Bankruptcy Companies Prediction

Supervised Machine Learning Methods

Yu Tian - Data Science first year; Chia-Yi Liaw - Data Science first year
Shuo Yang - Computer Science second years; Qiankun Huang - Applied Mathematics second year

1. Introduction

1.1 Background and Dataset

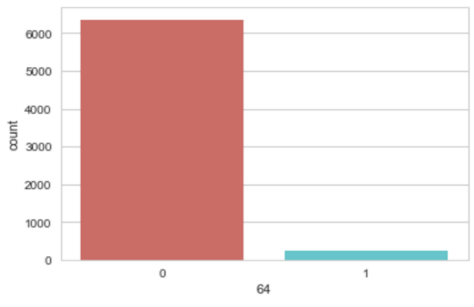
Bankruptcy is a process for business or individual to declare the fact of unable paying outstanding debts. With the financial difficulties, companies choose to file bankruptcy would influence the local community, industry, investor or even global economy. Bankruptcy has been critical matters to the economics, which lead to financial short and long-term distress. It is essential to be able to foresee bankruptcy of firms from financial performance and accounting research, especially with the prediction, it helps on economic decision making significantly. As a result, the prediction of bankruptcy is a practical and demanding problem to be discussed.

For this research, using financial statement to correlate the bankruptcy of Polish companies from 2007. In terms of machine learning technique, classification is one of the most efficient way to evaluate the data. In addition, main questions to be answered from the data analysis are 1) Given business and operational data, predicts if the company is going bankruptcy 2) Find out the most influential factor among attribute so it could provide better advice to companies on financial operations. Furthermore, by exploring four types of model, SVM, KNN, logistic regression and random forest, achieve the in-depth data analysis, classification and prediction on company bankruptcy.

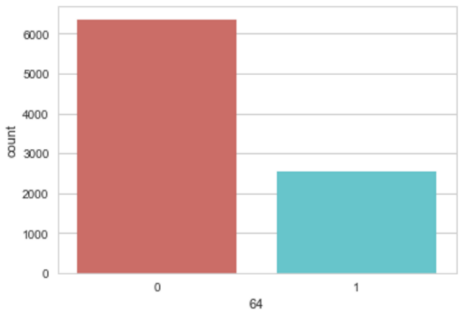
1.2 Data Exploration and Preprocessing

Unbalanced Distribution

For the model training, we are using the first-year data. There are two classes in the dataset: bankrupted companies and non-bankrupted company. There are 271 of companies which become bankruptcy while 6791 companies were not heading bankruptcy. The following histogram shows the distribution of two classes. From the graph, we find out that the data is highly unbalanced distributed. Because our project goal is to predict whether a company is going to bankrupt or not. The bankrupt companies' data is what we are really concerned. As a result, we use the over sampling method by replicating bankrupt company data by 10 times. After the over sampling procedure, the un-bankruptcy to bankruptcy companies' ratio become less skewed.



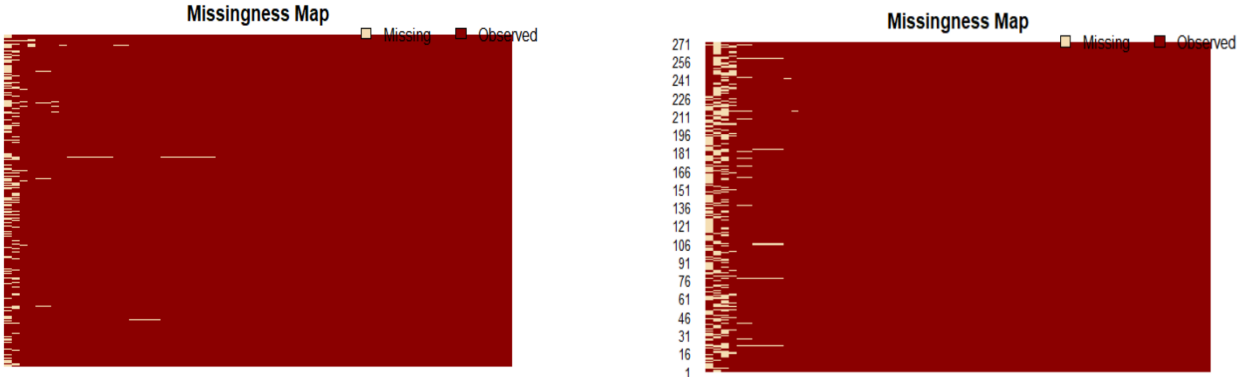
Distribution of Original Data



Distribution of Oversampling Data

Missing Value

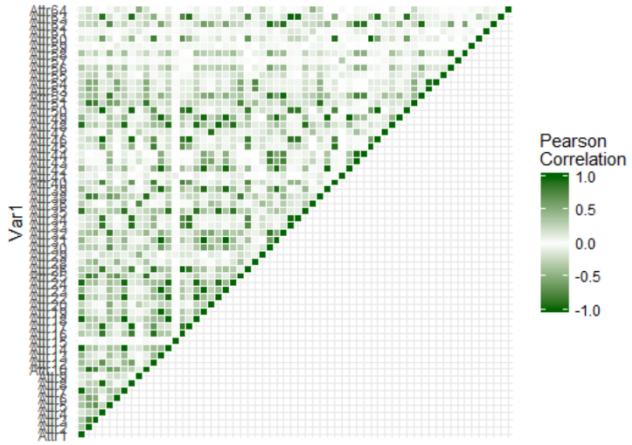
As the respect of missing data, we used missing map to find relationship between missing data and features. From the missing map, we find out that missing value is concentrated in serval features. For the bankruptcy data, there are four features containing more than 10% missing which are attribute 27, 21, 37, 11. As for the non-bankrupted companies, the top four predictors with most missing value are 21, 27, 37, 24. Because we have enough observation for non-bankrupted companies. As a result, delete features with most missing value based on the bankruptcy data. And after that, we delete the observations with missing value.



The left graph represents missing value based on each predictor in non-bankrupted companies, and the right graph represents the missing value based on each predictor in bankruptcy companies. As we can find in the graph, the predictors with high amount of missing value based on non-bankrupted companies and bankruptcy companies are highly overlapped.

Feature Correlation

By calculating the r-square to determine the linear correlation between features. By analysis correlation feature, we can get rid of highly correlated when we implement the random forest method and logistic regression for the sake of improving calculation efficiency. From the graph, easily tells the fact that there are many features are highly linearly correlated.



Heat map showing correlation between predictors. The darker green indicates the higher correlation.

2. Methodology and Result

2.1 KNN

K Value Selection

One major part of the KNN model is to find the K value. In general, a larger k suppresses the effects of noise, but makes the classification boundaries less distinct. In this project, it's unbalanced data, if k is too high, all prediction would be 0. The higher k is, the smoother the model would be. The boundary is smooth in KNN, our problem is not this case. On the other side, if k is too small, it will be easily affect by noise and be under fitting. In this project, 10-folds cross validation were used to pick the best K value and kernel function for the KNN model. After testing K ranges from 2 to 20, we found that when k = 5 we reach the highest accuracy.

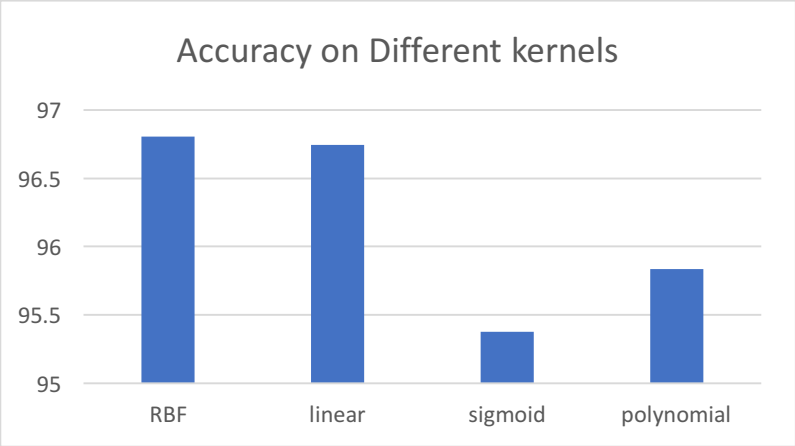
Distance Metrics

Euclidean distance $d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$, and Manhattan distance $d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k-y_k|}$ were compared to get a better model. After comparison, we got the best K=5, with Euclidean distance.
In our project, the number of 0 is much more than the number of 1, The k nearest neighbor would be more possible to be 0, so the prediction is more possible to be 0 and recall is higher but precision is lower.

2.2 SVM

Kernel Selection and Evaluation

The vector support machine (SVM) can be used to solve linear or Non-linear problems. The boundary can be a line, a circle or high order's graph. It should perform well in high-order problems, but it's hard to predict the actual boundary shape of a data with more than 60 features.
In this project, 10-folds cross validation were used to evaluate the kernel function for the SVM model. The accuracy of Radial Basis Function, in which we were using, on year 1 data is around 96.8%, that performs better than other kernel functions.



Cost and Regularization Parameters

To determine the parameter of c and σ2, package (e1071) with the tune function were used. The parameter γ=0.1 , which γ=1/2σ2 in $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$, and cost parameter c was set at 1.0 in regarding to the cost function of

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2 .$$

2.3 Logistic Regression

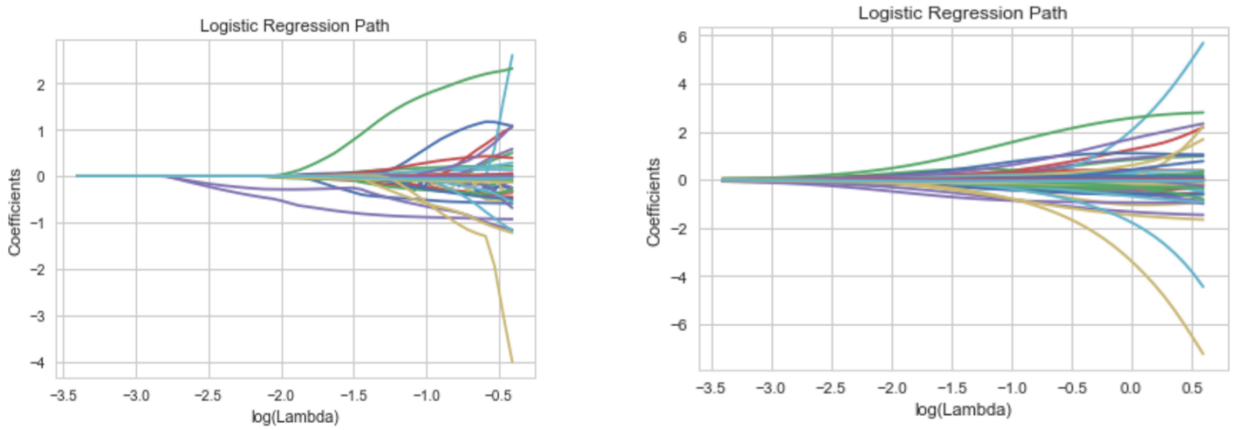
Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous. Generally, the logistic regression is using the sigma function $\sigma = \frac{1}{1+e^{-t}}$ to simulate events probability. By default, if the estimation value is larger than 0.5, we classify the event to class 1 and if is smaller than 0.5, we classify the event to class 0.

Implementation and Analysis of Logistic Regression

By using logistic regression, we are going to fit the classification model as well as try to pick up the most influential predictors for predicting classes for new observations. We used Lasso and Ridge logistic regression for the sake of regularization. Following graphs show the regularization path from logistic regression with L1 and L2 penalty. As we mentioned above, some of the predictor holds high linear relationship, so the result of lasso regularization is more interpretable and performs better on feature selection compared with ridge regularization.

According to the regression path graph, we chose λ=0.22 using L1 penalty. After the implementing the LASSO regression, 22 out of 60 features are chosen. The top 10 influential factors are shown in the flowing table. We will discuss the results at the end part of the report.



The left graph shows the regression path of logistic regression with penalty L1,

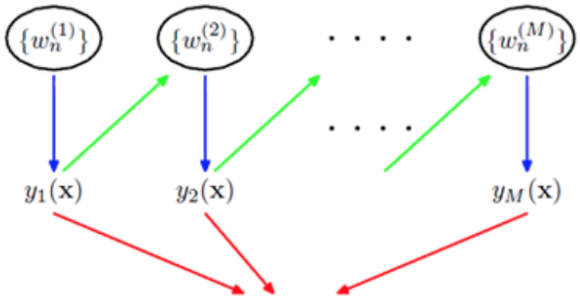
while right one shows the regression path of logistic regression with penalty L2.

According to the regression path graph, we chose λ=0.22 using L1 penalty. After the implementing the LASSO regression, 22 out of 60 features are chosen. The top 10 influential factors are shown in the flowing table. We will discuss the results at the end part of the report.

-5.05	gross profit (in 3 years) / total assets
3.38	EBITDA (profit on operating activities - depreciation) / total assets
0.72	net profit / total assets
-0.47	(current assets - inventory) / short-term liabilities
-0.39	sales / short-term liabilities
0.37	operating expenses / short-term liabilities
0.37	current assets / total liabilities
0.26	(current assets - inventory - receivables) / short-term liabilities
0.21	total sales / total assets
-0.15	sales / total assets

Top 10 influential predictor along with the coefficient given by LASSO logistic regression.

2.4 Random Forest



Ensemble learning including Boosting method and Bagging method. In ensemble learning. We train multiple learners to solve the same problem. Ensembles can combine multiple hypotheses to form a better hypothesis.

Random Forest use the bagging idea, it resamples the data and features which will decrease the variance in the model. This method is helpful for unbalanced data and solve the overfitting problem in complete decision tree because in random forest algorithm, we do not need do pruning for each "weak" decision tree We use out of bag error to find the most important feature in random forest but several features have same important number. It's hard to say which feature is the most important.

Importance(i) = performance(RF)- performance (RF(random(i)) = EOOB(RF)- EOOB(RF(random(i))

3. Result Analysis

3.1 Model Evaluation

In this specific project, as we are using classification algorithms, confusion matrix seems to be a good evaluation tool for us to evaluate our model. General we can get recall, precision and accuracy score from the confusion matrix.

Consider a Risk Manager wish to evaluate the risk the company faces bankruptcy. He would like to have a model with higher Recall value rather than the Precision. Because he wants to make sure the company to be invested are not going to bankrupt based on given operation data. Meanwhile, the precision cannot below a threshold to make sure the he will not take every company as a company which is going to bankrupt.

In another case for shareholder of a company who wishes to initiate a certain bankruptcy procedure for his partner's company, the decision has to be made with good precision rather than a good recall. Or the decision will result in loss of money and break the relationship in advance. Thus, we may need a model that is more conservative, which means a higher precision.

Depends on the different condition, the importance of precision and recall may varies. But in most case, we would prefer the first case, which we consider recall as more important while precision remains higher than a certain threshold.

3.2 Model Result Comparison

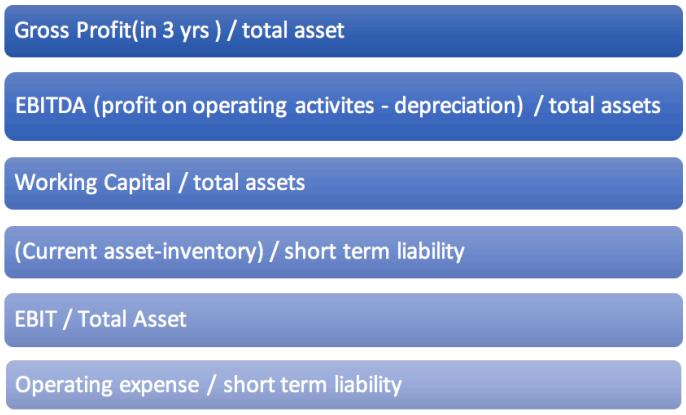
For the evaluation, we calculated precision, recall, and accuracy for each model we used above. As we can find from the following table, the Random Forest method perform best in according to the recall. Logistic regression performs relative well in term of recall but not very accurate when it comes to the precision and accuracy.

	Recall	Precision	Accuracy
SVM	0.846666666667	0.998270712152	0.847757671125
KNN	0.89203976688	0.885238170099	0.877655389457
Logistic Regression	0.911462450592	0.682203019909	0.65814416965
Random Forest	0.999051383399	0.993397264581	0.993705743509

3.3 Interpretation for Influential Predictors

From the Lasso logistic regression, we find the most relevant features indicating bankruptcy. First three features are indicating the asset turnover, which means how efficient the company is using their asset (cash, investment etc.). Last three features are indicating the company's liquidity, suppose a company can't cover the short term (within one year) liability (account payable, interest expenses) using resources, it might get bankrupted. Base on accounting principle, Asset = liability + equity. The higher the asset may lead to higher the liability. When the ratio between assets and liability is low, there chance of bankruptcy is high, which means a company's profit is not enough to cover the liability.

In conclusion, we would recommend a company to be aware of the ratio of gross profit and liability as well as paying attention to the short-term debt, which may more likely to lead to bankruptcy.



Most influential predictor chosen based on result from L1 logistic regression.

4. Discussion and future work

KNN

The KNN distance measurements function may have more distance metrics like cosine distance, Minkowsky distance, chi-square, etc. Some of which may perform better than the one we are using.

Data Preprocessing

In this project, we only tried some simple over sampling methods by replicating bankrupt data. However, we can use more advanced data preprocessed method such resampling skewed data by the scale of companies or other information we can get inside the data.

SVM

In this Project, only 4 kernels were tested with given value. Actually, there're more different types of kernels that may potentially fit in this project. We can keep working on different kernel function that may perform better than RBF.

Logistic regression

As we can see from the result of logistic regression model, although the recall of the is relatively good, the precision and accuracy of overall prediction does not give a positive feedback. If we care about prediction bankruptcy companies as well as the non-bankrupted companies, the model still need to be improved. The reason that the logistic model does not perform well on accuracy and precision is may because the non-linear relationship between predictors and the probability of each class. For the future, we may try high order term as well as the interaction between different predictors.