CSE/ECE 343: Machine Learning Project Interim Report

Title: Company Bankruptcy Prediction

Ayushi Jain ayushi19031@gmail.com

Darsh Parikh darsh20560@iiitd.ac.in

Gurmehak Kaur gurmehak20298@iiitd.ac.in

Sumit Kumar sumit20249@iiitd.ac.in

1.Abstract

In the current age of startups, there is a huge increase in the number of companies coming up. But not all companies manage to stay successful over time and burn out pretty fast. Bankruptcy can be considered a curse for the organization and the investors. It is expressed as the inability of a company to pay its debts to its creditors. Effective bankruptcy prediction is crucial for companies to make appropriate business decisions.

It is important for investors, so that they can decide if investing in the company is profitable or not. With the high dimensional dataset with 96 features, we plan to use advanced machine learning models and neural networks to give accurate predictions and meaningful insights.

2. Introduction

Machine Learning Algorithms have improved so much in recent years especially for classifying which class a data point belongs to. Using these classification algorithms, we can classify a company and identify the factors which significantly affect the financial health of a particular company. Early warnings of bankruptcy help not only the investor but also public policy makers to take proactive steps to minimize the impact of bankruptcies. Bankruptcy prediction is also significant for Better allocation of resources, Input to policy makers, Corrective action for business managers, Identification of sector wide problems and Signal to Investors.

The goal of our project is to select and find the key features that helped to predict bankruptcy. Secondly, visualize interesting patterns present in the data.Since the target output is binary, we put forward the idea of binary classification models to classify a company as bankrupt or not. way to allocate a popularity score using regression models. We aim to find the suitable features that affect bankruptacy using different feature learning techniques; test and compare multiple machine learning models for our proposed task. This will help the organizations to make wise business decisions, invest wisely and reap profits.

3. Literature Survey

The School of Expertness and Valuation, Institute of Technology and Business in Czech Republic used Support Vector Machine and artificial neural network(multilayer perceptron artificial neural networks and radial basis function artificial neural networks) to create a model for predicting potential bankruptcy of companies.They lay it out that a very useful tool to predict the development of companies going to bankrupt is by using artificial neural networks but the disadvantages of ANNs include possible illogical behavior of networks and required high quality data.

Machine learning models and bankruptcy prediction: In this study, Flavio Barboza, Herbert Kimura and Edward Altman have tested machine learning models such as support vector machines, bagging, boosting and random forest to predict bankruptcy one year prior to the event, and compare their performance with results from discriminant analysis, logistic regression, and neural networks.They use data from 1985 to 2013 on North American firms.

Bankruptcy Prediction Using Machine Learning, by Nanxi Wang: In this paper, three relatively new methods have been proposed for predicting bankruptcy based on real-life data. The usage of the three models (support vector machine, neural network with dropout, autoencoder) in economics or finance is comparatively hard to find. So, the paper aims to find out if they work well in the economic field, by predicting company bankruptcy.

4. Dataset

4.1 Data Description

The dataset was collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. This dataset has around 96 parameters with mainly continuous features in 6819 rows. This would help us conduct a detailed analysis. It would also help us get realistic conclusions as we would almost be considering all the major and minor features.

Few feature descriptions are below:

●Operating Expense Rate: The operating expense rate shows the efficiency of a company's management by comparing the total operating expense (OPEX) of a company to net sales.

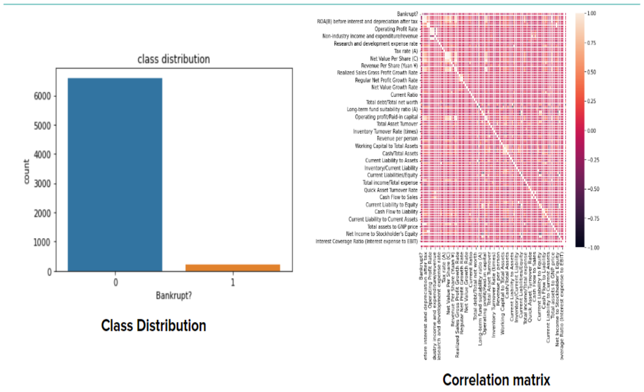
● Interest Bearing Debt Interest Rate: The interest-bearing debt ratio is significant because it gives a window into the financial health of a company. The interest-bearing debt ratio, or debt to equity ratio, is calculated by dividing the total long-term, interest-bearing debt of the company by the equity value.

● Tax Rate: A tax rate is the percentage at which an individual or corporation is taxed.

- Revenue per share: Amount of revenue over common shares outstanding. Answers the question, what's the ownership of sales to each share? Increasing revenue per share (RPS) over time is a good sign, because it means each share now has claim to more revenues.
- Total Asset Growth Rate: Total Asset Growth Rate defined as year-over-year percentage change in total assets

4.2 Exploratory Data Analysis

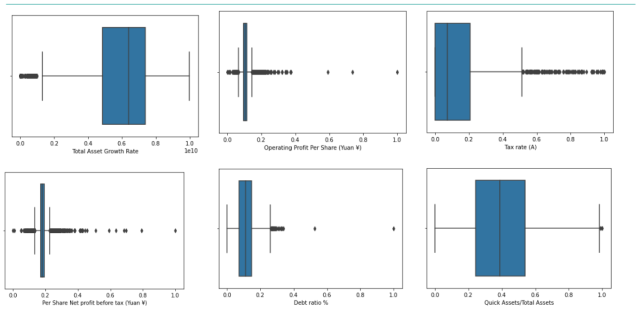
We have done EDA on the entire dataset to get a good idea of the nature of the dataset. We observed a huge imbalance between the majority and minority classes, i.e, there were significantly more rows with class=0 compared to rows with class=1. This is seen below in the plot.



Class Distribution

Correlation matrix

Box-plots of some features of raw data

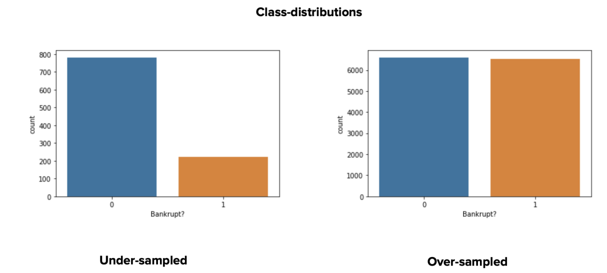


We also checked for any null values.

4.3 Preprocessing

As the first step of Data Preprocessing, we first normalized our training and test data-set so that each of 96 feature values of our data-set were in the range of 0 to 1. Our Exploratory Data Analysis revealed some of the issues we needed to take care of before fitting, namely, data imbalance and existence of correlated features. We have used different sampling methods to handle the imbalanced data. For oversampling, we used SMOTE and ADASYN techniques respectively. For undersampling the data, we used Tomek links and CNN(Condensed nearest neighbors). In case of oversampling, ADASYN worked better and in the case of undersampling, CNN worked better.

Re-sampling the Dataset

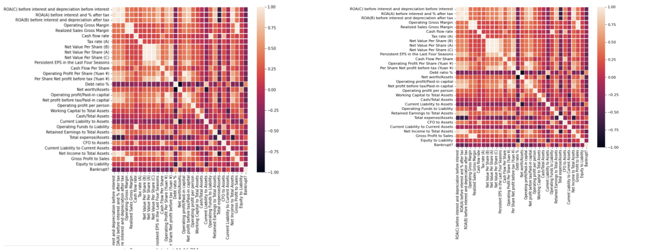


Under-sampled

Over-sampled

4.4 Data Visualisation

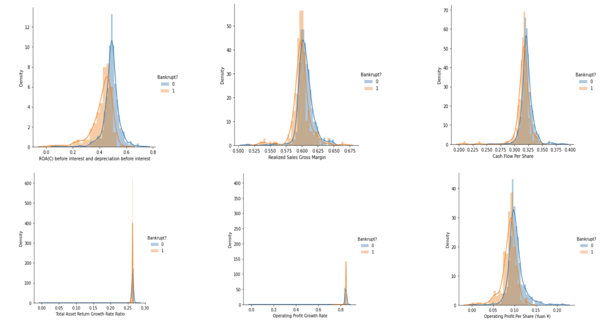
Correlation Heat-maps of selected features



Under-sampled

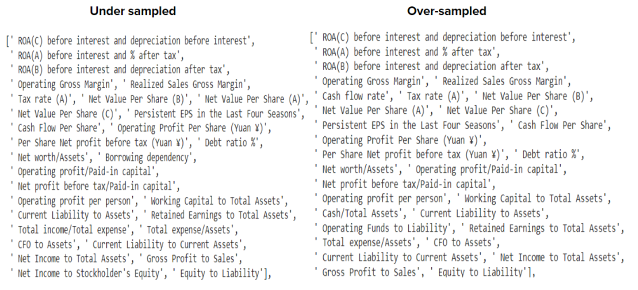
Over-sampled

Univariate plots of some features



4.5 Feature Selection

We performed feature selection on both over-sampled and undersampled data to select the best 30 features.



Under sampled

```
[' ROA(C) before interest and depreciation before interest',
 ' ROA(A) before interest and % after tax',
 ' ROA(B) before interest and depreciation after tax',
 ' Operating Gross Margin', ' Realized Sales Gross Margin',
 ' Tax rate (A)', ' Net Value Per Share (B)', ' Net Value Per Share (A)',
 ' Net Value Per Share (C)', ' Persistent EPS in the Last Four Seasons',
 ' Cash Flow Per Share', ' Operating Profit Per Share (Yuan ¥)',
 ' Per Share Net profit before tax (Yuan ¥)', ' Debt ratio %',
 ' Net worth/Assets', ' Borrowing dependency',
 ' Operating profit/Paid-in capital',
 ' Net profit before tax/Paid-in capital',
 ' Operating profit per person', ' Working Capital to Total Assets',
 ' Current Liability to Assets', ' Retained Earnings to Total Assets',
 ' Total income/Total expense', ' Total expense/Assets',
 ' CFO to Assets', ' Current Liability to Current Assets',
 ' Net Income to Total Assets', ' Gross Profit to Sales',
 ' Net Income to Stockholder's Equity', ' Equity to Liability'],
```

Over-sampled

```
[' ROA(C) before interest and depreciation before interest',
 ' ROA(A) before interest and % after tax',
 ' ROA(B) before interest and depreciation after tax',
 ' Operating Gross Margin', ' Realized Sales Gross Margin',
 ' Cash flow rate', ' Tax rate (A)', ' Net Value Per Share (B)',
 ' Net Value Per Share (A)', ' Net Value Per Share (C)',
 ' Persistent EPS in the Last Four Seasons', ' Cash Flow Per Share',
 ' Operating Profit Per Share (Yuan ¥)',
 ' Per Share Net profit before tax (Yuan ¥)', ' Debt ratio %',
 ' Net worth/Assets', ' Operating profit/Paid-in capital',
 ' Net profit before tax/Paid-in capital',
 ' Operating profit per person', ' Working Capital to Total Assets',
 ' Cash/Total Assets', ' Current Liability to Assets',
 ' Operating Funds to Liability', ' Retained Earnings to Total Assets',
 ' Total expense/Assets', ' CFO to Assets',
 ' Current Liability to Current Assets', ' Net Income to Total Assets',
 ' Gross Profit to Sales', ' Equity to Liability'],
```

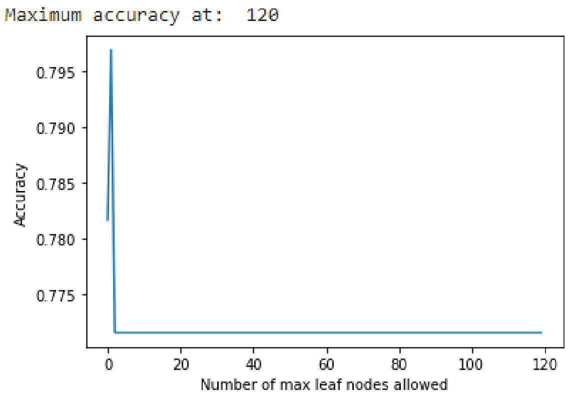
5. Methodology

5.1 Classification

Logistic Regression: It is a classification model that's simple to implement and delivers excellent results with linearly separable classes. The logistic regression model does not classify data; instead, it models the likelihood of output in terms of input. It can, however, be used to build a classifier by simple cutoff-based rules.

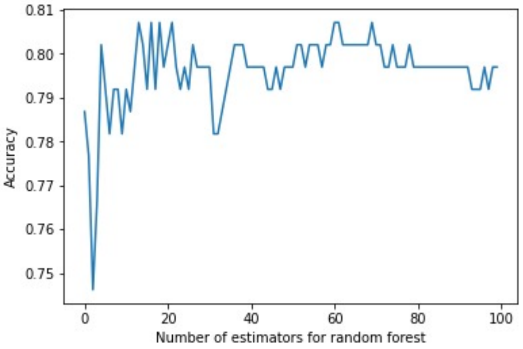
Naive Bayes: Naive Bayes is a classification method which combines Bayes Theorem of conditional Probability and the assumption that all the features are independent of each other. We have used Binomial and Gaussian Naive bayes.

Decision Tree:A decision tree uses a tree-like model of decisions and their possible consequences. We have tested the decision tree. We have explored how accuracy varies with the definition of maximum number of root leaves.



```
Maximum accuracy at:  120
```

Random Forest: A meta estimator that fits a number of decision tree classifiers on various sub samples of dataset and uses averaging to improve the predictive accuracy and control over fitting. The sub-sample size is always the same as the original input size but the samples are drawn by boosting.We have checked how accuracy of the model varies as we increase the number of estimators. We find that the accuracy becomes nearly constant approximately 80 percent as the number of estimators reach 100.



5.2 Dimensionality Reduction

We have used Principal Component Analysis, Factor Analysis, Factor Analysis with Varimax Rotation, Truncated SVD, LDA and Kernel PCA with linear, poly and rbf kernels.

We find that the accuracy with using just 10 components is just slightly lesser than for using 96 attributes for the oversampled and undersampled dataset in all techniques with random forests and decision trees, except for the matrix generated by rbf kernel PCA.

Principal component analysis aims to reduce the dimensions of the data by preserving maximum amount of variance. Factor Analysis tries to reduce interrelated components to one combined component. Varimax rotation changes the coordinates of the datapoint to the maximum sum of their variances. Truncated SVD works efficiently with sparse

matrices, decomposing original matrix to multiple interesting matrix components. LDA is used for separating different classes. Kernel PCA projects dataset into a higher dimensional feature space where it can be separated into the different classes linearly.

5.3 Clustering

Clustering is an unsupervised learning algorithm which divides data points into a number of groups such that data points of the same group are similar to each other and dissimilar to data points of other groups. For clustering, we have used the K-means clustering algorithm. In this, the data points are grouped in k clusters, in our case 2. Since it is a centroid based clustering, the algorithm takes unlabeled input, and assigns it to the centroids and repeats this process until the euclidean distance between the centroid and data point is not minimized, hence forming clusters.

5.4 Support Vector Machines

Specifically used for classification in Machine Learning, it uses kernels to project the data to higher dimensions and attempts to form a linear boundary between the data of two or more classes. Here, we have a data of 95 features which is already high dimensional and we observe that SVM with rbf kernels perform the best on this data.

We have applied Support Vector Machines on undersampled and oversampled data generated using SMOTE and ADASYN methods respectively, as well as on undersampled and oversampled data generated using simple random sampling.

We observe that SVC works most effectively with Oversampled dataset with Factor Analysis. It works better when the oversampling and the undersampling techniques are sophisticated as compared to randomness in the same. Also, given that the undersampled dataset gives exactly same results with No DR, PCA and FA as in the table above, we find that DR, PCA and FA applications have no effect on the undersampled data. The model requires more data to work effectively as we can see in the random oversampling and random undersampling part.

5.5 Neural Network

The MultiLayer Perceptron is a type of neural network that classifies datasets which are not linearly separable. They do this by using a more robust and complex architecture to learn regression and classification models for difficult datasets. In a MLP, the inputs are multiplied with weights and fed to the activation function and in backpropagation, they are modified to reduce the loss. We have run the MLP classifier on the undersampled and oversampled data generated using simple random sampling. For a deeper analysis, we also changed the activation functions.

5.6 Gradient Boosting

Gradient boosting is a popular boosting algorithm. In which each predictor corrects its predecessor's error.the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of the predecessor as labels.

6. Results and Analysis

We obtained varying results from different classifiers. MLP showed the best accuracy with relu activation function on oversampled data.The Gradient boosting classifier performed better than MLP with an accuracy of 95% on the oversampled data as we increased the tree depth, the model started to overfit. The accuracy of the K-means model reaches about 54-55% for both the undersampled and oversampled data and gives an accuracy of 69% for the whole dataset.

UNDERSAMPLING

Model	Accuracy	Precision	Recall
Logistic Regression	0.8407	0.7372	0.8407
Binomial NB	0.8208	0.4104	0.8208
Gaussian NB	0.8009	0.6690	0.8009

OVERSAMPLING

Model	Accuracy	Precision	Recall
Logistic Regression	0.8792	0.8798	0.8792
Binomial NB	0.4979	0.2489	0.4979
Gaussian NB	0.8624	0.8656	0.8624

Accuracies with SVM classifier

	No DR	PCA	FA	FA with rotation
Oversampled	0.82133333333333334	0.8514285714285714	0.9081904761904762	0.8441904761904762
Undersampled	0.8265306122448898	0.8265306122448898	0.8265306122448898	0.8214285714285714

Random Oversampling	0.651893939393939
Random Undersampling	0.1931818181818181

Accuracies for Multi layer perceptron

MLP with varying activation functions	Accuracy with random undersampled dataset	Accuracy with random oversampled dataset
Sigmoid	0.6022	0.5053
Linear	0.5568	0.5625

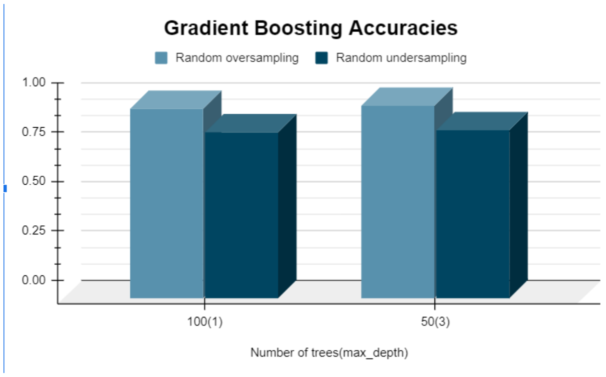
Relu	0.6136	0.9045
Tanh	0.6818	0.6159

K-Means Clustering Results

Data	Accuracy	Rand Index	Silhouette score
Undersampled	0.55	0.5038	0.2269
Oversampled	0.54	0.5034	0.2169
Whole	0.69	0.5717	0.2285

Gradient Boosting Classifier.

Sampling	Number of trees	max depth	Prediction score
Random undersampling	100	1	0.8409
	50	3	0.8522
Random oversampling	100	1	0.9598
	50	3	0.9776



7. Conclusion

7.1 Learnings from the Project

1. The importance of dimensionality reduction and feature selections - In case of dimensionality reduction we were able to reduce from 96 attributes to just 10 attributes with slightly decreased accuracies.

2. The data is imbalanced for Bankruptcy class with an imbalance factor of nearly 30:1,Resampled the data to reduce the imbalance factor.

3. The imbalance factor after random oversampling is 1:1 , Imbalance factor after random under-sampling is 1:1.

4. The results gained by the Gradient Boosting classifier were significantly better than the results gained by MLP classifier and KMeans clustering. Furthermore, we should explore other methods like synthetic feature generation for better performance and try to propose a novel solution.

5. We noticed that the supervised learning algorithms are performing comparatively better than the unsupervised learning algorithms on this dataset.

7.2 Discussion

We were able to successfully illustrate how machine learning can be exploited in the field of economics. We were able to follow the proposed timeline tentatively. As mentioned in our timeline, we trained several classification models. The classification models predict whether a company is likely to go bankrupt or not. Interesting Future Research subtopics can be configured by incorporating new types of data. For example, unstructured data for example the text of the annual reports of companies. Using a neural network after adding such complexities to the dataset would be a very wise option. We could also extend this problem to giving companies personalized feedback that would be even more helpful in preventing bankruptcy. If time permits, we would love to explore more about this topic in the near future.

7.3 Member Contributions

All the team members plan to contribute equally to the project. Few details about the contribution:

• Ayushi Jain: Application of models of Dimensionality Reduction Techniques, Random Forests, Support Vector Machines.

• Darsh Parikh: Data Preprocessing, Analysis, and inference of the preprocessed data. Feature selection, K-means Clustering

• Gurmehak Kaur: Literature review, Data visualization, Logistic regression, Naive Bayes, Multilayer perceptron, discussions.

• Sumit Kumar: Data Extraction and Collection Analysis and inference of the Raw data. Data Preprocessing(Under and Over sampling), Gradient Boosting, and Model analysis.

8. References

- A study on bankruptcy prediction, by the Institute of Business and Technology in Czech Republic.
- A study on bankruptcy prediction using ML models by Flavio Barboza, Herbert Kimura and Edward Altman.