AN6807
Data Analytics for Credit and Related Risk

Take-Home Assignment

Name: Shih Hsiao Ju
Matriculation No.: G2400458E

Objectives

Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies using directed acyclic graph (DAG). Each node in the graph represents a variable, and each edge shows a direct influence between variables. The strength of these relationships is captured using conditional probability tables (CPTs). By leveraging Bayesian network, business users will be able to understand what characteristics of loan applicants entail more default risk and therefore better mitigate the credit risk of their organizations.

Dataset

1. Source: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset
2. Description

   This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

3. Instances: 30,000 rows
4. Data dictionary

Variables	**Description**
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit
SEX	Gender (1=male, 2=female)
EDUCATION	Education Level (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-2=no paying balance in the first place, -1=pay duly, 0=pay part of the due balance, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)

BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
default payment next month	Target variable, default payment (1=yes, 0=no)

5. Quick view of the dataset

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0 ...	PAY_6	BILL_AMT1 ...	BILL_AMT6	PAY_AMT1 ...	PAY_AMT6	default payment next month
1	20000	2	2	1	24	2	-2	3913	0	0	0	1
2	120000	2	2	2	26	-1	2	2682	3261	0	2000	1
3	90000	2	2	2	34	0	0	29239	15549	1518	5000	0
4	50000	2	2	1	37	0	0	46990	29547	2000	1000	0
5	50000	1	2	1	57	-1	0	8617	19131	2000	679	0

Approach

1. Tools used
    a. Programming language: Python
    b. File type: Jupyter Notebook
    c. Library: Pgmpy, Pandas, Sci-kit learn
2. Data Preparation
    a. Converted continuous features into discrete bins

    For the variables with greater than 15 distinct values, discretize it into bins and encode it in an order related to its quartile. Other than age, which is binned based on common demographic statistics standard, rest of the continuous variables is separated into equal size of 10 bins, with higher data assigned higher code (e.g., top 10 percentile of data will be assigned code 10).

    b. Quick view of the cleaned data

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0 ...	PAY_6	BILL_AMT1 ...	BILL_AMT6	PAY_AMT1 ...	PAY_AMT6	default payment next month
1	1	2	2	1	1	2	-2	3	1	1	1	1

2	5	2	2	2	1	-1	2	3	4	1	4	1
3	4	2	2	2	1	0	0	6	5	3	6	0
4	2	2	2	1	2	0	0	7	7	4	2	0
5	2	1	2	1	3	-1	0	4	6	4	2	0

c. Train-test split

Split dataset into training and testing sets with the ratio of 70:30, with the aim to test the efficacy of the model prediction.

3. Model Building
   a. Search the optimal graph structure

The number of possible DAGs grows exponentially with the number of variables, and the scoring functions used can get stuck in local optima. Considering the size of the data used, that means it's nearly impossible to exhaustively search for the best structure. And even local optimization methods can fall short of finding the ideal solution. That said, heuristic search strategies like hill climbing search do well in practice and will be adopted in this report.

Hill climb search starts with an initial network structure and evaluates the effectiveness of the model using scoring method like BIC or BDeu. Next, it will start to grow the network. For each move it made, the respective score will be calculated and be compared to the original one. This process will continue until there's no change to made to improve the score.

As for the scoring method, the most common two choices are BIC and BDeu. Each approach has their pros and cons, and the key differences come from their ability to handle different size of data. BIC is more suitable for large datasets, as it can better manage overfitting, which leads to simpler structure and better predictive accuracy. On the other hand, BDeu uses Drichlet distribution to assume prior probabilities, making it suitable with smaller dataset with missing values. It can also help create a more complex structure to provide more insights about the relationships among the variables.

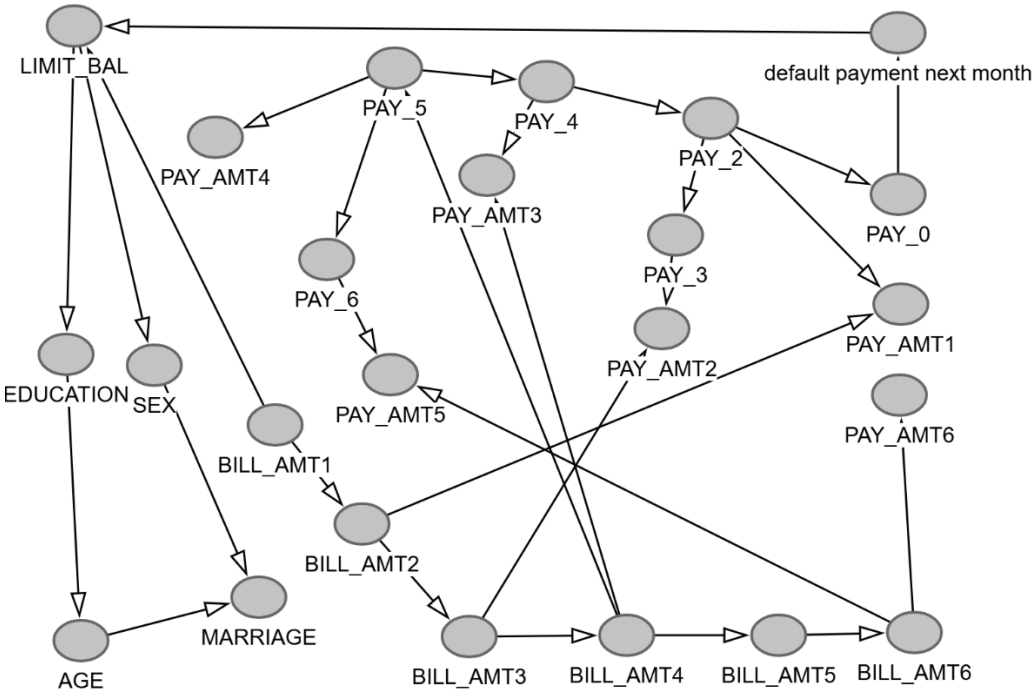
In the end, since the goal of this project is providing business with an effective risk mitigation tool, prediction accuracy will be more important than relationship exploration. As a result, hill climb search and BIC are selected as the tools to search for the optimal structure.

   b. Estimate CPT for each node in the graph

The Bayesian estimator is often considered superior to the maximum likelihood estimator (MLE) in scenarios involving uncertainty or limited data. Unlike MLE, which relies solely on the observed data and may assign zero probability to unobserved outcomes, the Bayesian estimator incorporates prior beliefs through a prior distribution. This allows it to

make more realistic and robust inferences, especially when data is sparse or incomplete. By blending prior knowledge with observed evidence, the Bayesian approach avoids overfitting and provides smoother, more stable probability estimates that better reflect the underlying uncertainty in real-world situations. As such, Bayesian estimator is selected to calculate the CPTs in the graph for further analysis and prediction.

4. Inference Examples
    a. DAG structure



    b. Output CPT examples
        i. Target node (default payment next month)

	default payment next month (Target Node)	
PAY_0 (Parent Node)	0	1
-2	0.87	0.13
-1	0.83	0.17
0	0.87	0.13
1	0.66	0.34
2	0.31	0.69
3	0.24	0.76
4	0.32	0.68
5	0.61	0.39
6	0.40	0.60
7	0.26	0.74
8	0.54	0.46

        ii. Node with multiple parents

BILL_AMT1 (Parent Node)	default payment next month (Parent Node)	LIMIT_BAL (Leaf Node) 1	2	3	...	10
1	0	0.07	0.06	0.02	...	0.08
1	1	0.08	0.05	0.01	...	0.10
2	0	0.08	0.06	0.02	...	0.08
...	...	...	...	...	...	...
10	0	0.00	0.00	0.00	...	0.22
10	1	0.00	0.00	0.00	...	0.16

c. Inference Example

Input		Graph Direction	Output	
EDUCATION	4	... → PAY_2 → PAY_0 → default payment next month	default payment next month	1
PAY_2	6			
BILL_AMT1	9			
PAY_AMT1	1			

In the CPT of P (PAY_0 | PAY_2), P (PAY_0 = 7 | PAY_2 = 6) = 0.77 is the highest conditional probability, so the model infers this applicant with PAY_0 = 7. Next, if we look up the CPT of P (default payment next month | PAY_0), we can discover P (default payment next month | PAY_0 = 7) = 0.74 as the highest conditional probability. As a result, the model predict the applicant is going to default payment next month (=1), even without all the info in the graph.

5. Model Performance:
   a. Evaluation metrics
      i. Confusion matrix

		Predicted	
		0	1
Actual	0	6254	755
	1	1796	195

      ii. Classification report

	Precision	Recall	F1-score	Support
0	0.78	0.89	0.83	7009
1	0.21	0.1	0.13	1991
Accuracy			**0.72**	9000

      iii. ROC-AUC Score: 0.50

Discussion

1. Strengths and Limitations
    a. Strengths
        i. Interpretable

        Unlike common machine learning models such as neural network or tree-based model, which is hard to interpret due to its complex structure. Bayesian Network, on the other hand, is relatively easy to understand thanks to its graphical nature. If curious, user can look into the CPT for each node to understand how does different variables affect the prediction.

        ii. Handles missing data well

        Another great capability of Bayesian Network comes from how it tackles missing data. Leveraging Bayes Theorem and some other probability distributions, it can infer the missing data based on the assumption of prior probability, which expands the choice of dataset. Also, when making predictions, it is also allowed to have only part of the input variables thanks to its inference capability. This feature is especially helpful since in real-life scenario, it's likely that applicants will only provide partial data due to privacy concern or administrative issue. Instead of searching for correct way to infer the missing data, Bayesian network can take care of it based on the theorem behind, making it a safer and more intact approach.

        iii. Good for causal analysis and "what-if" queries

        Graphical models provide a clear relationships among the input variables, making causal analysis more feasible and understandable. It can also serve as a simulation tools if user want to understand how will the prediction outcome change if one/some of the input is modified, and what will be the potential business implications. For example, business will be interested to see what will happen if they increase the credit limit balance of customer, and how will it affect the overall credit risk of the company. By doing so, business can have a better grasp before conducting costly strategy changes.

    b. Limitations
        i. Scales poorly with large continuous datasets

        One of the main criticisms of Bayesian network is its dependency on discrete data. Although discretization can help, transforming continuous data into discrete will still lose possibly important

information, and this risk amplifies as the range of data increases. Also, the techniques to bin the data is also subjective. Separating data into equal size of groups makes certain sense, but without the qualitative consideration, the likelihood of improper grouping is still expectable. For example, binning age into equal batches is not a wise choice since people with different age group will generally act differently due to lots of social factors. Therefore, important information will lost if it is discretized based on the same approach.

ii. Structure learning is computationally expensive

Once the number of variables and their categories increase, the possible combinations of paths and nodes will grow exponentially. Even with the assist of heuristic approach like hill climb search, it will still take large computing resources and huge amount of time to train the model.

iii. Not necessarily the best pure prediction performance

Bayesian Network generative model instead of discriminative model, which is purely designed to make the most accurate prediction. As such, it's performance will generally worse than the other choices like neural network of random forest. This fact can also be verified by the report, where the model's accuracy is only ~72%. On top of that, the options to choose or the parameters to tune are also limited, which indicates its low upside as a prediction tool.

Conclusion

This report has demonstrated the application of Bayesian networks for credit default risk analysis using credit card client data from Taiwan. Despite achieving only moderate predictive accuracy (72%) compared to other machine learning approaches, the Bayesian network model offers significant advantages in interpretability, handling of missing data, and support for causal reasoning. The graphical structure provides business users with transparent insights into how different client characteristics influence default risk, making it valuable for decision-making and risk mitigation strategies. While the model has limitations—including challenges with continuous data discretization, computational complexity in structure learning, and lower predictive performance than discriminative models—its ability to answer "what-if" queries makes it particularly useful for testing potential risk management policies before implementation. For credit risk analysis applications where interpretability and causal understanding are also considered as essential as predictive power, Bayesian networks represent a valuable analytical approach that balances statistical rigor with practical business utility.