Proposal of B5: Risk Analysis and Fraud Detection

Qiaoge Zhu (qz2383)
Xinyi Zhang (xz2862)
Feb.10th, 2020

Abstract: Risk analysis and fraud detection have been a hot topics for years. However, most of the work has been done by auditing, which is both human and time expensive. This project is meant to build a risk analysis model based on the Bayesian network, so as to recognize the possible frauding reasons and recommend better policy protection.

1. Background (Review of Related Literature)

Risks are inevitable, especially in the financial world where huge profits hidden huge loses. However, there is a huge information gap between every player. Unless by customers themselves, banks and insurance companies wouldn't know their customers, which will lead them to be exposed to unpredictable losses. Therefore, risk analysis and fraud prediction, with potential ability to detect abnormal signals and possible causes, are of great importance.

Risk analysis and fraud detection has been a hot topic for years. Traditionally people diagnose risks of companies through financial year reports and most of the job is done by auditors. C.verbano and K.Venturini summarized risk management (RM) application into nine mean streams including Financial RM, Insurance RM, Entreprise RM, Project RM etc. They also presented the ratio between studies done on risk types and study types, leading our project into a conceptual modeling study focus on ERM and FRM. Deeper analysis has been done by P.Pornprasitpol, they decompose the risks of enterprise into five layers: Jurisdiction, Strategy, Deployment, Operation and Events.

The Bayesian Network has been developed from Naive Bayesian, and is now one of the most important probability graph models in risk analysis. Concha Bielza and his colleagues introduced the method in detail while scholars presented many examples in modeling in industrial sectors. D.E.Nordgård developed the Bayesian network risk model for personal injury and line interruption. Yamine B and Hans J.P predicted food fraud type with the help of Bayesian network. Each node in the model has multiple classes which affect the outcome. The model would return to the maximum likelihood fraud result based on new data. And Vahid Khodakarami carried out a Bayesian network model to predict the project' total cost, which was a regression model for continuous response variables.

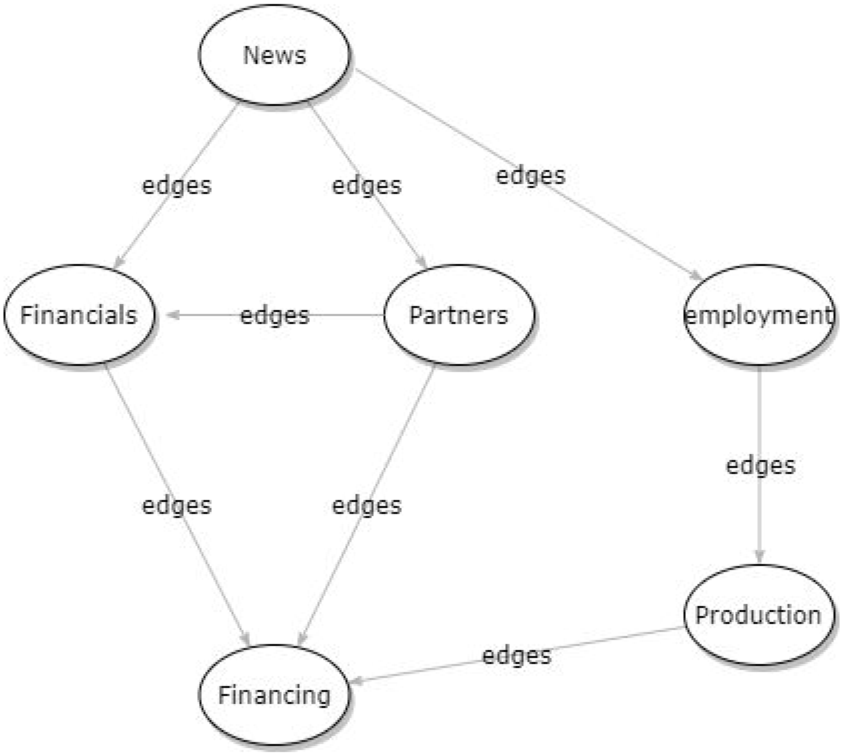
2. Introduction to the Project

We would like to know how the different factors would affect the financing of companies. Based on the data collected, visualization will be carried out regarding each variable. Model preparation including data cleaning and organizing will focus on dirty data as well as labeling continuous variables, e.g. <10,000, between 10,000 and 20,000 .etc. Next, we plan to build models in Python. Based on each model's performance, we will focus on improving each model's performance by tuning hyperparameters. After that, we would like to share our findings by building a website. We would like to have a feature that allows users to select a company, and the platform will show the information and our model predictions. We are also considering a feature that allows users to input datasets on different companies, and the platform will impute the missing values and make predictions automatically. The results could be downloadable as .csv files. In this step, we plan to use RShiny since it is convenient to import data and our models.

2.1 Methods

Models including random forest, SVM, LightGBM would likely be introduced in prediction. For fraud analysis, we do want to introduce the Bayesian Network, which is ideal for reasoning. The idea is clear by graph 1. And our task is to identify if these edges exist and the power of it.

graph 1



2.2 Potential User

This project is beneficial for investors who would like to know the potential depreciation and appreciation. Also for stockholders as well as employees to have a prospect of what may happen.

3. Introduction to the Dataset

We would like to focus the study on the U.S market. To study companies' financing situation and ability, we would like to collect information on products, employment, financial data, partners' behavior and news. See Table 1.

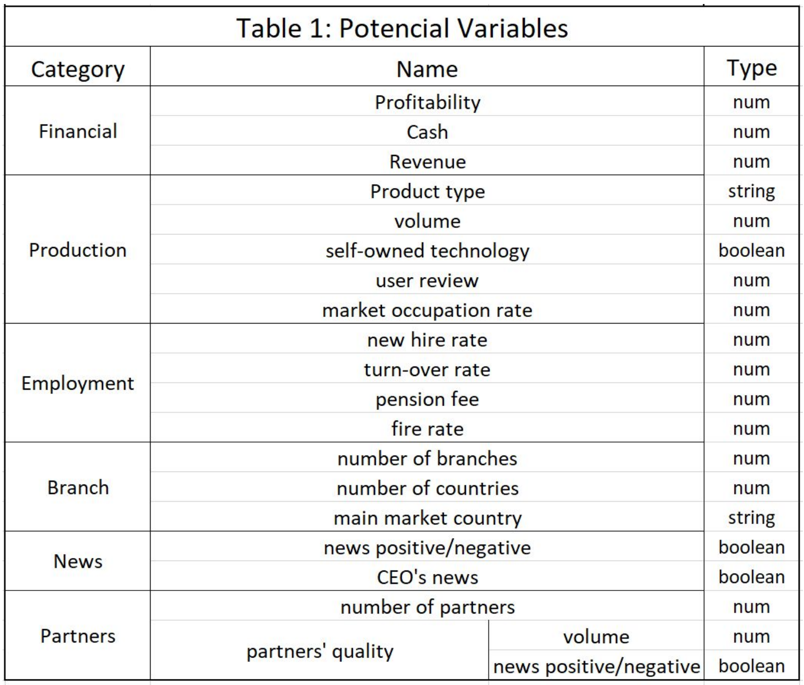


Table 1: Potencial Variables

Category	Name		Type
Financial	Profitability		num
	Cash		num
	Revenue		num
Production	Product type		string
	volume		num
	self-owned technology		boolean
	user review		num
	market occupation rate		num
Employment	new hire rate		num
	turn-over rate		num
	pension fee		num
	fire rate		num
Branch	number of branches		num
	number of countries		num
	main market country		string
News	news positive/negative		boolean
	CEO's news		boolean
Partners	number of partners		num
	partners' quality	volume	num
		news positive/negative	boolean

We have asked Professor Lin for help in data collecting. He told us to collect data by crawling, and will update with us some new ideas on Feb 14. Sources including U.S Securities and Exchange Commission (https://www.sec.gov/data), Companies' Year report, New York Times (https://www.nytimes.com/section/business) and also wikipedia (https://www.wikipedia.org/). Data will be cleaned and would be divided into classes for better classification.

The potential obstacles are the workload of collecting data. We have not decided how many companies will be involved in the analysis. And if some of the data is unavailable for free. We might lose important variables. Also, we searched data on the internet (kaggle, LendersClub.etc), but these data have anonymous company names, which do not allow us to do further investigation on potential risks.

4. Plan

Milestone 1: By milestone 1, we have investigated the background of the topic, such as why the topic is important and how the analysis of the topic could help in decision-making in real markets. We tried to get some datasets online, but the datasets we found would only allow us to do modeling. They do not reveal much about the underlying information of companies. We have come up with several variables that could be important in the project, and we have sought the professor's help and he will update with us on Feb 14. We have also determined the goal of the project and what methods we are going to use (see Section 2 for the details of the methods)

Milestone 2: In milestone 2, we will create some visualizations towards different variables in the datasets to get some initial insights on how each variable could influence the risks. We then aim to build some models to analyze the financing of each company to further understand the risks of investing.

Milestone 3: In milestone 3, we plan to focus on improving existing model performances by tuning hyperparameters. We will compare the accuracy and performance of each model and decide which model to use in our platform. We will also start building the platform to demonstrate our models and insights.

Final: We will mainly focus on building our platform. We will focus on both functionality and aesthetics to provide both information on the companies in our dataset and also allow them to get predictions on the features of the company based on their selection. We will also provide the insights we found during our analysis by adding both textual and graphical explanations.

Reference

1. C. Verbano, K. Venturini, "Managing risks in SMEs: A literature review and research agenda", Journal of technology management & innovation, vol. 8, no. 3, pp. 186-197, 2013.
2. P. Pornprasitpol, D. Ye and M. Sun, "An approach to risk events analysis for ERM using AHP and cluster analysis," 2010 IEEE 17Th International Conference on Industrial Engineering and Engineering Management, Xiamen, 2010, pp. 1009-1013.
3. Concha Bielza and Pedro Larrañaga. 2014. Discrete Bayesian Network Classifiers: A Survey. ACM Comput. Surv. 47, 1, Article 5 (July 2014), 43 pages. DOI:https://doi.org/10.1145/2576868
4. D.E. Nordgård, K. Sand,Application of Bayesian networks for risk analysis of MV air insulated switch operation,Reliability Engineering & System Safety,Volume 95, Issue 12,2010,Pages 1358-1366,ISSN 0951-8320
5. Yamine Bouzembrak, Hans J.P. Marvin,Prediction of food fraud type using data from Rapid Alert System for Food and Feed (RASFF) and Bayesian network modelling,Food Control,Volume 61,2016,Pages 180-187,ISSN 0956-7135
6. Vahid Khodakarami, Abdollah Abdi,Project cost risk analysis: A Bayesian networks approach for modeling dependencies between cost items,International Journal of Project Management,Volume 32, Issue 7,2014,Pages 1233-1245,ISSN 0263-7863