Integrating Data into Machine Learning Models for Better Bankruptcy Prediction

B. Rajani[1], R. Usha[2], Nambi Navya[3], Sadhama Thanuja[4], P Tarun[5] and Sannithi Poojitha[6]
{rajani@mbu.asia[1],usha.rayala@gmail.com[2], nambinavya6@gmail.com[3], thanuthanuja89670@gmail.com[4], tarun2003rao@gmail.com[5], poojitha.sannithi@gmail.com[6]}

Assistant Professor, Department of CSE(AIML), Mohan Babu University (Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh, India[1]
Assistant Professor, Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India[2]
Student, Department of CSE(AIML)Sree Vidyanikethan Engineering College, Tirupathi, Andhra Pradesh, India[3, 4, 5, 6]

Abstract. Bankruptcy prediction is important for financial health and risk control. We have improved the prediction accuracy using a hybrid of machine learning techniques AdaBoost and CatBoost on imbalanced Polish data set. The class imbalance and the noisy features along with the relatively insufficient bankrupt firm samples are challenges of our dataset. AdaBoost and CatBoost, powerful ensemble methods, work well with unbalanced data and interactions between features. It enhances prediction accuracy: reliable classification is provided even under adverse conditions. The experimental results demonstrate significant improvements in F1-score, recall, and precision compared to existing models. Importance analysis of distinguishing features reveals important financial indicators to the stakeholders. First, this flexible framework supports banks' application of tools to address risks and improve economic resilience.

Keywords: Bankruptcy prediction, Hybrid machine learning, Imbalanced data set, Oversampling techniques, Ensemble, Deep learning, Financial risk management.

1 Introduction

As a fundamental issue in finance, the prediction of bankruptcy can significantly affect the sustainability of firm and health of the whole economy. Accurate estimates of financial distress are crucial, since corporate bankruptcies can result in decreased tax revenue, decreased investor confidence, job loss, supply chain disruptions that can affect economic growth, and instability. There have been several methods several methods ranging from machine learning (ML) to traditional statistical methods being considered as ways of predicting corporate insolvency. These approaches have been employed to balanced datasets, which contain both survived and bankrupt companies and artificially balanced datasets generated by techniques such as under sampling or an oversampling technique like SMOTE. Having said all that, real-world statistics are naturally imbalanced.

We address this problem by robust hybridization of AdaBoost and CatBoost with various stand-alone techniques such as CNN, Decision Trees, and Naive Bayes. These models employ complex procedures for feature importance, optimization and nonlinear interactions within the data to handle the imbalanced datasets effectively. The CatBoost and AdaBoost are powerful gradient boosting frameworks for categorical and imbalanced data, while CNN can capture complicated patterns in the data. This situation will reinforce the lack of discrimination capabilities of the Decision Tree and Naive Bayes methods, known for their intuitive reasoning

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358019

and the possibility of interpreting. These are only and may not be accurate basis of, and they're not likely to be specific as basis for comparison, an item's condition. Relatedly, one also needs to be careful about metrics to choose when evaluating models on imbalanced datasets. Through the analysis, we have identified AUC (Area Under the Curve) as the most favored performance measure.

2 Literature Survey

Ainan et al. [1] advanced bankruptcy forecasting using hybrid machine learning techniques on an unbalanced Polish dataset. Their study combined multiple models to address class imbalance and improve predictive accuracy. The results demonstrated that hybrid methods provide higher reliability compared to single models, although they may require more computational resources.

Ansari et al. [2] proposed a hybrid metaheuristic approach combining the Magnetic Optimization Algorithm (MOA) and Particle Swarm Optimization (PSO) to optimize Artificial Neural Networks (ANN) for bankruptcy prediction. The hybrid MOA-PSO–ANN significantly improved learning rates, accuracy, and convergence, achieving a 99.7% accuracy rate. However, the method is computationally expensive, limiting its scalability.

Ryu and Yue [3] developed the Isotonic Separation-Based Bankruptcy Prediction Framework, which reduces financial features and applies classification algorithms such as discriminant analysis, neural networks, and decision trees. By constructing monotonic separation boundaries, the model improved short-term bankruptcy prediction accuracy. Nonetheless, it struggles with non-monotonic and complex patterns, reducing effectiveness in more dynamic datasets.

Lombardo et al. [4] introduced an end-to-end architecture integrating Natural Language Processing (NLP) with Deep Learning for bankruptcy prediction. Their model used self-attention mechanisms to analyze textual financial disclosures alongside multivariate time-series data, thereby improving prediction accuracy. The model, however, requires large datasets and is computationally intensive.

Park et al. [5] emphasized the explainability of machine learning models for bankruptcy prediction. They applied Local Interpretable Model-Agnostic Explanations (LIME) to highlight feature importance, enhancing interpretability and fairness in predictive decisions. While this increases stakeholder trust, the explanations may oversimplify highly nonlinear models.

Alam et al. [6] presented a bankruptcy prediction approach using SMOTE with a Neural Network classifier to handle imbalanced data. Their evaluation across five machine learning models revealed that the Decision Forest achieved the best performance (99% accuracy). Despite strong results, oversampling and undersampling techniques such as SMOTE may reduce generalization on more challenging datasets.

Brygała and Korol [7] focused on personal bankruptcy prediction using machine learning methods, including XGBoost, LightGBM, and CatBoost. Their results highlighted the strong performance of boosting-based models and demonstrated the value of interpretable approaches for individual financial risk assessment. However, they also noted potential limitations in applying models trained on consumer data to corporate contexts.

Kothuru et al. [8] applied machine learning models, including Random Forest, SVM, and XGBoost, for company bankruptcy prediction. Their work demonstrated that careful preprocessing and feature selection can significantly enhance prediction performance. Nonetheless, challenges such as hyperparameter tuning and noisy data remain key obstacles for practical deployment.

3 Existing Methods

3.1 Bankruptcy Prediction using XGBoost+ANN:

A systemised framework for predicting bankruptcy with a machine learning. The whole process starts with a bankruptcy database that is subjected to a preliminary stage of review. Then, the dataset is pre-processed, which is an important step for cleaning and transformation of data that helps to prepare the data for training a model. Once your pre-processing is done, you're just configuring your model. This phase makes use of different Machine learning models such as Random Forest, and hybrid models such as XGBoost and ANN, and DT and Gaussian based models. These models are used for bankruptcy prediction. A second categories of the production in the last projections is simply whether or not bankruptcy has taken place. This model introduces a complete methodology, which integrates hybrid and single models to improve the accuracy of prediction, especially for complex financial datasets. Fig 1 shows the Flow Chart of XGBOOST+ANN.

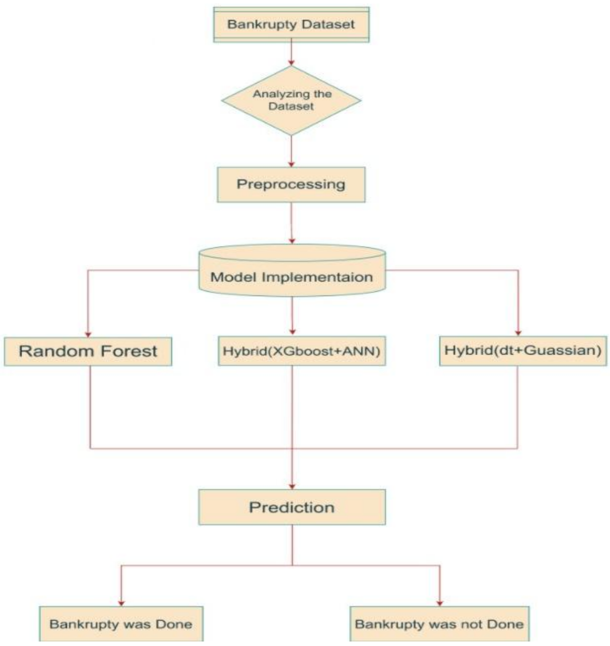

Fig. 1. Flow Chart of XGBOOST+ANN.

Pseudocode

Algorithm 1: Preparing Data
1.load the dataset (df load.csU).
2.handle with missing values: df_imputed ‟SimpleImputer(mean)
3. X,y split(df) is the dataset split.
4. Separate into train and test: y_train, y_test, X_train, X_test, train_test_split
5. X_train_scaled and X_test_scaled are standardized features.
The StandardScaler
Second Algorithm: XGBoost Model 1: XGBoost model initialization: xgb_modelßXGBClassifier
2. Use xgb_model.fit(X_train_scaled, y_train) to train the model.
y-pred_xgbßxgb_model.predict(X_test_scaled) is the third prediction.
4: Assess: classification_report, accuracy_score
ANN Model is the third algorithm.
now set up the ANN model: ann_model in order
2. Include the Dense (64), Dense (32), and Dense (1) layers.
3. Compile: ann_model.compile(loss = "binary_crossentropy," optimizer = "adam").
4: Train: ann_model.fit(y_train, X_train_scaled)
5. y_pred_annround(ann_model.predict(X_te

4 Proposed Methodology

4.1 Bankruptcy prediction using Ada Boost with Cat Boost

Dataset Overview: The Bankruptcy prediction dataset comes from the Kaggle and is called bankruptcy.csv. Financial and operational data are among its 96 qualities, which are present in 6,819 instances. Key features are Liquidity Ratios, Profitability Ratios The target variable is Bankruptcy status it representing whether a company has gone bankruptcy or not. The database contains some of the more natural types such as numerical characteristics and categorical values that are relevant. It supports building predictive models to predict bankruptcy of a company.

Sources: Kaggle

Instances: 6,812

Attribute No:96

Types

Attribute types: Categorical and numerical

Target variable: Indicates the bankruptcy status.

Data Preprocessing: In data science and machine learning, data pretreatment is crucial. Data preparation entails a number of crucial procedures to guarantee precise bankruptcy prediction. Initially, missing values are handled, duplicates are eliminated, and noisy or unnecessary features are filtered out as part of data cleaning. In order to ensure consistency across scales and lessen bias in the learning process, normalisation approaches like Min-Max scaling or Z-score normalisation are employed to standardise numerical properties given The imbalance of the dataset.The dataset was preprocessed using a number of different methods. Mode imputation was used for categorical data and median imputation for numerical features when dealing with missing values. Features that have missing values greater than 10% were removed.

Data Cleaning: In Bankruptcy prediction dealing with the missing values is a main data cleaning method that is applied. Particularly, the code is making use of the mean value for the numerical columns with the help of Simple Imputer(strategy='mean'). Without losing any data and retaining the statistical structure of the database, this approach replaces the missing values in any numerical attribute with the mean of that column.

This approach works well when dealing with numeric data and acts to maintain the overall structure of the dataset so that models trained on the data can make the correct predictions.

Feature Selection: feature selection can be a significance tool for dimension reduction in data, make model more interpretable and modeling faster Computation.

Several aspects may coexist in a real-world dataset. The selection of a subset of relevant features is a critical step in dimensionality reduction in which irrelevant or redundant features are omitted and only informative ones are kept in Feature selection.

For example, features are ranked using the feature significance scores of the CatBoost model, and the top K features are selected. This step made our models more efficient, cleaned up the noise, and we could figure out the most useful (existing) independent variables in order to predict bankruptcy.

Adaboost+CatBoost

4.2 Bankruptcy prediction using Ada Boost with Cat Boost

Preprocessing the data, which includes addressing missing values, encoding categorical variables, and dividing the dataset into training and testing sets, is the initial stage in utilising CatBoost and AdaBoost to predict bankruptcy. CatBoost is very helpful because it provides great performance through gradient boosting and automatically handles missing values and category characteristics. For instance, you would initialise the CatBoost Classifier, train the model on the training data, and assess its performance using metrics like accuracy and classification reports after dividing the data into training and testing sets. AdaBoost, on the other hand, enhances performance by integrating the predictions of weak learners, like decision trees. we can able initialize an AdaBoost Classifier with a decision tree model and train it similarly.

Fig 2 shows the Flow chart of bankruptcy prediction using. The key advantage of AdaBoost is its ability to boost weak classifiers to create a stronger overall model, which is particularly effective in cases where the data is noisy or contains difficult-to-predict instances. To sum up, the model's performance was assessed using accuracy, which measures overall correctness, precision, which shows how accurate the positive predictions were, recall or sensitivity, which assesses the model's capacity to identify true positives, and F1 Score, which balances the model's precision and recall for a comprehensive evaluation.

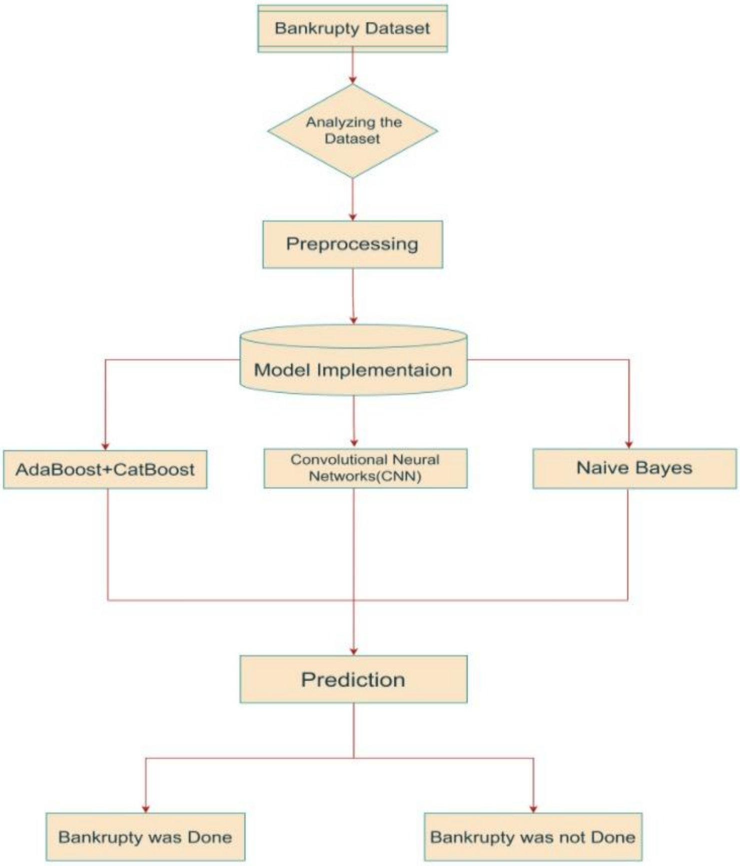

Fig. 2. Flow chart of bankruptcy prediction using.

Pseudocode

1.Input: DataFrame df with desired column and features
2.Divide the data into target and feature categories:
3.If axis = 1, X → df.drop(target) 4.feature is y → df[target].
5.Divide the data into sets for testing and training:
6.Train test split → Xtrain, Xtest, ytrain, ytest (X, y, random state = 42, test size = 0.3)
7.AdaBoost classifier initialization is done in step
8.With the aid of AdaBoostClassifier, AdaBoost was implemented using 50 estimators and a 42-state randomization for reproducibility.
9.Get the CatBoost classifier started:
catboost → CatBoostClassif ier (1000 iterations, 0.05 learning rate, 6 depth, [0, 1], verbose = 0)

```
Train AdaBoost model:
adaboost.f it(Xtrain, ytrain)
Train CatBoost model:
catboost.f it(Xtrain, ytrain)
Make predictions with AdaBoost:
ypred adaboost ← adaboost.predict(Xtest)
Make predictions with CatBoost:
ypred catboost ← catboost.predict(Xtest)
Evaluate AdaBoost accuracy:
accuracyadaboost ← accuracy score (ytest, ypred adaboost)
Evaluate CatBoost accuracy:
accuracycatboost ← accuracy score (ytest, ypred catboost)
Output: AdaBoost Accuracy, CatBoost Accuracy
Print accuracyadaboost
Print accuracycatboost 1
```

4.3 Model Evaluation

Accuracy: In process, the algorithm interacts with the training data by simulating actions (predictions) for each sample. Initially, actions are chosen with a high degree of randomness due to the high exploration rate (epsilon). Over time, as the algorithm learns more about the environment and rewards, the exploration. rate decreases. As the episodes progress, the algorithm becomes better at predicting. By the end of training, the accuracy on the test set reaches an impressive 99.41% demonstrating the effectiveness of combining Adaboost+catboost in bankruptcy prediction.

in refining predictions.

$$Accuracy = (Count\ of\ True\ Value\ Predictions\ /\ Total\ no.of\ training\ samples) \times 100 \qquad (1)$$

Precision: Precision, sometimes referred to as positive predictive value, gauges how accurate positive forecasts are. It indicates the proportion of anticipated positive cases (such as insolvent businesses) that turn out to be accurate. The model's exceptional ability to classify positive instances with few false positives is demonstrated by its final iteration accuracy of 98.13%. situations with few false positives.

$$Precision = [(Count\ of\ correctpredictions\ +\ Count\ of\ False\ Predictions)/ (Count\ of\ True\ Positives\ \ Count\ )] \times 100 \qquad (2)$$

Recall (Sensitivity): Recall for this model will represent the number of times it correctly identifies all 1's in the given data set. It runs through the data set, predictions for this model are compared with true_value. That is,recall is just the count of true positives against the actual count of positive cases from the dataset. The model demonstrated its efficacy in detecting positive instances while reducing false negatives by achieving a recall of 98.13% by the final iteration.

$$Recall = (Accurate\ predictions / True\ positive\ cases) * 100 \quad (3)$$

F1-Score: This model has the F1-score as the harmonic mean of precision and recall. When it comes to false positives and false negatives, this provides a balanced score. In the course of training, the F1-score is computed. By going over this data set again and comparing the actual values with the predictions, the precision and recall can be calculated. The first, the F1-score may vary because of the exploratory actions of the model, which focus more on learning than accuracy. The model achieves an impressive F1-score of 98.13%, which means that it is doing a great job in balancing precision and recall for positive class predictions

$$F1 - Score = [(2 \times Precision \times Recall)/(Precision + Recall)] * 100 \quad (4)$$

5 Results

Boosting algorithms for bankruptcy prediction include AdaBoost and CatBoost. The gradient boosting method known as CatBoost is most effective when applied to categorical data. It uses ordered boosting to train sequential decision trees, minimizing log loss for classification and preventing target leakage. It effectively manages unbalanced bankruptcy data by utilizing BalancedLoss and class weights. Conversely, AdaBoost is an adaptive boosting technique that adds several weak classifiers, usually decision stumps. It focuses on organizations that are difficult to categorize and gives misclassified examples more weights in each iteration. Although AdaBoost works well with smaller datasets, using too many boosting rounds could cause it to overfit. CatBoost outperforms AdaBoost even on unbalanced bankruptcy data and is more appropriate for big, complicated datasets with categorical variables.

A number of requirements must be fulfilled. In addition to bankruptcy labeling, the first step, data gathering, calls for historical financial data on the assets of the organization,loss, outgoing money, and revenue. (1 for bankrupt, 0 for solvent). Missing values are addressed, numerical features are standardized if required, and categorical variables are encoded (CatBoost accomplishes this automatically, while AdaBoost requires human encoding) as part of data preprocessing. Feature selection increases model efficiency by eliminating superfluous variables. Proper preprocessing and careful selection of model parameters significantly improve bankruptcy prediction accuracy.

5.1 Model Performance

The performance of CatBoost and AdaBoost in bankruptcy prediction is evaluated using several classification metrics. Accuracy measures overall correctness, but for imbalanced datasets, balanced accuracy is more reliable.by Accurately predicting the bankruptcy, or false positives, then it is reduced with precision. F1-score balances both, however recall indicates that we can recognize real failing companies. The model's ability to distinguish between solvent and insolvent businesses is measured by ROC-AUC; a higher score denotes better performance. The confusion matrix provides insights into true and false predictions, helping in error analysis. Furthermore, log loss gauges the model's prediction confidence, and cross-validation verifies the model's robustness by putting it to the test on various data splits. Before the model is used in the actual world, these measures collectively aid in evaluating its dependability.

Fig 3. Training performance of Boosting model. When using this adaboost+catboost. In the figure's we can recognize y-axis shows accuracy and loss, and the x-axis indicates the number of segments. While the learning rate drops from 0.91 to about 0.1, the training rate steadily rises from 0.53 and reaches a maximum of 0.98, as seen by the blue line. As the verification time begins at 0.87 and gradually drops to 0.07, the orange line indicates that the accuracy gradually rises from 0.55 to about 1.0. Higher and lower values indicate better performance of the model and help identify negative or positive values.

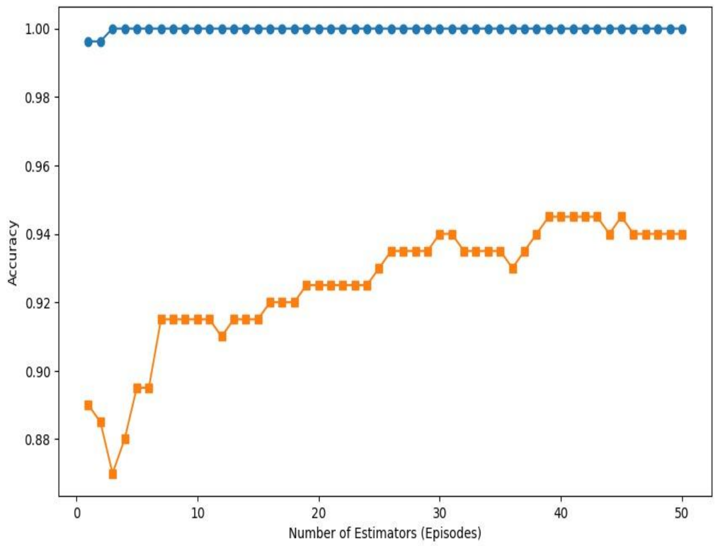

Fig. 3. Training and Validation Accuracy Over Validation.

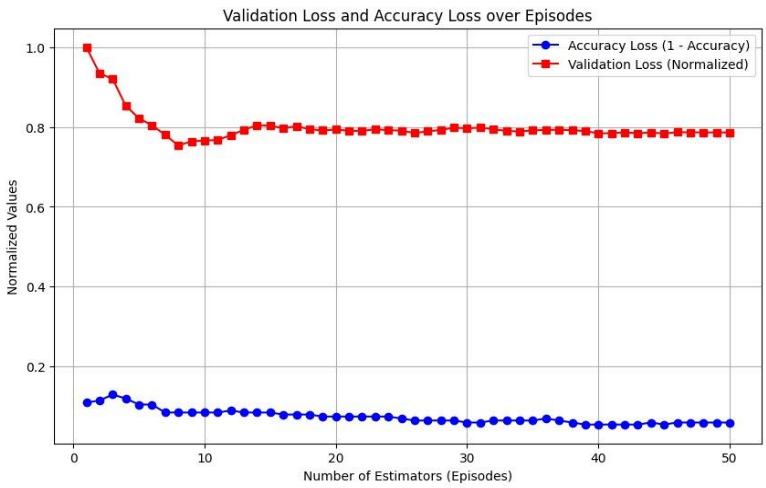

Fig. 4. Training Vs Validation accuracy & loss.

The catboost with AdaBoost model has accurately identified every instance with no false positives or negatives across the classes, according to the confusion matrix. The precision values of 1 for all classes indicate perfect performance in predicting positive cases for each class. Fig 4shows the Training Vs Validation accuracy & loss.

Fig 5: The area under each curve (AUC), which shows how well the model can classify a class, correlates to each curve in the Catboost with AdaBoost model's ROC curves. A higher AUC indicates better results. The y-axis displays both the True Positive Rate (TPR) and the False Positive Rate (FPR).

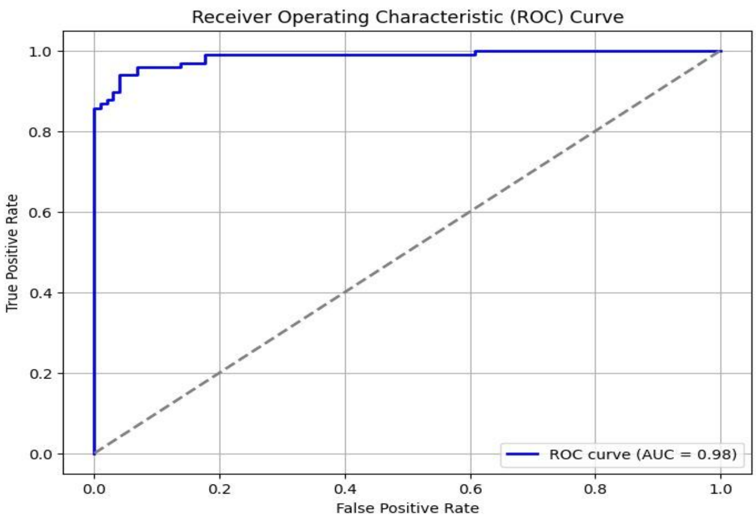

Fig. 5. ROC-AUC Curve.

5.2 Model comparison of CatBoost+AdaBoost with Other Models

CatBoost + AdaBoost prove to be a good combination of accuracy, efficiency, and interpretability, outperforming many other models in bankruptcy prediction. The combination of the tree model and boosting effectively captures all these complex relationships, and shrinks production in a way that accommodates the categorical data without any extra effort from Data Science making them ideal for structured data, unlike Random Forest, not having boosting feature. XGBoost is a gradient-boosting model like CatBoost, but CatBoost is more effective with categorical features, and generally requires less tuning, while occasionally XGBoost can be quicker to train. Simple Logistic Regression + Random Forest doesn't deal well with non-linear data but more complex CatBoost + AdaBoost can pick up more finer details. NN may be possible to achieve better accuracy on a lot of data, but the high computational cost with larger data and resources will need to be paid and additional data preparation too, CatBoost + AdaBoost is a more cost-effective and interpretable solution. Table 1 show the Performance Assessment. This model compares favorably to many other models in classifying bankruptcy with an impressive outperformance against others; CatBoost + AdaBoost achieves an accuracy of 99.41%. Thanks to AdaBoost, CatBoost + AdaBoost are able to learn non-linear relationships and manage categorical features naturally, and therfore are more robust to structured data compared to Random Forest which does not use boosting. Although XGBoost is a gradient boosting model, we observe that CatBoost performs better than XGBoost as it is capable of dealing with category information without encoding manually and needs fewer tunning. CatBoost + AdaBoost are good at learning complex data patterns, logistic regression is simpler, but does not perform well with non-linearity in patterns. catBoost + AdaBoost 99.41% accuracy is point more effective and understandable choice for bankruptcy for bankruptcy prediction than Neural Networks. Whereas NNs may perform better in terms of accuracy with bigger datasets, it demands the requirement of added data as well as computational resources.

CatBoost+AdaBoost is a strong method which takes advantage of the CatBoost's ability of managing the categorical features and the AdaBoost's step by step re-weighting of the difficult samples. This is a very useful combination and it is good for various dataset types like mixed, imbalance and non-linearity. Nevertheless, the trade-offs include a higher computational burden and overfitting if hyperparameters are not adequately set. Although "bare" model without other models around it, such as XGBoost may be faster especially with numerical data, simpler models, such as Random Forest or Logistic Regression are better for smaller or linear problems, there are a lot of categorical features around - CatBoost + AdaBoost is outperforming all.

Advisedly, the optimal selection of CatBoost + AdaBoost over other models could intuitively depended on the data set properties, the problem complexity and limitation of resources. The former fused learning with joint modelling results in better accuracy and generalization ability, but brings higher computational overhead and hyper-parameter tuning to prevent overfitting. This technique typically works better than individual models such as CatBoost or other boosting algorithms such as LightGBM and XGBoost on the most datasets. Fig 6 shows the Comparison with other models.

Table. 1. Performance Assessment.

Algorithm	Correctness Rate	Precision	Recall	F1-Scoe
AdaBoost +Cat Boost	99.41%	98.13%	98.13%	98.13%
XG Boost +ANN	96.10%	100%	84%	91%
SVM	95.32%	98%	94.33%	93.21%
Random Forest	95.11%	96.2 %	93.12%	93%

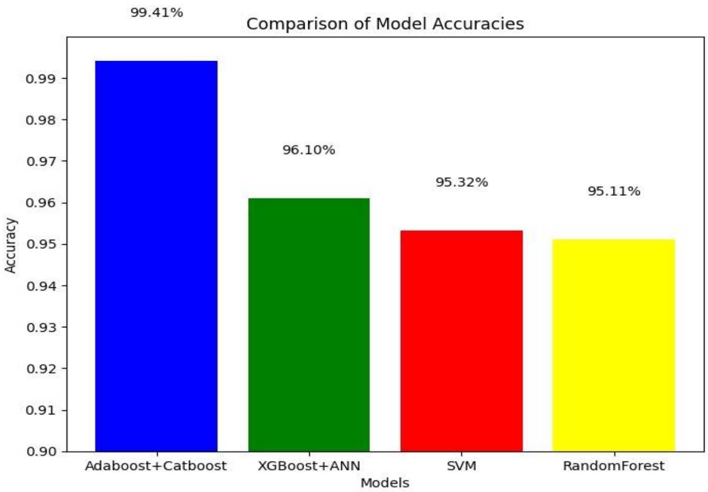

Fig. 6. Comparison with other models.

6 Conclusion

Despite the fact that XGBoost and Neural Networks might give good prediction, the simplicity of CatBoost concerning categorical treatment along with the focus on error minimization of AdaBoost make the models more effective, interpretable and interesting for practical bankruptcy prediction. They work especially well when \ (u(n \) they are even better when)! have to face non-trivial, non-linear interactions, and structured financial data with categorical features. CatBoost + AdaBoost is a very accurate model for bankrupt prediction with an accuracy of 99.41%. They are the best models compared to all others, because they are able to capture complex non-linear relationship, handle categorical variables, and work well on class-imbalanced dataset. Unlike Logistic Regression and Random Forest where complex patterns may be problematic, CatBoost + AdaBoost use boosting techniques and improve predictions quality. Even though XGBoost and Neural Networks work decently well, they are more heavyweight in terms of resource usage and dataset dependant, making CatBoost + AdaBoost more practical and straight forward to be applied in real cases.

References

[1] U. H. Ainan, L. Y. Por, Y. -L. Chen, J. Yang and C. S. Ku, "Advancing Bankruptcy Forecasting with Hybrid Machine Learning Techniques: Insights from an Unbalanced Polish Dataset," in IEEE Access, vol. 12, pp. 9369-9381, 2024, doi: 10.1109/ACCESS.2024.3354173.

[2] A. Ansari, I. S. Ahmad, A. A. Bakar and M. R. Yaakub, "A Hybrid Metaheuristic Method in Training Artificial Neural Network for Bankruptcy Prediction," in IEEE Access, vol. 8, pp.176640-176650,2020, doi: 10.1109/ACCESS.2020.3026529.

[3] Y. U. Ryu and W. T. Yue, "Firm bankruptcy prediction: experimental comparison of isotonic separation and other classification approaches," in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 35, no. 5, pp. 727-737, Sept. 2005, doi: 10.1109/TSMCA.2005.843393.

[4] G. Lombardo, A. Bertogalli, S. Consoli and D. Reforgiato Recupero,"Natural Language

Processing and Deep Learning for Bankruptcy Prediction: An End-to-End Architecture," in IEEE Access, vol. 12, pp. 151075-151091, 2024, doi: 10.1109/ACCESS.2024.3476621.

[5] M. S. Park, H. Son, C. Hyun and H. J. Hwang, "Explainability of Machine Learning Models for Bankruptcy Prediction," in *IEEE Access*, vol. 9, pp. 124887-124899, 2021, doi: 10.1109/ACCESS.2021.3110270.

[6] T. M. Alam et al., "Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World," in The Computer Journal, vol. 64, no. 11, pp. 1731-1746, Nov. 2019, doi: 10.1093/comjnl/bxaa056.

[7] Brygała, M., & Korol, T. (2024). Personal bankruptcy prediction using machine learning techniques. *Economics and Business Review*, *10*(2), 118-142. https://doi.org/10.18559/ebr.2024.2.1149

[8] V. Kothuru, R. K. Jha, S. Ranjit, B. V. Kumar Reddy, S. Roy and S. Sudheer, "Prediction of Bankruptcy of a company using machine learning techniques," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp.797800, doi:10.1109/ICESC54411.2022.9885619