Proposed title of the project

Pei-Ling Tsai (pt2534) & Yuan-Hsi Lai (yl4305)

Risk Analysis and Default Prediction for Taiwan Companies

2020/02/12

Abstract
In this project, we are going to do the risk analysis for companies in Taiwan. The sources we used for risk analysis are from financial statements, public opinions toward the industry in the global market, and toward a single company and the role of one company in the industry chain. After getting these sources, we will use NLP methods for sentiment analysis on public opinions and news for companies to train weights for different sources and multiply the weight with the score in each data source to obtain a risk value. With the risk value, we will compare it to the company's credit rating to check the result. After training the model, we will use the model to deduce the default prediction.

1. Background (Review of Related Literature):

This topic is about using risk analysis methods to get the default prediction for companies.

In the beginning, we considered using the open source dataset from UCI machine learning repository. However, we found that the data is out of date and cannot represent the current economic environment now. Moreover, we found that the risk analysis methods used in banks typically gain information from historical data and this data is insufficient to represent the instant risk of one company.

Therefore, we start to look for companies in our real world. With this in mind, we want to seek different ways to help us gain the latest and more information needed to estimate the week. Thus, we decided to take the public opinions into consideration. The idea is: if customers are optimistic to one company, they are more likely to invest in this company and then the company can gain more cash flow. Nevertheless, evaluating one public opinion and determining it is positive or not need analysis of words and multiple training sessions. Our goal in this project is generating a useful model to do the prediction.

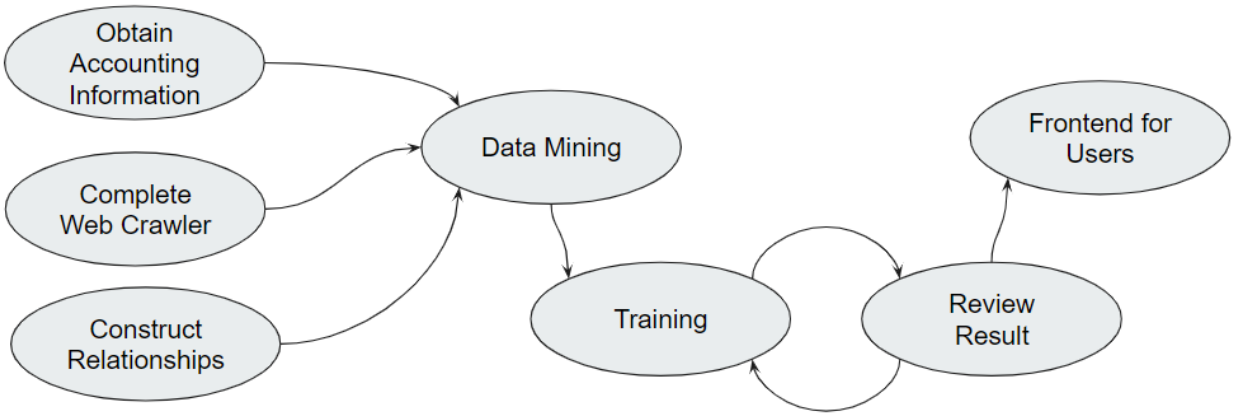
2. Introduction to the Project:

To analyze the risk of a company, we believe there are four important factors: accounting information, industry prospect, reputation, and relations.

1. Accounting information:
   In this part, we will evaluate numbers which come directly from financial statements. The formulas we will use are as the following:
   a. Profitability Ratios (e.g., ROE, ROA)
   b. Leverage Ratios (e.g., Debt to Assets Ratio)
   c. Coverage Credit Analysis Ratios (e.g., Interest Coverage Ratio)
   d. Liquidity Ratios (e.g., Cash Ratio)

2. Industry prospect:
   To achieve the industry prospect, we will do data crawling on certain industrial keywords to find news about the industry. For example, if we want to do analysis on semiconductor industries, we can crawl news about the including the word “semiconductor” and perform sentiment analysis to know how this industry is performing recently. Also, we might assign some weight to the sentiment score according to the date of the news.
3. Reputation:
   The reputation includes reputation of the company, reputation of the president/CEO, reputation of the management inside the company. To get this data, we same perform data crawling on news about the certain company name, the president/CEO name, to achieve news from the internet. In addition to the information from the news, we also want to get gossip from social media. We believe the social media contains some information that public news wouldn’t. For example, the management risk or the culture in the company can only be achieved this way, so we also do crawling on social media. After crawling, we also perform same sentiment analysis as above to get positivity of the comment/post/news.
4. Relation:
   The relation between companies and products are also important. For example, if two companies produce the same product, when A company is performing well and takes a lot of parts of the market, the B company might be in some risk. Also, if A company is the client of B company, when B is not performing well, A is also affected. Moreover, when certain products are suddenly profitable, the company producing them are also benefited.
   To produce the relation and do analysis on it, we are planning to find out the relationship between companies and products and use the graph database for our calculation.

The overall process will be like below:



We first do crawling on news, social media and accounting information. Then, we do sentiment analysis on crawled news and comment to form features. We also find relationships and give weight between them and combine them with the features that we chose from the accounting information and the sentiment. Using the combined features of a company, we can train a linear model for scoring the risk of certain company. The potential methods and tools we will use are XGBoost, LightGBM and Multilayer Perceptron. These are machine learning methods that we will try, and we will also use SnowNLP, an open source NLP toolkit for Chinese sentiment analysis.

3. Introduction to the Dataset:

Since we obtain data sources from four parts, there are four different datasets:

1. Accounting information: These data can be crawled from Taiwan Market Observation Post System.
2. Industry prospect: Parsing sentences from news, analyzing adjectives, adverbs, verbs and nouns. Collect those words that can represent positive or negative opinions.
3. Reputation: The data will be crawled from news and social media including: news of CEO/president, news of company, social media posts of company, comments on the posts.
4. Relation: We will construct the relation based on text mining on articles, news and also by hand crafting.

4. Plan:

Plans between every milestone and final. Please be specific.

Milestone	Finish Date	Items
1	2/14	a. Complete the proposal
2	2/28	a. Obtain financial statements b. Complete the Web crawler to obtain public opinions c. Construct relationship between companies in one industry
3	3/13	a. Data Mining
3	3/27	a. Training the model
4	4/10	a. Review the result b. Fix errors
4	4/24	a. Design user interface

Reference:

- Credit Risk Analysis and Prediction Modelling of Bank Loans Using R: https://www.researchgate.net/publication/309626126_Credit_Risk_Analysis_and_Prediction_Modelling_of_Bank_Loans_Using_R
- Understanding Default Risk in Bond Investing: https://smartasset.com/investing/understanding-default-risk-in-bond-investing
- UCI ML Repository: https://archive.ics.uci.edu/ml/datasets.php
- Credit Analysis Ratio: https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-analysis-ratios/
- Taiwan Ratings: https://www.taiwanratings.com/portal/index.gsp