iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47099

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Company Bankruptcy Prediction with SMOTE

Mr. Dhruv Khara
Dept. of Computer Engineering Universal College of Engineering, Mumbai university

Abstract: ***Effective bankruptcy forecasting is essential for financial firms to make sound loan decisions. In general, the two most essential aspects influencing prediction performance are input variables (or features) such as financial ratios and prediction approaches such as statistical and machine learning techniques. While numerous relevant publications have presented innovative prediction algorithms, relatively few have examined the discriminating potential of bankruptcy prediction variables. In addition to financial ratios (FRs), corporate governance indicators (CGIs) have been identified as an important form of input variable in the literature. The prediction performance produced by merging CGIs with FRs, however, has not been thoroughly investigated. Only a subset of CGIs and FRs were employed in related investigations, and the characteristics used may vary from study to study. As a result, the goal of this work is to evaluate the prediction performance obtained by combining seven distinct FR categories and five different CGI categories. Based on a real-world dataset from Taiwan, the experimental results reveal that the FR categories of solvency and profitability, as well as the CGI categories of board structure and ownership structure, are the most critical variables in bankruptcy prediction. The best prediction model performance is determined by combining prediction accuracy, Type I/II errors, ROC curve, and misclassification cost. However, these findings may not be applicable in certain countries, such as China, where the definition of distressed enterprises is vague and the features of corporate governance measures are not explicit.***

Keywords: Bankruptcy prediction, Data mining, neural networks, Decision trees, Support vector machines, SMOTE

I. INTRODUCTION

Bankruptcy or corporate failure may be detrimental to both the organization and the global economy. Long have business practitioners, investors, governments, and academic scholars investigated techniques to identify the possible danger of business failure to limit the economic loss caused by bankruptcy (Balleisen, 2001, Zywicki, 2008).

In short, predicting insolvency is a critical undertaking for many financial firms. In general, the goal is to forecast the chance of a company going bankrupt. Effective prediction models are required by financial organizations to make suitable loan decisions.

A significant amount of study has been devoted to the prediction of bankruptcy, including the use of data mining. Although neural networks, support vector machines, and other algorithms frequently fit data well, they are termed black box technologies due to their lack of comprehensibility. Although numerous researchers have attempted to propose innovative machine-learning strategies that would improve model prediction performance, relatively few have focused on the influence of input variables (or features) on prediction performance. Financial ratios (FRs), which are considered one of the most important elements influencing bankruptcy prediction, are commonly employed to construct prediction models. Solvency, profitability, cash flow ratios, capital structure ratios, turnover ratios, growth, and others are the seven categories of FRs. Corporate governance indicators (CGIs) are important predictors of bankruptcy. CGIs are categorized into five types: board structure, ownership structure, cash flow rights, retained key personnel, and others. However, only a subset of these CGIs has been studied to demonstrate that they increase model performance. The goal of this research is to find the optimal mix of CGIs and FRs so that using them in our model gives the optimal result to use them in predicting bankruptcy. The data included in the bankruptcy prediction model contains about 6800 enterprises. The bankruptcy instances of these firms in the data are indicated as 1 (bankrupted) and 0 (failed to go bankrupt), and 95 financial ratios are used to forecast if they would go bankrupt.

II. EXISTING APPROACHES

A. Existing Approaches/Studies

A significant amount of study has been devoted to the prediction of bankruptcy, including the use of data mining.

Although neural networks, support vector machines, and other algorithms frequently fit data well, they are termed black box technologies due to their lack of comprehensibility. Human users, on the other hand, find decision trees more understandable. However, having far too many regulations might lead to another type of incomprehensibility. The amount of rules produced by decision tree algorithms can be limited to some extent by specifying different minimum support levels. This research compares the accuracy and amount of rules using a range of data mining technologies on bankruptcy data.

Decision trees were shown to be more accurate than neural networks and support vector machines for this data, however there were more rule nodes than expected. Adjusting the minimal support resulted in more tractable rule sets.

According to Altman[1][2], bankruptcy could be explained quite completely by using a combination of five (selected from an original list of 22) financial ratios. Altman utilized a paired sample design, which incorporated 33 pairs of manufacturing companies. The pairing criteria were predicated upon size and industrial classification. The classification of Altman's model based on the value obtained for the Z score has a predictive power of 96% for prediction one year prior to bankruptcy. These conventional statistical methods, however, have some restrictive assumptions such as the linearity, normality and independence among predictor or input variables. Considering that the violation of these assumptions for independent variables frequently occurs with financial data (Deakin, 1976), the methods can have limitations to obtain the effectiveness and validity.

B. Project Scope

Unbalanced dataset is a common difficulty in classification problems that happens when the class distributions are very widely apart. This issue emerges because in machine learning methods, the dominant class outnumbers the minority class. As a result, algorithms frequently forecast the full data set for the minority class quite poorly, indicating closeness to the majority class. Although there are several metric selection and resampling strategies for similar issues, the SMOTE sampling approach is the simplest and most effective to use. The SMOTE oversampling methodology begins with minority class samples and creates synthetic new observations in the feature space at random using the interpolation method. As a result, it strikes a balance between the majority class and the quantity of observations. However, other than increasing the amount of samples, it has no effect on the model and provides no additional information to it. According to some authors, while using SMOTE, the approach should only be used on the train data set and the original test data should be used for evaluating the data. Nevertheless, in some models, after all the data is balanced with the SMOTE method, the split of train and test data and the application of algorithms in this way also stand out in practice In this paper, we provide a method and demonstrate how classification models can be utilized to find business failure and prediction. This approach has the advantage of being flexible in terms of feature selection, as well as being capable of extracting complex rules for people to understand, comparable to expert systems.

The rest of this paper is structured as follows.

Section 3 presents a summary of the current implementation and proposed system.

Section 4 describes the model development and experiment outcomes.

Section 5 covers the findings and potential research directions.

III. PROPOSED SYSTEM

This chapter includes a brief description of the proposed system and explores the different modules involved along with the various models through which this system is understood and represented.

A. Analysis/Framework/Algorithm

1) Datasets for training and testing, determining target variables and predictors. We determine the target variable/column in the dataset before the train-test split. The goal for our dataset would be Bankrupt, which will have a binary value of either 0 or 1, indicating yes or no.

2) Splitting the training and testing, the train-test split is one of the most critical elements in any modelling process. Here is a train-test split dataset with a ratio of 80/20 and random state set to 123.

3) We choose the features that have the most influence on the classification algorithms. For this, we calculate the mutual information between the variables and the target; the lower the value of the mi, the less information we can infer about the target from the feature.

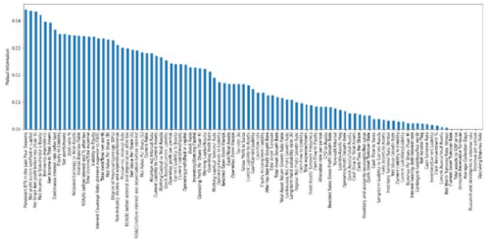


Fig -1: Features vs Mutual Information

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

4) Choose the top k features based on MI. We have over 96 features to pick from, which are a combination of FRs and CGIs. We select only the top 10 features with the highest MI.

The selected features are:

Continuous interest rate (after tax),Persistent EPS in the Last Four Seasons, Per Share Net profit before tax, Debt ratio %, Net worth/Assets, Borrowing dependency, Net profit before tax/Paid-in capital, Net Income to Total Assets, Net Income to Stockholder's Equity, Equity to Liability

5) I trained models such as Logistic Regression, Naive Bayes, K-nearest Neighbors, Decision Trees, and Support Vector Machines to fit and test the model. catboost, xgboost

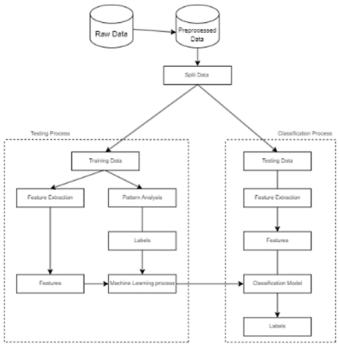


Fig -2: Flow chart illustration of classification modeling.

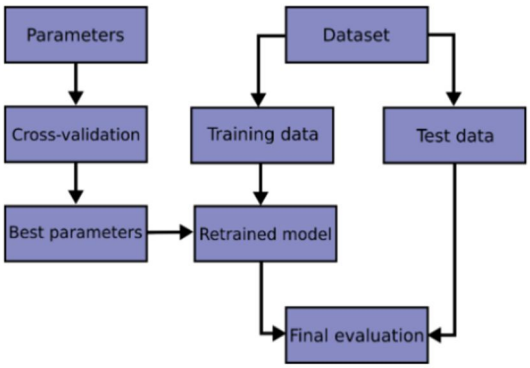


Fig -3: System Architecture

A number of recent research have shown that artificial intelligence systems that are less subject to these assumptions, such as inductive learning, NNs, GAs, and case-based reasoning (CBR), can be used as alternatives to standard statistical methods for classification issues. Inductive learning is a system that automatically extracts information from training samples, in which induction algorithms such as ID3 and classification and regression trees (CART) build a tree type structure to arrange examples in memory. Thus, the distinction between a statistical technique and an inductive learning approach is that various assumptions and procedures are employed in order to produce knowledge.

Messier and Hansen [6] retrieved bankruptcy rules using a rule induction system that categorizes things based on observable characteristic ratios. They started with 18 ratios and used data from two previous investigations. Their system created a bankruptcy prediction rule using five of these ratios. This approach properly classified 87.5% of the holdout data set.

Altman [2] was among the first researchers to foresee corporate insolvency. Altman employed the basic multivariate discriminate analysis (MDA) approach, which was based on the Bayes classification procedure and assumed that the two classes had Normal distributions with identical covariance matrices. Working capital/total assets; retained profits/total assets; earnings before interest and taxes/total assets (ROA); market capitalization/total debt; and sales/total assets were utilized as inputs by Altman (asset turnover). The MDA model and the logistic regression model (hereinafter LR) have both been widely employed in practice and in several academic investigations. They have used as standard reference points for the loan default forecast problem.

Kyung-Shik and Yong-Joo [3] used a genetic algorithm approach. The starting point for the optimization step is a population of genetic structures called chromosomes that are randomly dispersed in the solution space. Each chromosome is assessed using a user-defined fitness function that quantitatively encodes the chromosome's performance. The mating convention for reproduction is set up in such a way that only the highest scoring members will maintain and transmit their desirable traits from generation to generation. Genetic-based algorithms might be utilized to systematically solve bankruptcy prediction difficulties. They used GAs to extract criteria for predicting company failure. The findings indicate that a rule extraction strategy based on GAs for bankruptcy modelling is promising. More informative features will almost certainly lead to better results.

Some observation made while going through the data were;

There were minimal bankruptcies in the ten years between 1999 and 2000. Several firms have a lot of assets, which is usually a positive thing for a company. Despite holding several assets, a company cannot ensure that it will not go bankrupt. The companies in the dataset have been losing money for the past two years, since their net income is expected to be negative. Very few of the organizations that have experienced negative income in the last two years have filed for bankruptcy. It has been discovered that "Debt Ratio%, Current Liability to Assets, Current Liability to Current Assets" qualities have a good association with the target attribute. An increase in the values of the qualities "Debt Ratio%, Current Liability to Assets, Current Liability to Current Assets" leads an organization to incur significant losses, eventually leading to bankruptcy. A rise in the values of qualities with a negative association to the goal attribute aids a company in avoiding insolvency. There appears to be a link between qualities having a high correlation with the target attribute and attributes with a low correlation with the target attribute. We discovered numerous connections among the top 12 traits, one of which is a negative correlation between "Net Worth/Assets and Debt Ratio%."

Now in the analysis section, our work begins with determining the targeted class. As a result, this entire procedure is referred to as classification. An algorithm is a process or formula for solving problems in mathematics and computer science that is based on doing the stages in the defined order. The computer model may be seen as a comprehensive algorithm. Our target class here is bankruptcy. I decided to go with classification models like Logistic Regression, KNN, SVM, Naive Bayes, XG Boost, Cat Boost, RFC, and DTR. Also reducing the dimensions

Dimension reduction is the process of lowering the amount of characteristics in a dataset. Dimension reduction can be used to make the model less complex since the existence of many characteristics might make it harder to comprehend and maintain. Furthermore, the usage of too many features results in poor performance in machine learning algorithms, and models run longer by consuming more memory. Although there is no one "optimal" way for dimension reduction, it may be accomplished using a variety of techniques such as missing value ratio in features, low-high correlation filter, random forest algorithm, factor analysis, and principal component analysis. In summary, dimension reduction simplifies the model and ensures that it may be produced more readily with simple variables.

B. Models Used

Logistic Regression is intended for categorization and is particularly beneficial for comprehending the impact of several independent factors on a single outcome variable. As our result is binary it can be trained to obtain desired outcome.

Naive Bayes assumes predictor independence, or the Bayes theorem. To estimate the appropriate parameters, this sort of method requires a modest quantity of training data. When compared to more cosmopolitan ways, this procedure is extremely quick. The only issue is that the forecast may be incorrect.

SVM (Support Vector Machine) aids in the coordination of groups with varying characteristics. For example, if we simply knew two characteristics, such as a company's EPS and EBITA, we would have to first plot these two-dimensional spaces where each point has two coordinates, known as support vectors. It performs well in high-dimensional spaces and employs a subset of training points in the decision function, making it memory economical. It would take into account any of the categories specified by the users, even if they are irrelevant.

CatBoost employs ordered target encoding, which allows you to maintain the feature/column in its original condition, allowing for easier collaboration. Not to be concerned with matching one-hot-encodings of multiple features, and to interpret the features as they were intended. Furthermore, this encoding allows for greater feature relevance. Training is more efficient. Categorical characteristics are more important. The model is more precise.

XGBoost is a fantastic algorithm. It works well with little amounts of data, subgroup data, large amounts of data, and problematic data. It does not, however, function well with sparse data, and widely spread data may also cause issues. However, it beats most supervised learning algorithms on specific types of data problems. The black box nature is most likely the most significant limitation. XGBoost will not offer effect sizes if you need them.

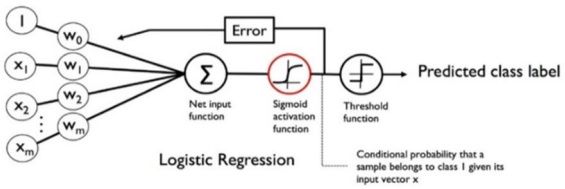


Fig -4: Logistic Regression Architecture

IV. EXPERIMENTAL RESULTS

The result received by the pipeline was as follows; the assimilated results vary significantly based on the model used.

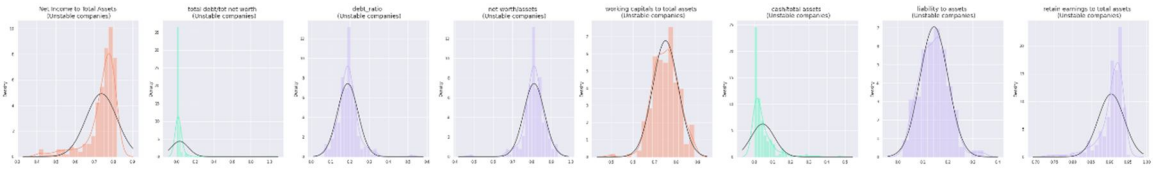


Fig -5: feature distributions for close to bankruptcy companies

We can see that eliminating the extreme outliers certainly aids in obtaining more "bell shape" distributions.

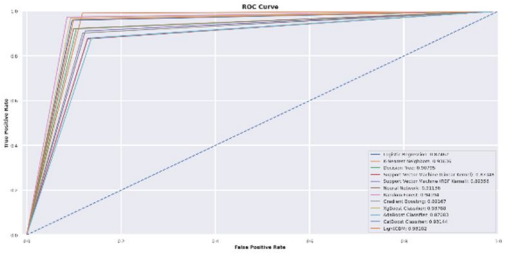


Fig -6: ROC curve for different models

Following dimension reduction, the new data with 7 variables was ran using the SMOTE model (in which the test data was also SMOTE). When compared to the original model, the ROC curve and confusion matrices produced pretty poor results; nevertheless, despite the fact that the model could be described with just 7 variables rather than 74, no significant performance reduction was found. The ROC curves and confusion matrices of the "reduced model" are shown in the Fig 6.

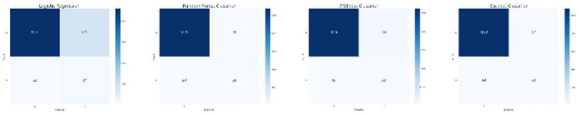


Fig -7: Confusion Matrix for different models

As seen by the validation set findings, all of our models continue to struggle in recognizing firms on the verge of bankruptcy. Logistic Regression is the approach that detects more minority class observations, although at a substantial cost in terms of precision (high presence of false negatives). Despite the flaws, I believe that it is better to identify a non-near to bankruptcy observation as close to bankruptcy than vice versa in this situation, therefore it might be a helpful model.

A. Result based on Models

	precision	recall	f1-score	support
Fin.Stable	0.99	0.84	0.91	605
Fin.Unstable	0.15	0.82	0.26	22
accuracy			0.84	627
macro avg	0.57	0.83	0.58	627
weighted avg	0.96	0.84	0.88	627

Fig -8: Metrics for Logistic Regression

	precision	recall	f1-score	support
Fin.Stable	0.98	0.98	0.98	605
Fin.Unstable	0.41	0.41	0.41	22
accuracy			0.96	627
macro avg	0.69	0.69	0.69	627
weighted avg	0.96	0.96	0.96	627

Fig -9: Metrics for Cat Boost

Using validation data, we can see that the measure under consideration (F1) is greater when catboost is used. Nonetheless, in this scenario, the best choice is to utilize Logistic regression since it can better comprehend the minority class, even misclassifying some enterprises that are not near to bankruptcy as close to bankruptcy.

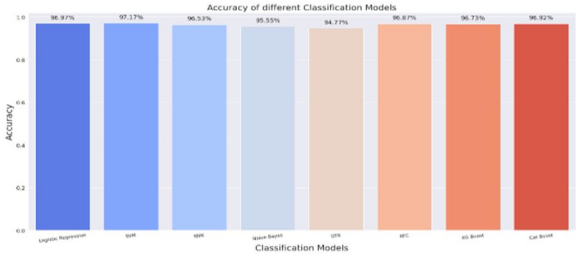


Fig -10: Accuracy of different Classification Models

One key observation was restricting the data characteristics does not necessarily produce better outcomes. However, there is always the potential that the findings will be a few points better than the overall dataset characteristics.

B. Future Scope and Implementation

1) Better handling of class imbalance

2) Increasing accuracy by adding more by standards, also implement lazy classifier

V. CONCLUSIONS

We were able to create three models with an accuracy of 0.97 despite drastically lowering the amount of features (just 10). As a result, we were able to save running time (from 5.47 seconds to only 0.5)

We were also able to articulate how a corporation may go bankrupt or not with the reduced characteristics, which helped us explain the model better.

1) high "Interest-bearing debt interest rate" tend to go bankrupt (≈ 0.000499)

2) high "Total debt/Total net worth" tend to go bankrupt (≈ 0.015723)

3) high "Fixed Assets Turnover Frequency" tend to go bankrupt (≈ 0.001225)

4) low "Cash/Total Assets" tend to go bankrupt (≈ 0.023755)

5) low "Equity to Liability" tend to go bankrupt (≈ 0.018662)

6) companies with a low 'Net profit before tax/Paid-in capital', 'Persistent EPS in the Last Four Seasons' and 'Net Value Per Share (A)' tend to go bankrupt

REFERENCES

[1] Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23, 589–609.

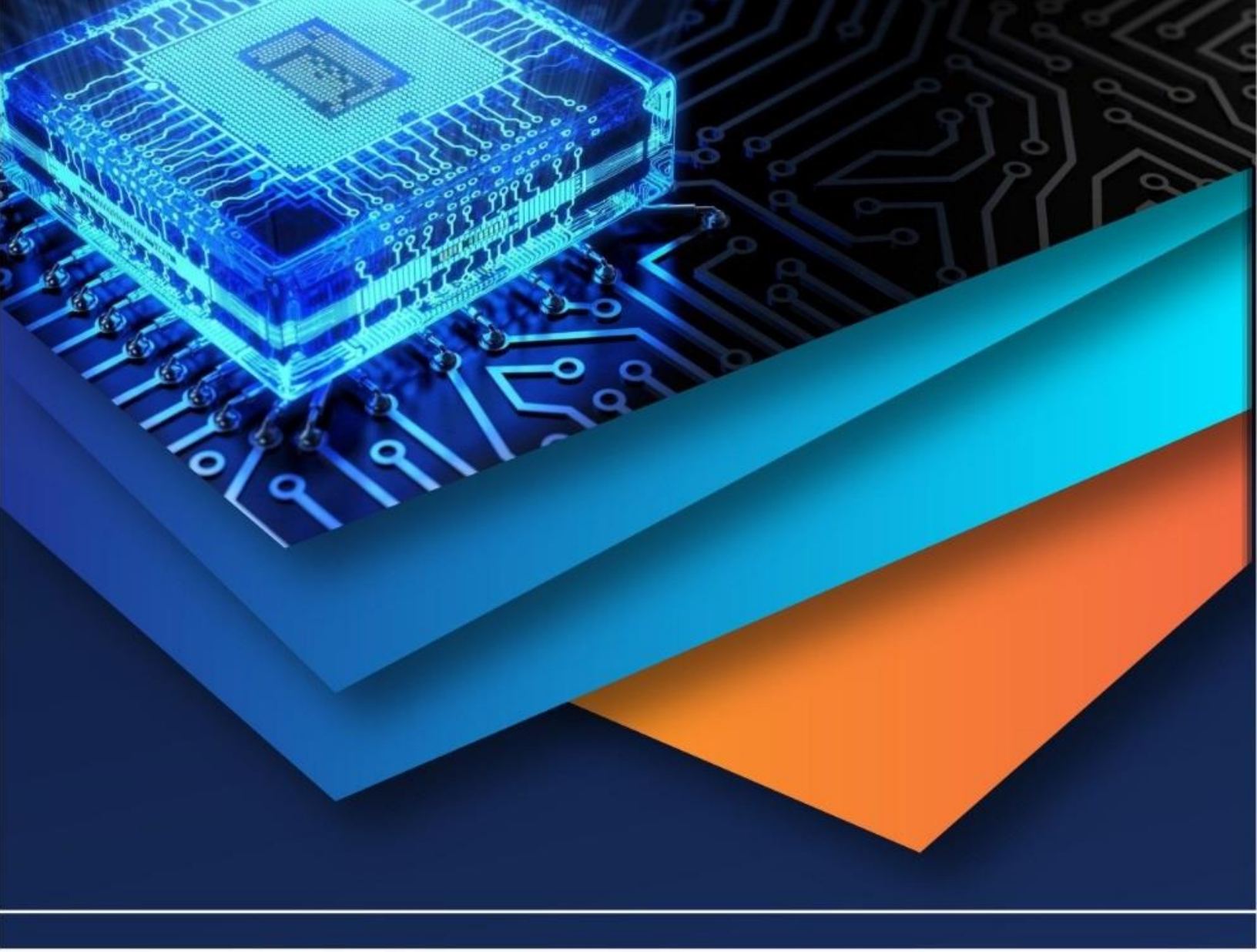
[2] Altman, E. (1983). Corporate financial distress—A complete guide to predicting, avoiding and dealing with bankruptcy. New York: Wiley.

[3] Shin, K., & Lee, Y. (2002). A genetic algorithm application in bankruptcy prediction modeling. Expert Systems with Applications, Expert Systems with Applications, 23(3), 321-328.

[4] Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems, 52(2), 464-473. https://doi.org/10.1016/j.dss.2011.10.007

[5] Baldwin, J., & Glezen, G. W. (2016). Bankruptcy Prediction Using Quarterly Financial Statement Data. Journal of Accounting, Auditing & Finance. https://doi.org/10.1177/0148558X9200700301

[6] Messier, W., & Hansen, J. (1988). Inducing rules for expert system development: An example using default and bankruptcy data. Management Science, 34(12), 1403–1415.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY