Employee attrition and its main causes

Case study on the most probable reasons of the loss of workforce

Salvatore Guarrera, Riccardo Monaco
Master's Degree in Artificial Intelligence, University of Bologna
{salvatore.guarrera, riccardo.monaco2}@studio.unibo.it

January 20, 2023

Abstract

Employee attrition is defined as the departure of employees from the organization where they work, for any reason. The main causes of this loss of workforce are here considered and deeply analyzed, making use of the studies done about probability and uncertainty. It has been considered a dataset, pre-processed in order to be analyzed in the best possible way, created a Bayesian Network and made exact and approximate inferences, exploiting, respectively, *variable elimination* and *likelihood weighted sampling*. It has been possible to complete this project thanks to the usage and support of means such as the *pgmpy* library and its methods.

Introduction

Domain

The dataset considered in this project has been selected from Kaggle. It has been used in order to study employee attrition, which is the diminishing workforce of the employees of a company. The dataset takes into account 14710 observations and 13 variables, which translate into 14710 rows and 13 columns. The 13 variables are the following:

- Age of employees
- Attrition
- Department of work
- Distance from home
- Education
- Education Field
- Environment Satisfaction
- Job Satisfaction
- Marital Status
- Monthly Income
- Number of Companies Worked
- Work Life Balance
- Years At Company

In this project, causes of employee attrition have been examined, also considering what can be found in Singh, K., Singh, R. (2019). A Study on Employee Attrition : Effects and Causes (we'll simply refer to it as *paper* from now on). The 13 variables present in the considered dataset find a match in the paper, thanks to which it has been possible to confront the results of our analysis with a real-world case.

Aim

The main purpose of this project is to find out which are the variables which influence the most the attrition of the employee and, on the other hand, which are those that make the employee want to keep his/her job. This can be done through one of the three *reasoning patterns* that have been taught during the course, i.e. the Causal Reasoning. Moreover, Evidential Reasoning and Intercausal Reasoning are also taken into consideration, in order to further examine the matter.

Method

Once the dataset has been selected, the variables have been compared to attrition causes present in the scientific paper and a first sketch of what the Bayesian Network was going to look like has been made. The causal connections made follow both the ideas of the paper and our intuition (Figure 1).

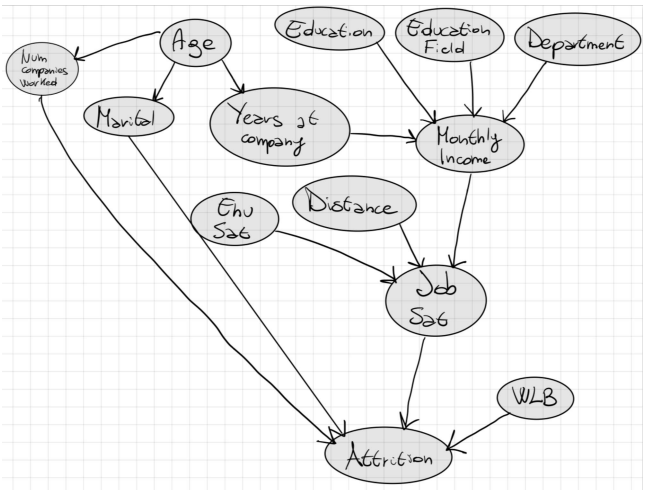

Figure 1: Handwritten Bayesian network

Then, thanks to the help of python library *pgmpy*, the Bayesian Network could be modeled. Through pgmpy, CPDs (Conditional Probability Distributions) of the variables have been fitted to the dataset, and CPTs have been constructed. The next step was inference, exact and approximate. For the first one, it has been taken into consideration Variable Elimination. For the latter, Likelihood Weighted Sampling has been used.

Results

The realization of this project taught us how Bayesian Networks can be related to real-world situations and can be useful to gather information about a certain topic. The usage of pgmpy really helped us put into practice what learned during the course, especially how can change the influence that some variables have on other ones, depending on whether exact or approximate inference is used.

Model

The dataset considered for this project was fundamental for constructing the Bayesian Network, since it gave us all the variables, which later on became the nodes of the network. In order to make it more readable, abbreviations were used (Figure 2).

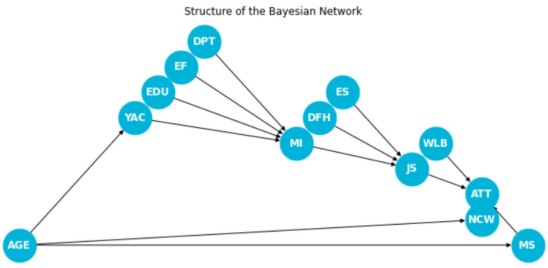

Figure 2: Structure of the Bayesian Network

The causal links between each node were instantiated, as said before, looking at the paper taken into account and considering our own ideas of the relations between the variables, binding causes to parents and effects to children nodes. Getting to the bottom of the network, the main variable, *Attrition*, can be found. Pre-processing made it possible to work with the data provided, transforming string values into numerical values and defining the ranges in which they occur. In order to show how conditional probabilities are related, Conditional Probability Tables (CPTs) are shown in the notebook, obtained exploiting the relative pgmpy method.

Analysis

Experimental setup

Exact Inference As said, the main purpose of this project is understanding what are the most relevant causes that determine *employee attrition*. In order to do so, the *Attrition* variable is taken as query variable considering the *Yes* value, and all the variables except *Attrition* are taken individually as evidence, each of which with all the values that it can take, making use, firstly, of Exact Inference (exploiting *Variable Elimination*). The highest probability value for each single variable value is saved and, then, sorted from highest to lowest. The same thing is done using *Attrition* variable with the *No* value, in order to determine what are the main causes that lead an employee to keep his/her job. Evidential Reasoning is exploited in order to ask the network, given *Attrition* as evidential variable with the *Yes* value, in which age range is an employee most likely to be and what level of education he/she is most likely to have. Eventually, Intercausal Reasoning is used to determine in which monthly income range the employee is most likely to be, given the age range.

Approximate Inference Following the same process used for Exact Inference, Approximate Inference can be done, this time, though, exploiting *likelihood weighted sampling* using *Attrition* as query variable, and each variable except *Attrition* as evidence, each of which taken individually.

Results

According to Exact Inference, a poor work life balance, a low job satisfaction and a single marital status are the main causes that lead to employee attrition. The same causes are also the most relevant in making the employee to keep the job, obviously considering other values. Thus, a good work life balance, a high job satisfaction and a divorced marital status are the main causes that lead to an employee to keep his/her job, in line with what written in the paper. If *Attrition* is taken as evidence, it is more likely that an employee is between 33 and 40 years old and has a *bachelor* as education level. It was also observed that as the age of an employee increases, also the monthly income is more likely to be high.

Conclusion

In conclusion, Bayesian Networks turned out to be a good tool to study and analyze causal relations, given a set of data, and a great mean to, then, exploit both exact and approximate inference for seeking the probability of an event to occur or not. Following the results, a company can aim to improve work life balance giving employees the opportunity to have more space to personal life, or to improve job satisfaction through an increase of the monthly income and the enhancement of transportation that takes the employee from his/her home to the workplace, in order to reduce employee attrition and retain them. The Bayesian Network can be further improved in the future, with a better understanding of the knowledge domain, for example studying other scientific papers so that causal relations can be improved.

References and Links to external resources

- Singh, K., Singh, R. (2019). A Study on Employee Attrition : Effects and Causes
- https://www.kaggle.com/datasets/prachi13/employeeattritionrate?resource=download