Risk Analysis and Default Prediction on U.S Companies

Xinyi Zhang
Statistics
Columbia University
xz2862@columbia.edu

Qiaoge Zhu
Data Science
Columbia University
zq2383@columbia.edu

Abstract

Analyzing risks of investment have been a large topic over year. In this project, we quantified company risks and likelihood of default by market capitalization. A prediction model based on random forest was built with $R^2 = 0.65$. The result was interpreted with the example of Shutterfly Inc. Also, we established an interactive platform based on R Shiny App, which support visualization, sentiment analysis, prediction and downloading the results carried out by our model.

1 Introduction

Risks are inevitable, especially in the financial world where huge profits hidden huge losses. 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. Risk analysis and fraud detection has been a hot topic for years. Financial managers, fund institutions, company owners, stock investor and financial market players seek for an efficient method to access the likelihood of companies' defaults. Therefore, risk analysis and fraud prediction, with potential ability to detect abnormal signals and possible causes, are of great importance.

Traditionally people diagnose risks of companies through annual financial reports and most of the job is done by auditors. The process is both time and labor expensive. And may neglect important signals. As the development of machine learning techniques, scholars have tried different models to address the problem. Former trials including logistic regression, discriminant analysis, etc. On top of that, in this report we discuss methods including lasso, random forest, elastic net and light GBM, as well as non parametric model such as KNN. To quantify the risk exposed to a particular company, our target variable would be market capitalization, referring to the total value of a companies' share.

This study is focused on building the model of predicting market capitalization. More insights are found by interpreting the model we build and visualization. We also implemented a sentiment analysis to know how sensitivity and objectivity people are about the companies. More than 500 companies operating in the United States are included in the sample.

2 Related Work

Studies have dealt with the discussion of risk analysis in different angles. C.verbano and K.Venturini summarized risk management (RM) application into nine mean streams including Financial RM, Insurance RM, Entreprise RM, Project RM etc [3]. 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 [7]. Our study would focus on financial risk management, and we collected annual financial report 2017 for each of the company.

For machine learning models, we would like to highlight the paper Flavio Barboza and her colleagues [2], in which they summarized the work done by previous studies and discussed the advantages and disadvantages of each model. In all tests without exceptions, boosting, bagging and random forest appear to have the best accuracy ratio compare to traditional models (logistic regression, ANN and MDA). While this paper gave us a lot of inspiration on model choosing, we find more detail on feature selection. "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment"[1] test the stability of feature selection methods. However, as the Pal et al. said that feature selection in the finance context "depends upon the individual judgment of the analyst or group decision-making. This makes the theoretical basis for the feature selection limited and less reliable"[8]. In short, financial modeling is highly connected with sample and selected features. In this project, we studied companies running their business in the United States, carried out feature selection and built our machine learning model with our data.

We also interested in how dose the public opinion would affect the market capitalization of a company. Sentiment analysis is carried out on company related news topic and articles. Similar method has been utilized on stock price prediction, but mostly the dataset would be on social media. An example would be "The impact of social and conventional media on firm equity value: A sentiment analysis approach"[9]. Study by Anjeza Kadilli[6] also showed a close relationship between stock returns and investors sentiment. Mu-Yen Chen et al[4] collected data from ChinaTimes.com, cnYES.com, Yahoo stock market news, and Google stock market news over an 18 month period for their analysis.

3 Data

3.1 Data Description

We collected 576 companies operating in the United States. There are 50 variables covering basic information and financial status and 8 were computed as ratios. All of the information is derived from structured data filed with the Commission by individual registrants as well as Commission-generated filing identifiers[1]. After removing redundant and irrelevant information, 38 variables are selected. An overview of all of our data would be as table 1.

The first column indicate the catalog of the information. Basic information are mostly characters for different labels, and act as the primary keys when we merging all the tables. Notice the name and Ticker here are both referred to the companies. Countryba indicates the registered office of the company rather than the operating area. Ratios are inspired by Costs of debt, tax benefits and a new measure of non-debt tax shields: examining debt conservatism in Spanish listed firms [5].

Table 1: Variables

	name	type
basic info	adsh	character
basic info	name	character
basic info	sic	character
basic info	countryba	character
basic info	cityba	character
financial	value	numeric
financial	Ticker	character
financial	Standard Poor Rating	character
financial	Mkt cap	numeric
financial	sales	numeric
basic info	country	character
basic info	industry	character
financial	foreign solvency capital	numeric
financial	free cash flow	numeric
financial	net income	numeric
financial	ltm sales gr	numeric
financial	ltm ebitda	numeric
financial	ltm net income growth	numeric
financial	sales gr	numeric
financial	sales 3yr average gr	numeric
financial	ebitda gr	numeric
financial	net income growth	numeric
financial	net income 5yr gr	numeric
financial	tangible assests	numeric
financial	tangible equity	numeric
ratios	debt/total book value	numeric
financial	cash from operations	numeric
financial	debt	numeric
financial	interest	numeric
financial	return on capital	numeric
financial	capital expenditure	numeric
financial	ebit	numeric
financial	ret on cap	numeric
ratios	debt/ebitda	numeric
ratios	cfo/interest	numeric
ratios	ebit/sales	numeric
ratios	debt/total assest	numeric
ratios	cfo/capex	numeric

3.2 Prepossessing: Missing Values

Prepossessing includes dealing with missing values, encoding and scaling. Figure1 shows the percentage of the missing values for each variable. We can see that some variables have high percent of missing values, especially those related to debt. While selecting features, we omitted some of these variables with a large number of missing values. For the remaining features, we used K Nearest Neighbors to impute them.

[1] https://www.sec.gov/dera/data

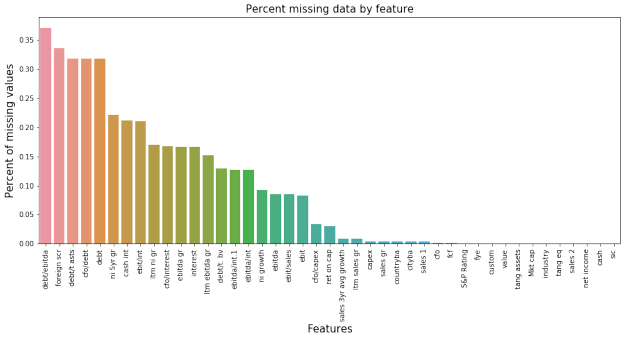


Figure 1: Missing Values

3.3 Prepossessing: Encoding and Scaling

3.3.1 Categorical Variables

For categorical variables, we used variables encoding to combine levels with few observations. For example, figure2 shows the number of companies in each country.

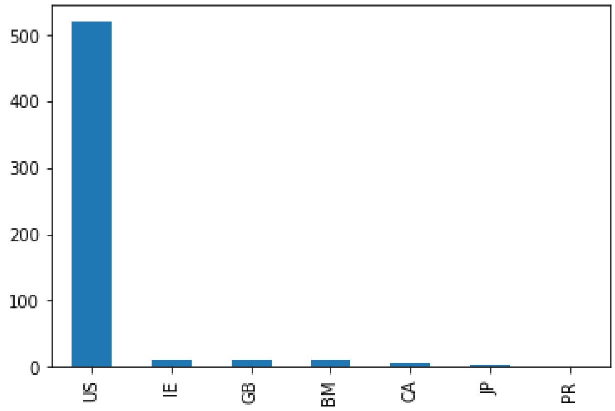


Figure 2: contryba

We can see that most of the companies are in the United States. Very few of them are in other countries. This variable is very imbalanced. When we do train-test split, there are chances that some levels are included in the training set but not in the test set. We would like to combine them to form a new level called *other*. Figure3 shows the histogram of the variable after combining several levels.

3.3.2 Continuous Variables

There are too many continuous variable in the dataset and we choose several of them to visualize to get some insights.

As shown in the figure4, we see that these variables are in different ranges, we would like to scale them in the modeling. We decided to use Standard Scaler since there are both positive and negative values without a bound for maximum and minimum.

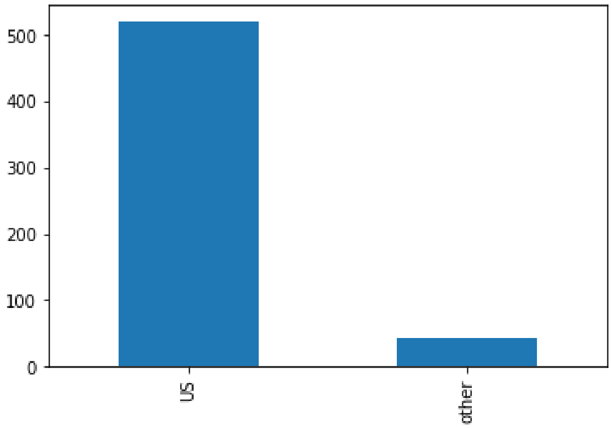


Figure 3: after combining

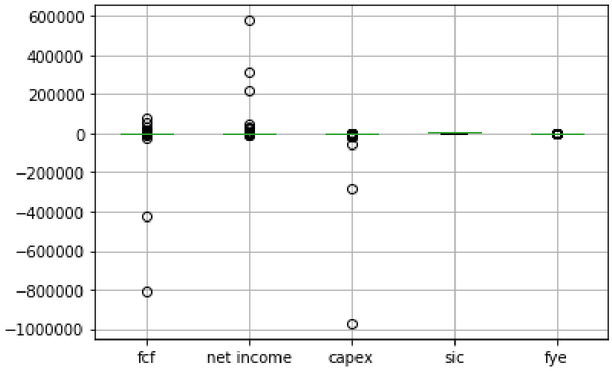


Figure 4: continuous variables

4 Methods

4.1 Machine Learning Algorithm

We tried linear regression, elastic net, light GBM, K Nearest Neighbors, random forest, and gradient boosting for our model part.

Linear regression is a method to linearly model the relationship between response and explanatory variables.

Elastic Net is a regularized regression approach which linearly combines the L1 and L2 penalties of lasso and ridge models.

Light GBM is a gradient boosting framework based on decision trees. It is known for fast and high-performance. Unlike other boosting algorithms, it split the tree leaf-wise rather than level-wise.

K Nearest Neighbors uses the K-nearest points to predict the value. For regression, it uses the average of the K-nearest points.

Random forest is a technique that works well with both regression and classification. It could also be used to reduce dimensions and treating missing values.

Gradient boosting is a method that can be used for both regression and classification. It produces a prediction model using an ensemble of weak prediction models, typically decision trees

4.2 Sentiment Analysis

Sentiment analysis is a method to analysis text data. A polarity score and a subjectivity score will be given. It helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences.

Process of sentiment analysis acts is shown in figure5.

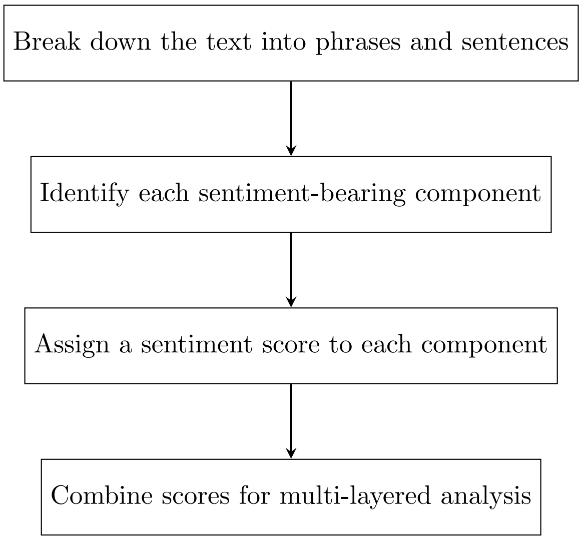

Break down the text into phrases and sentences
Identify each sentiment-bearing component
Assign a sentiment score to each component
Combine scores for multi-layered analysis

Figure 5: Process of Sentiment Analysis

We need to have a sentiment library in order to assign the score for each phrases and component in the text. The method is well-developed and the library TextBolb[2] in python would be ideal for the project. We didn't train the bag-of-word specifically for company words. The library allow us to compute subjectivity and polarity, also able to check the components cut down by the algorithm.

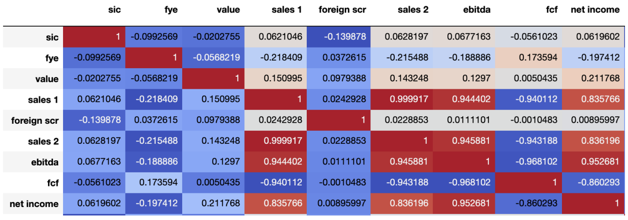
Yahoo Finance is a leading comprehensive website for finance information. It's ideal to represent public opinions, or media opinions of a company. One benefit is that the website will return the latest news about the company so that the result can be calculated with steaming data instead of historical data. With this advantage, investors would know exactly how subjectivity and sensitivity public are about the company. Moreover, the news on Yahoo finance are highly correlated with the companies' finance and operation status. Unlike results return by Google and other searching engine, data collected on finance website will provide more accurate information.

For scripting work, we used python libraries requests[3] to get the information and beautiful soup[4] to interpret the web pages.

5 Experiments

5.1 Correlation Matrix

Correlation matrix are introduced to remove features that are highly correlated to each other by looking at their correlations. Figure6 shows the correlation of some of the variables.



	sic	fye	value	sales 1	foreign scr	sales 2	ebitda	fcf	net income
sic	1	-0.0992569	-0.0202755	0.0621046	-0.139878	0.0628197	0.0677163	-0.0561023	0.0619602
fye	-0.0992569	1	-0.0568219	-0.218409	0.0372615	-0.215488	-0.188886	0.173594	-0.197412
value	-0.0202755	-0.0568219	1	0.150995	0.0979388	0.143248	0.1297	0.0050435	0.211768
sales 1	0.0621046	-0.218409	0.150995	1	0.0242928	0.999917	0.944402	-0.940112	0.835766
foreign scr	-0.139878	0.0372615	0.0979388	0.0242928	1	0.0228853	0.0111101	-0.0010483	0.00895997
sales 2	0.0628197	-0.215488	0.143248	0.999917	0.0228853	1	0.945881	-0.943188	0.836196
ebitda	0.0677163	-0.188886	0.1297	0.944402	0.0111101	0.945881	1	-0.968102	0.952681
fcf	-0.0561023	0.173594	0.0050435	-0.940112	-0.0010483	-0.943188	-0.968102	1	-0.860293
net income	0.0619602	-0.197412	0.211768	0.835766	0.00895997	0.836196	0.952681	-0.860293	1

Figure 6: Correlation table

We see that many features are highly correlated with each other. We need to remove one of the highly correlated features in our modeling. We used a threshold of 0.75, so if the absolute value of the correlations between two variables. For those two variables, we removed the one with with more missing values or the one that is highly correlated with more variables. In this step, we removed 18 variables.

5.2 Sentiment Analysis

We collected both headings of the most recent news and the article of the latest one showed on Yahoo finance. Sentiment score varies from -1 to 1, with -1 represents the most negative and 1 represent the most positive. Subjectivity is a score from 0 to 1, with 0 represent most subjective and 1 represent the most objective. In the table, title_sentiment is the sentiment

[2] https://textblob.readthedocs.io/en/dev/quickstart.html

[3] https://requests.readthedocs.io/en/master/

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

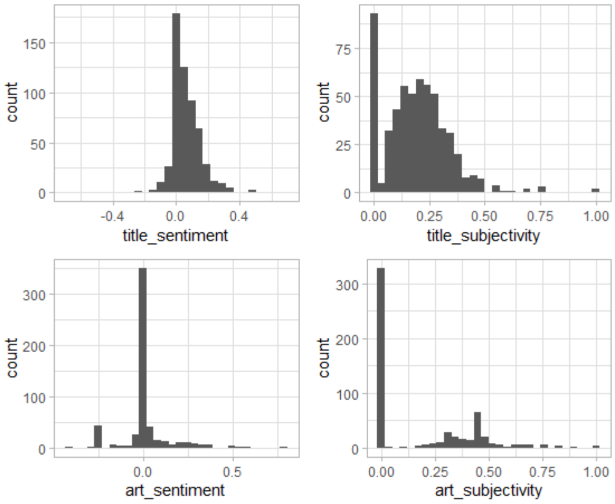


Figure 7: Result

score for recent news and art_sentiment is the sentiment score for the latest article. Title_subjectivity and art_subjectivity is the subjectivity score for the two collections. An overview of the results is showed in figure7.

5.3 Modeling

We used R^2 as are metric to evaluate the models. Linear models and light GBM do not have good model performances. Actually, their R^2 is less than 0, which means that they are even worse than simply using the mean to predict the values. Random forest and gradient boosting work well, each with an R^2 of about 0.65. We used random forest as our model to build the R Shiny app.

Random forest uses a strategy that generates feature importance by how well each feature could improve the purity of the node, therefore, is a good way to select important features. Figure8 shows the feature importance for each variable.

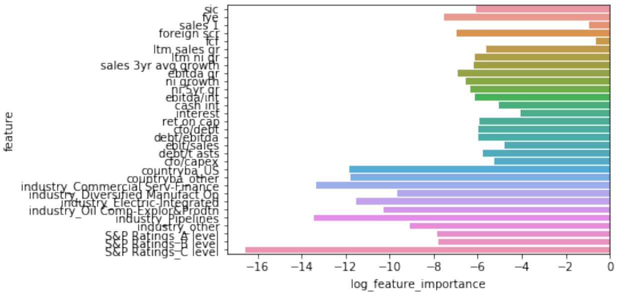


Figure 8: Feature Importance

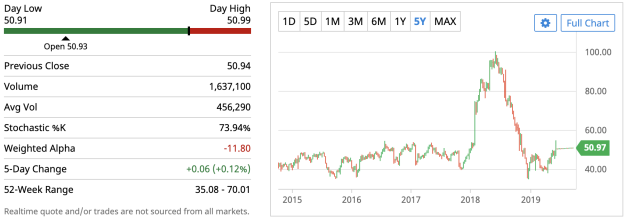
5.4 Interpretation

Market capitalization refers to the total value of all of a company's shares of stock. We researched on the topic and found stock risks could be reflected from market capitalization. Companies are categorized based on the value of their market capitalization.

Small-cap companies have a market capitalization of between 150 million to 500 million. These companies are bounded with large risks, but they are also more likely to have rapid growth. Mid-cap companies have market capitalization from 500 million to 5 billion. Stock prices of these companies are relatively stable. Large-cap companies have market capitalization of 5 billion or more, and their stock prices are stable and not likely to drop or increase drastically.

We used machine learning models to predict the market capitalization of each company with a pretty good R^2 achieved (0.65). We have checked several companies with our prediction and it seems that the theory works well in reality.

Take ShutterFly Inc as an example to support our model. As shown in the figure9, the real market capitalization is 3,116,754,944.0. Our model predict the market capitalization of ShutterFly Inc to be around 3,379,205,841.92, which is pretty similar. By the definition, this company should be categorized as a small-cap company and the stocks should be very unstable. So we also checked its stock prices, and found the 52 week range is 35.08 to 70.01, which is a pretty large range with the stock prices.



Day Low 50.91 — Day High 50.99

Open 50.93

Previous Close	50.94
Volume	1,637,100
Avg Vol	456,290
Stochastic %K	73.94%
Weighted Alpha	-11.80
5-Day Change	+0.06 (+0.12%)
52-Week Range	35.08 - 70.01

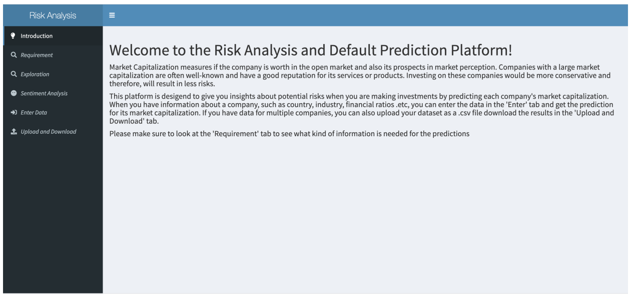
Realtime quote and/or trades are not sourced from all markets.

Figure 9: Stock Price of ShutterFly Inc

Additionally, from the figure9, we can see that the stock prices are experiencing drastic ups and downs, which further demonstrates the correctness of our methods.

6 System Overview

Shiny App is an interactive platform supported by R. We insert our models and data into the Shiny App. There are five functional pages in our app including visualization, data explanations, prediction, downloading and uploading datasets.



Welcome to the Risk Analysis and Default Prediction Platform!

Market Capitalization measures if the company is worth in the open market and also its prospects in market perception. Companies with a large market capitalization are often well-known and have a good reputation for its services or products. Investing on these companies would be more conservative and therefore, will result in less risks.

This platform is desigend to give you insights about potential risks when you are making investments by predicting each company's market capitalization. When you have information about a company, such as country, industry, financial ratios .etc, you can enter the data in the 'Enter' tab and get the prediction for its market capitalization. If you have data for multiple companies, you can also upload your dataset as a .csv file download the results in the 'Upload and Download' tab.

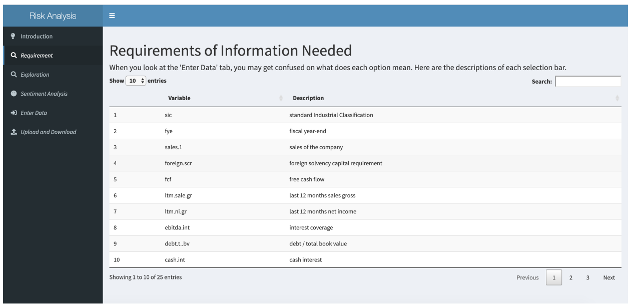
Please make sure to look at the 'Requirement' tab to see what kind of information is needed for the predictions

Figure 10: Welcome Page

Figure10 works as the welcome page of our platform. It provides description of the goal of the study as well as some instructions for users on how to use the platform.

6.1 Requirement

The requirement tab illustrates what kind of information to enter or provide in order to make the prediction work. The table shows the name and meaning of each variable. There is also a search bar so that users could search for a specific variable by keywords.



Requirements of Information Needed

When you look at the 'Enter Data' tab, you may get confused on what does each option mean. Here are the descriptions of each selection bar.

Show 10 entries

Search:

	Variable	Description
1	sic	standard Industrial Classification
2	fye	fiscal year-end
3	sales.1	sales of the company
4	foreign.scr	foreign solvency capital requirement
5	fcf	free cash flow
6	ltm.sale.gr	last 12 months sales gross
7	ltm.ni.gr	last 12 months net income
8	ebitda.int	interest coverage
9	debt.t..bv	debt / total book value
10	cash.int	cash interest

Showing 1 to 10 of 25 entries

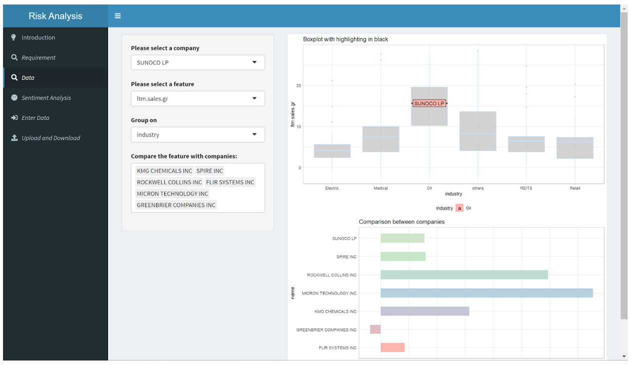
Figure 11: Requirement

6.2 Exploration

Data exploration will be donw by visualization, as shown in figure12. The upper boxplot shows the percentage position of the chosen company. The boxes can be grouped by industry, country and SP ratings. The lower bar chart is designed for comparison. Usrs are allowed to choose at most 20 countries in the list to compare the chosen variable.

6.3 Sentiment Analysis

This part shows historical data as well as real-time result of sentiment analysis. The table above presents the data we used in the prediction model (collected in April 2020). Four variables are included, indicating polarity and subjectivity of news topics and the latest articles. Users are allowed to search certain companies. Notice that 0 will exist as some of the companies do not have record on Yahoo Finance, or their article is forward from another website.



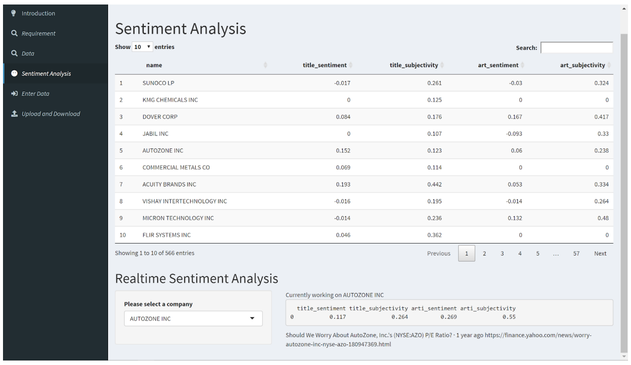
Please select a company: SUNOCO LP

Please select a feature: ltm.sales.gr

Group on: industry

Compare the feature with companies: KMG CHEMICALS INC, SPIRE INC, ROCKWELL COLLINS INC, FLIR SYSTEMS INC, MICRON TECHNOLOGY INC, GREENBRIER COMPANIES INC

Figure 12: Visualization



Sentiment Analysis

Show 10 entries

Search:

	name	title_sentiment	title_subjectivity	art_sentiment	art_subjectivity
1	SUNOCO LP	-0.017	0.261	-0.03	0.324
2	KMG CHEMICALS INC	0	0.125	0	0
3	DOVER CORP	0.084	0.176	0.167	0.417
4	JABIL INC	0	0.107	-0.093	0.33
5	AUTOZONE INC	0.152	0.123	0.06	0.238
6	COMMERCIAL METALS CO	0.069	0.114	0	0
7	ACUITY BRANDS INC	0.193	0.442	0.053	0.334
8	VISHAY INTERTECHNOLOGY INC	-0.016	0.195	-0.014	0.264
9	MICRON TECHNOLOGY INC	-0.014	0.236	0.132	0.48
10	FLIR SYSTEMS INC	0.046	0.362	0	0

Showing 1 to 10 of 566 entries

Realtime Sentiment Analysis

Please select a company: AUTOZONE INC

Currently working on AUTOZONE INC

```
  title_sentiment title_subjectivity arti_sentiment arti_subjectivity
0           0.117              0.264          0.269              0.55
```

Should We Worry About AutoZone, Inc.'s (NYSE:AZO) P/E Ratio? · 1 year ago https://finance.yahoo.com/news/worry-autozone-inc-nyse-azo-180947369.html

Figure 13: Sentiment Analysis

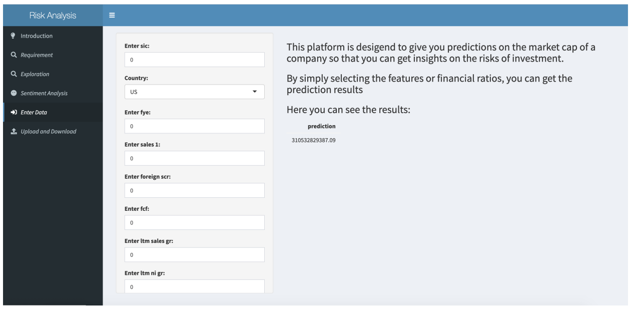
In order to achieve real-time analysis, we insert the python code into our R Shiny app. Users are allowed to search among the company list and the results would shown on the right side. The result table is in the same format as the upper table. More detailed information on the news topics, published date and link will also be presented.

6.4 Enter Data

The Enter Data tab is designed for a single prediction. In this tab, the users could enter the data on their own, and this platform will predict the market capitalization based the inputs. If we change the values of the features, the prediction will also change automatically.

6.5 Upload and Download

The Upload and Download tab is designed for multiple predictions. The users could upload a .csv file containing all the variables described in the requirement tab, then the platform will automatically predict the market capitalization for all of the observations.



Enter sic: 0

Country: US

Enter fye: 0

Enter sales 1: 0

Enter foreign scr: 0

Enter fcf: 0

Enter ltm sales gr: 0

Enter ltm ni gr: 0

This platform is desigend to give you predictions on the market cap of a company so that you can get insights on the risks of investment.

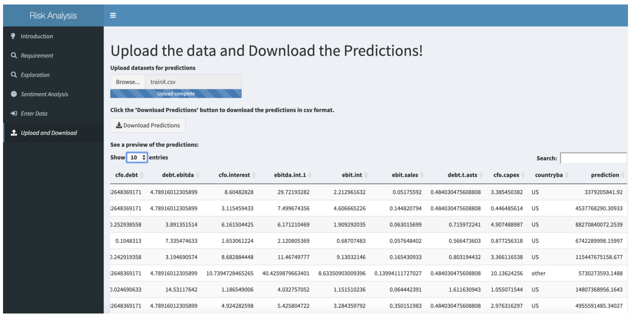
By simply selecting the features or financial ratios, you can get the prediction results

Here you can see the results:

prediction
310532829387.09

Figure 14: Single Prediction

After the process of prediction is finished, the table containing the data uploaded and the predictions can be viewed below. The users are able to search and turn pages to view the results. They could also download the predictions as a .csv file using the Download button.



Upload the data and Download the Predictions!

Upload datasets for predictions

trainX.csv

Upload complete

Click the 'Download Predictions' button to download the predictions in csv format.

See a preview of the predictions:

Show 10 entries

Search:

cfo.debt	debt.ebitda	cfo.interest	ebitda.int.1	ebit.int	ebit.sales	debt.t.asts	cfo.capex	countryba	prediction
2648369171	4.78916012305899	8.60482828	29.72193282	2.212961632	0.05175592	0.484030475608808	3.385450382	US	3379205841.92
2648369171	4.78916012305899	3.115459433	7.499674356	4.606665226	0.144820794	0.484030475608808	0.446485614	US	4537768290.30933
.252938558	3.891351514	6.161504425	6.171210469	1.909292035	0.063015699	0.715972241	4.907488987	US	88270840072.2539
0.1048313	7.335474633	1.653061224	2.120805369	0.68707483	0.057648402	0.566473603	0.877256318	US	6742289998.15997
.242919358	3.194690574	8.682884448	11.46749777	9.13032146	0.165430933	0.803194432	3.366116538	US	115447675158.677
2648369171	4.78916012305899	10.7394728465265	40.4259879663401	8.63350903009396	0.13994111727027	0.484030475608808	10.13624256	other	5730273593.1488
.024690633	14.53117642	1.186549006	4.032757052	1.151510236	0.064442391	1.611630943	1.055071544	US	14807368956.1643
2648369171	4.78916012305899	4.924282598	5.425804722	3.284359792	0.350151983	0.484030475608808	2.976316297	US	4955591485.34027

Figure 15: Multiple Prediction

7 Conclusion

7.1 Challenges

1. Historical news are extremely hard to find. We are only able to collect recent news, which would serve as an indicator for current market cap. However, after combing the result of sentiment analysis with the financial data, the result actually got worse. Thus we separate the part.

2. Since R Shiny is a new topic for us, we have encountered some difficulties while writing the code. For example, the structure of R Shiny is very unique. When we are writing the codes, unexpected bugs appear frequently. So we took some time to research on it.

7.2 Further Extensions

1. More information around companies should be collected and organized, such as subsidiaries, supply chain etc.

2. This study would be a good source studying company financial structures, and how different operations may affect the value of the company.

3. For the R Shiny part, we could work towards a more user-friendly interface. We could also work on enabling the platform for automatic feature selection and missing value imputation.

References

[1] Nisha Arora and Pankaj Deep Kaur. “A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment”. In: *Applied Soft Computing* 86 (2020), p. 105936.

[2] Flavio Barboza, Herbert Kimura, and Edward Altman. “Machine learning models and bankruptcy prediction”. In: *Expert Systems with Applications* 83 (2017), pp. 405–417.

[3] K. Venturini C. Verbano. In: *Journal of technology management innovation* 8.3 (2013).

[4] Mu-Yen Chen and Ting-Hsuan Chen. “Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena”. In: *Future Generation Computer Systems* 96 (2019), pp. 692–699.

[5] José A. Clemente-Almendros and Francisco Sogorb-Mira. “Costs of debt, tax benefits and a new measure of non-debt tax shields: examining debt conservatism in Spanish listed firms”. In: *Revista de Contabilidad* 21.2 (2018), pp. 162–175.

[6] Anjeza Kadilli. “Predictability of stock returns of financial companies and the role of investor sentiment: A multi-country analysis”. In: *Journal of Financial Stability* 21 (2015), pp. 26–45.

[7] D. Ye P. Pornprasitpol and M. Sun. In: *2010 IEEE 17Th International Conference on Industrial Engineering and Engineering Management* (2010).

[8] Rudrajeet Pal et al. “Business health characterization: A hybrid regression and support vector machine analysis”. In: *Expert Systems with Applications* 49 (2016), pp. 48–59.

[9] Yang Yu, Wenjing Duan, and Qing Cao. “The impact of social and conventional media on firm equity value: A sentiment analysis approach”. In: *Decision Support Systems* 55.4 (2013). 1. Social Media Research and Applications 2. Theory and Applications of Social Networks, pp. 919–926.