Bankruptcy Prediction with Statistical Learning

STAT432 Final Group Project

March 21, 2019

Group Information

Group Members

Mary Liu (NetID: zliu203)
Ziqiao Hua (NetID: ziqiaoh2)
Yixin Zhang (NetID: yzhng224)
Sian Liu (NetID:sianliu2)
Jiahui Zhao(NetID:jzhao71)

Introduction

Project Description and Problem of Interest

Prediction of firm bankruptcies have been extensively studied in the field of accounting to monitor the financial performance by all shareholders. With the introduction and expansion of statistical learning methods on financial analysis, bankruptcy rate estimation has taken on a new importance [1]. This paper uses an expanded database with more than fifty econometric attributes. The aim of this project is to examine the relationships between these parameters and develop an effective prediction model which allows forecasting the bankruptcy condition of a firm in the near future. In the project, we apply and compare some widely known statistical imputation techniques, such as Decision Tree, K-nearest Neighbor, Logistic Regression and K-Fold Cross Validation and evaluate the performance of these techniques by their accuracy rates.

Data Source and Description

The dataset we use is called `Polish Companies Bankruptcy Data Set` which is hosted by UCI Machine Learning Repository and collected from EMIS, a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013 [4]. In this project, we will use partial data called `3year` for bankruptcy prediction. It contains financial rates from 3rd year of the forecasting period and corresponding class label that indicates bankruptcy status after 3 years.

The data `3year` contain contains 64 variables and 10503 observations in total. Below are all the variables which are worth studying. The dependent variable is the class variables with levels 0 or 1, indicating the company bankruptcy or not. Some variables as financial ratio could affect the company be classified as bankruptcy or not. For example, the first variable is "net profit/total assets" which is return on assets (ROA), a financial ratio that shows the percentage of profit a company earns in relation to its overall resources. It is possible that the higher the ROA, the less likely the company will be bankrupt.

Variables

Independent Variables:
attr1 - net profit / total assets
attr2 - total liabilities / total assets
attr3 - working capital / total assets
attr4 - current assets / short-term liabilities
attr5 - [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
attr6 - retained earnings / total assets
attr7 - EBIT / total assets
attr8 - book value of equity / total liabilities
attr9 - sales / total assets
attr10 - equity / total assets
attr11 - (gross profit + extraordinary items + financial expenses) / total assets
attr12 - gross profit / short-term liabilities
attr13 - (gross profit + depreciation) / sales
attr14 - (gross profit + interest) / total assets
attr15 - (total liabilities * 365) / (gross profit + depreciation)
attr16 - (gross profit + depreciation) / total liabilities
attr17 - total assets / total liabilities
attr18 - gross profit / total assets
attr19 - gross profit / sales
attr20 - (inventory * 365) / sales
attr21 - sales (n) / sales (n-1)
attr22 - profit on operating activities / total assets
attr23 - net profit / sales
attr24 - gross profit (in 3 years) / total assets
attr25 - (equity - share capital) / total assets
attr26 - (net profit + depreciation) / total liabilities
attr27 - profit on operating activities / financial expenses
attr28 - working capital / fixed assets
attr29 - logarithm of total assets
attr30 - (total liabilities - cash) / sales
attr31 - (gross profit + interest) / sales
attr32 - (current liabilities * 365) / cost of products sold
attr33 - operating expenses / short-term liabilities
attr34 - operating expenses / total liabilities
attr35 - profit on sales / total assets
attr36 - total sales / total assets
attr37 - (current assets - inventories) / long-term liabilities
attr38 - constant capital / total assets
attr39 - profit on sales / sales
attr40 - (current assets - inventory - receivables) / short-term liabilities
attr41 - total liabilities / ((profit on operating activities + depreciation) * (12/365))
attr42 - profit on operating activities / sales
attr43 - rotation receivables + inventory turnover in days
attr44 - (receivables * 365) / sales
attr45 - net profit / inventory
attr46 - (current assets - inventory) / short-term liabilities
attr47 - (inventory * 365) / cost of products sold
attr48 - EBITDA (profit on operating activities - depreciation) / total assets
attr49 - EBITDA (profit on operating activities - depreciation) / sales attr50 - current assets / total liabilities
attr51 - short-term liabilities / total assets

attr52 - (short-term liabilities * 365) / cost of products sold) attr53 - equity / fixed assets
attr54 - constant capital / fixed assets
attr55 - working capital
attr56 - (sales - cost of products sold) / sales
attr57 - (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
attr58 - total costs /total sales
attr59 - long-term liabilities / equity
attr60 - sales / inventory
attr61 - sales / receivables
attr62 - (short-term liabilities *365) / sales
attr63 - sales / short-term liabilities
attr64 - sales / fixed assets

Dependent Variable:
class - the response variable Y: 0 = did not bankrupt; 1 = bankrupt

10 Observations

The head ten observations are listed below:

Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8
0.1741900	0.412990	0.143710	1.3480	-28.9820	0.6038300	0.2194600	1.12250
0.1462400	0.460380	0.282300	1.6294	2.5952	0.0000000	0.1718500	1.17210
0.0005953	0.226120	0.488390	3.1599	84.8740	0.1911400	0.0045718	2.98810
0.0245260	0.432360	0.275460	1.7833	-10.1050	0.5694400	0.0245260	1.30570
0.1882900	0.415040	0.342310	1.9279	-58.2740	0.0000000	0.2335800	1.40940
0.1820600	0.556150	0.321910	1.6045	16.3140	0.0000000	0.1820600	0.79808
0.1150300	0.036659	0.923450	112.6300	55.5800	0.0000000	0.1421500	26.27900
0.0098323	0.670660	0.135460	1.2393	-107.7300	-0.0014006	0.0134220	0.49108
0.2389500	0.554730	0.406970	1.7609	-22.9070	0.0000000	0.2971400	0.80268
-0.1198600	0.607330	0.041368	1.0688	-37.5950	-0.4479700	-0.1198600	0.64655

Attr9	Attr10	Attr11	Attr12	Attr13	Attr14	Attr15	Attr16
1.19610	0.46359	0.2194600	0.531390	0.142330	0.2194600	592.240	0.616300
1.60180	0.53962	0.1757900	0.383170	0.126470	0.1718500	829.460	0.440040
1.00770	0.67566	0.0045718	0.020219	0.030966	0.0045718	2094.100	0.174300
1.05090	0.56453	0.0245260	0.069747	0.036812	0.0245260	3299.400	0.110630
1.33930	0.58496	0.2388100	0.633170	0.187800	0.2335800	602.310	0.606000
1.81260	0.44385	0.2076600	0.341880	0.271630	0.1820600	412.300	0.885280
0.44339	0.96334	0.1435000	17.183000	0.325330	0.1421500	92.759	3.934900
1.18250	0.32934	0.0557170	0.023711	0.047491	0.0134220	4359.100	0.083733
2.33170	0.44527	0.3306100	0.555530	0.131980	0.2971400	657.950	0.554750
3.00000	0.39267	-0.1198600	-0.199300	0.011424	-0.1198600	6467.900	0.056432

Attr17	Attr18	Attr19	Attr20	Attr21	Attr22	Attr23	Attr24
2.4213	0.2194600	0.1227200	37.573	0.99690	0.295100	0.0974020	0.75641
2.1721	0.1718500	0.1072800	60.954	5.08890	0.175710	0.0912950	NA
4.4225	0.0045718	0.0035921	53.881	0.67451	0.040610	0.0004677	0.23470
2.3129	0.0245260	0.0188760	86.317	0.62795	0.055446	0.0188760	0.56944
2.4094	0.2335800	0.1744100	140.860	1.20390	0.234930	0.1405900	0.00000
1.7981	0.1820600	0.1004500	43.168	1.27240	0.000000	0.1004500	NA
27.2790	0.1421500	0.3206100	19.801	0.82872	0.082265	0.2594300	10.57300
1.4911	0.0134220	0.0113510	44.614	1.01300	0.046120	0.0083151	0.00000
1.8027	0.2971400	0.1274300	81.972	1.04670	0.330150	0.1024800	NA
1.6466	-0.1198600	-0.0399530	28.419	0.85800	0.000000	-0.0399530	NA

Attr25	Attr26	Attr27	Attr28	Attr29	Attr30	Attr31	Attr32
0.46359	0.506690	1.97370	0.32417	5.9473	0.224930	0.1227200	100.8200
0.17523	0.384420	44.59300	1.04860	4.0792	0.243840	0.1097400	105.0900
0.67566	0.156720	0.32153	1.71070	4.6220	0.036196	0.0035921	65.3450
0.56453	0.110630	0.44844	0.73869	4.2600	0.286240	0.0188760	103.8100
0.57250	0.496870	44.94700	1.18530	4.6033	0.306910	0.1778400	122.0900
0.41600	0.885280	0.00000	2.21180	3.8104	0.299260	0.2287700	241.4100
0.96082	3.195000	61.21000	13.52600	4.3390	-0.055356	0.3222200	7.2995
0.31050	0.078381	1.09040	0.45384	5.9673	0.567170	0.0113510	217.8100
0.43913	0.449850	9.86330	6.99830	3.9110	0.202110	0.1383600	97.4000
-0.52078	0.056432	NA	0.11581	3.3322	0.193860	-0.0799060	99.4560

Attr33	Attr34	Attr35	Attr36	Attr37	Attr38	Attr39	Attr40
3.6203	0.71453	0.2951000	1.80790	1.2314e+05	0.46359	0.1650100	0.212820
3.4733	3.38360	0.0440760	1.60180	NA	0.53962	0.0275160	0.164060
5.5857	0.17960	0.0406100	1.34250	NA	0.67566	0.0319070	0.844690
3.5161	0.12824	0.0554460	1.30680	3.9624e+00	0.64524	0.0426730	0.178260
2.9897	2.65740	0.2363500	1.33930	4.5490e+00	0.62769	0.1764800	0.013769
1.5120	1.44780	0.0295380	1.81260	4.9017e+01	0.45691	0.0162960	0.185600
50.0030	11.28400	0.0297240	0.44339	5.8907e+01	0.97875	0.0670380	108.440000
1.7455	1.47330	0.2338900	1.18250	1.1749e+01	0.37675	0.1978000	0.462850
3.7474	3.61330	0.3273200	2.33170	2.1065e+01	0.46512	0.1403800	0.175070
3.6700	3.63420	0.0069493	3.00000	NA	0.39267	0.0023164	0.079719

Attr41	Attr42	Attr43	Attr44	Attr45	Attr46	Attr47	Attr48
0.041124	0.165010	95.682	58.1090	0.9462100	0.90221	44.941	0.2600300
0.074333	0.109690	149.750	88.8010	0.5466900	1.03300	62.678	0.1449700
0.098528	0.031907	150.130	96.2510	0.0031684	2.32900	54.296	0.0057691
0.180500	0.042673	158.550	72.2370	0.0798190	0.90954	90.707	0.0321410
0.054712	0.175420	192.450	51.5850	0.3642900	0.52685	171.050	0.2169900
0.059746	0.000000	152.160	108.9900	0.8493000	1.20190	97.177	-0.3102800
0.014485	0.185540	28.536	8.7355	4.7821000	109.72000	21.224	0.0801690
0.251590	0.039003	135.670	91.0580	0.0680290	0.98398	53.392	0.0033856
0.054265	0.141590	132.770	50.8020	0.4563000	0.78182	95.359	0.3195500
0.131340	0.000000	72.372	43.9530	-0.5131500	0.68040	38.627	-0.1541300

Attr49	Attr50	Attr51	Attr52	Attr53	Attr54	Attr55	Attr56
0.1454000	1.3480	0.4129900	0.276220	1.0457	1.0458	1.2728e+05	0.1639600
0.0905030	1.5874	0.4484900	0.287910	2.0044	2.0044	3.3878e+03	0.0275160
0.0045328	3.1599	0.2261200	0.179030	2.3667	2.3667	2.0453e+04	0.0076387
0.0247370	1.4504	0.3516400	0.284400	1.5139	1.7303	5.0126e+03	0.0483980
0.1620300	1.7136	0.3689100	0.334490	2.0256	2.1735	1.3730e+04	0.1764800
-0.1711900	1.5364	0.5325400	0.661400	3.0496	3.1393	2.0806e+03	0.5557700
0.1808100	25.4160	0.0082728	0.019999	14.1100	14.3360	2.0158e+04	0.0670380
0.0028632	1.0460	0.5660600	0.572910	1.1034	1.2622	1.2393e+00	0.1978000
0.1370400	1.6978	0.5348800	0.266850	7.6569	7.9983	3.3152e+03	0.1403800
-0.0513770	1.0584	0.6014200	0.272480	1.0993	1.0993	8.8899e+01	0.2642800

Attr57	Attr58	Attr59	Attr60	Attr61	Attr62	Attr63	Attr64	class
0.375740	0.83604	0.0000065	9.7145	6.2813	84.2910	4.3303	4.0341	0
0.271000	0.90108	0.0000000	5.9882	4.1103	102.1900	3.5716	5.9500	0
0.000881	0.99236	0.0000000	6.7742	3.7922	64.8460	5.6287	4.4581	0
0.043445	0.95160	0.1429800	4.2286	5.0528	98.7830	3.6950	3.4844	0
0.321880	0.82635	0.0730390	2.5912	7.0756	100.5400	3.6303	4.6375	0
0.410190	0.46957	0.0294210	8.4553	3.3488	107.2400	3.4036	12.4540	0
0.119400	0.75777	0.0159950	18.4330	41.7830	6.8102	53.5960	6.4942	0
0.029854	0.83478	0.1439400	8.1813	4.0084	174.7300	2.0889	3.9616	0
0.536630	0.87292	0.0445840	4.4527	7.1847	83.7270	4.3594	40.0970	0
-0.305250	0.73749	0.0000000	12.8440	8.3043	73.1720	4.9883	8.3984	0

Methods

In this project, various methods which based on statistical hypothesis testing, statistical modeling and statistical learning techniques will be explored. We will start with logistic regression since the dependent variable is binary with only two classical levels. Then, we will apply some of the widely used classification models: decision trees, random forest, K-Nearest Neighbors and Bagging. By applying K-Fold Cross Validation, we will be able to analyze with imputed and resampled datasets. Finally, we will evaluate and rank the performance of models we used on the validation datasets by accuracy (error) rates (with confusion table). We will be able to find the model with best performance on predicting future bankruptcy condition for a firm by the end of the project. Other analysis methods which we have learned from class (e.g. Ridge Regression, Lasso Regression and Hierarchy Clustering) may be also applied in this project.

Challenges

We have discovered some challenges. The first challenge is value Missing: by looking through the data, we found missing values for some observations, indicating data imputation techniques like Expectation-Maximization or KNN are needed. The second challenge is data imbalance: the summary of dependent variable suggests among the 10503 observations, only 459 of them reported bankruptcy and the rest 10008 firms did not bankrupt in the future 3 years. This ratio indicates a need to oversample the minority categorical class. The next challenge is risk of overfitting: the data contains more than 60 predictor variables, indicates there is a great risk/potential of overfitting models and tuning parameter is needed in decision tree. Finally, since the data set contains 10503 observations in total, there is a challenge of extensive data visualization and the computational burden may increase.

Appendix

Code Chunks

If code chunks are used within the document, this information can be dynamically retrieved and embedded.

```
knitr::opts_chunk$set(echo = TRUE)
# install and load packages
pkg_list = c('ggplot2', 'tidyr', 'stringr', 'dplyr', 'foreign', 'knitr')
to_install_pkgs = pkg_list[!(pkg_list %in% installed.packages()[,"Package"])]
if(length(to_install_pkgs)) {
  install.packages(to_install_pkgs, repos = "https://cloud.r-project.org")
}
```

```
sapply(pkg_list, require, character.only = TRUE)

# Sets default chunk options
knitr::opts_chunk$set(
  fig.align = "center",
  echo = FALSE,
  message = FALSE,
  warning = FALSE
)
# load in data
year3 = foreign::read.arff('3year.arff')

# top 10 observations
knitr::kable(head(year3,10)[, 1:8], format = 'latex')
knitr::kable(head(year3,10)[, 9:16], format = 'latex')
knitr::kable(head(year3,10)[, 17:24], format = 'latex')
knitr::kable(head(year3,10)[, 25:32], format = 'latex')
knitr::kable(head(year3,10)[, 33:40], format = 'latex')
knitr::kable(head(year3,10)[, 41:48], format = 'latex')
knitr::kable(head(year3,10)[, 49:56], format = 'latex')
knitr::kable(head(year3,10)[, 57:65], format = 'latex')
```

Reference

[1] Sudheer Chava and Robert A. Jarrow, Bankruptcy Prediction with Industry Effects, Review of Finance 8: 537-569, 2004, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.4409&rep=rep1&type=pdf

[2] Risck L. Wilson, Ramesh Sharda, Bankruptcy Prediction Using Neural Networks, ScienceDirect, Volume 11, Issue 5, June 1994, Pages 545-557, https://doi.org/10.1016/0167-9236(94)90024-8

[3] Sai Surya Teja Maddikonda and Sree Keerthi Matta, Bankruptcy Prediction: Mining the Polish Bankruptcy Data, https://github.com/smaddikonda/Bankruptcy-Prediction/blob/master/Bankruptcy%20Prediction%20Report.pdf

[4] Polish Companies Bankruptcy Data Set, UCI Machine Learning Repository,http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data