Bach Quang Minh

Le Ngoc Thach

Nguyen Nhat Quang

Jon Von Neumann Institute – ICT2017

HACKATHON II - 2018

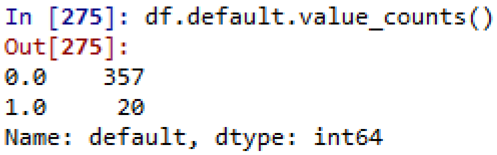
Problem statement

Measure the default level of a company based on its financial profile.

Data exploration

The whole data set contains 12 features and 2 classes: ‘default’-1 and ‘not-default’-0. We identify that this is a binary classification problem.

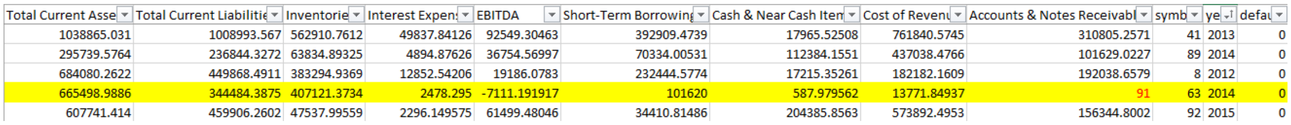
Here is the distribution of the data class:



```
In [275]: df.default.value_counts()
Out[275]:
0.0    357
1.0     20
Name: default, dtype: int64
```

As seen, the data suffered from imbalanced-class issue. The class percentage of 1 and 0 are 5.3% and 94.7% respectively.

We also notice that there is one company was having an unusual ‘Account & Notes receivables’ value:



Total Current Asse	Total Current Liabilitie	Inventorie	Interest Expen:	EBITDA	Short-Term Borrowing	Cash & Near Cash Iten	Cost of Revenu	Accounts & Notes Receivabl	symb	ye	defau
1038865.031	1008993.567	562910.7612	49837.84126	92549.30463	392909.4739	17965.52508	761840.5745	310805.2571	41	2013	0
295739.5764	236844.3272	63834.89325	4894.87626	36754.56997	70334.00531	112384.1551	437038.4766	101629.0227	89	2014	0
684080.2622	449868.4911	383294.9369	12852.54206	19186.0783	232444.5774	17215.35261	182182.1609	192038.6579	8	2012	0
665498.9886	344484.3875	407121.3734	2478.295	-7111.191917	101620	587.979562	13771.84937	91	63	2014	0
607741.414	459906.2602	47537.99559	2296.149575	61499.48046	34410.81486	204385.8563	573892.4953	156344.8002	92	2015	0

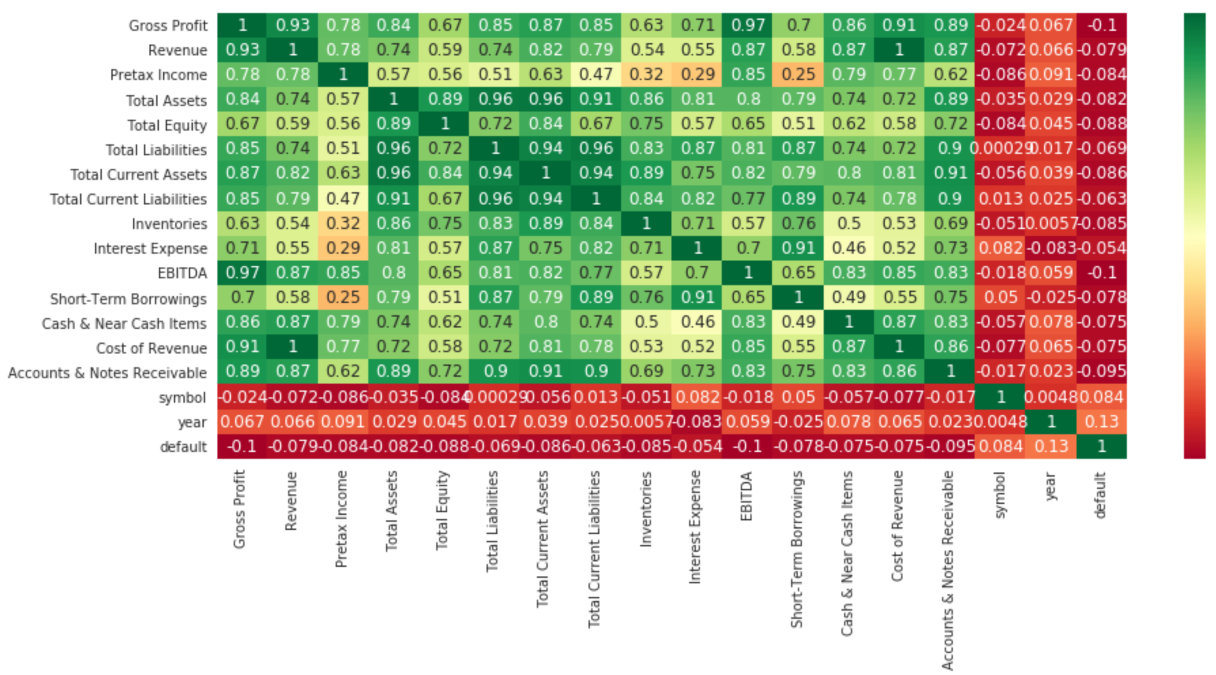
Since, this is the company data, we decided to drop this row before further analysis.

Proposed method

Since most columns from the data are real, continuous values. We decided to use Logistic Regression – a classification algorithm which could deal with continuous data.

The measure that we use for evaluation is the True Positive rate.

Feature selection



	Gross Profit	Revenue	Pretax Income	Total Assets	Total Equity	Total Liabilities	Total Current Assets	Total Current Liabilities	Inventories	Interest Expense	EBITDA	Short-Term Borrowings	Cash & Near Cash Items	Cost of Revenue	Accounts & Notes Receivable	symbol	year	default
Gross Profit	1	0.93	0.78	0.84	0.67	0.85	0.87	0.85	0.63	0.71	0.97	0.7	0.86	0.91	0.89	-0.024	0.067	-0.1
Revenue	0.93	1	0.78	0.74	0.59	0.74	0.82	0.79	0.54	0.55	0.87	0.58	0.87	1	0.87	-0.072	0.066	-0.079
Pretax Income	0.78	0.78	1	0.57	0.56	0.51	0.63	0.47	0.32	0.29	0.85	0.25	0.79	0.77	0.62	-0.086	0.091	-0.084
Total Assets	0.84	0.74	0.57	1	0.89	0.96	0.96	0.91	0.86	0.81	0.8	0.79	0.74	0.72	0.89	-0.035	0.029	-0.082
Total Equity	0.67	0.59	0.56	0.89	1	0.72	0.84	0.67	0.75	0.57	0.65	0.51	0.62	0.58	0.72	-0.084	0.045	-0.088
Total Liabilities	0.85	0.74	0.51	0.96	0.72	1	0.94	0.96	0.83	0.87	0.81	0.87	0.74	0.72	0.9	0.00029	0.017	-0.069
Total Current Assets	0.87	0.82	0.63	0.96	0.84	0.94	1	0.94	0.89	0.75	0.82	0.79	0.8	0.81	0.91	-0.056	0.039	-0.086
Total Current Liabilities	0.85	0.79	0.47	0.91	0.67	0.96	0.94	1	0.84	0.82	0.77	0.89	0.74	0.78	0.9	0.013	0.025	-0.063
Inventories	0.63	0.54	0.32	0.86	0.75	0.83	0.89	0.84	1	0.71	0.57	0.76	0.5	0.53	0.69	-0.051	0.0057	-0.085
Interest Expense	0.71	0.55	0.29	0.81	0.57	0.87	0.75	0.82	0.71	1	0.7	0.91	0.46	0.52	0.73	0.082	-0.083	-0.054
EBITDA	0.97	0.87	0.85	0.8	0.65	0.81	0.82	0.77	0.57	0.7	1	0.65	0.83	0.85	0.83	-0.018	0.059	-0.1
Short-Term Borrowings	0.7	0.58	0.25	0.79	0.51	0.87	0.79	0.89	0.76	0.91	0.65	1	0.49	0.55	0.75	0.05	-0.025	-0.078
Cash & Near Cash Items	0.86	0.87	0.79	0.74	0.62	0.74	0.8	0.74	0.5	0.46	0.83	0.49	1	0.87	0.83	-0.057	0.078	-0.075
Cost of Revenue	0.91	1	0.77	0.72	0.58	0.72	0.81	0.78	0.53	0.52	0.85	0.55	0.87	1	0.86	-0.077	0.065	-0.075
Accounts & Notes Receivable	0.89	0.87	0.62	0.89	0.72	0.9	0.91	0.9	0.69	0.73	0.83	0.75	0.83	0.86	1	-0.017	0.023	-0.095
symbol	-0.024	-0.072	-0.086	-0.035	-0.084	0.00029	0.056	0.013	-0.051	0.082	-0.018	0.05	-0.057	-0.077	-0.017	1	0.0048	0.084
year	0.067	0.066	0.091	0.029	0.045	0.017	0.039	0.025	0.0057	-0.083	0.059	-0.025	0.078	0.065	0.023	0.0048	1	0.13
default	-0.1	-0.079	-0.084	-0.082	-0.088	-0.069	-0.086	-0.063	-0.085	-0.054	-0.1	-0.078	-0.075	-0.075	-0.095	0.084	0.13	1

After studying carefully the correlation matrix and combining with our understanding about these finance indicators, we decided to omit the following features:

1. Total equity (because this is just an index calculated from two others indices)
2. EBITDA
3. Cost of Revenue (= revenue – gross profit)
4. Symbol (Just an encoded for each company)
5. Year (How well a company perform in a year is reflected by other index)

We will keep the following features for our first attempt:

'Gross Profit'

'Revenue'

'Pretax Income'

'Total Assets'

'Total Liabilities'

'Total Current Assets'

'Total Current Liabilities','Inventories'

'Interest Expense'

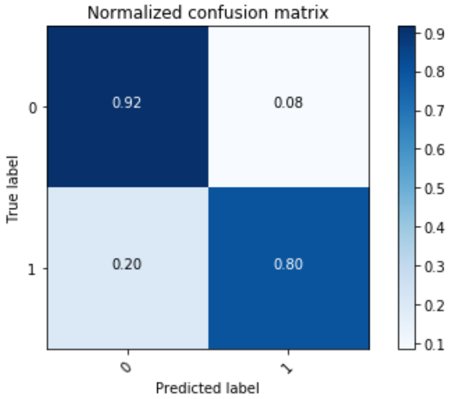
'Short-Term Borrowings'

'Cash & Near Cash Items'

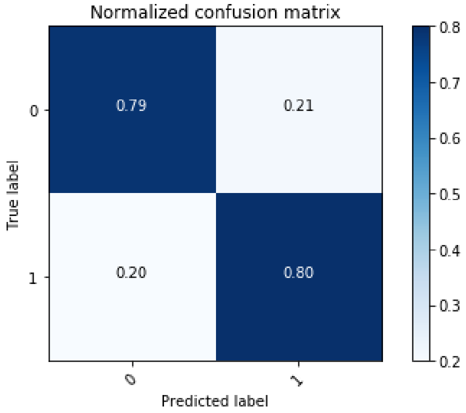
We use the train set and split it into train set and validation set with the ratio of 5:1

Here is the result:

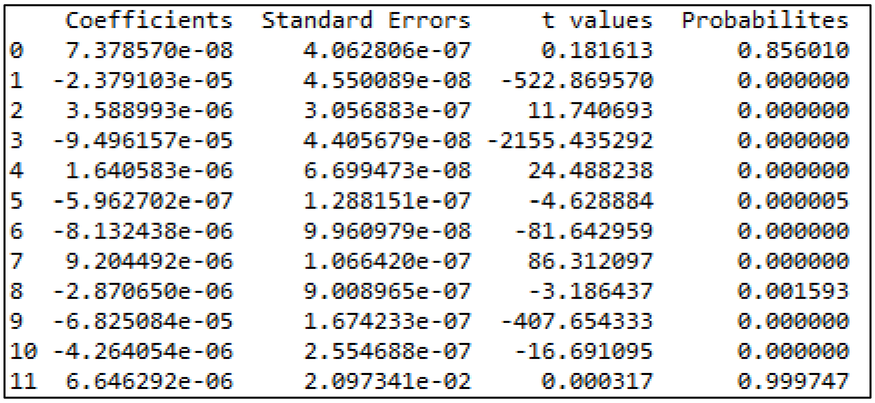
Confusion matrix on the train set



Confusion matrix on the test set



To know if our features are statistically significant, we investigate these statistics:



```
    Coefficients  Standard Errors      t values  Probabilites
0    7.378570e-08     4.062806e-07      0.181613      0.856010
1   -2.379103e-05     4.550089e-08   -522.869570      0.000000
2    3.588993e-06     3.056883e-07     11.740693      0.000000
3   -9.496157e-05     4.405679e-08  -2155.435292      0.000000
4    1.640583e-06     6.699473e-08     24.488238      0.000000
5   -5.962702e-07     1.288151e-07     -4.628884      0.000005
6   -8.132438e-06     9.960979e-08    -81.642959      0.000000
7    9.204492e-06     1.066420e-07     86.312097      0.000000
8   -2.870650e-06     9.008965e-07     -3.186437      0.001593
9   -6.825084e-05     1.674233e-07   -407.654333      0.000000
10  -4.264054e-06     2.554688e-07    -16.691095      0.000000
11   6.646292e-06     2.097341e-02      0.000317      0.999747
```

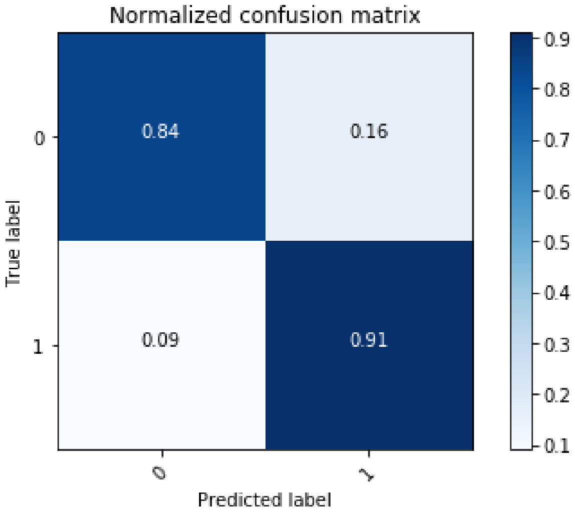
The result shows that some features does not have small p-values. We decided to seek for further feature selections.

First, we use the train-validation ratio as 0.3 and construct **a for loop** for **the number of features** to select the model with **the highest true positive values.**

The loop gives us these features:

['Gross Profit', 'Revenue', 'Pretax Income', 'Total Assets', 'Total Current Assets']

This is the confusion matrix for the validation set



The result obtained is higher and the features are suitable to our understanding.

We decided to use these features to train the model and produce the result for the test set.