FAIKR Module 3 Project

Nanni Gabriele
Master's Degree in Artificial Intelligence, University of Bologna
gabriele.nanni4@studio.unibo.it

March 25, 2024

Abstract

This project aims to define (in an approximated way) a Bayesian Network capable of estimating the Credit Score of individuals and offering a default risk evaluation. For this purpose, I've used a Kaggle dataset considering 1000 customers, their earnings, debts, expenses and some other factors like having dependants or gambling. The results of the project are sometimes counter-intuitive but they overall align with common belief. It is obvious that a better knowledge of the matter, a more complex network and more data can offer a more solid result.

Introduction

Domain

To acknowledge the risk associated with a possible loan the banks utilize an algorithm to determine the capability of someone to repay the debt, outputting a numerical value called Credit Score. The Credit Score has no objective ranges, but it will be considered an approximate splitting to discretize the elements.
In this project, we try to build a network that can predict the Credit Score of a customer. The model is drawn starting from a Kaggle dataset and some basic knowledge about financial management. The elements we are going to consider are the amount of money the person is entitled to spend, the debt they have, some notable expense categories, assets and credit sources.
It is important to stress that, starting from an already structured dataset, the network has been based on the element offered by the starting data. Credit Score evaluation is a much more complex process and a lot of other factors are considered, for example, the payment history is an important element in determining if the customer is a trustworthy person, that pays on time, or if they are not. Credit history and other factors are kept in consideration when evaluating the Credit Score, but in this experiment the network considered is a simplified and small prototype that has no purpose of being precise in a real Credit Score prediction.

Aim

The aim of the project is to define and implement a Bayesian Network that could describe the domain of interest, starting from the elements present in the Kaggle dataset found. After defining the model and fitting it to the dataset the objective is to query the system to observe if the results can be considered satisfactory and what are the most important factors in the evaluation of the default risk and credit score.

Method

For the preprocessing phase (deleting non-relevant columns, renaming, etc.) it has been chosen to use the pandas library. To implement the network and to run queries it has been used the pgmpy library instead.

Results

The network implemented in this project seems to reflect the expectations about the topic considered, although some results are not aligned with the starting hypothesis. Some of the outlying results could be caused by the lack of representative data, but the network could be improved analysing the structure with an expert.

Model

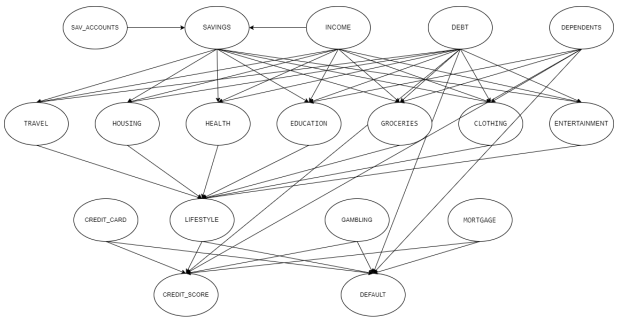

Figure 1: Structure of the Bayesian Network

The network is structured as follows:

- INCOME and SAV ACCOUNT will determine the range in which SAVINGS is.
- INCOME, DEBT and SAVINGS will determine the range of all the expenses considered in the network (in some cases DEPENDENTS will influence, for example, considering the groceries, having someone else to feed will influence the expenses).

- All the expenses will determine the LIFESTYLE (the actual value of lifestyle was determined considering the expenses in relation to the disposable money of an individual, but it was just a parameter to define a more abstract concept).
- Then the LIFESTYLE, the attitude towards GAMBLING of the customer, having a MORTGAGE, possessing a CREDIT CARD, the DEBT and having DEPENDANTS will influence the CREDIT SCORE and the DEFAULT.

This structure was constructed analysing the domain without a really deep knowledge of the financial process to determine the Credit Score, but thanks to information already known or easily retrievable for anyone. The probability distributions were obtained by fitting the model on the dataset already mentioned.

Analysis

Experimental setup

The queries executed on the system have the purpose to test the network to understand if the modelling was correct or if there are other factors or other links to be considered in the building of the network.
Specifically:

- The first query wants to observe if the system can successfully observe that (in an indirect way) the disposable money a person have influence the CREDIT SCORE and DEFAULT values. The expected result is that the influence must be noticeable because a bigger money flow, even if not directly influent for the CREDIT SCORE, offers more possibilities to develop credit sources.
- The second query wants instead to determine the best set of "input variables" (even if SAVINGS is not properly an input variable) INCOME, DEBT and SAVINGS, to achieve the best possible result ('Very High' Credit Score and no Default).
- The aim of the third query is to analyse the influence of gambling on an already risky lifestyle.
- The fourth query wants to observe the influence of the variables MORTGAGE and DEPENDENTS on the Credit Score, to observe if the network considers positively other assets.
- The fifth query, with a more sociological view, wants to observe considering the average American income, which are the most expensive categories.
- The sixth query wants to determine if owning a Savings Account influences, even in a slight proportion, the capability of avoiding Default.
- The last query wants to observe the behaviour of the CREDIT SCORE variable based on owning a credit card.

All the queries aim to consider quite simple (even if maybe indirect) causal connections between the variables defined.

Results

The results of the queries are quite satisfactory in general. The first query gave the expected results, in fact observing the CPDs generated it is clear to see a higher probability of avoiding Default and a distribution of Credit Score that tends toward higher values considering the 'Very High' values of the evidence.
The second query instead offered some unexpected results: the DEBT value was aligned with expectations (high debt can indicate several investments and credit sources by the customer), however the INCOME and SAVINGS values obtained are surprising. No explanation has been found apart from the skewed data.
The results of the third query were really interesting, showing that, considering a risky lifestyle, gambling has a slightly positive impact on credit score; it is possible that, considering the unlikely scenario where a loan is offered to the customer, the possibility of gaining money from gambling has a stronger impact than losing that money on gambling.
The fourth query, as expected, showed that other credit sources impact positively the probability of obtaining a high credit score.
The fifth query gives an interesting insight into the categories of expenses preferred by the average American customer.
The sixth query shows that, even if the increase is slight, owning a Savings Account (a very generic assumption) influences the default risk.
Lastly, the seventh query shows that, as one could expect, the probability of a 'Fair' Credit Score decreases, that is due to the two possible behaviours adopted with a credit card: using it reasonably and cautiously will increase the Credit Score with time, instead of maxing out the credit limit or having reckless behaviour will penalize the Credit Score significantly.

Conclusion

The idea behind the experiment was to define a Bayesian Network starting from a predefined dataset and to try to use the data to build solid inferences. The results obtained are quite satisfactory even if the dataset did not contain precisely the information used in the Credit Score evaluation. Obviously, the network could be improved in many ways: interacting with an expert on this topic could define more precise relations between the nodes, with more data and a more even distribution it would be possible to fit the model in a more proper way and considering other determining factors in Credit Score evaluation (as the payment history and the credit history) could lead to a more precise and complete network.

Links to external resources

Kaggle Credit Score Dataset: `https://www.kaggle.com/datasets/conorsully1/credit-score/data`

References

[1] *Average American Debt in 2024: Household Debt Statistics*. https://www.businessinsider.com/personal-finance/average-american-debt?r=US&IR=T. Accessed: 2024-03-10.

[2] *INCOME AND WEALTH IN THE UNITED STATES: AN OVERVIEW OF RECENT DATA*. https://www.pgpf.org/blog/2023/11/income-and-wealth-in-the-united-states-an-overview-of-recent-data. Accessed: 2024-03-02.

[3] Abby M Kern. "Credit score analysis". In: (2017).

[4] *Money Spent On Clothes Statistics*. https://gitnux.org/money-spent-on-clothes-statistics/. Accessed: 2024-03-11.

[5] *New to America: What is the Average Monthly Cost of Living in USA?* https://www.upwardli.com/resources/new-to-america-what-is-the-average-monthly-cost-of-living-in-usa. Accessed: 2024-03-10.

[6] *Statistics About The Average Monthly Grocery Bill*. https://gitnux.org/average-monthly-grocery-bill/. Accessed: 2024-03-11.

[7] *What is a Credit Score?* https://www.myfico.com/credit-education/credit-scores. Accessed: 2024-03-10.