Description

Our program is named “Corporate Investment Decision-making”, aiming to employ the F early warning model to recognize magnitude of risks based on principal component analysis and logistic regression analysis, and to make wise decisions about which company to invest in as an investor.

Business Computing Project

Corporate Investment Decision-making

成员：冯圣飒、杭慧丽、乐雅馨、陆之东、王棠

```
lines=open("we are the best.txt","r").readlines()
with open("we are the best","w")as fout:
    for line in lines:
        line = line.strip().split()
        yidiandian= line[-1]+"\t"+line[0]+"\t"+line[1]+"\t"+
line[3]+"\n"
        fout.write(yidiandian)
fout.close()

F_score = open("we are always the best.txt","r").readlines()
with open("we are always the best","w")as fout:
    for number in sorted(F_score,reverse=True):
```

Table of Contents

Division of work

乐雅馨 responsible for PART1 coding

杭慧丽 responsible for PART2&3 coding

王莹 responsible for slide

冯圣飞 responsible for the content of the report

陆之东 responsible for the proposal& designing the cover of the report

Preparation

Data to Collect

What is EDGAR?

The SEC's EDGAR database provides free public access to corporate information, allowing us to quickly research a company's financial information and operations by reviewing registration statements, prospectuses and periodic reports filed on Forms 10-K and 10-Q. We also can find information about recent corporate events reported on Form 8-K but that a company does not have to disclose to investors.

EDGAR plays an essential role in our project in the process of collecting data, which we'll mention later.

Preparations

First of all, we learn about the structure of HTML Document and HTTP communication roughly. Then, we do some research about Python Library to parse HTML Document and realize HTTP communication (including “requests” and “Beautiful Soup”).After that, we tried to have a rough understanding of the regular expression (but not use it in the program).At last, we review the tools and calculation methods we need to analyze the corporations’ financial statements.

The Model we used

An early warning system (EWS) is a system which is used for predicting the success level, probable anomalies and is reducing crisis risk of cases, affairs, transactions, systems, phenomena, firms and people. Furthermore, their current situations and probable risks can be identified quantitatively. Financial EWS is a monitoring and reporting system that alerts for the probability of problems, risks and opportunities before they affect the financial statements of firms. Nearly, all of the financial EWSs are based on financial statements. Balance sheets and income tables are the data sources that reflect the financial truth for early warning systems. In essence, the early warning system is financial analysis technique, and it identifies the achievement analysis of enterprise due to its industry with the help of

financial ratios. As we can see from above, an early warning system can serve as a strong financial tool for investors.

The Model we used — — F-score model(More details in“F-score model.doc”)

As we know, liquidity can measure the ability of an enterprise to meet the cash requirements at any time and we can compare the “Current Ratio” among companies to get the result. Profitability is another important factor to measure the performance of a company, which we can use the “Return On Equity” to evaluate it. Considering these two factors into account, we use the F-score model(W1=CR,W2=ROE):

F=-0.1774+1.1091W1+1.9271W2+0.1074W3+0.0302W4+0.04961W5

In our project, we give a definition of a simplified F-score model. Given the different impact each variable has on the total value, we take current ratio and ROE as primary variables and neglect the secondary ones, which leads us to the formula:

F=-0.1774+1.1091W1+1.9271W2.

The Development Process

Part1

Skeleton

As some unpredictable errors may appear when Python program runs for a long period, we decide to divide this part of program into three steps, so that the running time of each step can be decreased and this kind of error can be avoided. The first step is to search the keyword of "CHINA" in the SEC.gov | EDGAR | Search Tools and get the CIK codes of the corporations which have 10-Q or 10-K statements. We save those CIK codes in the file"10-XList".The second step is to get the URLs corresponding to the 10-Q or 10-K statements through the CIK codes in "10-XList" and save them into the file "chartUrls".

The third step is to visit the statements through the URLs in "chartUrls" and from which get the data and the information we need.

Some Details

Step one: We first operate the search by ourselves to obtain the URL needed and analyze its composition. We found that the argument "country=F4" corresponds to CHINA, "start" corresponds to the starting item and "count" corresponds to the amount of companies shown on each page. After several tries, we found that the valid range of count is 10-100, and then we set it to 100 for convenience. We put the corresponding CIK code of each company into the CIK list, after that, we visit the URL each CIK corresponds to. We searched through the web

page obtained and saw whether there exists a 10-K or 10- Q statement. If that exists, we wrote the corresponding URL into 10-X list. Only 1500 companies were found under the tag of CHINA. We repeated the process until all the companies were searched through and judged. We could get the URL of all the companies that have a 10-K or 10-Q statement.

Step two: After Step one, we found that the URL we obtained corresponds to an index. We searched through the index and found the URL corresponding to the files in the list and saved it in 'chartUrls'.

Step three: After analyzing several files, we found many formats. Data are mainly saved in <font>, <p> or <td>, and those tags are usually surrounded by <tr>. We judged every <tr> in the file and got the <tr> we needed. We then took the first two valid <td>, take the average value of them and saved them in data.txt in the form of "CIK\tTCA\tTCL\tTSE\tNI\n" for the following process.

Comments: CIK——CIK code; TCA——Total Current Assets; TCL——Total Current Liabilities; TSE——Total Stockholders' Equity; NI——Net Income.

Part2

Step one: Read the data. The data we need is already saved in the data.txt. Therefore, the first step to do is to open and read the data.txt. And there

are five lists in the text, including CIK—CIK code, TCA—Total Current Assets, TCL—Total Current Liabilities, TSE—Total Stockholders' Equity and NI—Net Income.

Step two: Process the original data. On account of commas leading to errors when we compute the data via python(for example: 123,456,789),first we remove the commas by using string's built-in function——replace. And then we can turn them into "float" type.

Step three: Choose the rational financial ratios and compute them out. Formulas: Current Ratio equals to Total Average Current Assets divided by Total Average Current Liabilities. Return On Equity equals to net income divided by average total equity. Obviously, python can use previous data to accomplish this task and we save the calculated ratios in "ratio.txt". For the calculated ratios in "ratio.txt": The first column is companies' names; the second column is CR; the second column is ROE; the fourth column is F-score.)

Part3

Step one: Try to sort the F-score in the descending order. Because we compare the size of the F-score, we move F-score to the first column, second as company names, third as CR, fourth as ROE(— — "ratio_sorted"). By using the function of "sorted", we can effectively sort a whole column(with other columns removing accordingly) and then we

save the sorted data in the "ratio_sorted2.txt".As the function is to sort data type of string, there is a problem like that "5" will be bigger than"44" - And we finally solve the sorting problem by using Excel.

Step two: Prepare for machine learning — The data model is KNeighborsRegressor. We use CSV format files to do machine learning. Hence, we output the "ratio_sorted2.txt" column by column (output three times in total——line_example.csv) and then respectively copy F, CR, ROE in the file of “ratios_sorted.csv”

Step three: Utilize KNN model.

Feature_columns = [col, for, col, in, data.columns]

X = data[feature_columns]

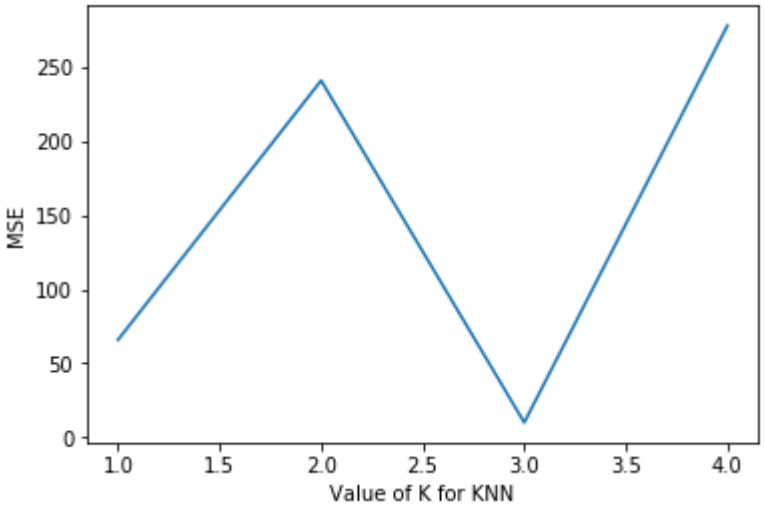
Y = data['F-score']

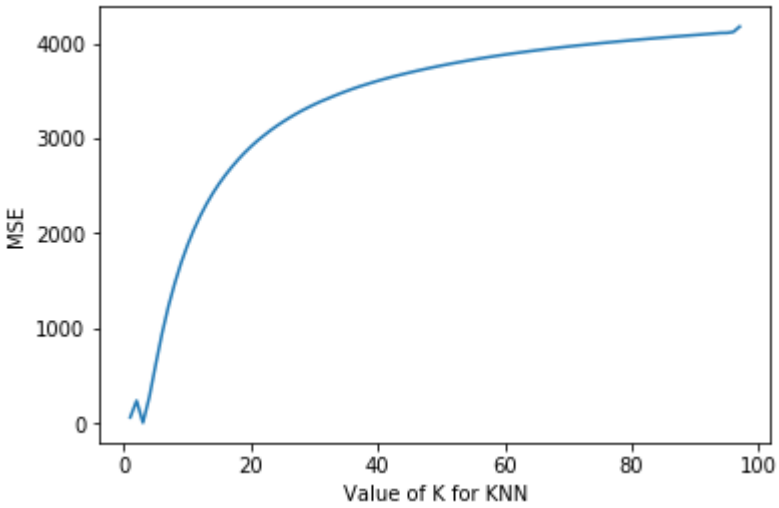
Test_size=0.2

At the beginning, we let K equal to 50.Due to the limit quantity of our data, the value of Mean Square Error and Root Mean Square Error is large (3765.50072067 and 61.3636759057).

Then we use a loop in the jupyter notebook.First we compare the output of the K with different values ranging from (1, 98) export the graph,and look for the most accurate one. The X axis is Value of K for KNN and the Y axis is the value of MSN (Mean Square Error).According to the graph, we find that the minimum error exists when K is in the range of (1, 5).Therefore, We change the range to (1, 5).Output the chart and we

have observed that when K=3,we get the highest degree of accuracy. When K is less than 3,It's over fitting.When K is larger than 3, it's under fitting.When K=3, Mean Square Error (KNN): 9.92373352601, Root Mean Square Error (KNN): 3.15019579169





Optimization

1. efficiency optimization

Due to the use of BeautifulSoup in this project (external library), which is used to markup language (includes html/xml), we need to choose document parser. After trying to use the "html.parser", "html5lib"and "lxml" parser, we finally found the efficiency of ”html.parser” is much higher than that of "html5lib". Though the efficiency of "lxml" is high, there are too much external dependence. The final selection is to choose "html.parser" as the HTML parser.

In the first step and the third step, first go through all <a> or <font>, to find out what we need. On average, each page needs to loop for 8000-25000 times. Later changing the object to the <TR>, it is reduced to 400-800 times, greatly improving operation efficiency. By going through strings, searching and skipping the inevitable failure of the cycle in time, we reduce the times of looping. By using linear search instead of the regular expression, the program improves the efficiency of operation.

2. Debugging process optimization

In the third step, there are some failures. In the following improvements, the failed URL is saved in "errlog.txt"for following processing, which is to use multiple-nested "try-except-block" to judge the documents, use if-else block to choose corresponding processing method and meanwhile add similarities when finding <TR>.It improves the rate of success of matching. The success rate increases from 12/150 to 123/376.

3. Output optimization

By using formatted strings,it generates the unified format strings. By saving the document, we pass the CIK in the first step to the third step and output the txt document.Then we can reduce the following workload. We choose \t as separator so that we can use string.split () to make them into LIST.

Outcome

Companies ranking among the top ten

① 790.9032840708696 0000042136

② 527.2005619089598 0001380706

③ 332.1362294007186 0001527675

④ 291.9217215654966 0001104904

⑤ 132.56385340352097 0000029952

⑥ 44.25995652844528 0001346352

⑦ 43.87003891991425 0001445196

⑧ 39.813375206461075 0001451264

⑨ 33.10389048404004 0001650101

⑩ 32.18423921312216 0001378270

In the end, through outputting companies ranking among the top 10.We can use this kind of information to assist investors in making investment decisions(In this situation, 0000042136 seems to be the best choice).However, the investors can't make their decisions merely depending on F-score.Given that the financial market is complex and changeable.Investors had better consider some other factors when investing companies. In this way, they can make wiser decisions and gain more profits.

Problems

- F-score isn't always the bigger the better. The F-score of small scale companies sometimes is higher comparing to some big companies, but this can’t directly illustrate that these companies do better.
- We finally choose excel to solve sequencing problem.Because the “sort”function in python may come out the result that 5 is bigger than 44.When we turn string to float for sorting, “CIK” can't move with F-score automatically,which makes it difficult to figure out which company's performance is better.
- F-score is a very simple reference.Different companies of different kinds have different developing potential,which can't reflect in the F-score.When making investment decisions, we should also consider other factors.

Conclusion

In our project, we make use of machine learning and Python library to take a first step at predicting the financial risk of a company. In the process of optimization, we have tried different methods to improve the efficiency and accuracy.

This project can be served as a strong tool for business investment. Investors can use this project to rationalize their investment decision-making. We believe that the computing project is able to assess the operation situation and predict the prospect of companies.