**Risk Analysis and Default Prediction for Taiwan Companies**

|  |  |  |
| --- | --- | --- |
| Yuan-Hsi Lai  Dept. of Electrical Engineering  Columbia University  yl4305@columbia.edu | Pei-Ling Tsai  Dept. of Computer Science  Columbia University pt2534@columbia.edu |  |

**Abstract**

*Nowadays, the world becomes more complicated than before. Many companies that perform well may still default on their commitments to pay the money back. The reason is: the status of a company is not only affected by its financial statements, but also the social events and the confidence its investors have. For example, during the period affected by COVID-19, Boeing’s stock price dropped significantly than ever. Therefore, instead of using traditional financial ratios (e.g., Altman Z-score), we are going to use machine learning techniques to do the default prediction. Here we use the Taiwanese companies as a case study. Why we choose them as the dataset is because the financial statements are available on the “Market Observation Post System”[1] and we are more familiar with the social media in Taiwan.*

*Before training, we collect data from financial statements, news and social media, and company relationships in the industry. After that, we use sentiment analysis to analyze the public opinions, transferring them into concrete scores. Then we put all the numbers into vectors. Finally, we use XGBoost [2], a distributed gradient boosting library, to do the default prediction.*

*After training, we do the evaluation and get R2 score [3] equals to 0.9, which means that the prediction’s direction is highly consistent with the true results. Thus, we use the model to do the default prediction and represent the result on our Webpage (*[*http://146.148.63.155:5001/index*](http://146.148.63.155:5001/index)*).*

# **Introduction**

# In this project, we aim at obtaining information from many facets. It is important because the more useful information we get, the more precise the model can achieve. The first question here is: Which data should we get? And which key words should we use for searching?

# In financial statements, they contain many numbers and we can find many ratios which use these number to represent different facets. Since we need to do default prediction, we choose ratios that can show a company’s cash flow and ability to pay money back. For news and social media, we target at four categories: company names, products, industry and CEO names. With these categories, we can know how people think about them. For example, we can read from news that SSD is more popular than disks these years. Thus, we can expect companies that produce SSD would be more lucrative than companies produce disks. However, this method only works given the presumption that this company is famous enough and we can find it mentioned in many discussions. For those that are not well-known, we need to find another way to evaluate them. Thus, we decided to use the relationship among companies to complement the shortage of public posts. For example, for a laptop firm, if the company that sales memory to it performs well this year, then we can reasonably assume that the laptop firm performs well this year since it means that there is a large demand of laptops.

# After having these data, there is another problem: how to transfer the data into concrete numbers to compare? With this in mind, we adopt the sentiment analysis, which gives us scores regarding how positive or how negative a post is.

# **Related Work**

There are quite a lot of match result prediction methods, L.M. Hvattum and H. Arntzenin in [4] used ELO ratings for match result prediction, F. Owramipur in [5] used Bayesian Network for Spanich League football match prediction, and in [13]. In [4], the work simply calculated the two teams’ ELO difference and used a regression model to predict the result. However, due to the changing environment of football league nowadays, this method could fail by not considering a team’s recent performance. In [5], it used several reasonable factors including weather, psychological state, player age, injured player… etc., to predict the match results. The factors used are pretty good but however, factors like weather, injured players are very detailed data that was hard to collect. Most importantly, the 92% accuracy that it reached was trained and tested without splitting the data and was only based on a single season. For [5], it claimed that it reached an average of 75% accuracy in predicting three different season’s matches. However, the attributes that it used was team shots, team fouls, yellow cards, and red cards, which are useless inputs for an unplayed upcoming match since the only data we know about an unplayed match are only the teams and the players.

# **Dataset**

As mentioned in previous section, our data is composed of accounting data, news and PTT [5] posts, and relationship between companies.

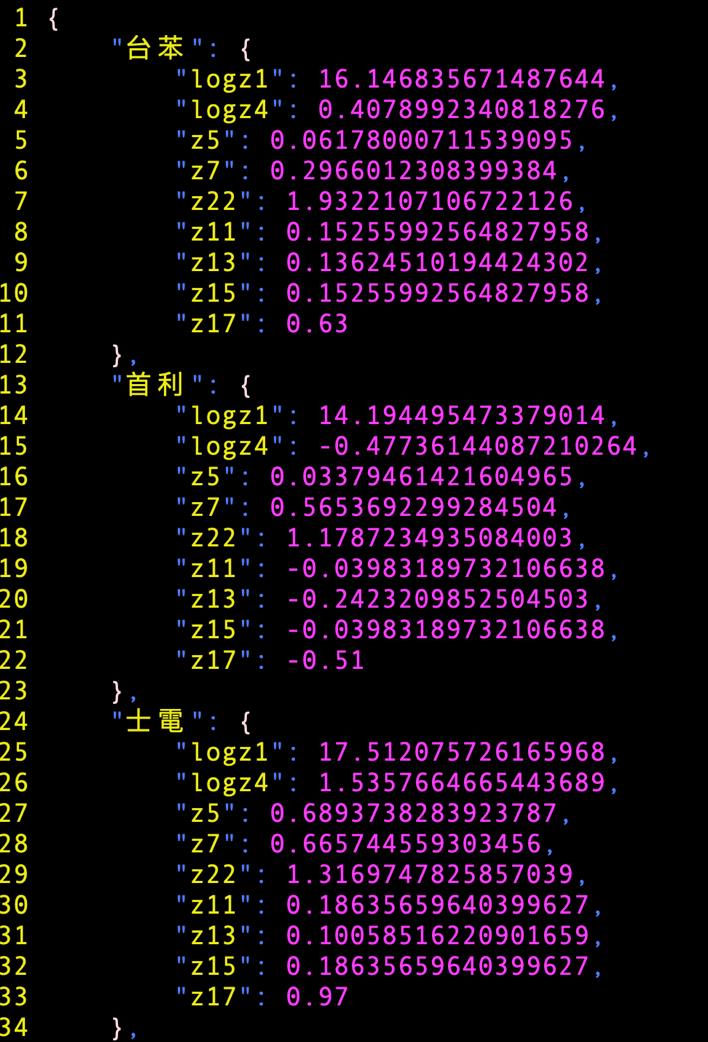
## 3.1 Accounting Data

We used nine financial ratios to represent the ability a company can pay money back. ***Table 1*** below shows these ratios and directions that show if the numbers mean good or not. After counting these ratios, we used JSON format to save company name with its vector into a file (***Figure 1***.). Each JSON file saves all the vectors of one quarter.

Table .

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Variable | Ratios/Numbers | Direction (Good) |
| Firm size | Z1 | Total asset value |  |
|  | Z4 | Book-to-market value |  |
| Financial leverage | Z5 | long-term debts/ total invested capital |  |
|  | Z7 | Total debt/ total capital |  |
| Profitability | Z11 | Operating income/ received capitals |  |
|  | Z13 | Net income before tax/ received capitals |  |
|  | Z15 | Gross profit margin |  |
|  | Z17 | Earnings per share (EPS) |  |
| Liquidity | Z22 | Quick ratio |  |

*Table 1. Financial ratios in four categories.*



*Figure 1. Ratios saved in dictionary format in ratios2018Q4.json.*

## 3.2 News and PTT Posts

### For news, we obtained from the Google search engion;

### for social media, we tried to use Twitter. However, we

### found that only premium member of Twitter can get

### posts over 7 days. Thus, we decided to change the social

### media to PTT.

### Searching key words

We divided them into four categories as the following, which can reveal the status of one company:

* Company names
* Products
* Industry
* CEO names

### Searching period

To align to the duration of financial statements, we use one quarter (3 months) as our searching duration. And we will put the posts of one year (4 quarters) together via the dictionary format in one JSON file. The format is: {“{DURATION}”: “……”} as ***Figure 2*** below.

A picture containing indoor, white, room

Description automatically generated

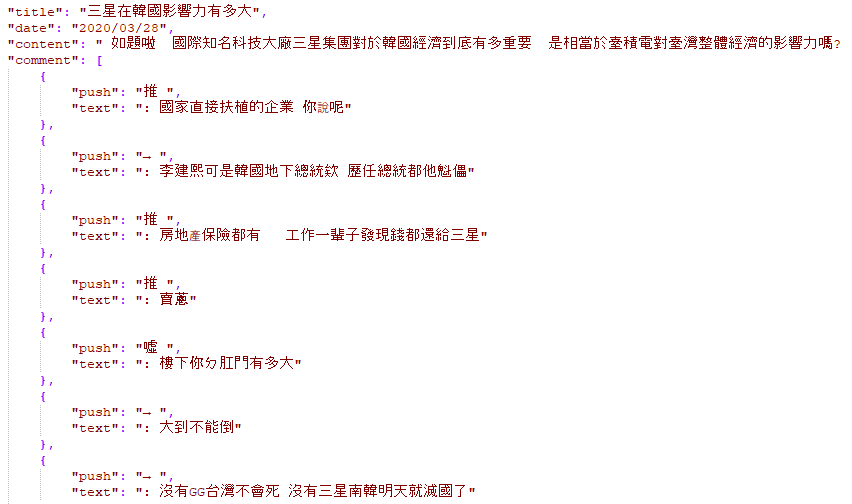
*Figure 2. The dictionary format for news posts*

### Preprocessing for news posts

In news searching, some small companies that are not famous, so the Google search engine would fill in unrelated articles to the search result, especially when the date is far from now. Thus, we used the filter to eliminate those articles. Although it would make the result list to be empty sometimes, it can prevent those unrelated articles from affecting the sentiment score.

### Preprocessing for the PTT posts

We managed to crawl posts by keyword and save the result posts by their dates. For each post, we also keep the comments and the like or dislike they have left. ***Figure 2*** shows an example of a PTT post with keyword “三星”(Samsung).



*Figure 3. The format saved for PTT posts.*

### Relationship between Companies

In this part, we used Ardi [6] to build the graph database. We obtained “distances” among companies by referring to the following facets:

* Upstream
* Downstream
* Competitors.

As mentioned before, these three numbers can help us evaluate companies that are not well-known by referring to the scores their “neighbors” get.

A picture containing fireworks, light, large, sitting

Description automatically generated

*Figure 4. Relationship between companies*

# **Methods**

In this section, we are going to introduce the data preprocessing, the machine learning model, problems happened during training and the evaluation of the model. ***Figure 5*** below shows the pipeline of the prediction model.

A picture containing screenshot, remote, sitting, white

Description automatically generated

*Figure 5. The pipeline of the prediction model*

## **Data Preprocessing**

For financial ratios, we simply put these nine ratios into the vector. For news and PTT posts, we use sentiment analysis to convert the public opinions into sentiment scores to represent the positive or negative reaction toward companies. For relationships in graph database, we……(TBD)

### Sentiment Analysis

#### SnowNLP

It is a Python package that used to get the sentiment score of one sentence. First, it needs to cut a sentence into pieces (words or phrases). Then it would use the “part-of-speech” concept, giving each piece a “tag” (e.g., noun, verb, adverb, … etc). After that, it counts the probability of the piece given the tag. Later, it uses the Naïve Bayes classifier to count the overall probability of the sentence. In the end, it will know the extent of positive or negative of the sentence.

#### Jieba

It is a Python package that is used for Chinese text segmentation. The can cut a sentence into words or phrases. In this model, we used Jieba to cut sentences before the sentiment analysis

With the help of SnowNLP, we got 81.8 % accuracy. The result is not bad; however, we found it failed to resolve the semantic and polysemy problem in Chinese. Thus, we used BERT instead.

#### Bidirectional Encoder Representations from Transformers (BERT)

Bert is a Pre-trained language model using Wikipedia data. It used word embedding in the begin, then it did the pooling regarding the results. After that it did the sentence embedding with random forest [7]. We got 92.4% accuracy in the end.

### Methods to get “distance” from graph database

## Model

During the training, we first chose the LSTM model, since it takes the timestep information into consideration. However, it suffered an issue, so we adopted XGBoost in the end. We will have more discussions in the following part.

### Long short-term memory (LSTM)

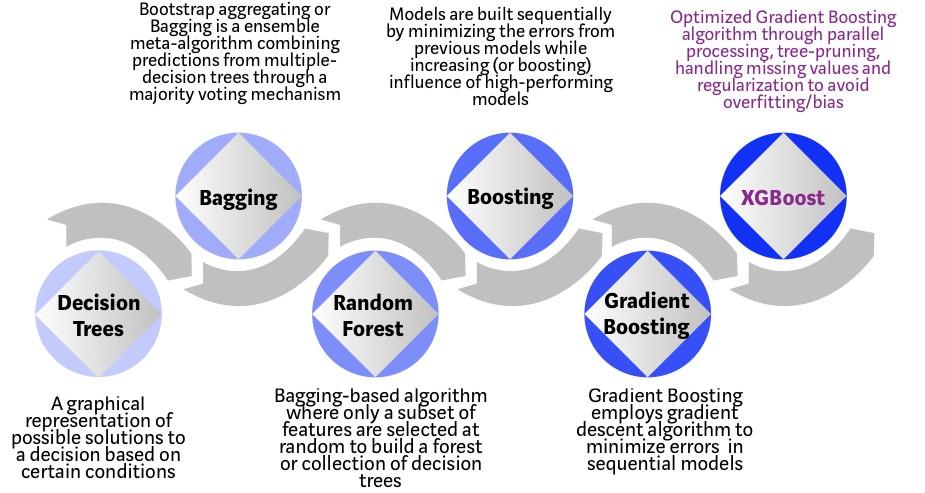
Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. It is developed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs.

In this project, we used [“Q1”, “Q2”] together for training data and [“Q3”, “Q4”] as testing data. Since our labels are not binary or classes; instead, they are numbers that are proportion to the numbers in the vector, we used ReLu [8] as the activation function. We got accuracy of 94% in evaluation. However, we found that LSTM is more suitable to data with many timesteps. For example, the trend of the everyday stock price. In our case, we only have 1 timestep (2020Q1) in our data for prediction because we use quarter as the unit of one timestep. Thus, we changed to XGBoost model.

### XGBoost

XGBoost model evolves from decision tree. It is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. It is a scalable and accurate implementation of gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms, as it was built and developed for the sole purpose of model performance and computational speed.



***Figure 6: Evolution of XGBoost Algorithms***

# **Experiment**

In this section, we will demonstrate how we train and evaluate the XGBoost model.

## **Prepare Labels**

For training and evaluation, we both need labels to know how good the result is. In the begin, we tried to get credit analysis done by financial institution for companies. But it needs to pay for the report. Thus, we use to stock prices, debt-to-equity ratios and sentiment scores, normalizing these numbers and giving equal weights to them. After that, it forms a sequence of numbers as our labels.

## **Training**

For the XGBoost model, the parameters are as the following:

* training data: 0.67 of ["Q1”, “Q2”, “Q3”]
* testing data: 0.33 of ["Q1”, “Q2”, “Q3”]
* regression function: linear
* learning rate: 0.6

During the training, we found that some companies didn’t have financial ratios. The reason is: these companies register in other countries or do not register in the open market.

The workaround we used is : Fill in the average of each factor (the average of values in one column) and time a penalty of 1.2. We put the penalty here is because we want to convey the concept that we should be more conservative to those companies who don’t provide sufficient information.

## **Evaluation**

We used the training model to predict our testing data. After that, we use R2 score to judge the performance. The score we got is 0.9. This means that the prediction is highly align to labels.

## **Prediction**

# **System Overview**

Webpage introduction

# **Conclusion**

In this paper, we designed several models to analyze the English Premier League data. Observing the results, we could see that there are still room for improvement. The main difficulty of our work is that the change in recent football league are huge. Every season, there are team leaving and joining, player leaving and joining, and good teams don’t always win games. Therefore, the possible improvement of our work relies on more specific details in the data or possibly more useful data. Though we have only four seasons of data and we could have collected more, it is useless for data too long ago since the players and teams will be completely different. We tried our best on the current dataset and we realized that it is still hard to design a model that effectively predict the result for today’s matches.

# **References**

1. Premier League Football News, Fixtures, Scores & Results [*https://www.premierleague.com*](https://www.premierleague.com)
2. English Premier League News, Stats, Scores – ESPN