PROJECT ON LOGISTIC REGRESSION

COMPANY BANKRUPTCY PREDICTION

WORK DONE BY- SUNIL KUMAR

Content

- Problem Statement.
- Project Steps.
- Visualization.
- Steps and challenges.
- Model and Ensembles.
- Imbalanced classification.
- Model Calibration.
- Model Accuracies and differences
- Challenges.
- Conclusion.

Problem Statement

Prediction of bankruptcy is a phenomenon of increasing interest to firms who stand to loose money because on unpaid debts. Since computers can store huge dataset pertaining to bankruptcy making accurate predictions from them before hand is becoming important.

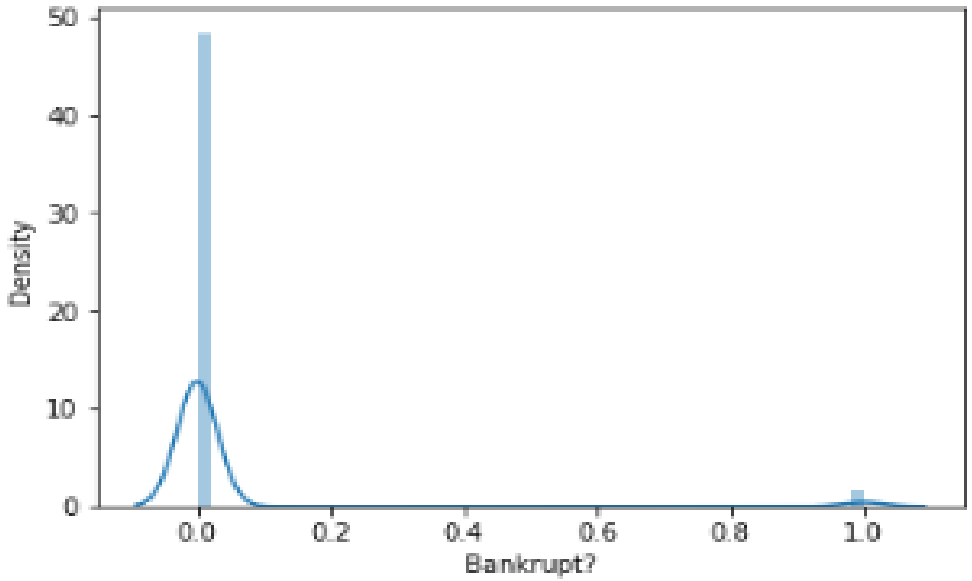
The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.

In this project you will use various classification algorithms on bankruptcy dataset to predict bankruptcies with satisfying accuracies long before the actual event.

Project Steps

- Data collection and data Understanding.
- Loading to tool and importing Libraries.
- Data Preparation and Data cleaning.
- Data Transformation.
- Feature selection and removal.
- Modeling- importing models, Defining the dependent and independent.
- Imbalanced classification.
- Model Ensembles.
- Model Calibration.
- Model Evaluation and selection.
- Deployment.

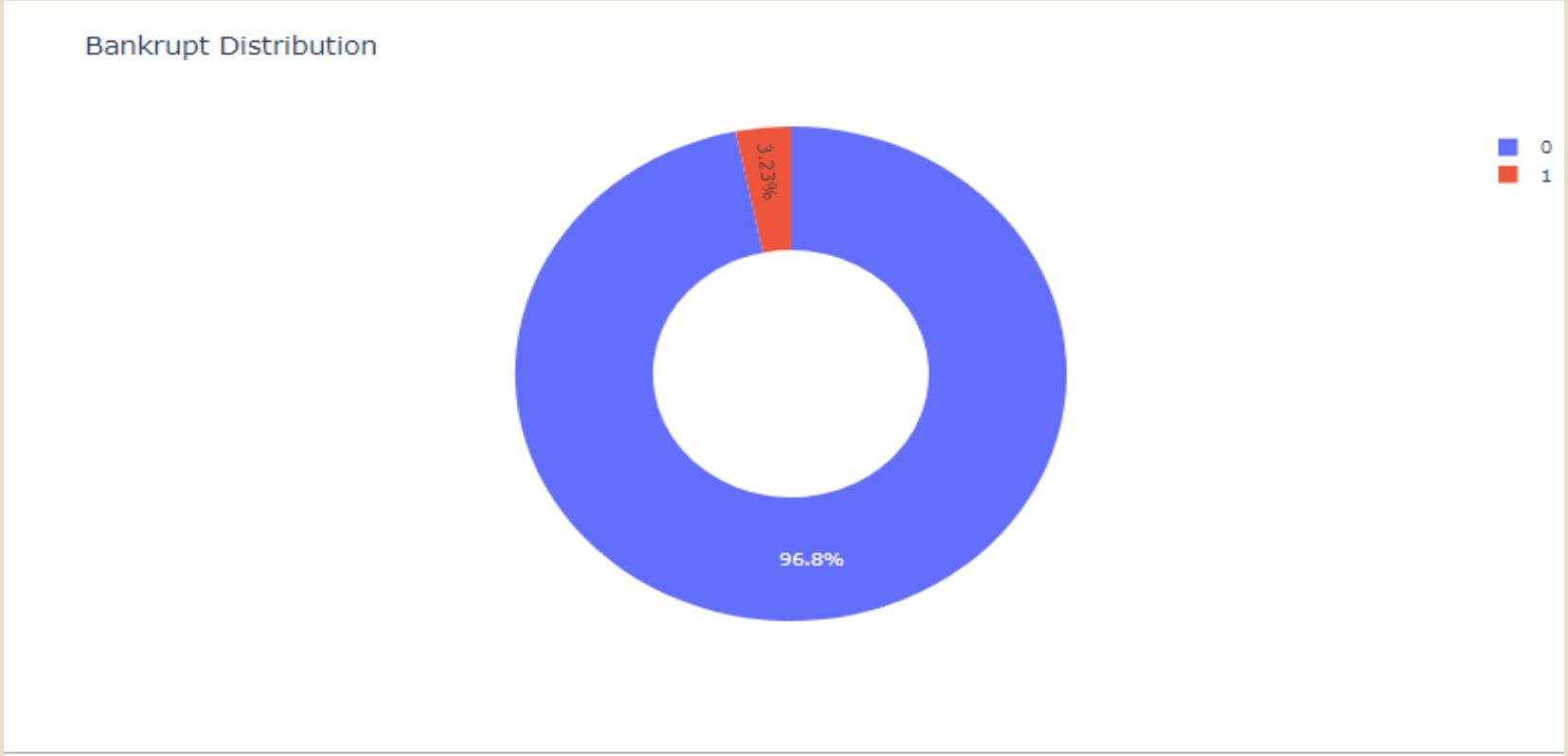
Visualization



From above distplot we can conclude that we have: *Bankrupts? column is normally distributed

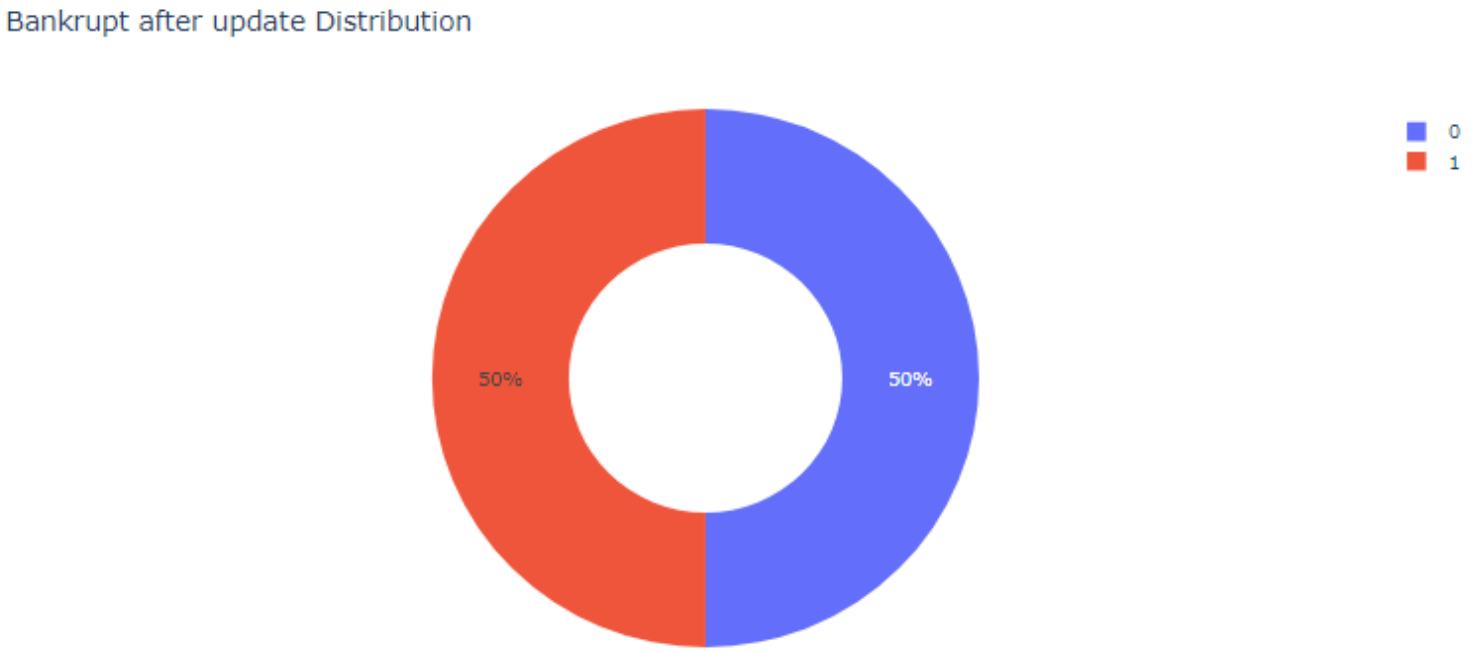
- Bankrupts? is having normal distribution.
- Mean is equal to mode.
- Mean is equal to median as well.

Visualization



- there is a huge difference between bankrupt and non-bankrupt companies.
- As we can see that 96.8% of companies are non-bankrupt and 3.2% are bankrupt.
- Data seems to be imbalanced

Visualization continued..



Its clear that data has been equally distributed now in order to avoid errors in the prediction.

Steps and challenges

- Finding out missing values.
- Finding out duplicates.
- Finding out outliers.
- Data Transformation.
- Feature selection and feature dropping.

Model and ensembles

- Logistic Regression.
- Decision tree.
- Model ensembles.
- Random forest.
- Ada Boost.
- Gradient Boost.
- Voting Boost.

Imbalanced Dataset

The number of observations belonging to one class is significantly lower than those belonging to the other classes.

Fraudulent transactions in banks, identification of rare diseases or in this case of bankrupt company or not etc. In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate.

This happens because Machine Learning Algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution / proportion or balance of classes.

Imbalanced Dataset

Imbalanced dataset resembling techniques:

Random Under sampling.

Random Over sampling.

SMOT(Synthetic Minority Oversampling Technique.

Techniques and accuracies

1ST Approach- Resembling-Minority Accuracies

Logistic Regression accuracy is : 0.5454545454545454

	precision	recall	f1-score	support
0	0.55	0.72	0.62	46
1	0.54	0.36	0.43	42
accuracy			0.55	88
macro avg	0.54	0.54	0.53	88
weighted avg	0.54	0.55	0.53	88

Gradient Boosting Classifier 0.9318181818181818

AdaBoost Classifier Model Accuracy: 0.9204545454545454

Decision Tree accuracy is : 0.8295454545454546

Model Accuracy is: 0.8977272727272727

Voting Accuracy Score is: 0.9431818181818182

Techniques and accuracies continued..

2nd Approach- SMOT Accuracies

Logistic Regression accuracy is : 0.5454545454545454

	precision	recall	f1-score	support
0	0.55	0.72	0.62	46
1	0.54	0.36	0.43	42
accuracy			0.55	88
macro avg	0.54	0.54	0.53	88
weighted avg	0.54	0.55	0.53	88

AdaBoost Classifier Model Accuracy: 0.9306818181818182

Gradient Boosting Classifier Accuracy is: 0.9575757575757575

Decision Tree accuracy is : 0.8295454545454546

Voting Accuracy Score is:
0.9727272727272728

Random Forest Model Accuracy is: 0.9810606060606061

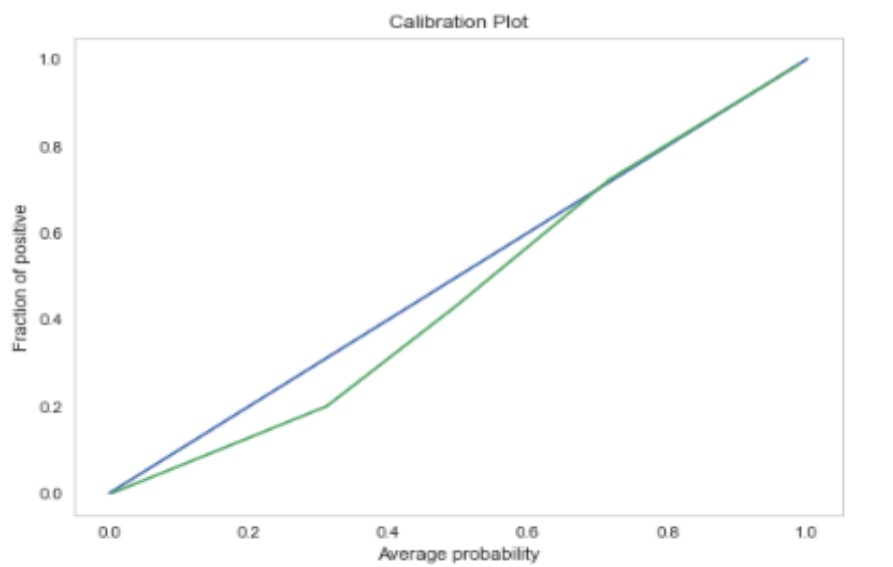
Model Calibration

1)Predicted probabilities that match the expected distribution of probabilities for each class are referred to as Calibrated.

2)comparison of the actual output and the expected output given by a model.

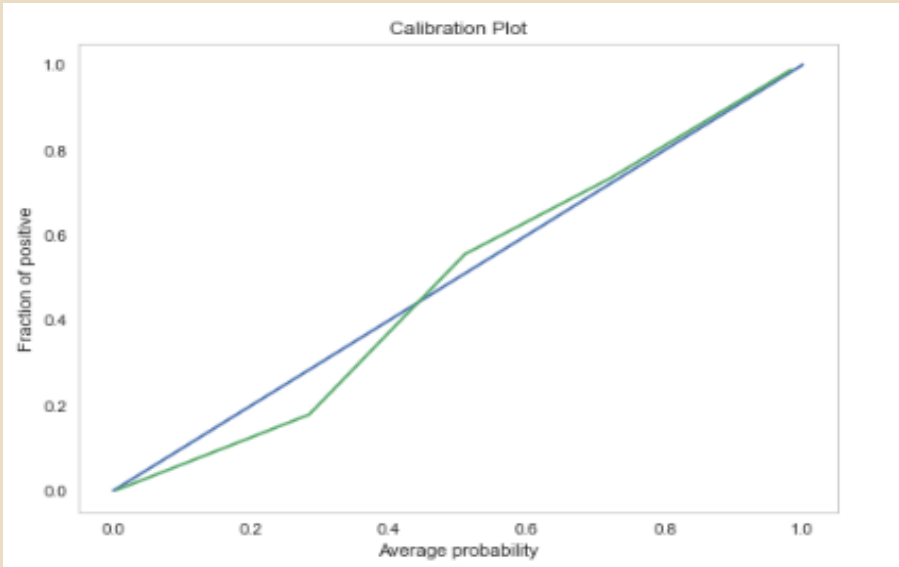
3) Assuring reliable benchmarks and results.

Sigmoid




```
calibrated sigmoid model Accuracy Score is:  0.9840909090909091
```

Isotonic




```
calibrated isotonic model Accuracy Score is:  0.984469696969697
```

Conclusion

Started with data loading and importing the libraries and then started with the exploring the data and looking into columns and rows. It was seen that there were no missing values.

While looking into Bankrupt? Column It was seen that column was following almost normal distribution. While exploring Bankrupt? column further it was evident that data is imbalanced and there is a huge difference between bankrupt and non-bankrupt companies.

I have used 2 techniques to overcome this problem. 1st was to create a dataframe and divide into equal rows that was equal (220,220) rows and 96 columns.

Once I was done here then applied modeling techniques started with Logistic regression and the other ensembles.

Conclusion continued..

I achieved the consistently better quality with a better model every time, I also have used voting classifier which basically takes ensemble of numerous models and gives the best predicted output/accuracy. Voting classifier have given me the best quality. 2nd I have used Synthetic Minority Oversampling Technique(SMOT)in which I oversampled the minority class and balanced the data. I also performed the normalization on non-fractional columns to make sure data following the same scale. I have seen and found that after SMOT results for accuracy were better than the 1st technique. Finally I have used calibration, it basically assures the reliable benchmark and accuracy results when features are very important, by this way I have achieved the best accuracy for the model prediction.

Challenges

The major challenges I have faced in this project are mentioned Below:

I handling the imbalanced data.

Non-fractional column containing uneven values.

Choosing right Model for this problem, explored the Voting classifier.

While oversampling I was unable to create a dataframe and results were coming out different even after few iterations.

Calibration of the Model.

Thank you!