data before model testing, instead of doing this during this process when we found out that it was important. We would also use a correlation test instead of the chi-square test during independence testing in order to mitigate the problems caused by a lack of data points for certain combinations of variable instantiations. Lastly, we would have liked to introduce an ordering between bins before model testing, instead of before learning the SEM, in order to increase the amount of information captured by the data, to increase the significance of the independence tests, and to make applying the tested model easier.

7 Conclusion

Looking back at our project it is a surprising / serendipitous result that the state of the economy, as opposed to who someone is, is most important in the result of whether or not someone will take out a loan at the bank. This is not to say that personal variables do not matter, but the extent to which they do is considerably smaller.

Variables such as time and day were found to be bad predictors. This was already found before applying the model using a SEM.

We are pleased to find that our conclusions conform to those of Moro et al. [2], from whence our data set originates: They also obtain the result that economic factors are the most relevant attributes. Moro et al. find that personal variables are also good predictors, more so than we found, but their dataset contains different, and more personal variables than the dataset that is publicly available, and that we used.

We have found that, independent of economic variables, the bank should target older single people that do not have a loan in default, with marginal preference for people with a higher education. Ad-

ditionaly, we've found no preference for a moment to call someone, independent of who this person is. We've found that the current state of the economy is very important with respect to the likelihood that someone takes out a loan. Lastly we've found that we cannot conclude whether to target different people when the economy changes, as a result of bias present in the data.

A Initial DAG

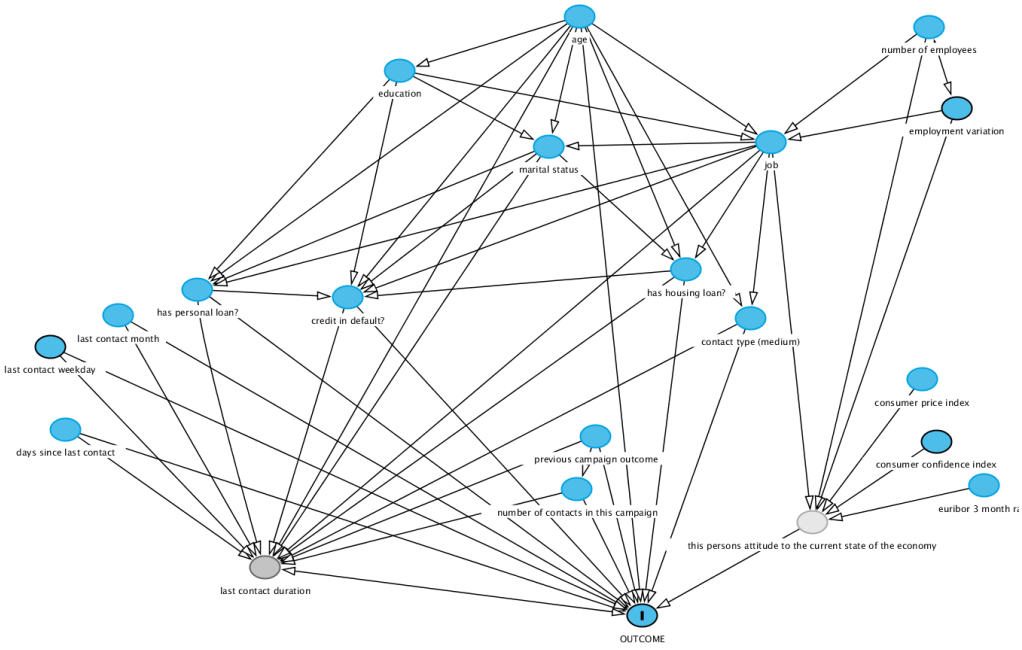


Figure 10: Initial DAG as proposed in exposee.

B Final DAG

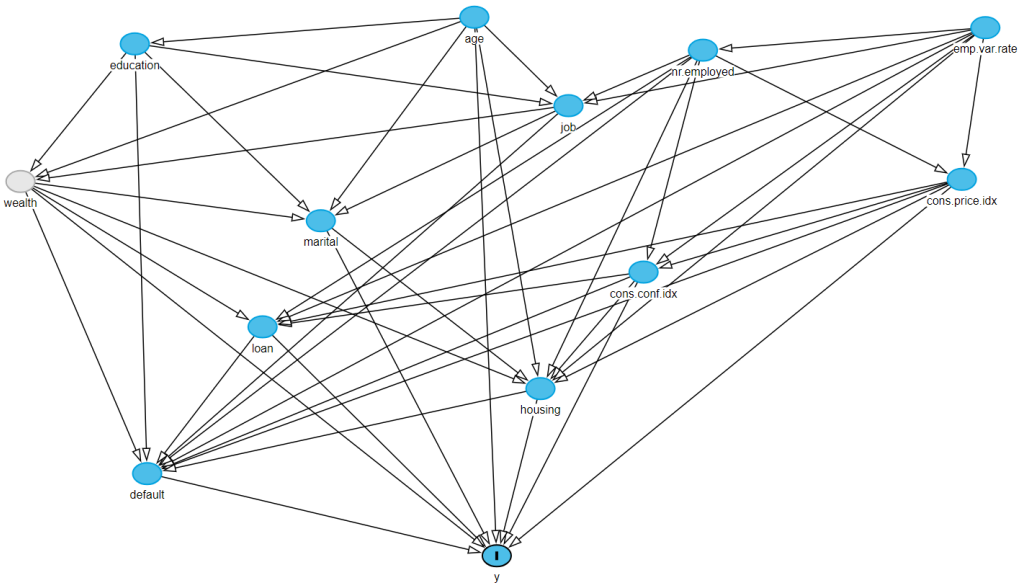


Figure 11: Final DAG after model testing.

C Learned DAG

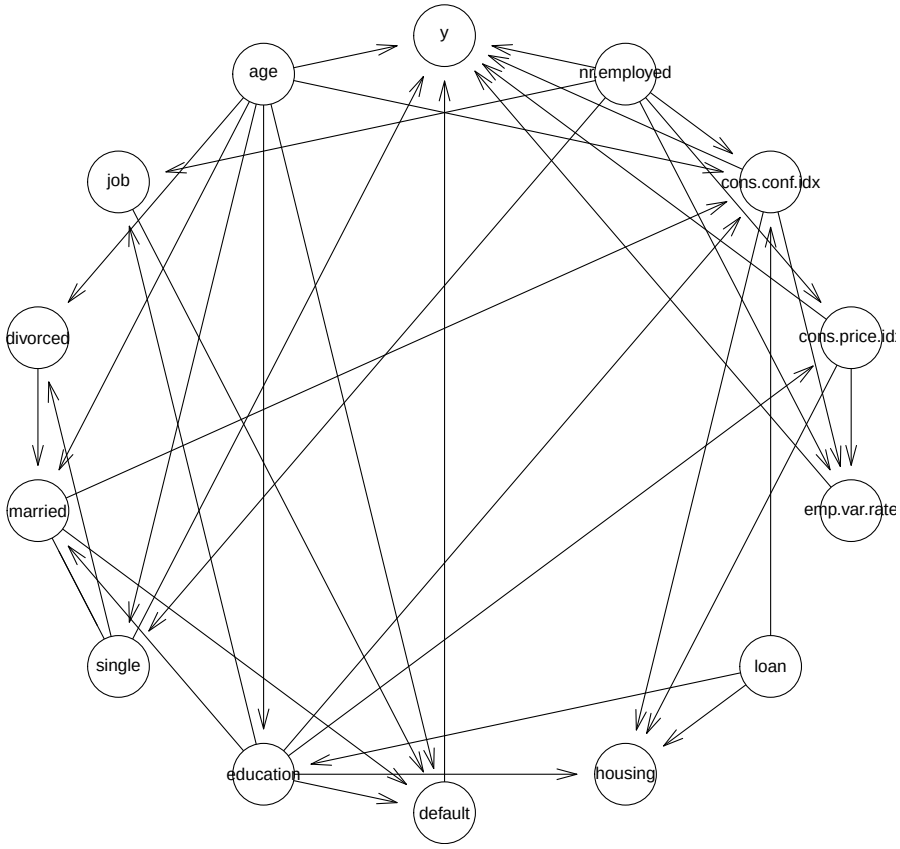


Figure 12: Learned DAG with PC algorithm.

D Initial conditional dependencies

```
consumer confidence index _||_ consumer price index
consumer confidence index _||_ contact type (medium)
consumer confidence index _||_ credit in default?
consumer confidence index _||_ days since last contact
consumer confidence index _||_ employment variation
consumer confidence index _||_ euribor 3 month rate
consumer confidence index _||_ has housing loan?
consumer confidence index _||_ has personal loan?
consumer confidence index _||_ last contact duration
consumer confidence index _||_ last contact month
consumer confidence index _||_ last contact weekday
consumer confidence index _||_ marital status
consumer confidence index _||_ number of contacts in this campaign
consumer confidence index _||_ number of employees
consumer confidence index _||_ previous campaign outcome
consumer confidence index _||_ age
consumer confidence index _||_ education
consumer confidence index _||_ job
consumer price index _||_ contact type (medium)
consumer price index _||_ credit in default?
consumer price index _||_ days since last contact
consumer price index _||_ employment variation
consumer price index _||_ euribor 3 month rate
consumer price index _||_ has housing loan?
consumer price index _||_ has personal loan?
consumer price index _||_ last contact duration
consumer price index _||_ last contact month
consumer price index _||_ last contact weekday
consumer price index _||_ marital status
consumer price index _||_ number of contacts in this campaign
consumer price index _||_ number of employees
consumer price index _||_ previous campaign outcome
consumer price index _||_ age
consumer price index _||_ education
consumer price index _||_ job
contact type (medium) _||_ credit in default? | age, job
contact type (medium) _||_ days since last contact
contact type (medium) _||_ employment variation | age, job
contact type (medium) _||_ euribor 3 month rate
```

contact type (medium) _||_ has housing loan? | age, job
contact type (medium) _||_ has personal loan? | age, job
contact type (medium) _||_ last contact month
contact type (medium) _||_ last contact weekday
contact type (medium) _||_ marital status | age, job
contact type (medium) _||_ number of contacts in this campaign
contact type (medium) _||_ number of employees | age, job
contact type (medium) _||_ previous campaign outcome
contact type (medium) _||_ education | age, job
credit in default? _||_ days since last contact
credit in default? _||_ employment variation | age, education, job
credit in default? _||_ euribor 3 month rate
credit in default? _||_ last contact month
credit in default? _||_ last contact weekday
credit in default? _||_ number of contacts in this campaign
credit in default? _||_ number of employees | age, education, job
credit in default? _||_ previous campaign outcome
days since last contact _||_ employment variation
days since last contact _||_ euribor 3 month rate
days since last contact _||_ has housing loan?
days since last contact _||_ has personal loan?
days since last contact _||_ last contact month
days since last contact _||_ last contact weekday
days since last contact _||_ marital status
days since last contact _||_ number of contacts in this campaign
days since last contact _||_ number of employees
days since last contact _||_ previous campaign outcome
days since last contact _||_ age
days since last contact _||_ education
days since last contact _||_ job
employment variation _||_ euribor 3 month rate
employment variation _||_ has housing loan? | age, job, marital status
employment variation _||_ has housing loan? | age, education, job
employment variation _||_ has personal loan? | age, education, job
employment variation _||_ last contact duration | age, credit in default?, has housing loan?, has personal loan?, job, marital status
employment variation _||_ last contact duration | age, education, job
employment variation _||_ last contact month
employment variation _||_ last contact weekday
employment variation _||_ marital status | age, education, job
employment variation _||_ number of contacts in this campaign
employment variation _||_ previous campaign outcome
employment variation _||_ age
employment variation _||_ education
euribor 3 month rate _||_ has housing loan?
euribor 3 month rate _||_ has personal loan?
euribor 3 month rate _||_ last contact duration
euribor 3 month rate _||_ last contact month
euribor 3 month rate _||_ last contact weekday
euribor 3 month rate _||_ marital status
euribor 3 month rate _||_ number of contacts in this campaign
euribor 3 month rate _||_ number of employees
euribor 3 month rate _||_ previous campaign outcome
euribor 3 month rate _||_ age
euribor 3 month rate _||_ education
euribor 3 month rate _||_ job
has housing loan? _||_ has personal loan? | age, job, marital status
has housing loan? _||_ last contact month
has housing loan? _||_ last contact weekday
has housing loan? _||_ number of contacts in this campaign
has housing loan? _||_ number of employees | age, education, job
has housing loan? _||_ number of employees | age, job, marital status
has housing loan? _||_ previous campaign outcome
has housing loan? _||_ education | age, job, marital status
has personal loan? _||_ last contact month
has personal loan? _||_ last contact weekday
has personal loan? _||_ number of contacts in this campaign
has personal loan? _||_ number of employees | age, education, job
has personal loan? _||_ previous campaign outcome
last contact duration _||_ number of employees | age, education, job
last contact duration _||_ number of employees | age, credit in default?, has housing loan?, has personal loan?, job, marital status
last contact duration _||_ education | age, credit in default?, has housing loan?, has personal loan?, job, marital status
last contact month _||_ last contact weekday
last contact month _||_ marital status
last contact month _||_ number of contacts in this campaign
last contact month _||_ number of employees
last contact month _||_ previous campaign outcome
last contact month _||_ age
last contact month _||_ education
last contact month _||_ job
last contact weekday _||_ marital status
last contact weekday _||_ number of contacts in this campaign
last contact weekday _||_ number of employees
last contact weekday _||_ previous campaign outcome
last contact weekday _||_ age
last contact weekday _||_ education
last contact weekday _||_ job
marital status _||_ number of contacts in this campaign
marital status _||_ number of employees | age, education, job
marital status _||_ previous campaign outcome
marital status _||_ OUTCOME | age, credit in default?, employment variation, has housing loan?, has personal loan?, job, number of employees
marital status _||_ OUTCOME | age, credit in default?, education, has housing loan?, has personal loan?, job
number of contacts in this campaign _||_ number of employees
number of contacts in this campaign _||_ age
number of contacts in this campaign _||_ education
number of contacts in this campaign _||_ job
number of employees _||_ previous campaign outcome
number of employees _||_ age
number of employees _||_ education
previous campaign outcome _||_ age
previous campaign outcome _||_ education
previous campaign outcome _||_ job
OUTCOME _||_ education | age, credit in default?, employment variation, has housing loan?, has personal loan?, job, number of employees

E Final test results

	rmsea	x2	df	p.value
age _||_ cons.conf.idx	0.08114638	4083.0823	15	0.000000e+00
age _||_ cons.price.idx	0.05653150	2652.5172	20	0.000000e+00
age _||_ emp.var.rate	0.07828864	3801.5956	15	0.000000e+00
age _||_ nr.employed	0.10749941	4769.6201	10	0.000000e+00
cons.conf.idx _||_ education	0.02926247	761.6302	21	1.223205e-147
cons.conf.idx _||_ job | emp.var.rate, nr.employed	0.02830802	490.7039	66	3.618113e-66
cons.conf.idx _||_ marital | age, education, job	0.12343419	2427.4274	1272	6.928513e-75
cons.conf.idx _||_ marital | emp.var.rate, nr.employed	0.02646795	159.7641	18	9.266606e-25
cons.price.idx _||_ education	0.04324455	2184.6563	28	0.000000e+00
cons.price.idx _||_ job | emp.var.rate, nr.employed	0.03749213	2051.0141	77	0.000000e+00
cons.price.idx _||_ marital | age, education, job	0.12079377	2896.2705	1557	1.759695e-83
cons.price.idx _||_ marital | emp.var.rate, nr.employed	0.02793123	122.8179	21	2.166818e-16
education _||_ emp.var.rate	0.02872860	734.8521	21	5.688697e-142
education _||_ nr.employed	0.03035075	545.1620	14	2.425708e-107
emp.var.rate _||_ marital | age, education, job	0.12642528	2417.9671	1208	3.182673e-83
marital _||_ nr.employed | age, education, job	0.12932179	1626.0104	831	6.335085e-54

F SEM: Markov blanket - Learned DAG

Regressions of y with:

Column	Estimate	Std.Err	z-value	P(>|z|)
age	0.032	0.005	6.197	0.000
cons.conf.idx	0.104	0.005	20.093	0.000
cons.price.idx	0.132	0.011	12.328	0.000
default	-0.041	0.005	-8.788	0.000
emp.var.rate	-0.149	0.022	-6.647	0.000
nr.employed	-0.287	0.016	-17.778	0.000
single	0.032	0.005	6.407	0.000

G SEM: Markov blanket - Learned DAG incl. Education, Job

Regressions of y with:

Column	Estimate	Std.Err	z-value	P(>|z|)
age	0.034	0.005	6.687	0.000
cons.conf.idx	0.103	0.005	19.889	0.000
cons.price.idx	0.137	0.011	12.759	0.000
default	-0.037	0.005	-7.832	0.000
education	0.028	0.005	5.478	0.000
emp.var.rate	-0.155	0.022	-6.946	0.000
job	-0.013	0.005	-2.615	0.009
nr.employed	-0.283	0.016	-17.516	0.000
single	0.030	0.005	5.849	0.000

job ~				
education	0.399	0.005	88.264	0.000

References

[1] "Bank marketing data set," 2012. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[2] P. R. S. Moro P. Cortez, "A data-driven approach to predict the success of bank telemarketing," *Elsevier*, vol. 62, pp. 22–31, Jun. 2014.